使用新式注音鍵盤及複合馬可夫語言模型之中文輸入系統

A Chinese-character Inputting System Using a New Type of Phonetic Keyboard and a Compound Markov Language Model

古 鴻 炎　　陳 志 耀
Hung-yan Gu　and　Jr-yiau Chen

國立台灣工業技術學院 電機系
Department of Electrical Engineering
National Taiwan Institute of Technology
Taipei, Taiwan, R. O. C.
E-mail: gu@mouse.ee.ntit.edu.tw

摘要

我們設計、製作了一個實際的中文輸入系統，並把一些文字編輯、查詢等輔助功能整合進來，使操作更方便。系統裡採用我們設計之宜韻注音鍵盤來輸入國語音節之聲、韻、調符號。此外，輸入者不需逐音選字，而由系統自動作音轉字的轉換處理，且即使到校對者編輯時才發現錯字，仍可方便地操作同音字、詞查詢功能而加以更正。由於系統的詞典所收錄的詞是有限的，因此，我們製作提供了線上詞彙學習的功能，包括新詞登錄，舊詞出現頻率調整，及舊詞刪除。此外，考慮到中文字裡有許多不常用的字，一般人可能不會唸而無法輸入，因此，我們提出近形字群線上查詢與建立的想法，並加以實現。

關於自動音轉字的問題，我們提出一種複合式馬可夫語言模型來解決，它不但支援線上詞彙學習功能，也把句子裡相鄰兩詞間的相關性考慮進去；然後，我們應用動態規畫演算法於模型機率之計算，以做到即時之音轉字處理。對於此模型，實際以報紙社論文章(6,291音節)及小學生作文文章(9,118音節)來測試後，分別得到了92.6及93.9之正確轉換率，這比起個別之零階、詞爲狀態之馬可夫模型，或者一階、字爲狀態之馬可夫模型的轉換率，分別高出了1.0% 及 4.4% 以上。

關鍵詞：中文輸入、注音鍵盤、馬可夫模型

國科會補助專題研究計劃編號： NSC 82-0408-E011-209

ABSTRACT

A practical system for Chinese-character inputting is designed and implemented. Also, a few auxiliary functions has been integrated into the system to make it more friendly. To input a Chinese character in terms of its three pronunciation components (i.e., syllable initial, syllable final, and lexical tone), a new design of phonetic keyboard, named "yi-yen", is adopted and implemented. For each syllable entered, the user doesn't need to select the desired Chinese character among the homonym characters because the system will automatically select the most likely sentence. Later when proof reading the text, if there are incorrectly inputted characters, the user can still correct them easily through a mechanism provided for homonym character/word replacement. Because the vocabulary words collected in the system's dictionary are limited, an on-line mechanism is therefore implemented for the user to add fresh words to the system's dictionary, to adjust an existing word's frequency count, and to delete an existing word. Also, consider that there are many less-frequently used Chinese-characters, whose pronunciations may not be known, that cannot be inputted directly. Therefore, the concept of setting up similar-profile characters' group is proposed, and on-line character-group lookup and constructing mechanisms are implemented to help the user to indirectly input those characters.

To solve the problem of automatic syllable-to-character decoding, a compound Markov language model is proposed. It not only supports the function of on-line word learning, but also makes use of correlation between adjacent words in a sentence. In addition, the algorithm of dynamic programming is elaborately used to evaluate all candidate sentences' modeled probabilities in real-time. To test the proposed model's performance, a collection of editorials from news papers (totally 6,291 syllables) and a collection of children composed articles (totally 9,118 syllables) are used in the experiments. The decoding rates obtained are 92.6% and 93.9% respectively. These numbers are at least 1.0% and 4.4% higher than the rates obtained from the zero-order word-as-state Markov model and the first-order character-as-state Markov model, respectively.

Key words: Chinese-character inputting, Phonetic keyboard, Markov model

一、導言

使用電腦來處理中文資訊的一個嚴重瓶頸在於中文之輸入，爲了提高電腦在中文社會的普及率，並加速中文資訊之電腦化，我們覺得各種中文輸入方法的發展都是很有意義的。目前可見的中文輸入方法，它們所採用的輸入媒介包括鍵盤、滑鼠[1]、感應筆(線上手寫輸入)[2,3]、掃描器(光學文字辨識)[4,5]、語音(語音辨識輸入)[6]等，其中，鍵盤仍是目前在輸入效率、操作省力性、費用等因素同時考慮下，一種較佳的中文輸入媒介，不過，由於中文字的個數是數以萬計的，如目前廣泛被採用的big-5中文內碼[7]，有一萬三千左右的字數，所以，我們很難以鍵盤來直接輸入中文，而實際上被採用的作法有：(1)賦予各個中文字一個數字代碼，而以鍵盤來輸入數字代碼；(2)依據一些規則將各個中文字拆解成字根，然後以鍵盤來輸入字根，例如倉頡、大易、行列等輸入法；(3)依據各個中文字之讀音，以鍵盤輸入此音之注音符號，例如漢音[8]、國音[9]、忘形[7]等輸入法。雖然目前已有各式各樣的中文輸入方法，但是，並沒有一種方法能夠說是十全十美的，可能被抱怨的缺點如：需花一段時間學習輸入方法且隔一段時間不用又可能忘記(如字根類之中文輸入方法)；原本之輸入速度就不快，連續使用後會因手或眼疲勞而輸入更慢(如滑鼠、線上手寫之輸入方式)；設備費用較爲昂貴(如光學文字辨識)等。

考慮數種輸入方式的優缺點後，我們決定使用原本就已配備的鍵盤來作爲輸入中文的媒介，如此，可不必多花額外費用，熟練後可盲目打鍵(打鍵時不看鍵盤)，較不易讓手、眼疲勞，並且輸入速度之上限比滑鼠與手寫方式高。此外，我們也選擇以中文字之國語讀音作爲間接輸入中文之依據，這是考慮大多數人都有學過國語注音符號，而可省去學習使用的時間，並去除可能的恐懼、疑慮。由於原始之注音輸入法需由使用者逐音選取螢幕上出現的同音字，而使輸入速度快不起來，並且無法盲目打鍵，因此，我們建造的系統將具備自動選取同音字的功能，而選錯時可由使用者方便地加以更正，如此，就可使輸入速度及盲目打鍵問題獲得舒解。

整體來看，我們的系統是由如圖1裡的幾個功能方塊所組成，其中，鍵盤 aa

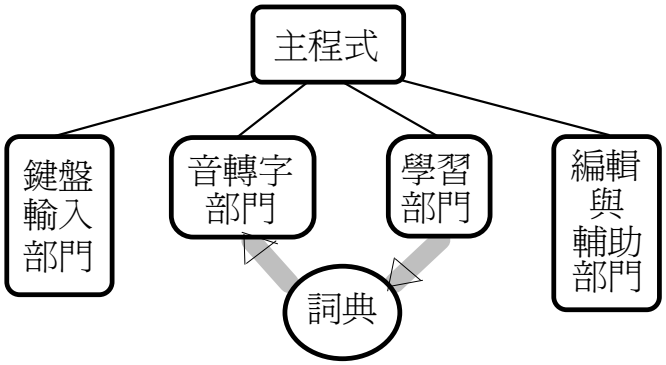

圖1 系統架構

輸入部門負責按鍵之讀取與解釋；音轉字部門，負責將使用者輸入的音節序列轉換成最可能的中文句子；學習部門負責接受使用者的指示去登錄新詞、消掉舊詞、或調整舊詞之出現頻率，此外，它也提供線上近形字群建立的功能；編輯與輔助部門，除了提供文字編輯的功能之外，也能在收到特定按鍵後去呼叫其它輔助功能，如查單一(或全篇)字之注音，查某一字之同音字或近形字等。在以下各節將會對各功能方塊作較詳細之說明。

關於系統之製作，我們採用 Watcom C/386 Ver. 9.01 之 C 語言編譯器，它支援個人電腦內延伸記憶體(extended memory)之使用，不受 640 Kbytes 的限制。當執行所建造的系統後，大約會佔據 1.55 Mbytes 的記憶體(其中 1.15 Mbytes 是延伸記憶體)，此時，螢幕上會顯示出如圖2的畫面，畫面上方的"A"視窗，我們稱它爲文字區，用以顯示已輸入(由檔案或鍵盤)的文字，並且編輯的動作也在本視窗進行；下面的"C"視窗稱爲緩衝區，是使用宜韻注音鍵盤來輸入中文字的視窗，使用者可隨時隨意在"A"、"C"兩視窗之間作切換；另外，"B"視窗是系統用以顯示警告、動作完成等訊息的，而"D"視窗是用以顯示使用者所查詢的同音字、詞，及近形字群等，這是指在緩衝區內所進行的查詢，如果是在文字區進行查詢時，則以機動調整位置之視窗來顯示。所以本系統已將中文輸入有關的編輯、查詢功能整合進來，再配合視窗式的使用者介面，以使中文輸入之操作更方便。

圖2　螢幕畫面

二、鍵盤輸入部門

這個部門負責處理使用者由鍵盤輸入的按鍵，它內部有兩個處裡模式，其中一個處理模式(於圖2之文字區時生效)將鍵盤當成是平常之英文鍵盤，對所獲得之輸入碼不另外作解釋，因此，當我們的系統是在進入某一個基礎中文系統(如倚天中文系統)後再執行時，就可利用此模式來輸入英文或憑藉基礎中文系統所提供的中文輸入方法來輸入中文字。然而，當被切換至另一個處理模式(於圖2之緩衝區)時，此部門就將鍵盤當成是宜韻注音鍵盤[10]，並對所獲得之輸入碼加以解釋，以轉換成宜韻鍵盤上對應之注音符號。

在本系統裡實現(寫作驅動軟體)之宜韻注音鍵盤，它是由本文第一作者先前所設計的，同時考慮了三項鍵盤設計的準則，即鍵盤效率，人體工學原則，及符號至按鍵對應的規律性。鍵盤效率指的是輸入一個音節的平均按鍵次數；而人體工學原則是要盡量減少手指頭的運動量，以避免疲勞；至於符鍵對應之規律性，是要讓使用者很輕鬆地建立符號和按鍵位置的聯想對應關係。在這三個準則需同時考慮的條件下，宜韻鍵盤可說是一個相當不錯的設計。

宜韻注音鍵盤把一個實體鍵盤當成兩個虛擬鍵盤來使用，分別稱爲聲母鍵

盤與韻調鍵盤，聲母鍵盤如圖3所示，是依一定規則將國語音節裡用到的聲母放上去，而韻調鍵盤如圖4所示，也是依一定規則將國語音節用到的韻母及聲

aa

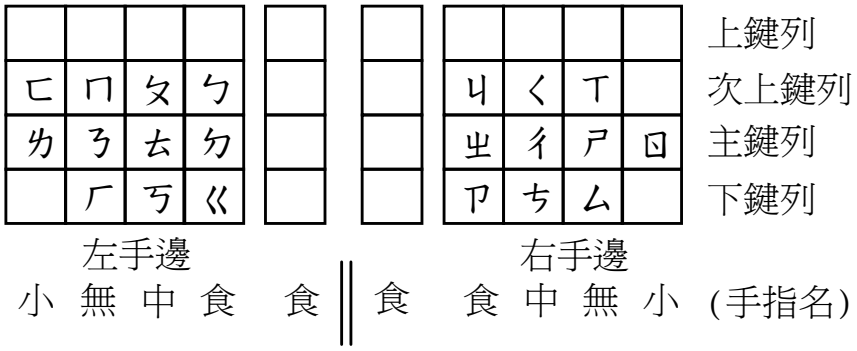

圖3　聲母鍵盤

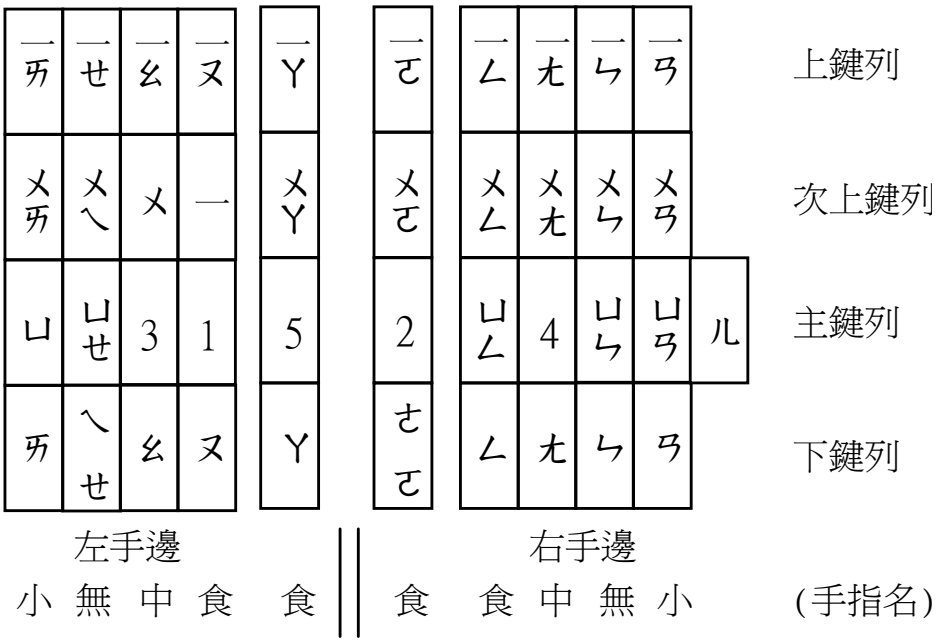

圖4　韻調鍵盤

調佈放上去，這裡所說的規則是，在圖3裡，各鍵列上的聲母都是由中間向外邊排列開來，並且左右手兩邊都是由上而下排下來，而在圖4裡，帶介音"ㄧ"的韻母都排在最上列，帶介音"ㄨ"的韻母都排在次上列，帶介音"ㄩ"的韻母都排在主鍵列，並且，各列是按照ㄧ、ㄨ、ㄩ的順序由上而下排下來，其它不帶介音的韻母大多排在最下列，左手四手指由外而內分別負責打以ㄞ、ㄟ、ㄠ、

ㄡ爲韻尾的韻母，右手四手指由外而內分別負責打以ㄢ、ㄣ、ㄤ、ㄥ爲韻尾之韻母。前述之排列規則，看似簡單，實際上是經過精心的設計，所以在實行上才不會產生困難，即機器可追蹤使用者按鍵的次序來確定他要輸入那一個注音成分(聲母，韻母，或聲調)，並且使用者不需按特殊鍵去作鍵盤面的切換、或告知系統已輸入一個音節了(即可連續輸入一序列的音節)，至於空韻音節(如/ㄓˇ/)之輸入，所採取的補救辦法是將這類音節的聲調按鍵從主鍵列提到上鍵列，因爲和空韻連接的聲母不會和帶有介音/ㄧ/的韻母連接。如此，輸入一個音節只需按2至3鍵，而由統計分析得知，輸入一個音節平均需按2.76次按鍵[10]，比一般注音鍵盤之3.12次按鍵少了0.36次，這意謂每輸入1000個音節可少按360次按鍵，也就節省了相當多的時間及力氣。

另外，宜韻鍵盤所遵循的兩個人體工學原則是，各鍵列的打鍵負擔分佈中，主鍵列的百分比應該是最大的，而離主鍵列最遠的上鍵列，它的百分比應該最小；各手指的打鍵負擔分佈中，食指的百分比應該比中指的大，中指的應比無名指的大，而無名指的應比小指的大，這是因爲各手指的靈活程度有差別，其中食指最靈活。

三、音轉字部門

以注音鍵盤來輸入中文的輸入方式，早期需由使用者在輸入每一個音節後，立刻從螢幕上顯示出的同音字中去挑選一個所要的字，這樣，眼睛需交替盯著螢幕(以選字)及鍵盤(以輸入注音符號)，自然使輸入速度快不起來。因此，本系統的音轉字部門提供了自動音轉字的功能，以便讓使用者能夠一口氣把一句話的音節注音整批輸入進去，事實上我們的音轉字部門是採取動態規劃演算法來作漸進的的處理[11,12]，如此，使用者每輸入一個音節後，音轉字部門就能夠立刻把目前最可能被對應的中文句子顯示出來，不過，還是希望使用者在輸入一個句子後才去檢查有無挑到非想要的字，然後再加以更正(這樣

的機會一般說來低於10%)，如此就可讓輸入速度提高許多。

(1)、複合馬可夫模型

目前被提出來解決自動音轉字問題的方法，至少有直接查詞典法[13]、統計法[14,15]、語法剖析法[16]等，而我們所提出的方法可算是一種統計作法，其想法是要把詞頻資訊和相鄰字間的相關性資訊加以組合運用，正式說來就是把以詞爲狀態之零階馬可夫模型(簡稱 MW0 模型)和以字爲狀態之一階馬可夫模型(簡稱 MC1 模型)組合成一個複合馬可夫語言模型，希望藉以提高自動音轉字的正確率。根據我們過去所做的研究[11,12]得知，以詞爲狀態之馬可夫模型(零階或一階)，其轉換率比以字爲狀態之馬可夫模型較高且穩定，依據這樣的觀察，原本應選擇採用以詞爲狀態之一階馬可夫模型，但是考慮到我們的系統將提供線上新舊詞學習的功能，其中一項是允許使用者增加新詞到詞典去，需要加入一個新詞是預期它將經常被用到，但此時有關這個新詞與其它舊詞連接時的相關性資料系統裡卻沒有，而使得MW1模型不會選到此新詞，例如增加新詞"漢堡"到詞典去以後，MW1 模型對條件機率 P(漢堡|吃) 的估計值仍然是很小(如果訓練語料裡，"吃"與"漢堡"沒有連在一起出現過的話)，因此，我們只得採用 MW0 模型，以使新增的詞有較大的影響力。接著，爲了使句子裡詞與詞連接的相關性資訊也能用於作音轉字決擇時的參考，且要免除新詞造成的困難，所以，我們就提出了一個變通的作法來估計某二詞相鄰的可能機率，其實就是用前一詞之詞尾字與後一詞之詞頭字相鄰的機率來取代，而這樣的機率值可以 MC1 模型來估計。

令一個中文句子是由 $W_1,W_2, ..., W_n$ 等詞語所串接而成的，並且C_i 代表W_i之詞頭字，D_i 代表W_i 之詞尾字，則我們提出的複合式模型將以下式

$$P(W_1,W_2,...,W_n) = P(W_1)\,[P(C_2|D_1)]^u\,P(W_2)\,[P(C_3|D_2)]^u \ldots [P(C_n|D_{n-1})]^u\,P(W_n)$$
$$= [P(W_1)P(W_2)...P(W_n)] * [P(C_2|D_1)P(C_3|D_2)...P(C_n|D_{n-1})]^u \quad (1)$$

來估計此句子之出現機率，其中$P(W_i)$表示詞語W_i之出現機率，$P(C_i \mid D_{i-1})$表示D_{i-1}後緊跟著C_i 之條件機率，而 u 則是加權常數，其數值要由實驗來決定。如果只是以 MW0 模型來估計，則計算式子應爲

$$P(W_1,W_2,...,W_n) = P(W_1)\,P(W_2)\,...\,P(W_n) \quad (2)$$

另外，如果只是以 MC1 模型來估計，且句子是由 $V_1,V_2,...,V_m$ 等中文字所串成，則計算式子爲

$$P(V_1,V_2,...,V_m) = P(V_1)\,P(V_2|V_1)\,P(V_3|V_2)\,...\,P(V_n|V_{n-1}) \quad (3)$$

前面的式子(2)與(3)是假設中文具有馬可夫特性而推導得到[11,17]，至於式子裡的條件機率項 P(Y|X)，並不是要在模型訓練時就將所有可能的 X、Y 組合之 P(Y|X) 值算出來，而事實上也不可能，因爲可能的組合多於 10,000*10,000 ，並且即使收集了非常大量的語料後，仍然會有許多組合不曾看過，這就是所謂的零出現頻率問題，對於這樣的問題已有一些專家提出了不錯的解決方法[18,19]，使得式子(1)至(3)在實做上不成問題。所以，P(Y|X)之估計值是等到要用時才去計算的，而我們修正過再採用之公式是：

$$\begin{aligned} P(Y|X) &= (1-Pe) * N(X,Y) / N(X), \quad \text{if } N(X,Y) > 0 \\ &= Pe * (N(Y)+1) / (Nt+10000), \quad \text{if } N(X,Y) = 0 \end{aligned} \quad (4)$$

$$Pe = (Ns(X)+1) / (N(X)+2) \quad (5)$$

其中 N(X,Y) 表示在訓練語料中 Y 緊跟著 X 出現的次數，N(Z)表示 Z 在訓練語料中出現的次數，Ns(X)表示具有 N(X,Y)=1 之不同的 Y 的個數，Nt表示訓練語料的總字數，而Pe則表示逃脫機率。另外，關於 P(X) 項之數值，本研究是以 X 之字頻(X代表字時）或詞頻(X代表詞時）加上 1 再除以一個常數值 1,000,000 來估計。

(2)、以動態規劃找最大機率之句子

當使用者輸入一個句子的音節序列後，音轉字部門首先要依據音節序列去查詞典，將所有可能被對應到的詞語找出來，例如輸入 /ㄗㄥ/、/ㄐㄧㄣˋ/、/ㄐㄧˋ/、/ㄧˋ/、/ㄌㄧˋ/ 等五個音節後，單字詞之 {增、曾、...，進、近、...，記、計、...，...}，雙字詞之 {增進、禁忌、技藝、記憶、毅力}，以及三字詞之 {記憶力} 都會被找出來。然後，音轉字部門必須在這些詞語的所有可能之連結路徑中去找出一條具有最大機率的路徑(稱爲最佳路徑)，一條路徑代表一個可能出現的句子，例如圖5裡，可連結出來的路徑包aaaa

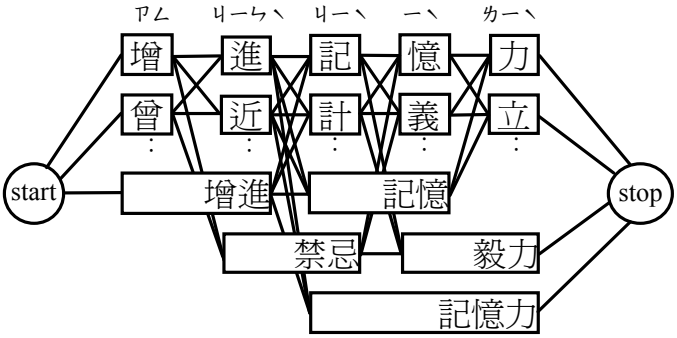

圖5 可能連結之路徑

括：(增)(進)(記)(憶)(力)、(增進)(記憶力)、(增)(禁忌)(毅力)、...等。一般說來，存在的路徑的數目是隨著音節個數成指數成長的，因此，有一些啓發式搜尋方法爲了做到即時處理，只能找到次佳的路徑，不過，本研究採用我們過去提出的一種動態規劃(dynamic programming)演算法[11,12]去作搜尋，此法不但將時間複雜度降到多項式時間(也就能做到即時處理)，也保證會找出最佳路徑，並且它還允許作漸進式的處理，也就是每輸入一個音節後，就可作一部份的處理(包括查詞典及找當時的最佳路徑)，因爲我們可把不同長度的詞以右邊切齊來當作動態規劃[20]裡的處理階段(stage)，如圖5裡，輸入/ㄗㄥ/後查詞典得到{增、曾、...}等詞，就當作動態規劃裡的第一個階段，再輸入/ㄐㄧㄣˋ/後查詞典得雙字詞{增進}及單字詞{進、近、...}，就當作動態規劃裡的第二個階段，其餘的可由此類推。下一段將敘述動態規劃處理之較詳細公式。

令 $S_1,S_2, ..., S_t$ 代表使用者已輸入的音節序列，W_{tk} 代表由k個音節 $S_{t-k+1}, S_{t-k+2}, ..., S_t$ 去查詞典所得到的詞語的集合，W_{tkj} 代表 W_{tk} 的第j個元素，接著令 $Q(t,k,j)$ 表示從起始點到詞語 W_{tkj} 爲終點之間的一條最佳路徑的機率值，而所謂的最佳路徑是以我們提出的複合馬可夫模型來評定的，另外，K 表示詞典裡最長詞的詞長，則 $Q(t,k,j)$ 可以如下之遞迴公式

$$Q(t,k,j) = P(W_{tkj}) * \left[\underset{h=1}{\overset{K}{MAX}} \; \underset{i=1}{\overset{|W_{t-k,h}|}{MAX}} Q(t-k,h,i) * [P(C_{tkj}|D_{t-k,h,i})]^u \right] \quad (6)$$

去求值，公式(6)裡，$|W_{tk}|$ 表示集合 W_{tk} 的元素個數，C_{tkj} 代表 W_{tkj} 之詞頭字，而 D_{tkj} 代表 W_{tkj} 之詞尾字。接著，我們可依如下公式

$$Q(t) = \underset{h=1}{\overset{K}{MAX}} \; \underset{i=1}{\overset{|W_{t,h}|}{MAX}} Q(t,h,i) \quad (7)$$

去找出具有最大機率值 $Q(t)$ 之終點詞，然後由此詞循最佳路徑回溯(backtrack)回去到起始點，去將音節序列 $S_1,S_2, ..., S_t$ 最可能對應的中文句子找出來。

四、學習部門

由於系統裡原有之詞典不可能將各行各業會用到的詞語都收錄進來，並且使用者才最清礎那一個字、詞會經常出現於要輸入的文章中，因此，本部門提供了線上詞彙學習功能，以讓使用者將新的詞彙插入系統之詞典，也可作消除舊詞的處理，此外，不管新詞或舊詞，也都可學習其詞頻和在同音詞中的排名次序，與一般以注音輸入中文的系統，只能學習新詞(不含詞頻)的情況有很大的不同。如果能夠修改一個詞的詞頻，則使用者可以控制音轉字部門之動態規劃處理的選字動作，例如可把某一不想要的同音詞之詞頻降低很多，而讓本次要輸入的文章的一個常用到的同音詞之詞頻增加，如此，就可減少音轉字後的錯誤字。

此外，本部門也提供了近形字群線上建立的功能，當建立近形字群後，就

可使用近形字查詢的功能。近形字的觀念是要來解決一些不常用的字，因不會唸而無法輸入的問題，也就是說，如果我們將一些部首不同而主體部分相同的字(如：者、著、賭、堵、都、諸、睹… 等)建立出一個近形字群的關係放在詞典裡，則當要輸入"睹"但不會唸時，可先輸入"者"或同一群裡的其它近形字，再使用近形字查詢功能，將"者"或其它近形字換成"睹"，這種操作程序類似於同音字查詢的操作，用於更改自動音轉字時被轉換錯誤的字。近形字之觀念，我們未曾在其它以注音輸入中文的系統中看過，因此，這樣的觀念算是本系統一項重要特色，它對採用注音之中文輸入方法的推廣有很大的幫助，因爲許多潛在的使用者(如中小學生)當他不知道所要輸入字的注音時，就可先輸入所要輸入字的形狀相似字，然後按近形字查詢的按鍵，就可由此字的近形字群中，去選取他所要輸入的中文字。

五、編輯與輔助部門

本部門提供了一些文字編輯功能及其它的輔助功能，以便讓使用者在輸入一段或一篇文章後(或任何時候)，都可暫停輸入，而去對先前所輸入之文字作編輯或有關的處理動作。先前輸入的文句顯示於螢幕上的文字區，不過，使用者也可在緩衝區操作部份的編輯與輔助功能。本部門所提供的編輯功能包含鍵盤上之編輯按鍵所代表的之外，還有區塊(block)之設定、複製、與消除，以及讀取、寫出文字資料檔(可選擇是否要存注音資料)等。由於存檔、讀檔功能，可把文字對應的注音資料也一起寫出及讀入，所以可等到下一次進入本系統時，再來操作同音字查詢功能去改正上次輸入時轉換錯的字，如此，校對者(可以不是原輸入者)操作起來就方便多了。

除了一般的編輯功能之外，本部門也提供了一些很有用的輔助功能，亦即它能夠在收到特定的按鍵後去呼叫：同音字、詞查詢之功能(依據系統內保存的注音資料，若無注音資料時會要求使用者輸入)；近形字查詢的功能；自動

查詞典以設定全篇中文字之注音的功能(即自動字轉音，可用以設定由其它軟體輸入的文字的注音)；全篇字的注音資料轉換成Big-5碼並存到檔案去；全篇自動音轉字的功能；以及其它的功能。所以，本部門已將文句編輯功能與中文輸入有關之輔助功能整合在一起，使得使用者可隨意變換去螢幕上的緩衝區進行輸入，或者去文字區編輯文句與執行輔助功能，這種操作上的方便性，一般的中文輸入系統並沒有提供。

六、音轉字實驗

當應用馬可夫模型來解決自動音轉字的問題時，在實做的考慮下，可被選用的具體模型包括MW0(以詞爲處理單位，但不作預測)、MW1(以前一個詞來預測下一詞)、MC1(以前一個字來預測下一字)、或MC2(以前二個字來預測下一字)等。可是，再考慮本系統提供的線上新詞學習功能時，MW1、MC1、與MC2模型的應用就碰到困難了，因此，我們才提出一種稱爲 MW0/MC1 之複合模型，來把原先的MW0與MC1模型加以結合運用。

這一節我們就拿 MW0、MC1、及 MW0/MC1 模型來進行自動音轉字之實驗，看所提出模型的音轉字正確率，和基本模型MW0與MC1的正確率有無差別，爲了使情況單純，在實驗進行中並不做線上新詞學習的動作。實驗所用的測試文章分成兩組，一組取自於晚報社論，共6,291個音節，節錄出來的兩篇如附錄一所示；另外一組則取自於小學生的作文[21]，共9,118個音節，節錄出來的兩篇如附錄二所示，這兩組測試文章都未被用於訓練語言模型。系統裡的詞典約有 43,500 個詞，它們的詞頻可用以估計公式(1)(2)(3)裡的機率項P(X)。至於條件機率項 P(Y|X) 的估計，我們使用了國小國語課本之課文(大約 90,200 個中文字)，報紙社論文章(大約 75,000 個中文字)，以及短篇小說、寓言(大約 105,500 個中文字)，來統計出那些有關的計次參數的數值，以便代入公式(4)及(5)去估計所要的條件機率值，對於前述的三種訓練語料，我們細分

成兩種訓練情況，稱爲 train-a與 train-b，train-a 表示只用國語課本課文與報紙社論，而 train-b 表示三種都用。另外，關於公式(1)裡的加權常數 u 的值，經過初步實驗之觀察後，我們設定它的數值爲 1.0。

將測試文章所對應的音節注音送給各個模型去作自動音轉字處理，然後比對轉換結果與原始文章，我們就得到了如表1所示的音轉字正確率值。由此表

aa

表1 數種模型之音轉字正確率

\ 音轉字方法 測試文章 \	MC1 (train-a)	MC1 (train-b)	MW0	MW0/MC1 (train-b)
小學生作文(9,118字)	88.1	89.5	92.4	93.9
晚報社論(6,291字)	85.2	86.4	91.6	92.6

可知，複合馬可夫模型(MW0/MC1)對兩組測試文章各得到 93.9 與 92.6 之正確轉換率，比 MW0 模型的 92.4 與 91.6 高 1.0%以上，也比 MC1 模型的 89.5 與 86.4 高 4.4% 以上，這說明複合馬可夫模型的確可用來改進個別的 MC1 與 MW0 模型。在這三種模型當中，MW0 與 MW0/MC1 模型的轉換率都超過 90%，而 MC1 模型的轉換率最低且未超過 90%，此外，當把訓練語情況從 train-a 變成 train-b 時，MC1 模型的轉換率也才上升約 1.3%，因此我們預期再增加訓練語料後，MC1 模型的轉換率也不會上升很多，所以，在選用馬可夫語言模型來處理中文時，至少可採用 MW0 模型，而在作中文輸入的系統裡，則可考慮採用我們提出的複合馬可夫模型，以提升音轉字之正確率。另外，由表1可發現一個共同現象，即三種模型對小學生作文的轉換率都比晚報社論的高，不過，就轉換率變化的幅度來看，MC1 模型的變化幅度(約 3%)仍是比其它兩個模型的 0.8% 與 1.3% 高許多，所以，MC1 模型較不穩定，而另外二者較不受文章難易、種類之影響。

七、結語

我們在設計、製作本中文輸入系統時，是採取實用的觀點，希望所製作的系統可以讓使用者更方便地去操作使用，所以花了許多時間精力於設計演算法與寫作程式，以提供大大小小之各種功能。雖然我們的系統必需進入一個基礎中文系統(如倚天中文系統)後才可操作使用，但是我們的系統和基礎中文系統是相輔相成而不會衝突的，即可用我們的系統來輸入中文外，也可切換使用基礎中文系統所提供的中文輸入方法。

本系統採用之宜韻注音鍵盤，考慮了三項鍵盤設計的重要準則，即鍵盤效率，人體工學原則，及符號至按鍵對應的規律性，其中，鍵盤效率所指的是輸入一個音節的平均按鍵次數；而人體工學原則是要儘量減少手指頭的運動量，以避免疲勞；至於符鍵對應規律性的目標是，讓使用者很輕鬆地在建立符號和按鍵位置的聯想對應關係。

關於自動音轉字的處理，本系統採用了新提出的複合馬可夫語言模型的作法，這樣的模型除了可支援線上新詞學習的功能外，也兼顧了句子裡相連兩詞間的相關性。測試實驗的結果顯示，複合馬可夫模型的確可改進個別之 MC1 與 MW0 模型的音轉字正確率。

本系統也提供了近形字群線上建立的功能，以及近形字查詢的功能，這是考慮許多潛在的使用者(如中小學生)可能不知道所要輸入字的注音，這時他就可先輸入所要輸入字的形狀相似字，然後透過近形字查詢的功能，去選取他所要輸入的中文字，所以，近形字的觀念非常有助於以注音輸入中文之系統的推廣。

參考文獻

[1] Microsoft Corporation, Microsoft Windows 中文版，1992。
[2] 倚天資訊有限公司，倚天神雕筆手寫辨識系統使用手冊，1992年。
[3] 蒙恬科技有限公司，蒙恬中國筆使用手冊，1991年10月。

[4] 范欽雄、李豐壽，「應用類神經網路辨認常用中文字5401字」，全國計算機會議論文集(嘉義)，第619-627頁，1993。
[5] 黃雅軒等，「印刷體光學中文字形辨識系統」，電子發展月刊，第141期，第16-26頁，1989年9月。
[6] Lee, Lin-shan, chiu-yu Tseng, Hung-yan Gu, *et al.*, "Golden Mandarin(I) -- A Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary", IEEE Trans. Speech and Audio Processing, pp. 158-179, 1993.
[7] 倚天資訊有限公司，倚天中文系統使用手冊，1992年5月。
[8] 松下電器開發有限公司，漢音詞彙輸入法使用手冊，1991年12月。
[9] 長諾資訊圖書股份有限公司，國音輸入法，1993年5月。
[10] 古鴻炎，「一個同時考慮鍵盤效率人體工學原則及符鍵對應規律性之國語注音輸入鍵盤的設計」，電工雙月刊，第35卷，第2期，第123-132頁，1992年4月。
[11] Gu, H. Y., A Study on a few Relevant Problems about Machine Dictation of Mandarin Speech, Ph. D. Dissertation, Department of CSIE, National Taiwan University, Jan. 1990.
[12] Gu, H. Y., C. Y. Tseng and L. S. Lee, "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters", Computer Speech and Language, Vol. 5, No. 4, pp. 363-377, 1991.
[13] Kuo, J. J., J. H. Jou, M. S. Hsieh, and F. Maehara, "The Development of New Chinese Input Method -- Chinese Word-string Input System", Proceedings of International Computer Symposium (Tainan, Taiwan), pp. 1470-1479, 1986.
[14] Lin, M. Y. and W. H. Tsai, "Removing the Ambiguity of Phonetic Chinese Input by the Relaxation Technique", Computer Processing of Chinese and Oriental Lanuages, pp. 1-24, 1987.
[15] 季震寰，結合詞與統計的注音中文輸入系統，國立台灣大學資訊工程系碩士論文，1991年7月。
[16] Hsieh, M. L., T. T. Lo and C. H. Lin, "Grammactical Approach to Converting Phonetic Symbols into Characters"， Proceedings of National Computer Symposium (Taipei)， pp. 453-461，1989.
[17] Ross, S. M., Introduction to Probability Models, third edition, Academic Press, Inc., 1985.
[18] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE trans. Acoust., Speech, and Signal Processing, pp. 400-401, March 1987.
[19] Witten, I. H. and T. C. Bell, "The Zero-Frequency Problem: Eetimating the Probabilities of Novel Events in Adaptive Text Compression", IEEE trans. Information Theory, Vol. 37, pp. 1085-1094, 1991.
[20] Horowitz, E. and S. Sahni, Fundamentals of Computer Algorithm, Computer Science Press, Inc., 1978.
[21] 鄭博真編著，小學生作文寶典習作篇，小叮噹圖書公司，1993年六月。