Research plan for Feb. ~ June 2022

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

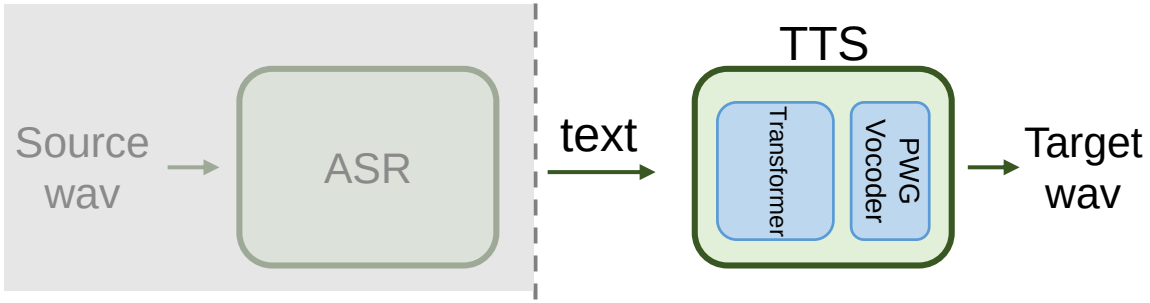


Figure 1: System structure

Action item

規劃自己的半年研究，並配合王老師於2月底將舉辦的海報展進行演練

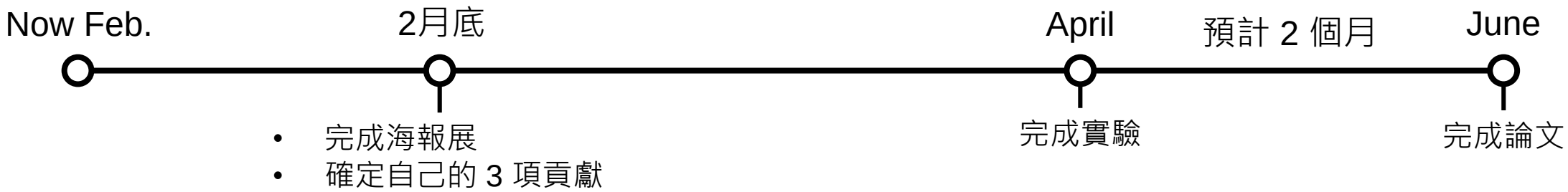
Introduction

我目前的研究使用 VCC2020 的 VC baseline，架構是由 ASR 與 TTS 兩個 model 組成，兩個 model 除了互相獨立之外，ASR 辨識出的結果只要音譯正確就不太會影響 TTS 的生成品質，因此轉換重點放在 TTS model 上。

TTS model 中分成兩個部分，一個是神經網路，另一個則是將神經網路學習到的特徵生成語音，稱為 Vocoder
其中神經網路所使用的是一個近代語音處理熱門的技術 – Transformer

Research plan

1. Problem to be solved
2. Related works (how the problem was solved)
3. Innovations (key ideas of your approach)
4. Experiment setup (how you verify your idea & why your approach has better results)
5. Plan for Feb. ~ June 2022



Problem to be solved & innovations

1. Vocoder: 目前使用的是 Parallel WaveGAN (PWG) 是一種非自回歸模型，犧牲音質提升運算效率的技術，如果只考慮提升音質，可以將 PWG 更換成自回歸模型，像是 WaveNet

2. Transformer:
   - Transformer 完全依賴 self-attention 並完全拋棄 RNN，雖然有不錯的效果，但這下降了它對於局部特徵的抓取能力。
   - Transformer 中的 self-attention 因為是所有 input 都參與訓練，因此它並沒有捕捉輸入順序的能力，需要額外的 position embedding，但論文中所使用的位置編碼雖然可以知道距離，卻無法得知順序。

   目前使用的 TTS model 為預訓練模型，因此希望不更動現有架構，從現有的步驟中添加額外的處理，使 Transformer 的表現提升。

3. 目前程式會將語料降頻至 16KHz 進行操作，除此之外，實驗室的語料也都是 16KHz，若將輸入的 sample rate 提升或許可以提升生成品質。

4. 在訓練之前會下載 x-vector 預訓練模型直接使用，可以嘗試使用 i-vector 是否能改善音質

5. 程式中在處理空白音檔時容易切到頭尾，可以嘗試將其改善

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