以英語寫作輔助爲目的之語料庫語句檢索方法

劉吉軒、洪培鈞
國立政治大學資訊科學系
jsliu@cs.nccu.edu.tw

李金瑛
國立台灣師範大學英語系

摘要

鑑於現有語言資源工具對於 ESL/EFL 學習者於英語寫作上提供的協助功能有限，本研究致力於提出一種語言資訊檢索方法，能在不同的語言認知程度條件下，從語料庫中找出對使用者之寫作表達需求有參照作用之例句。此方法提供準確字詞、單字的開頭/結尾、英語詞性、一個單字的萬用字元、不限定的子句等表達元素，並使用完整比對或部分比對的兩種比對方式選取例句，而後對於選取的例句使用多重序列排列技術進行相關性評估，最後推薦最符合參照需求的例句。

Abstract

Current language resource tools provide only limited help for ESL/EFL writing. This research proposes a language information retrieval approach to acquire referential sentences from corpus under various levels of users' language cognition. The approach includes a set of expression elements, such as exact words, prefix, suffix, part-of-speech, wildcard, and subsequence. Sentence retrieval involves both exact match and partial match with user query. Finally, the set of retrieved sentences are evaluated and ranked by multiple sequence alignment for relevance to user expression needs.

關鍵詞：資訊檢索，語料庫，寫作輔助

Keywords: Information Retrieval, Corpus, Writing Assistance.

一、緒論

對於大部分的寫作目的而言，寫作是一種深度的表達過程，需要詳盡而深入的描述能力，精準而嚴謹的語意呈現，及適當的詞語選擇安排。對於英語爲非母語的學習者(English as Second Language - ESL/English as Foreign Language - EFL)來說，英語寫作尤其是一個困難的過程，常常會因爲單字、搭配字(collocation)、詞性組合、常用片語、句型結構等方面的語言知識不足，或是受到母語認知與習慣的牽制影響，而造成用詞、語法、甚至語意上的錯誤[1]。同時，英語寫作也是一個費時的過程， ESL/EFL 的作者常常要花費許多時間，在自身不足的語言知識中搜尋適當的表達方式，或是要借助於字典、辭典等各種語言資源工具，取得初階的詞語資訊。但是，大量時間的投入往往僅能獲得有限的寫作品質的提升，而造成許多 ESL/EFL 作者的挫折與障礙。總結而言，若把 ESL/EFL 作者的英語寫作視爲一個資訊處理的過程，這項工作的困難癥結在於資訊量的不足、資訊取得成本的過高、及資訊使用的效益有限。

近年來，語料庫的發展帶來語言研究上的新面向。語料庫是大量語言使用情形的紀錄，不論是以文字或是以口語方式的語言表達內容，都可以被大量蒐集彙整、描述標註，而

形成了自然而真實的語言使用現象呈現[2]。因此，語料庫代表著豐富的語言資源，不僅可以對語言學的研究，提供許多統計分析上的資訊與理解詮釋上的依據，也可以做爲語言學習上各種語言使用方式的參照資料[3][4]。同樣的，以英語寫作爲目的的觀點而言，語料庫的存在提供了資訊處理工作上相當大的利用空間，若能發展一個有效的語言資訊使用工具，在 ESL/EFL 的英語寫作過程中，針對作者的語言表達需求，提供足夠的、適當的、參照的語言使用資訊，必能有效輔助 ESL/EFL 的英語寫作過程，降低寫作障礙，提升寫作品質。

目前使用語料庫的語言資訊使用工具以 concordance 技術爲主，提供特定字詞的檢索，將特定字詞在語料庫中出現的上下文，以特定字詞爲基準排列呈現，讓使用者觀察特定字詞可能的使用方式及相關字詞的搭配情形。這種工具一般叫做 concordancer，不僅可以做爲語言學者的研究輔助，也可以做爲英語學習者的諮詢參考[5][6][7]。但是 concordancer 並非針對非母語人士的英語寫作需求，對 ESL/EFL 的寫作困難所提供的協助仍然不足[8]，其主要限制包括：(1)較有限的查詢機制，只提供詞彙與詞性的條件檢索，無法充分對應 ESL/EFL 作者在語言知識上感到不足或不確定的情形下，所需要的查詢方式；(2)較無彈性的檢索機制，只提供完全符合條件的檢索，在使用者認知錯誤而使用不適當的查詢條件時，將無法提供任何查詢結果；(3)檢索結果的呈現以上下文爲主，而非以完整的例句呈現，同時，在結果的提供上，並未考量使用者在寫作上所需資訊的優先次序。

以資訊處理技術的觀點而言，concordancer 是一種相當初階的資訊檢索功能，只能針對有限形式的查詢條件，進行簡單的比對檢索，對檢索結果也沒有評估與排序的概念。因此，在語料庫的使用上，concordancer 只能針對明確的檢索目標提供少量的有用資訊。這樣的功能特性是無法滿足不同程度的 ESL/EFL 作者，在寫作上所需的多樣化資訊指引與參照協助。寫作是一個語言知識使用與輸出的過程，作者必須將其描述的意圖，明確的用正確的及適當的語彙組合排列，產生具體的文字內容與結構。而 ESL/EFL 作者常因語言知識的不足，在文字內容的生產過程中，常常必須針對疑惑進行探索或尋求答案，甚至在可能的認知錯誤下，也希望得到適當的引導而獲取正確的語言使用資訊。因此，一個以寫作協助爲目標的語料庫使用工具，必須提供更有彈性的查詢與檢索機制，呈現更能針對作者語言資訊需求的檢索結果，才能有效利用語料庫，協助作者提升寫作品質。

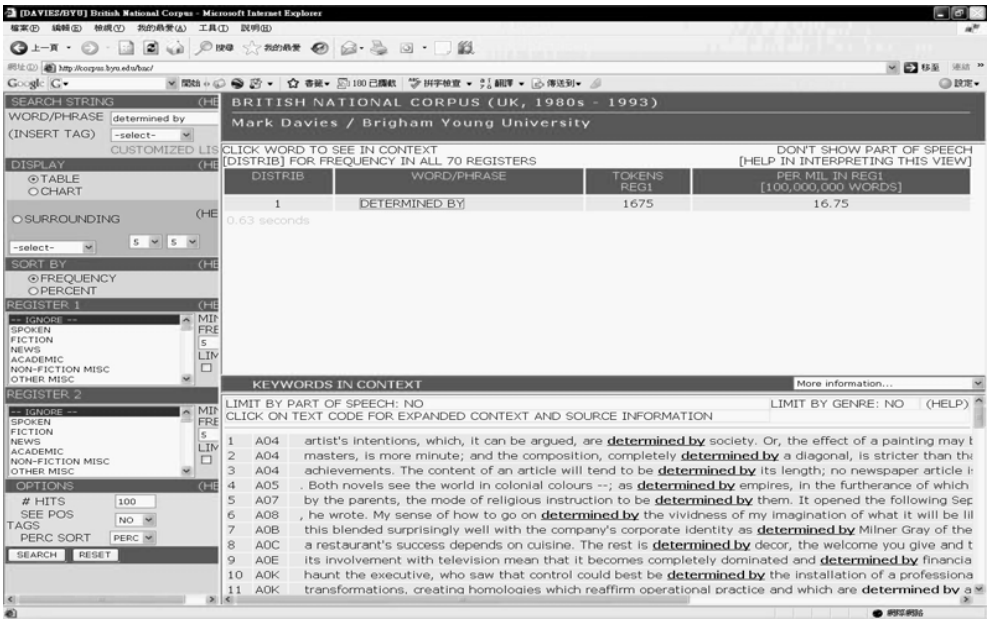
本研究提出一種語言資訊檢索方法，針對 ESL/EFL 作者在寫作過程上的語言知識不足或不確定，從語料庫中尋找出可用的語言使用範例，進而提供針對性的範例協助。我們的語言資訊檢索方法包含三個模組，第一個模組是一個多元的表達元素模型，能針對寫作過程中的語言資訊需求，包括單字、搭配字、常用片語、句型結構等，提供作者以更多樣的部份資訊的方式來表示其資訊需求。第二個模組是一個彈性的檢索機制，具備精確比對與部分比對兩種功能，能依照作者的資訊掌握程度而調整，從語料庫中比對尋找符合作者語意表達需求的範例。第三個模組是一個評估排序的機制，針對找出的範例，評估其符合作者語意表達需求的程度，並將之排序呈現，以提升範例參照的使用效率。我們以上述方法爲基礎，發展了一個以寫作協助爲目的的語料庫使用工具 - SAW (Sentence Assistance for Writing)系統，並以客觀的指標量測和主觀的問卷調查兩種評估方式來評量 SAW 系統的成效。這兩種評估方式都驗證 SAW 系統能針對 ESL/EFL 作者在其寫作過程中的語言使用資訊需求，給予一定程度的滿足而達到諮詢指引的效果，證明本研究的語言資訊檢索方法確實能達成寫作協助的既定目標。

二、英語寫作及語言資源工具

英語學習與使用的「聽、說、讀、寫」四個面向中，「寫」往往是較爲困難的部分，除了需要較爲廣泛、深入、及精準的語言知識使用之外，也常會受到母語認知語習慣的牽制影響，而無法使用正確的英語詞彙及結構。例如對於「創作音樂」的表達方式，母語爲中文的學生多以 make、create、produce 和 music 搭配，然而正確的英語詞彙應該是 compose music。研究指出 ESL/EFL 作者所需的前提知識有三個，分別爲英語單字、英語搭配和英語文法句型[8]。英語單字指作者對於英語字彙的認知能力；英語搭配指作者對於特定詞彙搭配上的認知能力；英語文法句型指作者對於英語文法句型本身組成結構的認知能力。

相關研究亦指出[9][10]，在英語的使用上，詞語搭配(collocation)是 ESL/EFL 作者較爲薄弱的一環，需要較多的語言資訊協助。詞語搭配是指共同出現的情形高於一般機率的特定詞彙組合。例如，在英語的使用習慣中，單字 problem 常和 cause、create、solve 一同搭配，而較少用 make，原因爲詞語搭配爲一種特定的結合方式，近似於語言使用上的約定與習慣。Ilson 等學者將複雜的英語詞語搭配分爲文法搭配(grammatical collocation)與詞彙搭配(lexical collocation)兩類[11]。文法搭配是指包含一個主要單字(dominant word)及一個介係詞、冠詞或連接詞的片語，例如， determined by。而詞彙搭配是指名詞、動詞、形容詞、副詞等特定字詞之間的慣用組合，例如，compose music。詞彙搭配沒有主要單字，也是 ESL/EFL 作者感到較爲困難及容易發生錯誤的部分。

語料庫使用工具泛指結合語料庫資源對於語言使用上提供資訊的工具。語料庫資源，是學習語言上最有成效的參考資源，語料庫使用工具能針對指定的詞語，提供許多的使用方式與情境範例，讓使用者有較佳的參照與學習成效。以 Word Sketch 爲例[12]，該系統藉由語料庫的使用，提供英語詞語使用情形描繪(lexical profiling)與搭配詞(collocation)顯著性的資訊。另一個語料庫使用工具的範例爲美國 Brigham Young University 研發的 VIEW[13]系統，使用者可輸入完整的或部分的字詞及詞性標籤(POS)等，查詢的結果則以 concordance 的方式呈現。圖一顯示以 determined by 的詞語查詢的結果。



圖一：VIEW concordancer 的語言資訊提供範例

由相關研究可知，語言資源工具對於 ESL/EFL 作者的寫作協助是相當重要的，然而不論是字典/辭典或是目前的語料庫使用工具對於寫作需求上的滿足仍然是相當有限的。不同的 ESL/EFL 作者因其語言知識程度的不同及認知的不同，往往對於相同的表達意圖會使用不同的查詢條件，而目前的語言資源工具仍然缺乏足夠彈性的表達方式對應於各種 ESL/EFL 作者可能需要的查詢條件。舉例而言，當 ESL/EFL 作者想要諮詢有關 by and large 的使用方式，有的作者能正確的以完整的片語組合查詢。但是更多的作者卻對此片語的認知不足或不確定，甚至有所誤解，而嘗試使用 large 或 by large 等查詢條件。這些不正確的或是模糊的查詢條件，在目前的語言資源工具中並無法得到直接的及有效的協助。這也正是 ESL/EFL 寫作需求與目前的語言資源工具之間矛盾與吊詭之處，對英語的語言知識愈不足的 ESL/EFL 作者愈需要語言資源工具的協助，而使用語言資源工具的成效卻以語言知識掌握的程度爲門檻，造成中等程度以下的 ESL/EFL 作者並無法在寫作的過程中，善用語言資源工具以提升寫作品質，甚至形成了學習進步上的瓶頸。

三、句子檢索與推薦

爲了解決在語言知識不足的情形下使用現有語言資源工具的成效不彰的問題，我們提出了一個語言資訊檢索方法，能接納不正確的或是模糊的查詢條件，盡可能的找出可以參照的使用範例，讓使用者探索或確認其真正需要的表達方式。同時，在檢索結果的提供上，我們也對選取出來的使用範例，進行與使用者語言資訊需求的相關性評估，再根據其個別相關程度建立排序，以推薦的諮詢參照先後次序，嘗試讓使用者能以最有效率的方式確認其適合的表達方式。此方法讓使用者以彈性的查詢條件表達其不同程度的語言認知，搜尋檢索的對象是語料庫，查詢結果是以語料庫中的個別句子爲單元排序呈現，提供使用者在一個完整的例句中觀察、學習及確認特定的表達方式。

我們將語言資訊檢索方法定義如下：給定使用者的語言資訊需求，句子檢索與推薦機制的目的是回傳一個經過排序的多個例句所組成的例句集合。此方法可分爲三個模組：表達元素模組、檢索模組、及排序模組。表達元素模組是讓使用者能將其語言資訊需求及完整或部分認知，轉換成多個表達元素(expression elements)，並藉由彈性的組合搭配，提供寬廣的需求與認知的對應。檢索模組從表達元素的組合中轉換成相對應的查詢條件，從語料庫中比對選取可能符合使用者語言資訊需求的例句。而排序模組則針對檢索模組所選取的例句，評估其個別符合使用者語言資訊需求的程度並建立排序，再回傳此排序過的例句集合給使用者。

(一) 表達元素模組

表達元素模組必須彈性的允許不同語言程度的使用者表示其完整或部分認知，我們的方法目前涵蓋下列表達元素(expression elements)：

1. 準確字詞(exact words)：準確字詞讓使用者提供拼字無誤的英文單字或單字的組合表達其語言資訊需求的認知。
2. 單字的開頭/結尾(prefix/suffix)：單字的開頭/結尾讓使用者以單字的部份資訊來表達其不完整的認知。在 ESL/EFL 寫作過程中，使用者常因單字的完整拼字方法遺忘而無法有效率的進行諮詢。我們以"%"符號來代表單字開頭或結尾的表達方式。
3. 一個單字的萬用字(wildcard)：一個單字的萬用字讓使用者表示不確定或不特定的一個單字，當成欲諮詢的表達方式的一部分。我們以"#"符號來代表一個單字的萬用字元的表達方式。

4. 英語詞性(POS)：當語料庫含有英語詞性標籤時，使用者可以運用英語的詞性標籤當成諮詢條件。英語的詞性標籤種類繁多，本研究將語料庫的詞性標籤大致分爲六類，分別爲介係詞(PREP)、形容詞(ADJ)、名詞(N)、副詞(ADV)、動詞(V)以及其他詞性(Other)。
5. 不限定英語子句(subsequence)：不限定英語子句代表零個至多個不限定的英文單字序列，其考量爲英語句型中常有特定片語結構結合不限定子句的情形，如 either … or 或 rather … than。我們以"*"來代表不限定單字序列的英語子句。

上述這些表達元素部分是以 regular expression 之概念爲基礎，但提供了更多元的語言表達需求之空間，可以讓使用者依照其語言認知情形與表達需求，選用適當的表達元素，代表部份確認而部分設定範圍的查詢條件。例如，使用者要表達中文之「濃茶」之意思，但不確定英語中表示「濃」之形容詞爲何，就可以用 ADJ tea 來表達其認知與需求。另外，這些表達元素也可以被彈性的組合，形成一個表達元素序列，例如，a pro% P 或 would rather V than V。這代表使用者在較多的語言認知下，所提供的較充分的表達意圖，而構成更強的查詢條件限制。

(二)檢索模組

檢索模組的目的爲從語料庫中選取符合使用者表達需求的例句。本研究同時採用完整比對和部分比對的選取法則，完整比對是要求語料庫的例句選取和使用者限定的查詢條件必需完整的吻合。部分比對則容許料庫的例句選取和使用者限定的查詢條件有部分的差異。我們預期 ESL/EFL 作者經常會有語意認知不完整或不正確的情形，若只依照使用者的限定條件進行完整比對，將導致於諮詢成效取決於使用者認知程度的矛盾。部分比對的選取方式提供了較大的彈性空間，容許使用者在不完整或不正確的認知下，也能獲得有用的參照資訊。

舉例而言，一個程度不高的 ESL/EFL 作者對於 not only … but also 的片語結構可能只約略有 only 及 also 的認知。使用者可以提出 only also 的需求表達方式，經由部分比對的選取法則，也可以檢索出使用 not only … but also 的例句。而在使用者的認知是錯誤的情形下，部分比對也可以選取出可能相關的例句，提供使用者判斷其認知是否正確的機會。例如，使用者以 an university 爲查詢條件，其中"an"是錯誤的用詞，正確的用詞應爲"a"，若以完整比對的方式，是無法搜尋出任何例句的，而部分比對則能找出含有 an 開頭而 university 在後的例句。使用者可以從例句的檢索結果評斷其原先認知的 an university，可能含有詞語用法搭配上的錯誤，以致於選取出的結果並無 an university 緊密相連的例句，而間接得到諮詢的協助。

(三)排序模組

排序模組的目的爲針對選取出來的例句，評估其符合使用者的語意表達需求程度，並將之排序，以提升使用者參照例句的成效。本研究採用多重序列排列(Multiple Sequence Alignment – MSA)的技術[14]，進行查詢條件與選取例句之間的相關程度評估與排序。給予多個長度不同的序列(sequence) S_1、S_2、… S_n，$n \geq 3$，一個多重序列排列爲相同長度的序列 A_1、A_2、… A_n，其中 A_1 對應 S_1、A_2 對應 S_2，依此類推到 n。而 A_1、A_2、… A_n 序列中允許元素之間出現「間隔」，以"–"表示。

假設有兩個序列 S_1= CCAATA、S_2= CCAT，序列元素爲 Σ={A,C,T}，則排列的結果可能爲 S_1 = – – – – C C A A T A，S_2 = C C A T – – – – – –，或 S_1 = C C A A T A，S_2 = C C A –

T –。兩種結果均爲 S_1、S_2 的 MSA，兩種排列方式之間的優劣或合宜程度，可經由計算所有位於同一相對位置元素的比對分數總和來評估。通常我們以一個置換矩陣(substitution matrix)來代表上述所有元素間的比對分數。而兩個序列之間的最優排列結果(optimal MSA)，可以透過 Needleman-Wunsch algorithm[14]求得。然而排序模組的工作需要對兩個以上的選取例句進行排列與評估，在運算時間的考量下，我們採用 the center star algorithm，以一個參考序列當基準用以比對其他的序列，尋求較佳解而不是最佳解。對於長度均爲 k 的 n 個序列，the center star algorithm 的效能爲 $O(k^2n^2)$[15]。本研究以學習者輸入的 query 爲參考序列，可省略尋找中心序列的步驟。

本研究將 MSA 的參數矩陣依照表達元素的模組分爲 11 個比對單位，分別爲英語詞性標籤 6 個：P、J、N、D、V、O(介係詞、形容詞、名詞、副詞、動詞、其他詞性)；以及準確字詞(exact)、單字開頭/結尾(prefix/suffix)、萬用單字(wildcard)和用以對應於不限定字串的間隔(gap)。比對參數上的取捨以準確字詞最高(100)、單字開頭片尾其次(50)、詞性標籤次之(25)、萬用單字最低(5)，其考量爲查詢條件的精準程度區分。若以 = 代表準確字詞，% 代表單字開頭/結尾，# 代表萬用單字，– 代表不限定字串的間隔，X 代表無關之單字。當學習者輸入的 query 爲 S_0：a pro% P，S_0 的 MSA 代表元素爲"= % P"，選取的例句爲 S_1、S_2、S_3：

S_1. This(X) posed(X) **a(=)** particular(X) **problem(%) for(P)** an(X) agent(X).
S_2. Listening(X) to(X) all(X) these(X) personal(X) accounts(X) has(X) had(X) **a(=) profound(%)** effect(X) **on(P)** us(X).
S_3. Increasingly(X) acid(X) rain(X) is(X) **a(=) problem(%) in(P)** Europe(X) too(X).

其中 S_1、S_2、S_3 粗體的單字依序爲選取的"a"(exact)、"pro%"(prefix)和"P"(介係詞)，括弧內的符號爲 S_1、S_2、S_3 的 MSA 代表元素。以代表學習者的語意表達需求之查詢條件(S_0)爲中心點，依據 the center star algorithm 排序，可得等長的序列 A_0、A_1、A_2、A_3，並以 sum-of-pair score 來計算排序後的分數，假設 A_1 相對於 A_0 的分數爲 C_1、A_2 相對於 A_0 的分數爲 C_2、A_3 相對於 A_0 的分數爲 C_3，其中 S(x,y) 代表 x 元素比對 y 元素的分數，我們可得以下的結果：

A_0	–	–	–	–	–	–	–	–	=	%	P	–	–	–
A_1	–	–	–	–	–	–	X	X	=	X	%	P	X	X
A_2	X	X	X	X	X	X	X	X	=	%	X	P	X	–
A_3	–	–	–	–	X	X	X	X	=	%	P	X	X	–

C_1=S(–,–) + … + S(=,=) + S(%,X) + S(P, %) + S(–,P) + ... + S(–,X) = 93
C_2=S(–,X) + … + S(=,=) + S(%,%) + S(P, X) + S(–,P) + ... + S(–,–) =139
C_3=S(–,–) + … + S(=,=) + S(%,%) + S(P, P) + S(–,X) + ... + S(–,–) = 169

根據 sum-of-pair score 的分數，SAW 系統會推薦給學習者的例句依序爲 S_3、S_2、S_1：

S_3. Increasingly acid rain is **a problem in** Europe too.

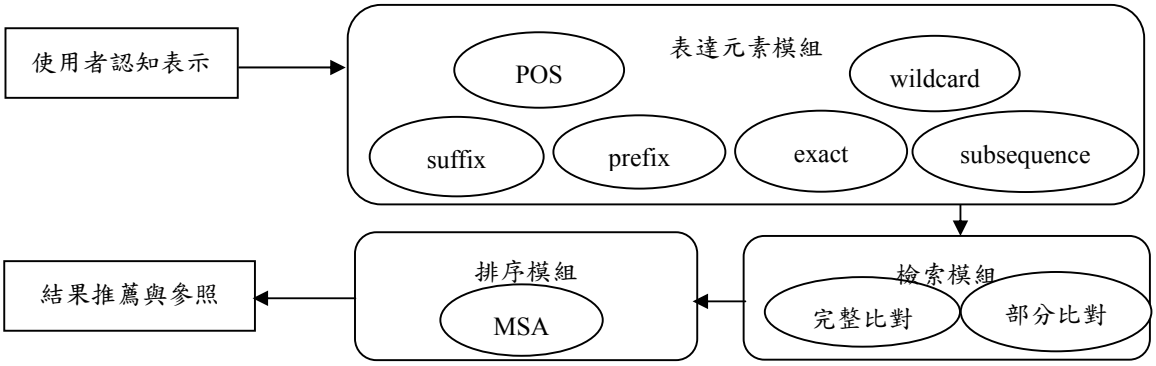
S_2. Listening to all these personal accounts has had **a profound** effect **on** us.

S_1. This posed **a** particular **problem for** an agent.

(四)實作系統與處理流程範例

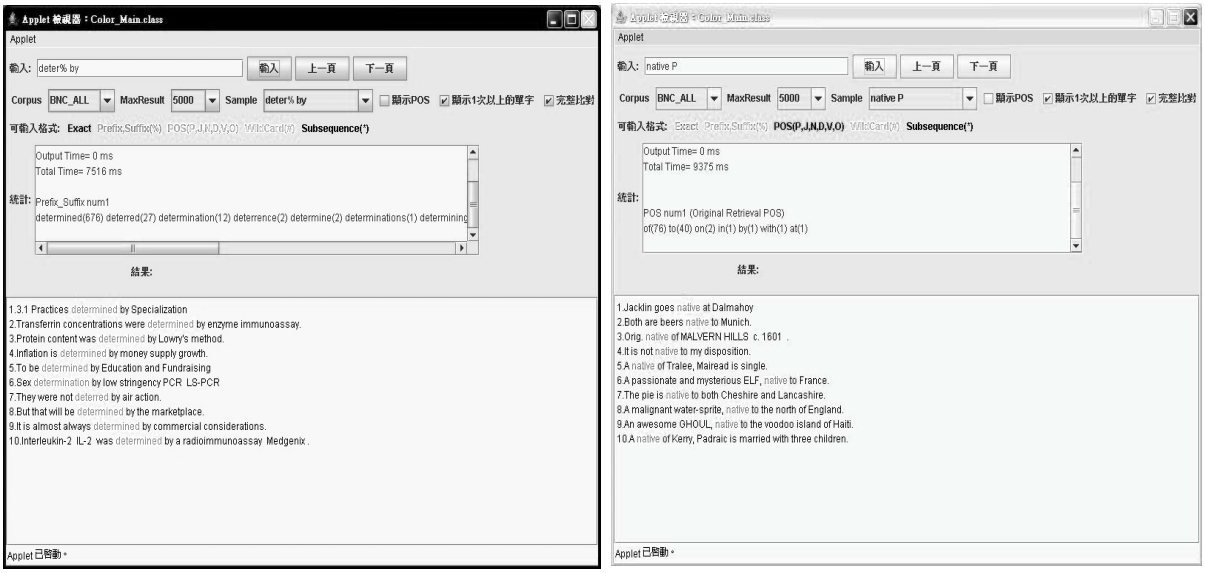
我們以上述的語言資訊檢索方法實際開發完成一個稱爲 SAW(Sentence Assistance for Writing)的雛型系統，其表達元素模組的部分包含了六種認知表示的基本方式；檢索模組則可分別進行完整比對與部分比對，以選取符合查詢條件的例句；最後，排序模組則

採用 MSA 演算法進行檢索出來的例句與查詢條件之間的排列方式並評估其相關程度，再依照相關性依序推薦呈現給使用者。SAW 系統的整體架構如圖二所示。



圖二：SAW 雛型系統架構圖

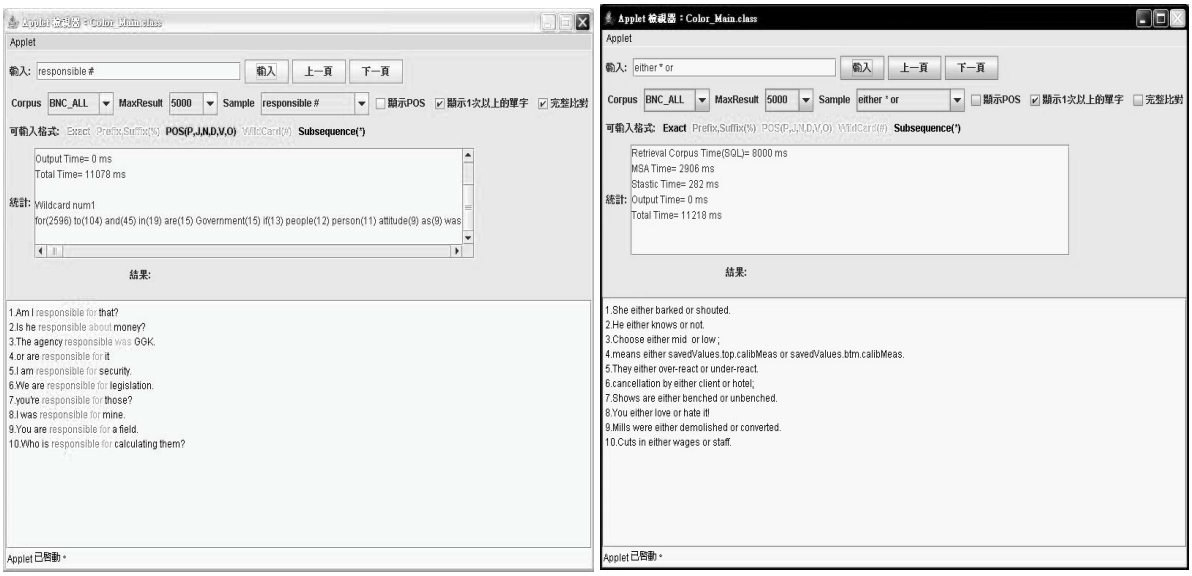
假設使用者想要搜尋 not only … but also 的例句，而其認知的表示方式爲 not only but also，四個緊密相連的英語單字，而語料庫中有三個相關例句，分別爲 S_1：We must also make sure that future generations not only read, but also have a real enthusiasm for visiting bookshops and libraries. S_2：This was not only humiliating but also very awkward for Baldwin. S_3：This is not only easier, but also more fun. 首先，使用者藉由準確字詞表達其所認知的表示方式；檢索模組中的完整比對因語料庫中並無 not only 和 but also 緊密相連的例句，故無法搜尋出任何例句；而部分比對因容許不同單字間可以有間隔，而可選取出 S_1、S_2、S_3 例句。而排序模組再根據整體例句的結構，例如以句子的長短爲考量因素，將 S_1、S_2、S_3 的例句排序爲 S_3、S_2、S_1 的順序以供使用者參照。



圖三：SAW 系統分別以 deter% by (左)及 native P(右)的查詢結果

最後，我們展示 SAW 系統的介面及運作結果，查詢的條件分別爲 deter% by、native P、responsible #、either * or，亦即以 prefix、POS、wildcard、subsequence 等表達元素所組成的查詢條件當測試樣本。其中 deter% by、native P、responsible #將採取完整比對的機制，either * or 則採取部分比對的機制，其結果畫面分別由圖三與圖四呈現。使用者可以選擇特定語料庫進行檢索，可以選擇完整比對或部分比對。檢索結果是精簡而完整的

例句，而非如圖一所示以上下文爲主的 concordance，系統提供每頁 10 個例句，使用者可視需要檢視下頁的例句。另外，SAW 系統也提供了檢索結果的統計資訊。例如，deter% by 的檢索結果例句中，使用 determined by 的數量有 676 個，使用 deterred by 的數量有 27 個，使用 determination by 的數量有 12 個，使用 deterrence by 的數量有 2 個，使用 determine by 的數量有 2 個，使用 determining by 的數量有 1 個，等等。這些統計資訊也將提供給使用者更進一步判斷或探索的指引。



圖四：SAW 系統分別以 responsible # (左)及 either * or (右)的查詢結果

四、實驗評估

本實驗採取 BNC(British National Corpus)語料庫，該語料庫收錄的文章類別包括自然科學、社會科學、應用科學、哲學、藝術、科幻、金融、休閒、世界九大類，年代範圍爲 1974-1994，共有約 350 萬個例句。語料庫中有內定的 POS 細目分類，共計 62 個。本實驗將 BNC 的 62 個細目 POS 分類對應爲 6 個 POS 的大類。

(一)評估方法

爲了客觀評估 SAW 系統對於 ESL/EFL 作者的參照成效，我們進行了兩種評估方法，分別爲模擬測試(實驗客觀評估)和問卷調查(使用者主觀評估)。在模擬測試方面，我們從英語能力測驗的試題中，選取重要的英語表達方式(如單字、片語、句型結構等)，模擬 ESL/EFL 作者於寫作上可能遭遇的困難，再由實驗者提供數種不同的查詢方式，測試 SAW 系統的檢索結果，並觀察評估其效能。試題來源包括大學聯招和大學學測的考題，共計 16 期，及 TOFEL 的考題，共計 11 期。總共選取 45 個單字片語、12 個句型結構 (含有 subsequence 表達元素的句型)當試題樣本。本實驗從二個角度分析成效，第一個角度爲單字片語的評估：針對 18 個單字片語，以 SAW 系統不同的表達元素(exact、prefix、POS、wildcard)，模擬不同認知程度的使用者，以評估對檢索結果所能得到的參照諮詢成效。第二個角度爲句型結構和認知錯誤的評估：針對 12 個句型結構，模擬使用者在認知有部分錯誤的情況下，SAW 系統能給予使用者參照諮詢上的成效。

問卷調查的主要目的爲以使用者的觀點評估 SAW 系統所提供的參照諮詢成效。我們選

取部分英語能力測驗試題，包括單選題 16 題及翻譯題 3 題，由受測者模擬英語能力測驗，再以分組測驗的結果，比較受測者於使用 SAW 系統協助和未使用 SAW 系統協助之間測驗成績上的差異。另一部份的比較爲同一組的受測者，在未使用 SAW 系統前和使用 SAW 系統後的測驗成績差異。

爲客觀並且確實的評估 SAW 系統的成效，本研究參考推薦系統的評估方式[16]，擬定 5 項觀察指標，詳述如下：

1. 試題測驗分數：本研究藉由試題模擬學習者於寫作上的困難，而試題具備了錯誤選項及正確的答案，因此，可藉由學習者的測驗分數來評估系統的參照諮詢成效。
2. 推薦度：對於系統提供的語言使用資訊，推薦度是使用者所認爲真正實用的程度。本研究對於推薦度採取 4 種高低程度的量度，即高推薦度、中推薦度、低推薦度、無推薦度。推薦度是由使用者主觀評估的，本研究對於推薦度的衡量將以使用者的回應方式取得。
3. 滿意度：對於系統提供的語言使用資訊，滿意度是使用者對其整體滿意的程度。本研究對於滿意度採取 5 個高低程度的量度，即非常滿意、滿意、普通、不滿意、非常不滿意。滿意度亦是由使用者主觀評估的，本研究對於滿意度的衡量將以使用者的回應方式取得。
4. 符合度：符合度是依據使用者提供的查詢條件，系統比對選取出的結果和使用者預期答案上的相似度。本研究對於符合度的評估方式爲在一定選取數量的例句中，真正符合學習者預期答案的比值(number of relevant retrieved / number of retrieved)，也就是一般資訊檢索中的準確度(precision)之定義。
5. 例句成本：例句成本爲使用者閱讀完一定數量的單字序列所需花費的成本，實驗上以所需閱讀的英語單字字數爲成本量。例句成本的評估爲反應簡潔有力的例句，對於參照諮詢上的效果可能較好。

(二)實驗結果

1. 模擬測試之單字片語評估

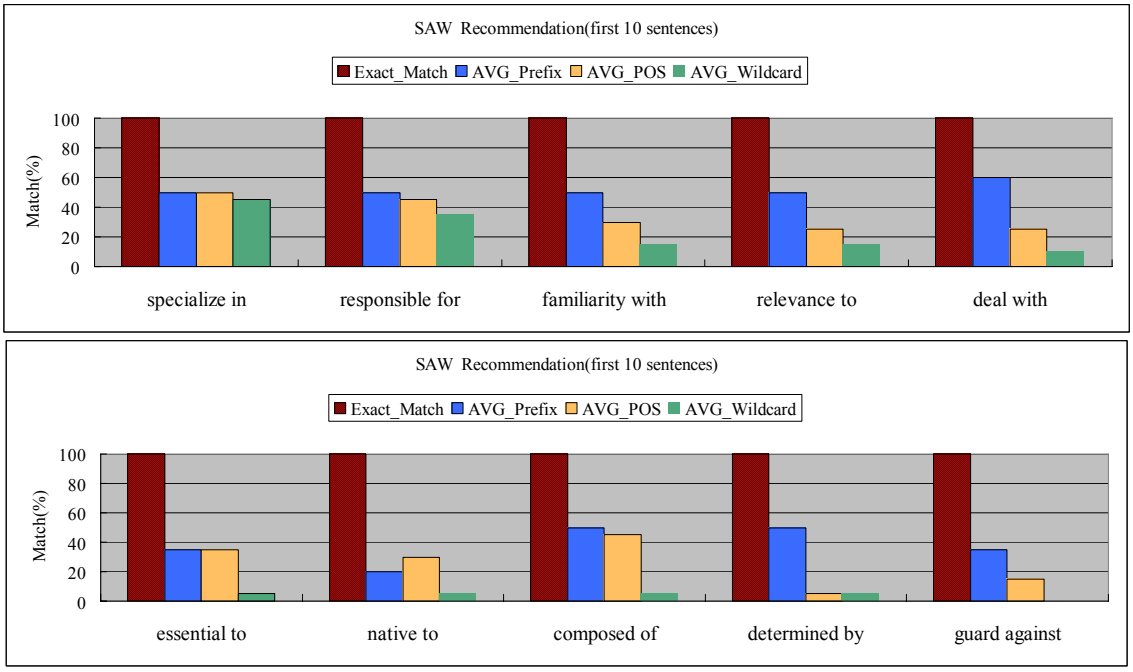
單字片語的評估主要藉由試題模擬學習者於寫作上所需的協助，並假設不同符合度的查詢條件，評估對認知程度不同的學習者所能提供的協助成效。實驗以 10 個由兩個英語單字組成的片語及 8 個由多個英語單字組成的片語當測試樣本，並採用 exact、prefix、POS、wildcard 等表達元素形成不同的查詢條件，模擬學習者可能的不同認知及所得到的不同參照結果，再以結果的符合度進行評量。

本實驗採用 4 大類型的查詢方式：(1)exact：完整提供兩個英語單字，如 specialize in。(2)prefix：將兩個英語單字其中一個的第一個英語字母當成 prefix 的引導原則，如以 specialize in 爲例，將分爲 s% in 和 specialize i% 兩種可能的查詢條件，再取此兩種查詢條件所得檢索結果的平均符合度。(3)POS：將兩個英語單字其中一個以 POS 表達元素查詢，如以 specialize in 爲例，將分爲 V in 和 specialize P 兩種可能的查詢條件，同樣的，取兩種不同查詢條件所得檢索結果的平均符合度。(4)wildcard：將兩個英語單字其中一個以 wildcard 表達元素查詢，如以 specialize in，將分爲 # in 和 specialize # 的兩種可能的查詢條件，再取兩種不同查詢條件所得檢索結果的平均符合度。

實驗設定 SAW 系統每次提供 10 個例句，亦即須推薦出前 10 個例句供使用者參照。圖

五顯示 10 個由兩個英語單字組成的片語的資訊需求，對於 SAW 系統所提供的參照資訊之符合度評估。一如預期的，exact 查詢條件的結果符合度最高，爲 100%。另外，prefix 查詢條件於此 10 個片語中符合度高於 POS 條件；而 POS 查詢條件符合度高於 wildcard 查詢條件，其中有一個例外爲 native to 片語。這些結果約略說明 prefix 表達元素爲限制較強的查詢條件，而 POS 次之，wildcard 最低。

除了兩個英語單字組成的片語之外，本研究亦測試了三個和四個英語單字組成的片語，評估上述不同查詢條件的檢索結果之符合度。和上述兩個英語單字片語查詢方式唯一不同的是，三個英語單字的 prefix 查詢方式可能的查詢條件有三個，例如以 a proportion of 爲例，可能有 a% proportion of、a p% of、a proportion o% 三種不同的查詢條件，此時，實驗是取此三種查詢條件所得檢索結果的平均符合度。同樣的，以 a proportion of 爲例，POS 查詢方式可能有 O proportion of、a N of、a proportion P 三種不同的查詢條件；而 wildcard 查詢方式可能有 # proportion of、a # of、a proportion # 三種不同的查詢條件，此時，實驗是取此三種查詢條件所得檢索結果的平均符合度。另外，關於四個英語單字的 prefix、POS、wildcard 查詢方式可能的查詢條件有四個，此時，實驗是取此四種查詢條件所得檢索結果的平均符合度。



圖五：單字片語的符合度評估

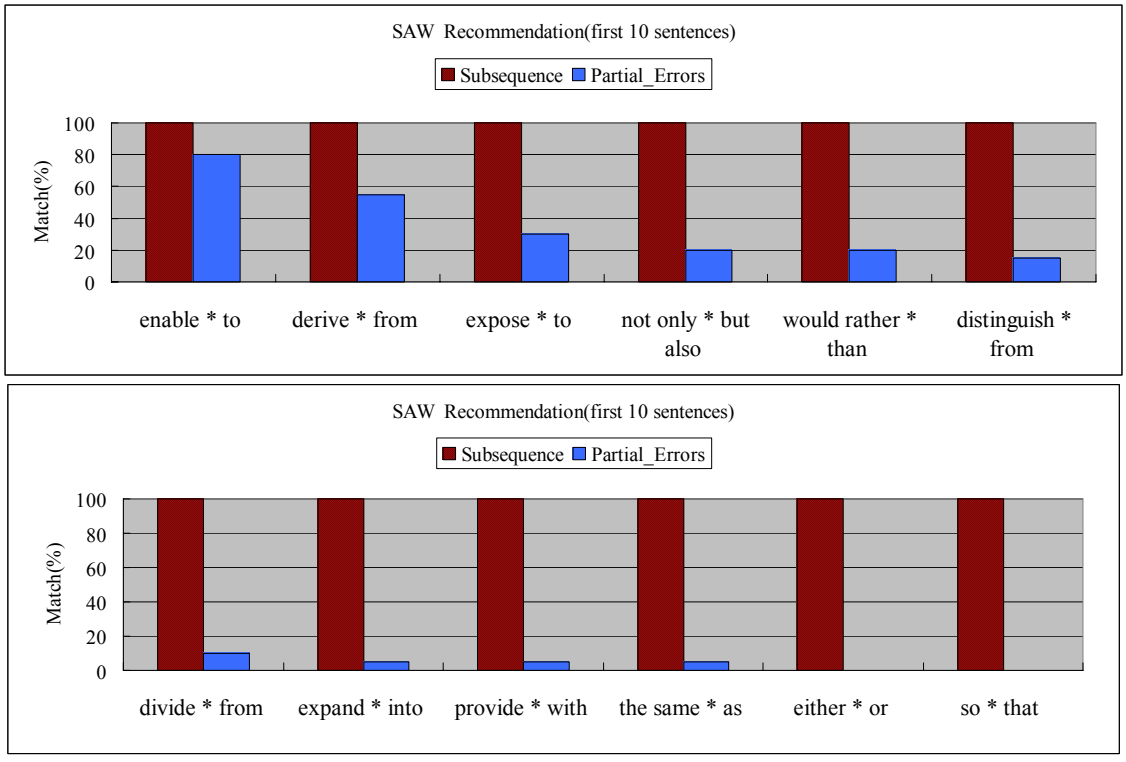
實驗結果顯示，查詢條件中包含的字數越多，SAW 系統提供的參照資訊的符合度越高，原因在於多個英語單字組成的片語，其本身結構性較強，在其中一個字爲模糊的情況下，學習者還是能得到較針對其表達需求的參照協助。而結果亦顯示，prefix 查詢方式的成效高於 POS 查詢方式，POS 查詢方式成效高於 wildcard 查詢方式，顯示 prefix 表達元素爲限制條件較強的查詢條件，而 wildcard 表達元素的限制條件最爲寬鬆。

2. 句型結構和認知錯誤的評估

本實驗選取 12 個句型結構，包括：enable * to；derive * from；expose * to；not only * but also；would rather * than；distinguish * from；divide * from；expand * into；provide * with；

the same * as；either * or；so * that ，以評估 subsequence 表達元素的查詢方式之參照成效，並刻意對於 12 個句型結構中提供部分錯誤的查詢條件。舉例而言，對於 enable … to 的片語，刻意以 enable * and 和 enable * but 的部分錯誤查詢條件模擬學習者認知上的錯誤，並以平均符合度來描述系統的參照協助成效。此 12 個子句中，均有一個重要的關鍵介係詞，依序爲 to, from, to, but, than, from, from, with, into, as, or, so。本實驗將部分錯誤的認知均以 and 和 but 來取代上述的關鍵字，唯一例外的是 not only * but also 中的 but，因 but 是正確用法，實驗將 but 取代爲 or 當部分錯誤認知的查詢條件。

實驗結果如圖六顯示，在句型結構查詢條件下 SAW 系統均能對學習者提供針對性的參照協助(符合度 100%)。在學習者部分錯誤的認知方面，查詢結果會因句型結構關鍵字彼此間的強度而有不同的成效，例如 enable * to 因本身關鍵字較強，故即使將 to 輸入成 and 或 but，亦能藉由系統的部分比對得到一定的參照協助。相反的，諸如 either * or 和 so * that 子句中的關鍵字彼此的關連強度不強，縱使經過部分比對的選取方式，提供給學習者的參照符合度亦有限。但是，由問卷調查的結果顯示，學習者亦能於低符合度的情況下，觀察推薦例句結果而評估本身是否有錯誤(false query)或者不夠嚴謹(vague query)的查詢條件發生，可以說間接讓學習者得到一定的參照協助。



圖六：subsequence 和 partial errors 查詢條件結果之符合度評估

3. 問卷調查評估

試題回應有 19 道試題(16 題單選 3 題翻譯)， SAW 系統分別提供與個別試題相對應的 BNC 語料庫中的相關例句，做爲受測者作答時的參照。藉由試題測驗分數上的差異，以及受測者提供的參照推薦度與滿意度，評估 SAW 系統給予受測者寫作參照協助的程度。實驗將受測者分爲四組：第一組填答 19 道試題時，並無任何參照例句；第二組填答同樣的 19 道試題，而且有高符合度的參照例句；第三組填答同樣的 19 道試題，但提

供的爲低符合度的參照例句。第四組施測時先在無參照例句的情形下填答同樣的 19 道試題，收回試題後，再於提供高符合度參照例句的情形下，填答同樣的 19 道試題。問卷施測的對象爲 125 位的五專生。而依據英語字彙程度測試問卷[17]，可瞭解受測者對於英語字彙的認知度，亦約略代表其英文程度。受測的五專生對於英語常用 1000 單字約有 80.44%的認知度；對於英語常用 2000 單字有 46.10%的認知度。而根據 Nation 的研究[18]，英語常用 1000 單字出現於英語文章的頻率爲 74%，英語常用 2000 單字頻率爲 81%。換而言之，問卷調查測試對象平均不超過 2000 單字的字彙程度，亦即平均每五個單字中受測者約有一個單字不熟識。

問卷回收的份數爲 125 份，其中無效的問卷有 12 份，包括繳交白卷和猜題作答(如選項只有 1 和 2，受測者選 3 或 4；抑或受測者對於選項全作答 1、2、3 或 4)。總計有效問卷爲 113 份，其中有無使用系統上的區分問卷共 86 份，細分爲無使用系統參照的 44 份、使用高符合度參照例句的 20 份、使用低符合度參照例句的 22 份問卷；而使用系統前和使用系統後的問卷共 27 份。

在有無使用系統上的區別有效問卷 86 份中，單選試題有 44 份問卷爲受測者沒有使用系統的協助而填答試題、20 份爲使用系統較高符合度的參照例句填答試題(平均符合度爲 62.48%)、22 份爲使用系統較低符合度的參照例句填答試題(平均符合度爲 34.69%)。其中單選試題共有 16 題，每題爲 0.625 分，滿分爲 10 分。而由結果發現有使用系統參照的分數均高於未使用系統協助的分數，其中無使用系統參照的分數爲 4.42 分，標準差爲 1.95；而使用高符合度的參照例句其分數爲 6.69 分，標準差爲 2.23，比無使用系統參照的分數高 2.27 分；而使用低符合度的參照例句其分數爲 5.97 分，標準差爲 1.85，比無使用系統參照的分數高 1.55 分，證實 SAW 系統能給予學習者一定程度的寫作上的參照協助。此部分的相關問卷分析結果彙整於表一中。

表一：有無使用系統之問卷結果分析

	無使用系統	符合度高的參照	符合度低的參照
單選試題分數(0~10)	4.42	6.69	5.97
翻譯試題分數(0~10)	4.2	5.83	
單選試題推薦度(1~4)	---	2.78	2.70
翻譯試題推薦度(1~4)	---	2.80	
句型長度(1~3)	---	2.09	
滿意度(1~5)	---	2.76	

爲了驗證上述結果並非偶然發生，我們以統計學的顯著水準(significance level)說明 SAW 系統確實有一定的協助成效顯著性。假設母體的受測分數爲常態分佈，並以無使用系統參照的試題分數當母體的受測分數。假設 Null Hypothesis (H_0) 爲 SAW 系統沒有成效(平均分數爲 4.42)，Alternative Hypothesis (H_1)爲 SAW 系統有成效(平均分數大於 4.42)，在顯著水準 5%下，對於高符合度參照例句的 20 份樣本，可得拒絕 H_0 的臨界分數爲 $4.42+(1.65*1.95/(20)^{1/2})$ =5.14，因實驗所得的平均分數爲 6.69，大於臨界分數 5.14，所以可以否定 H_0，亦即在 5%顯著水準下 SAW 系統對於受測者是有協助成效的。依據同樣的統計分析方法，對於低符合度參照例句的 22 個樣本，其臨界分數爲 $4.42+(1.65*1.95/(22)^{1/2})=5.10$，而使用 SAW 系統得到的分數爲 5.97，因 5.97 大於臨界分數 5.10，證明在 5%顯著水準下 SAW 系統對於受測者仍然是有協助成效的。

除單選試題外，問卷有三題翻譯試題，翻譯試題能更確實地模擬學習者於寫作上所遇到的困難。在上述 86 份問卷中，有 44 份問卷爲受測者沒有使用系統的協助而填答試題，42 份爲系統給予同樣的參照例句而填答試題。翻譯試題的評分標準爲句型結構 50%、用詞 25%、其他 25%(包含語意、文法等)，三題試題以滿分 10 分爲標準。結果顯示，無使用系統參照的平均分數爲 4.20，標準差爲 2.56；而有使用系統參照的平均分數爲 5.83，標準差爲 2.46，分數高於無使用系統的問卷 1.63 分。同樣依據統計的顯著性分析，假設以無使用系統參照的對象爲母體，而使用系統參照的對象爲樣本，在顯著水準 5% 下臨界分數爲 4.97，因 5.83 大於 4.97，即在 5%顯著水準下 SAW 系統對於受測者翻譯試題是有協助成效的。

推薦度是由使用者評估系統是否有實用性(usefulness)，問卷以分數 1~4 分爲標準，分數越高代表受測者越覺得有實用性。而結果顯示在使用系統參照的 44 份問卷中，其中單選試題使用高符合度參照例句的問卷平均爲 2.78、使用低符合度參照例句的問卷平均爲 2.70，而翻譯試題平均爲 2.80。整體而言，有使用系統的平均推薦度介於 2.7~2.8 之間，表示受測者認爲 SAW 系統是有一定的協助實用性。然此數值不算太高的原因是受限於受測者平均字彙程度不到 2000 字，BNC 語料庫的例句的程度較深，造成受測者發生不甚瞭解例句的狀況，以致於受測者認定 SAW 系統的參照結果推薦度不高，而上述的假設亦經由受測者的問卷回饋證實此情況是發生的。

表二：使用系統前後之問卷結果分析

	使用系統前	使用系統後
單選試題分數(0~10)	4.99	6.53
翻譯試題分數(0~10)	4.56	5.42
單選試題推薦度(1~4)	---	2.59
翻譯試題推薦度(1~4)	---	2.53
句型長度(1~3)	---	2.16
滿意度(1~5)	---	3.16

在使用系統上前後的區別有效問卷 27 份中，單選試題共有 16 題，每題爲 0.625 分，滿分爲 10 分。相關問卷分析結果彙整於表二中。使用系統前受測者的平均分數爲 4.99，標準差爲 2.49；使用系統後的平均分數爲 6.53，標準差爲 2.00，分數增加 1.54 分。同樣的，若以使用系統前的結果爲母體，使用系統後的結果爲樣本，在 5%顯著水準下的臨界分數爲 5.71。因 6.53 大於 5.71，證實 SAW 系統對於單選試題是有協助成效的。

在使用系統上前後的區別中，翻譯試題同樣以句型結構 50%、用詞 25%、其他 25%(包含語意、文法等)爲評分標準，滿分爲 10 分。而使用系統前的平均分數爲 4.56，標準差爲 2.59；使用系統後的平均分數爲 5.42，標準差爲 2.60，分數增加爲 0.86 分。同樣的，以使用系統前的結果當母體，而使用系統後的結果當樣本，在 5%顯著水準下的臨界分數爲 5.38。因 5.42 大於 5.38，證實 SAW 系統對於翻譯試題是有協助成效的。

(三)結果討論

在問卷調查中，研究證實無論從有無使用系統的角度或者使用系統前後的角度，SAW 系統如從試題分數的觀點切入，能給予學習者分數上的協助；而從推薦度的觀點切入，學習者亦認爲能從 SAW 系統中得到一定的參照協助，證實 SAW 系統對於寫作有其一

定的協助成效。然因本問卷施測的對象普遍字彙程度不超過 2000 字，而 BNC 的例句對於受測者普遍偏難，容易造成受測者的推薦度和滿意度下滑，而此部分可藉由採取較適宜的語料庫讓結果更完善。

我們另外也以學生實際的寫作錯誤進行測試，如 I would rather going shopping than staying home. 此學生所犯的錯誤爲對於 would rather … than 的句型不太熟悉，其中 would rather 後若接動詞，應以原形表達，亦即正確的句型應爲 I would rather go shopping than stay home. 我們使用 SAW 系統以 would rather V than V 和 would rather * than * 模擬學習者可能的查詢條件，而由結果發現，使用 would rather V than V 的查詢條件可以於前五句推薦例句中都可以得知正確的動詞形式，符合度爲 5/5；而使用 would rather * than *的查詢條件可以於前十句推薦例句中的七個例句得知正確的動詞形式，其符合度爲 7/10，證實以上述兩種查詢方法，SAW 系統均能給予學習者相關的參照協助。

五、相關研究

近年來，運用資訊技術協助英語學習的研究引起極多的興趣，我們僅討論與本研究較爲直接相關者。首先是 CMU 的 REAP 系統[19]，其主要目的爲提供學習者英語閱讀的協助，研究方法包括依照學習者認知程度，以測驗的方式評估學習者的詞彙程度，在使用 REAP 系統前後，是否有顯著性的差異，並以問卷調查和統計數值整合分析 REAP 系統的成效。本研究參考了 REAP 系統的評估方式評估 SAW 系統的成效。另外，Gsearch 系統主要目的爲在語料庫中建立句法準則(syntactic criteria)，而學習者可以依據建立的句法準則有效率地從語料庫中選取合宜的例句[20]。Gsearch 系統提供的句法準則較爲彈性，對於不同的學習者能依據多樣的句法準則有效率地查詢。本研究延續其理念，但特別關注於認知不足情形下之表達方式查詢，而這也是 ESL/EFL 作者最爲關鍵的困難。最後，TANGO 系統[21]提供中英雙語的詞語搭配關係抽取與查詢，尤其著重在動詞與名詞的相互搭配資訊。本研究提供較多元的需求表達與查詢方式，並對檢索結果進行評估與排序，以提升使用者參照的效率。同時，本研究以更嚴謹的方式評量系統的成效。

六、結論

本研究鑑於現有語言資源工具對於 ESL/EFL 學習者於英語寫作上提供的協助功能有限，致力於提供一種語言資訊檢索方法，能有效率地根據不同學習者的認知程度予以一定的參照諮詢協助，而此方法分爲表達元素模組、檢索模組、排序模組三個部份。爲了證實此方法的成效，我們實作完成 SAW 系統。SAW 系統可使用準確字詞(exact)、單字的開頭/結尾(prefix/suffix)、英語詞性(POS)、一個單字的萬用字元(wildcard)、不限定的子句(subsequence)等做爲表達元素(expression elements)，並可使用完整比對或部分比對的兩種比對方式，於檢索模組中選取例句，而後對於選取的例句使用 MSA 技術爲排序模組的評估方法，最後提供推薦的相關例句供學習者參照。

爲了評估 SAW 系統的成效，實驗分爲模擬測試(實驗客觀評估)和問卷調查(使用者主觀評估)兩種方式，藉由試題分數、推薦度、滿意度、符合度、句型成本五項指標，驗證在上述兩種評估方法下，SAW 系統均能得到一定程度的協助成效。除此之外，我們並依據學生的作文範例，將範例模擬輸入於 SAW 系統中，發現 SAW 系統亦能給予學習者一定的參照協助，更加驗證 SAW 系統的協助成效。綜觀以上幾點，本研究的語言資訊檢索方法確實能針對 ESL/EFL 學習者於英語寫作時給予適度的協助。

參考文獻

[1] A. Aghbar, Fixed Expressions in Written Texts: Implications for Assessing Sophistication, East Lancing, MI: National Center for Research on Teacher Learning, ERIC Document Reproduction Service No. ED 352808, 1990.

[2] T. McEnery and A. Wilson, *Corpus Linguistics*, Edinburgh University Press, Edinburgh, 1996.

[3] S. Conrad, The Importance of Corpus-based Research for Language Teachers, *System* 27: 1-18, 1999.

[4] A. B. M. Tsui, ESL Teachers' Questions and Corpus Evidence, *International Journal of Corpus Linguistics* 10(3): 335-356, 2005.

[5] I. de O'Sullivan and A. Chambers, Learners' Writing Skills in French: Corpus Consultation and Learner Evaluation, *Journal of Second Language Writing* 15(1):49-68, 2006.

[6] Jean-Jacques Weber, A Concordance and Genre-informed Approach to ESP Essay Writing, *ELT Journal* 55(1): 14-20, 2001.

[7] Y. C. Sun, Learning Process, Strategies and Web-based Concordancers: a Case Study, *British Journal of Educational Technology* 34(5): 601-613, 2003.

[8] H. Yoon, An Investigation of Students' Experiences with Corpus Technology in Second Language Academic Writing, Ph.D. Dissertation, The Ohio State University, USA, 2005.

[9] C. C. Shei and H. Pain, An ESL Writer's Collocational Aid, *Computer Assisted Language Learning* 13(2): 167–182, 2000.

[10] B. Altenberg and S. Granger, The Grammatical and Lexical Pattern of MAKE in Native and Non-native Student Writing, *Applied Linguistics* 22(2):173-195, 2001.

[11] R. Ilson, B. Morton, and B. Evelyn, *The BBI Dictionary of English Word Combinations,* Amsterdam: John Benjamin Publishing Company, 1997.

[12] A. Kilgarriff and D. Tugwell, WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proceedings of the Collocations Workshop, ACL 2001*, pp. 32-38.

[13] M. Davies, The Advantage of Using Relational Databases for Large Corpora, *International Journal of Corpus Linguistics* 10(3): 307-334, 2005.

[14] S. B. Needleman and C. D. Wunsh, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology* 48: 443-453, 1970.

[15] Y. L. Francis, N. L. Ho, T. W. Lam, W. H. Wong, and M. Y. Chan, Efficient Constrained Multiple Sequence Alignment with Performance Guarantee, *IEEE Computational Systems Bioinformatics* 2: 337-346, 2003.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, Analysis of Recommendation Algorithms for E-Commerce, *Proceedings of ACM Conference on Electronic Commerce*, Minneapolis, Minnesota, U.S.A., 2000.

[17] English Word Tests, [Online]. Available: http://www.lextutor.ca/ [Accessed: May. 15, 2007].

[18] Paul Nation, Measuring Readiness for Simplified Material: A Test of the First 1,000 Words of English, *In Simplification: Theory and Application RELC* 31:193-203, 1993.

[19] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, Classroom Success of an Intelligent Tutoring System for Lexical Practice and Reading Comprehension, *Proceedings of the Ninth International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006.

[20] S. Corley, M. Corley, F. Keller, M. Crocker, and S. Trewin, Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System, *Computers and the Humanities* 35(2): 81-94, 2001.

[21] T. C. Chuang, J. Y. Jian, Y. C. Chang, and J. S. Chang, Collocational Translation Memory Extraction based on Statistical and Linguistic Information, *Computational Linguistics and Chinese Language Processing*, Vol. 10, No. 3, September 2005, pp. 329-346.

貝氏主題混合資訊檢索模型

Bayesian Topic Mixture Model for Information Retrieval

吳孟淞 許軒睿 簡仁宗
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
mswu@chien.csie.ncku.edu.tw

摘要

在自動文件處理之相關研究中，我們常利用機率主題模型從字詞相互關係推斷並建立潛在主題變數。在機率潛在語意模型(PSLA)裡，文件中的每一個字詞在混合模型即視爲一個樣本，其混合成分是使用多項分佈來表示的。然而，多項分佈方式沒有考慮到文集中發生的突發現象。雖然 PLSA 模型可以顯示多重主題樣式，但是每個主題模型都十分簡單。在本研究中，我們提出一種新型之貝氏主題混合模型來解決多項分布固有的一些問題。使用 Dirichlet 分佈表示每一個主題的條件機率分佈，在相同種類內的不同的文件經由不同的多項分布來產生。在 TREC 文件集之資訊檢索實驗上，利用文件檢索及文件模組化之評估來驗證貝氏主題模型的優越性。

Abstract

In studies of automatic text processing, it is popular to apply the probabilistic topic model to infer word correlation through latent topic variables. Probabilistic latent semantic analysis (PLSA) is corresponding to such model that each word in a document is seen as a sample from a mixture model where mixture components are modeled by multinomial distribution. Although PLSA model deals with the issue of multiple topics, each topic model is quite simple and the word burstiness phenomenon is not taken into account. In this study, we present a new Bayesian topic mixture model (BTMM) to overcome the burstiness problem inherent in multinomial distribution. Accordingly, we use the Dirichlet distribution for representation of topic information beyond document level. Conceptually, the documents in the same class are generated by the associated multinomial distribution. In the experiments on TREC text corpus, we show the results of average precision and model perplexity to demonstrate the superiority of using proposed BTMM method.

關鍵詞：貝氏機率模型，圖形模型，機率潛在語意模型，Dirichlet 事前機率，資訊檢索

Keywords: Bayesian model, Graphical model, PLSA, Dirichlet Prior, Information Retrieval

一、緒論

隨著資訊大量氾濫，各種數位文件（digital documents）的遽增，使得資訊檢索精確度和文件模型的建立日顯重要。在資訊檢索和機器學習研究上，統計型本文模型(statistical text model)已逐漸成爲一個重要的議題。就資訊檢索的研究者而言，大多數將

文件視爲是 bag-of-word 的表示法，嘗試用統計的方法，擷取文字的特徵以建構資訊檢索的模式，此類方法亦稱爲向量空間模型[32]。Bag-of-word 的缺點是不考慮人類語言的同義字詞（synonym）以及多義字詞（polysemy）。再者，此方法的空間維度表示相當於字典個數的大小。這意謂有許多的參數必須被估計，容易導致效能的降低。在文獻上，已有一些文件表示法被提出解決 bag-of-word 方面的一些問題。首先，潛在語意分析(Latent Semantic Analysis, LSA)[10]，是將文件以“字詞－文件”矩陣表示的方法。透過奇異值分解(Singular Value Decomposition, SVD)將文件投射到一個低維度的語意空間，並假設每一奇異值及其對應的奇異向量(singular vector)代表其潛在主題或概念，且每一文件可由右奇異矩陣轉置的行向量表示。在資訊檢索和語音辨識上已證明是有價值的分析工具[2][3][24]。第二，機率模型(Probabilistic Model)的基本假定爲觀測資料下的一個生成模型，此模型反應資料本身的架構。目前，已有一些機率模型的技術被廣泛地使用。例如，機率潛在語意分析(Probabilistic Latent Semantic Analysis)[16][17]以及 Latent Dirichlet Allocation[6]。PLSA 模型作法是擷取與文件關聯的意向模型(Aspect model)[18]。PLSA 模型有幾項缺點[6]，首先，是沒有直接的方法將機率分配給先前未出現(unseen)的文件。其次，參數數量會隨著文件數量線性擴增。LDA[6]爲一個較完整的生成模型，其方法是將每一篇文件的機率視爲潛在主題中隨機字詞機率的混合模型，進而求得該篇文件出現的機率值。然而，其近似推論演算法並不容易實現。再者，文件以多項分佈表示法，無法有效取得字詞在文件中的突發現象(burstiness phenomenon)[12][25]。所謂「突發現象」意指，字詞在文件中出現過一次之後，很有可能會再出現的情形[22]。一般而言，字詞在文集裡一般分爲三種範疇，即常見(common)、一般(average)和稀有(rare)。雖然多項式表示能獲得常見字詞的突發性，但是對於一般和稀有字詞的突發性並未被正確的模組化。而透過 Dirichlet 分佈來替代多項分佈，可以趨緩突發現象的問題[25]。在本研究中，對於機率和主題混合模型問題感興趣，將探討幾個較先進的圖形模型[6][16][23][25]，期望藉由相關背景，來改善現有的文件模型架構。本文中以 PLSA 機率模型爲基礎，在混合模式的結合上，透過貝氏方法使用 Dirichlet 分佈決定各個分配所佔的比例，稱之爲貝氏主題混合模型(Bayesian Topic Mixture Model, BTMM)。透過 Gibbs 抽樣法來估計所需的參數。Gibbs 抽樣法的優勢是不需要明確地表示模型參數，可以在字詞分配到代表的潛在主題方面，簡單定義模型。本研究利用貝氏主題混合模型進行資訊檢索相關研究，所獲得成果對於改善搜尋系統檢索較易具有相當的應用價值。此外也可提供相關領域如資料探勘、機器學習等領域進行深入探討。本文接下來章節組織如下。第二章探討目前文獻中各種相關的文件模型研究方法。第三章將說明本文所提出的方法，並比較幾種主要模型的差異。第四章爲本文提出的方法和其他作法比較實驗效能分析的結果，用以證明本研究方法的效益及結果討論。最後，第五章爲本文的結論以及未來的研究方向。

二、相關文獻探討

在許多的應用上，資訊檢索和機器學習可以說密不可分。本章，我們將探索一些較具體、熟知的機率統計模型。首先，簡單描述在資訊檢索中較常見的文件表示法[10][32]。接著，針對廣泛的生成模型做更深入的探討，其中包含一些機率模型和混合模型等圖形模型表示式[6][16][17][23] [31]。

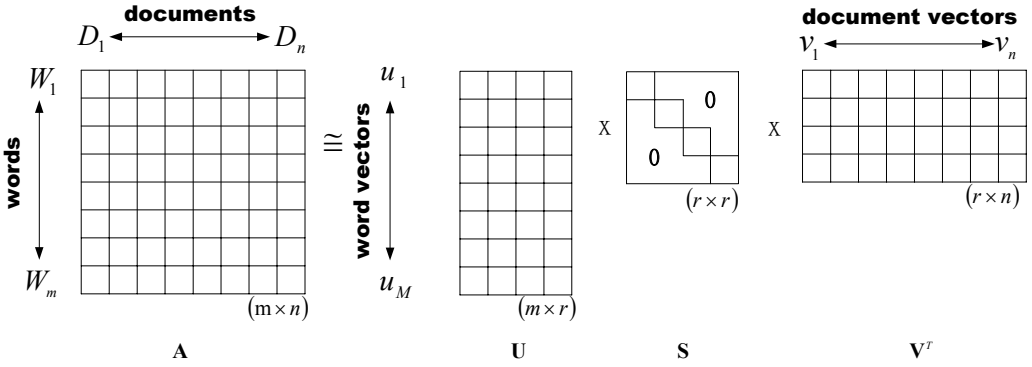
(一)、文件表示法

在資訊檢索系統中，文件通常由向量表示，意指視爲特徵的字詞出現在每篇文件的現象，此種表示法稱爲 bag-of-word 或者向量空間模型(Vector Space Model)[32]。其中

w_i 表示字典中的字詞在文件中出現的頻率值，而字典通常由文件集中的訓練集合所擷取得到。整個文件集可以透由字詞文件矩陣來表示，如下所示

$$\mathbf{A} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \tag{1}$$

其中 w_{ij} 表示字典中的第 i 字詞在第 j 篇文件中出現的頻率值。在上述表示法中，缺乏任何有關字詞之間的語意訊息。因此，有其他學者考慮此類相關訊息來描述文件，稱爲潛在語意分析（Latent Semantic Analysis, LSA）[10]。LSA 基本的概念是以低維度的共同語意因子呈現原先文件和字詞之間的關聯。利用奇異值分解（Singular Value Decomposition, SVD）找出字詞對應文件的語意結構，可將高維度的矩陣資料降低爲 r 維度大小之特性。其奇異值分解之架構示意圖，如圖一所示。



圖一、奇異值分解之架構示意圖

(二)、文件混合模型之探討

1、Mixture of Unigrams

Mixture of Unigram (MU)模型是將 Unigram 模型經由離散隨機主題變數而擴增[31]。在此混合模型下，每份文件經由所選擇的主題所產生，接著，從主題相關的多項式獨立產生字詞。其文件的機率表示如下

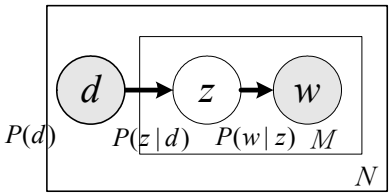
$$\begin{aligned} P(w) &= \sum_z P(w \mid z) P(z) \\ P(d) &= \prod_w P(w) = \prod_w \sum_z P(w \mid z) P(z) = \sum_z P(z) \prod_w P(w \mid z) \end{aligned} \tag{2}$$

當整個文集(corpus)被估計時，字詞分佈可以視爲在每一個文件對應一個主題的假設之下的主題表示。

2、Probabilistic Latent Semantic Analysis

先前所提 LSA 模型在文件和字詞上的呈現，並非以統計觀點出發。因此，Hofmann 提出機率潛在語意分析模型（Probabilistic Latent Semantic Analysis, PLSA）[16][17]，其

模型如圖二所示。PLSA 模型不同於 LSA 將文件和字詞向量投射至潛在語意空間的做法，其方法是以 Aspect Model 作爲主要架構[18]，使用機率密度函式作爲已觀察到的文件和字詞之間潛在語意關聯性的呈現方式，並利用最大相似度估測法則，結合了 EM 演算法[11]推估出隱含的模型參數。PLSA 模型目前被廣泛應用在多個領域，包括文件分段[7]、網頁探勘[19]、語音辨識技術及語言模型調適[1][8][9][29]等應用。



圖二、PLSA 模型示意圖

PLSA 模型主要的特徵，是針對字詞和文件共同事件尋求一個生成模型[16][17]。本文資料集是由字詞-文件對 (d,w) 所組成，文件以 $\mathbf{d} \in \{d_1,\ldots,d_N\}$ 表示，其個數爲 N；另外，字詞以 $\mathbf{w} \in \{w_1,\ldots,w_M\}$ 表示，字典相當於是 M 個字詞所形成之集合。假設每一字詞在給定的文件中潛在主題 $\mathbf{z} \in \{z_1,\ldots,z_K\}$ 下產生。將字詞-文件對 (d,w) 共同出現(co-occurrence)的聯合機率以式(3)表示

$$\begin{aligned} P(d,w) &= \sum_z P(z)P(w \mid z)P(d \mid z) \\ &= P(d)\sum_z P(w \mid z)P(z \mid d) \end{aligned} \tag{3}$$

在 PLSA 模型中，文件則經由 $P(w \mid z)$ 的因子的混合描繪其特性。將 z 視爲潛在變數，可以容易地對 PLSA 模型利用 EM 演算法來學習參數。最大化對數相似度可以表示成：

$$L_{\text{PLSA}} = \sum_d \sum_w n(d,w) \log P(d,w) = \sum_d \sum_w n(d,w) \sum_z P(z)P(d \mid z)P(w \mid z) \tag{4}$$

其 $n(d,w)$ 表示字詞在文件中的數量。在 E-step 中，利用目前估計的參數來計算潛在變數的事後機率，其式子如下

$$P_{\text{PLSA}}(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_z P(z)P(d \mid z)P(w \mid z)} \tag{5}$$

在 M-step 中，利用潛在變數在 E-step 中的估測，使得觀察的聯合對數相似度的期望最大化。其所有參數的更新如下

$$\hat{P}_{\text{PLSA}}(w \mid z) = \frac{\sum_d n(d,w)P(z \mid d,w)}{\sum_w \sum_d n(d,w)P(z \mid d,w)} \tag{6}$$

$$\hat{P}_{\text{PLSA}}(d \mid z) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_d \sum_w n(d,w)P(z \mid d,w)} \tag{7}$$

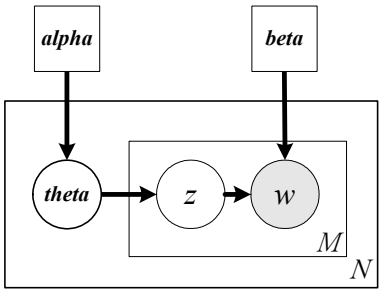
$$\hat{P}_{\text{PLSA}}(z) = \frac{\sum_d \sum_w n(d,w)P(z \mid d,w)}{\sum_d \sum_w n(d,w)} \tag{8}$$

PLSA 在資訊檢索中，可以藉由低維的”潛在”空間代替原始文件的表示。在 Hofmann[16][17]裡，以 $P(z \mid d)$ 作爲在低維空間之文件的組成，對於未看見(unseen)之文件或查詢句，經由最大化對數相似度和固定 $P(w \mid z)$ 及計算而得。

3、Latent Dirichlet Allocation

近幾年來，Latent Dirichlet Allocation (LDA)被提出來模組文集的潛在主題[6]。在大詞彙自動語音辨識系統下使用在語言模型的調整[30][33]，以及其他機器學習應用上皆有不錯的成效[4][5]。LDA 主要是克服 PLSA 模型中上述的缺點，比較 LDA 與 PLSA 模型相異之處，在於 LDA 將每一篇文件的機率都視爲潛在主題中隨機字詞機率的混合模型，藉此取得該篇文件出現的機率值。LDA 模型使用隨機變數 θ 來代替 PLSA 模型中 $P(z \mid d)$ 參數。θ 和 z 有相同的維度，表示文件中主題的混合。θ 對每一文件從 Dirichlet 分佈取樣，代替估計每一訓練文件的混合機率 $P(z \mid d)$ ，對 PLSA 模型而言，LDA 所需要的參數量較少。在 PLSA 模型中，有 $K*N$ 個 $P(z \mid d)$ 參數，而 LDA 模型，對文件的取樣，θ 只需 K 個參數。

在 LDA 模型裡，假設文件從潛在主題上隨機混合取樣，透過字詞上的分佈描繪每一主題的特性。在此模型中，文件爲觀察變數，視爲字詞的集合，$\mathbf{d} \in \{1,\ldots,M\}$ ，每一字詞取決於未觀察變數(也就是 topic) z，表示在 $\{1,\ldots,K\}$ 的可能值，並且 K 超參數 (hyperparameter) 必須被決定。在文件空間裡，LDA 模型存在未觀察變數，$\theta = (\theta_1,\ldots,\theta_K), \theta_k > 0$ 且 $\sum_k \theta_k = 1$。其模型如圖三所示，α 表示爲主題混合 θ 之 Dirichlet priori，而字詞機率透過 $K * M$ 矩陣 β 參數化，其中 $\beta = P(w \mid z)$ 。



圖三、LDA 模型示意圖

文件 d 和主題混合 θ 的聯合分佈爲

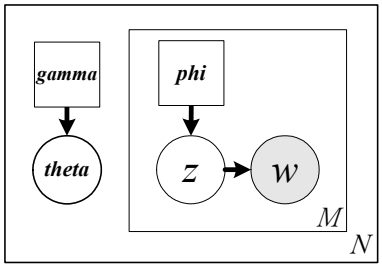
$$P_{\mathrm{LDA}}(\theta, \mathbf{d} \mid \alpha) = P(\theta \mid \alpha) \prod_w \left[\sum_z P(w \mid z) P(z \mid \theta) \right]^{n(d,w)} \tag{9}$$

其中，$n(d,w)$ 表示字詞 w 在文件 d 中出現的個數，$P(\theta \mid \alpha)$ 爲 θ 的 Dirichlet 機率分布。我們可以得到文件的邊際分佈

$$P_{\mathrm{LDA}}(\mathbf{d} \mid \alpha) = \int P(\theta \mid \alpha) \prod_w \left[\sum_z P(w \mid z) P(z \mid \theta) \right]^{n(d,w)} d\theta \tag{10}$$

然而，爲了估計這些參數必須計算事後分佈 $P(\theta, z \mid d)$ ，通常這些推論是不易實現的。在文獻上，一些基於變化方法的近似推論技術被提出，如 Variational Methods[6][21]、Expectation Propagation[28]和 Gibbs 抽樣法[14]。在此，對 Blei et al.[6]所提出的方法做說明，在圖形模型來說，Variational Method 是將一個複雜的圖形模型轉換爲一個簡單的

圖形模型，如圖四所示，期望簡化過後的模型能夠用正推論(exact inference)解決。



圖四、近似事後 LDA 模型之 Variational 分佈示意圖

Blei et al. [6] 定義一個分佈 $q(\theta, \mathbf{z} \mid \gamma, \phi)$ 的近似群，並且選擇 Variational Parameters γ 和 ϕ 接近真實的數值。Variational 分佈定義爲

$$q(\theta, z \mid \gamma, \phi) = q(\theta, \gamma) \prod_{z} q(z \mid \phi_z) \tag{11}$$

對於這新模型，可以經由 Variational Distribution 和 True Posterior 之間的 KL Divergence 最大化得到 $P(\theta, z \mid \mathbf{d}, \alpha, \beta)$ 的近似，

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D(q(\theta, z \mid \gamma, \phi) \,||\, P(\theta, z \mid \mathbf{d}, \alpha, \beta)) \tag{12}$$

參數估測過程利用 variational EM，使得對數相似度最低界限(lower bound)最大化，基於近似事後分佈 $P(\theta, z \mid d)$ 的一種變化分佈來更新參數，透過下列兩個步驟迭代過程。在 E-step 中，使用變化的事後分佈近似，對每份文件找到多變參數 $\{\gamma, \phi\}$ 的最佳化值，

$$\phi_n \propto \beta \exp\{E[\log(\theta) \mid \gamma]\} \tag{13}$$

$$\gamma = \alpha + \sum_n \phi_n \tag{14}$$

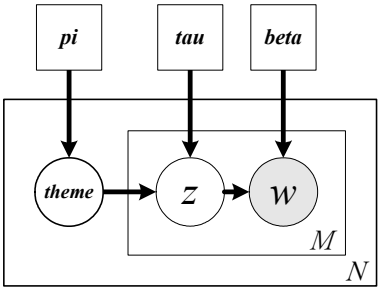
在 M-step 裡，使得有關模型參數對數相似度最小界限最大化，對條件多項參數的更新可以表示如下

$$\beta \propto \sum_d \sum_n \phi_{dn} w_{dn} \tag{15}$$

而參數 α 可以透過 Newton-Raphson 演算法求得[27]。Girolamin 和 Kaban [13]說明當 Dirichlet 分佈相同時，PLSA 模型實際上是 LDA 的一個特例。

4、Theme Topic Mixture Model

如前所述，LDA模型近似推論演算法並無法得到正解(Exact Solution)且計算複雜度增加。爲了克服這個問題，Keller和Bengio[23]提出一個正推論且易處理的模型，稱之爲 Theme Topic Mixture Model (TTMM)。在TTMM裡，文件空間的變數稱爲Theme，不同於LDA，TTMM對於topic的混合程度所佔的比例利用離散有限集(discrete finite set)來代替連續空間的使用。如圖五所示，此模型的觀察變數爲文件d，可視爲字詞w的集合，而未觀察變數爲theme，以 $\mathbf{h} \in \{1, \ldots, J\}$ 表示，以及topic，以 $\mathbf{z} \in \{1, \ldots, K\}$ 表示。其參數 π, τ 和 β 個別表示theme的混合程度所佔的比例 $P(h = j)$ 、topic給定theme的混合程度 $P(z \mid h = j)$ 以及每一字詞給定每一主題的機率值 $P(w \mid z)$ 。



圖五、TTMM 示意圖

每個文件可以視爲 theme h 的混合，表示爲

$$P(\mathbf{d}) = \sum_j P(h=j)P(d \mid h=j) = \sum_j P(h=j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h=j)\right]^{n(d,w)} \tag{16}$$

其中，$P(d \mid h=j)$ 表示給定一個主題 $h=j$ ，其文件的生成機率，而 $n(d,w)$ 表示字詞在文件中的頻率，且 $\sum_w n(d,w) = n(d)$ 。假定文集 D 爲 N 篇文件的集合，給定文件模型，其文集 D 的對數相似度可以表示爲

$$L_{\text{TTMM}} = \sum_d \log\left[\sum_j P(h=j)\prod_w \left(\sum_z P(w \mid z)P(z \mid h=j)\right)^{n(d,w)}\right] \tag{17}$$

如同 PLSA 一樣，參數估計亦可經由 EM 演算法使得對數相似度最大化。在 E-step 中，潛在變數的事後機率被估計，如下所示

$$P(h=j \mid d) = \frac{P(h=j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h=j)\right]^{n(d,w)}}{\sum_j P(h=j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h=j)\right]^{n(d,w)}} \tag{18}$$

$$P(z \mid w, h=j) = \frac{P(z \mid h=j)P(w \mid z)}{\sum_{z'} P(z' \mid h=j)P(w \mid z')} \tag{19}$$

在 M-step 下，其對數相似度期望值是使用在上一階段估測的事後值，使得在標準化限制(normalization constraint)條件下最大化。模型參數的重新估測，可以表示爲

$$P_{\text{TTMM}}(h=j) = \frac{\sum_d P(h=j \mid d)}{\sum_{j'}\sum_d P(h=j' \mid d)} = \frac{\sum_d P(h=j \mid d)}{N} \tag{20}$$

給定條件限制 $\sum_{j'} P(h=j' \mid d) = 1$ ，可得

$$\hat{P}_{\text{TTMM}}(z \mid h=j) = \frac{\sum_d P(h=j \mid d)\sum_w n(d,w)P(z \mid w, h=j)}{\sum_d n(d)P(h=j \mid d)} \tag{21}$$

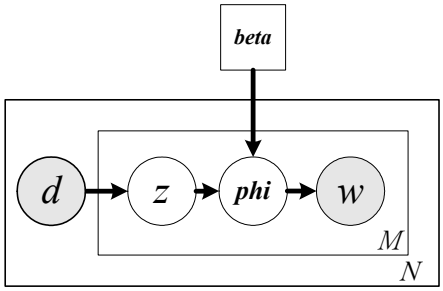
同理，給定條件限制 $\sum_{z'} P(z' \mid w, h = j) = 1$ ，可得

$$\hat{P}_{\text{TTMM}}(w \mid z) = \frac{\sum_{d}\sum_{j} n(d,w)P(h = j \mid d)P(z \mid w, h = j)}{\sum_{w'}\sum_{d}\sum_{j} n(d,w')P(h = j \mid d)P(z \mid w', h = j)} \tag{22}$$

三、貝氏主題混合模型

(一) 模型定義

Madsen et al. [25]提到多項模型比較適合常見的字詞，但是對於其他較一般或是稀有的字詞無法有效獲得其突發現象。且多項式產生的計數分佈(counts distribution)基本上不同於自然本文的計數分佈。對於本文的模組化，Dirichlet 分佈在先前研究已被廣泛使用[6][12][25]。本文所提出之貝氏主題混合模型，同時擁有 PLSA 以及 Dirichlet 分佈的主要概念。使用 Dirichlet 模型於每一個主題的條件分佈，且文件裡每一個字詞可以由不同的主題所產生，使得在文件模型的表示上更豐富。模型架構如圖六所示



圖六 BTMM 示意圖

在 BTMM 裡，假設文件集 D 包含文件數 N 篇，文件表示爲 $\mathbf{d} \in \{d_1, \ldots, d_N\}$ ，而字典 V 相當於是 M 個字詞所形成的集合，字詞以 $\mathbf{w} \in \{w_1, \ldots, w_M\}$ 表示。未觀察變數爲主題，以 $\mathbf{z} \in \{z_1, \ldots, z_K\}$ 表示。假設文件 d 和字詞 w 條件獨立於給定的未觀察主題變數 z，對於所產生的模型參數，字詞是經由主題的多項分佈 ϕ 所產生，而對於字詞分佈的具體主題多項分佈 ϕ ，可以從 Dirichlet priori 參數 β 對應的主題 z 得到。另外，文件是在 K 個潛在主題上使用 N 個混合數的多項分佈來表示，且 $\sum_{z} P(z \mid d) = 1$ 。在模型裡，參數集以集合 $\{\phi, \beta, P(z \mid d)\}$ 來表示，在推演過程中，使用 Dirichlet 分佈於主題多項分佈之上，因此隱藏參數 ϕ 可以被在外結合而不需要明確地被估計，此簡化過程，不需要在對 ϕ 取樣。如此一來，所需要的參數量共有 $KN + K$ 個。依據生成過程，字詞和主題的聯合分佈可以表示爲

$$P(w \mid z, \beta) = \int_{\phi} P(w \mid \phi)P(\phi \mid \beta, z)d\phi \tag{23}$$

而文件-字詞對 (d, w) 的聯合機率可以寫成

$$\begin{aligned} P(d, w \mid \beta) &= P(d)\sum_{z} P(z \mid d)P(w \mid z, \beta) \\ &= P(d)\sum_{z} P(z \mid d)\int_{\phi} P(w \mid \phi)P(\phi \mid \beta, z)d\phi \end{aligned} \tag{24}$$

(二) 推論

在我們的模型裡，潛在變數爲 $z_{d,w}$，即文件的字詞 w_d 所出現的主題。首先，觀察計算 $P(\mathbf{z} \mid \mathbf{w}_d)$ 的複雜度，此分佈和聯合分佈成正比。根據貝氏規則，對潛在變數 z 之條件事後分佈(conditional posterior distribution)給定爲

$$P(\mathbf{z} \mid \mathbf{w}_d) = \frac{P(\mathbf{w}_d, \mathbf{z})}{\sum_z P(\mathbf{w}_d, \mathbf{z})} \tag{25}$$

在此，我們以 $\{\mathbf{w}_d, \mathbf{z}\}$ 代表完整的觀察資料，$\mathbf{w}_d$ 表示文件向量。計算 $P(\mathbf{z} \mid \mathbf{w}_d)$ 意味著在大的離散狀態空間評估機率分佈。遺憾的是，這分佈無法直接被求得，主要是因爲分母部分加總比較難評估[14]，並且包含 Z^M 個離散隨機變數，其 M 表示在文件集中字詞的總數。就這觀點，在本研究中，嘗試以馬可夫蒙地卡羅法[20]以模擬參數的事後分佈(posterior distribution)來估計未知參數。如同文獻[14]一樣，使用 Gibbs sampling 來估計參數。對本研究提出的模型來說，Gibbs sampling 演算法容易實現，需要較少的儲存記憶體空間，並且在數度和執行上和現有的演算法具有競爭性。對於每一個字詞標記而言，Gibbs sampling 從對應的條件分佈中，給定其他字詞對於主題的分配，來估計目前字詞分配到主題的機率。然後目前字詞可以被分配到主題上，並且將此分配儲存下來，以便 Gibbs sampling 著手其他字詞計算時使用。爲了模擬 $P(\mathbf{z} \mid \mathbf{w}_d)$，Gibbs sampler 使用充分條件 $P(z \mid \mathbf{z}_{\neg i}, \mathbf{w}_d)$ 執行 Markov chain。充分條件透過估算等式(25) 隱藏變數的方法可以寫成

$$P(z \mid \mathbf{z}_{\neg i}, \mathbf{w}_d) = \frac{P(\mathbf{z}, \mathbf{w}_d)}{\int_z P(\mathbf{z}, \mathbf{w}_d) dz} \tag{26}$$

其中，$\mathbf{z}_{\neg i}$ 定義爲 $\mathbf{z} - \{z_i\}$，表示除了目前的字詞 w_i 之外，對所有字詞的主題分配。在 BTMM 中，聯合分佈可以被分解爲

$$P(z, w \mid d, \beta) = P(w \mid z, \beta) P(z \mid d) \tag{27}$$

等式右邊的兩個元素能夠被分別處理，第一項 $P(w \mid z, \beta)$ 可以由給定相關主題的被觀察字詞總數之多項式導出，如式(28)所示

$$\begin{aligned} P(w \mid z, \beta) &= \int_\phi P(w \mid \phi) P(\phi \mid \beta, z) d\phi = \int_\phi \frac{n!}{\prod_w n_w!} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_z^{n_z^{(w)} + \beta_w - 1} d\phi \\ &\cong \frac{\Gamma(\sum_w \beta_w)}{\Gamma(\sum_w \beta_w + n_z^{(w)})} \prod_w \frac{\Gamma(n_z^{(w)} + \beta_w)}{\Gamma(\beta_w)} \end{aligned} \tag{28}$$

其中，n_z 定義爲字詞 w 被分配到潛在主題變數 z 發生的次數。在式(28)中，$\prod_w \phi_w^{n_w}$ 和 $\prod_w \phi_w^{\beta_w - 1}$ 結合是 Dirichlet 分佈 $P(\phi \mid n_w + \beta_w)$ 的未正規化變化形式，並利用 $\int P(\phi \mid n_w + \beta_w) d\phi = 1$ 推導所得。過程中，不需導入參數 ϕ，因爲他們只是在被觀察資料 (d,w) 和對應主題 z 之馬可夫鏈的狀態變數之間的關聯統計。考慮式(28)中的分佈，只對包含索引 i 之潛在變數 z 乘積項保留，其他全部消去。更進一步地，利用等式

$\Gamma(x)=(x-1)\Gamma(x-1)$。因此，式(28)可以重寫爲

$$\hat{P}_{\text{BTMM}}(w\mid z,\mathbf{z}_{\neg i})=\frac{P(w\mid \mathbf{z})}{P(w\mid \mathbf{z}_{\neg i})}=\frac{\frac{\Gamma(n_z^{(w)}+\beta_w)}{\Gamma(\sum_{w'}n_z^{(w')}+\beta_w)}}{\frac{\Gamma(n_z^{(w)}-1+\beta_w)}{\Gamma([\sum_{w'}n_z^{(w')}+\beta_w]-1)}}=\frac{n_{z,\neg i}^{(w)}+\beta_w}{[\sum_{w'}n_z^{(w')}+\beta_{w'}]-1}\propto\frac{n_{\neg i,z}^{(w)}+\beta_w}{n_{\neg i,z}^{(\cdot)}+V\beta_{w'}} \tag{29}$$

同理，潛在主題分佈 $P(z\mid d)$ 可以被推得以下結果

$$\hat{P}_{\text{BTMM}}(z\mid d,\mathbf{z}_{\neg i})=\frac{n_{d,\neg i}^{(z)}}{\sum_{z'}n_{d,\neg i}^{(z')}} \tag{30}$$

其中，$n(d,w)$ 表示字詞 w 在文件 d 中出現的個數。最後，對於潛在變數，由式(29)、(30)我們可以推導出更新等式，其結果爲

$$P(z_i\mid \mathbf{z}_{\neg i},w,d)\propto P(w\mid z,\mathbf{z}_{\neg i})P(z\mid d,\mathbf{z}_{\neg i})\propto\frac{n_{\neg i,z}^{(w)}+\beta_z}{n_{\neg i,z}^{(\cdot)}+V\beta_z}\cdot\frac{n_{d,\neg i}^{(z)}}{\sum_{z'}n_{d,\neg i}^{(z')}} \tag{31}$$

其中，$n_{\neg i,z}^{(w)}$ 表示字詞 w 分配給主題 z 的次數 ，$n_{d,\neg i}^{(z)}$ 包含主題 z 在文件 d 裡被分配到一些字詞 w 的次數，而 $n_{\neg i,z}^{(\cdot)}$ 表示所有字詞分配給主題 z 的總數，標記 $\neg i$ 表示當前字詞 w_i 在這些計數已被移去，不被列入計算考慮。β表示 Dirichlet priori，在本模型裡，對全部字詞β假設是相同的，亦即 β 的所有組成部分都相同。z_i 的初始被設定介於值 1 到 K 之間，決定馬可夫鏈(Markov chain)的初始狀態。然後執行幾個迭代次數，直到鏈接近目標分佈， z_i 目前值將會被記錄下來。

(三) 不同模型之關聯和比較

在本章節中，我們將討論並比較前面章節所描述的幾個模型。從主要的方程式看來，模型之間差異大同小異。爲了容易理解文件模型生成的差異。針對本文所提出的方法和第二章所提到的模型，如 PLSA、LDA 以及 TTMM 等，對其組成元素(字詞、主題及文件)之生成機率/分佈表示，簡單歸納如下表一所示。

表一、不同方法之各組成元素機率分佈表示

	Word	Topic	Document
PLSA	$P(w\mid z)$	$P(z\mid d)$	$P(d,w)$
LDA	$w\mid z,\beta\sim Mult(\beta)$	$z\sim Mult(\theta)$	$\theta\sim Dir(\alpha)$
TTMM	$w\mid z,\beta\sim Mult(\beta)$	$z\sim Mult(\tau)$	$h\sim Mult(\pi)$
BTMM	$w\sim Mult(\phi_z),\phi_z\sim Dir(\beta)$	$P(z\mid d)$	$P(d,w)$

假設在文件集裡有 N 篇文件，字典數大小爲 M，$|d|$表示文件長度，亦即在文件的字詞個數，K 爲主題(Topic)個數，J 爲 theme 數目以及群組個數爲 C。對於模型的空間複雜度比較，以表三做一簡單的闡述。各個模型所需的參數量，從表二可以得知，TTMM 需要 $J(1+K)+KM$ 個參數，而 LDA 只需 $K+KM$ 個參數。主要是由於連續分佈使用一

個參數，在 LDA 產生混合比例 θ 參數，取代在 TTMM 兩個離散分佈。除此，當文件透過主題(theme)被群聚在一起，如此 $J < N$，則 TTMM 的參數量可能少於 PLSA 的參數量 $KN + KM$。在 BTMM 模型中，字詞是經由主題 z 的多項分佈 ϕ 所產生，而對於字詞分佈的具體主題多項分佈 ϕ，可以從 Dirichlet priori 參數 β 對應的主題 z 得到，其參數量比 PLSA 少，只需 $KN + K$ 個。

表二、對不同模型之空間複雜度比較

	PLSA	LDA	TTMM	BTMM
Parameters	$O(KN+KM)$	$O(K+KM)$	$O(J+JK+KM)$	$O(KN+K)$

四、實驗

(一)、實驗文集及設定說明

在本文的實驗中，我們使用 TREC 所收集的文集，分別爲 Associated Press newswire (AP) 88 和 Wall Street Journal (WSJ) 89，資料的統計資訊，如表三所示。我們所使用測試的查詢句子爲 Topics 101-150，主要取各個主題中的標題(title)和敘述(description)部分作爲查詢句，每個查詢句的平均長度爲 14.48 個字。文件會先經過 stop word 和 stemming 的前處理。本文分別對此兩文集以文件檢索和文件模組化驗證本文方法的正確性和可行性。在實驗中主要是針對 Language Model (LM)、PLSA、LDA 及本文所提出的 BTMM 做比較。對於潛在變數 k 的個數，初始實驗設定爲 16。實驗分爲兩個部分，第一評估各個模型應用在文件檢索上的效能，以 Precision-Recall curve 和 mAP 作爲評估的準則 [15]。第二個是以 perplexity 評估文件模型的效果。

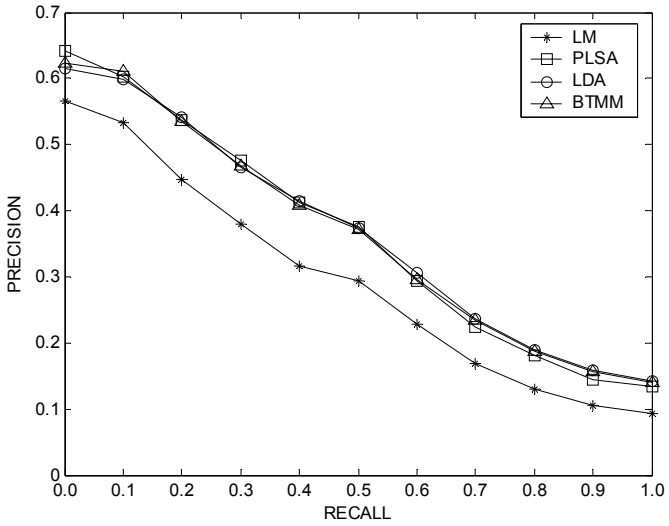
表三、TREC 文集的統計資訊

Collection	Description	Size (MB)	#Doc.	Vocabulary Size
WSJ89	Wall Street Journal (1989), Disk2	36.5	12,380	17,732
AP88	Associate Press (1988), Disk1	237	79,908	8,783

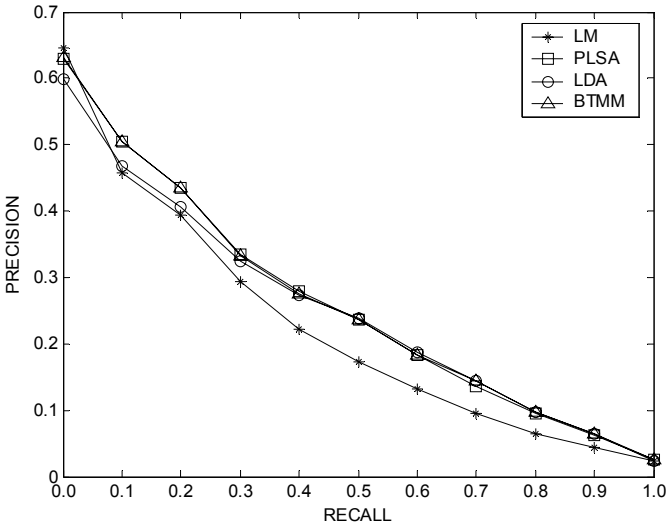
(二)、實驗結果

1、不同模型在檢索效能的影響

首先，比較不同的方法對 TREC 文件集在文件檢索上效能的比較。從圖七和圖八表示不同模型之 Precision-Recall 曲線，分別 WST89 和 AP88 的結果，而表四爲 mAP 在不同模型所計算的結果。從這些圖表當中，可以看出以主題爲基礎的文件模型，皆比語言模型有更好的效能。BTMM 的效能雖然比 PLSA 好，然而效果並不明顯。分析其原因，其影響的因素可能來自潛在主題變數 k 值的設定和參數值初始的設定。另外，文件前處理 stemming 亦可能造成影響。因此，在未來的實驗，將針對這些部分更進一步探討。



圖七、Precision-recall curves 對不同方法在 WSJ89 文集上的比較



圖八、Precision-recall curves 對不同方法在 AP88 文集上的比較

表四、LM、PLSA、LDA 以及 BTMM 在不同文集中 mAP 之比較

	LM	PLSA	LDA	BTMM
AP88	0.2128	0.2507	0.2411	**0.2536**
WSJ89	0.2761	0.3448	**0.3507**	0.3486

2、不同模型在文件模組化的評估

在文件模組化的實驗過程裡，以 WSJ89 爲實驗資料，將文件分爲兩個部分，三分之二的資料量作爲基礎模型的訓練資料集，共 7,931 篇文件，另外，三分之一部份做測試的文件資料集合，包含 4,449 篇文件。初步實驗結果如表五所示。從表中可以看出 BTMM 比 LM 和 PLSA 模型有較好的結果，其 perplexity 分別由 257.59 和 251.8 降至 250.42。

表五、不同模型之間 perplexity 之比較

	LM	PLSA	LDA	BTMM
Perplexity	257.59	251.8	248.63	250.42

BTMM 主要是改進 PLSA 中，字詞和主題之間的表示型態，以 Dirichlet 分佈替代原始的多項分佈，在字詞的主題分佈上導入 Dirichlet 事前機率，使得資訊更完整和豐富。然而，從實驗結果我們可以發現比 LDA 略差。針對此部分，我們將對字典個數的影響更進一步的探討分析。其分析結果如表六所示。我們分別選取字典字數一萬、二萬及三萬字來做對照，潛在主題變數個數設定爲 8。

表六、不同字典個數對 perplexity 值的影響

	10,000	20,000	30,000
LM	247	380	511
PLSA	240	372	504
LDA	**205**	**365**	505
BTMM	232	369	**495**

從表六可以得知，當字典數增加時，模型針對文字發生機率的預測分支度越高，所以 perplexity 都呈現上升的趨勢。當字典大小約爲 3 萬字時，BTMM 的 perplexity 比 LDA 低。主要原因是因爲當我們過度對字典數做刪減時，突發現象對模型的影響變得輕微。而由於 LDA 模型對文件階層加入事前機率，使得估算文件的主題分佈時，較貼近真實的分佈情形。然而，在字典數較大時，從實驗數據，可以發現突發現象較爲顯著，使得在文件中較稀有但卻具有鑑別性的字詞對模型產生影響，由於 BTMM 模型對字詞的主題分佈導入 Dirichlet 事前分佈，使得在 perplexity 的評估上略比 LDA 佳。

五、結論

本文中主要是以機率模型爲基礎提出一個貝氏理論的文件模型，致力解決 bag-of-word 表示法的問題，並對現有模型做改進，以期達到更好的效能。其架構延伸原始 PLSA 模型的概念，對於一個主題的條件分佈以 Dirichlet 代替原有的多項分佈表示，在此稱之爲貝氏主題混合模型。文中利用 Gibbs 抽象法估計模型未知參數，此方法的優點是不需要明確地表達模型參數且實做上比較容易，對記憶體需求量也比較少。在主題混合模型中，雖然假設文件可由不同主題所產生，但文件與字詞彼此之間是獨立的。然而，在真實世界裡，文件之間通常是有關聯的。例如，在新聞的文件標題中，可以分爲主要主題和次要主題。在 Tam 和 Schultz[34]的研究中，以 Dirichlet Tree[26]代替 LDA 中 Dirichlet Prior，使得潛在主題可以表達更多關聯。在未來的研究方向，對於文件模型演算法，我們擬延伸至層級概念，將文件以少量的概念或是主題來呈現，使得模型更具有強健性。另外，目前文件的機率模型表示法，大致以 Unigram 爲主，如何結合 *n*-gram 語言模型，使得文件模型更具強健性，亦是未來研究工作。

參考文獻

[1] Y. Akita and T. Kawahara, “Language model adaptation based on PLSA of topics and speakers”, *Proceedings of International Conference on Spoken Language Processing*, pp. 1045-1048, 2004.

[2] J. R. Bellegarda, “Exploiting latent semantic information in statistical language modeling,” *Proceeding of the IEEE,* vol. 88, No. 8, pp. 1279-1296, 2000.

[3] M. W. Berry, S. T. Dumais and G. W. O’Brien, “Using linear algebra for intelligent information retrieval”, *SIAM Review*, vol. 37, no. 4, pp. 573-595, 1995.

[4] D. M. Blei and J. D. Lafferty, “Correlated topic model”, *Advances in Neural Information Processing Systems* (*NIPS*), vol. 18, pp. 147-154, 2006.
[5] D. M. Blei and J. D. Lafferty, “Dynamic topic model”, *Proceedings of the 23rd International Conference on Machine Learning*, pp.113-120, 2006.
[6] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993-1022, 2003.
[7] T. Brants, F. Chen and I. Tsochantaridis, “Topic-based document segmentation with probabilistic latent semantic analysis”, *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 211-218, 2002.
[8] J.-T. Chien, M.-S. Wu and C.-S. Wu, “Bayesian learning for latent semantic language”, *Proceedings of European Conference on Speech Communication and Technology*, pp. 25-28, 2005.
[9] J.-T. Chien, M.-S. Wu and H.-J. Peng, “On latent semantic language modeling and smoothing”, *Proceedings of International Conference on Spoken Language Processing*, vol. 2, pp. 1373-1376, 2004.
[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.
[11] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society*, Series B, vol. 39, no. 1, pp. 1-38, 1977.
[12] C. Elkan, “Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution”, *Proceedings of the 23rd International Conference on Machine Learning*, pp. 289-296, 2006.
[13] M. Girolami and A. Kaban, “On an equivalence between PLSI and LDA”, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433-434, 2003.
[14] T. L. Griffiths and M. Steyvers, “Finding scientific topics”, *Proceedings of the National Academy of Science*, vol. 101, pp. 5228–5235, 2004.
[15] D. Harman, Overview of the Fourth Text Retrieval Conference. 1995. Available at http://trec.nist.gov/pubs/trec4/overvies.ps.gz
[16] T. Hofmann, “Probabilistic latent semantic analysis”, *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 289-296, 1999.
[17] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis”, *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.
[18] T. Hofmann, “Unsupervised learning from dyadic data”, *Advances in Neural Information Processing Systems*, vol. 11. MIT Press, 1999.
[19] X. Jin, Y. Zhou and B. Mobasher, “Web usage mining based on probabilistic latent semantic analysis”, *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 197-205, 2004.
[20] M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambrige, MA, 1999.
[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Sail, “Introduction to variational methods for graphical models”, *Machine Learning*, vol. 37, pp. 183-233, 1999.
[22] S. M. Katz, “Distribution of content words and phrases in text and language modeling”, Natural Language Engineering, vol. 2, pp. 15-59, 1996.
[23] M. Keller and S. Bengio, “Theme topic mixture model: A graphical model for document representation”, *in PASCAL Workshop on Learning Methods for Text Understanding and Mining*, 2004.
[24] T. G. Kolda and D. P. O’Leary, “A semi-discrete matrix decomposition for latent semantic indexing in information retrieval”, *ACM Transactions on Information Systems*,

vol. 16, no. 4, pp. 322-346, 1998.
[25] R. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution", *Proceedings of the 22nd International Conference on Machine Learning*, pp. 545-552, 2005.
[26] T. Minka, "The Dirichlet-tree distribution", in http://research.microsoft.com/~minka/papers/dirichlet/minka-dirtree.pdf
[27] T. Minka, "Estimating a Dirichlet distribution", *Technical Report, MIT*, 2000.
[28] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model", *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 352-359, 2002.
[29] D. Mrva and P. C. Woodland, "A PLSA-based Language Model for Conversational Telephone Speech", *Proceedings of International Conference on Spoken Language Processing*, pp. 2257-2260, 2004.
[30] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription", *Proceedings of International Conference on Spoken Language Processing*, pp. 1961-1964, 2004.
[31] K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell, "Text classification from labeled and unlabeled documents using EM", *Machine Learning*, vol. 39, no. 2-3, pp. 103-134, 2000.
[32] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
[33] Y.-C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference", *Proceedings of European Conference on Speech Communication and Technology*, pp. 5-8, 2005.
[34] Y.-C. Tam and T. Schultz, "Correlated latent semantic model for unsupervised LM adaptation", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 41-44, 2007.

Korean-Chinese Cross-Language Information Retrieval Based on Extension of Dictionaries and Transliteration

Yu-Chun Wang[†‡], Richard Tzong-Han Tsai[†§], Hsu-Chun Yen[‡], Wen-Lian Hsu[†]

[†]Institute of Information Science, Academia Sinica, Taiwan
[‡]Department of Electrical Engineering, National Taiwan University, Taiwan
[§]Department of Computer Science and Engineering, Yuan Ze University, Taiwan

{albyu,thtsai}@iis.sinica.edu.tw
yen@cc.ee.ntu.edu.tw
hsu@iis.sinica.edu.tw

Abstract

This paper describes our Korean-Chinese cross-language information retrieval system. Our system uses a bi-lingual dictionary to perform query translation. We expand our bilingual dictionary by extracting words and their translations from the Wikipedia site, an online encyclopedia. To resolve the problem of translating Western people's names into Chinese, we propose a transliteration mapping method. We translate queries form Korean query to Chinese by using a co-occurrence method. When evaluating on the NTCIR-6 test set, the performance of our system achieves a mean average precision (MAP) of 0.1392 (relax score) for title query type and 0.1274 (relax score) for description query type.

摘要

本文描述我們所提出之韓中雙語跨語言檢索系統。我們採用韓中雙語辭典進行問題之翻譯，並利用線上維基百科以及韓國 Naver 網站來擴增我們雙語辭典的覆蓋率。此外，針對韓文中西方人名的翻譯，我們提出一音譯對應的搜尋方法。對於韓中翻譯時的歧義性問題，我們採用 Mutual Information 方法來解決。我們使用 NTCIR-6 之 test set 測試我們韓中跨語言檢索系統之效率，其使用標題部分進行查詢時之 Mean average precision (MAP) 之結果爲 0.1392；使用敍述部分進行查詢時之 MAP 爲 0.1274。

Keywords: Korean-Chinese cross-language information retrieval, query translation
關鍵詞：韓中跨語言資訊檢索，問題翻譯

1 Instroduction

The contents of whole Internet are growing explosively due to the improvement of the computer and web technology. Besides English, the web pages written in other languages also increase tremendously. In order to get the useful information from the Internet, many advanced modern search engines are developed, like Google[1], Yahoo[2], AltaVista[3], and so on. However, for the users that do not have any knowledge about other languages, it is impossible to get the information in other languages by current single-language web search engines.

Therefore, the research of cross language information retrieval (CLIR) is rising quickly. Cross language information retrieval systems allow the users to input the key words in their own languages and then the systems will retrieve the relevant documents written in the other language that the users want to search based on the queries the users inputted.

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://www.altavista.com

There are many different approaches of CLIR. The first is the translation methods. There are two kinds of approaches usually adopted: translation approach and statistical approach. The translation approach uses the bilingual dictionaries, ontology, or thesaurus to translate either queries or documents. The statistical approach uses pre-constructed bilingual corpora to extract the cross-lingual associations without any language translation methods [1–3]. The translation approach is restricted with the coverage and the precision of the dictionaries. The statistical approach can extract bilingual lexicons automatically; however, it bases on a well-constructed and large-scaled bilingual corpus which requires a lot of human effort.

In translation approach, there are two different targets to do the translation. One is document translation; the other is query translation. The document translation approach is to translate all the documents in the collection from the target language to the source language the users use. Then, while the users give the query, the system will do a monolingual information retrieval. The query translation translates the query in the source language that user inputted into the target language and then retrieval the documents which is written in the target language. The document translation approach is possible if there exists a high quality machine translation system. [4, 5] However, the document translation approach is not very practical when the documents are not stable or can be updated frequently, like the web text retrieval.

In this paper, we propose a Korean-Chinese cross-language information retrieval system. We adopt the query-translation approach because it is effective. Moreover, the translation method, which is dictionary-based, does not involve a great deal of work. In CLIR, the most serious problem is that unknown words cannot be translated correctly. To resolve the problem, we utilize Wikipedia, an online encyclopedia, to expand our dictionary to make higher coverage of vocabulary. Another difficult issue involves translating Western people's names written in Korean into Chinese. As a solution, we propose a transliteration mapping method to deal with the problem.

The remainder of the paper is organized as follows. In Section 2, we give an overview of our system and describe its implementation, including the translation and indexing methods adopted. In Section 3, we detail the evaluation results of our CLIR system based on the topics and the document collections provided by NTCIR CLIR task, and discuss the effectiveness of our method, as well as some problems that have to be solved. Finally, in Section 4, we present our conclusions and indicate the direction of our future work.

2 System Description

Figure 1 shows the architecture of our CLIR system. It is comprised of four stages. First, a Korean query is chunked into several key terms, which are then translated into Chinese by three dictionaries. In the third stage, we disambiguate the translated terms and transform them into a Lucene query. Finally, the query is sent to the Lucene IR engine and the answer is retrieved.

2.1 Query Processing

Unlike English, Korean written texts do not have word delimiters. Spaces in Korean sentences separate eojeols, which are composed of a noun and a postposition, or a verb stem and a verb ending. Therefore, Korean text has to be segmented. There are two types of queries that the users might make. One is composed of several key words; the other is an natural language sentence. Therefore, we use two different segmentation methods to deal with these two query types separately.

Due to the characteristics of the Korean language, the keyword-typed queries written in Korean are comprised mainly of nouns. We use spaces to split the title into several eojeols, and

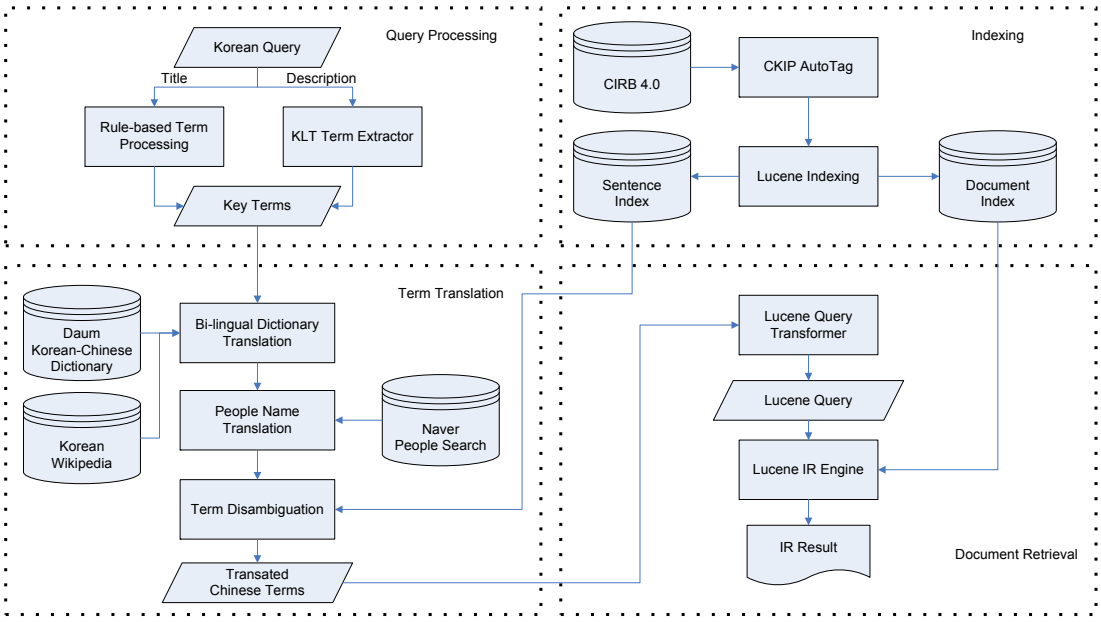


Figure 1: System Architecture of Our CLIR System

then remove the postpositions at the end of the eojeols according to our predefined rules.

For the natural language sentence of a Korean query, we use the KLT Term Extractor[4], developed by Kookmin University in Korea. KLT term extractor will do the word segmentation to extract vital key words which are useful for information retrieval and remove stop words.

2.2 Query Translation

2.2.1 Bilingual Dictionary Translation

Due to copyright restrictions, we use the free online Korean-Chinese dictionary provided by the Daum Korean web site[5]. We send the key terms obtained in the query processing stage to the online dictionary. However, as a general bilingual dictionary is not suitable for proper nouns, we use Wikipedia[6], an online encyclopedia, to expand our dictionary. In Wikipedia, an item might contain inter-language links to the same item in Wikipedia written in other languages. Therefore, we send a Korean term to Korean Wikipedia. If it contains an inter-language link to Chinese Wikipedia, we can find the corresponding Chinese word. This method is very efficient because it yields accurate Chinese translations of Korean words.

The Daum Korean-Chinese dictionary is written in simplified Chinese, as are many pages in Chinese Wikipedia. We use a simple mapping table to convert simplified Chinese characters to traditional Chinese characters.

If some terms cannot be found in the Daum dictionary or Wikipedia, we apply the maximal matching algorithm to split a long term into several shorter terms. Then, the shorter terms are sent to the Daum dictionary and Wikipedia to search for Chinese translations.

[4] http://nlp.kookmin.ac.kr/HAM/kor/index.html
[5] http://cndic.daum.net
[6] http://www.wikipedia.org

2.2.2 Person Name Translation

The person names may often appear in the query. Although Wikipedia contains many famous people's names around the world, some people's names are still excluded. Therefore, it is necessary to deal with this person name translation problem. Unlike Korean-English or Korean-Japanese CLIR, transliteration methods are not appropriate for Korean-Chinese CLIR because so many Chinese characters have the same pronunciation in Korean. Besides, to translate Japanese personal names, Korean uses the Hangul alphabet to pronounce the names of Japanese people; however, Chinese uses original Chinese characters with Mandarin pronunciation, instead of Japanese pronunciation of Chinese characters. Thus, transliteration methods are not useful in this context. To solve the problem, we use Naver People Search[7], a database containing the basic profiles of famous people, including their original names. We can submit person names in Korean to Naver people search and get their original names. If the original name is composed of Chinese characters, it is clearly Chinese, Japanese, or Korean; therefore, we can send it to next stage directly, i.e., the disambiguation stage. If, however, the original name is in English, we use the English name translation table provided by Taiwan's Central News Agency (CNA)[8] to translate it into Chinese and the proceed to the next stage.

2.3 Term Disambiguation

In the past, the Korean language adopted many Chinese words. More than half of its vocabulary comprises Chinese words. Now, however, Koreans use Hangul, an alphabet writing system, instead of Chinese characters, which is an ideograph writing system. As a result, many different Chinese loanwords have the same pronunciation when written in the Hangul alphabet. For example, the four different Chinese loanwords with different meanings: "理想" (ideal), "以上" (above), "異常" (unusual), and "異狀" (indisposition) are written in the same way as the Hangul word "이 상" because their pronunciation is the same in Korean. This creates a very serious ambiguity problem when Korean is translated into Chinese. Therefore, choosing the correct translation term among translation candidates is important.

For each term in a given query Q, there may be several possible translation candidates. To select the best translation term among all the candidates, we must not only consider the original query term qt but also consider all the other terms in Q and their translation candidates. We denote the j-th translation candidate for the i-th term qt_i in Q as tc_{ij}. We adopt the mutual information score (MI score) [6] to evaluate the co-relation between the tc_{ij} and all translation candidates of all the other terms in Q. The MI score of tc_{ij} given Q is calculated as follows:

$$\text{MI score}(tc_{ij}|Q) = \sum_{x=1,x\neq i}^{|Q|} \sum_{y=1}^{Z(qt_x)} \frac{Pr(tc_{ij}, tc_{xy})}{Pr(tc_{ij})Pr(tc_{xy})},$$

where $Z(qt_x)$ is the number of translation candidates of the x-th query term qt_x; $Pr(tc_{ij}, tc_{xy})$ is the probability that tc_{ij} and tc_{xy} co-occur in the same sentence; and $Pr(tc_{ij})$ is the probability of tc_{ij}. The values of the probabilities are obtained from Chinese Information Retrieval Benchmark (CIRB) Chinese corpus which is provided by NTCIR CLIR task [7]. The higher the translation candidate's MI score is, the higher weight is assigned to it in the retrieval module.

[7] http://people.naver.com

[8] http://client.cna.com.tw/name/

2.4 Chinese Document Indexing

The Chinese documents we use is CIRB 4.0 documents which is provided by NTCIR. The CIRB 4.0 documents are pre-processed to remove noise and then segmented by CKIP AutoTag [8] to obtain words and part-of-speech (POS). We use Lucene[9], an open source information retrieval engine, to index Chinese documents. Our index is based on Chinese characters.

2.5 Lucene Queries

After processing a Korean query into several terms and translating it into Chinese, we transform the Chinese terms into a Lucene Query. Different Chinese terms are separated by a space, which means an "OR" operator in the Lucene format. If a term has different translation candidates, the weight of the candidate with highest mutual information score will be increased by 1 by the boost operator. The other candidates are boosted by a weight that is the reciprocal of the total number of candidates. The boost operation affects the ranking of the documents the Lucene returns. The default boost value of each terms is 1, and we decrease the weight of the candidates with lower mutual information score to make them not affect the ranking so much.

3 Evaluation and Analysis

In order to evaluate our CLIR system, we use the topics and the document collections which is provided by NTCIR-6 CLIR task [7]. The topics contains 50 Korean queries composed of four parts: title, description, narration, and keywords. We use these topics as the queries that users inputted in our system.

The main metric to evaluate the performance of information retrieval is Mean Average Precision (MAP) [9]. Average precision is based on the whole list of documents returned by the system and emphasizes returning more relevant documents earlier. The Mean Average Precision is the mean value of the average precisions computed for each query. Besides, R-precision [10] is also a good metric which is the precision among the front of R relevant documents.

There are two kinds of relevance judgments: Rigid and Relax. A document is rigid relevant if it is highly relevant; a document is relax relevant if it is highly relevant or partial relevant. Our evaluation is based on the 50 topics which is selected by NTCIR-6 CLIR task to compute among all 140 topics they provided.

In order to evaluate the effectiveness of our CLIR system, we build a monolingual Chinese IR system for comparison. NTCIR-6 CLIR test set also contains the Chinese topics which meanings are the same as Korean ones. We use these Chinese topics as queries and apply CKIP AutoTag to do Chinese word segmentation and remove Chinese stop words. Then, we use Lucene search engine we use in our CLIR system to retrieve related Chinese documents.

We do the four different runs:

- **KC-title-run**: a run using a Korean title field to retrieve Chinese documents.
- **KC-description-run**: a run using a Korean description field to retrieve Chinese documents.
- **CC-title-run**: a monolingual Chinese run using Chinese title field to retrieve Chinese documents.

[9] http://lucene.apache.org/

Table 1: Evaluation Results

Run	Rigid		Relax	
	MAP	R-precision	MAP	R-precision
KC-title-run	0.1118	0.1420	0.1392	0.1781
KC-description-run	0.1022	0.1311	0.1274	0.1760
CC-title-run	0.1501	0.1961	0.2141	0.2747
CC-description-run	0.1567	0.2111	0.2157	0.2788

- **CC-description-run**: a monolingual Chinese run using Chinese description field to retrieve Chinese documents.

Table 1 shows the performance of our Korean-Chinese CLIR system and the monolingual Chinese IR system. The performance of Korean-Chinese CLIR is not as good as that of Chinese monolingual IR. We have investigated why it is difficult to retrieve high precision answers to some queries.

3.1 Problems of Bilingual Dictionaries

We use a general bilingual dictionary and Wikipedia to translate most of the words in 50 topics provided by NTCIR-6 CLIR task. Although we have used Wikipedia to expand our dictionary, there are some problems that cause translations to fail. The first problem is that there are still some unknown words. For example, the word "배아" (embryo) is not listed in the dictionaries. The other problem is that the dictionaries do not always have the proper translation candidates of the words and terms in queries. For instance, the word "감청" (monitor) is not translated correctly because the dictionary lacks the correct translation and provides another translation instead, i.e., "紺青" (deep blue). Also, the word "암" (cancer) in one topic is translated as "岩" (rock), "庵" (nunnery), and "雌" (female), but no correct translation, i.e., "癌" (cancer).

3.2 Different Phraseology Used in Taiwan and China

The Daum Korean-Chinese dictionary that we use was written people studying Mainland Chinese, i.e., Pinyin. However, the CIRB 4.0 document collection contains Taiwanese newspapers. Taiwanese people use traditional Chinese characters, whereas Mainland Chinese people use simplified characters. Besides the difference in characters, the vocabulary and grammar used in Taiwan and China are slightly different. The differences between Taiwanese Chinese and Mainland Chinese can make IR difficult.

The following are some examples of the difficulties we face. The term "휴대폰" (mobile phone) is translated into Mainland Chinese word as "移動電話" (the phone that can move); however, the correct word used in Taiwan is "手機" (the machine held in the hand). The word "유전자" (gene) is translated to "遺傳子" (the factor of heredity) , not to correct word "基因" (the Mandarin transliteration of the English word "gene") used in Taiwan. The word "인터넷" (internet) in some topics is translated to "互聯網" (the net connecting to each other) , but the correct word used in Taiwan is "網際網路" (cyber network).

3.3 The Limitations of Maximal Matching Algorithm

If a term is not defined in our dictionaries, we split it into several shorter terms by the maximal matching algorithm discussed in Section 2.2.1. In some cases, however, the algorithm do not

segment a term correctly. For example, for the term "비 접 촉 형"(contactless), the correct segmentation is 비(not)-접촉(contact)-형(type). However, it is segmented as 비 접-촉-형 so that the wrong word, "비 접"(convalescing), is retrieved.

3.4 Different Expressions Used in Korean and Chinese

In some topics, different expressions used in Korean and Chinese may cause translation problems. In one of the topic, the word "10대" refers to people aged between 10 and 19. Similarly, "20대" means people aged from 20 to 29. Therefore, the corresponding translation of the word "10대" in this topic is "青少年" (teenager). However, our system translates the numbers and the Hangul characters separately so that the final translation is "10代" (ten generations). This is a semantic problem that our system has difficulty coping with.

Another problem relates to abbreviations used in Chinese. For instance, in another topic, "외국인 노동자" (foreign worker) is translated into "外國人勞工" (foreign worker) by our system. However, in Taiwanese newspapers, the abbreviation "外勞" , which is composed of the first characters of the two words : "外國人" (foreigner) and "勞工" (worker), is used more frequently. Our translation in one of the topic for the phrases "원자능 반대" is "反對 核能" (anti-nuclear), but the abbreviation "反核" is frequently used.

4 Conclusions and Future Works

We have described our Korean-Chinese CLIR system. It is based on a query-translation approach and uses a general Korean-Chinese dictionary and Wikipedia to translate words and terms. To obtain person names, we use the Naver people search website and the CNA transliteration table to translate the names.

We have evaluated the performance of our Korean-Chinese CLIR system with the Korean topics and the Chinese document collection which is provided by NTCIR-6 CLIR task. Our translation method is effective, but there are still some cases where the precision is low. We believe the problems are due to the limitations of the dictionaries, the different phraseology used in Taiwan and China, and the expressions used in Chinese and Korean.

In our future work, we will apply a Chinese thesaurus to overcome the problem of different Chinese phraseology and use more bilingual dictionaries to reduce the number of unknown words. We will also incorporate a query expansion method into our CLIR system to improve its precision.

References

[1] S. Dumais, T. Letsche, M. Littman, and T. Landauer, "Automatic cross-language retrieval using latent semantic indexing", in *AAAI Symposium on CrossLanguage Text and Speech Retrieval*. 1997, American Association for Artificial Intelligence.

[2] Bob Rehder, Michael L. Littman, Susan Dumais, and Thomas K. Landauer, "Automatic 3-language cross-language information retrieval with latent semantic indexing", in *Sixth Text REtrieval Conference (TREC-6)*, 1997.

[3] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking, "Translingual information retrieval: Learning from bilingual corpora", *Artificial Intelligence*, vol. 103, pp. 323–345, 1998.

[4] Oh-Wook Kwon, I.S. Kang, J-H Lee, and G.B. Lee, "Cross-language text retrieval based on document translation using japanese-to-korean mt system", in *NLPRS*, 1997, pp. 101–106.

[5] Douglas W. Oard and Paul Hackett, "Document translation for the cross-language text retrieval at the university of maryland", in *the Sixth Text REtrieval Conference (TREC-6).*

[6] Hee-Cheol Seo, Sang-Bum Kim, Ho-Gun Lim, and Hae-Chang Rim, "Kunlp system for ntcir-4 korean-english cross-language information retrieval", in *NTCIR-4*, Tokyo, 2004.

[7] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, and Hsin-Hsi Chen, "Overview of clir task at the sixth ntcir workshop", *Proceedings of NTCIR-6 Workshop Meeting*, 2007.

[8] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, and Wen-Lian Hsu, "Asqa: Academia sinica question answering system for ntcir-5 qa", in *NTCIR-5*, Tokyo, 2005.

[9] Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison, "A study of information seeking and retrieving", *Journal of the American Society for Information Science*, vol. 39, no. 3, pp. 161–176, 1988.

[10] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

加成性雜訊環境下運用特徵參數統計補償法於強健性語音辨識

Feature Statistics Compensation for Robust Speech Recognition in Additive Noise Environments

謝宗學　Tsung-hsueh Hsieh
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
s94323532@ncnu.edu.tw

洪志偉　Jeih-weih Hung
國立暨南國際大學電機工程學系
Department of Electrical Engineering
National Chi Nan University
jwhung@ncnu.edu.tw

摘要

在自動語音辨識的研究上，如何有效地降低背景雜訊的影響，以增加語音辨識系統的強健性，一直是一大研究重點，其中語音特徵參數正規化法是廣為人用的強健技術之一。然而，對於多變的語音訊號該如何準確的估算出其不同時段的統計值，是影響語音特徵參數正規化法效果的一個重要因素。本論文主要是針對在加成性雜訊環境下，對不同的特徵參數提出更有效率且準確的統計值補償法，以降低加成性雜訊對語音特徵參數的影響。我們提出了運用虛擬雙通道碼簿為基礎之特徵參數補償技術，其中包含三種方法：倒頻譜統計補償法、線性最小平方回歸法與二次最小平方回歸法。我們將這些方法運用在四種不同的語音特徵參數的補償上，發現都能有效降低加成性雜訊對語音特徵的影響，進而大幅提升辨識率，同時，在與傳統以段落為基礎的特徵參數正規化技術比較下，我們所提出的方法可達到更佳的強健效果。

Abstract

In this paper, we propose several compensation approaches to alleviate the effect of additive noise on speech features for speech recognition. These approaches are simple yet efficient noise reduction techniques that use online constructed pseudo stereo codebooks to evaluate the statistics in both clean and noisy environments. The process yields transforms for noise-corrupted speech features to make them closer to their clean counterparts. We apply these compensation approaches on various well- known speech features, including mel-frequency cepstral coefficients (MFCC), autocorrelation mel-frequency cepstral coefficients (AMFCC), linear prediction cepstral coefficients (LPCC) and perceptual linear prediction cepstral coefficients (PLPCC). Experimental results conducted on the Aurora-2 database show that the proposed approaches provide all types of the features with a significant performance gain when compared to the baseline results and those obtained by using the conventional utterance-based cepstral mean and variance normalization (CMVN).

關鍵詞：自動語音辨識、虛擬雙通道碼簿、倒頻譜統計補償法、線性最小平方回歸法、二次最小平方回歸法

Keywords: automatic speech recognition、pseudo stereo codebooks、cepstral statistics compensation,

linear least squares regression, quadratic least squares regression

一、緒論

本論文主要重點是在加成性雜訊環境下，對語音特徵參數補償法的探討，目的是使測試語音的統計特性在經過補償後能更接近訓練語音的統計特性。

我們運用四種語音特徵參數擷取技術結合兩大類特徵參數補償法，並觀察兩類特徵參數補償法之間的差異與優缺點。本論文中所討論的兩大類特徵參數補償法分別爲：

(1)以段落爲基礎之特徵參數正規化法

即傳統的整段式倒頻譜平均與變異數正規化法[1](utterance-based cepstral mean and variance normalization，U-CMVN)與分段式倒頻譜平均與變異數正規化法[2](segmental cepstral mean and variance normalization，S-CMVN)。前者是以一整段語句爲基準去估算該維特徵參數的統計值，並執行特徵參數正規化法；後者則是將每段語句以一小段的片段爲基準，去估算該片段的統計值，然後執行特徵參數正規化。

(2)以碼簿爲基礎之特徵參數補償法

爲了更精確地估測語音特徵參數統計值，以執行特徵參數補償與正規化法進而消除雜訊影響，我們提出透過虛擬雙通道碼簿，來幫助我們更準確地估算出代表訓練語音與測試語音的統計值，並藉由較準確的統計值來執行特徵參數補償，以提升辨識效率。其中包含三種方法：倒頻譜統計補償法(cepstral statistics compensation, CSC)、線性最小平方回歸法(linear least squares regression, LLS)與二次最小平方回歸法(quadratic least squares regression, QLS)。

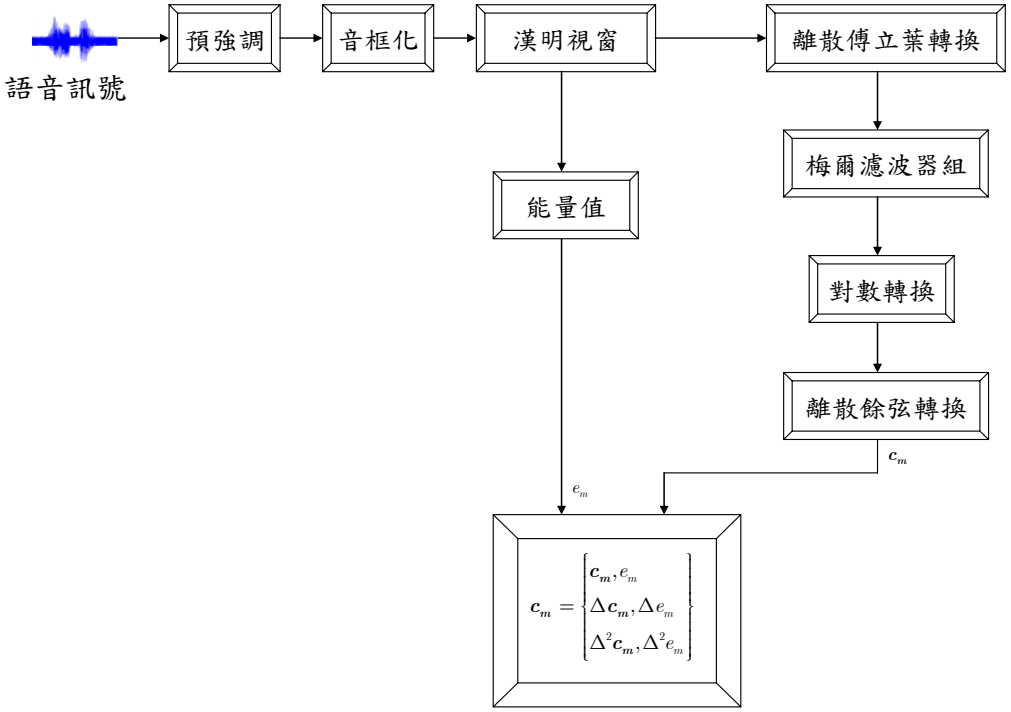
在之後的第二章裡，我們簡單介紹本論文所使用的四種語音特徵參數抽取流程。第三章介紹傳統之以段落爲基礎特徵參數正規化法，第四章之討論即爲本論文之重點：包括虛擬雙通道碼簿的建立方法，及三種以碼簿爲基礎之特徵參數補償法。第五章與第六章分別爲實驗環境介紹與實驗結果及討論。最後，第六章包含了簡要的結論。

二、各種語音訊號特徵參數抽取流程的介紹

本章節介紹在語音訊號處理中四種常用的語音特徵參數及其抽取流程，分別爲梅爾倒頻譜係數(mel-frequency cepstral coefficients，MFCC)、自相關梅爾倒頻譜係數[3](autocorrelation mel-frequency cepstral coefficients，AMFCC)、線性預測倒頻譜係數[4][5](linear prediction cepstral coefficients，LPCC)以及感知線性預測倒頻譜係數[6](perceptual linear prediction cepstral coefficients，PLPCC)。 我們將使用這四種語音特徵參數來驗證本論文所提出的強健性語音特徵參數技術，並且與其他強健性方法運用在這四種特徵參數上做比較。

(一) 梅爾倒頻譜係數 (mel-frequency cepstral coefficients，MFCC)

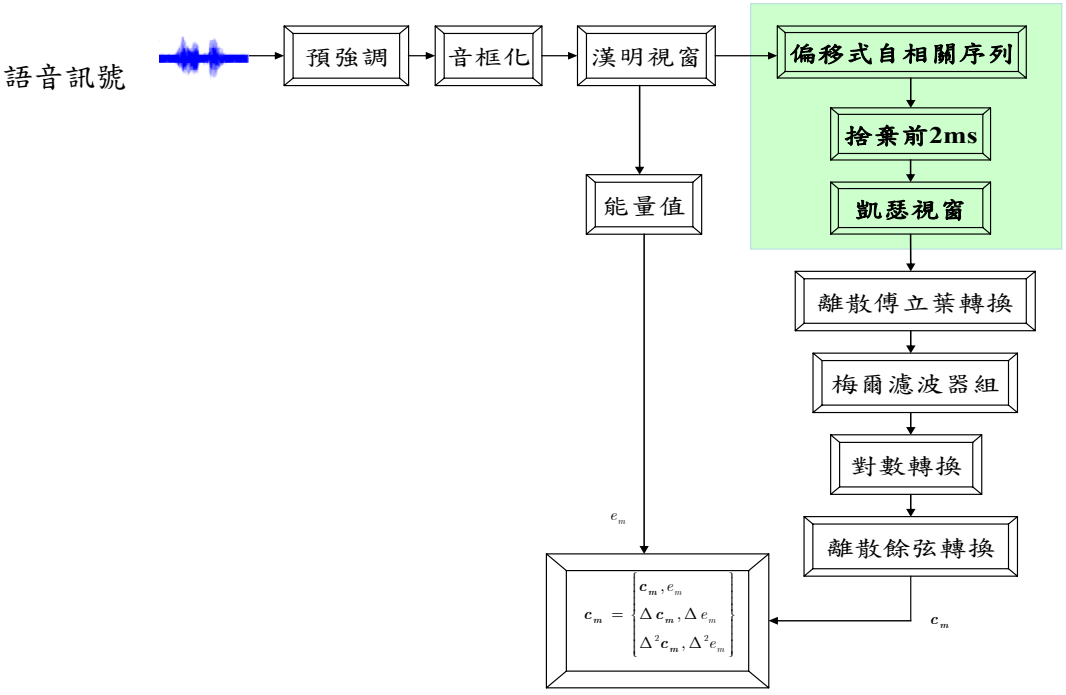
圖一爲梅爾倒頻譜係數擷取流程圖，梅爾倒頻譜係數結合了人在發音上與聽覺上的諸多性質，是目前語音研究上，最常被使用的特徵參數。



圖一、梅爾倒頻譜特徵擷取流程圖

(二) 自相關梅爾倒頻譜係數 (autocorrelation mel-frequency cepstral coefficients，AMFCC)

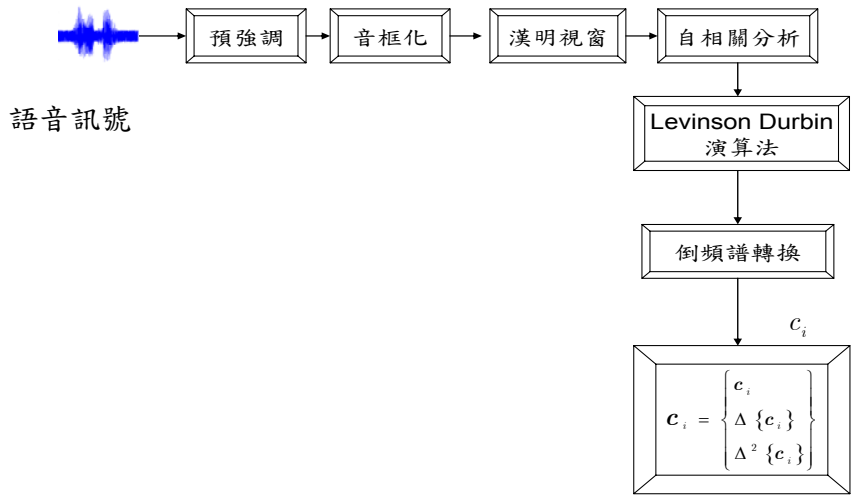
自相關梅爾倒頻譜係數是利用自相關序列結合梅爾倒頻譜係數求取步驟來做特徵參數抽取[1]，其流程即在一音框內的語音訊號經過漢明視窗處理後，取其偏移式自相關係數(biased autocorrelation coefficients)，並捨棄其前端約 2ms 係數後，再經過一凱瑟視窗以降低高頻效應。除了上述步驟外，其餘取頻譜、對數轉換及離散餘弦轉換等流程，皆與梅爾倒頻譜係數抽取流程相同。圖二為自相關梅爾倒頻譜係數之抽取流程。



圖二、自相關梅爾倒頻譜特徵擷取流程圖

(三) 線性預測倒頻譜係數(linear prediction cepstral coefficients，LPCC)

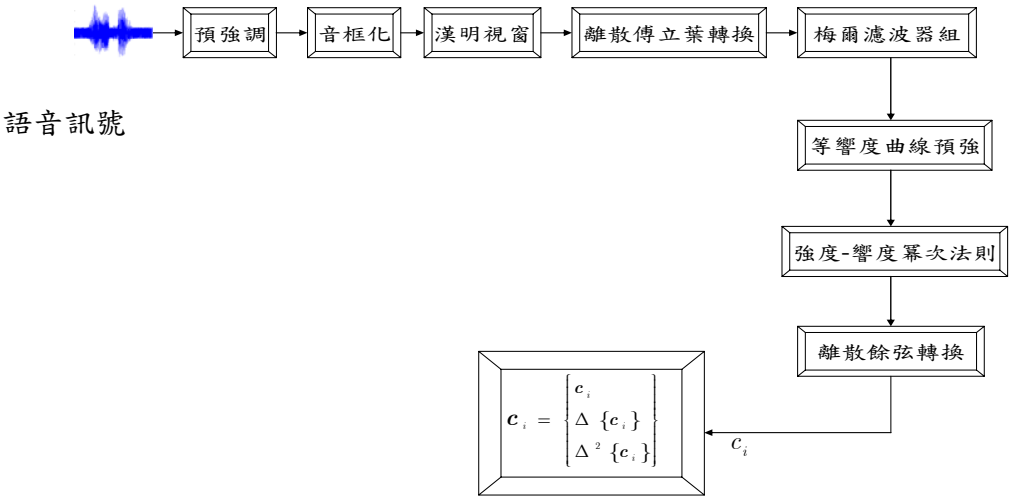
線性預測(linear prediction)的基本原理是假設目前的聲音取樣值可由在前面的 p 個取樣值，以線性組合來預測。圖三即爲線性預測倒頻譜係數之擷取流程圖。如同前兩種特徵參數擷取技術，我們將語音訊號經過預強調後，切割成許多一小段的音框與漢明視窗的處理後取其自相關係數，透過 Levinson Durbin 演算法求得線性預測係數，最後將線性預測係數轉換成倒頻譜，便得到線性預測倒頻譜係數(linear prediction cepstral coefficients，LPCC)。



圖三、線性預測倒頻譜特徵擷取流程圖

(四) 感知線性預測倒頻譜係數(perceptual linear prediction cepstral coefficients，PLPCC)

感知線性預測倒頻譜係數的擷取流程圖如圖四，與線性預測係數倒頻譜擷取流程不同之處在於：(1)它經過模擬人耳的梅爾濾波器組，對於頻譜作頻率校準(frequency warping)的處理。(2)它利用等響度曲線(equal loudness curve)對強度頻譜做預強調。(3)對於經過預強調後的強度頻譜取三次方根(cubic root)，相當於對強度與響度之間做校準(intensity-loudness warping)的動作。這些改變都是針對人的聽覺特性而做的。



圖四、感知線性預測倒頻譜係數的擷取流程圖

三、整段式與分段式之強健性語音特徵等化技術

倒頻譜平均值與變異數正規化法(CMVN)是常被使用來強健語音特徵參數的方法之一，其作法是將一連串語音資料中的每一維倒頻譜特徵參數做統計量的調整，以便消除雜訊對語音的影響；在作法上，整段式的倒頻譜平均值與變異數正規化法[1](U-CMVN)是利用一整段語音特徵求取平均值與變異數，因此所用的音框長度隨特徵係數序列長短而異，而分段式倒頻譜平均值與變異數正規化法[2](S-CMVN)的作法，是將每一維語音特徵參數，以當時的音框爲中心，對其前後數十個音框做分段統計量的計算，然後對當下的音框作正規化處理，以下將分別對這兩種方法詳加介紹。

(一) 整段式倒頻譜平均值與變異數正規化法(utterance-based cepstral mean and variance normalization，U-CMVN)

乾淨的語音訊號在經過加成性雜訊干擾後，其倒頻譜之平均值會和原本的乾淨語音倒頻譜的平均值之間會存在一個偏移量，而其變異數相對於乾淨語音特徵參數而言則會有壓縮性，因此會造成訓練與測試特徵的不匹配而降低辨識效果。而使用倒頻譜平均值與變異數等化法[1](CMVN)可將每一維倒頻譜特徵參數之平均值化爲零，並將其變異數正規化爲 1，這樣就能降低上述所謂的偏移量與壓縮性，進而提升倒頻譜參數的強健性。

整段式倒頻譜平均值與變異數等化法的作法如(式 3-1)，假設$\{Y[n], n = 1,2,\ldots,N\}$爲一由語音資料擷取所得到的某一維倒頻譜特徵參數序列，而經過整段式倒頻譜平均值與變異數等化法處理後，得到新的特徵參數$\{Y_{U-CMVN}[n], n = 1,2,\ldots,N\}$，其中的$\{Y_{U-CMVN}[n], n = 1,2,\ldots,N\}$平均值與標準差是經由整段語音的音框求取而得，如式(3-2)與式(3-3)。

$$Y_{U-CMVN}[n] = \frac{Y[n] - \mu_Y}{\sigma_Y}, \qquad n = 1,2,....,N \qquad (式\ 3-1)$$

其中

$$\mu_Y = \frac{1}{N}\sum_{n=1}^{N} Y[n] \qquad (式\ 3-2)$$

$$\sigma_Y = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Y[n] - \mu_Y)^2} \qquad (式\ 3-3)$$

(二) 分段式倒頻譜平均值與變異數正規化法(segmental cepstral mean and variance normalization，SCMVN)

如同整段式倒頻譜平均值與變異數正規化法，分段式倒頻譜平均值與變異數正規化法[2]目的亦是降低雜訊對語音的干擾，不同的是正規化時，其平均值與變異數是分段求得而非整段，如(式 3-4)、(式 3-5)與(式 3-6)。假設當第 *n* 個音框爲當時的正規化之特徵參數，則其前 *P*/2 個音框與其後 *P*/2 個音框的參數皆用以求取統計值。也就是說，長度爲 *P*+1 的視窗在一段語音特徵的時間軸上作橫移，視窗的中心點代表當時的音框，其前後的 *P*/2 個音框爲其求取統計值之片段，執行平均值與變異數正規化。至於語句中前段的特徵向量中，由於特徵向量數目少於 *P*/2 ，所以求取統計值之長度爲特徵參數的起始音框至該音框的後 *P*/2 音框；語句中後段之求

取統計值之長度亦同理可得，其表示法如圖五。

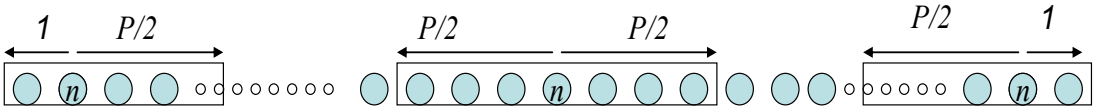
$$Y_{S-CMVN}[n]=\frac{Y[n]-\mu[n]}{\sigma[n]},,\qquad n=1,2,....,N \tag{式 3-4}$$

其中

$$\mu[n]=\frac{1}{P+1}\sum_{i=n-\frac{P}{2}}^{n+\frac{P}{2}}Y[i] \tag{式 3-5}$$

$$\sigma[n]=\sqrt{\frac{1}{P+1}\sum_{i=n-\frac{P}{2}}^{n+\frac{P}{2}}(Y[i]-\mu[n])^2} \tag{式 3-6}$$

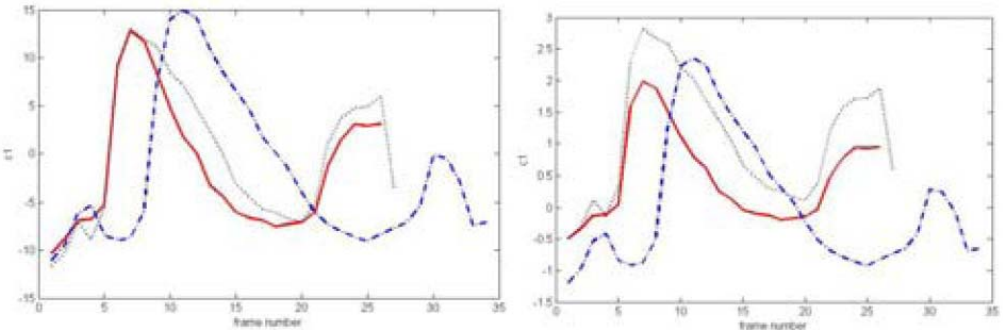
其中 $P+1$ 爲正規化片段的長度。



圖五、特徵參數經分段正規化視窗處理的示意圖

(三) 討論

本章所介紹兩種語音特徵參數正規化法目的都是希望藉由正規化倒頻譜統計值，來降低訓練語音與測試語音之間的不匹配。其中，整段式倒頻譜平均值與變異數正規化法(U-CMVN)是以一整段語音的所有音框，去計算特徵參數的統計值，執行上比較簡單，不過此方法存在以下幾個缺點：(1)在語音特徵參數擷取流程中，在得到最後一個音框前，我們無法執行 U-CMVN。因爲它是以一整段語調中所有特徵參數的音框，當做統計值的計算與進行正規化的基礎，所以 U-CMVN 是無法以線上方式(on-line manner)執行的。所謂線上方式(on-line manner)，即我們在擷取每一段倒頻譜特徵參數後，以即時(real-time)的方式計算出屬於該特徵參數之統計值，並執行正規化的動作。(2)一段語音的音框數通常是影響統計值正確性之因素，但我們無法控制原始一整段語音的長短。(3)因爲不同聲學單位(acoustic units)的長度或總數量，在不同段語音之間會有變化，所以一段語音中同一個聲學單位被正規化後的特徵參數，在另一段語音中可能會不同。圖六(a)是從 AURORA2 中乾淨的訓練語料庫中，三段不同語音"FAC_1911446"，"FAC_1473533A"與"FAC_1O1"擷取出聲學單位"one"之原始第一維倒頻譜參數 c1 之輪廓；圖六(b)則是圖六(a)經過 U-CMVN 處理後的版本。從圖六(a)可發現未經 CMVN 處理的 c1，其輪廓在([-10.3，12.8]，[-11.6，13.0]及[-11.0，15.0])這三個範圍內有相似的分佈狀況；反觀圖六(b)，因爲不同語音中的特徵參數是被不同的平均值與變異數作正規化，使正規化後三個 c1 的輪廓變得不太相同，它們的分佈狀況在([-0.5，2.0]，[-0.5，2.8]及[-1.1，2.3])，這三個範圍跟之前相比下是比較不同的。



(a) (b)

圖六：AURORA2 中乾淨的訓練語料庫中，三段不同語音"FAC_1911446"，"FAC_1473533A"與"FAC_1O1"擷取出聲學單位"one"之(a)原始第一維梅爾倒頻譜特徵 c1 輪廓(b)經 U-CMVN 處理後第一維梅爾倒頻譜特徵 c1 輪廓

另一方面，分段式倒頻譜平均值與變異數正規化法，它是以移動一固定長的分段視窗來計算正規化每個音框所要用到的統計值，並作統計值正規化法。假如我們使得該分段視窗夠短的話，在執行上它是可以較接近於線上方式(on-line manner)的。再者，因爲在一個短分段中的聲學單位總數目相對而言較少，所以使用分段式倒頻譜平均值與變異數正規化法，可降低相同聲學單位在不同段語音間的特徵參數變異性。而根據之後的實驗結果顯示，使用分段式倒頻譜平均值與變異數正規化法(S-CMVN)在本論文中所用之四種語音特徵參數上(MFCC、AMFCC、LPCC、PLPCC)，所得到的辨識效果，的確可比整段式倒頻譜平均值與變異數正規化法(U-CMVN)效果來得好。這也間接證明了，以分段式進行特徵參數統計值的估算，相對於整段式來的準確。而以前者估算出的統計值，進行特徵參數統計值正規化法所得到之特徵參數也能更加降低雜訊對語音的影響。

四、運用虛擬雙通道碼簿爲基礎之雜訊強健技術的介紹

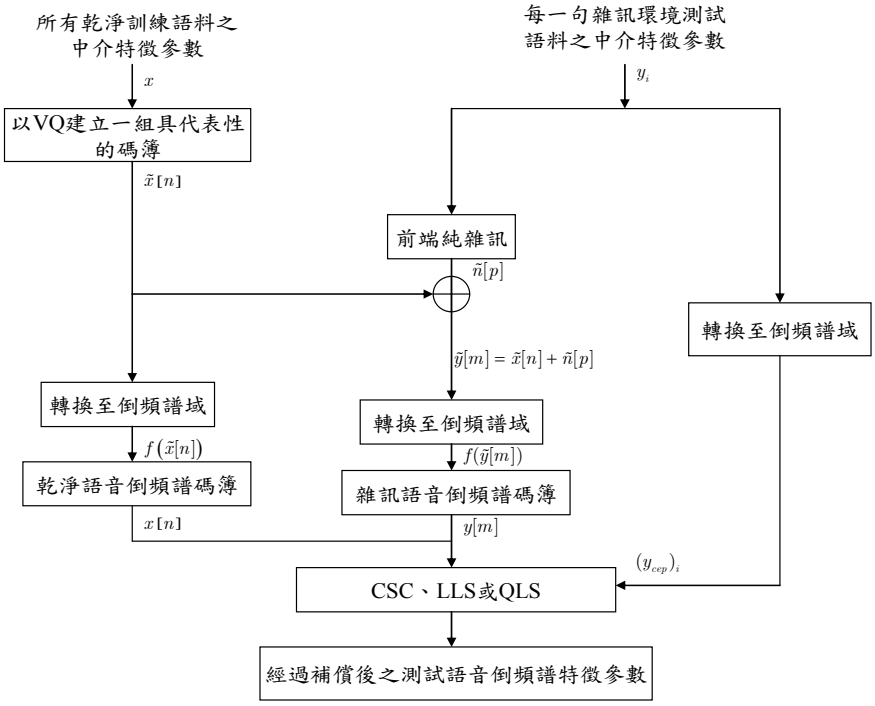
本論文所提出三種特徵參數補償法語音強健技術，依序爲倒頻譜統計補償法[7](cepstral statistics compensation，CSC)、線性最小平方回歸法[7](linear least squares regression，LLS)以及二次方最小平方回歸法[7](quadratic least squares regression，QLS)。在執行這幾個特徵參數補償法前，我們先訓練兩組分別代表乾淨語音與雜訊語音的碼簿(codebooks)，我們稱之爲虛擬雙通道碼簿(pseudo stereo codebooks)。藉由這兩組碼簿的使用，我們得以發展上述三種特徵參數補償法。

運用所謂的虛擬雙通道碼簿來計算乾淨語音與含雜訊語音之統計值，進而執行三種減低雜訊的技巧是簡單而又有效率的。倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)以及二次方最小平方回歸法(QLS)，這三種特徵參數統計量補償法的概念就是對含雜訊之語音倒頻譜係數做轉換(transformation)，使得經過轉換後的語音倒頻譜其統計值更相似於乾淨訓練語音倒頻譜的統計值，進而提升語音辨認強健性。

(一) 虛擬雙通道碼簿之建立方法

本論文中四種語音特徵參數擷取流程的虛擬雙通道碼簿之建立方法，其步驟如圖七所示，首先

必須在擷取所有用以訓練之乾淨語音訊號的 MFCC、AMFCC、LPCC、PLPCC 不同的倒頻譜特徵參數中，保留下具有語音與雜訊爲線性相加特性的中介特徵參數(intermediate feature)，將這些乾淨語音的中介特徵參數訓練成一組碼簿，接著將此乾淨語音碼簿中所有的碼字與測試語音中所選取的一段純雜訊之中介特徵參數作線性相加，如此便可得到兩組分別代表乾淨語音與測試語音在該中介特徵參數域中的碼簿，最後將這兩組中介特徵參數的碼簿轉到倒頻譜域中，以利之後特徵補償演算法的進行，此兩組分別代表乾淨語音與雜訊語音的倒頻譜特徵參數碼簿，我們稱之爲虛擬雙通道碼簿 (pseudo stereo codebooks)。



圖七：虛擬雙通道碼簿之建立架構及以虛擬雙通道碼簿執行特徵參數補償法之流程圖

虛擬雙通道碼簿之建立過程詳述如下：

首先將語料庫中所有乾淨語料的每一段語音，透過特徵參數擷取流程轉換成一序列的中介特徵向量，如表一所述。這些由所有乾淨語料的語調所得到的中介特徵向量，透過向量量化(vector quantization，VQ)後可建立成一組包含 N 個碼字(codewords)的集合，以 $\{\tilde{\mathbf{x}}[n], 1 \le n \le N\}$ 表示。這組在中介特徵參數域上的乾淨語音碼簿中所有碼字，都可經過剩下的特徵參數擷取步驟轉換至倒頻譜域，如(式 4-1)所示：

$$\mathbf{x}[n] = f(\tilde{\mathbf{x}}[n]) \quad , \qquad \text{(式 4-1)}$$

表一：本論文中所使用之四種語音特徵參數，及其具備語音與雜訊線性相加特性之中介特徵參數。

倒頻譜特徵參數型態	具備語音與雜訊線性相加之中介特徵參數
梅爾倒頻譜係數(MFCC)	梅爾頻譜(mel-spectrum)
自相關梅爾倒頻譜係數 (AMFCC)	梅爾頻譜(mel-spectrum)
線性預測倒頻譜係數 (LPCC)	自相關係數(autocorrelation coefficients)、強度頻譜 (magnitude spectrum)
感知線性預測倒頻譜係數 (PLPCC)	自相關係數(autocorrelation coefficients)、強度頻譜 (magnitude spectrum)

其中 $f(.)$ 爲轉換函數，它是隨著我們所選擇的特徵參數型態而不同。因此 $\{\mathbf{x}[n],1 \le n \le N\}$ 這組經轉換至倒頻譜域的碼簿，即稱之爲乾淨語音的倒頻譜碼簿。

至於含雜訊的測試語音方面，因爲要完全以每段簡短的測試語音爲基礎，去建立一組可靠的碼字是很困難的，所以我們試著藉由乾淨語音在中介特徵參數域上的碼字，來建立對應至該段的含雜訊之測試語音的碼簿。步驟如下：

對於一段測試語音，我們假設估計到的純雜訊在中介特徵參數域上可用一組向量來代表，以 $\{\tilde{\mathbf{n}}[p],1 \le p \le P\}$ 表示。因爲乾淨語音與雜訊在中介特徵參數域上是近似線性相加的，因此含雜訊語音的碼字可表示成(式 4-2)：

$$\tilde{\mathbf{y}}[m]\big|_{m=(n-1)P+p} = \tilde{\mathbf{x}}[n] + \tilde{\mathbf{n}}[p] \qquad \text{(式 4-2)}$$

接著我們將 $\tilde{\mathbf{y}}[m]$ 經過剩下的特徵參數擷取步驟轉換至倒頻譜域，如(式 4-3)所示：

$$\mathbf{y}[m] = f(\tilde{\mathbf{y}}[m]) \qquad \text{(式 4-3)}$$

因此，$\{\mathbf{y}[m],1 \le m \le NP\}$ 這組碼字即代表雜訊語音在倒頻譜域上的碼簿。$\{\mathbf{x}[n]\}$ 與 $\{\mathbf{y}[m]\}$ 這兩組碼字可分別代表乾淨訓練語音與雜訊測試語音，我們稱之爲虛擬雙通道碼簿(pseudo stereo codebooks)。所謂"虛擬"，即因爲雜訊語音的碼簿並非直接由雜訊語音得到的，而是透過結合乾淨語音碼簿與雜訊估算值所得到的。值得注意的是，我們在建立乾淨語音碼簿 $\{\mathbf{x}[n]\}$ 時是一次將語料庫中所有乾淨訓練語音做處理，這是屬於非線上方式(off-line manner)的處理。不過，當輸入每一段不同的測試語音，或雜訊環境改變時，雜訊語音碼簿 $\{\mathbf{y}[m]\}$ 必須隨之更新。因爲雜訊估算值 $\tilde{\mathbf{n}}[p]$ 可粗略地以每一段測試語音的前幾個音框得到，因此雜訊語音的碼簿 $\{\mathbf{y}[m]\}$ 可以在一個幾乎爲線上運算方式(on-line manner)，即在不會有太長的延遲時間的運算情況下建立。

在本論文中，我們以虛擬雙通道碼簿爲基礎來執行三種特徵參數補償法，以降低加成性雜訊的影響。以下，我們對三種特徵參數補償法做完整的介紹。

(二) 倒頻譜統計補償法(cepstral statistics compensation，CSC)

我們利用虛擬雙通道碼簿，可以算出分別代表乾淨語音與雜訊語音的統計值，如(式 4-4)、(式 4-5)所示：

$$\mu_{x,i} \approx \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}[n])_i, \sigma_{x,i}^2 \approx \frac{1}{N}\sum_{n=1}^{N}[(\mathbf{x}[n])_i - \mu_{x,i}]^2 \qquad \text{(式 4-4)}$$

$$\mu_{y,i} \approx \frac{1}{NP}\sum_{m=1}^{NP}(\mathbf{y}[m])_i, \sigma_{y,i}^2 \approx \frac{1}{NP}\sum_{m=1}^{NP}[(\mathbf{y}[m])_i - \mu_{y,i}]^2 \quad \text{(式 4-5)}$$

其中$(\mathbf{v})_i$代表一個任意向量$\mathbf{v}$第 i 維成份，$\mu_{x,i}$與$\sigma_{x,i}^2$分別代表乾淨語音特徵向量$\mathbf{x}$第 i 維的平均值與變異數；$\mu_{y,i}$與$\sigma_{y,i}^2$分別代表雜訊語音特徵向量$\mathbf{y}$第 i 維的平均值與變異數。以這些統計值來執行倒頻譜統計值補償法，我們轉換每一段雜訊語音之倒頻譜向量，如(式 4-6)：

$$(\mathbf{z})_i = \frac{\sigma_{x,i}}{\sigma_{y,i}} \times [(\mathbf{y})_i - \mu_{y,i}] + \mu_{x,i} \quad \text{(式 4-6)}$$

在理想的情況下，$(\mathbf{z})_i$與乾淨語音特徵向量$(\mathbf{x})_i$會有相同的平均值與變異量，由於雜訊語音倒頻譜的某些統計值被補償，使得補償過後的雜訊語音倒頻譜其統計值是近似於乾淨語音倒頻譜的統計值，因此我們將此方法稱爲倒頻譜統計補償法(cepstral statistics compensation，CSC)。我們可以用矩陣的形式改寫(式 4-6)的倒頻譜統計補償演算法，如(式 4-7)：

$$\mathbf{z} = \mathbf{\Psi}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}) + \boldsymbol{\mu}_\mathbf{x} \quad \text{(式 4-7)}$$

其中$\boldsymbol{\mu}_\mathbf{x} = [\mu_{\mathbf{x},1}, \mu_{\mathbf{x},2}, \cdots]^T, \boldsymbol{\mu}_\mathbf{y} = [\mu_{\mathbf{y},1}, \mu_{\mathbf{y},2} \cdots]^T$，$\mathbf{\Psi}$是一對角線爲$\{\sigma_{\mathbf{x},i} / \sigma_{\mathbf{y},i}\}$之對角矩陣(diagonal matrix)。

事實上，CSC 的概念是類似傳統的倒頻譜平均與變異數正規化法(CMVN)，因爲這兩種演算法的目的都是希望訓練語音與測試語音能得到相似的統計值。不過 CSC 擁有下列幾項優點：
(1)CSC 可以一個幾乎爲線上方式(on-line manner)的處理程序來執行，因爲乾淨語音的碼簿是事先建立好的，而在建立雜訊語音碼簿時所需的雜訊估算值，通常可以在每段測試語音的前幾個音框來得到。
(2)在 CSC 中，統計量是利用訓練語料庫中所有訓練語音所建立的碼簿所得；但在 CMVN 中，只利用單一語句去決定平均值與變異量。因此，我們可以預期碼簿幫助我們求得更準確的特徵參數統計值。
(3) 在 CSC 中，相同雜訊環境下不同語句的特徵參數接受相同的轉換，這使得不同語句之間的特徵參數，在對應至相同的聲學單位時，能保持特徵相似度。這是 CMVN 無法做到的，我們在上一章的圖六已做了說明。

(三) 線性最小平方回歸法(linear least squares regression，LLS)與 二次最小平方回歸法(quadratic least squares regression，QLS)

在這裡，線性最小平方回歸法與二次最小平方回歸法都是屬於多項式回歸法(polynomial regression approaches)，其概念就是希望雜訊語音的碼簿，在透過一個轉換函數的運算後能和乾淨語音碼簿的整體距離是最小的，如此我們便可預期，當雜訊語音倒頻譜經過相同轉換後，會更接近乾淨語音倒頻譜。以下我們做詳細的介紹。

在前面的介紹中，我們知道每個雜訊語音碼字$\mathbf{y}[m]$對應的乾淨語音碼字爲$\mathbf{x}[n]$，其中$n = \lceil m/P \rceil$($\lceil . \rceil$表示無條件進位運算，P爲純雜訊的向量數目)，$\{\mathbf{x}[n]\}$與$\{\mathbf{y}[m]\}$這兩組碼字分別代表乾淨語音與雜訊語音倒頻譜$\mathbf{x}$與$\mathbf{y}$。若我們能對每一個雜訊語音碼字$\mathbf{y}[m]$找到一個轉換函數$\mathcal{T}(\cdot)$，使得$\mathcal{T}(\mathbf{y}[m])$與$\mathbf{x}[n]$之間的整體距離是最小的，那我們可以合理的預期雜訊語音倒頻譜$\mathbf{y}$經轉換後$\mathcal{T}(\mathbf{y})$，會更接近乾淨語音倒頻譜$\mathbf{x}$。爲了簡單起見，我們假設轉移函數是執行在$\mathbf{y}$的每一維上。假設$\mathcal{T}_i(\bullet)$是$\mathbf{y}$的第 i 維成份的轉移函數，則定義一目標函數J_i將使得$\mathcal{T}_i\left((\mathbf{y}[m])_i\right)$與

$(\mathbf{x}[n])_i$的整體平方距離最小，如(式 4-8)：

$$J_i = \sum_{m=1}^{NP} \left[\mathcal{T}_i\left((\mathbf{y}[m])_i\right) - (\mathbf{x}[n])_i \right]^2 \quad \text{(式 4-8)}$$

其中$n = \lceil m/P \rceil$，假設$\mathcal{T}_i(\bullet)$是一個 K 次多項式，則(式 4-8)中以處理$\mathcal{T}_i(\bullet)$來最小化J_i，就變成一個典型的最小化平方(least squares)的問題，如(式 4-9)：

$$\mathcal{T}_i(u) = a_K^{(i)} u^K + a_{K-1}^{(i)} u^{K-1} + \cdots + a_0^{(i)} \quad \text{(式 4-9)}$$

(式 4-8)中的目標函數可以改寫成向量矩陣的形式，如(式 4-10)：

$$J_i = \left\| \mathbf{Y}_i \mathbf{a}_i - \mathbf{b}_i \right\|^2 \quad \text{(式 4-10)}$$

其中矩陣$\mathbf{Y}_i$的第(*m*,*n*)項如(式 4-11)所示：

$$(\mathbf{Y}_i)_{mn} = [(\mathbf{y}[m])_i]^{K-n+1}, 1 \le m \le NP, 1 \le n \le K+1 \quad \text{(式 4-11)}$$

且 $\mathbf{a}_i = [a_K^{(i)} \quad a_{K-1}^{(i)} \cdots \quad a_0^{(i)}]^T$ ，

$$\mathbf{b}_i = \left[\left(\mathbf{x}\left[\left[1/P\right]\right]\right)_i \left(\mathbf{x}\left[\left[2/P\right]\right]\right)_i \cdots \left(\mathbf{x}\left[\left[NP/P\right]\right]\right)_i \right]^T \text{。}$$

多項式$\mathcal{T}_i(\bullet)$中最小化J_i的係數向量$\mathbf{a}_i$即爲最小平方解，如下(式 4-12)：

$$\hat{\mathbf{a}}_i = \left(\mathbf{Y}_i^T \mathbf{Y}_i\right)^{-1} \mathbf{Y}_i^T \mathbf{b}_i \quad \text{(式 4-12)}$$

值得注意的是，多項式$\mathcal{T}_i(\bullet)$的次數 K 不可以設太大，以避免有過度擬合(over-fitting)情況或不良狀況的矩陣(ill-conditional matrix) $\mathbf{Y}_i^T \mathbf{Y}_i$產生。因此，我們只考慮 K = 1 與 K = 2 兩種情況：當 K = 1 時，轉移函數$\mathcal{T}_i(\bullet)$是一個線性函數，我們稱之爲線性最小平方回歸法(linear least squares regression, LLS)。當 K = 2 時轉移函數$\mathcal{T}_i(\bullet)$即爲一個二次函數，我們稱之爲二次最小平方回歸法(quadratic least squares regression, QLS)。

如同本節一開始所提到，用這兩種多項式回歸法的概念就是希望雜訊語音的碼簿，在透過一個轉換函數的運算後能和乾淨語音碼簿的整體距離是最小的，當雜訊語音倒頻譜經過相同轉換後會更接近乾淨語音倒頻譜，如此便可提升辨識效果。

五、實驗設定

(一) 語音資料庫簡介

本論文所使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute，ETSI)發行的 AURORA2 語音資料庫，它是一套連續的英文數字字串，內容是以美國成年男女所錄製的乾淨環境連續數字，再加上雜訊與通道效應。加成性雜訊共有八種，分別爲地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，前四種歸類爲 Set A，後四種歸類爲 Set B。

訊雜比(signal-to-noise ratio, SNR)則有七種，分別爲 20dB, 15dB, 10dB, 5dB, 0dB. -5dB 與完全乾淨狀態。

(三) 特徵參數的設定與辨識系統的訓練

本論文共使用四種特徵參數分別爲梅爾倒頻譜係數(MFCC)、自相關梅爾倒頻譜係數

(AMFCC)、線性預測倒頻譜係數(LPCC)及感知線性預測倒頻譜係數(PLPCC)，其相關設定與個別對應之中介特徵參數，如表二所示。對於每個欲辨識的數字模型而言，本論文使用隱藏式馬可夫模型工具(hidden Markov model toolkit，HTK)來訓練，包含 11 個數字模型(0~9 以及 oh 11 個數字模型)以及靜音模型，每個數字模型包含 10 個狀態，各狀態包含 4 個高斯密度混合。隱藏式馬可夫模型是一種運用統計理論推導出來的模型，用來描述語音產生的過程，相當適合用在連續語音的辨認。HMM 有很多種類型，本論文採用由左到右的形式，也就是每個狀態在下一個時間只能跳到此刻狀態或下一個鄰近的狀態，隨著時間的增加，狀態由左至右依序轉移。另外，模型中的狀態觀測機率函數是選用連續式的高斯混合機率密度函數(Gaussian Mixture probability density function，簡稱 GM)，因此我們也稱此模型爲連續密度隱藏式馬可夫模型(continuous density HMM，簡稱 CDHMM)。

表二、實驗中所用的特徵參數詳細資料

特徵參數種類	特徵參數維度	中介特徵參數維度
MFCC	12 維倒頻譜加上 1 對數能量維，並取其一階和二階差量，總共 39 維特徵參數。	23 維梅爾頻譜加上 1 對數能量維。
AMFCC	12 維倒頻譜加上 1 對數能量維，並取其一階和二階差量，總共 39 維特徵參數。	23 維梅爾頻譜加上 1 對數能量維。
LPCC	13 維倒頻譜，並取其一階和二階差量，總共 39 維特徵參數。	23 維強度頻譜， 或是 24 維自相關係數。
PLPCC	13 維倒頻譜，並取其一階和二階差量，總共 39 維特徵參數。	23 維強度頻譜，或是 23 維自相關係數。

(四) 強健性特徵參數技術實驗設定

在分段式平均值與變異數正規化法(S-CMVN)，我們令式 3-5 與式 3-6 中使用的分段長度爲 P+1 = 101 個音框，即大約爲 1 秒的長度。

虛擬雙通道碼簿的建立方法當中，乾淨語音碼簿爲 $\{\tilde{\mathbf{x}}[n],1 \le n \le N\}$，其中 N 值我們分別設爲 32、64、128、256、512、1024。在實驗結果中，我們將只呈現得到最佳辨識率時的 N 值之整體實驗數據。

對於純雜訊的估測值 $\{\tilde{\mathbf{n}}[p]\}$，我們是以在中介特徵參數域上，每一段測試語音的前 5 個音框當作該段語音的純雜訊音框。

在 LPCC 與 PLPCC 兩種特徵參數擷取過程裡，因爲其具備語音與雜訊爲線性相加的中介特徵參數有兩種，分別爲強度頻譜(magnitude spectrum)與自相關係數(autocorrelation coefficients)，因此在實驗結果中我們以 MS-與 AC-分別代表之。

六、實驗結果與分析

首先，表三與表四分別爲整段式倒頻譜平均與變異數正規化法(U-CMVN)與分段式倒頻譜

平均與變異數正規化法(S-CMVN)的辨識精確度，相對於原始未處理的各種倒頻譜特徵而言，U-CMVN 與 S-CMVN 皆能有效提升各種雜訊環境下的辨識率，這意謂這兩種方法的確具有提升特徵參數強健性的效能，而當我們將表四與表三的數據比較，可明顯看出 S-CMVN 相對於 U-CMVN 在辨識率能有更明顯的提升。這與我們之前分析的結果相吻合，即利用分段的方式估測特徵參數的統計值能比利用整段的方式更精確。

接下來，我們探討本論文所提出的三種以碼簿爲基礎的特徵參數補償法的效果，表五、表六與表七分別爲倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)與二次最小平方回歸法(QLS)的辨識精確度。爲了比較起見，我們將表四之 S-CMVN 的結果亦列於各表中。從這三個表的數據可知：

表三、整段式倒頻譜平均與變異數正規化法之辨識精確度(%)

Set A	subway	babble	car	exhibition	average	baseline
MFCC	72.91	69.71	68.71	69.22	70.14	61.99
AMFCC	68.60	72.02	68.94	65.58	68.78	65.52
LPCC	72.25	71.32	71.50	70.18	71.31	51.26
PLPCC	75.24	74.34	73.72	74.60	74.48	57.38
Set B	restaurant	street	airport	train station	average	baseline
MFCC	71.60	72.16	71.00	68.28	70.76	55.78
AMFCC	72.89	71.23	73.35	70.23	71.93	59.43
LPCC	73.44	72.82	74.68	70.69	72.91	49.58
PLPCC	76.48	75.79	77.01	73.23	75.63	54.51

表四、分段式倒頻譜平均與變異數正規化法之辨識精確度(%)

Set A	subway	babble	car	exhibition	average	U-CMVN
MFCC	75.71	73.42	72.36	72.63	**73.53**	70.14
AMFCC	72.12	74.82	73.49	70.21	**72.66**	68.78
LPCC	74.53	74.20	75.30	74.38	**74.60**	71.31
PLPCC	76.07	75.14	75.72	76.06	**75.75**	74.48
Set B	restaurant	street	airport	train station	average	U-CMVN
MFCC	75.65	75.23	74.97	71.93	**74.45**	70.76
AMFCC	75.58	76.11	75.02	72.14	**74.71**	71.93
LPCC	76.08	76.63	76.87	73.84	**75.86**	72.91
PLPCC	77.69	77.48	78.21	75.13	**77.13**	75.63

1. 相對於原始未處理的倒頻譜參數（數據列於表三）而言，這三種新的特徵參數補償法都能夠大幅提昇辨識精確度，意謂各種不同的特徵參數都能藉由這三種方法而提升其強健性。
2. 在大部分的情形下，這三種新的特徵參數補償法的表現都優於 S-CMVN 與 U-CMVN，這呼應了我們之前的推論：利用碼簿來估測特徵參數的統計值相較於利用整段或分段的方式估測更來的精確。
3. 雖然這三種特徵參數補償法作用於四種特徵參數上的實驗結果，所得到最佳辨識率時的 N 值都不相同，不過若個別觀察個特徵參數的實驗結果，可發現其具有規則性。如在特徵參數爲 MFCC 時，三種特徵參數補償法之最佳實驗結果都在 N 值爲 512 或 256 這些比較中段

的值；而特徵參數爲 LPCC 時則在 N 值爲較大值 1024 時，可得到最佳辨識率。

4. 一般而言，倒頻譜統計補償法的效果優於線性最小平方回歸法與二次最小平方回歸法，然而，其表現的差異並沒有十分明顯。

表五、倒頻譜統計補償法(CSC)之辨識精確度(%)，其中 N 表示乾淨碼簿的碼字數

Set A	subway	babble	car	exhibition	average	S-CMVN
MFCC (N=512)	78.71	75.84	80.54	77.40	78.12	73.53
AMFCC (N=512)	79.11	74.26	82.80	75.91	78.02	72.66
MS-LPCC (N=1024)	75.81	72.31	82.38	74.62	76.28	74.60
AC-LPCC (N=1024)	74.94	75.71	81.19	74.54	76.60	
MS-PLPCC (N=128)	77.79	76.14	81.25	78.71	78.47	75.75
AC-PLPCC (N=32)	78.64	76.57	78.55	77.70	77.87	
Set B	restaurant	street	airport	train station	average	S-CMVN
MFCC (N=512)	75.08	77.82	77.15	77.77	76.95	74.45
AMFCC (N=512)	73.15	79.36	77.28	79.10	77.22	74.71
MS-LPCC (N=1024)	73.57	76.55	78.48	79.16	76.93	75.86
AC-LPCC (N=1024)	76.35	75.95	80.24	79.61	78.04	
MS-PLPCC (N=128)	77.24	77.47	80.40	79.55	78.66	77.13
AC-PLPCC (N=32)	76.76	77.06	79.90	78.07	77.95	

表六、線性最小平方回歸法(LLS)之辨識精確度(%)，其中 N 表示乾淨碼簿的碼字數

Set A	subway	babble	car	exhibition	average	S-CMVN
MFCC (N=512)	78.92	76.09	80.01	76.57	77.90	73.53
AMFCC (N=512)	76.97	75.06	80.35	73.46	76.46	72.66
MS-LPCC (N=1024)	72.85	75.23	79.79	71.18	74.77	74.60
AC-LPCC (N=1024)	71.70	77.48	78.22	70.91	74.58	
MS-PLPCC (N=64)	79.74	77.70	81.37	79.22	79.51	75.75
AC-PLPCC (N=64)	76.63	77.15	76.02	75.93	76.43	
Set B	restaurant	street	airport	train station	average	S-CMVN
MFCC (N=512)	78.89	78.19	77.72	77.62	77.35	74.45
AMFCC (N=512)	74.45	77.55	77.29	76.95	76.56	74.71
MS-LPCC (N=1024)	75.29	74.91	79.42	77.97	76.90	75.86
AC-LPCC (N=1024)	77.82	74.32	80.66	78.12	77.73	
MS-PLPCC (N=64)	78.71	78.91	81.17	80.44	79.81	77.13
AC-PLPCC (N=64)	77.79	75.56	79.40	76.69	77.36	

七、結論與未來展望

本論文提出三種以虛擬雙通道碼簿爲基礎的特徵參數補償法，分別爲倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)與二次最小平方回歸法(QLS)，個別作用於四種語音特徵參數：梅爾倒頻譜係數(MFCC)、自相關梅爾倒頻譜係數(AMFCC)、線性預測倒頻譜係數(LPCC) 與感知線性預測倒頻譜係數(PLPCC)上。我們發現，以虛擬雙通道碼簿爲基礎之特徵參數補償法，

表七、二次最小平方回歸法(QLS)之辨識精確度(%)，其中 N 表示乾淨碼簿的碼字數

Set A	subway	babble	car	exhibition	average	S-CMVN
MFCC (N=256)	77.71	76.44	82.60	77.38	78.53	73.53
AMFCC (N=512)	72.61	76.94	81.08	71.02	75.41	72.66
MS-LPCC (N=1024)	69.82	74.99	80.80	71.04	74.16	74.60
AC-LPCC (N=1024)	67.35	75.45	77.81	69.54	72.54*	
MS-PLPCC (N=64)	76.83	76.74	82.80	77.94	78.58	75.75
AC-PLPCC (N=512)	71.48	73.05	74.29	72.48	72.82*	
Set B	restaurant	street	airport	train station	average	S-CMVN
MFCC (N=256)	73.88	77.19	77.70	79.18	76.99	74.45
AMFCC(N=512)	74.47	75.50	78.95	78.72	76.91	74.71
MS-LPCC (N=1024)	74.75	72.06	79.65	78.55	76.25	75.86
AC-LPCC (N=1024)	75.28	70.17	78.92	76.97	75.34*	
MS-PLPCC (N=64)	76.53	76.60	81.22	81.19	78.89	77.13
AC-PLPCC (N=512)	72.61	70.62	76.19	74.24	73.41*	

以線上方式即時地估算出雜訊語音特徵參數統計值，所估算出的統計值較爲準確，也使得執行特徵參數補償法後語音特徵參數更爲強健。相對於傳統特徵參數正規化法是以整段或分段語句爲基礎去估算語音特徵參數的統計值後，而執行特徵參數正規化，以虛擬雙通道碼簿爲基礎的特徵參數補償法，更能降低雜訊對語音的影響。

本論文只著重於加成性雜訊環境下的研究，因此在未來，我們期望能以虛擬碼簿爲基礎的強健性語音技術，藉由結合一些通道補償技巧如：相對頻譜法(RASTA) [8]，使這些虛擬碼簿爲基礎的強健性語音技術能延伸於消除通道失真的效應上。

八、參考文獻

[1] S. Tiberewala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition", Eurospeech97, 1997, pp. 107-110

[2] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization", in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels, Pont-a-Mousson, France, 1997, pp.107-110.

[3] Benjamin J. Shannon, Kuldip K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition", Speech Communication 2006.

[4] Atal, B.S. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", Journal of the Acoustical Society of America, 1974.

[5] J. Makhoul, "Spectral linear prediction: properties and applications," IEEE Transactions on Acoustics, Speech and Signal Processing, 1975.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[7] Jeih-weih Hung, "Cepstral statistics compensation using online pseudo stereo codebooks for robust speech recognition in additive noise environments", ICASSP 2006.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, 2, pp.578-589, 1994

Cyberon Voice Commander 多國語言語音命令系統

何泰軒、劉進榮
賽微科技股份有限公司
Cyberon Corporation
tai@cyberon.com.tw, AlexLiou@cyberon.com.tw

摘要

Cyberon Voice Commander (CVC) 是賽微科技自行研發的多國語言版本手機聲控軟體，應用於 Windows Mobile, WinCE, Symbian 等智慧型手機中，提供使用者語音撥號、語音指令、語音點歌、Text-to-Speech (TTS) 朗讀簡訊、電子郵件、和行事曆內容等功能。CVC 能支援 24 種語言版本，目前已有超過 30 款手機內建 CVC 銷售全球。本文描述 CVC 所使用的語音辨識和 TTS 技術、支援多國語言的作法、幾種語言的語音辨識實驗結果、以及我們對 CVC 未來改良的想法。

關鍵詞：語音辨識、語音合成、TTS、語音撥號、語音命令、多國語言語音技術

一、緒論

隨著來無線通訊、半導體技術的快速進步，以及手機製造商建立起完善的全球供應鏈大幅降低製造成本，促成了手機產業的快速發展，2006 年全球手機出貨超過 10 億台，Nokia 預估今年全球手機用戶數亦將突破 30 億，2010 年可達 40 億用戶，手機已成爲大部分現代人生活中不可或缺的裝置之一。

近年來語音聲控功能在手機上的重要性逐漸增加。小型化、攜帶方便爲多年來手機設計不變的趨勢，手機按鍵越來越少、螢幕大小有所限制，造成用手操作手機不夠方便，例如很多使用者電話簿中都有上百筆人名資料，用按鍵一筆一筆地瀏覽尋找聯絡人是非常沒有效率的，應用語音辨識技術就能很快的幫助使用者找到想查詢的聯絡人資料。另外，大部份人在開車中不可避免地會撥打手機，但這是相當危險的舉動，不僅危害自身安全，也帶給其他用路人威脅，爲避免撥打手機而造成交通事故，許多國家已經訂定嚴格的法規禁止開車時用手撥打手機，語音撥號功能就能讓使用者在開車的同時安全地撥號。自 1997 年 Philips 推出第一款語者相關 (speaker-dependent) 的聲控手機開始，手機上語音撥號、語音命令功能的需求即逐年增加，包括 Nokia、Motorola、Samsung、Sony Ericsson、LG 等前五大手機廠都開始採用此功能。2006 年底路透社 (Reuters) 粗略預估當年約有 1 億到 1.5 億台手機整合語音撥號功能，而 2007 年整合語音撥號的手機數量將成長一倍。

爲此，我們從 2002 年開始研究適合在手機平台上運行的語音技術，並開發 CVC 手機聲控軟體，以台灣手機製造商爲主要客戶。台灣手機製造商以代工爲主要營運模式，代工產品行銷全球，要打入手機製造商的供應鏈，產品和技術必須能支援多國語言，因此除了中文外，我們也開發多國語言的語音辨識與 TTS 技術，目前 CVC 可支援 24 種語言的語音辨識與 TTS，詳見表一。以 HP iPAQ 510 Voice Messenger [1][2] 爲例，即內建 13 種不同語言版本的 CVC 出貨全球。除此之外，包括 Nokia 6708、ASUS P535、

Fujitsu-Siemens Pocket Loox T830、HTC Touch 等多款暢銷手機都有內建多國語言版 CVC。

表一、CVC 支援的語言版本

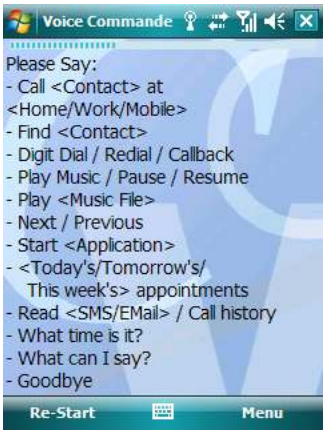
地區	語言版本
亞澳洲	台灣口音中文、大陸口音中文、粵語、韓語、日語、泰語、土耳其語、澳洲口音英語
美洲	美國口音英語、中南美口音西班牙語、巴西口音葡萄牙語
歐洲	英國口音英語、德語、法語、義大利語、西班牙語、葡萄牙語、俄羅斯語、荷蘭語、丹麥語、波蘭語、捷克語、瑞典語、希臘語

CVC 讓使用者能用語音和手機對話互動，而達成在 hand-free、eye-free 操控手機的目的。CVC 具備的功能包括語音撥號、語音指令、語音查詢聯絡人、語音點歌、語音朗讀簡訊、電子郵件、行事曆內容等。圖一爲英文版 CVC 畫面，以下是透過 CVC 進行語音撥號的範例：

CVC: 請說指令
User: 打電話給何泰軒
CVC: 打電話到何泰軒，住家、公司、手機、或取消
User: 公司
CVC: 公司，撥號中請稍候

另一個範例是使用者直接說「打電話給何泰軒公司」:

CVC: 請說指令
User: 打電話給何泰軒公司
CVC: 打電話到何泰軒公司，確認或取消
User: 確認
CVC: 撥號中請稍候



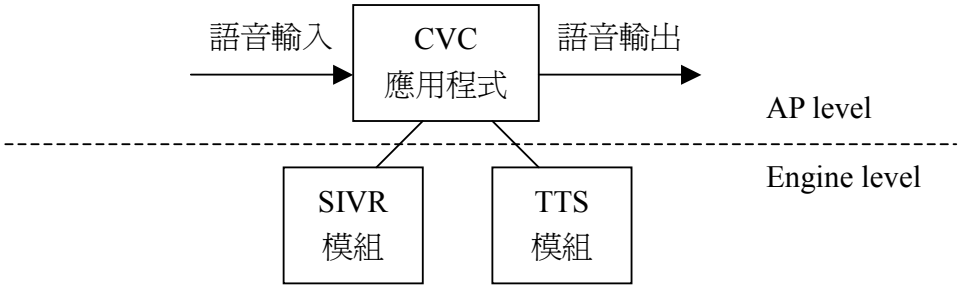
圖一、英文版 CVC 畫面

本論文將描述 CVC 的系統架構以及支援多國語言的作法，第二節將描述 CVC 系統架構，包括多國語言的語音辨識和 TTS 所使用的技術，第三節描述語音辨識實驗環境和幾個語言版本的實驗結果，第四節提出我們未來對 CVC 改良的幾個想法。

二、CVC 系統架構

CVC 包括一組應用程式，提供使用者介面、進行錄放音、控制整個和使用者對話的流程，底層爲 SIVR (speaker-independent voice recognition) 模組和 TTS 模組，如圖二所示。SIVR 模組可進行獨立詞彙辨識，或有文法限制的連續語音命令辨識，TTS 能將辨識的結果或甚至任意文字輸入轉換成 PCM 聲音，透過手機的喇叭播放放出來。由於手機資源較小，要能在上面運行，所有的運算必須改成整數運算，且使用的空間也不

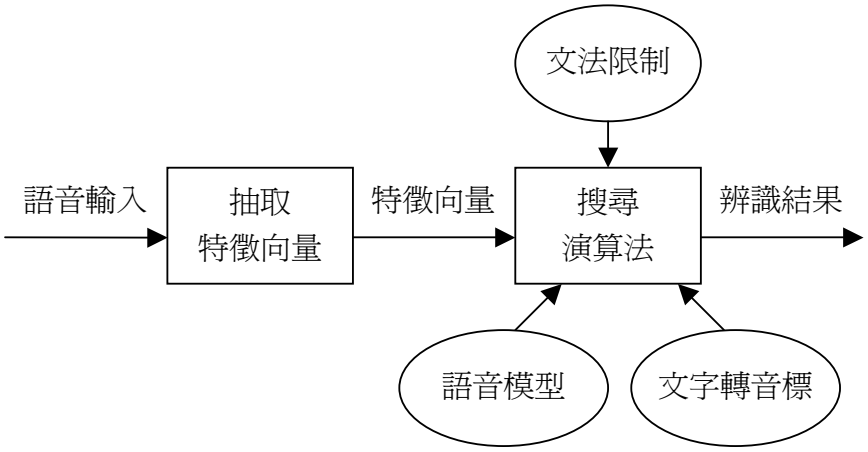
能太大，我們的 SIVR 和 TTS engine 一個語言加起來大約佔用 500KB 到 800KB，依語言不同而有所差異，RAM 大約需要 500KB。



圖二、CVC 系統架構

(一)、SIVR

SIVR 模組可分成幾個部份：一是從聲音中抽取特徵向量 (feature extraction)，二是辨識時所使用的語音模型(acoustic model)，三是將辨識的辭彙轉換成音標(word-to-phone conversion)，四是具文法限制的語音辨識搜尋演算法(search algorithm)，如圖三。



圖三、SIVR 各組成部分

1. 特徵向量

進行語音辨識時，會先將錄進來的聲音轉成特徵向量，聲音可來自手機上的麥克風、有線耳機、或藍芽耳機，由於藍芽耳機只能傳輸 8kHz, 16-bit PCM 的聲音，因此我們使用的聲音輸入爲 8kHz, 16-bit PCM。每秒中聲音抽取出 100 個特徵向量，每個特徵向量爲 16 個維度，由 8 維的 MFCC (Mel-Frequency Cepstral Coefficients) 和 8 維的 delta MFCC 所組成。MFCC 用 CMS(Cepstral Mean Subtraction)做通道效應補償。

2. 語音模型

語音模型爲傳統以音素爲單位的隱藏式馬可夫模型(Phoneme-based Hidden Markov Model)，每個模型由 3 個由左到右的狀態(state)組成，我們使用三聯音素模型(Triphone)加強對連音的辨識。相較於 Triphone 數量，我們所收集的訓練語料相對不足，此時可用決策樹(decision tree) [3][4]來決定哪些類似的狀態(state)可共用參數和訓練語料，我們可調整共用的程度來控制參數量，一方面避免某些參數的訓練語料不足，一方面可依手機的資源調整出最佳的模型大小。

訓練語音模型使用 forward-backward 演算法 [5]。每個語言我們蒐集 100 到 800 人不等的聲音，每人念 200 到 300 個句子，大約 25 到 30 分鐘的語料，每個人唸的文稿皆不同，另外再唸 40 到 60 個單詞，這些單詞涵蓋該語言所有的音標，用來初始化語音模型 (boot model)。初始化模型的步驟如下：我們先取約 10%的單詞語料，用人工標出每個組成音標的邊界，並訓練 CI model (context-independent model)，再用此 CI model 自動標出所有單詞語料的音標邊界，並以人工重新校正調整，之後再拿校正後的語料重新訓練一次 CI model。我們的經驗是初始化模型的品質對最後語音模型的好壞有很大的影響，因此在這個階段投入較多的人力做語料的處理。訓練完成 CI model 後，再將其逐步擴充訓練成 RCD model (right-context-dependent model)和 Triphone model，最後再進行共用參數的調整，依手機平台的計算資源和客戶的需求，調整出最佳大小的語音模型。

3. 文字轉音標

辨識時須將欲辨識的文字轉成音標，再由語音模型中取出對應的 Phoneme HMM，串接組成以詞爲單位模型(word model)，辨識時將輸入的語音特徵向量和 word model 比對計算相匹配的機率值。

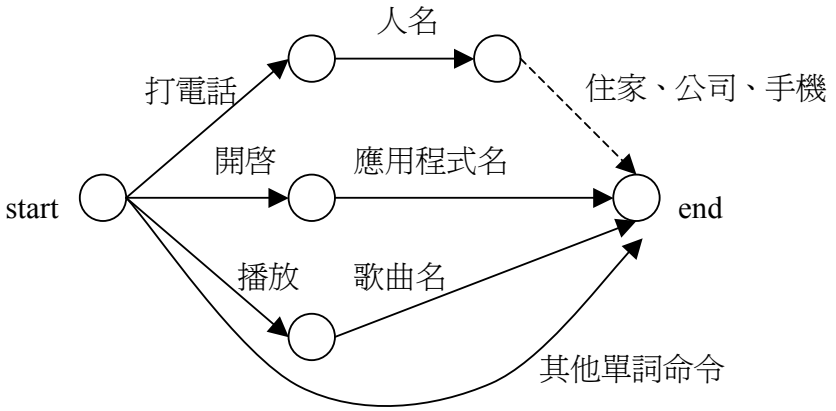
表二、CVC 各語言版本所使用的文字轉音標方法

方法	語言
發音規則加例外發音對應表	義大利語、西班牙語、葡萄牙語、捷克語、土耳其語、韓語、日語假名, 俄羅斯語、希臘語、泰語
文字音標對應表	中文、日語漢字、韓語漢字
決策樹演算法加例外發音對應表	英語、法語、德語、荷蘭語、丹麥語、瑞典語

依語言的不同我們使用三種文字轉音標的方法，第一種是有固定發音規則的語言，從文字本身即可對應出音標，例如西班牙文和義大利文，但有些外來語會出現例外的狀況，此時可建立一個例外發音的對應表來解決這個問題。第二種是完全沒有發音規則的語言，例如中文或日文的漢字，只能用一個對應表來儲存每個文字的發音。第三種是可從文字中大略猜出發音，發音有某種程度的規則性但不夠明確，例如英文，這類語言的發音難以用幾條明確的規則表達出來，我們則使用決策樹演算法 [6]來建立主要發音規則，同樣地，對一些例外的發音也可以用對應表來解決。表二是 CVC 現在所支援的語言所使用的文字轉音標方法。

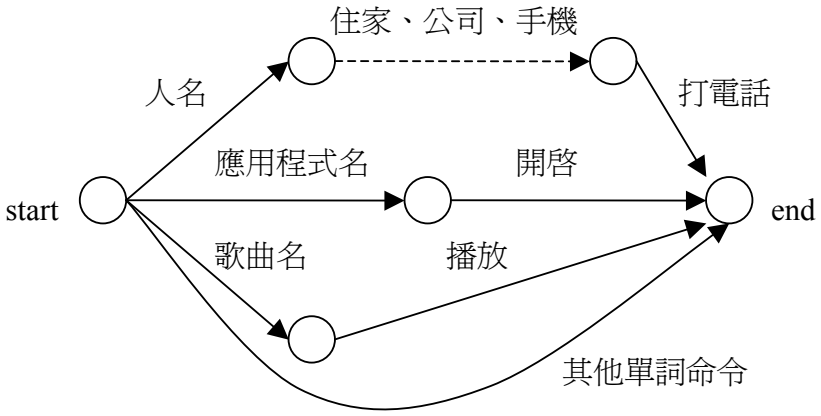
4. 搜尋演算法及文法限制

辨識使用的搜尋演算法是 Viterbi Search。SIVR 可進行獨立詞彙辨識，或有文法限制的連續語音命令辨識，以語音撥號、語音命令而言，輸入的連續語音命令通常是有規則的，例如「打電話給何泰軒住家」是由打電話(動作)、何泰軒(人名)、和住家(位址)所組成，「開啓 Windows Messenger」是由開啓(動作)、Windows Messenger(應用程式)所組成。以 CVC 而言，其語音命令的文法限制如圖四所示，其中虛線上的詞彙爲非必要的，使用者講「打電話給何泰軒」也是合法的語句。在辨識的過程中，Viterbi search 進行 word model 之間的狀態轉移(state transition)會參考此文法結構，只允許符合文法結構的狀態轉移，如此可降低連續語音辨識的搜尋複雜度，降低計算量，同時提高辨識的準確率。



圖四、CVC 連續語音命令的文法

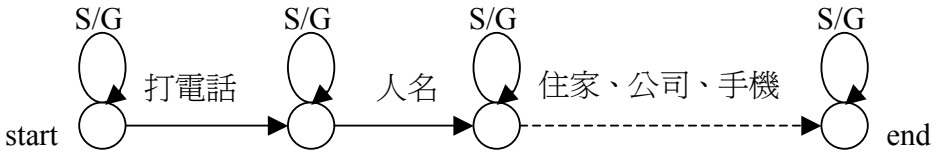
並不是所有語言版本都是使用圖四的文法結構，有些語言是後置動詞的，例如 SOV 語序的日文、韓文和土耳其文。以「打電話給何泰軒住家」爲例，韓文的念法爲「하태헌의 집으로 전화걸기」，對應到中文念法是「何泰軒（하태헌의）住家（집으로）打電話（전화걸기）」。對這類後置動詞的語言，我們使用如圖五所示的文法。



圖五、後置動詞語言版本 CVC 使用的文法

在 word model 之間的狀態轉移時，SIVR 允許先轉移至靜音模型(silence model)

或垃圾模型(garbage model)，圖六描述包含靜音和垃圾模型的打電話命令的文法，其中 S 代表靜音模型，G 代表垃圾模型。靜音和垃圾模型皆是 1 個狀態的 HMM，使用這兩個模型可以讓使用者的輸入語音中包含一些贅詞或稍微停頓不講話，例如使用者可以說「請幫我電話給…嗯…何泰軒他的…住家的…電話」，其中的「請幫我」、「給…嗯…」、「他的…」、和「的…電話」等不屬於辨識詞彙的語音有時可以被 S/G 過濾掉。不過我們實際使用的經驗發現，過濾贅詞的效果只有在辨識的詞彙量少的時候會比較好，如果某個人名和「給…嗯…」念起來的聲音很類似，就很容易發生誤判的情形。

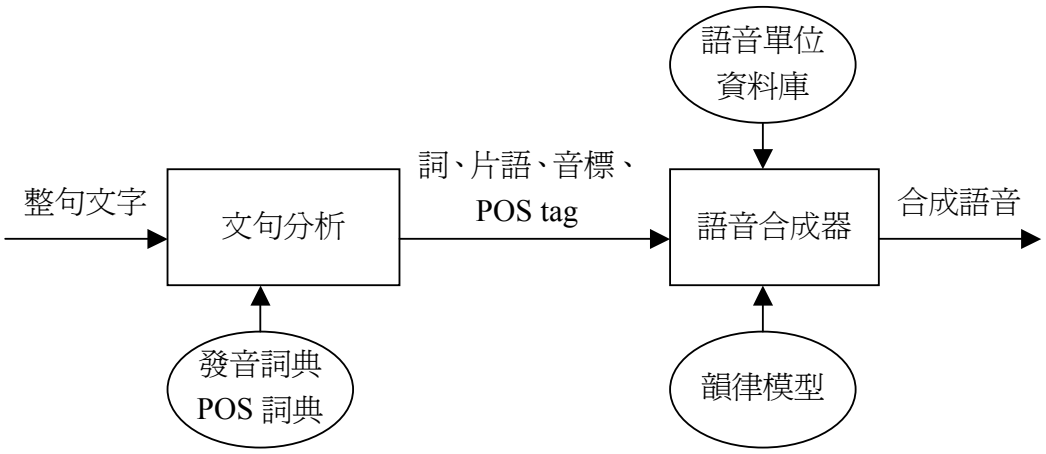


圖六、包含靜音和垃圾模型的打電話命令的文法

使用者有時會忘記語音指令的確切說法，要解決這個問題，我們可以爲固定的語音命令加入其他常用的說法，例如「打電話」也可以說成「撥電話」和「打給」，「開啓」可以說成「打開」和「執行」，把這些不同說法加入辨識的詞彙，可以讓使用者覺得產品更容易使用、更聰明、表現更穩定。

(二)、TTS

由於 CVC 主要應用的平台爲手機，相對計算量、記憶體比較小，因此我們必須選擇使用較精簡的 TTS 技術。在中文和粵語我們以音節(syllable)爲發音單位，只存放聲調爲一聲的音節，其他語言則以 diphone 爲單位。我們的聲音資料庫存的是 16kHz, 16-bit 的聲音，透過壓縮，一個語言的 TTS 大小可降低到 300KB 到 600KB，依語言不同而有所差異。我們的語音合成方法須對聲音做不少的處理，因此合成出來的聲音失真也比較嚴重，一般使用者對 CVC 語音合成品質的評價是「機械音較重，不過還是能聽得懂。」



圖七、TTS 各組成部分

TTS 的架構如圖七所示，整句文字輸入後先進行文句分析處理：對整句文字進行斷句(phrasing)，同時找出對應的音標和每個詞的 POS (part-of-speech) tag，有些語言還需要先進行斷詞處理。根據文句分析結果，韻律模型(prosodic model)建立每個片語的韻律參數，語音合成器(speech synthesizer)則一次合成一個片語的語音：由語音合成單位資料庫(speech unit database)取得片語中所需的合成單元語音資料，並參考韻律模型提供的資訊，調整片語的韻律(prosody)輸出合成語音。以下描述 TTS 的各個組成部分運作過程。

1. 文句分析

整句文字輸入後，我們須先將每個詞標註 POS，由於在詞典中每個詞可能有多種 POS，我們使用 Viterbi search 和 POS n-gram 找出機率最高的 POS tag 組合，POS n-gram 是事先由人工標示過 POS tag 的大量文字資料中訓練而得，再由此標註出的 POS tag 找出最可能的片語邊界(phrase boundary)，找尋的方法，同樣地也是使用 Viterbi search 和 boundary n-gram 找出機率最高的 boundary 位置 [7]，boundary n-gram 也是事先由人工標示過 POS 和 boundary tag 的大量文字資料中訓練而得。但有些語言我們的字典缺乏 POS，也缺乏標註好 POS tag 的文字資料，這類語言我們使用簡單的規則做斷句：先以標點符號斷句，若句子仍太長則設定片語音節數的最大值，超過最大值即強迫斷句，但斷句的位置必須落在詞邊界(word boundary)上，不能把一個詞分開放在兩個片語中。在歐洲語系詞和詞之間會用空格分開，沒有斷詞的問題，但亞洲語言如中文、泰文就必須先做斷詞，找出詞邊界，泰文一般不使用標點符號，斷詞在此處更爲重要。這種簡單規則斷句的品質並不好，不過在片語間的靜音停頓不要太長的情況下，合成語音聽起來還不至於太突兀。

2. 韻律模型

在韻律模型方面，中文和粵語都是具有聲調(tone)的語言，中文包含輕聲共有五種聲調，粵語則有九種聲調，我們用人工設計出各種聲調的 F0 形狀，透過合成器改變音節的 F0 形狀，即可合成出其他聲調的音節。除此之外，我們讓整個片語的 F0 由高到低變化，遇到有聲調的音節時將該聲調的 F0 形狀加入片語的 F0 中，組合出整個片語的 F0 變化。另一個韻律訊息是音節的長度變化，在中文和粵語我們使用固定的規則，根據音節在片語和詞內位置的變化給予不同的權重(weight)。

其他語言沒有聲調，但有重音的訊息，我們使用 CART (Classification and Regression Trees) [8]演算法預測片語中重音的位置。我們找一位語者唸大約 1 到 1.5 小時的語料，用人工標示出片語邊界、詞邊界、音節邊界、音標邊界、重音位置、重音種類、以及每個詞的 POS，從中取出若干的特徵參數，建立預測音節重音的 CART。常用的特徵參數包括：該音節在片語內的位置、在詞內的位置、該音節前後若干個音節的類別、該音節所在的詞的 POS 等，在不同的語言這些特徵參數對韻律的影響也不同，因此每種語言會使用適合該語言的特徵參數。合成時我們從片語的音標和 POS tag 中取得每個音節的特徵參數，由 CART 依據這些特徵預測哪個音節有重音、以及其重音種類，配合線性回歸法(linear regression)去訓練出預測整段片語的 F0 模型，即可組合出片語的 F0 變化，

我們也將 CART 應用在預測片語中每個語音單元的長度上，我們用 VR 的語音模型將所有訓練語料自動切出每個音標邊界，從中取出特徵參數建立 CART，常用的參數和預測重音用的 CART 類似，包括該音標在片語內的位置、在詞內的位置、在音節內的位置、該音節前後若干個音標爲何、該音標的重音種類、以及該音標所在的詞的 POS 等等。

預測重音的 CART 需要大量人力進行人工標註，目前由於人力不足，只有先針對英文和韓文建立，其他非中、粵、英、韓的語言則是將音標對應成英文的音標，用英文的 CART 預測該語言的重音位置和形狀，但每種語言會使用不同的片語 F0 變化規則，也有該語言的自己的音標長度 CART。這種方法合成的語音聽起來還可以被接受，不過韻律聽起來怪怪的，以德文爲例，德國的客戶曾提過 CVC 的 TTS 聽起來「像外國人在說德文」。

3. 語音合成器

語音合成器是基於線性預測編碼(LPC, Linear Predictive Coding)。每個語言我們找一位以該語言爲母語的女性語者進行錄音，從所錄製的語音中切割出所需要的語音合成單元(speech unit)，同時決定出每個語音單位的 pitch 位置。我們對每個 pitch 做 LPC 分析，再對 LPC 係數和 residual 壓縮儲存於資料庫中。合成時韻律模型(prosody model)會預測出整個片語的 F0 形狀和每個語音單位的長度(duration)，合成器可透過改變 LPC residual 的長度調整 F0，和透過增加或減少合成的 pitch 數目調整語音的長度。

雖然我們選擇使用精簡的 TTS 方法，不過我們的經驗發現，合成語音的品質和語音資料好壞、人工處理的品質有很大的關係，包括錄音者聲音的特性、所錄的聲音是否有瑕疵、音標邊界和 pitch 位置是否切得準確等。如果能仔細檢查資料和人工切音的結果，也是能合成出品質不錯的語音。例如我們的英語語音合成只佔用不到 400KB 的大小，但其合成英文單詞和 4 個詞以內組成的片語品質就還不錯，我們將其應用於手機英漢字典的英文發音功能中，於 2006 年下半年推出產品「賽微隨身典」，得到相當不錯的評價，終端使用者一般認爲已經達到接近真人發音的水準。

三、多國語言語音辨識實驗

本論文只有對語音辨識進行比較科學性的實驗。語音合成的驗證方法目前仍不夠嚴謹，驗證方法主要是找幾位以該語言爲母語的聽者試聽，以聽者主觀認定聽得懂爲合格的標準，目前 CVC 所支援的語言版本仍有幾個語言無法達到這個標準，細節則不在本論文中討論。

語音辨識的實驗分成兩種，第一種是模擬實驗(simulation)，我們收集以該語言爲母語的語者的聲音作爲測試語料，訓練語料中不包含測試者的聲音，我們以這些測試語料驗證我們系統的辨識率，並在開發過程中做爲改進技術的依據。測試語料爲人的名字，包含姓(last name)和名(first name)，每種語言至少收集 4 到 6 人的聲音，男女各半，測試者須把每個測試人名唸 2 次，以國內宏達國際電子公司(HTC)所生產的 Universal

PocketPC Phone 錄音，在安靜的辦公室環境中錄製。我們實驗的辨識詞彙量爲 200 個人名，由於手機聲控主要在開車時使用，我們把測試語料加入 AURORA 汽車噪音，進行 S/N 爲 15dB 到 0dB 的噪音實驗。有時候我們爲了配合客戶出貨時程而趕工，以致有些語言的實驗並不完整，在此我們選擇列出 13 種實驗做得較完整的語言，辨識率結果列於表三。

表三、各語言於 AURORA 汽車噪音下模擬實驗的準確率(%)

S/N / 語言	安靜	15dB	10dB	5dB	0dB
台灣口音中文	98.03	97.04	96.37	93.09	75.33
大陸口音中文	96.62	96.21	95.21	90.33	71.67
粵語	95.36	94.01	93.97	88.01	71.62
美國口音英語	98.90	97.90	96.68	92.58	79.40
英國口音英語	93.88	94.85	94.21	91.45	77.79
德語	95.17	95.17	93.65	87.81	75.29
法語	94.83	95.02	94.08	90.25	76.62
義大利語	95.77	94.15	93.64	91.56	81.73
西班牙語	96.18	95.37	92.83	89.28	78.00
巴西口音葡萄牙語	96.20	97.15	95.49	93.35	80.29
荷蘭語	94.25	93.12	92.62	88.12	74.75
日語	96.55	96.10	92.40	90.40	81.10
俄語	97.15	95.60	93.62	87.07	75.47
平均	96.07	95.51	94.21	90.25	76.85

在安靜環境下，各個語言基本上都可達到 95%以上的正確率，同時我們的辨識核心也有相當的穩健性，S/N 爲 10 dB 以內的噪音只會稍微降低辨識效果 2%左右。隨著噪音程度加強，辨識率隨之呈現較高的下降趨勢，在 S/N 爲 5 dB 時辨識率維持在 90%左右，0 dB 時平均辨識率則下降到約 76%，此時各個語言也呈現出較大的差異，我們推測可能和測試人名的長度有關(中文人名爲 3 個音節，其他語言人名大多高於 4 個音節)。

另一種是實地測試(field test)，CVC 每種語言版本出貨前，我們會找以該語言爲母語的測試者來實際使用，我們從旁觀察使用狀況，並記錄準確率。每個語言我們找 4 到 6 名測試者，男女各半。測試的環境爲辦公室內、馬路邊(北二高新店交流道旁新店市中興路上)、以及行進在高速公路上窗戶關閉的汽車內，汽車平均時速爲 90 到 100 公里，測試車輛是 2000 年生產的 Nissan Sentra 1.6。測試手機包括 HTC、ASUS、BenQ 所生產的 PocketPC Phone，手機的設定爲電話簿中有 200 個人名，手機系統內大約有預設 20 到 30 個應用程式，每個測試者在一種測試環境下唸 100 句的連續語音測試語句，包括下列 4 種語音命令：

1. 打電話給 <人名>
2. 打電話給 <人名> <住家、公司、手機>
3. 查詢 <人名>
4. 開啓 <應用程式>

測試結果列於表四。

表四、各語言實地測試的準確率(%)

場所 / 語言	辦公室	馬路邊	行進汽車內
台灣口音中文	98.6	92.8	93.5
大陸口音中文	96.2	90.4	92.3
粵語	94.8	89.7	91.5
美國口音英語	93.7	85.2	90.5
英國口音英語	93.2	83.7	88.5
德語	95.7	86.3	93.8
法語	96.5	91.4	92.6
義大利語	97.5	92.3	94.0
西班牙語	97.1	89.4	91.2
巴西口音葡萄牙語	95.3	87.6	88.7
荷蘭語	92.4	84.0	91.3
日語	96.2	88.3	91.2
俄語	96.3	88.4	92.8
平均	95.63	88.42	91.68

雖然和模擬測試的指令語法略爲不同，實地測試得到的結果和模擬測試相近，辦公室環境下有 95%的正確率，汽車內的辨識率 91%則介於模擬測試 S/N 爲 10 dB 和 5 dB 之間。馬路邊的噪音種類較多且較不穩定，因此辨識效果也略低一些，約爲 88%。實地測試使用的各款手機錄音頻率響應和通道效應或許不盡相同，但只要錄音音量控制得宜且無特殊的通道雜訊，都可得到相似的辨識結果。

四、結論與改良 CVC 的想法

本論文介紹 CVC 商用嵌入式語音命令系統，CVC 是國內唯一、也是全球少數能支援多國語音辨識與 TTS 的手機聲控軟體，本論文特別描述了 CVC 多國語言版本的開發經驗，和一些語言的語音辨識實驗結果。以手機聲控軟體而言，CVC 的辨識率已經能讓全球大部分語言的使用者所接受，特別是讓使用者在開車的時候，能更安全地用語音操作手機，保障自己和他人的生命財產安全。

我們於 2002 年中開始 CVC 產品的研發，於 2004 年初產品上市開始銷售，經過多年不斷地改良技術，語音辨識準確率和 TTS 發音品質才逐漸爲使用者所接受，並獲得國內外各大手機製造公司採用，內建於其手機產品中行銷全球。雖然如此，CVC 還是有很大的改進空間，在全球市場競爭仍然十分激烈的今天，我們絕對不能怠慢，以下我們提出一些 CVC 可以改善的地方：

1. 我們在製作某些語言版本時遭遇一些困難，例如阿拉伯文因爲文化的關係，很難找到女性的語者來錄音，有些語言我們收集的資料也明顯不足。我們未來將加強這些語料的收集，提升這些語言的語音辨識率。

2. 有時爲了配合客戶出貨而趕工，以致有些語言的實驗做得不夠嚴謹，可能造成消費者抱怨辨識率不好的風險，帶給客戶困擾。我們正積極招募更多具語音相關研究背景的人才，解決目前人力不足的問題，並爲每個語言建立更嚴謹更完善的實驗流程。

3. TTS 的品質還有很大的改進空間，目前有三個改進的方向，一是收集更多語言的訓練資料，爲每個語言建立韻律模型，二是使用更多的語音單位和 unit selection 的方法，減少對語音資料的調整，降低合成語音失真的程度，三是建立嚴謹的驗證方法和流程，確保產品的品質。

4. 我們發現很多使用者剛開始對產品不熟悉，唸出錯誤的語音指令而誤以爲 CVC 聽不懂他們說的話，幾次辨識失敗後產生的挫折感讓他們不願再使用。要讓產品更容易使用，我們須提升 CVC 語音辨識模組的容錯能力，甚至從使用者不完整的語音命令中，了解使用者可能的意圖，再透過對話的過程解決使用者不會操作的問題。

5. Barge-in，此功能可讓熟悉 CVC 的使用者直接切斷冗長的系統語音提示，讓 CVC 的操作更有效率。

參考文獻

[1] *HP iPAQ 510 Voice Messenger series - overview and features*. Available: http://h10010.www1.hp.com/wwpc/us/en/sm/WF05a/215348-215348-64929-314903-3352590-3360087.html

[2] Bonnie Cha, *HP iPaq 510 Voice Messenger*, CNET Asia Review, February 2007. Available: http://asia.cnet.com/reviews/mobilephones/0,39051199,40151061p,00.htm

[3] Hwang, M. Y., *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, PhD thesis, CMU-CS-92-230, Carnegie Mellon University, 1993

[4] Odell, J.J., *The Use of Context in Large Vocabulary Speech Recognition*, PhD thesis, Cambridge University, 1995

[5] Rabiner, L. and Juang, B., *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[6] Black, A., Lenzo, K., and Pagel, V., “Issues in building general letter to sound rules”, *ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia. 1998.

[7] Taylor, P., and Black, A. “Assigning Phrase Breaks from Part of Speech Sequences”, *Computer Speech and Language*. Vol. 12, pp. 99-117, 1998.

[8] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, P. J., *Classification and Regression Trees.* The Wadsworth Statistics/Probability Series, Wadsworth and Brooks, 1984.

改善以最小化音素錯誤爲基礎的鑑別式聲學模型訓練於中文連續語音辨識之研究

劉士弘, 朱芳輝, 陳柏琳
國立臺灣師範大學資訊工程學系
{ g93470185, g94470144, berlin}@ntnu.edu.tw

摘要

本論文探討改善最小化音素錯誤爲基礎的鑑別式聲學模型訓練於中文大詞彙連續語音辨識之研究。首先，本論文提出一個新的音框層次音素正確率函數來取代最小化音素錯誤訓練的原始音素正確率函數，此新的音素正確率函數在某種程度上能充分地懲罰刪除錯誤。其次，本論文提出一個以音框層次正規化熵值爲基礎的嶄新資料選取方法來改進鑑別式訓練，其正規化熵值是由訓練語料所產生之詞圖中高斯分布之事後機率所求得。此資料選取方法可以讓鑑別式訓練更集中在那些離決定邊界較近的訓練樣本所收集的統計值，以達到較佳的鑑別力。所使用的實驗題材是公視新聞外場記者語料。初步的實驗結果顯示，結合時間音框層次的資料選取方法和新的音素正確率函數在前幾次的迭代訓練中確實有些微且一致的進步。

關鍵詞：最小化音素錯誤訓練，鑑別式訓練，資料選取方法，大詞彙連續語音辨識

一、緒論

語音，是人與人之間最自然的溝通橋樑，倘若語音能夠成爲資訊產品的主要輸入形式，那麼人與機器之間的溝通就會變得簡單許多，並且可以盡量避免文明病的產生。因此自動語音辨識(Automatic Speech Recognition, ASR)的研究已變得非常重要，這也是目前語音與語言處理領域中熱門的研究議題之一。

大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)所使用的鑑別式訓練法則是不以最大化訓練語料的相似度爲目標，而以最小分類錯誤爲目標，進而增進辨識率。傳統在聲學模型之訓練上，大都使用最大化相似度(Maximum Likelihood, ML)法則，配合波氏重估演算法(Baum-Welch algorithm)來進行聲學模型的訓練，但此種訓練方法並沒有考慮語音辨識時聲學模型彼此間的關係，在調整聲學模型參數之後，可以使得相關的語音特徵落在此聲學模型的相似度(Likelihood)變大，卻也可能同時讓非相關的語音特徵落在此聲學模型的相似度更大，造成辨識上的混淆。因此，近來有不少研究針對此項缺點，提出鑑別式訓練(Discriminative Training)法則來加以改進。故本文著重於探討以最小化音素錯誤(Minimum Phone Error Training, MPE)爲基礎的鑑別式訓練法則，藉著提出一個新的音框層次音素正確率函數來改善原始音素正確率函數之缺點，同時也提出一個以音框層次正規化熵值爲基礎的嶄新資料選取方法來改進最小化音素錯誤訓練。

本論文接下來的安排如下：第二章將介紹貝氏風險與全面風險；第三章則介紹最小化音素錯誤聲學模型訓練；第四章探討最小化音素錯誤訓練之改進；第五章則探討資

料選取方法於改進最小化音素錯誤聲學模型訓練；第六章爲實驗與討論；第七章爲結論與未來展望。

二、貝氏風險與全面風險

語音辨識的過程可視爲一個分類的動作，將每句可能的詞序列都視爲一類，語音辨識即是要從所有可能類別(詞序列)中找出最佳的一類(一句)。若 O_z 爲一語句的語音特徵向量序列，將 O_z 歸類至詞序列 W 時，可以用函數 $R(W \mid O_z)$ 代表此歸類行爲的風險(Risk)；而語音辨識則可視爲找出此風險最低的詞序列。 $R(W \mid O_z)$ 可定義如下[1]：

$$R(W \mid O_z) = \sum_{W' \in \mathbf{W}} l(W, W') P(W' \mid O_z) \tag{1}$$

其中 **W** 爲所有可能詞序列所成的集合； $P(W' \mid O_z)$ 表示給定語音特徵向量序列 O_z 時，詞序列 W' 的事後機率(Posterior Probability)； $l(W, W')$ 爲一減損函數(Loss Function)，用以表示詞序列 W 與 W' 之間差異所造成的損失(Loss)， $R(W \mid O_z)$ 爲將 O_z 歸類至 W 時的期望損失(Expected Loss)，又稱爲貝氏風險(Bayes Risk)或條件風險(Conditional Risk)。在語音辨識或解碼上，需要最小化此貝氏風險來找最佳的詞序列 $\hat{W}$ ，即：

$$\hat{W} = \arg\min_{W \in \mathbf{W}} R(W \mid O_z) = \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} l(W, W') P(W' \mid O_z) \tag{2}$$

目前有許多辨識器根據貝氏決策定理(Bayesian Decision Theorem)，即最小化此貝氏風險(式(2))來設計其搜尋演算法，如標準最大化事後機率解碼方法(Maximum a Posteriori Decoding, MAP)[2]、ROVER(Recognizer Output Voting Error Reduction)[3]、最小化貝氏風險(Minimum Bayes Risk, MBR)[4]、最小化時間音框錯誤搜尋(Minimum Time Frame Error Search)[5]及詞錯誤最小化(Word Error Minimization) [6]等。

然而，若在聲學模型和語言模型的訓練上，則需要計算全面風險(Overall Risk)，並且最小化此全面風險 R_{all} [1]：

$$R_{all} = \int R(W \mid O) P(O) dO \tag{3}$$

其中 W 爲語音特徵向量序列 O 對應之正確轉譯詞序列， $P(O)$ 爲 O 的事前機率(Prior Probability)；全面風險 R_{all} 是在語句空間(語音特徵向量序列空間)上作積分，爲所有訓練語句(語音特徵向量序列)的期望條件風險(Expected Conditional Risk)。由於訓練語料有限，故全面風險可簡化爲 Z 個訓練語句的條件風險總和：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z) P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z) P(O_z) \tag{4}$$

若事後機率分布 $P(W' \mid O_z)$ 由聲學模型 λ 及語言模型 Γ 所決定，令 $\theta = \{\lambda, \Gamma\}$ ，所以事後機率我們將之表示爲 $P(W' \mid O_z; \theta)$ ，則全面風險可改寫成：

$$R_{all} = \sum_{z=1}^{Z} R(W_z \mid O_z) P(O_z) = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) P(O_z) \tag{5}$$

若假設 $P(O_z)$ 對所有 O_z 均有一致(Uniform)的機率，且此項與模型參數 λ 及 Γ 無關，則可將此項省略：

$$R_{all} = \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) \tag{6}$$

在估測聲學模型和語言模型時，希望估測之模型θ能將全面風險降至最低：

$$\hat{\theta} = \arg\min_{\theta} \sum_{z=1}^{Z} \sum_{W' \in \mathbf{W}} l(W_z, W') P(W' \mid O_z; \theta) \tag{7}$$

在此所表示的減損函數是一般化減損函數(Generalized Loss Function)，並沒有明確定義要如何計算，這也因此成爲一個開放的研究議題(亦即要如何去設計一個減損函數以期望訓練出較佳的模型θ，進而提高辨識率)。目前有許多的模型訓練的方法都是以風險最小化(Risk Minimization)爲基礎，並搭配其設計的減損函數來達成鑑別式之模型訓練，如最大化交互資訊估測(Maximum Mutual Information Estimation, MMIE) [7]、全面風險估測法則(Overall Risk Criterion Estimation, ORCE) [8]、最小化貝氏風險鑑別式訓練(Minimum Bayes Risk Discriminative Training, MBRDT) [9]、最小化音素錯誤訓練(Minimum Phone Error Training, MPE) [10]等。

三、最小化音素錯誤之聲學模型訓練

新近劍橋大學提出的最小化音素錯誤(Minimum Phone Error, MPE)聲學模型訓練，是以全面風險爲出發，以辨識出詞序列的原始音素正確率(Raw Phone Accuracy)函數$A(W_i, W_z)$來取代其中減損函數$l(W_i, W_z)$。因此，它的目標函數變成是最大化語音辨識器對所有訓練語句(語音特徵向量序列)O_z的可能辨識出候選詞序列W_i($W_i \in \mathbf{W}_z = \{W_1, W_2, W_3, \cdots\}$)的期望音素正確率(也就是最小化語音辨識器對所有訓練語句可能辨識出候選詞序列W_i的期望錯誤率)，最小化音素錯誤的目標函數可表示如下：

$$F_{MPE}(\lambda) = \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} p(W_i \mid O_z) A(W_i, W_z) = \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_z} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} A(W_i, W_z) \tag{8}$$

其中$p(O_z)$可用語音辨識器產生的詞圖$\mathbf{W}_{z,lattice}$來近似[11]，因此目標函數可進一步表示成：

$$F_{MPE}(\lambda) \approx \sum_{z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_\lambda(O_z \mid W_k) P(W_k)} A(W_i, W_z) \tag{9}$$

其中W_i與W_k分別表示詞圖$\mathbf{W}_{z,lattice}$上任兩條候選詞序列(假設O_z對應的正確詞序列W_z亦包含在詞圖裡)。

爲了對目標函數$F_{MPE}(\lambda)$進行最佳化，Povey 等人提出最小化音素錯誤的弱性(Weak-sense)輔助函數$H_{MPE}(\lambda, \bar{\lambda})$爲[12]：

$$H_{MPE}(\lambda, \bar{\lambda}) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \left[\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)} \Big|_{\lambda=\bar{\lambda}} \right] \log p_\lambda(O_z \mid q) \tag{10}$$

其中$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)} \Big|_{\lambda=\bar{\lambda}}$的值可爲正或負，取決於詞圖上通過此音素的候選詞序列的期望正確率$c_z(q)$是否大於詞圖上所有候選詞序列的期望正確率c_{avg}^{z}。也就是：

$$\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}} = \gamma_q^z\left(c_z(q) - c_{avg}^z\right) \tag{11}$$

其中：

$$\gamma_q^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{12}$$

爲詞圖上通過音素段落 q 的候選詞序列的事後機率和，而

$$c_z(q) = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}, q \in W_i} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}: q \in W_k} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{13}$$

爲詞圖上通過此音素段落的候選詞序列的期望正確率，而

$$c_{avg}^z = \frac{\sum_{W_i \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_i) P(W_i) A(W_i, W_z)}{\sum_{W_k \in \mathbf{W}_{z,lattice}} p_{\bar{\lambda}}(O_z \mid W_k) P(W_k)} \tag{14}$$

爲詞圖上所有候選詞序列的期望正確率。 γ_q^z 、 $c_z(q)$與 c_{avg}^z 的統計量可在詞圖上使用波氏重估演算法來求得[12]。

另一方面，針對對數機率函數 $\log p_\lambda(O_z \mid q)$ ，必需透過一個強性輔助函數 $Q_{ML}(\lambda, \bar{\lambda}, z, q)$ 來估測新的模型參數值，因此弱性輔助函數 $H_{MPE}(\lambda, \bar{\lambda})$ 可表示成：

$$H'_{MPE}(\lambda, \bar{\lambda}) = \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \left[\frac{\partial F_{MPE}(\lambda)}{\partial \log p_\lambda(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}} \right] Q_{ML}(\lambda, \bar{\lambda}, z, q) \tag{15}$$

若以 $\gamma_q^{z\,MPE}$ 來表示 $\frac{\partial F_{MPE}(\lambda)}{\partial \log p(O_z \mid q)}\Big|_{\lambda=\bar{\lambda}}$ ，且 $Q_{ML}(\lambda, \bar{\lambda}, z, q)$ 可表示如下:

$$Q_{ML}(\lambda, \bar{\lambda}, z, q) = \sum_{t=s_q}^{e_q} \sum_m \gamma_q^z(t) \log N(o_z(t); \mu_{qm}, \Sigma_{qm}) \tag{16}$$

其中 $o_z(t)$爲 O_z 的第 t 個語音特徵向量； $N(\cdot; \mu_{qm}, \Sigma_{qm})$ 是音素段落 q 的第 m 個高斯分布， μ_{qm} 與 Σ_{qm} 分別是它的平均值向量與共變異矩陣。因此弱性輔助函數 $H_{MPE}(\lambda, \bar{\lambda})$ 可進一步表示成：

$$H'_{MPE}(\lambda, \bar{\lambda}) = \sum_z \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{z\,MPE} \gamma_{qm}^z(t) \log N(o_z(t); \mu_{qm}, \Sigma_{qm}) \tag{17}$$

其中 s_q 與 e_q 分別爲音素段落 q 的開始與結束時間， $\gamma_{qm}^z(t)$ 爲語音特徵向量 $o_z(t)$在音素段落 q 上的高斯分布 m 的佔有機率。若把平滑函數 $H_{SM}(\lambda, \bar{\lambda})$ 加入弱性輔助函數 $H'_{MPE}(\lambda, \bar{\lambda})$ ，則 $H'_{MPE}(\lambda, \bar{\lambda})$ 可進一步表示成[12]：

$$H''_{MPE}(\lambda,\overline{\lambda}) = \sum_z \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{z^{MPE}} \gamma_{qm}^z(t) \log N(o_z(t), \mu_{qm}, \Sigma_{qm}) \\ -\sum_{q,m} \frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|) + (\mu_{qm}-\overline{\mu}_{qm})^T \Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm}\Sigma_{qm}^{-1})\right] \tag{18}$$

而平滑函數 $H_{SM}(\lambda,\overline{\lambda})$ 表示爲：

$$H_{SM} = \sum_{q,m} \frac{D_{qm}}{2}\left[\log(|\Sigma_{qm}|) + (\mu_{qm}-\overline{\mu}_{qm})^T \Sigma_{qm}^{-1}(\mu_{qm}-\overline{\mu}_{qm}) + tr(\overline{\Sigma}_{qm}\Sigma_{qm}^{-1})\right] \tag{19}$$

其中 $\overline{\mu}_{qm}$ 與 $\overline{\Sigma}_{qm}$ 爲舊有模型的平均值向量與共變異矩陣。我們可以對 $H'_{MPE}(\lambda,\overline{\lambda})$ 使用延伸波式(Extended Baum-Welch ，EBW)演算法得到聲學模型參數估測更新公式(當假設語音特徵向量維度間爲無關時，亦即共變異矩陣爲對角矩陣)[12]：

$$\mu_{qmd} = \frac{\{\theta_{qmd}^{num}(O) - \theta_{qmd}^{den}(O)\} + D_{qmd}\overline{\mu}_{qmd}}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}} \\ \sigma_{qmd}^2 = \frac{\{\theta_{qmd}^{num}(O^2) - \theta_{qmd}^{den}(O^2)\} + D_{qmd}(\overline{\sigma}_{qmd}^2 + \overline{\mu}_{qmd}^2)}{\{\gamma_{qm}^{num} - \gamma_{qm}^{den}\} + D_{qmd}} - \mu_{qmd}^2 \tag{20}$$

其中統計值資訊可分爲兩類，亦即 *num*(numerator)與 *den*(denominator)兩類，*num* 代表 $\gamma_q^{z^{MPE}}$ 爲正時的統計值資訊，而 *den* 則代表 $\gamma_q^{z^{MPE}}$ 爲負時的統計值資訊，詳細統計資訊可分別表示如下：

$$\gamma_{qm}^{num} = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) \tag{21}$$

$$\theta_{qmd}^{num}(O) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t) \tag{22}$$

$$\theta_{qmd}^{num}(O^2) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, \gamma_q^{z^{MPE}}) o_z(t)^2 \tag{23}$$

$$\gamma_{qmd}^{den} = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) \tag{24}$$

$$\theta_{qmd}^{den}(O) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t) \tag{25}$$

$$\theta_{qmd}^{den}(O^2) = \sum_{z=1}^{Z} \sum_{q\in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \gamma_{qm}^z(t) \max(0, -\gamma_q^{z^{MPE}}) o_z(t)^2 \tag{26}$$

其中式(20)中的 D_{qmd} 是一個常數，需要用來確保每一維度的變異數必須要是正數，同時它也會影響收斂速度。一般而言， D_{qmd} 的值都設爲最小確保變異數爲正數的兩倍。另外，爲了要增加正確詞序列的對於模型參數訓練時的貢獻，可以引入所謂的 I-Smoothing 技術[12]，其公式如下：

$$\theta'^{num}_{qmd}(O) = \theta^{num}_{qmd}(O) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}}\theta^{ML}_{qmd}(O)$$
$$\theta'^{num}_{qmd}(O^2) = \theta^{num}_{qmd}(O^2) + \frac{\tau_{qm}}{\gamma^{ML}_{qm}}\theta^{ML}_{qm}(O^2) \quad (27)$$
$$\gamma'^{num}_{qm} = \gamma^{num}_{qm} + \tau_{qm}$$

其中γ^{ML}_{qmd}、$\theta^{ML}_{qmd}(O)$及$\theta^{ML}_{qmd}(O^2)$爲使用傳統最大化相似度訓練法所求得的統計值資訊，τ_{qm}爲訓練時設定之常數。最後，對於詞圖$\mathbf{w}_{z,lattice}$上候選詞序列正確率可以以每一候選詞序列所組成音素的正確率加總來代表，原始最小化音素錯誤(MPE)聲學模型訓練的計算候選詞序列中每一個組成音素q正確率的公式是：

$$A(q) = \max_u \begin{Bmatrix} -1+2e(q,u) & \text{if } u \text{ and } q \text{ are same phone} \\ -1+e(q,u) & \text{if } u \text{ and } q \text{ are different phones} \end{Bmatrix} \quad (28)$$

其中$e(q,u)$爲音素q與正確詞序列中音素u的重疊比例(根據正確音素u的長度)，如果q與u爲同一音素則套用計算式$-1+2e(q,u)$，反之套用計算式$-1+e(q,u)$，最後$A(q)$取與所有重疊的正確詞序列中音素u計算式值最大者爲音素q正確率(介於-1 與 1 之間)，圖 1 爲計算原始音素正確率的一個範例。而$\gamma^{MPE}_{q^z}$則可以由前向後向演算法(Forward-Backward Algorithm)求得[12]。

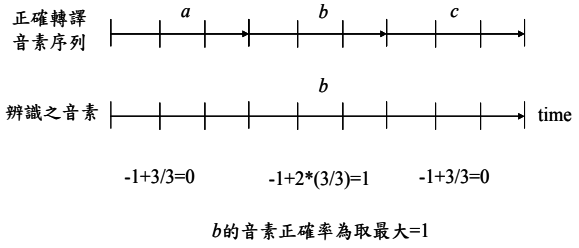


圖 1 最小化音素錯誤訓練之原始音素正確率函數的範例

四、最小化音素錯誤訓練之改進

最小化音素錯誤(MPE)訓練主要有兩個缺點:

1. 最小化音素錯誤中的原始音素正確率函數(Raw Phone Accuracy Function)並沒有給予刪除錯誤(Deletion Errors)適當的懲罰，只有對於插入錯誤(Insertion Errors)和取代錯誤(Substitution Errors)給予適當的懲罰。
2. 原始音素正確率函數是以音素爲單位(Phone-by-Phone)來做計算，如式(28)所示，且其值域範圍爲-1 到+1，這樣的範圍可能過於狹窄。每個音素段落(Phone Arc)所收集到的正確率統計值最大爲 1，因此遇到訓練語料不足的任務時，模型訓練所收集到的統計值會不夠強健。

爲了克服上述之問題，吾人提出了時間音框音素正確率(Time Frame Phone Accuracy Function, 記作 TFA)函數來取代原始音素正確率函數[13]：

$$TimeFrameAccuracy(q) = \frac{\sum_{t=s_q}^{e_q}\delta(q,u(t))}{e_q - s_q + 1} \quad (29)$$

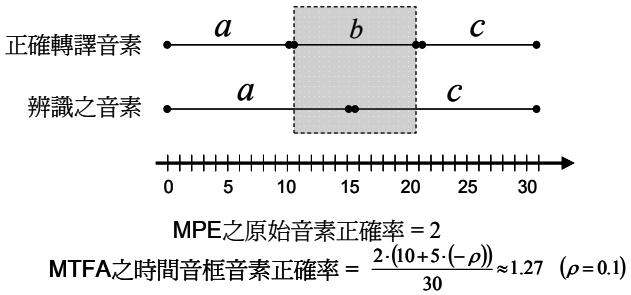


圖 2 最小化音素錯誤訓練之原始音素正確率及時間音框音素正確率對於刪除錯誤的影響

$$\delta(q,u(t))=\begin{Bmatrix}1 & ,if\ q=u(t)\\ -\rho & ,if\ q\neq u(t),0<\rho<1\end{Bmatrix} \tag{30}$$

其中 q 爲詞圖中某一音素段落，s_q 和 e_q 分別爲音素段落 q 的開始時間及結束時間，$u(t)$ 爲正確音素段落 u 在時間 t 時的音素標記(Phone Label)，ρ 爲刪除錯誤的懲罰權重(Deletion Penalty Weight)，用來懲罰某不完全正確音素段落 q 的正確率，因此某一音素段落在某個時間點 t 的正確率值域範圍爲介於 $-\rho$ 到 1 之間。時間音框音素正確率公式是看每一個音框的音素標記是否與正確音素標記一致來計算音素段落的正確率，因此對於一個完整的語句所對應的詞序列，就只要計算是否擊中(Hit)或取代(Substitution)，而不用考慮插入(Insertion)或刪除(Deletion)，因此在音素段落比對時比計算編輯距離(Edit or Levenshtein Distance)有效率，且時間音框音素正確率與我們要做評估的音素正確率有很大的正相關[5]，所以使用時間音框音素正確率的確可以去近似某個音素段落的音素正確率。圖 2 即爲計算時間音框音素正確率(TFA)的一個例子，假設某個語句有 30 個音框，此語句的正確轉譯音素共有三個，即 a、b 和 c；而此語句的辨識音素只有兩個，即 a 和 c，那麼 b 就是刪除錯誤。在圖 2 中灰色部份代表出現刪除錯誤，此刪除錯誤發生在第 11 個到第 20 個時間音框，我們理當給予這些錯誤的時間音框一些刪除錯誤的懲罰。而在詞圖中一整條路徑(詞序列) W_i 的時間音框音素正確率爲:

$$TimeFrameAcc(W_i)=\sum_{q\in W_i}TimeFrameAccuracy(q) \tag{31}$$

將式(31)取代式(28)，即本論文所提出的最大化時間音框音素正確率(Maximum Time Frame Phone Accuracy, 記作 MTFA)的目標函數:

$$\begin{aligned}F_{MTFA}(\lambda)&=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}}p(W_i\mid O_z)TimeFrameAcc(W_i)\\&=\sum_{z=1}^{Z}\sum_{W_i\in\mathbf{W}_{z,lattice}}\frac{p_\lambda(O_z\mid W_i)P(W_i)}{p(O_z)}TimeFrameAcc(W_i)\end{aligned} \tag{32}$$

另外，爲了更充分地懲罰刪除錯誤且使其值域與原始音素正確率同爲介於-1 到 1 之間，本論文使用了 S 型函數(Sigmoid Function)來正規化時間音框音素正確率函數(式(29))的分子項，稱之爲 S 型時間音框音素正確率函數(Sigmoid Time Frame Phone Accuracy, 記作 STFA):

$$SigTimeFrameAccuracy(q)=\frac{2}{1+\exp(-\alpha\cdot net+\beta)}-1 \tag{33}$$

其中

$$net = \sum_{t=s_q}^{e_q} \delta(q, u(t)) \tag{34}$$

其$\delta(\cdot)$的定義同式(30)，α及β爲S型函數中可調整的參數，α控制S型函數的曲度，β則控制S型函數的平移。故式(33)的值域範圍介於-1 到+1 之間。而在詞圖中一整條路徑(詞序列)W_i的S型時間音框音素正確率爲:

$$SigTimeFrameAcc(W_i) = \sum_{q \in W_i} SigTimeFrameAccuracy(q) \tag{35}$$

將式(35)取代式(28)，則本論文所提出的最大化 S 型時間音框音素正確率(Maximum Sigmoid Time Frame Phone Accuracy, 記作 MSTFA)的目標函數爲:

$$\begin{aligned} F_{MTFA}(\lambda) &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} p(W_i \mid O_z) SigTimeFrameAcc(W_i) \\ &= \sum_{z=1}^{Z} \sum_{W_i \in \mathbf{W}_{z,lattice}} \frac{p_\lambda(O_z \mid W_i) P(W_i)}{p(O_z)} SigTimeFrameAcc(W_i) \end{aligned} \tag{36}$$

其本論文所提出的時間音框音素正確率函數主要並非去逼近編輯距離，而只是有考量給予刪除錯誤一些適當懲罰，以改進最小化音素錯誤(MPE)鑑別式聲學模型訓練。至於有關如何在詞圖中正確地計算編輯距離，可以參考[14]。

五、資料選取方法於改進最小化音素錯誤聲學模型訓練

近年來，由於最大邊際分類器(Large Margin Classifier)[15]在機器學習(Machine Learning)的領域中已有高度的發展，且在分類(Classification)任務中都已達到非常不錯的分類效果。其設計理念就在於提升分類器的一般化能力(Generalization Ability)，以致能夠在未知的測試樣本中達到較好的分類效果。在觀念上，我們以二元類別且可分離的(Separable)訓練樣本爲例，因訓練樣本通常與測試樣本會有不一致(Mismatch)的現象，要提升分類器的一般化能力，就要使得訓練樣本在某種定義域中(如相似度定義域(Likelihood Domain))離此定義域的決定邊界(Decision Boundary)越遠越好，訓練樣本到決定邊界的最近距離我們一般會稱之爲邊際(Margin)，而此邊際越大且邊際內沒有其他的訓練樣本代表一般化能力及容錯能力會越好[16]。

最大邊際估測法(Large Margin Estimation, LME)[17]是以相似度(Likelihood)爲基礎的分離邊際(Separation Margin)來選取距離決定邊界(Decision Boundary)較近的語音特徵向量序列，依其選取門檻(Threshold)，可以定義出支持向量集合(Support Vector Set)，再利用最大邊際估測法則進而調整聲學模型。對於那些不在支持向量集合裡的訓練樣本(訓練語句)，因爲距離決定邊界較遠，所以較不具鑑別力，因此就沒有拿來調整聲學模型的參數。所以我們可以視最大邊際估測法爲以相似度作爲選取準則的資料選取方法，選出較爲重要的語音特徵向量序列(訓練語句)。在柔性邊際估測法(Soft Margin Estimation, SME)[18]中，也是以相似度作爲選取準則，藉由定義不同的門檻值，進而選取出較有影響力的訓練語句，而且從選取出來的訓練語句中，更進一步地用類別比對(Label Matching)的方式選取出重要的時間音框(Frame)。所以我們也可以視柔性邊際估測法(SME)爲以相似度和類別比對爲基礎的進階資料選取方法。

最大邊際估測法與柔性邊際估測法所使用的資料選取方法都是在相似度定義域

(Likelihood Domain)中來執行資料的選取。在本論文中，吾人提出以熵值(Entropy)爲基礎的時間音框資料選取(Data Selection)方法來改進最小化音素錯誤聲學模型訓練。其中是以給定在某語音特徵向量序列(訓練語句) O_z，某個狀態中的某個高斯分布出現的事後機率(Posterior Probability，此事後機率有考慮到詞與詞之間的轉移機率，即語言模型機率)來求得熵值，再利用事先所設定的門檻值來選取資料，故可視爲在事後機率定義域中來取選資料。因其熵值的計算是在事後機率定義域(Posterior Domain)中，故有別於以相似度定義域爲基礎的傳統資料選取方法。然而傳統熵值的值域爲 0 到 $\log_2 N$，其中 N 爲參與熵值計算的樣本個數，但爲了方便決定門檻值進而選取時間音框，故在此我們使用正規化熵值(Normalized Entropy)來使其值域介於 0 到 1 之間，其公式如下：

$$E_z(t) = \frac{1}{\log_2 N} \sum_{q=1}^{Q} \sum_{m \in q} \gamma_{qm}^{z}(t) \cdot \log_2 \frac{1}{\gamma_{qm}^{z}(t)} \tag{37}$$

其中 $E_z(t)$ 爲在第 z 句訓練語句時間 t 時的正規化熵值，$\gamma_{qm}^{z}(t)$ 爲在第 z 句訓練語句時間 t 時，在音素段落 q 中之高斯模型 m 的事後機率，Q 爲在時間 t 時所有的音素段落個數，N 爲在時間 t 中所有事後機率不爲零的高斯模型 m 。

然而在資料選取方法中，資料(或樣本)可以定義在不同的單位上，以語音辨識爲例，訓練樣本(Training Sample)可以定義在語音特徵向量序列(訓練語句(Sentence or Utterance))、詞圖中的某詞段(Word Arc)、音素段落(Phone Arc)或時間音框(Frame)等。在語音辨識的任務中，鑑別式訓練收集統計值時是以時間音框(Frame)爲最小單位，所以本論文將著重在時間音框之選取(Frame Selection)，並將每一個時間音框視爲一個訓練樣本(Training Sample)。鑑別式訓練時是將所有的時間音框所收集到的統計值都用來調整模型的參數，事實上有些時間音框對於鑑別式訓練是沒有幫助的，例如那些已經可以被分類器(在語音辨識中，通常使用連續密度隱藏式馬可夫模型(CDHMM)來當成分類器)很正確分類或很錯誤分類的時間音框，故本論文提出的以熵值(Entropy)爲基礎的時間音框選取方法(Frame Selection)旨在找出哪些時間音框是會被很正確或很錯誤地分類，哪些是不容易被分類正確，進而丟棄那些被很正確分類和被很錯誤分類之時間音框所收集到的統計值，且只利用這些被收集到的統計值來調整模型參數，以幫助鑑別式聲學模型訓練。因此使用此資料選取方法可以適用於所有的鑑別式聲學模型訓練，不僅能夠保持鑑別式訓練最小化訓練樣本分類錯誤率，還可以增進分類器的一般化能力。以最小化音素錯誤(MPE)訓練的統計值收集爲例，每個時間音框要先算正規化熵值，再由其門檻值決定是否累加統計值，則其數學式可表示爲(以 *num* 類爲例):

$$\begin{aligned}
\gamma_{qm}^{num} &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z^{MPE}}) \right] \cdot I(E_z(t) > \rho) \\
\theta_{qmd}^{num}(O) &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z^{MPE}}) o_z(t) \right] \cdot I(E_z(t) > \rho) \\
\theta_{qmd}^{num}(O^2) &= \sum_{z=1}^{Z} \sum_{q \in \mathbf{W}_{z,lattice}} \sum_{t=s_q}^{e_q} \left[\gamma_{qm}^{z}(t) \max(0, \gamma_{q}^{z^{MPE}}) o_z(t)^2 \right] \cdot I(E_z(t) > \rho)
\end{aligned} \tag{38}$$

其中 ρ 爲事先定義的門檻值(Threshold)，其值介於 0 到 1 之間， $I(E_z(t) > \rho)$ 可表示爲:

$$I(E_z(t) > \rho) = \begin{cases} 1, \ if \ E_z(t) > \rho \\ 0, \ if \ E_z(t) \le \rho \end{cases} \tag{39}$$

式(39)使用的是指示函數，其值非 0 即 1，故我們可將它視爲是一種硬性選取(Hard Selection)的資料選取方法。另一方面，我們亦可將每個時間音框所計算出的正規化熵值作爲權重(Weight)，用來強調(Emphasized)或非強調(Deemphasized)此時間音框的重要性，我們將此方法視爲另一種柔性選取(Soft Selection)的資料選取方法，其數學式如下所示:

$$\gamma_{qm}^{z}(t) = \gamma_{qm}^{z}(t)\left(1 + \omega \cdot E_{z}(t)\right) \tag{40}$$

其中ω爲一比例控制參數。

六、實驗與討論

（一）實驗架構與設定

本論文所使用的大詞彙連續語音辨識器爲臺灣師大目前所發展的新聞語音辨識系統[19]，主要包括前端處理、詞彙樹複製搜尋(Tree-Copy Search)及詞圖搜尋(Word Graph Rescoring)[11]等部分。

在前端處理方面，本論文所採用的是異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)[20]。且在做完鑑別分析之後還額外使用最大化相似度現性轉換(Maximum Likelihood Linear Transform, MLLT)[21]，其目的是爲了配合目前我們在連續密度隱藏式馬可夫模型所使用的對角化(Diagonal)之共變異矩陣。同時，爲了降低通道效應對語音辨識的影響，在此使用倒頻譜正規化法(Cepstral Normalization, CN)。

在聲學模型方面，我們採用 151 個連續密度隱藏式馬可夫模型作爲中文 INITIAL-FINAL 的統計模型，而每個模型的狀態數分別爲 3 至 6 個不等，每個狀態皆爲高斯混合分布，其中每個高斯混合分布的分布個數分別爲 1 至 128 個不等，本論文總共使用到約 14,396 個高斯分佈。另一方面，本論文所使用的詞典約含有七萬二千個一至十字詞，並以從中央通訊社(Central News Agency, CNA) 2001 與 2002 年所收集到的約一億七千萬(170M)個中文字語料作爲背景語言模型訓練時的訓練資料[22]。在本文中的語言模型使用了 Katz 語言模型平滑技術[23]，在訓練時是採用 SRL Language Modeling Toolkit (SRILM)[24]。在詞彙樹搜尋時，本系統採用詞雙連語言模型；在詞圖搜尋時，則採用詞三連語言模型。

（二）實驗語料

本論文實驗使用的訓練與測試語料爲 MATBN 電視新聞語料庫[25]，是由中央研究院資訊所口語小組[26]耗時三年與公共電視台[27]合作錄製完成。我們初步地選擇採訪記者語料作爲實驗語材，其中包含 25.5 小時的訓練集(5,774 句)，供聲學模型訓練之用，其中男女語料各半；1.5 小時的評估集(292 句，共 26,219 字)，供辨識評估之用。訓練集由 2001 及 2002 年的新聞語料所篩選出來的；評估集則均爲 2003 年的語料，由中研院的評估語料篩選出來，只選擇了採訪記者語料並濾掉了含有語助詞之語句。

（三）實驗評估方式

本論文採用美國國家標準與技術中心(National Institute of Standards and Technology, NIST)所訂立的評估標準來進行正確轉譯詞序列與辨識詞序列的比較。此評估標準需

要使用動態規畫(Dynamic Programming)來做詞序列比對。然而因在中文中存在著斷詞不一致的問題，故在本文的實驗中皆是以字爲比對單位。令 H 爲正確轉譯詞序列與辨識詞序列比對後相同(Match)的字元個數、I 爲辨識詞序列多餘插入(Insertion)的字元個數、N 爲正確轉譯詞序列的字元總數，則語音辨識系統之正確率(Accuracy)的計算方式爲 $\frac{H-I}{N}\times 100\%$ ，錯誤率(Error Rate)則爲 1-正確率。本文的實驗數據中，皆是以字錯誤率(Character Error Rate, CER)來呈現實驗結果。

（四）基礎實驗結果

於基礎實驗中，先利用最大化相似度(ML)估測法訓練 10 次，所得到的字錯誤率(CER)爲 23.64% (記作 Baseline)。接著進行最小化音素錯誤(MPE)訓練 10 次，最後所得到的字錯誤率(CER)爲 20.77%。故於接下來的實驗中，皆以這組實驗爲比較對象。

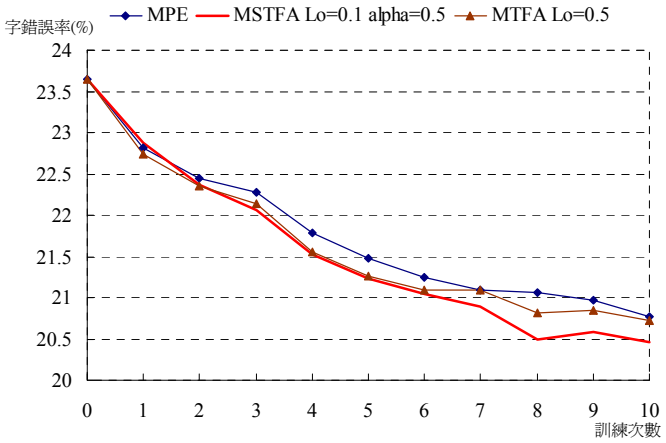


圖 3 時間音框正確率函數與最小化音素錯誤之比較結果

（五）改進最小化音素錯誤之實驗結果

本小節呈現本論文針對最小化音素錯誤訓練(MPE)的缺點而改進的最大化時間音框正確率函數(MTFA)訓練之實驗數據。事實上，本論文所提出的時間音框正確率函數並不是要減少刪除錯誤的個數(但實際上於共 26,219 字的評估集中，MPE 在第 10 次迭代訓練上有 359 個刪除錯誤，然而 MSTFA 則稍微減爲成 345 個)，而是去考量詞圖中某詞段受到刪除錯誤的影響，而減少其收集到的正確率統計值，以利聲學模型訓練之強健。實驗結果可參考表 1，其中刪除錯誤的懲罰權重(Penalty Weight)以 Lo(ρ)表示。由實驗數據顯示得知，在前幾次的迭代訓練中，時間音框正確率函數都會稍比最小化音素錯誤來得好。不同的刪除錯誤懲罰權重設定會有不同的刪除錯誤懲罰之效果。由表 1 的數據顯示，太大(ρ=0.8)或太小(ρ=0.1)的刪除錯誤懲罰權重設定，在一開始的迭代訓練上並無法得到明顯的效果，但都比最小化音素錯誤訓練來得佳，不過在第 10 次的迭代訓練上卻比最小化音素錯誤訓練的字錯誤率要高一點。最好的刪除錯誤懲罰權重設定(Lo=0.5)的時間音框正確率函數在第 10 次的迭代訓練上比最小化音素錯誤(MPE)訓練的辨識字錯誤率好 0.05%，相對字錯誤率將低約 0.1%，訓練次數 1 到 10 次的字錯誤率曲線圖請參考圖 3。

本論文亦使用一個常見的 S 型函數來平滑時間音框正確率函數，記作 MSTFA。其中 S 型函數有兩個參數可調整，在本實驗中只調整 α (alpha)，而 β 設爲零($\beta = 0$)。

實驗結果同樣參考表 1，實驗數據顯示出 MSTFA 在每次迭代訓練中都會比最小化音素錯誤(MPE)訓練來得好，最好的設定($\rho = 0.1$，$\alpha = 0.5$)在第 10 次的迭代訓練上可以比最小化音素錯誤的辨識字錯誤率好 0.31%，相對字錯誤率降低約 1.5%。

表 1 時間音框正確率函數之實驗結果

CER(%)	MPE	MTFA ρ =0.1	MTFA ρ =0.3	MTFA ρ =0.5	MTFA ρ =0.8	MSTFA ρ =0.1 α = 0.5	MSTFA ρ =0.5 α = 0.5	MSTFA ρ =0.1 α = 1	MSTFA ρ =0.5 α = 1
Baseline	23.64								
Itr01	22.82	22.85	22.73	22.74	22.80	22.88	22.82	22.83	22.77
Itr02	22.44	22.35	22.33	22.36	22.39	22.37	22.34	22.37	22.38
Itr03	22.28	22.07	22.13	22.14	22.19	22.06	22.10	22.02	22.05
Itr04	21.79	21.65	21.50	21.56	21.69	21.52	21.58	21.41	21.56
Itr05	21.48	21.26	21.14	21.26	21.34	21.23	21.47	21.3	21.52
Itr06	21.24	20.98	20.97	21.09	21.23	21.05	21.27	21.06	21.32
Itr07	21.10	20.91	20.87	21.09	21.19	20.89	21.11	20.80	21.19
Itr08	21.06	20.87	20.81	20.82	20.93	20.50	20.97	20.54	20.98
Itr09	20.97	20.84	20.74	20.85	20.90	20.58	20.82	20.57	21.03
Itr10	20.77	20.82	20.80	20.72	20.93	20.46	20.87	20.65	21.10

（六）資料選取方法之實驗結果

表 2 資料選取方法之實驗結果

CER(%)	MPE	MPE Random	MPE HS Thr=0.05	MPE HS Thr=0.08	MPE SS ω =1.0	MPE SS ω =0.5
Baseline	23.64					
Itr01	22.82	23.02	22.63	22.43	22.84	22.88
Itr02	22.44	22.62	22.05	21.80	22.40	22.43
Itr03	22.28	22.22	21.60	21.45	22.21	22.25
Itr04	21.79	22.16	21.40	21.34	21.65	21.73
Itr05	21.48	21.76	21.19	20.94	21.34	21.31
Itr06	21.24	21.66	20.92	20.82	21.33	21.18
Itr07	21.10	21.74	20.91	20.73	21.29	21.29
Itr08	21.06	21.62	21.22	20.74	21.00	21.06
Itr09	20.97	21.78	21.08	20.65	21.02	20.93
Itr10	20.77	21.84	21.29	20.63	20.94	20.89

本小節呈現資料選取方法於最小化音素錯誤(MPE)訓練之實驗結果。其中最小化音素錯誤的 I-平滑技術參數設定爲 10[28]。所使用的時間音框資料選取方法分爲硬性選取(HS)和軟性選取(SS)，其實驗結果皆可參考表 2。在軟性選取部分，所得到的結果跟基礎實驗結果不相上下，而其硬性選取最佳門檻值(記作 Thr)之設定爲 0.05(其時間音框總數爲 4,214,360 個，佔所有時間音框總數的 45.88%)。如實驗數據所顯示，資料選

取方法應用在最小化音素錯誤(MPE)訓練確實可以加快收斂速度，但在第 10 次的訓練上沒有明顯比最小化音素錯誤的字錯誤率來得低，其效果是差不多的。特別注意的是在 Thr=0.08 的這組實驗中，我們嘗試把門檻值隨著迭代訓練次數而遞減以企圖避免過度訓練之現象，所得結果亦符合我們所期望。故可以得知資料選取方法具有加快收斂速度之能力同時在第 10 次迭代訓練上與最小化音素錯誤訓練擁有差不多之結果。同時更說明了以正規化熵值爲基礎的資料選取方法確實能選出在事後機率定義域中離決定邊界較近的時間音框樣本，受惠於這些時間音框樣本本身比較具有鑑別力，故資料選取方法對於鑑別式訓練特別有幫助。

另一方面，吾人使用隨機選取(Random Selection)方法進行比較驗證以正規化熵值爲基礎的資料選取方法的確有效用的，而非亂選。實驗結果同樣參考表 2，其中隨機選取(記作 MPE Random)方法在每一次的迭代都隨機選取所有時間音框總數的 45.88%(與 MPE HS Thr=0.05 這組實驗的時間音框總數一致)。

（七）資料選取方法結合時間音框正確率函數之實驗結果

最後本小節將呈現資料選取方法於最大化 *S* 型時間音框正確率函數(MSTFA)之實驗結果。其最大化 *S* 型時間音框正確率函數的 I-平滑技術參數最佳化設定爲 10。所使用的時間音框資料選取方法爲硬性選取(HS)、軟性選取(SS)以及結合硬性和軟性選取(HS+SS)，實驗結果可參考表 3。其中最大化 *S* 型時間音框正確率函數的參數設定爲 $\rho = 0.1$、$\alpha = 0.5$；而各資料選取方法之參數設定皆列於表 3 中。由數據顯示得知，資料選取方法應用在最大化 *S* 型時間音框正確率函數依然保有加快收斂速度之成效，但在第 10 次迭代訓練上卻沒能比最大化 *S* 型時間音框正確率函數的字錯誤率來得低。此外，軟性資料選取方法比硬性選取效果來得好，在第 10 次迭代訓練上跟最大化 *S* 型時間音框正確率函數的字錯誤率差不多。最後，結合硬性與軟性選取(HS+SS)的實驗結果卻並沒有達到我們所預期的加成性效果。

表 3 資料選取方法結合時間音框正確率函數之實驗結果

CER(%)	MPE	MSTFA ρ=0.1 α = 0.5	MSTFA HS Thr=0.05	MSTFA SS ω=1.0	MSTFA HS+SS Thr=0.1 ω=0.5
Baseline	23.64				
Itr01	22.82	22.88	22.46	22.75	22.53
Itr02	22.44	22.37	21.87	22.25	21.72
Itr03	22.28	22.06	21.40	21.83	21.45
Itr04	21.79	21.52	21.38	21.45	21.38
Itr05	21.48	21.23	21.08	21.27	21.03
Itr06	21.24	21.05	21.03	20.94	20.90
Itr07	21.10	20.89	21.02	20.65	21.14
Itr08	21.06	20.50	21.15	20.78	21.14
Itr09	20.97	20.58	20.86	20.56	21.07
Itr10	20.77	20.46	21.43	20.86	21.37

七、結論與未來展望

鑑別式聲學模型訓練在大詞彙連續語音辨識的研究上一直扮演著重要的角色。本論文旨在改善最小化音素錯誤之聲學模型訓練，相關研究內容與成果可從下面兩個面向來作探討:

(1)首先，本論文提出了新的時間音框正確率函數來取代最小化音素錯誤訓練的原始音素正確率函數，進而充分地給予刪除錯誤適當的懲罰。在實驗結果上，最大化 *S* 型時間音框正確率函數(MSTFA)能達到比最小化音素錯誤(MPE)訓練約有1.5%的相對字錯誤率降低。

(2)其次，本論文提出以正規化熵值爲基礎之新的資料選取方法來改善鑑別式聲學模型訓練，由於正規化熵值是以給定某訓練語句的語音特徵向量序列中，某個狀態中的某個高斯分布出現的事後機率來求得的，所以可以視爲是在事後機率定義域中來選取訓練樣本，且所選出來的訓練樣本是比較混淆的，對鑑別式訓練來說，這些混淆的訓練樣本是較具有鑑別力的。根據初步的實驗結果顯示，此資料選取方法可以加快收斂速度，在前幾次的迭代訓練中，比最小化音素錯誤訓練有很大且一致的字錯誤率降低。最好的結果在第 6 次的迭代訓練上，比最小化音素錯誤訓練約有 1.5%的相對字錯誤率降低。

以全面風險爲基礎的鑑別式聲學模型訓練中，減損函數的設計一直都是一個重要的議題，如最流行的最小化音素錯誤訓練目標函數中，以類別比對爲基礎的原始音素正確率函數就還存在著改進的空間。已有學者提出以聲學模型間的關係來計算正確率以取代以類別爲基礎的正確率函數。類別比對爲基礎的音素正確率函數和聲學模型彼此間關係的減損函數皆各有其優缺點，未來吾人想要嘗試將這兩種不同的資訊結合，企圖改進鑑別式聲學模型訓練。

同時，吾人未來也想嘗試將以正規化熵值爲基礎的資料選取方法應用到其他的鑑別式訓練，如最小化分類錯誤、最小化貝氏風險鑑別式訓練等，以驗證此方法的一般性。事實上，由加最小化音素錯誤訓練的收斂速度來看，此以正規化熵值爲基礎之新的資料選取方法的確爲鑑別式聲學模型訓練提供了一個新的方向。

參考文獻

[1] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification, Second Edition*. New York: John & Wiley, 2000.

[2] Lalit R. Bahl, F. Jelinek and Robert L. Mercer, *A Maximum Likelihood Approach to Continuous Speech Recognition*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. PAMI-5, no.2, March 1983.

[3] J. Fiscus, *A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)*, in Proc. ASRU 1997.

[4] V. Goel and W. Byrne, *Minimum Bayes-Risk Automatic Speech Recognition*, Computer Speech and Language, Vol. 14, pp.115-135, 2000.

[5] F. Wessel, R. Schluter, K. Macherey and H. Ney, *Explicit Word Error Minimization Using Word Hypothesis Posterior Probability*, in Proc. ICASSP 2001.

[6] L. Mangu, E. Brill and A. Stolcke, *Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks*, Computer Speech and Language, Vol. 14, pp.373-400, 2000.

[7] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Ph.D Dissertation, McGill University, Montreal, 1991.

[8] J. Kaiser, B. Horvat and Z. Kacic, *Overall Risk Criterion Estimation of Hidden Markov Model Parameters*, Speech Communication, Vol. 38, pp.383-398, 2002.

[9] V. Doumpiotis, S. Tsakalidis and W. Byrne, *Lattice Segmentation and Minimum Bayes Risk Discriminative Training*, in Proc. Eurospeech 2004.

[10] D. Povey and P. C. Woodland, *Minimum Phone Error and I-smoothing for Improved Discriminative Training*, in Proc. ICASSP 2002.

[11] S. Ortmanns, H. Ney and X Aubert, *A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition*, Computer Speech and Language, Vol. 11, pp.11-72, 1997.

[12]D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D Dissertation, Peterhouse, University of Cambridge, July 2004.

[13] Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin and Berlin Chen, *Investigating Data Selection for Minimum Phone Error Training of Acoustic Models*, in Proc. ICME 2007.

[14] G. Heigold *et al*, *Minimum Exact Word Error Training*, in Proc. ASRU 2005

[15] A. J. Smola, P. Bartlett, B. Scholkopf and D. Schuurmans, *Advances in Large Margin Classifiers*, The MIT Press.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[17] Xinwei Li, Hui Jiang and Chaojun Liu, *Large Margin HMMs for Speech Recognition*, in Proc. ICASSP 2005.

[18] Jinyu Li, Ming Yuan and Chin-Hui Lee, *Soft Margin Estimation of Hidden Markov Model Parameters*, in Proc. ICSLP 2006.

[19] B. Chen, J. W. Kuo and W. H. Tsai, *Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription*, in Proc. ICASSP 2004.

[20] N. Kumar, *Investigation of Silicon-Auditory Models and Generalizaion of Linar Discriminant Analysis for Improved Speech Recognition*, Ph.D. Thesis, John Hopkins University, Baltimore, 1997.

[21] R. A. Gopinath, *Maximum Likelihood Modeling with Gaussian Distributions*, in Proc. of ICASSP 1998.

[22] LDC: Linguistic Data Consortium, http://www.ldc.upenn.edu

[23] S. M. Katz, *Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer*, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 35, No.3, pp. 400-401, 1987.

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

[25] H. M. Wang, B. Chen, J.-W. Kuo, and S.S. Cheng. *MATBN: A Mandarin Chinese Broadcast News Corpus*, International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 2, pp. 219-236, 2005.

[26] SLG: Spoken Language Group at Chinese Information Processing Laboratory, Institute of Information Science, Academia Sinica. http://sovideo.iis.sinica.edu.tw/SLG/index.htm

[27] PTS: Public Television Service Foundation. http://www.pts.org.tw

[28]郭人瑋, *最小化音素錯誤鑑別式聲學模型學習於中文大詞彙連續語音辨識之初步研究*, Master Thesis, NTNU, 2005.

[29]劉士弘, *改善鑑別式聲學模型訓練於中文連續語音辨識之研究*, Master Thesis, NTNU, 2007.

端點偵測技術在強健語音參數擷取之研究
Study of the Voice Activity Detection Techniques for Robust Speech Feature Extraction

杜文祥 Wen-Hsiang Tu
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s94323537@ncnu.edu.tw

洪志偉 Jeih-weih Hung
暨南國際大學電機工程學系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

由於發展環境和應用環境兩者之間的不匹配，導致於語音辨識系統效能經常會下降，而引起這不匹配的主要原因之一是加成性雜訊，處理加成性雜訊的方法我們可以分成三類，語音強化法、強健性語音特徵參數、以及語音模型調適法，而本論文所討論的方法主要是屬於強健性語音特徵參數之技術。

在本論文中，我們主要的重點在於探討不同的信號特徵對於語音端點偵測的影響，所利用的特徵分別爲低頻帶頻譜強度、全頻帶頻譜強度、累積量化頻譜、以及高通對數能量等。利用以上這些不同的特徵進行語音之端點偵測，所得之純雜訊的位置資訊可以提供頻譜消去法與靜音對數能量正規化法中所需的雜訊頻譜或能量的估測。

在實驗環境上我們採用 Aurora2 語料庫，在八種背景雜訊以及訊雜比 0~20dB 下做實驗。在第五章中所呈現的實驗數據與分析可證明以上所述的各種特徵顯然可用以有效的鑑別出一段語音中純雜訊部分與語音部分，使之後所使用的頻譜消去法與靜音對數能量正規化法等強健性語音特徵技術，得以明顯提升在雜訊環境下語音辨識的精確度，增加語音辨識系統的強健性。

關鍵詞：端點偵測法，能量特徵，頻譜消去法，自動語音辨認

Abstract

The performance of a speech recognition system is often degraded due to the mismatch between the environments of development and application. One of the major sources that give rises to this mismatch is additive noise. The approaches for handling the problem of additive noise can be divided into three classes: speech enhancement, robust speech feature extraction, and compensation of speech models. In this thesis, we are focused on the second class, robust speech feature extraction.

The approaches of speech robust feature extraction are often together with the voice activity detection in order to estimate the noise characteristics. A voice activity detector (VAD) is used to discriminate the speech and noise-only portions within an utterance. This thesis

primarily investigates the effectiveness of various features for the VAD. These features include low-frequency spectral magnitude (LFSM), full-band spectral magnitude (FBSM), cumulative quantized spectrum (CQS) and high-pass log-energy (HPLE). The resulting VAD offers the noise information to two noise-robustness techniques, spectral subtraction (SS) and silence log-energy normalization (SLEN), in order to reduce the influence of additive noise in speech recognition.

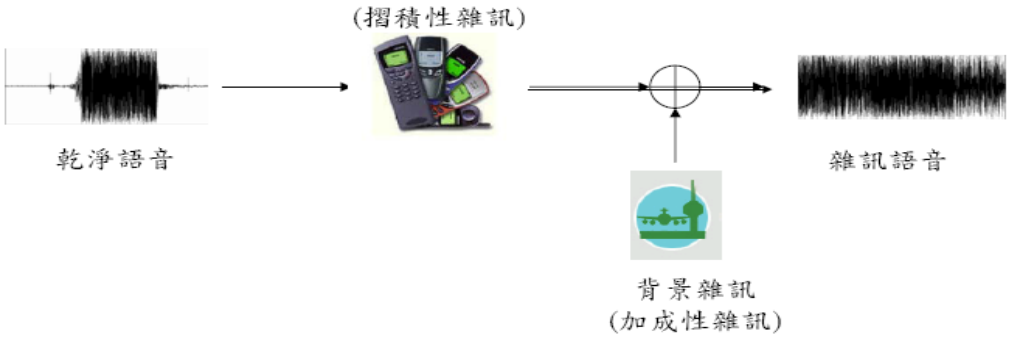
The recognition experiments are conducted on Aurora-2 database. Experimental results show that the proposed VAD is capable of providing accurate noise information, with which the following processes, SS and SLEN, significantly improve the speech recognition performance in various noise-corrupted environments. As a result, we confirm that an appropriate selection of features for VAD implicitly improves the noise robustness of a speech recognition system.

Keywords：voice activity detection, spectral magnitude, spectral subtraction, speech recognition

一、緒論

在我們的生活環境中，影響語音辨識結果的因素很多。其中一重要的因素爲語音辨識系統訓練與應用環境上的不匹配(environmental mismatch)，此不匹配的相關因素又包含了加成性雜訊(additive noise)、摺積性雜訊(convolutional noise)以及頻寬限制(bandwidth limitation) 等因素。其中加成性雜訊也可說是背景雜訊，這是因爲語音辨識系統所處在的環境，並非都像實驗室毫無其他干擾雜訊。也許系統是處於地鐵站中、餐廳、機場等這些具有其他干擾源的環境。甚至旁人呼吸喘息聲音都會混進語音裡面，造成辨識率的降低。摺積性雜訊也稱爲通道雜訊或是通道失真，主要是因爲麥克風的不同、傳輸線材的遮蔽效應不好而受外在電磁波影響所造成的。頻寬限制也是因爲收音通道的差異所帶來的影響。後面兩項的因素在電話語音辨識系統就非常的明顯。在有限頻寬的電話線，把通話者頻寬做限制以便利傳輸，這往往會造成語者的聲音變調，甚至通道失真的影響還會造成使用者兩端會發生吱吱的雜音，造成語音辨識很大的困擾。

爲了改善以上所述之環境上的不匹配，有眾多學者提出各種改進的方法，其中一類的方法爲強健性語音特徵參數技術，強健性語音特徵參數技術的主要目的是在抽取出不容易受到外在環境干擾而失真的語音特徵參數，進而突顯出語音變化的部分。在許多針對加成性雜訊所發展的強健性語音特徵參數技術裡，如何取得雜訊部分的資訊是很重要的，亦即在一段語音訊號中，我們通常必須偵測純雜訊所在的位置，以利於雜訊資訊的取得，其相關的技術則統稱語音活動偵測法(voice activity detection, VAD)，或簡稱爲端點偵測法(endpoint detection)。本論文的重點即在發展一系列使用端點偵測法的信號特徵，藉由這些特徵，使端點偵測的結果更精確，進而使之後的強健性語音特徵參數技術能達到更好的效果。在論文中。我們將介紹幾個用以語音偵測的信號特徵，分別爲低頻帶頻譜強度(low-frequency spectral magnitude, LFSM)、全頻帶頻譜強度(full-band spectral magnitude, FBSM)、累積量化頻譜(cumulative quantized spectrum, CQS)、以及高通對數能量(high-pass log-energy, HPLE)等。我們嘗試把這些端點偵測的方法與兩種強健式語音特徵參數擷取法結合，即頻譜消去法(spectral subtraction)與靜音對數能量正規化法(silence log-energy normalization, SLEN)等，發現皆有相當程度的提高辨識效率。



圖一、雜訊干擾語音之示意圖

本論文其餘部分共分爲五章，其中第二章詳細介紹所提出之端點偵測的各種信號特徵，第三章介紹本論文所用的兩種強健語音特徵技術，第四章爲實驗環境的設定，第五章爲實驗結果與討論，最後，第六章則包含了簡要的結論。

二、端點偵測所使用之信號特徵

端點偵測法(endpoint detection)或稱爲語音活動偵測法(voice activity detection, VAD)是指可以將一段語音中雜訊與語音的位置偵測出來的演算法。藉由一個有效的端點偵測法，我們可以利用所求得的純雜訊音框，準確的估測雜訊的資訊，例如其頻譜能量等。進而促成各種強健技術的使用，以達到雜訊的抑制，降低雜訊對語音訊號的影響。在以下幾節，我們將介紹用以端點偵測的各種信號特徵。

（一）低頻帶頻譜強度(low-frequency spectral magnitude, LFSM)

我們觀察到無論任何種類的雜訊，在頻帶[0,50Hz]之間都有相當比例的能量。同時，語音在此低頻帶的能量也具有一定程度的比例。因此我們根據此頻譜上的特性，去計算每個音框在此低頻帶的頻譜強度。根據此強度值，判斷此語音音框是否爲純雜訊音框，或是包含語音的音框。

首先，我們假設 $\{x_m[n], 1 \le n \le N\}$ 是語音訊號的第 m 個音框，將其取 K $(K \ge N)$ 點的離散傅立葉轉換，我們將可得到此音框所對應之頻譜如下式(1)：

$$X^{(m)}(f_k) = X^{(m)}[k] = \sum_{n=0}^{N-1} x_m[n] e^{-j\frac{2\pi nk}{K}}, \quad 0 \le k \le K-1 \tag{1}$$

其中 f_k 爲頻率，其值如下式：

$$f_k = \frac{F_s}{2K} k \tag{2}$$

其中 F_s 爲取樣頻率，因此我們定義出頻帶 $[F_L, F_U]$ 之頻譜強度計算方式爲：

$$Y^{(m)}_{[F_L,F_U]} = \sum_{F_L \le f_k \le F_U} \left| X_m(f_k) \right| \tag{3}$$

根據式(3)，我們可以計算每一音框之低頻帶頻譜強度，即 0 至 50Hz 以內的低頻帶頻譜強度如下：

$$Y^{(m)}_{Low} = Y^{(m)}_{[0,50]} = \sum_{0 \le f_k \le 50} \left| X_m(f_k) \right| \tag{4}$$

接著我們以一段語音前 P 個音框之低頻帶頻譜強度的平均爲參考值，其計算如下：

$$\theta = \lambda\left(\frac{1}{P}\sum_{m=1}^{P} Y_{Low}^{(m)}\right) \tag{5}$$

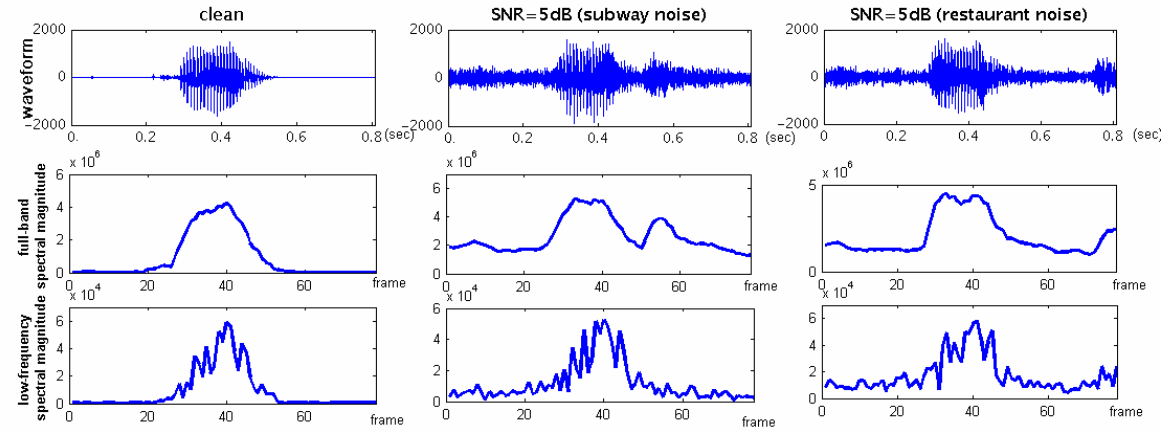
其中P表初步假設爲純雜訊之音框數。對於一段語音而言，前幾個音框通常是非語音的純雜訊音框，所以一開始我們假設前P個音框代表純雜訊。接著我們將每一個音框低頻帶內的頻譜強度$Y_{Low}^{(m)}$與門檻值θ做比較，若前者小於後者，則其歸類爲純雜訊音框，反之則爲語音音框。端點偵測的判斷式如下：

$$\text{第}m\text{個音框為}\begin{cases}\text{純雜訊音框，若 } Y_{Low}^{(m)} \le \theta \\ \text{語音音框 ，若 } Y_{Low}^{(m)} > \theta\end{cases} \tag{6}$$

（二）全頻帶頻譜強度(full-band spectral magnitude, FBSM)

利用全頻帶頻譜強度特徵進行端點偵測，其作法類似前一節，不同的是，我們並不針對特定頻帶來估測雜訊。假設取樣頻率爲 8kHz，我們是把每個音框的全部頻帶[0, 4kHz]的頻譜強度全部都考慮，如式(7)，我們計算每一音框全頻帶的頻譜強度：

$$Y_{Full}^{(m)} = Y_{[0,4000]}^{(m)} = \sum_{0 \le f_k \le 4000} \left|X_m(f_k)\right| \tag{7}$$



圖二、語音波形及頻譜強度分佈圖

圖二所示的三段語音，分別爲乾淨語音、受地下鐵雜訊干擾的語音(SNR=5dB)及受餐廳雜訊干擾的語音(SNR=5dB)。從語音波形及頻譜強度分佈圖看來，對於語音訊號我們可以很成功的區隔出語音的部分與非語音部分。但是在 SNR 很小的情況下，區隔語音和非語音部分變得比較困難。不同於[1]所提到，語音幾乎不分佈於[0,50Hz]的低頻帶，在第三列的低頻帶頻譜強度分佈圖中我們可看到，語音在此頻帶仍占有相當的比例，同時可看到此頻帶的頻譜強度可以有效用來區隔語音和非語音部分。觀察第二行的地鐵雜訊下的語音，大約在 0.5 到 0.6 秒之間所出現的較大能量之附加雜訊，但低頻帶頻譜強度幾乎不受其影響。而觀察第三行之餐廳雜訊下的語音，大約在 0.7 到 0.8 秒之間，也是出現頗像較大能量的附加雜訊，但我們發現，此附加雜訊在低頻帶頻譜強度的影響並不大。第二列的全頻帶頻譜強度可得正對純雜訊與語音音框也有很好的鑑別效果，爲其較容易受到較高頻譜強度影響而造成誤判。

（三）累積量化頻譜法(cumulative quantized spectrum, CQS)

在前兩小節裡，我們分別利用了低頻帶與全頻帶的頻譜強度來做爲端點偵測的標準，所用到的頻譜強度爲所用的頻帶內之每個頻率對應之強度總和。在這裡我們提出另一種利用頻譜性質來完成端點偵測的方法，所用的特徵稱爲累積量化頻譜(cumulative quantized spectrum, CQS)。其基本論點爲，將純雜訊與雜訊語音之離散頻譜比較，我們可以發現純雜訊音框所含較高強度之頻率個數，通常比雜訊語音音框之高強度的頻率個數少，因此我們可以藉由累積一音框中高強度之頻率的個數所得，來判斷此音框的種類。因爲這樣的做法，相當於將每個頻率的強度做量化(高者爲 1，低者爲 0)，再將這些量化後的強度值加總，因此我們稱所得之特徵值爲累積量化頻譜。我們假設每段語音的前 P 個音框爲純雜訊音框，利用此 P 個音框之強度頻譜的平均，我們定義每一頻率之強度的門檻值爲：

$$\theta(k)=\frac{1}{P}\sum_{m=1}^{P}\left|X_m(k)\right| \tag{8}$$

其中 $X_m(k)$ 表示第 m 個音框訊號取 N 點 DFT 後之頻譜。我們將每一個音框之離散頻譜強度 $\left\{\left|X_m(k)\right|,\ k=0,\ 1,\ 2,\ .....,\ \frac{N}{2}\right\}$ 與所得之門檻值 $\left\{\theta(k),\ k=0,\ 1,\ 2,.....,\ \frac{N}{2}\right\}$ 比較大小，所得到量化後的頻譜如下：

$$Y_m(k)=\begin{cases}1 & if\ \ \left|X_m(k)\right|>\theta(k)\\ 0 & if\ \ \left|X_m(k)\right|\le\theta(k)\end{cases}\quad k=0,\ 1,\ 2,.....,\ \frac{N}{2} \tag{9}$$

接著將每一音框的量化頻譜 $\left\{Y_m(k)\right\}$ 做累加，得：

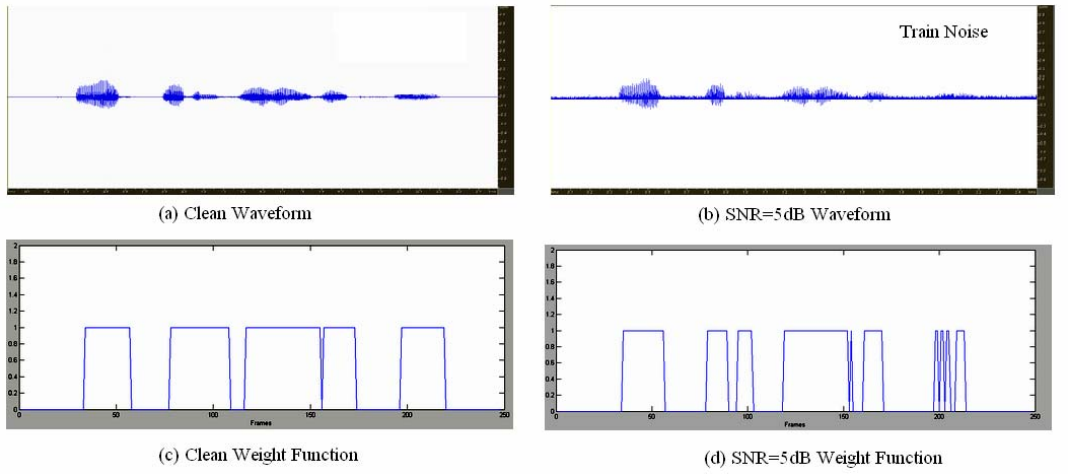
$$Z_m=\sum_{k=0}^{\frac{N}{2}}Y_m(k) \tag{10}$$

因此所得之累積量化頻譜 Z_m 即爲第 m 個音框中，強度大於門檻值的頻率個數，Z_m 越大，則意味著此音框有越大的機率是能量較大的音框(語音音框)；反之，則此音框爲一純雜訊音框。最後我們爲累積量化頻譜 Z_m 定一門檻值 $\frac{N}{4}$，即約爲頻率點數的一半，當 Z_m 小於 $\frac{N}{4}$ 時，則此音框歸類爲純雜訊音框；反之，則爲語音音框：

$$\text{第}m\text{個音框為}\begin{cases}\text{語音音框，若 } Z_m\ge\frac{N}{4}\\ \text{純雜訊音框，若 } Z_m<\frac{N}{4}\end{cases} \tag{11}$$

圖三所示的兩段語音，分別爲乾淨語音與受車站雜訊干擾的語音(SNR=5dB)。第一列兩圖爲語音的波形圖(圖三(a)與(b))，而第二列兩圖(圖三(c)與(d))意義爲音框判定的結果，若高強度之頻率的個數占多數，則爲雜訊語音音框，在圖中以 1 表示；反之，若高強度之頻率個數占少數，則此音框判定爲雜訊音框，圖中以 0 表示。在第一行的兩個組圖(圖七(a)與(c))，我們發現以累積量化頻譜分佈對於乾淨的語音中，對於語音音框與非語音音框，具有著很良好的辨別度。而在於 SNR 較小的語音環境(圖三(b)與(d))，在訊號較後段部分，其語音成分幾乎要被背景雜訊給覆蓋過去，但是經過累積量化頻譜做判定處理後，我們可以發現其對於鑑別出純雜訊部分以及雜訊語音部分比起完全乾淨的語音判定結

果(圖三(c))只是略差一些，對於語音部分整體來說並沒有太大的遺漏或是誤判。



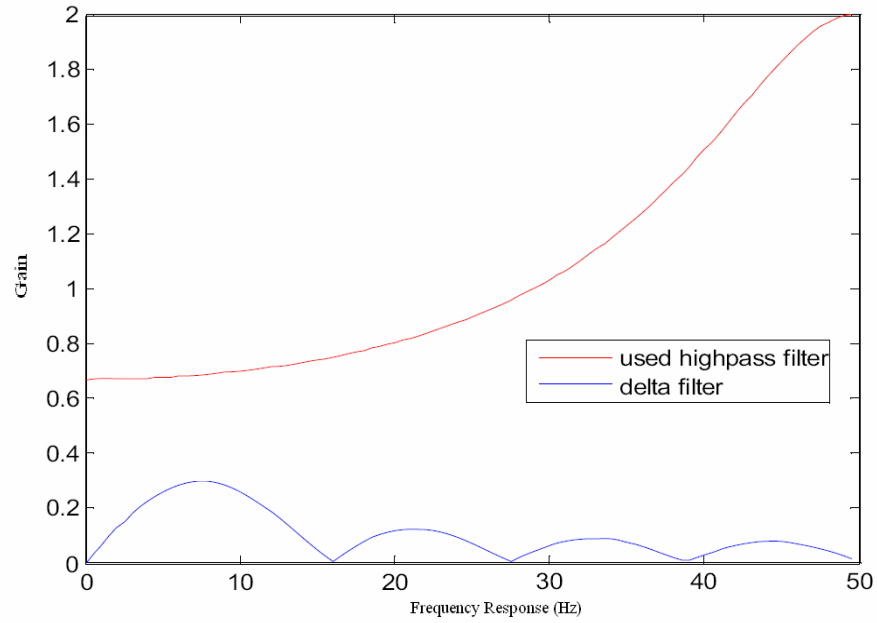
圖三、語音波形及累積量化頻譜法之權重圖

（四）高通對數能量(high-pass log-energy, HPLE)

能量大小一向是判定一音框是否爲雜訊或語音的重要指標，在[2-6]等諸多文獻裡提到音框能量或其他變化的形式(如對數能量，能量的差分等)可用作端點偵測的主要特徵。根據我們的觀察發現，將音框對數能量值通過一高通的時間序列濾波器，所得到之高通對數能量，可有效鑑別純雜訊與語音音框。相對於在一般使用之能量差分法中，差量濾波器(delta filter)捨棄能量之調變頻譜高頻成分，我們發現其實高頻成分也是包含著很多重要的語音資訊。因此我們所使用一高通時間序列濾波器來對能量做處理：我們做一無限脈衝響應(infinite impulse response, IIR)的高通時間序列濾波器來處理一連串的音框對數能量，其輸入與輸出的關係式爲：

$$E[n] = \frac{1}{2}\left(e[n+1] - E[n-1]\right) \tag{12}$$

其中$E[n]$是每個音框更新後的對數能量值，而$e[n]$爲每個音框的原始對數能量值。此高通濾波器之頻率響應如圖三。我們從圖中可知道，此高通時間序列濾波器並沒特別抑制低頻部分，而在高頻部分卻有著放大的效果。



圖三、高通濾波器與差量濾波器振幅頻率響應圖

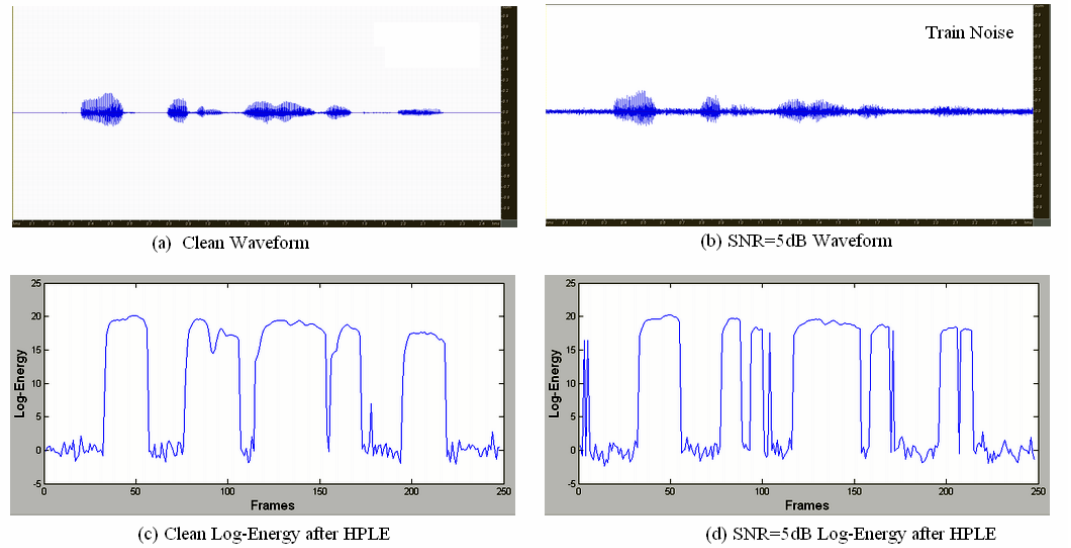
我們令整段語音音框高通對數能量的平均爲一門檻值，計算如下：

$$T = \frac{1}{N}\sum_{n=1}^{N} E[n] \tag{13}$$

其中N爲音框總數。語音音框與純雜訊音框判定的原則爲：

$$\text{第}n\text{個音框為一}\begin{cases}\text{純雜訊音框，若} & E[n] < T \\ \text{語音音框，若} & E[n] \geq T\end{cases} \tag{14}$$

圖四爲乾淨語音與雜訊語音之波形與高通對數能量圖，圖四(a)與(c)表示出，在乾淨情況下，高通對數能量法對於非語音與語音具有很不錯的鑑別力。而當我們在 SNR=5dB 時的雜訊環境，我們由圖四(b)與(d)發現，整段訊號末端的語音部分，幾乎被雜訊給覆蓋過去，然而經過高通對數能量法處理過後，這部分的語音有被鑑別出來，惟一誤判部分是在剛開始的一小部分。大體來說，整體語音部分幾乎都有被鑑別出來，因此可看出此方法的效能。



圖四、語音波形與其高通對數能量圖

三、強健式語音特徵參數擷取技術

本章節中，我們將介紹兩種強健性語音特徵擷取的方法，將語音中估測的雜訊消除，以還原乾淨的語音特徵。

（一）非線性頻譜消去法 (nonlinear spectral subtraction, NSS)

頻譜消去法[7-8]主要目的是估測出加成性雜訊的頻譜分佈，再將被附加雜訊干擾的語音訊號頻譜扣除所估測出的雜訊頻譜，以還原原始的乾淨語音頻譜。使用頻譜消去法中，有兩個基本假設：

（1）乾淨語音訊號與雜訊訊號在統計上是無關的(uncorrelated)，並且在時域 (time domain) 上是可線性加成的；

（2）雜訊訊號相對乾淨語音訊號而言是變化較爲緩慢的。

根據這兩個假設，如果我們要得到乾淨語音訊號，通常必須從非語音的區域估測出雜訊的頻譜，再將受雜訊干擾的語音頻譜減去雜訊的頻譜，式(15)說明了雜訊語音訊號、乾

淨語音訊號和雜訊訊號的關係：

$$y_i(t) = x_i(t) + n_i(t) \xrightarrow{\text{Fourier Transform}} Y_i(f) \approx X_i(f) + N_i(f) \tag{15}$$

其中$y_i(t)$、$x_i(t)$與$n_i(t)$分別代表第i個音框的雜訊語音訊號、乾淨語音訊號以及雜訊訊號，而$Y_i(f)$、$X_i(f)$與$N_i(f)$則是$y_i(t)$、$x_i(t)$與$n_i(t)$的強度頻譜(magnitude spectrum)值。

因此，理想上，我們若能精確得到雜訊之強度頻譜$N_i(f)$，則可從$Y_i(f)$直接扣除$N_i(f)$而得到乾淨語音強度頻譜$X_i(f)$。實際上，$N_i(f)$通常無法十分精確的估測，這會導致一個問題：若強度頻譜$Y_i(f)$比估測之雜訊訊號強度頻譜$N_i(f)$還小時，相減的結果會得到一個負值，而乾淨語音強度頻譜是不應該出現負值的，因此解決這問題的方法之一就是當估計到的乾淨語音強度頻譜值爲負時，我們就以一個極小的值代替，這樣的方法我們稱爲非線性頻譜消去法[8]，如式(16)所示：

$$X_i(f) = \begin{cases} Y_i(f) - \alpha N(f) & if \;\; Y_i(f) > N(f) \\ \beta Y_i(f) & otherwise \end{cases} \tag{16}$$

其中$N(f)$是估測而得的雜訊強度頻譜，α是過度估測因子(over-estimation factor)，用來控制估測雜訊功率頻譜被減去的程度，而β是底限因子(flooring factor)。在本論文中，$N(f)$是利用前一章之端點偵測法所得之純雜訊音框強度頻譜的平均，公式如下：

$$N(f) = \frac{1}{M}\sum_{j=1}^{M} N_j(f) \tag{17}$$

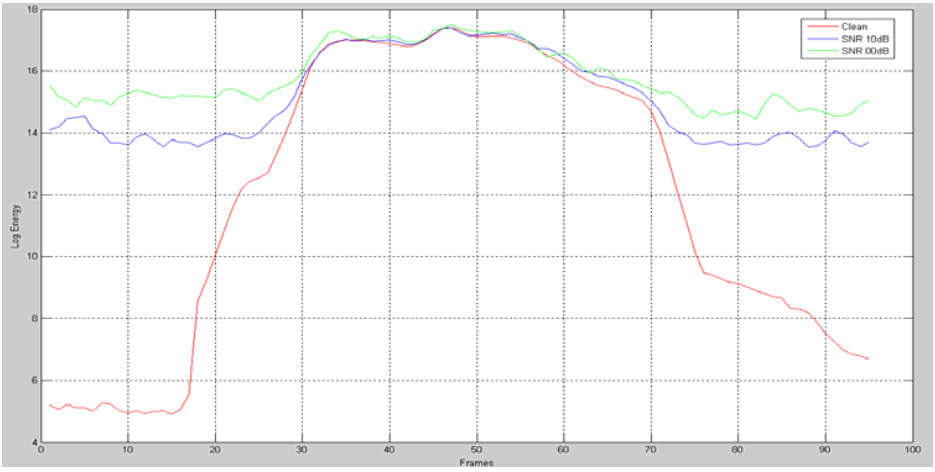
其中$N_j(f)$爲第j個被標示爲純雜訊音框之強度頻譜，M爲純雜訊音框總數。

（二）靜音對數能量正規化法(silence log-energy normalization, SLEN)

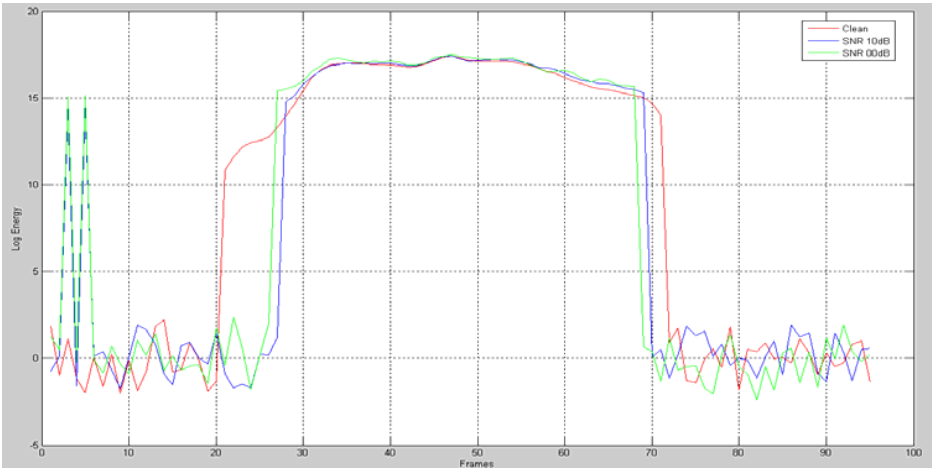
靜音對數能量正規化法[9]的原理在於觀察發現，在雜訊語音的能量曲線上，受雜訊干擾較爲嚴重部分爲能量小的波谷，而能量大的波峰部分比較不受影響。而且許多實驗發現，只保留能量波峰部分的音框，而捨棄能量小的音框，經由辨識依然可以得到不錯的辨識率。此外我們假設，對能量特徵而言，最重要的是原始語音能量的變化曲線之波形，是否被完整保留。也就是說一段語音整體的能量波形，比單一音框之能量的降低失真距離重要。若是保留原始整體能量波形的完整，即使原始語音能量曲線波形有小幅度的位移，最後的辨識效果也不會相差太多。依照上述的原理，我們找出能量曲線波形中非語音的部分，並且把它正規化處理過後，乾淨的語音訊號能量曲線波形將會與受雜訊影響的語音能量曲線波形十分的相似。根據四之一節的各種端點偵測法，我們將判定爲純雜訊音框之對數能量正規化爲一極小值，而語音音框的對數能量則維持不變：

$$\hat{E}[n] = \begin{cases} E[n] & \text{若第}n\text{個音框為語音音框} \\ \varepsilon & \text{若第}n\text{個音框為純雜訊音框} \end{cases} \tag{18}$$

其中ε爲一極小值。我們利用圖五來觀察靜音音框對數能量正規化法的作用。



圖五(a) 一段語音之原始對數能量軌跡圖



圖五(b) 一段語音經靜音音框對數能量正規化法後之對數能量特徵軌跡圖

由圖五可以很明顯發現，雜訊語音(SNR=10dB，SNR=0dB)之兩條對數能量曲線在較低值處與乾淨能量的曲線有很大的不匹配(圖五(a))，經由靜音對數能量正規化之後，此不匹配的現象改善了許多(圖五(b))。因此，靜音音框對數能量正規化法能有效降低受雜訊污染的語音與乾淨語音在能量上之不匹配，減少雜訊對語音能量的干擾。

五、實驗設定

(一) 語音資料庫簡介

本論文所使用的語音資料庫爲歐洲電信標準協會(European Telecommunication Standard Institute，ETSI)發行的 AURORA2 語音資料庫[10]，它是一套連續的英文數字字串，內容是以美國成年男女所錄製的乾淨環境連續數字，再加上雜訊與通道效應。加成性雜訊共有八種，分別爲地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，前四種歸類爲 Set A，後四種歸類爲 Set B。其訊雜比(signal-to-noise ratio, SNR)則有七種，分別爲 20dB, 15dB, 10dB, 5dB, 0dB. -5dB 與完全乾淨狀態。

(二) 特徵參數的設定與辨識系統的訓練

在本論文的語音辨識實驗中，我們使用的特徵參數包含了12維梅爾倒頻譜參數與1維對數能量，附加上其一階與二階差量（在部分實驗中，我們會省卻對數能量部分）。

特徵參數抽取之詳細設定，如表一所示。對於每個欲辨識的數字模型而言，本論文使用隱藏式馬可夫模型工具(hidden Markov model toolkit，HTK)來訓練，包含11個數字模型(0~9以及oh 11個數字模型)以及靜音模型，每個數字模型包含10個狀態，各狀態包含4個高斯密度混合。隱藏式馬可夫模型是一種運用統計理論推導出來的模型，用來描述語音產生的過程，相當適合用在連續語音的辨認。HMM有很多種類型，本論文採用由左到右的形式，也就是每個狀態在下一個時間只能跳到此刻狀態或下一個鄰近的狀態，隨著時間的增加，狀態由左至右依序轉移。另外，模型中的狀態觀測機率函數是選用連續式的高斯混合機率密度函數(Gaussian Mixture probability density function，簡稱GM)，因此我們也稱此模型爲連續密度隱藏式馬可夫模型(continuous density HMM，簡稱CDHMM)。

表一 本論文實驗所使用特徵參數抽取之設定

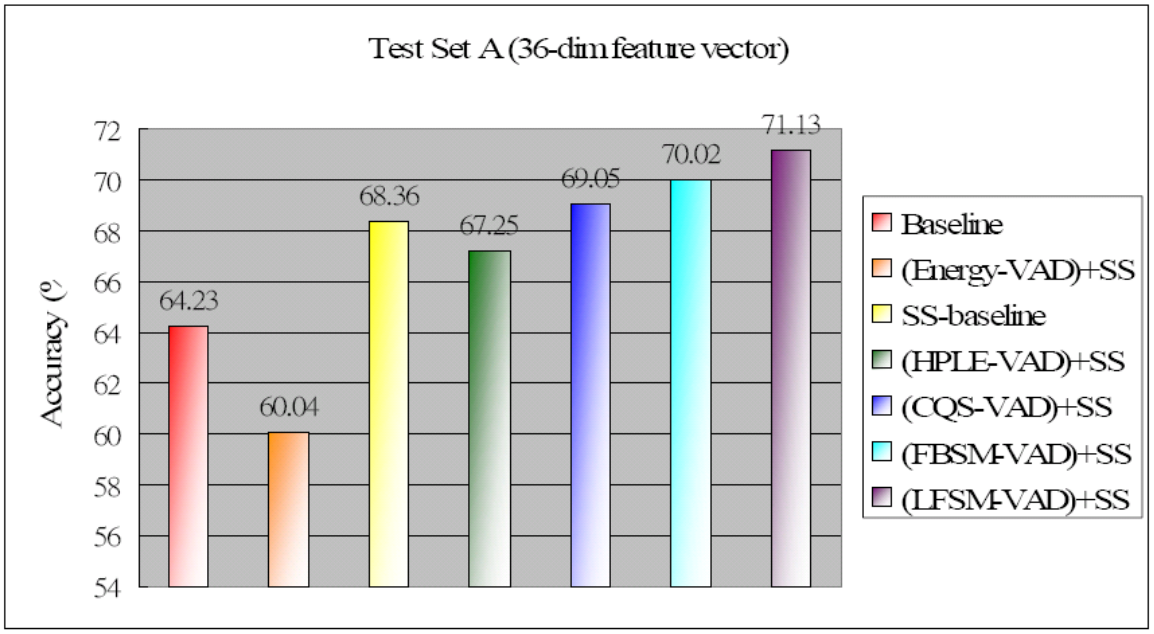
取樣頻率	8000 Hz
音框長度(frame size)	25 ms
音框平移(frame shift)	10 ms
預強調濾波器	1-(0.97) z^{-1}
視窗形式	漢明窗(Hamming window)
快速傅立葉轉換點數	256 點
濾波器組	梅爾刻度三角濾波器組(Mel-scaled triangular filter bank)，共 23 個濾波器
語音特徵參數	13 維 MFCCs(含對數能量)+Δ13 維 MFCCs +ΔΔ13 維 MFCCs，共 39 維

五、 端點偵測法配合強健性語音技術之實驗結果

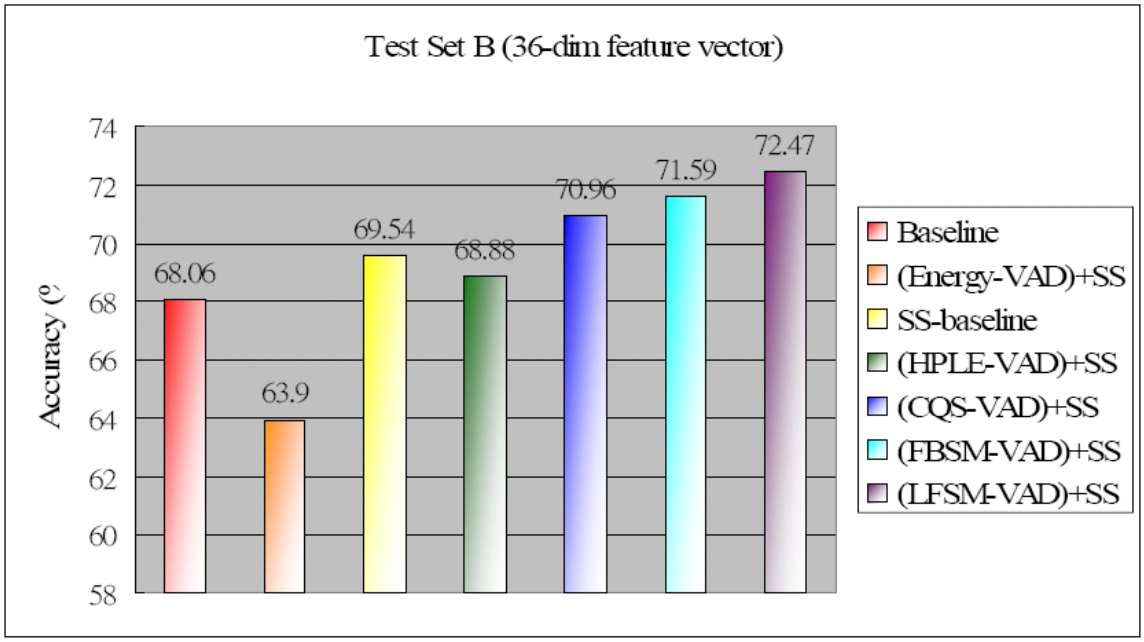
本章將會介紹第四章所有端點偵測法配合強健式語音的辨識實驗結果。我們將實驗區分爲「省略能量維特徵參數」以及「加入能量維特徵參數」這兩種，藉此分析能量維對於語音辨識上的影響。實驗結果可以證明本論文所提到之方法幾乎都可提升雜訊環境下語音辨識率，降低雜訊對語音的干擾。

（一）梅爾倒頻譜特徵參數之實驗結果

本章節所有實驗所使用的特徵參數爲 12 維梅爾倒頻譜參數及其一階和二階差量，總共爲 36 維特徵參數。圖六與圖七分別爲 A 組環境與 B 組環境下各種方法所得之平均辨識率。其中「baseline」是指沒有處理過的原始特徵參數、「SS-baseline」是指利用每段語句前五個音框作爲純雜訊音框所作的頻譜消去法，「(Energy-VAD)+SS」、「(HPLE-VAD)+SS」、「(CQS-VAD)+SS」、「(FBSM-VAD)+SS」與「(LFLE-VAD)+SS」則分別爲以音框能量、高通對數能量、累積量化頻譜、全頻帶頻譜強度與低頻帶頻譜強度作爲端點偵測特徵，執行端點偵測並與頻譜消去法作結合。



圖六、A 組環境下平均辨識率(%)比較圖



圖七、B 組環境下平均辨識率(%)比較圖

從圖六與圖七的辨識結果，我們有以下幾點的發現：

1. 以傳統的能量特徵作爲端點偵測的信號特徵，其效果不盡理想，其得到的端點偵測之結果配合頻譜消去法，所得到的辨識率甚至比基礎實驗還差。
2. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之頻譜消去法，相較於基礎實驗約可得到 2-4%的平均進步率。
3. 本論文所提出的四種端點偵測的特徵(HPLE, CQS, FBSM 與 LFSM)應用於端點偵測，配合頻譜消去法之下，都能得到明顯的進步。其中以高通對數能量(HPLE)表現稍差，但仍比傳統之能量特徵來的好。而這四種特徵又以低頻帶頻譜強度(LFSM)表現最

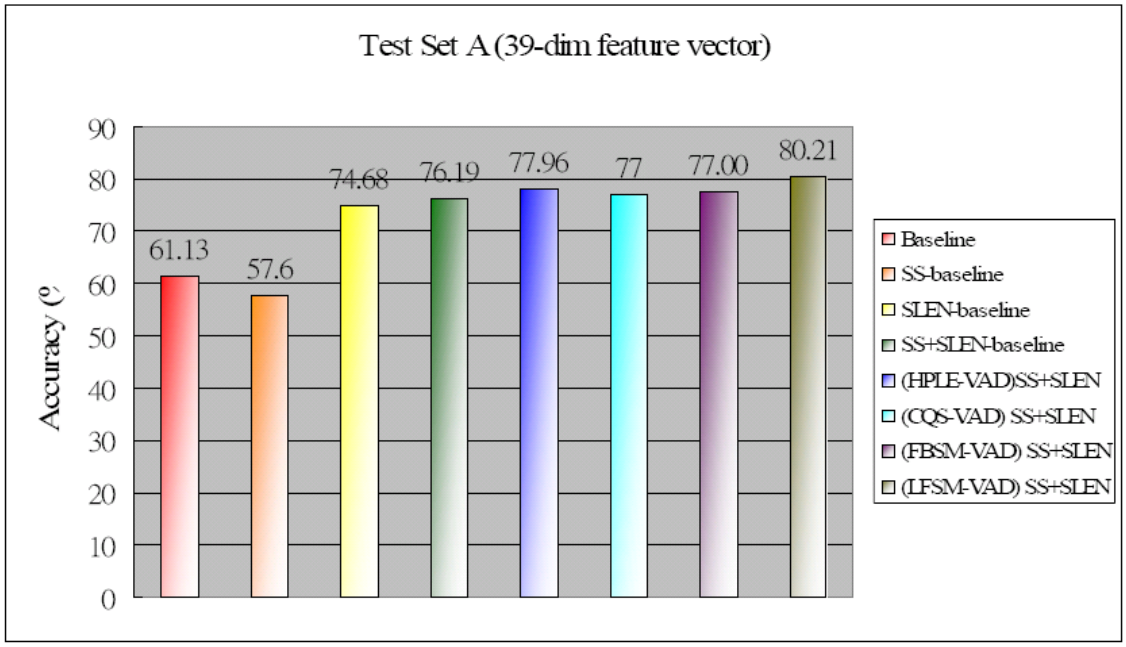
好，相較於基礎實驗約可得到 4-7%的平均進步率。

（二）梅爾倒頻譜系數與對數能量之特徵參數的實驗結果

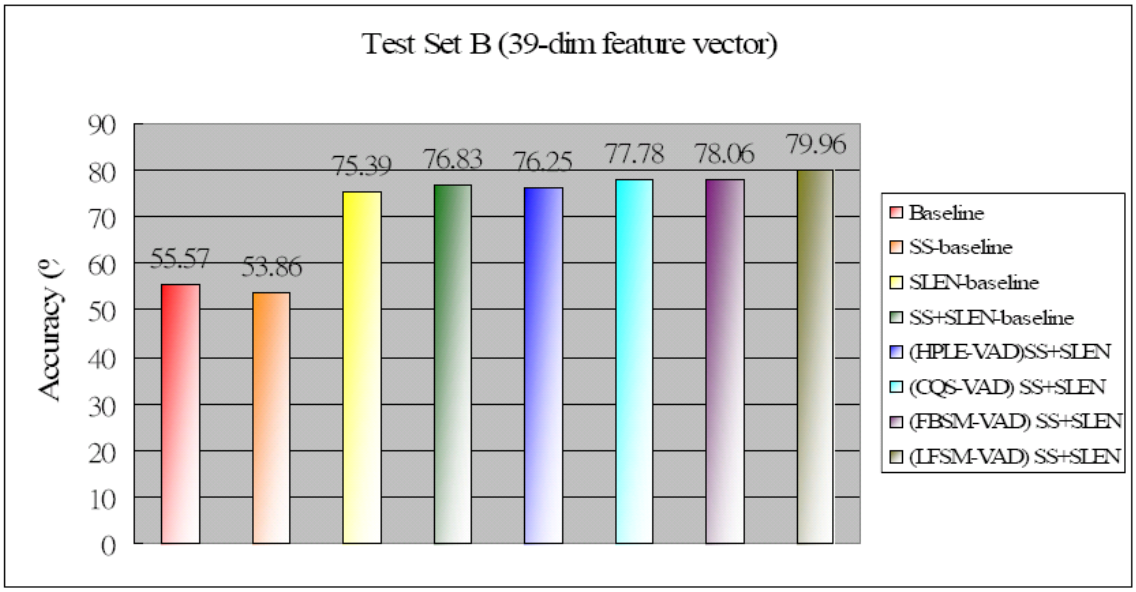
本章節所有實驗所使用的特徵參數爲 12 維梅爾倒頻譜參數與 1 維對數能量，附加其一階和二階差量，總共爲 39 維特徵參數。圖八與圖九分別爲 A 組環境與 B 組環境下各種方法所得之平均辨識率。其中「baseline」是指沒有處理過的原始特徵參數、「SS-baseline」、「SLEN-baseline」與「SS+SLEN baseline」是指利用每段語句前五個音框作爲純雜訊音框分別作頻譜消去法(SS)、靜音音框對數能量正規化法(SLEN)及 SS 和 SLEN 的結合，「(HPLE-VAD) SS+SLEN」、「(CQS-VAD) SS+SLEN」、「(FBSM-VAD) SS+SLEN」與「(LFLE-VAD) SS+SLEN」則分別爲以音框能量、高通對數能量、累積量化頻譜、全頻帶頻譜強度與低頻帶頻譜強度作爲端點偵測特徵，執行端點偵測，再使用頻譜消去法與靜音音框對數能量正規化法。

從圖八與圖九的辨識結果，我們有以下幾點的發現：

1. 相較於前一節的基本實驗，本節的特徵參數額外引進了對數能量及其一階與二階差量，然而其基本實驗辨識率反而較未引進這三個能量參數的結果還來的差，辨識率降低了 3%至 7.8%。這可能是因爲，雖然能量特徵它包含了很多語音的資訊，但是相對的它也深受雜訊的影響，反而不利於系統辨識。
2. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之頻譜消去法，相較於基礎實驗結果反而退步了約 3%，其可能原因如同前述，即能量特徵的失真所帶來的嚴重影響。
3. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊所作之靜音音框對數能量正規化法，在辨識率上有十分顯著的提升，相較於基礎實驗結果提升了 13%至 20%之多。這代表了在雜訊環境下，能量特徵強健性處理的重要性，也凸顯了靜音音框對數能量正規化法的明顯效能。
4. 未作端點偵測而純粹以每段語句前 5 個音框爲純雜訊，同時執行頻譜消去法與靜音音框對數能量正規化法，其辨識率比單純使用靜音音框對數能量正規化法可以再進步 1%-2%左右。
5. 本論文所提出的四種端點偵測的特徵(HPLE, CQS, FBSM 與 LFSM)應用於端點偵測，配合頻譜消去法與靜音音框對數能量正規化法，都能得到明顯的進步。其中除了高通對數能量(HPLE)在 Set B 環境下稍微退步外，其他情形下皆比未做端點偵測的結果進步 1%以上。類似前一節，這四種特徵又以低頻帶頻譜強度(LFSM)表現最好，相較於未做端點偵測的結果，可得到 3%-4%的平均進步率。



圖八、A 組環境下平均辨識率(%)比較圖



圖九、B 組環境下平均辨識率(%)比較圖

六、結論

在本論文中，我們提出了幾種端點偵測所用的信號特徵，包括低頻帶頻譜強度、全頻帶頻譜強度、根據頻譜強度分佈的累積量化頻譜、及根據能量調變頻譜特性的高通對數能量。其目的是偵測出純雜訊音框，進而結合頻譜消去法與靜音對數能量正規化法，達到強健語音消除雜訊的功能。其中低頻域頻譜強度之端點偵測法配合頻譜消去法與靜音對數能量正規化法，可以最有效地提升雜訊環境下的辨識率。而其他三種信號特徵也有不錯的端點偵測效果。由比較中可得到，這幾個方法比起基本實驗，都有著顯著的進步。

此外，我們發現能量維特徵蘊藏著很多語音鑑別資訊，但相對而言，雜訊對其影響也很大。但是經過靜音對數能量正規化法處理，可以明顯提升辨識效果。此外，低頻帶頻譜強度之端點偵測法配合頻譜消去法與靜音對數能量正規化法可達到平均 80%的優異表現。由實驗結果也發現，八種雜訊環境在雜訊比 0~20dB 的條件下，每個辨識率都很接近，即表示此方法不受雜訊型態的影響。在穩定與非穩定雜訊都有很好的辨識率。

參考文獻

[1] K. Yamashita,; T. Shimamura; " Nonstationary noise estimation using low-frequency regions for spectral subtraction", Signal Processing Letters, IEEE Volume 12, Issue 6, June 2005

[2] Tai-Hwei Hwang, “Energy Contour Extraction for In-Car Speech Recognition”, 9th European Conference on Speech Communication and Technology (Eurospeech 2003)

[3] Weizhong Zhu and Douglas O’Shaughnessy “Log-energy Dynamic Range Normalization for Robust Speech Recognition”, 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)

[4] Tai-Hwei Hwang and Sen-Chin Chang, “Energy Contour Enhancement for Noisy Speech Recognition”, 2004 International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)

[5] M. Ahadi, H. Sheikhzadeh, R. Brennan and G. Freeman, “An Energy Normalization Scheme for Improved Robustness in Speech Recognition” 8th International Conference on Spoken Language Processing (ICSLP 2004)

[6] R. Chengalvarayan, “Robust Energy Normalization Using Speech/nonspeech Discriminator for German Connected Digit Recognition”, 6th European Conference on Speech Communication and Technology (Eurospeech 1999)

[7] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction” IEEE Trans. on Acoustics, speech, and Processing, VOL. ASSP-27, NO. 2, April 1979

[8] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", Eurospeech 1991

[9] C.F. Tai, J.W. Hung, "Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments", INTERSPEECH 2006 – ICSLP

[10] H.-G Hirsch, D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” ISCA ITRW ASR 2000, Paris, France, September 18-20, 2000

從不同韻律格式驗證階層式韻律架構並兼論對語音科技的應用

鄭秋豫 蘇昭宇
中央研究院語言所語音實驗室
cytling@sinica.edu.tw, morison@gate.sinica.edu.tw

摘要

本論文以四種韻律格式語料的基頻變化，驗證鄭秋豫[1][2][3]所提出的「階層式韻律句群 HPG 架構」，結果證明一種基型即可解釋並產製不同的韻律格式。HPG 架構解釋：口語語流的韻律輸出，並非僅止於字調與句調的線性串接，而同時含有跨短句語篇上層訊息，各韻律階層對總體韻律輸出的貢獻度可驗證。所用語料爲朗讀古典文體三種（詩、詞二種韻文及古典散文）、氣象撥報語料一種。我們採 Fujisaki model 檢驗基頻（F0）參數，用 HPG 階層式模型進行分層分析與預測，結果發現：每類語體的韻律輸出都含有上層語篇訊息（higher level information）效應，唯韻律貢獻度依韻律格式呈不同分佈。三種古典文體語料的上層貢獻度均高於非氣象撥報語料，且文體結構越工整、語料的樂律性越高時，上層貢獻度所佔比例越重。因此韻律格式的差異，具體表現於階層貢獻度的分佈，而 HPG 架構可解釋韻律輸出的差異。我們以此結果更進一步證明：無論語體或韻律格式爲何，上層語篇效應所造成的跨短句韻律語境，都是語流韻律不可或缺的因素。我們也因此推論，語音合成語流韻律輸出時，只需一個韻律基型，配以不同的韻律層貢獻度，便可產生不同韻律輸出。

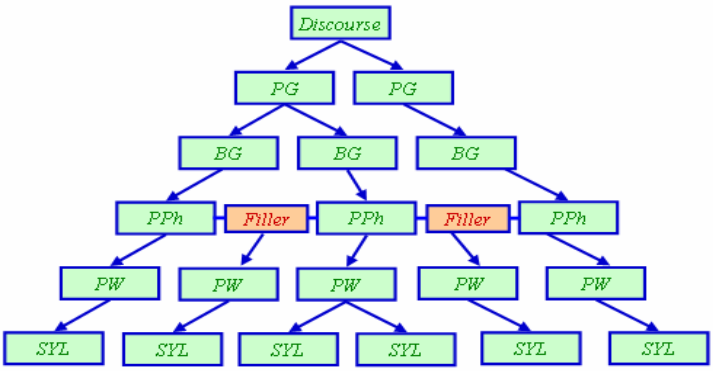
關鍵詞：階層式韻律句群架構，語流韻律，古典文體，Fujisaki model，上層效應，語體

一、緒論

與韻律訊息相關的聲學語音參數有四：基頻曲線(F0 contour)、音節時長(syllable duration)、能量分佈(Intensity distribution)、和停頓時長(pause duration)。與韻律相關的文獻中，最受重視的聲學參數一向是基頻，研究聲調語言時，字調基頻曲線的變化也一直是重點，本論文將以韻律短句的基頻曲線變化爲主要分析參數，聽感的韻律邊界停頓爲輔，探討不同韻律格式的韻律輸出，究竟如何不同？是否可以透過量化語料，加以驗

證。

傳統韻律研究一向著重於孤立音節的字調及孤立短句句調的分析，然而在長篇口語敘述段落的語料中，相同字調的音節或同類的句調，卻有相當大的變異，語音訊號的變化非常複雜。鄭秋豫從聽感出發，發現這些變異對聽者完全沒有影響，聽者輕易的同時處理不同韻律範圍的資訊，辨識音段、音節、字調、句調甚至更大單位的韻律訊息，更有效的應用短句以上的語篇語意訊息，以上層語篇訊息（higher level information）的效應，辨識出字調與句調以上的語段單位，作爲口語分析的基礎。她經由量化分析大批口語料，發現語流中的音節長短在時程上的長短分佈、能量分佈、停頓時長，都含有上層語篇訊息，語流韻律輸出不但具有系統性，而且是層層韻律效應疊加的總和[3]，也因此證明語段中語篇語意的連接與連貫性確實存在於語流語音信號中，若只考慮字調與語調效應，不但不足以說明多短語語流的韻律特性，且無法系統性的呈現語段中各短語間的關聯性（association），她以此基礎，提出「階層式多短句韻律架構 Hierarchical Prosodic Phrase Grouping　Framework (HPG)」及數學模型[3]。2006 年她進一步得到基頻走勢的證據[Tseng 2006]，至此結合基頻、音長、能量、停頓的證據，將 HPG 架構（如下圖所示）擴及語篇。



圖一、階層式多短語韻律句群架構圖。其中 SYL、PW、PPh、BG、PG 分別代表口語語流中的音節、韻律詞、韻律短語、呼吸句群，韻律句群。（Tseng, 2006）

在 HPG 架構中，下層韻律單位（如音節、韻律詞、韻律短句）接受上層韻律單位（如呼吸群、韻律句群、語篇）的管轄，因此音節與短句一旦進入多短語語段，除了字調（tone）與短語調(intonation)外，因爲來自上層語篇的語意訊息造成語流的連接性（cohesion）與連貫性（coherence），以至韻律短句間產生韻律關連性(association)，各下級韻律單位必須依照更語篇韻律（Discourse Prosody）的上層訊息進行系統性的調整，提供語段從

何處開始、維持到結束的韻律語境訊息，各語段下轄之次次級韻律單位需依此規範，做系統性的調整。由於此架構的提出，前文所提語流中，字調及孤立短句效應外的變異性，其實皆可由階層式架構韻律短句上層的資訊加以解釋，而不再屬於無法預測的變異。由此得知，進行語音合成時，語音信號必須兼顧架構中各層的韻律效應，而不僅只是做字調與句調的線性串接及平滑，才能將完整地模擬出連續語流的特性。

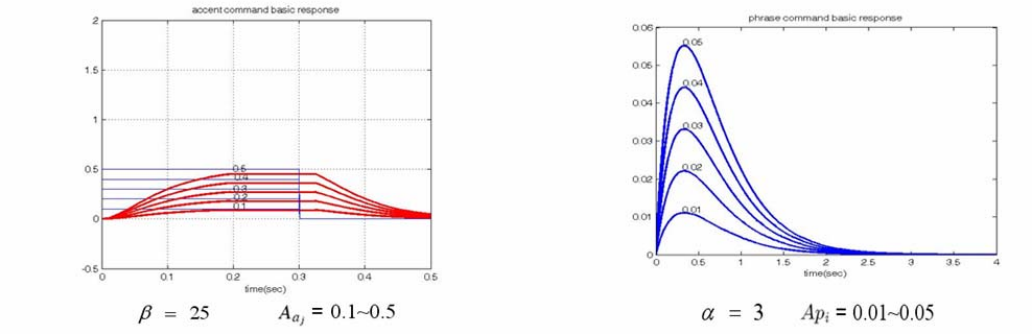
二、Fujisaki 特徵參數自動擷取系統

(一) Fujisaki model 簡介

與韻律相關的文獻中，前文所提四項聲學語音參數（基頻曲線、音節時長、能量分佈和停頓時長）又以基頻曲線最受重視，基頻曲線的變化也是研究聲調語言字調的重點，本論文即以韻律短句的基頻曲線變化爲主要分析參數。傳統研究中，是以直線、二次方或三次方函數曲線趨近基頻曲線，但 Fujisaki 等 1984 年[4]提出基於發聲器官的模型：產生基頻主要器官爲聲道（vocal tract）及相鄰的隨意肌，而應於語言單位在時間程上產生高低的變化而控制隨意肌時，有一定物理特性的限制，所以基頻曲線並非任意直線或曲線，而是必須受制於一定的規則在時間程所產生的變化，且此限制出現於對數尺度上，此爲基頻曲線的函數限制，此一模型一般通稱爲 Fujisaki model。Fujisaki model 由三個不同的元件（component）所構成，分別爲(1.)短語元件 (phrase component Ap)，反應發聲器官產製較大單位基頻曲線的控制與發聲限制；(2.)語氣元件(tone component Aa)，反應發聲器官產製較小單位基頻曲線的控制發聲限制；與(3.)基底直線(base frequency Fb) 代表基本音高。原 Fujisaki model 中 tone component Aa 中的 tone，泛指語氣變化（tone of voice），如強調、加重語氣對基頻曲線造成的影響。此模型本身對於大、小單位爲何並未規範，又因爲基於發聲器官，所以也與語言無關，迄今已被用來成功的模擬過多種語言。應用到輕重音語言與英語、德語時，大單位指的是片語或短語的語調、小單位則用來表示加重（emphasis）；應用到聲調語言模擬中文（國語）時，大單位表示短語語調的下傾、小單位被用來表示字調變化[5]。圖二分別列出此二元件的函數及輸出特徵，其中短語元件和語氣元件函數分別由一個函數能量 Ap 及 Aa 所控制，各自控制輸出的突起程度以及上升／下降斜率，Ap 及 Aa 越大，元件函數曲線凸起程度越明顯，下降斜率也更陡峭。

$$A_{a_j} G_t(t) = \begin{cases} \min\left[1-(1+\beta t)\exp(-\beta t), \gamma\right], & t \geq 0 \\ 0, & t < 0 \end{cases}$$

$$Ap_i\ G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases}$$



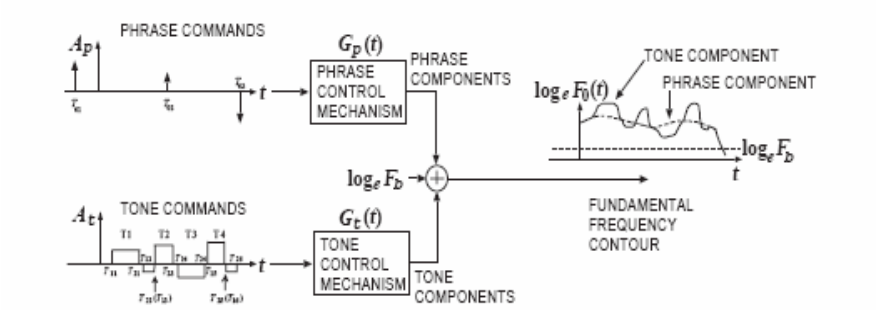
$\beta = 25$ A_{a_j} = 0.1~0.5 $\alpha = 3$ Ap_i = 0.01~0.05

(a) (b)

圖二、(a)Aa (b)Ap 隨函數能量變化的時間響應圖

(二) 自動擷取 Fujisaki 特徵參數的方法

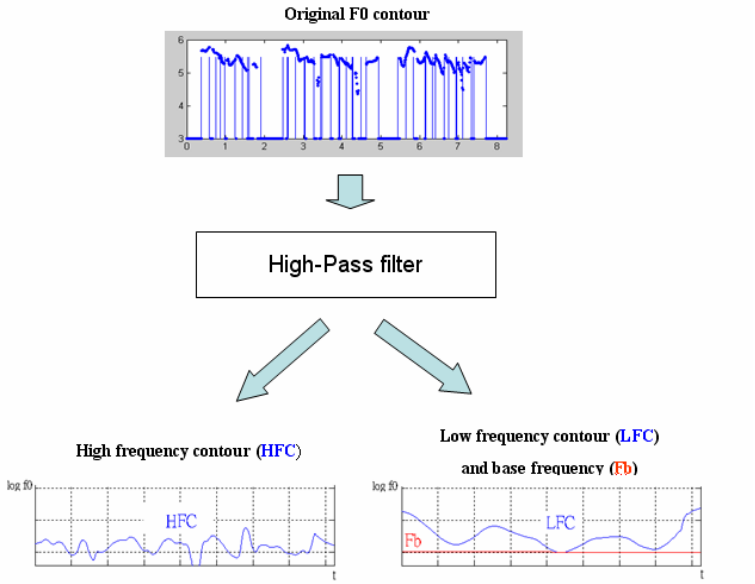
Fujisaki model 假設，基頻曲線可由短語元件、語氣元件和基底直線疊加而趨近，圖三表示 Fujusaki model 三個基本元件的疊加輸出。



圖三、短語元件、語氣元件與基底直線疊加後的基頻曲線（Fujisaki, 1984）

換句話說，基於 Fujisaki Model 的國語基頻曲線，可視爲語調下傾、字調變化與基本音高三元件疊加所成，而該模型泛指語氣變化的語氣元件（tone component），很巧合的剛好可用來表示字調（tone）的變化。雖然有了 Fujisaki model 可以來趨近原始語音，但由於 Fujisaki model 係由三個元件構成，所以趨近原始基頻曲線必須以三個元件的最佳組合爲參考指標。此外，根據 Fujisaki model 之定義，相同的語氣元件必須在語流中各位置保持大致相同的權重，因此傳統的作法是以目測原始基頻曲線，手動調整決定這三個元件的最佳組合，這樣的作法相當耗費人力。我們則採用了 Mixdorff 2000、2003 年

[6][7]提出的方法並加以改善：以高通濾波器(high-pass filter)來分離基頻曲線，自動提取出基頻曲線中變化劇烈的部份即爲語流的基本單位，可對應 Fujisaki model 中的語氣元件；而變化和緩的部份，則爲語流中語調的下傾趨勢，可對應 Fujisaki model 中的短語元件，因此採用一組截止頻率爲 0.5Hz 的高通濾波器分離原有基頻曲線爲三個部份(1) 高通濾波器的輸出定義爲高通曲線(HFC)，爲語氣元件逼近的目標曲線，(2) 扣掉高通部份剩餘平滑曲線則定義爲低通曲線(LFC)，找出此低通曲線的最低點並定義爲通過此最低點的直線爲基底直線(Fb)，(3) 扣掉基底直線後的曲線視爲短語元件的目標曲線，必須用短語元件函數來加以逼近。圖四是 Mixdorff 自動擷取 Fujisaki 參數架構。

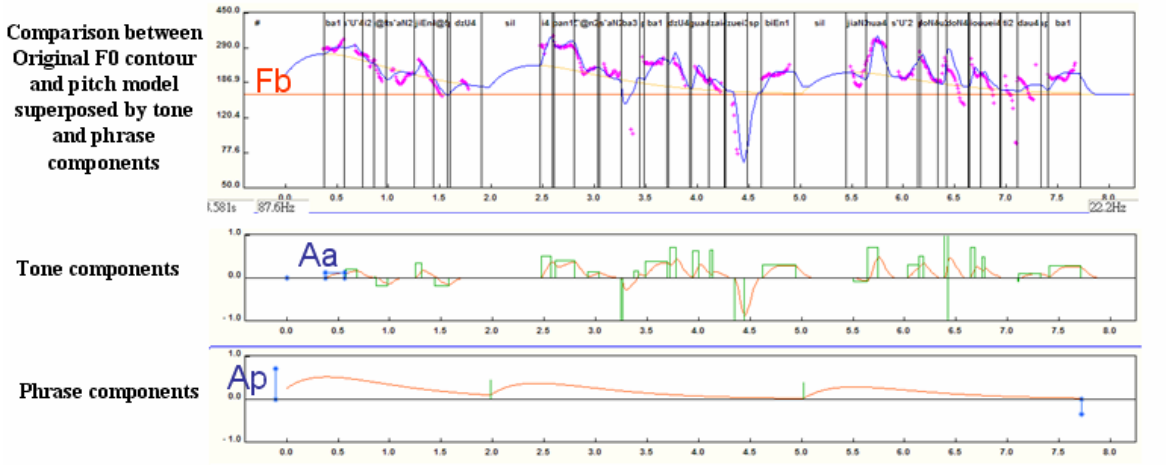


圖四、利用濾波器將基頻曲線分離爲高通及低通曲線，再分別以字調元件與短語元件趨近高通與低通曲線，進行自動擷取程序 (Mixdorff, 2000)。

至於自動提取 Fujisaki 參數的方法，近來又有 Bu 等[8]提出以一階微分曲線來，不過，Mixdorff 處理的語言是國語，而 Bu 等處理的是日語，所以我們仍維持 Mixdoff 的方法。

(三) 加入聽感邊界的方法與自動擷取結果

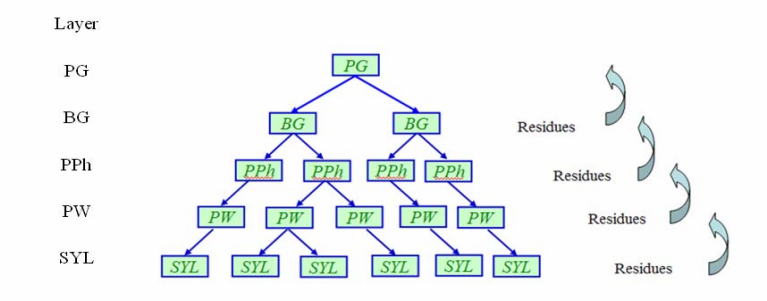
在本研究中，我們除驗證 HPG 韻律架構，也希望探討聽感獲得的韻律邊界邊界與基頻曲線的關係，因此我們加入聽感的邊界資訊作爲自動擷取 Fujisaki model 的 Ap 參數。初步觀察顯示，局部最低點通常就出現韻律短語的聽感邊界上，因此加入聽感邊界的自動擷取方法，可減少 Mixdorff 方法在搜尋整條低通曲線的範圍的運算時間。圖五爲加入聽感邊界爲輸入參數後，自動擷取 Fujisaki model 的 Ap 參數結果。



圖五、加入聽感的邊界資訊後自動擷取 Fujisaki 參數的結果

三、階層式多元線性迴歸模型與計算

Nakai 等於 1994，1995 [9][10]提出歸類 Ap 的方式來建立語流中的短語模型，然而並未將語流中的階層性關係列入考慮，我們採用了 Keller 等[11]與 Zellner 等[12]用於法語的階層性線性迴歸分析，並考慮國語的音韻結構加以改良，用此適用於國語的模型來分析語音信號。階層性線性模型是簡單線性迴歸的衍生，最大的不同點在於：每個輸入都附有多層次的停頓標註，每一標註分別代表著在每一韻律層的參數與特性，可進行逐層分析，分析步驟如下: (1) 分析及預測某一韻律層時，只利用當層的參數做爲線性迴歸的預測變項並得到此層的預測模型，(2)由於有更上層的標註，我們假設，原始值與預測模型的差距並不被視爲實驗誤差，而是視爲來自更上層的效應，因此將分析及預測某一韻律層的殘差（原始值與預測模型的差距）作爲分析及預測下一韻律層的輸入，以更上層的標註進行線性迴歸分析，因此可得到更上層的預測模型與貢獻度，(3) 逐層分析、預測後計算出各韻律層的貢獻度。圖六以圖式表示階層式線性回歸逐層分析。



圖六、以階層式線性迴歸逐層分析示意圖 （Tseng et al, 2004）

本研究以 Ap 爲輸入參數，由於 Ap 是對應韻律短語的參數，因此，分析的步驟由階層式架構中韻律短語層(PPh)開始、之後對上層的呼吸句群層(BG)及更上層韻律短語句群(PG)進行逐層的線性迴歸，分析參數如下: (1)PPh 層: 以目前 PPh 長度(Current PPh Length)與前一 PPh 長度(Preceding PPh Length)的組合做爲分析參數，進行分析，經過線性迴歸後的殘差定義爲 Delta1，並輸入 BG 層進行分析，(2)BG 層: 以目前 PPh 在 BG 中的位置(BG Sequence)做爲分析參數，若 BG Sequence=1，表示目前 PPh 爲此 BG 之起始 PPh，以此類推，進行線性迴歸，(3)PG 層: 跟 BG 層輸入參數相同，其數學函數表示如下：

PPh:

$$\mathrm{Ap} = f(precedingPPhLength, currentPPhLength) + Delta1$$

BG:

$$Delta1 = f(BGSequence) + Delta2$$

PG:

$$Delta2 = f(PGSequence) + Delta3$$

四、實驗語料

文本部份，採用古典文體（詩、詞等韻文及古典散文）與氣象播報文本，共計 26 篇古典文體語篇段落（包含：4 篇古典散文，1 首賦，１首民歌，6 首古詩，6 首唐代樂府詩和 8 首宋詞），以及 34 則氣象播報的語篇段落——以此分別代表四種韻律格式。古典文體各自承載不同文體的韻律變化特性，氣象播報文本則當作白話文本。我們依文本結構的工整規則性，定義四類韻律格式： 規則(R)、半規則(SMR)、不規則(IR)與氣象播報(WIR)。其中詩詞等韻文主要爲規則及半規則的文體，而沒有固定重複特徵的古典散文則被歸於不規則類；而氣象播報文本亦屬不規則類，但另以一類計，以示與古典文體有別，並可與古典文體比較。分類範例見附註一。古典文體文本包括的三種韻律格式(R、SMR、IR) ，採人工方式判別詩歌的工整性，依據準則爲：(1)標點符號。採用較大的段落標點符號，如：句號，將文本分成包含多短語的段落區塊；(2)短語長度。挑出相同文本中重複的多短語段落區塊，人工檢視重複區塊的工整性比例，加以分類。

語料部分，由三位發音人朗讀以上文本，使用 Sony ECM-77B 迷你麥克風、以及 Cool Edit 2000 在隔音室進行錄音。發音人爲(1)女性發音人一位，f054， 朗讀所有文本，(2)男性

發音人兩位，一位 m056 朗讀三種風格的古典文體語料（26 篇），另一位 m054 朗讀氣象語料（34 篇）。表一統計三種古典韻律格式加總後的音節、韻律短句數目及篇數，並以音節平均時長代表平均語速；表二統計氣象撥報語料的音節、韻律短句、語篇數目語料及平均語速。

表一、古典文體語料的音節、韻律短句、語篇數目語料及平均語速統計

韻律單元數/平均語速 / 發音人	音節數	韻律短句數	語篇數	平均語速（毫秒/音節）
女 f054	3502	710	26	271
男 m056	3510	711	26	202

表二、氣象撥報語料的音節、韻律短句、語篇數目語料及平均語速統計

韻律單元數/平均語速 / 發音人	音節數	韻律短句數	語篇數	平均語速（毫秒/音節）
女 f054	7054	720	34	193
男 m054	7096	747	34	165

從表一、表二可看出，本研究語料設計的主要控制組是韻律短句和語篇的數目，而非音節的數目。古典文體語料只有三千五百餘音節，氣象撥報語料則有七千餘音節，二者差異頗大；但古典文體語料含韻律短句七百餘、語篇二十六，氣象撥報語料則亦含韻律短句七百餘、語篇三十四，數目差異並不大。

五、實驗結果與分析

(一) 聽感標註的重疊性

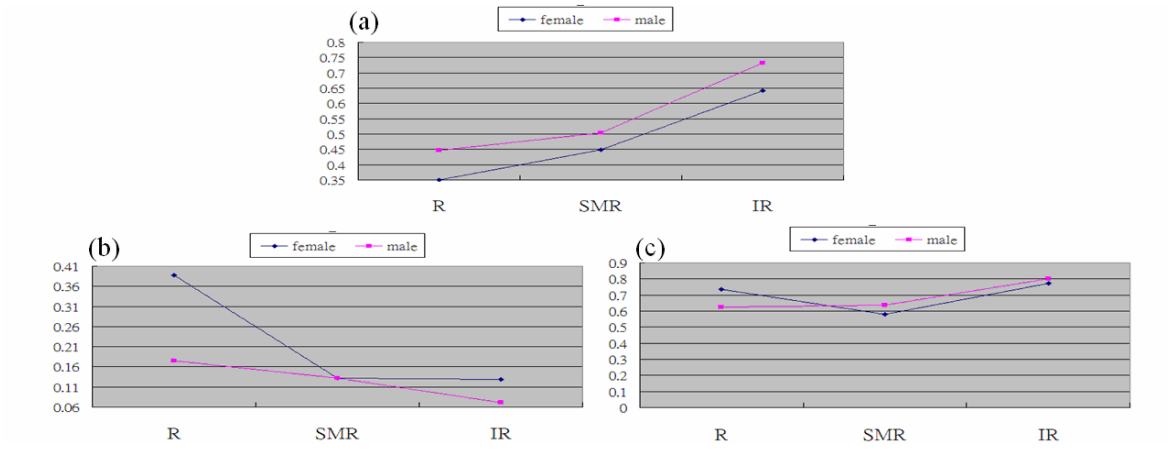
我們分析古典體文語料中二位發音人（f054 與 m056）聽感標註韻律邊界的重疊性後，將結果列於表三。表三中 B1、B2、B3、B4、B5 分別代表音節、韻律詞、韻律短語、呼吸句群，語段（韻律句群）之邊界。結果顯示二人的差異性並不大，而且從表三列出 HPG 韻律階層上層到下層的重疊比例，可發現上層標註的重疊率雖然不如下層，但在上層都能維持一定的重疊比例，可見在古典文體中，語篇的韻律效應仍存有相當的一致性。

表三、比較兩語者朗讀古典文體韻律邊界的一致性結果

韻律邊界 / 發音人	B1 音節邊界	B2 韻律詞邊界	B3 韻律短句邊界	呼吸群 B4	語段（韻律句群）B5
f054	97.98%	86.46%	82.79%	76.76%	62.30%
m056	96.71%	88.69%	80.19%	76.76%	80.85%

(二) Ap 分析

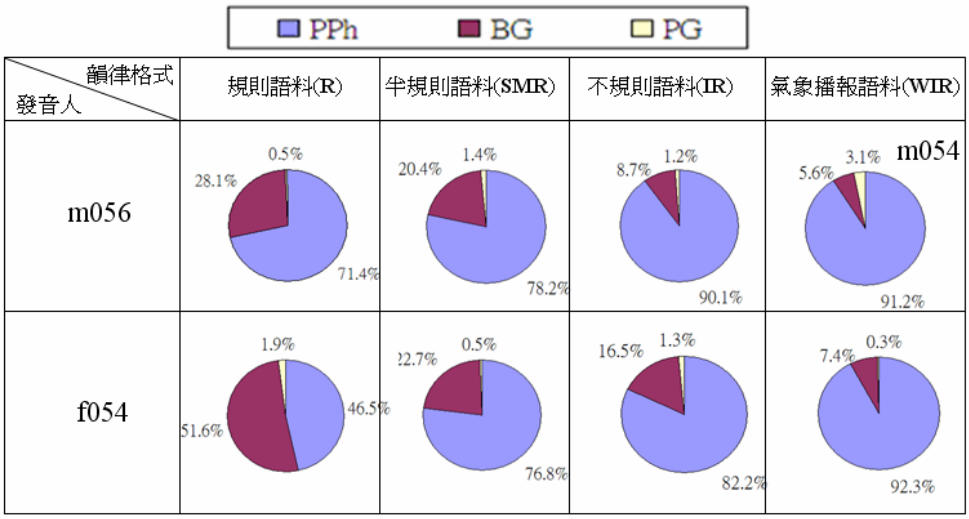
我們比較古典文體語料在單一韻律層中 Ap 正確率隨文體（規則、半規則到不規則）的變化趨勢後發現：在韻律短語層(PPh)的預測，隨著文體從規則到不規則的分類，兩語者 Ap 的正確率的預測皆呈現相同的趨勢。也就是說，預測的正確率隨文體的不規則程度上升，顯示出越不規則的短語結構變化，越是提供了較多的預測資訊，使不規則語料的 Ap 參數在 PPh 層獲得較準確的預測。然而，在呼吸句群層(BG)的預測，卻呈現相反的趨勢，即 Ap 預測正確率隨著文體工整性的提升，呈現上升的趨勢，顯示在文體越工整，上層韻律效應越明顯。因此在 BG 層中，預測度隨著文體工整性與 Ap 呈正相關。若將兩層的 Ap 預測正確率相加，可發現相加後的 Ap 預測率，對總預測率具有互補效應。在規則語料中，BG 貢獻度高，PPh 層貢獻低；不規則語料中，PPh 貢獻度高而 BG 貢獻度低;而半規則語料的 PPh 及 BG 貢獻度則介於規則跟半規則間。兩者互補的結果，總預測率曲線則趨於平緩。圖七比較二發音人的三種古典文體語料古文語料 Ap 預測正確率，在特定的韻律層隨三種不同韻律格式(R、SMR、IR)變化曲線，(a) PPh 層 (b) BG 層 (c) PPh 層與 BG 層疊加後的結果，顯示不同韻律階層的差異(b)，加總後因相抵，最後呈現相當一致的趨勢(c)。



圖七、古典文體語料 Ap 預測正確率在特定的韻律層隨三種不同韻律格式(R、SMR、IR)

變化曲線，(a) PPh 層 (b) BG 層 (c) PPh 層與 BG 層疊加後結果。

各層的貢獻度在總預測率的比例分佈，我們以圖八圓餅圖表示。靛色表示 PPh 層的貢獻比例，紫色表示 BG 層的貢獻比例，而黃色表示 PG 層的貢獻比例。上列表示男性發音人，下列表女性發音人，由左至右的分類分別爲規則(R)，半規則(SMR)，不規則(IR)，最後則爲比較用的氣象語料(WIR)。相對於韻律格式的規則性，二位發音人的語料都顯示，最上層 PG 的貢獻度並不明顯，我們認爲，這是因爲此三類古典文體的語篇長度偏短。不過，在 BG 層則顯示，語料格式越規則，BG 層的貢獻度越高，且 BG 貢獻度的比重隨語料格式的規則度由規則向不規則遞減。BG 的貢獻度在規則語料(R)分別是 28.1%和 51.6%；半規則語料(SMR)分別是 20.4%和 22.7%；不規則語料(IR)分別是 8.7%和 16.5%；氣象撥報最低，分別是 5.6%和 7.4%。



圖八、Ap 預測正確率在不同語者與不同韻律格式的分佈比例總比較

分析四種語料後的總體趨勢顯示：韻律格式的差異，可以從不同韻律階層對整體韻律輸出的貢獻度分佈具體表示。韻律格式越規則、內含樂律性越高時，上層資訊的貢獻度越高、韻律規劃的範圍越大。韻律格式越不規則，下層韻律韻律短句的貢獻度越高，表示韻律規劃的範圍越小。基於的結果，我們認爲：鄭所提出的階層式韻律句群 HPG 架構，可系統性的解釋並產製不同的韻律格式，因此可視爲語流韻律的基型，其作用及意義，與字調對應於音節、句調對應於短句相同。總括而言，HPG 是對應於多短語語流韻律的基型。

六、結論與未來展望

本文以韻律短句的基頻曲線變化爲主要分析參數，聽感的韻律邊界停頓爲輔，從文體工整性及內建的韻律性來檢視語料、剖析朗讀古典文體語料的韻律格式，並從階層式韻律句群及架構分析語流韻律中的韻律成分，我們得到的結果顯示，上層語篇訊息提供跨短語韻律語境，是語流韻律不可或缺的成分。韻律格式越規則，上層訊息的貢獻度越高。隨著工整性變化造成不同的貢獻度分佈的研究結果可知，語者在朗讀大量重複結構的過程中，會更強調語意的銜接與語篇的轉換效應，因此大範圍的上層語意資訊及語段、語篇的效應更明顯。更重要的是，韻律格式的不同，從 HPG 韻律階層貢獻度而言，只不過是不同的分佈模式而已，完全可系統性的表示。而氣象撥報語料在四種語料中規則性最低，分析結果顯示，上層效應越不明顯。因此不論古典文體或氣象撥報語料，都提供階層式多短語韻律句群架構 HPG 階層韻律貢獻的證據。本研究中基頻曲線的實驗證據證明了多短語韻律句群架構具有一定的強健性，適用不同語體的語料，因此以 HPG 做爲長篇語段的基本架構，但具有一定的自由度，可依照語體而變化調整，若能適當的運用 HPG，只需一種跨短語基型，藉由操弄 HPG 各層的貢獻度，便能達到韻律格式轉換的結果，不但能將連續語流中韻律基本特性表現出來，更能使語音合成系統的韻律輸出更具豐富的變化性。

本研究限於篇幅，僅以韻律短句的基頻曲線變化爲主要分析參數，不過，我們強調，韻律的研究並非僅止基頻，其他的聲學參數也具有韻律效應，全面的韻律研究，應考慮語音信號中所有的聲學參數，以及它們的互動模式。鄭較早的研究結果[2, 3]顯示，語流韻律的節奏，也是重要的聲學參數，她也獲得音節時長在時間層的分佈的效應，得到對應於 HPG 架構系統性的節奏模版。未來我們將繼續分析不同韻律格式的語料中，進一步求取語流韻律的節奏特性，並與基頻研究的結果整合，以期對語流韻律有更全面性的瞭解。

參考文獻

[1] Tseng, C. "Prosody Analysis", *Advances in Chinese Spoken Language Processing*, World Scientific Publishing, Singapore, pp. 57-76, 2006.

[2] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, Y. "Fluent Speech Prosody: Framework and Modeling", *Speech Communication, Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation*, Vol. 46:3-4, pp. 284-309, 2005.

[3] Tseng, C. and Lee, Y. "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese", *Proceedings of the International Conference on Speech Prosody 2004*, pp. 251-254, 2004.

[4] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J.Acoust, J.Acoust. Soc*. Jpn.(E), 1984; 5(4), pp. 233-242, 1984.

[5] Wang, C., Fujisaki, H., Ohno, S. and Kodama, Tomohiro. "Analysis and synthesis of the four tones in connected speech of the standard Chinese based on a command-response model", *Proceedings of EUROSPEECH'99*, pp. 1655-1658, 1999.

[6] Mixdorff, H. "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", *Proceedings of ICASSP 2000*, vol. 3, pp.1281-1284, 2000.

[7] Mixdorff, H., Hu, Y. and Chen, G. "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech* 2003, pp. 873-876, 2003.

[8] Bu, S., Yamamoto, M. and Itahashi, S. "An Automatic Extraction Method of F0 Generation Model Parameters", *Trans. IEICE*, TRANS. INF. & SYST., VOL.E89–D, NO.1, Jan 2006.

[9] Nakai, M., Singer, H., Sagisaka, Y. and Shimodaira, H. "Automatic prosodic segmentation by F0 clustering using superpositional modeling" , *Proceedings of ICASSP95*, pp. 624–627, 1995.

[10] Nakai, M., Shimodaira, H. and Sagayama, S. "Prosodic Phrase Segmentation Based on Pitch-Pattern Clustering", *Trans. IEICE*, J77-A, 2, pp. 206–214, Feb. 1994.

[11] Keller, E. and Zellner, K. "A Timing model for Fast French", *York Papers in Linguistics*, *17*, University of York, pp.53-75, 1996.

[12] Zellner, K. and Keller, E. "Representing Speech Rhythm" *Improvements in Speech Synthesis*. Chichester: John Wiley, pp. 154-164, 2001.

附註一、古典文體三種分類範例

1. 規則(R)
古詩十九首之九　(漢代五言古詩)
庭中有奇樹，綠葉發華滋。
攀條折其榮，將以遺所思。
馨香盈懷袖，路遠莫致之。
此物何足貴，但感別經時。

2. 半規則(SMR)
將進酒 （唐代樂府 李白）
君不見黃河之水天上來，奔流到海不復回；君不見高堂明鏡悲白髮，朝如青絲暮成雪。
人生得意須盡歡，莫使金樽空對月。
天生我材必有用，千金散盡還復來。
烹羊宰牛且爲樂，會須一飲三百杯。
岑夫子，丹丘生，將進酒，杯莫停。
與君歌一曲，請君爲我側耳聽。
鐘鼓饌玉不足貴，但願長醉不願醒。
古來聖賢皆寂寞，惟有飲者留其名。
陳王昔時宴平樂，鬥酒十千恣歡謔。
主人何爲言少錢，徑須沽取對君酌。
五花馬，千金裘，呼兒將出換美酒，與爾同消萬古愁。

3. 不規則(IR)
禮運大同篇　(先秦散文)
大道之行也，天下爲公；選賢舉能講信修睦。
故人不獨親其親，不獨子其子；使老有所終，壯有所用，幼有所長，矜、寡、孤、獨、廢疾者皆有所養。
男有分，女有歸。
貨惡其棄於地也，不必藏於己；力惡其不出於身也，不必爲己。
是故謀閉而不興，盜竊亂賊而不作；故外戶而不閉。

是謂大同。

多語聲學單位分類之最佳化研究

呂道誠　Dau-cheng Lyu
長庚大學電機工程學系
d9221003@stmail.cgu.edu.tw

呂仁園　Ren-yuan Lyu
長庚大學資訊工程學系
renyuan.lyu@gmail.com

江永進 Yuang-chin Chiang
國立清華大學統計學研究所

許鈞南　Chun-nan Hsu
中央研究院資訊科學所
chunnan@iis.sinica.edu.tw

摘要

由於全球化的形成，人與人之間的溝通不再限於同一種語言，因此多語的語音辨識也變的格外的重要。如何有效整合多語的聲學模型是一個關鍵議題，因爲一組好的多語聲學單位將影響辨識結果。本論文提出了一套整合專家背景知識與實際語音分析的方法，來產生一組新的聲學單位，並且對這組聲學單位的數目，使用差分貝式資訊法則來做最佳的處理。從訓練好的隱藏式馬可夫聲學模型中，計算其單位間的相似度矩陣，之後透過語音學和音韻學的知識，限定了各個聲學單位能群化的上限，根據不同限定的群化上限，使用聚合階層式分群法，來建立不同的結構樹。之後，利用差分貝式資訊法則，將每個結構樹中發音相近的聲學單位做合併，當差分貝式資訊法則的值小於零的時候，就停止合併，而新合併成一群的聲學單位則爲新的聲學單。我們將用 ForSDAT01 華台雙語語料庫來實驗評量，而實驗結果顯示，本論文所提出的新方法比只用專家知識所定義的聲學單位所訓練出的辨識器有較高的辨識效果。

關鍵詞：多語語音辨識、音素群化、差分貝式資訊法則

一、緒論

語音是人與人溝通最直接也是最原始的一種工具之一，透過語音的傳遞，能夠拉近彼此的距離。近來，由於地球村的形成，人與人之間的交談並不再限制單一的語言。這種現象，在國家彼此相鄰的歐洲和人種混合的亞洲都會時常發生。如台灣，目前華語和台語是最主要的兩種語言。因此，自動語音辨識的研究領域裡，從原本的單一語言的語音辨識，漸漸的，已朝向雙語或多語的方向了[1]。同樣的，如果一套自動語音辨識機器，能夠一次辨識兩種語言或兩種以上的語言，這會比只能處理單一語言的機器來的更強大。所以說，多語言的語音辨識，是個當前重要且必須探討的議題。

另外，在多語的語音辨識領域裡，對於不同語言間聲學單位的定義，是個值得研究的方向。這個議題，主要在探討如何將多個語言的發音符號做有效的定義，好讓最後，給定一句未知的語音，在根據這些定義好的發音符號所訓練出來的聲學模型，能夠有最佳的辨識效果。換句話說，如何定義一組發音符號，能夠有效的整合不同語言間具有相

同發音或相似的發音。而整合的觀念，也可以說是群類化(Clustering)的意思，這代表著說，藉由群類化的技術，將不同語言間發音接近的單位群化，讓這群化後的發音有共同的標音方式，而最後得到多語語言間整合的效果。根據之前學者的實驗結果[2,3]，一組好的多語聲學單位，將有助於最後語音辨識的結果，因此由此可知，聲學單位的選取與定義在多語的語音辨識裡，有著舉足輕重的地位。

依據之前的報告，有學者是使用國際音標系統(International Phonetic Alphabet，簡稱 IPA)[4]來統一標記不同語言間的發音。這套系統是透過語音學或音韻學專家的知識，將不同語言的發音都一對一的對應到 IPA 裡的標音符號，透過這套機制，可將而不同語言間專家所認定相同發音的語音都標記成相同的符號[16]。利用這種方法的聲學模型有個好處，就是相同標音符號但不同語言的訓練語料，能夠彼此分享，使得聲學模型更加強健。但，大部分的語料庫，其劇本都是事先規劃好的，之後再請語者進行錄音，由於資料的龐大，錄完音之後，並沒有做標音符號與真實發音之間的再確認，最後使得，真正所錄下來的發音，和原來的劇本裡所標的發音符號並不完全的一致。如以華語來說，”師”這個發音，往往語者會發成沒捲舌”斯”的情形。因此，有其他學者，用另一套方法，先分析既有的語音資料，而後根據發音相似度的量測，將相似的發音歸類成一群，最後找出最能符合這批語料的聲學單位[2]。近年來，也有學者利用聲學與文脈的分析，產生了多語語音的聲學單位[3]。可是，這些學者所提的方法，並沒有將這些所找到的多語語音聲學單位數目上做最佳化的分析，而是利用不同的門檻值，產生不同的聲學單位數目來做實驗。

本篇論文，提出了一套整合專家背景知識與實際語音分析的方法，來產生一組新的聲學單位，並且對這組聲學單位的數目做最佳的處理。聲學單位是以左右相關音素爲主，實驗在華語及台語的語料庫上。此方法，首先對每種語言，訓練以隱藏式碼可夫模型(Hidden Markov model，簡稱 HMM)爲主的聲學模型，之後透過相似度的量測，產生所有聲學單位的相似度矩陣。然後，加入了語音學以及音韻學的知識，限定了各個聲學單位能群化的上限，根據不同限定的群化上限，使用聚合階層式分群(Agglomerative Hierarchical Clustering，簡稱 AHC)，建立了不同的結構樹，此外，在這裡引進了差分貝式資訊法則 (delta Bayesian Information Criterion，簡稱 delta-BIC)[5]，將同一棵結構樹裡，由下而上，把兩群類的 delta-BIC 值大於 0 的聲學單位做融合，直到 delta-BIC 值小於 0 則停止融合，根據貝式理論，以現有的資料來說，這樣的方式可以找到最佳的聲學單位數目，因此我們才用這種方法來做聲學單位的最佳化。最後，將這些定義好的聲學單位做訓練 HMM 的模型訓練，在每個融合後的聲學單位裡，彼此分享訓練語料，這樣可達到多語聲學的整合效果。

本篇文章的架構如下，第二章：介紹多語聲學單位分類的相關研究，裡面包含了兩種(專家知識和資料驅動)方法。第三章是本論文所提出的新方法，將前章的兩種技術做結合，並且探討了如何把哪些聲學單位該分爲一群，以及最後該如何決定聲學單位的數目。實驗所用到的語料、環境設定和結果分析都在第四章做完整的描述。而最後一章是結論。

二、多語聲學單位分類相關研究

多語聲學單位分類的方法，大致上可分爲兩種：(一)以專家知識的方法； (二)從資料分析的角度(data-driven)，合併多語言之相似音素。現分別介紹如下：

（一）利用專家知識的方法

聲學單位利用專家知識的方法，可分爲 1.語言相關(Language Dependent)與 2.語言獨立(Language Independent)的方式。其中，語言相關的聲學單位是結合各自語言的音素而成的，依據此方法，聲學模型的訓練上，各個語言間具相同發聲的音素彼此之間並不共用訓練語料。如華語和台語，這兩種語言都有/a/的發音，透過專家的認定，將台語語音 /a/ 和華語語音 /a/ 的發音，標記成[a_T]與[a_M]，因此但每個音素符號會分別帶上各個語言的標記。但此作法的缺點是具相同發音的音素在不同的語言裡，彼此的語料並不能共用，而可能會使得某些訓練語料相對的不足，而造成聲學模型在做參數估計的時候會不夠強健。

相反的，語言獨立的方法，則是利用一套能包含所有語言的音標符號，如 IPA、SAMPA [6] 和 Worldbet [7]等，將不同語言但相同發音的音素標記成相同的符號，因此，如台語的 /a/ 和華語的 /a/ 的發音通通都標記成 [a] 的發音符號。此種作法可以有效地將相同發音部分的音素做合併，以減少語音音素的數目，相對的，在同樣多的多語訓練語料上，此種方法其每個聲學單位所能分配到的訓練語料會比其語言相關的方法來的多，因此，所訓練出來的聲學模型參數也會更加強健。<表一>裡記載了華台語用福爾摩沙標音系統(ForPA)[8]所定義出來的音素。ForPA 是根據專家知識所定義的一套可標記台灣主要三種語言(華語、台語和客語)的標音系統，同時，這套標音系統裡的每一個標音都可以和 IPA 做一對一的對應。

然而，使用語言獨立的方法的缺點爲：此方法沒有辦法反映出實際語料的發音特性。原因是，其音素的定義是完全建立在專家的知識上，而非從資料特性上做考量。理論上，假設所有的語料都發音的和標記的完全相同，這樣以專家知識上所定義的聲學單位分類是完美的。但，在真實情況下卻不是這樣的。往往，事先標記好的劇本，錄音員並不能百分之一百的完全照著劇本的發音音素正確的念出來，而造成發音和音素標記上會產生不匹配的現象。因此，我們也要從真實語料上做統計和分析，這樣才能確實地反應真實語料上發音音素的特性。

	子音	母音
OBT	[bh] [gh] [r]	[ah] [ak] [annh] [annp] [ap] [at] [eh] [ennh] [erh] [et] [ih] [ik] [innh] [ip] [it] [oh] [ok] [onnh] [op] [uh] [ut]
TM	[b] [c] [d] [g] [h] [k] [l] [m] [n] [p] [s] [t] [z]	[ann] [a] [enn] [e] [er] [i] [inn] [ng] [o] [onn] [unn] [u]
OBM	[ch] [f] [rh] [sh] [zh]	[ernn] [err] [ii] [yu]

表一、以 ForPA 爲標記的華台語裡的音素表。在這個表中，可分爲 3 個部分，其中 OBT 是台語特有的音素，OBM 是華語特有的音素，而 TM 則是華台語都有的音素

（二）利用資料驅動的方法

此方法是以真實語音資料的發音特性爲考量，根據現有的語料，定義出一組多語聲學單位。主要的觀念是運用群聚技術，將資料依據彼此的相關程度，分成不同的群組；而被凝聚在同一群的發音會有某些特性是相近的，也就是說透過真實語料的分析，有相同特性的發音會被標記成一致的發音符號。在此所用到的群聚技術是以階層式的爲主，

這個技術可分爲兩類，1. 分裂法(Divisive algorithm) [9] 2. 聚合法(Agglomerative Algorithm) [10] 二種。分裂法是先把整個資料集合看成一個群聚，然後逐次分裂，每次都會在其中一個群聚裡，切割相似度最低的連結，成爲二個較小的群聚，直到群聚數目達到事先所設定的數目爲止。而聚合法是先將每一筆資料視爲一個群聚，然後每次將特性最相近的二個群聚合而爲一，直到群聚數目達到事先所設定的數目爲止。以後者爲例，在做聚合階層式的群聚(AHC)技術的演算法之前，會算出所有將要群聚的聲學單位的相似程度矩陣，而此矩陣的數值是利用訓練好的聲學模型參數來算出彼此間的距離，而距離有兩種普遍的計算方法，分別爲 a. Bhattacharyya distance [2] 和 Kullback-Leibler divergence [11],其公式如下：

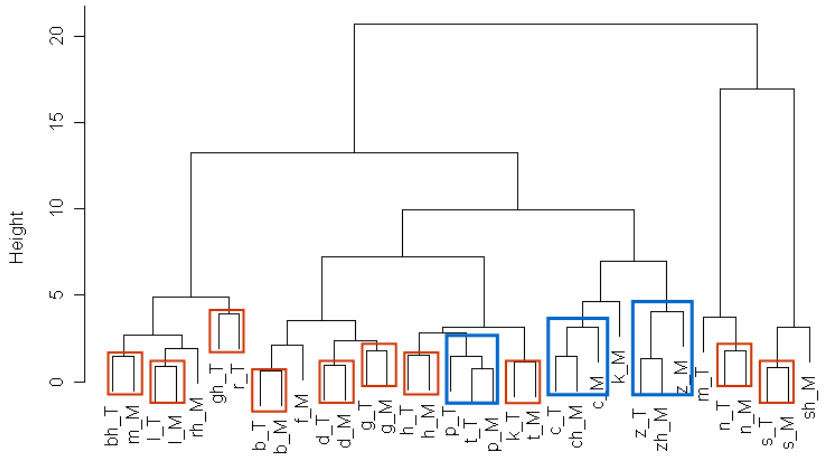
$$D_{bata} = \frac{1}{8}(u_p - u_q)^T \left[\frac{\Sigma_p + \Sigma_q}{2}\right]^{-1} (u_p - u_q) + \frac{1}{2}\ln \frac{\left[\frac{\Sigma_p + \Sigma_q}{2}\right]}{\sqrt{|\Sigma_p||\Sigma_q|}} \qquad \text{(式一)}$$

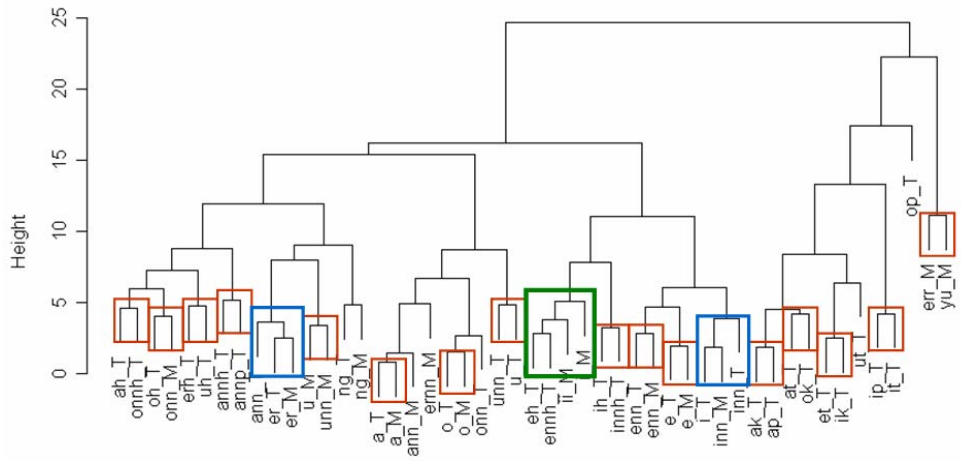
$$D_{KL} = \frac{1}{2}\left(\ln \frac{\Sigma_p}{\Sigma_q} + trce\left(\Sigma_p^{-1}\Sigma_q\right) + (u_p - u_q)^T \Sigma_p^{-1}(u_p - u_q) - d\right) \qquad \text{(式二)}$$

而其 AHC 的評估方式有四種[12]，分別爲 A.重心連結聚合演算法、B.平均連結聚合演算法、C.完整連結聚合演算法以及 D. 單一連結聚合演算法。

不論是由上而下(top-down)的分裂法或由下而上(button-up)的聚合法，其最後的聲學單位數目，都是事先決定好的，之後產生固定的聲學單位來訓練聲學模型。之前的學者使用這兩種階層式方法，並沒有對這些方法做出聲學單位數目的最佳化。除此之外，這些聚合方法，捨棄了專家知識，而直接只採用現有的語音資訊去作分析去定義最後的聲學單位，如果此批語料摻雜了許多的雜訊或某些聲學單位的訓練語料不足的時候，往往可能會產生比用專家知識方法還差的辨識結果。

<圖一>中爲用標準 AHC 所產生的華語與台語其母音與子音的樹狀圖，其中距離的計算是採用歐基里德距離，而 AHC 的評估方式爲完整連結聚合演算法。而在 AHC 之前的相似度矩陣值由每個聲學單位其 HMM 裡的三個狀態下的 Bhattacharyya 距離的平均值。



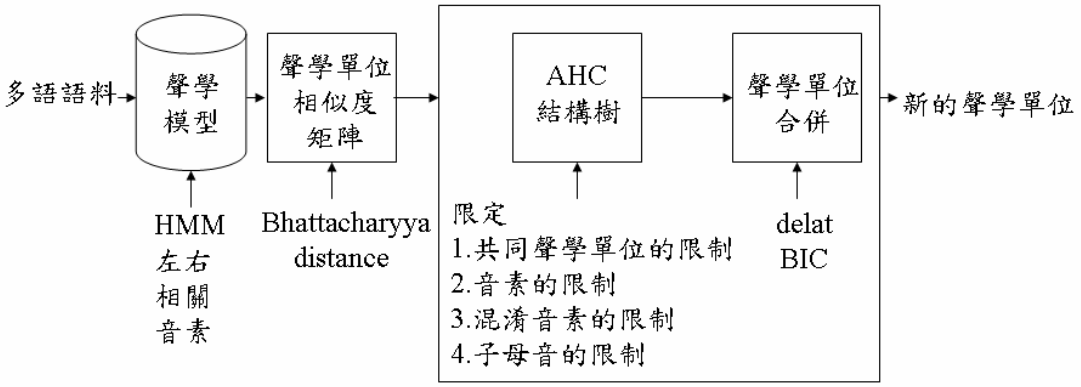


圖一、標準 AHC 所產生的華語與台語母音與子音的樹狀圖

三、聲學單位的分類與數目的最佳化

本篇論文提出了一套，結合了專家知識和資料驅動的方法，從既有的多語語料當中，尋找一組新的聲學單位且讓其數目最佳化。我們依據聲韻學和語音學的語音知識，限定了群聚技術裡的分類，讓發聲相近的聲學單位，透過相似度的篩選機制，建立了以 AHC 爲方法的結構樹，在相同的結構樹中，聲學單位有機會做合併。然而，不同的結構樹中的聲學單位彼此之間不能做合併。這樣的機制，避免了發聲差很多的聲學單位，因爲語料品質的影響（如雜訊），而相互融合。

另外，如何在相同的結構樹中，最後找出最佳的聲學單位數目，也就是說，如何決定哪些聲學單位該合併，哪些群不該合併。在此，我們引進了 delta-BIC 技術。根據貝式定理，從現有的訓練語料中，利用 delta-BIC 模型選擇的方法，可找出最佳的模型。相同的，我們將這個觀念，應用在如何找出最佳的聲學單位。而整個尋找聲學單位數目的最佳化過程顯示在<圖二>。



圖二、結合專家知識和資料驅動方法產生最佳聲學單位數目的流程圖

因此，由<圖二>可看出，要尋找出一組新的聲學單位，有兩元件程是必須探討的：AHC 結構樹的限制與聲學單位合併的機制，以下，我們就分別細說這兩項元件。

（一）限定結構樹範圍 - 依據專家知識

根據純資料驅動的做法，從聲學單位相似度矩陣到產生 AHC 結構樹時，並沒有做什麼機制來防止因語料品質的缺陷，而造成 AHC 在建立結構樹的時候，將一些發聲差別很大的聲學單位反而安排的很近，因爲標準的 AHC 是完全按照相似度矩陣來做聲學單位的分類，而相似度矩陣是由語料庫所建立的聲學模型所產生的，因此 AHC 結構樹長的好不好，有很重的部份是依賴語料庫的品質優劣。因此，我們在這一節將做一些限制，來改善因語料的錄音品質缺陷所造成 AHC 長的不好的問題。

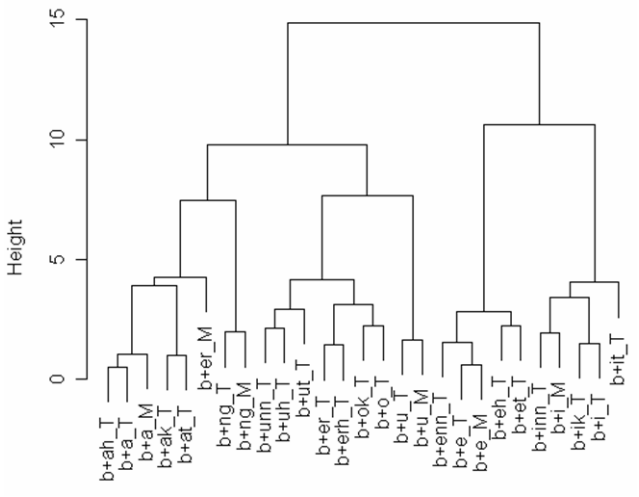
由 IPA 或 ForPA 等標音系統所定義出的各語言聲學單位，在語言學或聲韻學上是有道理的，因爲這些聲學單位，都是由一些專家所定義出來的。根據這些定義，我們可以將這些單位，粗分爲子音、母音、或由不同的發聲構造來分成，鼻音和喉頭音等。因此，在 AHC 之前，我們先將所有的聲學單位分成子群，而這些子群的成立，是根據以下的四種分群限定。我們依此道理來將我們從由 Bhattacharyya 所算出的每個聲學單位裡的狀態混淆矩陣，到由 AHC 所產生的分類樹之間，做四種階層的限制。而不同的限制，也將會產生不同結構的分類樹。

I. 共同聲學單位的限制：

在這個限定下，AHC 結構樹的數目，只針對各語言間的具有相同的 IPA 標音的聲學單位，而每個子群裡的聲學單位爲左右相關音素。以華台語舉例來說，[b-a+ng_T, b-a+ng_M]，這樣的結構樹共有 284 個。因此，這些結構樹，只有兩層，一層就是華台語的左右相關音素，第二層就是他們兩個的合併 [b-a+ng]。這個限制，主要是在觀察，華台語之間共同左右相關音素是否應該合併，如果完全合併，則是語言獨立的方法，相反的，如果全部不合併的話，就退回語言相關的方法了。

II. 音素的限制：

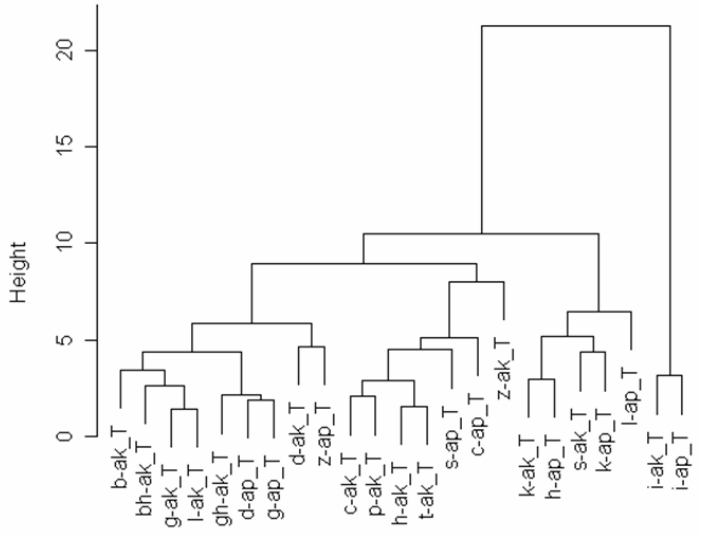
這個限定，是在觀察左右相關音素與左右獨立音素之間的關係。因此，每棵樹的最底層是左右相關音素，而最上層爲左右獨立音素，如<圖三>所示。因此，如果到最後在這棵樹裡的每個左右相關音素都能夠合併，則原本以[b+*]爲主的左右相關音素，就會退化成語言獨立的左右獨立音素的[b]



圖三、以[b]爲例的 AHC 結構樹

III. 混淆音素的限制：

根據[13]，我們了解到，有些音素是很容易混淆的，比如華語的捲舌音和不捲舌音，或台語的入聲音，如帶-p 和-k 結尾的音素，因此，這個限制裡，我們擴大可以合併的範圍到混淆音素。如<圖一>的字音和母音 AHC 的結構樹圖中，我們用 delta-BIC 的技術(下一節會談到)，將結構樹中，delta-BIC 值大於零的聲學單位群組化，而被群組化的聲學單位就被認定是一組混淆音素組，之後，以此組音素所衍生的左右相關音素就可用 AHC 的方式產生一棵結構樹。如<圖一>裡我們可以看出，子音共有 13 組混淆音素（用框框圍住的），而母音則有 19 組。這些混淆音素，最小的是由兩個音素所組成的，而最大的是四個音素。如-ap 與-ak 就是一組混淆音素，而以-ap 和-ak 爲主的左右相關音素所產生的 AHC 結構樹顯示在<圖四>中。



圖四、母音中帶-ap 和-ak 結尾音素的 AHC 結構樹

IV.子母音的限制：

在這個階段，我們只將分類樹的樹頭分爲兩類，子音類和母音類。而樹根則爲華台語所有的左右相關聲學單位。因此在最極端的情況下，華台語的聲學單位，最後則會分成只有字音和母音兩個。

（二）決定最佳數目 - 依據差分貝式資訊法則 (delta Bayesian Information Criterion)

貝式資訊法則(簡稱 BIC)，由 Schwarz 在 1978 所提出[2]，是一種非對稱模組選取的準則。其利用最大概似度(Maximum Likelihood)的方式從 p 個模型中找出最能代表 n 比資料 $X = x_1,...,x_n, x_i \in R^d$ 的最佳模型。假設，每筆資料都是相互獨立的。而第 p 個模型的 BIC 公式如下

$$BIC_p = \log L_p(X) - \frac{1}{2}\lambda d_p \log n \qquad \text{(式三)}$$

其中 L_p 是模型 p 的最大概似度，λ 是一個微調值，而 d_p 是 p 模型裡的參數數目。

delta-BIC 則是將兩群不同的貝式資訊法則值做相減，簡化之後，可以轉化成兩個不同群的最大概似度值做相加，之後減去兩群組合併後的最大概似度值，再加上模型參數目的函數值。這項技術常常被拿來當作語者交換點或語言交換點偵測的一個判斷[5]，通常一般法則是當 delta-BIC 大於 0 的時候，就判定兩群組的邊界資料爲一個交換點。而本篇論文將此方法拿來作爲多語聲學模型單位最佳數目的根據。根據[5]，針對第 p 個、第 q 個模型和兩個合併之後的 r 模型的 delta-BIC 公式如下所示：

$$\Delta BIC_{pq} = BIC_p - BIC_q$$
$$= -\frac{n_p}{2}\log|\Sigma_p| - \frac{n_q}{2}lpg|\Sigma_q| + \frac{n_r}{2}\log|\Sigma_r| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n_r \qquad \text{(式四)}$$

其中 n_p, n_q 分別爲模型 p 和 q 所對應的訓練語料數目，$n_r = n_p + n_q$ ，而 Σ_p, Σ_q 分別爲模型 p 和 q 的共變異數矩陣的行列式， d 爲參數數目。

因此，由(式四)我們可看出，如果模型 p 和 q 以及合併之後的 r 模型，其最大概似度還大於合併之前個別模型 p 和 q 的最大概似度總和，則我們相信，這兩個模型是可以合併的。所以，針對上述的聲學單位分類範圍，透過 delta-BIC 的方法，最後可找出最佳的聲學單位數目。而如果其中的 λ 值爲零的話，(式四)就退化成一般的最大概似度估計了。但一般在做模組選取時是採用 delta-BIC，因爲直接用大於或小於零就可判別，但如用一般最大概似度估計，則要對每個不同的狀況設下不同門檻值，才能作決定，因此本論文也採用 delta-BIC 來做多語聲學模型最佳化的判準。

根據之前依不同限定所產生的結構樹，我們從結構樹最下面兩兩聲學模型或群組間的 delta-BIC 值大於零的部分做合併，而合併後的聲學單位，利用[14]的技術，將原本未合併之前模型 p 和 q 所對應的訓練語料分享給新的合併之後的 r 模型，而達到聲學單位整合的目的。這樣的合併會採由下而上的方式一直進行，直到當 delta-BIC 值小

於零的時候才會停止。停止後所合併的群組則爲新的一組聲學單位，用來訓練聲學模型，做語音辨識。

四、實驗與結果

（一）實驗語料

本實驗，是爲了找出最佳的多語聲學單位，並且驗證此組聲學單位能達到最佳的語音辨識結果。因此，我們用了華台雙語語音資料庫 ForSDAT 中的 01 年麥克風語料[8]。這個語料庫裡包含了 100 人的訓練語料和另外 20 人的測試語料。在訓練語料中，每個語者都錄製了兩種語言，且男女比率平均。其相關的統計數字列在<表二>

	語言	人數	語音句數	總時間（小時）
訓練語料	華語	100	43078	11.3
	台語	100	46086	11.2
測試語料	華語	10	1000	0.28
	台語	10	1000	0.28

表二、ForSDAT01 年華台雙語語音資料庫的相關統計數字

（二）實驗設定

由之前的介紹，我們可將多語聲學單位的實驗依三種方法來分類，分別爲語言相關(Lang-De)、獨立語言(Lang-In)和本篇所提出的聲學單位數目最佳化，而在本論文所提出的方法中，又依不同的限定，可分爲 4 類，分別是共同聲學單位的限制(C-I)、混淆音素的限制(C-II)、音素的限制(C-III)和子母音的限制(C-IV)。

在特徵擷取上，我們採用了以梅爾倒頻係數(簡稱 MFCC)爲主的方法，而 20 毫取一個音匡，每 10 毫秒移動一個音框，每個音匡有 39 維度。每個聲學模型使用 HMM 來做訓練，而模型的單位爲左右相關音素，而每個 HMM 有 3 個狀態，每個狀態下的高斯分佈模型(簡稱 GMM)數目，則是依照每個狀態下所能對應到的訓練語料量來決定，根據[15]原則，所對應到的訓練語料量越多的狀態，其 GMM 數目也會越多，在此系列實驗中，我們設定每個 GMM 必須要有 30 個以上的音匡。所以如果每個狀態下的 GMM 數目的增加是根據訓練語料的多寡來決定，那稱之爲"*動態 GMM*"。而相反的，如果每個狀態下的 GMM 數目是以固定的倍數增加，那我們稱之爲"*固定 GMM*"。

由於 delta-BIC 的計算是依照 GMM 的單位來做的，而聲學單位又是以 HMM 爲單位，因此，在做聲學單位最佳化的過程中，除了以 HMM 爲合併單位之外，我們還用狀態爲單位來合併。以 HMM 爲單位的時候，是否合併，是觀察 HMM 下三個狀態的 delta-BIC 平均值來判斷。而根據我們觀察，有些 HMM 中前兩個狀態其 delta-BIC 都小於零，而第三個卻大於零，但平均之後還是小於零，因此不能合併。爲了讓合併更佳寬鬆，讓每個狀態都能自由的做合併與否的判斷，我們也採用了狀態爲單位的合併。

實驗的目的要觀察不同的聲學模型間的語音辨識率，因此，語言模型的機率在這一系列的實驗中，都讓每個不帶聲調的音節之間的機率設爲相同，也就是說，其爲均勻分

佈。而依此所產生的搜尋網路複雜度爲 924。

（三）結果分析

A. 以專家知識爲本的固定 GMM 與動態 GMM 辨識結果

我們將 Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的比較，結果顯示在<表三>，而相對應的 GMM 總數和平均數列在<表四>。

		8-mix	16-mix	32-mix	64-mix
動態 GMM	Lang-De (1503)	60.7	63.9	62.1	60.2
	Lang-In (1242)	62.5	**64.7**	64.3	63.0
固定 GMM	Lang-De (1503)	59.3	62.8	60.2	58.6
	Lang-In (1242)	61.4	63.1	62.5	61.6

表三、 Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的語音辨識正確率。

		8-mix	16-mix	32-mix	64-mix
動態 GMM	Lang-De (1503)	32,848(7.1)	54,824(12.1)	88,000(17.8)	109,905(24.3)
	Lang-In (1242)	26,328(7.1)	**46,360(12.4)**	69,328(18.6)	98,857(26.5)
固定 GMM	Lang-De (1503)	36,054(8.0)	72,144(16.0)	126,252(28.0)	261,522(58.7)
	Lang-In (1242)	29,787(8.0)	59,616(16.0)	115,506(31.1)	219,834(59.6)

表四、Lang-De 與 Lang-In 的聲學模型用來做固定 GMM 與動態 GMM 的 GMM 總數和平均數(括弧中的值爲平均值)

在<表三>中，可看出動態 GMM 的語音辨識結果比固定的方式的結果來的佳。這表示說，GMM 的增加，必須考慮到實際語音訓練量的狀況，而不能一昧著增加 GMM 的數目。另外，對於動態和固定的方法，其最佳辨識率都出現在每個狀態裡最高有 16 個 GMM 的設定下，之後再增加 GMM 的數目，辨識率反而會下降。而固定方法的下降率會比動態方法的下降率來的多。因此，我們在 Lang-In 用動態方式產生 GMM 於 16-mix 實有最佳的辨識率，爲 64,7%。由於我們所使用的軟體爲 HTK[14]，所以在估計 GMM 參數時，會將每個 GMM 的比重做調整，因此，有些比重較輕的 GMM 會被刪除，而造成固定方法的 GMM 平均數的下降。其中，Lang-De 的聲學數目爲 1503，而 Lang-In 的聲學數目爲 1242。

雖然出動態 GMM 的語音辨識結果比固定的方式的結果好，但<表四>裡，動態方法的總 GMM 數會比固定方法的總 GMM 數來的少，因此，這裡可看出，並不是總 GMM 多，辨識率就會跟著變好。反而倒是要顧慮每個狀態下平均 GMM 數要配合實際的訓練語音量，這樣才能得到好的辯識結果。

B. 以 HMM 爲單位的最佳化聲學模型的結果

由之前的結果得知，用動態的方式來增加 GMM 會有不錯的結果，本小節也是用此方法，但卻是使用本論文提出的結合專家知識與資料驅動的方法來產生最佳化的聲學模型單位。辨識率與 GMM 相關資訊分別列於<表五>和<表六>。第一欄位中的數字爲最

後每個不同限定下所產生的聲學單位數目。

	8-mix	16-mix	32-mix	64-mix
C-I (1242)	62.5	64.7	64.3	63.0
C-II (527)	51.7	56.7	60.4	59.4
C-III (1083)	59.5	64.2	**65.7**	**66.1**
C-IV (862)	56.4	59.6	61.8	61.5

表五、限制下最佳化聲學模型的辨識結果

	8mix	16mix	32mix	64mix
C-I (1242)	26,328(7.1)	46,360(12.4)	69,328(18.6)	98,857(26.5)
C-II (527)	12,578(7.8)	24,671(15.6)	41,459(26.2)	55,285(36.4)
C-III (1083)	24,618(7.5)	46,256(14.3)	**78,841(24.3)**	**98,754(30.4)**
C-IV (862)	19,590(7.6)	38,498(14.9)	63,784(24.6)	84,059(32.5)

表六、不同限制下最佳化聲學模型的 GMM 總數和平均數(括弧中的值爲平均值)

C-I 爲共同聲學單位的限制，其聲學數目和 Lang-In 的數目是一樣的，這表示說，具有相同標音符號但隸屬不同語言的聲學單位，其 delta-BIC 值都大於零，因此其聲學單位都會合併。在<表五>中，我們最佳的辨識結果是採用 C-III 的方法，辨識率爲 66.1%，此方法連產生混淆音素都是用 delta-BIC 來做決定的。值得一提的是，利用此方法，從 8-mix 到 64-mix 的結果都是節節上升，和其他方法不同，有的從 16-mix 或 32-mix 之後，辨識率就開始走下坡了。與 Lang-In 或 C-I 來比較，在 32-mix 與 64-mix 時，C-III 都有較佳的辨識率。此外，觀察<表六>我們也可發覺和<表四>有相同的結論，就是 GMM 總數多並不一定代表辨識率一定高。還要配合每個狀態下 GMM 的平均值，才能使辨識率上升。如比較 C-I 和 C-III 在 64-mix 時的辨識率與 GMM 的關係，可以發現這兩個的 GMM 總數差不多，但後者的每個狀態平均 GMM 數卻高於前者，這表示說，C-III 的方法不僅可將發音相似的發音合併，但合併的同時，也把那些本來訓練語料相對少的聲學模型聚合在一起，因爲 delta-BIC 的公式中也考慮到合併後和合併前的訓練語料出現的數目，因此在這兩個因素下，才能將辨識率向上提升。而 C-II 的方法，雖然在每個狀態下有最多 GMM 平均數，但最後的聲學模型數目是原始的三分之一左右，使得聲學單位在這麼少情況下能有良好的鑑別率，而造成在<表五>裡其辨識率是進陪末座。

C. 以狀態爲單位的最佳化聲學模型的結果

以上C-I到C-IV的聲學模型都是以HMM作爲合併的單位，但delta-BIC的計算是以狀態爲單位，因此這裡我們做了比較有彈性的變動，將合併單位從HMM轉爲狀態。語音辨識結果呈現在<表七>。其中C-I與C-II的結果和<表五>相同，這表示說，這兩個的delta-BIC值，其每個狀態是否要合併和以HMM爲單位是一樣的。第一欄位中的數字爲狀態數目。

而另外一方面，在傳統聲學模型的合併裡，有相關學者是採用決策樹(decision tree)的方法，用一些語音學、或語言學上的專家知識規則，來將聲學模型做做分類，相同類的聲學模型就互相分享訓練語料[16]。因此在這裡，我們也利用這樣的技術，來和本篇

所提出的方法來做比較。

而決策樹和 C-III 的方法最大差別就在於，1.前者是使用最大相似度法則而後者是採用 delta-BIC 來做分類以及尋找混淆音素。使用最大相似度法則時，需要制訂一個門檻值，好讓決策樹在分類的時候能停止分裂，不同的門檻值會影響到最後聲學模型的狀態數目，或說是狀態的解析度。而 delta-BIC 則不用到門檻值，而是採用一個固定的懲罰值來替代，而這個值的大小，根據我們的實驗經驗，並不會對最後的狀態結果有很大的區別。所以我們用 delta-BIC 來做聲學模型最佳化的時候，考慮到其 delta-BIC 值是否小於 0，並不用再設定其他的參考值來做最佳化。可是在決定不同決策樹的門檻值時，其最後結果（狀態數目）也會跟著不相同。因此我們試著用幾組不同的門檻值來做停止分裂的條件，把最好的辨識結果呈現在表七的最後一行。

	8-mix	16-mix	32-mix	64-mix
C-I (3726)	62.5	64.7	64.3	63.0
C-II (1581)	51.7	56.7	60.4	59.4
C-III (3569)	61.9	65.1	66.4	**66.7**
C-IV (2760)	59.2	61.5	62.3	62.5
DT(3374)	62.2	63.4	64.7	64.9

表七、以狀態爲單位的最佳化聲學模型的辨識結果，其中最後一行是決策樹的結果。

整體來說，使用狀態爲合併單位所訓練出來的聲學模型(C-III 與 C-IV)，其辨識率比以 HMM 爲合併單位所訓練出來的聲學模型來的好。另外使用決策樹方法所做出來的效果，在 8-mix 的時候有超過 C-III 的結果，但在增加 GMM 的數目之後，其效果就沒有 C-III 來的好，但也是比 C-I,C-II 和 C-IV 來的好。分析如下：決策樹裡的問題共分三大類，其分別爲：子音、母音、和語言議題，在這些議題下，我們總共產生 124 個問題來將這些聲學模型做分類。有了這些的問題，我們也就不用如 C-I 到 C-IV 的限制了。所以，在這些限定下，決策樹還是有他的優點，因爲是一次考慮到這四類的問題，而不是如 C-I 到 C-IV，是一次只考慮到單一種狀況。決策樹的樹頭是整個左右相關聲學模型(3726)，最後則剩下 3374 個狀態。但在最後結果上，還是 C-III 的 64-mix 勝出，同時這也產生了本論文的最佳辨識率 66.7%。

五、結論

本論文提出了一套整合專家背景知識與實際語音分析的方法，利用聲韻學和語音學的知識，將 AHC 的結構樹做分類的限定，再由 delta-BIC 的判斷來把同一棵樹中的聲學單位做合併，並且找出最佳數目的一組新的聲學單位，再將他們以 HMM 重新訓練，實驗在 ForSTDA01 的華台雙語語料庫。實驗結果驗證了，1.利用動態方式增加 GMM 的辨識率比固定方式的方法來的高。2.結合專家知識與資料驅動的方法所訓練的新聲學模型，其辨識率比只用專家知識所訓練的新聲學模型來的佳，在訓練的同時，能提高每個狀態下

平均 GMM 的個數。3.利用狀態爲合併單位比 HMM 爲合併單位，更能充分的反映出哪些聲學模型需要做合併，產生更有彈性的合併，而達到最佳的辨識效果。

參考文獻

[1] T. Schultz and A. Waibel, "Multilingual Cross-lingual Speech Recognition," in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] Brian Mak and Etienne Barnard, "Phone clustering using the Bhattacharyya distance," in Proc. of ICSLP 1996, pp. 2005-2008.

[3] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin, "PHONE SET GENERATION BASED ON ACOUSTIC AND CONTEXTUAL," in Proc. of ICASSP 2006

[4] Mathews, R. H., 1975. Mathews' Chinese-English Dictionary, Caves, 13th printing.

[5] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in Proc. EUROSPEECH 1999, pp. 679–682.

[6] J. C. Wells, "Computer-Coded Phonemic Notation of Individual Languages of the European Community," Journal of the International Phonetic Association, 1989, pp. 32-54.

[7] James L. Hieronymus, "ASCII Phonetic Symbols for the World's Languages: Worldbet," Journal of the International Phonetic Association, 1993.

[8] Ren-yuan Lyu, Min-siong Liang, Yuang-chin Chiang, "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin," Computational Linguistics and Chinese Language Processing, Vol. 9 (2), 2004, pp. 1-12.

[9] Y. J. Chen, C-H. Wu et al. "Generation of robust phonetic set and decision tree for Mandarin using chi-square testing," Speech Communication, Vol. 38 (3-4), 2002, pp. 349-364.

[10] T.S. Chen, C.C. Lin, Y.H. Chiu and R.C. Chen "Combined Density- and Constraint-based Algorithm for Clustering," In Proceedings of 2006 ICISKE 2006.

[11] Jacob Goldberger and Hagai Aronowitz, "A Distance Measure Between GMMs Based on the Unsented Transform and its Application to Speaker Recognition," in Proc. of EUROSPEECH 2005, pp. 1985-1988.

[12] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[13] LIU Yi and Pascale Fung, "Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech," in Proc. of EuroSpeech 2005.

[14] Phil C. Woodland, Steve J. Young, "The HTK Tied-State Continuous Speech Recogniser," in Proc. of EurpSpeech 1993.

[15] X. Anguera, T. Shinozaki, C. Wooters, and J. Hernando, "Model Complexity Selection and Cross-validation EM Training for Robust Speaker Diarization," in Proc. of ICASSP

2007, Honolulu, HI. April 2007.

[16] Dau-Cheng Lyu, Bo-Hou Yang, Min-Siong Liang, Ren-Yuan Lyu and Chun-Nan Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," in Proc. of SST 2002.

詞義辨識:機器學習演算法特徵的選取與組合

高紹航
denehs@gmail.com
臺灣大學資訊工程學系

高照明
zmgao@ntu.edu.tw
臺灣大學外國語文學系

摘要

詞義辨識(Word Sense Disambiguation, WSD)是自然語言處理中很重要的一環，本文利用 Naïve Bayes 的分類器(classifier)選用目標詞的詞性，相鄰詞及詞性，搭配語，語法依存關係及語義特徵等共九項特徵，以 Le and Shimazu (2004)所提出的 Forward Sequential Selection Algorithm 來得到最佳的特徵組合。我們以 Senseval-2 中的 English lexical sample 做爲訓練語料以及測試語料，得到 61.2%的正確率，與 Senseval-2 參賽第一名的隊伍 64.2%的正確率相差 3%，與第 4 名史丹福大學的參賽隊伍相差 0.5%。

關鍵詞：詞義辨識(word sense disambiguation)，搭配語(collocation)，語法依存關係(dependency relations)，sketch engine，Stanford parser，Hownet，Naïve Bayes，Forward Sequential Selection Algorithm

一、前言

一個英文詞可能有好幾個不同的意思，例如 bank 有銀行，河堤，庫等多個意義。詞義辨識的目的就是要讓電腦自動辨識一個岐義詞在某一個語境裡正確的意義。由於現有詞性標記的演算法正確率都相當的高，如果岐義詞的意義具有不同的詞性很容易透過詞性標記程式辨識出不同的意義。而像前面的例子 bank 不同的意義如銀行，河堤，庫都是名詞，辨識的困難度增高許多。我們所使用的訓練語料 Senseval-2 English lexical sample，是在 2001 年所發布，語料中包含了 73 個不同的目標詞，詞性有名詞、動詞、形容詞，但同一個目標詞的不同意義詞性都是相同的，對於詞義辨識的演算法形成很大的挑戰。

Senseval-2 的訓練以及測試語料是以 XML 的型式儲存，以下是一筆訓練語料的範例：

```
<instance id="art.40001" docsrc="bnc_ACN_245">
<answer instance="art.40001" senseid="art%1:06:00::"/>
<context>
Their multiscreen projections of slides and film loops have featured in orbital parties, at the
```

Astoria and Heaven, in Rifat Ozbek's 1988/89 fashion shows, and at Energy's recent Docklands all-dayer.
From their residency at the Fridge during the first summer of love, Halo used slide and film projectors to throw up a collage of op-art patterns, film loops of dancers like E-Boy and Wumni, and unique fractals derived from video feedback.
&bquo;We're not aware of creating a visual identify for the house scene, because we're right in there.
We see a dancer at a rave, film him later that week, and project him at the next rave.&equo;
Ben Lewis Halo can be contacted on 071 738 3248.
<head>Art</head>you can dance to from the creative group called Halo
</context>
</instance>

語料中目標字會用<head>以及</head>標出，測試語料格式與訓練與料相似，其差別在於沒有 senseid 的標記。

二、文獻回顧

早期詞義辨識的演算法大都利用利用辭典的定義、或同義詞辭典(thesaurus)的語義分類訊息。例如 Lesk (1986) 判斷目標詞的語境與辭典的哪一個意義的定義最接近，所採用的相似度計算方式以兩者相同的非功能詞的數目爲主。Walker (1987)則利用同義詞辭典(thesaurus)當中的語義類別。這些演算法跟目前常用的機器學習演算法相比正確率低許多（請參考表 16 Senseval-2 詞義辨識競賽各個方法的正確率）

機器學習方法主要可分爲監督式(supervised learning)及非監督式(unsupervised learning)。兩者的差別在於前者的訓練語料有標記答案的而後者沒有，我們所採用的方法是監督式的方法。無論是哪一種機器學習的詞義辨識演算法都需要利用語境的訊息。例如 Purandare and Pedersen (2004) 採用非監督式的方法，從沒有標示詞義純文字語料抽出語境並將機讀辭典 Wordnet 裡面不同詞義的定義去除功能詞後建立共現矩陣(co-occurrence matrix)，利用 Singular Value Decomposition (SVD)將維數降到 100，最後用 Latent Semantic Indexing (LSI)找出某一句中的目標詞最有可能的詞義。Jurafsky and Martin (2000)將常用的語境特徵分成兩類。一類是搭配語特徵(collocational features)，另一類是 bag of words information。兩者的最大差別在於後者只考慮某些詞在目標詞左右一定範圍的詞有沒有出現，不考慮這些詞彼此或跟目標詞前後的關係，而前者則納入與目標詞前後相對位置的訊息，甚至用語法剖析器得到語法依存關係。

詞義辨識方法除了可以利用 Semantic Concordancer 或 Senseval 這些有標示詞義的語料之外，還可以利用 pseudoword 或雙語語料。pseudoword 是 Gale et al. (1992)和 Schutze(1992)爲了省去標示詞義所需的大量人力與時間所創造出來的方法。透過人造的岐義詞如 banana-door，將語料中所有出現 banana 或 door 都代換成 banana-door，這樣就可以得到類似人工標記詞義的訓練語料。此外，某一個有岐義的詞在另一個語言通常沒有岐義，例如英文的 duty 有兩個意義，但在中文裡則由海關和責任兩個詞來表達。Brown et al. (1991) 及 Gale et al. (1992)利用這個特性，以英法雙語語料庫作爲訓練語料，採取目標詞左右若干詞（例如 50 個詞）構成一個語境向量(context vector),再利用 Bayesian classification 來選擇在某一個語境當中哪一個詞義的機率最大。我們也採用 Bayesian classification 但搭配不同的特徵。Bayesian classification 的概念是目標詞周圍的

詞會反映出目標詞的意義，因此將周圍的詞以及目標詞做統計再利用機率選擇詞義，在第三節中會有詳細的介紹。

Yarowsky (1995)注意到在某一篇文章中一個目標詞的詞義通常是固定某一個詞義(One sense per discourse)。且目標詞的搭配語提示了這個目標詞的詞義(One sense per collocation) 。本文所採用搭配語作爲機器學習演算的法特徵受到 Yarowsky (1995)的啓發。Lin (1997)有鑑於以機器學習分類器(classifier)來辨識詞義需爲不同的詞分別訓練出不同的分類器，頗不方便，因此提出一種使用同一種知識來源(knowledge source)的方法。他利用自己所發展的 MINIPAR 英文剖析器得到的語法依存關係(dependency relations)，如動詞與受詞的關係作爲機器學習演算法的特徵。比較特別的地方在於他的方法不需要標示詞義的語料，而是利用相同語意的詞會出現在具有相同的依存關係所組成的局部語境(local context)。Lin (1997) 的正確率達到與其它機器學習演算法相同水準。本文採用語法依存關係作爲詞義辨識的特徵源自於 Lin (1997)的想法。有關於特徵的選取，Le and Shimazu (2004)針對英文詞義辨識提出數個特徵並以 Forward Sequential Selection Algorithm 來得到最佳的特徵組合，本文採用 Le and Shimazu (2004)所提出的 5 個特徵另外加上 4 個特徵，並仿照 Le and Shimazu (2004)所使用的 Forward Sequential Selection Algorithm 得到最佳特徵的組合。

除了上面介紹的方法，還有許多詞義辨識的方法，例如利用 mutual information 的 Flip-Flop algorithm (Brown et al. (1991)),使用 decision list (Yarowsky (1994))等，限於篇幅無法一一介紹。近幾年詞義辨識的演算法除了 Naïve Bayes 之外，越來越多人使用 Maximum Entropy，Support Vector Machine，及 Conditional Random Field 等較新的機器學習演算法。本文所選取特徵和組合的方法也可以與這些方法一起使用。

三、我們採取的方法

（一）、Bayesian Classification

在我們的實驗中，我們採用 Bayesian Classification 搭配多種特徵的方法，下面簡述 Bayesian Classification。

假設我們現在要對一個目標詞做詞義辨認，該目標詞的詞義有 k 個，依序是$s_1, s_2, \ldots, s_k$，則目標就是要找出一個s'，使得$P(s'|c)$爲最大，c 是目標詞所含有的某種特徵。根據貝式定理，可以得到如下的等式：

$$P(s_k|c) = \frac{P(c|s_k)}{P(c)}P(s_k)$$

因此

$$s' = \arg\max_{s_k} P(s_k|c)$$

$$= \arg\max_{s_k} \frac{P(c|s_k)}{P(c)}P(s_k)$$

$$= \arg\max_{s_k} P(c|s_k)\,P(s_k)$$

$$= \arg\max_{s_k} [\log P(c|s_k) + \log P(s_k)]$$

我們所有的實驗都是使用這個方法來作詞義辨認，差別是在於選取的特徵的不同。

（二）、Forward Sequential Selection Algorithm

在特徵的選取方面，由於我們嘗試了很多種特徵，假如特徵有 7 種那麼特徵組合的種類就有 127 種，數量非常的可觀，一個一個將所有的組合做實驗非常沒有效率，因此使用 Le and Shimazu (2004)所提出的 Forward Sequential Selection 演算法來挑選特徵。這個方法大致上是先令一個特徵的集合 S 爲空集合，首先挑一個最好的特徵放進 S 中，接著將每一個特徵都放進 S 中看哪個得到的正確率最高來決定第二個要放入 S 中的特徵，如此反覆直到最後正確率不再增加爲止，最後集合 S 中的特徵就會是一個很不錯的特徵組合，雖然未必真的是最佳解但應用在英文詞義辨認的特徵選取上與真正最佳解的差異非常小。

（三）、特徵

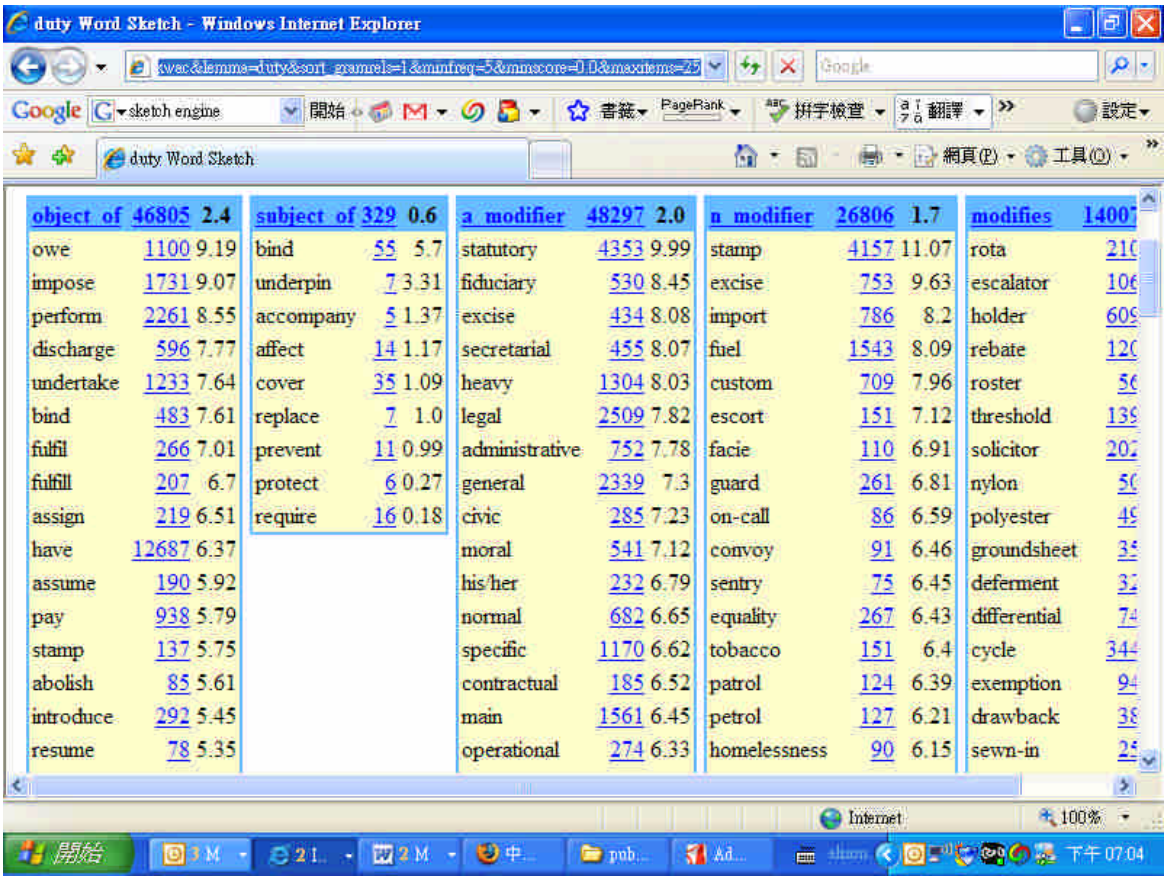
我們一共嘗試了 9 種特徵，分別以 F1 到 F9 命名之，前五個主要是針對目標詞周圍的詞以及其詞性，這 5 個特徵都是 Le and Shimazu (2004)所使用的特徵，第六個以及第七個則著重在詞的依存關係，例如主詞與動詞，動詞與受詞的關係等等，而最後兩個則是利用 HowNet 的取目標詞的前後兩個詞以及與目標詞有依存關係的 HowNet 義元(語義特徵)當作特徵，在此一一介紹。

F1 是直接把目標詞周圍的詞做爲特徵，但是會排除一些如 is, a 之類的功能詞(stop words)。
F2 也是目標詞周圍的詞，但會加上位置的資訊，例如目標詞是 art 時，〝The art of design〞中會被取出的特徵會是{(The, -1), (of, 1), (design, 2)}。
F3 跟 F2 類似，但不同的是 F3 是取出詞性。
F4 則是目標詞與周圍詞的組合，同樣以〝The art of design〞爲例，會被取出的特徵有{The-art, art-of, The-art-of, art-of-design, The-art-of-design}。
F5 與 F4 類似但取詞性組合。

F6 則是利用 Sketch Engine，將可能與目標詞有語法搭配關係的詞列爲特徵。
F7 是利用 Stanford Parser，將 Stanford Parser 所剖析出的與目標詞有依存關係的詞以及依存關係的類別列爲特徵。
F8 是取目標詞前後兩個詞的 HowNet 義元做爲特徵。
F9 是先利用 Stanford Parser 找出與目標詞有依存關係的詞，再取出其 HowNet 義元做爲特徵。

我們使用 Sketch Engine (Kilgarriff et al. (2004))找出跟目標詞具有語法搭配關係的所有詞。圖一是利用 Sketch Engine (http://www.sketchengine.co.uk/)的 word sketch 查詢 duty 個目標詞的輸出，object_of 這一欄表示目標詞可以作爲這些詞的受詞的搭配語，subject_of 表示目標詞可以作爲這些詞的主詞的搭配語, a_modifier 是可以修飾這個目標詞的形容詞，n_modifier 是可以修飾這個目標詞的名詞, modifies 則是可以被這個目標詞修飾的詞。我們選擇的英文語料超過 20 億，如此龐大的語料可以確保得到大部分的搭配語。

圖一 Sketch Engine Word Sketch 的輸出結果



Stanford Parser 是史丹福大學 Klein and Manning (2003)發展出來的多國語言剖析器，只要輸入符合 Pen Treebank 格式的語法樹庫，即可自動從語法樹庫中訓練得到該語言的語法剖析器。下面的例子是 Stanford Parser 的輸出結果。除了標示詞性，語法結構，

最特別的是還將語法的依存關係列出來,例如,nsubj 表示動詞和主詞的關係,dobj 表示動詞和受詞的關係,advmod 表示動詞和副詞修飾語的關係,amod 表示名詞和名詞修飾語的關係。必須強調的是 Stanford Parser 的語法依存關係是從語法樹庫歸納出來後利用 regular expression 抽取出來的,因此即使語法剖析的結果正確,語法依存關係不一定正確。下面是 Stanford Parser 的輸出結果。

The/DT government/NN first/RB established/VBD modern/JJ criminal/JJ investigation/NN system/NN in/IN 1946/CD ./.

```
(ROOT
  (S
    (NP (DT The) (NN government))
    (ADVP (RB first))
    (VP (VBD established)
      (NP
        (NP (JJ modern) (JJ criminal) (NN investigation) (NN system))
        (PP (IN in)
          (NP (CD 1946)))))
    (. .)))
```

```
det(government-2, The-1)
nsubj(established-4, government-2)
advmod(established-4, first-3)
amod(system-8, modern-5)
amod(system-8, criminal-6)
nn(system-8, investigation-7)
dobj(established-4, system-8)
prep(system-8, in-9)
pobj(in-9, 1946-10)
```

知網 Hownet(http://www.keenage.com)是由董振東所發展出來（參考 Dong and Dong (2006)）。Hownet 架構不同於 Wordnet，Wordnet 基本上是一個詞彙網路，同樣語意的詞屬於同一組的 synset，裡面的定義，例句都相同。Wordnet 裡面包含的詞彙語意關係包括上位詞，下位詞等。Hownet 則利用抽象的義元作爲表達所有概念的工具和單位。Hownet 包含的訊息相當的多，是一個中英雙語的知識庫，包括義元，語意角色，上下位關係，部件與整體關係等等語意訊息。義元類似一個語意特徵。Hownet 對於醫生(doctor)的義元表示法爲

{human| 人 :HostOf={Occupation| 職 位 },domain={medical| 醫 },{doctor| 醫

治:agent={~}}}

Hownet 裡面的訊息表示醫生是一個人，具有職位，屬於醫學領域，醫生在醫治這個事件裡扮演主事者的語義角色。在我們的實驗中，我們只使用 Hownet 表示法當中第一個義元，例如：doctor 的第一個義元是 human。對於名詞而言，第一個義元相當於這個詞的語意類別或本體 ontology。

四、實驗結果

在 F1~F6 中，都必須取一個 Window Size，否則會導致特徵和目標詞以及詞義的相關聯性表現不出來，因此這六種特徵都會有 Window Size 的實驗。而我們的實驗是對各種特徵先獨立的來做詞義辨認以得到各種特徵的最佳參數，每種特徵都最佳化以後，最後再使用 Forward Sequential Selection Algorithm 來決定要採用哪些特徵。處理語料以及辨認程式是以 Perl 及 C++寫成。

（一）、F1

F1 是最簡單的直接把目標詞周圍的詞做爲特徵，但是會排除一些如 is, a 之類的 stop words，原因是 stop words 通常對於辨認一個詞的詞義沒有什麼幫助。表一顯示 F1 的最佳 Window Size 爲 3。

表一、F1 Window Size 實驗結果

Window Size	正確率(%)
1	52.7
2	54.2
3	54.6
4	54.6
5	54.1

（二）、F2

F2 也是目標詞周圍的詞，但會加上位置的資訊，會記錄某個詞是出現在目標詞的什麼位置，表二顯示 F2 的最佳 Window Size 爲 1

表二、F2 Window Size 實驗結果

Window Size	正確率(%)
1	54.9
2	53.6
3	51.1
4	47.9

（三）、F3

F2 是目標詞周圍的詞，但會加上位置的資訊， F3 跟 F2 類似，但不同的是 F3 是取出詞性。

表三、F3 Window Size 實驗結果

Window Size	正確率(%)
1	44.6
2	35.5
3	30.7
4	27.7

因此 F3 的最佳 Window Size 爲 1

（四）、F4

F4 則是目標詞與周圍詞的組合，同樣以〝The art of design〞爲例，會被取出的特徵有{The-art, art-of, The-art-of, art-of-design, The-art-of-design}。

表四、F4 Window Size 實驗結果

Window Size	正確率(%)
1	48.2
2	56.9
3	57.8
4	57.8
5	57.8

因此 F4 的最佳 Window Size 爲 3。

（五）、F5

F5 則是目標詞與周圍詞性的組合。

表五、F5 Window Size 實驗結果

Window Size	正確率(%)
1	48.2
2	52.1
3	53.8
4	54.2
5	54.1

因此 F5 的最佳 Window Size 爲 4。

（六）、F6

F6 則是透過 Sketch Engine 將可能與目標詞有依存關係的詞列爲特徵。Sketch Engine 在使用時有數個參數可以做調整，分別是所要包含的依存關係種類、minimum salience，在做 Window Size 實驗時，minimum salience 設爲 0、依存關係種類爲全部，而在做 minimum salience 時 Window Size 設爲 5、依存關係種類爲全部，在做依存關係種類選擇時，minimum salience 設爲 0、window size 設爲 5。

表六、F6 Window Size 實驗結果

Window Size	正確率(%)	Window Size	正確率(%)
1	50.5	11	51.6
2	51.1	12	51.4
3	51.6	13	51.4
4	51.8	14	51.1
5	**52.0**	15	50.8
6	**52.0**		
7	51.8		
8	51.5		
9	51.4		
10	51.5		

因此 F6 的最佳 Window Size 爲 5。

表七、F6 minimum salience 實驗結果

Minimum salience	正確率(%)
0.0	52.0
1.0	51.8
2.0	51.8
3.0	51.3

因此 F6 的最佳 Minimum Salience 約爲 0.0。

對於依存關係的選擇，則是使用 Forward Sequential Selection Algorithm 來選出最好的組合。

表八、F6 依存關係組合選擇(第一步)

Type	正確率(%)	Type	正確率(%)
Object	49.2	and/or	49.4
object_of	47.5	pp*	49.4
Subject	48.4	possessor	46.6
subject_of	48.0	possessed	47.6
a_modifier	48.1	Modifier	48.4
n_modifier	49.0	part*	48.5
Modifies	**50.1**	*comp_of	48.3
		*comp	48.2

在這步中選擇了 modifies

表九、F6 依存關係組合選擇(第二步)

Type	正確率(%)	Type	正確率(%)
Object	**51.1**	and/or	50.8

object_of	50.1	pp*	50.9
Subject	50.2	possessor	50.1
subject_of	50.1	possessed	50.0
a_modifier	50.3	Modifier	50.2
n_modifier	50.7	part*	50.2
*comp	50.1	*comp_of	50.1

在這步中選擇了 object

表十、F6 依存關係組合選擇(第三步)

Type	正確率(%)	Type	正確率(%)
object_of	51.2	and/or	51.5
Subject	51.1	pp*	51.4
subject_of	51.1	possessor	51.1
a_modifier	51.3	possessed	51.0
n_modifier	**51.5**	Modifier	51.2
Comp	51.1	Part	51.1
		comp_of	51.2

在這步中選擇了 n_modifier

表十一、F6 依存關係組合選擇(第四步)

Type	正確率(%)	Type	正確率(%)
object_of	51.6	and/or	51.7
Subject	51.5	pp*	51.6
subject_of	51.5	possessor	51.5
a_modifier	**51.7**	possessed	51.5
comp_of	51.4	Modifier	51.4
Comp	51.4	Part	51.3

在這步中選擇了 a_modifier

表十二、F6 依存關係組合選擇(第五步)

Type	正確率(%)	Type	正確率(%)
object_of	51.6	and/or	**51.9**
Subject	51.7	pp*	51.8
subject_of	51.7	possessor	51.6
comp_of	51.7	possessed	51.7
Comp	51.8	Modifier	51.8
		Part	51.7

在這步中選擇了 and/or

表十三、F6 依存關係組合選擇(第六步)

Type	正確率(%)	Type	正確率(%)

object_of	51.9	pp*	51.9
Subject	51.9	possessor	51.9
subject_of	51.9	possessed	51.9
comp_of	51.9	Modifier	**52.0**
Comp	51.9	Part	51.9

在這步中選擇了 modifier

表十四、F6 依存關係組合選擇(第七步)

Type	正確率(%)	Type	正確率(%)
object_of	52.0	pp*	51.9
Subject	52.0	possessor	51.9
subject_of	51.9	possessed	51.9
		Part	52.0
		comp_of	52.0
		Comp	52.0

在這步中可以看到無論加進哪種依存關係正確率都不再上升了，因此最後所找到的最佳依存關係組合爲{ modifies, object, n_modifier, a_modifier, and/or, modifier}。這個結果顯示主詞的特徵對於詞義辨識而言不是很重要。最重要的是修飾語和受詞。

（七）、F7

F7 是利用 Stanford Parser 所剖析出與目標詞有依存關係的詞以及依存關係的類別（例如：object_of, modifies）列爲特徵。準確率是 54.6%

（八）、F8

F8 是取目標詞前後兩個詞的 HowNet 義元做爲特徵。準確率是 47.2%

（九）、F9

F9 是先利用 Stanford Parser 找出與目標詞有依存關係的詞，再取出其 HowNet 義元做爲特徵。準確率是 54.1%

（十）、特徵選取

採用 Forward Sequential Selection Algorithm 來做特徵選取。

表十五、特徵選取結果

Step	F1	F2	F3	F4	F5	F6	F7	F8	F9
1st	54.6	54.9	44.6	**57.8**	54.2	52.0	54.6	47.2	54.1
2nd	58.9	58.5	56.2		56.8	58.2	**59.6**	56.8	59.2
3rd	**60.7**	60.1	58.2		58.1	60.2		59.1	59.7
4th		**61.2**	60.1		58.8	60.6		60.4	60.7

5th			60.3		59.4	60.8		60.4	61.1

每個 step 會有一欄是粗體, 代表該 step 選取的特徵。例如，第一步選出 F4，接下選出 F7,換言之，第二步的 F1 其實是代表 F4+F1, F2 代表 F4+F2..依此類推。而第二步驟結束後選取的特徵就是 F4+F7。在第三步驟時的 F1 是代表 F4+F7+F1，以此類推。因此最好的特徵組合爲目標詞週圍的三個詞、目標詞週圍一個詞及其位置關係、目標詞與周圍三個詞的連續組合、以及利用 Stanford Parser 所得到與目標詞有依存關係的詞，正確率是 61.2%。

由實驗結果可看出，由於 senseval-2 的岐義目標詞詞性都一樣，採用詞性相關的特徵對於詞義辨認沒有什麼幫助，較有幫助的是目標詞周圍的詞以及與其有依存關係的詞。

五、結論

下表是 Senseval-2 當時的結果，這份結果所使用的訓練及測試語料和我們使用的是相同的：

表十六、Senseval-2 English Lexical Sample Result

準確度	系統	準確度	系統
64.2	JHU (R)	51.2	Baseline Lesk Corpus
63.8	SMUls	50.8	Duluth B
62.9	KUNLP	49.8	UNED - LS-T
61.7	Stanford - CS224N	47.6	Baseline Commonest
61.3	Sinequa-LIA - SCT	43.7	Baseline Grouping Lesk Corpus
59.4	TALP	42.7	Baseline Grouping Commonest
57.1	Duluth 3	41.1	Alicante
56.8	JHU	26.8	Baseline Grouping Lesk
56.8	UMD - SST	24.9	IRST
56.4	BCU - ehu-dlist-all	23.3	BCU - ehu-dlist-best
55.4	Duluth 5	23.0	Baseline Grouping Lesk Def
55.0	Duluth C	22.6	Baseline Lesk
54.2	Duluth 4	18.3	Baseline Grouping Random
53.9	Duluth 2	16.3	Baseline Lesk Def
53.4	Duluth 1	14.1	Baseline Random
52.3	Duluth A		

由表中數據可看出，如果對每個目標詞隨機選一個意義的話準確度是 14.1%、選最常見的意義的話是 47.6%。而我們的結果 61.2%遠大於 baseline 的數據而且也比大部分的系統好，表示這些特徵應用在詞義辨認中是有效的。我們實驗的結果顯示主詞的特徵對於詞義辨識而言不是很重要。最重要的是修飾語和受詞。雖然與 Senseval 2 參賽裡面最好的系統正確率還相差 3%,我們正在實驗其它重要的特徵（如 Wordnet 的 synset, lexicographical file, 及定義）並嘗試用其它機器學習演法如向量支撐機(SVM)或 CRF 希

望進一步提升正確率。

致謝

本研究得到下列國科會計畫經費補助，特此致謝。「詞彙語意關係之自動標注─以中英平行語料庫爲基礎(3/3)」 NSC 93-2411-H-002-013「中英平行句法樹庫的建立與英漢結構對應演算法的研究(I)(II)」NSC94-2411-H-002-043 NSC95-2411-H-002-045-MY2

參考文獻

Brown, Peter et al. (1991) Word sense disambiguation using statistical methods. In ACL 29, pp. 264-270.

Dong, Zhendong and Dong, Qiang. (2006) Hownet and the Computation of Meaning. World Scientific.

Gale, William, Church, Kenneth, and Yarowsky, David. (1992) A method of disambiguating word senses in a large corpus. Computers and the Humanties 26:415-439.

Jurafsky, Daniel, and James H. Martin. (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall.

Klein, Dan. and Manning, Christopher. (2003) Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Le, Cuong Anh and Shimazu, Akira. (2004) High WSD Accuracy Using Naïve Bayesian Classifier with Rich Features. PACLIC 18, Tokyo. http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/564/1/oral-8.pdf

Lesk, Michael. (1986) Automatic Sense Disambiguation: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, pp. 24-26, New York. Association for Computing Machinery.

Lin, Dekang . (1997). Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity In *Proceedings of ACL-97*, Madrid, Spain. July, 1997.

Manning, Christopher, and Schutze, Hinrich. (1999) Foundations of Statistical Natural Language Processing. MIT Press.

Patwardhan, Banerjee, and Pedersen (2005) SenseRelate::TargetWord - A Generalized Framework for Word Sense Disambiguation. Appears in the Proceedings of the Twentieth National Conference on Artificial Intelligence, July 12, 2005, Pittsburgh, PA. (Intelligent Systems Demonstration)

Purandare and Pedersen (2004) Improving Word Sense Discrimination with Gloss Augmented Feature Vectors. Appears in the Proceedings of the Workshop on Lexical

Resources for the Web and Word Sense Disambiguation, November 22, 2004, Puebla Mexico.

Yarowsky, D. (1994) Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French." In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, NM, pp. 88-95.

Hownet http://www.keenage.com/

senseval-2 http://193.133.140.102/senseval2/

Sketch Engine http://www.sketchengine.co.uk/

Stanford Parser http://www-nlp.stanford.edu/downloads/lex-parser.shtml

利用依存關係之辭彙翻譯

Word Translation Disambiguation via Dependency

Meng-Chin Hsiao[1], Kun-Ju Yang[2], and Jason S. Chang[2]
a871002@nthu.us;shanks22@gmail.com;Jason.jschang@gmail.com

[1] Institute of Information and Systems and Applications, National Tsing Hua University
[2] Department of Computer Science, National Tsing Hua University

摘要

本論文提出了一個利用依存關係解決詞彙翻譯的新方法。我們的方法包含了訓練階段及測試階段。在訓練階段，取得與實詞具依存關係的搭配字，並在這些依存關係的條件下，學習分辨翻譯歧義的決策表(decision list)。在測試階段，對於句子中每個實詞檢查跟其有依存關係的搭配字。在測試階段，比對決策表，給予這些字一個正確翻譯。我們實際撰寫了程式，並利用香港新聞及香港立法會議記錄作爲訓練資料。在實驗中我們用了五種不同的方法去處理測試資料並透過一個自動的擬似 BLEU 的評估方法去比較實驗結果。由實驗結果顯示，依存關係的確可以顯著的幫助詞彙翻譯，而實驗也證實某些依存關係是比其他的依存關係更具影響力的。

Abstract

We introduce a new method for automatically disambiguation of word translations by using dependency relationships. In our approach, we learn the relationships between translations and dependency relationships from a parallel corpus. The method consists of a training stage and a runtime stage. During the training stage, the system automatically learns a translation decision list based on source sentences and its dependency relationships. At runtime, for each content word in the given sentence, we give a most appropriate Chinese translation relevant to the context of the given sentence according to the decision list. We also describe the implementation of the proposed method using bilingual Hong Kong news and Hong Kong Hansard corpus. In the experiment, we use five different ways to translate content words in the test data and evaluate the results based an automatic BLEU-like evaluation methodology. Experimental results indicate that dependency relations can obviously help us to disambiguate word translations and some kinds of dependency are more effective than others.

關鍵詞 ： 翻譯選擇，統計式機器翻譯，平行語料庫，決策表，依存關係

Keyword: translation selection, statistical machine translation, parallel corpus, decision list, dependency.

1. Introduction

English is the major language in today's world; for this reason, the latest knowledge and information is mostly written in English. People who want to get new information have to be good at reading English. Although non-native speakers of English can consult a dictionary to understand the meanings of a word, it is still difficult

to find the suitable translation of m-context meanings of the words in the specific sentence. Hence, there are more and more machine translation systems on the web to help people overcome the language barrier. For example, BABEL FISH(http://babelfish.yahoo.com/translate_txt) and Google Translate (http://google.com/translate_t) are two representative machine translation services on the web.

The traditional machine translation systems mostly translate by word or phrase. However, such a word (phrase)-based approach may lead to problems for not considering the structure of the sentence. Consider the word *"motion"* in the given sentence. When the sentence containing it was submitted to BABEL FISH (Figure 1) for translation, the incorrect answer *"行動"* is returned. To improve the kind of limitation seen in BABEL FISH, many researchers consider cross-language phrasal information in statistical machine translation (SMT). At present, some machine translation systems (e.g., Google Translate) have been developed based on the idea to improve performance. However, we submit whole sentence containing "*motion*" and "*passed*" to Google Translate (Figure 2), we still cannot get the suitable translation like "*通過*" of the word "*passed*", especially when "*motion*" and "*passed*" are for apart. Sometimes, words are not translated at all. To obtain the proper translation of the words in a sentence, a promising approach is to consider the syntactic information of the sentence, and to use them for improving the performance of word translation disambiguation (WTD).

Figure 1. Submitting text containing "motion" and "passed" to BABEL FISH for translation in Chinese

We present a new method that automatically determines the translation of given words in the sentence by considering dependency relationships between the words in a sentence. Dependency information includes structure information and dependency can be established between two words that are far apart in the sentence. For example, consider the following sentence "I move that the motion on "*Education on media literacy" as set out on the Agenda be passed.*", "*motion*" and "*passed*" has a dependency of subject-complement. Intuitively, by conditioning probability of translations of "*motion*" and "*passed*" on the dependency pair, nsubjpass (passed-20, motion-5), we can find the correct translation of the words for the context.

The rest of the paper is organized as follows. We review the related work in the next section. Then we present our method in details for automatically training a word translation disambiguation system (Section 3).

Afterward we compare the quality of results between the proposed model and other models (Section 4). Finally, we discuss the results, make conclusion, and close with future work.

Figure 2. Submitting text containing *"motion"* and "*passed*" to Google Translate for translation in Chinese

2. Related Work

Word translation disambiguation has been an important problem in natural language processing. This problem is related to the WSD tasks and is one of the difficult issues in machine translation. In our work, we focus on finding the translations of each content word in the given sentence. The contexts would be English and the target words will be their translations in a second language (e.g., consider the word "*motion*" can be translated as "*行動*" or "*會議*" depending on the sentential content).

Dagan, Itai, and Ulrike (1994) presented an approach for resolving lexical ambiguities in one language using a statistical data on lexical relationship in another language. Yarowsky (1994) showed that decision list (Rivest, 1987) is a good way to model the relation between the words and their translations. We also use the decision list in our approach for estimating translation probability of the word. Yarowsky (1995) exploited two powerful properties that one sense per collocation and one sense per discourse for WSD. He also presented a bootstrapping approach for word sense disambiguation. We also exploit one sense per dependency relationship in our approach.

Pedersen (2000) presented a corpus-based approach to word sense disambiguation that builds an ensemble of Naive Bayesian classifiers, each of which is based on lexical features that represent co-occurring words in varying sized windows of context. Koehn and Knight (2000) present a novel approach to the WTD problem that can be trained using only unrelated monolingual corpora and a lexicon to estimate word translation probabilities using the EM algorithm. Zhou, Ding, and Huang (2001) also proposed an approach to training the translation model by using unrelated monolingual corpora. They parsed a Chinese corpus and an English corpus with

dependency parsers, and two dependency triple databases are generated. Then, the similarity between a Chinese word and an English word can be estimated using the two monolingual dependency triple databases with the help of a simple Chinese-English dictionary. Their translation model overcomes the long distance dependence problem to some extent. Their model can be used translate Chinese collocations into English. In our approach, we only parse the English sentences in a parallel corpus with a dependency parser and try to translate English in to Chinese.

Li and Li (2002, 2004) considered bilingual bootstrapping as an extension of Yarowsky's approach. When the task is word translation disambiguation between two languages, they used the asymmetric relationship between the ambiguous words in the two languages to significantly increase the performance of bootstrapping. They have developed a method for implementing this bootstrapping approach that combines the use of naive Bayes and the EM algorithm. Ng, Wang, and Chan (2003) considered WSD when manually sense-tagged data is not available for supervised learning. They evaluated an approach to automatically acquire sense-tagged training data from English-Chinese parallel corpora. Pham, Ng, and Lee (2005) have investigated the use of unlabeled training data for WSD, in the framework of semi-supervised learning. Empirical results show that unlabeled data can bring significant improvement in WSD accuracy. We used a bilingual corpus but we do not require sense annotation of the data, because we rely on word alignment tool to annotate translation information of the words in the source sentences.

In a study more closely related to our work, Carpuat and Wu (2005) purposed a state-of-the-art Chinese word sense disambiguation model to choose translation candidates for a typical IBM statistical MT system. However, they did not obtain significantly better translation quality than using statistical machine translation system alone. But Cabezas and Resnik (2005) purposed using target language vocabulary directly as "sense," leading to small improvement in translation performance over a state of the art phrase-based statistical MT system. In previous work, human judgment is required for evaluation of sample word tasks of WSD or WTD. In our research, our goal is to study all-word task of WTD and we propose an automatic evaluation methodology.

3. Word Translation Disambiguation Via Dependencies

Finding the appropriate translation of content words in a given sentence is important for machine translation as well as computer assisted language learning. State of the art phrase-based statistical MT systems do a good job if the word providing the "hint" is nearby. Unfortunately, a phrase-based MT system may fail to use of the word that is in a distant from the word we want to translate. To translate the words of the sentence, a promising alternative approach is to find the likely translation of each word through statistical analysis of its dependencies.

3.1 Problem Statement

We focus on a subtask of MT system; that is we focus on finding the appropriate translation of content words via dependencies. These dependencies provide recursive syntax structure information of the words in the sentence. We collect these dependencies and the relevant translations in a parallel corpus and find out the relationship between them. The goal is to find the proper translation of content words in the given sentence. Formal statement of the problem is as follows.

Problem Statement: We are given an English sentence *S* (e.g., "*A very big apple on the table was eaten by him.*") that we want to translate. Our goal is to give each content word, $w_1, w_2, \ldots, w_m$, in *S* a most appropriate Chinese translation relevant to the context of *S*. For this, we derive dependencies (e.g., advmod (big-3, very-2), amod(apple-4, big-3), nsubjpass(eaten-9, apple-4), etc.), $d_1, \ldots, d_p$, in *S*, then use the dependencies of the word *w* (i.e., dependency relationship (*w*, *w'*) or dependency relationship(*w'*, *w*)) to find the most appropriate translation for *w*.

In the rest of this section, we describe our solution to this problem. First, we define a dependency-based translation model for word translation disambiguation (Section 3.2). This training strategy relies on a set of dependency relationships derived from a dependency relationships collection. In this section, we also describe the other two strategies that we use when no dependency information is available. Finally, we show how our method handles a given sentence at run time by using a decision list (Section 3.3).

3.2 Training the Dependency-Based Translation Model

We take advantage of a word-aligned parallel corpus as training data to establish a decision list for word translations based on dependency relationships. For each word in a sentence, we obtain the translation and dependency relationships using word alignment tool (e.g., Giza++) and a general purpose parser (e.g., Stanford parser). With that information, we compute the word translation probability for all dependency relationships based on logarithmic likelihood ratio (LogL):

$$LogL = Log(\frac{P(t \mid w_t, d, w_d)}{P(\bar{t} \mid w_t, d, w_d)})$$

$$= Log(\frac{\frac{count(t, w_t, d, w_d))}{count(w_t, d, w_d)}}{\frac{count(\bar{t}, w_t, d, w_d)}{count(w_t, d, w_d)}}) = Log(\frac{count(t, w_t, d, w_d)}{count(\bar{t}, w_t, d, w_d)})$$

(1) Parse the source language using a dependency parser (Section 3.2.1)
(2) Use an alignment tool to align words in a parallel corpus (Section 3.2.2)
(3) Compute the decision list for translation and dependency (Section 3.2.3)
(4) Compute the probability of a translation for each word **(**Section 3.2.4)

Figure 3. Outline of the process used to train in our method

3.2.1 Parse the source language using a dependency parser

In the first stage of the training process (Step (1) in Figure 3), we use the English part of an English-Chinese parallel corpus as the input data. First, we utilize a tagger to tokenize the sentences, give each word in the source sentences a part of speech (POS) tag, and obtain dependency relationships from the source sentences via a dependency parser. We use an English sentence as an example to show the process. (Figure 4).

3.2.2 Use an alignment tool to align words in a parallel corpus

In the second stage of the training process (Step (2) in Figure 3), we use a word alignment tool to align words in a parallel corpus. First, we lemmatize the tokens obtained from the first stage. Words that are tagged proper

noun are not lemmatized. Then, target language sentences are segmented using a word segmentation tool. Finally, each pair of source and target sentence is word-aligned using an existing word alignment model to produce word alignment information. Figure 5 shows an example of the process.

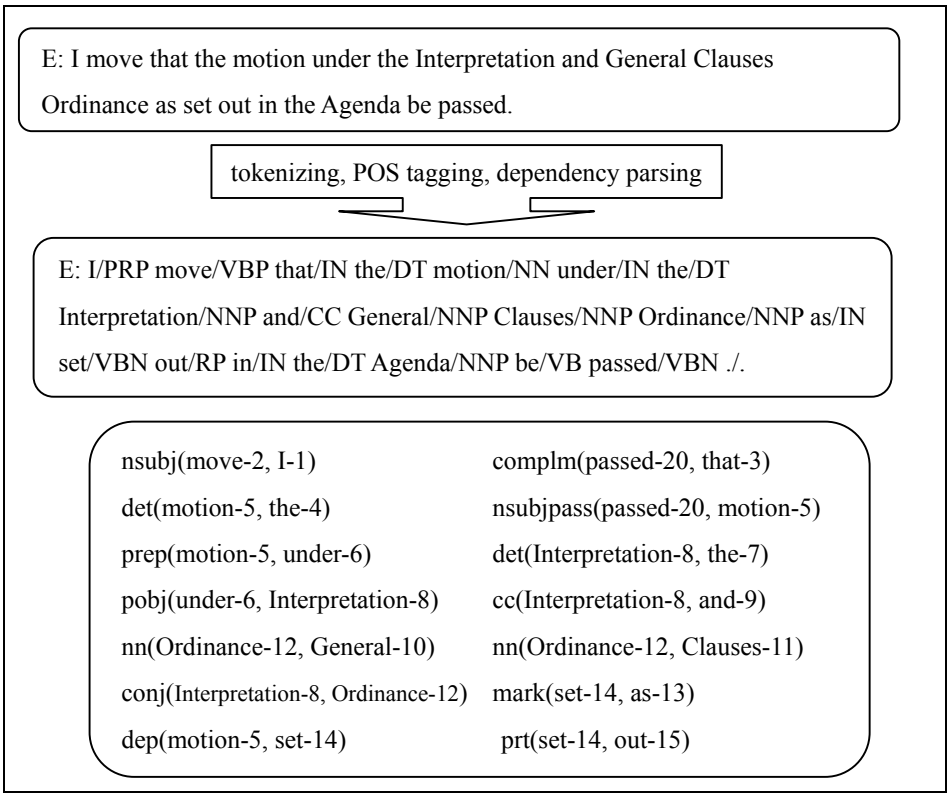


Figure 4. An example to show the result of the tagger and dependency parser

3.2.3 Compute the decision list for each translation and dependency

After deriving the translations and dependencies, we are in a position to train a classifier for WTD. We use the dependency relationship to condition the translation probability, and then we compute a score for each translation conditioned on one of the relevant dependency relationships. There are many different approaches to do this for various pattern recognition problems. We choose the decision list for simplicity and efficiency considerations. The algorithm we used is similar to the approach proposed in Yarowsky (1994) for WSD. For each possible translation of a given word, we compute the logarithmic likelihood ratio (LogL)

$$LogL = Log(\frac{count(t, w_t, d, w_d)}{count(\bar{t}, w_t, d, w_d)})$$

where t is the translation of the word w_t with dependency d with another word w_d and *count* (t, w_t, d, w_d) is the number of instance of word w_t aligned with the translation t under of dependency relationship d (w_t, w_d) d, and count ($\bar{t}$, w_t, d, w_d) is the number of instance of word w_t aligned with the other translations $\bar{t}$ under the same relationship.[1]

Sample output is shown in Table 1. The LogL in Table 1 are computed by count (t, w_t, d, w_d) and count ($\bar{t}$,

[1] Here d (w_t, w_d) and d (w_d, w_t) are treated as different dependency relationships.

w_t, d, w_d). In the experiment described in Chapter4, we smooth count (t, w_t, d, w_d) and count ($\bar{t}$, w_t, d, w_d) by held out data.

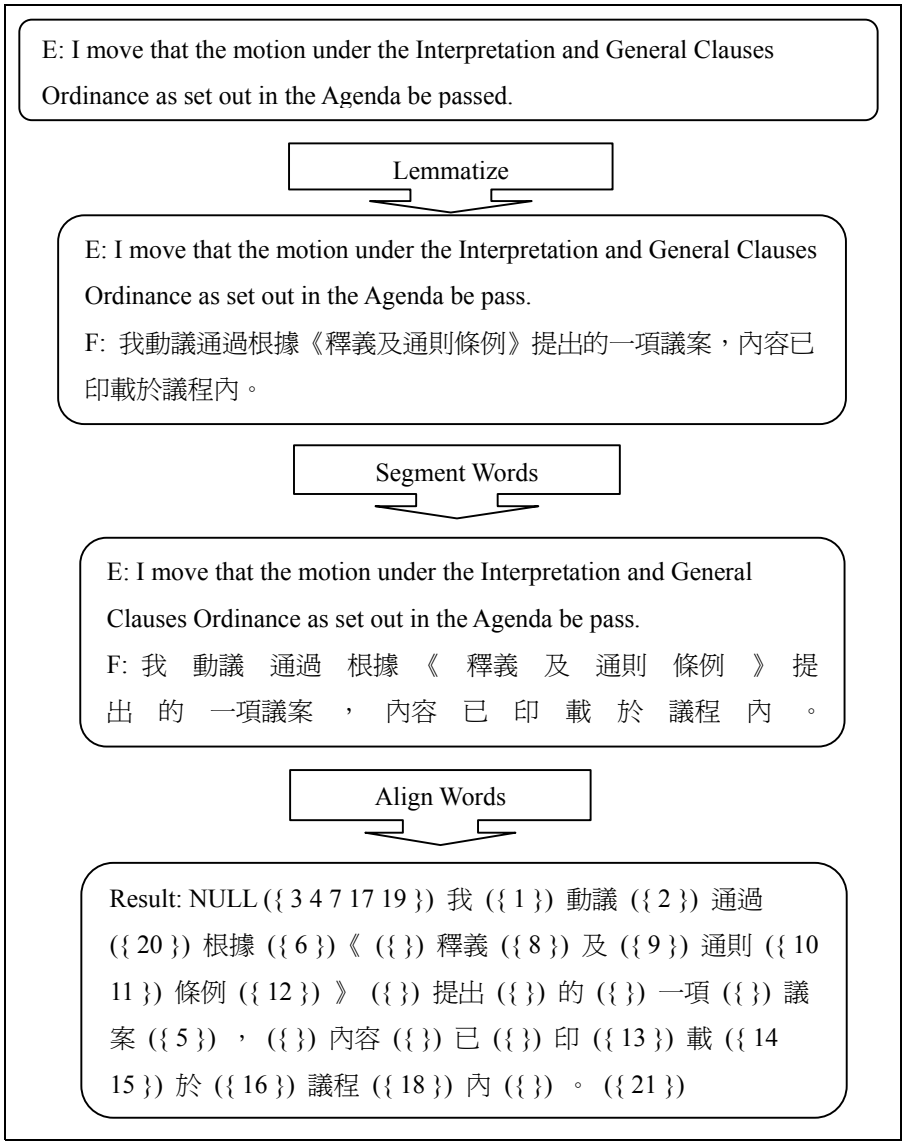


Figure 5. An example to show the data handling after word alignment

Table 1. Calculating LOGL with N = count (t, w_t, d, w_d), N'= count ($\bar{t}$, w_t, d, w_d), $\bar{t} \neq t$

dep	w_d	w_t	t	LogL	N	N'
nsubjpass	pass	motion	議案	1.608	294	58.8798
nsubjpass	pass	motion	動議	-1.975	43	309.8798
nsubjpass	pass	motion	…	…	…	…

3.2.4 Parse the source language using a dependency parser

In the last stage of the training process (Step (4) in Figure 3), we compute two types of word translation probability: P (t | w) and P (t | w, p). For unseen dependency relationships, we use P (t | w, p) to predict translation and for unseen word/POS combination, we use P (t | w) to predict translation for all POS's.

The word translation probability is calculated based on sentences where the source and target words are aligned.

We compute the word translation probability for all word aligned with a translation by the ratio of two counts:

$$P(t \mid w) = \frac{count(w,t)}{count(w)}$$

where *count* (*w*, *t*) is the number of instance of word *w* aligned with some translation, and *count* (*w*) is the number of w instances.

Table 2. An example to calculate P (t | w) for "plant"

w	t	Count
plant	核電廠	342
plant	植物	258
plant	種植	167
plant	…	…
P(核電廠|plant)=342/2172, P(植物|plant)=258/2172		

Table 3 Examples of P (t | w, p) for the noun "plant"

w	t	Count
plant	核電廠	341
plant	植物	245
plant	發電廠	132
plant	…	…
P(核電廠|plant, Noun)=341/1852		

We can then condition the translation probability using the POS information obtained from the first stage. Table 3 shows an example for the word "*plant*" that is tagged noun.

$$P(t \mid w, p) = \frac{count(w,t,p)}{count(w,p)}$$

where *count* (*w*, *t*, *p*) is the number of instances of word w with the POS *p* aligned with some translation, and *count* (*w*, *p*) is the number of *w* with the POS *p* instances.

3.3 Word Translation Disambiguation at Runtime

After the decision list and context-independent translation probabilities are obtained in the training process, we can then use them to disambiguate translations for the words in a sentence containing the words. The process of word translation disambiguation at runtime is shown in Figure 6.

Step 1: Parse the input sentence by a dependency parser
Step 2: Select the highest score translation by using dependencies
Step 3: Use P (*t* | *w, p*) to predict translation for unseen dependency relationship
Step 4: Use P (*t* | *w*) to predict translation for unseen word/pos combination

Figure 6. Outline of the process at run time

In Step 1 we exploit a parser to obtain word tokens, POS tags, and dependency relationships of given sentence, and then we lemmatize all tokens except for words that are tagged as "NNP". In Step 2 we determine the translation of words by using the most reliable piece of evidence. Figure 6 shows the process at runtime by using the sentence *"In accordance with the Rules of Procedure, the motion and the amendment will be debated together in a joint debate."* as an example. In some situation, there is no dependency relationships information available to help us translate the word. In Step 3 we use POS information to find the translation of the word for unseen dependency relationship. In Step 4 we take the highest frequency translation to be the translation of the

word. If the word is not in our training data, we cannot translate the word.

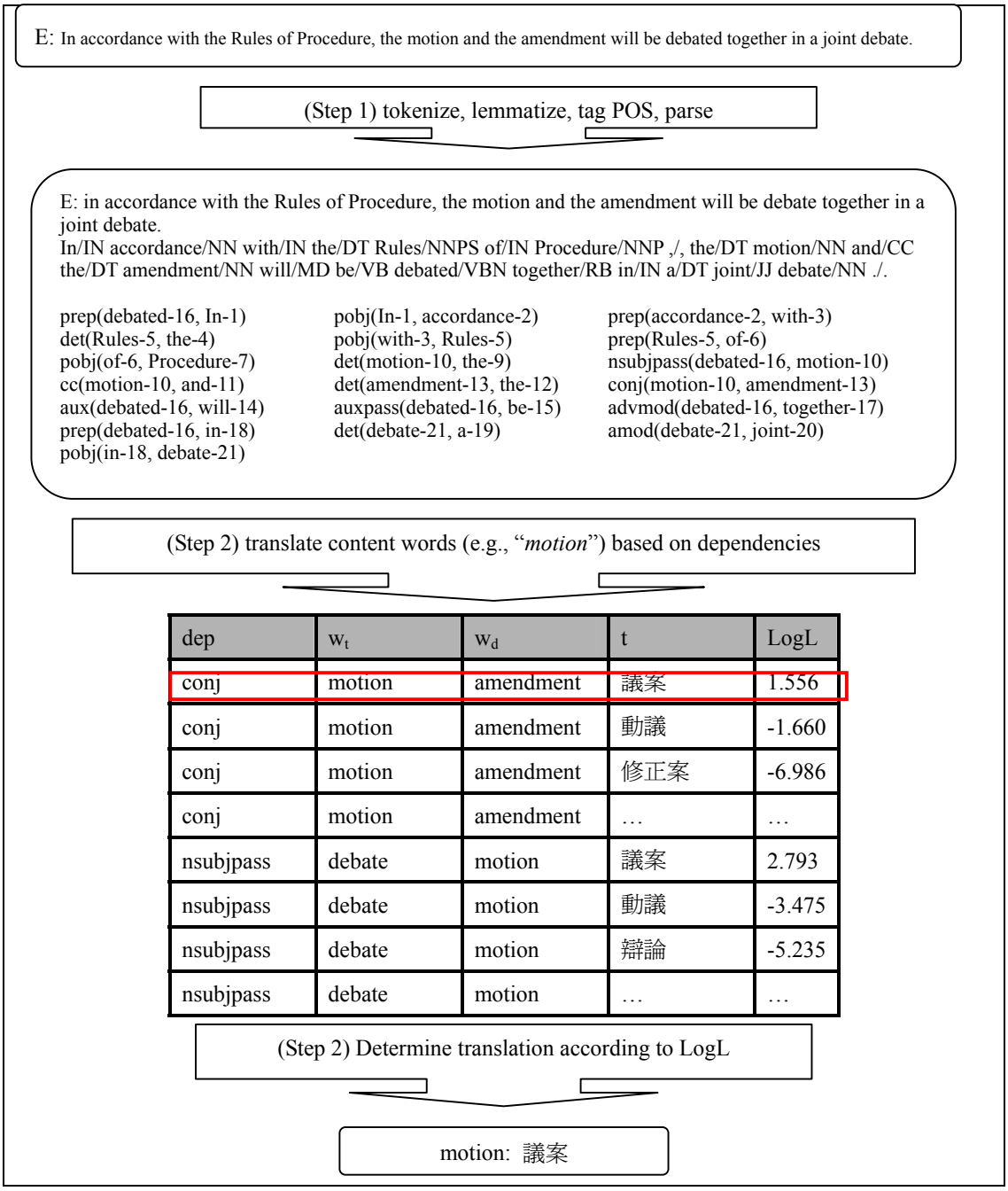


Figure 7 An example of Step 1 and Step 2

4 Experiments and Evaluation

This approach was designed to disambiguate the translation of the words in the given sentence, by using the statistical properties of the dependency relationships with word to translation. In this section, we first describe the details of the experiments for the evaluation (Section 4.1). Then, we introduce the test data and automatic evaluation methodology and results. (Section 4.2)

4.1 Experimental Setting

In this section we will describe the implementation and experiments of the method described in section 3. For training the proposed model, we used a collection of approximately 740,000 sentence pairs from Hong Kong News English-Chinese Corpus (HKNC1997~2003) and approximately 1,375,000 sentence pairs obtained from Hong Kong Hansard English-Chinese Corpus (HKLC1985~2003).

First, to preprocess the training data, we used Stanford Parser (Version 1.5.1) to implement tokenizing, POS tagging, and dependency parsing. We filtered out the English sentences with word length longer than forty or have some unusual letters. After filtering, we were left with approximately 630,000 sentence pairs from HKNC and approximately 1060,000 sentence pairs from HKLC. Then we use an in-house word lemmatization tool to lemmatize each word in the English sentences. We also segment each Chinese sentence using a word segmentation tool developed by CKIP in Academia Sinica. Finally, we use the Giza++v2 toolkit made available at (www.fjoch.com/GIZA++.html) to obtain word alignment information for the training data. In our experiment, we only use the direction of Chinese to English for word alignment information part. After filtering some errors that occurred in the word alignment process, we were left with about 556,000 sentence pairs from HKNC and about 983,000 sentence pairs from HKLC for training, and we reserved 3,500 sentence pairs obtained from HKNC and 9,500 sentence pairs obtained from HKLC for testing.

Second, we grouped POS's used in the Stanford Parser into nine groups. Table 4 shows the grouping of parts of speech. The grouping was done to reduce sparseness.

Table 4. The nine POS groups

Pos Group	Original tags	Notes
Light Verb	have, do, know, think, get, go, say, see, come, make, take, look, give, find, use	15 high-frequency verbs (Svartvil and Ekedahl 1995)
V	VB, VBD, VBG, VBN, VBP, VBZ, ask	verb but not light verb
N	FW, NN, NNS, PDT	noun
NNP	NNP, NNPS	proper noun
C	CD	quantifier
$	$	$, no., rule, section, …
J	JJ, JJR, JJS, a	adjective
R	RB, RBR, RBS, RP	adverb
F	the other tag	function word

Third, after calculating count (t, w_t, d, w_d) and count ($\bar{t}$, w_t, d, w_d) described in section 3.2.3, we smoothed the counts for the unseen translations for w_t and w_d that have dependency relationship d using held out estimator that is purposed by Jelinek and Mercer(1985). We split training data into two parts that have equal number of sentence pairs. One was used as training data and the other was used as held out data, and then we changed the role of two parts and did held out estimation again. Table 5 shows the final modified number N. Table 6 shows the results of smoothing. Every count ($\bar{t}$, w_t, d, w_d) had to add the counts for unseen translations.

Table 5. The average of two held out estimators

Count C	Obs. counts Set 1	Obs. counts Set 2	Smoothing counts
0	0.87995	0.87965	0.87980
1	0.25552	0.25562	0.25557
2	1.08762	1.08368	1.08565
…	…	…	…
8	7.14678	7.13740	7.14209
9	8.20037	8.15285	8.17661

Table 6. An example of smoothing

dep	wd	wt	t	LogL	N	smoothing N	N'	smoothing N'
nsubjpass	pass	motion	議案	1.608	294	294	58	58.8798
nsubjpass	pass	motion	動議	-1.975	43	43	309	309.8798
nsubjpass	pass	motion	議題	-5.100	3	2.1331	349	349.8798
nsubjpass	pass	motion	決議案	-5.778	2	1.08565	350	350.8798
nsubjpass	pass	motion	表決	-7.228	1	0.25557	351	351. 8798
nsubjpass	pass	motion	…	…	…	…	…	…

Table 7. The properties of test data

property	HKNC	HKLC
sentences	1,500	3,800
all words	31,569	84,290
content words	16,980 (53.79%)	42,343 (50.23%)
LV	585 (1.85%)	2,142 (2.54%)
be	858 (2.72%)	2,798 (3.32%)
F(Function words)	13,146 (41.64%)	37,007 (43.90%)

4.2 Evaluation and Discussion

In this section, we describe our test data and evaluation methodology (4.2.1). We then show the evaluation result of our experiment and give some discussions (4.2.2).

4.2.1 Test Data and Evaluation Methodology

We randomly choose 1,500 sentences out of 3,500 sentence pairs from HKNC and 3,800 sentences out of 9,500 sentence pairs from HKLC for testing. Then we translate the content words in the given sentences of test data. We did not consider the translation of the words that POS tagged in the group F and LV, also not did we consider the translation of the verb "*be*". Table 7 shows the properties of our test data.

The traditional WSD evaluation methodology relies on human judgment. In our experiment, we do not focus on the sense of the words, but rather the translation of the content words in the given sentences. Since it is

infeasible for human to evaluate such a large set of data, we developed a BLEU-like automatic evaluation methodology. We evaluate one sentence at a time. First, we combine all translations of content words that in the given sentence. Identical an overlapping translations of two neighboring word are combined and redundancy are removed. For example, see Figure 8 for more details.

EPD/環保署 and Transport/運輸署 Department/署 will also/亦 step/加強 up/加強 their programmes/計劃 to educate/教育 vehicle/車主 owners/車主 and mechanics/技師 to exercise/行使 their responsibility/責任 to maintain/維修 vehicles/車輛 properly/妥善 .

Combining translations

環保署 → 環保署運輸署 →環保署運輸署 → 環保署運輸署亦 → 環保署運輸署亦加強 → 環保署運輸署亦加強 → 環保署運輸署亦加強計劃 → 環保署運輸署亦加強計劃教育 → 環保署運輸署亦加強計劃教育車主 → 環保署運輸署亦加強計劃教育車主 → 環保署運輸署亦加強計劃教育車主技師 → … → 環保署運輸署亦加強計劃教育車主技師行使責任維修車輛妥善

Figure 8. An example to show how to combine the translations of the sentence

Second, we calculated unigram precision rate based on the aligned Chinese sentence as the reference translation. Figure 9 shows an example of the process. Third, we filtered the highest ten percentage and lowest ten percentage sentences for data balance, and then we average the score of middle eighty percentage sentence to be the result.

Result after combine:環保署運輸署亦加強計劃教育車主技師行使責任維修車輛妥善

Answer: 環境保護署和運輸署亦將加強宣傳，要求車主和技師妥善地保養和維修車輛。

$$\text{Score} = \frac{\text{the number of the Chinese characters matching the reference}}{\text{number of the Chinese characters produced}} = \frac{19}{27}$$

Figure 9. An example to show how to calculate the score of the sentence

4.2.2 The Results of WTD of Different Methods

We used five different methods to disambiguate word translations. Table 8 shows the results of WTD. Baseline is the result of using only P ($t \mid w$) to estimate the translation of the content words, while baseline with POS is the result of using both P ($t \mid w, p$) and P ($t \mid w$), dependency method (all) is the result of using the process described in 3.3, window size 1 is the result of using a window based co-occurrence (the word to the right or left of the word in question) instead of dependency, and dependency method (some) is the result of using the process described in 3.3 leaving out five kinds of dependency relationships, including *determiner, negative, possessive, coordinating conjunction, preposition*.

Table 8. Results of WTD in different methods

Method	HKNC	HKLC	HKNC+HKLC
Baseline	0.582	0.564	0.569
baseline + POS	0.589	0.569	0.575
window size 1	0.698	0.643	0.659
dependency method (all)	0.714	0.686	0.694
dependency method (some)	0.716	0.685	0.694

The results in Figure 10 indicate that the dependency method obviously outperforms baseline with POS and also outperforms window based co-occurrence approach. We also found that using POS can only improve slightly and ignoring some kinds of dependency relationships does not affect the results too much.

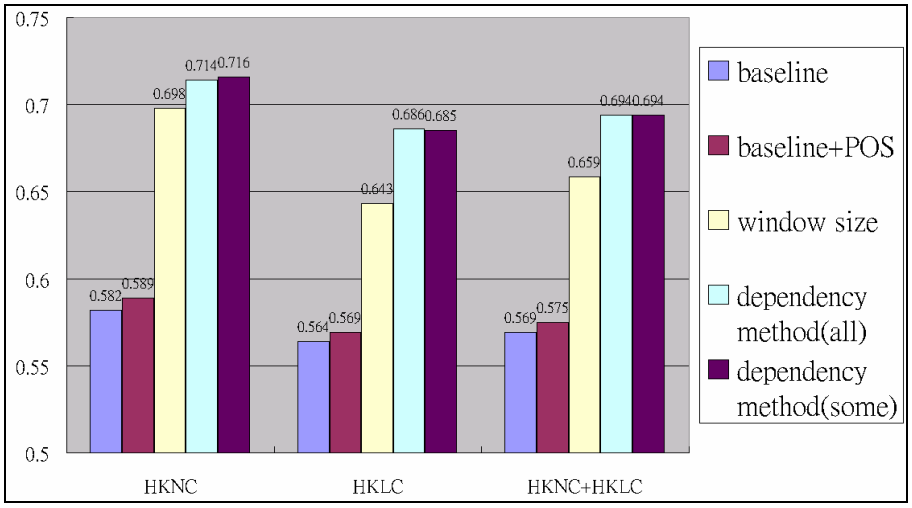


Figure 10. Results of WTD in different methods

However, we did not evaluate the performance of translations based on dependency. As shown in Table 9 that over ninety percent of the cases, words are translated via dependency.

Table 9. The percentage of the word translations when we used dependency method (some)

type	HKNC	HKLC
all content words	16,980	42,343
dependency method	15,698 (92.45%)	38,587 (91.13%)
baseline +POS	1,199 (7.06%)	3,610 (8.53%)
baseline	12 (0.67%)	34 (0.08%)
no answer	71 (0.42%)	112 (0.26%)

As shown in Figure 11, in different number of sentences, the results of HKNC are better than the results of HKLC. We believe this is a result of the different character of the corpora and not the different number of sentences in the two corpora.

Because of data sparseness, we may not calculate a suitable score for translations of the words with dependency relationships. If we use lager training set, we may improve the performance. Some word translation errors may be caused by word alignment errors. In addition, there also have some problems caused by incorrect segmentation. For example, "*吸煙者*" is segmented into "*吸煙*" and "*者*", but in our

module we only consider the one to one case, therefore the word "*smoker*" will be translate to "*吸煙*" and not "*吸煙者*".

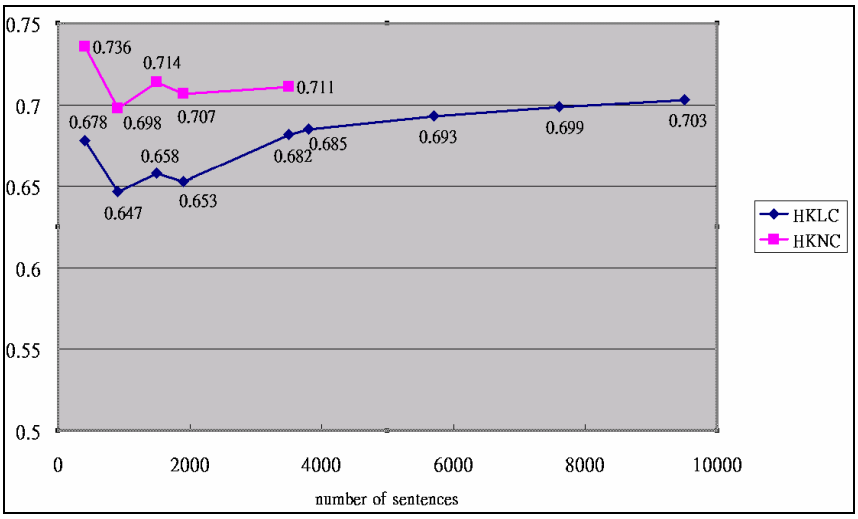


Figure 11. Results of WTD in different corpora

Table 10 an example to explain majority voting methodology

Dep	wt	wd	t	LogL
advmod	plant	at	植物	2.502
Advmod	plant	at	廠	-3.131
Dep	wt	wd	t	LogL
Amod	plant	chemical	處理廠	1.290
Amod	plant	chemical	一個堆	-3.568
Amod	plant	chemical	廠	-3.568

5 Future Work and Conclusion

In summary, we have introduced a method for word translation disambiguation, which improves the ability to disambiguate the translations of the content words in the given sentence using a dependency-based translation model trained as a parallel corpus. We have implemented and evaluated the method using a bilingual English-Chinese corpus. We have shown that the method outperforms the baseline. In addition, we also found some kinds of dependency are more effective than others. Moreover, we have purposed an automatic BLEU-like evaluation methodology for WTD. The results of word translation disambiguation can assist user in reading English, and also can be used as additional input information for an MT system to improve the performance.

Many future directions present themselves. First, it would be interesting to extend the method to translate all words in the sentence including function words. Second, we can give different weight to different type of dependency since we believe different type of dependency relationships have different level of effectiveness. Third, we are currently using the dependency relationships with the highest score, but we can also consider all dependency relationships of the word in the given sentence. Table 10 shows an example that "*plant*" that has two dependency relationships with "*at*" and "*chemical*". In the way we described in our approach, we will choose "*植物*" as the answer. If we combine scores of two dependency relationships to calculate a new score for

each translation, we may choose "*敵*" as our answer which seems to be more suitable.

References

1. Bengt Altenberg and Sylviane Grange. "The grammatical and lexical patterning of make in native and non-native student writing". Applied Linguisics, 22(2), 173-194, 2001.
2. Clara Cabezas and Philip Resnik. "Using WSD Techniques for Lexical Selection in Statistical Machine Translation". http://handle.dtic.mil/100.2/ADA453538 , July 2005.
3. Marine Carpuat and Dekai Wu. "Word Sense Disambiguation vs. Statistical Machine Translation". In 43th Annual Meeting of the Association for Computation Linguistics (ACL 2005), 2005.
4. Dagan, Ido, Alon Itai, and Ulrike Schwall. "Two Languages are More Informative than One". In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL91). Berkeley, 1991.
5. Philipp Koehn, and Kevin Knight. "Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm". In Proceedings of the 17th National Conference on Artificial Intelligence, pages 711–715, Austin, TX, 2000.
6. Cong Li and Hang Li. "Word translation disam-biguation using bilingual bootstrapping". In Proceed-ings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 343-351, 2002.
7. Yajuan Lü, Ming Zhou, Sheng Li, Changning Huang, Tiejun Zhao (2001b). "Automatic translation template acquisition based on bilingual structure alignment". International Journal of Computational Linguistics and Chinese Language Processing. 6(1), pp. 1-26, 2001
8. Hwee Tou Ng, BinWang, and Yee Seng Chan. "Exploiting parallel texts for word sense disambiguation: An empirical study". In Proceedings of ACL-03, Sapporo, Japan, pages 455–462, 2003.
9. K. Papineni, S. Roukos, T. Ward, and W. Zhu. "Bleu: a method for automatic evaluation of machine translation". In Proceedings of 40th Annual Meeting of the ACL, Philadelphia, 2002.
10. Ted Pedersen. "A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation". In Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, 2000.
11. Thanh Phong Pham, Hwee Tou Ng, and Wee Sun Lee. "Word sense disambiguation with semi-supervised learning" AAAI-05, The Twentieth National Conference on Artificial Intelligence, 2005.
12. Dan Klein and Christopher D. Manning. "Fast exact inference with a factored model for natural language parsing". In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, Advances in Neural Information Processing Systems 15, Cambridge, MA.MIT Press, 2003.
13. D. Yarowsky. "Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French". In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, 1994.
14. D. Yarowsky. "Unsupervised word sense disambiguation rivaling supervised methods". In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pages 189–196, 1995.
15. Ming Zhou, Yuan Ding, and Changning Huang. "Improving translation selection with a new translation model trained by independent monolingual corpora". Computational linguistics and Chinese Language Processing. Vol. 6, No. 1, pp 1-26, 2001.

Knowledge Representation for Interrogatives in E-HowNet

Shu-Ling Huang, You-Shan Chung, Yueh-Yin Shih, Keh-Jiann Chen

CKIP, Institute of Information Science, Academia Sinica

{josieh, yschung, yuehyin, kchen}@iis.sinica.edu.tw

Abstract

In order to train machines to 'understand' natural language, we proposed a universal concept representational mechanism called E-HowNet to encode lexical semantics. In this paper, we take interrogative constructions as examples, i.e. concepts or sentences for asking questions or making inquiries, to demonstrate the mechanisms of semantic representation and composition under the framework of E-HowNet. We classify the interrogative words into five types according to their semantic distinctions, and represent each type with fine-grained features and operators. The process of semantic composition and the difficulties of the representation, such as word sense disambiguation, will be addressed. Finally, we'll show how machine discriminates two synonymous sentences with different syntactic structures and surface strings to prove that machine understanding is achievable.

Keywords: semantic representation, interrogatives, E-HowNet

1. Introduction

To understand natural language by machines, lexical semantic representation and composition are the most important techniques. In this paper, we will take the interrogatives as examples to demonstrate the mechanism of lexical semantic representation and composition in E-HowNet (Chen et.al., 2004). E-HowNet uses the word sense definition mechanism of HowNet (Dong, 1988) and the vocabulary of WordNet (Fellbaum,1998) synsets to describe concepts.[1] Its goal is to achieve near canonical semantic representation,

[1] The advantage of using WordNet synsets is that each synset has unique sense and sense similarity between two synsets can be measured through WordNet ontology. But there is disadvantage of WordNet-like ontologies, for example, each concept class in WordNet has limited linking to other concepts. The major links are hyponymy relations which limit inheritance and inference capability to the classes on the taxonomy. So we adopt similar

that is, two sentences with different surface forms or in different languages may achieve similar E-HowNet representations. Take sentences (1) and (2) as examples:

(1) 我能否拍照？Is it OK for me to take pictures?

(2) 我可不可以照相？Can I take photos?

Although the syntactic structure and surface strings of (1), (2) are very different, by using lexical sense definitions in E-HowNet, we hope machine can 'understand' that they are synonymous sentences.

Analysis of interrogative constructions is of great interest to linguists, as well as to computer scientists, for example, those who engaged in QA techniques. Interrogative constructions have played a central role in the development of modern syntactic theory. Ginzburg & A.Sag (2000) have pointed that interrogative has been at the heart of work in generative grammar, also in government and binding (GB) and head-driven phrase structure (HPSG). Nonetheless, to date most syntactic work has taken place quite separately from semantic and pragmatic work on interrogatives. Taking questions in Mandarin Chinese as example, Shao (1996) has summed up the current study of interrogatives and listed the main research themes as follows: the types of question, interrogative particles, querying focus and its answer, degree of doubt and special interrogative sentences pattern etc.. Most of the above themes are purely grammatical analysis. To build a frame-based entity-relation knowledge representation model, we find interrogative construction a good and challenging example, for it is feature-structured, free formed, and demanding for story comprehension. In other words, it combines problems of syntax, semantics and pragmatics. Our approach is to find a framework to represent interrogatives, therefore, semantic distinction of interrogatives is our focus.

In E-HowNet, we made distinctions between content sense and relational sense and

---

mechanism in HowNet to define word sense, which represents concepts in more accurate way by not restricting the definition vocabulary to a closed set of primitives only, i.e., any well-defined concepts can be used to define a new concept. The detail discussions can be seen at (Chen et al., 2005).

represent senses of content words and senses of function words in different ways. For instance, we represent phrase ‘bathe with cold water’ as (3)

(3) bathe with cold water

bathe def: {clean|使淨: patient={body|身體}}
with def: instrument={}
cold water def: {water|水: temperature={cold|冷}}}.
composition: {clean|使淨: patient={body|身體},
instrument={water|水:
temperature={cold|冷}}}.

In this case, content words ‘bathe’ and ‘cold water’ are represented differently from function word ‘with’, for the latter plays the role of linking concepts. In much the same way, interrogative words have more relational sense than content sense, so they are defined by semantic role to denote relational sense and the operator ‘.Ques.’ to mark the querying focus that is the object or its discrimination features which speakers want to know.

In the following section, we’ll briefly describe the previous work for interrogatives. Then, we introduce our analysis of type classifications for interrogatives and their representation in E-HowNet. Next, we present the composition of interrogative sentences and the difficulties encountered. We conclude the paper by discussing our results and future works.

2. Background

Questions in Chinese studies traditionally attributed to mood category of syntactics.[2] Most of linguists consider there are four grammatical devices that explicitly mark an utterance as an interrogative. [3]First is question which can be answered by ‘yes’ or ‘no’, called factual question, true and false interrogative or yes/no interrogative. Second is question which includes Wh-words such as ‘who’, ‘what’ or ‘when’ and so forth, called Wh-word

[2] Ma (1989）wrote the first grammar book for Mandarin Chinese. He classified interrogative words into mood category. Later, Li (1930) and Lv (1942) have carried forward his viewpoint and influenced on modern research of questions deeply.
[3] Such as Lv (1942), Li & Thompson (1997), Tang (1983), Lu (1984), Shao (1996),etc..

interrogative, or information seeking interrogative. Third is question which mentions two or more possible alternative answers, called disjunctive interrogative or either/or interrogative. Fourth is question which is composed of a statement followed by an *A not A* form, such as *dui bu dui* 'right not right', *xing bu xing* 'Ok not Ok' etc., called *A not A* interrogative or tag interrogative.[4] From different analytical perspective, these four question devices may have different hierarchy. For example, Lv (1942) framed them as (4) while Lu (1984) structured them as (5):

(4) interrogative – Wh-word interrogative
– true/false interrogative – *A not A* interrogative
– disjunctive interrogative

(5) interrogative – true/false interrogative
– non true/false interrogative – Wh-word interrogative
– disjunctive interrogative
– *A not A* interrogative

Generally speaking, true and false interrogative and Wh-word interrogative are regarded as basic types.

3. Semantic Representation

3.1 Our classification of interrogatives

As we focus on knowledge representation, we are more concerned about semantic discriminations for different interrogative sentences. Therefore, we take a sense-based approach to create a hierarchical classification which is guided by a layered semantic hierarchy of answer types, and eventually classifies interrogative sentences into fine-grained classes, shown as (6):

[4] Tag interrogative is formed by adding a short *A* not *A* question form of certain verbs as a tag to a statement. In this paper, we regard it as a general *A* not *A* question device due to the same semantic performance.

(6) interrogative –(A) true/false interrogative

– (B) Wh-interrogative –(a) asking factual information

– (b) asking relationship

– (c) asking opinion

– (d) option choosing

According to different querying focus, we separate (A) true/false interrogative from (B) Wh-interrogative. Take sentences (7), (8) as examples:

(7) 你喜不喜歡這個遊戲？

Do you like this game?

(8) 有誰知道我可以在哪裡找到這個遊戲？

Who knows where I can find this game?

Sentence (7) belongs to the former question device because the entire statement is a querying focus. Dissimilarly, sentence (8) indicates two querying focuses by using different interrogatives 'who' and 'where'. In other words, the true/false interrogative asks truth value of the whole sentence. And the Wh-interrogative is used to ask information. By analyzing the querying focus, the latter can be further divided into four types: (B-a) asking factual information, such as time, location, quantity and so forth; (B-b) asking relationship, such as kinship; (B-c) asking opinion or attitude, such as possibility, capacity, volition etc.; and the last, (B-d) asking to choose an option. Sentence (8) refers to the type (B-a). For each of the remaining types, we give an example as follows:

(9) 她是你的什麼人？

What is the relationship between you and her?

(10) 他可不可以吃辣椒？

Can he eat hot peppers?

(11) 淘米水是酸性還是鹼性？

Rice washing water is acidic or alkaline?

Here, two distinctions have to be made. First, sentence (9) refers to type (B-b), but why need we separate it from type (B-a) when they both use 'what' to make questions? In sentence (9), the question word '什麼 *what*' ask for relationship but not the type of a frame element or the value of a semantic role.[5] Chen et. al. (2004) proposed a complex relation description, i.e. a representation denotes the relation between head and semantic role specifically. In general, E-HowNet semantic representation model presumes the relations variables is the head, for example, 'white cloud' will be define as (12):

(12) white cloud
def:{cloud|雲: color(~)={white|白}}

In the definition, '~' indicates the head 'cloud', and normally be omitted in the expression. Conversely, word indicates complex relation always has another relation variable apart from the head, so the variable needs to be marked clearly. For example, we express 'mother in law' as (13):

(13) mother in law
def:{human|人=mother(spouse(x:human|人))}

According to the representation model, when our querying focus is complex relation, we put question mark before the relation role, such as mother, spouse, parents etc. to make the interrogative definition. It makes the difference between interrogative type (B-a) and (B-b). See more examples in section 4.

Second, some may argue that there is no distinction between type (A) and (B-b). Comparing sentence (7) and (10), we find they both have a yes/no answer.

(7) 你喜不喜歡這個遊戲？

Do you like this game?

(10) 他可不可以吃辣椒？

Can he eat hot peppers?

But, further considering sentence (14):

[5] The disambiguation of '什麼 what', see section 4.1.1.

(14) 他吃不吃辣椒？

Does he eat hot peppers?

We can still find the slight difference between a yes/no question and an information seeking question. However, modal words like 'can', 'shall', 'will' bring richer meaning then general verbs. They are not used to ask truth value but ask opinion or attitude. In Mandarin Chinese, '可不可以 *can not can*' and '喜不喜歡 *like not like*' are both with a *A not A* form, in syntactical considerations, they are often assigned to the same type. But from the semantic point of view, we decide they belong to both types.[6]

3.2 Knowledge Representation for Interrogatives

According to the classification above, we represent each type of interrogative as follows:

(15) true/false interrogative def: truth={.Ques.}

Wh-word interrogatives

asking factual information def: role={.Ques.}

asking relationship def:. Ques.RelationRole()

asking opinion def: ModalityRole={.Ques.}

option choosing def: role={.Option.{{x}.or.{y}}}

We use two operators, .Ques. and .Option., to denote querying focus or optional items. The real examples are shown in the following table 1.

[6] Shao (1996) has classified *A* not *A* form into five classes according to *A*'s part of speech, shown as follows: (1) *A* is a copula. e.g. 是不是 'be not be' (2) *A* is a modal word e.g. 好不好 'ok not ok' (3) *A* is an auxiliary e.g. 肯不肯 'willing not willing' (4) *A* is a verb e.g. 懂不懂 'understand not understand' (5) *A* is an adjective e.g. 美不美 'beautiful not beautiful'. From the semantic perspective, we merge (1),(4),(5) and (2),(3) to re-divide these five categories into two categories, i.e. modal *A* not *A* interrogatives and other *A* not *A* interrogatives.

Table 1. The Type Classification and Semantic Representation of Interrogatives

Question devices	**Examples**
true/false interrogatives	嗎*ma*;[7]是否;有沒有;是不是;不是嗎;*A not A* def: truth={.Ques.};
Wh-word interrogatives: asking factual information	誰 def: participant={animate:.Ques.}; 幾點鐘 def: time={.Ques.}; 什麼 def: participant={inanimate:.Ques.}; 什麼(車) def: formal={.Ques.}; 爲何,何以 def: reason={.Ques.}; 哪裡,哪兒 def: location={.Ques.}; 哪些 def: quantifier={.Ques.}; 怎麼 def: manner={.Ques.}; 怎麼 def: means={.Ques.}; 多 def: degree={.Ques.}; 多少 def: quantity={.Ques.}
asking relationship	
asking opinion	可不可以 def: allowance={.Ques.}; 好不好 def: willingness={.Ques.}; 能不能 def: capacity ={.Ques. }; 莫非 def: possibility ={.Ques. }
choosing options	還是;或 def: role={.Option.{{x}.or.{y}}}

The interrogatives above are gathered from Li & Thompson's analysis, and integrated by checking over 1000 question titles manually in *Baidu knows* (http://zhidao.baidu.com/).

4. Semantic Composition

The previous discussion is about logical representation of events. To establish a formal system to handle the task requiring language understanding, we also need to address the issue of semantic composition. Through segmentation and parsing process, we get coarse-grained arguments and the head of the sentence. Take sentence (16) as an example:

(16) 資料因何漏失？

Why is the data missing?

[7] In this paper, our focus is semantic representation, so we don't discuss the interrogative words '啊*a*'; '吧*ba'* or '呢*ni*'. Because it depends on the tone to decide they are interrogative words or not.

The segmentation and parsing result of (16) is:

Theme[NP:資料 *data*]+ reason[Dj:因何 *why*]+Head[VJ3:漏失 *lose*]

Then, we try to map surface syntax onto semantic structure for establishing truly integrated semantic relations. In example (16), we get the head 'lose' from segmentation and parsing process, and base on E-HowNet, the arguments of event 'lose' are 'possessor' and 'possession', we thus know the 'data' here is the possession of 'lose'. Therefore, the composition is as follows:

def:{lose|失去:possession={information|訊息},reason={.Ques.}}

The other types of interrogative words also can be combined into different sentences, shown as follows:

true/false interrogative

(17) 他病了嗎？

Is he sick?

def:{sick|病:experiencer={he|他},truth={.Ques.}}

(18) 你是否曾說過謊？

Have you lied before?

def:{lie|說謊:agent={listener|聽者},time={past|過去}, truth={.Ques.}}

Wh-word interrogative:

asking factual information

(19) 衣服上的墨水怎麼洗掉？

How to wash away ink stains on cloth?

def:{wash|洗掉:patient={ink|墨水:place={clothes|衣服}},means={.Ques.}}

(20) 哪些桌子壞掉？

Which tables are broken?

def:{OutOfOrder|壞:theme={table|桌子:quantifier={.Ques.}}}

asking relationship

(21) 她是你的什麼人？

What is the relationship between you and her?

def:{he|他=.Ques.kinship(listener|聽者)}

asking opinion

(22) 莫非這是鬼城？

Is it possible a ghost town?

def:{be|是:relevant={this|這},content={ghost town|鬼城}, possibility ={.Ques. }}

(23) 由你開車行不行？

Is it ok for you to drive?

def:{drive|開車:experiencer={listener|聽者},willingness ={.Ques. }}

choosing options

(24) 他在這兒還是那兒住？

Does he live here or there?

def:{live|住:agent={he|他},location={.Option.{{here|這兒}.or.{there|那兒}}}}

(25) 他跪下來還是站在那裡求張三？

Does he kneel on the ground or stand there to beg ChangShan?

def:{beg|求:agent={be|他}, target={ChangShan|張三}, means={.Option.{{stand|站}.or.{kneel 跪}}}}

4.1 Disambiguation

To achieve the goal of automatic composition, we have to face the challenge of sense ambiguities. In Chinese, '什麼 *she me*', '怎麼 *ze me*' and '多 *duo*' are most frequently used interrogatives with ambiguous senses. Their sense disambiguation rules and representations are discussed below:

4.1.1 什麼

'什麼 *what*' plays the grammatical functions of adjective and pronoun. For each function, there are two senses. Accordingly, we generalize four rules to disambiguate the word sense of '什麼', and the details are shown in the table 2:

Table 2. Disambiguous Rules of '什麼'

Rules of disambiguation	**E.g. & E-HowNet representation**
adjective 1: 什麼 '*what*' +semantic role: ask the value of the semantic role	**什麼 N role={.Ques.}** 什麼時間 '*what time*' time={.Ques.} 什麼價錢 '*what price*' price/cost={.Ques.} 什麼地點 '*what place*' location={.Ques.} 什麼狀況 '*what situation*' condition={.Ques.} 什麼顏色 '*what color*' color={.Ques.}
adjective 2: 什麼 '*what*'+entity (nominalized verbs are included): ask the type/restriction of a frame element/ participant role	**什麼 N participant={entity:formal={.Ques.}}** 什麼人 '*what person*' participant={human|人: formal={.Ques.}} 什麼汽車 '*what car*' participant={car|汽車:formal={.Ques.}} 什麼變化 '*what change*' participant={change|變化:formal={.Ques.}} 什麼不同 '*what difference*' participant={difference|不同:formal={.Ques.}}
pronoun 1: use as an interrogative pronoun: function as NP	**V 什麼 {event:participant={.Ques.}}** 吃什麼 '*eat what*' {eat|吃: patient={.Ques.}} 說什麼 '*talk what*'{speak|說:content={.Ques.}}
pronoun 2: use as an indefinite pronouns: function as NP, the played role is either coindexed or univarsal quantified.	**V 什麼；什麼 V {event:participant={x}}** **or{event:participant={entity: quantity={all}}}** 拿什麼都可以 *'It's OK to get anything'* {hold|拿:patient={entity: quantity={all}}} 什麼也不怕 *'Be afraid of nothing'* {.not.fear|害怕:cause={entity: quantity={all}}}

4.1.2 怎麼

'怎麼*how*/*how come*' plays the role of adverb. It asks the value of an adverbial type of semantic role including mean/method (How) and reason (Why). Two meanings 'how' and 'why' can roughly be discriminated by the telicity of matrix verbs. That is, '怎麼'in a sentence with telic verb or verb phrase refers to event that have endpoints means 'why'. Contrarily, '怎麼'in an atelic sentence means 'how'. Take sentence (26), (27) as an example:

(26) 怎麼來

def:{come|來:means={.Ques.}}

(27) 怎麼來了

def:{come|來:reason={.Ques.}}

In short, we conclude the disambiguation rule as follows:

(28) 怎麼(How)+event[-telic]

怎麼(Why)+event[+telic]

4.1.3 多

'多 *how*' also plays the role of adverb. It's usually followed by an attribute value, such as '甜 *sweet*', '聰明 *smart*', '遠 *far*', '大 *big*' and so forth. It can be used to express the feelings of exclamation or doubt. We can not simply distinguish these two senses by the context, but need to rely on the tone. For this reason, we will deal only with the senses of doubt. Incidentally, it is always possible to turn a declarative statement into a question by using a slightly rising intonation pattern. For the same reason, we do not deal with such sentences and few interrogative words such as '啊 *a*'; '吧 *ba'* ; '呢 *ni*' as well.

'多 *how*' with interrogative sense can be represent as below:

(29) 多+attribute value def:{attribute value: degree={.Ques.}}

The real examples are:

(30) 多甜 *'how sweet'* def:{sweet|甜:degree={.Ques.}} = def: sweetness={.Ques.}

多聰明 *'how smart'* def:{smart|聰明:degree={.Ques.}}= def: smartness={.Ques.}

多遠 *'how far'* def:{far|遠:degree={.Ques.}} = def: distance={.Ques.}

多大 *'how big'* def:{big|大:degree={.Ques.}} = def: size={.Ques.}

5. Conclusion and Future Works

To achieve near canonical semantic representation, we study the semantic representation and composition of interrogatives. According to the semantic classification of interrogative, we

represent interrogatives in a hierarchy as follows:

true/false interrogative	def: truth={.Ques.}
Wh-word interrogative	
asking factual information	def: role={.Ques.}
asking relationship	def:. Ques.RelationRole()
asking opinion	def: ModalityRole={.Ques.}
choosing options	def: role={.option.{{x}.or.{y}}}

We have cited two examples (1),(2) earlier to illustrate what is the 'understanding' of machine towards natural language. After the discussion above, let's see the result of this work:

(1) 我能否拍照？Is it OK for me to take pictures?

Representation:

我	def:{speaker|說話者}
能否	def:allowance={.Ques.}
拍照	def: {TakePicture|拍攝}

Composition:

def: {TakePicture|拍攝:agent={speaker|說話者}, allowance={.Ques.}}

(2) 我可不可以照相？Can I take photos?

Representation:

我	def:{speaker|說話者}
可不可以	def:allowance={.Ques.}
照相	def: {TakePicture|拍攝}

Composition:

def: {TakePicture|拍攝:agent={speaker|說話者}, allowance={.Ques.}}

Although the syntax surface of (1),(2) is different, we find the result of composition is the same. It means through the analysis of E-HowNet model, machine can judge the similarity of sentences, i.e. machine understand the sentences. However, this is only an example with

simple sentence structure. For future researches, we will implement a parsing system incorporated with E-HowNet model to demonstrate semantic composition process for more complex sentences. To achieve this goal, apart from sense disambiguation, discordance between synthetic structure and semantic relations is another critical problem. Take sentence (31) as an example:

(31) 長途旅行不是很辛苦嗎？ Isn't it hard for Long-distance travel?

Its parsing result is:

Theme[VP:(manner[A: 長 途 *long distance*]+Head[VA4: 旅 行 *travel*])]+negation[Dc: 不 *not*]+epistemics[Dbaa:是 *be*]+degree[Dfa:很 *very*]+Head[VH16:辛苦 *hard*]+particle[Td:嗎 *ma*]

Let's see the E-HowNet definition of (31) first:

def:{hard| 辛 苦 :theme={travel| 旅 行 :distance={far| 遠 }},degree={very| 很},truth={.Ques.}}

Comparing the semantic representation with syntactic structure, we find rhetorical interrogative '不是嗎 *Isn't it* ' is segmented into three words in syntax analysis, but in semantic point of view, they are integrated into one word and represented as 'truth={.Ques.}'. There are still many types of discordance between synthetic structure and semantic relations need to be studied. That is, we have to find out the mapping rules and match coarse-grained syntactic arguments to fine-grained semantic relations in the future. These rules should be able to use both on declarative sentences and interrogative sentences, because most of interrogative sentences are transformed from declarative sentences. Additionally, this study is also useful to question-answering system for it not only represents the sense of question, but also marks the focused information to be answered. As for the application on QA technologies, it'll be our future task as well.

Acknowledgement:

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 95-2752-E-001-001-PAE and Grant NSC95-2221-E-001-039.

References

[1] Chen Keh-Jiann, Huang Shu-Ling, Shih Yueh-Yin, Chen Yi-Jun, *Extended-HowNet: multi-level concept definition and complex relation description*, Peking University, 2004

[2] Chen Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, *Extended-HowNet- A Representational Framework for Concepts*, IJCNLP-05 Workshop, Jeju Island, South Korea, 2005

[3] HowNet, *http://www.keenage.com/*, 1988

[4] Jonathan Ginzburg & Ivan A.Sag, *Interrogative Investigations*, Stanford University, 2000

[5] Li & Thompson, *Mandarin Chinese*, The Crane Publishing Co., LTD., 1997

[6] Sinica Treebank, *http://turing.iis.sinica.edu.tw/treesearch/*

[7] 石定栩(Shi, Dingxu)，‘疑問句研究’，收錄於‘*共性與個性—漢語語言學中的爭議*’，北京語言文化大學出版社，1999

[8] 呂叔湘(Lv, Shuxiang)，*中國文法要略*，1942

[9] 邵敬敏(Shao, Jingmin)，*現代漢語疑問句研究*，華東師範大學出版社，1996

[10] 馬建忠(Ma, Jianzhong)，*馬氏文通*，商務印書館，1935 (first edition:1989)

[11] 陸儉明(Lu, Jianming)，‘關於現代漢語裡的疑問語氣詞’，*中國語文*，第五期，1984

[12] 湯廷池(Tang, Tingchi)， ‘國語疑問句的研究’ ，*師大學報*，第二十六期，1983

[13] 黎錦熙(Li, Chinhsi)，*新著國語文法*(Chinese grammar on the national language)，商務印書館，1930

Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems

Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
{lwku, eagan}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Question answering systems provide an elegant way for people to access an underlying knowledge base. Humans are not only interested in factual questions but also interested in opinions. This paper deals with question analysis and answer passage retrieval in opinion QA systems. For question analysis, six opinion question types are defined. A two-layered framework utilizing two question type classifiers is proposed. Algorithms for these two classifiers are described. The performance achieves 87.8% in general question classification and 92.5% in opinion question classification. The question focus is detected to form a query for the information retrieval system and the question polarity is detected to retain relevant sentences which have the same polarity as the question. For answer passage retrieval, three components are introduced. Relevant sentences retrieved are further identified whether the focus (*Focus Detection*) is in a scope of opinion (*Opinion Scope Identification*) or not, and if yes, whether the polarity of the scope matches with the polarity of the question (*Polarity Detection*). The best model achieves an F-measure of 40.59% using *partial match* at the level of meaningful unit. With relevance issues removed, the F-measure of the best model boosts up to 84.96%.

1 Introduction

Most of the state-of-the-art Question Answering (QA) systems serve the needs of answering factual questions such as "When was James Dean born?" and "Who won the Nobel Peace Prize in 1991?". In addition to facts, people would also like to know about others' opinions, thoughts, and feelings toward some specific topics, groups, and events. Opinion questions (e.g. "How do Americans consider the US-Iraq war?" and "What are the public's opinions on human cloning?") revealing answers about people's opinions have long as well as complex answers which tend to scatter across different documents. Traditional QA approaches are not effective enough to retrieve answers for opinion questions as they have been for factual questions (Stoyanov et al., 2005). Hence, an opinion QA system is essential and urgent.

Most of the research on QA systems has been developed for factual questions, and the association of subjective information with question answering has not yet been much studied. As for subjective information, Wiebe (2000) proposed a method to identify strong clues of subjectivity on adjectives. Riloff et al. (2003) presented a subjectivity classifier using lists of subjective nouns learned by bootstrapping algorithms. Riloff and Wiebe (2003) proposed a bootstrapping process to learn linguistically rich extraction patterns for subjective expressions. Kim and Hovy (2004) presented a system to determine word sentiments and combined sentiments within a sentence. Pang, Lee, and Vaithyanathan (2002) classified documents not by the topic, but by the overall sentiment, and then determined the polarity of a review. Wiebe et al. (2002) proposed a method for opinion summarization. Wilson et al. (2005) presented a phrase-level sentiment analysis to automatically identify the contextual polarity.

Ku et al. (2006) proposed a method to automatically mine and organize opinions from heterogeneous information sources.

Some research has gone from opinion analysis in texts toward that in QA systems. Cardie et al. (2003) took advantage of opinion summarization to support Multi-Perspective Question Answering (MPQA) system which aims to extract opinion-oriented information of a question. Yu and Hatzivassiloglou (2003) separated opinions from facts, at both the document and sentence levels. They intended to cluster opinion sentences from the same perspective together and summarize them as answers to opinion questions. Kim and Hovy (2005) identified opinion holders, which are frequently asked in opinion questions.

This paper deals with two major problems in opinion QA systems: question analysis and answer passage retrieval. Several issues, including how to separate opinion questions from factual ones, how to define question types for opinion questions, how to correctly classify opinion questions into corresponding types, how to present answers for different types of opinion questions, and how to retrieve answer passages for opinion questions, are discussed. Note that the unit of a passage is a sentence in this paper, though a passage can sometimes refer to more sentences, such as a paragraph.

2 An Opinion QA Framework

Figure 1 is a framework of the opinion QA system. The question is initially submitted into a part of speech tagger (POS Tagger), and then the question is analyzed in three aspects: the question focus, the question polarity, and the opinion question type. The former two attributes are further applied in answer passage retrieval. The question focus is the query for an information retrieval (IR) system to retrieve relevant sentences. The question polarity is utilized to screen out relevant sentences with different polarities to the question. With answer passages retrieved, answer extraction extracts text spans as answers according to the opinion question types, and outputs answers to the user.

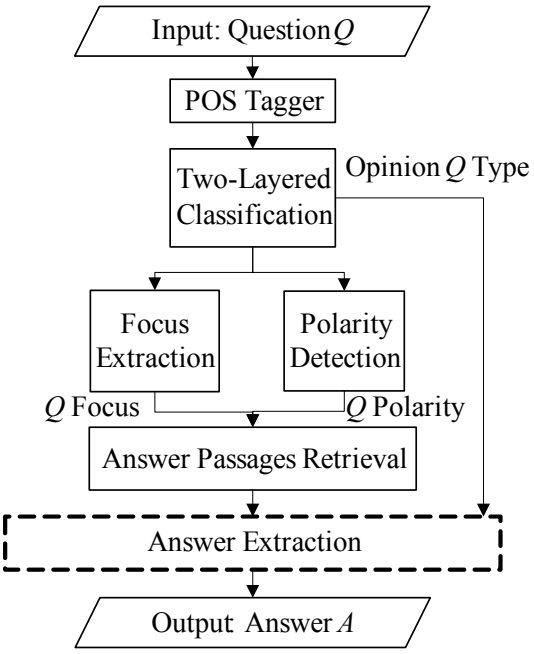


Figure 1. An Opinion QA System Framework.

3 Experimental Corpus Preparation

The experimental corpus comes from four sources, i.e. TREC[1], NTCIR[2], the Internet Polls, and OPQ. TREC and NTCIR are two of three major information retrieval evaluation forums in the world. Their evaluation tracks are in natural language processing and information retrieval domains such as large-scale information retrieval, question answering, genomics, cross language processing, and many new hot research topics. We collect 500 factual questions from the main task of QA Track in TREC-11. These English questions are translated into Chinese for experiments. A total of 1,577 factual questions are obtained from the developing question set of the CLQA task in NTCIR-5. Questions from public opinion polls in three public media websites, Chinatimes, Era, and TVBS, are crawled. OPQ is developed for this research, and it contains both factual and opinion questions. To construct the question corpus OPQ, annotators are given titles and descriptions of six opinion topics selected from NTCIR-2 and NTCIR-3. Annotators freely ask any three factual questions and seven opinion questions for each topic. Duplicated questions are dropped and a total of 1,011 questions are collected. Within these 1,011 questions in OPQ, 304 are factual questions and the other 707 are opinion questions.

Overall, we collect 2,443 factual questions and 1,289 opinion questions from four different sources. A total of 3,732 questions are gathered for our experiments, as shown in Table 1.

Q type / Corpus	Factual	Opinion	Total
TREC	500	0	500
NTCIR	1,577	0	1,577
Polls	62	582	644
OPQ	304	707	1,011
Total	2,443	1,289	3,732

Table 1. Statistics of Experimental Questions.

There are some challenging issues in extracting answers automatically by opinion QA systems. We categorize these challenges (indexed by numbers and enclosed by parentheses as follows) in question analysis into on holders, on opinions and on concepts.

On holders, (1) to automatically identify named entities expressing opinions is imperative. (2) Grouping opinion holders is another issue. Answers to the question, "How do Americans feel about the affair of the U.S. president Clinton?", consist of opinions from any American. To answer questions like "What kind of people support the abolishment of the Joint College Entrance Examination?", QA systems have to find people having opinions toward the examination and (3) classify them into correct category, such as students, teachers, scholars, parents, and so forth.

On opinions, (4) knowing whether questions themselves contain subjective information and deciding their opinion polarities is necessary. The question "Who disagrees with the idea of surrogate mothers?" points out a negative attitude and the answer to this question is expected to be a list of persons or organizations that have negative opinions toward the idea of surrogate mothers. Another issue is (5) whether the comparison and the summarization of positive and negative opinions are required. In the question "Is using the civil ID card more advantageous or disadvantageous?", opinions expressing advantages and disadvantages have

1 http://trec.nist.gov/
2 http://research.nii.ac.jp/ntcir/index-en.html

to be contrasted and scored to represent answers as "More advantageous" or "More disadvantageous" with evidence listed to users.

On concepts, it is essential (6) to understand the concepts of opinions and perform the expansion on concepts to extract correct answers. In the question "Is civil ID card secure?" it is vital to know the definition and expansion of being secure. Keeping public's privacy, ensuring system's security, and protecting fingerprints' obtainment are possible security points. For (7) the concept of targets, the idea is the same as the concept of opinions except that it is about targets. For instance, the question "What do Taiwanese think about the substitute program of Joint College Entrance Examination?" necessitates the comprehension of what the substitute program is or the alias of this program, and then the system can seek for text spans which hold opinions towards it.

Among the 707 opinion questions from OPQ corpus, answers of 160 opinion questions are found in the NTCIR corpus. These 160 opinion questions are analyzed based on the above seven challenges. Table 2 lists the number of questions (#*Q*) with respect to the number of challenges (#*C*).

#*C*	1	2	3	4	5	6	7	Total
#*Q*	19	47	39	30	13	12	0	160

Table 2. Challenge of Opinion Questions.

A total of 60 questions are selected for further annotation based on their challenges. Sentences are annotated as whether they are opinions (*Opinion*), whether they are relevant to the topic (*Rel2T*), whether they are relevant to the question (*Rel2Q*), and whether they contain answers (*AnswerQ*). If sentences are annotated as relevant to the question, annotators further annotate the text spans which contribute answers to the question (*CorrectMU*).

4 Two-layered Question Classification

A two-layered classification, i.e. with Q-Classifier and OPQ-Classifier, is proposed. Q-Classifier separates opinion questions from factual ones, and OPQ-Classifier tells types of opinion questions.

4.1 Types of Opinion Questions

According to opinion questions themselves and their corresponding answers, we define six opinion question types as follows.

(1) **Holder (HD)**
Definition: Asking who the expresser of the specific opinion is.
Example: Who supports the civil ID card?
Answer: Entities and the corresponding evidence.

(2) **Target (TG)**
Definition: Asking whom the holder's attitude is toward.
Example: Who does the public think should be responsible for the airplane crash?
Answer: Entities and the corresponding evidence.

(3) **Attitude (AT)**
Definition: Asking what the attitude of a holder to a specific target is.
Example: How do people feel about the affair of the U.S. President Clinton?
Answer: Question-related opinions, separated into support, neutral, and non-support categories.

(4) **Reason (RS)**
Definition: Asking the reasons of an explicit or an implicit holder's attitude to a specific target.
Example: Why do people think better not to have the college entrance exam?
Answer: Reasons for taking the stand specified.

(5) **Majority (MJ)**
Definition: Asking which option, listed or not listed, is the majority.
Example: If the government tries to carry out the usage of the civil ID card, will its reputation get better or worse?
Answer: The majority of support, neutral and non-support evidence.

(6) **Yes/No (YN)**
Definition: Asking whether their statements are correct.
Example: Is the airplane crash caused by management problems?
Answer: The stronger opinion, i.e. yes or no.

4.2 Q-Classifier

Q-Classifier distinguishes opinion questions from factual questions. We use See5 (Quinlan, 2000) to train Q-Classifier. Seven features are employed. The feature *pretype* (PTY) denotes types in factual QA systems such as SELECTION, YESNO, METHOD, REASON, PERSON, LOCATION, PERSONDEF, DATE, QUANTITY, DEFINITION, OBJECT, and MISC and extracted by a conventional QA system (*reference removed for blind review*). For example, the value of *pretype* in "Who is Tom Cruise married to?" is PERSON.

The other six features are *operator* (OPR), *positive* (POS), *negative* (NEG), *totalow* (TOW), *totalscore* (TSR), and *maxscore* (MSR). A public available sentiment dictionary (Ku et al., 2006), which contains 2,655 positive opinion keywords, 7,767 negative opinion keywords, and 150 opinion operators, is used to tell if there are any positive (negative) opinion keywords and operators in questions. Each opinion keyword has a score expressing the degree of tendency. The feature *operator* (OPR) includes words of actions for expressing opinions. For example, say, think, and believe can be hints for extracting opinions. A total of 151 operators are manually collected. The feature *totalow* (TOW) is the total number of opinion operators, positive opinion keywords, and negative opinion keywords in a question. The feature *totalscore* (TSR) is the overall opinion score of the whole question, while the feature *maxscore* (MSR) is the absolute maximum opinion score of opinion keywords in a question.

Section 3 mentions that 2,443 factual questions and 1,289 opinion questions from four different sources are collected. To keep the quantities of factual and opinion questions balanced, 1,289 factual questions are randomly selected from 2,443 questions and a total of 2,578 questions are employed. We adopt See5 to generate the decision tree based on different combinations of features.

With a 10-fold cross-validation, See5 outputs the resulting decision trees for each 10 folds, and a summary with the mean of error rates produced by these 10 folds. Table 3 shows experimental results. Only with feature *x* shows the error rate of using one single feature, while with all but feature *x* shows the error rate of using all features except the specified feature.

feature *x*	PTY	OPR	POS	NEG
only with feature *x*	19.6	38.5	34.9	35.3
with all but feature *x*	16.3	12.7	13.7	12.2
feature *x*	TOW	TSR	MSR	ALL
only with feature *x*	21.9	26.6	29.6	12.2
with all but feature *x*	14.8	12.4	12.8	

Table 3. Error Rates of Q-Classifier.

The features *pretype* (PTY) and *totalow* (TOW) perform best in reducing errors when used alone. They also cannot be ignored since the error rate increases more when they are excluded. The feature *totalow* shows that if a question contains more opinion keywords, it is more possible to be an opinion question. After all features are considered together, the best performance is 87.8%.

4.3 OPQ-Classifier

OPQ-Classifier categorizes opinion questions into the corresponding opinion question types. We first examine if there is any specific patterns in the question. If yes, then the rule for the pattern is applied. Otherwise, a scoring function is applied.

The heuristic rules are listed as follows.

(1) The pattern "A-not-A": Yes/No
(2) End with question words: Yes/No
(3) "Who" + opinion operator: Holder
(4) "Who" + passive tense: Target
(5) *pretype* (PTY) is Reason: Reason
(6) *pretype* (PTY) is Selection: Majority

A scoring function deals with those questions which cannot be classified by the above patterns. Unigrams, bigrams and trigrams in training questions are selected as feature candidates. These feature candidates are separated into two types. A topic dependent feature is only meaningful in questions of some topics, while general features may appear in questions of all kinds of topics. If a feature is topic dependent (e.g. human cloning and Clinton), it is dropped from the feature set. Only general features (e.g. is or is not, whether, and reason) are kept. Finally a set of features is obtained from the training questions. Then the discriminate power of these features is calculated as follows.

First, the observation probability of a feature *i* in the question type *j* is defined in Formula (1).

$$P_o(i,j) = \frac{NumQ(i,j)}{NumQ(j)} \qquad (1)$$

where *i* is the index of the feature, *j* is the index of the question type, and *NumQ* represents the number of questions. The observation probability shows how often a feature is observed in each type. It is then normalized by Formula (2).

$$NP_o(i,j) = \frac{P_o(i,j)}{\sum_{j=1}^{6} P_o(i,j)} \qquad (2)$$

Every feature has six normalized observation probabilities corresponding to the six types. With these probabilities, the score *ScoreQ* of a question can be calculated by Formula (3).

$$ScoreQ(j) = \sum_{i=1}^{n} NP_o(i, j) \qquad (3)$$

where n is the total number of features in question Q, and $ScoreQ(j)$ represents the score of question Q as type j. Since there are six possible opinion question types, the six *ScoreQ* represent how possible the question Q belongs to each type. These six scores form the feature vector of the question Q for classification.

Training instances are used to find the centroid of each type. The Pearson correlation is adopted as the distance measure. The distances between the testing opinion questions and the six centroids are calculated to assign the opinion questions to the closest type.

Number		Opinion question type					
		HD	TG	AT	RS	MJ	YN
Classified as	HD	27	0	0	1	0	0
	TG	0	5	0	0	0	0
	AT	0	0	68	0	0	0
	RS	1	0	4	17	0	0
	MJ	0	0	0	0	8	0
	YN	3	3	15	5	5	385
	Total	31	8	87	23	13	385

Table 4. Confusion Matrix (Number).

%		Opinion question type					
		HD	TG	AT	RS	MJ	YN
Classified as	HD	87.1	0.0	0.0	4.4	0.0	0.0
	TG	0.0	62.5	0.0	0.0	0.0	0.0
	AT	0.0	0.0	78.2	0.0	0.0	0.0
	RS	3.2	0.0	4.6	73.9	0.0	0.0
	MJ	0.0	0.0	0.0	0.0	61.5	0.0
	YN	9.7	37.5	17.2	21.7	38.5	100
	Total	100	100	100	100	100	100

Table 5. Confusion Matrix (Percentage).

We use the OPQ corpus in Section 3 for the evaluation of the OPQ-Classifier. The opinion types of these opinion questions are manually given. Among the 707 opinion questions, answers of 160 opinion questions are found in the NTCIR corpus. They are used as the training data for an intensive analysis of both questions and answers. The rest 547 opinion questions are used as the testing data. The confusion matrix of the OPQ-Classifier is shown in Table 4 and 5. The average accuracy is 92.5%. There are fewer questions of target (TG) and majority (MJ) types, 8 and 13 in testing collection respectively. The unsatisfactory results of these two types may due to the lack of training questions.

5 Answer Passage Retrieval

Figure 2 shows the framework of answer passage retrieval in an opinion QA system. The question focus supplied by the question analysis serves as the input to an Okapi IR system to retrieve relevant sentences from the knowledge base. Relevant sentences are further detected to identify whether the focus (*Focus Detection*) is in a scope of opinion text spans (*Opinion Scope Identification*) or not, and if yes, whether the polarity of the scope matches with the polarity of the question (*Polarity Detection*). The details are discussed in the following sections.

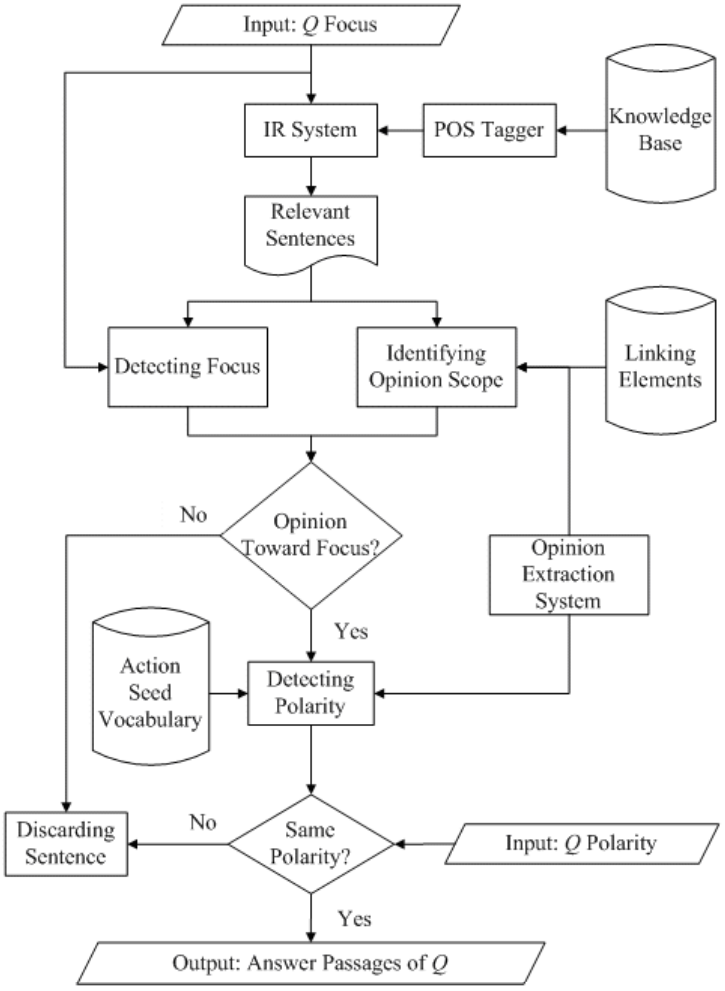


Figure 2. Answer Passage Retrieval.

5.1 Question Focus Extraction

The first stage of answer passage retrieval is to input the question focus as a query into an IR system to retrieve relevant sentences from the knowledge base. These retrieved sentences may contain answers for a question. A set of content words in one question is used to represent its focus. The following steps extract a set of content words as the question focus and formulate a query.

(1) Remove question marks.
(2) Remove question words.
(3) Remove opinion operators.
(4) Remove negation words.
(5) Name the remaining terms as focus.
(6) Use the Boolean OR operator to form a query.

Since question marks and question words are common in every question, they do not contribute to the retrieval of relevant sentences, and therefore are removed. Opinion operators and negation words are removed as well since they represent the question polarity instead of the question focus. Once we have the question focus, we use the Boolean OR operator rather than the AND operator to form a query. This is because we prefer the IR system to return sentences that have any relevancy to the question.

5.2 Question Polarity Detection

The polarity of the question is useful in opinion QA systems to filter out query-relevant sentences which have different polarities from the question. If the question polarity is positive, the sentences providing answers ought to be positive, and vice versa. The polarity detection algorithm is shown as follows.

(1) Determine the polarity of the opinion operator. 1 is for positive, 0 is for neutral, and -1 is for negative.
(2) Negate the operator polarity if there is any negation word anterior to the operator.
(3) Determine the polarity of the question focus. 1 is for positive, 0 is for neutral, and -1 is for negative.
(4) If one of the operator polarity and question focus is 0 (neutral), output the sign of the other; else output the sign of the product of the polarities of the opinion operator and the question focus.

We regard the polarity of the question focus together with the polarity of the opinion operator because the opinion operator primarily shows the opinion tendency of the question and different polarities of the question focus can affect the polarity of the entire question. A positive opinion operator stands for a supportive attitude such as “agree”, “approve”, and “support”. A neutral opinion operator stands for a neutral attitude such as “state”, “mention”, and “indicate”. A negative opinion operator stands for a not-supportive attitude such as “doubt”, “disapprove”, and “protest”. In the question “Who approves the Joint College Entrance Examination?”, “approve” is a positive operator, and “the Joint College Entrance Examination” is a neutral question focus. The overall polarity of this question is positive, so the opinion QA system needs to retrieve sentences that contain a positive polarity to “the Joint College Entrance Examination.” In contrast, in the question “Who agrees the abolishment of the Joint College Entrance Examination?”, the question focus “the abolishment of the Joint College Entrance Examination” becomes negative because of “the abolishment”. Even though the operator is positive, opinion QA systems still have to look for sentences that contain negative opinions toward “the Joint College Entrance Examination.”

5.3 Opinion Scope Identification

In Chinese, a sentence ending with a full stop may be composed of several sentence fragments *sf* separated by commas or semicolons as follows: ”sf_1，sf_2，sf_3，…，sf_n。”.

This paper (*reference removed for blind review*) shows that about 75% of Chinese sentences contain more than two sentence fragments.

An opinion scope denotes a range expressing attitudes in a sentence. It may be a complete sentence, a sentence fragment, or a meaningful unit (MU) based on different criteria. It is very common that many concepts are expressed within one sentence in Chinese documents. Therefore to identify the complete concept, which is denoted as a meaningful unit, in sentences is necessary for the processing of relevant opinions. As mentioned, a Chinese sentence is composed of several sentence fragments, and one or more of them can form a meaningful unit, which expresses a complete concept. This paper (*reference removed*) employed linking elements (Li and Thompson, 1981) such as "because", "when", *etc*. to compose MUs from a sentence. In *S* (in Chinese), "因此" (thus) is a linking element which links sf_2, sf_3, and sf_4 together, and sf_2 is a subordinate clause of the operator "表示" (indicate) in sf_1. Therefore, sf_1, sf_2, sf_3, and sf_4 form a MU in this case.

S: sf_1: 黃宗樂**表示**(indicate:operator)，
sf_2: 發行國民 IC 卡牽涉到基本人權，
sf_3: 因此(thus:linking element)，
sf_4: 在決策過程上必須相當嚴密，
sf_5: 例如日本就未發行國民身份證。

5.4 Focus Detection

The IR system takes a sentence as a retrieval unit and reports those sentences that are probably relevant to a given query. The focus detection aims to know which sentence fragments are useful to extract answer passages. Three criteria of focus detection, namely *exact match*, *partial match*, and *lenient*, are considered. In an extreme case (i.e. *lenient*), all the fragments in a retrieved sentence are regarded as relevant to the question focus. In another extreme case (i.e. *exact match*), only the fragment containing the complete question focus is regarded as relevant. In other words, *exact match* filters out the irrelevant fragments from the retrieved sentences. *Partial match* is weaker than *exact match* and is stronger than the *lenient* criterion. Those fragments which contain a part of the question focus are regarded as relevant.

There are three criteria for focus detection and opinion scope identification, respectively, thus a total of 9 combinations are considered. For example, a combination of *exact match* and meaningful units means there is at least one focus in meaningful units. Similarly, a combination of *partial match* and sentence fragments indicates that there is at least one partial focus in sentence fragments.

5.5 Polarity Detection

Given a combination of the above strategies, we have a set of opinion scopes relevant to the specific focus. Polarity detection tries to identify the scopes which have the same polarities as the question. How to determine the opinion polarity is an important issue. Two approaches are adopted. The opinion word approach employs a sentiment dictionary to detect if some words in this dictionary appear in a scope. The score of an opinion scope is the sum of the scores of these words.

People sometimes imply their feelings or beliefs toward a particular target or event by actions. For example, people may not say "Objection!" to disagree an event, but they may try to abolish or terminate it as possible as they could. On the other hand, people may not say "I'm loving it!" to show their delight to an event, but they may try to fight for it or legalize it.

In both circumstances, what people take in action expresses their opinions. Action words are those which indicate a person's willing of doing or not doing some behaviors. For example, *carry out*, *seek*, and *follow* are words showing willingness to do something, and we name these words as *do's*; *substitute*, *stop*, and *boycott* are words showing unwillingness to do something, and we name these words as *don'ts*. In the action word approach, we detect opinions in scopes with the help of a seed vocabulary of *do's* and *don'ts*, together with a sentiment dictionary.

5.6 Experiments on Answer Passage Retrieval

The F-measure metric is used for evaluation for the answer passage retrieval. To answer an opinion question, all answer passages have to be retrieved for opinion polarity judgment. Therefore, the conventional evaluation metric that uses the precision and recall at a certain rank, e.g. top 10, may not be suitable for this task. Since all answer passages, sentence fragments and meaningful units which provide correct answers are already annotated in the testing bed, the F-measure metric can be applied without questions. Tables 6 and 7 show the F-measures of answer passage retrieval using the opinion word approach and the action word approach, respectively. In these two approaches, adopting meaningful units as opinion scopes is better than adopting sentences and sentence fragments. Considering both opinion and action words are better than opinion words only. The best F-measure 40.59% is achieved when meaningful units and *partial match* are used.

Opinion Scope → / Focus Detection ↓	sentence	sentence fragment	meaningful unit
Exact Match	32.09%	36.06%	**36.25%**
Partial Match	27.32%	27.46%	**33.09%**
Lenient	19.91%	19.95%	**25.05%**

Table 6. F-Measure of Opinion Word Approach.

Opinion Scope → / Focus Detection ↓	sentence	sentence fragment	meaningful unit
Exact Match	28.75%	30.20%	**36.36%**
Partial Match	32.83%	35.09%	**40.59%**
Lenient	27.15%	29.19%	**32.87%**

Table 7. F-Measure of Action Word Approach.

5.7 Experiments on Relevance Effects

The previous experiments were done on sentences reported by the Okapi IR system. These retrieved sentences are not all relevant to the questions. This section will discuss how the relevance affects answer passage retrieval. Recall that the experimental corpus is annotated with *Rel2T* (relevant or irrelevant to the topic), *Rel2Q* (relevant or irrelevant to the question), *CorrectMU* (text spans containing answers to the question).

Assume meaningful units are taken as the opinion scope. Tables 8 and 9 show how relevance influences the performance of answer passage retrieval using the opinion word and action word approaches, respectively.

Rel Degree → / Focus Detection ↓	Rel2T	Rel2Q	CorrectMU
Exact Match	**36.69%**	36.73%	50.43%
Partial Match	34.79%	47.15%	70.15%
Lenient	28.03%	**48.35%**	**80.73%**

Table 8. Relevance Effects on Answer Passage Retrieval Using Opinion Word Approach.

Rel Degree → / Focus Detection ↓	Rel2T	Rel2Q	CorrectMU
Exact Match	36.88%	36.92%	48.99%
Partial Match	**41.90%**	50.37%	72.84%
Lenient	37.04%	**53.06%**	**84.96%**

Table 9. Relevance Effects on Answer Passage Retrieval Using Action Word Approach.

Rel2T shows the performance of using answer passages relevant to the six topics, that is, the original relevant documents from NTCIR CLIR task. *Rel2Q* shows the performance of using answer passages relevant to the questions, while *CorrectMU* shows the performance of using correct opinion fragments, which are relevant to the question focus, to decide opinion polarities. *Rel2T* is similar to the relevant sentence retrieval, which was shown to be tough in TREC novelty track (Soboroff and Harman, 2003). From *Rel2T* to *Rel2Q* and *CorrectMU*, the best strategy for matching the question focus switches from *partial match* to *lenient*. This is reasonable, since the contents of *Rel2Q* and *CorrectMU* are already relevant to the question focus. In *Rel2Q*, doing focus detection doesn't benefit or harm a lot (50.37% vs. 53.06%). It shows that the question focus will appear exactly or partially in the relevant sentences. However, focus detection lowers the performance in *CorrectMU* (72.84% vs. 84.96%). It tells that the question focus and the correct meaningful units may appear in different positions within the sentence. For example, the first meaningful unit talks about the question focus, while the third meaningful unit really answers the question but omits the question focus since it is mentioned earlier. From *Rel2T* to *Rel2Q*, the F-measure does not increase as much as that from *Rel2Q* to *CorrectMU*. This result shows that finding the correct fragments of passages to judge the opinion polarity is very crucial to answer passage retrieval. The F-measure of *CorrectMU* shows the performance of judging opinion polarities without the relevant issue. Using either the opinion word approach or the action word approach achieves an F-measure greater than 80%. As a whole, including action words is better than using opinion words only.

6 Conclusion

This paper proposes some important techniques for opinion question answering. For question classification, a two-layered framework including two classifiers is proposed. General questions are divided into factual and opinion questions, and then opinion questions themselves are classified into one of the six opinion question types defined in this paper. With both factual and opinion features for a decision tree model, the classifier achieves a precision rate of 87.8% for general question classification. With heuristic rules and the Pearson correlation coefficient as the distance measurement, the classifier achieves a precision rate of 92.5% for opinion question classification.

For opinion answer passage retrieval, we concern not only the relevance but also the sentiment. Considering both opinion words and action words is better than considering opinion words only. Taking meaningful units as the opinion scope is better than taking sentences. Under the action word approach, the best model achieves an F-measure of 40.59% using *partial match* at the level of meaningful unit. With relevance issues removed, the F-measure of the best model boosts up to 84.96%. Understanding the meaning of the question focus is important for the relevance detection, but some foci are quite challenging in the experiments. Query expansion and concept ontology will be explored in the future.

References

Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. In *Proceedings of AAAI Spring Symposium Workshop*, 20-27

Kim, S.-M. and Hovy, E. 2004. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367-1373.

Kim, S-M and Hovy, E. 2005. Identifying Opinion Holders for Question Answering in Opinion Texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*.

Ku, L.-W., Liang, Y.-T. and Chen, H.-H. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report*, 100-107.

Li, C.N. and Thompson, S.A. 1981. *Mandarin Chinese: A Functional Reference Grammar*, University of California Press.

Pang, B., Lee, L. and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on EMNLP*, 79-86.

Quinlan, J. R. 2000. Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html

Riloff, E. and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003* Conference *on EMNLP*, 105-112.

Riloff, E., Wiebe, J. and Wilson, T. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proceedings of Seventh Conference on Natural Language Learning*, 25-32.

Soboroff, I. and Harman, D. 2003. Overview of the TREC 2003 novelty track. In *Proceedings of Twelfth Text REtrieval Conference,* National Institute of Standards and Technology, 38-53.

Stoyanov, V., Cardie, C. and Wiebe, J. 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proceedings of HLT/EMNLP 2005*, 923-930.

Wiebe, J. 2000. Learning Subjective Adjectives from Corpora. In *Proceeding of 17th National Conference on Artificial Intelligence*, 735-740.

Wiebe, J., Breck, E., Buckly, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E. and Wilson, T. 2002 NRRC Summer Workshop on Multi-Perspective Question Answering. *ARDA NRRC Summer 2002 Workshop*.

Wilson, T., Wiebe, J. and Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT/EMNLP 2005*, 347-354.

Yu, H., and Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of HLT/EMNLP 2003*, 129-136.

基於統計方法之中文搭配詞自動擷取

張翠芸、柯淑津

東吳大學資訊科學系
Department of Computer Information Science
SooChow University
ms9513@sun.cis.scu.edu.tw
ksj@cis.scu.edu.tw

摘要

本研究採取以下四個步驟擷取出雙連詞、三連詞、四連詞之詞彙或詞性組合之搭配詞。首先採用 Smadja's Xtract 的平均數及變異數的方法，擷取具有變動距離模式所共同出現的詞彙或詞性的組合，接著使用搭配詞顯著性的衡量方法：相互資訊值及 T 檢定值。通過以上檢驗的候選搭配詞，經由對照中央研究院詞義標示語料庫之目標詞的結果，在同樣的跨距下，若同爲一個詞義者，則我們以此搭配詞作爲詞義標示知識。並且，本研究將產出之搭配資訊應用於詞義自動標示處理，達到 20.07%的應用率及 90.83%的正確率。

Abstract

We take the four following steps to extract collocations made of combinations of 2, 3, 4 words and/or part of speech, respectively. First, we use "Smadja's Xtract" to extract the co-occurrence combinations of words and/or part of speech of varying distance by computing means and variances. Second, we evaluate the significances of collocation candidates by 2 metrics: mutual information and t-test value. At last, we compare the head words of tagged word sense corpus made by Academic Sinica with the collocation candidates. If in the same distance, the head words of collocation candidates match the ones made by Academic Sinica, we say they are collocations. In addition, we apply the collocation information produced from this research to word sense disambiguation. It reaches application rate of 20.07% and precision rate of 90.83%.

關鍵詞：中文搭配詞，相互資訊值，自然語言處理，統計方法，T 檢定值，詞義辨識

Keywords: Chinese collocation, mutual information, natural language processing, statistical method, t-test, word sense disambiguation.

一、簡介

不同民族的歷史文化知識背景以及人們的思考邏輯模式不同，看待同樣的人事物、同樣的行爲情境過程，在語言的描述上也會有所不同。每個地區的語言都有其習慣性的用

法，而所謂的搭配詞 (collocation) 廣義而言，就是指兩個或多個詞依照語言習慣性結合在一起表示某種特殊意涵的詞彙現象。搭配詞在不同的研究領域上各有不同的解讀，尚未有一致性的定義。研究搭配詞著名的學者 Smadja [1] 定義搭配詞有以下四個特徵：1、搭配詞是任意詞的組合；2、搭配詞和領域相關；3、搭配詞是重複出現的；4、搭配詞具有詞彙的互相吸引性。母語使用者對於搭配詞的判定也許相當容易，但對於外國人的語言學習，常會誤用搭配語詞。以往對於搭配詞自動擷取的研究，大多是針對英語系語料做處理。至於擷取中文搭配詞的相關文獻仍然是相當稀少的，因此本研究利用統計的方式對大規模的中文資料進行分析以擷取出中文搭配詞。其產出的結果將可以應用在自然語言相關處理上，例如：詞義自動標示、資訊檢索、機器翻譯以及辭典編纂。

本研究提出將周邊詞彙及詞性皆作爲擷取搭配詞的重要特徵，採用 Smadja's Xtract [1] 基於統計上的平均數及變異數之方法，直接擷取出具有變動距離模式所共同出現的詞彙或詞性之組合，再使用搭配詞顯著性的衡量方法：相互資訊值 (Mutual Information) 和 T 檢定值。通過以上檢驗的候選搭配詞，在最後判定搭配詞的基準，是基於每個搭配詞僅一個詞義的原由 [2]，我們採取經由對照中央研究院詞義標示語料庫 SSTC (Sinica Sense-Tagged Corpus) [3]，在相同目標詞彙和周邊詞彙資訊的跨距下，目標詞在語料庫訓練資料的所有詞例中，若僅具唯一詞義，則我們就將此搭配詞擷取爲詞義標示知識；進一步再以相同的方式進而擷取三連詞及四連詞之搭配詞。最後我們將產出之搭配資訊應用於詞義自動標示處理。

本文組織如下，第二節是有關搭配詞擷取技術之相關文獻探討。第三節說明本研究提出之擷取搭配詞方法。第四節爲實驗設計與結果評估。最後，是本文的總結。

二、相關文獻

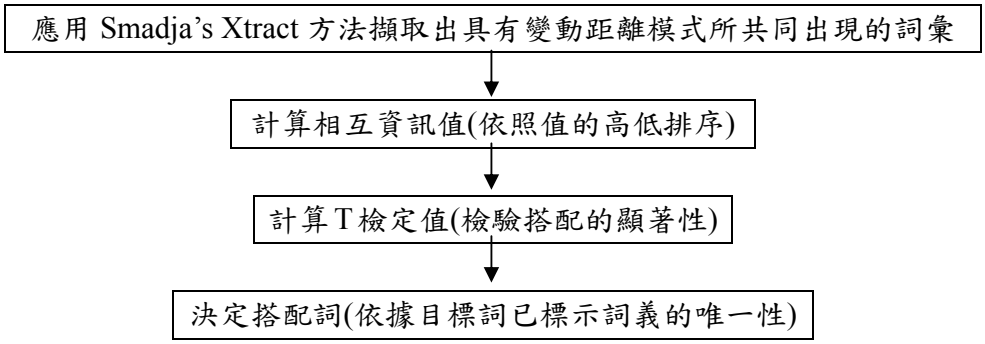
根據統計方法擷取搭配詞的相關文獻中，Smadja's Xtract [1] 採用平均數及變異數的方法於英文的語料中擷取雙連詞，並由雙連詞之結果擴增擷取 n 連詞，此方法被認爲是擷取搭配詞的經典方法。Breidt [4] 將相互資訊值及 T 檢定結合使用於德文的語料中擷取動詞-名詞的搭配詞。在中文的搭配詞研究中，Lu [5] 等人的 CXtract 研究中應用 Smadja's Xtract 的方法於中文語料中，但其研究過程所設置的門檻值會將一些極有可能爲搭配詞的周邊詞過濾掉。將搭配詞應用於其它自然語言處理的相關領域之研究，車方翔 [6] 等用平均數、變異數及 T 檢定的方法得到詞與詞之間搭配強度係數，並將此結果應用於搜索引擎中縮減檢索句子中的歧義度。全昌勤 [7] 等利用搭配詞典的輔助獲取最優種子，再由最優種子自動學習擴充指示詞集，有助詞義辨識之處理。有關詞義辨識的相關研究中，其中以語料爲基礎的監督式學習法是最爲成功的方法，主要是依據上下文的特徵來區別歧義詞，但因上下文共同出現的詞彙數量太多，若全都做爲訓練的樣本會使得雜訊很多，在標示歧義詞時則容易標示錯誤。Li [8] 提出縮小上下文的範圍，使用搭配詞作爲特徵，並且基於搭配詞的歧義詞詞義唯一性的概念，在標示歧義詞時，當上下文擷取到搭配詞時，上下文中其它詞彙的影響性將被減少。國內針對擷取搭配詞的相關研究，主要使用的資源分爲兩大類，第一，將網路視爲具有時間性的語料庫資源，Chen [9] 等人利用網路流量紀錄和 Google 搜尋引擎以擷取搭配詞，Teng [10] 等人利用網路部落格觀察時間性和搭配詞之間的關聯；第二，利用平行語料庫 [11, 12, 13]，根據語言的特徵和統計分析的方法，取得英文的搭配詞結構，進而擷取雙語搭配詞。

有別於過去的研究僅能擷取出詞彙的搭配詞或是固定樣式的搭配詞結構，如動詞與

名詞、形容詞與名詞等。本研究提出考量視窗範圍內周邊詞彙或其詞性之組合，基於 Smadja's Xtract 的演算法和相互資訊值、T 檢定值之統計檢驗的方法，以及大規模中文詞義標示語料庫 SSTC [3] 的輔助，以擷取出雙連詞、三連詞、四連詞之搭配詞。

三、自動擷取搭配資訊方法

本研究所提出之自動擷取搭配資訊處理方法如圖一所示，首先採用 Smadja's Xtract 的演算法 [1, 5]，擷取出詞語間間隔其它詞彙所共同出現的候選搭配詞，接下來採用相互資訊值及 T 檢定值的方式檢驗所擷取出的候選搭配詞在語料庫中共同出現的顯著程度，最後爲搭配詞結果的判定，我們對照中央研究院 SSTC 詞義標示語料庫，若目標詞詞義具有唯一性者，則認定其爲搭配詞。



圖一、自動擷取搭配資訊處理方法之流程圖

（一）擷取具變動距離模式之共現詞彙

首先設定目標詞，設置以句子爲單位，編輯目標詞之周邊詞彙跨距爲 d 的視窗內周邊資訊。在目標詞的 $\pm d$ 跨距內的周邊詞稱作 w_i $(1 \le i \le n$，n 爲所有周邊詞的個數)；設定 w_i 在第 j 個位置 (與目標詞的距離) 出現的次數定義爲 $f_{i,j}$；周邊詞 w_i 在目標詞 $\pm d$ 跨距內總共出現的次數定義爲 $f_i = \sum_{-d}^{d} f_{i,j}$；$f_i$ 的平均次數爲 $f_i = \sum_{-d}^{d} f_{i,j} / 2d$；針對每一個目標詞，平均次數 $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ 和標準差爲 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2}$；周邊詞在目標詞 $\pm d$ 跨距內總共出現的次數經標準化後定義爲 $k_i = \frac{f_i - \overline{f}}{\sigma}$；周邊詞在目標詞 $\pm d$ 跨距內出現次數之變異數定義爲 $U_i = \frac{\sum(f_{i,j} - \overline{f})}{2d}$，表示周邊詞分佈的特徵。爲了過濾不太可能爲搭配詞的組合，設定 (K_0, K_1, U_0) 的經驗門檻值，以下列三個條件 [1] 作爲過濾的依據：

$$C_1 : k_i = \frac{f_i - \overline{f}}{\sigma} \geq K_0 \quad (1)$$

$$C_2 : u_i \geq U_0 \quad (2)$$

$$C_3 : f_{i,j} \geq \overline{f_i} + (K_1 \cdot \sqrt{u_i}) \quad (3)$$

針對上述三個條件判斷分述如下：C_1 條件是衡量周邊詞在目標詞 $\pm d$ 跨距內所出現的次數，過濾掉共現次數太低的周邊詞；C_2 條件是衡量周邊詞在目標詞 $\pm d$ 跨距內各個位置的分佈情形，若周邊詞在各個位置分佈過於分散且次數平均，則將其過濾掉，留下出現次數在各個位置上具有變異性較大的周邊詞。C_3 條件則擷取出周邊詞在目標詞 $\pm d$ 跨距內出現次數較爲突出的位置。並且基於搭配詞必須出現於唯一且固定位置之原由，所以經 Smadja’s Xtract 門檻值過濾後的候選搭配詞，若是針對同一個目標詞，相同周邊詞出現於不同位置者，我們則將此候選搭配詞刪除，認定其不爲搭配詞。

（二）相互資訊值

接著，我們採用衡量兩個事件相關程度的相互資訊值 [14]，其用來表示兩個詞彙間，一個詞出現所帶給另一個詞出現的資訊量。相互資訊值的計算方式如公式 (4):

$$MI(x,y) = \log \frac{P(x,y)}{P(x) \cdot P(y)} \quad (4)$$

經由第一步驟的方法過濾後，我們再計算目標詞彙與周邊詞彙之相互資訊值，並將相互資訊值太低者自搭配候選行列中排除。

（三）T 檢定值

爲了確定搭配詞的顯著程度，我們採用假設檢定中的 T 檢定值 [14] 來檢驗候選搭配詞在語料庫中共現的顯著程度。首先需設定虛無假設：兩個共同出現的詞彙之間互爲獨立，不能形成搭配。T 檢定值的計算方式如公式 (5)，其中 $\overline{x}$ 為樣本平均數；若虛無假設爲真，事件受到伯努力試驗 (Bernoulli trial) 的影響，則平均數 $\mu = p$；變異數 $s^2 = p(1\text{-}p) \approx p$。

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \quad (5)$$

若 T 檢定值大於臨界值，則我們將會拒絕虛無假設，而得出結論：候選搭配詞在語料庫中共同出現是具有顯著性的。假設 T 檢定值小於臨界值，則我們沒有充分證據顯示其爲搭配詞，所以將會過濾掉此候選搭配詞。

（四）決定搭配詞

基於搭配詞的單一詞義特性 [2]，進一步我們利用由中央研究院 SSTC 詞義標示語料庫 [3] 的資源，將前三階段檢驗後的結果，去判斷目標詞與周邊詞在相同位置的結合下，目標詞在已標示詞義語料庫中是否皆爲同一個詞義，若目標詞具有歧義者，我們再作進一步的過濾；由於語言的多樣性，SSTC 詞義語料庫的資源仍屬有限無法概括所有檢驗詞彙，因此，若是目標詞在該語料中未找到標示詞義者，我們也暫時將其過濾。

四、實驗

本節首先介紹我們所使用的語料，以及實驗的設計、參數的設置，實驗結果，並針對實驗結果進行評估。

（一）實驗語料

本研究實驗語料使用《中央研究院語料庫現代漢語平衡語料庫（3.0 版）》，其包括 9,227 篇文件，每個句子已經斷詞處理，且標記詞性，全部約 500 萬詞，語料平衡分佈在不同的媒體（如報紙、學術論文、視聽媒體、演說等）、語式（如書面體、演講稿、會議記錄等）以及主題（科學、哲學、社會、藝術、生活、文學）上，是個適於中文相關處理的代表性語料庫。

（二）實驗設計

本實驗依據圖一的處理方法流程圖之步驟，考量詞彙或詞性的組合，擷取出雙連詞、三連詞、四連詞之搭配詞。擷取搭配詞的範圍設置，跨距指的是目標詞左右距離所跨的範圍，其可依照語料的不同作調整，本實驗因基於搭配詞不跨越標點符號之原則，依據我們的語料中顯示，標點符號與標點符號間的詞彙平均約爲 5.56 個，因此我們將跨距 d 設定爲 5。在擷取三連詞及四連詞的前置處理必須分別先將目標詞 (-5, -1) 和 (1, 5) 跨距內的周邊詞組合成雙連詞 (bigram) 及三連詞 (trigram)，周邊詞可以任意取詞彙或是詞性。

首先採用 Smadja's Xtract [1] 的演算法，在條件過濾時採用的門檻值 (K_0, K_1, U_0) 設定爲 (1, 1, 10)，而在 Lu [5] CXtract 研究中，設定 (K_0, K_1, U_0) 爲 (1.2, 1.2, 12)，但根據我們透過 SSTC 語料的觀察，當參數 (K_0, K_1, U_0) 設定爲 (1.2, 1.2, 12) 時，會過濾掉一些極有可能爲搭配詞的周邊詞。所以，最後我們仍採用 Smadja's Xtract 研究的門檻值 (1, 1, 10)。由於使用 Smadja's Xtract [1] 過濾方法後的周邊詞，仍有許多出現次數極高的停用字 (stopword) ，例如「的、了」。因此，我們再考量其他統計衡量方法，如：相互資訊值、假設檢定 (T 檢定、卡方檢定、對數概似率檢定等)，我們最後採用相互訊息值衡量兩個詞彙間的搭配強度，並依照相互資訊值之大小作排序，可以降低停用字在候選搭配詞的排名，我們將相互資訊值小於 3 者之候選搭配詞排除，而且在我們的實驗中已使用 Smadja's Xtract [1] 的條件 C_1 過濾低頻詞，所以本實驗使用相互資訊值的法，並不會受到相互資訊值容易將語義相近而並非爲搭配的低頻詞組合在一起之缺點的影響。本實驗再利用 T 檢定值檢驗候選搭配詞在語料庫中共同出現是否顯著，而其顯著水準設定爲 $\alpha = 0.005$，若 T 值 < 2.576，則我們沒有充分證據顯示其爲搭配詞，所以將會過濾掉候選搭配組合。表一爲以目標詞「關係」(詞性 Na) 擷取詞彙或詞性之雙連詞

搭配爲例，經過 Smadja's Xtract [1] 的條件過濾及進一步計算相互資訊值、T 值的結果。

表一、經過條件過濾及計算相互資訊值、及 T 值的結果

目標詞	特徵	位置序	Ki 值	Ui 值	目標詞與周邊詞共同出現次數	目標詞在語料出現次數	特徵在語料出現次數	MI 值	T 值
關係 Na	人際	-1	1.96	2732.36	175	2945	246	7.25	13.22
關係 Na	兩岸	-1	1.55	1189.04	117	2945	601	5.95	10.79
關係 Na	密切	-2	1.33	189.84	42	2945	305	5.61	6.46
*關係 Na	沒有	-1	1.60	764.20	95	2945	9775	2.95	9.24
*關係 Na	的	-1	21.79	57239.05	817	2945	296183	1.69	23.33
*關係 Na	A	-1	1.80	438.56	76	2945	31426	1.56	6.89
*關係 Na	T	1	1.47	652.69	88	2945	46697	1.31	6.86
*關係 Na	Nv	-1	5.16	2372.21	178	2945	95563	1.30	9.71
*關係 Na	也	1	1.11	117.61	34	2945	29759	0.81	3.24
*關係 Na	Nep	-1	2.68	395.65	61	2945	67325	0.58	3.44
*關係 Na	Na	-1	38.68	52626.69	858	2945	1025464	0.50	11.55
*關係 Na	個	-4	1.21	66.64	27	2945	41143	0.26	1.18
*關係 Na	VK	-5	1.13	85.49	31	2945	57286	0.07	0.35

(*代表 MI 值或 T 值低於門檻值)

最後判定搭配詞，我們將 SSTC 詞義標示語料的資源，隨機取 4/5 爲訓練資料，1/5 爲測試資料，在此以表二目標詞「關係」(詞性 Na) 擷取搭配詞之雙連詞、三連詞、四連詞之部分結果爲例，表二中的搭配組合「人際 關係」在 SSTC 詞義標示語料裡的訓練資料中佔有 3 筆，且目標詞「關係」皆相同詞義，我們判定其爲搭配詞；「兩岸 關係」因在訓練資料中，缺乏已標示資源，我們無法判定其是否爲搭配詞，在此，也將其暫時排除；而在搭配組合「與 Na DE 關係」在訓練資料中，目標詞「關係」有 2 個詞義（具有歧義），則我們判定其不爲搭配詞。

（三）實驗結果

本實驗以目標詞「關係 Na」、「好 VH」、「看 VC」、「講 VE」、「說 VE」爲例，依據 SSTC 詞義標示語料裡的訓練資料，擷取出詞彙或詞性結合之搭配詞完整結果置於附錄。部分結果如表三所示。

表二、由 SSTC 詞義標示語料判定是否爲搭配詞

目標詞	特徵	位置序	目標詞詞義	筆數	*是否爲搭配詞
關係 Na	人際	-1	1.普通名詞。人和人之間在社會或群體中的關聯性。	3	∨
	兩岸	-1	無標示資源		?
	密切	-2	1.普通名詞。事件之間的關聯性。	1	∨
	VC_人際	-2	1.普通名詞。人和人之間在社會或群體中的關聯性。	1	∨
	也_沒有	-2	1.普通名詞。事件之間的關聯性。	6	∨
	因_Na	-2	1.普通名詞。事件發生的原因。	2	∨
	很_大	-3	1.普通名詞。事件之間的關聯性。	1	∨
	密切_DE	-2	1.普通名詞。事件之間的關聯性。	1	∨
	Caa_Na_DE	-3	1.普通名詞。事件之間的關聯性。	2	∨
	Dfa_大_的	-3	1.普通名詞。事件之間的關聯性。	3	∨
	很_大_的	-3	1.普通名詞。事件之間的關聯性。	1	∨
	與_Na_DE	-3	1.普通名詞。人和人之間在社會或群體中的關聯性。	1	×
			2.普通名詞。事件之間的關聯性。	2	

*「是否爲搭配詞」欄位－標記 ∨ 代表經由我們的實驗結果判定爲搭配詞；
標記 × 代表經由我們的實驗結果判定爲不是搭配詞；
標記 ？ 代表因語料詞義標示資源不足，所以無法判別。

表三、依據 SSTC 詞義標示語料裡的訓練資料，擷取搭配詞之部分結果

目標詞	搭配詞
關係 Na	人際 關係、VC 人際 關係、密切 DE 關係、Caa Na DE 關係
好 VH	更 好、心情 D 好、Cbb Na Dfa 好、D 有 多 好、好 DE 方法、做 DE D 好
看 VC	看 書、看 電視、看 Di Nb DE、Nh 來 看、看 Di Nh Neu、看 著 Na Ncd
講 VE	聽 Nh 講、講 DE SHI Na、對 我 D 講、講 得 很 VH、D 跟 Nh 講
說 VE	說 不 出 話、跟 Nh 講 說、說 Di Neu 句、他 笑 Di 說、Nh 告訴 Nh 說

（四）評估

本實驗從 SSTC 詞義標示語料隨機取出 1/5 作爲測試資料，並且採用應用率及正確率作爲評估的準則，公式如(6)：

$$\text{應用率}=\frac{\text{標上的筆數}}{\text{測試資料中包含目標詞的筆數}}\times 100\% \quad \text{正確率}=\frac{\text{正確的筆數}}{\text{標上的筆數}}\times 100\% \qquad (6)$$

我們從訓練資料中利用搭配詞的約束力區別目標多義詞之詞義，並將上述判定爲搭配詞的結果及目標詞詞義標示於測試資料中，實驗結果如表四所示，在 543 個測試句中標出 109 個句子，其中有 99 個標示結果爲正確標示，因此總計達到 20.07%的應用率及 90.83%的正確率。

表四　以搭配知識進行詞義自動標示之實驗結果

詞彙	詞性	測試句數	標示句數	正確句數	應用率	正確率
好	VH	119	19	19	15.97%	100.00%
看	VC	121	17	11	14.05%	64.71%
說	VE	206	34	33	16.50%	97.06%
講	VE	75	32	31	42.67%	96.88%
關係	Na	22	7	5	31.82%	71.43%
		543	109	99	20.07%	90.83%

五、結論

本研究擷取詞彙或詞性組合之搭配詞。首先爲了擷取雙連詞之搭配詞，利用 Smadja's Xtract 的平均數及變異數的方法擷取被其它詞彙間隔之共同出現的詞彙資訊，進而計算相互資訊值，依照搭配強度高低排序，再由 T 檢定值檢驗候選搭配詞的顯著性。我們將這三種方法結合使用，若在這三種方法下通過考驗的候選搭配詞，我們再經由對照中央研究院人工標示詞義語料庫之目標詞的結果，若在同樣目標詞和周邊詞的跨距下，目標詞均爲同一個詞義者，則我們就判定其確實爲搭配詞；同樣，我們也以同樣的方式擷取三連詞及四連詞之搭配詞。最後，我們將擷取出的搭配詞資訊應用於語義辨識，達到 20.07%的應用率及 90.83%的正確率。

參考文獻

[1] F. Smadja, "Retrieving Collocation From Text: Xtract," *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177, 1993.

[2] D. Yarowsky, "One Sense Per Collocation," In *Proceedings of the ARPA Human* Language Technology Workshop, 1993, pp. 266-271.

[3] 柯淑津、黃居仁、洪嘉馡、劉詩音、簡卉伶、蘇依莉，「中文詞義全文標記語料庫之設計與雛形製作」，In ROCLING 2007.

[4] E. Breidt, "Extraction of V-N-Collocations from Text Corpora: A feasibility Study for German," In *Proceedings of the First Workshop on Very Large Corpora*, 1993, pp.

74-83.

[5] Q. Lu, Y. Li and R. F. Xu, “Improving Xtract for Chinese Collocation Extraction,” In *IEEE 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 2003, pp. 333-338.

[6] 車方翔、劉挺、秦兵、李生，「面向依存文法分析的搭配抽取方法研究」，哈爾濱工業大學信息檢索研究室論文集， 第一卷，2003。

[7] 全昌勤、何婷婷、姬東鴻、劉輝，「從搭配知識獲取最優種子的詞義消歧方法」，中文信息學報，第十九卷，第一期，2005，第 30-37 頁。

[8] W. Li, Q. Lu and W. Li, “Integrating Collocation Features in Chinese Word Sense Disambiguation,” In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005, pp. 87-94.

[9] H.-H. Chen, Y.-C. Yu, and C.-L. Li, “Collocation Extraction Using Web Statistics,” In *Proceedings of 4th International Conference on Language, Resources and Evaluation*, 2004, pp. 1851-1854.

[10] C.-Y. Teng and H.-H. Chen, "Analyzing Temporal Collocations in Weblogs." In *Proceedings of International Conference on Weblogs and Social Media*, 2007, pp. 303-304.

[11] C.-C. Wu and J. S. Chang, “Bilingual collocation extraction based on syntactic and statistical analyses,” *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 1 , 2004, pp. 1-20.

[12] J.-Y. Jian, Y.-C. Chang and J. S. Chang, “TANGO: bilingual collocational concordancer,” In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 2004, pp.166-169.

[13] J.-Y. Jian, Y.-C. Chang and J. S. Chang, “Collocational Translation Memory Extraction Based on Statistical and Linguistic Information,” In ROCLING 2004.

[14] C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. The MIT Press, 1999.

附錄

依據 SSTC 詞義標示語料裡的訓練資料，擷取搭配詞之完整結果

目標詞	目標詞 詞義	搭配詞	採用 的方法
關係 Na	普通名詞。人和人之間在社會或群體中的關聯性。	人際 關係	雙連詞
		VC 人際 關係	三連詞
		P Nh V_2 × 關係、P Nh 有 × 關係、 與 Nc DE 關係、與 Nc 的 關係	四連詞
	普通名詞。事件之間的關聯性。	密切 × 關係	雙連詞
		VJ 什麼 關係、V_2 Dfa × × 關係、V_2 Nep 關係、 V_2 什麼 關係、V_2 很 × × 關係、也 沒有 關係、 大 DE 關係、大 的 關係、有 Dfa × × 關係、 有 Nep 關係、有 什麼 關係、有 很 × × 關係、 很 大 × 關係、密切 DE 關係、密切 的 關係	三連詞
		Caa Na DE 關係、Caa Na 的 關係、Dfa 大 DE 關係、 Dfa 大 的 關係、Ng VH DE 關係、Ng VH 的 關係、 V_2 Dfa VH × 關係、V_2 Dfa 大 × 關係、 V_2 很 VH × 關係、V_2 很大 × 關係、 之間 VH DE 關係、之間 VH 的 關係、 有 Dfa VH × 關係、有 Dfa 大 × 關係、 有 很 VH × 關係、有 很 大 × 關係、 和 Na DE 關係、和 Na 的 關係、很 VH DE 關係、 很 VH 的 關係、很 大 DE 關係、很 大 的 關係	四連詞
	普通名詞。事件發生的原因。	因 Na 關係	三連詞
		Cbb Na DE 關係、Cbb Na 的 關係	四連詞
看 VC	監視管理。	看 DE 很 VH、看 得 很 VH	四連詞
	用眼睛察覺。	看 Di Nd、Ncd 一 看、VA 一 看、看 了 Nb、 看 了 Nd、再 D 看、看 著 Nb、看 著 Nc、 看 著 Nep、看 著 Nh、看 著 我	三連詞
		看 Di Na Ncd、Nh VA 一 看、看 著 Na Ncd	四連詞
	透過視覺來理解或欣賞。	看 書、看 電視	雙連詞
		看 Di D、VC 給 × 看、看 了 D、看 不 VC、 坐 P × × 看、坐 在 × × 看	三連詞
		DE Na 去 看、看 Di Nb DE、看 Di Nb 的、 看 Di Nh DE、看 Di Nh 的、VC 給 Nh 看、 的 Na 去 看	四連詞
	以特定態度對待。	看 Di Nh Neu	四連詞
	拜訪、探望後述對象。	Nh 來 看、到 Nc D 看、到 Nc 去 看	三連詞
	病人接受診治。	看 Di Neqa	三連詞

好 VH	形容對特定對象的正面評價。	更 好	雙連詞
		DE 不 好、好 DE 方法、DE 很 好、DE 最 好、 D 太 好、D 比較 好、D 更 好、Dfa D 好、 Dfa 不 好、Na 真 好、Na 最 好、Nep Dfa 好、 Nep 不 好、Nf Dfa 好、Nf 很 好、SHI 最 好、 VC 得 × 好、VJ Dfa 好、V_2 更 好、不 Dfa 好、 不 太 好、什麼 D 好、什麼 不 好、心情 D 好、 心情 不 好、有 Dfa 好、有 多 好、有 更 好、 好 的 方法、的 很 好、的 最 好、是 最 好、 個 Dfa 好、個 很 好、做 DE × 好、做 得 × 好、 得 D 好、得 不 好、得 很 好、都 Dfa 好、就 不 好、 會 Dfa 好、會 比較 好	三連詞
		Cbb Na Dfa 好、DE Na 不 好、好 DE Nv 方式、 D D 太 好、D D 很 好、D Dfa D 好、D Dfa 不 好、 D SHI 最 好、D VC 得 × 好、D VJ Dfa 好、 D V_2 更 好、D 不 Dfa 好、D 不 太 好、D 有 Dfa 好、 D 有 多 好、D 有 更 好、D 很 D 好、D 很 不 好、 D 是 最 好、D 做 DE × 好、D 做 得 × 好、 Na DE Dfa 好、Na DE 最 好、Na Dfa D 好、Na Dfa 不 好、 Na SHI Dfa 好、Na SHI 很 好、Na SHI 最 好、 Na VC 得 × 好、Na 不 SHI × 好、Na_不_是 × 好、Na 的 Dfa 好、Na 的 最 好、Na 是 Dfa 好、 Na 是 很 好、Na 是 最 好、Nf Na Dfa 好、 Nh VC 得 × 好、Nh VK Dfa 好、Nh 覺得 Dfa 好、VC DE D 好、VC DE 不 好、VC DE 很 好、 VC 的 很 好、VC 得 D 好、VC 得 Dfa 好、 VC 得 不 好、VC 得 很 好、VE DE D 好、 VJ Nep D 好、VJ Nep 不 好、VJ 什麼 D 好、 VJ 什麼 不 好、的 Na 不 好、好 的 Nv 方式、 個 Na Dfa 好、做 DE D 好、做 DE Dfa 好、 做 DE 很 好、做 得 D 好、做 得 Dfa 好、 做 得 很 好、該 有 Dfa 好、該 有 多 好、 說 DE Dfa 好、Na D Dfa 好、Na D 很 好、 P Nh Dfa 好、P Nh 不 好、P Nh 很 好、P 我 D 好、 P 我 Dfa 好、對 Nh D 好、對 Nh Dfa 好、對 Nh 很 好	四連詞
	表示同意或允許。	DE Na 好 × 好、DE VH 不 好、DE 好 不 好、 D VH D 好、Nh VH D 好、VA VH D 好、 VA VH 不 好、VA_好_D、VA 好 不 好、 VE VH D 好、VE VH 不 好、VE 好 D 好、 VE 好 不 好、的 Na 好 × 好	四連詞
	形容態度親切的。	好 DE 朋友、好 的 朋友、對 我 × 好	三連詞
	表示結束前一個話題,開始新的話題。	VE 得 Dfa 好	四連詞

	問候語,常用於對話的開場。	會 D 好	三連詞
講 VE	以口語媒介引述或陳述訊息。	講 DE D、講 DE Nep、講 DE 很、講 DE 都、 講 DE 話、D 我 × × 講、D 這樣 講、D 跟 × 講、Na 跟 × 講、講 × Nf 話、Nh P × 講、Nh 剛剛 講、Nh 跟 × 講、Nh 聽 × 講、P 她 講、P 你 講、 講 T T、講 了 一、她 D 講、你 剛剛 講、我 P × 講、我跟 × 講、講 的 D、講 的 Dfa、講 的 Nep、 講 的 都、講 的 話、講 得 很、就 P × 講、 就 跟 × 講、跟 Na 講、跟 你 講、跟 我 講、 講 說 Nh、聽 Nh 講	三連詞
		講 DE D VH、講 DE Nep Nf、講 DE 很 VH、 D D 跟 × 講、D P Nh 講、D P 他 講、D P 我 講、 D 跟 Na 講、D 跟 Nh 講、D 跟 他 講、 D 跟 我 講、講 Di Neu 個、講 Di 一 個、 Na 跟 Nh 講、Nh D P × 講、Nh D 跟 × 講、 Nh Na D 講、Nh P Nh 講、Nh P 你 講、Nh P 我 講、Nh VE Nh 講、Nh 就 P × 講、Nh 就 跟 × 講、 Nh 跟 Nh 講、Nh 跟 你 講、Nh 跟 我 講、 講 一 Nf 話、講 了 Neu 個、講 了 一 Nf、 我 D P × 講、我 D 跟 × 講、我 P Nh 講、 我 P 你 講、我 跟 Nh 講、我 跟 你 講、 講 的 Dfa VH、講 的 Nep Nf、講 得 很 VH、 就 P Nh 講、就 跟 Nh 講	四連詞
	以文字媒介引述或陳述訊息。	Na 上 D 講	四連詞
	描述後述內容。	講 DE SHI、講 DE 是、講 的 SHI、講 的 是	三連詞
		講 DE SHI Na、講 DE 是 Na、講 的 SHI Na、 講 的 是 Na	四連詞
	評價後述對象。	Na 來 講、Nc 來 講、Nh 來 講、我_來 講、 對 Nh × 講、對 我 × 講	三連詞
		P Nh D 講、P Nh 來 講、P 我 D 講、P 我 來 講、 對 Nh D 講、對 Nh 來 講、對 我 D 講、 對 我 來 講	四連詞
說 VE	以口語媒介引述或陳述訊息。	來 說	雙連詞

		說 DE 話、DE 對 × 說、說 D 出、D 這麼 說、 D 跟 × 說、Na 常 說、Nb 就 說、Nh 跟 × 說、Nh 講 說、SHI 覺得 說、VA 著 說、VH 地 說、V_2 人 說、V_2 話 × 說、說 不 出、他 就 說、 有 人 說、有 話 × 說、告訴 我 說、我 跟 × 說、 說 的 話、的 對 × 說、是 覺得 說、笑 Di 說、 笑 著 說、高興 DE 說、問 Nh 說、問 他 說、 媽媽 D 說、話 D 說、跟 Nh 說、跟 他 說、 跟 你 說、跟 我 說、對 Nh 說、對 他 說、聽 Na 說	三連詞
		DE P Nh 說、DE P 他 說、說 DE VH Dfb、 DE 對 Nh 說、D D 這麼 說、D D 跟 × 說、 D P 你 說、D VA 著 說、說 D VE 話、 D V_2 人 說、說 D 出 Na、說 D 出 話、 D 有 人 說、D 笑 Di 說、D 跟 Na 說、 D 跟 Nh 說、D 跟 你 說、D 對 Nh 說、 說 Di Neu 句、說 Di 一 句、Na D 跟 × 說、 Na P 我 說、Na 跟 Nh 說、Na 對 Nh 說、 Na 對 我 說、Nb VH DE 說、Nb VH 的 說、 Nh D 這麼 說、Nh D 對 × 說、Nh P 你 說、 Nh SHI VE 說、Nh SHI VK 說、Nh SHI 覺得 說、 Nh VA 著 說、Nh 告訴 Nh 說、Nh 是 VE 說、 Nh 是 VK 說、Nh 是 覺得 說、Nh 笑 Di 說、 Nh 笑 著 說、Nh 跟 Nh 說、Nh 跟 你 說、 Nh 跟 我 說、P Nh D 說、P Nh 講 說、P 他 講 說、 P 你 D 說、P 我 講 說、說 SHI 這 個、 說 × VE Na T、說 × VE Na 來、說 × VE 話 T、 說 × VE 話 來、VH DE 對 × 說、 VH 的 對 × 說、V_2 話 D 說、也 V_2 人 說、 也 有 人 說、說 不 VE 話、說 不 出 Na、說 不 出 話、 他 VA Di 說、他 VA 著 說、他 VE Nh 說、 他 VE 我 說、他 笑 Di 說、他 笑 著 說、 說 × 出 Na T、說 × 出 Na 來、說 × 出 話 T、 說 × 出 話 來、有 話 D 說、我 D 對 × 說、 我 P Nh 說 、我 P 你 說、我 SHI VE 說、 我 SHI VK 說、我 SHI 覺得、我 是 VE 說、 我 是 VK 說、我 是 覺得 說、我 跟 Nh 說、 我 跟 你 說、的 P Nh 說、的 對 Nh 說、 很 VK DE 說、很 VK 的 說、說 是 這 個、 說 得 VH Dfb、就 VE Nh 說、話 D D 說、 跟 Nh Na 說、跟 Nh VE 說、跟 Nh 講 說、 跟 他 VE 說、跟 他 講 說、跟 我 VE 說、 跟 我 講 說	四連詞
	以口語進行打招呼、道謝、拒絕等言談行爲。	Caa SHI 說、Caa 是 說、或者 SHI 說、或者 是 說	三連詞
		Nh D 覺得 說、Nh 就 VK 說、Nh 就 覺得 說、 我 D 覺得 說	四連詞

以部落格語料進行情緒趨勢分析

楊昌樺　高虹安　陳信希

國立台灣大學資訊工程學系

{d91013, r95116, hhchen}@csie.ntu.edu.tw

摘要

部落格提供大量具有時間標記的文本，為語言處理所需豐富語料來源。本文針對文本的時間標記特性，將其切分成不同時間域(Time Domain)的語料子集合，綜合個別時間域所提供的語料，觀察目標觀點(sentiment，通常包含意見與情緒)在橫跨時間域的變化，作為觀點趨勢分析的基礎。為了獲得不同時間域的部落格文本，本研究提出部落格資訊系統，收集跨時間域的文本。同時以情緒分析為例，以特定查詢在部落格資訊系統反饋的相關文本，獲得各時間域的情緒特徵，藉以解讀網路空間人們對特定議題所反應的情緒變化。

1. 緒論

近年來，全球資訊網(World Wide Web; Web)上各種資訊系統的發明，持續改變人們對資訊吸收與處理的方式。以 2005 年作為一個概括的分水嶺，分析使用者的習性，可發現：2005 年之前，人們從接觸網路，到逐漸習慣閱讀 Web 上的內容，包含新聞報導、旅遊情報、生活資訊、投資消息、工作機會等。2005 年之後，人們開始廣泛地創造 Web 上的內容，包含製作自己部落格、相簿遊記、影音紀錄等。Web 服務提供者，為了能滿足前者的閱讀需求，激發了內容網站(如 nyyimes.com、cnn.com)、與服務網站(包括入口網站，如 Yahoo![1]，及搜尋網站，如 Google[2])的興起。而為了能涵蓋後者的創造需求，也帶

[1] http://www.yahoo.com/

[2] http://www.google.com/

動了部落格網站(如 Blogger[3])、相片網站(如無名小站[4])、影音網站(如 YouTube[5])等的蓬勃發展。以媒體的觀點分析之，前者透過特定企業將 Web 視為大量資訊的媒介，提供的站台延續大眾媒體的角色。後者藉由 Web 使用者社群參與，創造新式資訊發佈型態，通稱為社群媒體。

近來社群媒體所創造的資源吸引很多學者的注意，本文針對部落格(或稱網路日誌、Weblog、Blog)所提供的文本，進行語言處理方面的探討。部落格系統提供簡單的介面，讓使用者發表具時間標記的文章，因此有越來越多的人們開始使用部落格在網路上分享每天的生活經驗、發表對事物的看法與心情。根據部落格搜尋引擎Technorati[6]的報告指出，全球部落格的數量已超過7,000萬個，並且平均每天有超過12萬個部落格成立，因此整個部落格空間(Blogosphere)每天所能貢獻出的新文本更在此數量之上。這份報告中同時也指出，目前部落格空間以日文及英文使用者居多，各佔37%及36%，而中文目前所佔比例是8%，但有增長的趨勢，本文即以中文部落格文本為主要的研究對象。

在社群媒體的框架下，人們在使用部落格搜尋引擎時，不但想找到較專業或具代表性的部落格，同時也想找到一般使用者所提出的心得及想法。TREC自2006年開始舉辦Blog Track[7](Macdonald, de Rijke, Mishne, and Soboroff, 2006)，其競賽項目說明了使用者上述的資訊需求。其中Opinion Retrieval Task是針對特定議題找出使用者表達意見的文章，並判斷該意見文章的正負面傾向。另外，Blog Distillation (Feed Search) Task是找出持續對某特定議題關注的部落格。舉例來說，使用者可能想在一個著名的歌唱大賽結束後，瀏覽各部落格最新發表的相關文章，並挑有興趣的閱讀。使用者也可能剛接觸古典音樂，想找專門討論古典音樂的部落格，並在肯定某部落格的豐富內容後，持續訂閱該部落格。

---

[3] http://www.blogger.com/

[4] http://www.wretch.cc/

[5] http://www.youtube.com/

[6] http://www.sifry.com/alerts/archives/000493.html

[7] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

相較於從Web語料判斷使用者的意見和情緒(Ku and Chen, 2007; Lin, Yang, and Chen, 2007)，從部落格文本中挖掘使用者觀點有如下的好處：個人化導向和時間性。部落格是個人在網路世界最直接的發聲工具，使用者經認證登錄後，就可以特定的虛擬身分進行發布文本的行為。相較之下，Web是零零總總資訊的巨大集合，所提供的各式文本在結構或型態上都可能不一致。當Web網頁一經發布後，除非進行更動或刪除，就會一直存在於網路空間裡，儘管能透過搜尋引擎取回文本進行判斷，也較無機制獲得網頁發布或更動的時間。相較之下，部落格文本通常具有時間標記，在呈現上也是依照時間從最近到最舊排列，因此透過部落格收集使用者觀點，除了能根據所有的文本進行分析外，也有機會分析出不同時間域(如最近、上周、去年)使用者的特徵。

儘管由如上所述的特點，透過部落格擷取使用者觀點，仍存在一些語料分析時必須考量的議題，例如嚴(2007)曾指出，部落格文本中有一半的文章是來自轉錄，而不是部落格作者自己撰寫。這些轉錄文章的內容大部份是從新聞網站、或一般的官方網站，經由複製、轉貼到使用者自己部落格上。這是因為在便利的網路環境下，人們很容易取得其他資訊，也很方便再把資訊傳播出去，因此傳播出去的不見得就是該使用者本身的立場或意見。在此情況下，系統可能收集到不同於使用者觀點的雜訊。

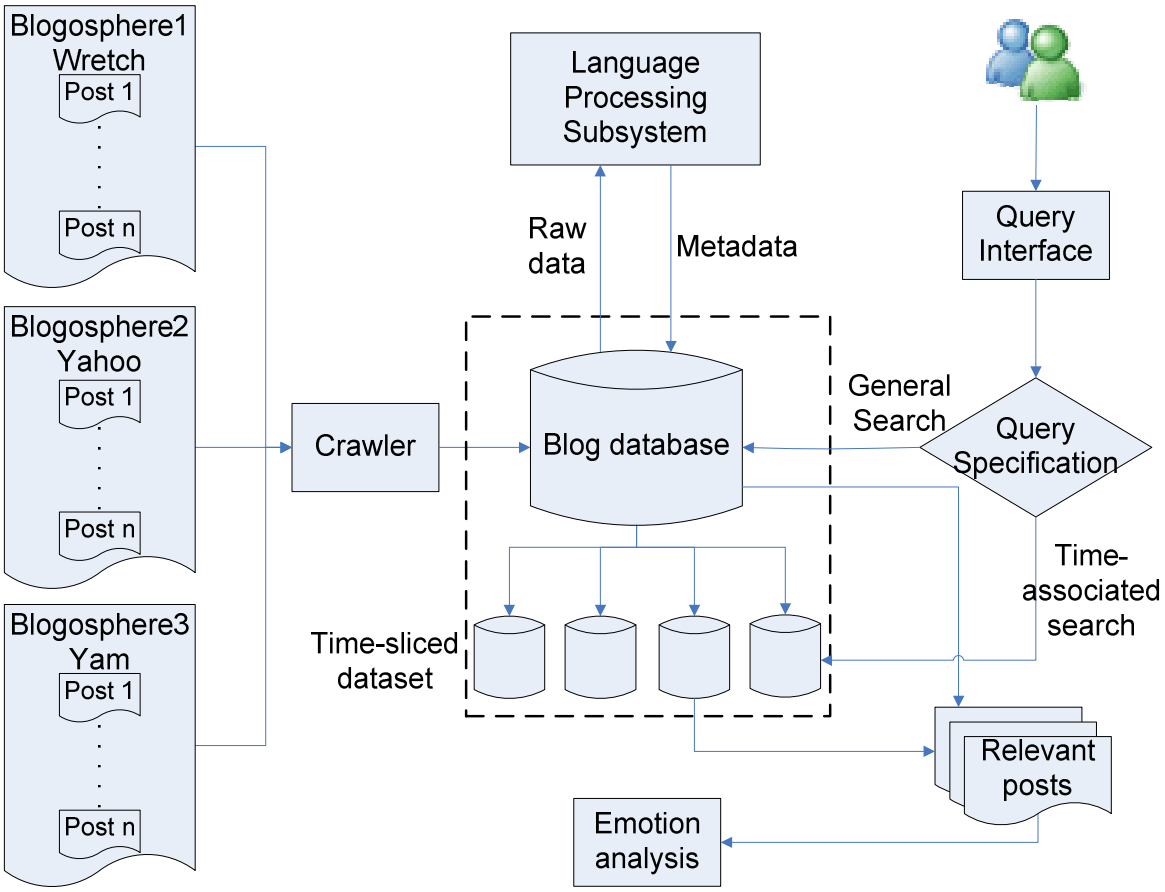
雖然透過部落格容易達成溝通，但相對來說，也較容易產生衝突。由於在部落格世界中，人們的身分通常是由一個虛擬的ID或暱稱所代表，因此在與社群意見不一致時，較容易產生加油添醋、謾罵，甚至言語攻擊等行為。倡議Web 2.0的Tim O'Reilly在2007年曾提出一個「部落客行為守則」 (Blogger's Code of Conduct[8])，希望部落客能自我維護言論自由的環境。然而從語料分析系統的觀點，無論是好或是壞的言論，都是一種文本的呈現，會一視同仁地收集回來，但在日後分析處理時，是否能過濾掉部落客不良動機所帶來的雜訊，也是一個需要重視的議題。

本文嘗試在不同時間域上對部落客情緒進行分析，建置部落格資訊系統，以獲取具

[8] http://radar.oreilly.com/archives/2007/03/call_for_a_blog_1.html

時間特性的文本語料。內容安排如下：第二節列出系統架構，第三、四節針對系統兩個重要核心－文本收集與整合、情緒分析與趨勢－加以說明，第五節是舉出一些應用範例，並在第六節提出結論。

2. 部落格資訊系統架構



圖一、部落格情緒趨勢分析系統架構圖

本研究所探討的部落格情緒趨勢分析系統包括四部分：收集與整合部落格文本資訊、資料庫與語言處理子系統、使用者查詢流程、以及情緒分析模組。第一部分及第四部分將於第三、四節再作詳細探討。

關於資料庫與語言處理子系統部分，在 Crawler 經由查訪部落格文本單位後，會先剖

析出部落格欄位資訊，將其存放在對應的部落格資料庫(Blog Database)。另外，各欄位所包含的語言資訊，如須經斷詞處理以獲得進一步索引時，我們的語言處理子系統(Language Processing System)會利用 Stanford Natural Language Processing Group[9] 所開發的 Stanford Chinese Word Segmenter[10]加以處理，並回傳至原資料庫，以 Metadata 的型式協助往後搜尋的進行。

部落格資料庫本身可透過文本的時間資訊，在實體資料集(Physical Dataset)上產生不同時間域子集合(Time-sliced Dataset)，上述兩種模式的資料集可支援以下不同的查詢應用：第一種是基本查詢(General 查詢)，例如於使用者介面可以提供搜尋系統，讓使用者鍵入擬查詢的關鍵字後，由系統回傳相關的部落格文本資訊，使用者可以瀏覽分析部落格文本資訊後，透過永久網址(Permalink)鏈結到該部落格文本原先的網頁。第二種資料集可支援時間相關的查詢(Time-associated Search)，例如透過關鍵字查詢相關文本在不同時間域的情緒呈現趨勢，系統可依照時間域與情緒分類繪製情緒波動的列表或趨勢圖。

3. 部落格文本資訊之收集與整合

部落格文本位於不同的部落格伺服器，這些伺服器可能是使用者自行架設、委託代管之程式或機器，或是由企業廠商所提供之服務平台。基於語料平衡性與完整性的考量，一個部落格資訊系統需要能收集不同來源的部落格文本，並能將各式文本整合成為一個資料集合。如同收集 Web 的材料時所面臨到的 Deep Web (He et al., 2007)問題，這裡我們也相對綜合出一個 Deep Blogosphere 的概念，與探索 Deep Blogosphere 的因應之道。

首先在剖析 Blogosphere 的結構後可了解到，每一篇部落格文本都有所謂的 Permalink，一個完整 Deep Blogosphere 的文本探索，簡言之就是能拜訪過世界上所有的 Permalink。然而這些 Permalink 大部分都由各部落格伺服器的資料庫所維護，其網址通常不代表一個真正的實體網頁，而根據其網址所瀏覽到網頁上的鏈結(Hyperlink)，大部分也

[9] http://nlp.stanford.edu/

[10] http://nlp.stanford.edu/software/segmenter.shtml

是由各部落格伺服器自動產生，如導引首頁鏈結、圖片鏈結、廣告鏈結、作者資訊鏈結、社群推薦鏈結等。

透過這樣的剖析，可以發現透過傳統 Hyperlink 樹狀拜訪所有網頁結點模式，將難以探索完所有 Deep Blogosphere 的文本。解決之道是在文本之上先維護一個有關所有部落客(Bloggers)的人口普查，意即如果能夠知道世界上所有 Bloggers 的列表，再定時查訪各 Bloggers 所發表的最新文本，以獲得新 Permalink，便能保證在一個實行時間點之後，能夠收集到所有部落格文章。

為了實驗上述概念，本研究將 Deep Blogoshpere 的範圍簡化，將「Yahoo!奇摩部落格」、「無名小站網誌」、「yam 天空部落」視為三個虛擬的 Bloggers，並從六月份開始定期查訪該三個"Bloggers"的最新文章。我們選自 6/20 至 7/8 為止收集到的 322,792 筆部落格文本資訊，供本研究分析。虛擬 Blogger 文章數，及相關統計資料如表一所示。其中「無名小站網誌」每天所能查訪的部落格文本資訊數為最多，接近 1 萬筆，「yam 天空部落」最少，僅約 250 筆。

表一、部落格文本資訊查訪統計

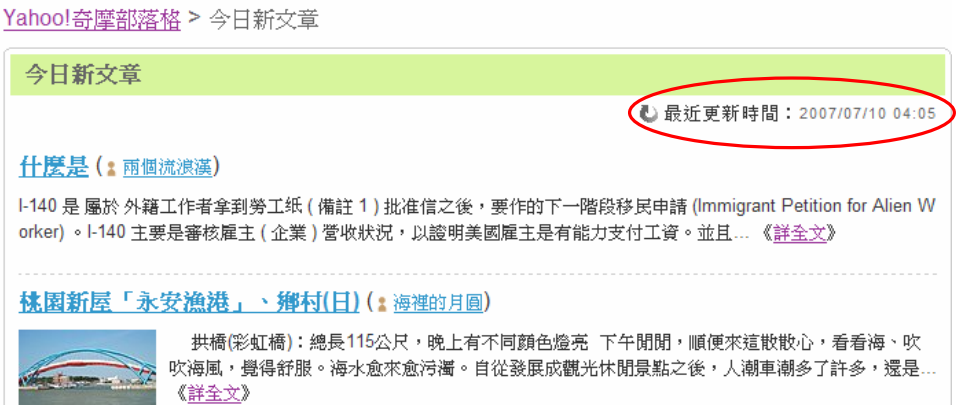
虛擬部落客	文本分析數	每日平均	查訪最新文章方式
Yahoo!奇摩部落格	132,661	6,982	RSS、動態網頁
無名小站網誌	185,234	9,749	Ping Server
yam 天空部落	4,897	258	動態網頁(首頁)

為了查訪Bloggers的最新文本資訊，得因應各部落格伺服器不同的特性，本研究歸納出兩種不同的方式來進行部落格文本資訊的收集，以下分述之。

3.1 查訪部落格網頁列表或 RSS

表一列出各虛擬部落客最新文章查訪方式，以「Yahoo!奇摩部落格」為例，其在首頁「最新文章」區塊提供如圖二紅圈處所標示的RSS按鈕，透過此RSS按鈕可導引到特定的URL[11]，RSS (Resource Description Framework)是部落格空間常用作為資料交換的XML格式規範，本例的URL即以XML的形式列出100篇最新文章。另外該伺服器亦提供「今日新文章」動態網頁[12]，如圖三所示，同樣也可作為查訪部落格文本資訊之依據。然而此網頁較前者不同之處在於部落格文本涵蓋率，以及即時性的比較上。因為其列表列出當天凌晨之前所發布的所有文章，故在部落格文本資訊的涵蓋率上，會較前者完整許多，然而其在即時性上的表現上較前者來得差，原因在於其文章列表僅在當天凌晨更新一次，如圖三紅圈標示處所示。

圖二、Yahoo!奇摩部落格之「最新文章」區塊



圖三、Yahoo!奇摩部落格之「今日新文章」網頁

[11] http://tw.blog.yahoo.com/rss/newarticle.xml

[12] http://tw.blog.yahoo.com/newarticle/newarticle.php

類似的方法亦適用於「yam天空部落」。「yam天空部落」首頁[13]即有一區塊列出新發布文章，例如可每隔十分鐘查訪其首頁，對此區塊作剖析以收集部落格文本資訊。

3.2 利用 Ping Server 即時更新文章列表

前一小節所述的RSS規範在部落格空間產生出一種推播(push)的資料流向，因此衍生出一種所謂Ping的資料交換方式，例如當部落客發布一篇新文章時，部落格常會提供自動Ping一個或多個伺服器的服務，亦即發送一個XML-RPC(遠端程序呼叫）信號給一個或多個所謂的 “Ping Servers”。這些“Ping Servers”會藉由收到的信號來產生一個列表，列出有新文本發布或更動的部落格網址。目前開放的Ping Server像是VeriSign公司的Weblogs.com[14]或是Yahoo!公司blo.gs[15]，皆允許網路服務者訂閱其部落格列表，例如部落格搜尋引擎可以藉由查訪最新更新的部落格，來提供使用者較新的搜尋結果。

本研究利用Ping Server的概念，選擇Weblogs.com所提供的“weblog change list”[16]作為部落格文本資訊收集之依據。該列表列出最近五分鐘更新的部落格網址，經由隨機取樣調查發現，中文語料的部落格文本以「無名小站網誌」數量最多，探討原因應為「無名小站網誌」提供自動Ping到Weblogs.com的功能，因此可以預期在透過定時下載此列表的方式，從各「最近五分鐘更新」的部落格文本資訊，累積成從特定時間點之後「所有」的部落格文本資訊。

4. 情緒分析模組

以部落格文本作為語料，Mishne (2005)使用Livejournal[17]標記文本發表時心情的情緒符號，訓練Support Vector Machine (SVM; Cortes and Vapnik, 1995)在文本層次的心情分類

[13] http://blog.yam.com/

[14] http://Web.weblogs.com/

[15] http://blo.gs/

[16] http://rpc.weblogs.com/shortChanges.xml

[17] http://www.livejournal.com/

器。Mishne (2006)更進一步在所有觀察的時間域上，以圖型闡釋部落格世界整體的心情指數，例如在情人節前後感受到愛(Love)，在夜間有喝醉(Drunk)的感受。楊和陳(2006)則收集部落格中帶有表情符號的句子，訓練SVM在句子層次的情緒分類器。在以上所述的研究中，部落格文本因在網頁呈現方式上所特別能包含的情緒符號，皆用來當作心情或情緒標記(Taggings)，而文本或文句所包含的關鍵字則形成了所謂的特徵值(Features)。Yang等人(2007)使用了更大規模代有表情符號的部落格語料，進行自動化情緒字典的抽取。本研究參考其情緒字典抽取方式，將文句中具有情緒詞彙的出現作為成特徵值，訓練出包含喜(joy)、怒(angry) 、哀(sad)、樂(happy) 共四類情緒的文句情緒分類器。各情緒分類在正負面傾向和能量波動的程度有所區隔，其區別與相關詞彙如表三所示，如「怒」類除了屬於負面情緒外，其能量波動的程度也較大。四類情緒在文句的呈現上，其中「喜」類情緒出現在感到愛慕、喜好、狂熱的語句中，例如：

「我最喜歡Jolin了!」

「怒」類情緒出現在感到生氣、憤怒、咒罵的文句中，例如：

「可惡!早知道就不要出門了。」

「哀」類情緒出現在感到低潮、痛苦、難過、同情的文句中，例如：

「嗚嗚...可魯真的很可憐」

「喜」類情緒出現在感到高興、開心、有趣的文句中，例如：

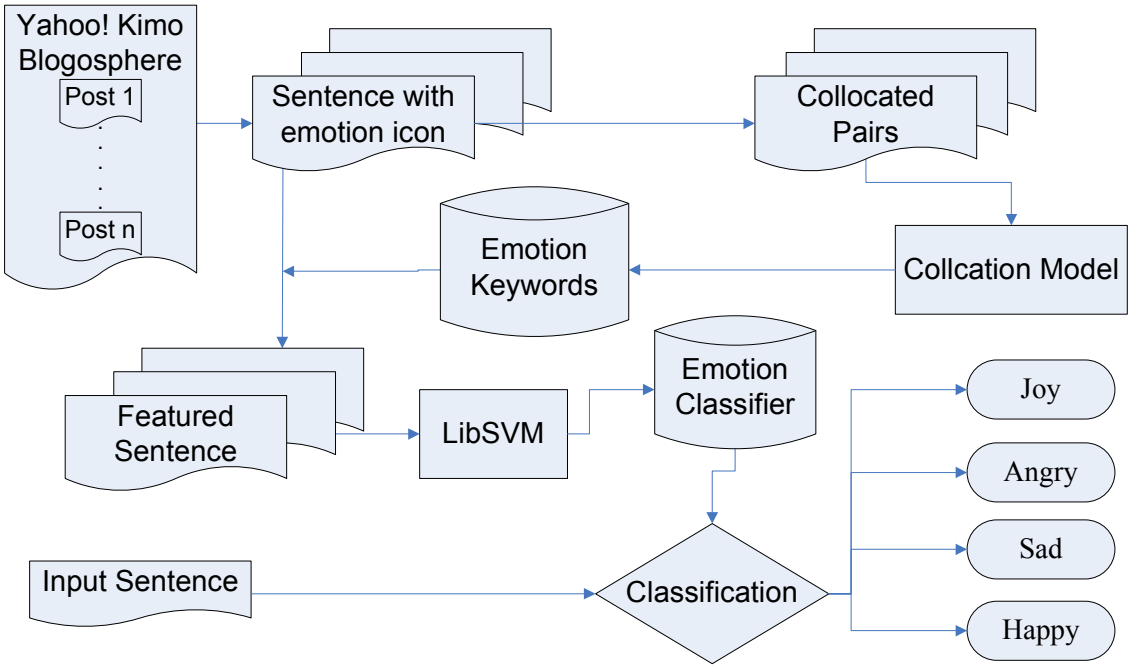
「計畫出遊又碰到好天氣實在很開心。」

表三、情緒分類器訓練時間分析

情緒分類	說明	相關詞彙
喜(joy)	情緒傾向：正面 能量波動：較大	愛、幸福、可愛、喜歡、 謝謝、害羞、感動、寶貝
怒(angry)	情緒傾向：負面 能量波動：較大	生氣、討厭、氣死、可惡、 幹、煩、罵、哼
哀(sad)	情緒傾向：負面 能量波動：較小	哭、痛、嗚、難過、 淚、傷、慘、可憐
樂(happy)	情緒傾向：正面 能量波動：較小	哈哈、開心、好笑、高興、 不錯、加油、很好、好玩

對於情緒的分類與與詞彙的使用有基本的定義後，情緒分析模組的實作流程如圖四所示。首先以Yahoo!奇摩部落格2006年1月到6月所有帶有表情符號的句子作為語料，取出帶有表情符號標記的文句共560,127筆，透過Collocation Model選出500個情緒詞彙。包含該情緒詞彙的文句經由特徵值轉換形成訓練資料，該資料用以訓練情緒分類器，分類器的實作使用Fan等人(2005)所提出的LibSVM[18]工具，以LibSVM工具所訓練的分類器模型則整合入部落格資訊系統，藉以判斷相關文章其構成文句的情緒傾向。

關於訓練資料量部分，於四類情緒各挑選30,000個文句，計12萬句進行情緒分類器的訓練。相較於楊和陳(2006)僅使用約4,000筆訓練語料，本實驗使用了30倍大的訓練集，訓練資料集數量雖然可以繼續嘗試擴大，但是在現今Intel Xeon 5320工作站環境下，30倍大的訓練集已增加了4,120倍的訓練時間。不同訓練資料大小與相對訓練時間如表三所示，當訓練資料僅有4,000筆時，所需的訓練時間僅需6秒，但訓練資料變成兩倍時，所需要的訓練時間竟增加成4倍之多，而本文所採用的分類器則需約7小時才能訓練完成。

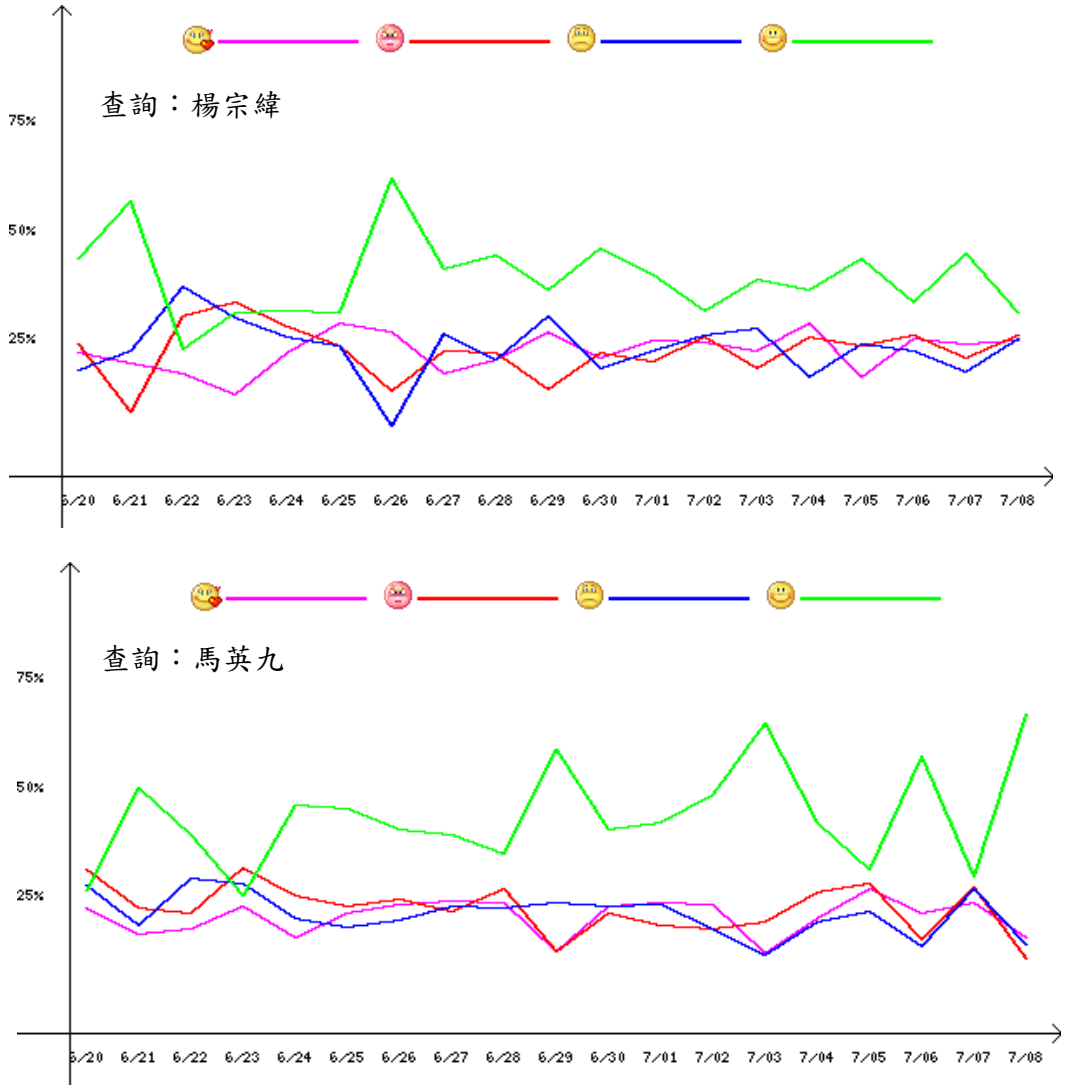


圖四、情緒分析流程圖

[18] http://rpc.weblogs.com/shortChanges.xml

表三、情緒分類器訓練時間分析

訓練句數	訓練倍數	訓練時間	時間倍數
4,000	1	約 6 秒	1
8,000	2	約 25 秒	4
40,000	10	約 31 分	310
120,000	30	約 6 小時 52 分	4,120



圖五、情緒趨勢分析圖

5. 部落格情緒趨勢分析

本文第三節提出一個虛擬部落客「人口普查」概念，可根據部落客名單，以查訪最新文章列表的方式來獲得文本資訊。本節中再提出一個部落客「民意調查」的應用，運

用第四節說明的情緒分析模組，針對個別文本中的各文句進行情緒分析，偵測該文本有無出現喜、怒、哀、樂等情緒，各情緒分數以該情緒的文句數的對數計算之。根據第二節所述，部落格資料庫所提供的文本資訊具有時間標記，因此我們在對文本作情緒分析時，也可以為不同時間域的文本集合計算出對應的情緒分數。

本研究觀察 2007 年 6 月 20 日至 7 月 8 日所發布的文本，個別時間域以天為單位，在對每天的文本集合算出個別情緒的分數比例後，可藉此觀察個別情緒在跨越時間域時的變化走向。圖五顯示以不同關鍵字查詢所獲得的相關文本在跨越時間域時的情緒變化，查詢以不同人物的姓名為例，分別是「楊宗緯」、「馬英九」。圖五所包含的三個子圖，其共通之處在於「樂」情緒類是屬於最大宗的情緒呈現，而「樂」情緒類趨勢的下降通常代表著「哀」情緒類的上升；而「喜、怒、哀」三種情緒類在不同時間域上會互有領先。說明個別的趨勢圖以「查詢：楊宗緯」為例，在 6/20-6/21 呈現較高的快樂的情緒後，6/22-6/23 負面情緒上揚，直到 6/26「樂」情緒類大幅上升後，便持續領先到 7/8 止。

表四、最近時間域情緒分析統計

查詢	喜	怒	哀	樂	說明
楊宗緯	12%	14%	11%	63%	星光幫偶像
林宥嘉	14%	13%	12%	61%	星光幫偶像
股票	7%	11%	9%	72%	台灣股市大漲
貓空	10%	22%	19%	48%	貓纜初期營運負面感受
纜車	11%	21%	18%	50%	貓纜初期營運負面感受
謝長廷	7%	14%	7%	73%	總統參選人
馬英九	14%	15%	10%	62%	總統參選人
王建民	10%	18%	10%	62%	旅美棒球選手
日本	14%	17%	12%	57%	國家、旅遊、演藝娛樂
韓國	17%	21%	15%	48%	國家、旅遊、演藝娛樂

所謂趨勢概念，最令人關心的部分是屬於「目前」、「最近」的趨勢，因此如果設定觀察的時間域為「最近」這個概念時，實作方法可以以時間排序針對各查詢選出若干最新文本，並統計其情緒分數比例。表四即針對各查詢選出最近 150 篇相關文本，在計算文本個別的情緒分數後，列出各查詢綜合相對的情緒比例。

各查詢之間的相對值或極大值，可用來觀察部落客對各項議題不同的感受，以極值的觀察為例，可解讀成部落客最近對股市的表現、或總統參選人謝長廷最感到高興、對貓空纜車的營運最感到不悅與生氣、對韓國則混有嚮往及生氣兩種情緒。以相對值的觀察為例，如楊宗緯和林宥嘉同屬於熱門歌唱比賽出身的偶像，其相關文章較為一致，情緒表現僅有些微的差距。而謝長廷與馬英九雖然同屬於總統參選人，但是其情緒「喜」、「樂」的比例則互有勝負，或許可解讀成儘管部落客普遍對謝長廷的表現較感到滿意，而情感上則較容易喜愛馬英九。

5.1 討論

6. 結論與未來方向

本文探討如何從部落格空間獲得文本資訊，實作出一個資訊系統，並應用在情緒趨勢的分析上。關於資訊系統未來發展方向，一方面可以擴充更多的查訪部落格空間獲得更豐富的文本資訊，一方面可支援如部落格搜尋、部落格摘要等研究議題所需用到的語料。在情緒趨勢的分析上，則可以朝向自動偵測出變動的情緒，並掌握相關的話題的自動機制發展。

參考文獻

Corinna Cortes and V. Vapnik. 1995. “Support-Vector Network,” *Machine Learning*, Vol. 20, pp. 273–297.

Rong-En Fan, Pai-Hsuen Chen and Chih-Jen Lin, “Working Set Selection Using Second Order Information for Training Support Vector Machines,” *Journal of Machine Learning Research*, Vol. 6, pp. 1889–1918, 2005.

Bin He, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang, “Accessing the Deep Web,” *Communications of the ACM*, Vol. 50, 5, pp. 94–101, 2007.

Lun-Wei Ku and Hsin-Hsi Chen, “Mining Opinions from the Web: Beyond Relevance Retrieval.” *Journal of American Society for Information Science and Technology*, Special Issue on Mining Web Resources for Enhancing Information Retrieval, accepted.

Kevin Hsin-Yih Lin, Changhua Yang and Hsin-Hsi Chen, "What Emotions Do News Articles Trigger in Their Readers?" *Proceedings of 30th Annual International ACM SIGIR Conference*, pp. 733–734, Amsterdam, Netherland.

Gilad Mishne. 2005. Experiments with Mood Classification in Blog Posts. *Proceedings of 1st Workshop on Stylistic Analysis of Text for Information Access*.

Gilad Mishne and Maarten de Rijke. 2006. Capturing Global Mood Levels using Blog Posts. *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 145–152.

I. Ounis, Macdonald, M. de Rijke, G. Mishne, and I. Soboroff, "Overview of the TREC 2006 Blog Track," *Proceedings of the 15th Text REtrieval Conference*, Gaithersburg, Maryland, 2006.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen, "Building Emotion Lexicon from Weblog Corpora," *Proceedings of ACL-2007*, poster, Prague, Czech, pp. 133–136, 2007.

Sheng-Chuan Yen, *A Study of Identifying Implicit Trackback in Weblogs and Its Application on Weblog Search*, Master Thesis, National Taiwan University, 2007.

楊昌樺、陳信希。"以部落格文本進行情緒分類之研究，"第十八屆自然語言處理與語音處理研討會論文集，253–269 頁，2006。

混合語言之語音的語言辨認

Language Identification on Code-Switching Speech

朱晴蕾 1, 呂道誠 2 , 呂仁園 1
1. 長庚大學資訊工程研究所
2. 長庚大學電機工程究所
E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967

摘要

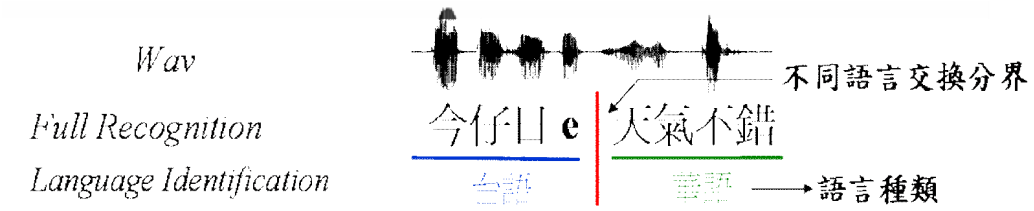
本論文主要針對台灣地區所出現的華語與台語的混合語言語音(Code-Switching Speech) ，研究其語音的語言辨識。藉由自動語音辨識技術得到語音音節序列，經語言詞典比對產生斷詞後，以字詞在不同語言中出現之頻率和字詞組合機率爲判別語言共同字詞準則，進而得到語音之語言種類辨識。最後以語言標籤和語言時間資訊兩種不同的評估方式，分別對語言辨別進行進一步正確率評估，約可達到 83.4%及 78%的語言辨識率。

一、緒論

傳統 LID 主要用於判別出一段語音是由何種語言所建構成，例如：「今天天氣不好」、“How are you？”…等，所判別的語音本身是僅由一種語言所組成。但由於多語的社會環境，因大量吸收外來語，及方言接受度的提高，導致現代語言結構規則發生改變，或爲因應不同需求或習慣，而擷取不同語言片段（/特徵）以創造新詞，使得越來越多不同語系的語詞夾雜應用於陳述上。在如今全球溝通交流普及的環境下，語言轉換（Code-Switching）技巧不僅在報章雜誌大眾傳媒上頻繁出現，也在日常生活上被一般大眾普遍使用。

所謂的混合語言語音（Code-Switching Speech）指的就是一段由 2 種或 2 種以上語言交替組合而成的語音。說話者從一種語言轉用另一種語言，轉換原因諸如遇到談話主題、對象或場合的改變，或是說話者雖有某事物的概念、卻僅能以某種語言形式表達時，語句便會在中途產生轉換，局部轉切至該語言形式，例如：「Star Bucks 的咖啡很好喝」爲華英 2 種語言的轉換用法，而「今天晚上有“夜市仔”」，則爲地方方言與國語混合使用，爲華台 2 種語言的轉換用法。

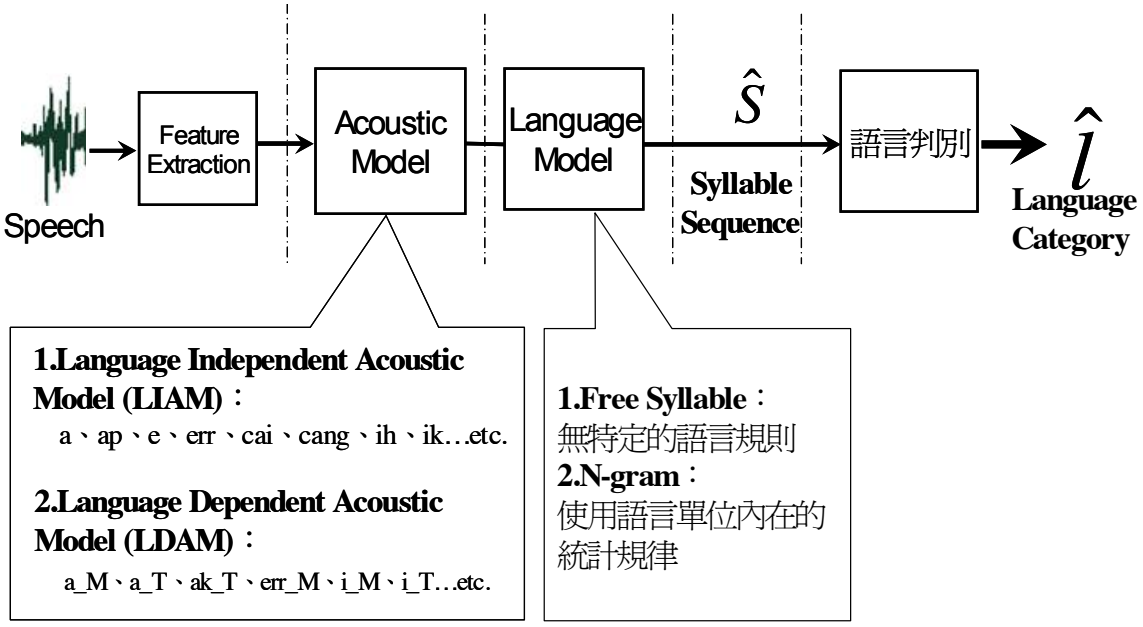
我們的研究目標，主要便是針對在台灣地區最普遍出現的語言交雜狀況—也就是以華語爲語言主幹，在語句中穿插其他種類語言的混合語言語音（Code-Switching Speech）者—爲研究對象。期望能找出一個較佳的策略，以區分一段未知語音中的不同語言的交換分界處，進而有效判斷出語言種類。如圖一所示，我們期望能找「今仔日 e」爲一個台語語音片段，而「天氣不錯」爲一個華語語音片段的結果。



圖一、Code-Switching Speech

目前在自動語言辨識的課題上，大部份仍是以純單一語言構成之語音爲主，混合語言語音之研究則較爲少見。香港的 Joyce Y. C. CHAN 等人曾在 2004、2006 發表關於廣東話與英語混雜的混合語言語音研究[7][9]，分別嘗試使用 Knowledge Base 和 Data Driven 兩種不同的語言混合音素集合辨識語音序列，並以 Bi-phone 出現在廣東話中的機率做爲語言判別準則，其辨識結果約 69.62%、76.54%。而在國內，成功大學林俊憲、Chi-Jiun Shia、Chung-Hsien Wu 等人對於華語與英語混雜語音的相關研究[4-6]，使用 BIC 預先切割語音，再以加入隱含式語意索引(Latent Semantic Indexing) [2,3] 所訓練而成的高斯混合模型來辨別語音音序，以達到混合語言辨別的成效，其辨識結果約爲 74%。前者在語言辨識的架構較爲簡單且亦可得到一定語言辨識效果，故我們採取類似 Joyce Y. C. CHAN 等人的實驗系統架構，並於後端語言辨別處理部份則加以改進，期望能達到更高的語言辨識率。

我們系統架構流程如圖二所示：當一個語音進來時，先將其進行特徵的截取，經由聲學模型和語言模型兩部份，將得到一個 syllable 序列；接著將這個 syllable 序列給予語言判別區塊，以詞典斷詞、語言共同詞處理判斷，得到這整段語音的語言類別；最後再使用語言標籤和語言時間資訊爲兩種不同的評估方法，評鑑整體語言辨認正確率。



圖二、系統架構流程圖

本篇論文內容分別爲：第二節音節辨識，介紹所使用之聲學和語言模型，第三節語言判別策略，第四節實驗用語料、正確率評估方式和最後的結果討論。

二、音節辨識

在我們的語言辨認系統中，是基於自動語音辨識所得到的音節序列爲基礎，再做後續的語言辨識。而音節序列的正確與否將會影響後續語言辨識的正確性。

（一）聲學模型

在語言的標音上，我們使用福爾摩沙音標(Formosa Phonetic Alphabet, ForPA)。ForPA 音標是一個台灣地區三語(華台客)共用標音系統，在華語的音素有 37 個，而台語有 56 個，兩種語言音素聯集共有 63 個，而交集的有 32 個。

以 ForSDAT (Formosa Speech Database) 台灣地區多語言語音資料庫做爲訓練資料（Training Data），華語部份使用 ForSDAT 中 MD01 語料庫，寮由 100 人所錄製成的華語句型語料，總長度爲 11.3 小時。台語部份使用 ForSDAT 中 TW01 語料庫，爲由 100 人所錄製成的台語句型語料，總長度爲 11 小時。每個聲學模型皆以音節內左右相關的方式的三個連續音素爲單位，使用 3 個狀態的隱藏式馬爾可夫模型（Hidden Markov Model：HMM）。

在聲學模型方面，分爲語言獨立的聲學模型（Language Dependent Acoustic Model, LDAM）：在聲學模型中所有的音標沒有語言上的分別，如 a、ap、e、err、cai、cang、ih、ik…etc，整個聲學模型語言是由華語與台語共同訓練而成。另外，我們對於每種語言分別訓練其獨立的聲學模型，並在每個聲學模型中標記其所屬的語言類別，再將不同語言的聲學模型結合在一起，形成一個混合語言的聲學模型[11]，稱之爲語言相依之聲學模型（Language Independent Acoustic Model, LIAM）：在聲學模型中所有的音標皆帶有所屬語言的標籤，如 a_M、a_T、ak_T、err_M、i_M、i_T…etc.。

（二）語言模型

N-gram 的語言模型可以提供一種語言中其文字的序列規則，並以統計和機率的方式來呈現。以下是 N-gram 的表示式，W 爲 n 個 w 的集合，其中每個 w 代表一個字詞。假設第 n 個字詞的出現只與前面 N-1 個字詞相關，而與其他任何字詞都不相關，整句的機率就是各個字詞出現機率的乘積。這些機率可以通過直接從語料中統計 N 個字詞同時出現的次數得到。當 N 愈大時所需統計語料將愈多。

$$P(W) = P(w_1 w_2 w_3 \ldots w_n) = P(w_1) P(w_2 \mid w_1) \cdots P(w_n \mid w_1 w_2 \ldots w_{n-1})$$

$$= \prod_{i=1}^{n} P(w_i \mid w_{i-n-1} \ldots w_{i-2} w_{i-1})$$

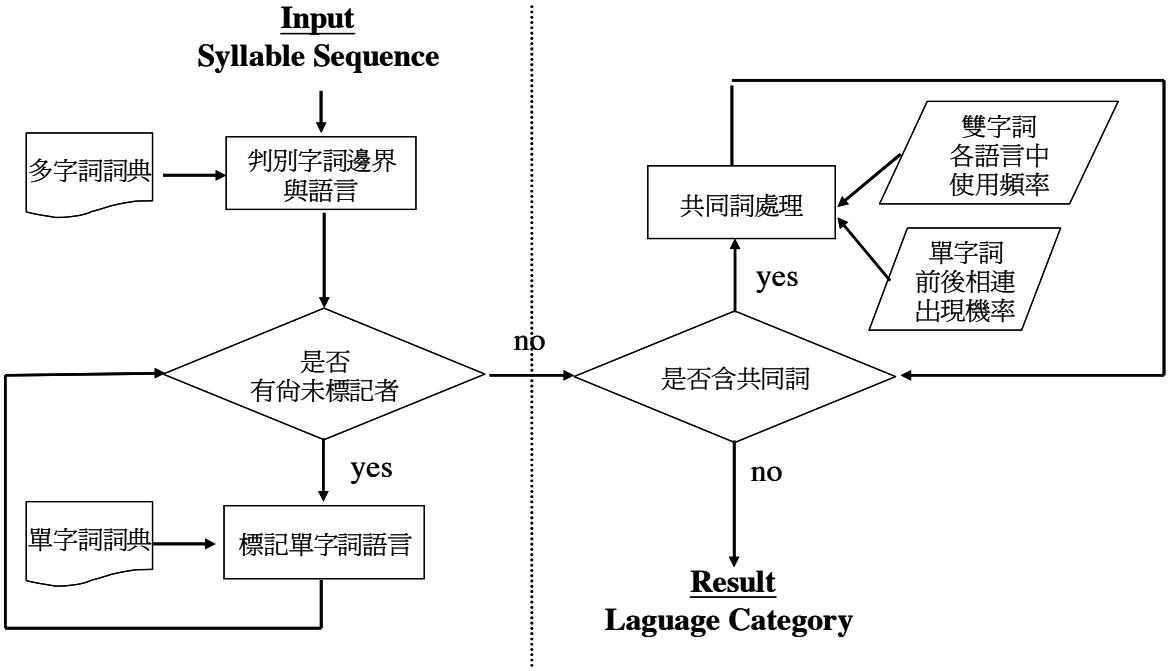
對正確的語言現象，字與字之間的共同出現機率較高，對一些較不符合語法者，字與字之間的共同出現機率較低。此機率的大小可以反映出一個語言的局部規律。例如：今天和今仔日有相同的意義，但“今+日”這種組合在華語中出現機率相對比在台語中出現機率高，而“今+仔+日”這種組合在台語中出現機率相對會比在華語中出現機率高。

依據現有已標音完成文字語料資料量，我們使用的 N-gram 語言模型上取 N=2，亦稱之爲 bi-gram，並以 Syllable 爲單位量。我們所使用來訓練 Bi-gram 機率的語料，在華語文章部份總句數約有 1 萬 7 千（17203）句，全部約 23 萬字（230236）左右；台語文章部份總句數約 9 千（9539）句，全部約 10 萬字（104324）左右。其中，所有文章中的句子皆僅由一種語言組成。

在 Bi-gram 的情況下，一個字詞出現的機率只會依據上個字詞而出現。Bi-gram 的機率表示法爲： $P(w_t \mid w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})}$ ， w_{t-1}, w_t 表示時間上相連的兩個 Syllable，$C(w_{t-1}, w_t)$ 爲 w_{t-1}, w_t 在文章中共同出現的次數，$C(w_{t-1})$ 則爲 w_{t-1} 在文章中出現的次數。

三、語言判別

一段語音進來後經由聲學模型和語言模型後可以得到一 syllable sequence，將這個序列再進行語言判別。以流程圖圖三所示，當一個 syllable sequence 進來後，會先判別字詞邊界與語言，若序列中每個 syllable 皆已標記了可能語言種類後，再確認所標記的語言種類中是否有包含了語言間共有的字詞後，將這些共有字詞再進行第 2 次判別，最後將可得到這個序列的語言類別。這是我們的語言判別大略的步驟，詳細的內容在之後會逐一說明。



圖三、語言判別流程圖

（一）字典與詞典制作

首先，先介紹在一開始判別字詞邊界與語言這個步驟所用到的多字詞詞典與單字詞詞典。多字詞詞典部份，我們使用中研院的華語詞典與實驗室既有的台詞詞典混合成一

新的華台混合詞典，總詞數約 14 萬(143705)詞，並將每個多字詞後標記所屬語言種類。整個詞典將被分成三種語言類別；僅會在華語中出現的多字詞詞，標示爲 M，約 8 萬詞(88675)；僅會在台語出現的多字詞詞，標示爲 T，約 5 萬詞(53499)；最後一種爲兩種語言中皆可能出現者，標示爲*M/*T，約一千五百詞(1531)。

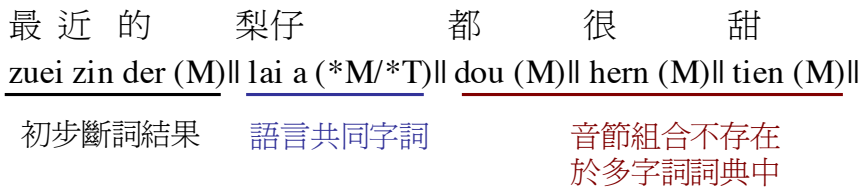
單字詞詞典的製作與多字詞詞典雷同，在單字詞詞典中將含有兩種語言所有可能出現的發音，總單字詞數爲 1010 個，並同樣分爲三種語言類別。僅在華語中使用的單字詞數爲 230 個，僅在台語中使用的單字詞數爲 580 個，華語與台語同時使用的單字詞數爲 200 個。

（二）判別字詞邊界與語言

我們實際觀察一些混合語言的語音文句，發現語言轉換時機似乎經常發生在詞（word）上，因此我們在這裡先假設語言轉換的時機在於 word 上，故 word 與 word 的邊界便可能爲一個語言交換的邊界處。在判別字詞邊界與語言這個步驟，便是將 syllable sequence 與詞典中詞句最長匹配比對，將比對到的部份做爲一詞與詞的分界點，並依詞典中每個詞後所帶的語言標記做爲其語言種類。例如：“uo zuei si huan cyu ia ci a chii dong si”「我最喜歡去（夜市仔）吃東西」這句話，若詞典中分別包含有“uo zuei si huan cyu”、“ia ci a”、“chii dong si” 三個詞，那判別的結果將會爲：uo zuei si huan cyu（M）‖ ia ci a（T）‖ chii dong si（M）‖ 。

然而，於此方式的初步成果中，我們可以發現一些問題存在。

首先是在 syllable sequence 中的 syllable 組合有可能不存在於詞典中。以圖四爲例，最後的 dou hern tien（都很甜）這個 syllable 組合並不存在於詞典中，而這種問題我們將改以直接使用字典判斷單一 syllable 的語言種類來解決，故最後會得到 dou（M）‖hern（M）‖ tien（M）‖這樣的結果。另外一個問題在於，有些字詞是同時存在於不同語言間共同使用，也就是原先分類爲*M/*T 的字詞，如例句中的 lai a（*M/*T）‖，而這個部份，我們將以另一套方法來處理判斷。



圖四、判別字詞邊界與語言初步結果

（三）共同詞處理

在共同詞的處理上，會分爲多字詞與單字詞的兩種處理方式。在多字詞的處理上，我們使用中研究 CKIP 語料庫與實驗室語料庫所統計得到的在不同語言所出現的字詞與其出現次數表，利用詞在不同語言出現的相對頻率多寡來決定所屬語言。例如，共同發音字詞 ker ci，在台語中出現次數爲 149 次，而華語僅 79 次，分別除以所統計的華語、台語總詞數，取兩者相對頻率較高者，而在此我們可以發現得到的在台語詞中出現機率

較高，故將 ker ci 標記爲一個台語發音詞。

而在單字詞方面，由於我們使用了兩種不同的聲學模型，故將以兩種不同聲學模型所得到的 syllable sequence 分別做討論。

1、 Language Independent Acoustic Model（LIAM）：

若聲學模型爲 LIAM，則得到的字串中每個 syllable 本身即會帶有語言標籤。在這種 syllable sequence 中發現的語言共有單字詞，便直接以 syllable 本身所帶的語言標籤做爲語言依據。

2、Language Dependent Acoustic Model（LDAM）：

若爲另一種聲學模型 LDAM，所得到的字串中 syllable 將不帶有語言標籤。因本身沒有語言標籤，故我們無法以直接的方式得到語言種類。考慮到在我們假設轉換爲 word 的前提下，應不會以一個 syllable 做一次快速轉換，若前後兩個相鄰詞爲相同語言時，則此 syllable 應與前後兩相鄰詞爲同一語言。

以上爲第一種方法，但這個想法並無法完全解決所有語言共有單字詞的問題，若語言共有單字詞出現在前後兩相鄰詞爲不同語言類別間時，便會無法處理。爲此，我們提出了第 2 種方法，利用與前後兩相鄰單字詞同時出現的機率大小做爲判別依據。與 N-gram 的概念相同，當機率值相對較高時，表示愈有可能爲同種語言。在加入這個方法後，將可以成功處理所有語言共有單字詞。

四、實驗評估

（一）語料庫設計

由於我們在現有的所有語料庫中，並沒有任何混合語言的語音資料。故在進行實驗前，我們將先行錄製建立一混合語言的語料庫做爲我們整個實驗的來源依據。

在錄製語音前，我們需要有混合語言的文句做爲錄音劇本。在混合語言的文句收集上，首先先由各式文章中，例如：電子新聞、部落格…等，尋找日常生活中可能使用到的混合語言文句來做爲我們的錄音劇本。由於台語爲一種方言（Dialect），在文字的呈現上，可能與華語使用文字上相同，或使用不常使用較爲特別的文字，較不易直接在一般文章中發現。因此，除了在各式文章中直接尋找華語與台語混合語言文句外，我們參考[8]的作法，我們將兩句由不同語言構成但意義完全相同的文句做一比對，可得到兩種語言在句法節構與用字上的相對情況，並將文句依相對情況切分成數個文字區塊。選取原文句中某文字區塊，替換成另一種語言文句中相對應的文字區塊，如此便可得到一混合語言文句。

在華語與台語混合語言文句上，我們收集了約 75 句。表一爲其中的一些例句。我們所錄製的華台混合語言語料庫，參與錄音的人數總數爲 10 人，約 750 句，平均每句長度爲 2 秒(0:02.391)，由麥克風及個人電腦的 32k 音效卡，在安靜無噪音的環境底下錄製 16KHz，16bits 的聲音訊號，並以 WAV 格式音檔儲存。

表一、華台語混合語言語音例句

Filename	Text	Transcription
MT_001	我最喜歡去（夜市仔）吃東西	uo3_zuei4_si3_huan1_cyu4_ia2_ci2_a4_chii 1_dong1_si1
MT_005	(歹勢！)我遲到了	painn4_se3_uo3_chii2_dau4_ler0

（二）、實驗結果

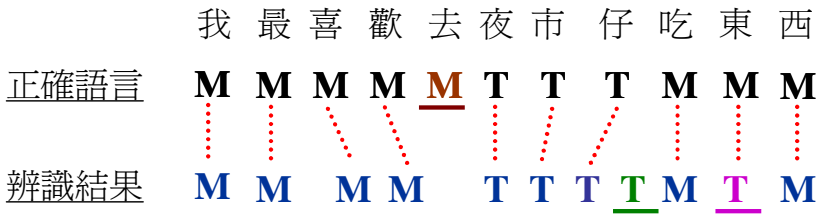
我們使用了兩種不同的聲學模型，以及兩種不同的語言模型來做比較，共分爲四種不同的實驗組合：語言獨立聲學模型與 Free Syllable 語言模型、語言獨立聲學模型與 Syllable Bi-gram 語言模型、語言相依聲學模型與 Free Syllable 語言模型、語言相依聲學模型與 Syllable Bi-gram 語言模型。將四種組合所得到的 Syllable Sequence 交予我們的語言判別策略做語言上的辨認，可得到四種不同的語言辨認結果。

在混合語言語音 LID 的實驗中，由於沒有語言轉換邊界的資訊，故無法直接得到如單一語言語音 LID 的語言區段。在混合語言語音 LID 所得到的結果會是以音節序列的方式做爲初始呈現，因此在這裡我們對原先在單一語言語音 LID 正確率計算上做了些改進。我們將辨識結果中相鄰爲相同語言的語言標籤合併，形成數個單一語言區塊，而每個單一語言區塊的語言標籤便爲我們計算正確率的單位量。這種單位的計算方式與在單一語言語音 LID 計算正確率的單位量是相同的，而做爲在混合語言語音 LID 的語言辨認正確率計算單位時，可同時顯示出語言轉換邊界的正確程度。另外，我們增加了另一種計算單位量，以組成單一語言區塊的基本單位－音節爲主，當完成語言辨識後一個音節即有一個對應的語言標籤，也就是最原始的結果單位。以這樣的計算單位，在語言辨認正確率評估上可同時顯示我們在音素辨認上的正確程度。

在評估語言辨認結果正確率上，我們同樣使用了兩種不同的評估方式。第一種是和單一語言語音 LID 正確率判別相同的評估方式，以語言標籤爲評估單位。另外我們在原本正確率的評估上再加入時間上的資訊，也就是第二種評估方式，以語言時間資訊做爲評估單位。

1、語言標籤評估法

第一種評估方法是以語言標籤來做爲評估單位。如圖五爲例來說，若有一句話：「我最喜歡去夜市仔吃東西」，正確答案的語言標爲：MMMMMTTTMMM，若我們辨識出來的結果爲 MMMMTTTTMTM，將正確答案與語言判別結果做比對，可以得到 4 種不同情況數值，分爲別：Hit：正確答案的語言標籤與語言判別結果的語言標籤完全相符。刪除錯誤（Deletion）：語言判別結果所缺少的語言標籤部份。插入錯誤（Insertion）：語言判別結果所多餘的語言標籤部份。替換錯誤（Substitution）：正確答案的語言標籤與語言判別結果的語言標籤不相符。



圖五、語言標籤做爲評估單位範例

在得到這四種不同的數值後，以現在最常使用的兩個衡量標準來評估我們的方法：精確率（Precision Rate）：在語言判別結果的語言標籤總數中，正確語言標籤所佔比率。N_H 爲語言判別結果中正確的語言標籤數(Hit 總數)，N_L 爲語言判別結果全部的語言標籤數(Hit 總數+Substitution 總數+Insertion 總數)。

$$precision = \frac{N_H}{N_L}$$

召回率（Recall Rate）：在正確答案中有被判別出來且爲正確的語言標籤佔正確答案語言標籤總數比率。N_H 爲正確答案中有被辨識出且爲正確的語言標籤數(Hit 總數)，N_C 爲正確答案中的語言標籤總數(Hit 總數+Substitution 總數+Deletion 總數)。

$$recall = \frac{N_H}{N_c}$$

通常，當有高召回率時，便很難能有高精確率；反之，有高精確率時，將難有高召回率。故我們另使用 F-Measure 做評估。F-Measure 是依據上面的精確率和召回率兩個衡量標準，加以綜合而成的另一個評估指標。

$$\frac{1}{F-measure} = \frac{1}{2}\cdot\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

$$\Rightarrow \quad F-measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2\cdot\frac{precision \cdot recall}{precision + recall}$$

在 F-Measure 中，需要同時滿足精確率和召回率皆較高時才能得到高的數值，而這樣的方法可以在精確率和召回率上取得一個平衡。

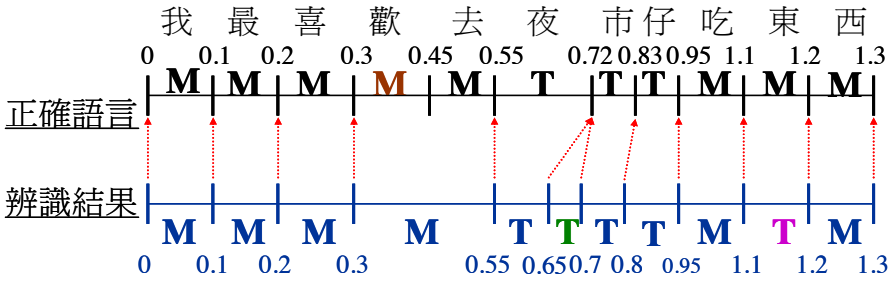
表二、以語言標籤做爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LIAM+Free Syllable	72.63%	68.55%	70.53%	58.67%	93.75%	72.17%
LDAM+Free Syllable	69.9%	65.7%	67.6%	69.11%	86.82%	76.96%
LIAM+Syllable Bi-gram	72.73%	68.71%	70.66%	82.26%	76.26%	79.14%
LDAM+Syllable Bi-gram	78.1%	68.51%	**73.02%**	82.35%	79.0%	**80.67%**

表二爲我們以語言標籤做爲評估單位的實驗結果，在這種評估單位下，我們可以發現在 LDAM+Syllabe Bi-gram 的實驗組合並以合併後的語言區塊爲計算單位的情況下能得到最高的正確率。

2、語言時間資訊評估法

第二種評估方式是使用語言時間資訊來做爲評估的單位，語言的時間資訊所指是每個語言標在語音中的起始時間、結束時間與持續時間長度(Duration)。這種評估方式，是參考[10]在辨認語音中不同語者(Speaker)的正確率計算方法，由於在語音中有一個以上不同語者交替出現與語音中有不同語言交替出現相類似，故我們使用[10]的評估方式來評估我們的正確率。首先，將語言判別結果的語言時間邊界與正確答案語言時間邊界做對位（Alignment）的動作，以判別結果語言區塊時間佔正確答案語言區塊時間比例爲主要對位上的準則。以圖六爲例，在完成對位後再比較兩者的語言標籤，同樣可以得到 Hit、Deletion、Insertion、Substitution 4 種數值，但在這邊數值的計算方式是取用時間爲單位來表示。



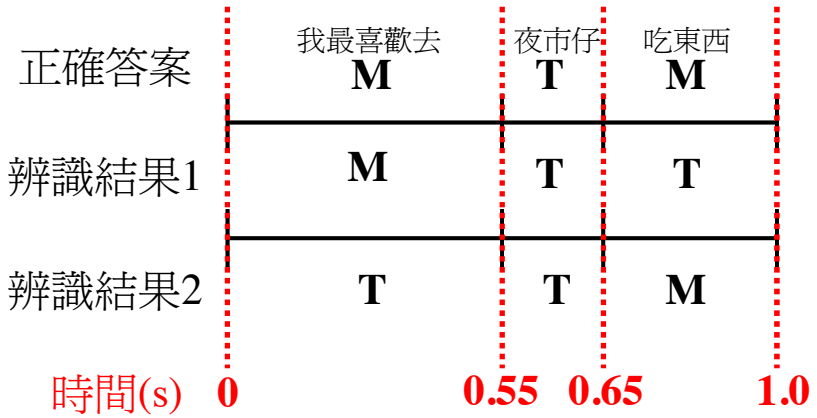
圖六、以語言時間資訊爲評估單位

同樣的我們計算其精確率、召回率與 F 測度，在這個方法下精確度爲在所有語言判別結果的語言標籤中，是正確語言標籤者的時間長度占辨識結果總時間長度的比率。而

回現率也變更爲在所有正確的語言標籤中，有被辨識出且爲正確語言標籤者的時間長度占正確答案總時間長度的比率。因此，若當每個語言區間時間長度相同時，其實會退化回成以語言標籤爲評估單位元同樣的方式。

$$precision = \frac{\sum_{i=1}^{N_H} \sigma_i}{\sum_{i=1}^{N_L} \sigma_i} = \frac{N_H * \sigma}{N_L * \sigma} = \frac{N_H}{N_L}$$

依語言時間資訊爲評估單位時，可以較清楚瞭解判別錯誤的部份對整句語音的影響程度。以圖七「我最喜歡去夜市仔吃東西」爲例，使用單一語言片段做計算單位，可分爲三個單一語言片段 MTM。若有二種不同的語言辨識結果，第一種辨識結果爲 TTM，而第二種辨識結果爲 MTT 時，以語言標籤爲單位的正確率計算方式上，兩種結果皆爲三個語言標籤中有一個語言標籤爲錯誤，所得到的正確率會相同，皆爲約 67%；若以語言標籤時間資訊爲單位時，第一種結果錯誤的語言長度爲 0.35 秒，而第二種結果錯誤的語言長度爲 0.55 秒，此時便可分出第二種辨識結果其實較第一種辨識結果好。



圖七、兩種評估方法比較範例

表三、以語言時間資訊爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段 爲語言計算單位		
	P	**A**	**F**	**P**	**A**	**F**
LIAM+Free Syllable	75.3%	68.68%	71.84%	60.2%	52.78%	56.25%
LDAM+Free Syllable	77.1%	68.87%	72.75%	68.8%	58.05%	62.97%
LIAM+Syllable Bi-gram	84.1%	70.76%	76.8%	79%	60.57%	68.57%
LDAM+Syllable Bi-gram	84.6%	71.76%	**77.3%**	79.1%	61.67%	**69.31%**

表三爲我們以在以語言時間資訊爲評估單位所得到的結果。將兩種不同評估單位的結果做相互對應，在以音節爲計算單位時，使用時間資訊爲評估單位的情況下能得到的正確率 81.4%，而以語言標籤爲評估單位的方式爲 73%，故可估測在語言辨識錯誤的音節，可能多數爲持續時間較短的音節。

在以合併的語言區塊爲計算單位時，我們可以發現使用語言標籤來做爲評估單位的情況下可得到的語言辨識結果約 80.7%。但在使用語言時間資訊來做爲評估單位的方式中，我們可以發現以語言區塊爲計算單位的情況可得到約 74.8%的語言辨識率。這表示我們辨識正確的語言片段佔所有的語言片段約八成左右，但正確語言的時間佔總體時間僅約七成，故可估測在爲正確語言標籤的語言區段時間中，含有部份的錯誤時間段。

（三）、**Word** 與 **Syllable** 混合之 **Bi-gram**

在我們的假設中，混合語言語音中語言轉換可能較常發生於 word 上，故若能得知 word 與 word 間之相應機率，對混合語言語音的語言辨識應能有所幫助。但由於在各種語言中，word 的數量皆相當多，若想以 word 爲單位訓練出其 Bi-gram 機率，則所需的訓練語料資料量將會相當龐大。因此我們折衷計算部份較常出現的 word 與 syllable 間的 Bi-gram 機率。由於我們既有的已標音完成的語料庫資料不足，故我們需要重新收集並製作新的語料庫。首先，收集大量文章文字，我們由新聞稿、鄉土文學、散文、生活小品等文章中，收集結成約 10 萬句總字數約 100 萬字的文章，但這些文章皆無標音。接著，我們將這些文章中的文字做標音，由於我們期望能得到類似 code-switching 此類較貼近我們的主題的文句，故我們先以約 2 萬較常用台語詞（僅雙字詞以上）爲主，將有出現在文章中的詞均先標記爲台語的標音，其餘尚未標音的文字再以同樣方式，以約 7 萬華語字詞爲做標音。

在整個標音過程中，真正有使用到的 word 個數約 22731；其中華語字詞數爲 15881 個，台語字詞數爲 6765 個，而華語與台語共同字詞數 85 個。另外，在標音的同時，我們計算所使用到的每個詞所出現的次數，並重新分別製作新的華台語字詞出現頻率表，以做爲之後語言共同字詞處理時使用。

而在標音完成後便可將文章轉爲一個華台語混合的音標序列，並使用完成標音的混合語言音標序列來做爲 word 與 syllable 混合 bi-gram 語言模型的訓練語料。最後將系統原本的 syllable Bi-gram 語言模型替換成新制作的 word 與 syllable 混合 Bi-gram 語言模型，再以與之前相同的方法，進行語言辨認的實驗。

我們在混合語言語音實驗中，以音節 Bi-gram 的語言模型目前可得到約 74.8%的辨識率。在我們假設中，詞可能爲混合語言語音轉換的單位，且猜測 word-base 的語言模型應能具有更佳的語言鑑別率，因此我們製作了以詞與字節混合 Bi-gram 語言模型，取代原本的音節 Bi-gram 語言模型，並進行語言辨識實驗。在正確率計算上，一樣使用兩種計算單位和兩種評估方式評測整體的語言辨識率。最後以目前擁有最佳語言辨認結果的音節 Bi-gram 架構做爲新語言模型的實驗對照組，比較詞與音節的 Bi-gram 語言模型對於語言辨識上是否有益，而實驗的結果如表四與表五所示。

表四、以語言標籤爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段 爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LDAM+Syllable Bi-gram	78.1%	68.5%	73.0%	82.4%	79.0%	80.7%
LDAM+Word&Syl Bi-gram	86.4%	63.3%	**73.0%**	88.4%	78.9%	**83.4%**

表五、以語言時間資訊爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段 爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LDAM+Syllable Bi-gram	78.8%	84.4%	81.5%	69.2%	79.0%	73.8%
LDAM+Word&Syl Bi-gram	78.2%	86.7%	**82.2%**	72.8%	84.1%	**78.0%**

由實驗統計可發現，無論以語言標籤爲評估單位或以語言時間資訊爲評估單位，詞與音節混合的 Bi-gram 方法所得到的結果皆比音節 Bi-gram 進步了約 3%左右。由此可知，以 word-base 的 Bi-gram 語言模型的確比音節 Bi-gram 的語言模型更具有語言鑑別能力。

五、結論

在我們的語言辨認系統中，由於是基於自動語音辨識所得到的音節序列爲基礎，音節序列的正確與否將會影響後續語言辨識的正確性。故我們使用了不同的聲學模型和語言模型做爲音節辨識上正確率提升的嘗試。由實驗結果得知，在特徵參數擷取中，直接使用 39 維 MFCC 在我們的系統能有較好的語音識別表現。在聲學模型上，以語言獨立的聲學模型能得到較好的結果，而在語言模型上，使用 Bi-gram 可表現出語言的規律性，有助於語言的鑑別，且以詞與音節混合 Bi-gram 又可比使用音節 Bi-gram 得到更高的語言鑑別度。而在評估方面，以語言時間資訊爲評估方式在語言辨識正確率的計算上，比以語言標籤爲單位的評估方式多考慮了時間因素，可得到更加精確的語言辨識率，使得評估結果上可更加客觀有力。

參考文獻

[1] 林奇嶽、王小川,"自動語言辨認簡介" ,計算語言學通訊第十七卷第四期,2006

[2] Haizhou Li,Bin Ma,Chin-Hui Lee,"A Vector Space Modeling Approach to Spoken Language Identification",IEEE TRANSACTIONS ON AUDIO,SPEECH,AND LANGUAGE PROCESSING,VOL.15,NO.1,JANUARY 2007,P271-P284

[3] Sheng Gao,Bin Ma,Haizhou Li,Chin-Hui Lee,"A Text Categorization Approach to

Automatic Language Identification",INTERSPEECH 2005

[4] Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsieh and Chung-Hsien Wu,"Language Boundary Detection and Identification of Mixed-Language Speech Based on MAP Estimation",ICASSP 2004

[5] Chung-Hsien Wu, Senior Member, IEEE, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin,"Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs",IEEE Transactions on audio, speech, and language processing, Vol.14, No.1, January 2006, p266-p276

[6] 林俊憲，＜應用隱含式語意索引與語言模型於中英夾雜語音之語言鑑別＞，國立成功大學，碩士論文，民國 91 年

[7] Joyce Y. C. Chan, P.C. Ching, Tan Lee and Helen M. Meng,"Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities", ISCSLP, 2004

[8] Joyce Y.C. Chan, P. C. Ching and Tan Lee,"Development of a Cantonese-English Code-mixing Speech Corpus", in Proc. of Eurospeech 2005,pp.1533-1536, Lisbon,2005

[9] Joyce Y.C. Chan, P.C. Ching, Tan Lee and Houwei Cao ,"Automatic speech recognition of Cantonese-English code-mixing utterance",INTERSPEECH 2006

[10] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, Member, IEEE,"Multistage Speaker Diarization of Broadcast News", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, 2006

[11] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu,"Language Identification by Using Syllable-based Duration Classification on Code-switching Speech", In Proceedings of Lecture Notes in Artificial Intelligence, Kent Ridge, 2006

基於 HNM 之國語音節信號的合成方法

An HNM Based Method for Synthesizing Mandarin Syllable Signal

古鴻炎　　周彥佐
Hung-yan Gu　and　Yen-zuo Zhou

國立臺灣科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
{guhy, m9315058}@mail.ntust.edu.tw

摘要

本文提出一個基於 HNM (Harmonic-plus-noise model) 的國語音節信號的合成方法，使用此方法時，一種音節只需錄、存一遍發音，就可用以合成多種韻律特性的發音，並且不易查覺出信號品質的衰退。在這個方法裡，一個欲合成的音節的音長，首先被分割成它的組成音素的音長，依據原始和合成音節裡各音素的音長，可建造一個片斷線性的時間對映函數，如此合成音節時間軸上的一個控制點，就可經由對映至原始音節上找出和它對應的兩個音框。然後依據兩音框的 HNM 參數作時間上的內差，再進一步在音色一致性的條件下作基週軌跡調整的內差，來求得該控制點上的 HNM 參數。當各個控制點上的 HNM 參數值都決定之後，就可使用我們重新公式化的 HNM 合成公式，來計算出各個信號樣本的值。接著我們作聽測實驗來評估合成語音的清晰度，初步結果顯示，本文所提的 HNM 擴充的方法所合成出的信號，非常清晰且無迴音，明顯地比先前提出的 TIPW 法的好很多。

關鍵詞：語音合成，諧波加雜音模型，音色一致性

Keywords: Speech Synthesis, Harmonic-plus-noise Model, Timbre Consistency.

一、前言

自從 PSOLA 合成法被提出之後[1]，它已廣泛地被應用於作語音信號的合成，不過使用 PSOLA 所合成出的語音信號的品質，並不穩定，例如對一個錄製的語音單元的基周軌跡(pitch contour)或音長(duration)作較大幅度的改變時，則合成語音的品質會衰退很多[2]。雖然在採取語料庫(corpus based)爲基礎的研究方向[3, 4]時，韻律特性通常不需要作大幅度的改變，並且 PSOLA 是一個容易製作、運算量很少的信號合成方法，如此情況下 PSOLA 可說是一個不錯的選擇。但是，基於語料庫之研究方向的前提條件是，語料要錄得夠多，否則合成出的語音信號，在某些音節之間會出現音調銜接得不平順的問題，並且某些語句內會出現說話速度忽快忽慢的問題，那麼就仍然需要作較大幅度的韻律特性的調整了。此外我們考慮到，所研究的語音合成技術希望可以很經濟地(節省人力、時間)移轉到其它語言(如閩南語、客語)去使用，因此我們傾向於採取信號模型(signal model)爲基礎的研究方向，而不願意採取語料庫爲基礎的研究方向。

由於國語是一種有聲調的語言，且國語聲調之間的差異，主要表現於音節基週軌跡的高

度和形狀。因此當採取信號模型爲基礎的研究方向時，我們需要對一個合成單元的基週軌跡的高度及曲線形狀作大幅度的改變，如此 PSOLA 就不適合使用了，而必需另外尋求或研發適合的信號合成技術。最近我們發現 HNM 是一個不錯的基礎，可對它作改進而用來合成國語語音的信號。此外我們覺得語料庫爲基礎的研究方向裡，若語料錄得不夠多，則也可以考慮以 HNM 來取代 PSOLA。

HNM 是由 Y. Stylianou 所提出的一種語音信號的模型[5，6]，希望在作語音處理(編碼，合成)時，仍能保持信號的清晰度與自然度。HNM 可看成是弦波模型[7]的改進，它對於語音高頻部分的雜音(noise)信號成分，建立了較好的模型。HNM 的模型參數分析程序裡，提供了最大有聲頻率 MVF(maximum voiced frequency) 的一個偵測方法，依 MVF 值可將一個語音音框(frame)的頻譜(spectrum)分割成低頻、高頻之兩個部分，對於低頻部分的信號成分，採取以諧波成分(harmonic partials)的加總來模式化(modeling)，而對於高頻部分的信號成分，則採取以平滑的頻譜包絡(spectral envelope)來模式化，實際上是以少數的倒頻譜(cepstrum)係數來代表此頻譜包絡。

當應用 HNM 來合成國語語音時，我們發現有幾個議題，其解決方法並未能在 HNM 的文獻上找到，第一個議題是，如何讓合成的音節信號，保持音色(timbre)的一致性(consistency)，即音色一致性議題。由於我們只希望對各種國語音節錄製一次發音，然後透過修改一個音節的基週軌跡的高度及形狀，來合成出其它聲調的音節信號，因此當一個欲合成音節的基週軌跡被指定時，我們必需使用一種適當的方法來調整 HNM 各諧波成分的參數值，以同時滿足基週軌跡及音色一致性的要求。第二個議題是，對於一個放置於欲合成音節之時間軸上的一個控制點(control point) [8, 9]，如何決定此控制點上的 HNM 參數的數值，即參數值設定之議題。在合成一個音節的信號時，我們需要調整該音節原始錄音的音長以滿足韻律單元所指派的合成音節之音長，因此在合成音節時間軸上的一個控制點，當它被對映(mapping)至一個位於原始音節兩分析音框之間的時間點時，我們必需使用一種適當的內差方法來計算此控制點上的 HNM 參數值。此外，第三個議題是，如何校正(warp)合成音節的時間軸，以合成出較爲流暢(fluent)的音節及語句的信號，也就是時間軸校正的議題，此議題應是和語音合成較爲相關，而和 HNM 的相關性較少，當一個合成音節的音長需要伸長或縮短時，一個簡單的時間軸校正方法是線性校正(linear warping)，但此種作法通常會得到較差的流暢度。

本文研究了前述三項議題的解決方法，然後再據以建造出一個 HNM 改進、擴充的國語音節信號合成系統，此系統的主要處理流程如圖 1 所示。當要合成一個音節的信號時，很明顯地此音節的各個韻律參數值已經由韻律單元訂定、指派好了，因此圖 1 裡的第一個方塊首先作的是，將合成音節的音長規劃、分割成此音節的組成音素(phoneme)的時長(duration)，接著依據相連音素的時長來建造一個片斷線性(piece-wise linear)的時間校正函數，以便將合成音時間軸上的時間點對映至原始音的時間軸上；在圖 1 裡的第二個方塊，先均勻地在合成音的時間軸上佈放控制點，然後對各個控制點求取該點上的 HNM 參數值；接著在圖 1 裡的後面三個方塊，將信號分類成三種形態分別去作合成處理，對於短時間的無聲(unvoiced)聲母(syllable initial)，其信號片斷直接由原始音裡複製到合成音裡，對於長時間的無聲聲母，其信號則當作是 HNM 的雜音信號成分來作合成，至於音節的有聲(voiced)部分，包括有聲子音及母音，則先分別合成出諧波和雜音成分，再作相加。

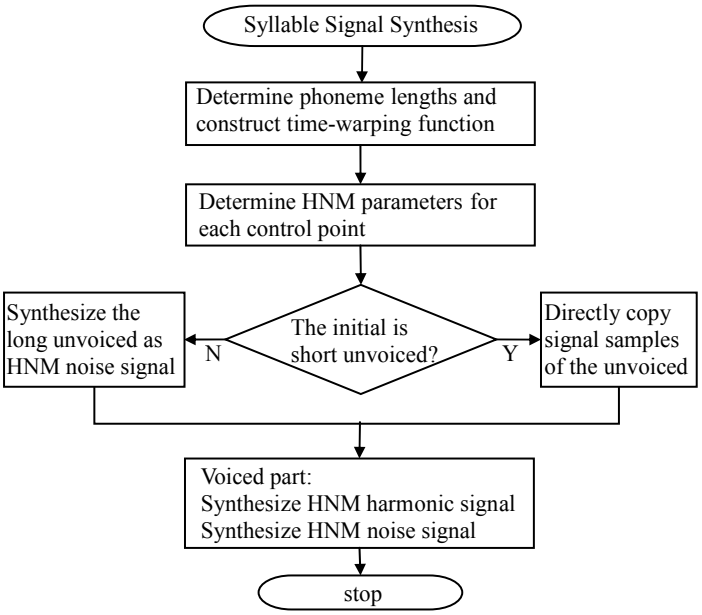


圖 1、基於 HNM 之音節信號合成方法的主要處理流程

二、音素時長規劃及時間校正函數

國語音節的結構可以看成是 C_xVC_n，其中 C_x 可以是有聲子音、無聲子音、或是空音(null)，而 C_n 可以是鼻音/n/或/ng/、或是空音，V 則可以單母音、雙母音、或是三母音。當 C_x 是一個無聲子音時，它可再被分類成短無聲(如/b/)或長無聲(如/p/)，然後進行圖 1 菱形塊右邊或左邊方塊的處理；此外當 C_x 是一個有聲子音(如/m/)或空音時，則我們把 C_x 或 V(當 C_x 是 null 時)的起始部分的信號當作是短無聲來看待及作處理(也就是走到圖 1 菱形塊右邊的方塊去)，這樣的對待方式是合理的，因爲以有聲音素開頭的音節，其起始部分的一、二個信號音框，在作 HNM 的參數分析時，也經常被判斷成非週期性的。

當一個音節是以廣義的短無聲子音開頭時，如/bau/及/man/，則此子音的時長規劃，就是直接依據原始錄音裡此子音的長度。相反地，如果音節是以長無聲子音開頭時，則此子音的時長規劃是，將原始錄音裡此子音的長度乘上一個比例值 *Fu*，*Fu* 的值基本上設定成合成音節的音長除以原始音節的音長，不過當 *Fu* 大於 1.4 時就改設爲 1.4，而當 *Fu* 小於 0.6 時就改設爲 0.6，這是因爲音節音長的伸長或縮短，主要是在母音部分進行，而非等比例的方式。在規劃了音節起始的無聲子音的時長 *Du* 之後，音節後面有聲部分的長度 *Dv*，很明顯地就是音節音長減去 *Du*。

接著，考慮音節有聲部分裡各音素的時長規劃，這裡以音節/man/爲例來說明，令/man/的原始錄音裡，音素/m/、/a/、/n/分別佔據的長度是 *Rm*、*Ra* 和 *Rn* 毫秒(ms)，且有聲部分的總長度是 *Rv* = *Rm* + *Ra* + *Rn*，另外令合成音節裡，這三個對應的音素的時長值分別是，*Dm*、*Da*、*Dn*，且令 *Dv* = *Dm* + *Da* + *Dn*，則我們規劃 *Dm*、*Da*、*Dn* 數值的作法就如下列的程序：

```
r = 0.6;
while ( r >= 0.1 ) {
   Dm = (Rm/Rv) * r * Dv;
   Dn = (Rn/Rv) * r * Dv;
```

```
    Da = Dv - Dm - Dn;
    if (Da > Dv*0.4) break;
    r = r - 0.05;
}
Db = Dm + Dn;
if (Dm > 0 && Dm/Db < 0.35) { Dm = 0.35*Db;  Dn=Db-Dm; }
if (Dn > 0 && Dn/Db < 0.35) { Dn = 0.35*Db;  Dm=Db-Dn; }
```

如果一個音節的結構是和/san/或/an/一樣的，也就是缺少有聲開頭的子音，則上列程序裡 *Rm* 和 *Dm* 的值可直接設爲 0；相同地如果音節的結構是和/ma/或/sa/一樣，即沒有結尾的鼻音，則上列程序裡 *Rn* 和 *Dn* 的值也可直接設爲 0。當 *Dm*、*Da*、*Dn* 的值設定好之後，就可以將合成音裡的有聲音素依序對應至原始音裡的有聲音素，而建造出如圖 2 所示的片斷線性之時間校正函數。

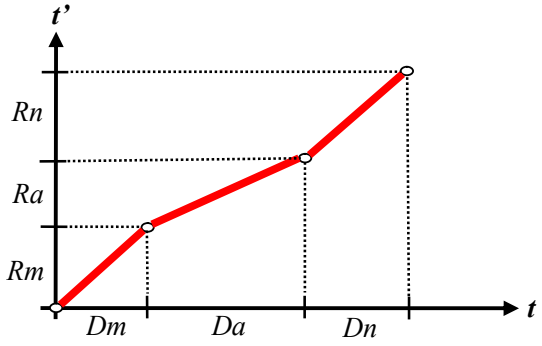


圖 2、片段線性之時間校正函數

上面的音素時長之規劃程序，其演算法是基於一個觀察，即有聲子音長度對於母音長度的比值 (*Rm* + *Rn*) / *Rv*，在句子裡發音節/man/的，會比單獨/man/音節發音的小許多。目前我們僅以簡單的程序來模擬此現象，並且時間校正函數也簡單地設定爲如圖 2 的片段線性之型式，未來我們將採取另一種較爲系統化的方法，來研究合成音節和原始音節之間的時間對映問題。不過，我們認爲即使只用簡單的時間校正函數，仍然可讓流暢度獲得明顯的改進。

三、控制點及其 HNM 參數的設定

3.1 控制點之佈放

我們錄製國語音節信號的取樣率是 22,050Hz，對於信號作 HNM 的參數分析時，設定音框的長度爲 512 個樣本點(23.2ms)，且音框每次前進 256 個樣本點。另一方面在作音節信號的合成時，我們採取了電腦音樂合成裡常用的觀念--控制點[8, 9]，這裡分別使用〝音框〞和〝控制點〞兩名詞，藉以指出在合成音節有聲部分的時間軸上所佈放的控制點，它上面的 HNM 參數，並不是從某一音框所分析出的 HNM 參數直接複製過來，而是拿兩個對應音框的 HNM 參數來作內差而求得的，不過在合成長時間的無聲子音時，我們就只有簡單地把一個音框分析出的 HNM 參數指派給一個對應的控制點。這兩種 HNM 參數的設定方式，以分別對付有聲部分和長時間無聲部分，就如圖 3 裡的圖形所表示的。

從圖 3 也可看出，在合成音的無聲子音部分，所佈放的控制點的數量，其實就是原始音裡無聲子音部分的音框的數量，因此合成音的無聲部分的時長調整，只是簡單地以線性

方式作伸長或縮短。然而在合成音的有聲部分，相鄰的控制點永遠是間隔 100 個信號樣本(4.5ms) ，因此控制點的數量是由所指派的有聲部分之時長來決定，至於選擇以 100 個樣本(而不用較大的數值)作間隔，是因爲我們希望對頻譜的演進(progressing)作較精細的控制。

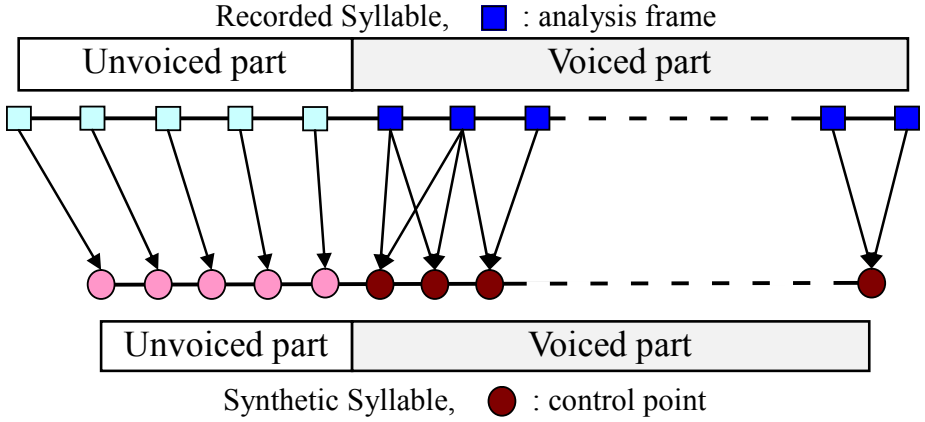


圖 3、控制點至分析音框之對映

3.2 音高(pitch)未變之 HNM 參數

要爲一個位於合成音有聲部分的控制點決定它的 HNM 參數值，第一步是依據如圖 2 所示的時間校正函數，將此控制點所在的時間位置 *ts*，對映至原始音時間軸上一個以音框爲單位的時間位置 *tr*；然後依據第⌊ *tr* ⌋和⌊ *tr* ⌋+1 音框兩者所分析出的 HNM 參數來作內差，以求得此控制點上的 HNM 參數，在此我們提出以線性內差的方式來求取各諧波的振幅、頻率、相位等三項參數，詳細的公式爲

$$\begin{aligned} \overline{A}_i &= (1-w)\cdot A_i^n + w\cdot A_i^{n+1} \\ n &= \lfloor t_r \rfloor, \quad w = t_r - n \end{aligned} \tag{1}$$

$$\overline{F}_i = (1-w)\cdot F_i^n + w\cdot F_i^{n+1} \tag{2}$$

$$\overline{\theta}_i = w\cdot(\hat{\theta}_i^{n+1} - \theta_i^n) + \theta_i^n \tag{3}$$

其中 A_i^n、F_i^n、θ_i^n 分別代表第 *n* 個分析音框裡第 *i* 個諧波的振幅、頻率與相位，而 $\overline{A}_i$、$\overline{F}_i$、$\overline{\theta}_i$ 則分別代表此控制點上第 *i* 個諧波的振幅、頻率與相位；此外，$\hat{\theta}_i^{n+1}$ 表示 θ_i^{n+1} 對於 θ_i^n 作反包裹(unwrap)運算後的相位值，也就是 $\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n)$，相位 θ_i^{n+1} 必需作反包裹運算以保證相位差值會落於-π到π之間，在此我們修正過的相位反包裹運算的作法是

$$\hat{\theta}_i^{n+1} = puw(\theta_i^{n+1}, \theta_i^n) = \theta_i^{n+1} - M\cdot 2\pi \tag{4}$$

$$M = \left\lfloor \frac{1}{2\pi}\left(\theta_i^{n+1} - \theta_i^n + \theta_c\right)\right\rfloor, \quad \theta_c = \begin{cases} \pi, & \text{if } \theta_i^{n+1} \geq \theta_i^n \\ -\pi, & \text{otherwise} \end{cases}$$

另外，由於一個音框作 HNM 參數分析後，會得到 10 個代表雜音成分的倒頻譜係數，因此，我們就拿兩個被對映到的相鄰音框所分析出的倒頻譜係數來作線性內差，而求得此控制點上的 10 個倒頻譜係數。

3.3 音高改變之 HNM 參數

在一個控制點上計算出原始音高(pitch)的諧波參數 $\bar{A}_i$、$\bar{F}_i$、$\bar{\theta}_i$ 之後，下一步要作的是，計算音高改變後的諧波參數 $\tilde{A}_k$、$\tilde{F}_k$、$\tilde{\theta}_k$。由於諧波頻率值 $\bar{F}_i$ 所定義的音高，其實是在音節錄音時就決定了，因此我們必需作音高的調整，以使連續的控制點各自的音高所串成的音高軌跡，能夠滿足韻律單元所指派的基週軌跡要求。在此假設諧波頻率值 $\bar{F}_i$ 所定義的音高爲 100Hz，而韻律單元要求的音高是 150Hz，那麼一個簡單的調整作法是，令 $\tilde{F}_k = \bar{F}_k \cdot 150/100$，而 $\tilde{A}_k = \bar{A}_k$ 且 $\tilde{\theta}_k = \bar{\theta}_k$，這就如圖 4 所畫的情況，由此圖可看出，音高的確可由 100Hz 調整到 150Hz，但是共振頻率(formant frequency)值也被調高了 1.5 倍，例如圖 4 的第一共振頻率 240Hz 被調高成爲 360Hz，這樣的共振頻率的改變，其後果是音色也被改變了，如果各個控制點上的頻率調整倍率高高低低不一致，則音色也會變來變去地不一致。

要在音色保持一致的前提下，調整一個控制點的音高，我們必需遵守的原則是，要讓頻譜包絡保持不變[8]，也就是頻率值被調到 $\tilde{F}_k$ 的第 k 個諧波的振幅 $\tilde{A}_k$，它的值必需從原始音高之諧波振幅值 $\bar{A}_i$ 所構建的頻譜包絡曲線中去內差出來，較詳細的作法是，先從舊的諧波頻率序列 $\bar{F}_1$，$\bar{F}_2$，$\bar{F}_3$，…中找出最靠近 $\tilde{F}_k$ 且比 $\tilde{F}_k$ 小的頻率值，令找出的是 $\bar{F}_j$，接著，就以 $\bar{F}_{j-1}$，$\bar{F}_j$，$\bar{F}_{j+1}$，$\bar{F}_{j+2}$ 四個原始音高之諧波頻率值和它們對應的振幅值，來作階數 3 之 Lagrange 內差，以求出頻率值 $\tilde{F}_k$ 上所應對應的振幅值 $\tilde{A}_k$，也就是依我們提出的公式(5)作計算，

$$\tilde{A}_k = \sum_{m=j-1}^{j+2} \bar{A}_m \cdot \prod_{\substack{h=j-1 \\ h\neq m}}^{j+2} \frac{\tilde{F}_k - \bar{F}_h}{\bar{F}_m - \bar{F}_h} \tag{5}$$

一個說明上述觀念(即在保持頻譜包絡不變的前提下調整音高)的圖形如圖 5 所示，在此圖裡，各諧波的頻率被調高了 1.25 倍，但舊諧波(直的實線)和新諧波(直的虛線)所構建的頻譜包絡曲線是重疊在一起的，如此就可確保音色的一致。

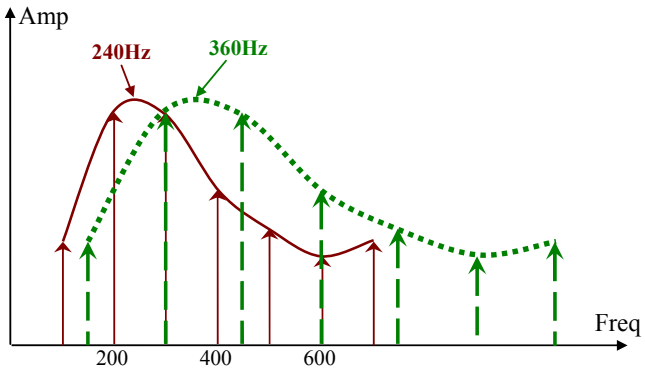


圖 4、音高和頻譜包絡一起被調升頻率值

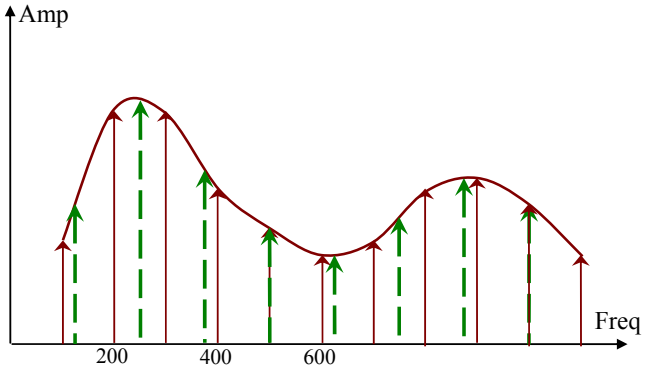


圖 5、音高調升頻率值而頻譜包絡不變

接著考慮頻率爲 $\tilde{F}_k$ 之新諧波的相位參數 $\tilde{\theta}_k$ ，在此我們一樣以頻率爲 $\bar{F}_{j-1}$， $\bar{F}_j$， $\bar{F}_{j+1}$，$\bar{F}_{j+2}$ 之四個舊諧波的相位值，來作 Lagrange 內差而計算出 $\tilde{\theta}_k$ 的值。不過，舊諧波的相位值在作內差之前必需先經過相位反包裹的運算，以避免相位值發生不連續的情況，詳細作法是，計算 $\hat{\theta}_{j-1} = \bar{\theta}_{j-1}$ ， $\hat{\theta}_j = puw(\bar{\theta}_j, \hat{\theta}_{j-1})$ ， $\hat{\theta}_{j+1} = puw(\bar{\theta}_{j+1}, \hat{\theta}_j)$ ， $\hat{\theta}_{j+2} = puw(\bar{\theta}_{j+2}, \hat{\theta}_{j+1})$ ，再拿計算出的相位值去作內差。

四、信號波形合成

在圖 3 裡合成音節的有聲部分，合成的信號 $S(t)$是由諧波信號 $H(t)$和雜音信號 $N(t)$相加而得到，也就是 $S(t) = H(t) + N(t)$。但是詳細說來，$H(t)$其實表示多個諧波信號成分的加總，而且 $N(t)$也是表示多個雜音信號成分的加總，在以下的兩個子節就分別說明 $H(t)$和 $N(t)$的合成方法。

4.1 諧波信號合成

對於位於第 n 和第 n+1 個控制點之間的諧波信號 $H(t)$，可以如下我們修正過的公式來計算它的樣本值，

$$H(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t)\right) \quad , \quad t = 0,1,...,99 , \tag{6}$$

$$a_k^n(t) = \tilde{A}_k^n + \frac{t}{100}(\tilde{A}_k^{n+1} - \tilde{A}_k^n) , \tag{7}$$

$$\phi_k^n(t) = \phi_k^n(t-1) + 2\pi f_k^n(t)/22,050 \quad , \quad \phi_k^n(0) = \hat{\theta}_k^n , \tag{8}$$

$$f_k^n(t) = \tilde{F}_k^n + \frac{t}{100}(\tilde{F}_k^{n+1} - \tilde{F}_k^n) , \tag{9}$$

其中 L 表示諧波成分(弦波)的個數，100 是相鄰控制點之間間隔的樣本數，22,050 是取樣率，$a_k^n(t)$ 表示第 k 個諧波在時刻 t (從第 n 個控制點算起之第 t 個樣本) 的振幅，$\phi_k^n(t)$ 表示第 k 個諧波累積到時刻 t 時的相位， $f_k^n(t)$ 表示第 k 個諧波的時變頻率，而

$\hat{\theta}_k^n = puw(\tilde{\theta}_k^n, \hat{\theta}_k^{n-1})$ ，也就是 $\tilde{\theta}_k^n$ 對 $\hat{\theta}_k^{n-1}$ 作反包裹運算後的相位。

當使用公式(6)來合成信號樣本時，累積的相位值 $\phi_k^n(t)$ 通常會在邊界的時間點(即 t=0 或 t=100 時)上發生不連續的現象，亦即 $\phi_k^n(100) \neq \phi_k^{n+1}(0)$ ，這會導致信號波形的不連續，而讓人聽到喀噠聲(click)。爲了避免此種相位的不連續，我們可先計算出在邊界時間點 t=100 時，兩者相差的相位量 ξ_k^n ，然後把差異的相位量平均地分配給相鄰控制點之間的 100 個樣本點，如此隨時間累積的相位就會變得連續了。在此我們所提用以計算 ξ_k^n 的公式是，

$$\xi_k^n = puw\left(\phi_k^n(100), \phi_k^{n+1}(0)\right) - \phi_k^{n+1}(0) \tag{10}$$

其中的相位反包裹運算 $puw(x,y)$就如公式(4)所定義的，而 $\phi_k^n(100)$ 可直接以如下我們所提的公式來計算出，

$$\phi_k^n(100) = \phi_k^n(0) + \frac{\pi}{22,050}(101\tilde{F}_k^{n+1} + 99\tilde{F}_k^n) \tag{11}$$

此公式是從公式(8)和(9)作反覆地疊代而得到。當把相位差異量 ξ_k^n 作平均分配之後，公式(6)就可以改寫成爲

$$H'(t) = \sum_{k=0}^{L} a_k^n(t) \cos\left(\phi_k^n(t) - \frac{t}{100} \cdot \xi_k^n\right) \quad , \quad t = 0,1,...,99 \tag{12}$$

關於 L 的值，令 L_n 表示第 n 個控制點上的諧波成分的個數，一般來說 L_n 和 L_{n+1} 的值可能會不相等，此時我們就令公式(6)和(12)中的 L 值爲 L_n 和 L_{n+1} 的較大者。當 L_n 比 L_{n+1} 小時，我們就需爲第 n 個控制點擴增出 L_{n+1} - L_n 個諧波成分及定義這些諧波的參數值，以便能夠套用前述的信號合成之公式，在此我們基於信號波形連續性的考慮，就簡單地定義 $\tilde{A}_k^n$=0, $\tilde{F}_k^n = \tilde{F}_k^{n+1}$, $\tilde{\theta}_k^n = \tilde{\theta}_k^{n+1}$, k = 1+L_n, 2+L_n, …, L_{n+1}。

4.2 雜音信號合成

對於位於控制點 n 和 n+1 之間的雜音信號 $N(t)$，在合成處理上我們以固定頻率間距的多個弦波信號成分的加總來計算 $N(t)$的樣本值[5]。令 G_k 表示第 k 個弦波的頻率，由於 G_k 值不會隨著時間 t 改變，因此不需區分是在那一個控制點上的，在此令 G_k 代表的值是 100·k (Hz)，這樣 G_k 的下標 k 的起始值就不是 1 了，且各控制點上的起始值也可能會不一樣，因此我們以 K_S^n 表示第 n 個控制點上下標 k 的起始值，它其實可由控制點的 MVF 值來計算得到，即 $K_S^n = \lceil \text{MVF}(n) / 100 \rceil$；此外，下標 k 的終值是一個固定值，$K_e = \lfloor 11,025 / 100 \rfloor$，因爲 G_k 不可大於取樣率的一半。

另外，在第 n 個和第 n+1 個控制點之間，第 k 個弦波的振幅我們以 $b_k^n(t)$ 表示，這表示它的值會隨著時間 t 在改變，實際上它是依據時間邊界點(即第 n 和第 n+1 控制點)上的振幅參數 B_k^n 來作線性調整的，至於 B_k^n 參數的求取方式是，先將第 n 個控制點上的 10 個倒頻譜係數補上一序列的零，使成爲 2048 個數值的序列，然後作反向 DFT(discrete Fourier trandform)轉換，取指數，而得到頻譜係數，X_j, j=0, 1, …, 2047，接著依 G_k 找出

兩個相鄰的 X_j，其下標 j 代表的頻率值會包含 G_k，然後對這兩個 X_j 作線性內差，來求出 B_k^n 的值。

當相鄰的兩個控制點上的弦波起始編號 K_s^n 和弦波振幅值 B_k^n 都算出之後，接著就可用這些參數值來合成這兩個控制點之間的雜音信號，我們修正後的計算公式爲，

$$N(t)=\sum_{k=K_s}^{K_e} b_k^n(t)\cos\left(\gamma_k^n + t\cdot 2\pi G_k/22{,}050\right) \quad , \quad t=0,1,...,99\,, \tag{13}$$

$$b_k^n(t)=B_k^n+\frac{t}{100}(B_k^{n+1}-B_k^n)\,, \tag{14}$$

$$\gamma_k^n=\gamma_k^{n-1}+100\cdot 2\pi G_k/22{,}050\,, \tag{15}$$

其中 K_s 設定成 K_s^n 和 K_s^{n+1} 中的較小者，γ_k^n 表示第 n 個控制點上第 k 個弦波的初始相位值。

關於圖 1 和 3 裡長時間無聲子音信號的合成，公式(13)、(14)和(15)仍然可被使用，不過，公式(13)裡的下標 k，它的起始值就要改成 1 了，這相當於設定 MVF 之值爲 0Hz。

五、信號合成實驗和聽覺測試

由於數年前我們曾提出一個改進的 PSOLA 變種之合成方法，稱爲 TIPW[10]，因此我們想要比較 TIPW 法和本文研究的 HNM 擴充之合成法，兩者在信號清晰度上的表現，如果一個合成的語音信號聽起來較不吵雜(noisy)、較無迴音(reverberant)，則它可說是具有較高的清晰度。這裡我們主要關心的是合成信號的清晰度，因此在文句分析和韻律參數產生方面就使用相同的程式模組[11, 12]。至於語音單元，兩種合成法一樣都使用國語音節爲單元，並且都不作單元選擇，因爲每一種國語音節都只存了一遍平調的發音。當使用一台 CPU 爲 Pentium 2.6GHz 的電腦來作語音信號的合成處理時，這兩種合成方法都可以被即時地執行，不過執行速度是有差異的，HNM 擴充之合成法的執行速度只有 3 倍的即時速度，即合成 30 秒的語音信號需花費 10 秒的 CPU 時間，而 TIPW 合成法的速度可以快到 20 倍的即時速度，即合成 30 秒的語音僅需花費 1.5 秒的 CPU 時間。

首先以觀察聲紋圖(spectrogram)的方式來比較這兩種合成方法。兩方法分別去合成出國語短句〝旋轉力〞/syuen-2 zhuan-3 li-4/的語音信號，然後以聲紋分析軟體(在此使用 wavesurfer)作分析來得到聲紋圖，圖 6 的聲紋是對 HNM 擴充法所合成的信號作分析而得到，圖 7 的則是對 TIPW 法合成的信號作分析而得到。從圖 6 和 7 我們可觀察倒，圖 7 裡的諧波紋路顯得比圖 6 裡的較爲零碎、較多斷裂的地方，並且圖 6 裡的諧波條紋較爲平滑，而不像圖 7 裡的顯得有一些毛燥、扭曲，因此 HNM 擴充法合成出的信號應會比 TIPW 法的清晰。

此外，我們選了一篇短文來讓這兩種方法去合成出語音信號，並且存成波形檔案，短文是一篇小學生的作文，有 132 個音節。接著我們將這兩個波形檔以隨機次序播放給 15 位參加聽覺測試者聆聽，然後請他們對前、後播放的檔案作清晰度的比較，評分的規則是，兩者無法區分時給 0 分，如果後者(前者)比前者(後者)稍好一些，則給 1 分(-1 分)，

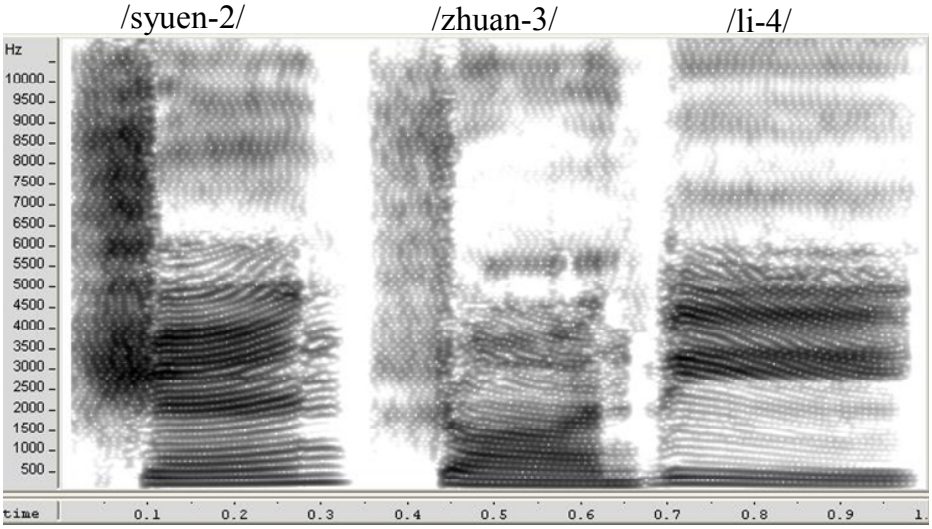


圖 6、HNM 擴充法所合成信號的聲紋圖

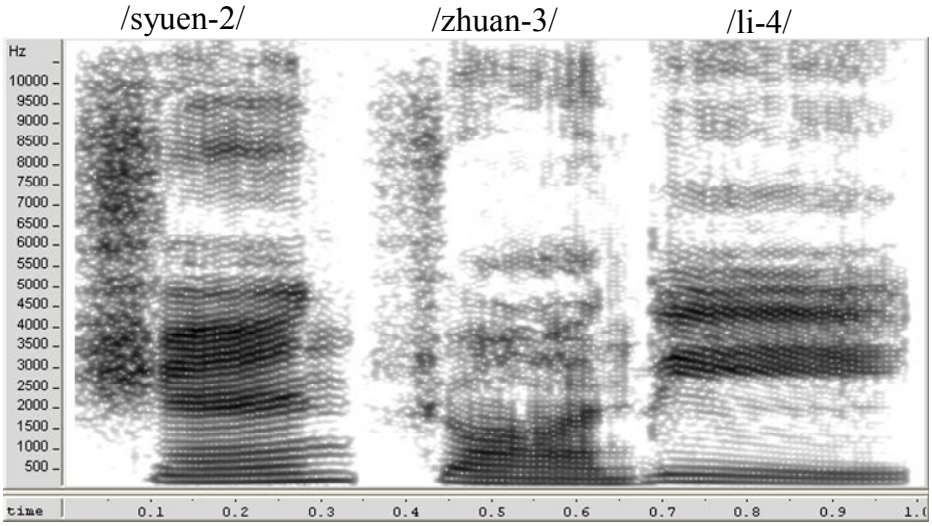


圖 7、TIPW 法所合成信號的聲紋圖

而如果是明顯地好或好很多則給 2 分(-2 分)，結果我們得到的平均分數是 1.53 分，也就是 HNM 擴充法會合成出比較清晰的語音。另外，爲了讓有興趣者能夠試聽這兩種方法所合成出的語音信號，我們設定了一個網頁以供人瀏覽，其網址是 http://guhy.csie.ntust.edu.tw/hmtts/hnm-demo.html。

六、結語

本文以 HNM 爲基礎，考慮了三個議題，即(1)如何保持音色的一致性，(2)如何決定控制點上的 HNM 參數值，(3)如何校正合成音節的時間軸。由於每一種國語音節(不區分聲調)，都只錄、存一次發音而已，所以必需作基週軌跡的調整，這就引發了音色一致性的問題；再者，合成音節的音長可能和原始音節的音長相差很多，如此在合成音節時

間軸上均勻佈放的控制點，各點上的 HNM 參數就會有數值訂定的問題；此外，要提升合成音節的流暢度，就會牽涉到合成音節時間軸的校正問題。本文對前述的三個議題，分別提出了可行的解決方法，在此稱它們整體爲 HNM 擴充之音節信號合成法。

爲了檢驗所提出的方法的效能，我們已把它製作成可即時執行之軟體，然後將它合成出的語音波形，和另一種 TIPW 法所合成的，去作聲紋比較和聽覺測試，初步結果顯示，本文提出的合成方法的確可明顯地提升合成語音的品質(較清晰、無迴音)。

雖然本文所提的 HNM 擴充之合成法，已可明顯地提升合成語音的清晰度，但是所合成的、供聽測用的語音信號檔案，聽起來仍然令人覺得有很強的機器味道，其主要原因應是韻律參數值的產生模型不佳所造成，例如音節的音量和音長數值只是依據幾個簡單規則來作決定而已，因此未來我們將再研究韻律模型改進的問題。

參考文獻

[1] Moulines, E. and E Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, *Speech Communication*, Vol. 9, pp. 453-467, Dec. 1990.

[2] Dutoit, T., *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[3] Chou, Fu-chiang, *Corpus-based Technologies for Chinese text-to-Speech Synthesis*, Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[4] 張唐瑜, *以大量詞彙作爲合成單元的中文文轉音系統*, 碩士論文, 國立中興大學資訊科學研究所, 2004.

[5] Stylianou, Yannis, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[6] Stylianou, Yannis, “Modeling Speech Based on Harmonic Plus Noise Models”, *Nonlinear Speech Modeling and Applications*, Springer-Verlag, Germany, 2005.

[7] Quatieri, T. F., *Discrete-Time Speech Signal Processing*, Prentice-Hall, NJ, USA, 2002.

[8] Dodge, C. and T. A. Jerse, *Computer Music: Synthesis, Composition, and Performance*, second edition, Schirmer Books, New York, 1997.

[9] Moore, F. R., *Elements of Computer Music*, Prentice-Hall, 1990.

[10] Gu, H. Y. and W. L. Shiu, “A Mandarin-Syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control”, *Proceedings of the National Science Council ROC(A)*, Vol. 22, No. 3, pp. 385-395, 1998.

[11] Gu, H. Y. and C. C. Yang, “A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech”, *International Symposium on Chinese Spoken Language Processing*, Beijing, China, pp. 125-128, 2000.

[12] Gu, H. Y., Y. Z. Zhou, and H. L. Liau, “A System Framework for Integrated Synthesis of Mandarin, Min-nan, and Hakka Speech”, to appear in *International Journal of Computational Linguistics and Chinese Language Processing*.

The Role of Sound Change in the Speech Recognition System: A Phonetic Analysis of the Final Nasal Shift in Mandarin

楊孝慈　James H. Yang
國立雲林科技大學應用外語系
Department of Applied Foreign Languages
National Yunlin University of Science and Technology
yanght@yuntech.edu.tw

摘要

過去十幾年來，電腦語言學家一直嘗試要設計一個能辨識標準語的語音系統，並能兼顧個別使用者的語音差異。目前，一些語音辨識系統已能針對使用者的年齡和性別，處理不同的音頻問題。然而，語音變化的問題還沒納入考慮。因此本文提議將語音變化的現象，整合到語音辨識系統的設計。本文以中文的字尾鼻音變化爲例，分析它的語音特質，並討論它對語音辨識系統的影響，最後提出一個能提升辨識字尾鼻音的解決方案。

Abstract

Over the past decade, computational linguists have been striving to design a speech recognition system that is able to identify standard speeches and to accommodate sound variables caused by different individual accents. Furthermore, some speech recognition programs have been able to learn and identify distinctive sound frequencies due to the user's age and gender. Nevertheless, regular sound alterations that occur in different varieties of a language have never been seriously considered in the design of the speech recognition system. Accordingly, this study proposes to incorporate the socio-phonological information about regular sound modifications to enhance the performance of Automatic Speech Recognition. To illustrate this point, this study investigates and analyzes the acoustic variation of the syllable-final nasal shift from the velar to the dental, which has been discovered to be one of the distinctive sound features that makes the variety of Mandarin spoken in Taiwan differ from that spoken in China. Following the phonetic analysis, this study discusses the effect of the nasal merger on the development of phonology-dependent speech technologies. It concludes by proposing a preliminary resolution to the identification of syllable-final nasals for the design of Automatic Speech Recognition.

關鍵詞：中文語音學，語音變化，鼻音，語音辨識系統
Key words: Mandarin phonetics, sound change, nasal, speech recognition system

1. Introduction

My motivation to explore the nasal merger of Mandarin spoken in Taiwan originates from an incident in my life. My brother's first baby was born on August 31st, 1999. He gave his son a name called *Geng-ren* /kəŋ.ʐən/[1] (耕仁, meaning "to cultivate benevolence"). Interestingly, I found, as a native speaker of Mandarin in Taiwan, that I would easily mispronounce his name as *Gen-ren* [kən.ʐən] (跟人, meaning "to follow people"), rather than its standard pronunciation. Since the name was subject to mispronunciation and misunderstanding, I suggested to my brother that he should change the name; hence, he later selected another name *Jia-he* (家和, meaning "harmony in the family"). Because of this interesting incident, I came to realize that many native speakers of Mandarin in Taiwan seem to merge the syllable-final velar nasal /ŋ/ with the dental nasal /n/, hence neutralizing such minimal pairs as *geng* /kəŋ/ (耕, "to cultivate") and *gen* /kən/ (跟, "to follow"). To explore this possible sound change, I later conducted a speech production experiment, which is discussed in the subsequent section.

2. Speech Production Experiment

To investigate the possible nasal merger observed above, I addressed three research questions:

1) Is the syllable-final nasal modification a free variation or a conditioned alteration?
2) Does it occur in Mandarin spoken in Taiwan, China, or both?
3) Is it an ongoing or complete sound change?

To address these questions, I invited 30 native Mandarin speakers to participate in the speech production experiment, including 11 males and 19 females, who were students of the University of Hawaii at Manoa. Fifteen of them were from Taiwan, and another fifteen were from China. They were all young adults with the average age of 27, the eldest subject being 36 years old and the youngest one being 21.

For this experiment, I designed a questionnaire to understand the subjects' basic sociolinguistic backgrounds. In addition, I also created sixty easy and interesting riddles to elicit spontaneous speech data. The answers to the riddles included the test words needed for this study--that is, words which end with three types of rhymes: –ing, -eng, and –ang. Three samples of the riddles are displayed below:

[1] The character *geng* (耕) is *phonemically* transcribed as [gəŋ] in *Guó-tái Shuāngyǔ Cídiǎn* [Mandarin-Taiwanese Dictionary], edited by Xing-chu Yang (1992).

Table 1. some examples of the riddles used for speech data collection

Riddle	Answer	Pinyin
晚上會發光的昆蟲，是什麼蟲? (What kind of insect lights in the evening?)	螢火蟲 (Firefly)	Yíng-huǒ-chóng
在海邊指引船出入港的，是什麼塔? (What kind of building stands on the coast and guides boats in and out of a harbor?)	燈塔 (Lighthouse)	Dēng-tǎ
全壘打是什麼球類運動的用語? (“Home run” is a term of what sports?)	棒球 (Baseball)	Bàng-qíu

The test words were randomly mixed with 10 irrelevant words in order to avoid the subject's awareness of what words were being examined. I interviewed each participant at a time and tape-recorded each interview. The subject was first asked to answer the questionnaire which included questions concerning his or her basic sociolinguistic background. Next, the interview proceeded with a riddle game, which was carried out in a relaxed atmosphere to collect data from the subject's virtually spontaneous utterances. The informant was told to give the answer to every riddle as soon as possible. If having no idea about the answer, the informant would be given clues to say the test word. Furthermore, if the response was perceived to be not loud enough for sound analyses, the informant would also be asked to say the answer once more and aloud. The recorded data were later analyzed on the computer using Praat, a sound-editing program. The findings are presented in the ensuing section.

3. Findings

Regarding the first research question as to whether the final nasal shift from the velar to the dental (/ŋ/>/n/) occurs without syllabic constraints or appears only in certain environments, the results show that the nasal merger is not a free variation, but a conditioned sound change, which can be formulated by the following phonological rule:

(1) Nasal Fronting:
/ŋ/→[n]/{i, ə}_____.

For example, the word 經 *jing* (/tɕiŋ/, “pass”) is regularly pronounced by the Taiwanese respondents as 金 *jin* (/tɕin/, “gold”) according to Rule 1. Furthermore, this change in the rhyme also causes lexical neutralization. For instance, the word *jing-yu* (鯨魚, whale) is recurrently pronounced by the Taiwanese respondents as *jin-yu* (金魚, goldfish). Consequently, this nasal merger leads to lexical neutralization, creating homophones sharing the nasal endings with different meanings.

Notably, the nasal merger in Mandarin spoken in Taiwan is not only conditioned by the preceding vowel (Rule 1) but is also blocked by the bilabial onset, which is regularized by the following rule:

(2) Vowel Labialization:
/əŋ/→[oŋ]/[labial] _____

To my knowledge, such discovery has not been specifically analyzed in any previous studies. For instance, the sound *meng* (/məŋ/) is pronounced as *mong* (/moŋ/), rather than *men* (/mən/) according to Rule 2, the vowel labialization rule. Obviously, this sound modification displays articulatory assimilation because the vowel is labialized due to the influence of the initial bilabial consonant /m/. This sound change, however, might not merely occur because of sound assimilation, but it might also exist to constrain the creation of homophones; for example, if the word *meng* (/məŋ/, 夢, dream) is changed according to the nasal fronting rule, it would become neutralized with another word *men* (/mən/, 們, depressed), thus resulting in homophones. By contrast, if the word *meng* is changed according to the vowel labialization rule, it would be pronounced as *mong* (/moŋ/), which does not appear in Mandarin vocabulary. Therefore, the vowel labialization rule does not simply occur for ease of articulation but may also fill the vocabulary gap while avoiding creating homophones.

In addition, the results demonstrate that the syllable-final nasal shift described above occurs mainly in Mandarin spoken in Taiwan, instead of China. Specifically, when the preceding vowel is /i/, the final nasal merger occurs 96 percent of the time in the native speakers of Mandarin from Taiwan (MT). By contrast, the final nasal alteration takes place only 38 percent of the time in the native speakers of Mandarin from China (MC). Moreover, when the preceding vowel is /ə/, the nasal shift occurs 95 percent of the time in MT, while only 3 percent of the time in MC. Nevertheless, the nasal merger never occurs when the preceding vowel is /a/. The occurrence percentage of the nasal merger in Mandarin is displayed as follows:

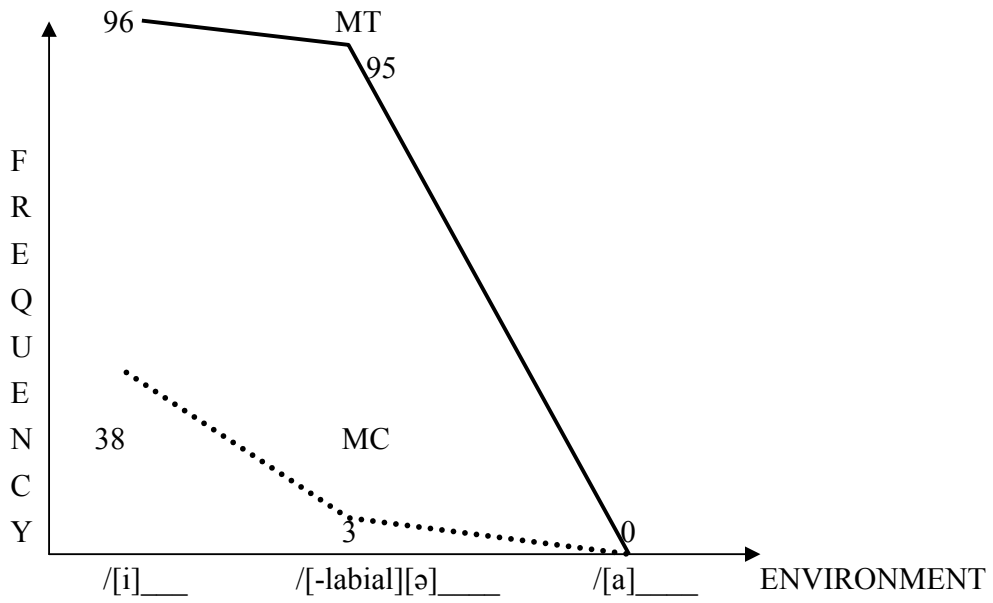


Figure 1 Occurrence percentage of the syllable-final nasal merger in three environments[2]

As Figure 1 displays, the syllable-final velar nasal /ŋ/ in MT merges nearly completely with the dental nasal /n/ when preceded by /i/ or /ə/. By comparison, MC in general does not undergo the nasal shift.

Taken together, all of the Taiwanese respondents underwent the nasal fronting (Rule 1). More than 95 percent of the time they displayed the nasal shift when the final nasal was preceded by the vowel either /i/ or /ə/. By comparison, Tse's 1992 survey suggests that 73% of his Taiwanese informants could not distinguish the syllable-final nasal minimal pairs. Accordingly, the final velar nasal merger with the dental has evolved into a nearly complete status in MT.

However, the final nasal modification described above only appears sporadically in MC. Specifically, Rule 1 occurs only 20 percent of the time in the Chinese responses. Comparatively speaking, Rule 1 occurs more than 95 percent of the time in the Taiwanese responses, making MT differ significantly from MC ($P<0.05$). While some of the Chinese informants from southern China also consistently undergo the nasal shift , the findings do not allow this study to conclude that the final nasal modification occurs regularly in southern China. First, although three of the speakers from southern China recurrently displayed the nasal merger, a couple of them were able to pronounce the test words according to the standard pronunciations without changing the final velar nasal into the dental one. Additionally, the number of the speakers from southern China was very few; only five informants participated in this study. Accordingly, the nasal shift from the velar to the dental seems to present an ongoing phonological process of confusion and interchange in

[2] The percentage of the nasal merger is obtained by excluding the pronunciations that follow Rule 2.

southern China. To confirm whether the nasal merger is common in southern China, an empirical and quantitative study is needed in the future.

In summary, the results demonstrate that the nasal shift (Rule 1) is nearly complete in MT, leaving very few lexical "residues," which are usually high frequency words, such as *sheng* (生, life) and *qing* (清, clear), instead of *shen* (身, body) and *qin* (親, kin), respectively. Theoretically speaking, this nasal shift is consistent with Zee's prediction of the nasal shift from –ng to –n in Chinese dialects (1985) and is opposed to M. Chen's theory of unidirectionality of the nasal shift from –n to –ng (1972, 1973, 1975). However, the syllable-final velar nasal remains unchanged when preceded by the non-palatal low vowel /a/.

4. Phonetic Analyses

The perceived transcriptions of the nasal shift in question are also supported by acoustic analyses. First, the spectrogram of the word *jing-yu* (鯨魚, whale) pronounced by the Taiwanese Mandarin speaker shows that the phonological change is a syllable-final nasal shift, rather than a nasal deletion with the preceding vowel nasalized, as shown below:

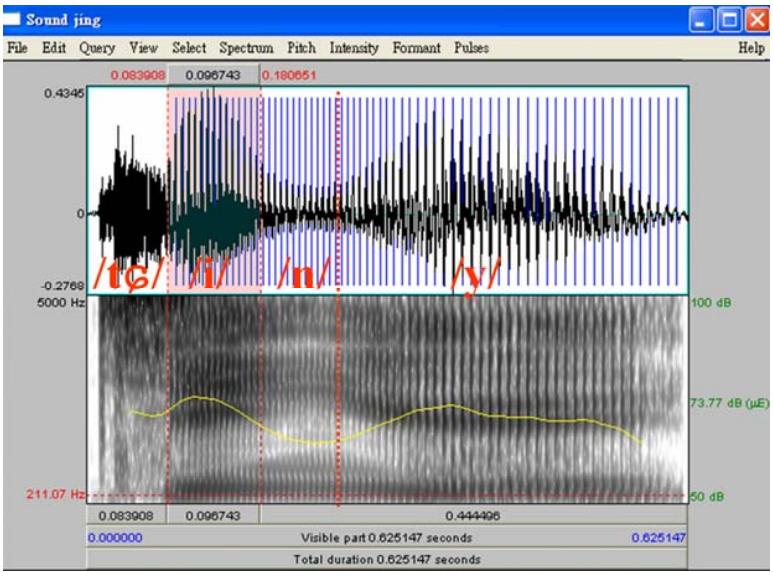


Figure 2. Spectrogram of the test word *jing-yu* (鯨魚, whale) pronounced by the Taiwanese informant

The darker parts of the sound reflect the spectrograms of the vowels, while the whiter parts signal those of the consonants (Ladefoged, 2003). In addition, the sound analysis via Pratt (a computer program for sound analyses) demonstrates that the syllable-final nasal still remains as shown in the middle whiter part of Figure 2.

The speculation that the sound alteration in question is a nasal deletion with the preceding vowel nasalized can also be cleared by comparing the spectrograms of the two

sounds *ǰī-yu* and *ji-yu*, as shown below.

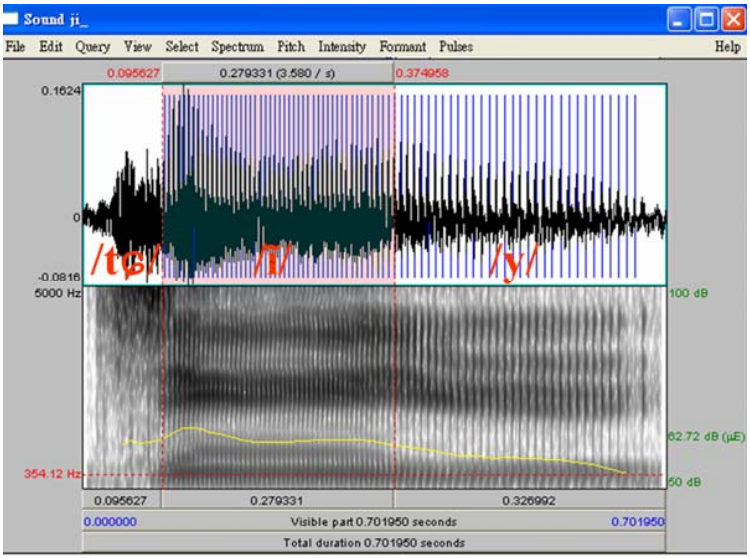


Figure 3. Spectrogram of the sound *ǰī-yu*

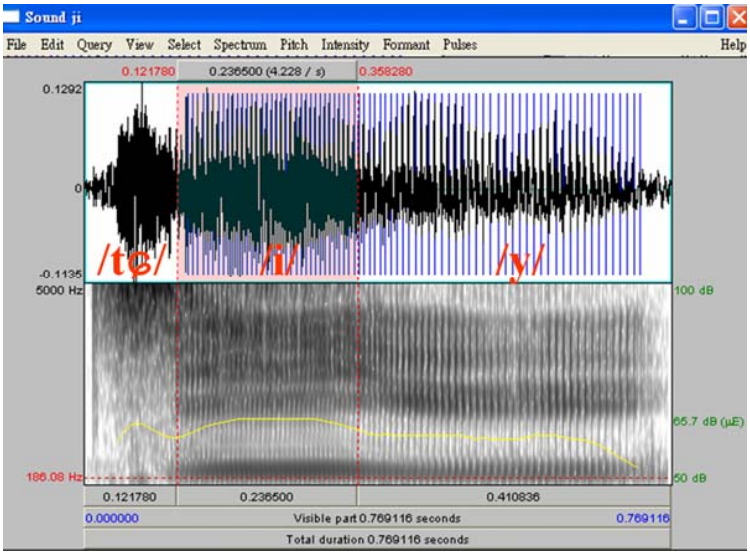


Figure 4. Spectrogram of the sound *ji-yu*

At a glance, none of Figures 3 and 4 look similar to Figure 2. Figure 3 (the spectrogram of the sound *ǰī-yu*) does not display any whiter part between the vowels /i/ and /y/, but only parallel lines equally black present before the vowel /y/. By contrast, Figure 2 exhibits a whiter part between the two vowels /i/ and /y/. Therefore, the nasal does not disappear but remains.

Furthermore, Figure 4 (the spectrogram of the sound *ji-yu*) also looks distinct from Figure 2. Because the spectrogram of the word *jing-yu* (鯨魚, whale) pronounced by the

Taiwanese informant neither look similar to the spectrogram of the sound *ji̍-yu*, nor does it look similar to that of *ji-yu*, the sound modification in question should not be the nasal deletion.

Nonetheless, the spectrogram of the word *jin-yu* (金魚, goldfish) pronounced by the same Taiwanese informant looks similar to Figure 2, although their durations and intensities are slightly different, as presented below:

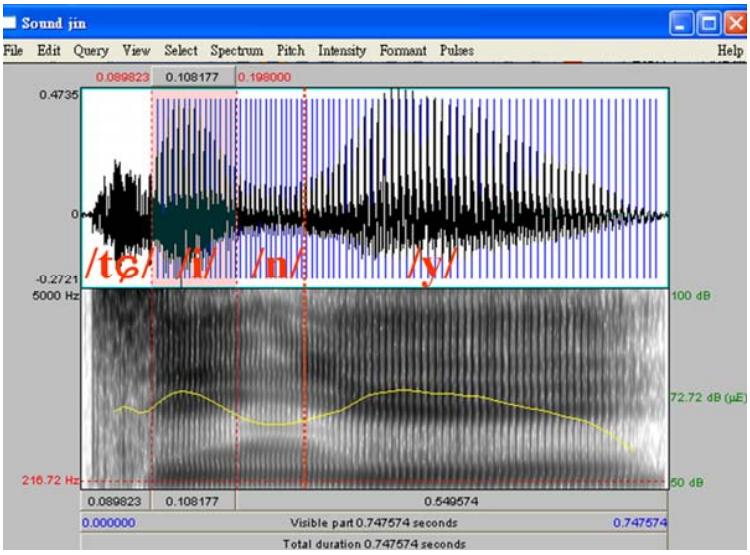


Figure 5. Spectrogram of the test word *jin-yu* (金魚, goldfish) pronounced by the same Taiwanese informant

The spectrogram similarity between Figures 2 and 5 indicates that the Taiwanese Mandarin speaker does not distinguish the minimal pairs *jing-yu* (鯨魚, whale) and *jin-yu* (金魚, goldfish), but pronounces them the same, changing the final velar nasal into the dental. Furthermore, this acoustic analysis is also supported by the discovery from my interview with the Taiwanese informant after the riddle game.

By comparison, the spectrogram of the word *jing-yu* (鯨魚, whale) pronounced by the Chinese informant is distinct from the spectrogram of the same word pronounced by the Taiwanese informant, as shown below:

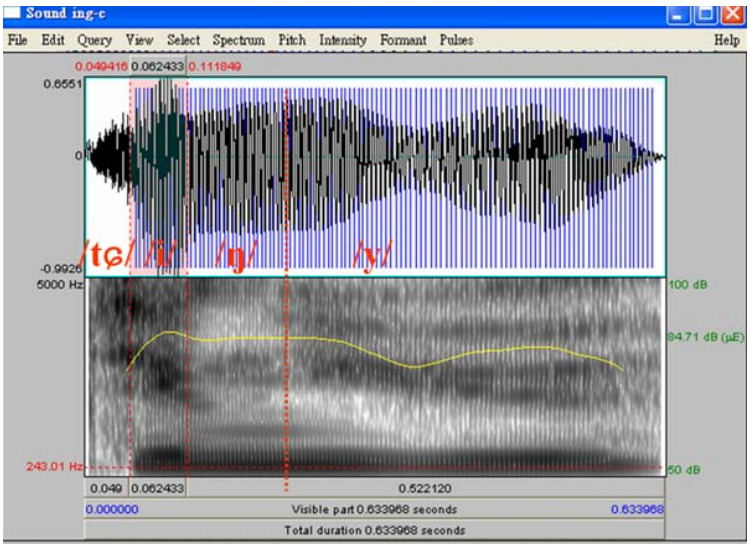


Figure 6. Spectrogram of the test word *jing-yu* (鯨魚, whale) pronounced by the Chinese informant

Apparently, Figure 2 does not look similar to Figure 6; accordingly, it is safe to infer that the final velar nasal does not retain in MT, but must shift into another, either the bilabial or the dental.

Finally, the comparison between Figure 2 and the spectrogram of the bilabial nasal demonstrates that the nasal does not shift to the bilabial because the two spectrograms are distinct from each other, as shown in Figure 7:

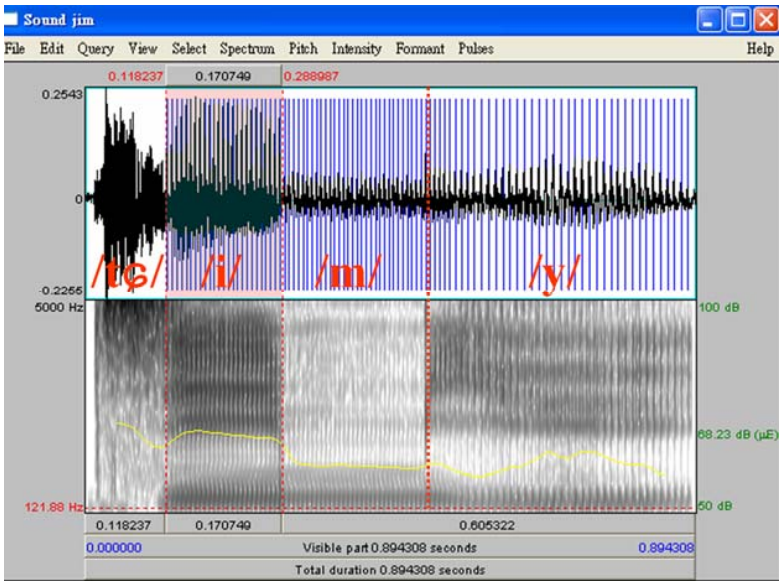


Figure 7. Spectrogram of the sound jim-yu

The spectrogram comparison between Figures 2 and 7 eventually assures us that the syllable-final velar nasal does not switch to the bilabial but the dental.

To summarize, the spectrograms presented above indicate that the syllable-final nasal changes from the velar to the dental when the preceding vowel is /i/ or /ə/, instead of vanishing with the preceding vowel nasalized.

5. Implications for Automatic Speech Recognition

As phonetically presented above, the syllable-final velar nasal shift to the dental regularly occurs in the variety of Mandarin spoken in Taiwan. While this nasal merger in phonologically independent words might not influence computer intelligibility, it might interfere with the computer word-identification process of the syllable-final nasal minimal pairs. For instance, because of the nasal shift, Taiwanese Mandarin speakers tend to say the word *gao-xing* as *gao-xin*, changing the velar nasal into the dental one. This sound alteration might not mislead the speech recognition system to misidentify *gao-xing* (高興) as *gao-xin* (高信) because no such word as *gao-xin* (高信) exists in Mandarin. Nonetheless, this nasal shift has resulted in considerable homophones, neutralizing many minimal pairs, such as *jing-yu* (鯨魚, whale) and *jin-yu* (金魚, goldfish), and *sheng-gao* (升高) and *shen-gao* (身高). Consequently, when a Taiwanese Mandarin speaker says *jing-yu* (鯨魚, whale) as *jin-yu* (金魚, goldfish), the system might misidentify the intended word as *jin-yu* (金魚, goldfish) rather than *jing-yu* (鯨魚, whale), hence retrieving the wrong word. To resolve word-identification problems like this, a possible approach might be to train the user to produce the standard pronunciations of the minimal pairs differing only in nasal endings. Specifically, the user might be trained to articulate the minimal pairs as the standard pronunciations so that the recognition system is able to retrieve the right words.

Such sound training, however, is neither easy nor realistic. While the speech recognition program may provide the standard pronunciations of the syllable-final nasal minimal pairs for the user to imitate and articulate, there is evidence that most people tend to mishear new sounds and mispronounce them according to their habitual articulation (Ohala, 1992, 2001). Furthermore, some recent research has demonstrated that Taiwanese Mandarin speakers tend to preserve their distinctive and unique sound features as their Taiwanese identities, instead of following the standard pronunciations associated with China (Hsu, 2005). Accordingly, a viable resolution might program the automatic speech recognition to distinguish syllable-final nasal minimal pairs, as represented in Table 2.

Table 2. Example of the minimal pairs that differ in the syllable-final nasals following the vowels /i/ or /ə/

-ing/-in	-eng/-en
qing-xin/qin-xin 輕信/親信	cheng-jio/chen-jio 成就/陳舊
ying-qi/yin-qi 英氣/陰氣	sheng-gao/shen-gao 升高/身高
jing-ying/jin-yin 經營/金銀	seng-qing/shen-qin 聲請/申請
xing-xiang/xin-xiang 星象/新象	zheng-zhi/zhen-zhi 整治/診治
jing-yu/jin-yu 鯨魚/金魚	zheng-feng/zheng-fong 政風/陣風

Table 2, however, is not a complete list. Further research is needed to include as many minimal pairs of the syllable-final nasals as possible in the corpus of the speech recognition system.

Most crucially, research into the context where minimal pairs differing in nasal endings is needed to enhance the performance of word identification. For example, when the user says the word *ying-qi* (英氣) as *yin-qi* (陰氣), the speech recognition system might retrieve both of the words for the user to choose. Nonetheless, if we would like to advance automatic speech recognition, we need to investigate the discourse where each of the words occurs. For instance, the word *ying-qi* (英氣) often collocate with such words as *huan-fa* (煥發) and *lin-ran* (凜然), while the word *yin-qi* (陰氣) does not. Furthermore, the word *ying-qi* (英氣) often occurs when the discourse includes words like *zhen-pai* (正派), *hao-shuang* (豪爽), and *guang-ming* (光明). Therefore, future research into the discourse where syllable-final nasal minimal pairs occur will facilitate the speech recognition system to retrieve the intended words.

To conclude, this study has analyzed the phonetic attributes of the syllable-final nasal shift from the velar to the dental that occurs nearly completely in Mandarin spoken in Taiwan. To accommodate this nasal shift, a computational linguist needs to include as many minimal pairs ending with the nasal rhymes as possible in the corpus of the speech recognition system. In addition, to improve the efficiency of the word identification, this study has also proposed to investigate the discourse where each of the minimal pairs appears. In short, future research into the syllable-final nasal minimal pairs and the context of their usages is needed to enhance the identification accuracy of the speech recognition system.

References

[1] J. K. Tse, "Production and perception of syllable final [n] and [ŋ] in Mandarin Chinese: An experimental study.," *Studies in English Literature,* pp. 143-56, 1992.

[2] E. Zee, "Sound change in syllable-final nasal consonants in Chinese," *Journal of Chinese Linguistics,* vol. 13, pp. 291-330, 1985.

[3] M. Chen, "Nasals and nasalization in Chinese: Exploration in phonological universals," University of California at Berkeley, 1972.

[4] M. Chen, "Cross-dialectal comparison: A case study and some theoretical considerations," *Journal of Chinese Linguistics,* vol. 11, pp. 38-63, 1973.

[5] M. Chen, "An areal study of nasalization in Chinese," *Journal of Chinese Linguistics,* vol. 3, pp. 16-59, 1975.

[6] P. Ladefoged, *Phonetic data analysis: an introduction to instrumental phonetic fieldwork*. Oxford: Blackwells, 2003.

[7] J. J. Ohala, "What's cognitive, what's not, in sound change," *Lingua e Stile,* vol. 27, pp. 321-362, 1992.

[8] J. J. Ohala, "An Account of Sound Change," University of California, Santa Barbara: The LSA Institute, 2001.

[9] H.-J. Hsu, "The leveling of neutral tone in Taiwan Mandarin and its new identity," in *2005 international conference of applied linguistics* Chia-yi, Taiwan, 2005.