“The sequence-to-sequence baseline for the voice conversion challenge 2020: cascading ASR and TTS”

W.-C. Huang et al., “Speech resynthesis from discrete disentangled self-supervised representations,” *arXiv preprint arXiv:2010.02434* ,2020.

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Introduction

主辦人提供語料、指標還有基準系統 (baseline) 希望參與者可以提升語音轉換的技術而舉辦了 Voice conversion challenge (VCC) 這樣兩年一次的挑戰賽。

從 2016 開始挑戰平行語料的轉換，直到 2020 年為第三屆挑戰賽
第三屆的挑戰賽有兩個主要想要挑戰的任務 (兩個任務的主要概念都是非平行的)：

1. Semi-parallel VC：同種語言內的語音轉換，訓練資料中有 30% 是平行語料，而剩下的為非平行語料
2. Cross-lingual VC：跨語言的語音轉換，訓練資料 source 與 target 完全不同，並希望在轉換過後，聲音和內容與轉換前保持不變

提供的 baseline 基準系統有三種：

1. Top system of VCC 2018
2. CycleVAE + Parallel WaveGAN
3. Seq-to-Seq based on Cascade ASR + TTS w/ ESPnet

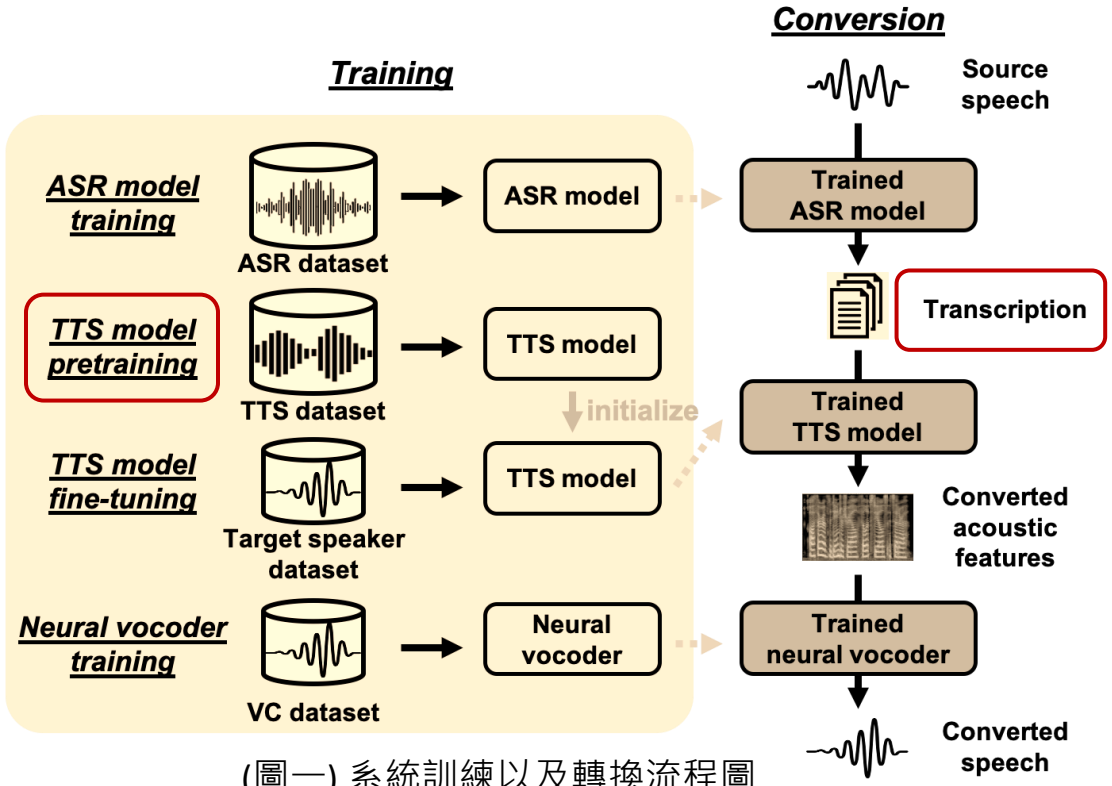
今天要介紹的則是第三種 baseline system：
The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS
會使用 Sequence-to-Sequence VC 是因為它具有好的 prosody 轉換效果
此系統是基於 Sequence-to-Sequence 的基礎下，將 Automatic Speech Recognition (ASR) 與 text-to-speech (TTS) 串接在一起。
雖然人們普遍地認為單純的串接系統它的效果會比 end-to-end (E2E) model 還要差，但作者在本篇論文中展示了，在競賽中被限制使用 public datasets 的情況下，系統不僅易於使用，並且在 VCC2020 中還具有強大的競爭力

System Overview

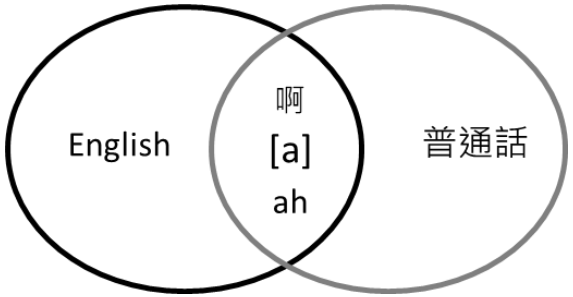
右圖是論文中的系統訓練以及轉換流程圖
其中總共包含了三個 modules，分別是 ASR、TTS 和 Neural vocoder
在轉換過程中，使用 ASR model 將 source speech 辨識成文字，再透過 TTS model 得到聲學特徵，最後再使用一個神經網路聲碼器合成語音。

其中針對兩個點做特別處理 (紅框)

1. 在 TTS model pretraining 的部分，為了避免是用商業或私人的語料，因此選擇不多，使用的語料庫只有 1~5 位語者，從而選擇使用預訓練以及微調的方式做處理。
2. 在任務二跨語言的挑戰中，如果使用 A 語言訓練並轉換 B 語言，用 A 訓練出來的 model 會不知道 B 語言如何發音，導致語音品質很差，為了解決這個問題，做出了一個假設，假設兩輸入語言之間存在重疊 (圖二)，因此在 ASR 辨識的部分就將輸出結果使用 phoneme 表示



(圖一) 系統訓練以及轉換流程圖



(圖二) 假設語料存在重疊

Implementation

ASR model：Transformer [1-3] w/ hybrid CTC/attention [4] + RNNLM (provided by ESPnet)
ASR data：Librispeech 960h, which contained 960 hours of English speech data from over 2000 speakers

TTS model：Multi-speaker Transformer-TTS w/ x-vertor
TTS data：

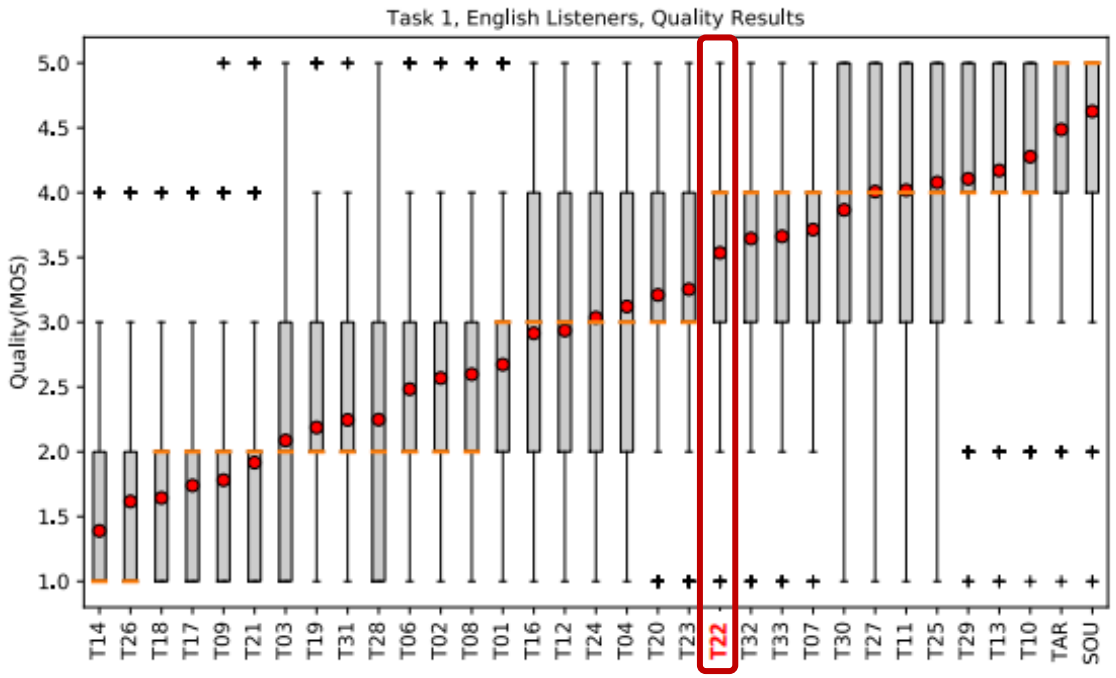
- Task1：因為輸入一直是英文，因此使用 LibriTTS dataset 250h
- Task2：利用英文語料庫以及以下目標語言建立 x-vector-based bilingual TTS models，並使用目標語言做微調

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

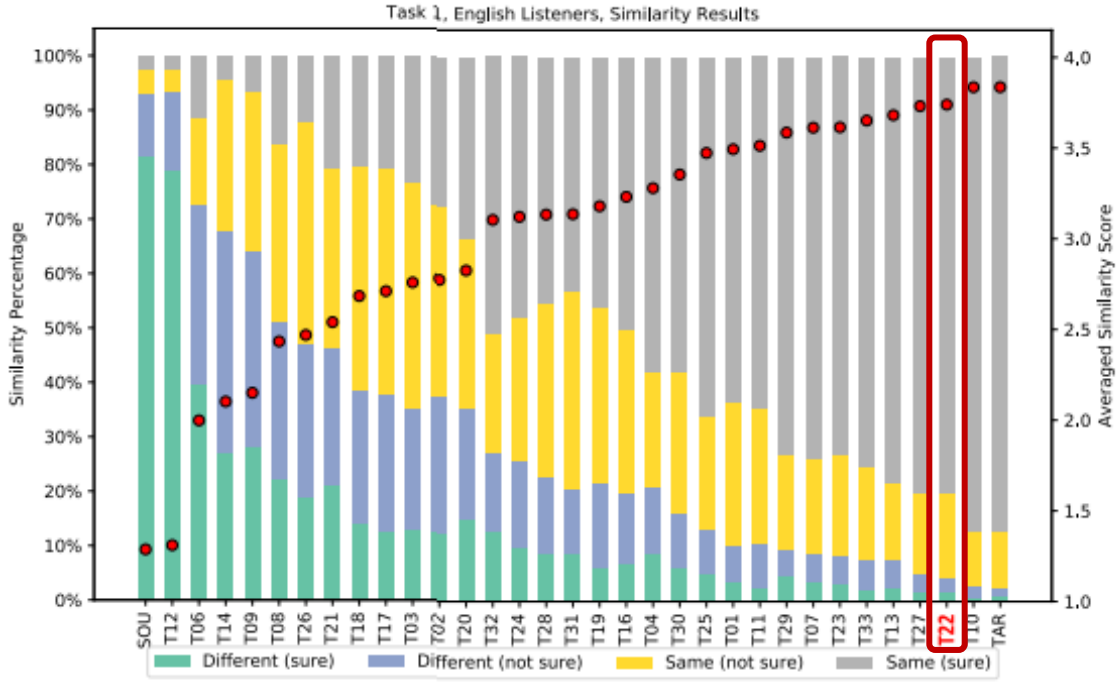
Neural Vocoder：Parallel WaveGAN，為一個 non-autoregressive model，使用它是因為它是一個即時聲碼器

[1] A. Vaswani et al., “Attention is all you need,” in *proc. NIPS*, 2017, pp. 5998–6008.
[2] L. Dong, S. Xu, and B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” *IEEE ICASSP*, 2018, pp. 5884–5888.
[3] S. Karita et al., “Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration,” in *Proc. Interspeech*, 2019, pp. 1408–1412.
[4] S. Watanabe et al., “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE JSTSP*, vol. 11, no. 8, pp. 1240–1253, 2017.

Task1 Results



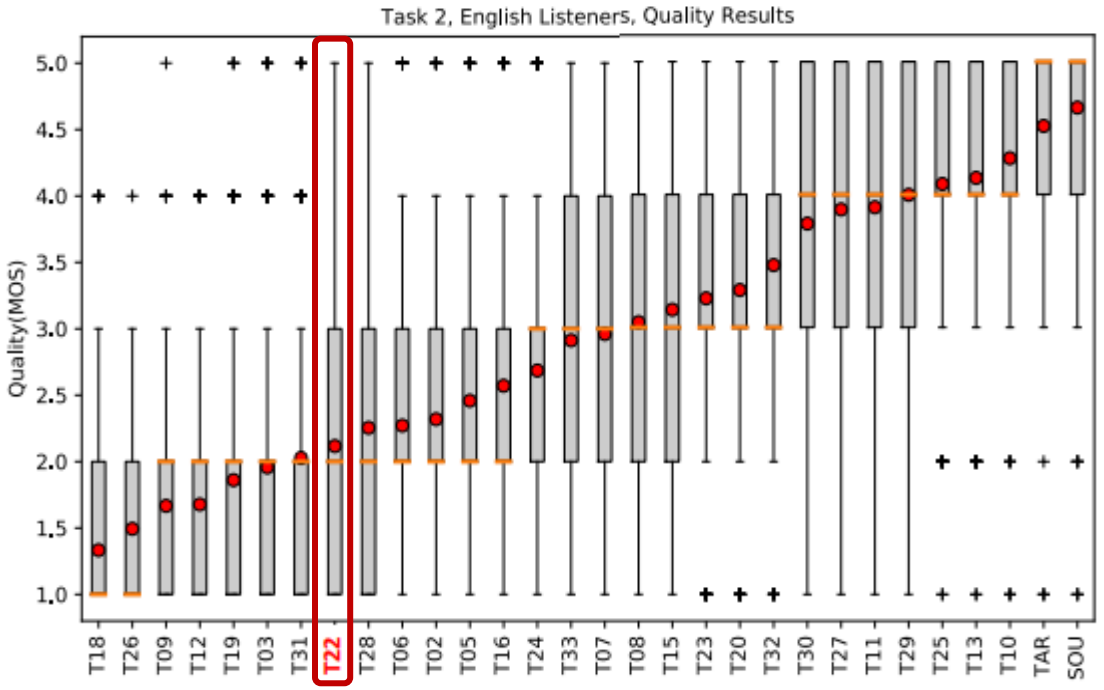
(a) *Naturalness results for task 1.*



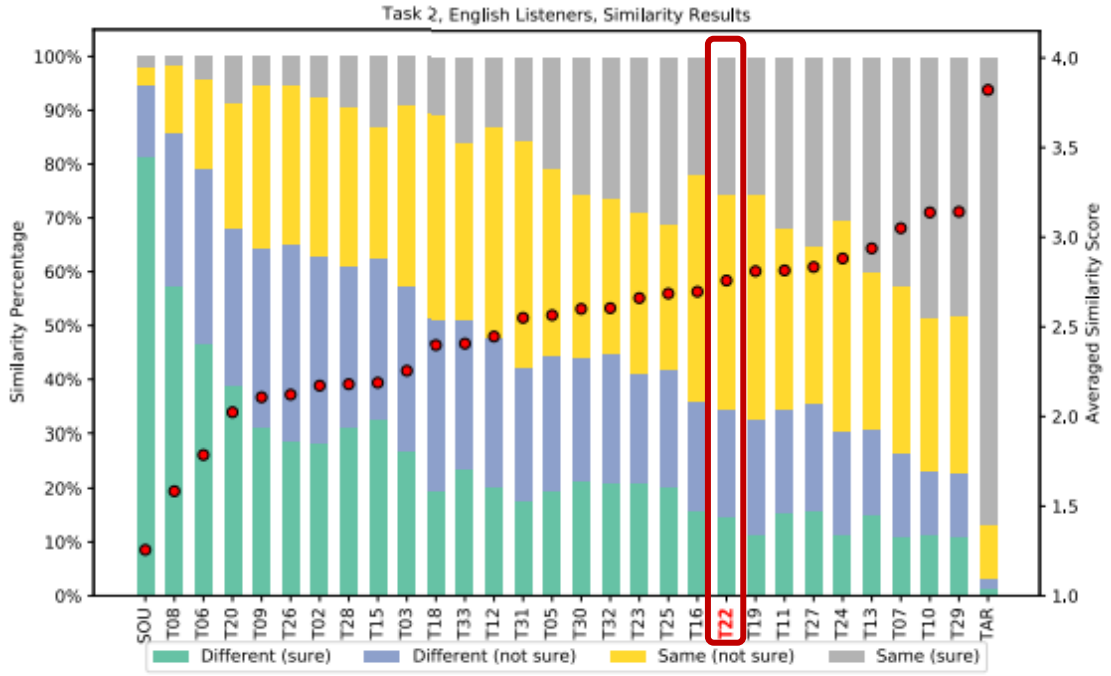
(b) *Similarity results for task 1.*

這是官方對於所有參賽者的評估結果，作者的系統為編號 T22。
在任務一中，此系統在自然度方面在排名的前三分之一，而在相似度方面則在 31 組參賽者中獲得了第二
這展現了 sequence to sequence model 對於語音處理具有良好的效果。

Task2 Results



(c) *Naturalness results for task 2.*



(d) *Similarity results for task 2.*

在任務二中，雖然在自然度方面的表面趨於落後，但在相似度上仍在排名的前三分之一。
對於任務二有著較差的表現，可能由於訓練資料有限，缺乏預訓練資料，導致實作出來的 TTS 過於簡單而無法處理處理跨語言資料。

Conclusion

雖然這次 ASR + TTS 在 VCC2020 中表現突出，但仍有很大的發展空間，以下為其他未來可能的方向：

1. Enhance the pretraining data
   - 任務 2 下的預訓練數據在開源約束下沒有太多選擇
   - 任務 1 中使用更多語者的語料可能會提高性能
   - 使用更高采樣率 (作者使用 16kHz、24kHz) 的數據集也可以提高聲碼器的品質
2. Utilize linguistic knowledge
   - End-to-end (E2E) 學習的一個原則是盡可能少地使用特定領域的知識，也就是說，當這些知識被利用時，系統性能有望得到提高
   - 例如：使用音素輸入可以極大地改進多語言 TTS 系統，但由於不熟悉芬蘭語和德語等目標語言，我們無法在任務 2 中這樣做
3. Select an advanced multi-speaker TTS model
   - 我們採用的 multispeaker TTS model [1] 是一個簡單的模型，像 [2] 這樣更先進的模型可能會提高性能
4. Improve the neural vocoder
   - 我們採用了 non-AR (non-autoregressive) 神經聲碼器進行快速生成，但普遍認為 AR (Autoregressive) 聲碼器仍然更勝一籌
   - 微調聲碼器可以進一步提高性能，像是 mel filterbank 與採樣的 waveform 不匹配導致品質下降

[1] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *proc. NIPS*, 2018, pp. 4480–4490.
[2] W.-N. Hsu et al., "Hierarchical generative modeling for controllable speech synthesis," in *proc. ICLR*, 2019.