會聽國語的電腦

● 李琳山

一 前 言

根據統計，1993年個人電腦之使用比率，美國為51%，日本為59%，而台灣只有7%。台灣人一向喜歡高科技產品，台北市手提電話密度及成長率之高就是證明，為何偏偏這麼少人買電腦呢？究其原因，中文電腦輸入的困難應是主要因素之一。電腦鍵盤是拼音文字的產物，因此在其上作中文輸入有先天困難；有些系統速度太慢，有些需專業訓練，同時又都幾乎必須打斷原有文思。無怪很多人在十多年前就問：我們有可能用國語語音來輸入中文嗎？

要了解這個問題，我們要先指出，電腦語音辨認技術能力目前還十分有限。因此，如果要求以連續自然語音輸入任意字彙、任意文句、任何使用者的聲音，而且正確率達到100%，那麼這只能在科幻小說中才會出現。如此說來，就中文電腦語音輸入而言，是否有技術上可行而實際上有用的領域呢？從14年前(1982年)開始，我們就一直在思考、研究這個問題，並且把可以用國語語音輸入的中文電腦命名為「國語聽寫機」，這主要是因為以語音作為電腦輸入法不僅速度快而且不需專業訓練，人人都可以輕鬆應用。一旦打通了輸入瓶頸，電腦就會更普遍，就可以掃除中文社會全面資訊化的最大障礙。

二 問題簡介

國語聽寫機的構想雖然十分吸引人，但涉及的技術問題卻是極其困難而複雜的。困難的基本原因，在於聽寫機必須能「即時」(real-time) 聽寫由極多字彙組成的任意文句的語音。這個要求難度極高，對中文來說，其中最主要的問題包括：中文常用字至少5千以上，常用詞至少10萬以上，這龐大數字造成語音辨認的高度困難。而且因為中文單音節混淆音極多，不易辨認；即使辨音正

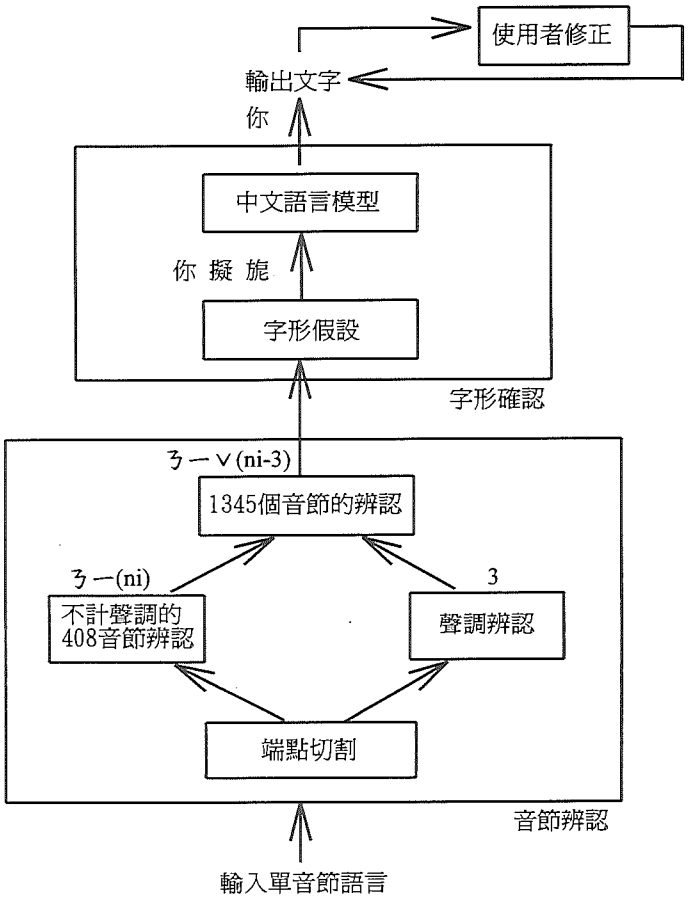
* 本文所述研究成果是歷年來數十位在台灣大學「語音實驗室」的學生的努力，結合中央研究院中文自然語言研究小組的研究成果共同締造的結晶。中央研究院同仁參與研究的尚包括資訊所研究員陳克健先生，歷史語言研究所研究員鄭秋豫小姐、黃居仁先生，及資訊所助理研究員簡立峰先生。作者特別在此向所有歷年來貢獻心力參與研究的同仁及學生表示由衷感謝。

確，同音字又極多，必須靠上下文才能確認每一個音代表甚麼字。加上中文句型千變萬化，缺乏統一規律，難以用人工智慧技術分析。此外，系統化、科學化的中文語文資料整理工作一向不足，遠遠落後於西方，這又是一個問題。以上舉列的，可說是中文特有的困難。至於聽寫機必須能「即時處理」語音，又能用於個人電腦上，其中牽涉了資訊科技、訊號處理、乃至中文語言和語音學等等，涵蓋範圍遼闊，是所有語言聽寫機共同面對的困難。

我們在12年前開始着手這一項研究時，為聽寫機增加一些限制條件，希望它能在技術上可行，並且實際有用。這包括三項限制。第一，聽寫機為「語者特定」，亦即機器只需一次聽懂一個人的國語即可。換言之，使用者先「訓練」機器聽他的國語，並保留「訓練資料」。這樣，不同的人仍可共用一台機器。其次，中文是一字一音的，因此可以用單音節為輸入單位。這在語音辨認技術上，就比較容易了。在這限制下，輸入時只能唸斷開不連續的單音節，因為在正常連續語音中，每一音都會受前後音影響而發生變化，分開唸可以大大減低這影響。最後，機器聽寫的結果允許有相當錯誤，而使用者可以用鍵盤或滑鼠更正在螢幕上所發現的錯誤。

在上述構想下，原始聽寫機的基本原理如圖1，它基本上可分為兩部分：(一)單音節辨認：國語單音節約有1,345個，不計聲調(四聲及輕聲)，則有

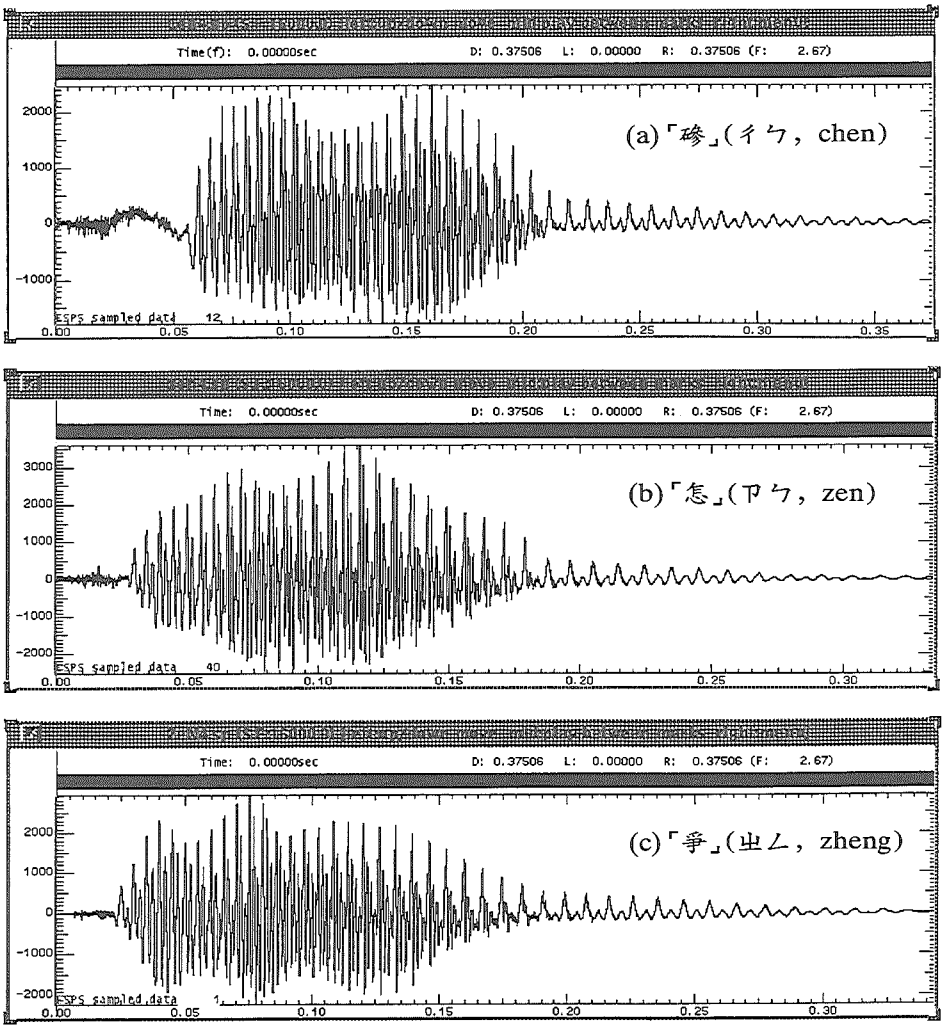
圖1　國語聽寫機的整體架構



408個。所以，單音節辨認可再細分為不計聲調的408個單音節辨認以及聲調辨認這兩個步驟。(二)字形確認：即根據上下文決定每一個辨認出來的音節對應那一個字。中文同音字那麼多，為甚麼人能知道所說的是甚麼字呢？這主要是靠上下文來判斷；我們當然希望機器也能做到這點。因此，在圖1的架構中，實在有三個問題需要解決：聲調辨認、不計聲調的單音節辨認、以及字型確認。這裏面除了聲調辨認在技術上不算太困難之外，後兩項問題都很不容易克服。

我們先談音節辨認。大家都知道，每一個國語音節(syllable)都是由聲母和韻母兩個語音元素構成，聲母共22個①，韻母共38個。要辨認不計聲調的408個單音節，相當於決定每個音節的聲母和韻母。這一工作的主要問題是：聲母在發音中只佔很短促的一小段，因此韻母相同的音節構成所謂「混淆音組」(圖2)。例如同屬「阿」(ㄚ或a)這個韻母的，就有阿、渣、差、沙、咱、擦、

圖2 國語淆混音訊號曲線的比較

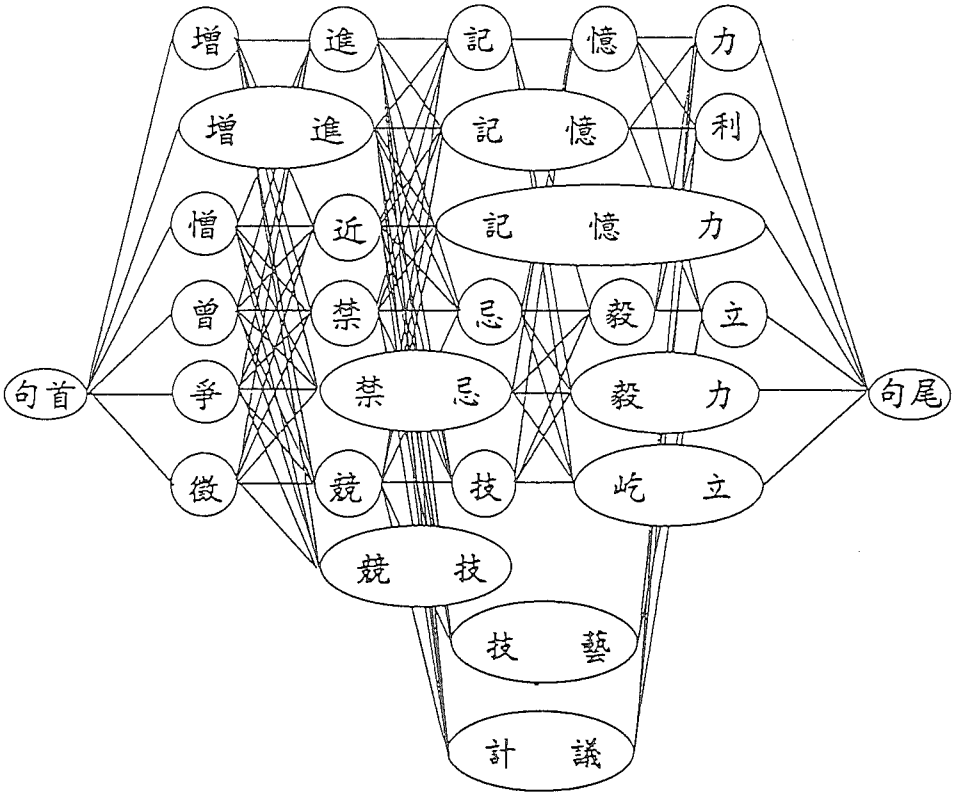


(a)「碜」和 (b)「怎」兩個音屬於同韻母淆混音組，它們的音訊曲線除了起始10%的一段是極相似的。「怎」和 (c)「爭」雖然聲母、韻母都不同，但音訊也高度相似，極易淆混。

撒、尬、卡、哈、搭、他、那、拉、把、啪、媽、發等一共18個不同音節，它們的語音訊號都是高度相似的。而全部408個單音節就是由38組這樣的「混淆音組」構成，要在同一混淆音組之中把不同音節分辨開來，這在語音機器辨認的技術而言，其難度之高是世界上少見的。

就字形確認的部分而言，圖3是一個簡單舉例。假設所輸入的5個字是「增進記憶力」，而且所有單音節都已正確辨認出來。但由於每一個音有一大堆同音字，他們就有可能組合成許多同音的多字詞，如「計議」、「技藝」、「屹立」、「毅力」等，這些單字詞、多字詞(包括正確的及一大堆不正確的)構成一個我們稱為「格狀詞組」的圖形，在這上面有多條可能途徑，每條途徑都可能是答案，而電腦必須找出那一條途徑是正確答案。現在再假設其中「增」(ㄗㄥ，zeng)、「進」(ㄐㄧㄣ，jin)這兩個字在辨認單音節時不完全確定，各分別帶有可能的混淆音「爭」(ㄓㄥ，zheng)及「竟」(ㄐㄧㄥ，jing)，那問題立刻就進一步複雜化。事實上圖3只是列舉了「爭」及「竟」的少數幾個同音字而已。因為前述408個

圖3 「字形確認」問題舉例



測試文句	增	進	記	憶	力
候選音節	ㄗㄥ (zeng) ㄓㄥ (zheng)	ㄐㄧㄣˋ (jin-4) ㄐㄧㄥˋ (jing-4)	ㄐㄧˋ (ji-4)	ㄧˋ (yi-4)	ㄌㄧˋ (li-4)

音節高度混淆，因此通常每遇到一個音節就必須同時考慮10–15個可能的混淆音，用以確保正確的音不會被遺漏。這樣每個音有10–15個選擇，而每個選擇有一大堆同音字，再由同音字構成一大堆同音詞，所以整體問題要比圖3所顯示的還要複雜困難得多。

三　雛型系統：『金聲一號』

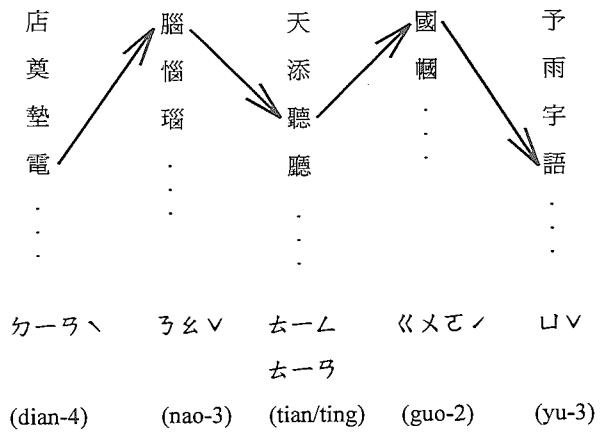
我們的研究是由1984年開始的。就音節辨認而言，早年想的方法是把音節切成聲母和韻母，先設法讓機器辨認韻母，待韻母確定後，可以接在前面的聲母就只有十幾個了。困難是如何自動把音節訊號正確切開，因為聲母、韻母之間實在並不存在明顯界線。後來決定引進當時辨認西方語言語音十分成功的「隱藏式馬可夫模型」(hidden Markov models)，並且用好幾年功夫加以改造，直到最後發展出特別為國語408個音節量身裁製的「加強聲母、考慮聲調特性的兩段式隱藏式馬可夫模型」，才算是找到初步答案。

這樣的馬可夫模型分別由一系列狀態(每一狀態代表一小段有特徵的聲音振盪曲線)構成，並由一組統計機率來描述。我們先建立99個聲母模型及38個韻母模型②，再把它們的各種組合串聯成408個單音節模型。未知的單音節進入電腦時不再被切成聲母和韻母，而是直接和這408個單音節模型作比對，從而找出答案。由於聲母模型是事先分開建立的，因此可以仔細強調它們的特性，避免相似的韻母把聲母間細微的差異遮蓋掉。因為這樣的單音節並未考慮其聲調，故還須把不同聲調都包含進去：例如單音節「巴」、「把」、「拔」、「爸」、「吧」五種聲調變化的情形都要包含在一個單音節模型中。

至於同音字的字形確認問題，早期的辦法是先構詞，即將多個同音字有可能建構的詞都列出來，成為圖3的格狀詞組，然後用語法分析根據上下文關係來判斷中間是甚麼字。但這並不成功，因為中文語法千變萬化，要電腦分析任意句型有高度困難。而且，每個音有多個可能候選音節，每個音節還有一大堆同音字，電腦需要搜尋的空間太大了。

後來發展的新方法，是所謂「中文語言模型」。其基本觀念是：人腦是憑經驗而不是靠分析來認識語法。對電腦來說，經驗就是統計。既然在圖3中每一途徑都是可能的答案，就可以利用大量文字檔案來計算字、詞相連的機率(即頻率)，這些機率網(或矩陣)就構成一個「中文語言模型」。從運作上說，這就是把圖3中所有可能途徑上的字、詞的相連機率相乘，再選擇機率最高的途徑為答案。以圖4的例子來看，假設輸入的是相當於「電腦聽國語」這五個字的音，那麼其中「電」有許多同音字如「店」、「奠」、「墊」、「電」等，「腦」也有許多同音字如「腦」、「惱」、「瑙」等。但「店腦」、「奠惱」、「墊瑙」等在文字檔案中很少有機會(當然不一定絕不可能)相連在一起，所以相連機率極低，乃至等於0，只有「電腦」是常常連在一起的，連接機率很高。同樣，在音節辨認不十分確定的

圖4 「中文語言模型」操作原理舉例說明



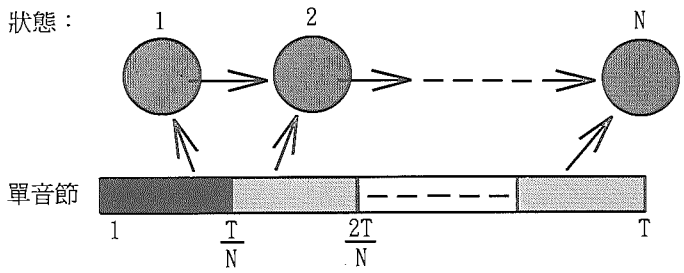
時候，也可能用語言模型根據機率從一堆聲音相類的字之中把正確答案挑選出來。

我們於1991年3月完成的第一代國語聽寫機雛型系統，基本上就是把上述經過特殊改造的馬可夫模型的單音節辨認技術、以及「中文語言模型」的同音和近音字選取技術在一台機器上整合起來，當時命名為「金聲一號」，以取「金聲玉振」之義。「金聲一號」證實了國語聽寫機的技術可行性，但事實上它與實用產品相距仍遠，這主要由於它所需的運算量極大，所以必須使用具有10個中央處理器的平行式電腦。它不但龐大、昂貴，而且速度緩慢，平均需4、5秒才能輸入一個字。況且，它缺乏強健性(robustness)及彈性調整能力，因此新使用者「訓練」機器需時冗長，機器對環境雜訊敏感；此外，它的語言模型所根據的文字檔案只是小學課本，那也不適當。凡此種種，都顯示它只不過是聽寫機的一個雛型而已。

四 改良智慧型系統：「金聲二號」

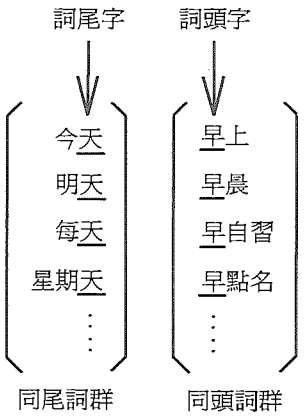
自1991年起我們繼續研究更進步的技術，獲得了兩個重大突破。第一個突破是決定放棄辨認西方語言的經典模型，即「隱藏式馬可夫模型」，因為它需要針對國語特性作許多特別剪裁，雖然成果不錯，但計算太複雜緩慢，並且缺乏彈性。我們重新設計了一個「分段機率模型」(圖5)。它和原來的模型很相近，只是每個音節訊號都依據固定比例分段，每一段對應一個狀態，而不再用統計法去估計狀態之間的轉移。這是由於要分辨國語的408個單音節，最大的困難只在前面的聲母造成混淆，其他部分的聲學結構並不太複雜。固定分段之後，模型簡化計算量大為減少，因而增加了強健性及調整彈性。

圖5 國語單音節的「分段機率模型」



另一項重要的突破則是在中文語言模型上。「金聲一號」的語言模型以字為基礎，但中文文句其實是以詞為構造單位。關鍵是中文常用詞數目極大，要建立詞與詞相連的統計模型極為困難。我們的新方法是讓許多詞構成一個「詞群」，並以「詞群」為基礎建立「詞群」與「詞群」相連機率的「詞群中文語言模型」。

圖6 詞群中文語言模型的基本概念



以「天」為詞尾和以「早」為詞頭的這兩個詞群的連接機率即「天」「早」的連接機率。

構成詞群最簡單的方法如圖6所示，是把「詞尾字」相同的詞集合構成「同尾詞群」。例如：「今天」、「昨天」、「星期天」、「每天」……等，就構成「天」為詞尾的同尾詞群。同樣，詞頭相同的詞例如「早上」、「早晨」、「早自習」、「早點名」……等等，也就構成以「早」為詞頭的「同頭詞群」，然後我們就可以進而計算同尾詞群和同頭詞群兩兩相連的機率。這時，「今天早上」和「每天早點名」等等的相連機率就計算在一起了。換句話說，機器只要學會「今天早上」，也就學會「每天早點名」了。這樣的模型不但十分「強健」，且調整彈性也大大增加。

第二代聽寫機，即「金聲二號」，把上述「分段機率模型」以及「詞群中文語言模型」整合起來，在1993年9月製成。和「金聲一號」相比，它除了因計算簡化而使軟硬體要求降低、速度加快、正確率提高之外，更大的好處是發展了智慧型學習功能，即能夠迅速學習新使用者的聲音、環境雜訊，甚至應用領域及遣詞、構句習慣等等。那也就是說，這一系統已經能夠適應使用者和環境的特徵與需要，這樣，「金聲一號」原有的諸多困難就迎刃而解了。

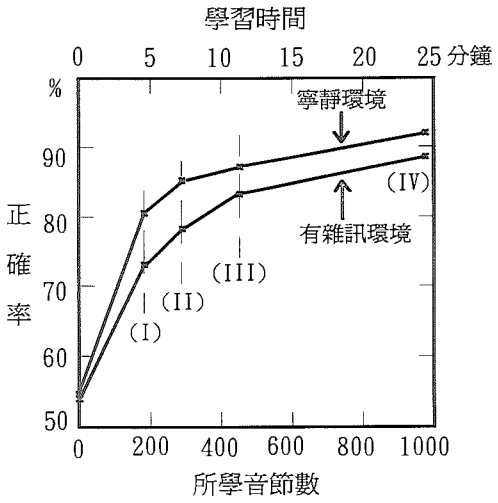
「金聲二號」的基本操作方式是：使用者先要訓練機器聽他的國語，以把語音特徵建檔。使用者唸斷開的單音節，中文字即自動輸入電腦，在螢光幕上發現錯誤後，可用滑鼠更正，完全不用鍵盤即可輸入中文。「金聲二號」不但可以接近「即時」的速度（平均約0.6秒／字）輸入相當正確的中文字，而且只需加上

一片數位訊號處理電路卡(DSP card)，插入任何AT級以上的個人電腦即可運作。

「金聲二號」更重要的特色，是它具備豐富的學習功能。這裏面最重要的突破，就是它會快速學習新使用者的聲音。未開始訓練之前，機器已有平均55%的正確率，這是因為它已先由許多在實驗室工作的同學訓練過，所以知道不同使用者的大致聲音特性。我們另外發展了讓機器同時學習聲母、韻母的技術，也就是新使用者唸了一聲「ㄅㄚ」(ba)，機器不但學到「ㄅㄚ」(ba)的唸法，也學到了他的「ㄅ」(b)這一聲母和「ㄚ」(a)這一韻母的唸法。因此新使用者只需要唸少數幾個音，機器就可以學到一大堆音。

為了增進機器的學習效率，我們發展了一套均衡涵蓋國語各種語音特性的「學習例句」，這是根據語音出現頻率在大量語料庫中挑出來的，雖然只包括很少字句，但卻涵蓋所有必需的音(如聲母、韻母、單音節等)。新使用者在機器引導下循序唸出這些句子，就可以把機器訓練好，過程並不枯燥、乏味、冗長，反而十分有趣。如圖7所示，這訓練程序分四個階段，它在不到半小時之內，就可以使機器的辨音準確率從55%提高到91%。在這既定程序進行的同時，使用者並可隨時脫離程序，直接使用機器，只要一面用一面更正，機器就會繼續學習，正確率可不斷提昇。機器還會同時自動學習和適應環境的雜訊特徵。從圖7可見，訓練25分鐘後，有雜訊情況比安靜環境中的準確率只差2.5%，因為機器已經適應環境雜訊了。此外，一般使用者都有自己的特別應用領域以及用字、遣詞、構句習慣，只要一面使用、一面更正，機器也同樣可以自動適應使用者的這些文化特性。綜合上述學習功能，使用者可在訓練5分鐘之後即初步使用，在短短25分鐘之後機器已經變為相當方便的工具了。經過相當時間的應用之後，機器還可以高度個人化，最後達到約95%的正確率。

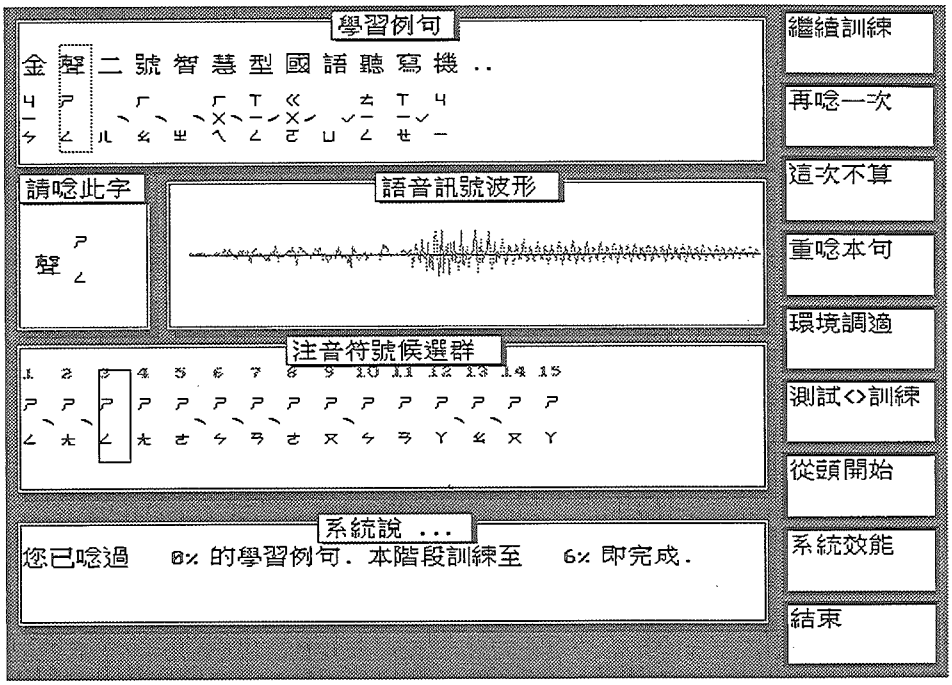
圖7 「金聲二號」的學習曲線



圖中括弧內數字為學習階段。

圖8為「金聲二號」的學習畫面。在畫面上方機器自動列出一句學習例句，並引導使用者唸出逐個單音節，然後即時列出辨認結果，這包括前15名候選單音節(進入前15名算是有辨認效果)，正確的音名次越前，效果越好，第1名表示辨認正確。使用者可以看到，一路唸下去，機器越來越能聽懂他的聲音的情形。畫面右方則是一系列用滑鼠操作的功能鍵。「金聲二號」另有一個聽寫畫面，供正式聽寫時用。它上方空間供編輯輸入文字之用，下方是當時唸進去的句子。使用者每唸一個字，辨認出來的音和字就顯示出來；如果有錯，使用者

圖8 「金聲二號」在監測屏幕上的學習畫面



可以用滑鼠按「訂正」鍵，在上方打開視窗，看到當時辨認的前15名候選音：如果其中有一個是正確的，可以立刻把它選進來；如果正確音不在前15名中，可以再開一個注音鍵盤，直接在上面用滑鼠選注音符號。同樣，如果要改字，也可用視窗進行。使用者還可按「學習」鍵，機器就會把剛才唸進去的聲音唸法、新詞及句型都學習進去。

五 連續語音聽寫機：『金聲三號』

雖然「金聲二號」已經十分方便，但使用者仍然必須每個字斷開來唸才能輸入，造成極大不便。今年3月剛剛完成的「金聲三號」，最大的技術突破就是克服了這個必須每個字斷開來唸的障礙。不要小看這簡單的一小步，它其實代表了在技術上的重大突破。在輸入極多字彙和任意文句的條件下，連續語音的辨認其實非常困難。原因是多方面的：每一個音節的特性受前後連接音的影響，不再像「斷開單字」的語音特性那麼穩定：在連續音中不但那一段是一個單字不易判斷，甚且一段之中究竟共有幾個字也不易判斷，因此極可能誤辨出「插入字」或發生「漏失字」，而錯誤又很容易向兩邊延伸：此外，在連續音中每個字的快慢可以有很大的變化，有些字會連在一起，不易分辨。例如「西洋」極易被誤認為「詳」，「答案」被誤認為「蛋」等等。「金聲三號」是克服了這連串困難，才成為可以聽寫「連續語音」的聽寫機。

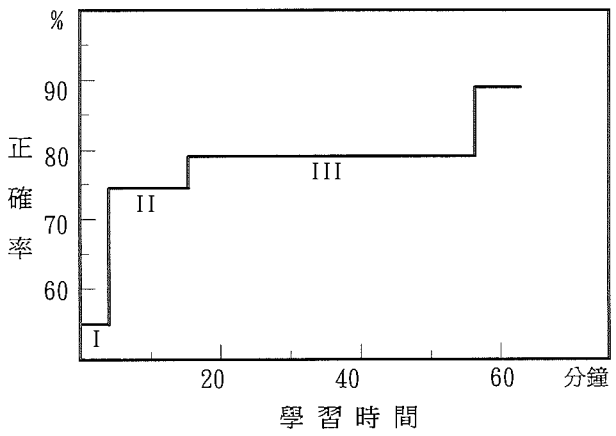
克服這些困難的關鍵有二，其一是不再用單音節作為辨認單位，而用更小的次音節單位(subsyllabic units)來作為辨認單位。在這個架構中，每一個音節是由數個次音節組成，而次音節的聲音特徵可以容許因前後連接的次音節單位之不同而有變化。我們嘗試了各種可能的次音節單位，發現前述的聲母、韻母相當可用，而音素(phoneme)也十分理想。例如相當於「交」的單音節「ㄐㄧㄠ」(jiao)可以用四個音素「ㄐ」、「ㄧ」、「ㄚ」、「ㄨ」(即j, i, a, u)組合而成。

其次，由於連續語音變化多端及不穩定，錯誤率顯然會大增，此外還要對付「插入音」、「漏失音」等問題。因此第二個關鍵是必須有更強健的「中文語言模型」來根據上下文更正錯誤和選出正確同音字。上一節所說的「詞群語言模型」還不夠好，因為同一詞群的詞未必有相同語言特性。例如前面圖6的「今天」、「昨天」、「星期天」、「每天」……構成一個「同尾詞群」，但是此一詞群也有諸如「老天」、「青天」等其他詞在其中。由於它們的特性及用法和「今天」、「每天」等並不一樣，如都混在一起就會造成錯誤。因此我們根據詞的語法、語意特徵以及統計特性，另外發展了組合詞群的技術，把10萬個以上的中文詞分成約1,000–3,000個各自有相當一致語言特性的詞群，在這些詞群上面建立起來的「詞群語言模型」就強健有效得多了。把上述兩種技術整合起來所造成的新聽寫機，就是「金聲三號」。

「金聲三號」有應用於工作站和個人電腦的兩個版本。工作站版可以直接輸入連續語音，而且長短不拘。使用者即使一口氣唸很長(例如超過30個字)的句子，機器也可以一口氣辨認出來，不會被難倒。有錯誤時，可以很方便的用滑鼠和聲音迅速作線上修正。此外，「金聲三號」不但男女聲均適用，而且還可以線上學習使用者的遣詞構句習慣、詞彙以及環境雜訊特性。「金聲三號」工作站版的速度是「即時」，亦即計算所需時間與輸入語音長度幾乎相同，可以立刻獲得辨認結果。目前，「金聲三號」是用報紙新聞作訓練工具的，輸入新聞的正確率達92%以上。估計未來變成產品後，使用者專心快速唸入文句，每分鐘至少可輸入正確文字100個以上。

至於「金聲三號」個人電腦版，則只使用486個人電腦加插一片普通的數位訊號處理卡，可說達到低價位、大眾化的目標。它與工作站版的最大不同，是不宜於一口氣輸入長句，而適合輸入十個字以內的短句。由於使用者撰稿時要思考，發音時需要換氣，短句輸入事實上是十分實用的。除此之外，它和工作站版分別不大：兩者都具備在視窗畫面上用滑鼠和聲音修正錯誤、線上學習使用者的聲音和語言習慣等功能。和「金聲二號」一樣，新使用者只要分三個階段唸「語音特性平衡詞」，並稍候機器訓練，即可獲得相當好的正確率。從圖9可見，只需前後不足1小時的訓練與適應，即可獲得約89%的正確率，此後機器仍可線上學習，使正確率繼續提高。此外，在特定應用領域下，常用字數、詞數及句型大為減少時，此時「金聲三號」個人電腦版可能不需訓練，即可直接供任何人使用。

圖9 「金聲三號」的學習曲線



「金聲三號」三個不同學習階段所需時間(唸詞及等候機器訓練各佔一半)及所唸詞數分別為I：4分，53；II：12分，186；III：40分，665。

如果使用者的國語不標準怎麼辦？這不用擔心：原則上機器只是在學習使用者的聲音，對它而言無所謂「標準國語」，所以這並不是問題。當然，使用者如果不能分辨一些常易混淆的音，例如「師」、「司」或「京」、「斤」等，則機器自然也可能誤認這些為同音字而發生混淆。但因為機器會根據上下文找出可能的同音字，故也會根據前後文刪去不可能的混淆音。例如使用者把「老師好」唸成「老司好」，那麼「師」、「司」都會列入候選，而「老師」構成一個常用詞且後面接以「好」很正常，「詞群中文語言模型」仍會把「老師好」挑選出來。

六 結論及展望

我們相信未來各種中文輸入法都各有其應用空間：鍵盤輸入需要專業人員操作，但其快速、大量的輸入能力是無可取代的；手寫輸入很方便，但永遠只限於少量字的輸入；印刷掃描輸入也很好，但限於已有現成文件的輸入。至於國語聽寫機呢？它的重要性將在於開拓新的中文電腦應用空間。很多人平常由於電腦使用量不夠大，所以不願練習鍵盤輸入，甚至因此在日常生活及工作中放棄使用電腦，他們都是聽寫機的理想使用者。因此聽寫機很有潛力提高中文社會的電腦使用率，加速中文社會的資訊化。目前已經有相當多國內外電腦業者有興趣將「金聲三號」發展成產品投入市場，有關的技術移轉以及合作研發工作已在積極進行了。

註釋

① 在通行於台灣的國語注音系統中，共有21個聲母，加上「空聲母」(即音節不帶聲母)共22個。通行於大陸的漢語拼音系統沒有空聲母，但在音節是由介音和元音組成的時候要用半元音聲母代替介音(例如「煙」字注音為「ㄧㄢ」，拼音則是yan，'y'是半元音聲母)，因此共有23個聲母。

② 聲母模型有99個的原因，是把聲母作了更精細的分類，例如「渣」和「朱」的聲母都是「ㄓ」(zh)，但因語音特性相當不同，要算是兩個不同的聲母；而「渣」和「摘」的聲母卻十分接近，可以算是同一個聲母，等等；如此細分起來聲母就增加到共99個，而不再只是22個。

李琳山 中央研究院資訊科學研究所所長、台灣大學電機系教授