# 行政院國家科學委員會專題研究計畫　期中進度報告

## 中文自發性語音語料庫之建立(2/3)

計畫類別：個別型計畫
計畫編號：NSC91-2219-E-009-039-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學電信工程學系

計畫主持人：陳信宏
共同主持人：王小川，王駿發，鄭秋豫，吳宗憲，王新民，李琳山

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 6 月 2 日

# 中文自發性語音語料庫之建立(2/3)

# Spontaneous Mandarin Speech: Corpus and Processing

期中報告
計畫編號：NSC-91-2219-E-009-039
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：陳信宏　國立交通大學電信工程學系，schen@cc.nctu.edu.tw
共同主持人：鄭秋豫，李琳山，吳宗憲，王駿發，王新民，王小川

## 一、中英文摘要

本三年計畫擬建立中文自發性語音語料庫，以提供國內學術界進行先進語音辨認科技研究及產業界發展實用語音辨認系統之用。本報告說明在第二年度我們的成果，包括：(1) 新聞廣播語音之錄製及文字標示、切割等處理；(2) TAICAR-汽車環境下語音收集；(3) 新竹 IC 電台語料收集處理；(4) 中央社訪談式錄音語料處理。
**關鍵詞**：自發性語音語料庫、語音辨認、新聞廣播語音、對話語音、文字標示、切割

The three-year project aims at constructing a spontaneous Mandarin speech database to be used in academic and industrial researches for the development of advanced speech recognition technologies. In the second-year progress report, we describe the recording and processing (transcription and segmentation) of four databases: broadcast news, and in-car speech, IC Broadcast Station speech, and Central Broadcast Station dialogue speech.
**Keywords**: Spontaneous Mandarin speech database, Speech recognition, Broadcast news speech, Dialogue speech, Transcription, Segmentation.

## 二、緣由與目的

近年來朗讀語音辨認技術已有長足進步，一些實用系統陸續被開發出來，但語音辨認科技之實用化關鍵在於進一步發展自發性語音辨認技術。為因應此趨勢，本計畫結合中研院、台大、清大、成大、交大、工研院、中華電信研究所，合力建立一個中文自發性語音語料庫，以提供國內學術界進行先進語音辨認科技研究及產業界發展實用語音辨認系統之用。計畫在三年內錄製及處理大量的新聞廣播語音、對話語音及車內語音。

## 三、結果與討論：

### (一) 新聞語音語料庫之建立

本計畫準備利用三年的時間收集及處理 220 小時的新聞語音資料。預計第一年處理 40 小時的語料，第二、三年分別處理 80 小時及 100 小時的語料。

第一年度計畫執行之初的數個月主要進行各項準備工作，包括聯繫電視及廣播公司洽談授權、準備標註軟體、決定標註方式等。經與公共電視洽談後，公視同意授權我們使用其新聞節目，並建議我們採用『公視新聞深度報導』節目及願意協助我們錄音（影），錄音工作自 90 年 11 月 7 日起正式展開。『公視新聞深度報導』於每週一至五晚間 21:00-22:00 播出一個小時，自 91 年 7 月起，變更節目名稱為『公視晚間新聞』，自 91 年 9 月起，播出時間改為晚間 21:00-21:45，播出 45 分鐘，另於 21:45-22:00 播出 15 分鐘的『公視手語新聞』，92 年 1 月 31 日起，『公視晚間新聞』移至 19:00-20:00 播出，21:00-21:45 則播出『公視全球現場』，21:45-22:00 仍播出『公視手語新聞』。自 90 年 11 月 7 日起至 92 年 2 月底止，錄音時間固定為 21:00-22:00，92 年 3 月起，錄音時間則包括 19:00-20:00 及 21:00-22:00 兩個時段。本計畫錄音工作預計進行至 92 年 6 月底結束，可以收錄約 300 個小時的新聞節目，主要內容為國內新聞，也有一小部分為國際新聞。語料收集及語料保存方式簡單說明如下：：

**A. 語料收集**

1. 錄音採 TASCAM DA-40 DAT 錄音座，經由主控台在新聞播放時利用 AES/EBU 平衡式類比輸入同步錄音。
2. 錄影採 SONY SLV-ED88 錄放影機，利用一般 RCA 接頭同步錄影，錄影帶採用 TDK HS-160 型號。
3. 錄音/錄影格式
   (1) DAT tape: 格式：44.1kHz、16bit、stereo
   (2) VHS tape: stereo

**B. 語料保存**

**1. 聲音資料**

(1) 公視取回的 DAT(數位錄音帶)，經 USB 介面直接將錄音帶內的數位信號讀進 PC 內轉為格式為 44.1kHz、16bit、stereo 的聲音檔 (windows PCM、.wav)，並燒錄於光碟中以便保存。

(2) 標註使用的聲音檔，因考量檔案傳輸及讀取速度的問題，將原始的檔案，利用聲音編輯軟體 — CoolEdit 2000 將已轉為 windows PCM 的聲音檔進行格式轉換。轉換為 16kHz、16 bit、mono 後，為便利日後管理及利用，每週的公視新聞深度報導儲存於同一光碟中保存。

**2. 影音資料**

公視取回的VHS 錄影帶，經由UPMOST 301BTR類比影像擷取卡，擷取avi 格式的影像，並由影像編輯軟體 – 會聲會影(友立出品)即時壓縮成MPEG1格式保存。

**3. 語料標註**

本計畫採用 LDC(Linguistic Data Consortium)提供的Transcriber系統[1]標註電視新聞錄音資料，請參考圖一。在標註過程中，舉凡雜訊、背景環境、發音不標準、方言、說話者性別、主播/記者/被採訪者等資訊都盡量鉅細靡遺標註下來，標註的結果以XML檔案儲存，請參考圖二。標註的基本架構已於第一年度的期中報告中說明，在此不再贅述。

第一年度預計完成的第一階段40小時的語料庫已於91年7月底如期完成，本年度截至目前為止已完成約60小時的語料庫，預計第二階段的80小時的語料庫7月底前可如期完成。本年度除進行語料標註工作之外，我們也將第一年度完成的40小時語料庫進行第一次的完整修訂工作，主要是將語料庫中原標註不一致處訂定統一標準，盡量達到語料標註的一致性，另外，我們也針對此一40小時語料庫進行初步的統計分析，並寫成一篇會議論文發表於2003 IEEE&ISCA Workshop on Spontaneous Speech Processing and Recognition[2]。第一階段的40小時語料共包括40集新聞錄音及其標註檔案，每五集儲存在一片光碟，共有8片光碟，此一語料庫已送交本計畫其他共同執行單位測試中，相信很快可以授權學術單位或產業界使用。

**C. 成果討論：**

第一年度計畫執行之初的數個月主要進行各項準備工作，包括聯繫電視及廣播公司洽談授權、準備標註軟體、決定標註方式等，加上標註人需要時間熟悉標註工具及學習標註方法，所以只預計完成40小時新聞錄音語料的標註工作。第二年度因為有了前一年度的經驗，所以標註工作進行的比較順利，目前已完成約60小時的語料標註，7月底前應可如期完成80小時的語料標註。語料標註是一件非常繁複累人的工作，再加上隨時會碰到不知如何標註的特殊情況，幸好中研院語言所的鄭秋豫博士及曾淑娟博士在標註應注意事項及標註符號方面提供很多的協助，並提供咨詢服務，在此一併致謝。

## (二) TAICAR-汽車環境下語音收集

**(此項工作由本計畫、教育部 ITS 卓越計畫及成功大學共同合作進行)**

語料的收集乃是語音辨識、語言模組的一項重要工作，有大量的語料才能提供訓練模組訓練出符合實際情況的語音模型以及語言模型。而汽車環境下的語料收集尤其重要，因為在台灣目前尚未有此種資料庫，因此藉由本計畫的執行，我們進行汽車語料的收集。合作的單位有台灣大學、清華大學、交通大學、成功大學、工研院、電信研究所等六個單位，收集的語料內容主要有兩種，第一種是含有汽車環境噪音的提示卡語料，另一種是純粹的汽車環境噪音。前者可供訓練語音模組，而後者可提供為雜訊消除之有效資料。

### 1. 麥克風

我們收集的汽車語料包含了「麥克風陣列」及「高指向性麥克風」,「麥克風陣列」放置在擋風玻璃上方，為顧及安全由前方乘客來錄音模擬駕駛者，其前方放置一高指向性麥克風，另有一頭戴式高指向性麥克風。

### 2. 錄音內容

在汽車環境下的語音辨識需要考慮到其語音特性跟在室內的語音特性有相當大的差異，因此需要重新訓練語音模型。而訓練語音模型得要有所謂的「平衡語料」，這部分我們參考國內執行過的大型計畫「MAT 語料收集」之作法，先由程式從 100 萬字的文字庫中挑選出能夠涵蓋所有國語基本音節的短詞、單字等，並加上英文、數字部分，總共這樣的語料有 360 份，所佔硬碟空間有 2.65M 之多。

為了噪音環境下的語音分析所需，我們也錄製車子內的噪音，亦即，當錄音進行時，人員不得交談、說話，錄製的噪音可供日後評估噪音消除演算法所需。

- 以車輛為單位，怠速噪音錄製

60 秒

- 市區路段噪音每人錄製 30 秒
- 快速道路噪音每人錄製 30 秒

為了實際記錄各種不同路況，錄音時我們分兩種路段：市區路段以及快速道路路段。市區路段下，時速為 0～50 公里；快速道路則需維持在 70～100 公里。

之前提到的語料在錄音時將發給錄音者，我們稱之為『提示卡』。合作的單位必須負責找來 40 個人，每個人分別於不同路段各講上一節的提示卡語料一次。因此，我們將收集到各種不同車種、路況、語者的平衡語料。

而汽車本身的噪音，對於雜訊消除也是一個重大的依據，我們也同時請錄音者錄下汽車單純的環境噪音。這分成三個部分：怠速狀態下、市區路段行駛中、快速道路行駛中的汽車噪音。

### 3. 語料錄製結果

整個錄音的結果如下表所示：

| 單位 | 人數 | 語料音檔數 | 語料大小 |
|---|---|---|---|
| 台灣大學 | 21 | 15,687 | 1,558 M |
| 清華大學 | | | |
| 交通大學 | 40 | 29,880 | 3,918 M |
| 成功大學 | 40 | 29,880 | 3,347 M |
| 電信研究所 | 40 | 29,880 | 2,936 M |
| 工研院 | 40 | 29,880 | 2,939 M |
| TOTAL | 181 | 135,207 | 14,658 M |

## (三) 新竹 IC 電台交談節目語料

經過和新竹 IC 電台洽談後，我們獲得他們的許可錄製語料，經處理後去除牽涉個人隱私語料後，我們可以使用處理後之語料。因此我們開始由廣播直接錄製訪問性語料，經先處理一小部份語料進行語音辨認後，確認如此錄製之語料可以使用。

我們已錄製一些語料，目前完成處理兩個小時的節目，預計至七月底完成 10 個小時的節目語料處理。至於語料之標註處理將採用和新聞語料標註處理相同的軟體及格式。

## (四) 中央社訪談式錄音語料

從 2002 年底開始到 2003 前五月，陸續從中央社取得電台訪談的語料（屬於自發性 Spontaneous Speech），有九片光碟（共 77 個檔案，合約九小時又三十分鐘，內容以第一片為範例見附錄一），實付中央社八萬元台幣。在我們的要求下，每個檔案錄音格式為標準的 Windows wav (Linear PCM'‘)，也附有描述每段訪談的主題和對話者的性別的文字檔。訪談內容以電台廣播方式呈現，內容大多為男、女主播採訪著名專業人士，剖析其相關領域的專業知識，有的檔案有獨立的主題，有些則是具連貫性主題的系列訪談。大部分的訪談錄音中，說話速度適中，語調清晰，但是部份檔案音量偏小，主播或受訪者聲音含混，也有穿插笑聲、彼此搶話等現象。

由於主播及被訪者的交談屬於自發性語音（並非照著稿子所唸），因此常常出現各式各樣的語氣詞、停頓、口語不清、喘氣、搶話、笑聲、聲量忽大忽小、發音不清等現象，例如連續說"對對對對對"來贊同對方所講的話，發出"嗯"、"哦"、"那"、"哇"等語氣詞；在多人訪談時，時常會有搶話的現象；而節目開始前通常會先來段音樂緩和氣氛或是訪談到一半大家想休息一下時，也會播放音樂。

我們請工讀生利用中研院推薦的語料標註軟體(Transcriber)，將從這些訪談的原始錄音資料聽取到的內容標記下來，除了將說話的文字內容以繁體中文標註之外，也仔細地標註每則訪談的語氣詞、停頓、口語不清、喘氣、搶話、笑聲、聲量忽大忽小、發音不清等現象，使我們日後在自發性語音辨識研究方面能有豐富的研究資料。

除了將自發性語音中會發生的種種現象也清楚標記之外，我們也正在請工讀生將這些語音及標記作進一步的分析整理。我們計畫根據 transcriber 的標記，將每則訪談的語音檔切割成一句話一個語音檔，然後以句子為單位在資料庫(MS SQL)中分別記錄其標記特性。因為當我們想要找所有有搶話特性的句子時，如果靠人一句一句地慢慢找的話並不可行，所以我們將語氣詞、停頓、口語不清、喘氣、搶話、笑聲、聲量忽大忽小、發音不清等這些特性在資料表中分別分配一個欄位，有此特性這欄位就標成 true，沒有的話則標成 false，例如某句子有夾雜笑聲的話，就將這句子的笑聲欄位設成 true。這些欄位建立好後，我們只要利用 SQL 的 query language 就可以在短短幾分甚至幾秒中查到我們想要的檔案或語料長度等。這部分的分析整理預計在七月底完成。

# 四、計畫成果自評：

本計畫經第一年完成 40 小時廣播新聞語料之處理，本年度除繼續進行廣播新聞語料之處理外，新增車內語音、及廣播交談語料之處理，目前計畫進行順利，與預定時程相符。

# 參考文獻


[1] Barras, E. Geoffrois, Z. B. Wu, M. Liberman, "Transcriber: Development and Use of S tool for Assisting Speech Corpora Production," *Speech Communication*, 33, pp. 5-22, 2001.

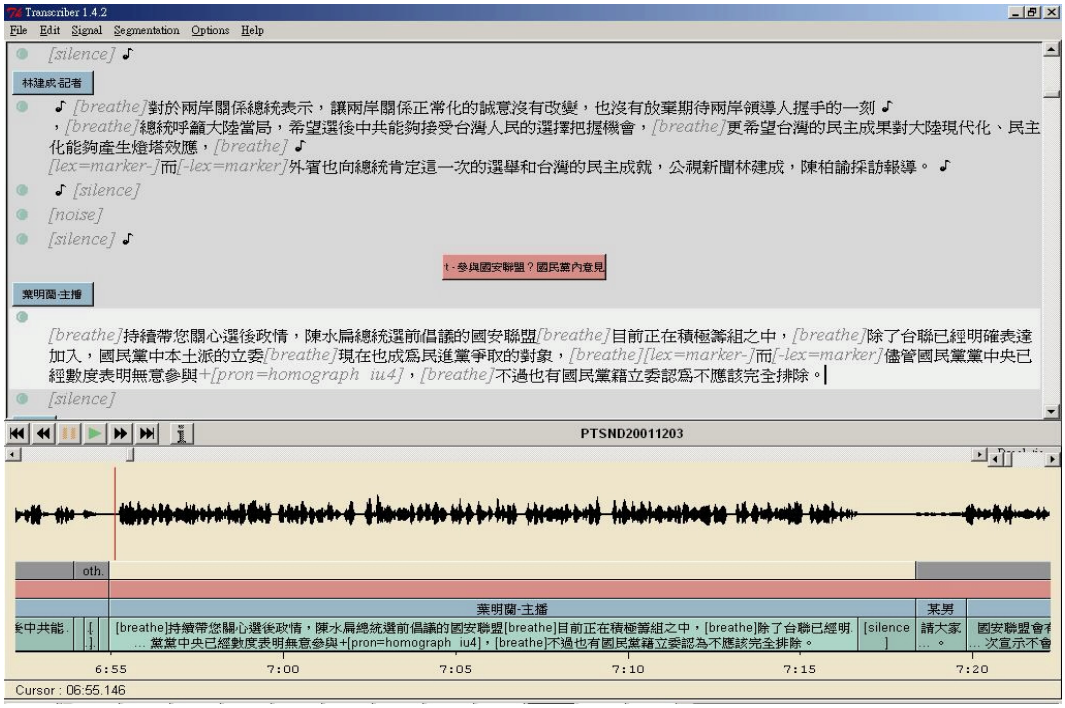
[2] Hsin-min Wang, "MATBN 2002: A

Mandarin Chinese broadcast news corpus," in Proc. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, Tokyo, April 2003.

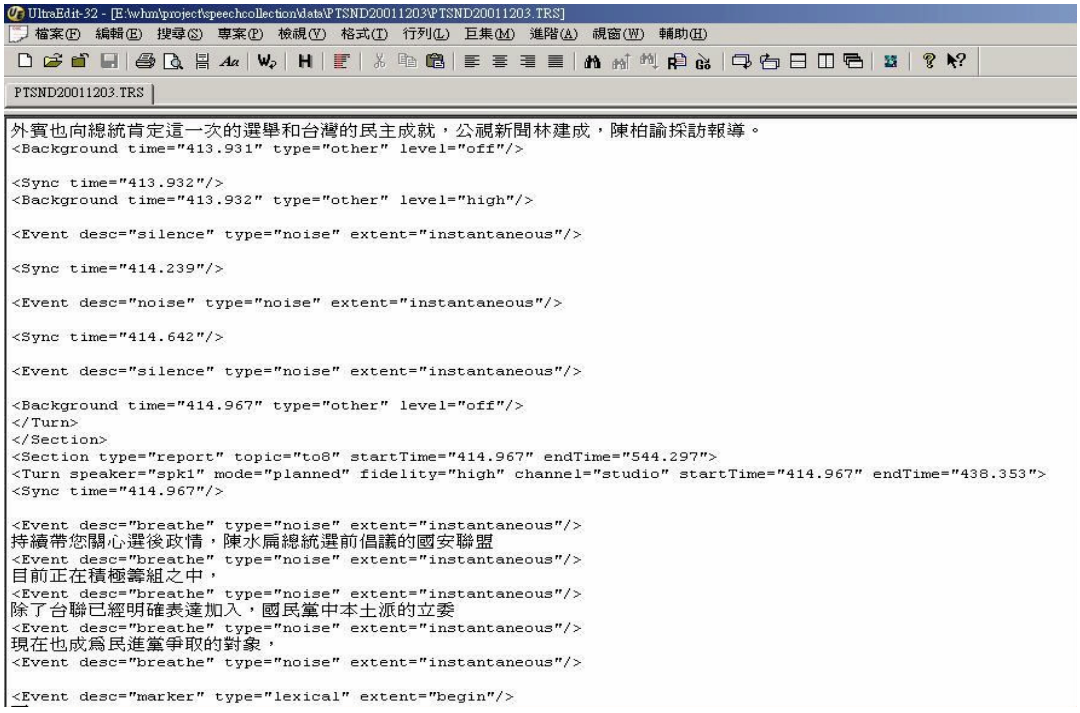
## 附錄一：中央社訪談式錄音語料第一片光碟中檔案內容資訊

**此光碟中所有檔案錄音長度共計 52'8"：**

- **主要工作經驗談(16'40"):**
  女主播採訪專業立委，對話中穿插許多語氣詞，受訪者語調平穩，聲音清晰，片尾因剪接導致說話者言論中斷。
- **國外見聞(14'42"):**
  兩位女主播的聲音不太清楚，受訪者說話速度偏快，偶爾女主播搶話導致言論中斷。
- **搭飛機經驗個人生涯規劃(20'46"):**
  片頭有非關採訪內容的對白，此檔案承續上則國外見聞，受訪者說話速度偏快，女主播的語調略嫌含混。




圖一：利用 Transcriber 標註新聞語音的實例




圖二：Transcriber 的 XML 標註檔案