玉振金聲築夢記

—— 十八年電腦語音研究的夢想與現實

文／李琳山（資訊系與電機系教授）

緣起

我們在台灣大學開始以電腦處理中文語音的研究是民國72年，是十八年前的事了。研究中文電腦的前輩們當時已有相當成果，中文輸入的方法有用大鍵盤的，有用四角號碼的，及用注音符號的等等，只是都不普及，很少人真的使用。我們當時就想：有可能用語音來輸入中文嗎？這個想法真是有夠精采，我們就把構想中的可以用國語語音輸入的中文電腦命名爲「國語聽寫機」（Mandarin Dictation System），因爲以語音作爲電腦輸入不僅速度快而且不需專業訓練，人人都可以輕鬆使用，就好像一個專業的聽寫人員替我們輸入。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。

國語聽寫機的構想確然吸引人，但仔細想想就知道那實在是一個遙遠的夢想，因爲所涉及的技術問題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：中文常用字至少五千以上，常用詞至少十萬以上，這龐大數字造成語音辨認的高度困難；而且中文單音中極易混淆的音極多，不易辨認；即使辨音正確，同音字又極多，必須靠上下文才能確認每一個音代表什麼字；加上中文句型千變萬化，缺乏統一規律，用人工智慧技術分析有高度困難；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。至於聽寫機必須要能「即時處理」語音，又要能用於個人電腦上，其中牽涉到資訊科技、訊號處理、乃至中文語言和語音學等等，涵蓋範圍遼闊，則是所有語言聽寫研究共同面對的困難。

電腦說國語－語音合成

由於這個問題太過困難，我們決定先由語音合成開始研究，也就是反向研究；先讓電腦將輸出的任意中文文句用聲音說出來。打開國語字典可以發現，中文字雖然千千萬萬，中文字的聲音只有一千三百多個；似乎只要把這些聲音都存入電腦，理論上就可以拼成任意文句的國語。這個想法就是中文的「文句翻語音」（text-to-speech）系統。這比語音輸入容易多了，因此我們由這個問題作起。當時電腦記憶體的容量十分有限，這一千三百多個音要存進電腦還需相當的壓縮技術，但最後發現拼出來的句子都根本不能聽。原來人說話時每一個音的聲音特性受到前後音的影響而改變，只要前後接的字不一樣，聲音就是不同的，因此硬拼的結果是完全不能聽的。後來與中研院的的語言學家合作，發展出一系列的聲音的音高、音調、音長、音量等受前後文影響調整的規則，才終於在民國73年完成第一台中文的「文句翻語音系統」，也就是全球第一台能說任意文句國語的中文電腦。之後在民國75年再完成第二台效果更佳的系統，並在76年在台北舉辦的「全國計算機會議」中讓兩台系統表演「電腦說相聲」，以豐富的戲劇效果，創造傳統文化與現代科技結合的少見紀錄。

金聲一號

在研究「文句翻語音」合成技術的同時，我們有了初步經驗，也在74年開始「國語聽寫機」的研究。當時所構想的基本原理事實上一直沿用到今日，雖然其基本技術及內涵已有了許多改變。這個原理可分爲兩個部份。一是音節（syllable）辨認，國語音節約有一千三百多個，不計聲調（四聲及輕聲），則約有四百多個。所以音節辨認可再細分爲不計聲調的四百多個音節辨認以及聲調辨認這兩個步驟。二是字形確認：即根據上下文決定每一個辨認出來的音節可對應到哪一個字。中文同音字那麼多，爲什麼人能知道所說的是什麼字呢？這主要是靠上下文來判斷。我們因此希望機器也能做到這點。這就有三個問題需要解決：聲調辨認、不計聲調的音節辨認、以及字型確認。這其中僅僅聲調辨認在技術上即已相當困難了，而後兩項問題更不容易克服。

經過多年的努力，終於在民國80年3月完成第一代國語語音聽寫機的雛型系統，基本上是把針對斷開的國語音節所專門設計的音節辨認技術，以及以單字爲基礎的同音和近音字選取技術，在同一台機器上整合起來，當時命名爲『金聲一號』，以取「金聲玉振」之義。「金聲玉振」在中國古典文獻中被用來描述世上最美的金玉之聲；對參與研究的人而言，國語的聲音確是極爲美麗的。金聲一號首度讓世人看到可以用國語語音輸入中文字，證實了國語聽寫機的構想在技術上的可行性，但事實上它與實用性相距仍遠。它只能辨認把每個字都斷開來唸的聲音，任何兩個字都不能連起來。它所需的運算量極大，所以當時必須使用具有十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，平均四、五秒才能輸入一個字。它缺乏強健性(robustness)及彈性調整的能力，因此新使用者『訓練』機器需時冗長，機器對環境雜訊相當敏感；此外，它的語言模型所根據的文字檔案只是小學課本，超出範圍就錯誤百出。

金聲二號

由於金聲一號並不夠理想，我們乃積極研究第二代的技術，到了民國82年9月，完成第二代聽寫機，即『金聲二號』。和金聲一號相比，它仍然只能辨認把每個字斷開來唸的聲音，但是因計算簡化而使軟硬體要求降低，速度加快，正確率提高，而且更大的好處是發展了初步的智慧型學習功能，即擁有學習新使用者的聲音、環境雜訊，甚至應用領域及遣詞、構句習慣等等的初步能力。使用者要訓練機器聽它的國語，所以需把語音特徵建檔。經過約二十五分鐘的建檔工作後，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）輸入相當的正確的中文字，而且只需加上一片數位訊號處理電路卡（DSPcard），任何AT級以上的個人電腦即可運作。

金聲三號

雖然金聲二號已經有相當成果，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。經過諸多努

力，我們在84年3月完成了第一版的『金聲三號』，其中最大的技術突破，是克服了必須每個字斷開來唸的障礙，可以直接用連續說的語音輸入。這是全球首台真的可以直接辨認連續國語語音的系統。不要小看這簡單的一小步，它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，連續語音的辨認事實上非常困難。原因是多方面的：每一個音節的特性受前後連接音的影響，不像「斷開單字」的語音特性那麼穩定，而是變化萬千的；在連續音中不但哪一段音是一個單字不易判斷，甚且一段音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」發生「漏失字」，而錯誤又很容易向兩邊傳播開來；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』極易被誤認爲『詳』，『答案』被誤認爲『蛋』等等。金聲三號是克服了這連串的困難，才成爲可以聽寫「連續語音」的聽寫機。

金聲三號第一版所需的計算量及記憶體容量都相當高，所以當時必須建在SunSpare20工作站上。它證實了連續語音輸入的可行性，但是因爲在昂貴的工作站上，一時的實用性並不高。它可以直接輸入連續語音，而且長短不拘。有錯誤時，可以用滑鼠和聲音作線上修正。它的速度是「即時」，亦即計算所需時間與輸入語音長度幾乎相同，可立刻獲得辨認結果。第一版的金聲三號是用報紙新聞訓練的，因此輸入新聞的正確率最高。它還有一個最大的困難，是語者特定(Speaker Dependant)的。使用者必須花數十小時的時間去訓練它，機器才能聽他的聲音。

之後在85、86、87年，我們再進一步完成了金聲三號的諸多新版本，最大的進步是由工作站轉移到個人電腦上，不需再附加任何硬體，掛在視窗下以純軟體操作；爲此必須大幅減少記憶體及計算量的需求；爲了讓一般大眾使用，必須大幅減少使用者訓練機器的時間；昂貴的高品質麥克風必須換成廉價的一般麥克風，並需適度抗拒環境雜訊等問題。這些都是爲了大眾化使用所必須採用的步驟，但也無可避免的小幅降低了它的正確性。這些版本也陸續透過國科會移轉給產業界，並成功的推出產品。幾乎同一時間，IBM公司也推出了它的中文語音聽寫軟體並大力行銷。值得一提的是，我們在學術界的文化下，所有研究成果均一路公開發表，IBM公司的研發人員都承認他們一路研讀我們所發表的所有論文，並聘請我們的研究生前往工作，而他們的研究成果卻不常公開。他們在這方面投入的人力物力及市場行銷和我們不成比例。但在87、88、89年連續三年由資策會主辦的全國使用者愛用資訊產品票選活動中，我們都大勝IBM成爲最受國內使用者肯定的產品，雖然資策會自己是IBM產品的代理人。

無線網路時代多元化應用的新挑戰

在努力研發一系列國語聽寫技術的同時，整個資訊世界也在快速演變中。不知不覺我們已由「個人電腦」時代進入了「後個人電腦」(Post-PC)時代，或是網路時代。在個人電腦時代中人類最主要的資訊活動是使用個人電腦，而個人電腦最大的用途是文件處理，輸入文字乃成爲最主要的語音應用。到了網路時代，人類最主要的資訊活動變成是上網，在網路上搜尋、瀏覽，取得資訊並進行各種業務例如電子商務。此時輸入文字不再是最主要的動作，因爲網頁上常有大量文字按鈕，只要按滑鼠就行了；即使要輸入文字，年輕一代的使用者早已熟習相當進步的中文輸入法，用語音反而會有錯誤，未必吸引人。倒是因無線通訊及無線網路的進步，人們上網也不見得再喜歡用個人電腦，而希望用無線手機，人手一隻，可以隨時隨地上網。由於手機輕巧，鍵盤滑鼠不再方便，語音可能成爲未來真正方便的輸入方式。全球語音研究界都體察到了這個新趨勢，幾年前就開始把注意力由語音聽寫文字轉向網路世界的多元化應用。我們也一樣。我們發現國語聽寫核心技術可以拆解成許多模組，不同的組合可以有不同的應用，當然還要加上網路環境下各種新的難題的克服，但應用到網路上去是水到渠成的事。

我們在86年發展完成全球第一套以國語語音搜尋網路上中文文件的技術，87年再進一步精緻化，可以自動抽取網路上中文文件的關鍵詞使語音瀏覽搜尋更爲精確，87年也發展出全球第一套以國語語音和網路對話找尋資訊的技術，當時是以電話查號爲例，是一個小型的電腦104查號台。88年完成新一代聲音品質大幅提升的中文「文句翻語音」系統，可用在手機上以聽取電子郵件及網頁內容。90年再完成以國語語音搜尋網路上的國語語音資訊(例如廣播新聞，沒有文字資訊)的技術，這不但在中文而言是全球首見，至少在公開領域中包括英文在內的其他任何語言，也尚無這樣的系統問世。我們的國語語音技術可說已成功的由個人電腦時代單一目標的語音聽寫技術轉型爲無線網路時代多元化目標的語音上網技術。由於各種技術成果豐碩，不易再以『金聲』編號，我們就以『金聲系列』爲名，作爲所有這一系列研究成果的統稱。

結語

時代不斷在進步，資訊科技的進步更是一日千里，瞬息萬變。國語的聲音卻始終如金玉之聲般悅耳美麗，萬古不變，歷久彌新。國語語音技術在資訊世界中的角色，也永遠是一個美麗的夢，在前面引導我們，一步步克服現實的困難，向前努力。Ω

這張照片出現在McGraw-Hill 2000年出版的新版教科書"Using Information Technology A practical introduction to computers and communications"(作者：B.K. Williams, S.C. Sawyer, S.E. Hutchinson)，3rd Edition，p.215，該節標題是"Input and Output"，書中在相片下的說明文字是："Taiwanese scientist Lee Lin-shan displays a computer that can listen to continuous speech in Chinese (Mandarin) and then print out the words at the rate of three characters a second"。這張相片及文字均取材自路透社(Reuters)在民國84年3月派員前往台大實驗室專訪後所發佈的新聞稿，當時刊載歐、美、亞洲各大媒體。路透社當時新聞稿的標題為"Computer Listens to , Writes Chinese"。