

Wang TTS (based on VCC 2020 ref. design): zhenwei (震威) results & TTS training process

W.-C. Huang, T. Hayashi, S. Watanabe, T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: cascading ASR and TTS," arXiv preprint arXiv:2010.02434, 2020.

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

Action item

使用震威的語料進行 TTS 微調並報告 TTS 的訓練流程

Status report

背景概要

- 我使用 VCC 2020 reference 的 VC baseline，架構由 2 個 model 組成，分別是 ASR + TTS，而兩個 model 互相獨立，ASR 將文字辨識出來後就可以與整個系統切開來，因此我們著重在 TTS model 的部分 (如圖一)。
- 先前使用不同的語料，有不同的合成品質，因此再使用不一樣的 input 作為輸入，這次使用震威的語料進行實驗。

本周進度

- 整理 TTS 微調訓練步驟
 1. 使用的 TTS 規格是什麼？
 2. 使用的 TTS 厲害的点在於哪裡？
 3. 微調前要準備什麼？
 4. 微調步驟

後續實驗規劃

- 在論文中的 references 看到使用多個 speaker 訓練 TTS 模型可以生成比單一語者還有更好的品質與穩定性 [1]，因此想嘗試使用此方法是否會更好。

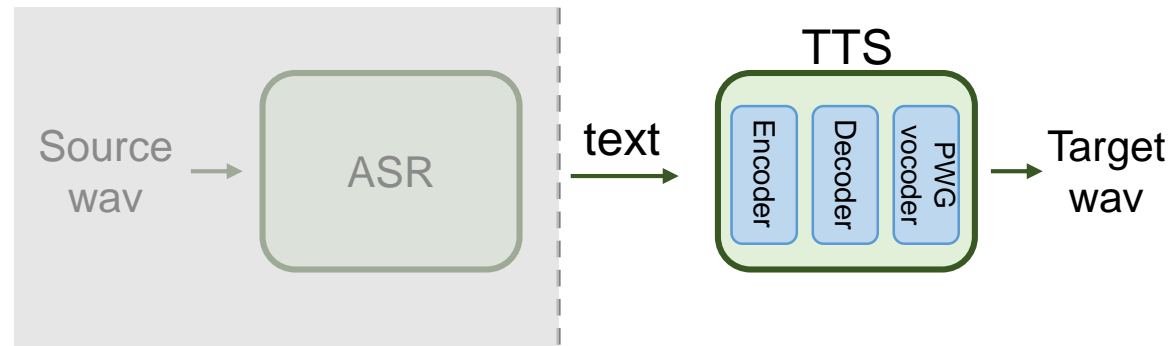


Figure 1: System structure

[1] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, "Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora," in *Proc. Interspeech*, 2019, pp. 1303–1307.

TTS - Transformer

VCC2020 Baseline **English pre-training model**: multi-speaker, x-vector Transformer-TTS model

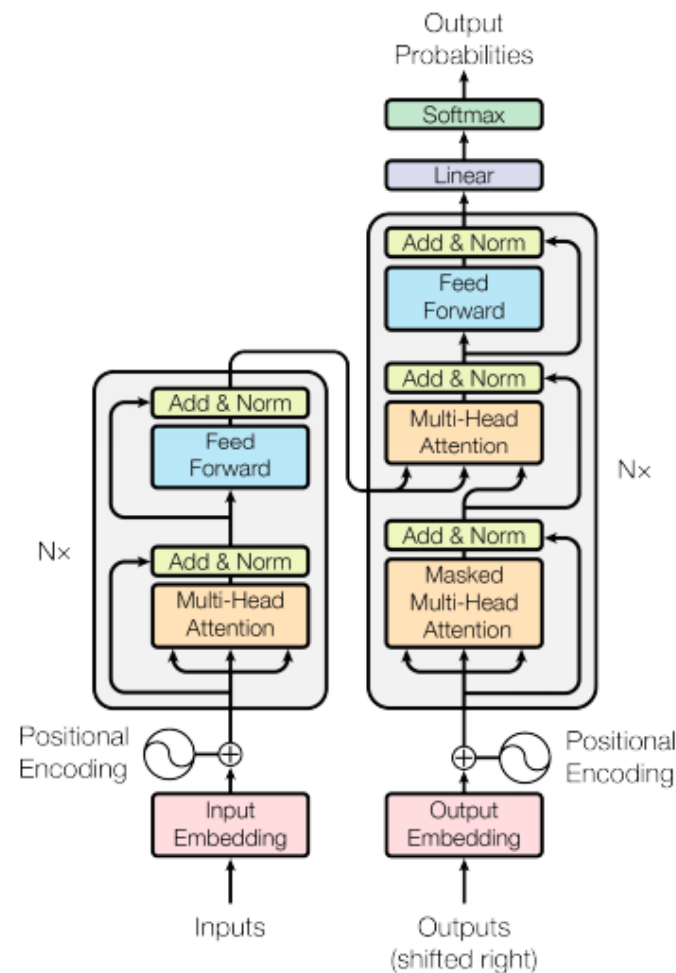
Chinese pre-training model: Initial Transformer

x-vector 也是一個神經網路，作用為可以接受任意長度的輸入，然後轉換成固定長度的特徵表示。

Transformer 是近代語音處理的一個熱門神經網路，在 2017 年由 Google 提出 “**Attention Is All You Need**” 這篇論文中出現，它由一組 encoder 與 decoder 組成。

Transformer 比 RNN 厲害的地方：

待補



The Transformer - model architecture.

Before training

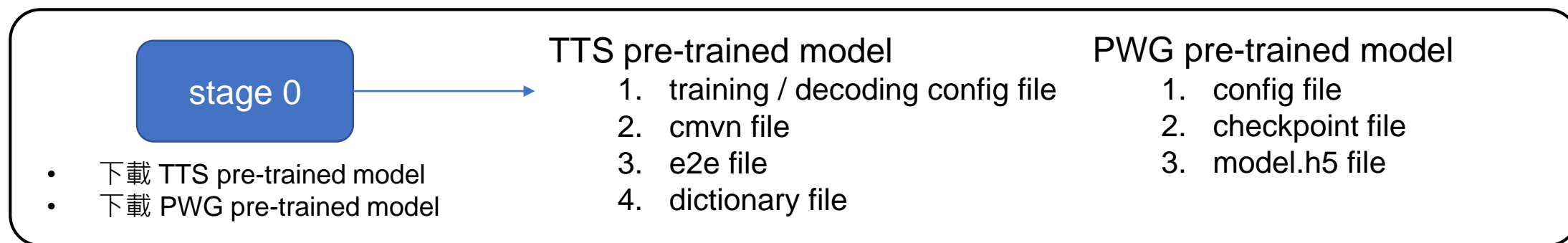
論文與 ESPnet 沒有標示至少要多少語句訓練效果會好，而論文遵從 VCC2020 競賽中，給予 70 句語句的限制，使用 60 句作為訓練，10 句作為驗證。

Baseline: 60/10

Currently version: 320/10

在訓練前我們要先下載 TTS (Text-to-speech) 與 PWG (Parallel WaveGAN) 的預訓練模型，會得到以下 7 個檔案。

並準備要訓練的語句與文本。



Training processes

TTS 訓練流程

1. Data preparation

- 選擇語料
- 建立語料與文本的關聯檔
- 降頻至 16kHz
- 檢查準備的資料目錄、格式是否正確

2. Feature Generation

(使用 TTS 預訓練中計算的統計資料，對特徵進行標準化)

- 切除空白音檔
- 生成 fbank
- 生成指定的 train、test 語句列表
- 使用預訓練 cmvn 取 train、test feature

3. Dictionary and Json Data Preparation

- 使用 TTS 預訓練中內置的字典對標記進行索引

4. x-vector extraction

- 生成 MFCC 並計算 energy-based VAD
- 對於 Kaldi-based X-vector pretrained model 提取 X-vector

5. fine-tuning

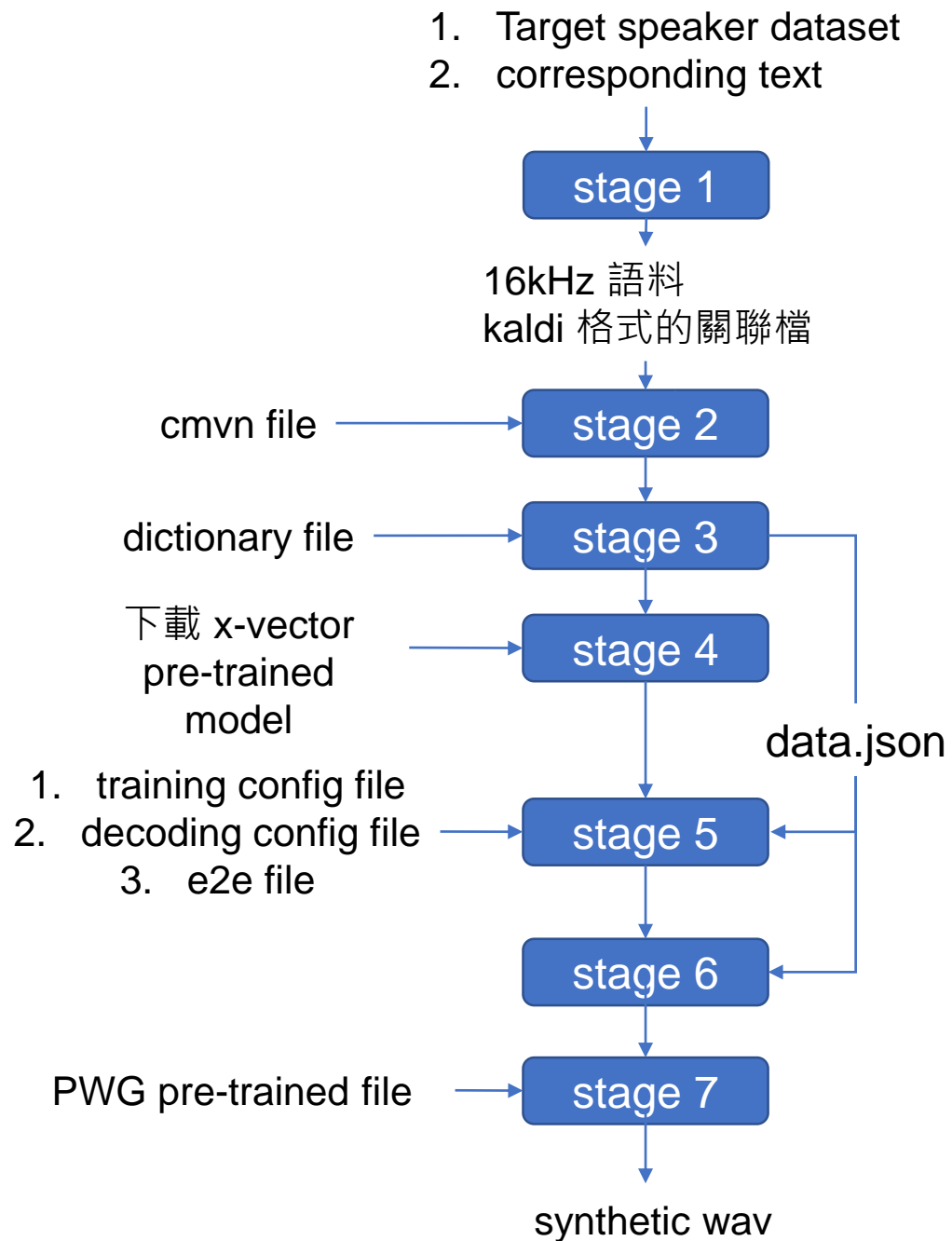
- Train E2E-TTS model (encoder)

6. Decoding

- 對於 test set 做 TTS 解碼 (decoder)

7. Synthesis


- 使用訓練過的 Parallel WaveGAN 將生成的 mel filterbank 轉回波形




Conversion result


版本 1、2 使用腳本中切除訓練用音檔前後的無聲區域，但會有部分音檔沒有切乾淨，以致**部分**合成結果音檔前會有無聲的問題 (圖二)，因此延伸三種方法在執行腳本前去除訓練音檔前後空白



版本	版本區別	訓練 / 測試 語句數量	音質表現
1	在執行腳本前沒經過處理	60 / 10	4
2	在執行腳本前沒經過處理	310 / 10	2
3	使用 Audacity 的截斷靜音	310 / 10	3
4	使用 IA 的 DTW	310 / 10	3
5	手動切除前後空白音檔	310 / 10	2
6	手動切 + 額外錄製語料	320 / 10	1



音色：  這學期學校有書法比賽

 全家人都非常  爸爸戴老花眼

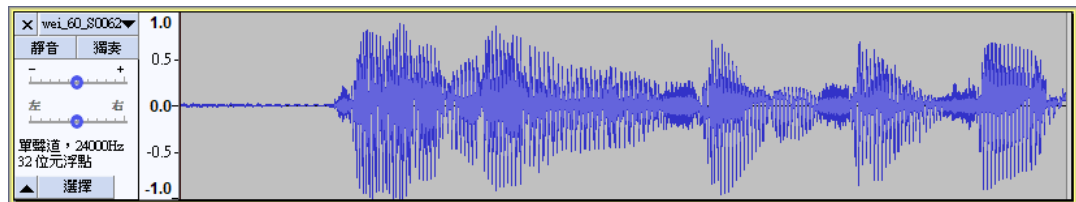
 他不小心把茶杯  他做完功課才

 他不小心把茶杯  他做完功課才

 他不小心把茶杯  他做完功課才

 他不小心把茶杯  他做完功課才

 這禮拜的天氣早  人生就像騎腳踏車想保持平衡



圖二：合成結果音檔前有空白

使用 Audacity 截斷靜音功能

使用 IA Lab 提供的 DTW 功能

手動切除前後空白音檔

Step 3 - Dictionary and Json Data Preparation

這個例子為 test 語句的 json 格式，此範例使用編號語句 311~320 作為測試語句，下圖為其中一筆標號 311。
重點在於會使用左方字典將語句作編號

```
1 <unk> 1
2 a1 2
3 a2 3
4 a3 4
5 a4 5
6 a5 6
7 ai1 7
8 ai2 8
9 ai3 9
10 ai4 10
11 ai5 11
12 air2 12
13 air4 13
14 an1 14
15 an2 15
16 an3 16
.
```

Dictionary

```
data.json
~/chullin/espnet/egs/vcc20/vc1_task1_change_Mandarin/dump/wei_hand_batchChange_dev

1 {
2   "utts": {
3     "wei_hand_batchChange_S0311": {
4       "input": [
5         {
6           "feat": "/home/ec1/chullin/espnet/egs/vcc20/vc1_task1_change_Mandarin/dump/wei_hand_batchChange_dev/feats.1.ark:27",
7           "name": "input1",
8           "shape": [
9             290,
10            80
11          ]
12        },
13        {
14          "feat": "exp/xvector_nnet_la/xvectors_wei_hand_batchChange_dev/xvector.1.ark:27",
15          "name": "input2",
16          "shape": [
17            512
18          ]
19        }
20      ],
21      "output": [
22        {
23          "name": "target1",
24          "shape": [
25            1
26          ],
27          "text": "zh uo1 z i5 sh ang4 b ai3 l e5 i1 d a4 p an2 g ual z i5",
28          "token": "zh uo1 z i5 sh ang4 b ai3 l e5 i1 d a4 p an2 g ual z i5",
29          "tokenid": "335 303 334 152 253 98 113 85 227 121 148 116 81 248 91 146 261 334 152"
30        }
31      ],
32      "utt2spk": "wei_hand_batchChange"
33    }
34  }
```