Wang TTS (based on VCC 2020 ref. design): survey traditional & speaker style TTS

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

Action item

1. 介紹 TTS 是什麼？情境題：試官問：「TTS做了什麼東西？有個人風格的TTS與一般的TTS差別又在哪裡？」
2. 傳統的TTS是什麼？帶個人風格的TTS是什麼？有哪些種類？我使用的是哪一種？
3. paper中所使用TTS的技術是什麼？它突破了什麼事情

Status report

1. 「TTS為文字到文本的語音合成，大致可以分為傳統TTS，與現在主流的技術，端到端神經網路的語音合成。傳統TTS像是發音合成、共振峰合成、拼接合成、統計參數合成，這些被歸類到傳統方法的TTS都需要具備足夠的先備知識，像是語言學、聲學、訊號處理等，而神經網路的TTS不僅大幅下降先備知識的需求，合成的品質也更接近真人的聲音，在神經網路TTS中，大部分的TTS都是追求高音質、低運算量，而我所做的研究除此之外還希望TTS可以發出指定對象的音調、音色。」
2. 傳統 TTS 可分為以下四種 **(p.3)**
   - 發音合成 (Articulatory Synthesis)
   - 共振峰合成 (Formant Synthesis)
   - 拼接合成 (Concatenative Synthesis)
   - 統計參數合成 (Statistical Parametric Synthesis)

   現今主流TTS
   - **Tansformer**
   - Tacotron
   - Fastspeech
3. Transformer 的定義、attention 公式、x-vector 作用、在 paper 中突破的事情 **(p.4)**

Category of TTS

傳統 TTS 可分為以下四種

- **發音合成 (Articulatory Synthesis)**：為模擬人類發聲器官的行為來產生語音，理想情況下應該是最好的語音合成方法
  - 缺：因為難以搜集咬合等發聲的聲音，因此合成語音品質最差
- **共振峰合成 (Formant Synthesis)**：根據一組控制簡化的源過濾器模型 (source-filter model) 的規則生成語音，以盡可能接近地模仿語音的共振峰結構和其他頻譜特性
  - 優：可以產生高度可理解的語音，計算資源適中，適合於嵌入式系統
  - 缺：合成結果聽起來很假不太自然，除此之外沒有標準合成規則
- **拼接合成 (Concatenative Synthesis)**：依賴於存儲在數據庫中的語音片段的串接
  - 優：音質清晰度高、音色接近原始講者的聲音
  - 缺：需要龐大的錄音數據庫、合成聲音因為平滑度降低等原因導致不太自然也沒有感情
- **統計參數合成 (Statistical Parametric Synthesis)**：生成語音所需要的聲學參數，然後透過數學方法恢復語音，其中包含了文本分析、參數預測 (聲學模型)、聲碼器分析/合成 (聲碼器)三部分。
  - 優：生成音檔更自然、更靈活方便修改參數、比串接式合成成本更低，不須要大量資料庫
  - 缺：生成的語音具有較低的理解性、很容易與人聲作區別，還是像機器人的聲音

現今主流TTS

在 [1] 中表示以下三種 TTS 在 MOS 評分上，Transformer 分數最高

- **Tansformer**
- Tacotron
- Fastspeech

[1] T. Hayashi, et al, ”ESPNET-TTS: unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” *arXiv preprint arXiv:1910.10909* , 2019

Transformer

- Transformer 的定義：第一個完全依賴self-attention去計算input、ouput之間關係並不使用序列對齊的RNN或CNN的transduction model。

(Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution)

- Attention 公式：公式內容為將輸入轉換成q、k、v三種向量，並利用q與k計算相似度，最後得到匹配的內容 q

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

- X-vector：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，目前認為paper引入x-vector作用為將不同長度的input轉換成固定長度的特徵向量，解決在test時輸入比training時長度還要長的時候效果差的問題

- 在 paper 中突破的事情：paper中突破了cascading式的轉換架構，確認了seq2seq建模的有效性，並展示這樣簡單的架構卻具有強大的競爭力

附錄

The evolution of neural TTS models.