VCC 2020 reference design & its application in Mandarin VC

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

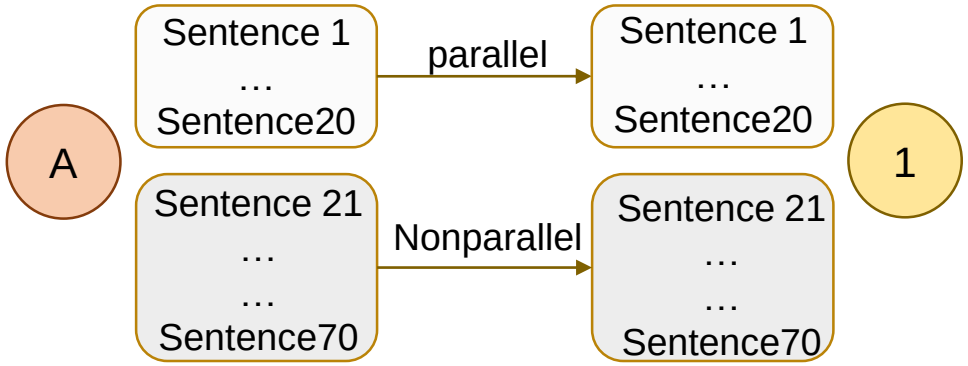
Action item

- 說明參與 VCC2020 應注意的事項，並規劃如何贏得比賽、如何應用到中文轉換

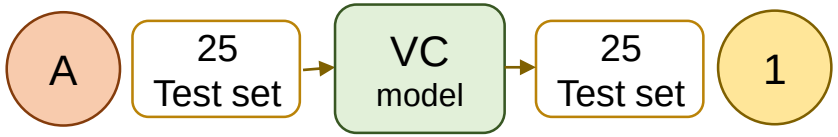
Status report

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

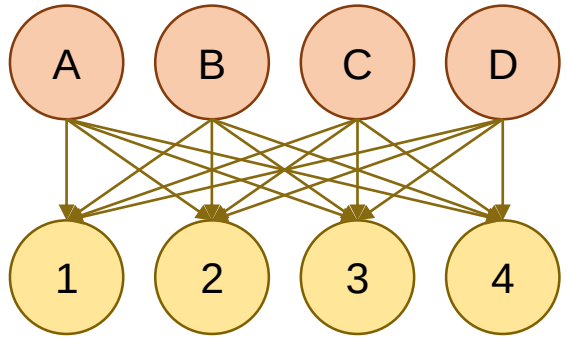
The VCC 2020 Rules



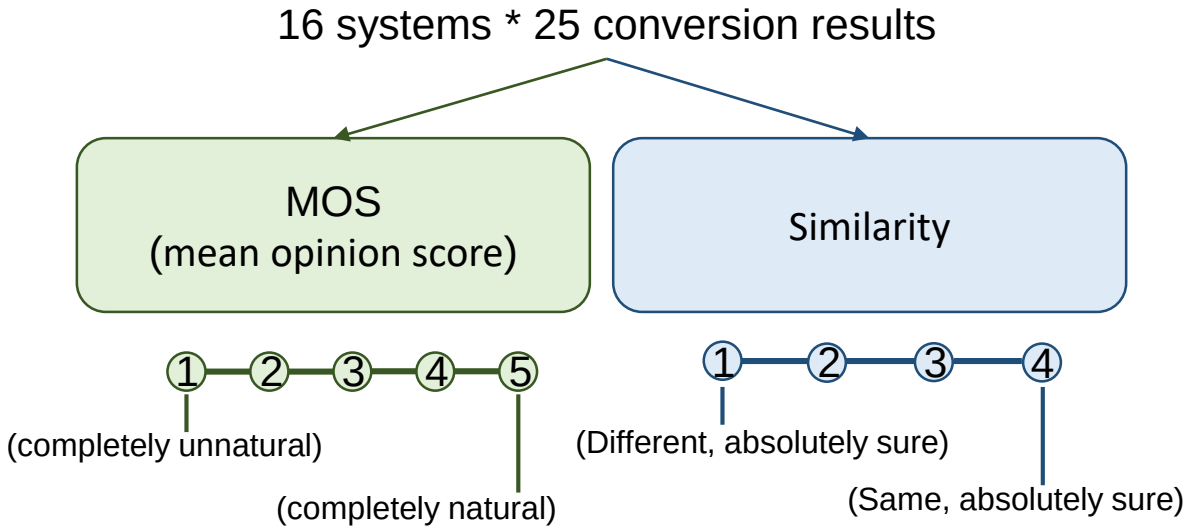
Same language conversion in task1



Conversion system



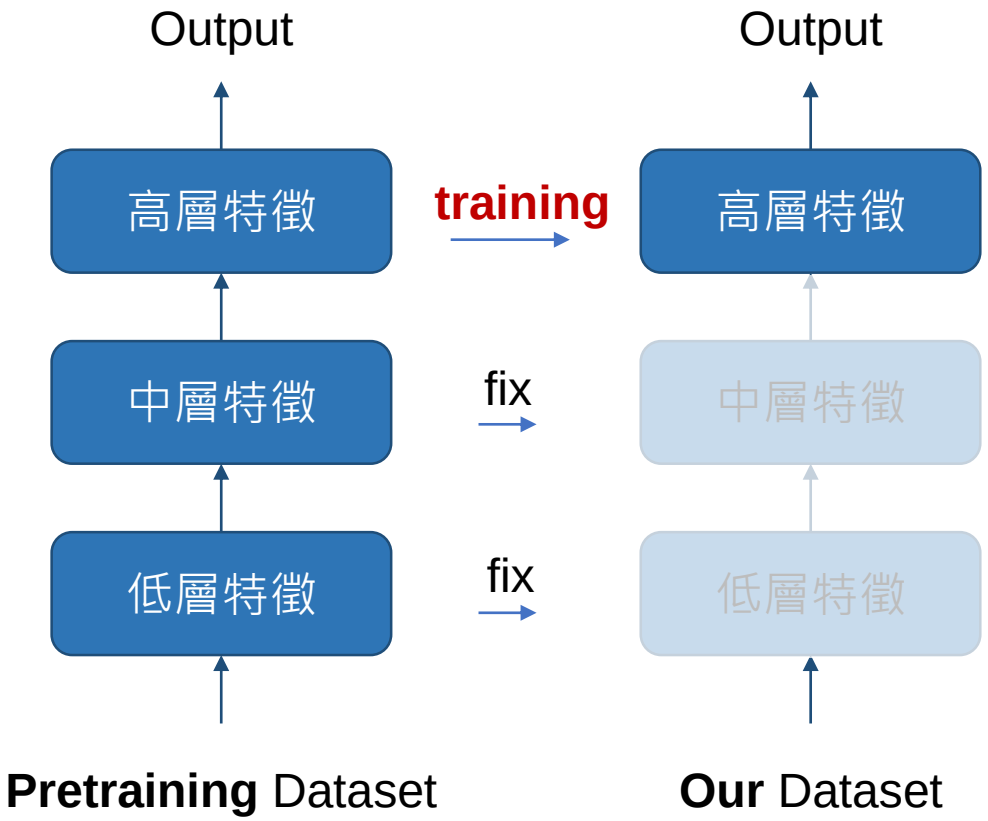
Total: 16 systems



Evaluation methodology

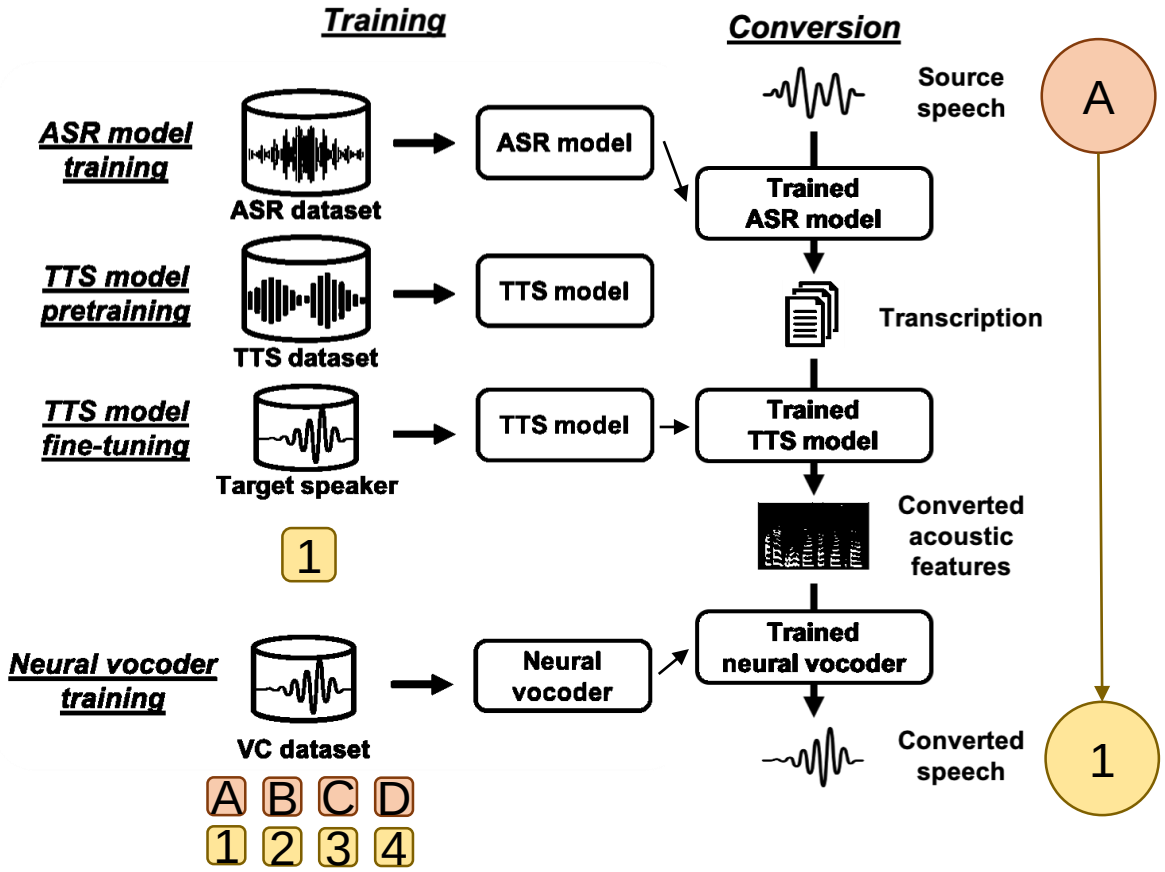
Transfer learning and fine-tuning

- 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.

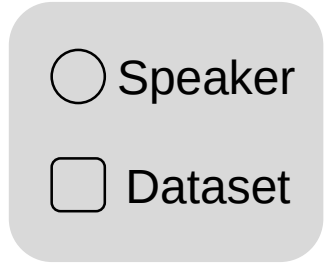


How Baseline Works

- 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.

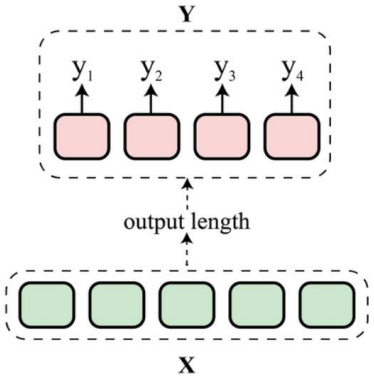


Pre-training model	Language	Dataset	speakers	Hours
ASR	English	LibriSpeech	Over 2000	960
TTS	English	LibriTTS	Over 2000	250



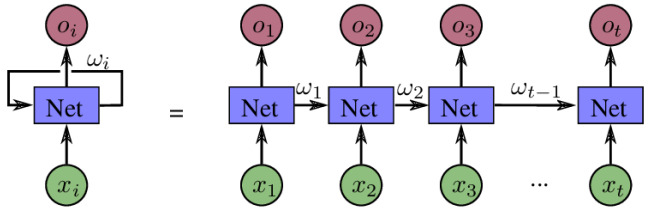
Strategy to win

- 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).



Non-Autoregressive (non-AR)

Definition "produces all of its outputs in parallel"



Autoregressive (AR)

Definition "Each output of the network is generated based on previously generated output"

[Auto-regressive decoding vs non -autoregressive decoding | by Mithilesh kumar singh | Medium]

Its application in Mandarin VC

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

Lang.	ASR pre-training model	Dataset	Speaker	Hours	Test set corr.
Mandarin	Conformer + SpecAugment	aidatatang_200zh	600	200	**95.2**
	Conformer				94.1
	Transformer				93.6
	Conformer + SpecAugment	AISHELL	400	178	**95.0**
	Transformer				93.4
	Transformer	AISHELL2	1991	1000	91.8
	transformer	HKUST電話語音	無標示	約 200	79.1

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					