利用多重連語規則之
電腦語音行事曆系統

王駿發　劉誠勇　吳宗憲　張哲賓
國立成功大學資訊工程研究所

摘要

本文提出一套結合辨識、合成、編碼與解碼等中文語音技術的電腦語音行事曆系統。其中辨識部份採用音中仙中文語音輸入系統，並且利用多重連語規則連接數個詞，使得語音輸入可以一次完成。編碼與解碼部份可選用 ADPCM或CELP兩種語音編碼技術。語音合成部份則採用以CELP爲基礎的文句翻語音系統。結合上述各種中文語音技術，本系統在486-66以上相容個人電腦上整合出DOS與視窗兩種版本。

一、簡介

語音一直是人類溝通最自然且最直接的方式，透過聲音的傳遞，可自然地表達出內心對人事物的感受與見解。在電腦發展的過程中，由於人機使用者界面的限制，只能透過鍵盤或滑鼠等手動的輸入裝置。因此人與電腦的語音溝通一直都只是科幻電影上的場景，而無法在日常實際生活中實現。

早期由於中文語音技術不夠成熟，再加上需要大量記憶體與快速的數位信號處理器，因此只有在實驗室中有一些正在研究而無法大眾化的成品。近年來，隨著電腦科技的長足進步與普及，再加上中文電腦語音技術漸趨成熟，因此漸漸的有一些中文語音產品的出現，如：音中仙中文語音輸入系統及蘋果中文語音輸入系統等。

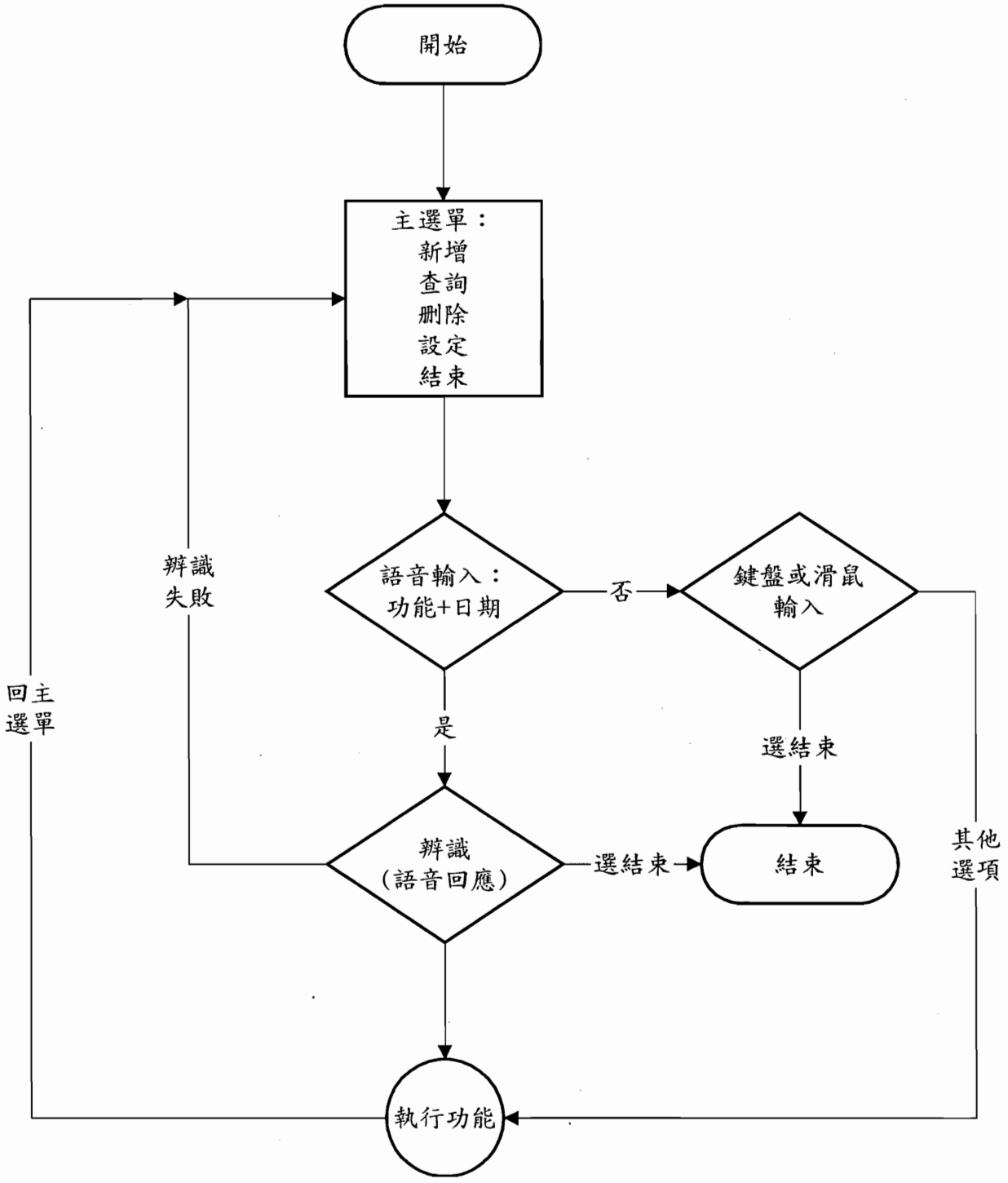
中文語音技術可大致分爲辨識、合成、編碼與解碼等幾部份。中

文語音辨識是人與電腦溝通的關鍵技術，也就是讓電腦可以聽得懂國語。由於目前中文鍵盤輸入法一直都非常不方便，除了必須牢記鍵盤上各種符號的相關位置外，更要拆字根或拼注音等，除非經過專業的訓練，否則不容易做快速的中文輸入。因此預期實用的中文語音輸入系統將有相當大的市場機會。中文語音合成是讓電腦講話的技術，讓電腦講話雖然不難，但要很自然地講卻不容易。而中文語音合成的應用也相當廣泛，如：有聲圖書、簡報系統、電話總機服務等。由於語音的資料量相當大，因此資料壓縮非常重要，而語音編碼與解碼就是這方面的技術。如何能達到最大的壓縮量又能即時而不失真，是中文語音編碼與解碼的研究方向。隨著網際網路的普及，在網路上語音交談要求快速又不失真，因此語音編碼與解碼的重要性與日俱增。

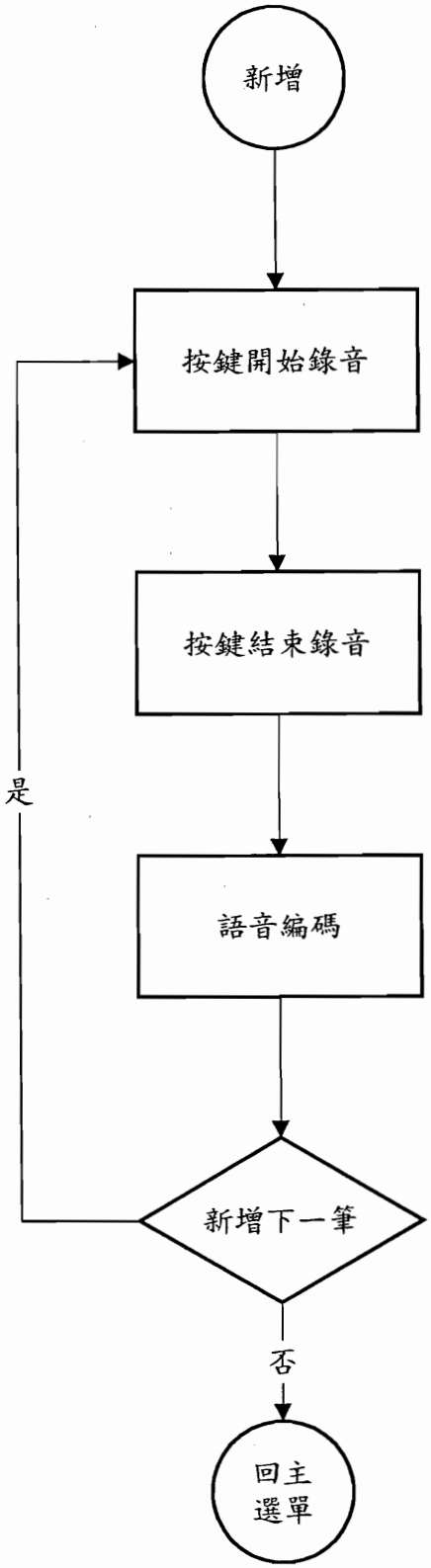
二、系統架構與流程圖

電腦語音行事曆是一套口語記事、查詢的系統，結合電腦語音辨識、合成、編碼與解碼等技術而成。其主要的流程架構如圖一所示。進入系統後會顯示主選單，其中有新增、查詢、刪除、設定與結束等功能。此時可用語音輸入功能與日期，如：新增六月十七日上午，系統會對輸入的語音加以辨識，並以語音回應讓使用者確認是否正確，如果沒有問題，就可執行各項功能。而其中的日期是來分別各事件用的，即爲各事件的編號，方便區別與查詢每個不同的記事。除了用語音輸入外，系統還可以用鍵盤或滑鼠輸入。

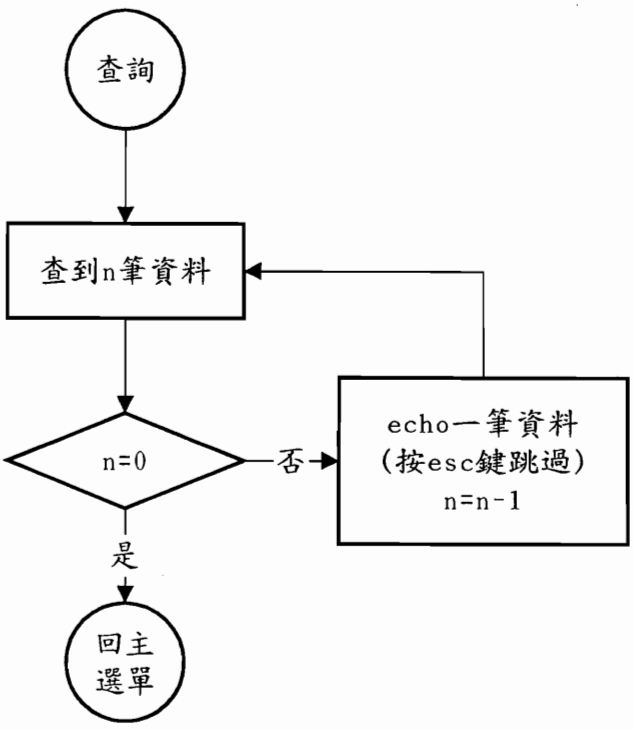
主選單中的新增功能是讓使用者能將想要記錄的事情用口語記錄下來，其流程架構如圖二所示。查詢功能是讓使用者找尋某一段時間內的有什麼事件，播放出來聽看看有什麼待辦事情，其流程架構如圖三所示。刪除功能是讓使用者清除某一段時間內的事件，讓系統佔用的記憶體減少，其流程架構如圖四所示。設定功能是讓使用者切換一些選擇，例如：語音編碼是採用CELP 或ADPCM。若想結束系統則可在主選單選取結束功能即可。



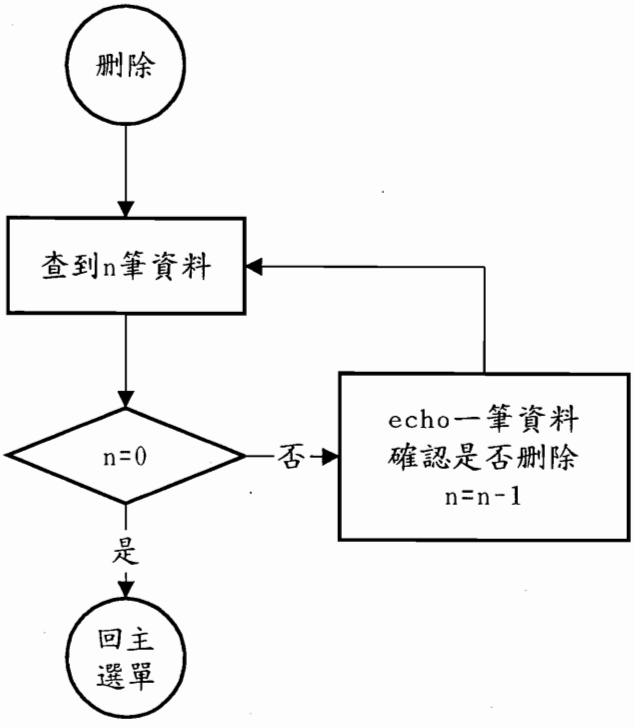
圖一　電腦語音行事曆基本流程圖



圖二　新增功能流程圖



圖三　查詢功能流程圖



圖四　刪除功能流程圖

三、各部份子系統

本系統採用了一些中文語音技術，以下分別加以介紹。

3.1 電腦語音辨識

在辨識部份本系統使用音中仙語音辨識系統[1]，其辨識模組是利用拜氏網路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。圖五是拜氏網路的基本架構圖，拜氏網路基本是以拜氏定理爲理論基礎，在架構上可分爲輸入層、高斯層、混合層與歸納層。輸入層爲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，再根據此一輸出距離來取辨認音節輸出。

除了辨識模組外，音中仙系統還有一些配詞計分的機構，其目的是將上述的辨識模組所得到的候選注音，轉換爲可能對應的詞或短語。圖六是音中仙的系統流程圖，其基本的特性如下所示：

a.以詞爲基本單位的中文語音辨識系統

b.使用規則庫可產生短語或句子

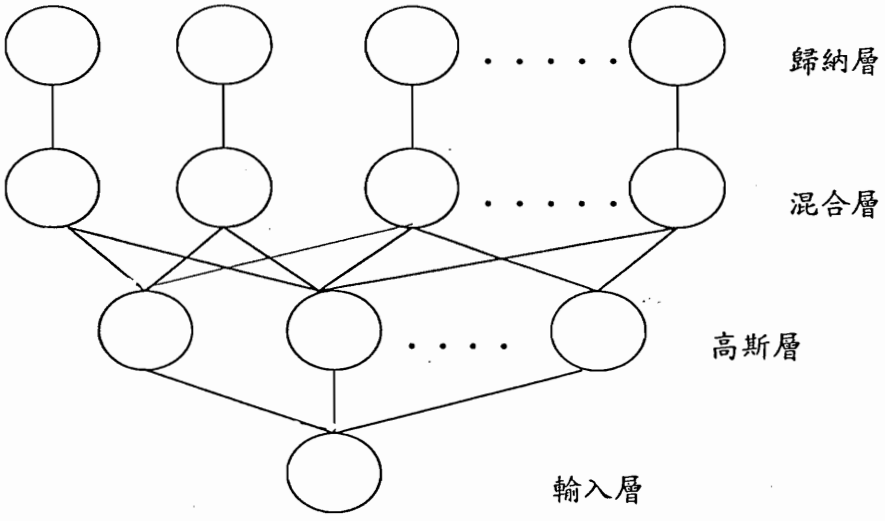
c.辨識的詞彙達到數萬詞，且可依需要再加以擴充

d.語者相關相連音系統，使用者訓練二十分鐘就可使用

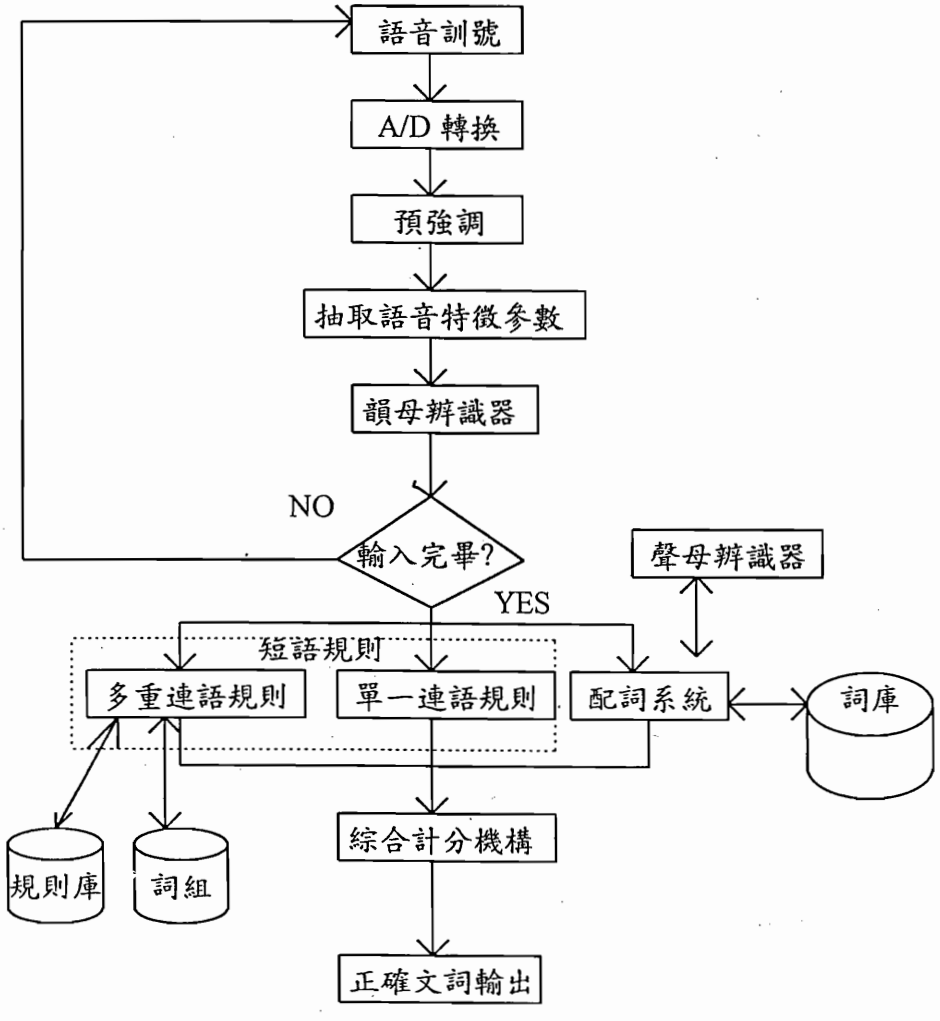
e.具有線上訓練的功能，可使得辨識越唸越準

3.2 多重連語規則

在句子中各種詞類間都有其前後連接的關係，根據這種關係發展出了自然語言的剖析器(natural language parser)[2]，又在各種雙連(bigram)[3]資訊的統計中，發現在雙連資訊中大部分的參數值都是零。也就是說存在於人類所習慣使用的自然語言中，所謂的連語現



圖五拜氏網路



圖六 音中仙之系統流程

象(Collocation)[4]是十分常見的。雖然在語法上，中文沒有西歐文字一般的嚴謹結構，但在連語的特性仍是非常明顯的。所以我們以音中仙爲架構發展了一套多重連語規則，如圖六左下角所示。如此可串接更多的詞，輸出爲短語或句子，加速了輸入中文的速度。

首先介紹可分支線性狀態機

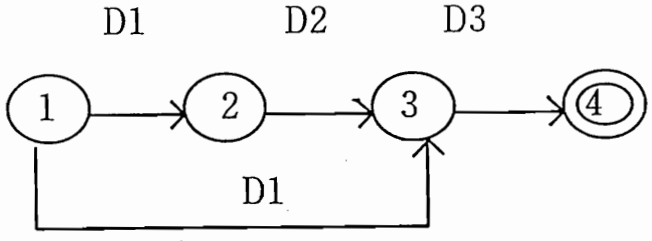
M = (S, I, v)

S = {S1, S2, ..., Sn}

I = {D1, D2, ..., Dn}

v : S x I -> S

其中S是在M中狀態的集合，且在M中沒有自身迴圈(self loop)及回饋(feedback)的情況。I爲輸入的詞的集合，稱爲詞組。v爲下個狀態的映射(the next state mapping)。如下圖是四個狀態的可分支線性狀態圖：



S = {1, 2, 3, 4}

I = {D1, D2, D3}

v : (1, D1) -> 2

(2, D1) -> 3

(3, D2) -> 3

(4, D3) -> 4

在本例中可看得出狀態圖中D2是可忽略的，也就是說不論D2是否有輸入，當到達最後狀態4(final state)時皆可被接受。以下是上述狀態圖的實際例子：設D1爲量詞，D2爲形容詞，D3爲名詞

"一個好學生"

"一個學生"

因爲形容詞"好"可忽略，故上兩短語皆可接受。

爲了讓使用者能描述上述的狀態圖，以便能在電腦上執行，我們定義了一個多重連語規則格式如下：

D1 A1B1 D2 A2B2 ・・・ Dn AnBn

其中D1，D2，・・・，Dn即爲上述狀態圖的詞組。對於j從1到n，Aj=1表示Dj爲必須輸入之詞組，Aj=0表示Dj爲可輸入或不輸入之詞組，Bj=1表示經Dj輸入後轉換至最後狀態，Bj=0表示經Dj輸入後不轉換至最後狀態。例如前面的狀態圖就可用下面的規則來表示：

D1 10 D2 00 D3 11

要使用上述的多重連語規則，首先定義本系統所需要的詞組：

D1 ＝ {新增，查詢，刪除，設定，結束}

D2 ＝ {一，二，・・・，九十九}

D3 ＝ {年}

D4 ＝ {月}

D5 ＝ {日}

D6 ＝ {上午，下午，晚上}

D7 ＝ {到}

本系統用到的規則是："功能"+"日期"+"到"+"日期"，其中的"功能"就是D1，"日期"則是由D2至D6所組合而成，而"到"則是D7。例如：查詢六月一日到六月六日、新增七月五日下午...等。最後依此建立了三十一條多重連語規則，如此就可將上述例子一次唸完辨識出來，而不必分成數個詞來唸。

3.3 語音編碼與解碼

語音編碼技術可分爲三種：波形編碼(waveform coding)、參數編碼(parameter coding)及混合式編碼(hybrid coding)。波形編碼一般應用在時域(time domain)上，如：DM、DPCM、ADM及ADPCM等，

都是在時域上模擬語音波形，可以產生較高音質的語音。但壓縮率較低，對於減少儲存空間及降低傳輸率並無太大幫助。參數編碼是應用在頻域(frequency domain)上，以模擬人類聲道特性爲基礎，藉著合成濾波器將聲音還原，如線性預測編碼(LPC)等。參數編碼器所需要的位元很低，可以有較高的壓縮率，但音質比較不清晰。混合式編碼是結合上述兩種編碼方式而成，有較低的位元率及較佳的音質，如多脈衝激發編碼(MPE)和碼本激發線性預測編碼(CELP)等。但因需要大量的運算，故要達到即時的要求，便得借助硬體設備及改良演算法。

本系統可選用ADPCM或CELP兩種語音編碼技術。ADPCM是CCITT的標準[4]，有四種不同的位元率：40、32、24及16kbps。本系統是採用32kbps的ADPCM，若錄音是以8kHz(Byte)取樣的，則壓縮率爲2倍。因爲是波形編碼，因此音質相當清晰，並且編碼速度相當快。CELP是由美國國防部和AT&T貝爾實驗室聯合發展的編碼技術，並爲美國國防部訂定爲美國國家語音編碼標準[5]。CELP是一個以音框(frame)爲導向的編碼器，對輸入的語音訊號以8kHz取樣，每240個樣本(30ms)當成一個主音框，作爲參數編碼的單位。並把主音框分爲四個各60個樣本且相互不重疊的次音框，作爲編碼運算的單位。CELP是架構在先合成後分析的處理方法上，藉著搜尋預存的各項參數量化表而形成語音訊號，並把此合成語音訊號與欲編碼的輸入語音訊號作波形上聽覺遮蔽，而求得最適當的編碼參數。最後以4.8kbps編碼輸出，因此其壓縮率高達13.3倍。其音質比ADPCM明顯較差，但仍能聽的清楚，不過其編碼速度也較慢。

3.4 電腦語音合成

電腦語音合成的方式主要可分爲兩類：第一類的作法是將可能使用到的語音信號事先錄製下來，直接或經編碼壓縮後儲存於系統記憶體之中。當系統欲說出某一文句時，僅需至記憶體之中，找出相對應的語音信號，將其輸出即可。這一類語音合成方式的複雜性低、運算

量較小，合成之語音易於達到自然、流利、清晰的要求，適合應用在少量文句的語音合成系統之中。然而，其應用範圍也受到相當的限制，對於文句數量多或文句變化大的系統，就無法符合實用性的需求。

第二類語音合成方式，是先將基本語音合成單元及合成規則存放於記憶體之中。當系統處理輸入文句時，便從記憶體中取出相對應的語音合成單元加以組合，並配合語音合成規則調整音韻特徵，使得所合成的語音更爲自然。這一類語音合成方式的複雜性較高，但所需之記憶體空間較小，可以合成任意文句，因此應用範圍極爲廣泛。

本系統的視窗版採用的是第一類方式。系統預先錄製一月到十二月、一日到三十一日、上午、下午、晚上、新增、查詢...等語句，當系統要回應時，再加以組合連續播放。例如：要回應"新增六月七日下午"，是連續播放"新增"、"六月"、"七日"和"下午"四句預先錄好的語音。

在DOS版部份採用以CELP爲基礎的文句翻語音系統[6]。此系統主要是以408個國語單音音節，配合聲調(tone)的變化，作爲基本的語音合成單元，所以預先錄製了1410個由女性發聲的國語單音(含四聲及輕聲)。爲了降低所有語音資料所佔的龐大記憶空間，利用CELP的13.3倍高壓縮率及其合成音質幾近原音之特性，將所有語音資料編碼壓縮後儲存。

四、系統整合結果

電腦語音行事曆系統包含了數種個別發展的系統，因此最重要的是系統整合的工作。由於發展平臺的不同，先後有DOS及Windows兩種版本，以下分別就這兩種版本加以討論。

4.1 DOS版

由於系統所用到的各種語音技術是不同的人用不同的程式語言所發展的，因此首先得決定發展工具。因爲所用到的語音辨識、編碼與

解碼是用C語言寫的，而語音合成是用C++語言寫的，基於相容性的考量，因此決定採用C++語言爲系統的發展工具。

接下來是模組化的問題。因爲每一個子系統原本都是完整獨立且可獨自執行的系統，並且都有各自人機界面。因此，必須先去除各個子系統的界面。再來原本是要將每一個子系統都簡化爲數個功能獨立的執行檔，由系統依需要叫用。但由於辨識與合成都有音檔存在記憶體中，必須機動地取用，因此僅能簡化成幾個功能獨立的函式。而編碼與解碼就很順利地簡化成執行檔。模組化的結果如下：

辨識部份有五個函式：

(1)將辨識的音檔載入記憶體中

(2)開啓辨識

(3)將聲音辨識成注音

(4)關閉辨識

(5)將辨識的音檔從記憶體中釋放

合成部份有三個函式：

(1)將語音合成單元的音檔載入記憶體中

(2)將文字字串合成爲聲音

(3)將語音合成單元的音檔從記憶體中釋放

編碼與解碼部份有四個執行檔：

(1)ADPCM編碼

(2)ADPCM解碼

(3)CELP編碼

(4)CELP解碼

系統整合的主要工作是要結合上述各個語音技術模組，以建立一個第二節所示的流程架構，透過適當的人機界面，而達到語音行事曆的功能。除此之外，還要處理一些整體性的管理工作，如音效卡的收放音控制與記憶體的管理。

在音效卡的收音控制方面，由於口語錄音和語音辨識都要用到收

音的裝置，因此系統採取"要用到才打開收音裝置，用完就馬上關閉收音裝置"的策略，而其控制的方法爲按空白鍵。例如在錄音前必須按一下空白鍵才能開始錄音，而且錄完音後必須再按一下空白鍵才會關閉錄音裝置。語音辨識也是同樣的方法，如此就不會造成資源衝突。在音效卡的放音控制方面，則整合成一個放音函式，統一由系統控制放音裝置的開關。利用參數的控制，可分別播放口語音檔、語音合成音檔與解碼後的音檔，並且可以按Esc鍵中斷正在播放的聲音。

由於語音資料非常龐大，因此記憶體的取用及釋放十分地重要。系統有幾個記憶體控制函式，用來管理系統所用到的傳統與延伸記憶體。系統啓動後，首先用語音合成的第一個模組函式將語音合成單元的音檔載入延伸記憶體(XMS)中，約900多K。接下來再用語音辨識的第一個模組函式將辨識用的音檔載入延伸記憶體中，約200多K。整個系統含辨識與合成的函式模組約300K，佔用傳統記憶體，還有系統呼叫編碼或解碼執行檔時會用到20到60K不等的傳統記憶體。再加上音效卡與中文驅動程式等，因此系統對記憶體的要求至少要2M的RAM。系統結束前，會先用語音辨識的第五個模組函式將辨識用的音檔從延伸記憶體中釋放，接著用語音合成的第三個模組函式將語音合成單元的音檔從延伸記憶體中釋放，最後再將系統從傳統記憶體全部移除。

此一版本可定位爲盲人語音行事曆，系統除了少數按鍵外，絶大部分的操作都用語音來控制，並且有語音回應及簡單的操作說明，因此可試用盲人來操作。

4.2 視窗版

視窗版的電腦語音行事曆的基本流程與DOS版差不多，但由於視窗作業系統的事件驅動導向(event-driven)與訊息傳遞方式的不同，因此這一版本比較適合明眼人來操作。事實上，只要麥克風加上滑鼠就能完全操作完畢，完全不必用到鍵盤。

由於目前還沒有合適的視窗版語音合成、編碼與解碼版本，因此

採用了一些和DOS版不同的方式。在語音合成方面使用3.4節提到的第一類語音合成方式，效果還不錯。因爲沒有編碼可壓縮語音資料，視窗版直接將所錄下來的口語事件原音音檔播放出來。而辨識部份則是另外開啓音中仙，辨識完後再將字串送到系統處理。

五、結論

由於中文語音技術漸趨成熟，我們嘗試將其結合，並在個人電腦上作出一套實用的電腦語音行事曆。這是結合中文語音辨識、合成、編碼與解碼等技術的系統。其中辨識模組是使用音中仙中文語音輸入系統，而配詞機構則發展多重連語規則，將數個詞連接成一個句子，使得語音輸入可以一次完成。編碼與解碼部份可選用ADPCM或 CELP兩種語音編碼技術。語音合成部份有兩種方式，DOS版是用以 CELP爲基礎的文句翻語音系統。視窗版則預先錄製所需的詞句，再加以組合播放。

將上述各種中文語音技術加以整合，本系統發展出 DOS與視窗兩種版本。直接的語音輸入及輸出是本系統的基本特性，但由於流程控制上的需要，尚須一些鍵盤或滑鼠的操作。完全與順暢的語音控制是未來改進的方向。另外視窗版的語音編碼與解碼技術的加入，也是未完成的工作。最後若想擴充功能，可加入名片管理系統，使得系統功能更加完整。

六、參考文獻

[1]葉瑞峰，王駿發，"應用於音中仙國語聽寫機之短語規則分析與建立," 碩士論文，成功大學資訊所

[2]謝子陵，吳宗憲，"A Linguistic Decoder for Mandarin Speech Recognition Using a Score Parser," 碩士論文，成功大學資訊所

[3]張元貞，林頌堅，簡立峰，陳克建，李琳山，"國語語音辨識

中詞群語言模型之分群方法與應用," ROCLING VII， pp.17-34，1994

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

[5]Furui, Sondhi,"Advances in Speech Signal Processing," Marcel Dekker Inc., New York, 1992.

[6]梁智能，王駿發，"FS1016語音編碼器在定點化及即時化之研究," 碩士論文，成功大學資訊所

[7]莊欣中，吳宗憲，"以CELP爲基礎之文句翻語音中韻律訊息產生與調整之研究," 碩士論文，成功大學資訊所