Summary

YUAN-FU LIAO
NATIONAL TAIPEI UNIVERSITY OF TECHNOLOGY

Universal Human-Machine Interface

Voice Assistant

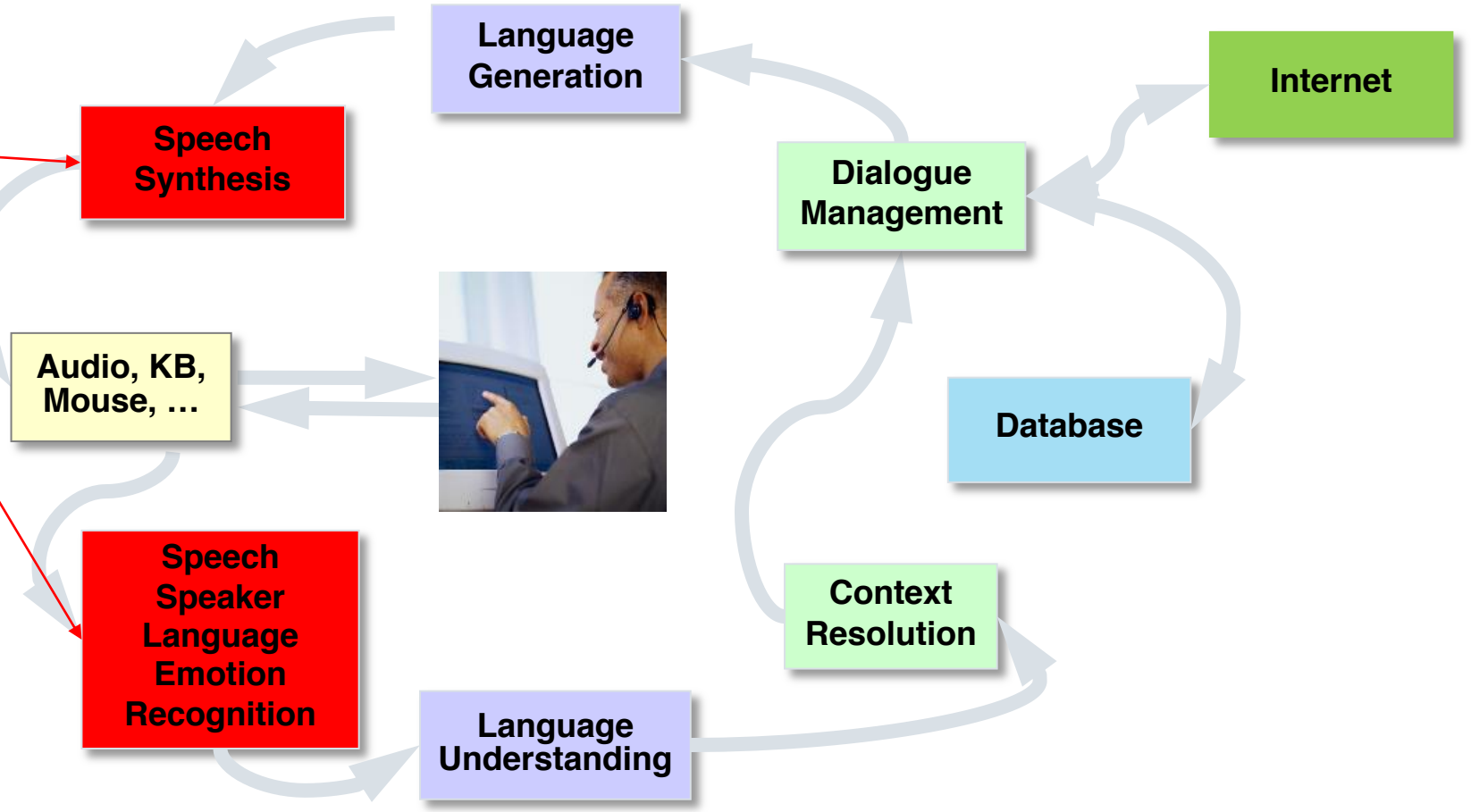
- Who (Jarvis?) always listens for your call, anticipates your every need, and takes action when necessary

VS.

Toward Universal Human-Machine Interface

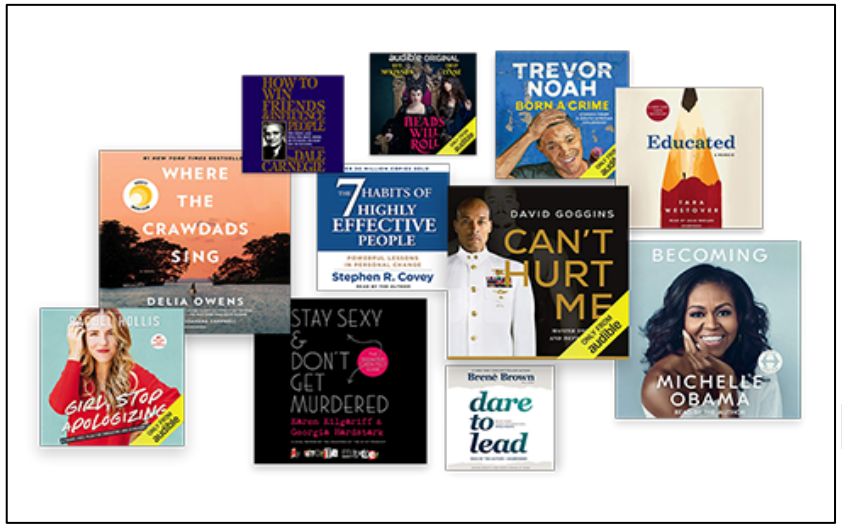
Speech Recognition

Speech Synthesis



Voice Bank 未來推廣

- iVoice.tw
  - 與漸凍人協會合作
  - 公司支援
  - 即時字幕



聲音日記社交網站

Indexing

Google

網路搜尋檢索

個人語音電子書下載
即時字幕應用
會議記錄電子化

自動語音逐字稿轉寫

合成語音音檔與模型下載

即時配音
視障輔具

網路媒體
網路課程
會議記錄

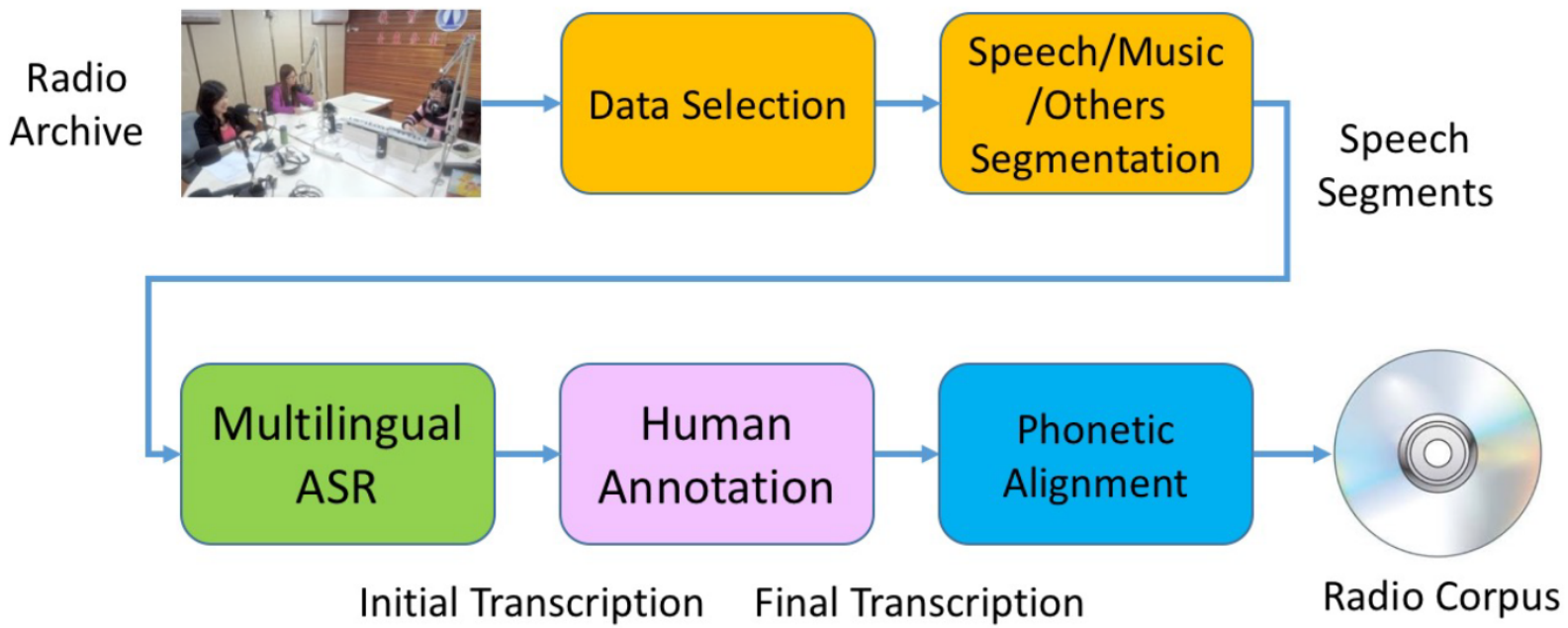
個人語音日記

個人合成語音應用

- URL: https://sites.google.com/speech.ntut.edu.tw/fsw

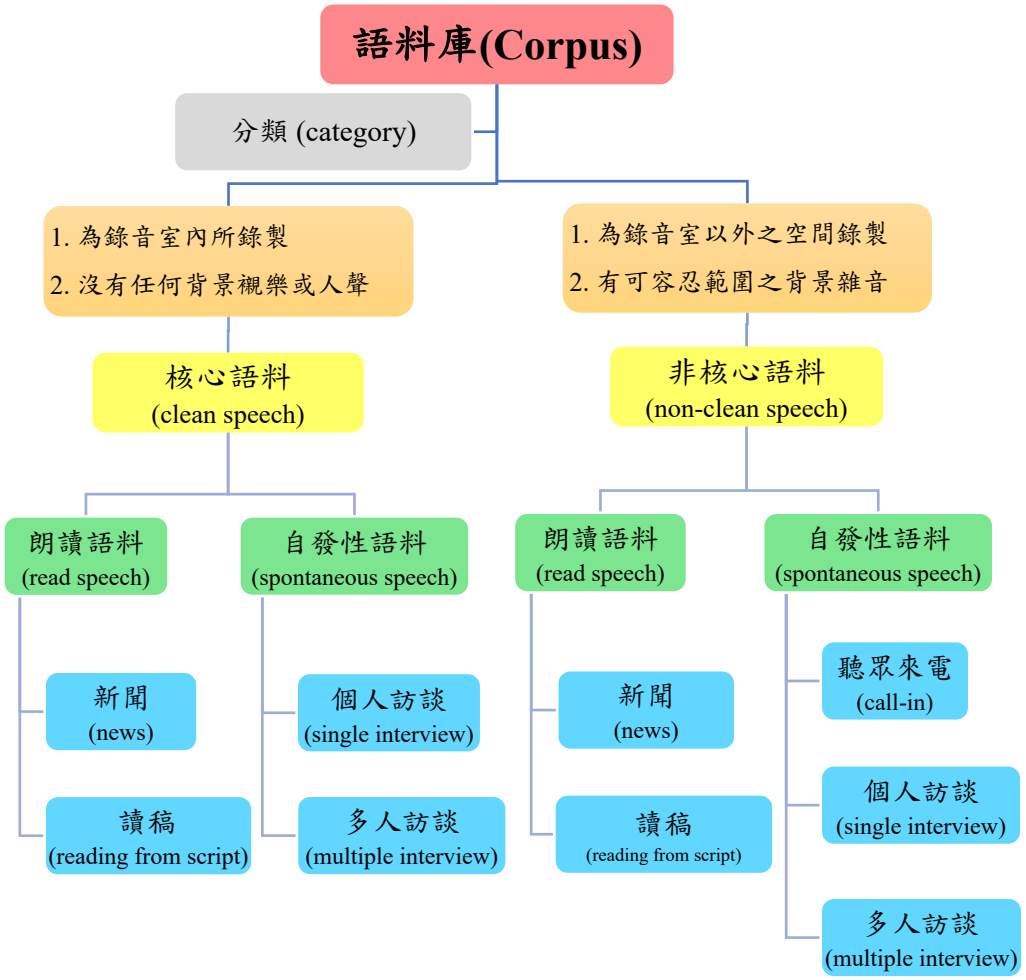
Semi-Automatic Corpus Processing

- National Education Radio Archive
- Reduce annotation effort as much as possible



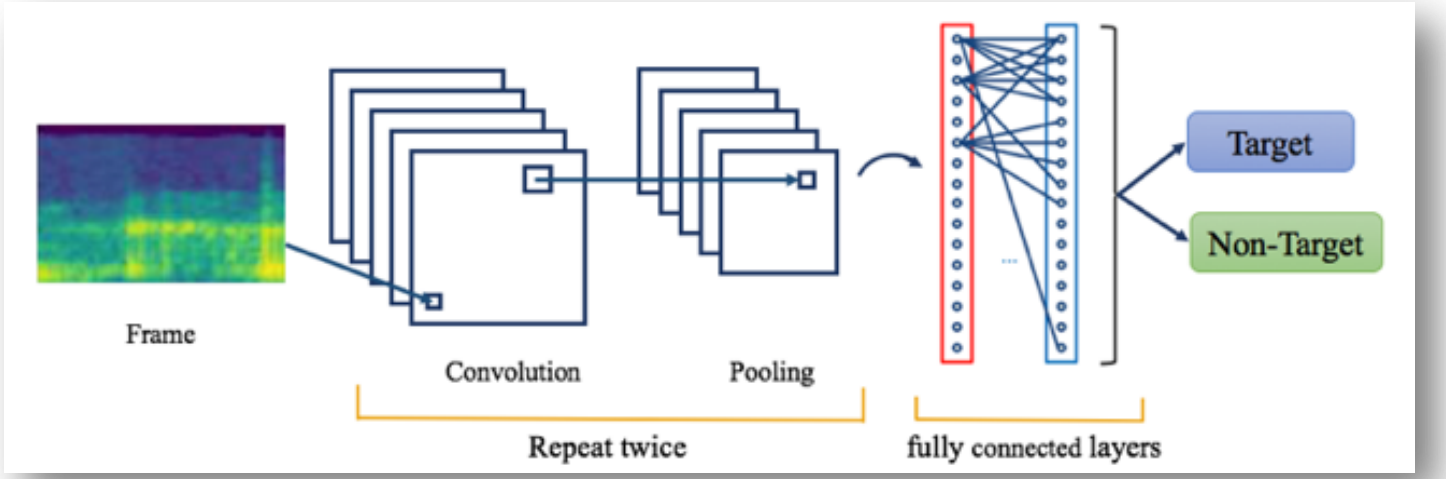
Corpus Design

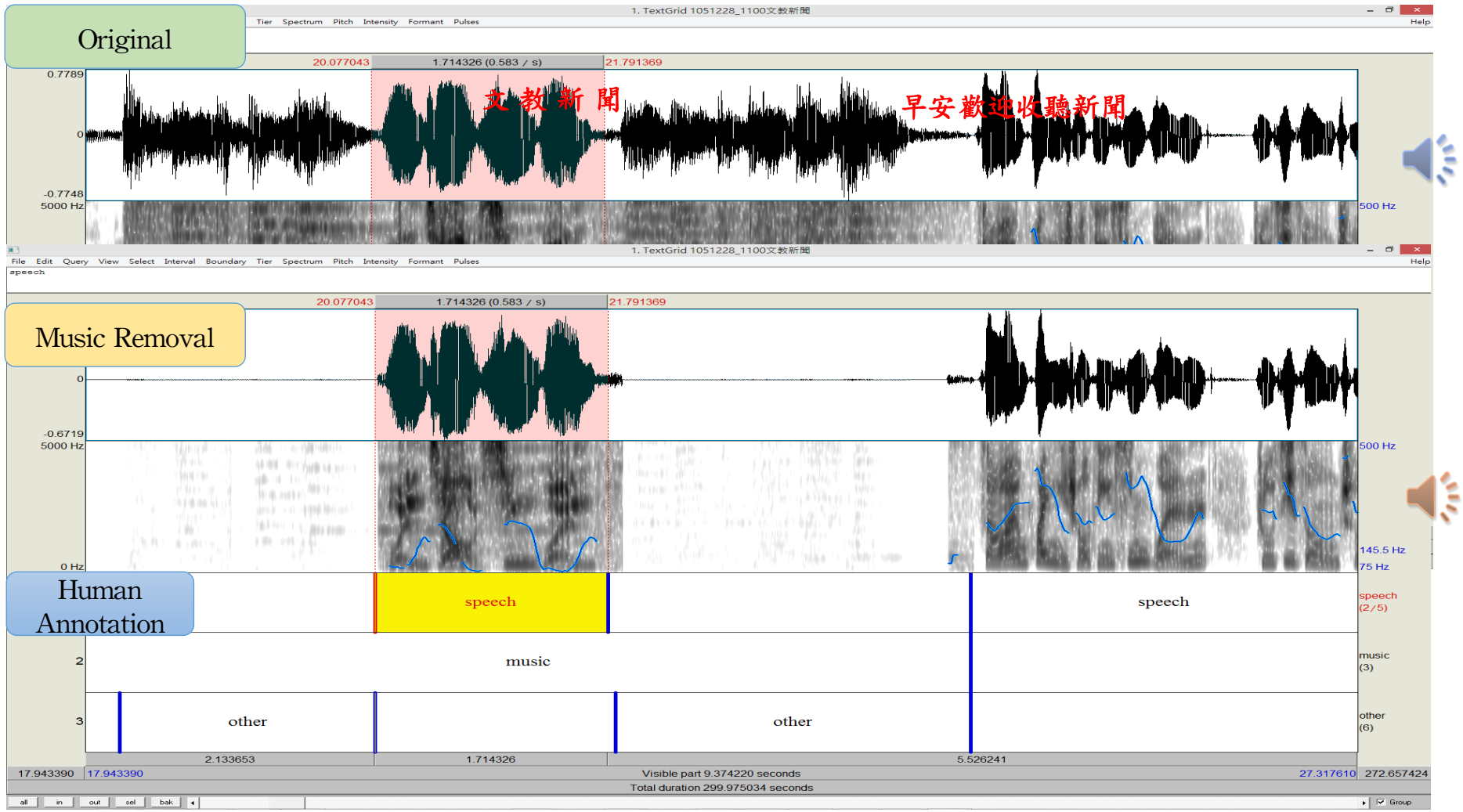
- Core (clean)
  → Human Correction and Annotation
- Non-Core (non-clean)
  → Semi-supervised training



Audio Event Detection

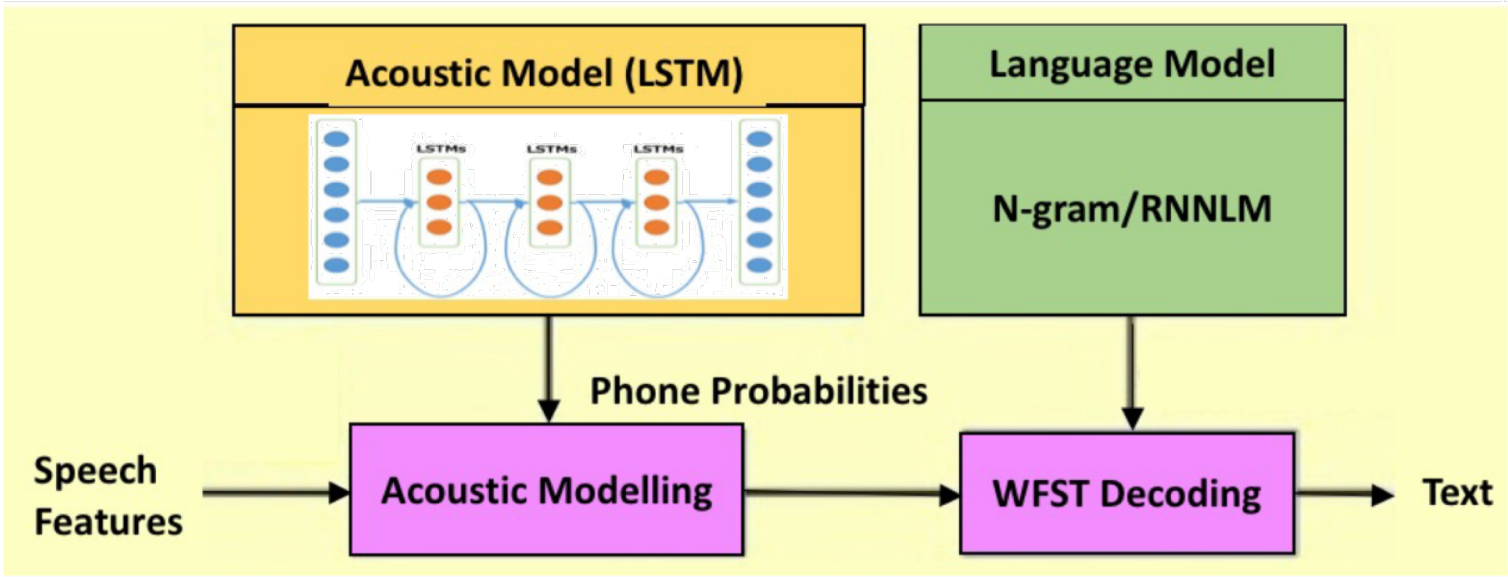
- Convolutional Neural Network (CNN) + raw spectral features
- Audio Event Database (manual annotation, ~60 hours)




Original
1. TextGrid 1051228_1100文教新聞
Tier Spectrum Pitch Intensity Formant Pulses
Help
20.077043
1.714326 (0.583 / s)
21.791369
0.7789
0
-0.7748
5000 Hz
文 教 新 聞
早安歡迎收聽新聞
500 Hz
1. TextGrid 1051228_1100文教新聞
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
speech
Music Removal
20.077043
1.714326 (0.583 / s)
21.791369
0
-0.6719
5000 Hz
0 Hz
500 Hz
145.5 Hz
75 Hz
Human Annotation
speech
speech
speech
(2/5)
2
music
music
(3)
3
other
other
other
(6)
2.133653
1.714326
5.526241
17.943390
17.943390
Visible part 9.374220 seconds
27.317610
272.657424
Total duration 299.975034 seconds
all
in
out
sel
bak
Group

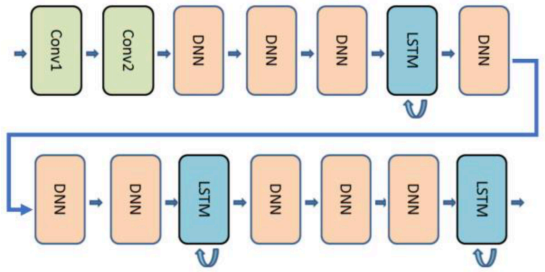
Automatic Radio Speech Transcription

- Mixed Chinese/English ASR
  - Trained using about 22 million words, 400 hours speech

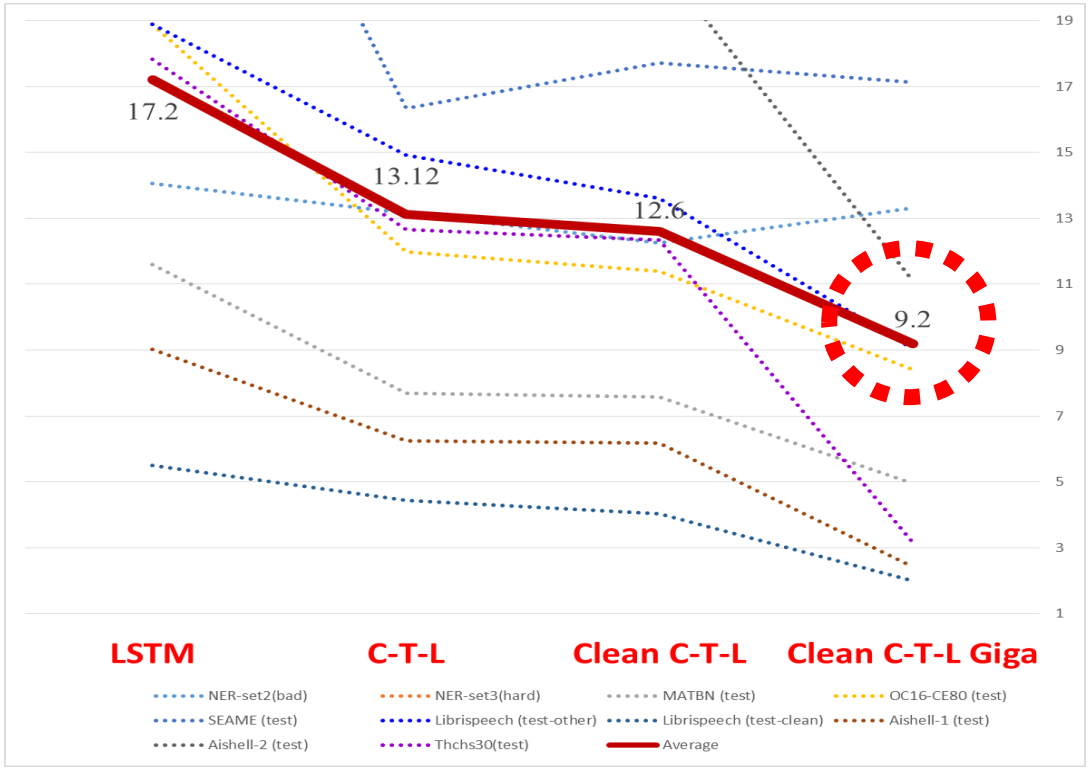


Performance of NTUT's Speech Recognizer

- CNN+(DNN+LSTM)*3



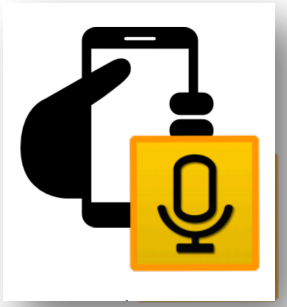
- 800 hours Training data
  - MATBN
  - NER-Trs-Vol1
  - SEAME
  - LibriSpeech
  - AiShell
  - OC16-CE80
  - Thchs30

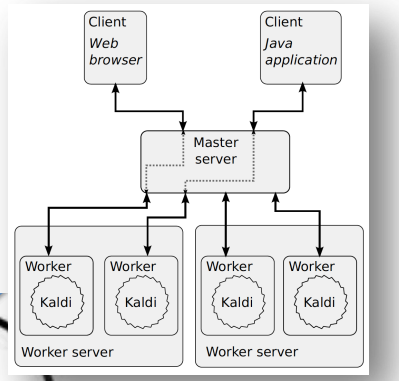


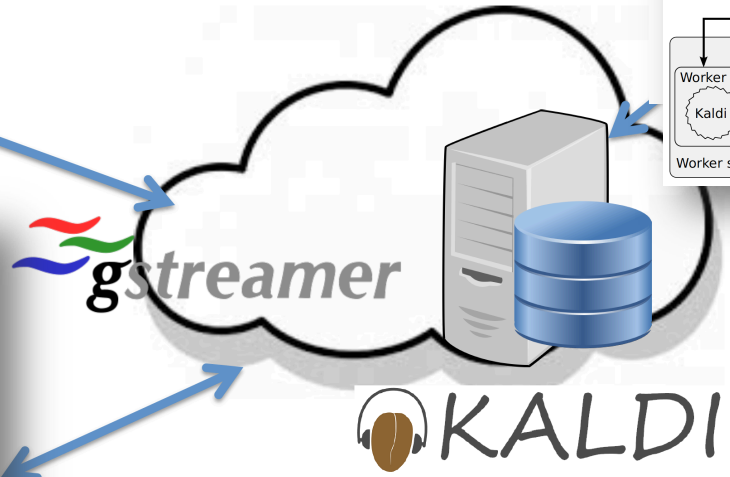
NTUT's Online Speech Recognition Server

Mobile APP Demo: https://youtu.be/NMYZ_UJDdzI

https://github.com/Kaljurand/K6nele





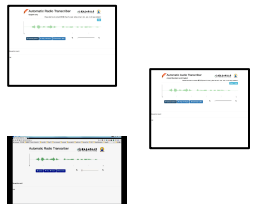


https://github.com/alumae/kaldi-gstreamer-server

https://youtu.be/m0boxASof0s

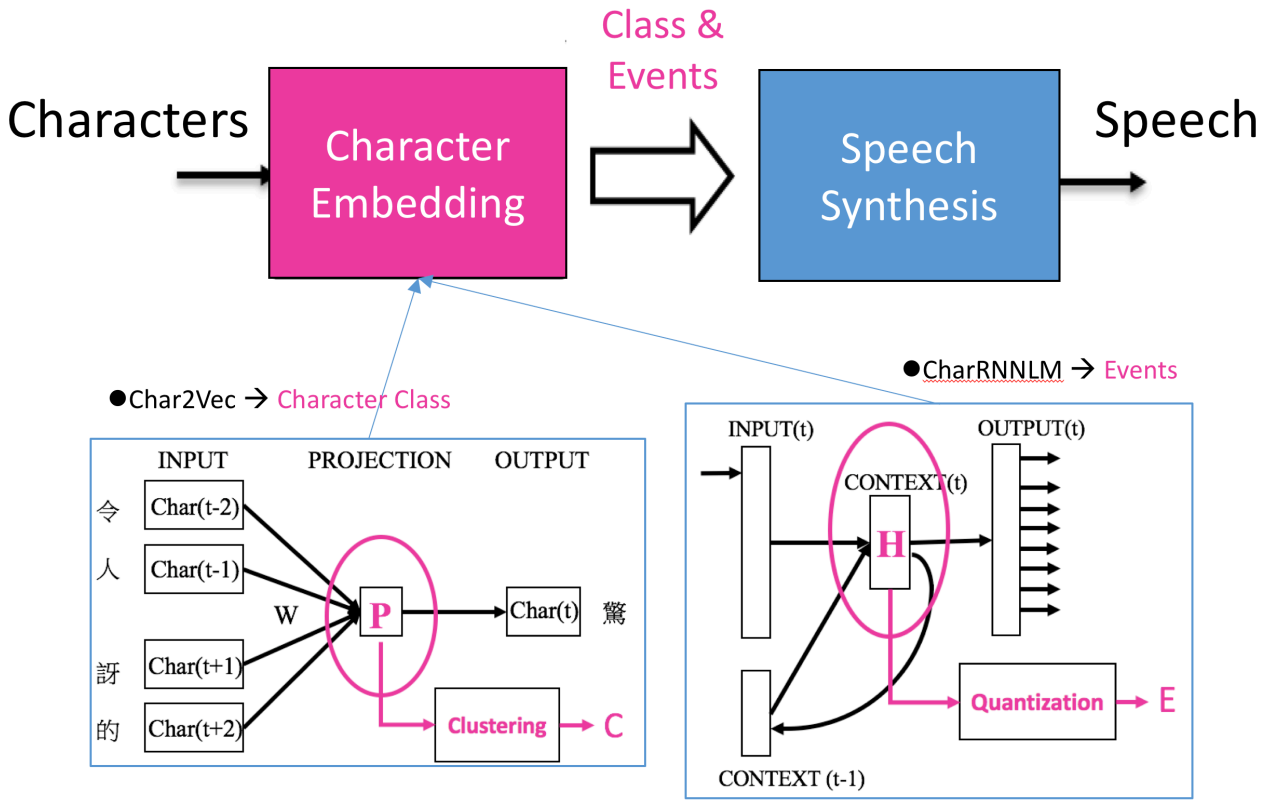
https://youtu.be/wUkfLn0tRbw

Online Service Demo: https://youtu.be/Mf4qvLRSa-Y



NTUT's Speech Synthesizer

- Mixed Chinese/English TTS



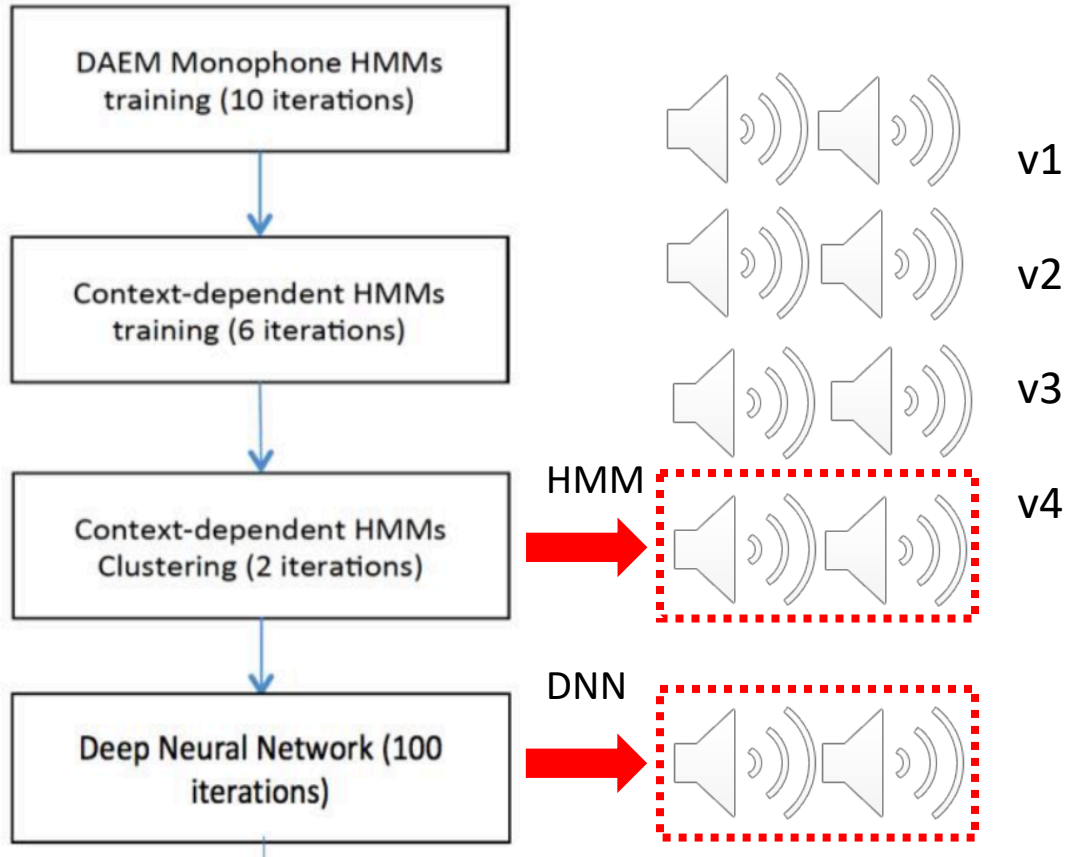
●Chinese

●Mixed Chinese-English

Speech Synthesis Backend (Improvement, 1/2)

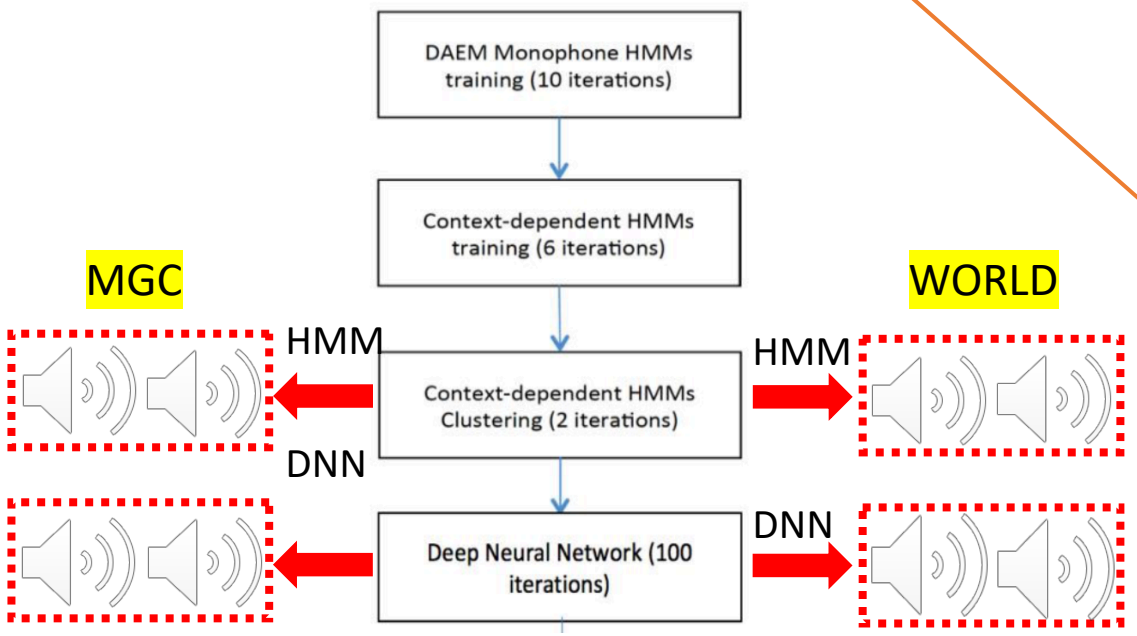
- Chinese TTS
  - HMM, 34-dim. MGC+F0, MLSA, Speaker-Dependent ASR
  - DNN, 60-dim. World Vocoder (on-going)

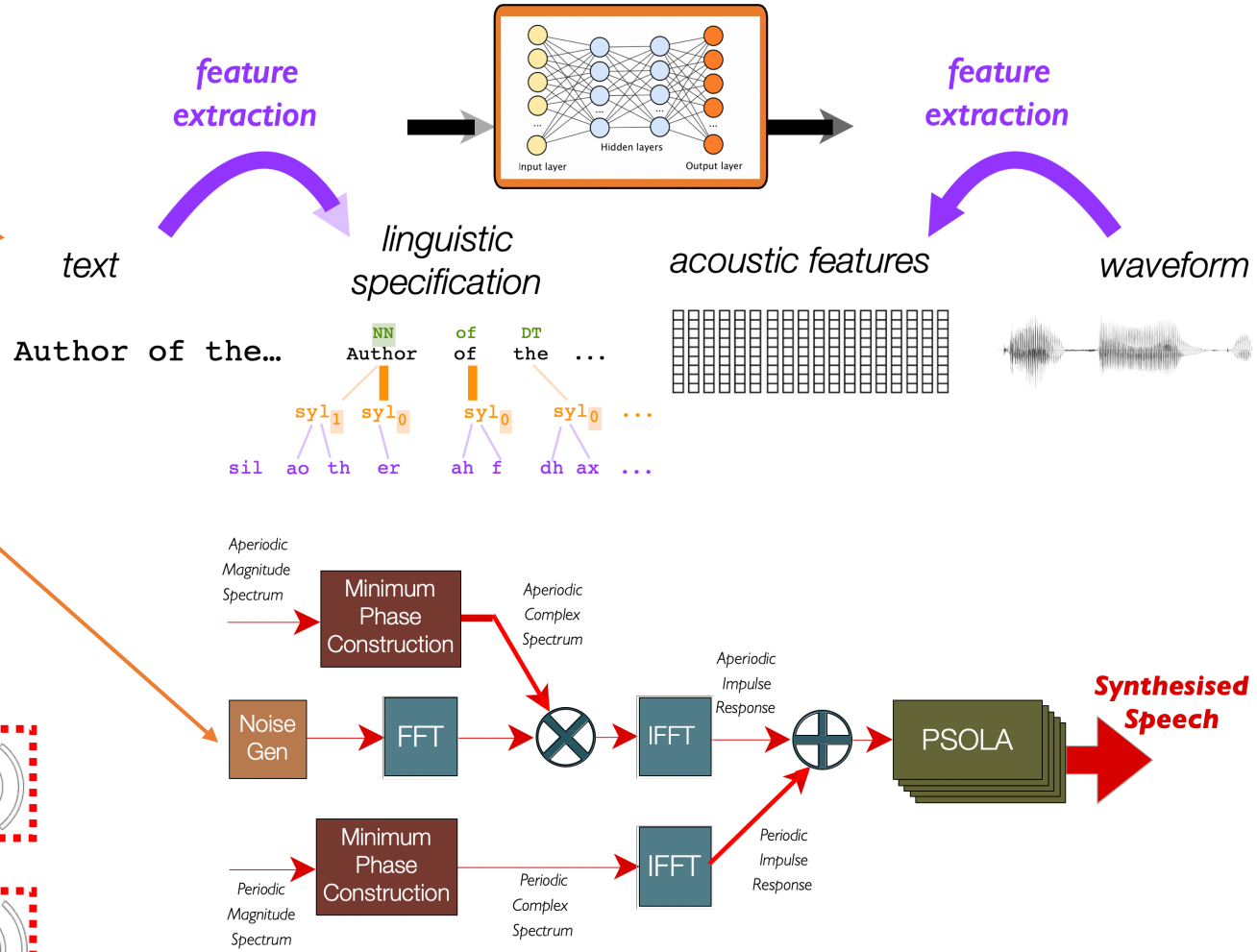
前一段时间我看到一篇文章，说上世纪六零年代的“大逃港”时，香港人真是好，全民动员救助逃过去的吃不饱饭的大陆人，可见香港人的精神境界有多高。有一次，我问一位老香港人，说这事是真的吗？他犹豫了一下说：真的，但是没有他们说的那么崇高。怎么回事呢？你想，香港原来就是个小渔



Speech Synthesis Backend (Improvement, 2/2)

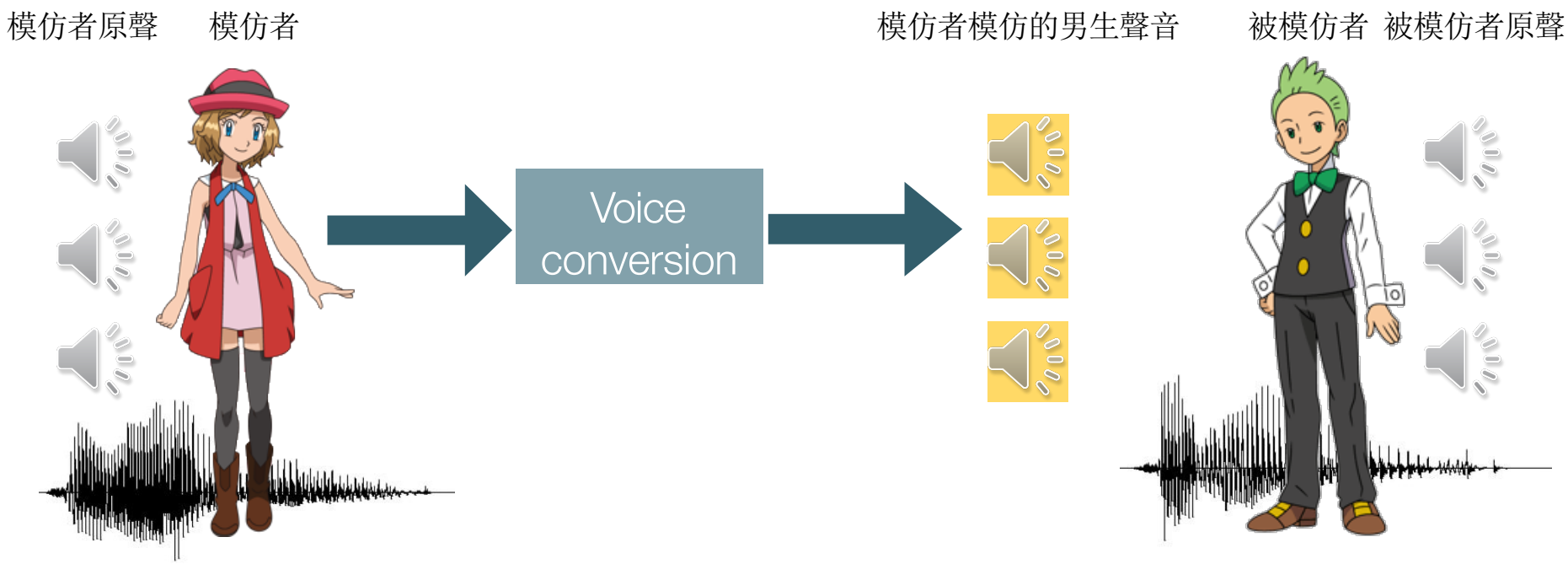
- CHT English TTS
  - DNN
  - WORLD vocoder





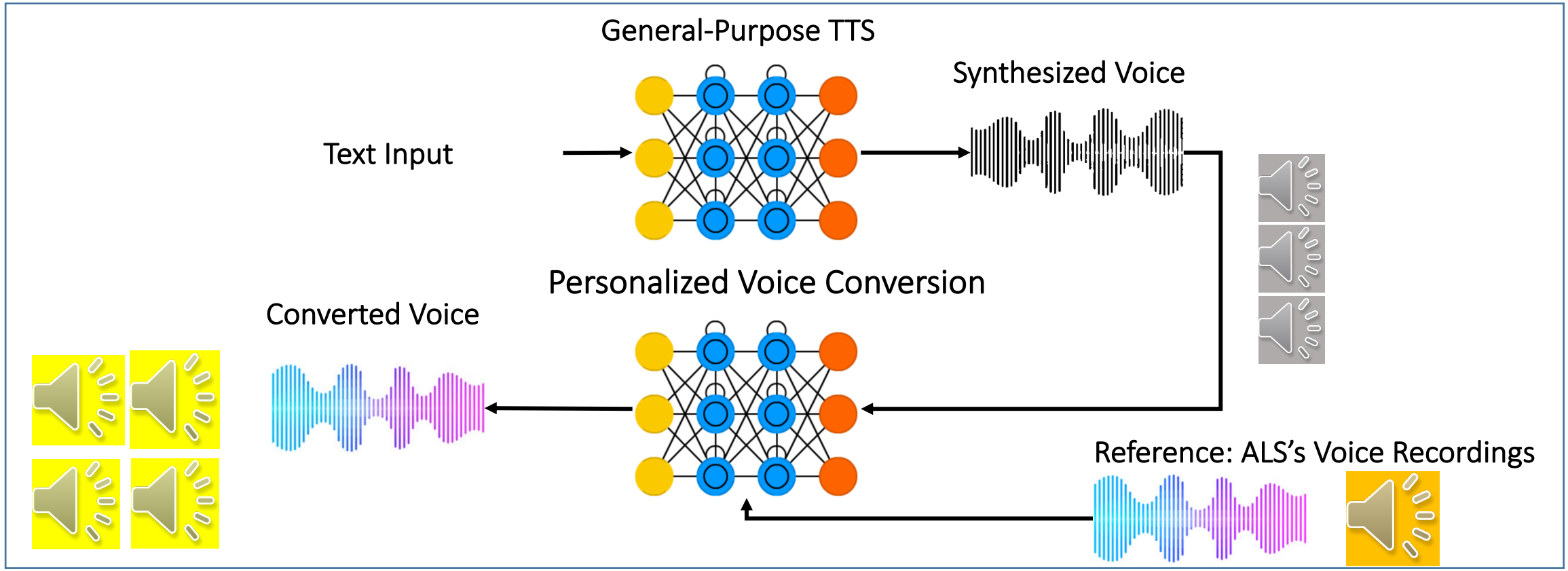
NTUT's Voice Conversion

- Manipulate source speaker's voice to sound like target without changing language content



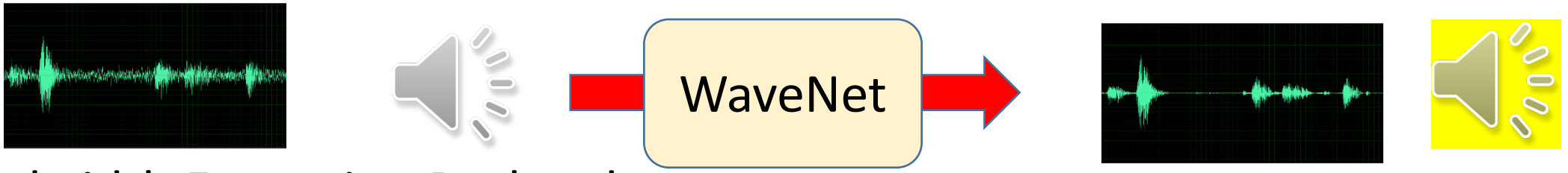
Personalized Voice Output Communication Aid (VOCA) for ALS (2/2)

- Real Case
  - Only Few Minutes Speech, Low-Quality Recordings (8 kHz, Noisy, MP3)

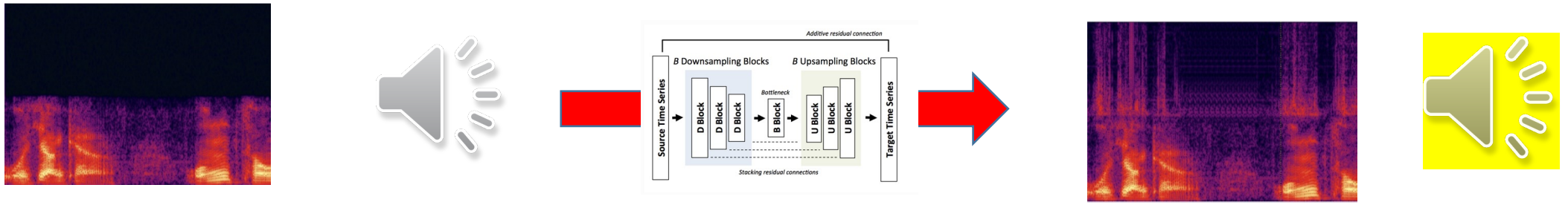


Approaches for Low-Quality, Noisy Samples

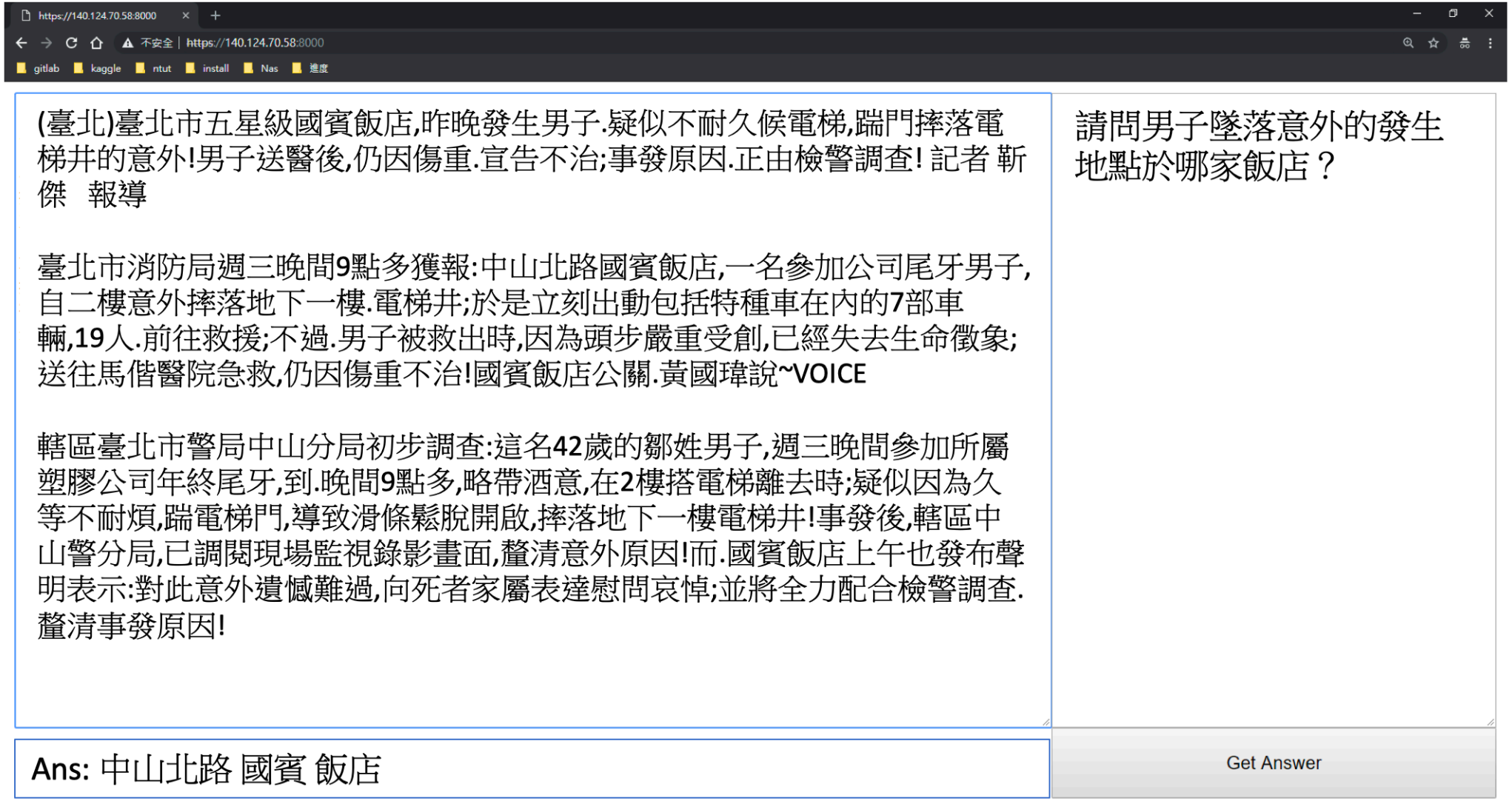
- Auxiliary Corpus
  - High-Quality Speech Corpus (20 hours, ~60 speakers) + Background Noises
- Denoising Frontend
  - WaveNet (noisy vs. clean speech pairs)



- Bandwidth Expansion Backend
  - U-Net (8 vs. 16 kHz speech pairs)



NTUT's Question/Answering System

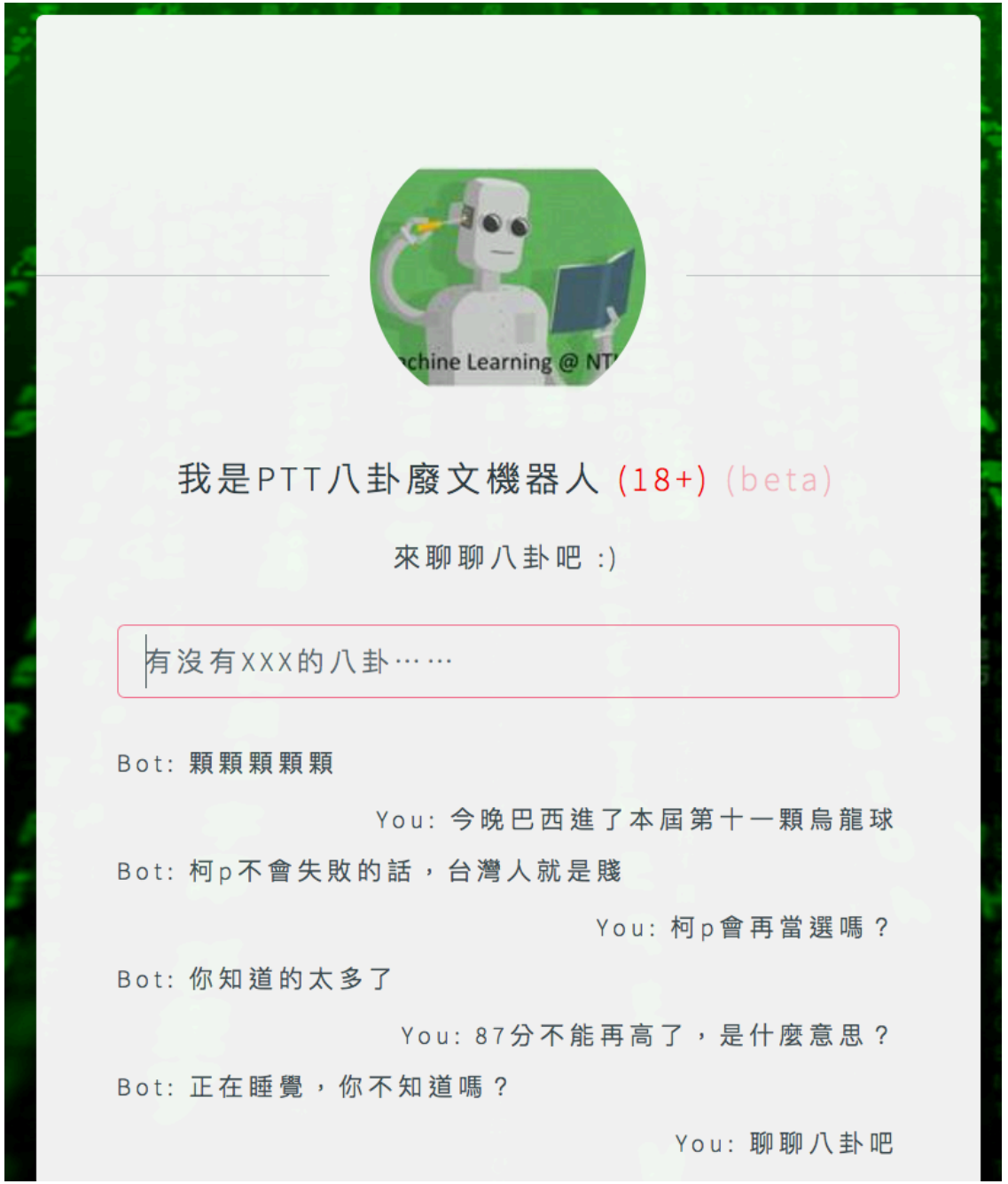


NTUT’s Toy #1

- English ChatBot
  - http://140.124.72.159:8888

NTUT's Toy #2

- Chinese ChatBot
  - http://140.124.72.159:8889



NTUT’s Toy #3

- Chinese Text to Taiwanese Romanization Translator
- https://140.124.72.159:8990