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

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

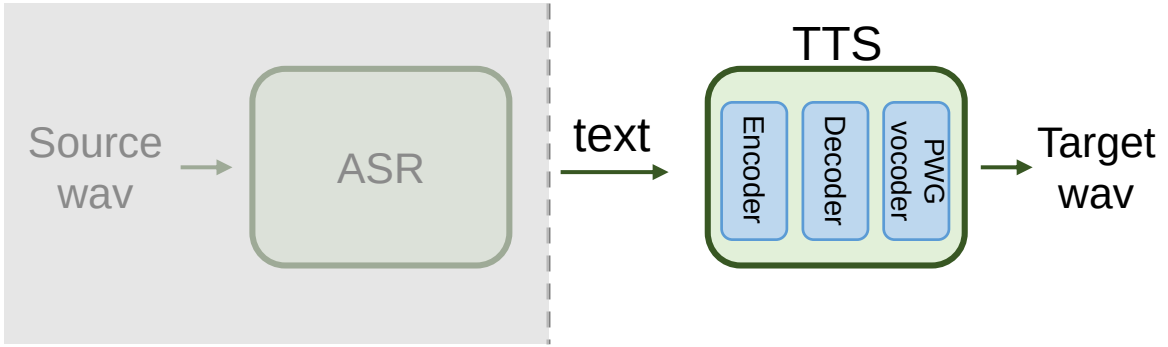


Figure 1: System structure

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。

目前狀況

- 對於震威的語料，總共做了 6 個版本，表格與音檔整理於次頁 (p. 3)。
- 訓練流程總共分為 7 個步驟 (p. 4)。

結果&結論

- 同樣的訓練語句數量，女生合成的品質較男生好，男生的合成結果會有一些雜音，聲音聽起來較不乾淨。
- 使用的 TTS pretrained model 是透過 csmsc 中國標準女聲音庫訓練而成，因此猜想預訓練模型對於女聲 (高頻) 的學習效果較好，而女聲大部分都是高頻，男聲則是在氣音的部分才有較多高頻，所以男聲語料的訓練結果才會出現一些雜音。

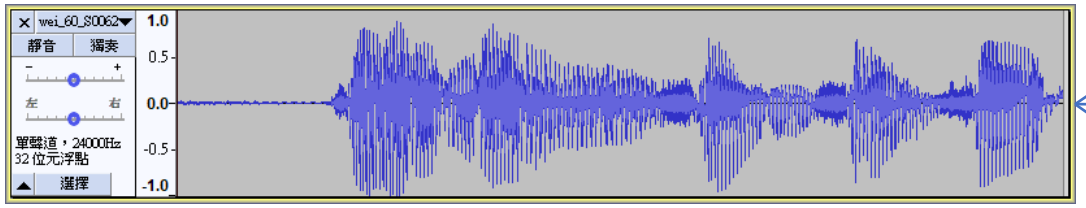
後續實驗規劃

- 找變聲軟體，使用變聲器將男女聲的語料作調換，查看是否由男變女的語料會表現的比較好。

Conversion result

版本 1、2 使用腳本內切除訓練用音檔前後無聲區域，會有**部分**音檔合成出音檔前有無聲的問題 (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

編號	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手動處理	60 / 10	4	全家人都非常	爸爸戴老花眼
2	在執行腳本前沒經過手動處理	310 / 10	2	他不小心把茶杯	他做完功課才
3	使用 Audacity 的截斷靜音	310 / 10	3	他不小心把茶杯	他做完功課才
4	使用 IA 的 DTW	310 / 10	3	他不小心把茶杯	他做完功課才
5	手動切除前後空白音檔	310 / 10	2	他不小心把茶杯	他做完功課才
6	手動切 + 額外錄製語料	320 / 10	**1**	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

Training processes

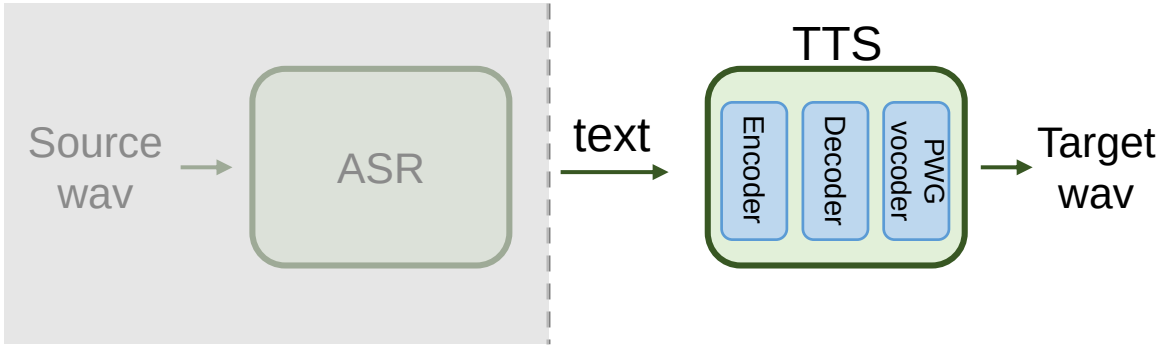
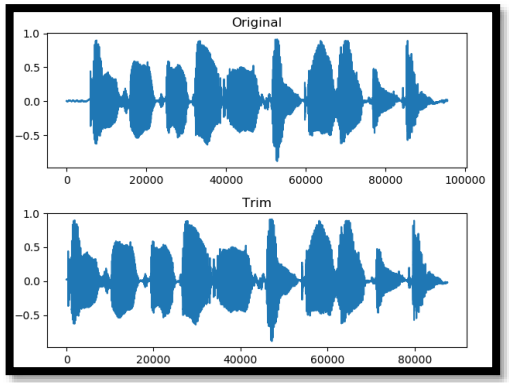


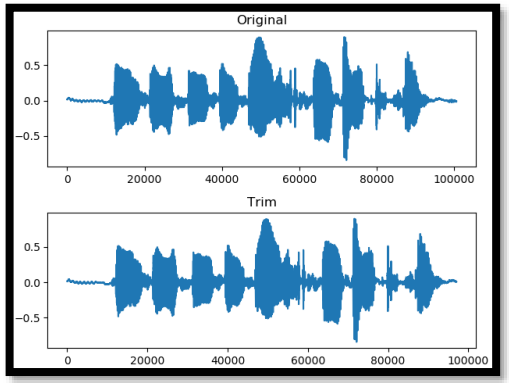
Figure 1: System structure

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 16K
   - 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數據對特徵進行歸一化)
   - 切除空白音檔
   - 生成 fbank
   - 生成指定的 train、test 語句列表
   - 使用 cmvn dump pretrained model 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、計算 energy-based VAD
   - 對於 Kaldi-based X-vector pretrained model 提取 X-vector
   - 在 config 中加入 TTS pretrained model 資訊
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



大部分音檔可以正常切除



部分音檔前無聲區沒有被切除