National Taipei University of Technology Audiobook Vol. 1 (NTUT-AB01)

廖元甫

台北科技大學電子系
yfliao@ntut.edu.tw

目的：

爲了研究語音的篇章韻律現象，此語料庫設計以整本書籍爲文本，並以文章段落爲單位，錄製整本書籍的語音語料庫，以呈現篇章韻律現象。此外，在此語料庫中，亦補充錄製從網路上收集到的常見中英夾雜文句，以呈現中英夾雜時的語音韻律變化，與錄製一份做爲對比用的英文音素平衡語料庫。此語料庫已被用於實現一個中英夾雜語音合成系統，有相關資料可供參考。

語料包括：

1. 中文－幾乎全中文的書籍《理所當爲：成就公平正義的社會》，作者爲洪蘭，遠流出版公司。
2. 英文－重新錄製 CMU（Carnegie Mellon University）-ARCTIC 語料庫，總共 1132 句全英文句子，此語料庫其音素資料量很平均，是一個標準英文語料庫。
3. 中英夾雜－文本由工研院提供，主要是爲了彌補中英轉換間的連貫文脈資訊，他們擷取網路上中英夾雜的句子，在此我們以人工選用 3050 句來錄製。

錄音資訊：

此語料庫的中文、英文及中英夾雜之語料皆由同一語者所錄製（台灣數位有聲書推展學會－秘書長　李秀鳳小姐），詳細語料錄製資訊如下。

語者性別	女性	語者總數	1 位
母語	國語	語料錄製媒介	麥克風
原始取樣頻率	44.1kHz	語料錄製環境	錄音室
切割取樣頻率	16kHz	合成取樣頻率	48kHz
語料錄製工具	DAISY 製作軟體 My Studio PC		

處理過程：

1. 將文字正規化（人工），將所有符號、數字轉成 Big-5 編碼。
2. 利用 Stanford Word Segmenter 工具，將語料庫中文文本斷詞。
3. 將斷詞文本轉爲漢語拼音串及英文音素串。
   - 中文部分：比對中文辭典與漢語拼音對照表（chinese.hanyu.wpa.txt），將斷詞結果轉成漢語拼音串。
   - 英文部分：經由 Flite（TTS 系統）工具，產生以 Alphabet 爲標準之英文音素串。
4. 將中文的漢語拼音串及 Flite 產生之 Alphabet 英文音素串整合，並依 X-SAMPA 編碼（音素對照表），統一轉換成用 X-SAMPA 系統標註的音素串，最後轉成 HTK 可接受之內碼格式（音素對照表）。
5. 利用中英雙語辨認器做 Forced Alignment 取得音素切割資訊。
   - 中文部分：使用TCC300的語料來做辨認器，這個語料所使用的音素單位爲子母音，其漢語拼音音節錯誤率爲31.82%，經過轉換改成以X-SAMPA音素爲單位的整合音素集，而後訓練成中文音素辨認器，其X-SAMPA音素錯誤率爲56.33%。
   - 英文部分：使用了 TIMIT 的語料來做辨認器，TIMIT 語料庫內容共 6300 句，將原始 Alphabet 音素集對應到 X-SAMPA 整合音素集後訓練一個英文辨認器出來，其原始 Alphabet 音素集的辨識錯誤率爲 36.89%，轉換後的 X-SAMPA 音素辨認錯誤率爲 42.62%。

光碟內容：

依據中文、英文及中英夾雜分別燒成三片光碟，即 NTUT-AB01-CH、NTUT-AB01-EN、NTUT-AB01-CE。其目錄資料如下：

Label-forced-alignment 資料夾：
透過強制對位（Forced Alignment）來取得各 X-SAMPA 音素的邊界資訊，每個 label 檔存放內容爲對應音檔內所有音素的切割時間，單位爲 10^{-7}sec。

Label-checked 資料夾：
修改完之 X-SAMPA 音素的切割時間，單位爲 10^{-7}sec，我們利用 Wavesurfer 軟體來讀取原始的 label 檔做修改，針對波型及頻譜在 unvoiced 的地方增加 sp 標籤，以及參考頻譜與波形圖，調整子音與母音的初始位置及結束位置。此外，在文本轉音素錯誤的地方（念法不同：破音字或英文縮寫）做音素的修改。

Pitch-checked 資料夾：
修改完之音檔 pitch，我們利用 Wavesurfer 軟體來讀取原始的音檔，取樣頻率選取 48kHz，對音檔的 pitch 做計算，音高設定爲 80~340Hz 之間，取樣間隔爲 0.005sec，計算完後做人工修改，並依據波形與頻譜，將 unvoiced 處的 pitch 做刪除。

Txt 資料夾：
音檔對應之文字檔，與音檔內容一致，採用人工正規化的方式，將所有符號、數字改成全形符號（Big-5）。

Wav 資料夾：

- NTUT-AB01-CH－中文語料之原始音檔，全書錄製約 500 分鐘，每一個音檔是一個文章段落，所以長度不一。共 85 個音檔。
- NTUT-AB01-EN－英文語料之原始音檔，語料錄製約 105 分鐘，每一個音檔長度約 40 句，大概 3-4 分鐘。共 25 個音檔。
- NTUT-AB01-CE－中英夾雜語料之原始音檔，語料錄製約 376 分鐘，每一個音檔長度不一。共 32 個音檔。

Toolboxes 資料夾：
存放 Flite 及 Stanford-Segmenter 工具。

- flite-1.4-release：將英文斷詞部分轉成音素串。
- stanford-segmenter-2012-03-09：斷詞器工具，對正規化完之中文文本做斷詞。

辭典與音素表：

a chinese.hanyu.wpa.txt：中文辭典與漢語拼音對照表。
b dict-mono-sp：拼音與音素對照表。
c 音素對照表：X-SAMPA phone set 與 HTK 使用之 Code set 對照表。

References 資料夾：

參考資料。