VCC 2020 reference design for Mandarin VC

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Time line

Before 11/30:

- 先前在更換 ASR 時，一次性的將 ASR model 與 corpus 更換成中文預訓練以及中文語料，因此預計將更換步驟拆解成兩步驟

The day:

- 完成 ASR model 更換、處理 TTS model 資料集前處理
  1. 根據 fig. 1，使用英文語料更換中文預訓練模型
  2. 使用中文預訓練模型更換輸入語料並輸出中文結果 (p. 7)

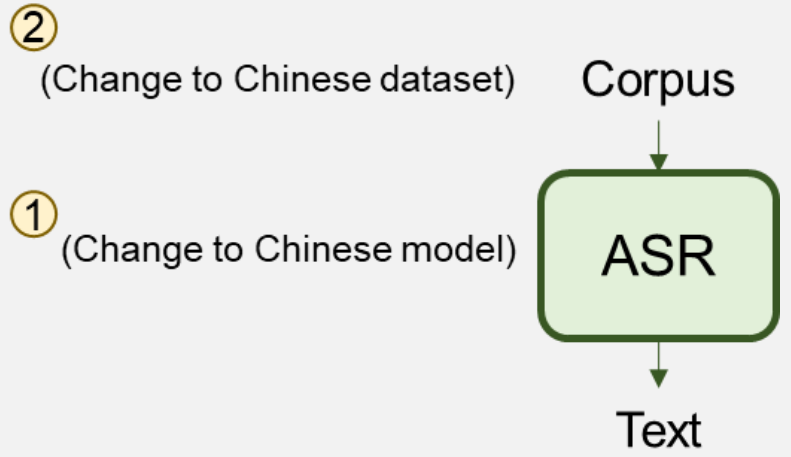


Fig. 1: Change ASR step

12/07:

- 將 TTS model 分成三部分 (table 1)，其中第二步驟又分為 6 小步
  1. Feature Generation
  2. JSON format data preparation
     - 在 TTS model 使用字典標記進行索引
  3. X-vector extraction based on the pre-trained, Kaldi
  4. Model training
  5. Decoding
  6. Synthesis speech
- 依執行腳本執行至第四步驟時遇到模型參數大小不匹配問題，因此檢查各步驟執行正確性時，認為自己在步驟二任意更換字典是錯的 (p. 5)

	Change step	date
1	Data set pre-processing	11/30
2	Fine-tuning TTS model	12/07
3	Cascade ASR + TTS	12/14

Table 1: Change TTS step & date

12/14

- 嘗試解決在 fine-tuning 中遇到的不匹配問題，最後將預訓練模型更換成 VCC2020 task2 中所使用的 TTS pre-trained model

Outline

Action item

- 將 VCC2020 baseline 從 English VC 更換成 Mandarin VC

Status report

- Before
  - 上週在更換 TTS model 中的微調部分發現錯誤，發現其中一個原因是發生在使用字典對語料進行標記與索引

- Now
  - 完成事項：
    1. 微調時，在 baseline task2 中也有將中文轉換成拼音的步驟，將其所需應用到 task1
    2. 使用 baseline task2 預訓練模型取代 ESPnet 提供的預訓練模型，並完成 TTS 微調
       (ESPnet 提供的預訓練模型會有模型參數大小不匹配問題)
    3. Cascade ASR + TTS
       (合成出奇怪的音檔，原因為 ASR 輸出為簡體中文，需轉換為拼音，目前紅色框為為手動更改成拼音，尚未完成腳本自動執行)

- After
  1. 完成腳本自執行
  2. 尋找問題是否為預訓練模型導致使用女聲語料效果較男聲語料好

	Change step
1	Data set pre-processing
2	Fine-tuning TTS model
3	Cascade ASR + TTS

Table 1: Change TTS step

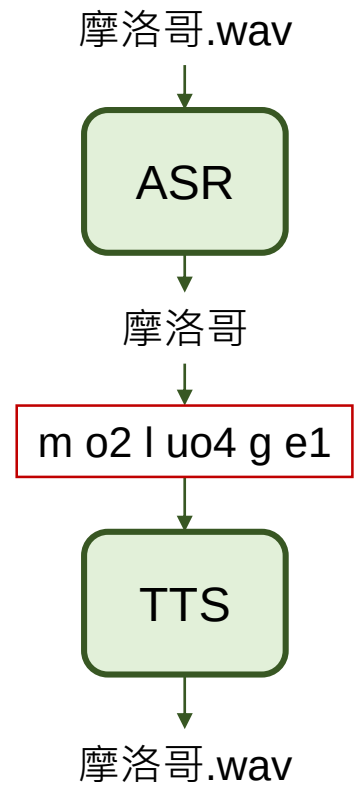
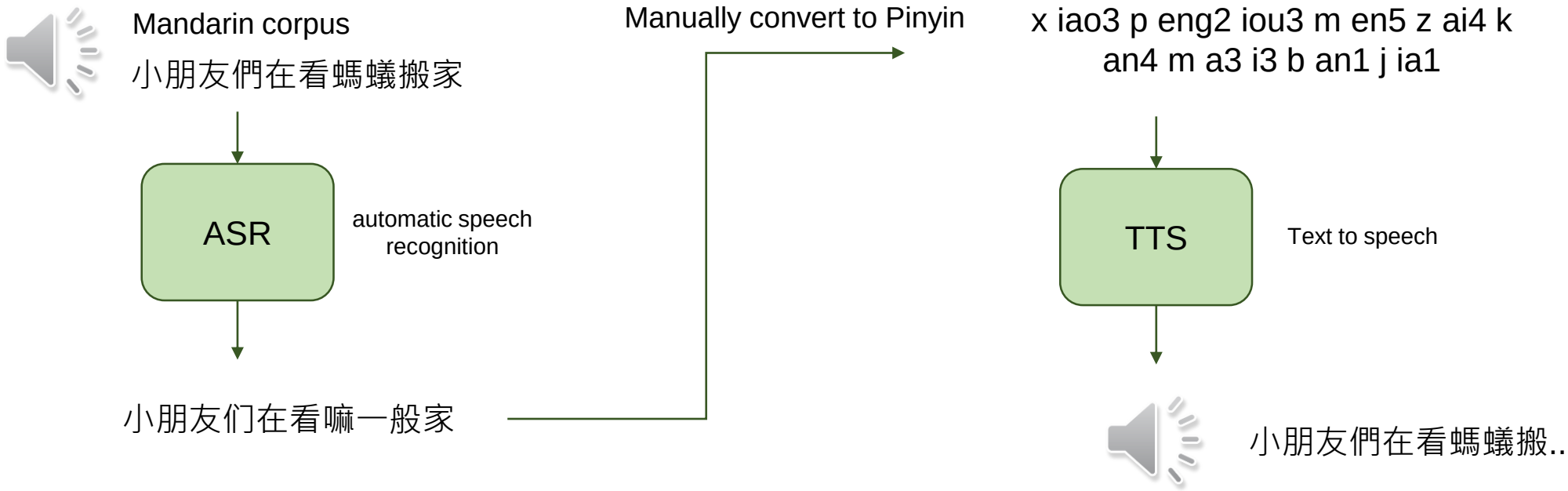


Fig. 2 Conversion example

Chinese Mandarin VC example



12/07 outline

Action item

- 繼上週更換完成 ASR model 繼續將 TTS model 更換成中文轉換模型

Status report

- Before：
  - 更換 ASR 發生錯誤，因此拆解成小步驟，完成更換後可以辨識中文語料輸出中文。
- Now：
  - 更換 TTS model，我分成三大部分
    1. Data pre-processing (**已完成**)
    2. **Fine-tuning**
    3. Cascade ASR + TTS
  - 其中第二部分又細分成 6 個子步驟
    1. **Feature Generation**：指定所需語句、語句編號，並對語料做 normalized
    2. **JSON format data preparation**：在 TTS model 使用字典標記進行索引
    3. **X-vector extraction**：基於預訓練與 Kaldi 進行 x-vector 提取
    4. **Model training** (**進行中**)
    5. **Decoding**
    6. **Synthesis speech**
  - 原先在第四子步驟錯誤，因此回頭檢查覺得在第二步驟我任意為了繁體中文更換了字典內容是錯誤的，目前做法為將中文轉換成拼音再利用字典標記成數字，細節於次頁做說明。
  - 確保先前步驟的正確性後，**第四步驟訓練時仍遇到問題**，問題與第二步驟更換字典有關，仍在找解決方法。

Problems encountered in the second substep

Baseline JSON (for English model)
- 字典形式為將所有**字母**編號
- **for example**：Sentence: “There are”
  - 34 50 47 60 47 13 43 60 47

T h e r e | a r e

<space>

Mandarin pretrain JSON (**for task1**)
- 字典形式為將所有**拼音**編號
- 但發現此字典有錯誤的地方，所以我打算不使用這張表
- 像是在字典中出現 air 或是 anr 這樣沒辦法唸的拼音

Mandarin pretrain JSON (**for task2**)
- 在任務二中同樣也有將中文轉換成拼音的步驟，因此我將任務二轉換的步驟搬移到任務一完成
- 目前第四步驟執行錯誤其中一個原因就是此字典大小不相同，導致不匹配

```
1 <unk> 1
2 ! 2
3 " 3
4 ' 4
5 ( 5
6 ) 6
7 , 7
8 - 8
9 . 9
10 / 10
11 : 11
12 ; 12
13 <space> 13
14 ? 14
15 A 15
16 B 16
.
.
.
```

總共為 76 個

Baseline JSON (for English model)

```
1 <unk> 1
2 a1 2
3 a2 3
4 a3 4
5 a4 5
6 a5 6
7 ai1 7
8 ai2 8
9 ai3 9
10 ai4 10
11 ai5 11
12 air2 12
13 air4 13
14 an1 14
15 an2 15
16 an3 16
.
.
.
```

總共為 259 個

Mandarin pretrain JSON (**for task1**)

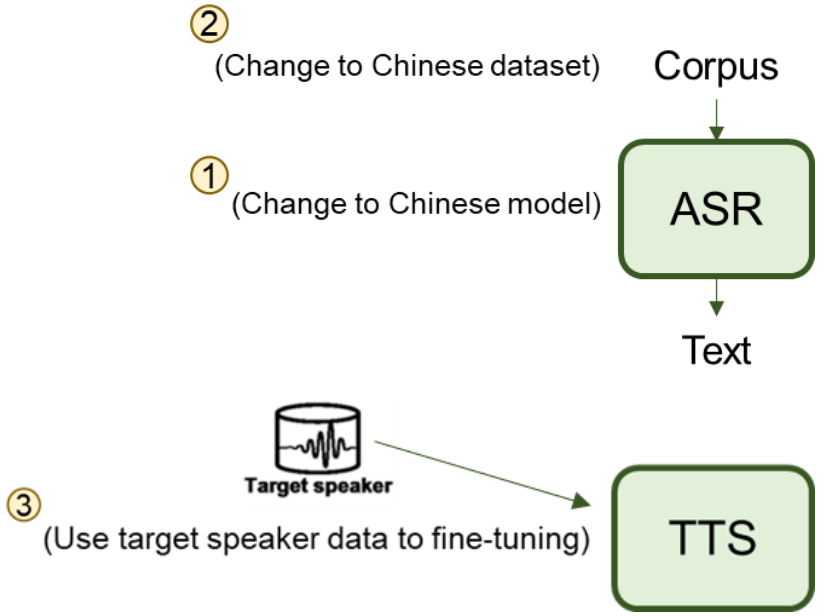
```
1 <unk> 1
2 ! 2
3 ' 3
4 , 4
5 . 5
6 <en_US> 6
7 <zh_ZH> 7
8 ? 8
9 AA0 9
10 AA1 10
11 AA2 11
12 AE0 12
13 AE1 13
14 AE2 14
15 AH0 15
16 AH1 16
.
.
.
```

總共為 335 個

Mandarin pretrain JSON (**for task2**)

最後認為字典中的錯誤並不會影響到拼音索引的部分，因為字典中仍然包含所需的對應值，因此使用原先提供的檔案確保標記結果大小與預訓練相同

11/30 outline



Action item

- 將 ASR 與 TTS 英文轉換模型更換成中文轉換模型

Status report

- Before：
  - 在更換 ASR model 時出錯，因此將更換步驟拆解成兩步驟
- Now：
  - 更換 ASR 步驟：(**完成**)
    1. 更換 ASR model，其餘不變 (使用英文語料進行中文辨識)，確保預訓練模型更換成功
    2. 更換成中文語料，輸出中文

    目前**已經成功輸入中文語料，辨識輸出中文 (結果於次頁)**，原先猜測上週錯誤點為是語料取樣率導致錯誤，最後驗證原因是檔名帶有特殊符號而抓取不到，導致輸出的關聯檔錯誤
  - 更換 TTS 步驟 (較細節步驟於 p.4)：
    1. 語料集處理
    2. 特徵提取 (**進行中**)
    3. TTS 微調
    4. Cascade ASR + TTS model 合成語音

    此步驟在 baseline 中使用 shell 腳本實現，細分為 10 個 stages，進行到第 3 個 stage，目前預到問題較多都是檔案中取值問題或是從 windows 上傳文件到 Linux 導致的格式錯誤，逐一更改關聯檔時都順利解決
- After：
  - 預期下週將 TTS 更換完成

ASR result

Try & error step 1：

English corpus

In reality, the European Parliament is practising dialectics.

→ Chinese Mandarin VC →

Data text

你不要了KTVPN爬了
们就是反正是你的
LISA的词

Try & error step 2：

Mandarin corpus

小朋友們在看螞蟻搬家

→ Chinese Mandarin VC →

Data text

小朋友们在看嘛一般家

- 改善想法：目前 ASR 預訓練模型是 Chinese Mandarin VC，但我們的語料是 Taiwanese Mandarin dataset，猜想這樣的差別(用詞習慣不同 e.g., **S0307** 警察 vs. 公安) 會導致辨識率下降
  1. 將 input Mandarin corpus 直接做繁簡轉換
  2. 使用 Mandarin corpus dataset 對 ASR model 做 fine-tuning

編號	Input	Output
S0301	小朋友們在看螞蟻搬家	小朋友们在看嘛一般家
S0302	這個池塘裡養了很多魚	这个词汤里养了很多余
S0303	這裡將要建一座發電廠	这里将要建议做发电厂
S0304	他下山時被蛇咬了一口	她下算时被子咬了一口
S0305	他計畫今年買一台電腦	他计划今年买一台电脑
S0306	這間公司要招聘工程師	这间公司要招聘工程师
S0307	便衣警察抓到一名小偷	便宜检查抓到一名小偷
S0308	他是那座大樓的設計師	__是那做大楼的设计时
S0309	他們擺好姿勢準備拍照	他们白好只是准备拍照
S0310	今天老師宣布提前放學	今天老师宣布提前放学
S0311	桌子上擺了一大盤瓜子	说__上百了一大盘瓜子
S0312	中國萬里長城中外名聞	中国万里长城中外名文
S0313	那片原始森林發生大火	那片延死森林发生大火
S0314	棒球飛過來打破了窗戶	榜车飞过来打破的仓库
S0315	他的哥哥養了一群白鴿	他的哥哥养了一群百哥
S0316	他做完功課才上床睡覺	他做晚公课才上床睡觉
S0317	一群孩子在那裏玩跳繩	一群孩子在那里玩跳省
S0318	這裏的風俗習慣很特別	这里的风俗习惯很特别
S0319	他腰痛的老毛病又犯了	他腰痛的老毛病又犯了
S0320	他不小心把茶杯碰翻了	它不小心把茶被碰烦的

■ 錯字 or 缺字　　使用中文語料並使用 Chinese Mandarin VC 轉換

TTS change stages

串接 ASR 前，要先對 TTS 微調，微調前須處理 dataset (Taiwanese Mandarin dataset)

Data pre-processing

stage 1: Data and Pretrained model download
- 在腳本中更改選擇的預訓練模型

stage 2: Data preparation
- 指定語料源、標記語料所對應的語句、降頻到 16 kHz，並產生一些關聯檔

Fine-tuning

stage 3: Feature Generation：指定所需語句、語句編號，並對語料做 normalized
stage 4: Dictionary and Json Data Preparation：在 TTS model 使用字典標記進行索引
stage 5: x-vector extraction：基於預訓練與 Kaldi 進行 x-vector 提取
stage 6: Text-to-speech model fine-tuning：訓練神經網路
stage 7: Decoding
stage 8: Synthesis speech

Cascade ASR + TTS

stage 9: Import ASR result and decoding
stage 10: Synthesis speech