應用於"音中仙"國語聽寫機之短語規則分析與建立

葉瑞峰 王駿發 許志興

e-mail address: wangjf@server2.iie.ncku.edu.tw

國立成功大學資訊工程研究所

摘 要

簡便而好用的人機介面是資訊研究的一個重要課題，鍵盤乃是針對歐美拼音文字而設計的，對於中文這種方塊文字，若非受過專業輸入訓練是很難普遍地利用鍵盤來做中文輸入，所以發展國語聽寫機技術對資訊中文化是有十分重大的影響。

而在聽寫機方面的研究已行之多年，在國外已完成的系統有Hearsay II、Harpy、BBN、TINA及Dragon 等系統，在國內則有台大與中研院聯合開發的"金聲系列"[1][2]以及成功大學所發展的"音中仙巨量詞彙輸入法"。

本文即是對於語音辨認後處理之自然語言處理提出方法，使中文的語音輸入技術在理論上及實用上都能兼顧的考量下發展，在本文中短語分析規則主要有兩類：

1.單一詞未知短語規則：指短語規則中，有一未知詞而其它詞已知稱之。例如"姓氏+校長"為一短語規則其中校長為已知，姓氏為未知。此類法則乃針對詞庫中未建之詞必需加以簡單組合之詞，利用大量語料庫做統計，再依據統計的輸出做為辨認系統構詞的法則權重，以解決斷詞含混與詞庫不足的問題。

2.多詞未知短語規則：指短語規則中有多詞未知，例如"某某市某某路某某巷"其中市、路、巷為已知，但縣市名稱、路名及巷名三個未知。此類法則所處理的主要對象是數量詞和住址或複合詞可以用狀態轉移表示的詞組。

一、 國語聽寫機概說

在本節中我們分別就訊號方面的辨認特性以及在中文語言特性的研究來討論。

語音在辨認上之特性

由於語音是一種高時變性的訊號，所以在辨認時會造成若干誤差。此外外在環境與輸入設備也可能對語音訊號造成相當程度的影響，我們就以語音透過電話線為例來測試語音訊號的穩定性。發現語音訊號中的穩定資訊是集中於母音部份，而子音部份不但在辨認上的特徵不穩定，同時也容易受到環境和輸出入介面的影響。

為了測試母音跟子音在辨認上的穩定性，我們利用電話的通道效應[25]來對語音作處理，之所以選擇電話作為干擾的原因主要是因為通過變動性大的電話除了背景噪音外，在頻譜上若干對應的增益衰減，此外每次不同的電話連接都會有不同的效應產生，所以是屬於比較客觀的測試環境。

首先我們利用麥克風輸入408個音節三遍，每遍盡量用不同的聲調（四聲變化），然後再經由電話線來測試，我們一樣透過電話線來唸一千多個單音節，不過我們是透過四個不同的電話機，六次不同的電話連接來測試的。

子音	麥克風輸入		電話線輸入	
	單項	累積	單項	累積
第一名	59.53	59.53	25.53	25.53
第二名	19.04	78.57	12.76	38.29
第三名	10.71	89.28	17.02	55.31
第四名	4.76	94.04	8.51	63.82
第五名	3.57	97.61	5.50	69.01

表 1 子音透過麥克風及電話線之辨識率

母音	麥克風輸入		電話線輸入	
	單項	累積	單項	累積
第一名	84.52	84.52	51.06	51.06
第二名	10.71	95.23	10.63	61.69
第三名	2.38	97.71	6.38	68.07
第四名	1.10	98.81	8.51	76.58
第五名	1.00	99.81	6.41	82.99

表 2 母音在透過麥克風及電話線之辨認率

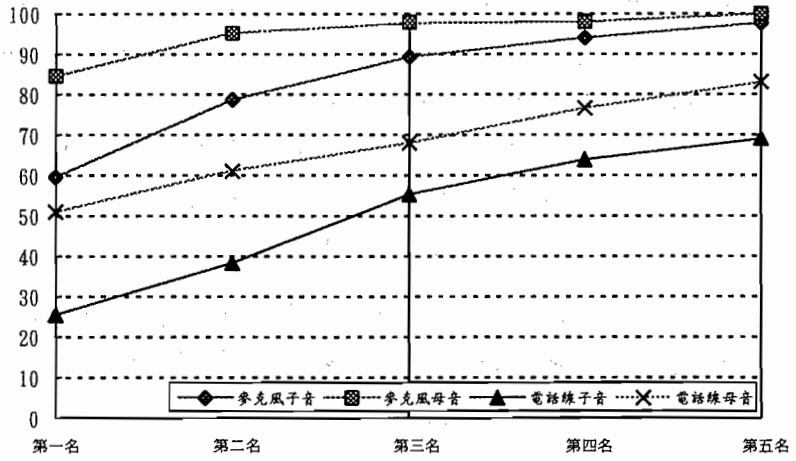

圖 1 透過麥克風及電話線後，子母音辨認率之比較

由以上實驗可知：母音在辨識上是十分穩定的，在未透過電話線的時候子母音的辨識率在前五名都達到九成五以上，可是在透過電話線通道效應干擾後，母音前五名累積辨識率仍能保持在八成以上，可是子音卻已經滑落至七成左右。所以在辨識上利用母音的穩定性是可以提高系統的包容性的。

國語的特性

中文語音是一字一音節，同音異形字[14]的字集相當多。此外由於中文語音存在有若干混淆集，如{八，搭，他，嘎，喀，....}，還有因發音習慣的不同所產生的音變現象 [13]如『倒塌』『他們』中『塌』和『他』皆是標示ㄊㄚ，但在台灣地區有相當比例的人唸出的音卻是不同的。

語言的最小基本語法單位都是詞，不過中文的詞在文句中並沒有像拼音文字一般以空格隔開，所以中文的自然語言處理便多了斷詞、配詞這個動作了。而且中文詞並沒有在字形或字音上表現出詞類變化的現象，所以很難直接從字形或字音得到所謂的詞類，所以中文詞性要絕對標示成為一件極為困難的事。

所以不論是國語語音辨認或是自然語言處理包容性(Robustness）都是十分重要的考量。

目前兩種實現國語聽寫機的方法

在目前關於國語聽寫機之操作模式大抵可以依處理之基本單元而分為兩類：

一.以句子為處理單元：由於句子含有最充足的語法訊息，所以有些聽寫機的設計是建構在整句處理的基礎上。以句子為基本輸入單元雖然可以避開『斷詞含混』的現象，但是由於以句子為處理單元，必須負荷比以詞為基本辨認單元更大的計算量，所以必須要有良好的硬體設備配合，而且只有在辨認正確率極高的情況，方能做有效處理，對於聲音較不穩定或者是發音習慣較為特殊者，以及一般無法提供相當硬體設備的情況下，可能都會造成實用上的不便。

二.以詞為處理單元：不論從語言學或心理學的觀點，詞皆是表示意念的最小單位，在自然語言的處理上，也是以詞為最小之語法單元，所以以『詞』為基本辨認單位看來也是十分合理，而且由於詞為處理單位可以容許辨識模組誤差，所以在現階段實用上是比較適當的。可是由於詞的定義一般人總是不能十分明確地定義出來，所以以詞為基本辨認單元的國語聽寫機必須要能克服斷詞含混的問題。

而在本文中即是就以詞為基本處理單元之運作模式加以研究，希望能透過一些方法在不影響系統即時處理以及保有系統實用性的原則下，將處理單元由詞提昇至詞組甚至未來可能由下而上發展至整句處理，以達成中文聽寫機的理想。

以下我們就以句子"在台灣大學生活像白紙"的詞絡（word lattice）來比較整句處理跟將句子作合理的斷句後其可能路徑的多寡，可以見得加入一些斷詞的知識是可以降低複雜度，使系統在處理上速度更快。

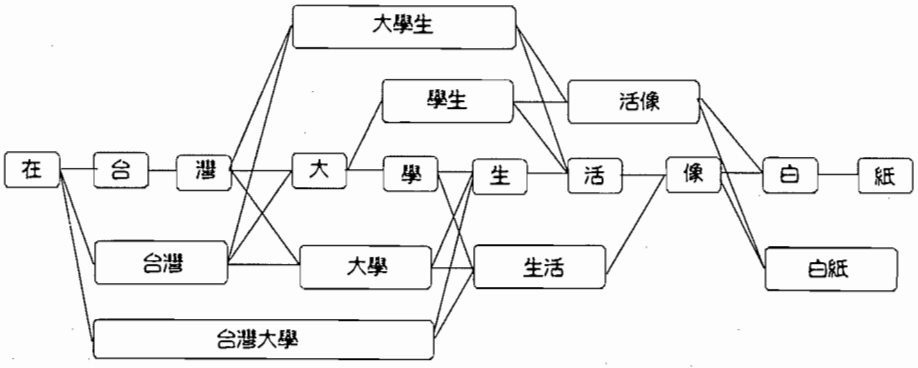

圖 2 以整句處理模式下的詞絡

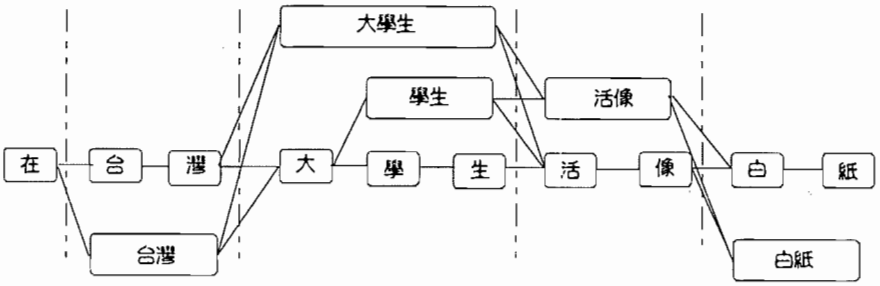

圖 3 存在合理斷點（詞）的詞絡

音中仙系統架構

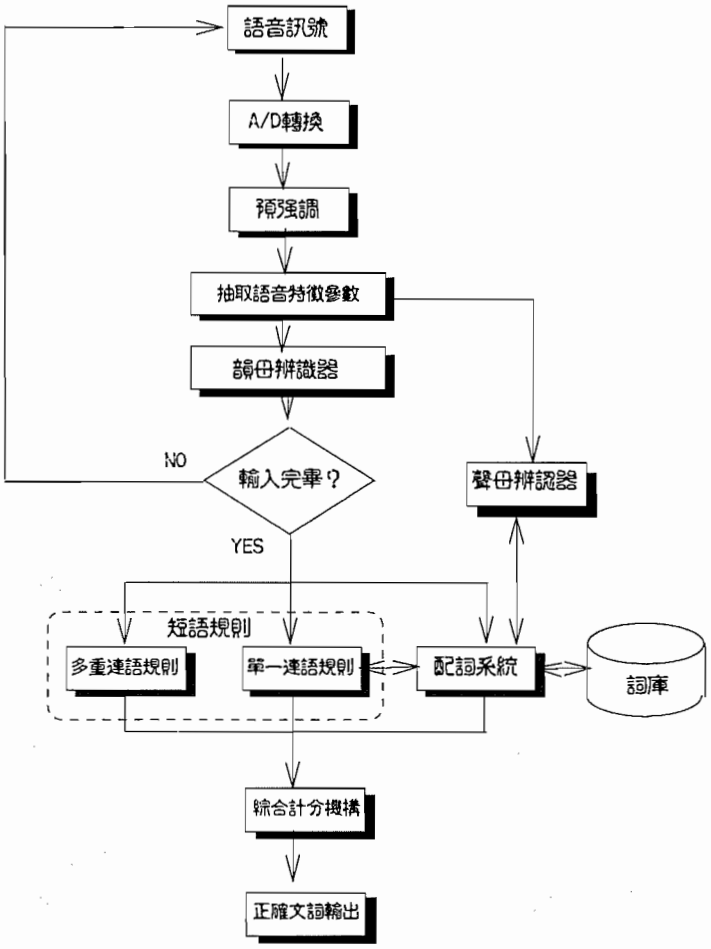

圖 4 音中仙國語聽寫機之系統流程

"音中仙"語音辨識器簡介

音中仙之辨識模組是利用拜氏網路[8]來估算參考樣本及測試樣本間的相異度，此網路是利用混合高斯機率密度之觀念來完成拜氏分類法則，以求得輸入特徵向量與參考類別間的相似機率值，其架構如下：

拜氏網路基本上是以拜氏定理為理論基礎，在架構上可分為輸入層、高斯層、混合層、歸納層。輸入層為待辨識的語音音框的特徵參數，高斯層是由統計訓練樣本的分佈情形所形成，混合層是一種混合的高斯機率分佈，而歸納層的輸出就是把混合層的機率轉換成距離輸出，在根據此一距離輸出來取辨認音節輸出。

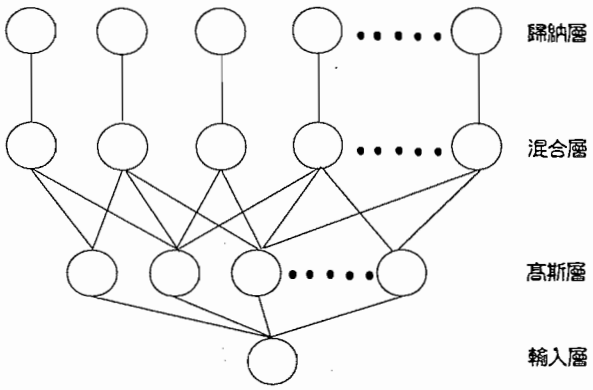

圖 5 拜氏網路

配詞機構

配詞機構的目的是在系統需要從辨識候選音節得到可能對應之文詞時做配詞的工作。由於音中仙具有大量詞彙處理能力的即時系統，所以必須具備快速準確的配詞能力。

在記憶體的考量下，我們不可能將所有整個詞庫載入系統中，所以音中仙的詞庫結構可分為兩部份，在圖中上方是用來配二字詞的結構，下方則是用來配長度三以上之長詞（含三字詞）。雖然在結構上有所不同，但是基本上還是以詞的前兩個音的韻母組合為指標開始搜尋，我們將這樣的指標以串列實現。而在系統執行時載入記憶體中也就只有一些串列跟指標而已，至於詞庫是儲存在磁碟中的。

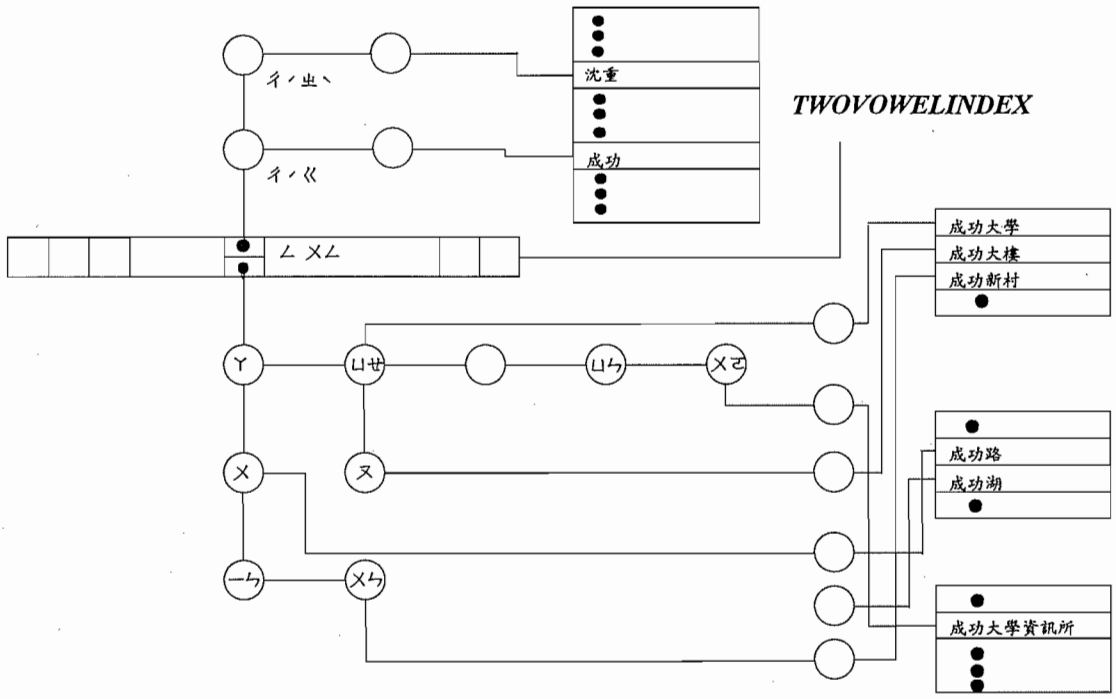

圖 6 音中仙系統之配詞機構

在一字詞語二字詞之辨認上，音中仙也做了一些處理，由於短詞語音所含的語言資訊有限，並且因為混淆音集的存在，我們要發展一套高包容性的系統，便需要對子音混淆集做歸納，而在系統處理時對辨認音節處理外也將其所屬混淆集的音節也一併處理，但是如此一來又可能會產生候選詞過多的窘態，所以對於音調的處理變成了重要的一環，不過音調處理在單字詞還算很不錯，不過在多字詞時由於轉折音的存在及發音習慣的不同，我們在做辨認後處理時也是必須考慮到音韻規則[10]才可以更準確地得到所想要輸入的字詞。關於四聲音韻在辨認上可以歸納如下：

聲調辨認結果	推測可能辨認結果
第一聲	第一聲，第二聲
第二聲	第二聲，第三聲
第三聲	第三聲，第四聲
第四聲	第四聲，第三聲

表 4 音韻處理表

二、 單一詞未知短語規則的分析與建立

短語規則主要是結合統計式[11][12]的文法觀念，利用電腦自動統計的方式產生，也就是在欲處理的規則中給予統計的頻率評分，如此可以便免統計式大量參數資料之不足[13]，也可以隨狀況分析隨時加入新的規則，而不像傳統剖析器一般必須在全盤考量後制訂，制訂後的修改又會往往造成系統維護的困擾。

在制訂規則時，必須考慮到一些問題如辨認模組的候選音過多，或輸入不是十分合法的字詞組合，新詞的產生或舊詞新用，甚至是隨著語言演化所產生之新文法新句型等，所以我們在制訂規則時，必須考慮到其包容性。

在雙連資訊的統計中，發現大部分的參數值都是零，連語現象(Collocation)〔15〕是十分強烈的。在作為語音辨認後處理時，可以利用樣本比對（Pattern matching）〔3〕的方法，逐步由下而上（bottom up）〔13〕地組織，如圖七所示為單一詞未知短語規則建立流程。

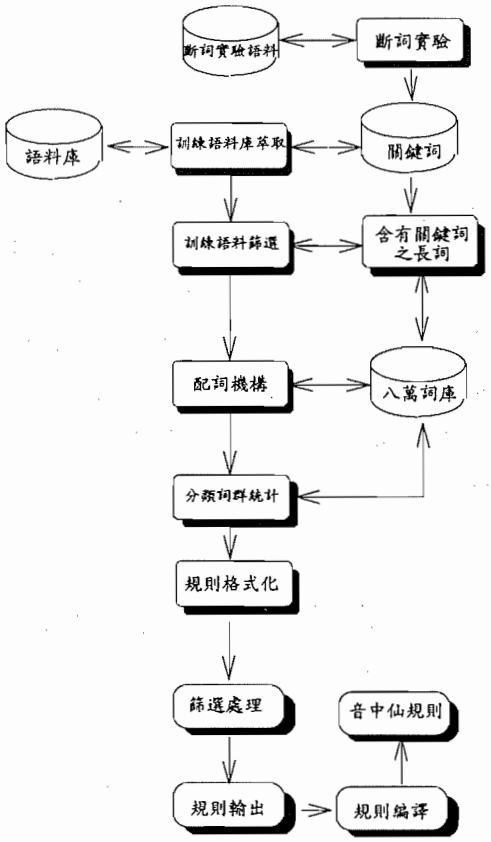

圖 7 單一詞未知短語規則之建立流程

斷詞實驗

誠如前面討論，以詞為辨認單元是可以降低系統複雜度以達成及時處理的程度，但是以詞為基本輸入單元可能因為個人斷詞習慣的不同而導致斷詞岐異的現象[4][6]，此外有些詞的數量是詞庫無法完全處理的如數量詞、複合詞等。針對斷詞歧異我們做了以下實驗：

我們分別就以下的五種文章，選取國小課本、報紙新聞部份（不含廣告）、小說、科學報導雜誌以及討論中國文學的『中國學術通義』的漢語文資料各約六百至八百字，分別讓國高中學生共六人，理工科大學生四人，文科大學生三人，還有研究生三人共十六人作斷詞分析。

以下所得之正確率乃指上述由人斷詞結果在音中仙系統詞庫中直接配詞，可以配得正確詞的漢字數除以所有測試漢字數總和。

	國小課本	報紙新聞	小說	科學雜誌	中國學術通義
正確率	89.76%	83.55%	81.03%	78.67%	79.98%

表 5 斷詞實驗之正確率

在實驗中，發現以詞為輸入單元，有相當之可行性。而發生斷詞錯誤之原因主要有下列幾個原因：

1.中文詞綴的引入，如『股票族』、『未登記』。

2.數量詞：定詞 + 量詞，如『三百五十公噸』。

3.介詞與連接詞：如『受折磨』，『戰爭與和平』。

4.複合詞：如 『國科會主委』。

5.人名：如 『王小明』。

6.古文：如 『猝然臨之而不驚』。

7.翻譯名詞：如 『關貿總協』。

8.外來語：如 『DNA』、『去氧核糖核酸』。

訓練語料庫萃取

根據斷詞實驗所得之結果，將一般斷詞容易發生錯誤的情況，將其關鍵詞訂出。根據所得的關鍵詞去訓練語料庫(計算語言學會提供之語料庫)中將含有關鍵詞的語料萃取出來，在此時我們所採的統計視窗大小為十二，即是在關鍵詞前後各有六個漢字，因為一般在四字詞以上的長詞不容易發生斷詞錯誤的情況。假如我們分析的關鍵詞是『受』這個詞則在語料庫萃取所得的訓練語料如下所示：

後桃園地檢署 受 理本案，關於

隔壁的老婆婆 受 無情的兒子拋

種不孝的行徑 受 大家指責。

訓練語料篩選

由於有些關鍵詞可能會是其他較長詞的一部份如：分析『受』這個字，訓練語料會將含有『感受』的語料也一併加入，所以我們必須將這些狀況排除，在這部份我們在系統詞庫中將含有關鍵詞的詞找出來，在根據上下文判斷是否是屬於我們分析的關鍵詞狀況。

另外也必須檢查關鍵詞位置是否為其他詞之局部所構成，這是補償前面可能的漏失。如關鍵詞為『經過』，而訓練語料為『...已經過去...』，此時我們將以『已經』及『過去』這兩個詞來和關鍵詞『經過』來比較，若相差不多，就以人力判斷，不過在實作的過程這樣的狀況非常少。

配詞

這是根據我們分析的關鍵詞前接詞跟後接辭去詞庫中配詞，這一階段配詞，首先在關鍵詞兩側設立絕對斷點然後對訓練語料其他部份斷詞，而以長詞優先，高頻詞優先的順序作為斷詞依據。

分類詞群

此外我們還必須對詞庫裡的詞作分類，我們以分類詞群[14]做為馬可夫統計模式的基礎主要的原因有二：

一、避免由於詞庫內詞數過多造成統計結果過於龐大。

二、可以藉分類詞群來調整語料庫不夠平衡的弊病。

而分類的標準主要是根據計算語言學會的詞庫之語法標示以及語意標示共分為兩百類。在分類詞群的類別數目上，我們發現若分類太多會造成短語規則過於冗雜，包容性會降低，但是所得的短語規則較為精確。若分類太少，會造成合於規則的輸出過多，容易造成混淆現象。由我們實驗的結果，發現有一些混淆的現象，可見分類數還不夠多。

馬可夫統計模式的引用

在自然語言處理中，馬可夫統計模式已經十分廣泛地被使用。在馬可夫模式中也是根據字與字之間或詞與詞之間的連接機率，在節省記憶體和處理速度的考量下，一般來說，一階馬可夫語言模式也就是雙連關係（bigram）由於在實作上容易，且在語音辨認後處理上能夠維持相當的水準，所已被廣泛接受。

輸入詞串： $W = w_1 w_2 w_3 \ldots\ldots w_n$

W的發生機率為：$P(W) = \prod_{i=1}^{n} P(w_i | w_1 w_2, \ldots. w_{i-1})$

雙連關係 ：$P(W) = \prod_{i=1}^{n} P(w_i | w_{i-1})$

其中在參數的訓練有人用純粹的機率統計模式，也有人用相對資訊（mutual information）來處理，而在本文中我們是用計算連接次數來處理，主要的原因是為了在日後有所調整時，不至於喪失原有的一些資訊。

單一詞未知短語規則之輸出格示

經過上述步驟之處理，系統的輸出格式如表 6 說明包括終止項(terminal term）以及非終止項(nonterminal term)，所謂終止項即是只在系統詞庫中可以配到詞的輸入，而非終止項所稱的就是經由短語規則處理之後可以得到的詞組輸入，也就是說短語規則的作用就是將非終止項轉為終止項的組合。至於混合項所描述的便是可以直接由配詞得到，也可以由多個詞經由短語規則所生成的單元。

【例一】

可愛的綿羊

{@可愛的$$ㄎㄜˇ ㄞˋ ㄉㄜ•〔(2)<53><893>〕〔(3)<53><122>〕}

【例二】

中華民國84年

{@中華民國$$ㄓㄨㄥ ㄏㄨㄚ ㄇㄧㄣ ㄍㄨㄛ〔(1)<0><23>〕〔(2)<0><210>〕〔3〕<0><189>〕#年$$ㄋㄧㄢˊ}

表 6 短語規則中輸出格式之符號定義:

{ }	：規則開始與結束
@ #	：關鍵詞國字部份起始位置
$$	：關鍵詞國字部份結束，注音部份開始
〔 〕	：非關鍵詞之分類詞群區塊
	〔(A)<x,y,z,...><X,Y,Z>〕
	其中 A 代表詞長
	x,y,z,.... 代表分類詞群
	X,Y,Z,.... 代表統計分數
	x,y,z,....與 X,Y,Z,....有相對應之關係

篩選處理

這樣產生出來的規則，是具有統計上的意義。不過系統可能會因規則產生的合理組合太多而導致速度或記憶體上的負擔，所以在這裡我們做了一些篩選動作。篩選的對象即是在規則中所佔分數比重較低者，以及在口語習慣上較不會出現之組合，例如在詞頭詞綴我們在規則中就專對其後接詞統計來做處理，而將其前接詞產生的規則略去。

音中仙短語規則之產生

在前一節所討論的都是屬於靜態單一連語規則的建立來討論，在本節是針對音中仙辨認模組與短語規則結合後的運作，加以觀察討論。

在音中仙國語聽寫機中所用的詞庫，是由計算語言學會提供的八萬詞目詞庫加以整理所得到的。而如此大量的資料，假如規則沒有加以處理，在做搜尋（serching）跟比對（matching）時都會造成系統在速度上嚴重的負擔，在詞庫方面我們是根據詞的音來做索引，而在規則庫方面我們用了一個規則庫編譯器，這個編譯器的作用主要是將我們所制訂的規則編譯成音中仙辨認模組所處理的型態，而這種型態的改變主要的目的便是加速搜尋跟比對。

而這個編譯器運作的原理主要可以分為兩部份說明：

關鍵詞語：在關鍵詞語的處理上主要是根據關鍵詞的注音與字的位置，將類似的規則放置在一起。而且如果關鍵詞語越長的，我們在評分機構中將會得到越高的評分。

非關鍵詞語：在非關鍵詞的比對中，首先我們的短語規則編譯器會把在規則中含有相同位置、相同詞長以及類似的分類詞群之非關鍵詞語的規則集中在一起，在辨認模組運作時，便可以省去大量比對的時間。

實驗結果

我們從中國時報二至四月份的報導中，取得1791個含有介詞之測試樣本，在原有系統中加入以六十四類共一百餘個介詞為關鍵詞的規則（按語言學會提供之八萬詞庫所標示詞性）共一千零七十五條，是否可以合理地組合出我們所要輸入的測試詞組，以下便是利用規則庫組合的辨認率：

	前三名	前五名	前十名
正確率	58.42	64.30	72.33

然後我們討論規則庫的加入對原有系統詞的影響，分別在有使用規則庫與沒有使用規則庫兩種情況來輸入中文的詞，這些測試的詞是由原詞庫中取得，而兩種情況所辨認的音串輸入是屬於同一個聲音，下面是實驗結果：

未加入介詞規則前原有系統詞之累積辨識率：

	第一名	第二名	第三名	第四名	第五名	第六名	第七名	第八名	第九名
三字詞	60.32	76.40	82.40	86.77	90.36	92.04	94.39	96.78	98.02
四字詞	62.12	76.33	82.02	85.52	89.98	91.03	94.00	96.20	97.18

加入介詞規則詞後原有系統詞之累積辨識率：

	第一名	第二名	第三名	第四名	第五名	第六名	第七名	第八名	第九名
三字詞	40.23	52.02	61.32	70.89	73.22	78.32	84.02	87.96	91.08
四字詞	57.08	67.01	75.18	81.56	86.00	87.26	91.06	93.01	94.63

由實驗發現在短語規則處理下，詞組輸出是十分可行的方法雖然有些詞的正確輸出排名會被擠到後面，原因分類詞群的多寡良窳對系統輸出的影響，這也說明了，我們分類詞群的數量還可以再區分為更多類，也就是說令每一類之間的特性更接近了，這樣便可以得到合理輸出又降低系統詞混淆的程度。不過在音很準(即每次皆只取第一名)的時候事實上這些問題都是不存在的，我們曾經就以正確音輸入，發現效果十分良好，不過在考慮實際狀況時，這樣的假設似乎有些嚴苛。

三、 多詞未知短語規則處理

不論是在詞庫或辭典的編定，有一些詞是不可能收錄完全，例如由定詞和量詞所合成的複合詞就具有無限制的常用性[9]。事實上我們可以利用一些法則從原有詞庫的詞來組合生成這些詞，在本節便是針對這一狀況利用狀態轉移的觀念來處理。

數字規則

數字可以分為"零"到"九"十個數字以及"十、百、千、萬..."兩類，而其中的轉移是有一定的規律的，在國語的數字系統中有其獨特的語法結構不同於西歐語言，經由我們觀察分析，中文數字是分為四位一小節，也就是說主要可以分析為兩階段，第一階段為在一萬以下的數字表示法，以及萬、億、兆...等，第二階段的表示法，也就是說我們可以將數字及千百十的組合規則定為一個狀態轉移，而這個狀態轉移將成為第二階段狀態轉移中的一個狀態，我們便可以利用兩階段的狀況轉移來表示中文中數字系統的組合。

而其中『零』會有兩種角色，一為單純數字中的零，一為十、百、千、萬、億、兆的位數省略表示，所以在狀態圖中將會對於這種省略表示的『零』特別給予一種狀態。，而這樣的觀念可以用下列狀態轉移圖來說明：

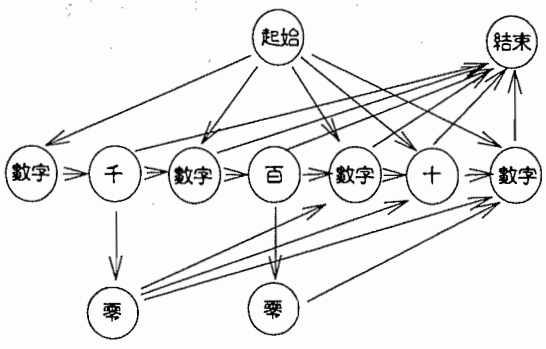

圖 8 第一階段數字狀態轉移圖

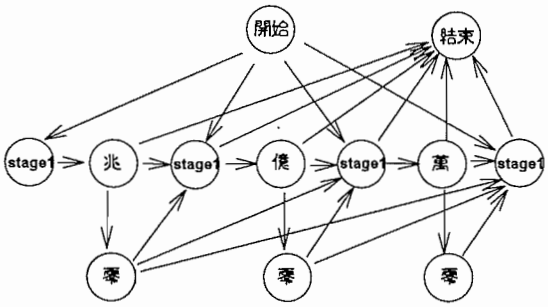

圖 9 第二階段數字狀態轉移圖

在前面所討論的數字系統是利用數字與十、百、千、萬等正規的寫法，事實上在日常生活中也有可能直接由數字來組合例如學號、電話號碼等等。

甚至有些狀況我們也習慣用數字串來取代前面所提的正規數字表示，而單純利用數字串的狀態轉移是十分簡單的如下圖所示：

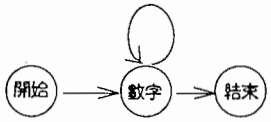

圖 10 數字串之狀態轉移圖

不過若要系統能兼顧正規數字表示以及數字串的表示方式，我們就要考慮到"四"跟"十"的混淆效應，基本上如果在數字表示所得的結果含有"十"之外的"百"、"千"、"萬"那就可以確定為正規數字表示，經由正規數字的狀態轉移便可以得到正確的結果。若在數字表示中找不到"百"、"千"、"萬"，也就是說我們並不確定輸入的是簡單數字串或是正規的數字表示時，我們就必須對四跟十混淆狀況討論。因為在四字以上的數字串我們可以明確地決定是屬於哪一種表示方式，所以首先我們在三個數字串的考量在三個數字串的合法組合中，如果在第二個數字產生混淆時，我們就依辨認音節給分，因為事實上"八十一"跟"八四一"都是合理的數字表示，如果"四"、"十"混淆不是產生在第二個數字也就是數字系統中的十位數，那我們就可以確定是四。而在二位數中如果不是十十的組合，基本上我們都認為是合理的，而一位數字更不用說四跟十都是合理的數字。而誠如前面所討論的作為語音辨認的後處理我們必須考慮到系統的即時處理能力，在本章中我們關於這類可以無限組合的詞組處理，我們採用最先最佳（first best）演算法來節省運算量，也就是我們從辨認模組中得到候選音節排名，根據該排名我們將音節應對至字，然後我們取屬於狀態的關鍵字前五名，之所以取前五名是因為在我們的實驗中，每個音取前五個字已經是綽綽有餘了，然後由每個音的第一名的字去組合，並進入狀態轉移去看看是否可以得到

合理的組合，如果合理便是我們要找的輸出（first best），而其他的組合因為速度的考量就不再進入狀態轉移中去運作了。如果在最佳路徑中我們無法得到合理輸出，我們會根據辨認音節權重去找第二條次佳運作的路徑，如此重複運作一直到找到一條合理輸出或是已經將關鍵字組合的路徑用完為止。不過若是屬於前面所提的"四"、"十"混淆，我們還是允許在輸出中可以得到兩條以上的輸出，因為事實上不僅訊號上類似，即使是人類也會常常誤判四跟十，所以再此我們特別將這一類混淆考慮進去。至於一般數字中常會混的"一七","六九"事實上由辨認模組輸出的加權便可解決，也就是說這樣的混淆在我們的辨認模組便以得到解決。

地址規則

在日常生活中地址是一項常用的資訊，舉凡個人資料的填寫、郵政金融等行政手續上，地址都是不可或缺的資訊，而數字更是一般常用的輸出入 項目。所以在這裡我們提出一套方法針對地址來處理。

在地址規則的制訂上，我們有必要瞭解在台灣地區的地址系統中可以區分為兩類，一類是必須項，如號；有一些是非必須項，如段、巷、弄以及幾號之幾。在非必須項中也有一定的關係，例如『弄』的出現，前面一定有『巷』，所以也是可以利用狀態轉移來處理的一種情況。

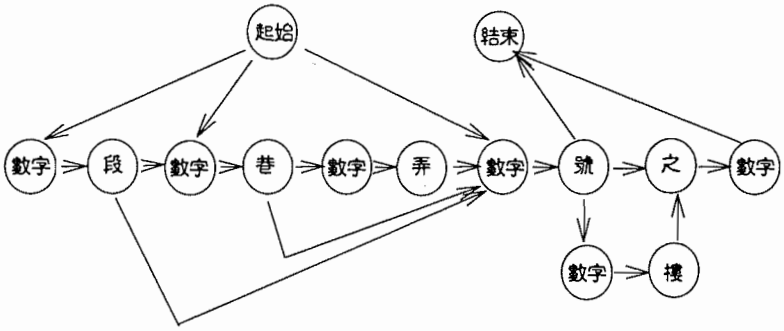

圖 11 地址狀態轉移圖

網狀規則

從人類的的斷詞經驗看來，有些句子可能會有多種合理斷詞，而這些合理斷詞中，一般而言是不會將一個詞切開，而是將幾個詞組合在一起。所以為了使我們的規則更具生命力，更具包容性，將短語規則所得之結果視為詞的種類，允許在規則中的非關鍵詞中還可以是其他的規則，這樣的網狀的規則(Rule Nesting），使得系統對於斷詞的包容性提升，也就是說原有規則產生詞組的生命力也隨之提升了。

而這觀念目前在系統中是用來實現單一連語規則以及多重連語規則的結合，而利用這樣結合的效果也解決了一些問題，未來如果規則庫建立完整後，便可以將規則分類，在歸入分類詞群，向上建構逐步增加處理的單位。

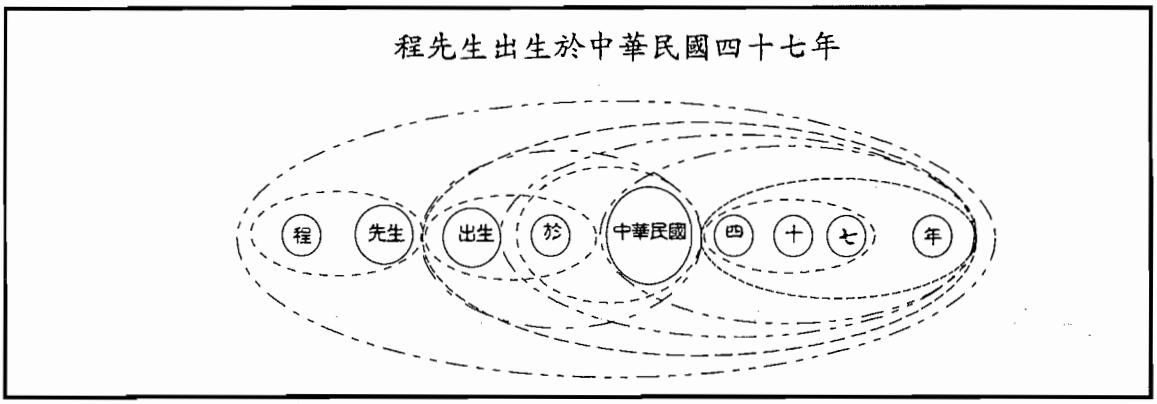

圖 12 網狀規則實例

定量複合詞

一般關於數量詞的描述，可以用定詞加量詞的語法分析來處理，事實上我們只要利用數字規則為非關鍵詞部份，就是說數字狀態規則是我們分類詞群標示中的一種，再利用量詞作為短語規則的關鍵詞，給予適當的分數，如此再依網狀規則的觀念，便可以處理定量式複合詞。例如我們說『五千元要買一頭牛』，其中『五千元』、『一頭牛』便可以利用定量式複合詞規則來組合輸出。而這樣的處理一來可以解決詞庫無法完全收錄定量式複合詞的問題，同時也較合乎一般人的斷詞習慣。

一般混合式規則

而在一般狀況下也有可能有某些字詞組合經常會夾雜數字串，例如"中華民國八十四年"、"第三課"等等，這一些我們也可能利用我們發展的兩種方式來產生其生成規則。首先還是針對前面第三章所描述的方法來處理所謂的關鍵詞部份，然後再利用數字狀態轉移方法來檢查非關鍵詞部份是否合乎條件，換句話說也就是說數字狀態轉移已經被列入我們的分類詞群中獨立為一類了。

實驗結果

【實驗一】我們分別就長度為三至九的正規數字，利用電腦亂數產生測試數字

所得之正確辨識率為：

	三字	四字	五字	六字	七字	八字	九字
第一名	88.10	85.76	86.02	85.30	83.76	81.14	78.02
第二名	6.52	6.06	1.24	2.32	1.68	1.02	1.68

表 7 正規數字狀態轉移辨識率

【實驗二】在一般純數字串的數字系統中，我們也是利用電腦亂數產生樣本，所測試的樣本數為

	三字	四字	五字	六字	七字	八字	九字
第一名	87.46	84.35	86.31	85.26	82.76	81.36	81.08
第二名	7.77	10.52	3.24	1.25	1.07	0.02	0.01

表 8 一般數字串之辨識率

從上面的實驗中我們可以得到幾個結論：

一）在音完全正確或是以鍵盤輸入的情況下，我們知道由狀態轉移的處理下，輸入串越長或是轉移過的狀態數越多，其出現在第一名的機率越大。不過在我們上面的實驗中發現在語音辨認中並不是如此，雖然隨著輸入音串的增長，合理的輸出會越少。不過也可能由於使用者在某個狀態的音不是準確的，而不能在狀態轉移中轉移至合法結束狀態，所以在這種情況下越長音串辨識率便越低了。我們曾經嘗試放寬條件來使得可以進入狀態轉移的組合多一點也就是在處理音的條件上放鬆了，不過這樣的結果雖然讓辨識率提升了，不過卻會使系統處理的速度降了下來，所以我們覺得在前面所提的條件在實用上是合理而可行的。

二）在上面兩個實驗中我們發現音串長度為四時，在第一名輸出的機率有降低的趨勢也就是說比音串長度為三小，同時也比音串長度為五來的小，不過在前二名的累積輸出機率便超越了音串長度為五的輸出。這樣的結果我們由實驗中的實作知道，這是因為在我們原有的詞庫中有些四字詞的成語是用若干個數字在詞當中，由於辨認音節權重的配分下，可能在配詞機構中會產生分數較高之輸出，所以就會把數字組合的名次往後擠了！這類的詞有『三頭六臂』、『四面八方』、『三人市虎』等，在我們去嘗試輸入這一些成語時，數字的組合也常常出現在我們的合理輸出中，不過由於辨認音節權重以及我們在狀態轉移時取最先最佳輸出（first best）的影響下，數字串的輸出並不會對系統詞造成妨礙。

四、 結論

在系統評估上，容錯率跟速度都是重要的考量，而我們在系統的發展上同時也考慮到這兩個特性，所以我們提出了將句子分為幾組詞組來處理的構想，希望從合理斷詞的想法上大量降低處理的複雜度，發展出一套在實用可行的系統。在單一詞未知短語規則的實驗中可以發現可以藉此來輸出詞組的組合，但是還是會有一些不良組合會影響系統輸出，這也表示分類詞群的研究還是有十分寬闊的發展空間。

而詞庫無法收錄的詞，從現實狀況觀察，歸納出一定的文法狀況，並藉狀態轉移的檢查篩選合理輸出。在這樣的考量下，我們從實驗中發現效果十分良好，而這一類的詞包括一些數字組合、一些複合詞甚至一些日常生活中時常會使用的樣板。

在這兩種短語規則的處理下，可以解決了中文詞綴、數量詞、介詞、連接詞以及一些複合詞的問題，關於古文、翻譯名詞、外來語以及新詞是可以利用整理詞庫的方式來解決，不過關於姓名的輸入一直是我們尚未能解決的問題。

參考文獻

[1] Ren-Yuan Lyu,Lee-Feng Chien,Shiao-Hong Hwang,Hung-Yun Hsieh , Rung-Chiuan Yang,Bo-Ren Bai,Jia-Chi Weng,Yen-Ju Yang,Shi-Wei Lin , Keh-Jiann Chen,Chiu-Yu Tseng,Lin-Shan Lee,"Golden Mandarin(III)- A User-Adaptive Prosodic-Segment-Based Mandarine Dictation Machine for Chinese Language with Very Large Vocabulary ,"ICASSP-95,pp.57-60,1995.

[2] Lin Shan Lee,Chiu-Yu Tseng,Hun-yan Gu,Fu-Hua Liu,Chen-Hao Chang, Yueh-Hing Lin,Yumin Lee,Shih-Lung Tu,Shew-Heng Hsieh,and Chian-Hung Chen, "Golden Mandarin (I)-A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary ,"IEEE Trans. Speech and Audio Processing, Vol 1,No. 2,pp158-179,1993.

[3] 許聞廉,陳克健,"『國音』智慧型輸入系統的語意分析",技術報告,中央研究院

[4] 陳克健,陳正佳,林隆基,"中文語句分析的研究-斷詞與構詞",技術報告,中央研究院,1986

[5] 謝子陵,"A Lingguistic Decoder for Mandarin Speech Recognition Using a Score Parser,"碩士論文,國立成功大學

[6] 張俊盛,陳志遠,陳順德,"限制式滿足及機率最佳化的中文斷詞方法",ROCLING IV,pp.147-165.

[7] Ming-Shih Chen,Tracy Yang and Hsiao-Chuan Wang,"Speaker Identification over telephone system based on channel-effect cancellation,"The Journal of the Chinese Institute of Engineers,IEEE,1993.

[8] Jhing-Fa Wang,Chung-Hsien Wu,Shih-Hung Chang,and Jau-Yien Lee, "A Hierarchical Neural Network Model Base on A C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition ,"IEEE Trans. Signal Processing,Vol 39,No. 9,pp.2141-2145,1991.

[9] 趙元任,"中國話的文法",中文大學出版社,香港,1980.

[10]羅肇錦,"國語學",五南出版公司,台北市,八十一年

[11]J.H. Wright,G.J.F.Jones and H.Lloyd-Thomas,"A Robust Language Model Incorporating A Substring Parser and Extended N-gram ," ICASSP,IEEE,Vol 1,pp.361-364,1994.

[12]J.H Wright,G.J.F.Jones and E.N.Wrigley,"Hybrid Grammar-bigram speech recognition system with first-order dependence model ," ICASSP-92,IEEE,Vol 1,pp.169-172,1992.

[13]Marie Meteer,and J. Robin Rohlicek,"Statistical Language Modeling Combining N-gram and Context-free Grammars ,"ICASSP- ,IEEE,1993.

[14]張元貞,林頌堅,簡立峰,陳克建,李琳山,"國語語音辨認中詞群語言模型之分群方法與應用",ROCLING VII,pp.17-34,1994.

[15]James Allen,"Natural Language Understanding,"The Benjamin/Cumming Publishment Comoany,Inc. ,Redwood City , CA, USA,1995.

[16]Lee-Feng Chien,Keh-Jiann Chen,and Lin-Shan Lee," A Best-First Language Processing Model Integrating the Unification Grammar and Morkov Language Model for Speech Recognition Application,"IEEE Trans. Speech and Audio Processing,Vol 1,No. 2,April 1993.

[17]McDonald, David D.,"An Efficient Chart-based Algorithm for Partial Parsing of Unrestricted Texts,"1992.

[18]Lee-feng Chien,K.J. Chen,and Lin-shan Lee,"An Augmented Chart Parsing Algorithm Integrating Unification Grammar and Markov Language Model for Continous Speech Recognition ,"ICASSP,IEEE, pp.585-589,1990.