VCC 2020 reference design & strategy to win

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

Action item

- 說明若作為一位參賽者參與 VCC2020 應該注意的事項，並規劃如何贏得比賽

Status report

1. The brief **introduction** of VCC2020 (p.3)
2. **Restriction** in the tasks of the Challenge (p.3)
3. Converted voice composition of the **dataset** (p.4)
   - 提供的語料分為 source 與 target
   - Target 又分為 task1 & task2
   - 命名規則、語者數量、語料時長
4. **Tasks** of the Challenge (p.4)
   - Task1：半平行的同種語言轉換
   - Task2：不同語言的非平行轉換
5. **Evaluation** method of the Challenge (p.5)
   - 使用 MOS(Mean Opinion Score) 評分
   - 對自然度、說話者相似性兩點做評估
6. **Baseline** Systems provide by organizers (p.6)
   - 提供了三種 baseline system
   - 我使用的系統為 Cascade ASR and TTS
7. Cascading **ASR and TTS** introduction (p.6)
   - ASR 與 TTS 的 model and data
   - Vocoder model
   - 系統訓練以及轉換流程
8. Cascading ASR and TTS **task1** implement **demo** (p.7)
9. **Strategy** to win (p.8)
   - 參考 VCC 2020 大賽第一名的架構
   - 使用更先進的 TTS model
10. Strategy to Change ASR and TTS model (p.9)
    - 將英文轉換模型換成中文轉換模型

The VCC 2020 introduction and restriction

1. The brief **introduction** of VCC2020
   - Voice conversion 語音轉換指的是將 source 語音波形中的說話人身分轉換為不同的身分，並同時保留原本波形語言訊息的技術
   - 可以用於修改聲音的波形，轉換說話人的音色、音調的技術
   - 而這場比賽希望通過提供 common data, metrics and baseline systems 來促進 VC 技術的發展

2. **Restriction** in the tasks of the Challenge
   - 在轉換過程中不允許手動編輯或修改 (像是在轉換過程中手動調整系統參數)，但可以在訓練過程中優化轉換系統
   - 不允許在評估語料上手動標記標籤 (像是 phoneme information、phoneme boundary、linguistic information…等)，但可以在訓練中標記
   - 除了原始 EMIME 語料 (the VCC 2020 database is based on the EMIME bilingual corpus) 不可使用，其餘可以自由使用額外的語料進行培訓
   - 可以在訓練過程中自由地從語料中丟棄一些語句

The VCC 2020 dataset and conversion task

3. Converted voice composition of the **dataset**
   - 提供的語料分為 source 與 target，target 又被分為 task1 與 task2
   - 命名規則 (來源or目標/語言種類/男or女)
     - ex: 來源/英文/男性 **S**ource**E**nglish**M**an SEM
   - Source 語料為各 2 位男性與女性共 4 名 speaker 組成
   - Target1 語料一樣為 2 位男性與 2 位女性組成
   - Target2 語料分為芬蘭文、德文、中文各有1名男性與1名女性組成，每種語言共 2 位 speaker 組成

4. **Conversion tasks** of the Challenge
   - Task1
     - 非平行的同種語言轉換
     - 完成 16 個轉換系統 (i.e., 4 sources by 4 targets)
     - Male to Male、Male to Female、Female to Male、Female to Female
   - Task2
     - 不同語言的非平行訓練 (英文->芬蘭語、英文->德文、英文->中文)
     - 完成 24 個轉換系統 (i.e., 4 sources by 6 targets)

Training dataset

English
(70 sentences)
- Male：2名
- Female：2名

共有 70*4 = 280 句

English
(70 sentences)
- Male：2名
- Female：2名

共有 70*4 = 280 句

Finnish / **G**erman / **M**andarin
(70 sentences)
- Male：1名
- Female：1名

各有 70*2 = 140 句

有 20 句與 source 平行

Testing dataset

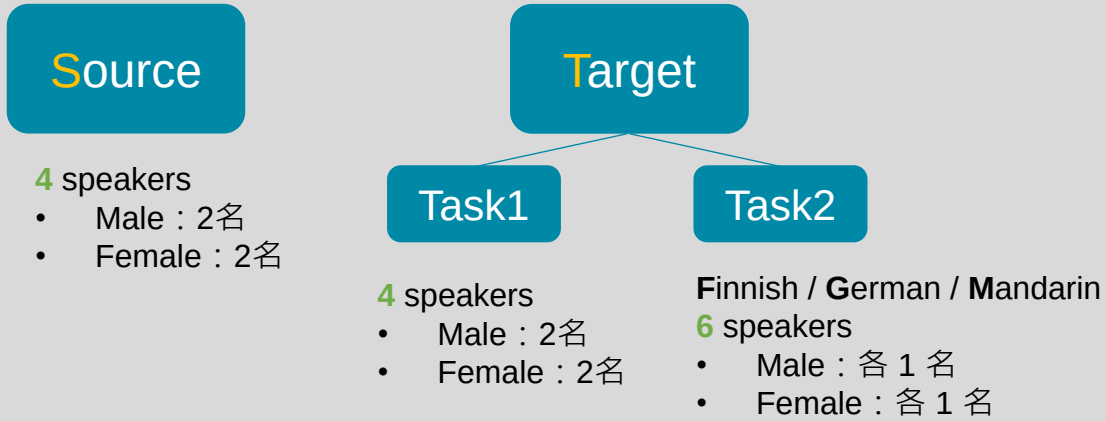
Each speaker provide 25 English sentences
- Male：2名
- Female：2名

共有 25*4 = 100 句

The VCC 2020 evaluation method

5. **Evaluation** method
   - Task1 會生成 400 句轉換語音 (**25** test data * **16** speaker pairs)
   - Task2 會生成 600 句轉換語音 (**25** test data * **24** speaker pairs)
   - 使用 listening tests in terms of naturalness、speaker similarity 進行客觀評估
     - Naturalness: five-point scale MOS (Mean Opinion Score)
     - Similarity: four-point scale score
       - 相同/非常確定、相同/不確定
       - 不同/不確定、不同/非常確定
   - English & Japanese listeners
     - 68 English listeners (33 male and 32 female, and 3 unknown)
     - 206 Japanese listeners (96 male and 110 female)

Training dataset



4 speakers
- Male：2名
- Female：2名

4 speakers
- Male：2名
- Female：2名

Finnish / **G**erman / **M**andarin
6 speakers
- Male：各 1 名
- Female：各 1 名

Testing dataset

Each speaker provide **25** English sentences

The VCC 2020 baseline and choose

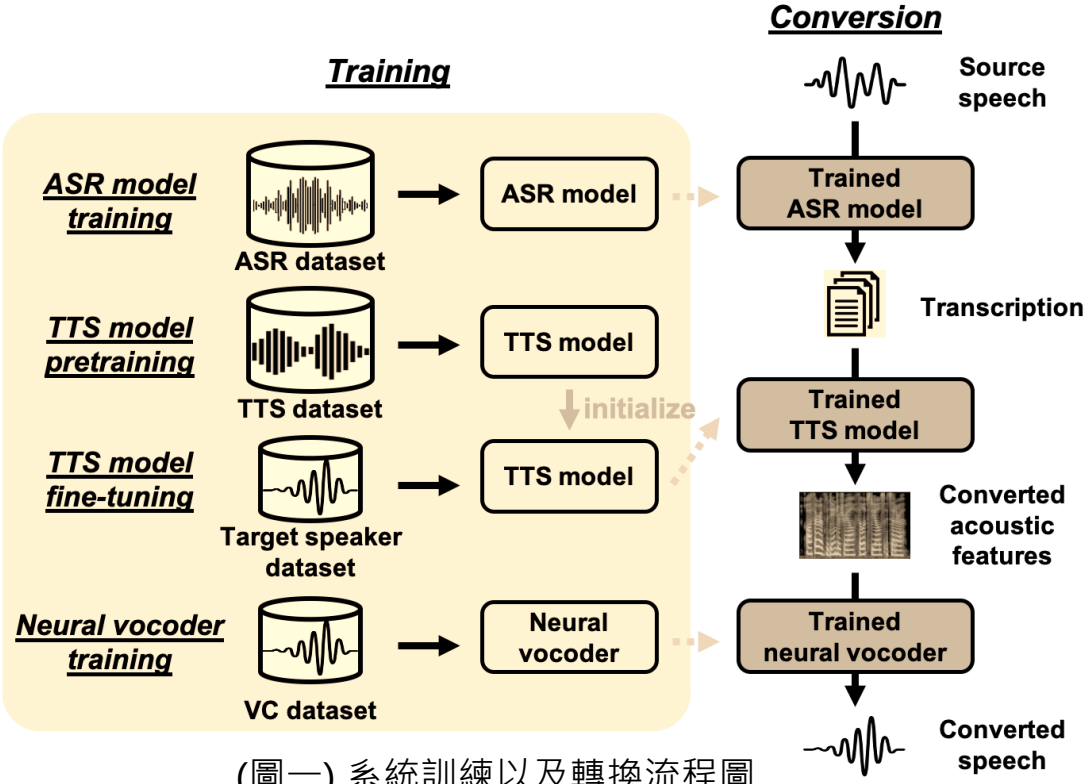
6. **Baseline** Systems provide by organizers
   - 提供了 3 種 baseline system
     1. Top system of VCC 2018
     2. CycleVAE + Parallel WaveGAN
     3. Seq-to-Seq based on Cascade ASR + TTS w/ ESPnet

   而我使用第三種 baseline system，串接 ASR 與 TTS model

7. Cascading **ASR and TTS** introduction
   - ASR
     - **Model**：Transformer w/ hybrid CTC/attention + RNNLM
     - **Data**：Librispeech 960h
   - TTS
     - **Model**：Multi-speaker Transformer-TTS w/ x-vertor
       - Task1
         - **Data**：LibriTTS dataset 250h
       - Task2 **(24kHz 下採樣到 16kHz)**
         - **Date**：使用英文與目標語言訓練 TTS model，並使用目標語言做微調

Lang.	Dataset	Spkrs	Hours	Input
Eng.	M-AILABS [27]	2	32	phn or char
Ger.	M-AILABS [27]	5	190	char
Fin.	CSS10 [28]	1	10	char
Man.	CSMSC [29]	1	12	pinyin

   - Neural Vocoder **(16kHz 映射至 24kHz)**
     - Parallel WaveGAN (PWG)



(圖一) 系統訓練以及轉換流程圖

ASR and TTS **task1** implement demo

輸出入音檔：各 25 個音檔
(取其中一個音檔作為 demo)

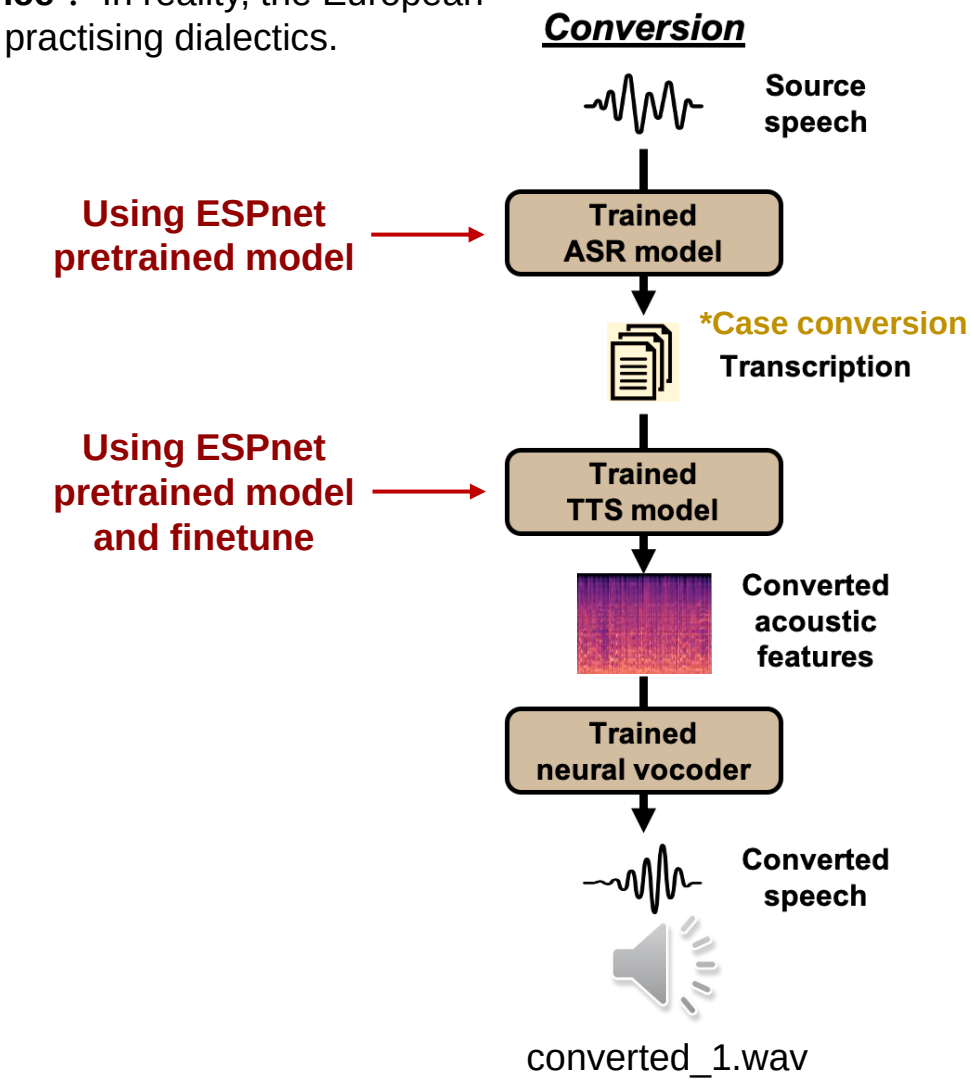
source_1.wav

8. Cascading ASR and TTS **task1** implement **demo**
   - **Example sentence**：Moroccan agriculture enjoys special treatment when exporting to Europe.

Train source voice

Train target voice

Demo sentence： In reality, the European Parliament is practising dialectics.



converted_1.wav

- 其中作者對於 ASR model 的輸出做了一個小改動
- ASR result case conversion
  - ASR 輸出結果皆為大寫，並且沒有標點符號
    - Ex：IN REALITY THE EUROPEAN PARLIAMENT IS PRACTISING DIALECTICS
  - 但 TTS model 訊料資料是 normalized text，因此將 ASR 結果全轉換為小寫，雖然會有部分不匹配，但作者認為是可接受的
    - Ex：Tom, the Piper's Son

Strategy to win

9. **Strategy** to win

1. 在 VCC 2020 大賽中，只有一組在自然度以及說話者相似度皆優於 baseline (Cascading ASR and TTS)
   - 對於任務一 (同種語言的轉換) 使用了兩種方法
     A. ASR-TTS + WaveNet
        - Baseline 為了快速生成語音而使用 non-autoregressive Vocoder，但普遍認為 autoregressive 生成效果會更好
          - Baseline 使用的 Vocoder 為 Parallel WaveGAN (PWG)
          - 自回歸模型：
            1. WaveNet
            2. WaveRNN
            3. LPCNet

     B. PPG–VC (LSTM) + WaveNet
        - 基於 PPG (Phonetic PosteriorGram) -VC 框架，改進 VCC 2018 編號 N10 的系統而來

2. 作者認為使用的 multi-speaker TTS model 是較為簡單的模型，如果使用最先進的模型效果可能更好，像是 [1]
   - [1] 同為 TTS model
   - 有 GitHub source code，但無說明文件

[1] W.-N. Hsu, et al., "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, 2019

Strategy to Change ASR and TTS model

10. **Strategy** to Change ASR and TTS model

- 對於 ASR model 進行以下操作：
  - Baseline (Cascading ASR and TTS) 使用英文資料集 (Librispeech) 作為預訓練的訓練資料，因此是英文的 ASR model
  - 如今要將中文轉換成中文，應該需要使用中文訓練語料的預訓練模型，ESPnet 有提供使用中文語料訓練的預訓練模型
  - 將 Baseline 中的 ASR model 換成同樣為 ESPnet 提供的中文預訓練模型，提供的 pre-trained dataset 有 4 種

Lang.	Dataset	Speaker	Hours
Mandarin	aidatatang_200zh	600	200
	AISHELL	400	178
	AISHELL2	1991	1000
	HKUST電話語音	無標示	約 200

- 同樣對於 TTS model 進行操作：
  - 若使用透過 Librispeech 預訓練的 ASR model，使用中文語料微調與訓練，效果很差就直接跳過
  - 更改使用 ESPnet 提供使用中文語料的預訓練 TTS model，有 1 種

Lang.	Dataset	Speaker	Hours
Mandarin	CSMSC 女聲	1	12