National Education Radio (NER) Corpus

Yuan-Fu Liao
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

Taiwanese Across Taiwan

1. Description

Taiwanese Mandarin has many notable differences from Putonghua in China, such as writing system, pronunciation, accent, wording, and vocabulary. Many of the differences can be attributed to the influences from Taiwanese, Hakka, Formosan, Dutch and Japanese languages. Therefore, it is well understood that a Taiwanese-specific automatic speech recognition (ASR) system is required for better speech-enabled human-computer interaction in Taiwanese people's daily life.

Therefore, we had built the National Education Radio (NER) corpus which is a real-life, multi-genre and spontaneous Taiwanese Mandarin broadcast speech corpus with manual transcription from the digital archive of Taiwan's National Education Radio. NER is the largest Taiwanese Mandarin spoken corpus that has 21-volume, 3200-hour speech data. Besides, it is also the largest Chinese spoken text (instead of writing text) database with about 60 million traditional Chinese characters.

2. Summary of the NER corpus

Sets	#. of hours	#. of files	#. of words	#. of characters
clean	624.71	89,908	6,638,286	12,983,842
other	2,581.07	400,495	19,362,579	47,800,272
Total	3,205.78	490,403	26,000,865	60,784,114

3. File Format

Audio file format: WAVE, 16 kHz sampling rate, 16 bits PCM

`YX_20170316_048.wav`

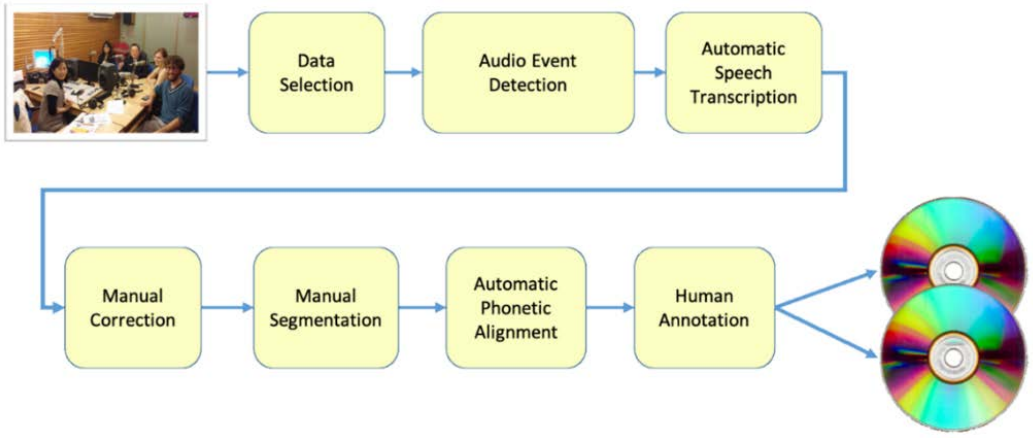
Transcription file format: text

YX_20170316_048.txt

那 其實 也 是 老師 一 個 很 大 的 挑戰 最後 想要 來問 一下 我們 的 小芬 老師 其實 工作 過程 裡頭 有 辛苦 當然 也 有 一 些 這 個 收穫 但是 您 覺得 啊 走 了 十一年 的 時間 可以 這樣 堅持 這麼 長 的 時間 同時 還可以 讓 你 繼續 想要 往下 走 繼續 投身 這 個 幼教 行業 跟 領域

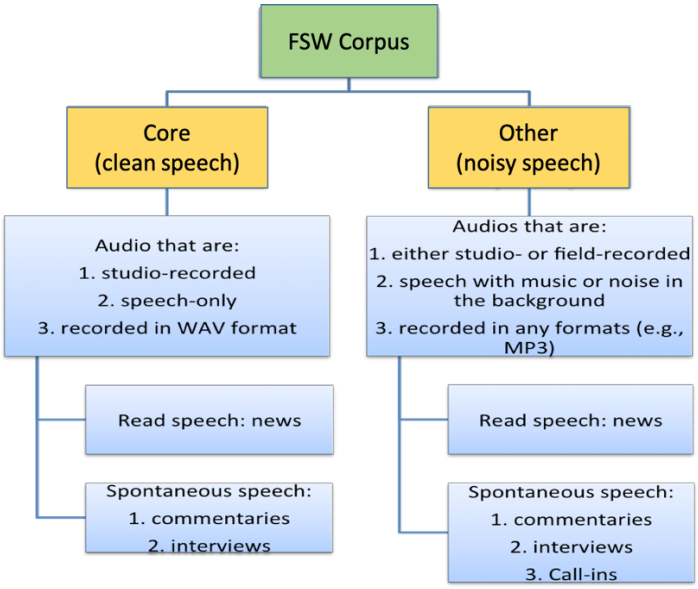
4. Corpus Construction

Seven steps are involved in processing NER's speech data including:
(1) data selection
(2) audio event detection and segmentation (optional)
(3) automatic speech transcription
(4) manual correction
(5) manual segmentation
(6) automatic phone alignment (optional)
(7) human annotation (optional)



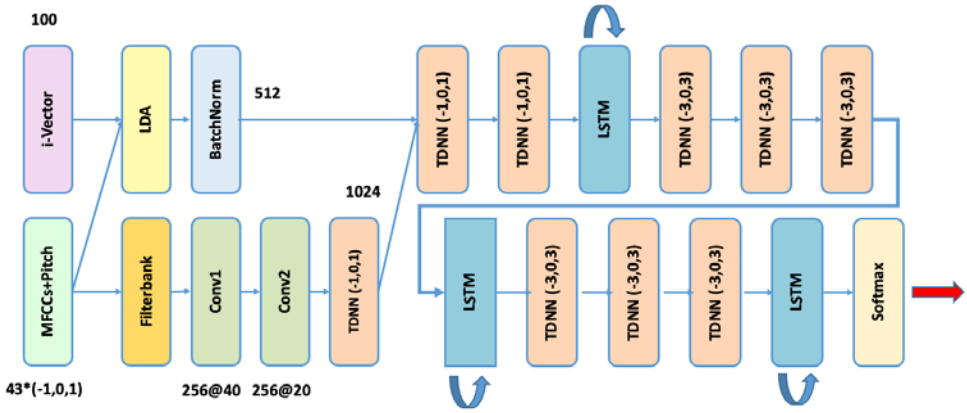
5. Data Selection

The NER has archived radio broadcast programs over decades. However, we only considered the audio recordings produced in the past five years and used the file format and quality to divide NER's archive into two subsets: the "Core" subset and the "Other" subset. The speech samples in the "Core" subset are studio-recorded and stored in the WAV format, while those in the "Other" subset may be mixed with field or telephone interviews or recorded in a compromised condition and processed with lossy compression algorithm. The detail selection method is as follows:

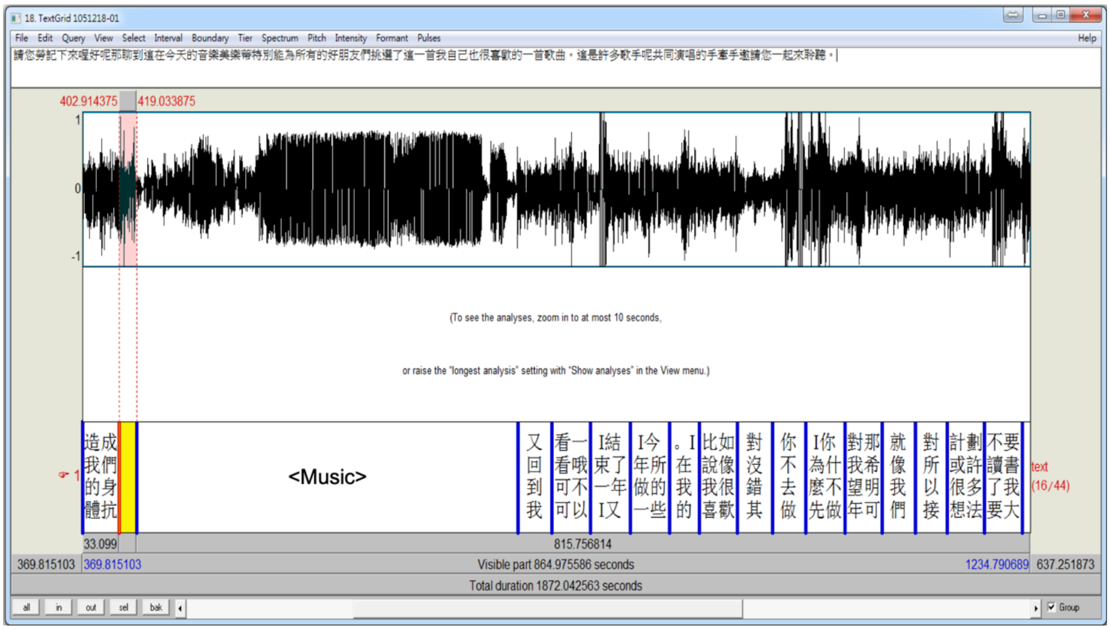


6. Semi-Automatic Transcribing

NTUT's multilingual (mixed Chinese and English) hybrid Hidden Markov Model/Deep Neural Network (HMM/DNN) system is utilized to transcribe NER's speech data. The Figure below shows the architecture of the underlying DNN-based acoustic model.



After automatic transcription generation, about 50 native speakers were recruited and divided into two groups. The first group was asked to listen to the extracted speech segments and correct the transcriptions word-by-word (first revision). The second group checked the revised transcripts again (second revision). Along with the corresponding audio, the outputs were segmented into about 15-second-long utterances. The final labelling output is as follows:


18. TextGrid 1051218-01
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
請您勞記下來喔好呢那聊到這在今天的音樂美樂蒂特別能為所有的好朋友們挑選了這一首我自己也很喜歡的一首歌曲。這是許多歌手呢共同演唱的手牽手邀請您一起來聆聽。
402.914375
419.033875
1
0
-1
(To see the analyses, zoom in to at most 10 seconds,
or raise the "longest analysis" setting with "Show analyses" in the View menu.)
<Music>
text
(16/44)
33.099
815.756814
369.815103
369.815103
Visible part 864.975586 seconds
1234.790689
637.251873
Total duration 1872.042563 seconds
all
in
out
sel
bak
Group