MATBN 中文廣播新聞語料庫 簡介

民國 90 年至 93 年間，國內從事語音研究之相關學校及研究單位聯合向國科會中提出一語料蒐集計畫－『中文自發性語音語料庫之建立』(Spontaneous Mandarin Speech: Corpus and Processing；計畫編號：NSC-92-2213-E-009-021)，參與的單位共有國立交通大學電信工程學系、國立台灣大學電機工程學系、國立清華大學電機工程學系、國立成功大學電機工程學系、中央研究院資訊研究所、工研院前瞻研究中心及中華電信研究所。在該項『中文自發性語音語料庫之建立』計畫中收錄了新聞語音資料，語料來源由公共電視台之新聞，計畫中並對所蒐集之197個小時節目之語音資料做人工文字標註(轉記；transcription）處理。會讓該項成果可與國內外從事國語語音研究之單位分享，今將申請將上述之研究成果申請技轉中華民國計算語言學學會(The Association for Computational Linguistics and Chinese Language Processing ; ACLCLP)。

<u>語音資料庫相關資訊</u>

1. 授權語音資料庫共包含197個小時之語音資訊及其內容之轉記與語音資訊之標註資料；共五片 DVD 光碟。

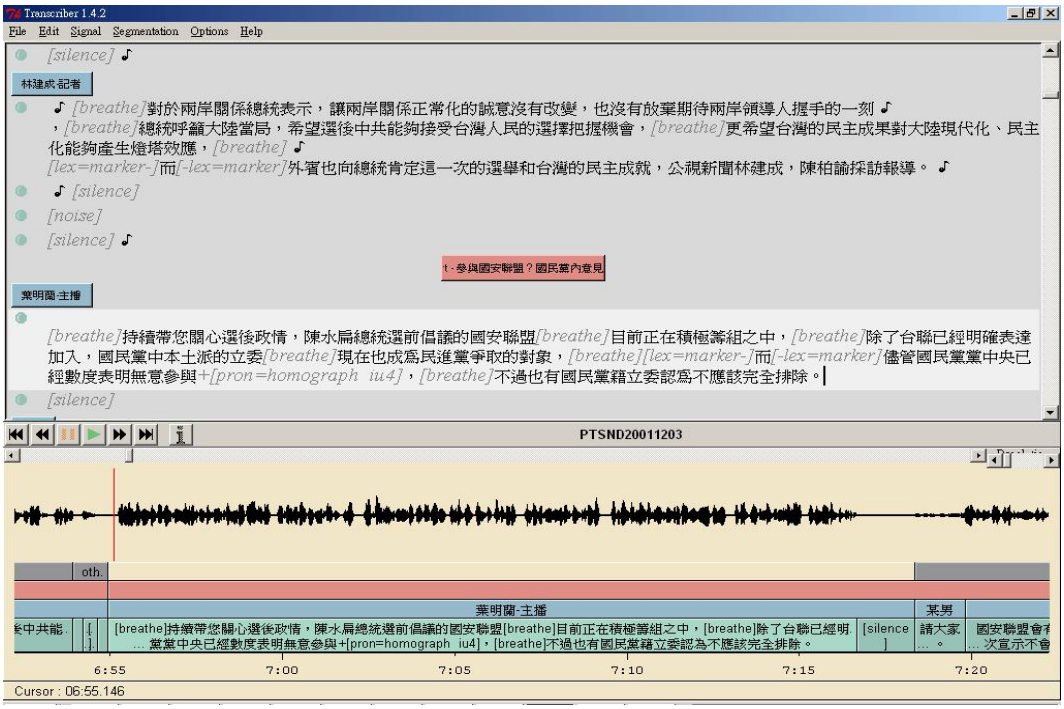
2. 錄製語料的相關資訊：

   所有的新聞語音都在公視錄製的，在第一年度計畫執行之初的數個月先進行各項準備工作，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。本計畫錄音工作進行至 92 年 6 月底結束，共

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。語料已轉為 windows PCM 的聲音檔，其規格為為 16kHz、16 bit、單軌聲音檔。

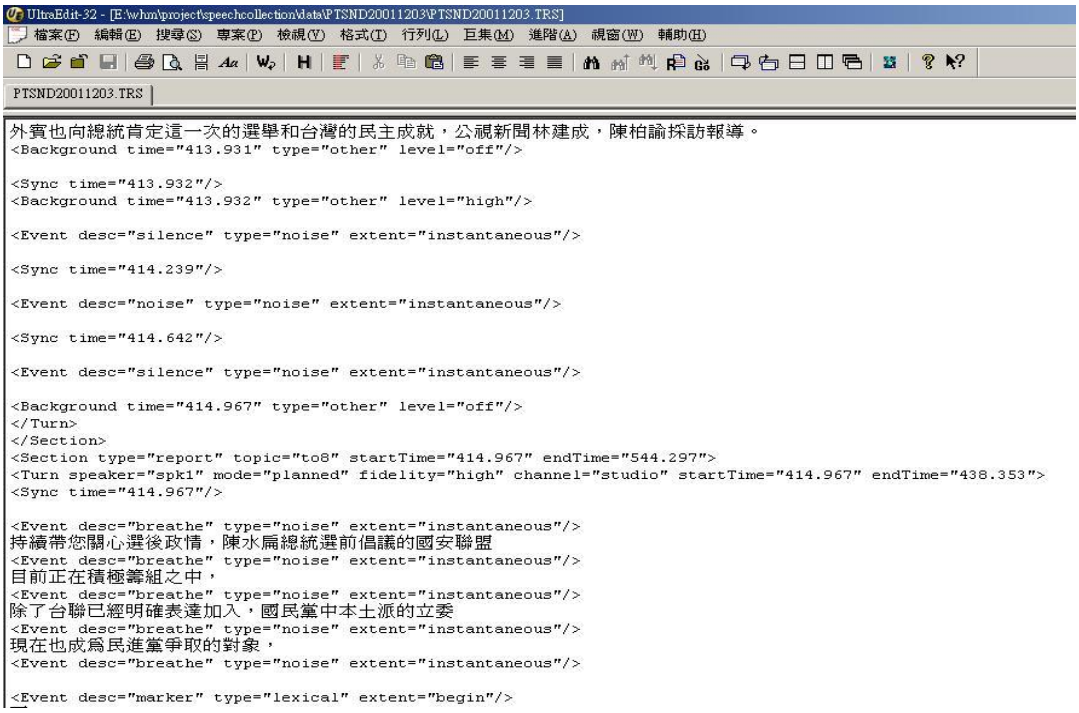
3. 語料資料之人工文字標註資料

語音的標註是由受過訓練的專任助理進行較準確的文字標註，並做 cross checking，此部分工作由中央研究院資訊所王新民博士統籌規劃及負責推動。我們採用 LDC (Linguistic Data Consortium) 提供的 Transcriber 系統來標註電視新聞錄音資料，請參考圖一。首先，將公視取回的 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。然後，將檔案轉成標註使用的聲音檔，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，每月的公視演講廳、客家新聞雜誌分別儲存於同一光碟中保存。



圖一：利用 Transcriber 標註新聞語音的實例

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。



圖二：Transcriber 的 XML 標註檔案

標註重點主要分為四大部分，分別為：

- 段落主題
- 說話者名稱
- 背景聲音
- 插入事件

『公視新聞深度報導』於 Transcriber 系統大致上之基本架構包含：

nontrans-空白
nontrans-廣告
filler-間隔音樂
filler-節目重點內容介紹
} *見圖三*

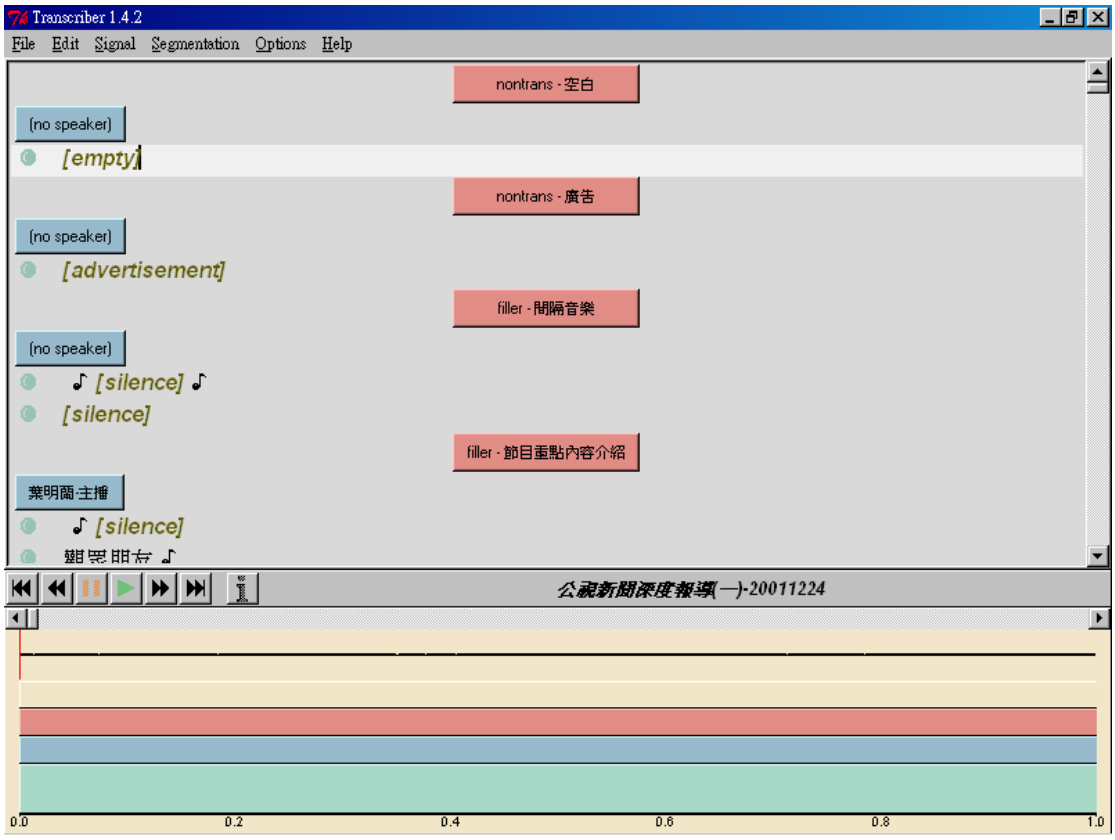
report-新聞主題
.
. (數則新聞)
.
report-新聞主題
} *見圖四*

filler-節目重點內容介紹
nontrans-廣告
} *見圖五*

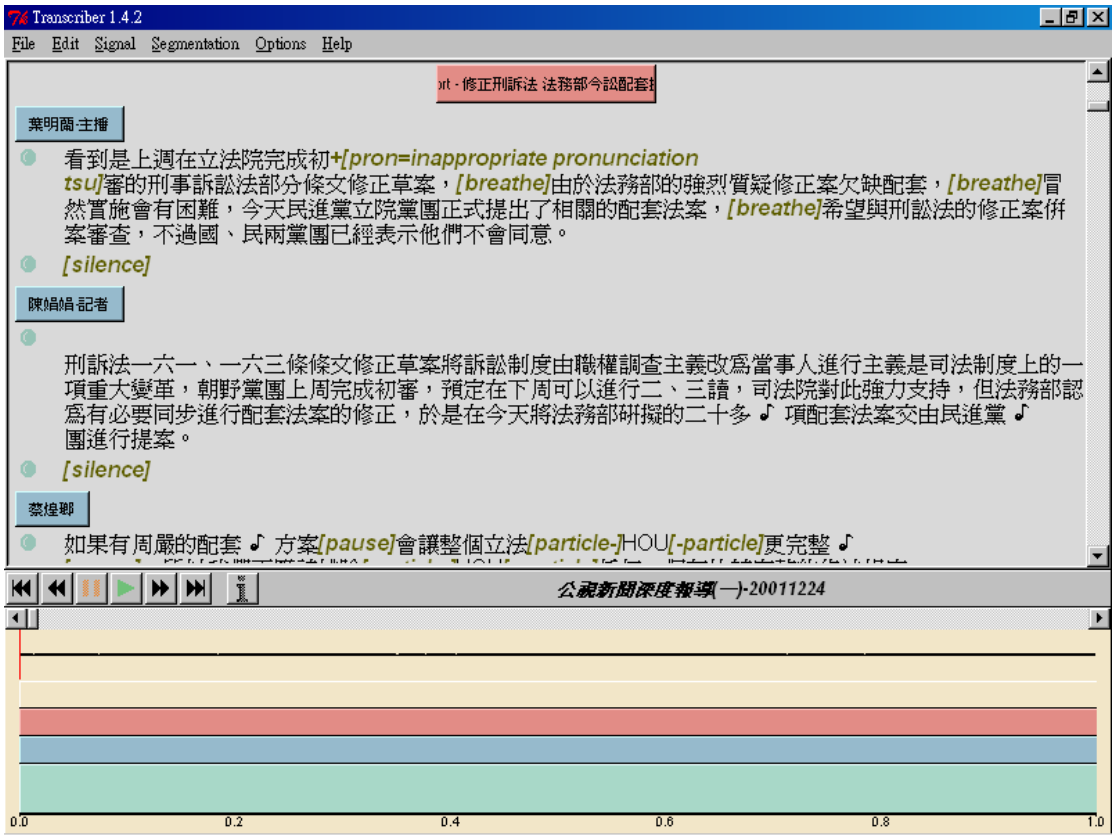
report-新聞主題
.
. (數則新聞)
.
report-新聞主題
} *同圖四*

report-氣象預報
filler-結尾
} *見圖六*

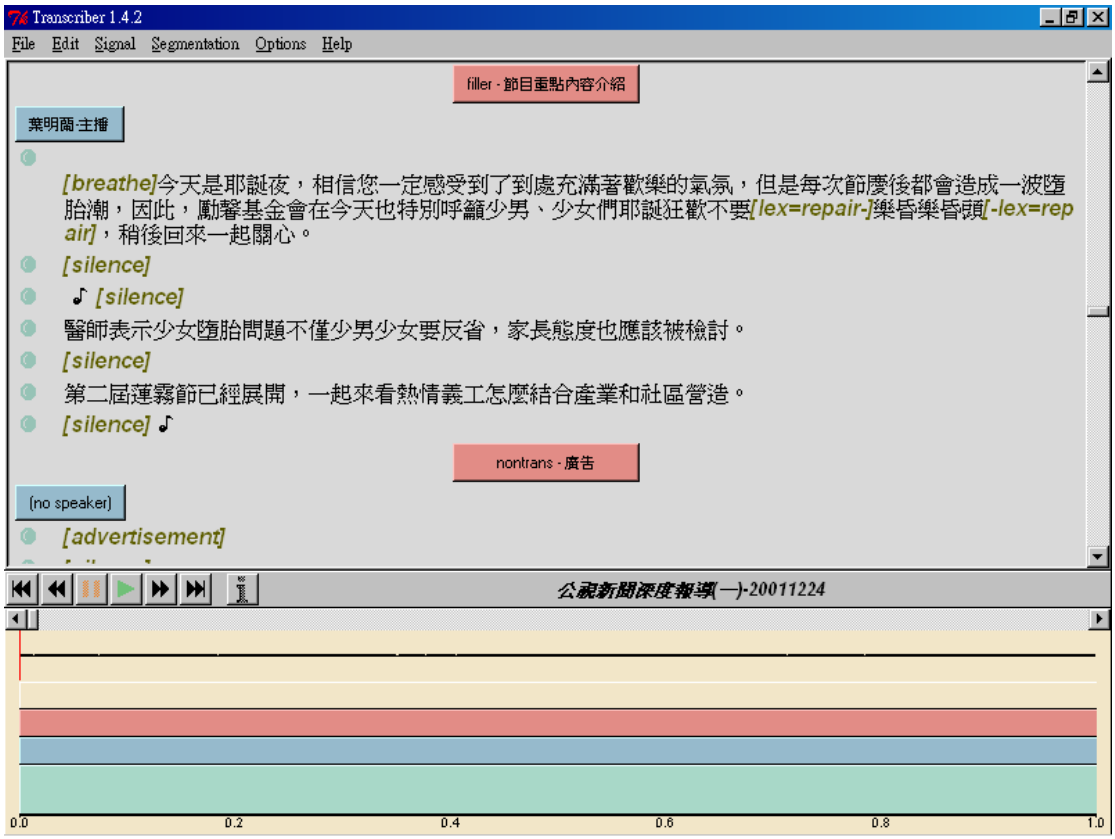
filler-片尾音樂
nontrans-廣告
nontrans-空白
} *見圖七*



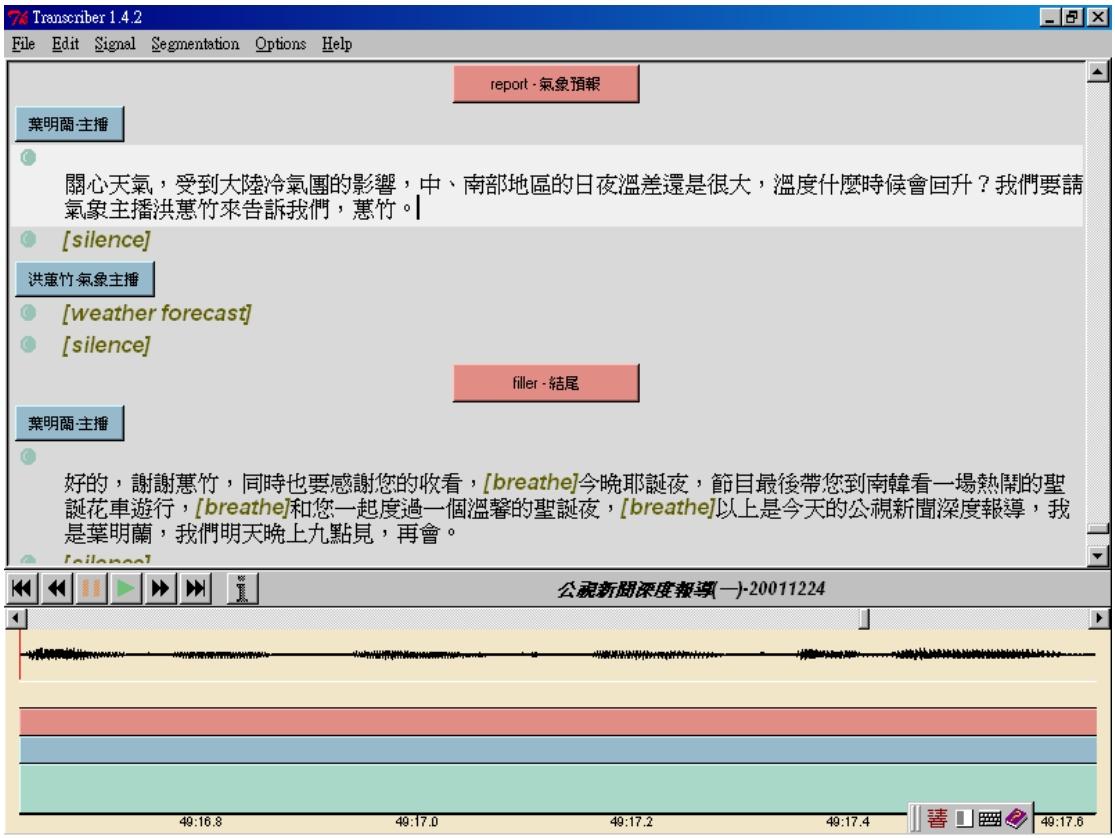
圖三


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
ort - 修正刑訴法 法務部今設配套
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]
陳娟娟-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

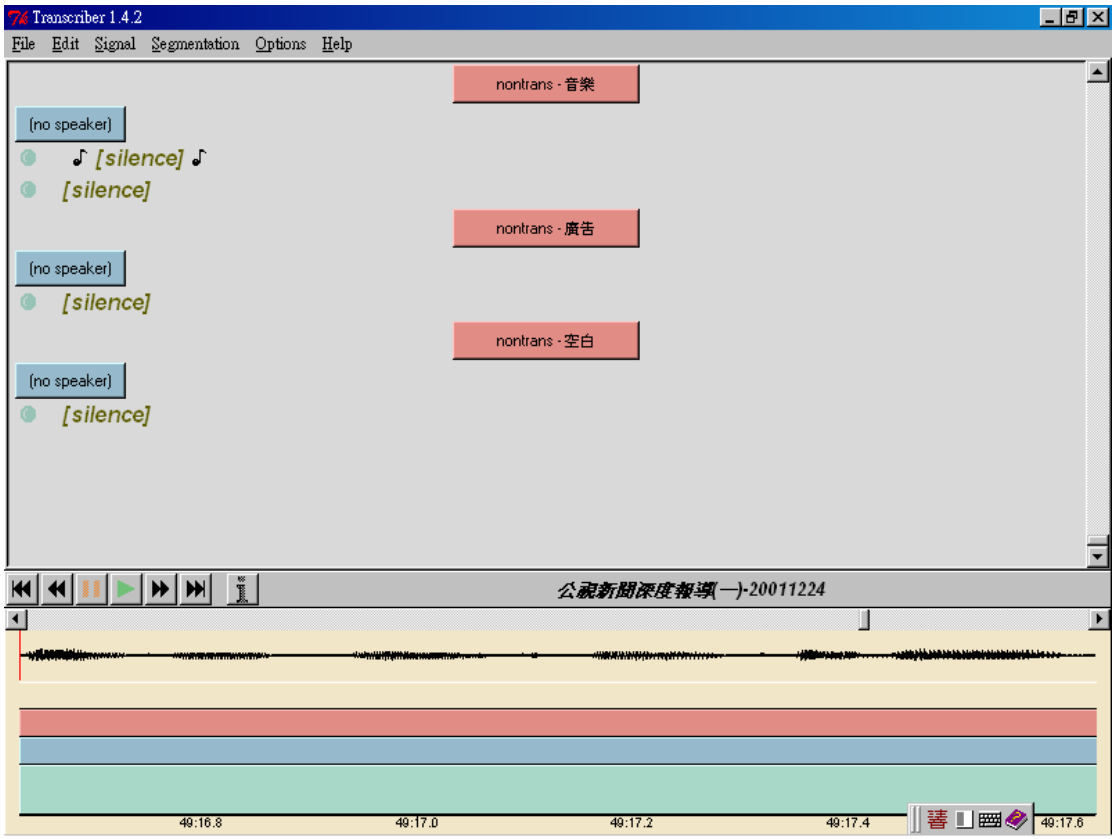
圖四


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

圖五


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

圖六


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
nontrans - 音樂
(no speaker)
[silence]
[silence]
nontrans - 廣告
(no speaker)
[silence]
nontrans - 空白
(no speaker)
[silence]
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

圖七