Fine-grained Style Control In Transformer-based Text-to-speech Synthesis

arxiv | code | sample

FINE-GRAINED STYLE CONTROL IN TRANSFORMER-BASED TEXT-TO-SPEECH SYNTHESIS

Li-Wei Chen, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University
{liweiche, air}@cs.cmu.edu

Model

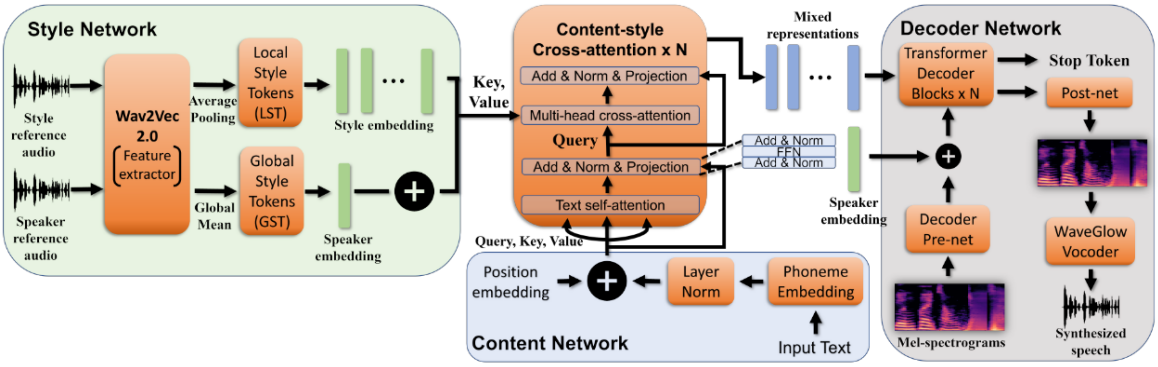

Fig. 1: Overview of our LST-TTS system.

基于TransformerTTS，增加了local style tokens (LST)，将TransformerTTS的content encoder替换为cross-attention block用于对content和style的对齐(alignment)以及融合(fusion)
花样设计Embedding和Attention

Style Network

Style Embedding

将Wav2Vec得到的结果经过LSTM以及pooling层，MHA没画出来，LSTM每一时间步（称为frame-level）的结果作为MHA的Q，和GST一样有可训练的K和V，得到结果作为Style Embedding

Speaker Embedding

GST这里的MHA也没画出来，将Wav2Vec得到的结果平均后作为Q，可训练的token作为K和V，得到结果作为Speaker embedding

Content-style cross-attention blocks

Text经过self-attention后做skip connections（其实就是残差），再经过Transformer结构（可堆叠多层），之后作为**Query**与Style Embedding的**Key**和Speaker Embedding的**Value**再做一次MHA，后进入TransformerTTS的Decoder

Training and inference

$$\mathbf{L}_{tts} = \| D\left(A\left(S\left(x_{sty}^{s}, x_{spk}^{s}\right), c\right), x_{sty}^{s}\right) - Mel\left(x_{sty}^{s}\right) \|_1$$

Training时，对Style Embedding取$\tilde{l}_{sty} \in [\alpha, l_{sty}]$，paper中$\alpha$为15，因为在inference阶段，$x_{sty}$会比合成的speech短很多，而且提供较少的reference speech信息有助于模型更多的从Text中获取信息，避免内容泄露（content-leakage）问题，这一点和之前组会汇报Meta-Stylespeech时，闫老师和蔡老师提出的挖掉一部分的vector再预测另一部分的想法是一样的，没想到真的可以这么做

Experiment

Dataset

1. Single speaker: **LJSpeech**
2. Multi speaker: **VCTK**
3. Emotional speech synthesis corpus: ESD database | paper，这篇paper使用了10个speaker每个5种情感共13小时

Metrics

1. WER
2. Emotional classification
3. MOS