中文自發性語音語料庫之建立(1/3)
Spontaneous Mandarin Speech: Corpus and Processing

期中報告
計畫編號：NSC-90-2213-E-009-109
執行期限：90 年 8 月 1 日至 91 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

一、中文摘要

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。本報告說明在第一年度我們的成果，包括：(1) 新聞廣播語音之錄製及文字標示、切割等處理；(2) 對話語料之規劃及錄製；(3) 其它語料之錄製及處理。

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

Abstract

The three-year project aims to construct a spontaneous Mandarin speech database to be used in academic and industrial researches for the development of advanced speech recognition technologies. In the first-year progress report, we describe the recording and processing (transcription and segmentation) of two databases: broadcast news speech and dialogue speech. The recording of other types of spontaneous speech, such as lecture and monologue, are also planned.

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、對話語音及演講語音。

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。第一年將處理 40 小時的語料，第二、三年分別處理 80 小時及 100 小時的語料。

經與公共電視洽談後，公視同意授權我們使用其『公視新聞深度報導』節目，並願意協助我們錄音（影），所以錄音工作自 90 年 11 月 7 日起正式展開，截至目前為止，已經收錄約 120 個小時的節目。以下就語料收集、語料保存及語料標註分別說明：

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

電視新聞錄音資料之處理是採用LDC(Linguistic Data Consortium)提供的Transcriber系統，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，限於篇幅，無法將標註細節於本報告詳述，以下僅就標註的基本架構說明：

nontrans-空白
nontrans-廣告
filler-間隔音樂
filler-節目重點內容介紹
} *見圖一*

report-新聞主題
.
. (數則新聞)
.
report-新聞主題
} *見圖二*

filler-節目重點內容介紹
nontrans-廣告
} *見圖三*

report-新聞主題
.
. (數則新聞)
.
report-新聞主題
} *同圖二*

report-氣象預報
filler-結尾
} *見圖四*

filler-片尾音樂
nontrans-廣告
nontrans-空白

C. 成果討論：

由於語料標註是一件非常繁複累人的工作，再加上計畫執行之初的數個月都在進行各項準備工作，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，截至目前為止，僅完成15捲新聞錄音語料的標註工作，預計到7月底第一年計畫結束時應可以如期完成40捲新聞錄音語料的標註工作。在計畫執行的過程中，中研院語言所的鄭秋豫老師及曾淑娟老師在標註應注意事項及標註符號方面提供很多的協助，在此一併致謝。依據第一年度的經驗，原預計第二年及第三年要完成的80小時及100小時的語料標註工作應可如期完成。

(二) 對話語料之規劃、錄製及處理

自發性對話語料種類繁多，無法進行廣泛收集，我們經過多次討論後決定本年度先收集廣播中的訪問性談話，再收集兩人交談的語料，在明年度再收集資料查詢式的對話語料。

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，我們擬先處理一小部份語料後，進行語音辨認，評估如此錄製之語料是否可以使用。

兩人交談的語料擬採由專人錄製，依選定主題自由交談，目前正接洽中廣人員參與中。

另外，我們已開始進行資料查詢式的對話語料的設計，選定以車上駕駛可能使用的資料查詢為主題，先規劃錄音方式，並試驗性錄製，以作為明年度錄製大量語料之參考。

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。

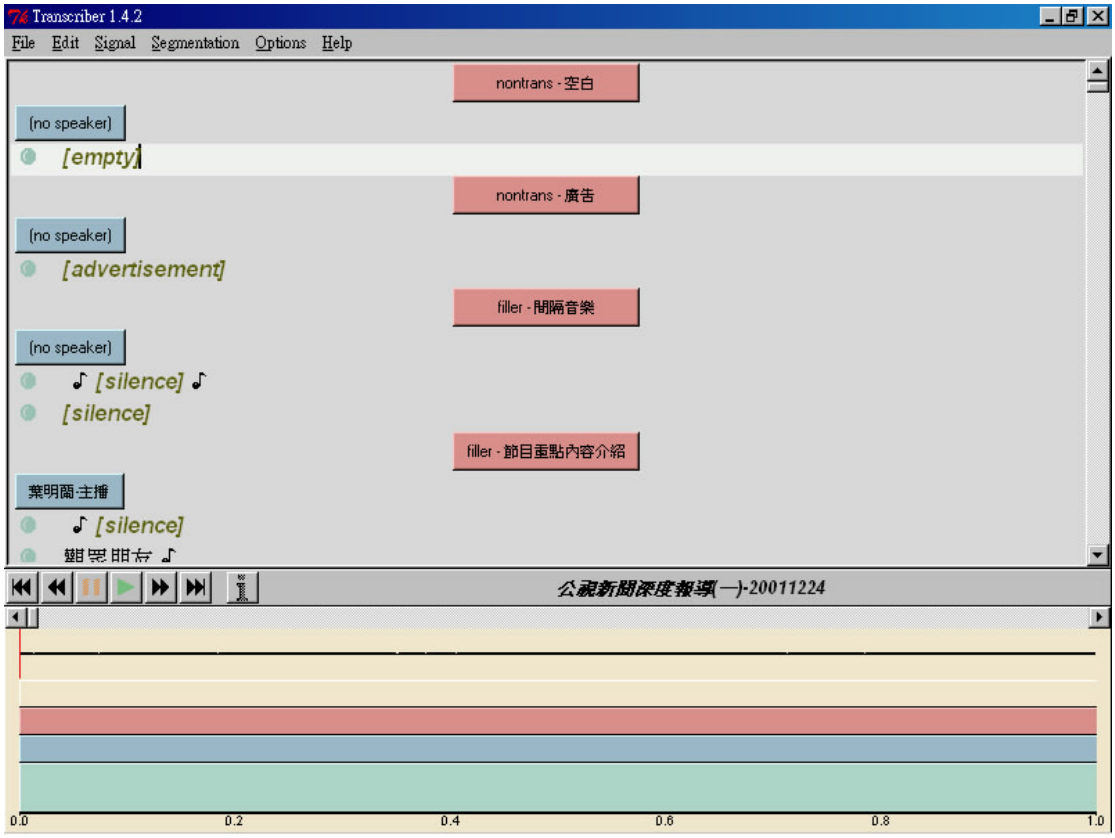
(三) 其它語料之錄製

另外，我們擬錄製一些其它種類的自發性語音，包括：專業性演講、獨白、新聞評論。本年度已開始進行專業性演講語料之錄製，由各校在 seminar 課程中邀請學者進行專業演講時，進行錄音。

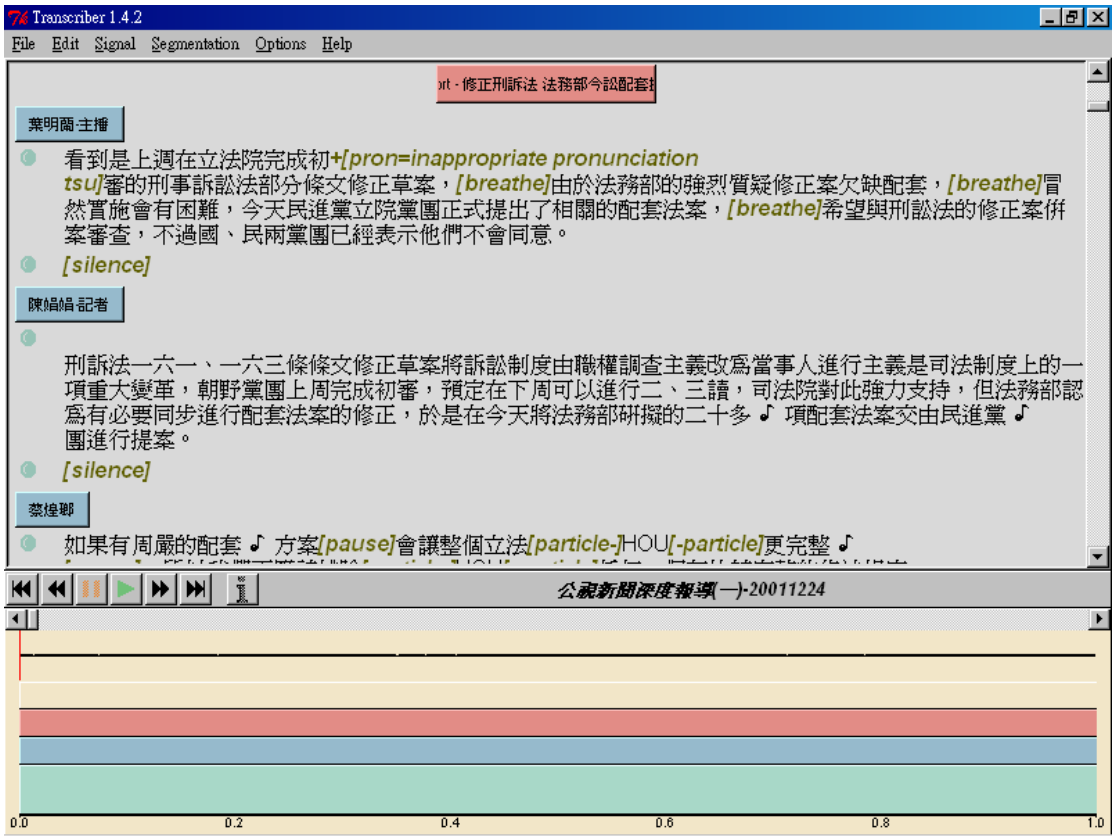
四、計畫成果自評：

本計畫擬進行之自發性語料收集及處理，由於語料種類繁多，需考慮未來應用之需求，同時需獲得語者之授權，因此在錄製語音之前，進行多次的討論以決定錄製語料的標的及方式，以及拜訪多家廣播公司以尋求授權，目前計畫進行順利，已開始錄製語料並進行處理，與預定時程相符。

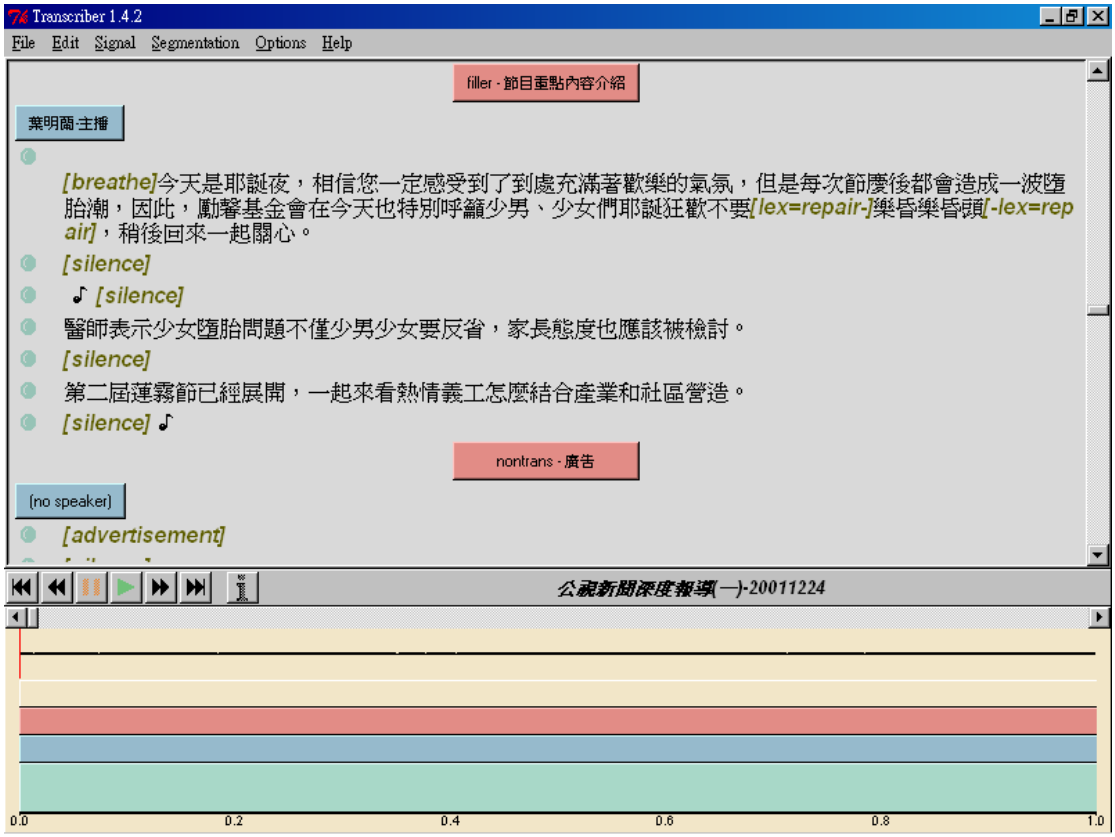
圖一


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
nontrans - 空白
(no speaker)
[empty]
nontrans - 廣告
(no speaker)
[advertisement]
filler - 間隔音樂
(no speaker)
[silence]
[silence]
filler - 節目重點內容介紹
葉明蘭 主播
[silence]
公視新聞深度報導(一)-20011224
0.0
0.2
0.4
0.6
0.8
1.0

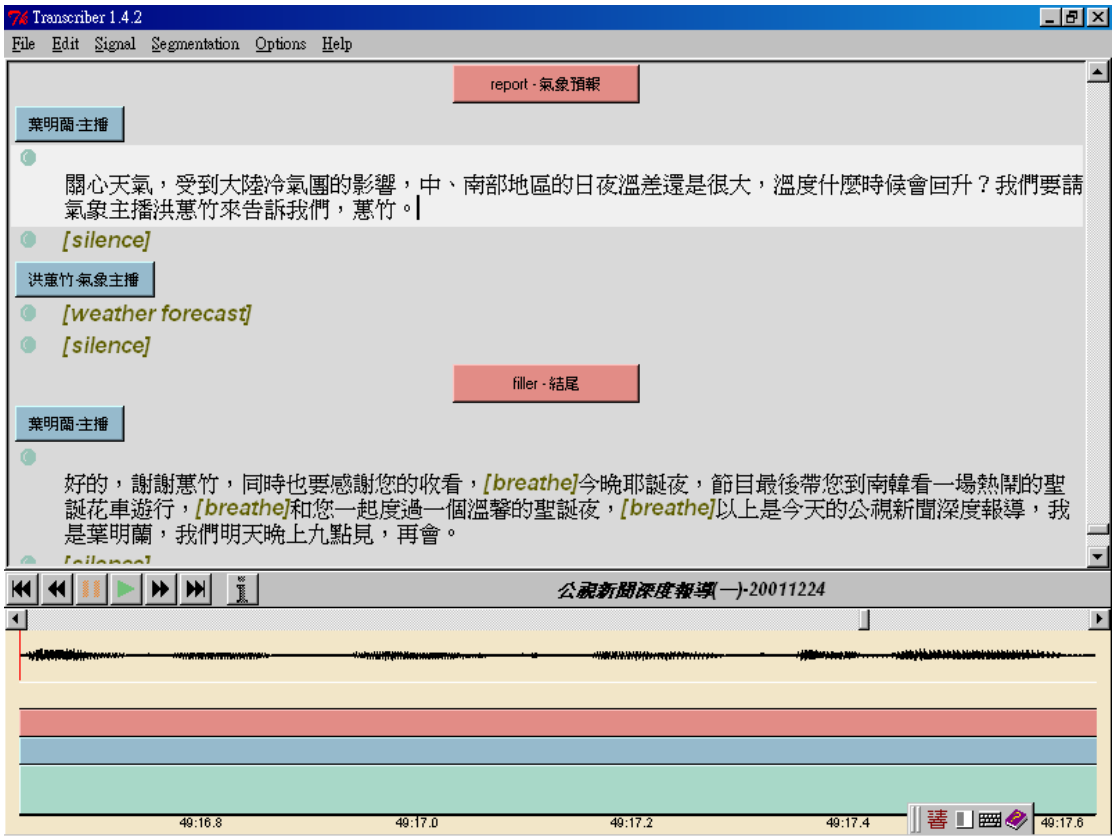
圖二：


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
葉明蘭 主播
看到是上週在立法院完成初+[pron=inappropriate pronunciation tsu]審的刑事訴訟法部分條文修正草案，[breathe]由於法務部的強烈質疑修正案欠缺配套，[breathe]冒然實施會有困難，今天民進黨立院黨團正式提出了相關的配套法案，[breathe]希望與刑訟法的修正案併案審查，不過國、民兩黨團已經表示他們不會同意。
[silence]
陳嫦娥 記者
刑訴法一六一、一六三條條文修正草案將訴訟制度由職權調查主義改爲當事人進行主義是司法制度上的一項重大變革，朝野黨團上周完成初審，預定在下周可以進行二、三讀，司法院對此強力支持，但法務部認爲有必要同步進行配套法案的修正，於是在今天將法務部研擬的二十多 ♪ 項配套法案交由民進黨 ♪ 團進行提案。
[silence]
蔡煌瑯
如果有周嚴的配套 ♪ 方案[pause]會讓整個立法[particle-]HOU[-particle]更完整 ♪
公視新聞深度報導(一)-20011224
0.0
0.2
0.4
0.6
0.8
1.0

圖三：


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
filler - 節目重點內容介紹
葉明蘭-主播
[breathe]今天是耶誕夜，相信您一定感受到了到處充滿著歡樂的氣氛，但是每次節慶後都會造成一波墮胎潮，因此，勵馨基金會在今天也特別呼籲少男、少女們耶誕狂歡不要[lex=repair-]樂昏樂昏頭[-lex=repair]，稍後回來一起關心。
[silence]
[silence]
醫師表示少女墮胎問題不僅少男少女要反省，家長態度也應該被檢討。
[silence]
第二屆蓮霧節已經展開，一起來看熱情義工怎麼結合產業和社區營造。
[silence]
nontrans - 廣告
(no speaker)
[advertisement]
公視新聞深度報導(一)-20011224
0.0
0.2
0.4
0.6
0.8
1.0

圖四：


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
report - 氣象預報
葉明蘭-主播
關心天氣，受到大陸冷氣團的影響，中、南部地區的日夜溫差還是很大，溫度什麼時候會回升？我們要請氣象主播洪蕙竹來告訴我們，蕙竹。
[silence]
洪蕙竹-氣象主播
[weather forecast]
[silence]
filler - 結尾
葉明蘭-主播
好的，謝謝蕙竹，同時也要感謝您的收看，[breathe]今晚耶誕夜，節目最後帶您到南韓看一場熱鬧的聖誕花車遊行，[breathe]和您一起度過一個溫馨的聖誕夜，[breathe]以上是今天的公視新聞深度報導，我是葉明蘭，我們明天晚上九點見，再會。
公視新聞深度報導(一)-20011224
49:16.8
49:17.0
49:17.2
49:17.4
49:17.6