

VCC 2020 reference design for Mandarin VC

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

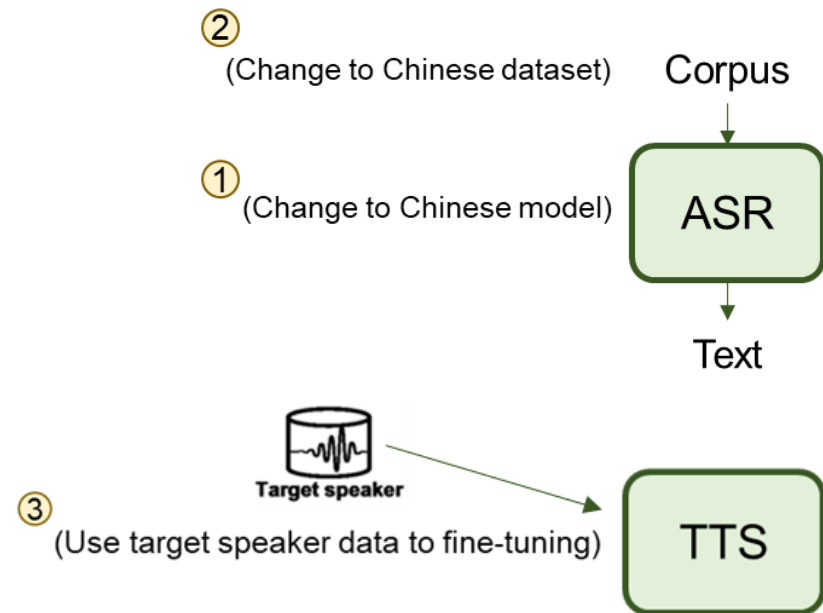
Outline

Action item

- 將 ASR 與 TTS 英文轉換模型更換成中文轉換模型

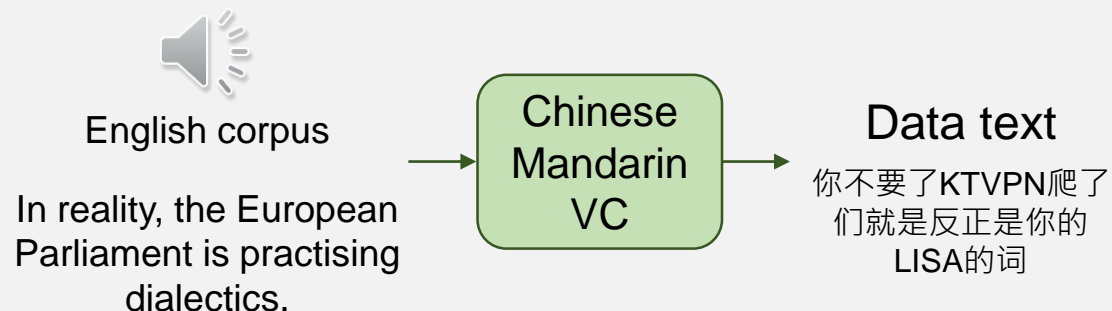
Status report

- 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

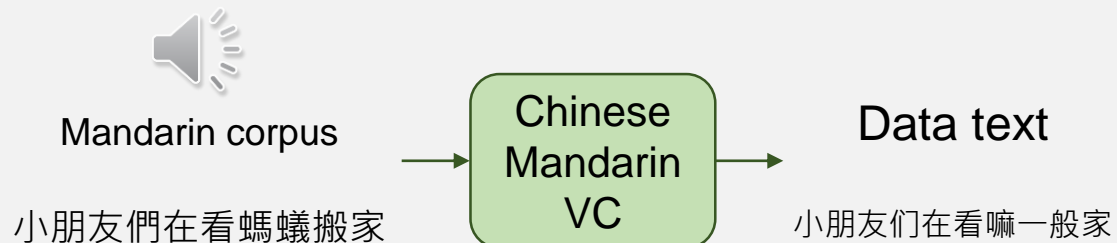


ASR result

Try & error step 1 :



Try & error step 2 :



- 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

- 將 input Mandarin corpus 直接做繁簡轉換
- 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 (完成)

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

stage 4: Dictionary and Json Data Preparation

stage 5: x-vector extraction

stage 6: Text-to-speech model fine-tuning

stage 7: Decoding

stage 8: Synthesis speech

Cascade ASR + TTS

stage 9: Import ASR result and decoding

stage 10: Synthesis speech

■ 附錄

1. The VCC 2020 Rules

- Dataset relationship
- Evaluation methodology

2. How Baseline (ASR+TTS) Work

- Transfer learning and fine-tuning
- Training and conversion process

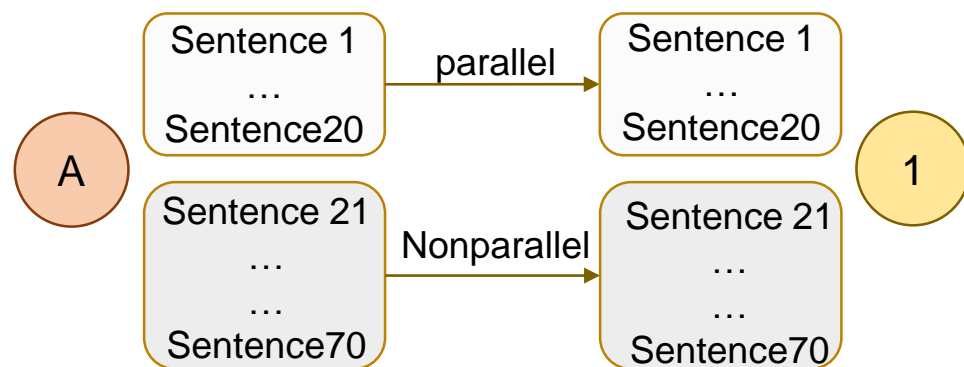
3. Strategy to win

- Why change vocoder from non-autoregressive to autoregressive

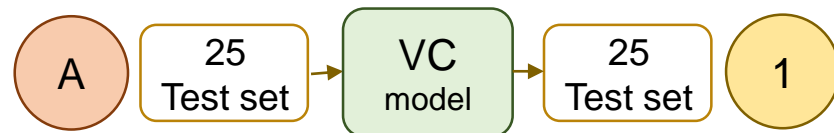
4. Its application in Mandarin VC

- ASR and TTS Mandarin pretraining model detail
- Replacement method

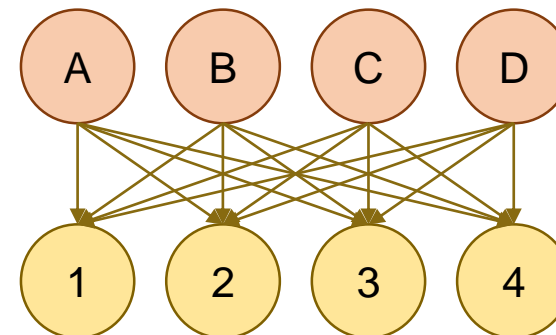
The VCC 2020 Rules



Same language conversion in task1

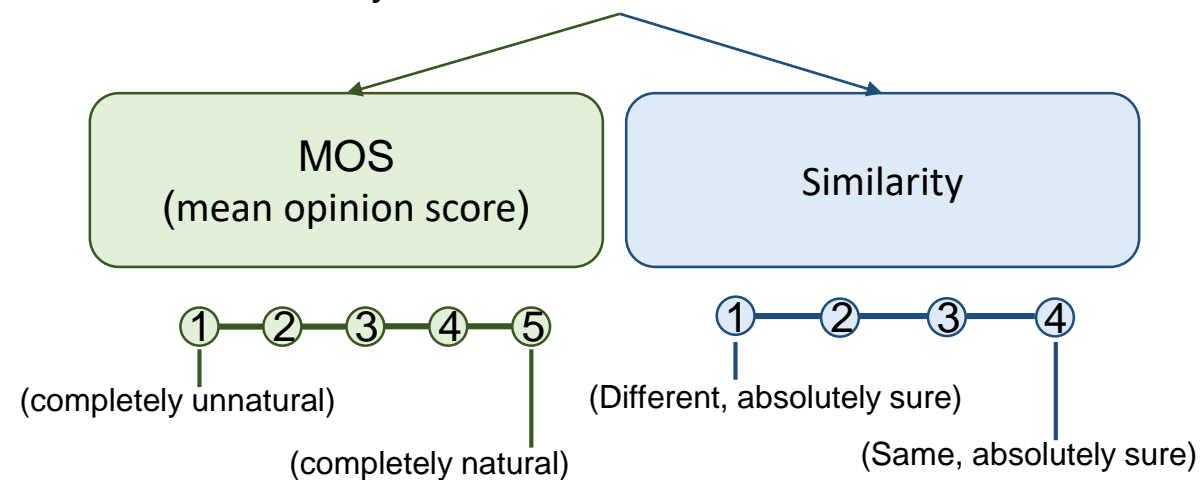


Conversion system



Total: 16 systems

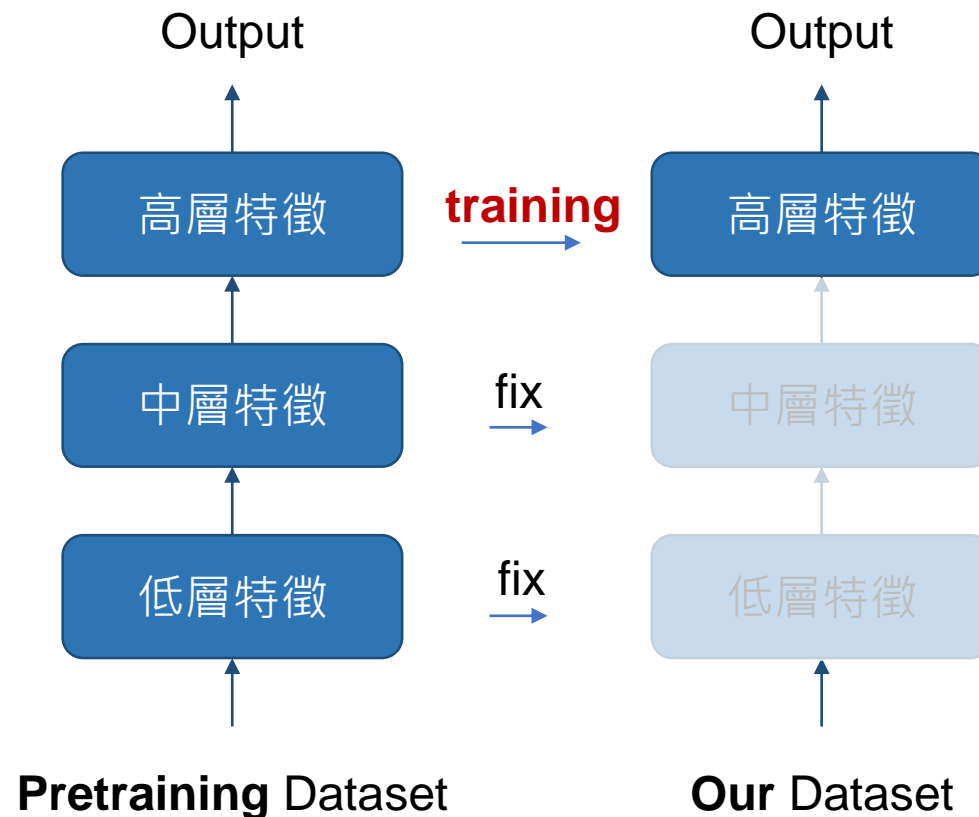
16 systems * 25 conversion results



Evaluation methodology

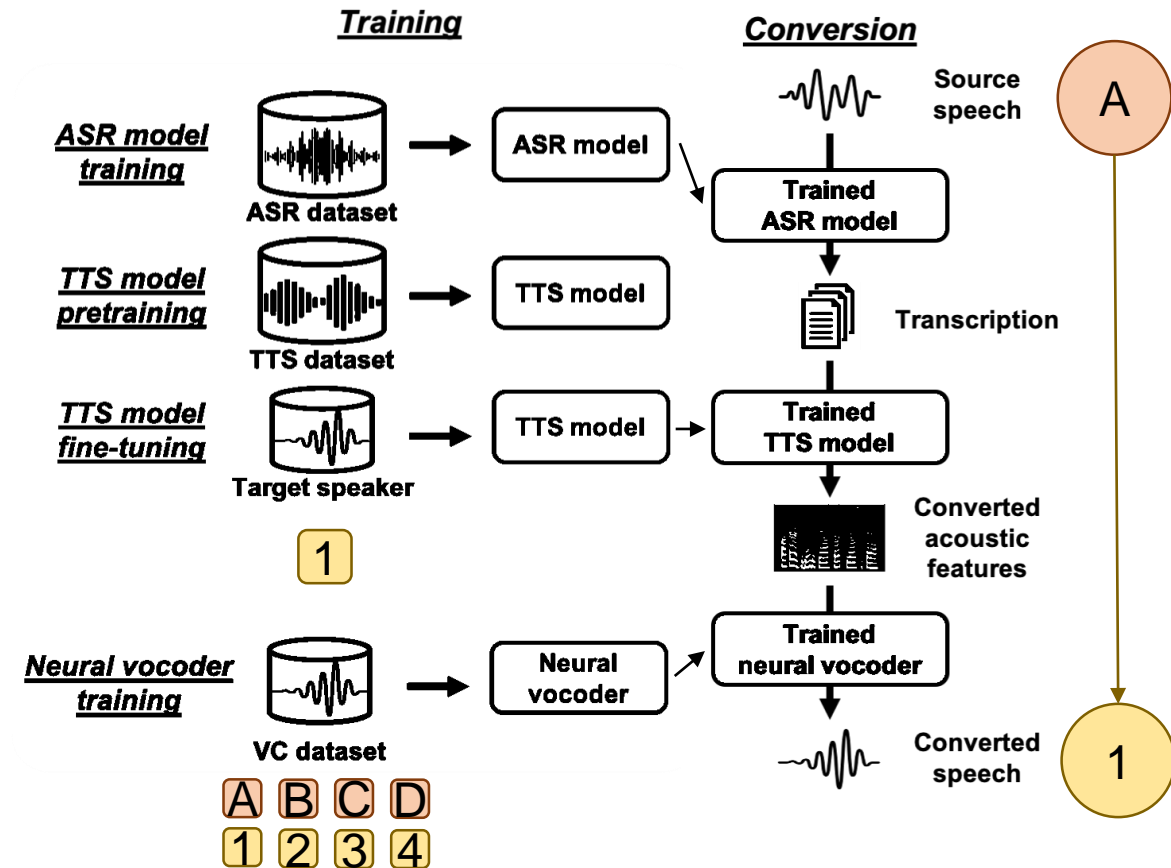
Transfer learning and fine-tuning

- The more dataset, the higher accuracy, but the longer training time.
- 70 sentences corpus approximately 5 minutes is not enough to train a good model.
- A pre-trained model is a model created by some one else to solve a similar problem.
- Advantages of the pre-training model can save a lot of computing time and resources, and can avoid some training risks.



How Baseline Works

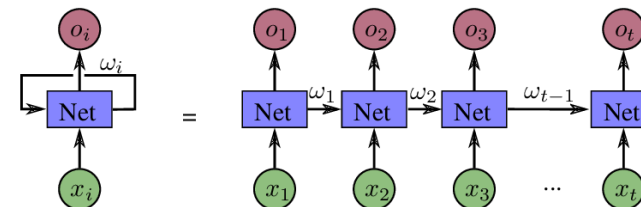
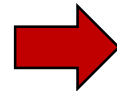
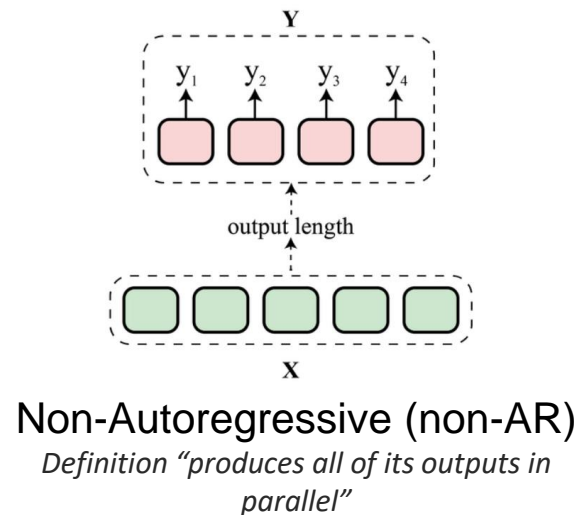
- Task1: voice conversion within the same language.
- A naive approach for VC is a cascade of an automatic speech recognition (ASR) model and a text-to-speech (TTS) model, and both model are using pre-training model.
- ASR models are usually trained with a multi-speaker dataset, thus speaker-independent in nature.
- TTS model output was the mel filterbank sequence extracted from the waveform and was also the input of the vocoder.



Pre-training model	Language	Dataset	speakers	Hours
ASR	English	LibriSpeech	Over 2000	960
TTS	English	LibriTTS	Over 2000	250

Strategy to win

- Baseline adopted a non-AR neural vocoder (Parallel WaveGAN (PWG)) for fast generation.
- Change vocoder from non-autoregressive to autoregressive, e.g., change from Parallel WaveGAN (PWG) to WaveNet.
- Voice conversion = VC model + Vocoder
- The champion in the VCC2020 use a AR neural vocoder (WaveNet).





Autoregressive (AR)
Definition "Each output of the network is generated based on previously generated output"

Its application in Mandarin VC










- The AST and TTS model used in the Baseline are provided by ESPnet.

Lang.	ASR pre-training model	Dataset	Dataset summary	Speaker	Hours	Test set corr.
Mandarin	Conformer + SpecAugment	aidatatang_200zh	speech data	600	200	95.2
		AISHELL	recording	400	178	95.0
	Conformer	aidatatang_200zh	speech data	600	200	94.1
	Transformer	aidatatang_200zh	speech data	600	200	93.6
		AISHELL	recording	400	178	93.4
		AISHELL2	recording	1991	1000	91.8
		HKUST	telephone speech	無標示	約200	79.1

similar

Lang.	TTS pre-training model	Dataset	Speaker	Hours	Sample sentence	Audio
Mandarin	FastSpeech	CSMSC 女聲	1	12	這場抗議活動究竟是如何發展演變的，又究竟是誰傷害了誰	
	Transformer					

TTS samples

TTS pre-training model	Sample sentence	Audio
FastSpeech	在雨中，張明寶悔恨交加寫了一份懺悔書	
	今天快遞員拿著一個快遞在辦公室喊，秦王是哪個有他快遞	
	李東王表示自己當時在法庭上發表了一次獨特的供訴意見	
	接下來，紅娘要求記者交費，記者表示不知表姊身分證號碼	
	小明搖搖頭說，不是，我只是美女看多了，想換個口味而已	
Transformer	在雨中，張明寶悔恨交加寫了一份懺悔書	
	今天快遞員拿著一個快遞在辦公室喊，秦王是哪個有他快遞	
	李東王表示自己當時在法庭上發表了一次獨特的供訴意見	
	接下來，紅娘要求記者交費，記者表示不知表姊身分證號碼	
	小明搖搖頭說，不是，我只是美女看多了，想換個口味而已	