畢業論文狀態回報

Sian-Yi Chen

Advisors : Tay-Jyi Lin and Chingwei Yeh

Outline

- **論文題目**
  - **關於自己論文內容的5個問題**
    1. 要解決什麼問題？
    2. 過去別人用什麼方式解決？
    3. 那些解法有什麼不足的地方？
    4. 提出什麼方法為什麼比他們好？
    5. 實驗結果說明好在哪裡？
  - **論文摘要**

- **突破點**
  - 實作Wang老師TTS
  - 提升音質
  - 做實驗證明串接ASR與TTS這種簡單架構是有競爭力的

- **目前進展**
  - **statistical parametric synthesis**
    - Synthesis part剩下MLSA filter看到一半
    - Training part還有單音素HMM擴展到三音素的過程與確認聚類方式
  - **multi-speaker, x-vector Transformer-TTS model**
  - **實作 Wang TTS**

E2E Wang TTS (based on VCC 2020 ref. design) & Conventional TTS (statistical parametric synthesis)

1. **要解決什麼問題？**
   - 使語音合成出來的音質接近人類的聲音。

2. **過去別人用什麼方式解決？**
   - 先透過文本分析（提取語句中的音節、音素、重音位置）取得語言學特徵，
   - 再將這些語言學特徵使用統計模型（Hidden Markov Model, HMM）生成聲學特徵，
   - 最後將聲學特徵使用Vocoder還原成波形。

3. **那些解法有什麼不足的地方？**
   - 在文本分析階段不僅需要具備語音學的知識，而且提取特徵的過程繁瑣。
   - 合成的聲音仍像機器人的聲音，易與人聲區別。

4. **提出什麼方法為什麼比他們好？**
   - 簡化傳統文本分析階段，語音學特徵交由神經網路（Transformer）自行學習訓練。
   - Transformer中使用了attention機制，與RNN相比，不僅減輕了運算效率，而且就算針對非常長的文本序列也不 會忘記前後文。

5. **實驗結果說明好在哪裡？**
   - 合成的聲音不僅像人聲，而且具備說話人的韻律，若要更改話說人僅須透過fine-tune就可達成。
   - 請實驗室所有人對音檔進行MOS評分，是否與對目標對象音色、聲調相似。

論文摘要

- 介紹什麼是TTS，並說明目標為追求合成出接近人類聲音的方法
- 介紹傳統中最具代表性的語音參數合成statistical parametric synthesis，並說明其優缺點
- 以參數合成為分界，進入end-to-end 神經網路的現代，為現在神經網路做分類
- 基於VCC2020選用Transformer作為現代神經網路，介紹架構並說明其優缺
- 傳統方法與Transformer各點比較
- 實作Wang TTS的過程，並比較Wang TTS與VCC2020 baseline的MOS分數
- 將fine-tune完成的TTS串接ASR，說明ASR辨識錯誤並不會影響TTS的效果，並再次證明ASR+TTS這種簡單架構對於語音合成來說是具有競爭力的

突破點

1. 使用自行準備(王老師)的語料，對現有的TTS模型做微調，並合成出帶有王老師說話風格的語句

2. 請實驗室所有人幫忙測試MOS分數，證明音質與baseline相比有所提升

3. 將完成的TTS串接ASR，並再次證明ASR+TTS這種以往認為過於簡單的架構合成出來的語音是具有競爭力的

目前進展

statistical parametric synthesis

- **Synthesis part**
  - Context-dependent Label
  - Conversion Process of Text-to-Label
  - Letter to Sound Rules
  - 持續時間模型
  - 參數生成spectral與excitation序列
  - X Excitation生成和語音合成MLSA濾波器
- **Training part**
  - Excitation & spectral 提取
  - Hidden Markov Model (HMM)介紹
  - 訓練HMM模型
  - X 單音素HMM擴展至三音素HMM
  - X 利用決策樹做上下文聚類取得最佳參數

multi-speaker, x-vector Transformer-TTS model

- Feature representation (MFCC)
- Embedding (x-vector)
- TTS model (Transformer)
- Attention mechanism
- Vocoder (Parallel WaveGAN)

實作 Wang TTS

- 前處理王老師音檔
- fine-tune
- MOS評分