專利雙語語料之中、英對照詞自動擷取

Automatic Term Pair Extraction from Bilingual Patent Corpus

曾元顯　Yuen-Hsien Tseng
國立臺灣師範大學
資訊中心
Information Technology Center
National Taiwan Normal University
samtseng@ntnu.edu.tw

劉昭麟　Chao-Lin Liu
國立政治大學
資訊科學系
Department of Computer Science
National Chengchi University
chaolin@nccu.edu.tw

莊則敬　Ze-Jing Chuang
威知資訊股份有限公司
WebGenie Information Ltd.
terry@webgenie.com.tw

摘要

針對 50 多萬筆的專利中英雙語語料，本文提出兩種翻譯對照詞彙的自動擷取方案，一種是精確導向、另一種是召回導向。在精確導向的方案中，我們提出了一種詞彙擷取方法，並比較了六種詞彙對列作法，以實際資料驗證，得出可供參考的經驗。我們發現 EM（Expectation Maximization）方法效果最好，但其最花時間，也難以找出多對多的同義翻譯。而即便是最差的 MI（mutual information）法，其排序在前頭的正確詞對跟 EM 法不同，因此可以作爲輔助的詞對擷取方法，爲後續合併或混用多種對列方式的研究，開啓了可能性。在召回導向的方案中，我們提出了簡單的想法與有效的實做，可從雙語對列語料庫中召回大量的新詞對，供後續應用，讓既有的上百萬條雙語詞庫，再增加約 20%的新詞對。

Abstract

This paper proposes two approaches to extract translation term pairs from Chinese-English bilingual corpus with more than 500,000 patents. One approach is precision-oriented, in which we compare six term alignment methods. Based on our experiments, we find that the EM (Expectation Maximization) method is the best. However, it is time-consuming and hard to extract many-to-many translations for the same concept. While the MI (mutual information) method performs worst, the term pairs extracted may be totally different from those by EM. This may inspire subsequent researches to study the possibility of hybrid term alignment methods. The other approach is recall-oriented, in which a simple idea was proposed. With an efficient implementation, 20% more term pairs were extracted based on an existing lingual lexicon which already has more than one million term pairs merged from several sources.

關鍵詞：專利語料庫，機器翻譯，專利跨語分析，詞彙對列，新詞擷取

Keywords: Patent Corpus, Machine Translation, Cross-lingual Patent Analysis, Term Alignment, Term Extraction.

一、緒論

世界各主要國家的專利機構（IPO, Intellectual Property Office，或是 PO, Patent Office），如世界智慧財產權組織的 WIPO、歐洲的 EPO、美國的 UPSTO、日本的 JPO、韓國的 KPO、大陸的 SIPO 以及我國的 TIPO，都有專利網站提供免費（或付費）的專利單語檢索服務。然而，在專利誇語言的服務上（特別是內容翻譯），最近才開始陸續提供。

目前台灣已是美國專利核准案中第四大外國申請國 [1]，且台灣專利申請案中有 40%爲外國申請案 [2]，其中比重最大的日本亦曾向經濟部反映希望增設專利英語檢索功能 [3]，顯見目前台灣相關市場上對於專利跨語言檢索與翻譯的迫切需求，但目前智慧財產局與民間尚無提供跨語言的專利自動翻譯服務，顯然與世界趨勢與國內需求嚴重脫節。由國科會登錄的研發計畫與人力來看，國內專利分析與翻譯方面的研究非常匱乏，更看不見技術與應用端的上下游整合，顯示出此一問題的嚴重性。

事實上，我國智慧局自民國 87 起即進行本國專利摘要英譯的業務，以便與世界專利機構交換資料，至今完成 33 萬餘件。雖然每年智慧局都編列預算進行專利摘要人工英譯的工作（平均每年約 3 萬篇），至今卻苦無相關的資訊系統提供英譯對照、專利近似文句的譯文範例，使得翻譯過程辛苦而緩慢、翻譯內容與品質難保一致、翻譯的知識無法累積，而不能爲廣大的使用者運用於跨語言的專利檢索與分析。

專利文件的特色，在於技術名詞繁多、艱深，使得翻譯者專業知識的差異，容易造成英譯名詞的不一致。另外，專利中新詞不斷衍生，且散落於不同領域文件，即便是專家，也難以知悉相同概念的所有相關譯名。因此，在能夠進行有效的專利自動翻譯之前，能否獲取大量的專利雙語對照詞彙（aligned term pairs），乃此項自動化服務的先決條件。

本文的目的，在試圖從上述既有的專利中英雙語語料中，擷取出中英對照詞彙，一方面可做爲人工翻譯的參考，以便提高後續翻譯的一致性，一方面也可做爲未來機器翻譯的基礎語料，讓翻譯專家的語言知識，得以變成機器翻譯可依賴的詞彙知識。

二、待解問題

在擷取中英對照詞彙的過程中，預計將遭遇兩種類型的問題，其各有不同的發生時機與解決目標：

問題一：給定一組特定領域的中、英雙語對照語料庫，自動找出該領域的中、英雙語對照詞彙。

問題二：給定一組特定領域的中、英雙語對照語料庫，以及一組專屬該領域的中、英雙語對照詞庫，自動找出未曾出現在詞庫中的詞對。

問題一發生於一開始還沒有該領域（專利）的專屬雙語詞庫時，需要以自力更生（bootstrap）、精確導向（precision-oriented）的方式，找出專屬領域的翻譯詞對（translation term pairs）。一旦專屬的雙語詞彙越來越多時，待解問題將變成問題二的類型，其目標在召回導向（recall-oriented）的來找出更多的翻譯詞對。

實務上，在解決上述問題時，所有可用的資源皆可使用，例如用到既有的非專屬領域雙

語詞庫、單語詞庫、甚至其他的語料庫。當用到這些資源時，問題一與問題二會變得有點模糊，因爲所謂專屬或非專屬領域的雙語詞彙，其實不易界定。在問題一的解法中使用到非專屬雙語詞庫，好像變成問題二。事實上，我們對問題一的解決方法，可以找出問題二解法所無法找出的翻譯詞對。這是因爲問題一的精確導向特性，與問題二的召回導向要求，使得其解法不同，而有相異的結果。

另外，專利文件其文句大多很長，非常不利於機器的自動翻譯。以統計式翻譯範例而言，其先透過翻譯模型（translation model）將每個字的可能翻譯列舉出來，組合成所有可能的英文候選句，再透過語言模型（language model）來評估每個英文候選句的英文符合程度。當文句很長，裡面的詞彙很多的時候，就容易碰到組合爆炸（combinational explosion）或是翻譯時間過長的情形。因此，在 Manning 與 Schutze 的書中（第 490 頁）[4] 甚至建議超過 30 個字的句子，就不給機器作翻譯訓練。但這對專利文件是行不通的。我們認爲，解決的辦法，在降低其組合數。亦即對付的句子雖然長，但只要其「已知如何翻譯的片段越長」，組合出來的候選句就會越少，而仍舊可以有效率的進行機器翻譯。在此所謂「已知如何翻譯的片段越長」，表示我們擷取出來的雙語詞對越長、越多，對長句的專利翻譯，會越有幫助。

基於上述特性，本文提出兩種解決方案，各別對付上述兩類問題。其中我們用到的語料，如表一所示。

表一、本文中使用到的語料庫

語料	數量	使用場合
專利中英摘要語料庫	506510 篇	擷取中英詞對
中文單語詞庫（一般領域用詞）	123280 條詞	中文斷詞
國立編譯館學術名詞	約 160 萬條中、英詞對（註）	驗證擷取的詞對

註：我們取得的國立編譯館的學術名詞有 77 個類別，共 170 個檔。每個檔案裡的格式不同，中、英詞對的標註習慣也不一致，例如：

熱乾 ⇔ Drying，heat	向熱性液晶⇔liquid crystal; thermotropic
黃[土]風 ⇔ yellow wind	液晶顯示 ⇔liquid crystal; display; LCD
向列的(液晶)，絲狀的⇔nematic	液晶顯示器⇔LCD (=liquid crystal display)
鼠【魚+銜】⇔ dragonet	液晶顯示器⇔liquid crystal display (LCD)
蒔蘿【火+青】⇔ cymenes	液晶顯示器⇔liquid-crystal display {= LCD}

在剖析成電腦可用的中英詞對時，常有剖析規則（在同一領域同一檔案中）互相衝突的情形，因而產生不少雜訊（即錯誤的詞對）。但由於數量太多，且專業領域知識有限，目前無法一一對其檢驗進行更正。因此，剖析規則的不同，最後的詞條數也不同。我們後來另有一個版本，約 110 萬條，其用到的剖析規則較多、較嚴、較精確，但可能遺漏較多。

三、解決方案一

針對問題一，我們提出的解決方案，如圖一所示，分下列步驟詳述。

<table>
<tr><td>步驟</td><td>芸香劑方中藥組成製造方法</td><td colspan="2">Method for producing rutin Chinese medicine composition</td><td>所需資源</td></tr>
<tr><td>斷詞</td><td>芸香劑方 中藥 組成 製造 方法</td><td colspan="2">method, producing, rutin, Chinese medicine, composition</td><td>片語擷取</td></tr>
<tr><td>過濾</td><td>芸香劑方 中藥 組成 製造 方法</td><td colspan="2">rutin, Chinese medicine, composition, producing, method</td><td>既有詞庫</td></tr>
<tr><td>對列</td><td colspan="2">Run 1:
Co(芸香劑方, rutin)
Co(中藥, Chinese medicine)
Co(組成, composition)

Co(芸香劑方, rutin)
Co(中藥, composition)
Co(組成, Chinese medicine)

Co(芸香劑方, Chinese medicine)
Co(中藥, rutin)
Co(組成, composition)

Co(芸香劑方, Chinese medicine)
Co(中藥, composition)
Co(組成, rutin)

Co(芸香劑方, composition)
Co(中藥, Chinese medicine)
Co(組成, rutin)

Co(芸香劑方, composition)
Co(中藥, rutin)
Co(組成, Chinese medicine)</td><td>Run 2:
Co(芸香劑方, rutin)
Co(中藥, Chinese medicine)

Co(芸香劑方, Chinese medicine)
Co(中藥, rutin)

Run 3:
Co(芸香劑方, rutin)</td><td>統計資料</td></tr>
</table>

圖一、中英文詞彙對列方法示意圖

（一）起始步驟：

首先，從專利中英摘要語料庫中，取出每一篇中英對列文件，根據標題內的中、英文用語，進行詞彙對列的計算工作。每一篇中英對列文件，除專利號外，只有標題與摘要。目前只用標題的文句，而不用摘要內的文句，來進行詞彙對列。這是因爲我國專利中英摘要的文句對照情況並不好。常常其中文摘要只有一句，但英文摘要卻有三句。因此，若要用到摘要的文句，來豐富可用的語料，需先進行文句對列的工作，而且不能只對到句子，還要對到正確的子句。然而這項工作並不容易。因此，在目前精確導向的考量下，我們決定只用標題的文句。

（二）斷詞處理：

斷詞處理的目標，是要將中英文對照的詞彙，擷取出來。像圖一中的範例，中文不能只斷出「芸香、劑方」兩個詞，而要整個「芸香劑方」都斷出來，才有可能匹配到英文的「rutin」一詞。反之，英文也是如此。否則就要考慮多對一、一對多、多對多的詞彙對列情況，而這樣會讓對列步驟過於複雜，成效也不見得較好。

因此，在斷詞處理中，要具備擷取片語、新生詞彙或未知詞彙的能力，否則不易找出既有雙語詞庫以外的詞對。爲此目的，我們採用 Tseng 的關鍵詞擷取演算法 [5-7]，其可擷取「最大的」（maximal）重複字串。在此所謂「最大的」，是指字串最長的或出現頻率最高的。由於此演算法沒有用到詞彙知識，僅憑字串的重複出現與其重複出現時左右字詞會有所不同的特性來擷取詞彙，因此，只要在文中出現兩次，即可被擷取出來，詞

彙擷取的門檻很低。另外，其也可以運用於 OCR 的錯誤詞彙擷取（具有正確左右邊界但內含辨識錯誤文字的詞彙） [8]，甚至是 MIDI 數位音樂的關鍵旋律擷取 [9]，可見其可擷取新詞的能力。

如同前述，我們只以標題進行詞彙對列，但在運用此演算法時，我們是將標題與摘要合併起來進行詞彙擷取，以便讓標題中的重要詞彙能夠重複出現，而被擷取出來。

然而，專利摘要的寫法，其第一句常包含整個專利標題。由於此演算法相當貪婪（greedy），造成整個標題常被擷取成單一個詞彙，而無法進行詞彙對列。我們的解決方法，是利用一般詞庫先對中文標題進行斷詞（英文則依空格斷詞），將斷詞後的單一字詞屬於連接詞、代名詞等停用詞者，代換成逗號，從而將整句標題斷開，而不會與摘要中的第一句完全相同。圖二的例子展示了此過程。其中，我們使用的「一般詞庫」，乃從網路上下載，並自行加入 3 千多個新聞詞彙而來，總共有 123280 個中文詞。

原始標題	紗線製法及其組成結構	method for making yarns and constitution structure thereof
斷詞及過濾後的標題	紗線, 製法, , , , 組成, 結構	, , making yarns , constitution structure ,
標題及摘要的重複字串	線捲:4,線紗:3,紗線製法:2,組成結構:2, 管狀:2	making yarns:2, winding:4, yarn:3, weaving:2, constitution structure:2
屬於標題中的重複字串	組成結構, 紗線製法	making yarns, constitution structure

圖二、標題中重要詞彙的擷取過程範例(冒後後面的數字爲該詞彙在該文件的出現次數）

（三）詞彙過濾：

此步驟目的，是要將不可能的詞對過濾掉，以節省後續對列計算的負擔。我們可以用單語詞庫，將一般性的詞彙刪除（因爲可以假設其對應的翻譯詞已知，或容易從他處取得），或是利用既有非專屬領域的雙語詞庫，將已知的詞對刪除，以減少後續不必要的詞彙對列。當然，若沒有任何詞庫資源，此步驟也可以不做，只是會浪費力氣去處理很多不可能的（implausible）詞對。

（四）對列計算：

一旦句對中的詞彙擷取出來，並運用既有資源將不可能的詞對排出後，接下來即可大膽的將中、英詞彙進行盲目的配對。如此累積一篇篇、一句句的詞對後，透過對列分析，正確的詞對，大多會和錯誤的詞對有統計上的區別。常用的對列分析列舉如下：

1. 相互資訊（MI，mutual information）

MI 計算公式爲 [4]：

$$MI(c,e) = \log_2 \frac{p(c,e)}{p(c)p(e)}$$

在我們的應用中，p(c,e)表示中文詞 c 與英文詞 e 一起出現在同一句對的機率，而 p(x)則表示單語詞彙 x 出現在某一句對的機率。這些機率可用最大可能性估計（maximum likelihood estimation）算出，亦即用其出現的句對數，除以全部句對數來求得。

MI 的值是對稱的（symmetric），亦即 MI(c,e)=MI(e,c)，而且 MI 的值都爲正數，但沒有上界。雖然 MI 很早就被統計式自然語言處理採用在這類應用上 [4]，但後續的研究，不斷的驗證出 MI 其實效果並不佳[10]。但由於其計算簡單，我們也納入此方法，以供比較驗證。

2. 相關分析（CC, correlation coefficient）

文獻中常用 Chi-square 來分析兩個事物的相關性 [4]：

$$\chi^2(c,e) = \frac{(f_{11} \times f_{22} - f_{12} \times f_{21})^2}{FcFc^* \, FeFe^*}$$

其中f_{11}、f_{21}、f_{12}、f_{22}分別代表中文詞c出現而英文詞e也出現的句對數、c出現但e沒出現的句對數、c沒出現但e出現的句對數、以及兩者都沒出現的句對數，如表二所示，其中 Fc、Fc*、Fe、Fe* 與N，分別代表各行與各列的合計。

表二、中文詞 c 與英文詞 e 在句對中出現次數的交叉分析

	c 出現	c 沒出現	合計
e 出現	f_{11}	f_{12}	Fe
e 沒出現	f_{21}	f_{22}	Fe*
合計	Fc	Fc*	N

Chi-square 可用來做相依性統計考驗（test for dependence），而不需假設事件是否常態分佈。此值介於 0 到 1 之間，其實是底下相關係數的平方：

$$CC(c,e) = \frac{(f_{11} \times f_{22} - f_{12} \times f_{21})}{\sqrt{FcFc^* \, FeFe^*}}$$

相關係數的值介於-1 到+1 之間，可顯示兩事物從負相關到正相關的程度。在本應用中，我們選用相關係數 CC 以分析對照詞彙的共現性（co-occurrence）。Chi-square 與 CC 都是對稱的指標。

3. 可能性比例（LR, likelihood ratios）

文獻上提到 Chi-square（或 CC）對事件發生次數不多的情形，效果較差，因此建議最好不要用於事件次數低於 20 次的場合 [4]。LR 則比較適合於資料稀少的情形。其計算公式，原爲負值 [4]，我們將其改爲正值，如下：

$$LR(c,e) = \log L(f_{11}, Fc, p(e \mid c)) + \log L(f_{12}, Fc^*, p(e \mid c^*)) - \log L(f_{11}, Fc, p(e)) - \log L(f_{12}, Fc^*, p(e))$$

其中$L(k,n,x)=x^k(1-x)^{n-k}$, p(e)=Fe/N, $p(e|c)=f_{11}/Fc$, $p(e|c^*)=f_{12}/Fc^*$。檢視其定義，並經過驗算，LR的值是對稱的。

4. Dice 係數（DC, Dice coefficient）

在資訊檢索中 Dice 係數常用來衡量兩事物的相似度：

$$DC(c,e) = 2f_{11}/(Fc+Fe) = 2f_{11}/(2f_{11}+f_{12}+f_{21})$$

其值是對稱的，且計算雖簡單，文獻上卻顯示其成效較不受出現次數多寡的影響 [10]。

5. 分數累積（FC, fractional count）

上述的各項分析，所依賴的資訊，都屬於跨句對的（inter-sentence），而沒有用到句對內的（intra-sentence）資訊。圖一中顯示，盲目的配對，有六種可能的組合，因此每一種配對組合裡的每一種翻譯，應該只獲得 1/6 的分數。在同一句對中累積這些分數，可以知道每一個詞對翻譯機率，例如「芸香劑方」翻成「rutin」的機率是 2/6。累積詞對在所出現句對中的翻譯機率（而非次數），則爲我們所謂的 FC 值。其計算公式表達如下：

$$FC(c,e) = \sum_{for\_all\_i, s.t.(c,e)\in sp(i)} p_{sp(i)}(c,e)$$

其中sp(i)表示第i個句對（且同時包含中文詞c與英文詞e），而$p_{sp(i)}(c,e)$表示在第i個句對中詞對(c,e)的翻譯機率。這樣的FC，其值是對稱的。

6. EM 分析（Expectation-Maximization analysis）

上述五種分析方式，都沒有考慮到一個因素，即：較差詞對的懲罰（penalization of implausible alignment）。我們當然可以利用上述五種方法之任一種，依照其指標數值排序，將排序在前面的詞對視爲正確的翻譯，然後回頭過濾不可能的盲目配對，如此反覆循環，如圖一的 Run 2 與 Run 3 所示。這樣，也可以刪除掉（懲罰了）較差的詞對。

然而，在我們反覆進行詞對的過濾與分析以前，利用 EM 方法，就可以在進行詞對分析時，懲罰出現次數較少的詞對，同時凸顯出現較多的翻譯詞對。亦即若某個中文詞與某個英文詞，在所有的句對中，有較多的配對關係，則該中文詞就不太可能再跟同句對中的其他英文詞配對。底下的 E 步驟與 M 步驟的反覆計算方法，可以自動達到這種特性：

E-step：在所有中文詞 c 與英文詞 e 同時出現的句對中，累積條件機率 p(e|c)的值

$$s(c,e) = \sum_{for\_all\_i, s.t.(c,e)\in sp(i)} p(e\,|\,c) = p(e\,|\,c) * f_{11}$$

其中f_{11}定義如前，而p(e|c)的初始值可任意設定爲 1，或設定爲FC(c,e)。

M-step：根據上述累積的結果，重新評估條件機率

$$p(e\,|\,c) = \frac{s(c,e)}{\sum_v s(c,v)}$$

其中v爲任意使得s(c,v)不爲 0 的英文詞。表三顯示一個中文詞c=「驅動電路」的範例。在所有的句對中，它與其他四個英文詞共同出現的次數如f_{11}一欄所示。在初始值p(c|e)都爲 1 的情況下，經過四次EM步驟後，c跟e4 互爲翻譯的機率越來越高，而其他則越來越低，果真達到我們的期望。

表三、EM 分析範例

		Loop 1		Loop 2		Loop 3		Loop 4	
c=驅動電路	f_{11}	s(c,e)	p(e|c)	s(c,e)	p(e|c)	s(c,e)	p(e|c)	s(c,e)	p(e|c)
e1= display devices	2	2	2/8	0.5	0.1818	0.3636	0.1081	0.2162	0.0584
e2= electroluminescent lamp	1	1	1/8	0.125	0.0455	0.0455	0.0135	0.0135	0.0036
e3=lamp driving circuit	1	1	1/8	0.125	0.0455	0.0455	0.0135	0.0135	0.0036
e4= driving circuit	4	4	4/8	2	0.7273	2.9090	0.8649	3.4595	0.9343

四、成效分析

將前述的 506510 篇專利中英標題與摘要，共 463MB 純文字檔案，以起始步驟、斷詞處理、詞彙過濾三步驟在桌上型電腦上計算（Pentium 4, 3.20GHz, 2GB RAM），總共花費 5945 秒，統計結果如下：

表四、專利中英標題與摘要之處理結果統計

Number of bilingual texts processed (463MB)	506510
Number of empty Chinese Title	0
Number of empty Chinese Abstracts	281
Number of empty English Title	0
Number of empty English Abstracts	166532
Number of Doc. with both abstracts empty	2
Number of bilingual texts yielding empty term pair	322309
Number of term pairs generated	1180281
Number of unique term pairs after merging	455696

將上述 455696 個詞對，以前一節中的所有對列分析方式計算，總共花費 1340 秒（Intel CPU T2500 2.0 GHz, 2 GB RAM），其中讀取所有詞對花費 15 秒，以 EM 分析同時求 p(c|e) 與 p(e|c)用五個迴圈所花費的總時間爲 1296 秒，而其他五種分析方法總花費時間則爲 29 秒，平均每一種方法不到 6 秒鐘。將此對列分析結果，以 FC 由大到小排序，檢視其前 10 名的詞對，如表五所示：

表五、以 FC 排序的前 10 個詞對

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor device	2078.30	1.00	1.00	0.79	5.58	0.79	-1	2455	2764	3429
顯示裝置	display device	867.83	1.00	1.00	0.70	6.22	0.70	-1	1225	1909	1589
液晶顯示裝置	**liquid crystal display device**	772.08	**0.98**	**1.00**	0.64	6.21	0.65	-1	1078	2072	1292
電連接器	electrical connector	750.25	1.00	1.00	0.80	7.15	0.79	-1	830	1045	1029
電子裝置	electronic device	326.45	1.00	1.00	0.57	6.90	0.59	-1	541	1180	706
背光模組	backlight module	323.25	1.00	1.00	0.74	7.70	0.74	-1	498	807	546
液晶顯示裝置	**liquid crystal display**	292.15	**0.01**	**1.00**	0.27	5.03	0.27	-1	451	2072	1227
半導體記憶裝置	semiconductor memory device	237.50	1.00	0.98	0.57	7.74	0.59	-1	302	394	661
* **液晶顯示裝置**	**liquid crystal**	218.15	**0.01**	**1.00**	0.27	5.10	0.27	-1	431	2072	1119
薄膜電晶體	thin film transistor	197.08	1.00	1.00	0.71	8.43	0.72	-1	280	467	320

* 註：錯誤的翻譯詞對

在表五中，中文詞「液晶顯示裝置」對列到三個英文詞「liquid crystal display device」、「liquid crystal display」與「liquid crystal」，其中前兩是正確的，第三個是錯的，但只有第一個的翻譯機率 p(e|c)接近 1，其餘接近 0。從 EM 法的計算公式可知，由於競爭性懲罰的關係，對同一個中文（或英文）而言，只有一個對列會存活而獲得較高的翻譯機率，其他的對列，其機率都會退化成接近 0。亦即其無法擷取多對多的同義翻譯詞對。

至於表中的 LR，其計算過程有 log(0)的情形，但因該值沒有數學定義，我們遂將其值設爲-1。事實上，若以 LR 排序，其前 10 名的詞對如表六所示。

表六、以 LR 排序的前 10 個詞對

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
環氧樹脂組成物	epoxy resin composition	44.62	0.97	1.00	0.87	10.47	0.87	757.59	99	114	113
照明系統	illumination system	67.64	1.00	1.00	0.82	10.32	0.83	731.76	98	133	106
記錄載體	record carrier	50.93	1.00	1.00	0.87	10.58	0.87	703.34	91	105	104
感光性樹脂組成物	photosensitive resin composition	65.58	1.00	1.00	0.72	9.86	0.72	697.28	102	137	148
資料處理系統	data processing system	52.29	1.00	1.00	0.87	10.65	0.87	675.94	87	100	100
熱交換器	heat exchanger	60.58	1.00	1.00	0.81	10.40	0.81	662.40	89	119	102
基地台	base station	38.96	1.00	1.00	0.80	10.42	0.80	638.43	86	111	104
資訊儲存媒體	information storage medium	48.62	1.00	1.00	0.83	10.59	0.83	633.72	83	103	96
半導體晶片	semiconductor chip	55.06	1.00	1.00	0.59	9.33	0.59	630.94	101	186	155
矽晶圓	silicon wafer	51.56	1.00	1.00	0.80	10.55	0.81	602.65	80	106	93

一般而言，我們可以依照各個對列方法的結果一一排序，然後檢視其前N名的詞對品質，以瞭解該分析方法的成效。然而實做結果顯示，有好幾個分析方法，其最大值的詞對很多，不是唯一，以致於前N名沒有區別性。如表七所示，若依照EM的結果以平均翻譯機率排序，其機率最大值 1 者，就有 1 萬 8 千多個。排除出現次數少於 5 次的詞對，還有 94 個並列第一。爲了再加以區別，可用其他的數值，例如用該詞對的出現次數（欄位f_{11}），做爲第二個排序條件。

表七、各項排序方式其最大值的詞對個數

最大數值範圍	詞對數
(p(c|e)+p(e|c))/2=1.0	18322
(p(c|e)+p(e|c))/2=1.0 and f_{11}>=5	94
DC=1.0	156353
DC=1.0 and f_{11}>=5	51
MI=17.49= maximum	152907
MI=17.49 and f_{11}>=5	33
CC=1.0	156353
CC=1.0 and f_{11}>=5	51
LR>186.13（註 1）	249
LR>186.13 and f_{11}>=5	249
FC>311.01（註 2）	6
FC>311.01 and f_{11}>=5	6

註 1：此數爲其平均數＋3＊標準差=26.07+3*53.35

註 2：此數爲其平均數＋3＊標準差=28.20+3*94.27

經由上述二階排序後，我們人工檢視了各個對列分析法前 n=50 個詞對，結果如表八所示。另外，我們也分析各個方法前 50 個正確詞對的重疊比率，結果如表九所示。這兩個表裡面的數據，印證了我們的觀察：{EM>LR>FC} > {DC=CC} >> MI，亦即 EM 與 LR、FC 的排序效果最好，惟三者中仍有些爲差距，其次是 DC 與 CC，兩者特性非常相似，最差的是 MI。

表八、各個方法前 n 個詞對人工判斷結果

排序方法	FC	EM	DC	MI	CC	LR
人工檢視的詞對數	50	50	50	50	50	50
錯誤的詞對數	3	0	6	39	6	1

表九、各個方法前 n=50 個正確詞對的重疊個數與重疊率

	FC	EM	DC	MI	CC	LR
FC						
EM	9 (18%)					
DC	0	0				
MI	0	0	0			
CC	0	0	45 (90%)	0		
LR	2 (4%)	10 (20%)	0	0	0	

表九透露出各個方法擷取的詞對重疊率不高（DC 與 CC 除外），顯示其能找出各有特色的正確詞對。爲了進行大規模的驗證，我們以國立編譯館學術名詞，共約 160 萬個中英詞對，來過濾前述的 455696 個詞對，得出其中的 7050 個，已出現在此雙語詞庫中，我們稱其爲「舊詞對」，其範例如表十所示；而另有約 25963 個詞對，雖不在既有詞庫中，但詞對裡中、英文詞彙的每個子字串都互有翻譯的關係，我們稱其爲「新詞對」，其範例如表十一所示。

表十、「舊詞對」範例

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor device	2078.3	1	1	0.79	5.58	0.79	-1	2455	2764	3429
半導體裝置	semiconductor devices	28.583	0	1	0.03	4.93	0.09	130	48	2764	105
顯示裝置	display device	867.83	1	1	0.70	6.22	0.70	-1	1225	1909	1589
顯示裝置	display devices	8.0595	0	1	0.02	6.20	0.08	61.7	16	1909	21
電連接器	electrical connector	750.25	1	1	0.80	7.15	0.80	-1	830	1045	1029
電子連接器	electrical connector	4	1	0	0.01	7.00	0.06	-1	5	7	1029
電氣接頭	electrical connector	1	1	0	0.00	7.48	0.03	-1	1	1	1029
電子裝置	electronic device	326.5	1	1	0.57	6.90	0.59	-1	541	1180	706
電子裝置	electron device	2.667	0	1	0.01	6.80	0.05	21.1	5	1180	7
薄膜電晶體	thin film transistor	197	1	1	0.71	8.43	0.72	-1	280	467	320
薄膜電晶體	thin-film transistor	28	0	1	0.15	8.39	0.27	215	39	467	46
記錄媒體	recording medium	162.4	1	1	0.65	7.98	0.66	-1	315	556	414
記錄媒體	recording media	15.87	0	1	0.09	7.88	0.19	135	27	556	38
記錄媒體	record medium	7.667	0	1	0.04	8.13	0.13	58.4	11	556	13
記錄媒體	record media	0.143	0	0.25	0.00	8.37	0.04	5.80	1	556	1
記錄介質	recording medium	5.333	1	0	0.04	8.10	0.11	-1	8	13	414
電腦系統	computer system	159.9	1	1	0.85	8.74	0.85	-1	313	374	361

表十一、「新詞對」（註）範例

c	e	FC	p(e|c)	p(c|e)	DC	MI	CC	LR	f_{11}	Fc	Fe
半導體裝置	semiconductor assembly	1.5	0	0.56	0.0014	5.06	0.02	5.66	2	2764	4
半導體裝置	semiconductor equipment	0.5	0	0.0	0.0007	3.06	0.01	1.29	1	2764	8
半導體裝置	semiconductor arrangement	0.5	0	0.0	0.0007	3.06	0.01	1.29	1	2764	8
半導體器件	semiconductor device	3	0.99	0	0.0017	4.75	0.02	-1	3	6	3429
顯示器裝置	display device	88.2	1	0	0.1347	6.39	0.23	-1	118	163	1589
顯示器設備	display device	0.5	0.5	0	0.0013	6.86	0.03	-1	1	1	1589
顯示器器件	display device	0.5	1	0	0.0013	6.86	0.03	-1	1	1	1589
液晶顯示裝置	liquid crystal display device	772	0.98	1	0.6409	6.21	0.66	-1	1078	2072	1292
液晶顯示裝置	lcd device	21.9	0	1	0.0265	5.82	0.09	97.1	28	2072	44
液晶顯示裝置	lcd apparatus	9.3	0	1	0.0105	5.85	0.06	38.4	11	2072	17
液晶顯示裝置	lcd equipment	0.5	0	0.5	0.001	6.47	0.02	4.49	1	2072	1
電連接器	electric connector	31.1	0	1	0.0879	6.95	0.18	212	49	1045	70
電連接器	cable connector	2	0	0	0.0037	3.18	0.01	2.67	2	1045	39
電氣連接器	electrical connector	81.9	1	0	0.1803	7.11	0.28	-1	105	136	1029
電接頭	electrical connector	2	0.97	0	0.0077	7.16	0.06	-1	4	5	1029
電連接器插座	electrical connector	1	1	0	0.0019	7.48	0.03	-1	1	1	1029
電路連接器	electrical connector	0.3	0.5	0	0.0019	7.48	0.03	-1	1	1	1029

註：以第一列而言，此詞對不在既有的雙語詞庫中，但中文「半導體裝置」中的「半導體」在其英文詞對中有「semiconductor」對應（亦即此詞對出現在既有的雙語詞庫中），而其中的「裝置」也有「assembly」對應，而且其中沒有任何中文沒有對應的英文，反之亦然，我們因此稱此其爲新詞對。

五、解決方案二

經過前一階段之處理，我們已經獲得一組可信的專業領域中英詞對，至少有 25963 條。在此階段主要的工作，是基於此一可信度較高的結果，自動擴充更多的中英詞對。基本的擴充概念很簡單：將已有的中英詞對組合加長後，再回到原專利文件中檢查是否出現在對應的文件中。此概念的範例，如圖三所示，其中既有的詞對以不同的底線與頂線顯示，當鄰近的詞彙找到既有詞對時，就可以將其組合出新的詞對。

標題：背光模組、串接模組及其導電塊	Title: Backlight modules with connector modules and conductive blocks thereof
摘要：一種適用於大尺寸平面顯示器之背光模組，包括一背板、複數個第一燈管、複數個第二燈管、一第一串接模組以及一光學構件組。…	Abstract: Backlight modules for large size flat panel displays are provided. A backlight module comprises a plate, a plurality of first and second lamps, a first connector and an optical assembly. …

圖三、基於既有詞對的翻譯詞對擷取示意圖

上述新詞對的擷取構想很簡單，也符合一開始提到的專利機器翻譯資源建構目標，亦即「擷取出來的雙語詞對越長、越多，對長句的專利翻譯，會越有幫助」。本階段的問題，在於如何提高執行效能。圖四是此構想的直覺演算方法。

我們目前有 506510 篇雙語專利摘要，並且有前一階段擷取的專利新詞對 25963 個，與國立編譯館的學術名詞約 160 萬條，再加上從一般性電子字典取得的中英詞彙，總共約 170 萬條對照詞。若直接使用上述演算法，其時間複雜度爲$O(NM^2)$，意謂著必需執行 1445000 兆次「詞對是否出現在某專利中」的檢查（50 萬 × 170 萬 × 170 萬）。顯然這並非是一個合理時間內，可執行完成的演算法。有鑑於此，我們利用檢索系統的索引結構與查詢功能，來提高效率，其演算法分三步驟，如下：

1. Set i = 1 to M；M 爲所有的中英詞對數
2. Set j = 1 to M；
3. Set $c_{ij} = c_i c_j$；c_i表示第i筆中英詞對中的中文詞（組合成較長的中文詞）
4. Set $e_{ij} = e_i e_j$；e_i表示第i筆中英詞對中的英文詞（組合成較長的英文詞）
5. Set k = 1 to N；N 爲所有的專利篇數
6. 若c_{ij} 出現在CP_k且e_{ij} 出現在EP_k，則 (c_{ij}, e_{ij}) 爲新增詞對；
   其中CP_k及EP_k分別代表第k篇專利的中文及英文部分。

圖四、新詞對擷取的基本演算法

1. 建立所有專利文件的反向索引檔 I，裡面記錄每個詞出現在哪些文件的資訊。

2. 建立一個記錄詞對編號的陣列 A，其維度爲專利資料的筆數 N。
3. Set i = 1 to M；M 爲所有的中英詞對數
4. 以 (c_i, e_i) 查詢索引I以檢索出所有包含 (c_i, e_i) 的專利編號集合$S = \{s_1,...,s_P\}$
5. Set $j = s_1$ to s_P
6. push i to A[j]；A[j]將記錄所有出現在專利 j 的詞對 i

7. Set p = 1 to N；N 爲專利資料的筆數
8. Set x = 所有包含在 A[p] 中的詞對編號
9. Set y = 所有包含在 A[p] 中的詞對編號
10. Set $c_{xy} = c_x c_y$；c_x表示第x筆中英詞對中的中文詞
11. Set $e_{xy} = e_x e_y$；e_x表示第x筆中英詞對中的英文詞

若c_{xy} 出現在CP_p且e_{xy} 出現在EP_p，則 (c_{xy}, e_{xy}) 爲新增詞對；其中CP_p及EP_p分別代表第p篇專利的中文及英文部分。

圖五、新詞對擷取的改良演算法

上述算法只針對「確定出現在同一篇專利文件的中英詞對」，才進行兩兩組合。根據實際處理結果，同一篇文件會出現的詞對數，平均約爲 4 組。因此只需檢查大約 200 萬組（50 萬 × 4）詞對是否在文件中即可，在速度上的增加非常明顯。

表十二爲演算法各步驟的執行時間。結果顯示，此演算法確實可在合理時間內自動擷取新的詞對。經此算法自動擴充的中英詞對數量爲 406184 筆，對既有的上百萬條雙語詞對而言，約可增加 20%（=40/(170+40)）的數量。不用驗證，**它們都是正確的**（錯誤的詞彙可用組合規則過濾掉），而且是**不在既有詞庫裡的專業領域新詞對**。

上述演算法只組合兩個既有詞彙，事實上也可以考慮組合三個、四個、甚至 n 個詞彙。然而，一方面合法的長詞會越來越少，二方面若既有詞對已夠多，可以先只組合兩個詞彙。若怕遺漏，將找出來的新詞對，以此方法再演算一遍，即可找出原來需要三個甚至四個詞彙組合的新詞。這種方式，比直接組合多個詞彙還要快。

表十二、新詞擷取演算法各步驟的執行時間

步驟	時間(秒)
索引檔建立	2971
查詢所有中英詞對	25920
檢查新增中英詞對	1560

六、結論

從相關計畫的分析得知，翻譯詞庫的持續更新，是不斷提升機器翻譯品質的必要任務，也是輔助人工翻譯或進行跨語檢索的基礎工作。本文以精確導向及召回導向角度的探討此議題，提出了因應不同目標的解決方法。

在精確導向的任務中，我們比較了六種詞彙對列方式，說明它們的優缺點，並以實際資料作驗證，從而得出合理的解讀結果。亦即就詞對的排序效果而言，{EM>LR>FC} > {DC=CC} >> MI。其中 FC 方法僅作局部（單篇）的分數計算再全部累加，計算最簡單；而 LR 的計算也不算太困難，效果卻更好；EM 則最花時間，但效果最好，只是難以找出多對多的同義翻譯，而 LR 與 FC 都可以（因爲其具有對稱性）；即便是最差的 MI 法，其排序在前頭的正確詞對跟這三種最好的方式不同，因此可以作爲輔助的詞對擷取方法，爲後續合併或混用多種對列方式的研究，開啓了可能性。

而一旦精確導向的詞對擷取結果出爐，立刻可用於召回導向的任務。我們的成果顯示，簡單的想法加上有效的實做，即可從雙語對列語料庫中召回大量的詞對。對後續不斷累積擴大的專利雙語語料而言，本文詳述的方法，可供後續一再地應用，以自動的方式，擷取出更多的專利新生中英詞對。

參考文獻

[1] Patents By Country, State, and Year - Utility Patents (December 2008). 2009; Available: http://www.uspto.gov/go/taf/cst_utl.htm. [Accessed: July 10, 2009].

[2] 96 年 專 利 統 計 . 2008; Available: http://www.tipo.gov.tw/ch/MultiMedia_FileDownload.ashx?guid=e24d3489-c729-4159-ab2a-eab9e6788d49.pdf. [Accessed: July 10, 2009].

[3] 台北市日本工商會 and 日本知的財產協會. 致經濟部建議書. 2007; Available: http://kousyoukai.japan.org.tw/tokusennchu.pdf.

[4] Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2001.

[5] Tseng, Y.-H. Multilingual Keyword Extraction for Term Suggestion. in 21st International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98. 1998. Australia.

[6] Tseng, Y.-H., Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology, 2002. 53(13): p. 9.

[7] Tseng, Y.-H., C.-J. Lin, and Y.-I. Lin, Text Mining Techniques for Patent Analysis. Information Processing and Management, 2007. 43(5): p. 1216-1247.

[8] Tseng, Y.-H., Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text. Journal of the American Society for Information Science and Technology, 2001. 52(5): p. 12.

[9] Tseng, Y.-H., "Content-Based Retrieval for Music Collections," Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99, Aug. 15-19, Berkeley, U.S.A., 1999, pp.176-182.

[10] W. Bruce Croft, Donald Emtzler, Trevor Strohman, *Search Engine: Information Retrieval in Practice*, Addison-Wesley, 2009.

On the Learning of Chinese Aspect Marker *le* through Interactive Multimedia Program*

多媒體互動課程對華語時貌標記「了」學習成效之研究

謝慈惠　Hsieh, Tzu-Hui
國立嘉義大學外國語言學系
Department of Foreign Languages
National Chiayi University
s0951072@mail.ncyu.edu.tw

吳俊雄　Wu, Jiun-Shiung
國立嘉義大學外國語言學系
Department of Foreign Languages
National Chiayi University
wujs@mail.ncyu.edu.tw

鐘樹椽 Chung, Shu-Chuan
國立嘉義大學數位學習設計及管理學系
Department of E-learning Design and Management
National Chiayi University
tschung@mail.ncyu.edu.tw

Abstract

The Chinese aspect marker *le* has long been considered very difficult for CSL learners. Therefore, we created an computer-based interactive multimedia CSL program of the perfective *le* based on the linguistic studies of the perfective *le* [3,25,26,28,29] and explored its effectiveness. Results of this study didn't show that the multimedia program as a self-learning tool outperform the printed materials significantly. Nevertheless, the result indicated that both the interactive multimedia program and the printed materials within their own groups do have significant effects on the members of the individual groups. This significance is the evidence supporting that the CSL program of *le* based on linguistic generalizations is effective.

* This research was based on the project funded by the National Science Council (NSC) of the Republic of China under Contract No. NSC 97-2631-S-415-002. This paper is a revised and developed version of the first author's master's thesis. The second author is the primary investigator of this project and the third author is the co-investigator. We thank Dr. Jenny Yi-Chun Kuo and Dr. Jung-hsing Chang for their insightful comments at the oral defense. We also thank the anonymous reviewers of the ROCLING conference for their precious comments. All remaining errors are ours.

摘要

時貌標記「了」一直爲中文爲第二外語學習者的困難點之一，所以本研究根據語言學研究設計「了」之多媒體互動課程，並探討此多媒體課程對中文爲第二外語學習者的學習成效。結果顯示，作爲自學工具的多媒體課程相較於紙本，並沒有顯著的學習成效，然而，多媒體課程和紙本在其各別所屬之組別中學習表現的顯著差異，說明這個「了」課程對中文爲第二外語的學習者有所助益。

Keywords: Chinese Aspect Marker *le*, Chinese as a Second Language Learners, Interactive multimedia program

關鍵詞：時貌標記「了」，中文爲第二外語學習者，多媒體互動課程

1 Introduction

The perfective *le* has a high frequency of occurrence in Mandarin Chinese [5]. The perfective *le* in Chinese is difficult to Chinese as Second Language (hereafter CSL) learners. Chao [5] investigated how much a 30 year-old British man comprehended the perfective *le* by a fill-in test and found that the subject made mistakes frequently. Kao [13] examined the usage of the perfective *le*, the durative *zhe* and the experiential *guo* of a corpus consisting of Chinese inter-language of students abroad and also found that most errors are about the perfective *le*, compared to the other two. Furthermore, Li [18] found that CSL students mistakenly treat the perfective *le* as English past tense.

According to [14], [19], [21], [25], [28], [29], etc., the perfective *le* can go with four situation types, which are Achievement, Accomplishment, Activity, and stage-level State, and the interaction of *le* with different types leads to different interpretations, such as completion, termination, and inception. In light of [5, 13, 18, 31], the perfective *le* is quite difficult to CSL learners. Li [18] and Yeh [31] further suggested that linguistic studies about the interaction of the perfective *le* with verbs be utilized in CSL learning and teaching. Thus, we created a CSL program of the perfective *le* based on the studies of the perfective *le* with four situation types [3, 25, 26, 28, 29].

Several studies [1, 4, 6, 12, 30] have noted that multimedia through technology can help with language learning and instruction. Although Computer Assisted Language Learning (CALL) has been widely accepted as a useful educational tool for four decades, the application of CALL on Chinese started late from 1995 [22]. According to Zhang [34], CALL programs on grammar need to be explored because it is a less common addressed area. Thus, we digitalized the CSL program of the perfective *le* as a computer-based interactive multimedia program in which we took Form Focused Instruction[1] (FFI for short) as our approach and Concise Narrated Animation[2] [23] as the concept of multimedia presentation. Few studies, if any, have reported the effects of Computer Assisted Language Learning on CSL grammar. Therefore, we examined whether the interactive multimedia program designed in the study is effective in CSL learning the perfective *le*. We predicted that this interactive multimedia program is effective and if applied correctly, it is far more efficient than the printed

[1] In terms of acquisition of language as a second language grammar, Ellis [9] points out "focusing on linguistic form aids the acquisition of grammatical knowledge". Thus, we used Form Focused Instruction and draws attention to the forms and structures of the perfective *le* as the approach of the CSL program.

[2] Concise Narrated Animation (CNA), a simple way to present multimedia, reminds us of ignoring the unneeded materials shown in multimedia and showing the most important part of ready-to-learn knowledge.

materials.[3] Two questions were explored. First, in contrast with the printed materials, is the interactive multimedia program designed in the study more effective on helping CSL learners with comprehension of the perfective *le*? Second, if the interactive multimedia program proves positive, is it more useful in terms of syntactic behavior or the semantics of the perfective *le* learning? We hope that the study can contribute to unraveling the effect of Chinese multimedia program on grammar as a self-learning tool and provide teachers with a teaching tool for efficient instruction.

The remainder of the paper is organized as follows. In Section 2, we present methodology including participants, instruments, data collection and data analysis. Section 3 reports results and discussion. Results were presented with various analyses following each of these descriptive sections. Discussion included the effect of the interactive multimedia program vs. the printed materials, assessment of the interactive multimedia program and the printed materials as multimedia in presentation modes [23]. Finally, Section 4 concludes this study.

2 Methodology

This study chose a quantitative method to investigate the effect of interactive multimedia program on Chinese Aspect Marker *le* learning. Based on the research goal, this study examined if the multimedia interactive program is more efficient and effective than the printed materials on Chinese Aspect Marker '*le*' learning.

In addition to the CSL program for the instruction and practice of *le*, a questionnaire, a pretest, and a posttest were used in the study. The CSL program included the interactive multimedia program and its printed materials. Multimedia in the interactive program means presentation using auditory and visual material. For example, the interactive multimedia program used audio narration and animation to present the content in the CSL program designed in the study. A questionnaire is to know the background of all participants. The result of the pre-test and posttest were analyzed, using Independent T-test to determine the effect of the interactive multimedia program.

All participants were divided into two groups. One is the control group and the other the experimental group. The control group studied the perfective *le* through the printed materials and the experimental group through the interactive multimedia program. The major difference between the multimedia program and the printed materials is the way to present the contents in the CSL program of the perfective *le*. The printed materials present the target sentences by the printed text and pictures, while the interactive multimedia by audio narration and animation. Figure 1 shows the research design.

[3] The printed materials contain hard copies of the interactive multimedia program designed in the study. They include target sentences and pictures retrieved from the animation of the program.

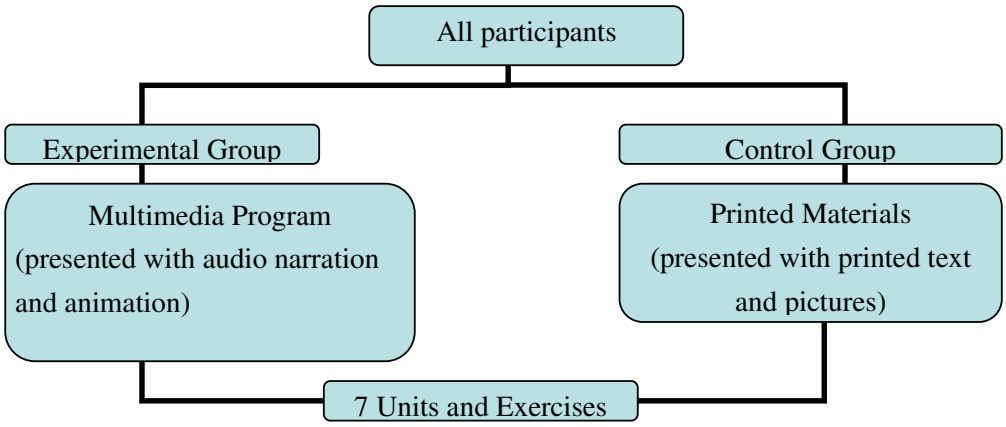


Figure 1. The Research Design

2.1 Participants

Thirty four participants for this research were selected from the population of CSL students enrolled at the CSL program in National Chiayi University and other CSL institutes in Taiwan, and overseas compatriots in Brunei and Australia[4]. All of them learned Mandarin Chinese as a second language. Twenty of the participants were female and fourteen were male. They came from various countries. Thirteen of them come are from Thailand, four Indonesia, four Korean, three the U.S.A., two France, one the U. K., one Japan, one Honduran, one Mongolia, and one Philippine. Besides, two overseas compatriots are in Brunei and one is in Australia. The amount of time that the participants have studied Mandarin Chinese ranged from half a year to three years. Also, all of them have learned the perfective *le*. Their ages ranged from eighteen to fifty-seven years.

These 34 participants were divided into two groups randomly, 17 students for each. One is the control group learning the perfective *le* through the printed materials and the other experimental group through the interactive multimedia program. In order to establish the homogeneity of the two groups, we calculated the average scores of the pretest of these two groups respectively out of a maximum score of 100 and got the mean score of the control group 48.63, and that of the experimental group 55.49, as shown in the following Table 1. Also, results of the independent T-test indicated no significant differences ($t = -1.845$, $p > .05$) between the two groups. Therefore, the level of comprehension of the perfective *le* with four situation types between the control group and the experimental group is roughly the same.

Table 1. Independent T-test Results on the scoring between the two groups in the pretest

Group	N	M	SD	t
Control	17	48.63	19.97	-1.845
Experimental	17	55.49	12.70	

Note: Maximum score=100, $p > .05$.

2.2 Materials and Instruments

[4] These three overseas compatriots don't speak Mandarin at home, and nor do they study in Chinese school. That is to say, Mandarin Chinese is not their native language. They study Chinese by themselves.

The experiment consists of two parts. One is learning materials, i.e. the CSL program of the perfective *le*, including interactive multimedia program and the printed materials. The other parts are the pretest and the posttest. In the following 2.2.1, the design of the CSL program and its content validity were discussed. Then in the following 2.2.2, the pretest and the posttest were introduced. In addition, their item difficulty (P) and discrimination (D) indexes and reliability were also examined through a pilot test.

2.2.1 The CSL Program of the Perfective *le*

The content in the CSL program is about the interaction of the perfective *le* with four situation types. The CSL consisted of seven units, seven to sixteen target sentences for each unit. Both Unit 1 and Unit 2 contained seven sentences showing the interaction of the perfective *le* with Achievement. Unit 3 and Unit 4 presented sentences with Accomplishment *le*, consisting of ten and eight sentences respectively. Both Unit 5 and Unit 6 were composed of eight sentences with Activity plus *le*. As to Unit 7, there are sixteen sentences with State *le*. The negative evidence column next to the target sentences showed error sentences, which were also included in the CSL program.[5] The patterns of target sentences in every unit came from the representative literature [3, 14, 19, 21, 25, 26, 28, 29]. Words in the CSL program are chosen from the Mandarin 800 words for beginner provided by the Steering Committee for the Test of Proficiency-Huayu (SC-TOP).[6] We picked up verbs for these 800 Chinese words and classified them into four situation types based on the following tests.

In the CSL program, we had two exercises following every unit. Every exercise contained five to eight items. Most of these items were from the target sentences. The form of these exercises was designed according to two content objectives as follows: (i) Learners are able to comprehend the interpretations of the interaction of the perfective *le* with four situation types, and (ii) Learners are able to know the collocation of the perfective *le* with four situation types. For example, *Kàn dònghuà, Xuǎn jùxíng* 'Choose the sentence that corresponds to the animation' could fit the objective (i) and *Jùxíngtiánkōng* 'Insert *le* in the right place' conform to (ii).

As above mentioned about the content of the CSL program, we digitalized it as the interactive multimedia program and collected it in a Compact Disc as the instrument of the experimental group. In the multimedia program, the target sentences were presented with animation and audio-narration.

Also, the multimedia program obeyed seven principles[7] for the design of multimedia message/presentation [23]. Thus, we showed every target sentence by words and animation for multimedia principle. The target sentences were near the animation for spatial contiguity principle and were presented with their corresponding animation at the same time for temporal contiguity principle. In order to meet coherence principle, on the screen are only animation and its corresponding target sentence. We used audio narration for modality

[5] According to [11], the positive input is not sufficient to reset a parameter for L2 learners and thus the alternative approach negative evidence can help.

[6] According to The Steering Committee for the Test of Proficiency-Huayu (SC-TOP), those who have learned Chinese more than half a year are familiar with these 800 words, see the following website. http://www.sc-top.org.tw/download/800Words_Beginners.pdf.

[7] The multimedia principle is to use words and animation rather than words alone, the spatial contiguity principle to place printed words near corresponding pictures, the temporal contiguity principle to present words and their corresponding animation at the same time, the coherence principle to avoid unneeded adjuncts, the modality principle to present words as narration rather than on-screen text, redundancy principle to avoid adding on-screen text to a concise narrated animation, and the individual difference principle to consider the students' background.

principle and avoid any redundant text to the animation for redundancy principle. This program was designed for CSL learners who have learned Chinese for at least half a year. In so doing it leads to fit individual difference principle.

On the other hand, the printed materials as the instrument of the control group represented target events by pictures. The pictures in the print materials were retrieved from animation of the interactive multimedia program. The major difference between the multimedia program and the printed materials is their presentation. The multimedia program presented the target sentences with animation and the audio narration on the screen while the printed materials presented them with pictures on paper. Besides, in Exercise, the interactive multimedia program offered the instant feedback. Once the users click on the right answer, the computer will give positive feedback immediately. On the contrary, the users who practice in Exercise of the printed materials have to check their answers with the Answer Sheet attached. They cannot get an immediate feedback.

In order to establish content validity of the CSL program designed in the study, we invited four scholars whose inputs and feedback were useful. Three of the scholars are linguists and the other one specializes in Computer Assisted Language Learning and E-learning. They were invited to review the instruments including the interactive multimedia program, the printed materials, and the following pretest and posttest. They judged the appropriateness of target sentences and their presentation. Thanks to their help, the content validity of the CSL program can be built.

2.2.2 Pretest and Posttest

Both the pretest and the posttest consisted of sixteen true/false and fourteen multiple-choice questions in Mandarin.[8] In terms of testing understanding of the perfective *le* with four situation types, true/false questions in each test was given to the subjects for grammatical judgment while multiple-choice for the semantic comprehension. When the test took place, these questions were shown on the projector/computer screen. The true/false questions were shown one by one while the multiple-choice were presented three choice items and each clips[9] based on which the participants answered for each question. These questions in each test were mostly picked up from the target sentences in the CSL program designed in the study. As previously mentioned, their content validity was established as a result of the review of the scholars. Next, the item analyses[10] of these two tests were performed to establish their internal consistency through a pilot test, discussed as below. In addition, their reliability was calculated by Cronbach's Alpha.

This pilot study was given to 20 CSL learners who study Mandarin in Taiwan. They are 14 males and 6 females. The amount of time they have studied Mandarin Chinese ranged from three month to six years. Also, all of them have learned the perfective *le*. Their age ranged from eighteen to fifty years. They were given the pretest and posttest simultaneously and also asked to complete a questionnaire for their L2 background. They spent 10 minutes on the pretest and then 10 minutes the posttest without any discussion based on the slides we designed in advance. They did these tests according to what they saw on the

[8] According to [10], the learning outcomes can be measured in a formal grammar test including a multiple-choice and a grammaticality judgment task.

[9] Thanks to two undergraduate students Xiao, Yu-Wun and Chao, Hui-Jiun in Department of Foreign Languages of Chiayi University for their performance in the video clips.

[10] When it comes to Item Analysis, test items are listed according their degrees of difficulty (easy, medium, and hard) and discrimination (good, fair, poor). These distributions provide a quick overview of the test, and can be used to identify items which are not performing well and which can perhaps be improved or discarded.

projector/computer screen on which questions were shown one by one. After the tests, the scores of the true/false and that of multi-choice in each test were calculated separately by the percentage of correct answer out of a maximum score of 100. Then, item difficulty (P) and item discrimination (D) indexes of the true/false and that of the multiple-choice were examined respectively in each test. For the item difficulty index (P), the higher the value is, the easier the question is.[11] For the item discrimination index (D), the higher the value gets, the better the item differentiates among participants on the basis of how well they know the materials being tested.[12] Regarding the result of item difficulty and discrimination on true/false questions in the pretest, the mean difficulty is .47 and the mean discrimination is .45. The results indicate that true/false questions in the pretest are difficult and determinable. The results of pretest on multiple-choice questions in which the mean difficulty is .48 and the mean discrimination is .53, present that multiple-choice questions in the pretest are difficult and distinguishable among students' performance on the test. As to the results of true/false in the posttest, the mean difficulty is .37 and the mean discrimination is .20 while on multiple-choice, the mean difficulty .31 and the mean discrimination is .24. The results indicate that the true/false and the multiple-choice in the posttest are quite difficult with positive discrimination.

Overall, both pretest and posttest on true/false or multiple-choice questions are difficult and differentiable to participants' knowledge of the perfective *le* with four situation types.[13] Regarding the reliability[14] of the pretest and posttest, it was established by the Cronbach's alpha. The greater the reliability is, the stronger the relative number of positive relationship among the questions would be. High reliability means that the questions of a test tended to "pull together." Participants who answered a given question correctly were more likely to answer other questions correctly. In the pilot study, the Cronbach's alpha of the pretest is .68 and that of the posttest is .60. Generally speaking, the results show that both the pretest and the posttest are reliable of testing comprehension of the perfective *le* with four situation types.

2.3 Procedures

Before the experiment, all participants were given the pretest for 10 minutes. The participants were tested on their knowledge of Chinese the perfective *le* with four situation types in terms of semantics and the syntactic behavior. Twenty-seven participants were tested in their Mandarin class, and seven participants online.[15] Those tested in the class were given the printed paper while those online were offered the website[16] on which they took the pretest. All of them were instructed to make a grammatical judgment for 16 True/False questions and semantics judgment for 14 multiple-choice questions. These questions were shown one by one on the projector/computer screen. Thus, they answered them according to questions presented on the screen. In addition, they are informed that it is not allowed to discuss in the test.

---

[11] In practice, item difficulty is classified as "easy" if the index is 85% or above; "moderate" if it is between 51 and 84%; and "hard" if it is 50% or below.

[12] Generally speaking, item discrimination is identified as "good" if the index is above .30; "fair" if it is between .10 and .30; and "poor" if it is below .10.

[13] Four questions of item discrimination indexes are negative. As suggested by scholars, the negative value of these four questions might be as a result of participants' uncertain about the answer. Thus, their answering these questions was not out of their understanding of the perfective *le*.

[14] The reliability of a test refers to the extent to which the test is likely to produce consistent scores. In practice, their approximate range is from .50 to .90 for about 95% of the classroom.

[15] Seven participants who had difficulty in coming in person for taking the test accepted the online test.

[16] We had our tests post on the website, see http://student.ncyu.edu.tw/~s0954185/LS.html.

After the pretest, participants were randomly divided into two groups, the experimental group and the control group, with an equal number of participants (N=17) in each group. To begin with the experiment, the experimental group received the CD which contains the multimedia program while the control group the printed materials. The period of learning the perfective *le* through the multimedia program for the experimental group or printed materials for the control group is seven weeks. After that, all participants took the posttest.

Those tested in the class got the printed paper while those online were provided the website on which they took the posttest. All of them were informed to make a grammaticality judgment for 16 True/False questions and semantics understanding for 14 multiple-choice questions. Being similar to the pretest, these questions in the posttest were shown one by one on the projector/computer screen. Meanwhile, participants answered them by reading the questions shown on the screen. Besides, the researcher avoided the discussion between participants.

2.4 Data analysis

After the data collection, the pretest was calculated, using the percentage of correct answers out of a maximum 100. As we noted before, we used an independent T-test to establish the homogeneity between the control group and the experimental group. The result showed that there was no significant difference between the two groups before using CSL program. In addition, the posttest was also computed by using percentage of correct answers out of a maximum 100.

For Research Question One, we examined if the multimedia program involved in the experimental group is more effective than the printed materials in the control group in terms of comprehension of the perfective *le* with four situation types. We used Paired T-test to examine the performance within these two groups. Then, the Independent T-test was employed to compare the performance between two groups.

For Research Question Two, if the result of Research Question One is positive, we investigated if the multimedia program is more useful in terms of the syntactic behavior or the semantic comprehension of the perfective *le* with four situation types. As has been discussed, the true/false questions tested understanding of the perfective *le* in terms of its syntactic behavior; the multi-choice questions addressed semantics. Thus, the score of the true/false and that of multiple-choice questions in the experimental group based on the pretest and the posttest were calculated respectively by the percentage of the correct answers out a maximum 100. The paired T-test was used to compare the experimental group's performances on the syntactic behavior to that on the semantic of the perfective *le*. The independent variable was the experimental group and the dependent variable was the difference between the pretest and posttest in terms of true/false and multiple-choice questions.

3 Results and Discussion

In this section, we answered Research Question One: In contrast with the printed materials, is the interactive multimedia program designed in the study more effective on helping CSL learners with comprehension of the perfective *le*? First, a paired T-test was used to examine performance of members within each group. The result showed the effectiveness either of the interactive multimedia program or of the printed materials although participants' responses through investigation by questionnaire[17] showed that not all of the participants used the CSL

[17] The purpose of the questionnaire is to understand how often participants used the interactive multimedia program or the printed materials and whether they used it.

program. For example, eight people used the multimedia program and eleven learned the perfective *le* through the printed materials.

Then, an Independent T-test was performed to compare performances between the interactive multimedia group and the printed materials group. The results revealed that the interactive multimedia program as a self-learning tool on 8 experimental subjects didn't outperform more significantly than the printed materials on 11 control subjects. That is to say, the answer to Research Question One is negative. However, the result of the paired T-test for the performance of members within each group proved that the CSL program designed in the study is effective no matter how it is presented: it can be represented in words and pictures, on the one hand, and audio-narration and animation, on the other hand. This implied that the CSL program created in the study on the basis of linguistic studies is useful for the CSL learning of the perfective *le* with four situation types. Also, the mean score obtained from the result suggested that the CSL learning of the perfective *le* is not easy.

Although the answer to Research Question One is negative, we still make further inquiry into Research Question Two: If the interactive multimedia program proves positive, is it more useful in terms of the syntactic behavior or the semantics of the learning of the perfective *le*? A Paired T-test was performed to compare the effect of the syntactic behavior and that of the semantics in terms of learning through the interactive multimedia program. Also, we investigated the difference between the performance of the syntactic behavior and the semantics in the printed materials since we found its effectiveness within group. Thus, two Paired T-tests were employed and the results revealed that there is no significant difference between the performance of the syntactic behavior and the semantics either in the multimedia program or in the printed materials. What these findings implied is that the CSL program designed in the study was equally involved the learning of the syntactic behaviors and the semantics of the perfective *le*.

Finally, we assessed the interactive multimedia program by the criteria of a good CALL program for CSL, proposed by Zhang [34] and discussed the specific questions of the program itself based on the participants' responses from the survey as previously mentioned. In addition, we considered the printed materials as a kind of multimedia in "presentation modes" based on Mayer [23] three views of multimedia. That may be the reason why there is the significant difference between the multimedia program and the printed materials.

3.1 Effects of interactive multimedia program and printed materials

For Research Question One, we used an Independent T-test to investigate if the interactive multimedia program is more effective than the printed materials as a self-learning tool in terms of the learning of the perfective *le* with four situation types. Before that, a Paired T-test was employed to examine the effect of the interactive multimedia program and the printed materials. We did a survey through questionnaires as previously mentioned in order to know the frequency of participants'' using the CSL program designed in the study as a self-learning tool. However, their responses were obtained: Not all of the participants used the CSL program. For example, 8 experimental subjects used the interactive program and 11 control subjects used the printed materials.

Thus, we showed the Paired T-test results of 8 experimental subjects on the learning of the perfective *le* through the interactive multimedia program and 11 control subjects through the printed materials. The Paired T-test results of 8 experimental subjects in Table 2 indicated there was a significant difference ($t = -3.845$, $p < .05$) between the mean scores of the pretest ($M=49.58$) and the posttest ($M=62.50$). The Paired T-test results of 11 control subjects in

Table 3 indicated there was a significant difference ($t = -4.042$, $p < .05$) between the mean scores of the pretest (M=45.15) and the posttest (M=58.18). In simple terms, these results revealed that the interactive multimedia program and the printed materials had a significant effect on 8 experimental subjects and 11 control subjects[18] respectively although their mean score of the posttest is not high. This implied that Chinese aspect marker *le* was not easy to learn for CSL learners.

Table 2. Paired T-test Results of 8 Experimental Subjects

Task	N	M	SD	t
pretest	8	49.58	14.52	-3.845*
posttest	8	62.50	7.51	

Note. Maximum score = 100, * $p < .05$

Table 3 Paired T-test Results of 11 Control Subjects

Task	N	M	SD	t
pretest	11	45.15	14.09	-4.042*
posttest	11	58.18	7.80	

Note. Maximum score = 100, * $p < .05$

Next, an Independent T-test was employed to answer Research Question One. There was a detailed descriptive statistics in the following Table 4. The figure indicated that there was no significant difference between the mean scores of the interactive multimedia program ($M = 62.50$) on 8 experimental subjects and the printed materials ($M = 58.18$, $t = -1.218$, $p > .05$) on 11 control subjects. That is to say, the interactive multimedia program didn't outperform the printed materials significantly.[19]

Table 4 Independent T-test Results between these Two Groups

Group	N	M	SD	t
Experimental	8	62.50	7.50	-1.218
Control	11	58.18	7.80	

Note. Maximum score = 100, $p > .05$

What is more, a Paired-sample T-test was employed to Research Question Two although the answer to Research Question One is negative. Research Question Two aimed to investigate if the interactive multimedia program is more effective in the learning of the syntactic behavior learning or the semantic. We also attempted to examine the performance of the syntactic behavior and the semantics on the learning of perfective *le* through the printed materials since

[18] The number of the experimental subjects may be too few to prove their effectiveness. Thus, we used the Wilcoxon Test, one type of the nonparametric methods which are most appropriate when the sample sizes are small, to reexamine the result. The results of Wilcoxon Test showed that there was a significant effect within these two groups ($p < .05$).

[19] A Mann-Whitney Test, one type of the nonparametric methods for small sample size, was performed to reassure the reulst of the Independent T-test. The result of the Mann-Witney Test also showed that there was no significant difference between these two small groups.

the multimedia program didn't show more effectiveness than the materials. The following Table 5 showed the Paired T-test result of the difference between the performance of true/false for the syntactic behavior and multiple-choice for the semantics of 8 experimental subjects. The figure indicated that there was no significant difference between the syntactic behavior and the semantics in terms of the pretest and the posttest in the multimedia group ($t = 1.35$, $p > .05$).

Table 5 Paired T-test Results of the difference between True/False and multiple-choice of 8 experimental subjects

Task	N	M	SD	t
True/False	8	17.97	11.298	1.35
Multiple-Choice	8	7.14	17.908	

Note. Maximum score = 100, $p > .05$

Table 6 below presented the Paired T-test result of the difference between the performance of the true/false and multiple-choice of 11 control subjects. The figure also indicated that there was no significant difference between true/false for the syntactic behavior and multiple-choice for the semantics in terms of the pretest and the posttest in the control group ($t = - .376$, $p > .05$). What the findings of the above Paired T-test in Table 5 and Table 6 [20] implied was that the CSL program designed in the study took account of the learning of both the syntactic behavior and the semantics of the perfective *le*.

Table 6 Paired T-test Results of the difference between True/False and multiple-choice of 11 control subjects

Task	N	M	SD	t
True/False	11	11.93	17.56	- .38
Multiple-Choice	11	14.29	11.07	

Note. Maximum score = 100, $p > .05$

All of these findings above could imply that that the CSL program of the perfective *le* created in the study on the basis of linguistic studies is useful for the CSL learning of the perfective *le* whether it is presented with words and pictures, comprised of the printed materials, or audio-narration and animation, of the computer-based interactive multimedia program. The CSL program designed in the study also considered both the syntactic behavior and the semantics in terms of the learning of the perfective *le*. Meanwhile, the mean scores obtained from the result revealed that the perfective *le* is not easy for CSL learners.

In the following, we assessed the interactive multimedia program by the criteria of a good CALL program for Chinese set by Zhang [34] and reviewed responses from the experimental group who used the interactive multimedia program as a self-learning tool.

3.2 Assessment of interactive multimedia program

[20] In order to make the result more reliable, we used a Wilcoxon Test for small size group to examine the difference between performance of the true/false for the syntactic behavior and true/false for the semantics of 8 experimental subjects and 11 control subjects. The results of the Wilcoxon Test showed that there was no significant difference between the performance of True/False and multiple-choice questions.

Zhang [34] argued that success or failure of a CALL program depends on the technology, language knowledge and language pedagogy. We took account of these three aspects while creating the multimedia program in our study. For example, for language knowledge, we adopted some well-established linguistic studies about generalization of the perfective *le* with four situation types for the content. Form-Focused Instruction, a type of language pedagogy known for teaching grammar, was used as our concept of arranging the content. As for technology, the well-known Macromedia Flash for animation design and Authorware integrating the components including sounds and clips.

Furthermore, Zhang [34] pointed out that the creativity and abstractness of grammar requires real ingenuity when designing CALL programs for CSL. Following this argument, we didn't give any grammatical explanation in our program. Instead, we used animation and audio-narration for target sentences. Animation can illustrate the abstractness and creativity of grammar. Audio-narration which comes out with the text is supposed to help beginning learners ease the anxiety when they don't recognize some words in these target sentences shown on the screen. It was said that the learner's emotional state plays an important part in student motivation and receptiveness to learning.

On the other hand, Zhang [34] proposed that some learners, especially those at the beginning level may feel overwhelmed when confronted with many choices. They may feel more comfortable following a teacher-suggested sequence of activities. In our study, we were unlikely to overcome the barrier because the multimedia program in the study was taken as a self-learning tool rather than an assisted-teaching tool. Learning through the interactive multimedia program on their own without teachers' assistance reflected the participants' responses received through questionnaires, as previously mentioned. From their responses in the questionnaires, we found that almost half of them didn't know how to start the multimedia program. For example, they had difficulty in downloading Flash Player required for them to see the animation even though we attached the guide sheet[21] showing instructions to use it. Thus, half of the experimental group never used it as a self-learning tool because of technical problem. This is the major reason resulting in that the effect of the interactive multimedia program didn't show more significantly than its printed materials.

Secondly, the cause for the failure of the multimedia program is those know how to use the multimedia program were not used to it while reading without any teacher stand-by. In this case, we learn that the tutorial sessions of using the multimedia program can be offered in the future studies as it is better to have a teacher stand-by to provide a sense of security. According to Zhang [34], learner's positive affect is the principle for language teaching, especially for CSL CALL programs.

In addition to the above-mentioned clues for effect of the multimedia program, two specific problems about the multimedia program itself we acquired from participants' responses through the questionnaire and interview are stated as follows. The first is the lack of easy access. The users cannot reach the multimedia program by just one click-on. Instead, they have to download Flash Player first if they haven't installed that in their computer and follow the instruction sheet to do the following three procedures including opening the specific folder as the first step, finding the index.swf file as the second step, and opening that file by Flash Player as the third step. Not until completing these three steps can users get start the multimedia program and skip to the menu page. Judging from the above procedures, starting the multimedia program in some way is not easy and simple, especially for users who don't get used to using computer.

[21] In the guide sheet, the procedures of using program were listed one by one.

The second cause induced from the participants' responses, is its instructions unfamiliar with the subjects. Take words shown on the button items or icons as example. It is somehow too complicated for users to understand its function. Take the word *HuíZhǔmùlù* 'back to the main menu' of the icon as the example, it is beyond the beginner-level. Even though we stated the functions of the icons on the instruction sheet, it was unlikely to be available for the beginning CSL learners, who simply cannot understand Chinese characters.Another illustration is interactive practices or games. The instructions for practices are too difficult to understand for CSL beginners. We also indicated how to do the exercises on the guide sheet. However, those wordy instructions likely resulted in learners' low motivation of keeping on the program.

So far, it seemed that the multimedia program didn't have a more apparent effect than the printed materials. We have discussed the reasons for less significant effectiveness of the multimedia program. The major reason is that not all of the participants use it as a self-learning tool. To make sure all subjects use the program in the future studies, we should provide tutorial sessions, in which they are asked to use it with teachers' assistance. Two more reasons related to the presentation of multimedia program are the lack of easy access and instructions of unfamiliarity with the experimental subjects, in particular those of the icons and practices. We learn that making access easy is the most important thing for activating users' motivation and reducing their anxiety while learning. The easier the access to the program is, the higher the motivation of learners has. Besides, instructions for the icons and practices are supposed to take students' background into consideration. Also, it is required to understand their real language proficiency. For example, some participants who said that they have learned Mandarin more than half a year do not know all the 800 words for Beginner. Thus, they may not comprehend some words in the program. For this point, we think the experienced TCSL teachers can be good supervisors or advisors for presentation of the interactive multimedia program. With their suggestion, the program can be more comprehensible for users.

3.3 The printed materials as multimedia in presentation modes

In this section, we argued that the printed materials in our study are considered as a kind of multimedia presentation. Based on Mayer [23], there are three views of multimedia - delivery media, presentation modes, and sensory modalities respectively.

In terms of the second view "presentation modes," multimedia means using two or more presentation types to present the materials. The focus is on more than one way to present teaching materials, such as words and pictures. This view is consistent with a cognitive theory of multimedia leaning [23], which assumes humans have separate information processing channels for verbal and pictorial knowledge. For example, a textbook which contains words and pictures is multimedia in this view because materials can be presented verbally as printed text and pictorially as static graphics.

The printed materials in our study are a kind of multimedia based on Mayer [23] presentation mode because words and pictures in the printed materials are more than one way to present them verbally and pictorially respectively. From the result mentioned above, its effectiveness was established by looking at the performance within the control group using it as a self-learning tool.

4 Conclusion

After a quantitative analysis of the data and discussion of the interactive multimedia program

and its printed materials, the findings were concluded as follows. The results of the present study may be summarized by pointing out that both the interactive multimedia program and its printed materials are able to show significant effect within their groups in terms of the learning of the perfective *le* with four situation types. What these findings implied is that the CSL program designed in the study is useful as a self-learning tool for the CSL learning of the perfective *le*. Also, the below 65 points of mean score of the posttest indicated that Chinese Aspect marker *le* is not easy to learn for CSL learners. This is in complete agreement with the studies [5, 13, 18, and 31] we mentioned in Section 1.

However, for Research Question One, the results did not reach our assumption in which the interactive multimedia program was more effective than the printed paper. These results may be explained by the following two causes. Firstly, participants' responses showed that only eight people used the interactive program and 11 people used the printed materials. Secondly, the printed materials composed of words and pictures are considered as multimedia [23]. Regarding Research Question Two, the results indicated that there was no significant difference between the learning of the syntactic behavior and the semantics of the perfective *le*. What these findings implied is that the CSL program of the perfective *le* as a self-learning tool designed in the study considered both the learning of the syntactic behavior and the semantics.

References

[1] Beatty, K, *Teaching and Researching Computer-Assisted Language Learning.* London: Pearson Education, 2003.

[2] Brandsford, J.D., Brown, A.L. & Cocking, R. R, *How People Learn*. Washington,DC: National Academy Press, 1999.

[3] Chang, Jung-hsing, State Eventualities and Aspect Marker 'le' in Chinese. *Taiwan Journal of Linguistic*, 1(1), 97-110, 2003.

[4] Chang, Y. D, Changing Learning Environments for Developing Listening Skills through a Multimedia and Interactive Model, APAMAIL: FL149-159, 2007.第十屆英語多媒體教學國際研討會/第四屆亞太地區多媒體語文教學研討會

[5] Chao, Li-Chiang, The Investigation of Foreign Students using *le*. *The Proceeding of the Fifth International Conference on Mandarin Chinese Teaching*, 300-308, 2002. (趙立江＜外國留學生使用"了"的情況與分析＞，《第五屆國際漢語教學討論會論文選》，300-308，2002。)

[6] Claybourne, T, The status of ESL, foreign language and technology. *Media & Methods*, 36 (1), 6-7, 2000.

[7] Comrie, Bernard, *Aspect.* Cambridge: Cambridge University Press, 1976.

[8] Doughty, Catherine, & Williams, Jessica, *Focus on Form in Classroom Second Language Acquisition*. Cambridge: Cambridge University Press, 1998.

[9] Ellis, Rod, *Understanding Second Language Acquisition*. Oxford:Oxford University Press, 1985.

[10] Ellis, Rod, *SLA Research and Language Teaching*. Oxford: Oxford University Press, 1997.

[11] Ellis, Rod, *Second Language Acquisition*. Oxford University Press, 2003.

[12] Jonassen, D. H., Peck, K. L., & Wilson, B. G, *Learning with Technology: A Constructivist Perspective*. Upper Saddle River, NJ: Prentice-Hal, 2003.

[13] Kao, Jui, *The Acquisition of Chinese Aspectual Marker 'le' 'zhe' and 'guo' to English speaking foreigners.* MA thesis, Beijing: Beijing Language University, 2006. (高蕊，＜歐美學生漢語標記"了" "著" "過" 的習得研究＞北京:北京語言大學碩士論文，2006.。)

[14] Klein, Wolfgang, Ping Li., & Hendriks Henriette, Aspect and Assertion in Mandarin Chinese. *Natural Language and Linguistic Theory*, 18(4), 700-723, 2000.
[15] Lightbown, P.M., & N. Spada, *How Languages Are Learned*. Oxford: OUP, 2006.
[16] Li, Charles, & Sandra Thompson, *Mandrain Chinese : A Functional Reference Grammar*, Berkeley: University of California Press, 1981.
[17] Li, Ch'i Lan, *Computer-Assisted Language Learning ib Brisbane Chinese Community Language Schools*. 5th International Conference on Internet Chinese Education ,92-108, 2007.
[18] Li, Hsiao-Ch'i, The Teaching of Chinese Aspectual Marker *le*. *Journal of Guangdong Normal University*, 4, 110-115, 1999. (李曉琪，＜漢語“了”字教學研究＞，《廣東師範大學學報》，第 4 期，110-115，1999。)
[19] Lin, Jo-wang, On the Temporal Meaning of the Verbal *le* in Mandarin Chinese, *Language and Linguistics*, 1(2), 109-133, 2000.
[20] Lin, Jo-wang, Aspectual Selection and Temporal Reference of –Zhe in Mandarin Chinese. *Tsinghua Journal of Chinese Studies, New Series Volume*, 32 (2), 257-296, 2002.
[21] Lin, Jo-wang, Temproal Reference in Mandarin Chinese. *Journal of East Asian Lin-guistics*, 12, 259-311, 2003.
[22] Liu, Meng, & Hong, Huo, Computer Assisted Language Learning (CALL) in China: Some Common Concerns. *US-China Foreign Language*, 5(1), 52-58, 2007.
[23] Mayer, R.E, *Multimedia learning.* New York: Cambridge University Press, 2001.
[24] Norman, D. A, *Things that make us smart.* Reading, MA:Addison-Wesley, 1993.
[25] Smith, Carlota, *The Parameter of Aspect*, 2nd ed., Kluwer Academic Publishers, Dordrecht, 1997.
[26] Soh, Hooiling, & Kuo, Yi-chun Jenny, *Perfective Aspect and Accomplishment Situations in Mandarin Chinese.* H. J. verkuyl, H. de Swart and A. van Hout (Eds.), Perspectives on Aspect: 199-216, 2005.
[27] Wang, Shuqin, Computer-assisted Language Learning (CALL): A Case Study. *Sino-US English Teaching*, 3(9), 17-20, 2006.
[28] Wu, Jiun-Shiung, *Modeling Temporal Progression in Mandarin: Aspect Markers and Temporal Relations*. Ph.D. Dissertation. University of Texas at Austin, 2003.
[29] Wu, Jiun-Shiung, The Semantics of the Perfective LE and Its Context-Dependency: an SDRT Approach. *Journal of East Asian Linguistics*, 14(4), 299-336. (SSCI, A&HCI), 2005.
[30] Yang, J. Peter, Networked Multimedia and Foreign Language Education. *Journal of Computer Assisted Language Instruction Consortium*, 15(1), 75-88, 1998.
[31] Yeh, Jung, The Instruction of Chinese Aspectual Marker le. *Journal of Xinan Chiao-Tung University*, 1(4), 89-92, 2000. (葉蓉. ＜關於"了"的教學＞，《西南交通大學學報》，第 1 卷，第 4 期，89-92，2002。)
[32] Yeh, Meng, The Stative Situaton and the Imperfective Zhe in Mandarin. *Journal of Chinese Language Teachers Association*, 29(1), 69-98, 1993.
[33] Zhang, Hong-yan, Computer-assisted elementary Chinese learning for American students, *US-China Education Review*, 4(5), 55-60, 2007.
[34] Zhang, Zheng-sheng, CALL for Chinese-Issues and Practice. *Journal of Chinese Language Teachers Association*, 33(1), 51-82, 1998.

A Framework for Machine Translation Output Combination

Yi-Chang Chen
Department of Computer Science and Engineering
National Sun Yat-Sen University
m963040046@student.nsysu.edu.tw
Chia-Ping Chen
Department of Computer Science and Engineering
National Sun Yat-Sen University
cpchen@cse.nsysu.edu.tw

摘要

本研究提供一個線上機器翻譯整合系統整合三個不同的線上翻譯引擎。該翻譯整合系統，利用了選擇、替換、插入及刪除等模組，針對線上翻譯假說進行修正。我們實際整合了GOOGLE、YAHOO、譯言堂的翻譯假說。在IWSLT07的測試語料進行中文至英文的翻譯整合。由實驗結果得知，該翻譯整合系統其 BLEU 分數由所整合的最佳翻譯系統的 19.15 進步到 20.55 。該翻譯整合系統相較於所整合的最佳線上翻譯系統進步了 1.4 BLEU 。

Abstract

In this paper, we propose a framework for combining outputs from multiple on-line machine translation systems. This framework consists of several modules, including selection, substitution, insertion, and deletion. We evaluate the combination framework on IWSLT07 in travel domain, for the translation direction from Chinese to English. Three different on-line machine translation systems, Google, Yahoo, and TransWhiz, are used in the investigation. The experimental results show that our proposed combination framework improves BLEU score from 19.15 to 20.55. It achieves an absolute improvement of 1.4 in the BLEU score.

Keyword: Machine translation, System combination

1 Introduction

The on-line machine translation is one application that is becoming popular nowadays. As each on-line translation system has its own strength and weakness, it is reasonable to expect that a framework capable of combining multiple on-line machine translation outputs may have the potential to produce translation results of better quality than the single-system outputs. In fact, this proposition has been shown to be true in certain published works, e.g., [1,2].

In this paper, we propose such a combination framework. The system is essentially sequential with the following basic components. First, one of the output sentence is selected as the raw best hypothesis. This raw best hypothesis is subjected to further post-processing modules

of substitution, insertion and deletion, based on the information provided by the unselected hypotheses and the source sentence.

This paper is organized as follows. In Section 2, we review related works of system combination for machine translation. In Section 3, we describe our proposed method for this problem. In Section 4, we present our experimental results. In Section 5, we draw conclusions.

2 Review

Our review of machine translation system combination is divided into three different categories: the sentence-level combination, the phrase-level combination, and the word-level combination.

2.1 Sentence-Level Combination

The sentence-level combination simply chooses one of the hypotheses as the combination output. That is, suppose the outputs from systems 1 through N are $H_1, \ldots, H_N$,

$$H^* = \arg \max_{H \in \{H_1, \ldots, H_N\}} S(H), \tag{1}$$

where $S(H)$ is a (re-)scoring function for hypothesis H. The design of the scoring function is the core problem in a sentence-level combination system. For example, different features with weights trained by the minimum error rate training (MERT) [3] can be used [4].

Note H^* is chosen *as is* without further processing. This approach renders the search space very limited. Such deficiency does need to be compensated by a rather sophisticated re-scoring mechanism for good performance. Still, that may not be enough, when the best hypothesis appears to be a *mixed-and-matched* solution.

2.2 Phrase-Level Combination

In the tribe of phrase-level combination, the phrase-level alignments are aggregated. The combined phrase translation table is used to re-decode the source sentence, generating a new hypothesis [2]. A phrase-based machine translation system, such as one based on GIZA++ [5], can be employed to generate phrase-level alignments. Potentially, the phrase-level combination can produce a final output sentence which is better than any of the input sentences.

2.3 Word-Level Combination

In the word-level combination approach, the candidate word for each word position is considered one by one. A consensus network [1] [6] can be constructed. As shown in Figure 1, the counts of word appearance in a given position based on the optimal word-alignment is used as the edge weights. For each section, the word with the maximum weight is then chosen, constituting the final hypothesis[1]. This idea actually comes from the automatic speech recognition [7].

To generate a consensus network, a skeleton (seed) has to be chosen as the reference for the optimal alignment. In [8], using the output of the consensus network as skeleton and re-aligning all hypothesis leads to a better accuracy.

[1] Note the edge weights may be fine-tuned to reflect the scores of each system.

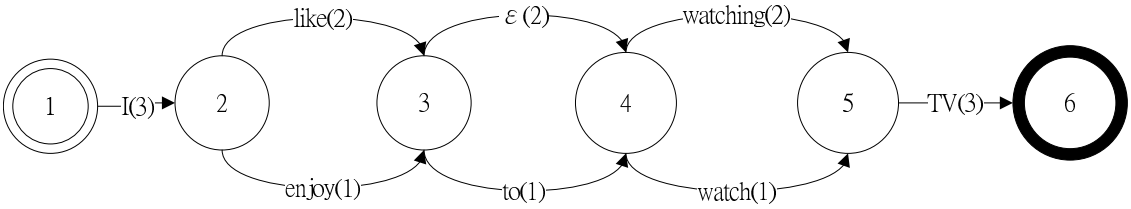


Figure 1: The consensus network of combining "I like watching TV", "I enjoy watching TV" and "I like to watch TV".

3 Method

Our proposed system combines three on-line machine translation systems. Let the source sentence be denoted by C. Given C, the target sentences from these systems are denoted by E_G, E_Y, E_{TW}, respectively for *Google*, *Yahoo*, and *TransWhiz*. With E_G, E_Y, E_{TW} and C as input, the combination system performs the following steps.

- **selection**: One of E_G, E_Y, E_{TW} with the highest language-model score is selected. We denote the selected sentence by E, and the unselected hypotheses as F and G.
- **substitution**: Some words in E are considered and may be substituted. The hypothesis after substitution is denoted by E'.
- **insertion**: Each position in E' is considered to insert an extra word. The hypothesis after insertion is denoted by E''.
- **deletion**: Each word in E'' is considered to be deleted. The hypothesis after deletion is denoted by E^*.

E^* is the final output sentence. The overall process is depicted in Figure 2. We next describe the implementation details.

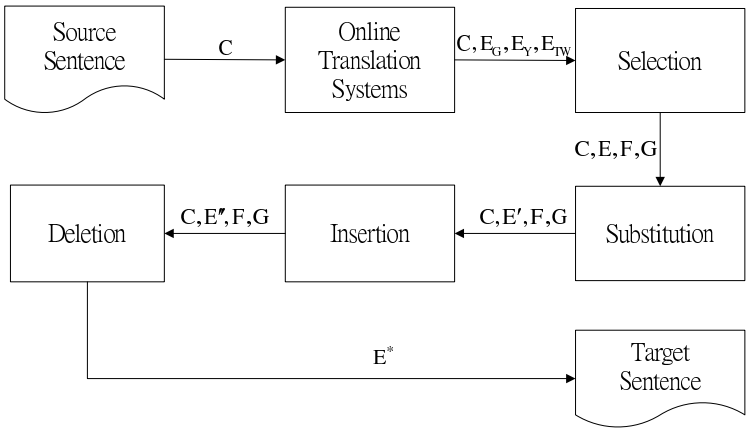


Figure 2: System Organization.

3.1 Selection

The selection is based on a language-model score,

$$E = \arg\max_{H \in \{E_G, E_Y, E_{TW}\}} \frac{1}{|H|} \log\ p_{5g}(H), \tag{2}$$

where H is the hypothesis, $|H|$ is the length of the hypothesis, and p_{5g} is the 5-gram language model probability. The language model used in the selection module is a 5-gram language model trained from the English side of the IWSLT07 training data. Note that in (2), We use the per-word log probability to avoid the (unfair) preference of short sentences. The per-word log probability is that the langauge model score of H devided by its length.

3.2 Substitution

The substitution of words in E is based on the following idea. If a word w appears in both F and G (the unselected hypotheses) but not in E, it is likely to be better to include w in the output. To safeguard against redundancy, we find a word w' in E to be replaced by w. To make sure that such a replacement is a sound operation, we compare the language model scores before and after the word substitution. A statistical machine translation system using MOSES [9] trained by the IWSLT07 data is used to determine the alignments between source and target sentences. The pseudo code for substitution is given in Algorithm 1, and an example for substitution is given in Example 1.

Algorithm 1 Substitution

Require: C, E, F, G
Ensure: E'

1: extract the set of candidate words for substitution;

$$\mathcal{S} = (\{F\} \cap \{G\}) - \{E\}\ ^{2}$$

2: **for all** $w' \in \mathcal{S}$ **do**
3: **if** find the word $c \in \{C\}$ which is aligned to w' **then**
4: **if** find the word $w \in \{E\}$ which is aligned to c **then**
5: compare the translation-model and bi-gram language-model scores to decide whether to replace w by w';
6: **end if**
7: **end if**
8: **end for**

Example 1 (*Substitution*)

The input is

- C : 我想要送這個特快專遞到日本。
- E : *I want to deliver this special delivery to Japan.*
- F : *I want to send this to fast and particularly pass Japan especially.*
- G : *I'd like to send this Speedpost to Japan.*

[2] We use notation $\{E\}$ to denote the set of words in sentence E.

$$\mathcal{S} = \{send\},\ w' = send,\ c = 送,\ w = deliver.$$

The system checks the translation-model score

$$p_t(send|送) > p_t(deliver|送),\ ^{3}$$

and the language-model score

$$\log\ p_{bg}(send|to) + \log\ p_{bg}(this|send) > \log\ p_{bg}(deliver|to) + \log\ p_{bg}(this|deliver),$$

and decides

- *E′ : I want to* send *this special delivery to Japan.*

The reference is

- *R: I want to send this by special delivery to Japan.*

3.3 Insertion

The insertion of words into E' is based on the following idea. If a word w in E' also appears in F or G, we check the adjacent words of w in F or G for possible insertion. The pseudo code for insertion is given in Algorithm 2. An example for insertion is given in Example 2.

Example 2 (*Insertion*)

The input is

- *C :* 你有地鐵地圖嗎？
- *E′ : Do you have subway map?*
- *F : You have subway map?*
- *G : You have a subway map?*

The set of words $\mathcal{I}$ in this example is

$$\mathcal{I} = \{you, have, subway, map\}.$$

The system checks language-model score

$$\log\ p_{bg}(a|have) + \log\ p_{bg}(subway|a) > 2\log\ p_{bg}(subway|have),$$

and decides

- *E″ : Do you have a subway map?*

The reference is

- *R: Do you have a subway map?*

[3] p_t is the translation probability.

Algorithm 2 Insertion

Require: C, E', F, G
Ensure: E''

1: extract the set of words;

$$\mathcal{I} = \{E'\} \cap (\{F\} \cup \{G\})$$ [4]

2: **for all** $w \in \mathcal{I}$ **do**
3: **if** find the word u immediately before w in F or G **then**
4: **if** the bi-gram language-model scores of inserting u before w in E' is larger than the orginal **then**
5: decide inserting u before w in E';
6: **else**
7: consider replacing the word before w in E' by u;
8: **end if**
9: **end if**
10: **if** find the word v immediately after w in F or G **then**
11: **if** the bi-gram language-model scores of inserting v after w in E' is larger than the orginal **then**
12: decide inserting v after w in E';
13: **else**
14: consider replacing the word after w in E' by v;
15: **end if**
16: **end if**
17: **end for**

3.4 Deletion

The deletion of words in $\{E''\}$ is based on the following idea. A word $w \in \{E''\}$ is a candidate for deletion if there is no word $c \in \{C\}$ with nonzero translation probability ($p_t(w \mid c)$). To avoid the deletion of the word in phrases, a candidate word w is deleted only when none of the bigrams formed by w and its immediate neighbors appear in the training data. The pseudo code for deletion is given in Algorithm 3. An example for deletion is given in Example 3.

Algorithm 3 Deletion

Require: C, E'', F, G
Ensure: E^*

1: extract the set of candidate words for deletion;

$$\mathcal{D} = \{w \in \{E''\} \mid t(w \mid c_j) = 0, \ \forall \ j\}$$

2: **for all** $w \in \mathcal{D}$ **do**
3: **if** none of the bigrams formed by w and its immediate neighbors in the training data **then**
4: w is to be deleted;
5: **end if**
6: **end for**

[4] we use the adjacent words of $\mathcal{I}$ as the candidate set for insertion.

Example 3 (*Deletion*)

The input is

- C : 那裡有手工藝品商店？
- E'' : *Where is the handicraft article store?*

The set of words $\mathcal{D}$ *is*

$$\mathcal{D} = \{article\}.$$

The system checks that "handicraft article" and "article store" are neither in the training data and decides

- E^* : *Where is the handicraft store?*

The reference is

- *R: Where is the handicraft store?*

4 Experiments

4.1 Setup

We use IWSLT07 C_E task to run this experiment. IWSLT07 contains tourism-related sentences. The test set consists of 489 Chinese sentences, each of which is accompanied by six reference sentences. Note that the Chinese sentences are word-segmented. The IWSLT07 C_E task we present in the Table 4.1.

Table 1: IWSLT07 C_E task.

	Sentences
Train	39953
Dev	2501
Test	489

We use the on-line machine translation systems of Google[5], Yahoo[6] and TransWhiz[7]. We input the 489 Chinese sentences of the test set to these engines, and get 1,467 English sentences back.

We use the training data in IWSLT07 task to train our 5-gram language model with SRILM [10]. We use MOSES to train the translation model from the training data in IWSLT07 task.

The BLEU [11] measure with six references per sentence is used in our evaluation. The answers are treated as case-insensitive.

[5]http://translate.google.com.tw/translate_t

[6]http://tw.babelfish.yahoo.com/

[7]http://www.mytrans.com.tw/mytrans/freesent.aspx

4.2 Results

The experimental results are presented in Table 4.2. The progressive improvements can be clearly seen in this table. Systems A to C are the three on-line machine translation systems ordered by their performance in BLEU.

- **selection (sel)**: The selection module leads to an absolute improvement of 0.58 BLEU score. Using the language model to select an output from multiple hypotheses is effective, as the selection module selects the most fluent sentence according the 5-gram language model. We think that the selection module can be further improved by joining other features to select the hypothesis.

- **substitution (sub)**: The substitution module leads to a small absolute improvement of 0.07 BLEU score. In this module, a rare word can be replaced by the common word. The candidate set of substitution is small, so we cannot achieve much improvement in this module. Yet it still fixes certain errors in the output. We think that the substitution module can be further improved by replaceing words not only from other hyptoheses but also from dictionaries.

- **insertion (ins)**: The insertion module leads to an absolute improvement of 0.29 BLEU score. It inserts the articles and the adjectives. Given *E* already contains most of the correct words, the improvement is somewhat limited. We think that the insertion module can be further improved by joining words from other source. For exmaple, phrase tables, dictionaries, and others.

- **deletion (del)**: The deletion module leads to an absolute improvement of 0.46 BLEU score. It deletes the redundant words, incorrect words, and out of travel domain words in the output. These words are error sources of our combination hypotheses.

The total improvement over the single best system is 1.4 BLEU absolute.

Table 2: Experimental results.

System	BLEU
System A	19.15
System B	12.39
System C	10.51
+sel	19.73
+sel+sub	19.80
+sel+sub+ins	20.09
+sel+sub+ins+del	20.55

5 Conclusion and Further Work

In this paper, we propose a combination framework that combines the outputs of multiple on-line translation systems. It uses selection module, substitution module, insertion module,

and deletion module. We evaluate out method with the IWSLT07 C_E corpus. The experiments show an overall improvement of 1.4 BLEU absolute.

Our proposed framework changes the hypothesis only locally. In the future, we plan to consider long-range information for better performance. Moreover, our system uses unselected hypotheses to decide which is the incorrect word in the selected hypothesis, but sometimes the wrong words are chosen. Therefore, we may work directly on the words in the selected hypothesis, and only use the unselected hypotheses after problematic words are spotted.

References

[1] B. Bangalore, G. Bordel, and G. Riccardi, "Computing consensus translation from multiple machine translation systems," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01*, 2001, pp. 351–354.

[2] A. Rosti, N. Ayan, B. Xiang, S. Matsoukas, R. Schwartz, and B. Dorr, "Combining outputs from multiple machine translation systems," in *Proceedings of NAACL HLT*, 2007, pp. 228–235.

[3] F. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics Morristown, NJ, USA, 2003, pp. 160–167.

[4] A. Stolcke, "Combination of machine translation systems via hypothesis selection from combined n-best lists," in *Proceedings of the Eighth Conference of the Association for Machine Translation*, 2008, pp. 254–261.

[5] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[6] K. Sim, W. Byrne, M. Gales, H. Sahbi, and P. Woodland, "Consensus network decoding for statistical machine translation system combination," in *Proc. ICASSP*, vol. 4, 2007, pp. 105–108.

[7] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding, 1997. Proceedings.*, 1997, pp. 347–354.

[8] N. Ayan, J. Zheng, and W. Wang, "Improving alignments for better confusion networks for combining machine translation systems," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 33–40.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "Moses: Open source toolkit for statistical machine translation," in *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, vol. 45, no. 2, 2007, p. 2.

[10] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*. ISCA, 2002.

[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation."

不需平行語料而基於共振峰與線頻譜頻率映對之語者特質轉換系統

A Voice Conversion System based on Formant and LSF Mapping without Using Parallel Corpus

吳嘉彧　Chia-Yu Wu
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
u921802@gmail.com

王小川　Hsiao-Chuan Wang
國立清華大學電機工程學系
Department of Electrical Engineering
National Tsing Hua University
hcwang@ee.nthu.edu.tw

摘要

語者特質轉換的研究已有廣泛的運用，早期使用的向量量化碼本對照，與目前被廣爲使用的高斯混合模型，都會使用經動態時軸校準的平行對應語句作訓練。近年來已有減少使用訓練語料與使用非平行句的語料進行語者特質轉換的方法。本論文提出一個不採用平行句的訓練方法，而依據語者音節共振峰映對，並結合線頻譜頻率映對，進行語者特質轉換。

Abstract

Voice conversion has been used in many applications. The methods based on vector quantization codebook and Gaussian mixture models need dynamic time warping on parallel sentence corpus for generating mapping functions. Recent study tries to use less training data, and even without parallel sentence corpus. This paper presents a voice conversion method without using parallel sentence corpus. It applies the formant mapping and line spectral frequency mapping to accomplish a voice conversion system.

關鍵詞：語者特質轉換，平行句語料，共振峰映對，線頻譜頻率映對

Keywords: voice conversion, parallel sentence corpus, formant mapping, LSF mapping

一、緒論

語音轉換和語者特質轉換已被探討多年，目前的研究除了提升轉換相似度以及保持語音品質，也要考慮實用層面會遇到的問題。例如爲了使用者的便利，訓練語料要減少，並要考慮跨語言語者特質轉換等沒有平行對應語句供訓練的情況，因此針對不同用途所使用的轉換方法和訓練語料都要有所調整。

語者特質轉換必須轉換來源語料的頻譜參數與韻律參數，使頻譜與韻律變成有目標語者的特性。最早被提出的頻譜參數轉換採用向量量化碼本對照(Vector Quantization Codebook Mapping)[1]方法，要面對不連續轉換造成音質不佳的問題。其後出現了其他使用類神經網路(Artificial Neural Networks, ANN)[2]和隱藏式馬可夫模型(Hidden Markov Model, HMM) [3]的方法，最被廣泛使用的則是高斯混合模型(Gaussian Mixture Model, GMM)[4]方法，但要面對頻譜過度平滑化的問題。之後有結合高斯混合模型(GMM)和動態頻軸校準(Dynamic Frequency Warping, DFW)[5]的方法被提出，能減低頻譜過度平滑化的情形。以上方法用於訓練的語料，需要是經動態時軸校準(Dynamic Time Warping, DTW)的平行對應語句。

近年來有依據語者共振峰特性所做的頻譜頻軸映對轉換(Formant Mapping)[6]方法，和從線頻譜頻率特性直接對高階數線頻譜頻率做轉換(LSF Mapping)[7]的方法，使用的語料數目減少，但能維持轉換品質。也有結合高斯混合模型於共振峰特性頻譜頻軸映對轉換[8]和多組合高斯混合模型線頻譜頻率轉換[9]等研究，使用的語料數目可以減到更少。

雖然使用經動態時軸校準的平行對應句能增進轉換正確性，但是在實際運用上難以取得平行對應句的語音資料，若是語料沒能準確對應，反而會造成誤差。通常語料內容要平衡音節出現機率，所以設計出的內容不一定符合語者說話習慣，或有些不自然的地方，若是要錄製的語料較多，在錄製者和語者較爲疲憊的情況下，容易有些語料錄製問題發生。

由於針對語言翻譯[10]的跨語言語者特質轉換研究[6][11]興起，如2002年歐洲提出的TC-STAR計劃[10]，辨識完英語後將其翻譯成西班牙文或華文，再用文字轉語音系統結合語者特質轉換系統，將其輸出爲近似原語者發出的西班牙語或華語。因此語者特質轉換要考慮在沒有平行對應語句的情況下，也能調整語者間轉換函式特徵參數，以減少非平行句轉換錯誤[12]。

韻律參數轉換常常使用基週同步疊加法(Pitch Synchronous Overlap and Add, PSOLA)[13]，可以彈性調整語句音調的高低起伏和說話速度。一般是分析語者正常說一句話時的韻律參數，如平均基頻、基頻標準差、音長、音量等等。再針對不同語者之間的韻律參數作轉換，以達到轉換語者特質的目的。

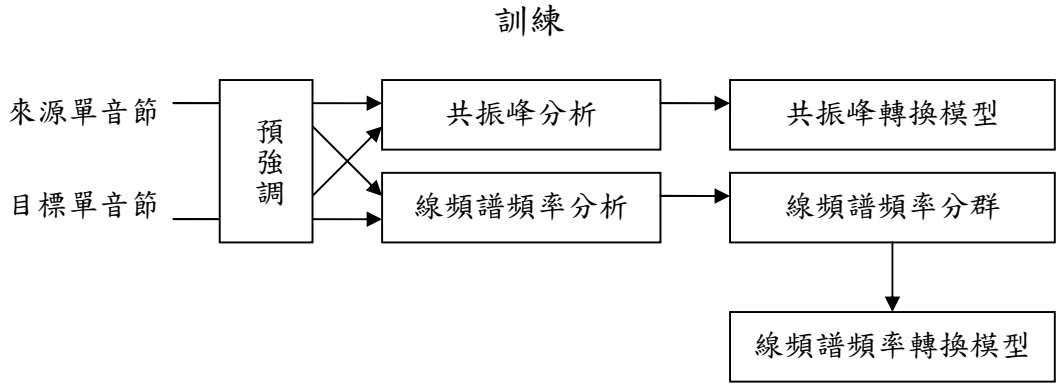
本研究的主要目標，是在不使用平行對應語句訓練的情況下做語者特質轉換，使轉換後的聲音相似於目標語者的聲音，並保持語音品質。在轉換頻譜參數上，本文僅根據國語單音節對共振峰特性做頻譜頻軸映對轉換，並結合線頻譜頻率特性做低階數線頻譜頻率轉換。由較不連續的分段共振峰轉換搭配分群加權平均後頻譜資訊較平滑的線頻譜頻率轉換，達到互補效果。由於單音節的基頻比較不穩定，因此在轉換韻律參數時還是使用短句(不平行)的基頻作爲基週同步疊加法(PSOLA)的基準，考慮單音節基頻變異性，僅將其作爲變異參數進行韻律轉換。

二、語者特質轉換的系統架構

一個語者特質轉換的系統，可以分成訓練和轉換兩個部份。

(一) 訓練部份

圖一展示頻譜參數的映對轉換模型訓練程序，在分別對來源和目標語者的七個主要韻母(ㄚㄛㄜㄝㄧㄨㄩ)單音節進行共振峰分析之後，建立共振峰轉換模型。再對來源和目標語者的全單音節線頻譜頻率做分群，建立線頻譜頻率的映對轉換模型。



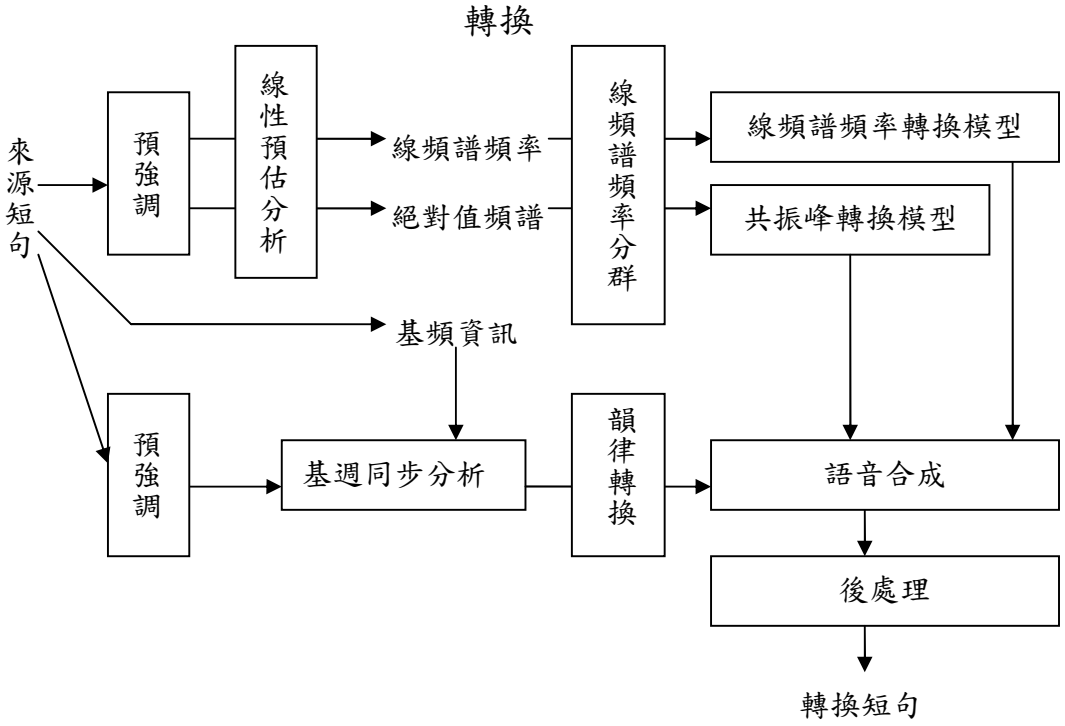
圖一、 訓練部份架構圖

(二) 轉換部份

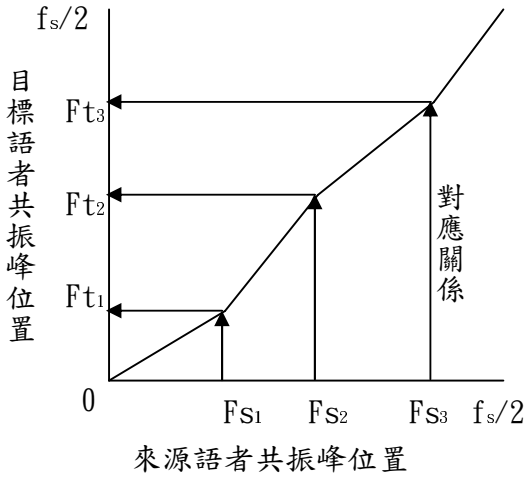
圖二說明轉換方式，對來源語句進行特徵參數抽取後，得到基頻參數，和線性預估絕對值頻譜與線頻譜頻率。基頻參數套進基週同步分析，抓取其基週標記，並取出欲轉換單元(兩倍標記週期)。線性預估絕對值頻譜套進共振峰轉換模型，線頻譜頻率則套進線頻譜頻率轉換模型的參數，搭配韻律轉換的資料進行語音合成，經過後處理就得到轉換後的結果。

三、以共振峰特性做頻譜頻軸映對轉換

由低頻到高頻的前三個共振峰 F_1、F_2與 F_3，常被用於描述語音中母音的差異。線性預估階數越高，能找到的共振峰位置越多，但是若過多也會干擾一對一對應的結果。本文抓取訓練用共振峰時所用的線性預估階數爲 16 和 20 (16 爲主，20 爲輔)。使用七個國語主要韻母(ㄚㄛㄜㄝㄧㄨㄩ)的前三個共振峰，做爲對各語者做頻譜頻軸映對轉換的基準。在 0Hz 到第一個共振峰、第一個共振峰到第二個共振峰、第二個共振峰到第三個共振峰、第三個共振峰到 8000Hz 的對應範圍內，來源語者以雙方共振峰頻率爲基準，頻譜形狀被進行線性壓縮或擴張。各區段內的轉換如圖三與(1)式所示。



圖二、 轉換部份架構圖



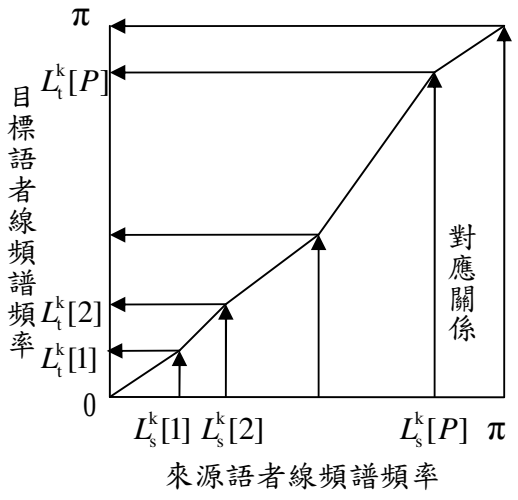
圖三、 共振峰轉換示意圖

$$F(f)=\begin{cases}\dfrac{f\times \mathrm{F_{t1}}}{\mathrm{F_{s1}}} & 0<f\le \mathrm{F_{s1}}\\ \mathrm{F_{t1}}+\dfrac{(f-\mathrm{F_{s1}})\times(\mathrm{F_{t2}}-\mathrm{F_{t1}})}{(\mathrm{F_{s2}}-\mathrm{F_{s1}})} & \mathrm{F_{s1}}<f\le \mathrm{F_{s2}}\\ \mathrm{F_{t2}}+\dfrac{(f-\mathrm{F_{s2}})\times(\mathrm{F_{t3}}-\mathrm{F_{t2}})}{(\mathrm{F_{s3}}-\mathrm{F_{s2}})} & \mathrm{F_{s2}}<f\le \mathrm{F_{s3}}\\ \mathrm{F_{t3}}+\dfrac{(f-\mathrm{F_{s3}})\times(\frac{\mathrm{f_s}}{2}-\mathrm{F_{t3}})}{(\frac{\mathrm{f_s}}{2}-\mathrm{F_{s3}})} & \mathrm{F_{s3}}<f\le \dfrac{\mathrm{f_s}}{2}\end{cases} \quad (1)$$

本研究在進行共振峰轉換後，搭配後續的線頻譜頻率轉換，藉著相對低階(和[7]相比)的線頻譜頻率線性對應，在線頻譜頻率轉換時所用的多個單音節分群加權平均，降低因爲不穩定單音節造成的影響，使頻譜表現較平滑，讓轉換能維持轉換相似度和語音品質的平衡。

四、以線頻譜頻率特性做線頻譜頻率轉換

線頻譜頻率爲線性預估分析中常用到的參數，可以穩定的表現出聲音的頻域特性。透過分析來源和目標語者的線頻譜頻率，將 404 個國語單音節語料分成 K 群，對這 K 群(K = 32)資料求出各階線頻譜頻率的對應轉換函式。



圖四、 線頻譜頻率轉換示意圖

圖四中的$L_s^k[1]$ 、 $L_s^k[2]$到$L_s^k[P]$是來源語者的 P階(P= 16)線頻譜頻率，直接對應到目標語者的 P階(P= 16)線頻譜頻率， $L_t^k[1]$ 、 $L_t^k[2]$ 、・・・、$L_t^k[P]$。0 和 π 是邊界條件，

$$L_s^k[0]=L_t^k[0]=0 \quad (2)$$

$$L_s^k[P+1]=L_t^k[P+1]=\pi \quad (3)$$

(4)式和(5)式描述前後階的線頻譜頻率應符合同一線性對應函式 $f_j^k(\ )$。

$$f_j^k(L_s^k[j-1]) = L_t^k[j-1] \qquad (4)$$

$$f_j^k(L_s^k[j]) = L_t^k[j] \qquad (5)$$

L_s^k 和 L_t^k 分別代表來源語者和目標語者的第 k 群線頻譜頻率值。來源語料線頻譜頻率值 $x[i]$ 若是在第 j-1 到 j 階範圍，就要套進轉換函式 $f_j^k(\ )$，$\tilde{y}[i]$ 即爲轉換後的線頻譜頻率。

$$x[i] \in \left\langle L_s^k[j-1], L_s^k[j] \right\rangle \qquad (6)$$

$$\tilde{y}[i] = f_j^k(x[i]) = a_j^k x[i] + b_j^k \qquad (7)$$

若使用的線性預估階數爲 P，要進行轉換必須先求出這 K 群轉換參數 a_j^k 與 b_j^k，j= 1, 2, …, P+1，k = 1, 2, …, K。每群資料可以使用 P+1 個最小平方差解進行，先將(2)式到(7)式的內容表示成 P+1 個式子，

$$\begin{cases} a_j^k \times L_s^k[j-1] + b_j^k = L_t^k[j-1] \\ a_j^k \times L_s^k[j] + b_j^k = L_t^k[j] \end{cases} \qquad j = 1, 2, \cdots, P+1 \qquad (8)$$

簡化成矩陣函式，

$$S_j^k A_j^k = T_j^k \qquad j = 1, 2, \cdots, P+1 \qquad (9)$$

其中 $S_j^k = \begin{bmatrix} L_s^k[j-1] & 1 \\ L_s^k[j] & 1 \end{bmatrix}$ $A_j^k = \begin{bmatrix} a_j^k \\ b_j^k \end{bmatrix}$ $T_j^k = \begin{bmatrix} L_t^k[j-1] \\ L_t^k[j] \end{bmatrix}$

利用公式 $\hat{A} = (S^T S)^{-1} S^T T$，即可求出第 k 群所需的參數 a_j^k 與 b_j^k，j= 1, 2, …, P+1。

來源語料每一個音框的線頻譜頻率進行轉換前，必須先將其經過一個分群加權，得到對這 K 群的個別加權值，$w_k(x)$，再乘上該群的轉換，這個新的 $\tilde{y}[i]$ 才是此部分轉換的輸出。

$$\tilde{y}[i] = \sum_{k=1}^{K} w_k(x) f_j^k(x[i]) \qquad (10)$$

$w_k(x)$ 是使用馬氏距離(Mahalanobis distance)計算得到。

$$w_k(x) = \frac{1/d_k(x)}{\sum_{n=1}^{K} 1/d_n(x)} \qquad (11)$$

(11)式中的 $d_k(x)$ 即爲馬氏距離，

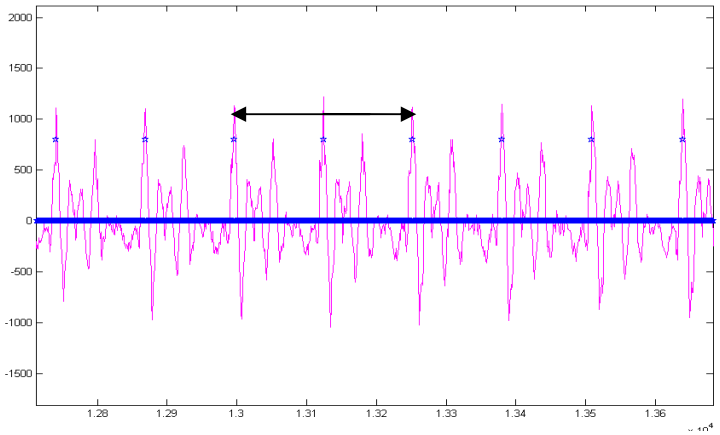
$$d_k(x) = [(x - L_s^k)^T (\sum_s^k\ )^{-1} (x - L_s^k)]^{\gamma} \qquad (12)$$

(12)式中的 $\sum_s^k$ 爲 $L_s^k[1]$ 到 $L_s^k[P]$ 的對角共變異數矩陣(diagonal covariance matrix)，γ 爲一可調整參數，其值越大，對距離越近的分群比重越大。本研究使用的 γ 爲 4，和[7]相同。所使用的線性預估階數爲 16，相較於使用高階數的做法[7]，在頻譜上的表現較爲平滑，用以補償經過分段共振峰轉換造成的不連續。

五、基週同步分析及韻律參數轉換

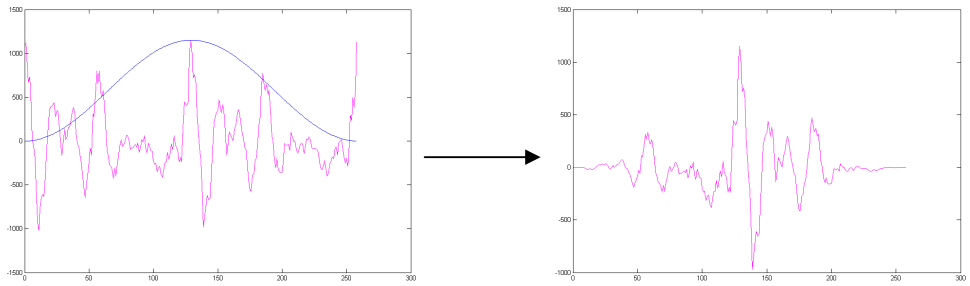
(一) 基週同步分析

爲了後續語音合成使用的基週同步疊加法(Pitch Synchronous Overlap and Add, PSOLA)，必須先對語料進行基週同步分析。以 ACF(Autocorrelation function)除以 AMDF (Average Magnitude Difference Function)的值估算基頻，完成基頻估算後，在每個有聲音段找到能量最大音框中振幅最大的取樣點，作爲第一個標記。從此標記依序往前與往後搜尋對應音框基本週期範圍，出現該區域最大值的位置，即爲下一個基週標記，重覆此動作可得到有聲音段的所有基週標記。



圖五、 基週標記

圖五爲基週標記示意圖，星號位置就是基週標記點，以基週標記前後兩個基週範圍(箭號所示)作爲運算單位，在此單位中的語音加上漢寧窗後，即爲韻律轉換所需的基本單元，如圖六。



圖六、 基本單元

(二) 發音腔道模型

利用線性預估分析，發音腔道模型可以表示成一個全極點系統(all-pole system)，

$$X(z) = \frac{\Theta_0}{1 - \sum_{j=1}^{P} a_j z^{-j}} E(z) \qquad (13)$$

$X(z)$是語音訊號，Θ_0是增益常數，$E(z)$是激發訊號。分母的$1 - \sum_{j=1}^{P} a_j z^{-j}$是一個逆向濾波器(inverse filter)，a_j (j= 1, 2, …, P)是線性預估係數(linear prediction coefficients)，P爲線性預估階數。

(三) 韻律參數轉換

將語音訊號經過其對應的逆向濾波器，就得到剩餘訊號。如果對剩餘訊號的基頻軌跡作轉換，得到新的基頻軌跡，這個轉換後的剩餘訊號，用以產生轉換後的語音訊號。對平均基頻做調整，可以轉換語者說話的音高。對基頻標準差做調整，可以轉換語者說話的起伏。本文用於實驗的平均基頻和基頻標準差，是在語料庫中任意選取三個短句算出的平均值。由於考慮不同音節的基頻變異性，在基頻標準差比值上多乘上一個變異參數C_{kst}。

$$f_{0c} = C_{kst} \times \frac{\sigma_{ft}(f_{0s} - \quad_{fs})}{\sigma_{fs}} + \quad_{ft} \qquad (14)$$

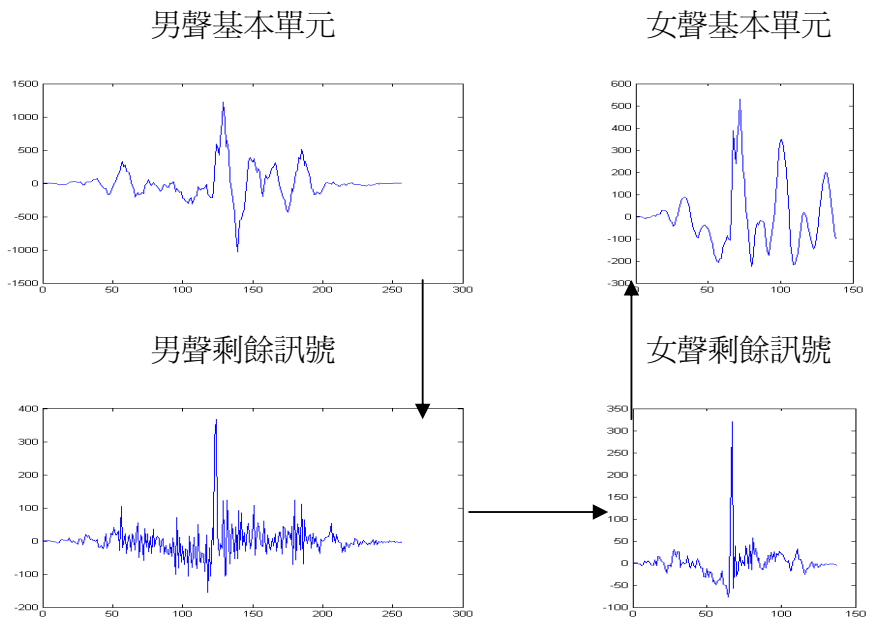
C_{kst}是以線頻譜頻率爲分群基準，計算「來源—目標」基頻比值F_{kst}得出：

$$C_{kst} = F_{kst} \Big/ \frac{1}{K}\left(\sum_{n=1}^{K} F_{nst}\right) \qquad (15)$$

六、語音合成

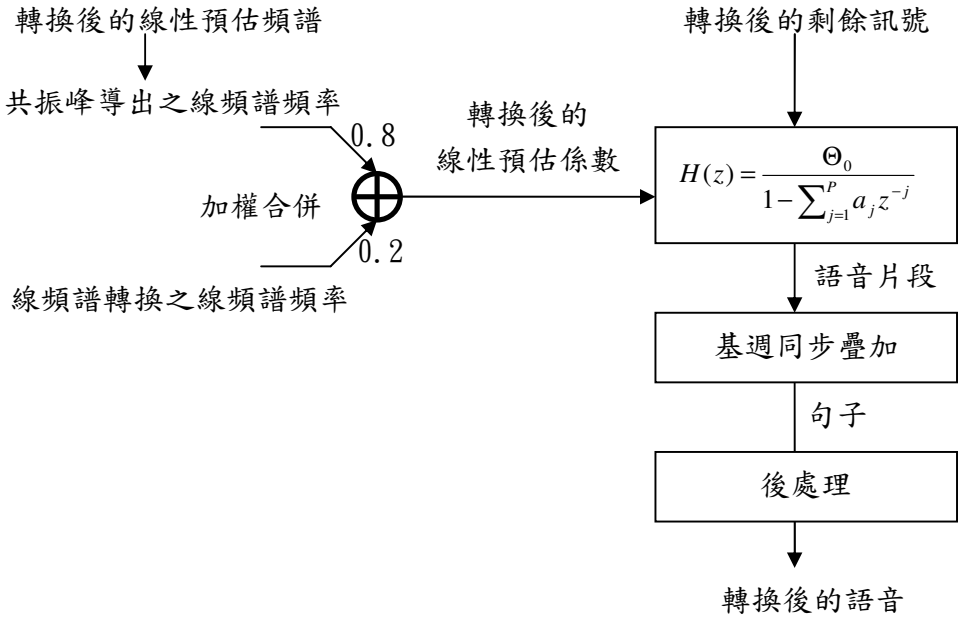
以 16ms 爲固定音框長度，8ms 爲音框移動距離，使用線性預估分析計算出該音框的線性預估頻譜和線頻譜頻率後，所得之絕對值頻譜形狀經由第三節共振峰頻譜分析轉換後，再將其表示成線頻譜頻率，稱之爲「共振峰導出之線頻譜頻率」。利用第四節線頻率頻譜轉換得到的線頻譜頻率則叫做「線頻譜轉換之線頻譜頻率」。將「共振峰導出之線頻譜頻率」和「線頻譜轉換之線頻譜頻率」兩者進行加權合併，產生新的線頻譜頻率，再轉爲線性預估係數，用以構成語音合成時的發音腔道全極點模型。每個不同長度的基本單元依其時間點，可以對應到進行頻譜資訊轉換時使用的固定音框序列。在時間上相對應的固定音框線性預估係數結合經過基週同步分析基本單元的剩餘訊號，可合成出一段語音訊號。

圖七爲基本單元的轉換示意圖(男聲→女聲)，將這些轉換完的語音片段經過基週同步疊加後，再經過後處理，即爲轉換後的語音。



圖七、 基本單元轉換

整個語音合成的流程見圖八。



圖八、 語音合成流程圖

七、實驗

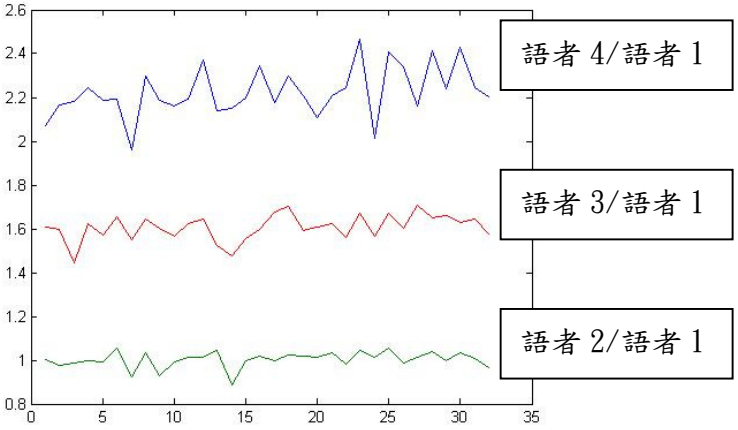
(一)實驗語料

所用語料經由麥克風錄製，取樣頻率爲 16kHz, 16 bits PCM。共有四位語者(兩男兩女)，每位皆有 404 個國語單音節及 110 句國語短句。本論文的轉換實驗訓練部份皆使用單音節進行，短句僅作爲韻律轉換的參考基準。使用語者資料如下：

表一、 語者資料-基頻

	性別	單音節平均基頻 (Hz)	短句平均基頻(Hz)
語者 1	男	126.07	125.22
語者 2	男	128.89	124.44
語者 3	女	206.79	188.10
語者 4	女	288.57	229.79

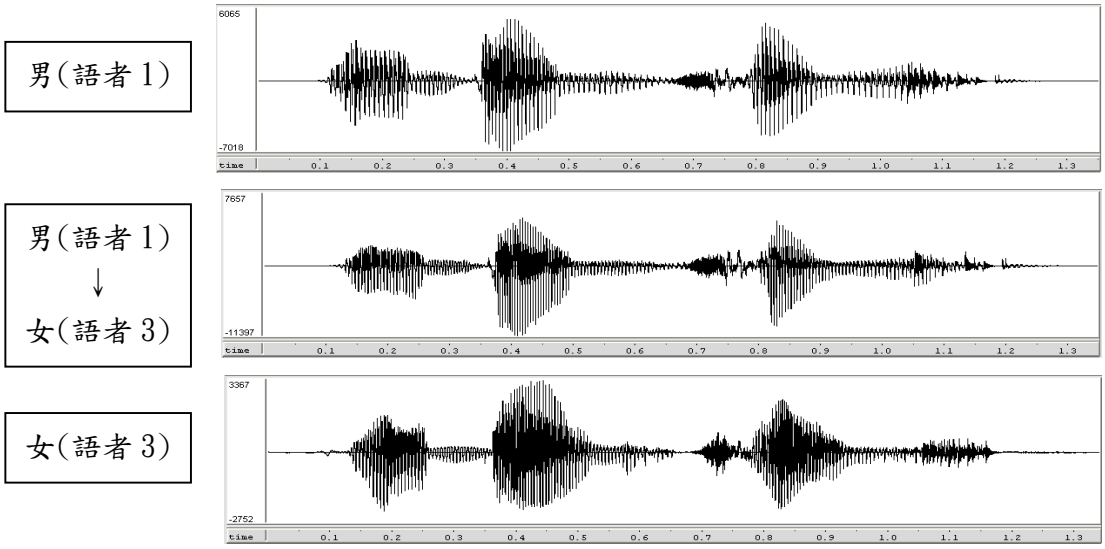
由於語者 4 發單音節的音高皆比平常說話時還高出許多，所以無法以平均基頻變異性做分段韻律轉換基準。但若以語者 1 的線頻譜頻率爲基準概分 32 群後，語者 2、3、4 的對應分群基頻和語者 1 的分群基頻比值，可以看出不同音節的基頻比值會有不同趨勢，考慮音節基頻變異性，可用於調整基頻標準差於改變音高起伏程度。



圖九、 單音節分群後的基頻比值趨勢圖

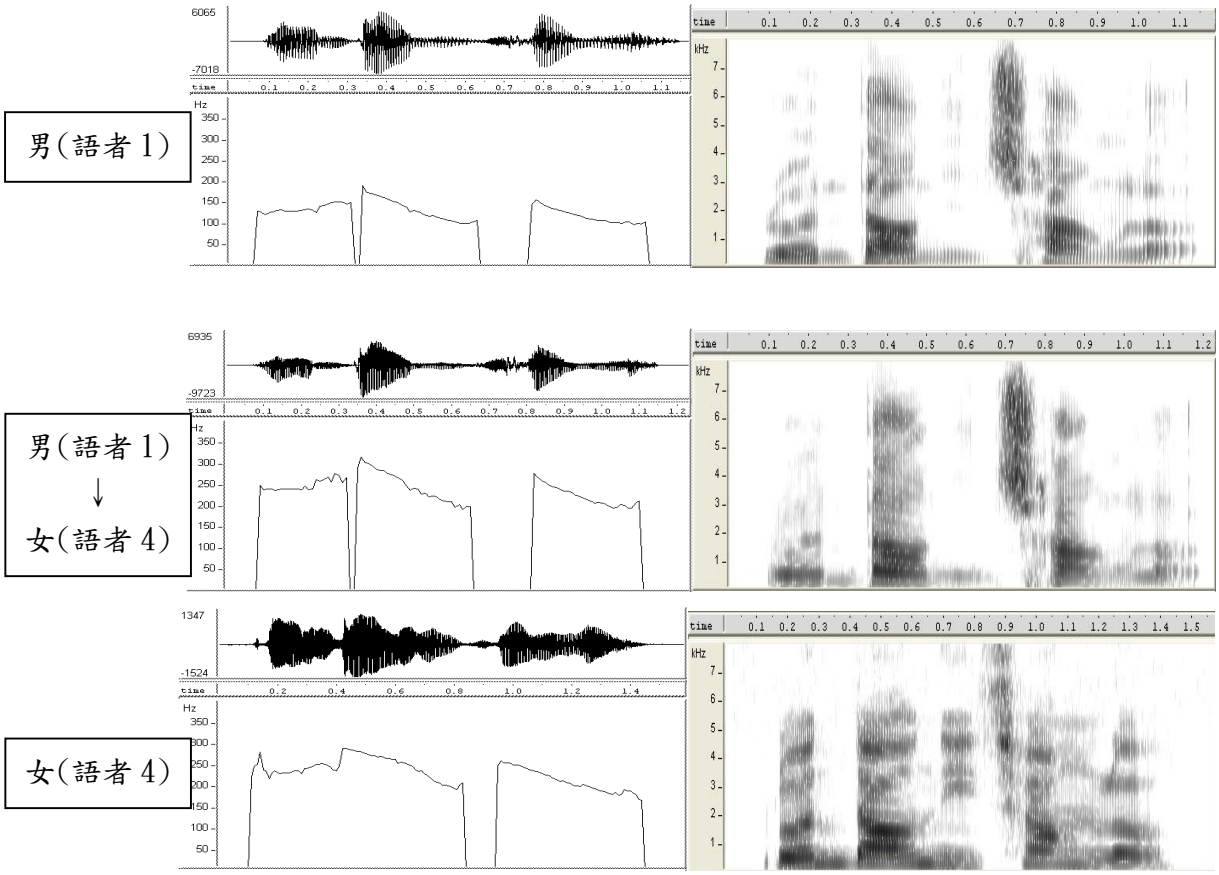
(二)轉換結果

圖十是語者 1(男)轉換到語者 3(女)的波形對照，轉換後的語音波形與來源語音波形比較，已有所改變。其中語者 3 的語料音量原本就較小，所以轉換的語音振幅強度對照語者 1 的強度做了調整。基本上轉換後的語音節奏受到來源語音的影響，但是音質接近目標語者的語音。



圖十、 語者特質轉換波形對照圖(輪到你唱了)

圖十一是語者 1(男)轉語者 4(女)的波形、基頻、聲譜對照圖，從圖中可看到將原本男生(語者 1)較低的基頻軌跡轉爲女生(語者 4)較高的基頻軌跡，而基頻軌跡的形狀由來源語音決定。



圖十一、 語者特質轉換波形、基頻、聲譜對照圖(輪到你唱了)

(三)主觀聽覺實驗

主觀聽覺實驗包含音質和相似度兩部份，兩個實驗都有 9 位未經過特別聽力訓練的受測者參與，其中包含了一位不熟悉國語，平常僅使用越南語和英語的受測者。受測者在實驗中各聽 35 個不重複的國語短句，其中混合了對不同目標進行的轉換結果及未轉換的語者原始語料。受測者不知道轉換內容及其是否爲轉換過的語料，藉此可得到語料經語者特質轉換前後的比較。評定方法如下：

(1) 音質

使用平均主觀分數(Mean Opinion Score, MOS)，依照語音品質由高到低分爲五級。

(2) 相似度

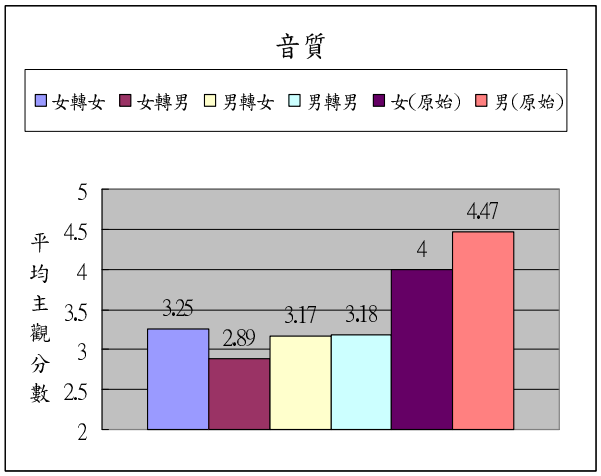
先讓受測者聽兩位語者(A 和 B，句子內容相同)的原始語料，再請受測者聽一個未知語料(X，句子內容不同於 A 和 B)，此未知語料可能是轉換後的語料(由 A 轉 B 或由 B 轉 A)，也可能是兩位語者的其他語料(A 或 B 的其他句子)，相似度部分有五個選項，不依選項大小判斷相似與否，而是要再從選項內容判斷。

表二、 相似度選項依據 (A 爲正確答案)

5	X 聽起來肯定是 A
4	X 聽起來接近 A，但不確定是否真的是 A
3	無法分辨 X 比較像 A 還是 B
2	X 聽起來接近 B，但不確定是否真的是 B
1	X 聽起來肯定是 B

(3) 實驗結果

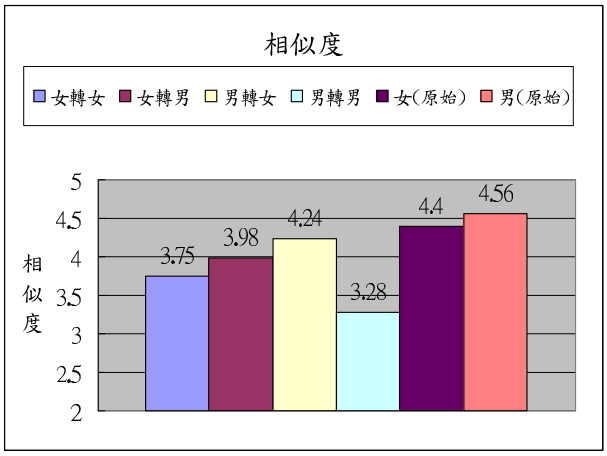
由於受測者都未接受過特別的聽力訓練，所以經過轉換的語料和語者原始的語料都進行音質評定，作爲轉換前後音質落差的參考。



圖十二、 各項轉換的音質比較

圖十二明顯看出經過轉換處理的語音，其音質變差，整體來說，平均主觀分數從 4.2 下變成 3.1，下降約 1.1。

由於本研究的語者特質轉換沒有使用平行對應句，所以相似度測驗也包含了各語者未經轉換的不同句子，作爲轉換前後相似度落差的參考。



圖十三、 各項轉換的相似度比較

圖十三展示經過語音轉換之後，與目標語者語音的接近程度，其中男聲轉男聲時轉換語音與目標語音的相似度最低，因爲這兩位男性語者的聲音比較相似。男聲轉女聲時，其效果最好。受測者並不知道是原始語音或是轉換語音，所以照樣評分，這兩個原始語音的相似度分數只作爲參考。

八、 結論

考量語料取得容易度對語者特質轉換系統實用上的影響，本研究不使用平行對應句，僅使用對應單音節建立轉換模型。對應單音節容易取得，但由於資料量小，對其穩定度的要求就要相對提高，才能使轉換相似度和聲音品質和使用平行對應句一樣好。期望將來的研究能再減低對特定語料的依賴性，並且維持良好的轉換相似度和轉換後的聲音品質。

參考文獻

[1] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., “Voice Conversion through Vector Quantization,” Proc. IEEE ICASSP pp.665-658, 1988.

[2] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” Speech Communication, vol. 16, pp. 207-216, 1995.

[3] E. K. Kim, S. Lee, and Y. H. Oh, “Hidden Markov model based voice conversion using dynamic characteristics of speaker,” Proc. EUROSPEECH, vol. 5, Rhodes, Greece, 1997.

[4] Stylianou, Y., Cappe, O., and Moulines E., “Continuous Probabilistic Transform for Voice Conversion,” IEEE Trans.on Speech and Audio Processing, vol.6, no.2, pp.131-142, 1998.

[5] Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, “Voice Conversion Algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT Spectrum,” Proc. IEEE ICASSP, pp. 841-844, 2001.

[6] Zhiwei Shuang, Raimo Bakis, and Yong Qin, “Voice Conversion Based On Mapping Formants,” TC-STAR Workshop on Speech-to-Speech Translation, pp.219-223, 2006.

[7] Zdenek Hanzlicek, Jindrich Matousek, “On Using Warping Function for LSFs Transformation in a Voice Conversion System,” Proc. IEEE ICSP 2008.

[8] Kun Liu, Jianping Zhang, and Yonghong Yan, “High Quality Voice Conversion through Combining Modified GMM and Formant Mapping for Mandarin,” IEEE ICDT 2007.

[9] Elina Helander, Jani Nurminen, and Moncef Gabbouj, “LSF Mapping for Voice Conversion with very small training sets,” IEEE ICASSP 2008.

[10] H. Höge, “Project Proposal TC-STAR - Make Speech to Speech Translation Real,” Proc. LREC’, Las Palmas, Spain, 2002.

[11] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, “TC-Star: cross-language voice conversion revisited,” TC-Star Workshop on Speech-to-Speech Translation, 2006.

[12] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, “Nonparallel Training for Voice Conversion Based on a Parameter Adaptation Approach, ” IEEE Trans.on Speech and Audio Processing, vol.14, no.3, 2006.

[13] H. Valbret, E. Moulines, and J. P. Tubach, “Voice Transformation using PSOLA Technique, ” Proc. IEEE ICASSP. San Francisco, USA, pp. 145-148, 1992.

[14] Yinqiu Gao, Zhen Yang, “Pitch modification based on syllable units for voice morphing system,” IFIP International Conference on Network and Parallel Computing Workshops 2007.

基於盲訊號分離語音增強技術之遠距離雜訊語音辨識

Speech Enhancement Technique Based on Blind Source Separation for Far-Field Noisy Speech Recognition

李聖捷 Sheng-Chieh Lee
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
n2897134@mail.ncku.edu.tw

王駿發 Jhing-Fa Wang
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
wangjf@csie.ncku.edu.tw

陳淼海 Miao-Hai Chen
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
n2696149@mail.ncku.edu.tw

摘要

語音辨識在語音處理領域中，爲其重要的一項研究領域項目之一，然而語音辨識中的辨識結果會隨著語音所在環境及語音距離而有所影響，在此本論文提出一套高適應性遠距離雜訊語音辨識系統，此系統結合獨立成分分析方法以及子空間語音增強方式，將雜訊語音進一步濾除噪聲並提升語音訊號強度以供辨識。實驗結果顯示，本論文所提出的語音辨識系統，可適用在各種噪聲環境當中，並有效的改善其辨識率，以及將原本信噪比 0dB 到 10dB 範圍中的帶噪語音，提升至 20dB 以上。

Abstract

Speech recognition is one of the im portant p arts of search field in speech processing. Nevertheless, the speec h environment and speec h distance w ill mainly affect the reco gnition result. In this paper, a high adaptation far-fie ld noise speech recognition system is proposed. This system is com bined with the m ethods of independent com ponent analysis and subspace speech enh ancement, and then furth er filteri ng the noise of speech to improve the speech quality for recognition. The experimental results show that the proposed system is suitable for several presented noisy environments, and it can effectively improve the recognition rate. For the SNR evaluation, this proposed sy stem can make enhanced speech SNR with 20dB higher than original corrupted speech which ranges from 0dB to 10dB.

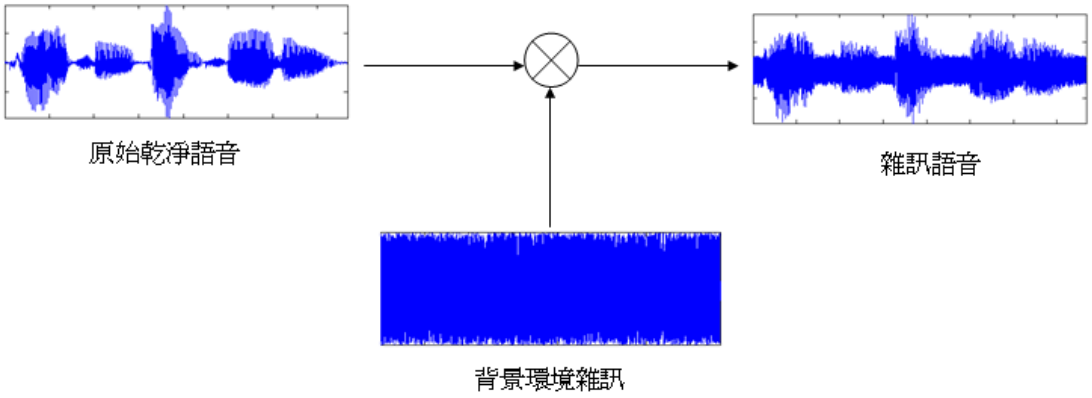
關鍵詞：語音辨識，盲訊號分離法，獨立成分分析，子空間語音增強，麥克風陣列

Keywords: Speech Recognition, Blind Source Separation, Independent Component Analysis, Subspace Speech Enhancement, Microphone Array.

一、緒論

語言爲人類彼此溝通時，最原始同樣也是最有效的方式，在科技蓬勃發展的現今，如何使電腦辨識人類語言也成爲語音處理上重要議題其中之一，因此對於語音辨識系統，如何達到有效且精確的辨識結果，也是目前語音處理領域中熱門的研究議題。

對於語音辨識結果，影響語音辨識結果的相關因素很多，這些相關因素都會造成語者語意和語音辨識結果的不匹配(mismatch)，其中影響辨識結果最重要的因素爲環境中所存在的背景雜訊，由於語音所存在的背景環境中，並非完全沒有遭受其他干擾雜訊影響，例如在餐廳環境、地鐵環境、車內行駛環境等，都有背景雜訊的干擾源存在，這些背景雜訊伴隨著語音進入辨識系統中，會嚴重影響到整體辨識結果，另外語者與辨識系統距離也是另一種影響辨識結果的因素，語音能量會伴隨著距離而逐漸衰減，因此衰減後的語音能量也會造成辨識率的降低。



圖一、背景環境雜訊干擾語音示意圖

爲了改善上述所提到之環境雜訊以及語者距離所造成的辨識結果不匹配，我們針對此雜訊語音做進一步分析，首先雜訊語音中包含了大量的雜訊資訊，因此如何取得雜訊部份並加以去除爲第一步重要的處理步驟，濾除相關的背景雜訊後，再來則是語者和辨識系統之間的距離問題，當距離相距越大時，語音辨識系統所接收到的語音能量則越小，因此對於濾除雜訊後的語音訊號，必須再進一步使用語音增強技術將加強語音訊號能量，以提升之後的辨識結果，最後在進行辨識之前，再將增強後之語音訊號做端點偵測處理，找出一段語音訊號中語音的實際位置再取得此語音資訊來進行辨識。

根據上述分析結果，在雜訊分離部份，我們採用盲訊號分離(Blind Signal Separation, BSS)的方法，使用獨立成分分析(Independent Component Analysis, ICA)方式來進行訊號分離，取出相近似語音成分較多的部份，再透過子空間語音增強方式(Subspace Speech Enhancement)，將取出的語音訊號進一步去除殘餘噪聲並加強語音訊號，使其可用來進行語音辨識之用，最後再利用語音活動偵測法(Voice Activity Detection, VAD)來偵測語音所在位置，藉此來提升辨識效率。最後在末端的語音辨識器方面，我們使用英國劍橋大

學所提供的 HTK(Hidden Markov Model T oolkit)語音套件來進行識別，並判斷所產生的結果是否正確。

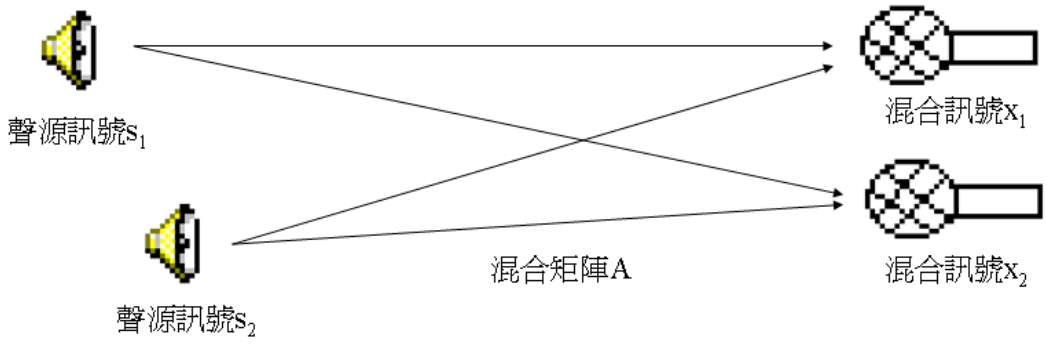
本論文總共分成五個章節，第一章節爲緒論，第二章節爲本論文針對此辨識系統所採用之各種研究方法並詳細加以介紹，第三章節則是介紹此語音辨識系統之系統架構，第四章節則是實驗環境評估和設定以及實驗結果，最後第五章節則是對此辨識系統做一精要結論及未來相關工作。

二、研究方法

本章節針對此遠距離雜訊語音辨識系統所採用的各種方法來加以詳述說明介紹。

（一）獨立成分分析法(Independent Component Analysis, ICA)

對於帶有噪聲的語音成分，由於原始語音成分和背景雜訊成分均爲未知，因此要分離此兩種未知訊號，我們可使用盲訊號分離方式，將此兩種未知訊號，分別從混合訊號中分離出來，一般盲訊號分離問題可由下面示意圖表示：



圖二、盲訊號分離問題示意圖

如圖二所示，兩未知聲源訊號 s_1 及 s_2，透過混合矩陣 A 後，在麥克風接收端則會接收到兩種混合訊號 x_1 和 x_2，此關係可由下列線性方程式表示。

$$\begin{aligned} x_1 &= a_{11}s_1 + a_{12}s_2 \\ x_2 &= a_{21}s_1 + a_{22}s_2 \end{aligned}, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{1}$$

因此若假設聲源訊號 s_1 爲語者的語音成分、s_2 爲噪音成分，我們可從所接收到之混合訊號 x_1 及 x_2 分離出原始的語音訊號以及噪音訊號，即可有效的去除雜訊，根據上述公式(1)，要求得原始訊號 x_1 和 x_2，必須找出一個解混合矩陣 $\overline{A}$ ，使得接收到的混合訊號經由 $\overline{A}$ 轉換後，可得到原來的聲源訊號，而此求解 $\overline{A}$ 之方法即爲獨立成分分析法。

在使用獨立成分分析法求得解混合矩陣前，必須先行假設訊號源彼此獨立，然而在真實情況下，訊號源並非都會彼此互相獨立，因此在進行獨立成分分析流程前，必須先經過前置處理後才能找尋解混合矩陣，在此我們前置處理方式爲集中變數(Centering)以及資料白色化(Whitening)處理，在此先針對集中變數及資料白色化來作爲說明。

1 集中變數(Centering)

集中變數的處理步驟主要是將混合訊號扣除其平均值，藉此簡化之後求得解混合矩陣之求解過程，其公式如下所表示。

$$\hat{x} = x - E\{x\} \tag{2}$$

此外我們將接收到的混合訊號做集中變數處理後，也同樣的對於聲源訊號做了集中變數處理，如公式(3)所示。

$$E\{s\} = E\{\overline{A}\hat{x}\} = \overline{A}E\{\hat{x}\} = 0 \tag{3}$$

2 資料白色化(Whitening)

至於前置處理的第二步驟就是資料白色化，資料白色化的目的在於將轉換後的資料彼此間具有非相關性(Uncorrelated)且變異數(Variance)數值爲一，在此假設轉換資料爲 z，則此資料之共變異矩陣(Covariance matrix)會成爲單位矩陣。因此資料白色化的方式爲找出一白色化矩陣 V，並將所接收到的訊號 x 做線性轉換且使其共變異矩陣爲單位矩陣。

$$z = Vx, E\{zz^T\} = I \tag{4}$$

3 解混合矩陣(De-mixing matrix)

做完前置處理後，再來則是計算最大非高斯分佈訊號，根據中央極限定理，將多個非高斯分布且彼此獨立的訊號個別加總後，會使得整體傾向於高斯分佈，因此若任兩個隨機訊號越傾向非高斯分佈，則此兩訊號彼此獨立的成分就越大，再此高斯分佈訊號具有疊加性，兩個高斯分佈訊號相加總後的訊號仍爲高斯分佈，所以若由高斯訊號線性混合而成的群集是無法分離出真正的原始訊號，因此在使用獨立成分分析來分離訊號時，必須事先假設只能允許其中一個訊號爲高斯分佈，在估算非高斯分佈訊號部份，本論文是採用負熵(Negentropy)來評估計算，其中熵(Entropy)的定義根據離散訊號或連續訊號可由下列公式所表示：

$$H(y) = -\sum P(y)\log P(y) \tag{5}$$

$$H(y) = -\int f(y)\log f(y)dy \tag{6}$$

在語音訊號部份中，當訊號 y 爲高斯分佈時，其熵爲最大值，因此爲了計算方便我們使用負熵來作爲依據，如公式(7)所示，y_{gauss} 爲和 y 有相同變異矩陣之高斯分佈訊號，因此當訊號 y 爲高斯分佈時，則負熵爲零，爲了簡化其計算，我們將公式(7)簡化爲公式(8)。

$$J(y) = H(y_{gauss}) - H(y) \tag{7}$$

$$J(y) \approx [E\{G(y)\} - E\{G(v)\}]^2 \tag{8}$$

其中 G 爲對照方程式，訊號 v 爲平均值爲零變異數爲一之高斯分佈訊號，一般來說對照方程式不能爲二次式函數或多項式函數，在此我們選擇的對照方程式如下所示。

$$G_1(y)=\frac{1}{a_1}\log(\cosh(a_1 y)) \quad ,a_1\text{爲一常數} \tag{9}$$

$$G_2(y)=-\exp(-\frac{y^2}{2}) \tag{10}$$

$$G_3(y)=y^4 \tag{11}$$

根據上面對照方程式，我們設定 E{G(y)}=E{G(W^Tx)}，W 爲解混合矩陣，x 爲混合訊號，因此公式(8)可改寫成公式(12)，當 E{G(W^Tx)}爲最大時，則可找到非高斯分佈性最高的語音訊號，最後再利用牛頓法疊代運算，將解混合矩陣 W 求解出來。

$$J(W)\propto[E\{G(W^T x)\}-E\{G(v)\}]^2 \tag{12}$$

$$W\leftarrow E\{xG(W^T x)\}-E\{G'(W^T x)\}W \tag{13}$$

（二）子空間語音增強法(Subspace Speech Enhancement)

經由獨立成分分析法，我們可將混合訊號分離出兩個訊號，其中一個訊號其語音成分較大，另一個則是雜訊成分較大，然而含語音成分較多的訊號中，仍舊會殘留些許雜訊部份，因此我們使用子空間語音增強法來進一步加強處理，濾除訊號中的噪聲雜訊。

在訊號子空間的假設中，可將觀測訊號的向量拆解成兩個子空間，一個爲由乾淨語音訊號組合而成的子空間，另一個是與乾淨語音空間正交(orthogonal)且由噪音所組成的子空間，由於噪音所組成的子空間沒有任何語音資訊因此可將此忽略，而乾淨語音訊號的子空間中，仍舊會有噪音成分與其並存，例如各頻帶皆有可能存在的白噪音(White noise)，因此要根據噪音成分的分佈來處理，還原出沒有雜訊訊號的語音子空間。

我們假設訊號子空間中乾淨的語音成分可由一線性模型組成，如公式(14)所示，其中 W_S 爲一 NxM 且 M 小於 N 的矩陣，x_S 爲 Mx1 的向量，則此訊號向量 y 爲一個由 W_S 所建立的歐基里德空間 R^N 裡的一個集合，而此空間就是訊號子空間。

$$y=W_S x_S \tag{14}$$

因此原始混合訊號即爲原本的語音訊號子空間 y 再加上另一個噪音訊號子空間 n_S，如(15)式所示，由於本論文的方法是採用在時間域(Time domain)下的估算，所以在此就直接探討在時間域下的相關估測。

$$z=W_S x_S+n_S=y+n_S \tag{15}$$

根據上式的混合訊號，我們必須找出一個 NxN 的濾波器 F，使得混合訊號經由濾波後能得到乾淨的訊號 y'=Fz，而濾波後的訊號與原訊號相比較可計算其濾波器 F 的誤差，其誤差值 δ 計算如下：

$$\delta=y'-y=(F-I)y+Fn_S=\delta_y+\delta_{n_S} \tag{16}$$

其中 δ_y 表示被濾波器濾除的語音訊號失真，δ_{n_s} 表示沒有被濾除的噪音所產生的失真，因此我們計算這兩種失真誤差的變異數當成強化後的誤差能量。

$$\bar{\delta}_y = E\{\delta_y^T \delta_y\} \tag{17}$$

$$\bar{\delta}_{n_S} = E\{\delta_{n_S}^T \delta_{n_S}\} \tag{18}$$

藉由(17)式和(18)式再和(16)式相比較，若要對訊號子空間中的濾波器作最佳化處理，對於語音訊號部份，語音失真的程度要最小，對於噪音訊號部份，殘留的噪音只要盡量抑制到不至於影響辨識結果的程度就好，而非要求完全沒有殘存的噪音成分存在，在如此折衷的條件下我們將此濾波器的最佳化條件以(19)式來表示。

$$\begin{aligned} &\min \bar{\delta}_y \\ &\bar{\delta}_{n_S} \le \gamma\sigma^2, 0 \le \gamma \le 1 \end{aligned} \tag{19}$$

其中 σ^2 爲噪音的變異數，γ 爲調整控制濾波器殘留噪音訊號的程度，因此我們使用 Lagrange 方法來計算此最佳化濾波器條件，推導結果如下，μ 爲 Lagrange m ultiplier，R_y 和 R_{n_s} 分別爲語音訊號和噪音訊號的共變異矩陣，若將 R_y 使用特徵值分解，假設 $R_y=PD_yP^T$，P 爲特徵向量矩陣且 D_y 爲特徵值對角矩陣，則(20)式可分解成(21)式，最後再將 $P^T R_{n_s} P$ 用噪音訊號的特徵值對角矩陣近似，得到最後訊號子空間濾波器最佳化處理結果。

$$\hat{F} = R_y (R_y + \mu R_{n_S})^{-1} \tag{20}$$

$$\hat{F} = PD_y (D_y + \mu P^T R_{n_S} P)^{-1} P^T \tag{21}$$

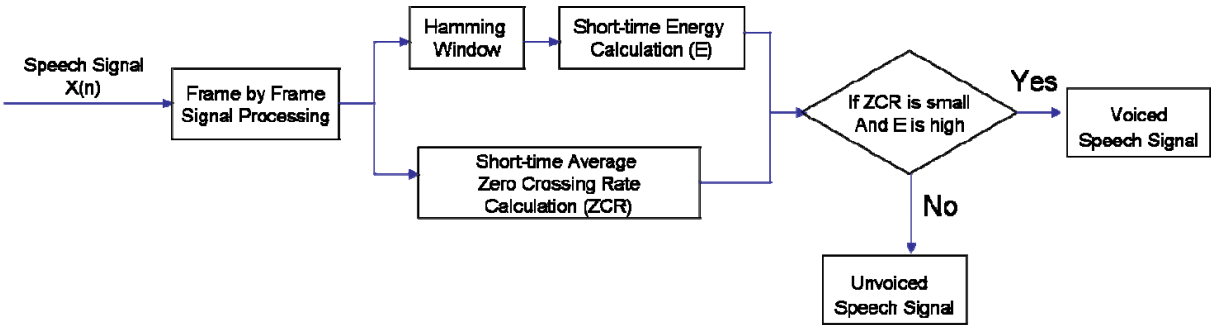
$$\hat{F} = PD_y (D_y + \mu D_{n_S})^{-1} P^T \tag{22}$$

（三）語音活動偵測法(Voice Activity Detection, VAD)

在語音活動偵測法上，我們利用語音訊號的能量曲線和過零率(Zero crossing rate)來進行語音訊號的端點偵測，一開始我們預先在語音訊號波形上設定一條基準線，當訊號振幅在此基準線上方定義爲正，反之定義爲負，再來則針對訊號中每個音框，個別計算振幅由正到負、以及由負到正的次數，若單位時間內越過基準線次數增多，表示訊號波形擺動越劇烈。對於一段含雜訊之語音訊號，雜訊或氣鼻音能量較小且過零率較高，而語音部份則是語音能量較高且過零率低，因此可藉由能量曲線以及過零率來針對每段語音訊號進行端點偵測處理。在此假設每個音框包含了 N 的樣本點，則過零率的計算方式如下所示。

$$\begin{aligned} &\mathrm{ZCR} = \frac{1}{2}\sum_{n=1}^{N-1} |\,\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]\,| \\ &\mathrm{sgn}[x(n)] = 1 \text{ if } x(n) \ge 0, \mathrm{sgn}[x(n)] = -1 \text{ if } x(n) < 0 \end{aligned} \tag{23}$$

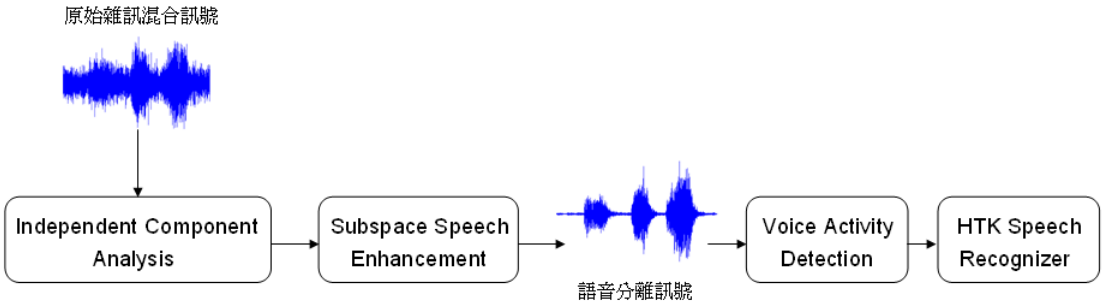
其中 x(n)表示第 n 個樣本點的振幅能量，x(n-1)表示爲前一個樣本點，因此過零率是指兩連續樣本間，具有不同的正負號次數。取出正確的語音訊號後即可開始進行辨識，下圖爲語音活動偵測法整體流程圖。



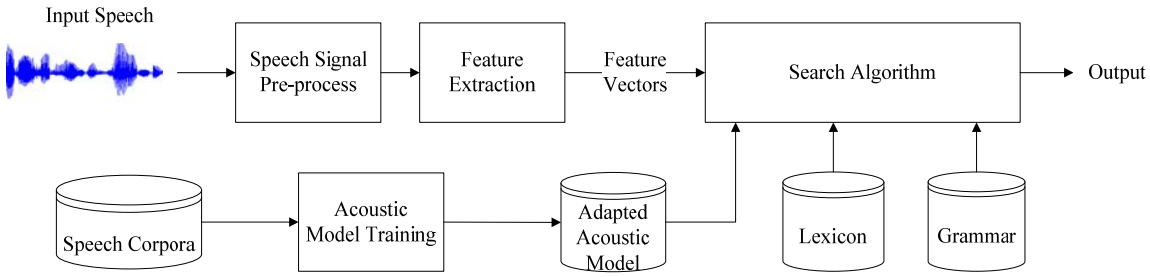
圖三、語音活動偵測法流程圖

三、系統架構

在上一章節中我們詳細敘述本論文所提出的遠距離雜訊語音辨識系統，所採用的各種研究方法，當收音系統接收到含有背景雜訊的語音時，首先會經由盲訊號分離所使用的獨立成分分析法將帶有雜訊的混合訊號進行分離，分離出兩個獨立訊號，再從這兩個獨立訊號中選取語音成分較多的獨立訊號，使用子空間語音增強法進一步濾除訊號中雜訊成分，最後再利用語音活動偵測法進行端點偵測，最後再使用 HTK 語音套件進行辨識，並判斷其辨識結果是否正確，下圖爲整體遠距離雜訊語音辨識系統整體架構流程圖以及 HTK 語音辨識器辨識流程圖。



圖四、遠距離雜訊語音辨識系統流程圖

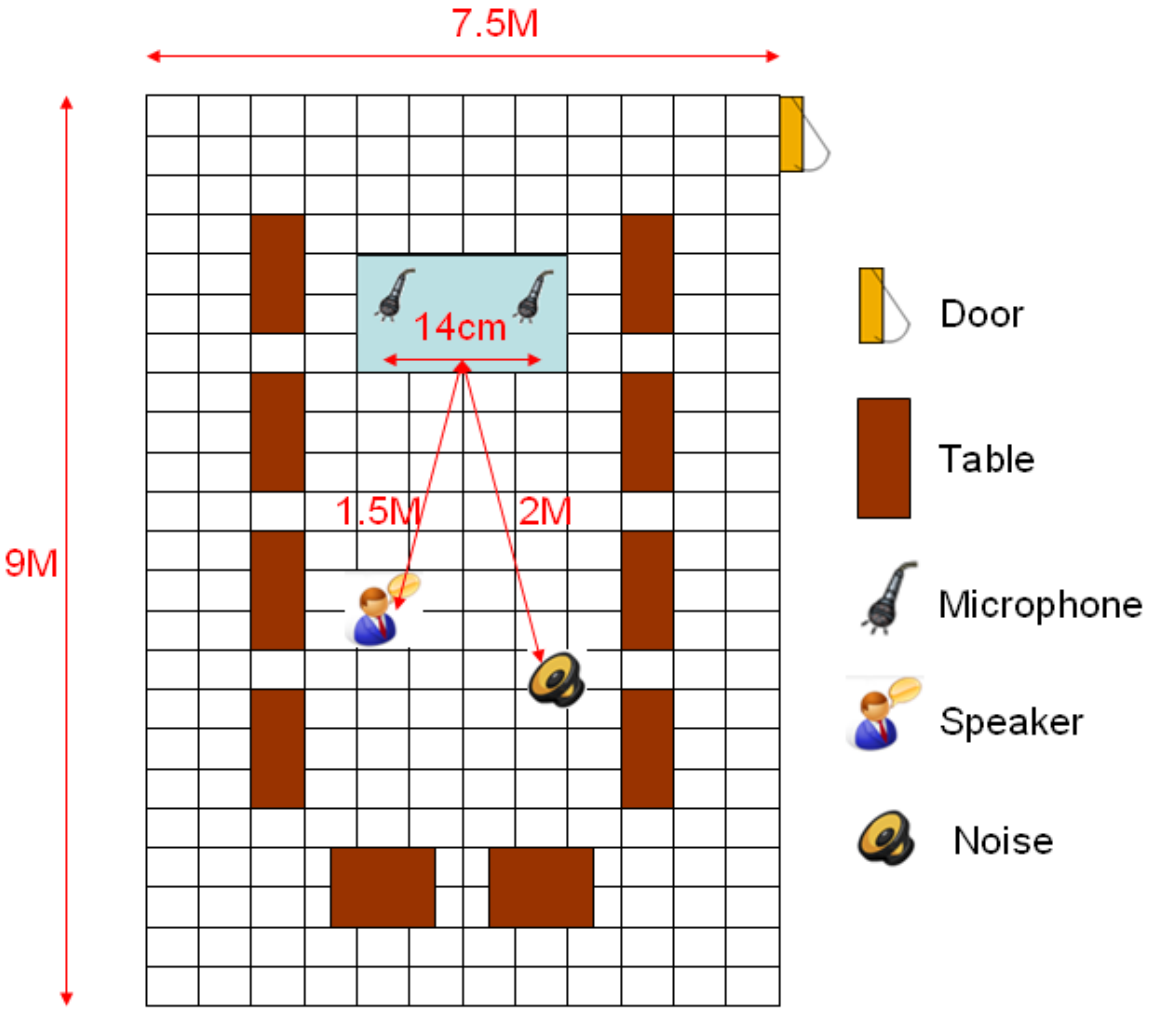


圖五、語音辨識器流程圖

四、實驗設定及辨識結果

（一）實驗環境評估和情境設定

在環境評估方面，實驗環境如下圖所示，會議室高度約爲 3 公尺，至於收音麥克風陣列，我們採用兩支麥克風進行收音，再根據不同噪聲環境及語者身分和語意內容進行辨識。



圖六、實驗環境示意圖

根據上圖實驗環境示意圖，我們設定語者距離麥克風陣列中心爲 1.5 公尺，噪聲源距離麥克風中心爲 2 公尺且高度爲 75 公分，且兩支收音麥克風間距爲 14 公分，高度爲 55 公分。在語者部份，我們採用三人進行錄音，且每人各說 10 句三字詞進行錄製，在此我們以人名做爲字詞來源；在噪音部份，我們採用 noise-92 所提供的噪聲資料庫作爲噪聲來源，在實驗中我們使用不同噪音段的 babble noise 和 car noise 當作噪聲種類。

語音和噪音混合後的訊號部份，我們根據 SNR(Signal-to-noise ratio)，分別產生各種噪聲情境下三種不同 SNR 值的雜訊訊號，分別是 0dB、5dB、以及 10dB，SNR 公式如(24)式所示，其中 P_{signal} 和 P_{noise} 分別指訊號和雜訊的平均能量，A_{signal} 和 A_{noise} 則是指訊號和雜訊振福大小，最後再將各種情境下的混合雜訊訊號進行語音分離和語音增強，最後再

進行辨識流程。

$$\mathrm{SNR(dB)} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \tag{24}$$

（二）實驗模擬辨識結果

最後在實驗辨識結果方面，我們以語音分離和增強後的訊號平均 SNR 值，以及與原始乾淨語音比較的 Segment SNR 值，還有辨識率(Recognition rate)當作我們辨識結果的主要依據，其中 Segment SNR 公式如(25)式所表示，其中 d(i)和 y(i)分別爲原始乾淨語音訊號和增益後語音訊號。

$$\mathrm{SegSNR(dB)} = \frac{1}{T}\sum_{t=0}^{T-1}\left[10\log_{10}\sum_{i=0}^{N-1}\frac{d^2(i)}{(d(i)-y(i))^2}\right] \tag{25}$$

在下列的實驗表格中，表一爲三種不同噪音段的 babble noise 和 car noise 與語音所混合而成的雜訊語音，再依據三種不同 SNR 值(0dB、5dB、10dB)情況下進行混合，最後再將此混合後的雜訊語音進行噪音分離及語音增強，並計算其增益後訊號的平均 SNR 值和 Segment SNR 值。在表一中我們可清楚看見，增益後的語音訊號，無論是在平均 SNR 值或是 Segment SNR 值，均比原始平均值提升不少，整體提升均超過 20dB。

表一、各種噪音情境下增益後訊號 SNR 值與 SegSNR 值

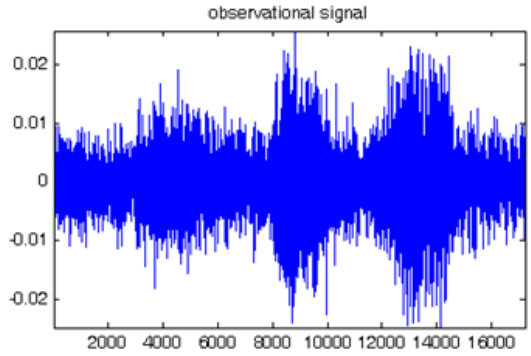
噪音情境	原始 SNR 值	增益後 SNR 值	增益後 SegSNR 值
Babble noise 1 (0dB、5dB、10dB)	5 dB	17.99 dB	30.27 dB
Babble noise 2 (0dB、5dB、10dB)	5 dB	21.81 dB	30.08 dB
Babble noise 3 (0dB、5dB、10dB)	5 dB	22.39 dB	31.21 dB
Car noise 1 (0dB、5dB、10dB)	5 dB	28.25 dB	31.79 dB
Car noise 2 (0dB、5dB、10dB)	5 dB	30.76 dB	32.50 dB
Car noise 3 (0dB、5dB、10dB)	5 dB	31.31 dB	33.17 dB

表二則是在無背景噪音及各種噪音情境下，原始雜訊語音訊號的辨識率與增益後的辨識率比較表，我們在表二中可看見，增益過後的語音訊號在辨識率上有一定的提升程度，與原始雜訊語音辨識率相比較，最高可提升 30%辨識率，整體而言約可提昇 22.96%辨識率。

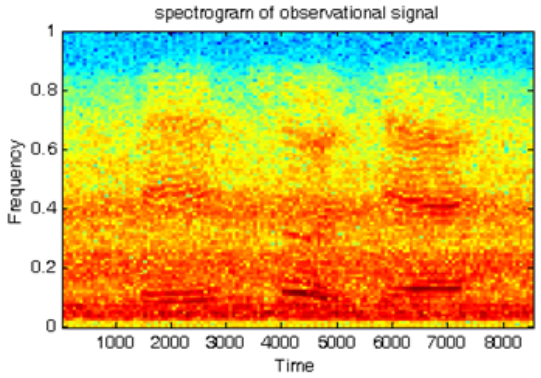
表二、各種噪音情境下增益後訊號辨識率

噪音情境	原始辨識率	增益後辨識率
無噪音	58.89 %	72.22 %
Babble noise 1 (0dB、5dB、10dB)	16.67 %	46.67 %
Babble noise 2 (0dB、5dB、10dB)	24.44 %	53.33 %
Babble noise 3 (0dB、5dB、10dB)	25.56 %	66.67 %
Car noise 1 (0dB、5dB、10dB)	61.11 %	67.78 %
Car noise 2 (0dB、5dB、10dB)	48.89 %	66.67 %
Car noise 3 (0dB、5dB、10dB)	58.89 %	72.22 %

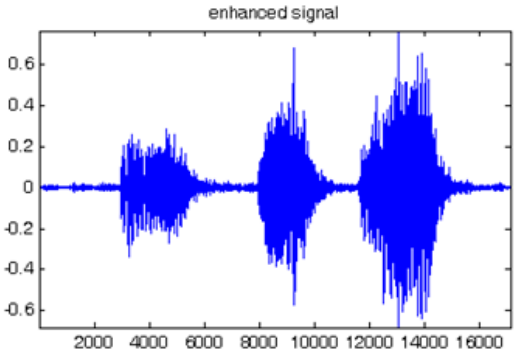
下圖爲原始雜訊語音訊號的波形圖和訊號聲譜圖(Spectrogram)、以及增益後語音訊號的波形圖和聲譜圖，在兩者訊號的波形圖比較中，可明顯看見增益後的語音訊號，在雜訊抑制上有顯著的提升；在聲譜圖比較中，除了可發現到非語音段訊號能量分布已降低不少，並且語音部份的訊號能量亦提昇許多，證實本論文所提出來的遠距離雜訊語音辨識系統，具有良好濾除噪聲雜訊和加強語音成分等功能，且能有效的提升其辨識率。



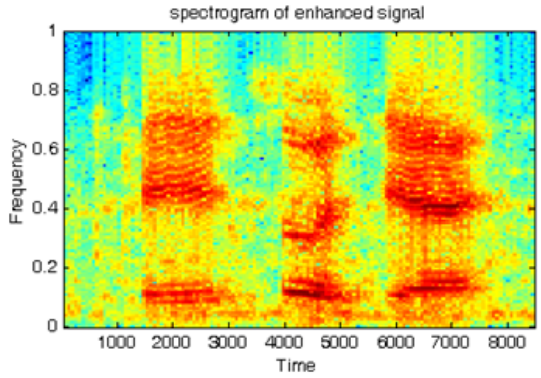
原始雜訊語音訊號波形圖



原始雜訊語音訊號聲譜圖



增益後語音訊號波形圖



增益後語音訊號聲譜圖

圖七、原始雜訊語音訊號和增益後語音訊號之波形圖和聲譜圖

五、結論

本論文所提出來的遠距離雜訊語音辨識系統，主要藉由盲訊號分離方式以及語音增強技術將雜訊語音分離出單一獨立語音訊號，再透過語音增強進一步濾除語音訊號中殘留噪音來提升辨識率，在實驗結果顯示，本論文所提出來之辨識系統，可明顯的有效提升語音能量以及辨識率，未來我們將模擬更多不同人聲及噪音情境、探討更多不同研究方法，發展出一套更具高音質解析且高辨識率之遠距離語音辨識系統。

參考文獻

[1] A. Hyvärinen., “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis,” *IEEE Transactions on Neural Networks* , Vol.10, No.3, pp.626-634, 1999.

[2] B.N. Gover, J.G. Ryan, and M.R. Stinson, “Microphone array measurement system for analysis of directional and spatial variations of sound fields,” *J. Acoust. Soc. Am.*, 112,

1980–1991 (2002).

[3] B.N. Gover, J.G. Ryan, and M.R. Stinson, “Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array,” *J. Acoust. Soc. Am. (in press).*

[4] Leukimmiatis, S., Dimitriadis, D., and Maragos, P, “An optimum microphone array post-filter for speech applications,” *ICSLP, 2006,* pp. 2142–2145.

[5] Yan Li, P. Wen and D. Powers, “Methods for the blind signal separation problem,” *in Proc. IEEE Int. Conf. Neural Network, Signal Processing,* Nanjing China, Dec. 2003, pp. 1386-1389.

[6] J. Herault and C. lutten, “Space or time adaptive signal processing by neural network models,” *In J. S. Denkcr (ed), editor, Neural Nehvorks For Computing: AIP Conference Proceedings 151,* American Institute for Physics, New York, 1986.

[7] G. Burel, “Blind separation of sources ~ a nonlinear neural algorithm,” *Neural Nehvorkr,* Vol. 5, No, 6, pp. 937-947, 1992.

[8] A. J. Bell and T. J. Sejnowski, “An information-maximisation approach to blind separation and blind deconvolution,” *Neural Computation,* Vol. 7, No. 6, 1004-1034, 1995.

[9] P . Smaragdis, *Information theoretic Approaches to source separation,* Master’s Thesis, MIT, Cambridge, MA, 1997.

[10] I. Lin, D. Grier, and J. Cowan, “Faithful representation of separable distributions,” *Neural Computation,* Vol. 9, pp. 1305-1320,1997.

[11] F. Tordini and F. Piazza, “A semi-blind approach to the separation of real world speech mixtures,” *in IJCNN'02*, Vol. 2, 2002, pp. 1293–1298.

[12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* Wiley, 1st edition, 2001.

[13] Roger L.berger, George Casella , *Statistical Inference*. 2nd edition , DUXBURY 2002.

[14] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[15] Aapo Hyvärinen, “New approcimations of differential entropy for independent component analysis and projection pursuit,” *Advance Neural Inform. Processing Syst.* 10. MIT Press, pp.273-279, 1998.

[16] W. Hu, and Z. Liu, “Partially blind source separation of continuous chaotic signals from linear mixture,” *The Institution of Engineering and Technology 2008*, Vol. 2, No4, pp. 424-430.

[17] Wang B.Y., and Zheng W.X., “Blind extraction of chaotic signal from an instantaneous linear mixture,” *Circuits Syst II, 2006*, 53, (2), pp.143-147.

[18] Paolo A., Arturo B., Luigi F., and Mattia F., “Separaion and synchronization of piecewise linear chaotic systems,” *Phys. Rev. E, 2006*, 74, p. 026212-1-026212-11.

Hierarchical Web Document Classification Based on Hierarchically Trained Domain Specific Words

Jing-Shin Chang
Department of Computer Science & Information Engineering
National Chi-Nan University
1, University Road, Puli, Nantou, Taiwan, ROC.

jshin@csie.ncnu.edu.tw

Abstract

Search engines return thousands of pages per query. Many of them are relevant to the "query words" but not interesting to the "users" due to different domain-specific meanings of the query terms. Re-classification of the returned documents based on domain specific meanings of the query terms would therefore be most effective. A *cross domain entropy (CDE)* measure is proposed to extract characteristic domain specific words (DSW's) for each node of existing hierarchical web document trees. Domain specific class models are built based on the respective DSW's. Such class models are then used for directly classifying new documents into the hierarchy, instead of using hierarchical clustering techniques. High accuracy can be achieved with very few domain specific words. With only the top 5~10% DSW's and a *maximum entropy* based classifier, 99% accuracy is observed when classifying documents of a news web site into 63 domains. The precision and recall of the extracted domain specific words are also higher than those extracted with conventional TF-IDF term weighting method.

Keywords: Domain Specific Words, Hierarchical Classification, Maximum Entropy Classifier, Cross-Domain Entropy.

1 Re-classification Issues

Search engines today return thousands of pages per query. Many of them are relevant to the "query words" but not always interesting to the "users". The major problem is that search engines do not distinguish query terms with multiple word senses. For instance, given the query term "Jaguar", most search engines are unable to identify whether it refers to an animal, a car, or an air gunship (strike fighter plane). Given a user name, most search engines cannot distinguish the person as a teacher, a technician or a government officer either. Re-classification of the web documents, so that users can quickly identify the interested documents, is therefore highly desirable.

For easy access, hierarchical classification of documents into a well-justified document hierarchy would be the most appropriate [2, 3, 4, 5, 6]. By "well-justified", we mean a hierarchy that was created or customized by human web masters, instead of an artificial hierarchy created with some automatic clustering approach. An effective classifier and a set of discriminative features for word sense disambiguation (WSD) are the key components for such purposes.

Many methods have been proposed to extract useful "features" for building classification models [13]. For hierarchical classification into an existing document hierarchy, the most effective approach would be to extract discriminative domain specific words (DSW's) for each node of the hierarchy, and build a classification model for each node directly based on such words.

A method for learning domain specific words from a web hierarchy, building associated domain-specific class model, and then classifying the documents directly into such a hierarchical document tree is therefore a key issue for re-classification.

2 Domain Specific Words as Discriminative Feature Words for Document Classification

For effective word sense disambiguation (WSD) in the reclassification process, characteristic domain specific words (DSW's) associated with each node of the document hierarchy will play an important role. The existence of some DSW's in a document will be strong evidence for the document to be a specific class and for the embedded words to be of specific senses. For instance, the existence of the domain specific word "basketball" in a document will strongly suggest the "sport" class of the particular document, even though it might have been accessed by a query term like "Pistons" (which normally has a machinery sense). With the DSW "basketball", the special usage of Piston as a basketball team, rather than its "machinery" sense, are likely to be recognized. Actually, both of them act as domain specific words of the sport domain, and enhance each other for supporting the "sport" class when both appear in the same document. DSW identification is therefore an important issue for document re-classification.

However, manually acquiring the associated domain specific words for each node in the hierarchy is most likely unaffordable in terms of time and cost. Therefore, learning domain specific words automatically from existing web hierarchy is the key step for web document re-classification. In addition, new words (or new usages of words) are dynamically produced day by day. For instance, the Chinese word "活塞" (piston) is more frequently used as the "sport" or "basketball" sense (referring to the Detroit Pistons) in Chinese web pages rather than the "mechanics" or "automobile" sense. It is therefore desirable to find an automatic and inexpensive way to acquire the whole set of domain specific words of the hierarchy directly from web corpora.

Actually, the directory hierarchy of a Web can be regarded as a kind of classification tree for web documents. Each node of the hierarchy is associated with a special class label, identified by the directory name, and a large number of documents with some kind of domain specific words. For instance, the documents under the "sport" hierarchy are likely to use a large number of domain-specific words for the "sport" class. Hence, the domain specific words for each node can be *extracted by removing "non-specific terms"* from the associated document sets provided a measure of "domain specificity" is well established. A cross-domain entropy (CDE) measure will be proposed for this purpose.

With these DSW's, associated with each node, a class model for each such node can be established for direct classification of web documents into the document hierarchy.

3 Classification by Clustering vs. Direct Classification into Web Hierarchy

As indicated, for easy access, hierarchical classification of the documents into a well-justified document hierarchy is important. Automatic construction of the document hierarchy by document clustering, however, might create a hierarchy that is not acceptable by users.

The conventional clustering approach to classify web documents into a hierarchical structure is to collect "important terms" in all web documents as their representatives and measure the distance between document pairs using some well-known distance metrics between documents. For instance, the vector space model [10] measures the cosine of the angle between two document vectors, consisting of weights for important terms, as a similarity measure. Documents with high similarity or short distances are then clustered bottom-up to build a hierarchy. The clustering hierarchy is then manually inspected for adjustment into a customized hierarchy if necessary.

There are several disadvantages with this approach. First of all, since the documents are clustered based on distance or similarity measures that do not have a direct link with any ontological criteria, the hierarchical relationship among the clustered web documents is not guaranteed to fit any naturally created hierarchy by most web masters. In particular, most clustering algorithms merge documents in a binary way, which is far from the way a human user would do. The mismatch of such hierarchical structures implies that the clustered one might not match human perception quite well. The clustering results may therefore not be acceptable by the users. Furthermore, due to such mismatch, the clustered hierarchy may not be adjusted comfortably by the web masters. As far as the computation cost is concerned, computation of document distances based on *pairwise* distance metrics will be time consuming. (Admittedly, clustering might be preferred in some circumstances [4].)

Fortunately, clustering is not the only option since a large number of web documents, which are natively arranged in a *hierarchical* manner, are created every day. One can readily learn to classify documents directly into a customized hierarchy by learning the domain specific words of each node from the training documents associated with each node, and creating a class model for each node.

4 Building Domain Specific Classifier Models with Domain Specific Words Detected from Web Hierarchy

Since the web documents already form an extremely huge document classification tree, we propose here a simple and automatic approach to acquire the domain specific words in the hierarchical web documents. The domain specific words for each node can then be used to build their respective classifier models for each node of the document tree.

This simple approach is inspired by the fact that most text materials (webpages) in websites are already classified in a hierarchical manner; the hierarchical directory structures implicitly suggest that the domain specific terms in the text materials of a particular subdirectory are closely related to a common subject, which is identified by the name of the subdirectory. If we can detect domain specific words within each document, by removing

words that are non-specific to the subdirectory, then the resultant DSW's would be good representative for building classifier models for the subdirectory where they reside.

For instance, a subdirectory entitled 'entertainment' is likely to have a large number of web pages containing domain specific terms like 'singer', 'pop songs', 'rock & roll', 'album', and so on. Since these words are good representatives of the 'entertainment' domain, they can be used to build a class model for the 'entertainment' domain. A new document accessed by the query term 'album', as well as other domain specific terms, can then be classified into the 'entertainment' class easily, instead of into the 'photo album' class or something else.

Since a large number of documents had been classified into various web hierarchies, and update of the hierarchies is a daily routine of the various webmasters, the training corpora is not sparse. As such, we will pay almost no cost in training the classifier models for various domains and customized document trees.

The idea might extend equally well to other hierarchically organized Internet resources, such as news groups and BBS articles. Extending the idea to hierarchically organized book chapters might also be possible.

The advantages of building classification models directly from the sets of DSW's associated with each node, by removing non-specific terms from documents, are many folds. First, the original hierarchical structure reflects human perception on which directory a document should be classified into. Therefore, one rarely needs to adjust the hierarchy; in the worse case, one may be more comfortable to adjust the hierarchy if necessary. Second, the computation cost is significantly reduced, since pairwise computation of document distance is now replaced by the computation of "domain specificity" of documents against domains. The reduction is significant, from $O(\sum|d|x\sum|d|)$ to $O(\sum|d|x|D|)$, where $|d|$ and $|D|$ represent the number of documents in individual domains and number of domains, respectively.

5 Domain Specific Word Acquisition: A Cross-Domain Entropy Approach

Since the terms in the documents include general terms as well as domain-specific terms, the only problem then is an effective model to exclude those domain-independent terms from the documents. The degree of domain independency can be measured with the cross-domain entropy (CDE) as will be defined in the following DSW (Domain-Specific Word) Extraction Algorithm. Intuitively, a term that distributes evenly in all domains is likely to be independent of any domain and therefore is unlikely to be a DSW. The CDE provides a way to better estimate domain independency than traditional inverse document frequency (IDF). The method is first revealed in [7] and is summarized as follows.

First of all, a large collection of web documents is acquired using a web spider while preserving the directory hierarchy. Since our target is Chinese documents, a word segmentation algorithm [12, 9] is applied in order to identify terms in the documents.

For each subdirectory d_j, we find the number of occurrences n_{ij} of each term w_i in all the documents, and derive the normalized term frequency $f_{ij} = n_{ij} / N_j$ by normalizing n_{ij} with the total document size, $N_j \equiv \sum_i n_{ij}$, in that directory. The directory is then associated with a set of $< w_i, d_j, f_{ij} >$ tuples, where w_i is the i-th words of the complete word list for all documents, d_j is the j-th directory name (refer to as the domain hereafter), and $f_{ij} = n_{ij} / N_j$ is the normalized relative frequency of occurrence of w_i in domain d_j.

Domain-independency of the terms are then estimated with the following **Cross-Domain Entropy** (CDE) measure [7]:

$$H_i^* \equiv H^*(w_i) \equiv -\sum_j P_{ij} \log P_{ij}$$

$$P_{ij} \equiv \frac{f_{ij}}{\sum_j f_{ij}}$$

Terms whose CDE is above a threshold is unlikely to be specific since such terms are evenly distributed in many domains.

To appreciate the fact that a high frequency term will be more important in a domain, the CDE is further weighted by the term frequency in the particular domain when deciding which terms are important DSW's. Currently, the weighting method mimics the conventional TF-IDF method [10] in information retrieval. In brief, a word with entropy *Hi* can be think of as a term that spreads in *2**Hi* domains on average. The equivalent number of domains a term could be found then can be equated to $Nd_i = 2^{H_i}$. The term weight for w_i in the *j-th* domain can then be estimated as:

$$W_{ij} = n_{ij} \times \log_2 \left(\frac{N}{Nd_i} \right)$$

where N is the total number of domains. Unlike the conventional TF-IDF method, however, the *expected number of domains* that a term could be found is estimated by considering its probabilistic distribution across all domains, instead of simple counting.

The directory tree, after domain independent terms are removed from the associated documents, now represents a hierarchical classification of the domain-specific terms of different domains. The lists of domain specific words in the subdomains can thus be used for building domain-specific class models for disambiguating ambiguous words in various contexts.

The Appendix shows a list of highly associated domain-specific words of low cross-domain entropies and high term weights (with literal English translation) in 4 special domains [7].

6 DSW-Based Document Classification Model

Given the DSW's for various domains, some discriminative classification models have to be developed in order to make the best use of such information source. This is particularly true for DSW's that are sense ambiguous since sense ambiguity will make a DSW less useful as a representative of a domain unless quantitative measures related to the domain, such as the posterior probabilities of the term under various domains, are known.

For instance, if we use the DSW's simply as the (sense-insensitive) index terms in conventional VSM-based (Vector Space Model-based) search engines, and use them to calculate the similarity between documents for clustering without distinguishing the distinct senses of the same term, then we may classify a document into a class that has "relevant documents" including the same index term, such as the keyword "bank", but not the "interested documents" relevant to "the organization where money is saved or withdrawn".

This ineffective use of the index terms (or DSW's) results from the fact that a VSM does not take domain specificity (such as "the probability of the index term in a domain") into consideration when calculating the similarity between the query keywords and the document. Such ineffective use is also a main reason why people are calling for document re-classification or web mining at the backend of those quick (but dirty) search engines.

Inspired partly by such an observation, a Bayesian classifier, which incorporates the domain specificity measure as a posterior probability into the model parameters, is adopted.

With a Bayesian classifier, the document classification task can be formulated as finding the most probable domain label of a document, given the set of words w_1^n in a document:

$$\begin{aligned} d^* &= \arg\max_d P\left(d \mid w_1^n\right) \\ &= \arg\max_d P\left(w_1^n \mid d\right) P\left(d\right) \end{aligned}$$

d : class label of document

(i.e., the "domain")

w_1^n: words in the document

If we assume that a document is representable by the domain specific words in the document, we can simply restrict w_1^n to the set of DSW's of all domains. If we further assume that each domain specific word is generated independent of others, then we can simply have:

$$d^* = \arg\max_d \prod_{w_i \in DSW} P\left(w_i \mid d\right) P\left(d\right)$$

where $P\left(d\right)$ is the prior probability that the domain d is addressed, which can be estimated by the number of documents in d normalized by the number of all document. Furthermore, $P\left(w_i \mid d\right)$ represents the posterior probability that a DSW will appear in the domain d. Intuitively, the more a DSW in a document matches the right domain, the higher score one will have for that domain. It is this probability factor that provides the domain specificity

information. With this factor included, one can retrieve "interested documents" better, since an interested document relevant to "money" will now be retrieved not merely because it has the ambiguous keyword "bank" but other associated DSW's.

The above Naïve Bayesian assumption, however, has the problem on how to smooth the empirically obtained probabilities if some domain specific words do not appear in some domains. To make this smoothing issue well resolved, a maximum entropy-based classifier [1, 11, 8], is, instead, adopted so that word uni-gram counts in each domain can be used as a feature to estimate the posterior probability of the document, subject to the constrains that the expected feature values will fit the empirical counts, and all other unseen features are smoothed equally probable.

Given such a classifier, the text version of the web documents can be filtered with domain specific words so that only domain specific words in the training documents are used for training the class models (i.e., the posterior probabilities for each domain). The testing documents, filtered with domain specific words, are then classified based on the class models.

7 Classification Performance with Domain Specific Words

To see how the DSW-based text classifier achieves high classification accuracy with very few DSW's, a large collection of Chinese web pages was collected from a local news site. The HTML web pages are about 200M bytes pre-classified into 138 proprietary hierarchical domains of the news site (including the most specific domains at the leaf nodes as well as their parent domains towards the root). About 16,000 unique words are identified after word segmentation is applied to the text parts.

Totally, there are 4,279 documents in the collection. On average, each domain has about 30 documents. For simplicity, a held out estimate of the classifier performance, based solely on the DSW's, is adopted. (An n-fold cross validation would be better though.) The training set for the classifier consists of about 9/10 of the documents in each domain; the other 1/10 of the documents is used as the test set. Since some domains have only a few documents, they are excluded from performance evaluation to factor out biases introduced by insufficient evaluation data. The final set of documents for training and testing consists of 3,322 documents in 63 domains. Each domain has at least 23 documents for training.

To train the classifier models, the training documents are first word segmented with a word-unigram based word segmentation model [12, 9]. Those words that are not identified as domain specific words are then filtered out from the documents, leaving a bag of domain-specific words, like the one shown in Table 1, for model training. Each row in this table is derived from a sentence (or phrase) of the un-filtered document.

The set of domain specific words depends on the threshold applied to the list of word types in each domain, sorted by decreasing term weight Wij. The simplified document representation as shown in Table 1 is filtered by the top 5% words that have the highest term weights (hereafter, "the top 5% domain specific words" for short) of each domain. Although the full text is not shown, it is not difficult for a person to classify it into the "US-stock" domain if one is given this domain as a candidate. This characterization by DSW's is exactly the basis for document classification without resorting to the full text.

理財　股市　美國股市　企業　裁員　元月
個股　查詢
理財　股市　美國股市
企業　裁員　元月
紐約　五日
公司
二月　大幅
公司　經濟
裁減　的工作
二月　裁減　百分之
公司
消費　者的
企業　更多
趨勢
公司
裁員　第三
裁員
福特　汽車　裁員　汽車　業的　裁員
單月
汽車　裁員　產業
美國　勞工　市場
企業　裁員
日報

Table 1. A Document about "US-stock" filtered with the top 5% DSW's.
(Each row is a filtered sentence.)

The evaluation is based on the bags of words filtered with the top 20%, 10%, 5% and 2% of the domain specific words, each sampling step uses approximately half of the entries of the next higher sampling threshold.

Furthermore, since the TF-IDF (term frequency inverse document frequency) approach [13, 10] is widely used in information retrieval applications for indexing important terms within documents, it can also be applied to identify domain specific words in various domains. To make a comparison, the TF-IDF term weighting method is also applied to the same corpus to see the differences.

The training set performance and the test set performance are depicted in Figures 1 and 2, respectively, showing the percentages of accurately classified documents by using different sets of DSW's of different sampling thresholds. Three term weighting strategies are shown in the three curves, where the "CDE" model refers to the proposed term weighting method based on the cross-domain entropy measure, "IDF" refers to the traditional TF-IDF approach, and "RAW" refers to the case where all the words in the unfiltered documents are used for

training the classifier. The details of the two figures are also shown in Tables 2 and 3 respectively.

Note that, the RAW model is equivalent to using the constant threshold of 100% for extracting domain specific words from the word list sorted by decreasing term weights. The performance is therefore a constant, shown as a horizontal line, in comparison with other models with different thresholds.

It is obvious from the curves that the performances of the other two models will eventually drop to the horizontal line if one tries to increase the number of DSW's to achieve "better" performance. This is not surprising since the model will then be contaminated by other non-specific words. And this is exactly what the domain specific words are valuable. *The performance curves at the low sampling thresholds are therefore the major criteria for comparing various models* [13].

In this area with low sampling thresholds, the CDE model is consistently better than the traditional TF-IDF approach, even though the differences are small. And, it breaks the performance of the RAW model much faster.

One must not be over-optimistic to the test set performance shown here, though, since the number of documents under testing is only 301 documents. With 5% of most heavily weighted terms, there are only 3 mis-classified documents. Hence it quickly reaches the highest performance when the domain specific words are doubled to 10%. Due to the small number of test documents, the performance figures may have a great variance. A more conservative performance might be acquired if the performance is evaluated against a larger data set. However, as a conservative estimate, it would be safe to say that with 5~10% most heavily weighted terms, a 99% accuracy is possible for this task.

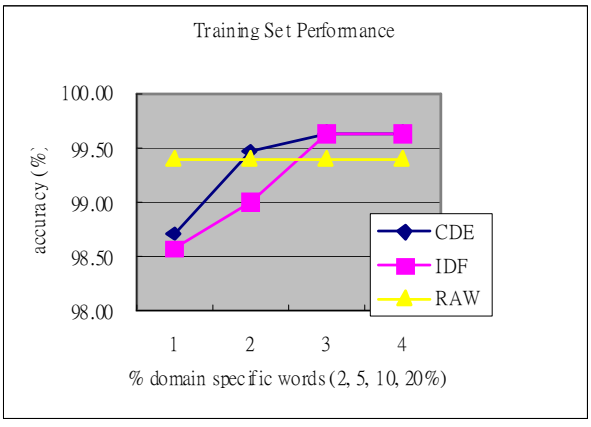


Figure 1. Training Set Performance

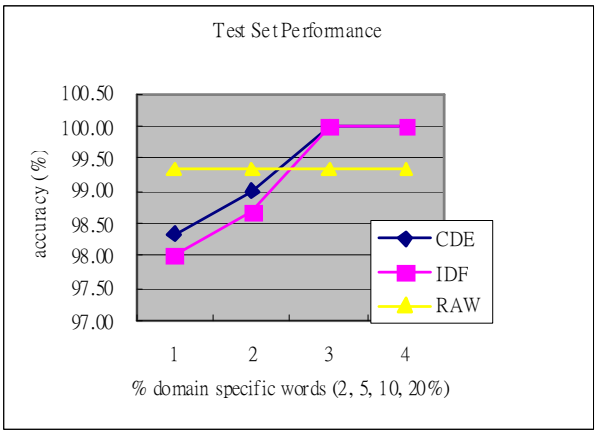


Figure 2. Test Set Performance

%DSW	CDE	IDF	RAW
2	98.71	98.58	99.40
5	99.47	99.01	99.40
10	99.64	99.64	99.40
20	99.64	99.64	99.40

Table 2. Training Set Performances

%DSW	CDE	IDF	RAW
2	98.34	98.01	99.34
5	99.00	98.67	99.34
10	100.00	100.00	99.34
20	100.00	100.00	99.34

Table 3. Test Set Performances

8 Domain Specific Words Acquisition Results

Although the DSW acquisition performance is not the main focus of this paper, it is interesting to summarize some of the results in [7].

In 5 sample domains of small sizes (with 300 word types or less), the current approach extracts domain specific words at an average precision of 65.4% and an average recall of 36.3%, corresponding to 45.8% F-measure, if the top-10% words with highest term weights are extracted as domain-specific words. In other words, by only gathering the first 10% of the whole word list, the current term weighting measure can identify about 36% of the embedded domain specific words, and one can identify significant amount of DSW's about every 1.5 entries from the top-10% list of low entropy words.

In comparison with the popular TF-IDF (term frequency inverse document frequency) term weighting approach, it is observed that the top-20% sampling threshold for the baseball domain results in 51.6% F-measure with the CDE-based approach, as opposed to 47.8 % with the TF-IDF weighting method.

From all the observations, the CDE measure does indicate the "degree of specificity" more informatively. In particular, the degree of domain specificity of a term is estimated by considering the cross-domain *probability distribution* of the term in the current CDE-based approach. In contrast, the TF-IDF approach simply counts the number of domains a term is found as a measure of randomness. The CDE approach is therefore gaining a little bit performance than a TF-IDF model.

The results partially confirm that one can extract a large domain specific lexicon at little cost from the web pages by removing non-specific words from web documents.

9 Concluding Remarks

In this paper, a method for learning domain specific class models for nodes of a hierarchical web document tree from domain specific words associated with each node is proposed for web document classification. With this approach, web documents can be easily re-classified into an existing web hierarchy directly, instead of being re-clustered with other documents to form a hierarchy that is unlikely to fit any human classification criteria.

With only the 5~10% of most heavily weighted DSW's and a maximum entropy based classifier, classification accuracy of about 99% is observed when the method is evaluated on a task that classifies web documents into the 63 domains of an existing news web site. Furthermore, the performance of the current term weighting method is consistently better than the conventional TF-IDF approach in the current task, in terms of classification performance and domain specific word acquisition performance. The current approach is therefore an appropriate candidate for applications of this kind.

Appendix: Sample Domain Specific Words

Baseball	Broadcast-TV	Basketball	Car
日本職棒 (Japanese professional baseball)	有線電視 (cable TV)	一分 (one minute)	千西西 (Kilo-c.c.)
棒球賽 (baseball games)	東風 (Dong Fong TV Station)	三秒 (three seconds)	小型車 (small car)
熱身 (warm up)	開工 (start to work)	女子組 (girl's teams)	中古 (used car)
運動 (athlete)	節目中 (on air)	包夾 (fold; clip)	引擎蓋 (engine cover)
場次 (time table)	廣電處 (radio-tv office)	外線	水箱 (tank)
價碼 (cost)	收視	犯規 (foul)	加裝
球團 (baseball team)	和信 (Ho-Hsin TV Station)	投籃 (shot)	市場買氣 (market atmosphere)
部長 (manager)	新聞局 (government information office)	男子組 (male team)	目的地 (destination)
練球 (practicing)	開獎	防守 (defense)	交車 (car delivery)
興農 (Hsin-Lung team)	頻道 (channel)	冠軍戰 (championship)	同級 (of the same grade)
球場(course; diamond)	電視 (TV)	後衛 (fullback)	合作開發 (co-development)
投手 (pitcher)	電影(movie)	活塞 (Piston team)	安全系統 (safety system)
球季 (season)	熱門 (hot)	國男 (national male team)	行李 (luggage)
賽程 (schedule)	影視 (video)	華勒(Wallace)	行李廂 (trunk)
太陽 (the Sun team)	娛樂 (entertainment)	費城 (Philadelphia)	西西 (c.c.)

Appendix. Sample domain specific words with low cross-domain entropies in 4 special domains [7].

References

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing." *Computational Linguistics*. 22(1):39-71, 1996.

[2] Dmitry Davidov, Evgeniy Gabrilovich, Shaul Markovitch, Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. *Proc. of SIGIR 2004*, pp. 250-257, 2004.

[3] Dumais & Chen, "Hierarchical Classification of Web Content", SIGIR-2000, 2000.

[4] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, "Scatter/ Gather: a cluster-based approach to browsing large document collections." In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318--329, June, 1992.

[5] Glover, Eric J., Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, and Gary W. Flake, "Using web structure for classifying and describing web pages." In *Proc. 11th WWW Conference*, pages 562—569, 2002.

[6] Huang, C.-C., S.-L. Chuang, L.-F. Chien, "LiveClassifier: Creating Hierarchical Text Classifiers through Web Corpora", pp. 184-192, WWW 2004, 2004.

[7] Jing-Shin Chang, "Domain Specific Word Extraction from Hierarchical Web Documents: A First Step Toward Building Lexicon Trees from Web Corpora," *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing,* pp. 64-71, International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Korea, 2005.

[8] Laird Breyer, "The DBACL Text Classifier," http://www.lbreyer.com/gpl.html, 2004.

[9] Ming-Yu Lin, Tung-Hui Chiang and Keh-Yih Su, "A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," *Proceedings of ROCLING VI*, pp. 119-142, 1993.

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval,* Addison Wesley, New York. 1999.

[11] Ronald Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech, and Language,* 10(3):187-228, 1996.

[12] Tung-Hui Chiang, Jing-Shin Chang, Ming-Yu Lin and Keh-Yih Su, "Statistical Models for Word Segmentation and Unknown Word Resolution," *Proceedings of ROCLING-V*, pp. 123-146, Taipei, Taiwan, R.O.C. 1992.

[13] Yang, Yiming and Jan O. Pedersen, "A comparative study on feature selection in text categorization." In *Proceedings of the International Conference on Machine Learning*, pages 412—420, 1997.

中文混淆字集應用於別字偵錯模板自動產生

Chinese Confusion Word Set for Automatic Generation of Spelling Error Detecting Template

陳勇志 Yong-Zhi Chen, 吳世弘 Shih-Hung Wu
朝陽科技大學資訊工程系
Department of Computer Science and Information Engineering
Chaoyang University of Technology
{9727602, shwu}@cyut.edu.tw

盧家慶 Chia-Ching Lu, 谷圳 Tsun Ku
資訊工業策進會
Institute for information industry
{gaty, cujing}@iii.org.tw

摘要

本研究透過常用字來產生混淆字集，自動產生能夠幫助錯別字偵測的模板，發展華語文錯別字偵測技術。本系統利用辭典爲基礎，使用辭典中的詞彙做爲正面用詞，透過混淆字集自動產生含別字的反面模板，能夠偵測的別字包含同音字、同部首字，並且透過斷詞軟體輔助擷取更正確的反面模板，用以協助華文教師進行大量華文作文的錯別字批改甚至輔助學生進行寫作，最後達到提昇寫作能力之成效。

關鍵詞：模板產生、模板探勘、正反面用語知識庫

Abstract

In this research, we proposed a system that can use automatically generated templates for detecting Chinese spelling error. At first, we use frequently used Chinese characters to produce the Chinese confusion set. Based on a dictionary, our system automatically generated negative vocabulary template with the help of Chinese confusion set. Error types include pronunciation-related errors and radical-related errors. And our system uses word segment to capture more accurately the negative template. We hope that such a system can help the teachers on the checking of students' essays, and also can help students learn to write effectively. Consequently, the students would improve their writing skill.

Keywords: Template generation, Template mining, Pragmatics Knowledge Base.

一、緒論

自民國 95 年起，教育部在國中基本學力測驗中加辦「寫作測驗」隨後列入升學計分，計分標準依據立意取材、結構組織、遣詞造句、錯別字給予 6 個等第的級分，華語學習中的作文能力備受重視。國中基本學力測驗每年約有三十萬學生應試，因此我們可以預見未來將有大量的作文輔助批改與輔助教學的需求，如何應用數位學習的技術來輔助教師批改作文並且幫助學生學習寫作，爲目前普遍研究之議題。

根據寫作測驗的評分依據，錯別字是個重要的評分標準，回顧以往中文錯別字的輔助學生系統的相關文獻有[1]與[2]，這兩篇文獻都是針對中文文章中進行偵錯與訂正的系統，其中[1]是利用替換五筆字型編碼來產生可能的別字，透過每次替換一個編碼即可達到產生多個可能的別字，而五筆字型輸入法主要用於使用簡體中文的中國大陸，五筆字型完全依據筆畫和字形特徵對漢字進行編碼，將漢字筆劃分爲橫 、豎、撇、捺、折五種，把字根或編碼按一定規律分佈在 25 個英文按鍵上。教育部也針對錯別字推出由人工編寫的常用國字辨似[3]，但是常用國字辨似只含有 1477 筆模板，並不敷大量的作文偵錯使用，而我們從書上蒐集常用的正反面用語模板含有 6,701 筆，並且在 2008 年發表中文作文的訂正與更正建議系統[4]，該系統利用學生所書寫的作文蒐集偵錯用模板藉以建立模板偵錯技術的正反面模板偵錯，並且透過統計 Corpus 的 uni-gram、bi-gram 建立語言模型做爲常用語偵錯，由於人工蒐集模板費時耗力且成本過高，所以隨後我們根據 QA 系統中的自動模板產生概念[5]、[6]，利用機器學習之技術並且大量探勘 Corpus 中可能的反面用語模板，最後利用學生所書寫得作文做爲 Training data，於 2009 年 5 月發表了中文作文錯別字偵錯模板自動產生[7]，該系統使用模板擴展演算法來取得大量的模板，並且透過卡方檢定做爲收納模板的檢定公式，但是自動模板產生還是依賴人工蒐集模板中之種子，於是我們使用混淆字集來大量產生蒐集模板用的種子，混淆字集爲一般人容易混用的字之集合，混淆字可能爲同音字、同形字、同部首字等等。

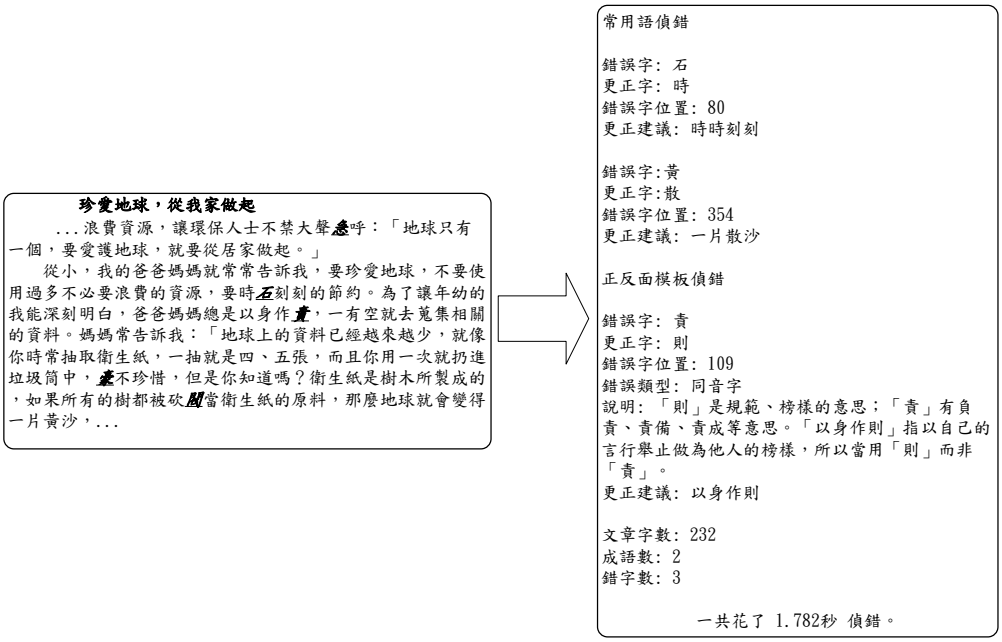
混淆字集是學生容易將正確的字書寫成錯誤的字之集合，根據劉昭麟教授的統計學生作文[8]、[9]，學生書寫的錯別字中同形字佔 30.70%、同音字佔 79.88%、同形同音字佔 20.91%、非同形同音字佔 2.43%，統計結果指出學生書寫的的錯別字大部分來自於同音字錯誤，同音字的混淆字集可以透過字典收集取得而同形字則較不易取得，不過劉教授依據倉頡輸入法發表[10]，利用替換倉頡輸入法字碼來取得同形字與[1]的替換五筆字型編碼相似之處。

二、系統設計與方法

（一）錯別字自動模板產生系統回顧

我們於 2008 於所發表的錯別字偵錯與訂正建議系統[4]是透過一個 Web 介面，讓學生輸入他們所書寫的作文而文章可能含有若干個別字如圖一左邊方塊，經過我們系統兩項功能常用語偵錯與正反面模板偵錯診斷後，常用語偵錯會提供學生的錯誤字、必須更正的字、錯誤字位置和更正建議的資訊，正反面用語偵錯除了常用語偵錯提供的建議之外還會提供詞語的說明如圖一右邊方塊，我們的系統可以偵測出常見的錯別字，並且明確指出學生錯別字在文章的何處，並且給予適當的建議與說明，讓學生瞭解自己何處寫錯字並且從錯誤中學習。

我們在 2009 年 5 月所發表的自動模板產生系統[7]是改進 2008 年發表[4]的系統，2008年的系統所使用的模板必須經由人工蒐集，由人工蒐集模板費時耗力且成本過高，自動模板產生系統確實能夠自動產生大量的偵錯模板不過卻有兩個缺點，1. 產生模板的正、別字種子必須經由人工蒐集。2. 其自動產生的部份模板不具可讀性且不符合詞彙的概念，如圖二。圖二中我們可以看出某些詞彙如“辯護律”、“視辯論”、“電視辯”等並不是完整的詞彙，如“辯護律”可能是從“辯護律師”擷取，這些不完整的詞彙並不適合當作更正建議資訊給使用者參考，因此下面我們根據以上兩個缺點進行改進。



圖一、2008 年所發表系統之偵錯功能

26749	會首長	會首常	7116	清潔隊長	清潔隊常
26750	會給予	會給于	7117	交通隊長	交通隊常
26751	辯論會	辨論會	7118	辯護律師	辨護律師
26752	辯護律	辨護律	7119	視辯論會	視辨論會
26753	的辯論	的辨論	7120	政策辯論	政策辨論
26754	視辯論	視辨論	7121	電視辯論	電視辨論
26755	電視辯	電視辨	7122	公開辯論	公開辨論
26756	半世紀	辦世紀	7123	半個世紀	辦個世紀
26757	半以上	辦以上	7124	一年半的	一年辦的
26758	半個小	辦個小	7125	的另一半	的另一辦

圖二、舊系統所產生的部份模板

（二）混淆字集

我們發表過的偵錯系統所使用的正、別字種子是經由我們所蒐集正、反面用語中擷取其中的正、別字所產生，例如：正反面用語爲“芭蕉”、“笆蕉”，而正、別字種子則爲“芭”、“笆”，由於正、別字種子也是必須經由人工蒐集，同樣也具有費時耗力

成本過高的缺點。根據劉教授的統計[8]、[9]，同音同形字的別字在學生所寫的錯別字中佔有 89.67%，其中同音錯別字高達 79.88%。

我們從字典蒐集所有用字的注音並且將同音同調視爲同音字，如“動”的注音爲“ㄉㄨㄥˋ”，因此“凍”“ㄉㄨㄥˋ”視爲同音字，“東”“ㄉㄨㄥ”視爲不同音字，而我們所蒐集的同音字表共有 1,351 音、15,160 字，如圖三。

漢字字形的構成要素是由筆劃、筆順、偏旁、六書、部首所構成，其中部首是東漢文字學家許慎所著之《說文解字》所創，此後漢字的檢字方式一般皆使用部首，而同部首字含有高相似度，因此我們使用部首資訊來產生同形字，根據《康熙字典》漢字的部首一共有 214 個，我們利用 214 個部首蒐集了 9,752 個漢字並且產生同部首字表，如圖四。

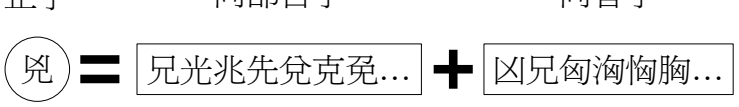
最後自動產生混淆字集方法則是使用正字檢索同音字表、同部首字表，如圖五概念圖利用“兇”即可找出同部首的“兄光兆先兌克兔…”與同音字的“凶兄匈洶恟胸…”等字。

1	ㄅㄚ	蚾扒八巴仢叭朳芭疤捌笆粑豝鈀吧
2	ㄅㄚˇ	伯靶霸猼把爸壩灞朳耙
3	ㄅㄚˊ	鈸茇拔朳跋菝詙軷魃鼥犮
4	ㄅㄚˇ	鈀把靶
5	ㄅㄚˋ	吧杷琶罷
6	ㄅㄛ	剝嚗波袚玻柀砵缽皤菠蟠撥嶓蹳鱍岥播潖
7	ㄅㄛˇ	播檗蘗亳擘譒北挀薜簸鑿
8	ㄅㄛˊ	爆伯犮潖挀蔔柏爬薄泊暴瀑鈹帛勃胉捋淳郭脖舶椤
9	ㄅㄛˇ	簸踣蚾
10	ㄅㄛˋ	蔔

圖三、部份同音字表

1	一	一丐丁七三下丈上万丌丑丐不丐丙世丕且丘丞丟並
2	丶	丸凡丹主
3	丿	乂乃久么之尹乍乏乎乒乓乘乖乘
4	乙	乙九乜也乞乩乩乳乾亂
5	亅	了予事
6	二	二于云井互五亓亙些亞亟
7	亠	亡亢交亦亥亨享京亭亮亳亶亹
8	人	人仁什仃仆仇仍今介仄仂仉以付仔仕他仗代令仙仞
9	儿	兀元允充兄光兕兆先兌克兔兕兔兒兗党兜兢
10	入	入內全兩

圖四、部份同部首字表

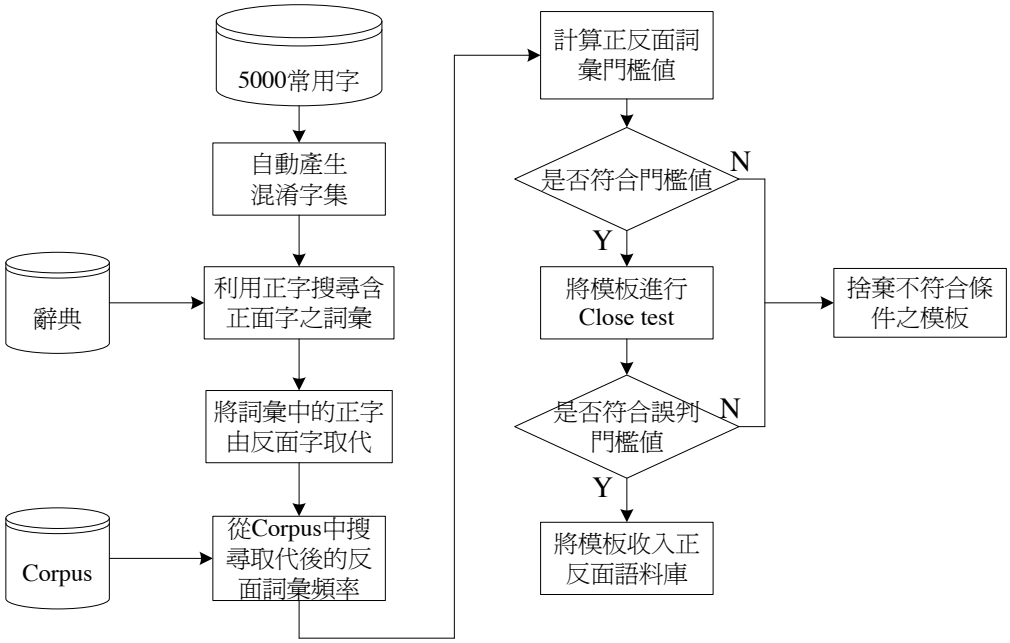


圖五、自動產生混淆字概念圖

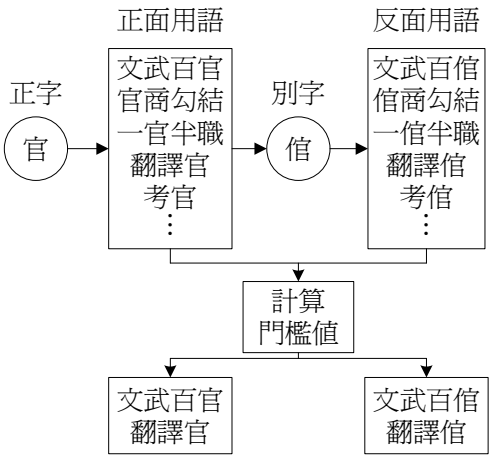
（三）自動化收集模板系統流程

圖六爲我們自動模板產生系統流程圖，我們的自動模板產生系統是基於正面用詞非常頻繁使用，而含別字的反面用語則會被使用很少的條件下，首先我們蒐集由國語推行委員會所公佈的八十七年常用語詞調查報告書[11]中的常用字共 4,998 字作爲正字種子，接著利用這些常用字自動產生同音、同部首的別字，最後將混淆字集輸入至我們的自動模板產生系統。

我們發表過的偵錯系統是使用演算法來產生正面用語模板但會有稍早所提之缺點，於是我們使用現有詞彙的基礎作爲正面用語模板，當系統取得正字後會去檢索辭典是否有包含該正字之詞彙，其中辭典爲教育部所公佈之教育部重編國語辭典修訂本[12]，經由我們濾掉單字詞彙共 145,608 詞，接著我們將檢索之詞彙的正字替換成別字做爲正、反面用語模板，接著到 Corpus 進行頻率統計，統計頻率如果符合門檻值的模板則收集起來進行 Close test，如果符合 Close test 誤判門檻值的模板最後則將此模板收入正反面語料庫，而自動模板產生的概念如圖七。



圖六、系統流程圖

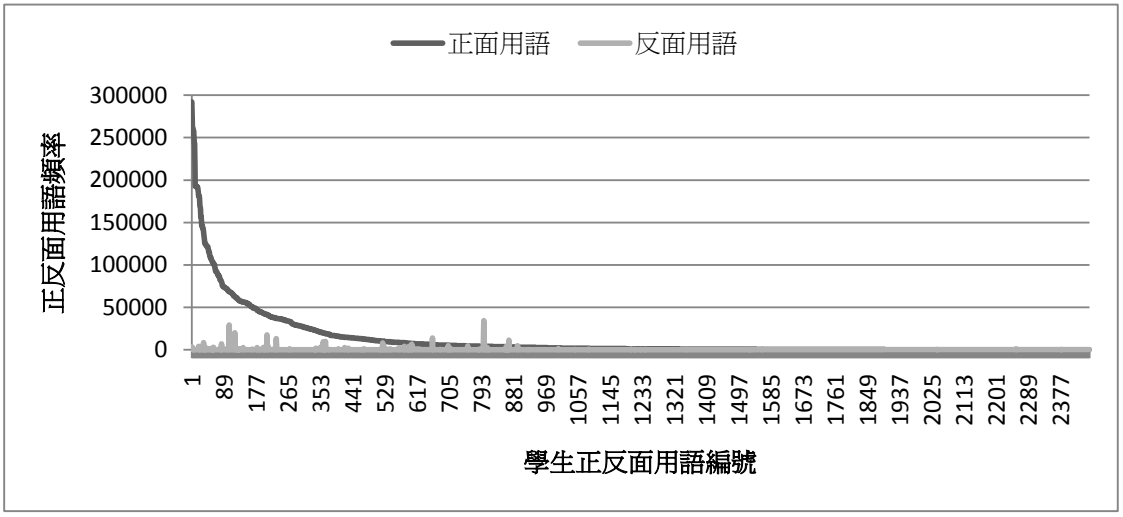


圖七、自動模板產生概念圖

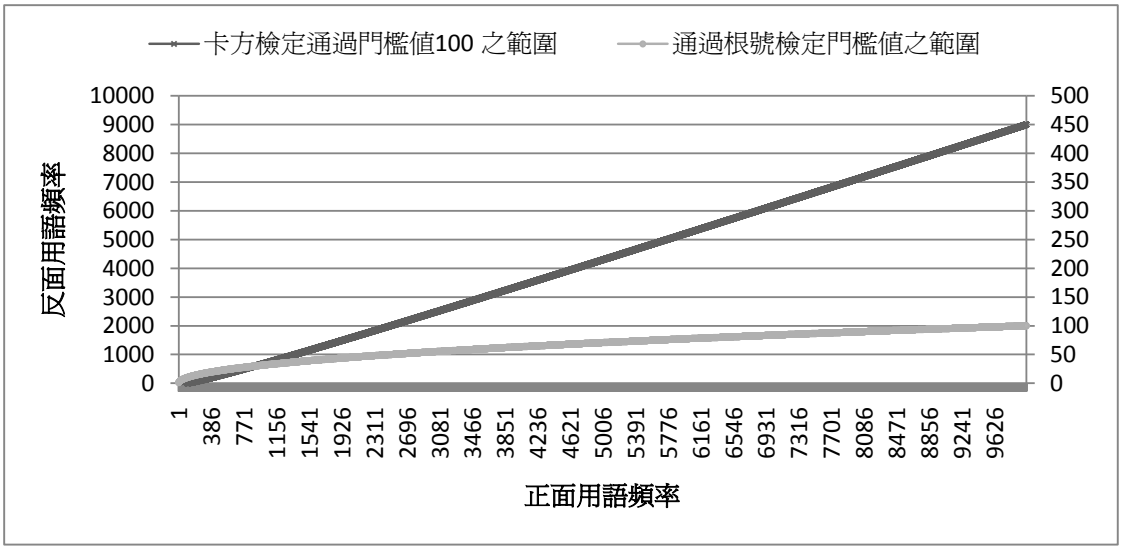
檢定公式方面，我們在 2009 年 5 月所發表的自動模板產生系統[7]是使用卡方檢定來檢定是否收納模板如(1)，其中 E 爲正面用語模板的出現頻率 O 爲反面用語模板的出現頻率，而中文中常有積非成是的用語或通用詞詞彙，爲了避免這樣的情況我們會限定 E > O。

$$X^2 = \frac{(O-E)^2}{E} \quad (1)$$

隨後我們觀察學生作文中學生所使用的反面用語與教師訂正後的正面用語，發現正面用語的頻率遠大於反面用語如圖八，而卡方檢定公式特性卻只學生正反面用語的頻率分佈不同其卡方檢定特性圖如圖九卡方檢定的門檻值範圍，在門檻值設定爲 100 的條件下，隨著正面用語頻率的提昇而反面用語也呈線性的提昇，也就是圖九上方線段以內的反面頻率皆會通過卡方檢定測試，這與我們從學生所使用的正反面用語有著非常大的差異，所以卡方檢定並不適合用來檢定模板是否收納。



圖八、正反面用語頻率分佈圖



圖九、卡方檢定與根號檢定門檻值分佈圖

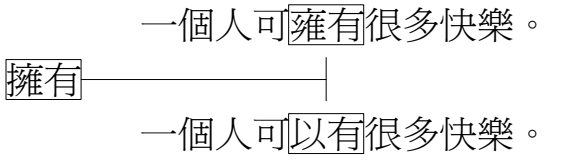
因此我們將否採納該模板的公式修改如(2) (3)，*Cfeq* 爲正面用語的頻率、*Wfeq* 爲反面用語的頻、*Threshold* 爲所有正面用語頻率之平均，而採納模板的條件是正面用語的頻率經過開根號的計算之後必須大於反面用語的頻率，且正面用語必須大於門檻值。使用此公式是依據圖八學生使用的正反面用語之特性所設計，依照每個正面用語得頻率取得相對之反面用語頻率，並且必須符合正面用詞非常頻繁使用，反面用語則被使用很少的條件，根號檢定的特性圖如圖九下方線段，根號檢定能夠針對每一個正面用語頻率去取得他的最佳反面用語頻率之門檻值，最後我們將學生正反面用語頻率共 2455 筆利用(2)公式做頻率分佈分析，其中共有 90.46%的模板符合根號檢定之測試，不符合此檢定測試的模板有“未來”、“爲來”，“已經”、“以經”，“但是”、“但事”等模板，這些模板的前後文資訊不足如果用來當作錯別字訂正模板，則會非常容易引入雜訊。

$$\sqrt{Cfreq} > Wfreq \quad , \quad Cfreq > Threshold \qquad (2)$$

$$Threshold = \frac{\sum_{i=1}^{n} Cvocabulary(i)}{n} \qquad (3)$$

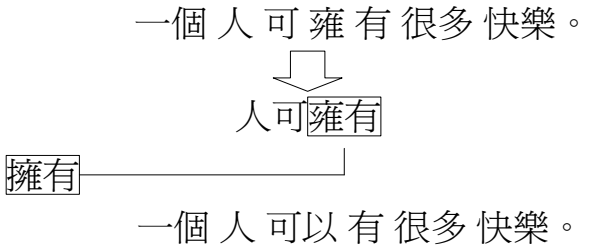
（四）斷詞軟體應用

上述檢定公式與 Close test 處理根據我們實驗與觀察，2 個字的詞彙仍然非常容易造成 False alarm，這個原因是 2 字的詞彙過短容易與其他詞彙發生重疊的現象範例如圖十，如果用一個正面詞彙如“擁有”去擷取反面用語，則會如圖十般將“雍有”、“以有”都收入，但在“一個人可以有很多快樂”這句範例中，“以有”中的“以”字應屬於“可以”這個詞彙。



圖十、詞彙重疊現象範例

斷詞軟體能夠將正確的詞彙斷詞，而含有別字的詞彙則無法正確斷出詞彙如圖十一，因此我們使用這個特性將 Corpus 利用斷詞軟體[13]斷詞，藉以用來擷取更正確的 2 字詞彙模板，我們會將正確斷詞的詞彙移除接著將剩餘單字詞合併用來擷取模板用。最後由我們系統自動產生的模板如圖十二。



圖十一、應用斷詞擷取反面用語模板範例

395	衝擊	衝急	437	絆腳石	伴腳石	879	逼不得已	逼不得已
396	檢視	機視	438	大部分	大不分	880	情非得已	情非得已
397	經濟	經紀	439	手電筒	手電桶	881	逼不得已	逼不得已
398	循環	循還	440	不經意	不經易	882	大勢已去	大勢以去
399	成績	成積	441	不願意	不願易	883	不能自已	不能自以
400	薪水	新水	442	董事長	懂事長	884	迫不得已	迫不得以
401	賺錢	購錢	443	三輪車	三軸車	885	情非得已	情非得以
402	關鍵	關建	444	腦震盪	腦振盪	886	萬不得已	萬不得以
403	老闆	老版	445	辦公室	辨公室	887	逼不得已	逼不得以
404	雖然	隨然	446	成績單	成積單	888	巡弋飛彈	巡曳飛彈

圖十二、經由我們系統所產生的部份模板

三、實驗結果與分析

（一）Corpus 與學生作文

由於統計模板頻率需要大量的語料資料，因此我們蒐集新聞語料庫做爲我們的 Corpus，資料整理如表一。

表一、Corpus 資料整理

資料年份	新聞社	文件數	檔案大小
1998-1999	Chinatimes	38,163	209MB
	Chinatimes Commercial	25,812	
	Chinatimes Express	5,747	
	Central Daily News	27,770	
	China Daily News	34,728	
1998-1999	United Daily News	249,508	320MB
2000-2001	United Daily News	172,421	1.03GB
	United Express	91,958	
	Ming Hseng News	168,807	
	Economic Daily News	463,873	

測試集是從學生作文中拆成兩個部份，一個部份做爲 Close test 用，另一部份則是用來 Open test 用，學生作文我們使用台北市某國中七、八年級考試作文、並且由教師校訂過錯別字共 3264 篇，每篇文章皆輸入成電腦可處理的格式如圖十三，而我們的系統並不處理注音文以及不存在於 Unicode 編碼中之錯字。

最後我們將蒐集到的作文做資料分析如表二，從表格中我們可以看出約 94%的作文用字皆爲常用字的範圍，而表三則爲學生作文的正別字分析同部首的正別字約在 15%同音正別字約 68%，而非同部首同音字約爲 19%，我們的作文統計分析結果與劉教授的分析結果[8]、[9]相近。另外我們也統計學生常犯錯誤之 Top 10 模板如表四。

```
<doc>
<class>七年一班</class>
<number>7</number>
<title>藉口</title>
<score>4.5</score>
<essay>
<p>人，有許多夢想，尼采說：「人因夢想而偉大。」雖然是這麼說，不過光「想」是不會有<revise><wr
<p>你是否曾找過一些<revise><wrong>冠冕唐荒</wrong><correct>冠冕堂皇</correct></revise>的藉口
<p>人非聖賢，誰能無過？知過能改，善莫大焉，摒除藉口，是一個需要決心、毅力、耐心的工程，我常常
<p>燕子去了有再來的時候，<revise><wrong>楊柳估了</wrong><correct>楊柳枯了</correct></revise>
</essay>
</doc>
```

圖十三、作文電子檔的格式

表二、學生作文基本分析

	作文數	平均級分	作文平均字數	平均別字數	常用字比例
Close test essay	2241	3.62	367.12	1.74	94.23%
Open test essay	1023	3.61	420.02	1.94	94.33%

表三、學生作文正別字分析

	正別字同部首比例	正別字同音比例	兩者皆有	兩者皆非
Close test essay	13.82%	70.27%	4.92%	20.81%
Open test essay	16.96%	66.31%	2.85%	19.58%

表四、常犯錯誤之 Top 10 模板

Close essay	正面用語	已經	變得	自己	景象	一旦	寄託	已經	畢竟	而已	根本
	反面用語	己經	變的	自已	景像	一但	寄托	以經	必竟	而己	跟本
Open essay	正面用語	自己	一旦	已經	選擇	煩惱	應該	已經	而已	選擇	後悔
	反面用語	自已	一但	己經	選則	煩腦	因該	以經	而己	撰擇	後侮

（二）實驗設計與評估

我們實驗的比較對象爲[4]所人工蒐集的模板與發表過的偵錯系統[7]所產生的模板，由於二字詞、三字詞、四字以上的詞彙出現頻率差異非常大，因爲我們針對這三組分別計算全部詞彙的平均頻率，依照平均頻率門檻值分別設定爲：2300、500、100，而 Close test 的過濾門檻值經由我們反覆實驗得到 0 爲最佳的設定。

評估的方式是使用 Precision 與 Recall 公式定義如下:

$$\text{Micro Recall} = \frac{\sum\left(\frac{dr}{r}\right)}{N} \qquad (4)$$

$$\text{Micro Precision} = \frac{\sum\left(\frac{dr}{sd}\right)}{N} \qquad (5)$$

$$\text{Macro Recall} = \frac{\sum(dr)}{\sum(r)} \qquad (6)$$

$$\text{Macro Precision} = \frac{\sum(dr)}{\sum(sd)} \qquad (7)$$

$$\text{False alarm rate} = 1 - \text{Precision} \qquad (8)$$

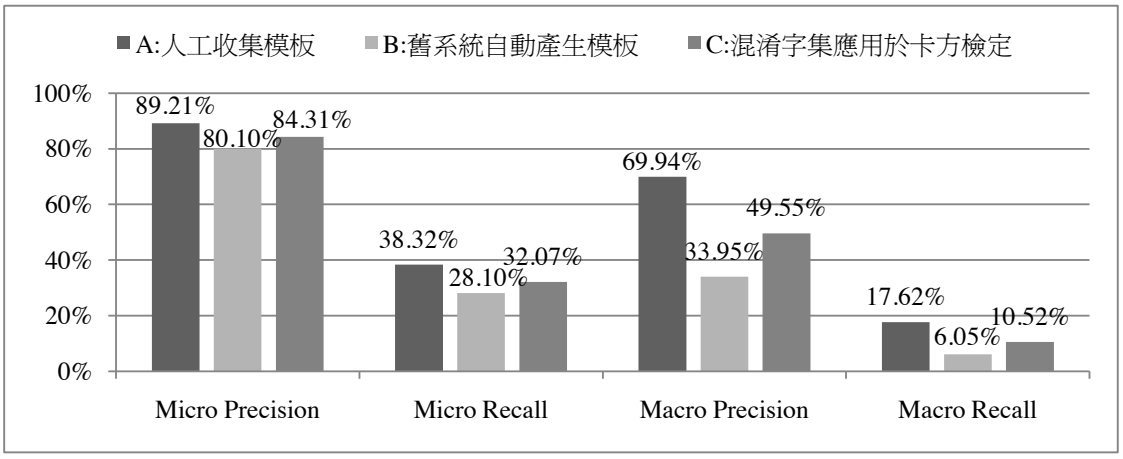
dr 爲每篇文章中偵錯正確的字數，r 爲每篇文章中真正的錯字數，sd 爲 每篇文章中系統偵測出的錯字數，N 爲所有文章的篇數。Micro Precision 與 Micro Recall 是以接近現實生活的偵錯情形，也就是以文章爲單位偵錯效能如何。除此之外還必須考量較

多的樣本，也就是將所有的資料視爲整個大集合，所以我們使用 Macro Recall 與 Macro Precision 來檢視系統的效能。最後我們系統要求的是在維持高 Precision 的情況下來提高 Recall 值，因爲我們不希望給使用者太多 False alarm。

（三）實驗結果

我們設計四組實驗第一組實驗爲混淆字集應用於卡方檢定時的實驗結果，第二組爲混淆字集應用於根號檢定的實驗結果，第三組爲根號檢定加入斷詞後的實驗結果，第四組爲根號檢定所自動產生的模板加入人工蒐集模板之後的實驗結果。

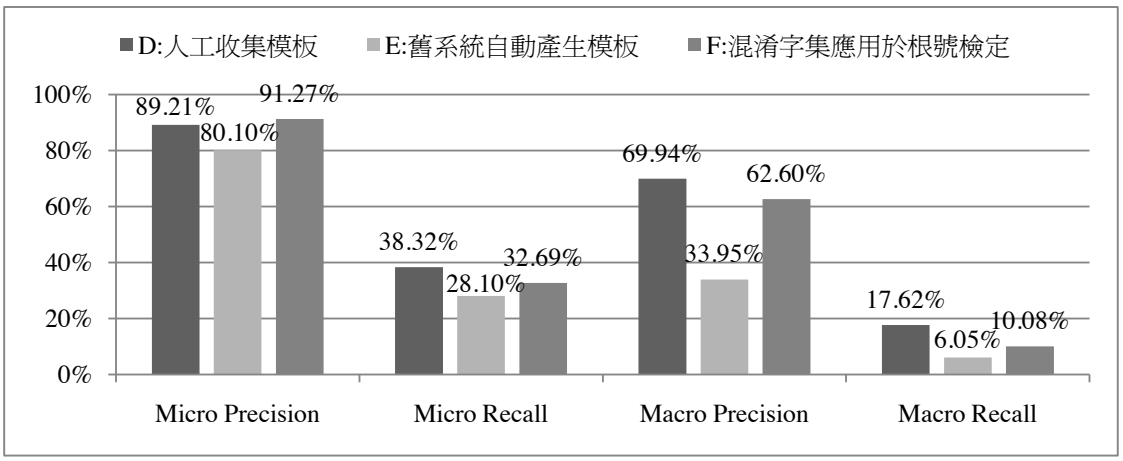
實驗一：混淆字集應用於卡方檢定



圖十四、卡方檢定實驗結果

實驗結果如圖十四，其中 A 組爲人工蒐集的模板共 6,701 筆，B 組爲我們發表過系統所產生的模板共 19,402 筆，C 組爲應用混淆字集後使用卡方檢定之系統產生的新模板共 54,253 筆，其中混淆字採取[11]中的常用字，不在常用範圍的字則不在此範圍。Precision 方面以人工蒐集的模板爲最佳，而 Recall 方面應用混淆字集卡方檢定所自動產生的模板優於過去我們所發表的系統，整體來說則還是以人工蒐集的模板爲最佳，不過自動產生的模板在 Recall 皆都逼近人工蒐集模板的數值。

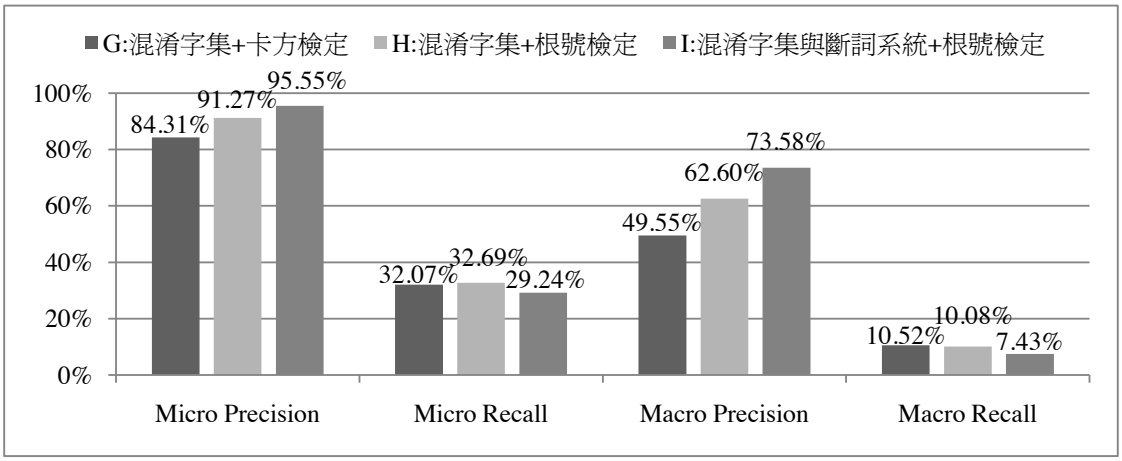
實驗二：混淆字集應用於根號檢定



圖十五、根號檢定實驗結果

實驗結果如圖十五，D、E 組與實驗一的 A、B 相同，F 組爲應用混淆字集後使用根號檢定之後系統產生的新模板共 50,467 筆。與實驗一最大的不同處是根號檢定在 Precision 方面皆比卡方檢定來得優異許多其中以 Macro 提昇最多，在 Recall 方面 Micro 些微提昇 Macro 則是些微下降。由此實驗可以得知過去卡方檢定的檢定模板方式讓許多 noise 進入造成 Precision 的下降，改用根號檢定的檢定方式可以改善以往的缺點。

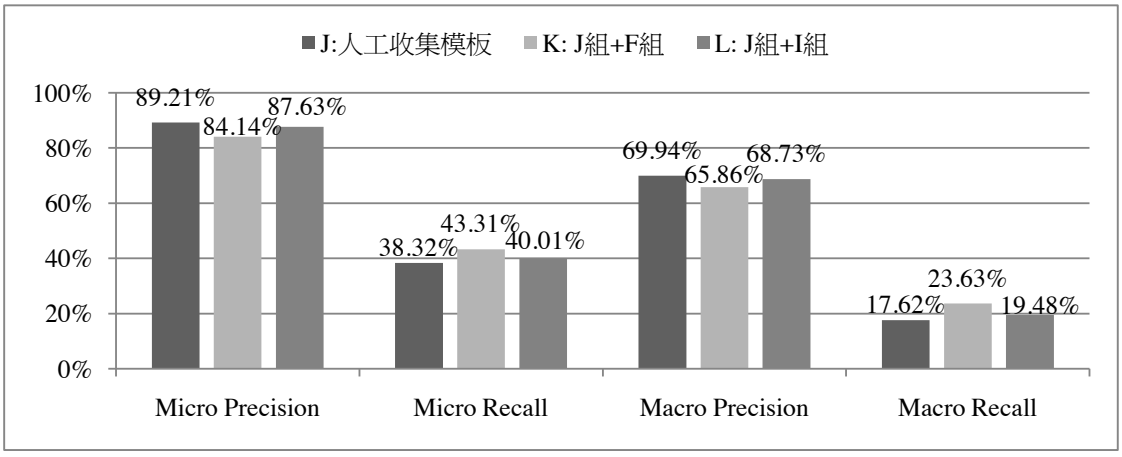
實驗三：加入斷詞系統



圖十六、應用斷詞之檢定比較

實驗結果如圖十六，G 組爲實驗一 C 組之卡方檢定模板，H 組爲實驗二 F 組之根號檢定模板，I 組爲 H 組應用斷詞軟體後自動產生的新模板共 9,013 筆。由於斷詞軟體的使用讓斷詞更準確，因此模板產生數跟前面兩個實驗相比較降低不少，在 Precision 方面可以發現不論 Micro 或 Macro 都比使用斷詞軟體前的模板在準確度有更進一步的提昇，但是在 Recall 部份則是下降 3%左右，這是追求 Precision 所犧牲的地方。

實驗四：混合人工蒐集模板



圖十七、比較混合人工蒐集模板之效能

實驗結果如圖十七，J 組爲實驗一 A 組人工蒐集的模板共 6,701 筆，K 組爲實驗二 F 組根號檢定模板與 J 組人工蒐集的模板混合使用共 57,167 筆，L 組爲實驗三 I 組應用斷詞軟體之根號檢定模板與 J 組人工蒐集的模板混合使用共 15,713 筆。Precision 方面混合人工蒐集模板後的自動產生模板皆下降到人工蒐集模板水平附近，而跟實驗三比較

Recall 方面在 Micro 與 Macro 部份皆有大幅度的提昇，這也表示我們的系統具有可擴充性，如果加入適當的模板能夠使用系統的偵錯範圍更進一步的提昇。

以混淆字爲基礎來產生模板的新系統理當可以掌握 70~80%的錯別字，但是經由我們的實驗卻發現自動產生的模板在 Recall 值方面的提昇非常有限，這個現象我們將在下節做數據分析。

（四）實驗結果分析

用來分析的模板我們使用實驗三中的應用斷詞軟體之根號檢定自動產生的新模板共 9,013 筆做爲分析模板，因爲這組模板比較符合我們當初所預期之在維持高 Precision 的情況下來提高 Recall 值。

1 Precision 方面

利用反面模板來偵測錯誤理當能夠讓 Precision 達成 100%，但是經由我們實驗結果發現卻不是如此，我們將 Open test 中系統偵錯部份造成 False alarm 提出討論如表五。根據[12]“垃圾桶”“垃圾筒”、“奇蹟”“奇跡”、“電線桿”“電線杆”、“銷聲匿跡”“消聲匿跡”，可以得知此四組模板爲通用詞，而“一再”“一在”則牽涉到語意層面再這邊並不適合使用模板的方式來偵錯，“放聲大哭”“放聲大叫”、“不用說”“不用講”、“讀書人”“讀書做”則是我們系統收納到不適合的模板，Precision 無法達到 100%就是上述原因所導致。

表五、部份 False alarm 模板

正面用語	垃圾桶	奇蹟	電線桿	銷聲匿跡	一再	放聲大哭	不用說	讀書人
反面用語	垃圾筒	奇跡	電線杆	消聲匿跡	一在	放聲大叫	不用講	讀書做

2 Recall 方面

我們將學生所書寫正反面用語模板與我們系統所產生的模板做個分析如表六。其中“沒產生到的模板”爲我們系統所沒有產生到的模板，“不在辭典”爲在沒有產生到的模板中其正面用語不在辭典中，“不在 Corpus”爲在沒有產生到的模板中其反面用語不在 Corpus 中，“兩者皆是”爲正面用語不再辭典中同時相對應反面用語也不在 Corpus 中。

從“沒產生到的模板”數值可以發現，絕大部分學生所書寫得模板並沒有被我們自動產生，再從“不在辭典”與“不在 Corpus”數值中相加並且扣除兩者皆有的部份，可得知 Close test essay 有 53.17%的模板與 Open test essay 有 32.97%，是我們的系統無法自動產生出來，因爲我們的系統自動產生模板是基於正確詞彙與 Corpus 曾經有人使用過該反面用語。

至於不存於辭典的詞彙如表七，可以 將這些詞彙加入辭典這樣便可以克服此問題，而不存在於 Corpus 的反面用語則必須蒐集更大量的 Corpus 語料庫，以便能夠蒐集到此類的反面模板。

表六、學生模板與系統模板分析

	沒產生到的模板	不在辭典	不在 Corpus	兩者皆是
Close test essay	91.53%	37.73%	35.64%	20.20%
Open test essay	93.15%	16.27%	23.94%	7.24%

表七、部份未收入辭典之詞彙

佈告欄	蒸飯機	值日生	作業本	辦派對	睡午覺	全班齊心	勤加練習	羞恥心	無厘頭
重拾信心	莽莽撞撞	淘汰	漆彈場	偶像劇	積陰德	融入團體	芬多精	燒炭	拉筋

四、結論及未來工作

根據我們應用混淆字集、根號檢定公式與斷詞軟體，我們能夠省去人工蒐集產生模板用種子的流程，並且能夠產生以詞彙爲基礎的模板如圖十二，改進過去發表過的系統模板如圖二非詞彙基礎的模板，給予使用者更明確的訂正資訊。使用根號檢定公式也經由實驗得知確實能夠比卡方檢定所自動產生的模板有較佳的 Precision，最後藉由斷詞軟體斷詞後的 Corpus 也經由實驗證實能夠更進一步提昇系統的 Precision，而 Recall 部份也能透過持續增加適合的模板來增加偵測率。

在未來我們會蒐集混淆字集所沒辦法產生的模板產生種子，也會持續蒐集更符合學生作文的文章來取代新聞語料庫，詞彙方面則會透過大型詞彙庫或線上資源如：維基百科，來增加我們辭典的詞彙數，最後我們預計使用學生的作文產生含別字之語言模型，利用該語言模型來智慧偵錯以輔助模板偵錯所沒有蒐集到的錯誤模板。

致謝

本研究依經濟部補助財團法人資訊工業策進會「98 年度智慧型網路服務技術與應用計畫(2/4) 」辦理。

參考文獻

[1] Lei Zhang, Chang ning Huang, Ming Zhou, Haihua Pan, *Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm*, Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pp: 248-254, 2000.

[2] Ren, F. , Shi, H., Zhou, Q., *A hybrid approach to automatic Chinese text checking and error correction*, In Proceedings of the ARPA Work shop on Human Language Technology, pp: 76-81, March 1994.

[3] MOE, *Common Errors in Chinese Writings (常用國字辨似)*, Ministry of Education, Taiwan, 1996.

[4] Ta-Hung Hung., & Shih-Hung Wu, *Chinese Essay Error Detection and Suggestion System*. Taiwan E-Learning Forum, 2008.

[5] Cheng-Lung Sung., Cheng-Wei Lee., Hsu-Chun Yen., Wen-Lian Hsu, *An Alignment-based Surface Pattern for a Question Answering System*, the IEEE International Conference on Information Reuse and Integration, pages pp. 172-177, 2008.

[6] D. Ravichandran., & E. Hovy, *Learning surface text patterns for a Question Answering system*, in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 41-47, 2001.

[7] 陳勇志, 吳世弘, 盧家慶, 谷圳, *中文作文錯別字偵錯模板自動產生*, The 13th Global Chinese Conference on Computer in Education, pp. 402-408, 2009.

[8] Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, Shih-Hung Wu, *Phonological and logographic influences on errors in written Chinese words*, Proceedings of the Seventh Workshop on Asian Language Resources, the Forty Seventh Annual Meeting of the Association for Computational Linguistics, August 2009.

[9] Chao-Lin Liu, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, Shih-Hung Wu, *Capturing errors in written Chinese words*, Proceedings of the Forty Seventh Annual Meeting of the Association for Computational Linguistics, August 2009.

[10] Chao-Lin Liu and Jen-Hsiang Lin, *Using structural information for identifying similar Chinese characters*, Proceedings of the Forty Sixth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2008.

[11] 國語推行委員會, *八十七年常用語詞調查報告書*, National Languages Committee, Taiwan, 1998.

[12] MOE, *教育部重編國語辭典修訂本*, Ministry of Education, Taiwan, 2007.

[12] CKIP, "Autotag," Academia Sinica, 1999.

Consolidation of Robust Speaker and Speech Recognition for Intelligent Doorway Application

Ta-Wen Kuan, Jhing-Fa Wang, Po-Yi Shih, Ta-Wei Sun and Miao-Hai Chen
Department of Electrical Engineering, National Cheng Kung University
Email:gwam.davin@gmail.com, wangjf@mail.ncku.edu.tw

Abstract

In this paper, integration of speaker identification and speech recognition for intelligent doorway application has been proposed. Two target speakers will be identified through an one-word speech utterance. Moreover, this utterance will be recognized to be a pre-defined speech command. The speaker identification in the proposed framework is based on support vector machine (SVM). The "one-versus-one" approach is applied in this paper to classify test point input utterance according to the number of votes. As for the speech recognition, we use confusion matrix to develop an efficient phonetic set for a command-based multi-lingual system, the confusion matrix calculates acoustic similarities between every two phonemes. The proposed framework has been realized in the intelligent doorway application and will be applied to many other daily life computer speech applications.

Keywords: SVM, confusion matrix, HTK, speaker identification, speech recognition.

1. Introduction

In the real world, there are three commonly applications in speech recognition system, such as "who is speaker?", "what is content?", and "where is speaker?". The contribution of this paper is to propose a practical consolidated framework to integrate both the speaker identification and speech recognition, with the aim at satisfaction of human computer interface in recognizing "who is speaker?" and "what is content?" at same time.

Support Vector Machine has been explored and proved in speaker recognition for many years [1][2]. SVM has many desirable attributes that can classify and robust to sparse data without over-training and to make linear and non-linear decision via kernel functions [3]. However, due to complicated algorithm and time-consuming process in training SVM, thus it still not gained widespread utilization in many applications. Ubiquitous Robot Companion (URC) proposed a text-independent speaker identification using microphone-array on a robot and intends to enrich the interaction between human and robot [4]. Far-field speaker recognition proposed two approaches to improve the robustness of speaker recognition. The first is to use the conventional method based on acoustic feature. The second approach is to make use of higher-level

linguistic feature. However, the adverse environmental condition and adverse training-testing conditions still need to be considered and conquered under proposed benchmark environment [5]. Ubiquitous and robust text-Independent speaker recognition [6] proposed a new microphone-array configuration of framework for benchmark. This framework is used a mixer to received speech signal from six microphones, then the six channel speech signal are mixed and output only one signal for feature extraction.

The mixed-language speech recognition has been researched for many years [7][8][9]. In this proposed consolidated of speech recognition system, the speaker independent voice command recognition is adopted, and with a string size of tens or more words. In addition, an acoustic and phoneme modeling based on confusion matrix for ubiquitous mixed-language speech of recognition system is integrated in proposed framework [10]. This system allows users to use given command to control electrical device via speech. The system is also flexibly applied in different command-based control applications by changing the dictionary description and grammar in each new work.

The reminder of this paper is organized as follows. In Section 2, the basic theories of SVM algorithm for data classification as well as confusion Matrix of acoustic model for bilingual speech recognition are described. In Section 3, the proposed framework of consolidated speech recognition system is presented. The experimental results of proposed architecture are shown in Section 4. Finally, we draw our conclusion in Section 5.

2. Literature Review

2.1 SVM based Speaker Identification

The main concept of SVM is to use a partition hyperplane to maximize the distance between support vectors of two classes features, and then to create a classifier between two clusters of sample. The gain of the SVM-based pattern recognition method is robust to sparse training data samples [11] [12]. This optimal hyperplane is obtained by minimizing the following constrained optimization problem as shown in Eq. (1).

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C\left(\sum_{i=1}^{N} \xi_i\right)$$

subject to , (1)

$$y_i(w\phi(x_i)+b)+\xi_i-1\geq 0, \quad 1\leq i\leq n$$

$$\xi_i \geq 0, \quad 1\leq i\leq n$$

where x_i is a training sample, y_i is the corresponding target value, $w \in R_m$ is a vector of weights of training instances, b is a constant, C is a real value cost parameter, and ξ_i is a penalty parameter (slack variable).

If $\Phi(x_i) = x_i$, the SVM finds a linear separating hyperplane with the maximal margin. If Φ maps x_i into a higher dimensional space, then it is called a nonlinear SVM. For the nonlinear SVM, the dimension of the vector w can be large or even infinite.

The constrained optimization problem in Eq. (1) can be handled by Lagrange multiplier approach. The Lagrange function is constructed as Eq. (2)

$$\begin{aligned} L(w,b,\xi_i,\alpha,\mu) = \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \\ -\sum_{i=1}^{N}\alpha_i[y_i(w^T x_i + b) + \xi_i - 1] - \sum_{i=1}^{N}\mu_i\xi_i \end{aligned} \tag{2}$$

where α_i, μ_i are the Lagrange multipliers.

Based on the duality theorem, Eq. (2) are the primal problem and its corresponding dual is formulated as in Eq. (3).

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i}^{N}{}_{i=1}\xi_i \\ &subject\ to \quad y_i(w^{\mathrm{T}}x_i - b) + \xi_i - 1 \geq 0,\ 1 \leq i \leq N \\ &\qquad\qquad \xi_i \geq 0, \qquad\qquad 1 \leq i \leq N \end{aligned} \tag{3}$$

where $C > 0$ is the upper bound of the Lagrange multipliers, ζ is penalty, b is bias, N is number of training data $\{(x_1,y_1), (x_2,y_2), (x_3,y_3),\ldots.., (x_i,y_i)\}$, w is coefficients vector and $y \in \{\pm 1\}$.

The objective function of the dual problem in Eq.(3), can be formulated and summarized as the Eq.(4) by vanished the primal variables of w, b and ζ.

$$\begin{aligned} &\max imize L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j \\ &subject\quad to \\ &\qquad 0 \leq \alpha \leq C \\ &\qquad \sum_i \alpha_i y_i = 0 \end{aligned} \tag{4}$$

Thus the solution of objective function is given as in Eq. (5)

$$w = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{5}$$

To train the SVM is to search through the feasible region of the dual problem and maximize the objective function, and the optimal solution can be checked using the KKT conditions. The further detail training algorithm is described in [13].

Alternatively, the classification approaches in SVM classifier is essential to be stressed as 1).One-versus-the rest approach. 2).One-versus-one approach [9]. The first approach is to construct K separate SVMs, in which the k^{th} model $y_k(x)$ is trained using the data from class C_k as the positive examples and the data from the remaining K-1 classes as the negative examples. The second approach is to train K(K-1)/2 different 2-class SVMs on all possible pairs of classes, and then to classify test point according to which class has the highest number of 'votes'. In this paper, the one-to-one approach is adopted in our proposed framework in testing phase for speaker identification.

2.2 Confusion Matrix of Acoustic Model for Bilingual Speech Recognition

The confusion matrix is basically a confusion matrix, which is a supervised learning skill in the field of artificial intelligence and pattern recognition [10]; the confusion matrix is also called a matching matrix as well in unsupervised learning. Each column of the confusion matrix is defined as the instances in a predicted class, while each row is defined as the instances in an actual case class. The advantage of confusion matrix is simple to be observed if the system is confusing two classes. The example of confusion matrix is shown in Table.2.1.

Table.2.1. the example of confusion matrix.

		Actual case class		
Predicted class		[m]	[d]	[b]
	ㄇ	90	5	5
	ㄉ	0	100	0
	ㄋ	10	0	90

The rows of the confusion matrix are always normalized by summarized number of total symptoms for evaluation. And the value of the confusion frequency in the estimated matrix is as the relative number of confusion. Eq.(6) is given by

$$\hat{s} = \frac{card\{k : \Omega_k^{k_i} \; is \quad classified \quad as \quad k_i\}}{card\{k : \Omega_k^{k_i}\}} \tag{6}$$

where card is defined as number of elements. $\hat{s}$: is a confusion matrix estimation which is obtained for the set of models, it contains the estimation of how likely it is that a given model is classified as other model.

The procedures of the mixed-language acoustic model based on Mandarin and English are shown in follow:

1). Clustering the similarity phones set acoustically and phonetically in English and Mandarin.

2). The monophonic sets are built by single Gaussian acoustic model.

3). For each phone in Mandarin, we calculate the dissimilarity of the phone set based on the confusion matrix to all the phones in the same group for English. If the value is below a threshold, the source phone in Mandarin would be mapped to that phone in English. Otherwise, both the phones would be modeled separately in the bilingual system.

4). If some phones in Mandarin can not map to phone cluster in English, in such cases will not try to map this phone in mandarin to English.

5). While the list of phones in bilingual system is finished, the lexicon for Mandarin is edited by using the mapping rules. The mixed-language phone set is shown in the Table 2.2.

Table 2.2. The mixed-language phone set

INX	MPA	Eng P.h.	Model	INX	MPA	Eng P.h.	Model	INX	MPA	Eng P.h.	Model
1	ㄅ	[b]	B	20	ㄘ		C	39		[ʤ]	JH
2	ㄆ	[p]	P	21	ㄙ	[s]	S	40		[θ]	TH
3	ㄇ	[m]	M	22	ㄚ	[ɑ]	AA	41		[h]	HH
4	ㄈ	[f]	F	23	ㄛ	[o]	OW	42		[w]	W
5	ㄉ	[d]	D	24	ㄜ	[ə]	@	43		[r]	R
6	ㄊ	[t]	T	25	ㄝ	[ɛ]	EH	44		[j]	Y
7	ㄋ	[n]	N	26	ㄞ	[aɪ]	AY	45		[ɪ]	IH
8	ㄌ	[l]	L	27	ㄟ	[e]	EY	46		[ʊ]	UH
9	ㄍ	[g]	G	28	ㄠ	[aʊ]	AW	47		[ʌ]	AH
10	ㄎ	[k]	K	29	ㄡ		OU	48		[ɝ]	ER
11	ㄏ		H	30	ㄢ		AN	49		[ɔ]	AO
12	ㄐ		J	31	ㄣ		EN	50		[æ]	AE
13	ㄑ		Q	32	ㄤ		ANG	51		[ð]	DH
14	ㄒ		X	33	ㄥ	[ŋ]	NG	52		[ʃ]	SH
15	ㄓ		Zh_m	34	ㄦ		ER	53		[ʒ]	ZH
16	ㄔ		Ch_m	35	ㄧ	[i]	IY	54		[v]	V
17	ㄕ		Sh_m	36	ㄨ	[u]	UW	55		[ɔɪ]	OY
18	ㄖ	[z]	ZR	37	ㄩ		YU				
19	ㄗ		Z	38		[ʧ]	CH				

3. Proposed Framework

3.1 Proposed Consolidation of Speech Recognition Framework

The appealing work of this paper is to propose a framework, which is integrated the speaker identification and speech recognition into a consolidated speech recognition system, this architecture is shown in Fig 3.1. This system is capable of dealing with one specified word of speech signal as prior defined, then using SVM based speaker identification procedure to find out the target speaker, at the same time the same speech signal is then used to examine second target speaker by confusion matrix based of speech identification system. The scenario is such as: The Speaker A pronounces a sentence as, "I want to leave message to Speaker B" to proposed system:, then the SVM based speaker identification will recognize Speaker A by the input speech utterance, and the speech identification system will recognize the Speaker B by the same word string of speech utterance.

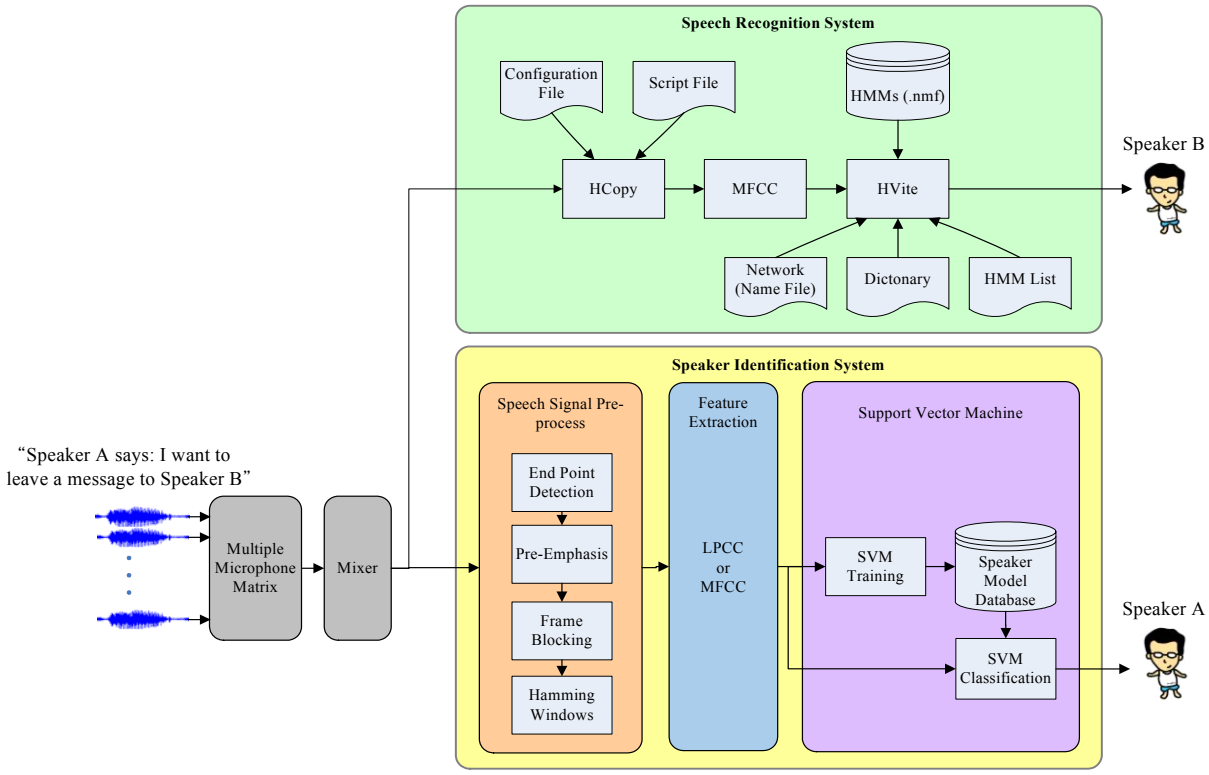


Fig.3.1 Proposed Consolidation of Speaker and Speech Recognition System

3.2 Feature Extraction and VAD Process of Consolidated System

The input speech signal of the consolidated speech system is collected from the ubiquitous speech environment [6]. The ubiquitous speech environment is a microphone-array framework which is composted of six microphones on the far-field space, when the speech signal are received then mixed to be only one speech signal by mixer and output for feature extraction. The 18th order of LPCC feature extraction

method is for SVM based speaker identification, while MFCC features are obtained by using HTK for speech recognition system.

Based on utilizing the information of energy, zero-crossing rate and the spectral flatness, then the informational signal frame is thus detected and non-information frame is ignored. If the silence is presented among two or more frames between signals, then that particular segments must be removed from the original speech signal.

VAD (Voice activity detection) method of EPD (end-point detection) is also adopted. The goal of EPD is to identify the important part of informational signals for further processing. The EPD is also known as "speech detection" or "voice activity detection", which usually play an important role in speech signal processing and recognition.

3.3 Multi-class SVMs of Testing Phase in Speaker Identification

The Multi-class (C classes) training process is to train C(C-1)/2 different 2-class SVMs on all possible pairs of classes, and then to classify test points according to which has the highest number of 'votes', this approach is called one-versus-one [12]. Table2.3 are shown that all possibly pair-wise classifier and listed as diagonal symmetric matrix, each pair-wise classifier is defined as a hyper plane between two classes. The repetitions of the pair-wise are presented from the diagonal symmetric matrix. According to the analysis and utilize the diagonal pair-wise, then only C(C-1)/2 "votes" , which also means that there are C(C-1)/2 hyperplanes are acquired for comparison.

Table 2.3. The method of C-classes for SVM training process

1 vs. 1	1 vs. 2	1 vs. 3	…	1 vs. C
2 vs. 1	2 vs. 2	2 vs. 3	…	2 vs. C
3 vs. 1	3 vs. 2	3 vs. 3	…	3 vs. C
C vs. 1	C vs. 2	C vs. 3	…	C vs. C

In testing phase, the extracted input sample using discriminate function to make a decision that whether the input sample are resided in Class 1 or Class 2. If the result of discriminate function $y \geqq +1$, then input data belongs Class 1, otherwise if $y \leqq -1$ then it is in Class 2. Eq.(7) is shown that symbol of $vote_{1,2}$ is the normalized result of input samples in Class 1, while $vote_{1,2}$ is the testing sample in Class 2.

$$
\begin{aligned}
vote_{1,2} &= \frac{test\ \ pattern\ in\ calss\ 1}{N} \\
vote_{2,1} &= \frac{test\ \ pattern\ in\ calss\ 2}{N} \\
vote_{1,2} &+ vote_{2,1} = 1
\end{aligned}
\tag{7}
$$

Based on Eq.(7), when the input sample data after undergoing C(C-1) hyper plane or $vote_{i,j}$ computations except diagonal $vote_{i,i}$ in Table.2.4. In order to find the high score of the target speaker, then to summarize *votes* from each row beside diagonal votes in Table.2.4. And then to compare values of summarized rows and to find out the maximum through each class. Finally, the maximum value of the class is represented the target speaker. The corresponding equation of evaluating target speaker is shown in Eq.(8).

Table 2.4. The $vote_{i,j}$ value of each class

Compare Class j / Value of type i	Class 1	Class 2	Class 3		Class C
Class 1	$vote_{1,1}$	$vote_{1,2}$	$vote_{1,3}$	…	$vote_{1,C}$
Class 2	$vote_{2,1}$	$vote_{2,2}$	$vote_{2,3}$	…	$vote_{2,C}$
Class 3	$vote_{3,1}$	$vote_{3,2}$	$vote_{3,3}$	…	$vote_{3,C}$
:	:	:	:	:	:
Class C	$vote_{C,1}$	$vote_{C,2}$	$vote_{C,3}$	…	$vote_{C,C}$

$$
\begin{aligned}
&class\ i\ value\ is\ vote_i = \sum_{j=1}^{C} vote_{i,j} \\
&target\ class = \arg\max_i vote_i
\end{aligned}
\tag{8}
$$

3.4 HTK based Speech Recognition for Testing Phase

The components of proposed consolidated system in speech recognition are included as 1) Tree lexicon, 2) The task grammar, and 3) Viterbi beam search [10]. The first component is to create a dictionary. The dictionary provides an association between words used in the task grammar and the acoustic models, in that may be composed of sub word (phonetic, syllabic etc.) units. The second component of task grammar is to constrain on what the recognizer can expect as input, as the system

built, then a voice operated interface is provided for name recognition, it is capable of handling the word strings. In order to limit the scope of this work, only the syllable to deal with name grammars is needed. The final component is the Viterbi beam search. This component is essentially a dynamic programming algorithm, consisting of traversing a network of HMM states and maintaining the best possible path score at each state in each frame. It is a time-synchronous search algorithm in that it is to process all states completely at time *t* before moving on to time *t + 1*.

After three components are finished, and the recognizer is complete and ready for evaluating the performance. The recognition process can be summarized as in the top part of Fig.3.1 related to Speak B. In the beginning, the input speech signal is transformed into a series of"acoustical vectors" (here MFCCs) by using the HTK tool HCopy, in the same way as what was done with the training data. The input observation is then processed by the Viterbi algorithm using the HTK tool HVite.

4. Experimental Results

In order to evaluate the performance of the consolidated system in real life, thus the experiments are tested in the Aspire Home, which is located in the NCKU Chi-Mei Building. The training and testing phase of the individual speaker identification system and speech recognition system as well as proposed consolidated speech recognition system are setup and evaluated, respectively.

4.1 Experimental Results of Individual Speaker Identification System

The component of speaker identification system is to use 18 order of LPCC feature. Ten seconds of speech utterances is for training, and two second of speech utterance is for testing. Total 10 persons are assessed in this case. The key elements of speech preprocess include the end point detection and voice activity detection. The parameters γ and C are setting to be 0.0005 and 50, respectively. The

	Single Microphone (Omni-directional)	Wireless Microphone (Omni-directional)	Microphone Array (Omni-directional)
Accuracy Rate (Silence Mode)	76.6%	91.6%	96.6%
Accuracy Rate (TV noise Mode)	66.7%	86.6%	73.3%

Fig. 4.1. The experimental results of individual component in speaker Identification system

sample rate is 16 kHz, and the frame size is 512 points. Three types of microphone configuration is to be assessed, such as single microphone, wireless microphone and microphone array. Two modes of background noise are also adopted, i.e. silence mode

v.s. TV noise mode are built in test environment. The experimental results of individual component in speaker identification system are shown in Fig. 4.1

4.2 Experimental Results of Individual Speech Recognition System

The training databases include English Across Taiwan (EAT) and Mandarin Across Taiwan-400 (MAT-400). Using man-made sifting way, then EAT are totally remaining 8375 wave files, including English long sentences, English short sentences and English words. The corpus contains 19221 words for training. In MAT-400, the MATDB-4 (1200) and MATDB-5 (400) category are adopted. By using man-made sifting way, there are totally remaining 15400 wave files, including words of 2 to 4 syllables and phonetically balanced sentences. The corpus contains 80903 words for training. There are one hundred of testing word strings, which can be regarded as voice command, and the content included Chinese movie name, English words and Chinese/English mixture sentence and etc. There are totally 10 people to test this system. The speaker randomly selected twenty sentences of testing string to evaluate the system. The experimental results of individual component in speech recognition system are shown in Fig. 4.2

Number of speaker	Positive Identification	Negative Identification	Testing Times	Accuracy Rate
8 male speakers	127	33	160	79.38%
2 Female speakers	24	16	40	60.00%
Total Speaker	151	49	200	75.5%

Fig.4.2. the experimental results in speech recognition system

4.3 Experimental Results of Consolidated System

The consolidated speech recognition system is examined in the real doorway of Aspire House in NCKU. In this experiment, totally six persons are joined test. The training and testing utterance period is the same as speaker identification system. In the speech recognition, there are six sentences of word string for testing, each testing pattern is formatted as “I want to leave message to XXX”, the sub-string “XXX” is represented as the name of six speakers in English or Mandarin. When the speaker pronounces the randomly selected testing pattern within six sentences, the system will identify two target speakers at the same time from the real speaker and word string speaker. Each speaker is to test the system for thirty times. In this test, the single microphone configuration is used for assessment. The experimental results of consolidated system are shown in Fig. 4.3.

	Positive Identification	Negative Identification	Consolidated System Accuracy
Speaker Identification Accuracy	157	23	87.22%
Speech Recognition Accuracy	164	16	91.1%

Fig.4.3. the experimental results of consolidated speaker and speech recognition system

5. Conclusion

For human-centric digital life, this paper has presented an integrated architecture of speaker identification and speech recognition for intelligent doorway application. This framework is capable of recognize two target speakers via only one-word utterance. The SVM is used for speaker identification and the confusion matrix is used for develop the multi-lingual speech recognition system. This integrated system has been realized in the intelligent doorway of our prototype digital house. The future work intends to integrate a sound localization technique to localize the speaker position.

References

[1] V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[2] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowl. Discov., vol. 2, no. 2, pp. 1–47, 1998.

[3] V. Wan and S.Renals, "Speaker verification using sequence discriminant support vector machines," IEEE trans. On Speech and Audio Processing vol.13, no. 2, Mar.2005.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-Field Speaker Recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, Sep. 2007.

[5] M. Ji, S. Kim, H. Kim, K. C. Kwak, and Y.J. Cho, "Reliable Speaker Identification Using Multiple Microphones in Ubiquitous Robot Companion Environment," 16th IEEE International Conference on Robot & Human Interactive Communication.

[6] J. F. Wang, T.W. Kuan, J.C. Wang, and G. H. Gu,"Ubiquitous and Robust Text-Independent Speaker Recognition for Home Automation Digital Life," UIC 2008, LNCS 5061, pp. 297–310, 2008.

[7] C.L. Huang, C-H Wu, "Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis", Department of

Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C. (2007)

[8] C. Y MA, Pascale FUNG, "Using English Phoneme Models for Chinese Speech Recognition" , The Human Language Technology Center Department of Electrical and Electronic Engineering Hong Kong University of Science and Technology (HKUST), Hong Kong

[9] F. Seide, J. C. Wang, 1998. Phonetic modeling in the Philips Chinese continuous-speech recognition system. In Proc.

[10] P.Y. Shih, J.F. Wang, H.P. Lee, H.J. Kai, H.T. Kao, Y.N. Lin ," Acoustic and Phoneme Modeling Based on Confusion Matrix for Ubiquitous Mixed-Language Speech Recognition," IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, DOI 10.1109/SUTC Jun,2008

[11] J.C. Wang, C.H.Yang, J.F. Wang, and H.P. Lee, "Robust speaker identification and verification," IEEE Compu. Intell. Mag., pp.52-59, May 2007.

[12] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY :Springer Science+Business Media, 2006, pp. 325-358

[13] J.C. Platt," Sequential Minimal Optimization for SVM," Published by Pennsylvania State University, http://citeseerx.ist.psu.edu/. 2007

[14] J. F. Wang, and G. H. Gu," Ubiquitous and Robust Text-Independent Speaker Recognition and FPGA Implementation for SMO algorithm of SVM", Master Degree Dissertation, July, 2008.

Voice Activity Detection Using Spectral Entropy in Bark-Scale Wavelet Domain

王坤卿 Kun-ching Wang, 侯圳嶺 Tzuen-lin Hou
實踐大學資訊科技與通訊學系
Department of Information Technology & Communication
Shin Chien University
kunching@mail.kh.usc.edu.tw

秦群立 Chuin-li Chin
中山醫學大學應用資訊科學學系
Department of Applied Information Sciences
Chung Shan Medical University

Abstract

In this paper, a novel entropy-based voice activity detection (VAD) algorithm is presented in variable-level noise environment. Since the frequency energy of different types of noise focuses on different frequency subband, the effect of corrupted noise on each frequency subband is different. It is found that the seriously obscured frequency subbands have little word signal information left, and are harmful for detecting voice activity segment (VAS). First, we use bark-scale wavelet decomposition (BSWD) to split the input speech into 24 critical subbands. In order to discard the seriously corrupted frequency subband, a method of adaptive frequency subband extraction (AFSE) is then applied to only use the frequency subband. Next, we propose a measure of entropy defined on the spectrum domain of selected frequency subband to form a robust voice feature parameter. In addition, unvoiced is usually eliminated. An unvoiced detection is also integrated into the system to improve the intelligibility of voice. Experimental results show that the performance of this algorithm is superior to the G.729B and other entropy-based VAD especially for variable-level background noise.

Keywords: Voice Activity Detection, Bark-Scale Wavelet Decomposition, Adaptive Frequency Subband Extraction.

1. Introduction

Voice activity detection (VAD) refers to the ability of distinguishing speech from noise and is

an integral part of a variety of speech communication systems, such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation [1]. In the GSM-based wireless system, for instance, a VAD module [2] is used for discontinuous transmission to save battery power. Similarly, a VAD device is used in any variable bit rate codec [3] to control the average bit rate and the overall coding quality of speech. In wireless systems based on code division multiple access, this scheme is important for enhancing the system capacity by minimizing interference. Common VAD algorithms use short-term energy, zero-crossing rate and LPC coefficients [4] as feature parameters for detecting voice activity segment (VAS). Cepstral features [5], formant shape [6], and least-square periodicity measure [7] are some of the more recent metrics used in VAD designs. In the recently proposed G.729B VAD [8], a set of metrics including line spectral frequencies (LSF), low band energy, zero-crossing rate and full-band energy is used along with heuristically determined regions and boundaries to make a VAD decision for each 10 ms frame.

In this paper we present a robust VAD algorithm for the detection of speech segment, which is based on the entropy of the spectrum domain of selected critical subband. First, the bark-scale wavelet decomposition (BSWD) is utilized to decompose the input speech signal into 24 critical subband signals. In contrast to the conventional wavelet packet decomposition, the BSWPD is designed to match the auditory critical bands as close as possible and has been applied into various speech processing systems [9, 10]. The entropy, on the other hand, a measure of amount of expected information, is broadly used in the field of coding theory. Shen *et al*. [11] first used it on speech detection and revealed that voiced spectral entropy is quite different from non-voiced one. Based on this character, the entropy-based approach is more reliable than pure energy-based methods in some cases, particularly when noise-level varies with time.

Since the frequency energy of different types of noise focus on different frequency subbands,

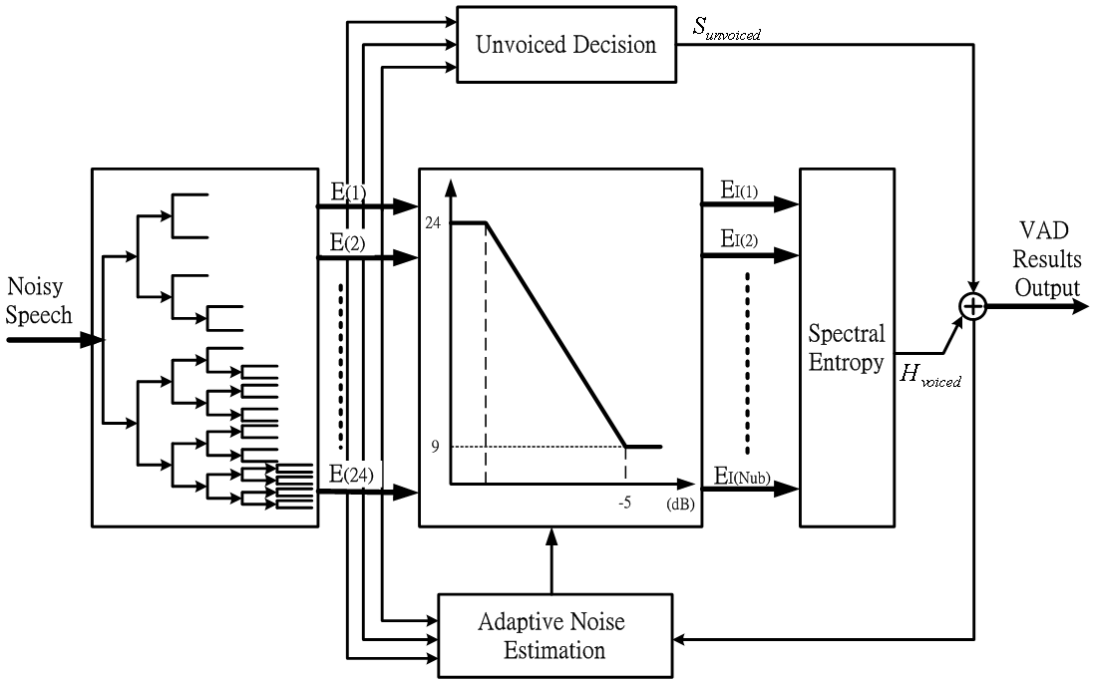


Figure 1. The Block Diagram of Proposed VAD Algorithm

the effect of corrupted noise on each frequency subband is different [12]. The seriously obscured frequency subbands have little word signal information left, and are harmful for detecting VAS. Based on the finds, we adopt the theory of adaptive frequency subband extraction (AFSE) to only uses the frequency subband which are slightest corrupted and discard the seriously obscured ones. The frequency subband energies are sorted and only the first several frequency subband with the highest energy are selected. Experiment results show that when more frequency subbands are corrupted by noise, the number of the selected frequency subbands decreases with the decrease of the SNR. A measure of entropy defined on the spectrum domain of selected frequency subband by the AFSE approach is proposed to refine the classical entropy-based VAD [12]. Finally, an unvoiced detection is integrated into entropy-based VAD system to improve the intelligibility of voice.

2. Implementation of the Proposed VAD Algorithm

In the block diagram shown in Fig. 1, the proposed VAD algorithm consists of five main parts:

bark-scale wavelet decomposition, adaptive frequency subband extraction, calculation of spectral entropy, adaptive noise estimation, and unvoiced decision. In this section, the five main parts are described in turn.

2.1 Bark-scale wavelet decomposition (BSWD)

Critical subband is widely used in perceptual auditory modeling [13]. In this section, we propose the wavelet tree structure of BSWD to mimic the time-frequency analysis of the critical subbands according to the hearing characteristics of human cochlea. A BSWD is used to decompose the speech signal into 24 critical wavelet subband signals, and it is implemented with an efficient five-level tree structure. The corresponding BSWD decomposition tree can be constructed as shown in Fig. 2. Observing the Fig.2, the input speech signal is obtained by using the high-pass filter and low-pass filter [14], implemented with the Daubechies family wavelet, where the symbol $\downarrow 2$ denotes an operator of downsampling by 2.

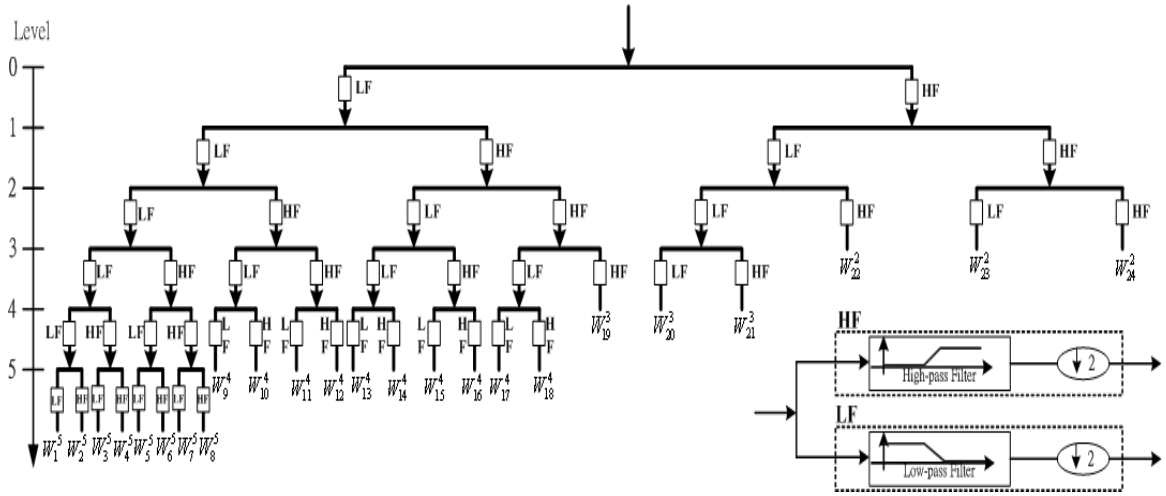


Figure 2. The Tree of Bark-Scale Wavelet Decomposition (BSWD)

2.2 Adaptive frequency subband extraction (AFSE)

In fact, the frequency energies of difference types of noise are concentrated on different frequency subbands. This observation demonstrates that not all the frequency subbands have

harmful word signal information. In our algorithm, we must use only the useful frequency subbands or discard the harmful subbands for detecting VAS. Since our goal is to select some useful frequency subbands having the maximum word signal information, we need a parameter to stand for the amount of word signal information of each frequency subband. According to Wu *et al*. [12], the estimated pure speech signal is a good indicator. The frequency subbands energy of pure speech signal is accomplished by removing the frequency energy of background noise from the frequency energy of input noisy speech.

For the mth frame, the spectral energy of the ξth subband is evaluated by the sum of squares:

$$E(\xi,m)=\sum_{\omega_{\xi,l}}^{\omega_{\xi,h}}\left|X(\omega,m)\right|^2, \tag{1}$$

where $X(\omega,m)$ means the ωth wavelet coeffience. $\omega_{\xi,l}$ and $\omega_{\xi,h}$ denote the lower boundaries and the upper boundaries of the ξth subband, respectively.

The ξth frequency subbands energy of pure speech signal of the mth frame $\tilde{E}(\xi,m)$ is estimated:

$$\tilde{E}(\xi,m)=E(\xi,m)-\tilde{N}(\xi,m), \tag{2}$$

where $\tilde{N}(\xi,m)$ is the noise power of the ξth frequency subband.

During the initialization period, the noisy signal is assumed to be noise-only and the noise spectrum is estimated by averaging the initial 10 frames. To recursively estimate the noise power spectrum, the subband noise power, $\tilde{N}(\xi,m)$, can be adaptively estimated by smoothing filtering and be discussed later.

It is found that the more the frequency subband covered by noise would result in the smaller the $\tilde{E}(\xi,m)$. Since the frequency subband with higher $\tilde{E}(\xi,m)$ contains more pure speech

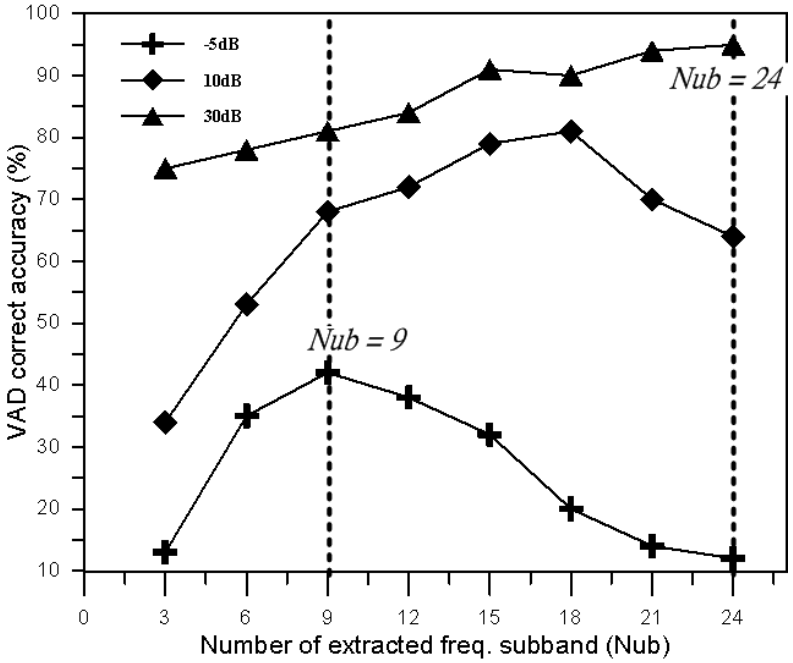


Figure 3. The Results of Correct Detection Accuracy with Number of Different Frequency Subband at –5dB, 10 dB and 30 dB under Three Types of Noise.

information, we should sort the frequency subband according to their $\tilde{E}(\xi,m)$ value.

That is,

$$\tilde{E}(I_1,m) \geq \tilde{E}(I_2,m) \geq \cdots \geq \tilde{E}(I_N,m), \tag{3}$$

where I_i is the index of the frequency subband with the ith max energy.

It means that the index of the frequency subband with higher energy is the more useful index of one. Moreover, we should only select the useful frequency subbands for VAD results output. That is, the first N frequency subbands $I_1,I_2,\ldots,I_N$ are selected and denoted as the useful number of frequency subband, N_{ub}, for the succeeding calculation of spectral entropy. According to the relation between the number of useful frequency subbands N_{ub} and SNR (shown as Fig. 3), we can see that the number of useful frequency subband increases with the increase of SNR under three types noises including white noise, factory noise and vehicle noise. $N_{ub}=9$ and $N_{ub}=24$ denote the boundary of N_{ub} among the range from -5dB to 30dB, respectively.

Based on the above finds, a linear function can be used to simulate the relationship between N_{ub} and SNR, and shown as Fig. 4.

$$N_{ub}(m)=\begin{cases}9 & ,SNR(m)<-5dB\\ [(24-9)\times\dfrac{(SNR(m)-(-5))}{30-(-5)}+9] & ,\text{-}5dB\le SNR(m)\le 30dB\\ 24 & ,SNR(m)>30dB.\end{cases} \tag{4}$$

where $[\cdot]$ is the round off operator, and $SNR(m)$ denotes a frame-based posterior SNR for the mth frame.

In addition, $SNR(m)$ is depended on the all summation of subbnad-based posterior SNR $snr(\xi,m)$ on the ξth useful subband and defined as:

$$SNR(m)=10\log_{10}\sum_{\xi\in N_{ub}}snr(\xi,m), \tag{5}$$

where

$$snr(\xi,m)=\frac{|X(\xi,m)|^2}{\tilde{N}(\xi,m)}.$$

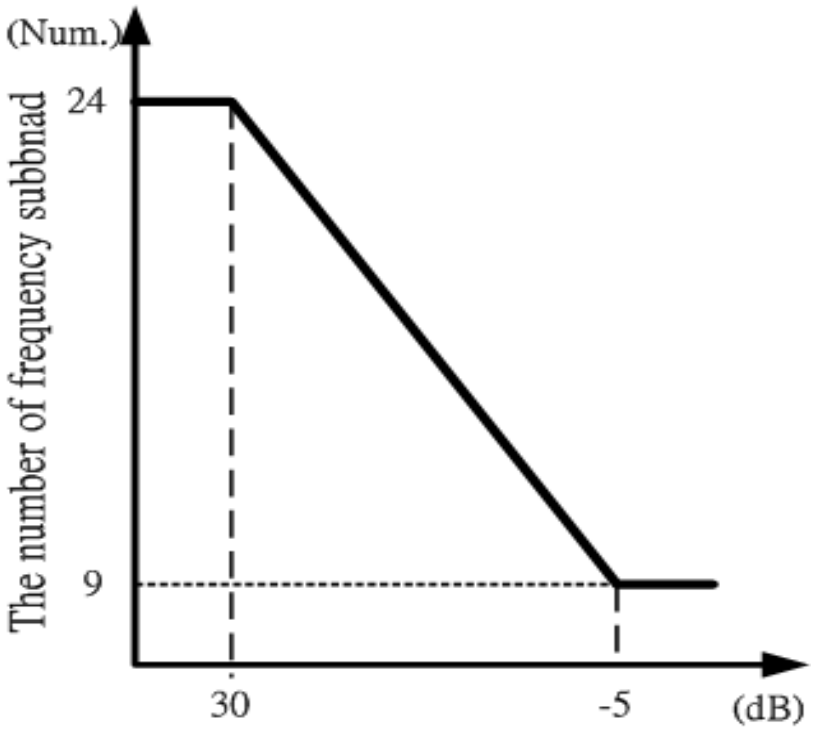


Figure 4. A Linear Function of the Relationship Between N_{ub} and SNR

2.3 Calculation of spectral entropy

To calculate the spectral entropy, the probability density function (pdf) and the entropy calculation are both necessary steps.

The pdf for the spectrum can be estimated by normalized the frequency componemts:

$$P(\xi,m) = E(\xi,m) \Big/ \sum_{\omega=1}^{N} E(\omega,m) \tag{6}$$

where $P(\xi,m)$ is the corresponding probability density, and N denotes the total number of critical subbnad divided by BSWD ($N = 24$ in this paper).

Some frequency subbands, however, are corrupted seriously by additive noise, and those harmful subbands may result in low performance of entropy-based VAD if those are extracted. Moreover, we use only the useful frequency subbands to calculate a measure of entropy defined on the spectrum domain of selected frequency subbands. The probability associated with subband energy modified from (6) is described as follows:

$$P(\xi,m) = E(\xi,m) \Big/ \sum_{\omega=1}^{N_{ub}} E(\omega,m), \tag{7}$$

where N_{ub} is the number of useful frequency subbands.

Having finishing applying the above constraints, the spectral entropy $H(m)$ of frame m can be defined below.

$$H(m) = -\sum_{\xi=1}^{N_{ub}} P(\xi,m) \cdot \log[P(\xi,m)]. \tag{8}$$

The foregoing calculation of the spectral entropy parameter implies that the spectral entropy depends only on the variation of the spectral energy but not on the amount of spectral energy. Consequently, the spectral entropy parameter is robust against changing level of noise.

2.4 Adaptive noise estimation

To recursively estimate the noise power spectrum, the spectral power of subband noise can be estimated by averaging past spectral power values using a time and frequency dependent smoothing parameter as following:

$$\tilde{N}(\xi,m) = \alpha(\xi,m) \cdot \tilde{N}(\xi,m-1) + (1-\alpha(\xi,m)) \cdot E(\xi,m) \tag{9}$$

where $\alpha(\xi,m)$ means the smoothing parameter and be defined as

$$\alpha(\xi,m) = \begin{cases} 1, & \text{if VAD(m-1)=1,} \\ \dfrac{1}{1+e^{-k\cdot(snr(\xi,m)-T)}}, & \text{otherwise.} \end{cases} \tag{10}$$

where T is used for center-offset of the transition curve in Sigmoid.

Observing (10), it is found that the smoothing parameter set one when previous speech-dominated frame, the spectral power of subband noise keep until noise-dominated frame. Otherwise, the smoothing parameter may be chosen as a Sigmoid functions when noise-dominated frame.

2.5 Unvoiced decision

More unvoiced information is eliminated from conventional VAD algorithm. In order to overcome this drawback, a method of unvoiced decision is proposed in this section. According to the structure of BSWD tree (shown as Fig. 2), the three sub-energies corresponding to the wavelet subband signals are defined as

$$E_{L0} = \sum_{j=1}^{8} W_j^5,\ E_{L1} = \sum_{j=9}^{12} W_j^4,\ E_{L2} = \sum_{j=13}^{18} W_j^4 + W_{19}^3. \tag{11}$$

The unvoiced segments are determined as:

$$S_{unvoiced} = \begin{cases} 1 & \text{, if } E_{L2} > E_{L1} > E_{L0} \text{ and } E_{L0}/E_{L2} < 0.99 \\ 0 & \text{, otherwise.} \end{cases} \tag{12}$$

2.6 Voice activity segment detection

Finally, the voice activity segment (VAS) is derived as:

$$VAS(m) = H(m) \bigcup S_{unvoiced}(m) . \quad (13)$$

3. Experimental Results

The speech database contained 60 speech phrases (in Mandarin and in English) spoken by 35 native speakers (20 males and 15 females), sampled at 4 KHz with 16-bit resolution. To set up the noisy signal for test, we add the prepared noise signals to the recorded speech signal with different SNRs range from – 5dB to 30 dB. The noise signals are all taken from the noise database NOISEX-92 [15]. Of the various noises available on the NOISEX database, white noise, factory noise and vehicle noise are selected as speech containment. Fig. 5 shows the VAD result of the proposed algorithm on the noisy speech signal "May-I-Help-you" under variable-level of noise. It is founded that the VAS of the proposed algorithm can correctly extract speech segments especially for unvoiced segment /H/ occurred at /Help/ sentence in Fig. 5(b). Conversely, in Fig. 5(c) the VAS of standard G729B performs fail during high variable-level of noise segment and unvoiced segment. In order to compare with other VADs specified in the ITU standard G.729B, we introduce three criteria: 1) the probability of correctly detecting speech frames P_{cS} is the ratio of the correct speech decision to the total number of hand-labeled speech frames. 2) the probability of correctly detecting noise frames P_{cN} is the ratio of the correct noise decision to the total number of hand-labeled noise frames. 3) the false-alarm P_f is the ratio of the false speech decision or false noise decision to the total hand-labeled frames. Under a variety of SNR's, the P_{cS}, P_{cN} and P_f of the proposed algorithm are compared with those of the VAD specified in the ITU standard G.729B [8] and other entropy-based VAD [11]. The experimental results are summarized in Table I. It is shown that. In high SNR, the result of Shen's VAD is comparable to proposed VAD. But, the proposed VAD has superior

performance to the Shen's VAD and G.729B particularly in low SNR.

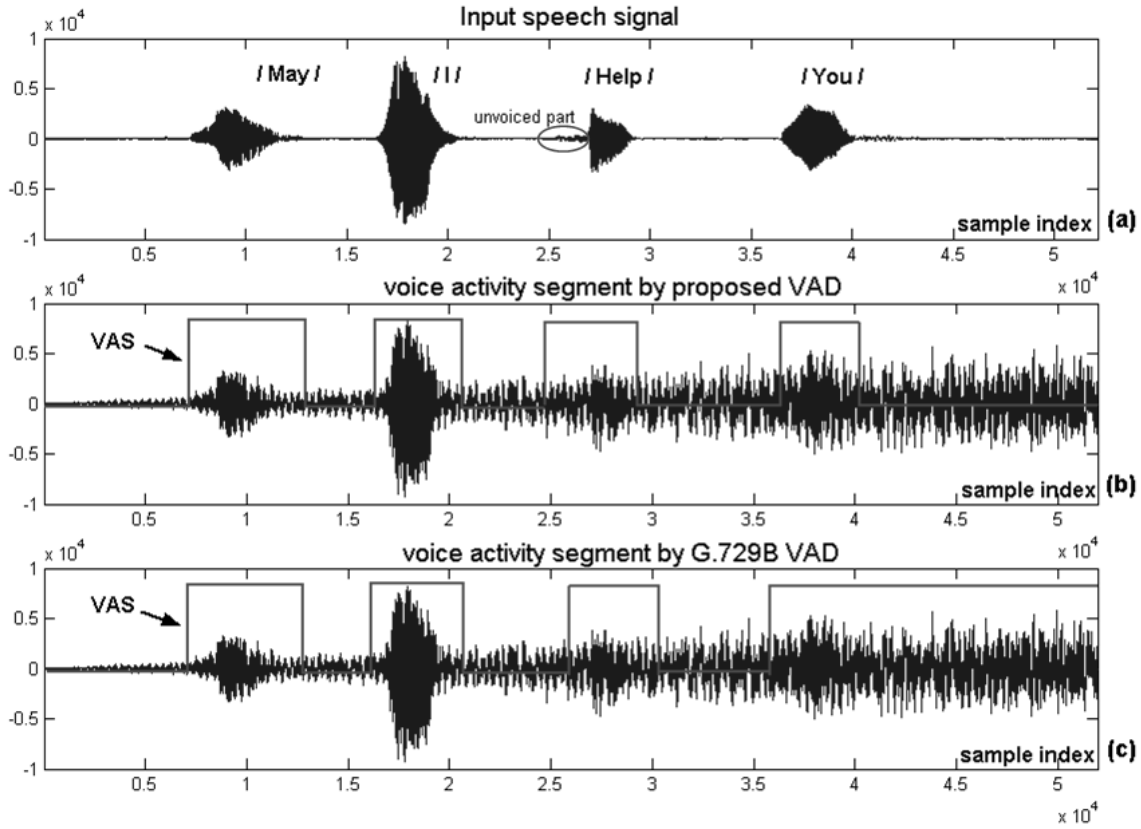


Figure 5. Comparison Between the Two VADs: (a) Waveform of Clean Speech, (b) The VAS of Proposed VAD, (c) The VAS of G.729B.

Table 1. Performance Comparisons for Three Noise Types and Levels

Noise Conditions		P_{cS} **(%)**			P_{cN} **(%)**			P_f **(%)**		
Type	**SNR(dB)**	**Proposed VAD**	**G.729B**	**Shen *et al.* [11]**	**Proposed VAD**	**G.729B**	**Shen *et al.* [11]**	**Proposed VAD**	**G.729B**	**Shen *et al.* [11]**
White Noise	***30***	99.8	93.1	99.1	99.2	84.6	99.8	1.5	12.9	1.6
	10	95.6	85.2	94.6	98.7	81.5	95.4	4.6	17.3	4.9
	-5	92.4	78.1	85.2	92.1	72.7	82.3	8.4	25.5	10.2
Factory Noise	***30***	94.6	92.9	94.3	93.1	88.9	93.0	10.2	13.6	10.8
	10	89.7	84.3	85.1	89.7	83.3	85.1	13.2	18.4	15.7
	-5	80.5	74.6	74.8	85.3	73.6	76.5	16.2	24.2	20.1
Vehicle Noise	***30***	96.8	95.3	96.5	94.2	92.3	93.1	6.3	14.3	6.5
	10	92.5	90.1	91.1	89.6	84.1	85.3	9.5	17.4	12.4
	-5	88.4	81.4	82.7	84.1	79.4	82.4	14.7	21.5	19.6

4. Conclusion

In this paper, a novel entropy-based VAD algorithm has been presented in non-stationary environment. The algorithm is based on bark-scale wavelet decomposition to decompose the input speech signal into critical sub-band signals. Motivated by the concept of adaptive frequency subband extraction, we use the frequency subband that are slightest corrupted and discard the seriously obscured ones. It is found that the proposed algorithm improves the classic entropy-based approach. Experimental results show that the performance of this algorithm is superior to the G.729B and other entropy-based approach in low SNR. The proposed algorithm has excellent presentation especially for variable-level background noise.

5. Conclusion

This work was supported by National Science Council of Taiwan under grant no. NSC 98-2221-E-158 -004.

References

[1] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan European digital cellular mobile telephone service," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1989, pp. 369-372.

[3] Enhanced variable rate codec, *speech service option 3 for wideband spread spectrum digital systems*, TIA doc. PN-3292, Jan. 1996.

[4] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1977, pp. 323-326.

[5] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *IEEE TEN-CON*, 1993, pp. 321-324.

[6] J. D. Hoyt and H. Wechsler, "Detection of human speech in structured noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, May 1994, pp. 237-240.

[7] R. Tucker, "Voice activity detection using a periodicity measure," in *Proc. Inst. Elect. Eng.*, vol. 139, no. 4, pp. 377-380, Aug. 1992.

[8] A. Benyassine, E. Shlomot, and H. Su, "ITU-T recommendation G.729, annex B, a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data spplications," *IEEE Commun. Mag.*, pp. 64-72, Sept. 1997.

[9] I. Pinter, "Perceptual wavelet-representation of speech signals and its application to speech enhancement," *Computer Speech and Language*, vol. 10, no. 1, pp. 1-22, 1996.

[10] P. Srinivasan and L. H. Jamieson, "High quality audio compression using an adaptive wavelet decomposition and psychoacoustic modeling," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 1085-1093, April 1998.

[11] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," presented at the *ICSLP*, 1998.

[12] G. D. Wu and C. T. Lin, "Word boundary detection with mel-scale frequency bank in noise environment," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 541-554, May 2000.

[13] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, New York, 1990.

[14] S. Mallat, "Multifrequency channel decomposition of images and wavelet model," *IEEE Trans. Acoust. Speech Signal Process.* 37, pp. 2091-2110, 1989.

[15] Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247-251, 1993.

讓格書寫 以及 台華互譯 初探

LangGeh Orthography and an Initial Study of Statistical Translation Between Taiwanese and Mandarin

江永進 Yuang-Chin Chiang
清華大學 統計所
Institute of Statistics, National TsingHua University
jjchiang1@hotmail.com

楊佩琪
清華大學 統計所

林淑卿
清華大學 統計所

張春凰
清華大學 通識中心

高明達
中央研究院 資訊研究所

呂仁園
長庚大學 資訊研究所

陳孟彰
中央研究院 資訊研究所

摘要

讓格書寫 是 新近提議的 書寫方式 [1]， 主張 在 語句中 適當的地方 加上 空白字元， 適合 台客華語等 使用漢字ｅ 文字系統。 大體上講， 讓格書寫 是 分 簡短詞組， 對比的是 英語的 分詞書寫， 傳統華語的 分句書寫， 以及 語言技術的 分詞技術。 讓格書寫 有 減少模糊、 方便閱讀、 利益 語言技術 等 優點； 我們 甚至認爲， 空白字母的 地位 如同 數字系統的 零。 我們 使用 讓格書寫， 製作 一套 主要是 台華語的 平行語料庫， 各約 15 萬字， 並且用來 初步探討 台華語 詞典製作、 以及 台華語互譯 問題； 在 此過程 中， 我們 利用了 台華語的 二大類似： 共同詞多、 詞序類似。 比較 現時 詞組爲基礎的 統計式 翻譯 趨向， 讓格書寫 實質上 有讓 台語的 語言技術 站在 較佳的 基礎上。

Abstract

LangGeh orthography is a new writing style proposed by [1]. For Han family languages such Taiwanese or Mandarin that uses Chinese character, *LangGeh* proposes writing with spaces in-between, using simple short phrase as a unit. This is in contrast to word-based orthography in English and sentence-based orthography in traditional Mandarin. Easy to add spaces, *LangGeh* has the advangtages of reducing ambiguity, easier to read, and easier for text processing in Chinese characters. Using the LangGeh orthography, we produce a parallel corpus in Taiwanese and Madarin, about 150 thousand characters each. We then explore the extraction of "phrase dictionary" from the parallel corpus, and begin the study of statistical translation between Taiwanese and Mandarin[7][8].

關鍵詞：讓格書寫、 翻譯詞組、 詞組典、 統計式翻譯、 台語、 華語

Keywords: *LangGeh* orthography, pharse-based translation pair, statistical translation, Taiwanese, Mandarin

一、 介紹

江永進等人[1] 提議 「讓格書寫」的 新 書寫形式， 有 幫助閱讀、 減少模糊 等 效果， 對 語言的 初學者、 外國人 有幫助； 不只 強勢語言 有利， 對 弱勢語言的 幫助， 更加明顯。

讓格書寫 採用 「四字左右 無模糊 原則」的 分 簡短詞組。 以台語爲例， 過去 受限於 斷詞正確率 不夠高， 台語 語言技術 常難以 做 進一步的 探討。 使用 讓格 書寫的 台語， 多少 避開了 斷詞的門檻， 似乎 建立了 自動 語言處理 技術的 新平臺。

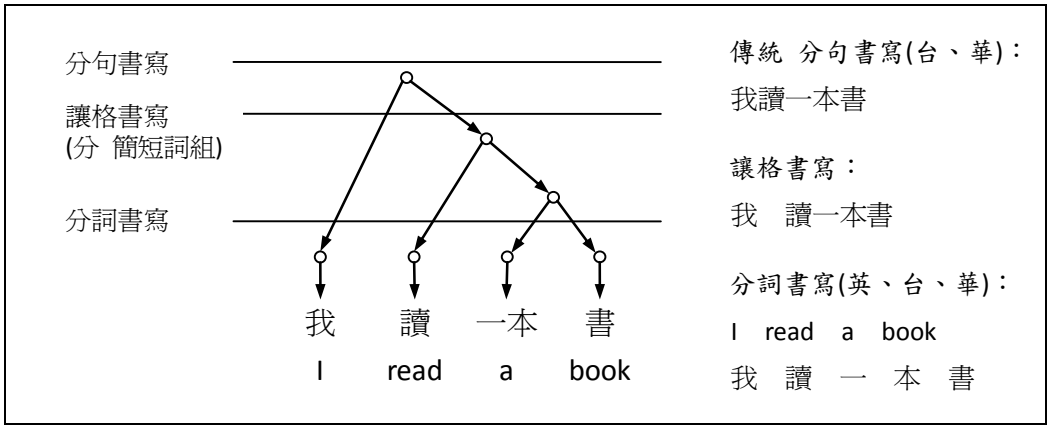
使用 讓格書寫， 本文 報告 「讓格 09 平行語料庫」。 這 是 一套 主要是 台華語的 讓格 • 平行語料庫。 同時 使用 此語料庫， 報告 台華語 對應詞組典的 抽取， 以及 使用 香腸詞組 針對 台華語對譯 做初步探討。

本文內容如下。 第二節 介紹 讓格書寫， 第三節 說明 「讓格 09 平行語料庫」 的 製作過程，包括 翻譯、 讓格、 詞組對齊。 第四節 是 對應詞組典 抽取之後的 統計數字， 第五節 探討 香腸詞組 在 台華語對譯的 實作。 最後 是結論。

二、 讓格書寫 簡介

語句 有階層性， 文字書寫的 時候 將 此階層關係 轉換做 線性形式， 因此 多少 損失了 某些消息， 造成 語意模糊。 英語等 拼音文字 以詞 爲單位， 不妨叫 分詞書寫。 相對的， 華語 主要是分句， 以句 爲單位， 句之間 以標點符號(，。？！等) 隔開；華語 不妨叫 分句書寫。當然 華語的 部份 標點符號 也有 分詞功能 (如，

頓號、連音號等)， 但是 現在 華語書寫 主要是 分句。 江永進等[1] 提議使用 分 簡短詞組 書寫的 讓格書寫。 圖一 說明 之間差別。



圖一、 語言 階層性 以及 三種 書寫方式： 分詞書寫、 分句書寫、 讓格書寫(分簡短詞組)。 (讓格的 台語發音 是 ànggēh)

讓我們 簡要敘述 讓格書寫的 「發現」過程， 然後 簡要說明 讓格規則。

一開始， 我們 持續 觀察到， 語句 若加上 適當空白， 閱讀 可以 較簡單， 因此 主張過 “space as a optional punctuation” ([2], p.141)。

然後最近， 我們 無意中 閱讀過 西方拼音文字 約自 第二世紀 開始 使用的是「連續書寫」 (scriptura continua)， 直到 第七世紀 愛爾蘭神父 爲書寫 愛爾蘭文 才於 句中 加上空白，方便 弱勢的 愛爾蘭文 的 閱讀，再經過 約四百年的 傳播，「連續書寫」 漸漸爲 「加空白書寫」 所取代， 變做現在的 分詞書寫 (Saenger [3])。 同時發生的 是， 西方文字 由「朗讀」到「默讀」， 加空白 幫助掌握 詞的界線， 效率閱讀、 快速閱讀 才 變爲可能， Saenger [3] 注重默讀 對 閱讀效率的 論述。

第三個因素 是 我們的 台語 斷詞技術 一直 無法進展。使用語料庫的 斷詞系統需要 大量語料， 華語斷詞的 正確率 可以到達 95%以上。 但是 書寫方式 類似華語的 台語， 書寫人口 少， 語料庫 不足， 斷詞的正確率 一直停留在 85%左右， 受限於此， 語句分析、 語意分析 等 進一步的 研究 一直受到 很大的限制。 最近 我們 才警覺： 傳統的 教會白話字， 使用全拼音， 沿用 西方文字的 分詞， 因此 根本不用 斷詞， 斷詞的正確率 可以說是 100%正確！

第四個因素 是 嚴格的分詞 常常過度 瑣碎、 難行。 以 圖一的例， 嚴格斷詞的結果 可以是 「我 讀 一 本 書」; 這 過度瑣碎， 跟不分詞 差別不大。

因此， 讓格 採取 分 簡短詞組 的 策略。 讓格書寫的 主要原則 是 「四字左右無模糊 原則」。 使用 四字左右 是 因爲 台華語的 雙字詞 很多， 二個詞 所合併的詞組， 模糊的機會 較小。 四字左右 只是原則， 不足四字 也可能 模糊， 超過四字也可能 不分詞組 較好：

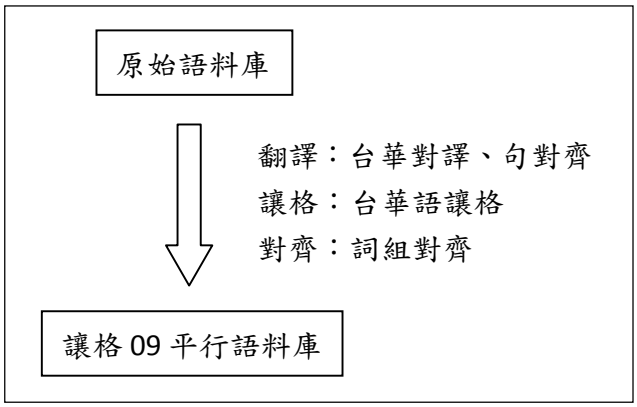
讀好書　　有模糊， 最好 作者自己 分詞組；

國際性交易　　有模糊， 最好 作者自己 分詞組；
南港區公所　　二種分法 似乎 差別不大， 建議 合寫。
其他 讓格 建議規則， 請參考 [1]。

本文 是 以讓格 書寫， 我們的經驗 顯示 閱讀 較容易， 比較 中研院 詞庫小組 的 分詞書寫， 讓格書寫 使用 也較簡單。

三、 讓格 09 平行語料庫

根基 讓格書寫， 我們 收集、 製作 一套 主要是 台華語的 平行語料庫。 從 原始語料， 經過翻譯， 再 經過讓格， 最後 再執行 詞組對齊。 過程 如圖二。



圖二、 讓格 09 平行語料庫 製作程序

三・一、 原始語料

我們的 原始語料庫 主要是 利用 三組資料， 分別爲：

(1) 自由時報 《中英對照讀新聞》 中， 2008 年的新聞 (每天一則， 共 366 則)。 原始語料 是 英文、華文。 [4]
(2) 《發明的故事》 台譯本， 譯者 爲 游政榮(2006)， 原作 爲 Hendrik · Willem · van · Loon (房龍)。 原始語料 算是 台文。 [5]
(3) 《青鳥》台譯本， 譯者 爲 林慧婷等(2009)， 原作 爲 Maurice · Maetrlinck。 原始語料 算是 台文。 [6]

三・二、 翻譯

第二個 工作是 翻譯。

原始語料 中， 新聞語料 所使用的 書寫語言 爲 英語 與 華語， 而 後兩本 名

著譯本 則爲台語。 所以， 新聞語料 主要是 翻譯成 台語， 而 其他 兩本名著 則翻譯成 華語。

對於 台華對譯 而言， 翻譯華語 要比 翻譯台語 來得 簡單多了。 因爲 大多數人，「台語書寫」 困難過 「華語書寫」， 所能使用的 台語詞彙 要比華語 來得少。

在翻譯時， 是 一段台語 翻譯 一段華語， 在這同時， 我們 也要 對語料做「句對齊」的 翻譯。這裡所說的「句」，是指 將 台語段落 或 華語段落，以「：，。；！？」此六種 標點符號， 分成 一句一句的 形式， 然後 翻譯時， 就 一句台語 翻譯 一句華語， 而 不要發生 二句台語 翻譯成 一句華語 等 情況。 如表一， 此例 是 語料庫中的 某一 台華對應的 平行段落，以及 其 對應的平行句，因爲 各句 各自對應，所以 叫 句對齊。

表一、 句對齊

台語段落	幾千年來，人 ganna 用空手去掠取活食，用空手 sa 起獵物，空手掠小動物 gah 飛禽，但是 suah ia 未想過可能 iau 有任何其他可行 e 辦法。	
翻譯成華語	幾千年來，人只有用空手去抓取活食，用空手拿起獵物，空手抓小動物和飛禽，但是卻也沒想過可能還有任何其他可行的辦法。	
句對齊	幾千年來，	幾千年來，
	人 ganna 用空手去掠取活食，	人只有用空手去抓取活食，
	用空手 sa 起獵物，	用空手拿起獵物，
	空手掠小動物 gah 飛禽，	空手抓小動物和飛禽，
	但是 suah ia 未想過可能 iau 有任何其他可行 e 辦法。	但是卻也沒想過可能還有任何其他可行的辦法。

做句對齊的 好處 就是， 可以方便 我們 在 學習語文 時， 清楚地 了解此句台語 就是 對應 此句華語， 而 不用再 費心思 去尋找； 也可以 幫助我們 做 之後的研究。 不過， 台華對譯的 句對齊 比較容易 實行， 若是 中英對譯 的話， 因爲 書寫的文法 不同， 比較 不容易 完成。

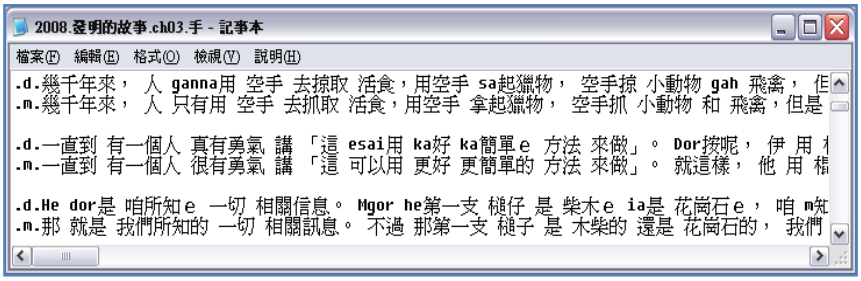
三・三、 讓格

翻譯完成 後，我們 再對語料 做讓格。而 讓格規則 已在 第二章 概述過了。值得注意的是， 我們 在做 台華讓格 時， 並不需要 像句對齊 那樣， 強制做 讓格詞組的 依照順序 對齊，句中詞組的 順序 應依照 各自語言的 自然順序。只要服從 讓格規則， 台華語 獨立讓格。 如 台語句 與 華語句：

hong 迫去 面對 非傳統 網路競爭

華語句中的 ***被迫面臨***， 是 不用特地 爲了 配合台語， 而 讓格成 ***被迫 面臨***。

讓格之後的 結果， 我們 以 utf-8 編碼 儲存成 普通文字 檔案， 如圖三。



圖三、 讓格之後的 台華語 平行語料 格式。
其中 .d.表示 台語段落， .m. 表示 華語段落。

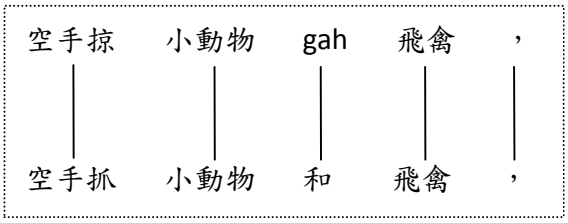
三・四、 對齊

所謂 「對齊」， 在此 是指「詞組對齊」， 簡稱「對齊」。

跟隨 Brown 等人[7][8]， 如圖四所示， 將 平行句 中 的 對應詞組 以 關聯線 連接， 每條 關聯線， 就稱爲 一個 連結 (connection)； 所謂 詞組對齊， 就是 這些 連結的 集合。 圖四中 的 平行句， 總共有 5 個連結。 連結 可以用 符號、數字 表示， 方便 機器閱讀 自動處理。 以 圖四之例 爲例， 台華語 對應句的 對齊 可表示成：

空手掠[1] 小動物[2] gah[3] 飛禽[4] ，[5]
空手抓(1) 小動物(2) 和(3) 飛禽(4) ，(5)

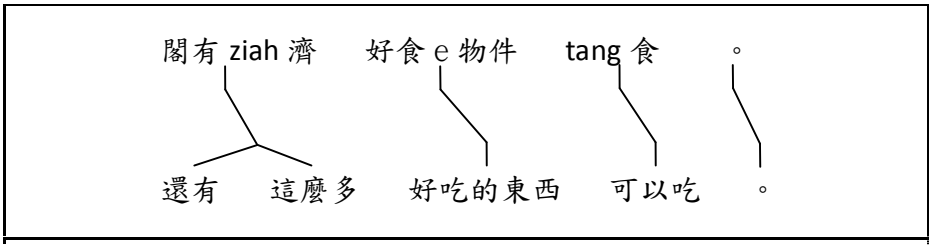
基本上 這 是[7]中的 記號， 我們 再加上 中括弧記號， 因爲 這 在 人工糾正 時，可立即 清楚知道， 華語詞組 所對應的 台語詞組 爲何者， 而 不用 再花時間 去數台語詞組的 位置。 簡言之， 前頭句的數字 代表「位置」， 後頭句的數字 代表「對齊」。



圖四、 台華讓格詞組的關聯線 以及 對齊的標記。

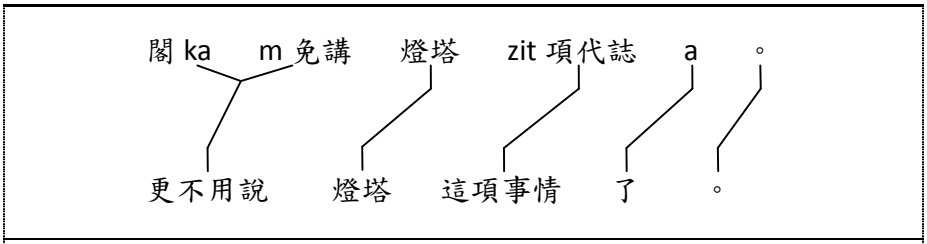
實務上， 詞組對齊的 標記 開始 是用 LCS(最長共同子序列) 爲基礎 先行自動標記， 然後 人工校正。 詳細 請見[9]。

圖四 是 對齊中 最簡單的 「1 對 1 對齊」， 圖五、六 是「1 對多對齊」、「多對 1 對齊」，其他「多對多對齊」以及 有時 無對應的 情形(「對應空詞組」)， 請見[10]。



閣有 ziah 濟[1] 好食 e 物件[2] tang 食[3] 。[4]
還有(1) 這麼多(1) 好吃的東西(2) 可以吃(3) 。(4)

圖五、「1 對多對齊」：台語詞組 1，華語詞組多

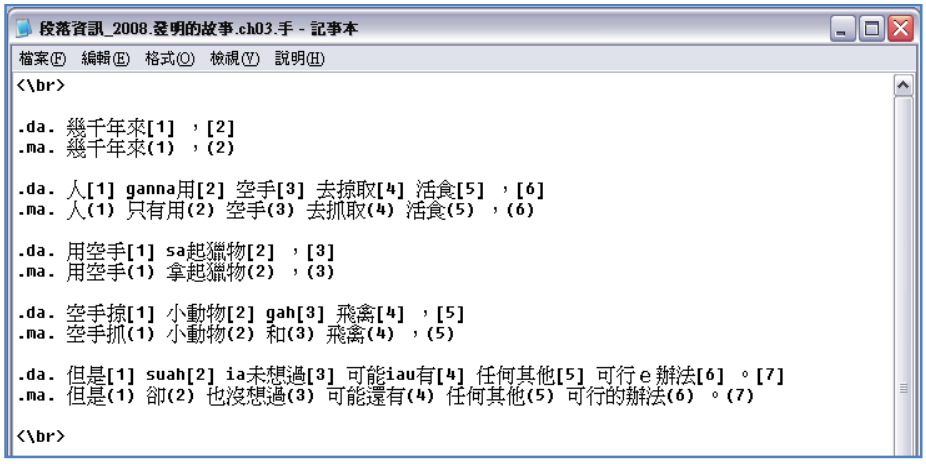


閣 ka[1] m 免講[2] 燈塔[3] zit 項代誌[4] a[5] 。[6]
更不用說(1,2) 燈塔(3) 這項事情(4) 了(5) 。(6)

圖六、「多對 1 對齊」：台語詞組多，華語詞組 1

將語料庫 做完翻譯、 讓格、 對齊 後， 即爲 讓格 09 平行語料庫。 此語料庫，是由 多人 合作完成。 利用 此語料庫， 我們 就可以 清楚知道， 台華對譯的 對應情形， 進而發展 後續的 研究工作， 如 讓格詞組典、 台華翻譯 實作。

讓格 09 平行語料庫 是 以 普通文字 檔案 儲存，只有加上 語言標記，如圖七；檔案編輯 可以使用 多樣的軟體。 儲存時， 使用 utf-8 編碼 ， 以便我們 使用 Python 程式[11] 做 讀取處理。 Python 3 全面使用 unicode， 處理文字 方便很多。



圖七、 讓格 09・平行語料庫 對齊後 格式。 .da. 表示 對齊的 台語句， .ma. 表示 對齊的 華語句， <\br> 表示 段落資訊。

四、 詞組對齊 與 台華語 對應詞組 抽取

有 讓格書寫的 平行語料 之後， 我們 可以 進一步 對齊詞組， 然後抽取 台華語 對應詞。

四・一、 台華讓格詞組典的抽取

讓格 09 平行語料庫 中 的 18363 組 對齊的句對， 容易抽取 對應的詞組對。 依照 這些詞組的 對齊情形 分類， 共 分成六類， 分別爲 「1 對 1 對齊」、「1 對多對齊」、「多對 1 對齊」、「多對多對齊」、「多種釋譯」、「跳位對齊」， 如表二， 舉些例子， 以便了解 這六種 對齊方式。

表二、 台華對齊類型

對齊類型	台語詞組	對齊	華語詞組
1 對 1 對齊	會知影	→	會知道
1 對多對齊	iau 無重視	→	還沒有 重視
多對 1 對齊	edang 防止 日頭曝傷	→	可防曬傷
多對多對齊	敢 edang pah 開	→	可以打開 嗎
多種釋譯	edang・ho	→	可以讓、能讓、能使、能夠讓
跳位對齊	ga~講	→	告訴、告訴~說

前 四種分類 在前面 已介紹過，而「多種釋譯」是 台語詞組 可以 被翻譯成 多個 不同的 華語詞組。 基本上， 前 四類對齊 都是 單一釋譯， 也就是 只有 一個 對應的 華語詞組； 而 「多種釋譯」 當中， 會有 前四類的情形 發生。 例如：

台語詞組：***一個月前***

對應的華語詞組：***一個月前***、***一個月 之前***

這當中 就有 1 對 1 對齊 (***一個月前→一個月前***)、 1 對多對齊 (***一個月前→一個月 之前***) 兩種情形。

而 「跳位對齊」 就是 台語詞組 或 華語詞組的 順序 是 不連貫的。 例如：

ga[1] zit 個警官[2] 講[3]

訴(1,3) 這名警官(2)

我們 會發現 華語詞組 ***告訴***， 所對應的 台語詞組 爲 ***ga~講*** 是 不連貫的； 又例如將 翻譯句子 改變：

ga[1] zit 個警官[2] 講[3]

告訴(1,3) 這名警官(2) 說(1,3)

則 台語詞組 和 華語詞組， 都有 不連貫的情形 (***ga~講→告訴~說***)。

然後 我們 去計算 這六種分類的 分佈情形， 如表三。 由 讓格 09 平行語料庫的 18363 組 句對齊 中， 我們 總共 得到了 108129 個 台語詞組， 其中 共有 37998 個

不同的 台語詞組。 由其分佈 可知， 大多數的 台華對齊 都是 1 對 1 的情形； 而 多種釋譯 雖然 只佔了 4.92 %， 但是 在 對齊類型 中 比例 是 第二高的， 所以 也是不容忽視的。 而 跳位對齊， 因 所佔的比例 只有 0.09 %， 所以 在做 後續的 台華翻譯 時， 我們 並沒有 針對 當 句子 發生跳位 時 的 特別處理， 即 當 句子 發生跳位 時， 其 翻譯出來的 結果， 一定會 發生錯誤。

表三、 台華讓格詞組典的分佈

對齊類型	詞組數	比例
1 對 1 對齊	34594	91.04 %
1 對多對齊	540	1.42 %
多對 1 對齊	879	2.31 %
多對多對齊	80	0.21 %
多種釋譯	1869	4.92 %
跳位對齊	36	0.09 %
總和	37998	100.00 %

四・二、 華台讓格詞組典的抽取

華台讓格詞組典 的 製作， 與 台華讓格詞組典 相同， 只是要對 讓格 09 平行語料庫 的 詞組對齊 做轉換。 原本語料的 詞組 對齊方式 是

.da. 只 edang[1] 做出[2] 幾種有限 e [3] 動作[4] ，[5]

.ma. 只能做出(1,2) 幾種有限的(3) 動作(4) ，(5)

現要將之 轉換成

.ma. 只能做出[1] 幾種有限的[2] 動作[3] ，[4]

.da. 只 edang(1) 做出(1) 幾種有限 e (2) 動作(3) ，(4)

再利用 轉換後的形式， 以 相同方法， 來製作 華台讓格詞組典。

由 讓格 09 平行語料庫 的 18363 組 句對齊 中，我們 總共 得到了 97940 個 華語詞組， 其中 共有 37647 個 不同的 華語詞組。 而 其 六種 對齊方式 分佈 如表四。 與表三 做一對照， 其分佈 與 台華讓格詞組典 差異不大。

表四、 華台讓格詞組典的分佈

對齊類型	詞組數	比例
1 對 1 對齊	33943	90.16 %
1 對多對齊	879	2.33 %
多對 1 對齊	543	1.44 %
多對多對齊	81	0.22 %
多種釋譯	2166	5.75 %
跳位對齊	35	0.09 %
總和	37647	100.00 %

我們的 讓格詞組典， 主要是 由 這些不同的 對齊類型 所組成。 可發現，就 我們的語料庫 而言， 主要是「1 對 1 對齊」 爲多，「多種釋譯」 次之。而 當中的「多種釋譯」 可以 被應用 來做 一個 「同義詞詞典」， 提供 較豐富的詞彙。

五、 台華語 互譯 初探

爲了 台華對譯的 需求， 本章 仿效[7][8]的 統計式翻譯方法， 並且 提出 簡化的 「香腸詞組」 翻譯法， 並報告 初步的結果。

在[7][8]， 翻譯的 語言對 是 英文法文。 當 給定 一法語句 F ， 以及 其可能的 英語翻譯句 E， 我們 給于 一機率 $Pr(E|F)$， 統計式 翻譯 是 在 所有可能的 E 中， 選擇 條件機率 最大者：

$$\hat{E} = \arg\max_{E} \Pr(E|F).$$

由於 $\Pr(E|F) = \Pr(F|E)\Pr(E) / \Pr(F)$， 而且 分母部份 與 E 無關， 因此

$$\hat{E} = \arg\max_{E} \Pr(F|E)\Pr(E).$$

跟隨[7]， 後面項 稱爲 語言模型(language model)， 前面項 稱爲 翻譯模型(translation model)。

關於 語言模型 $\Pr(E) = \Pr(E_1 \cdots E_n)$， 可以使用 n-gram 模型 逼近：

$$\Pr(E) \doteq \prod_{i=1}^{n} \Pr(E_i|E_{i-n+1} \cdots E_{i-1})$$

其中 $E_{i-n+1} \cdots E_{i-1} \stackrel{\text{def}}{=} E_1 \cdots E_{i-1}$ 如果 $i - n + 1 < 0$， $\Pr(E_1|E_{1-n+1} \cdots E_{1-1}) \stackrel{\text{def}}{=} \Pr(E_1)$。稍後的實驗 使用 bigram 語言模型， 或者 $n = 2$。

關於 翻譯模型， Brown 等 [8] 提出 一系列 逐漸複雜的 模型 1 到 模型 5， 並且 詳細討論 牽涉到$\Pr(F|E)$的 各項參數的 估計方法， 是 近來 以詞爲底的 統計式自動翻譯的 基礎， 而且 吸引了 很多 後續研究。

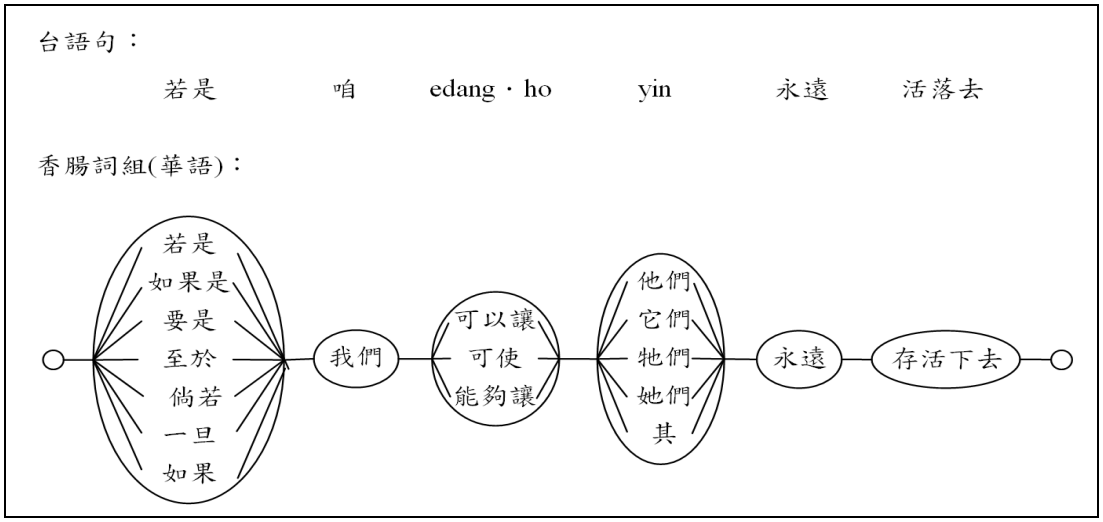
由於 我們的語料 不是 以詞爲底， 而是 以詞組爲底， 詞組數 遠大於 詞數； 而且 語料量 也不足， 此地 我們沒有 應用 Brown 等人的 方法， 而 另外提出 一個簡化的 香腸詞組的 翻譯法， 並且 初步報告 結果。

我們 舉例說明 香腸詞組。 圖八 是

台語句： “若是 咱 endang・ho yin 永遠 活落去”

(華語句： “如果 我們 可以讓 他們 永遠 活下去”)

所 對應 華語 香腸詞組串。 注意到 香腸詞組串 是 由 香腸詞組 組成， 每個香腸詞組 只是 每個 台語詞組的 可能翻譯的 包裹而已。



圖八、 香腸詞組 舉例。

如果 我們 限制 在 香腸詞組串 當中的 可能路徑 去尋找 機率最大者，那麼 最佳解的 搜尋範圍 可以 有效控制， 而且 翻譯模型的 機率估計 也可以簡化： 設 台語句 $D = (d_1, \cdots, d_I)$ 有 I 個詞組， S_{d_i} 表示 d_i 的 (華語)香腸詞組， $S_D = \{(m_1, \cdots, m_I): m_i \in S_i\}$ 表示 香腸詞組串的 所有可能的 路徑， 符號 m_i 表示 華語詞組， 那麼

$$(\hat{m}_1, \cdots, \hat{m}_I) = \underset{(m_1, \cdots, m_I) \in S_D}{\arg\max} \prod_{i=1}^{I} \Pr(m_i | d_i)$$ 不使用語言模型

$$(\hat{m}_1, \cdots, \hat{m}_I) = \underset{(m_1, \cdots, m_I) \in S_D}{\arg\max} \prod_{i=1}^{I} \Pr(m_i | d_i) \Pr(m_i | m_{i-1})$$ 使用語言模型

似乎可以 當做 最佳翻譯的 二種準則： 前者 完全不管 語言模型， 單純選擇 香腸詞組中 機率 最大者； 後者 再加上 語言模型 篩選。

對台華語 語言對 而言， 因爲 詞序接近， 共同詞 眾多， 使用 語言模型的 香腸詞組 翻譯模型 做爲 初步的 翻譯模型， 應 不算過分。 香腸詞組 有其優點： 實作 與 概念 簡單； 計算量 較少； 當然， 此地的 香腸詞組 也有限制： 譬如， 此地 沒有考慮 詞序不同的 情況； 詞組 若是 跳位對齊 則 無法處理 等等。 對本文而言， 主要是 初步探討， 我們 沒有追求 最一般的模型。 而且 實作上 有些 較小、較煩瑣的 細節， 如 句首、句尾的 機率、 多對多 詞組對應的 問題(使用 Brown 等人的 fertility 機率)， 我們就 不在此 詳述。

使用 訓練語料， 我們 估計各項 機率 如下：

$$Pr(w_1) \leftarrow \#(w_1) / N$$

$$Pr(w_i | w_{i-1}) \leftarrow \frac{\#(w_{i-1}, w_i) / N}{\#(w_{i-1}) / N} = \frac{\#(w_{i-1}, w_i)}{\#(w_{i-1})} \quad , \quad i = 2, \ldots, n$$

$$Pr(d \mid m) \leftarrow \frac{\#(d,m)}{\#(m)}$$

以下 報告 二種情況 之下的 台華互譯的 結果：(1) 使用 語言模型 (2)不使用 語言模型。 如表五，我們 從 讓格 09 平行語料庫 的 18363 組 對應句子 中，從中 隨機選取 90 %， 16527 個句子 來做 訓練語料， 其餘 1836 個句子 做爲 外部測試；再從 訓練語料 中， 隨機選取 約 5 % 826 個句子 做爲 內部測試。

表五、 各資料的總句數

資料	句子數	測試有效句數*
讓格 09 平行語料庫	18363	
訓練語料	16527	
內部測試	826	821
外部測試	1836	338

*有效句數 請見內文。

在 測試句 中， 我們 可以預期 以下的 情況：

(1) 香腸詞組 組合數 過多。 雖然 香腸詞組 已經減少 可能組合， 有時 測試語句 組合數 仍然 過於龐大。 因此 我們 爲 此組合數 設下 一門檻值—— 50 萬，若 組合數 大於 50 萬句， 則 我們 就 不翻譯 此句。

(2) 空的 香腸詞組。 受限於 有限 平行語料， 我們 不能永遠 找得到 對應詞組， 致使產生 空的 香腸詞組。這 在 outside test 極容易發生。 如果 有出現 空的 香腸詞組， 那麼 我們 就 不翻譯 此句。

(3) 有效測試句。 扣掉 組合數過多，以及 有 空的 香腸詞組 的 測試句，剩餘的 稱爲 有效測試句。

(4) 「也可」正確句。 翻譯的結果 可能 與 語料庫的答案 完全一樣， 也可能 雖與 答案不同， 但是 意思相同， 也可以 當做 翻譯正確。 Brown 等人 分別稱呼爲「Exact」、「Alternate」。 圖九 是 一個例。 此種 正確形式， 主要依靠 人工檢查， 算是麻煩。 實際的 正確率 應是 此兩種情形的 加總。

Exact

台語句 ： Mgor 最後 番仔火 勝利 a ，

標準答案： 不過 最後 火柴棒 勝利了 ，

翻譯句 ： 不過 最後 火柴棒 勝利了 ，

Alternate

台語句 ： 其他ｅ人 ma 攏 看 gah qang 去 。

標準答案： 其他的人 也都 看到 發愣 。

翻譯句 ： 其他的人 也都 看得 愣住了 。

圖九、 「exact」與「Alternate」正確句

在 以上的設定 之下，我們 使用 Python 3.0.1 [11] 實作 台華語互譯，結果如 表六 和表七。 其中 F-量度 是 正確率 以及 召回率 的 調和平均。

表六、 翻譯結果：比較 使用 語言模型的 效果 (台翻華)

語言模型	資料	有效句數	正確率 (Exact)	正確率 (Alternate)	正確率	召回率	F-量度
不使用	inside (826)	821	47.87% (393/821)	20.46% (168/821)	68.33% (561/821)	67.92% (561/826)	67.53
	outside (1836)	338	47.63% (161/338)	24.56% (83/338)	72.19% (244/338)	13.29% (244/1836)	22.45
使用	inside (826)	821	98.29% (807/821)	1.34% (11/821)	99.63% (818/821)	99.03% (818/826)	99.33
	outside (1836)	338	68.05% (230/338)	19.82% (67/338)	87.87% (297/338)	16.18% (297/1836)	27.33
註：其中的正確率 是以 有效句數 做分母， 非 句子總數。因為 我們的語料庫 過少， 而導致 台語詞組 無法找到 其對應的 華語詞組。							

表七、 翻譯結果-比較使用 語言模型的 效果(台翻華)

語言模型	資料	有效句數	正確率 (Exact)	正確率 (Alternate)	正確率	召回率	F-量度
不使用	inside (826)	826	31.60% (261/826)	23.24% (192/826)	54.84% (453/826)	54.84% (453/826)	54.84
	outside (1836)	343	37.03% (127/343)	28.86% (99/343)	65.89% (226/343)	12.31% (226/1836)	20.74
使用	inside (826)	826	97.94% (809/826)	1.57% (13/826)	99.52% (822/826)	99.52% (822/826)	99.52
	outside (1836)	343	61.52% (211/343)	29.74% (102/343)	91.25% (313/343)	17.05% (313/1836)	28.73
註：其中的正確率 是以 有效句數 做分母， 非 句子總數。因為 我們的語料庫 過少， 而導致 台語詞組 無法找到 其對應的 華語詞組。							

從 表六 及 表七， 我們 至少 可以結論：

(1) 沒有 對應詞組 當然 不能翻譯， 因此， 建立 對應詞組， 應當可以 提高 翻譯的效率。 這 不能單靠 詞組爲底的 平行語料， 應該可以 由詞爲底的 模型入手。 讓格書寫的 基礎單位 是 簡短詞組，小過 傳統翻譯的 大詞組(句)，也許會 較容易些。

(2) 由於採用 較大單位的 簡單詞組，詞組爲底的 語言模型 的 參數估計 更加容易 發生 轉移機率 估計爲 0 的 問題。 除了 傳統 語言模型的 平滑技巧，我們 值得研究 詞組 轉移機率 使用 詞轉移機率的 平滑方法。

六、 結論

讓格書寫 提供了 一個 新舞台。讓格書寫 實質上 讓 弱勢語言 如 台語客語 站在較好的 基礎上。 至少， 我們 可以不用 再受限於 斷詞， 可以 有效的 進行 進一步的 研究， 如翻譯等。 事實上， 讓格書寫 也同樣利益 強勢的華語， 使用 讓格書寫的 華文， 也可以 比較快速 使用 新想法， 得到 較佳的結果。 試想， 如果 英語現在規定 去掉 空白字元， 那麼 英語的 語言技術， 應該是 嚴重退步。 我們 沒有有必要 故意阻礙 台華客文的 語言技術。

從 技術的角度， 讓格的 分簡短詞組 接近分詞， 因此 比起 傳統的 分句書寫，讓格書寫的 分詞問題 較容易處理(如果 需要的話)； 另外一面， 比如 語言剖析的 需要， 讓格書寫的 簡短詞組 已經結合 前後詞， 實質上 語句的 單位數目 較少， 可以減低 剖析的模糊度， 因此 統計式的方法 也許 可以得到 較佳的結果。

過去 我們的 台語 自然語言 處理嘗試， 一直受限於 語料不足， 採用 讓格書寫之後， 在短時間中， 我們 也製作了 一套 句對齊、 詞組對齊 的 台華語 平行語料庫： 讓格 09・平行語料庫， 並且 用來探討 台華語 對應詞組問題， 構思出 香腸詞組， 初探 台華語 翻譯問題。

有趣的是， 同樣的方法 也利益 強勢華語。

致謝

讓格 09・平行語料庫 是 由多人完成， 除了 本文作者 參與 以外， 也受益 以下 諸位先生， 在此致謝： 呂菁菁、 游政榮、 吳德祥、 陳俊良。

參考文獻

[1] 江永進、張春凰、呂菁菁(2009). “讓格書寫： 意義、理由 gah 簡則”， *台灣風物* 59 卷 1 期，2009。

[2] 張春凰(1994). “母語寫作經驗”， *青春ｅ路途*， 台北：台笠。

[3] Saegner, Paul(1997). *Space Between Words: The Origin of Silent Reading*. Stanford University Press, Stanford, California, USA.

[4] 自由時報(2008). “中英對照 讀新聞”， 每日一則， 全年。

[5] 林慧婷、陳則伊、謝旻男(2009). *發明的故事*(台譯本)， 將出版(時行台語文會)，2009。

[6] 游政榮(2006)，*青鳥*(台譯本). 時行台語文會出版，2006。

[7] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990). “A Statistical

Approach to Machine Translation," *Computational Linguistics Volume 16, Number 2, June 1990.*

[8] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, (1993). "A Statistical Approach to Machine Translation," *Association for Computational Linguistics, 1993*.

[9] 林淑卿(2009). "從 台華平行 語料庫 擷取 對應詞組典" ， 國立 清華大學 統計所 碩士論文。

[10] 楊佩琪(2009). "讓格書寫下 統計式 台華翻譯 初探" ， 國立 清華大學 統計所 碩士論文。

[11] Python 3.0.1 (2009). http://www.python.org.