

# VCC 2020 reference design & its application in Mandarin VC

---

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

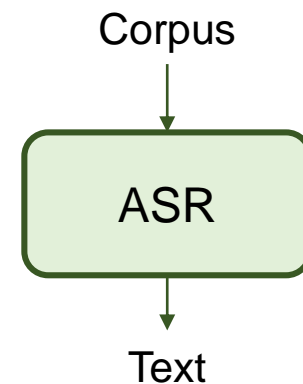
# Outline

## Action item

- 將 ASR 與 TTS 英文轉換模型更換成中文轉換模型

## Status report

- Before :
  - VCC2020 baseline 為英文轉換模型，使用的皆為 ESPnet (End-to-end speech processing toolkit) 所提供的預訓練模型，如今目標為更換為中文轉換模型，ESPnet 提供各種語言所訓練的預訓練模型，因此 ASR、TTS model 我打算使用與 baseline 相同架構 (Transformer) 不同語料的預訓練模型做替換
- Now :
  - 語料：使用碩一 IASoC 錄製的語料，明祥的語料作為 source、竣尹的語料作為 target
  - 替換步驟：
    1. 使用 ASR model 辨識，輸出文字 (進行中)
    2. 使用 target 語料微調 TTS 模型，並接續 ASR model 辨識結果完成轉換，輸出 mel-filterbank sequence
    3. 訓練 vocoder 並合成語音
- After :
  - 在使用 ASR 中文模型辨識時，在特徵提取的部分出錯
  - 解決方案：原先急著要輸出成果，因此除了將 ASR model 換成中文模型之外，還將語料換成中文
    - 使用原語料 (英文)，驗證成功將英文模型替換成中文模型再做更換 input 的動作 (可能輸入格式不正確、也可能替換過程有部分檔案缺失)



# ■ 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

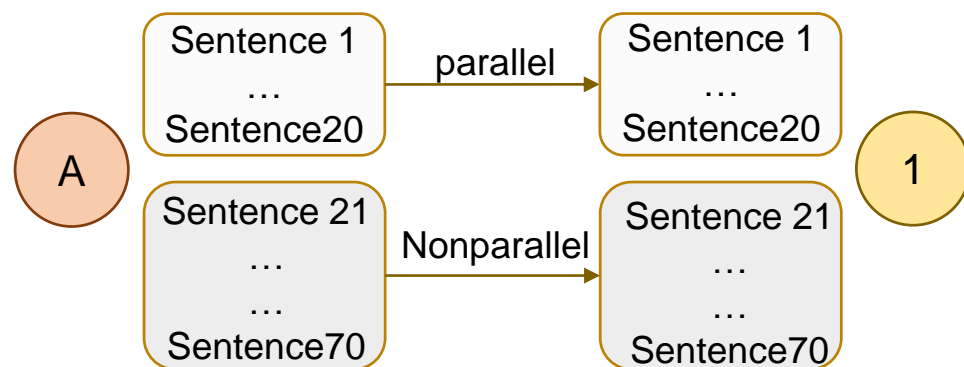
## 3. Strategy to win

- Why change vocoder from non-autoregressive to autoregressive

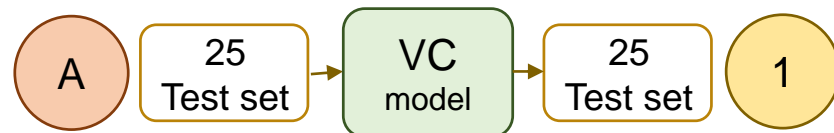
## 4. Its application in Mandarin VC

- ASR and TTS Mandarin pretraining model detail
- Replacement method

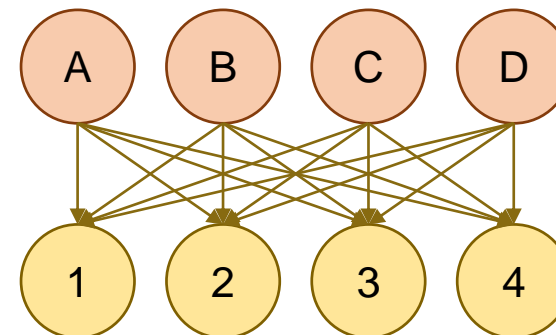
# The VCC 2020 Rules



**Same language conversion in task1**

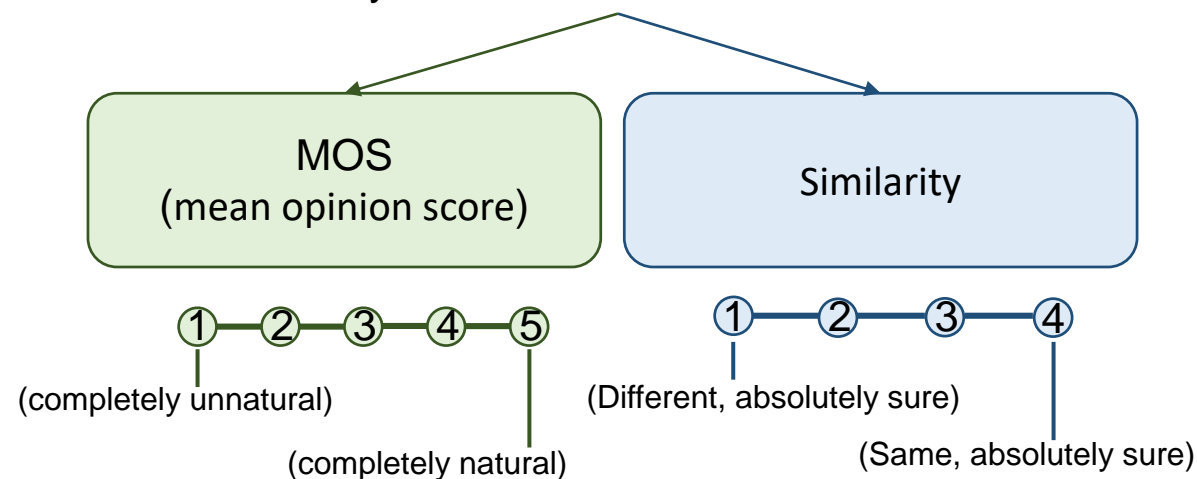


**Conversion system**



**Total: 16 systems**

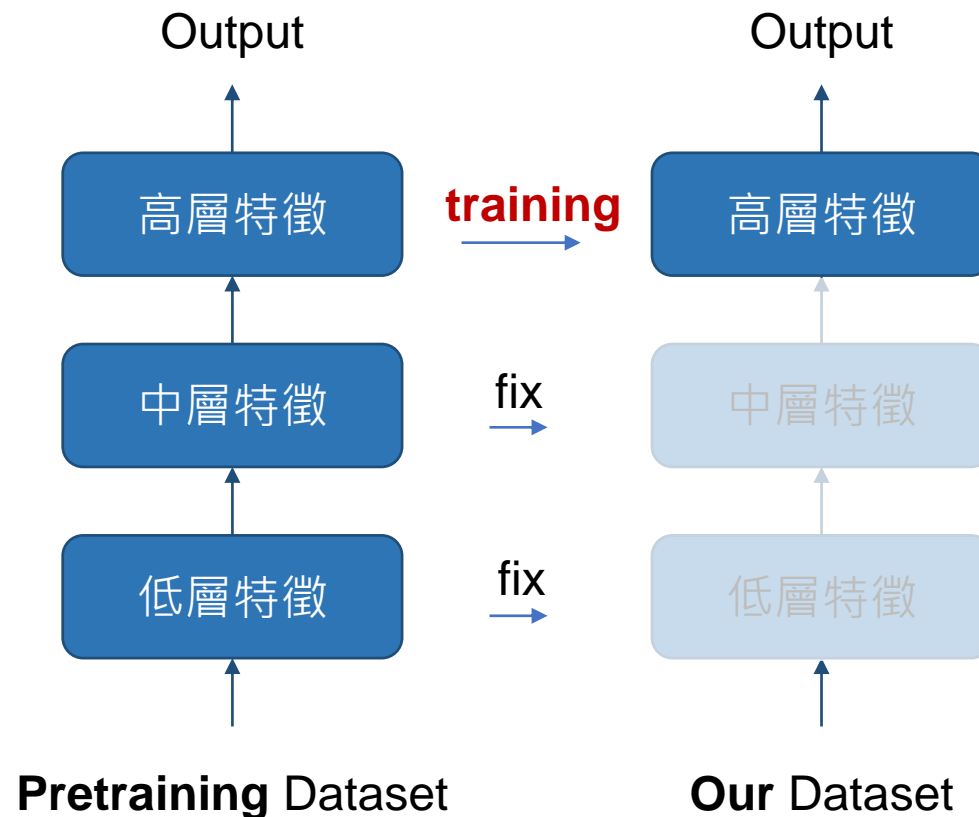
16 systems \* 25 conversion results



**Evaluation methodology**

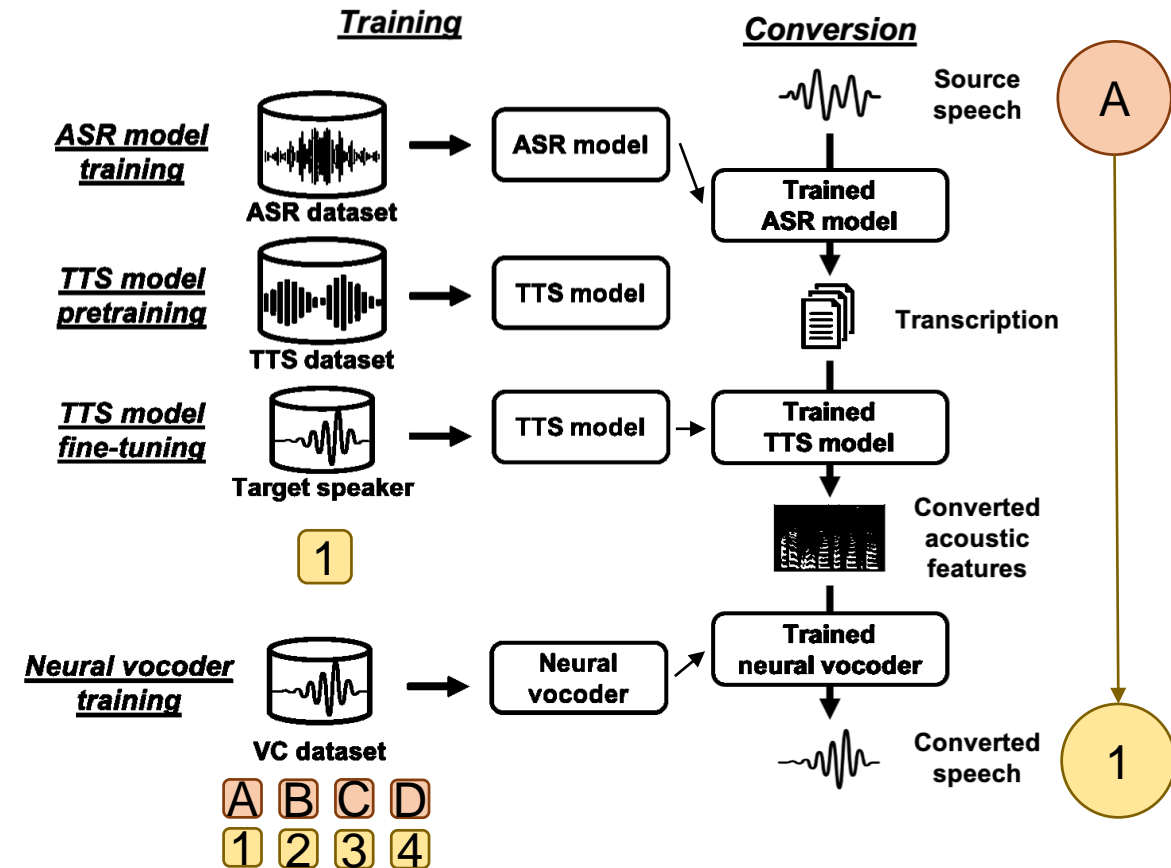
# Transfer learning and fine-tuning

- 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

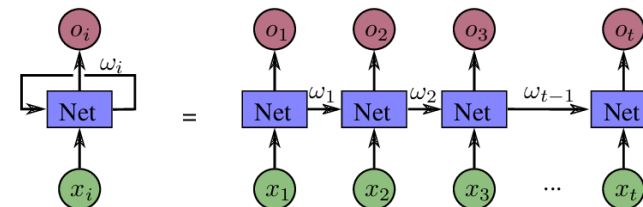
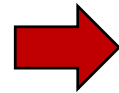
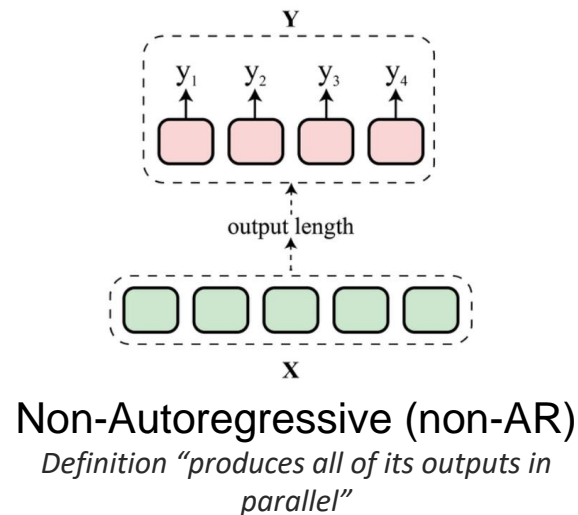
- Task1: voice conversion within the same language.
- A naive approach for VC is a cascade of an automatic speech recognition (ASR) model and a text-to-speech (TTS) model, and both model are using pre-training model.
- ASR models are usually trained with a multi-speaker dataset, thus speaker-independent 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	<b>95.2</b>
		AISHELL	recording	400	178	95.0
	Conformer	aidatatang_200zh	speech data	600	200	94.1
	Transformer	<b>aidatatang_200zh</b>	<b>speech data</b>	<b>600</b>	<b>200</b>	<b>93.6</b>
		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	