快速并行语音合成

2021-04-12 周志洋

目录

CONTENTS

章节

part

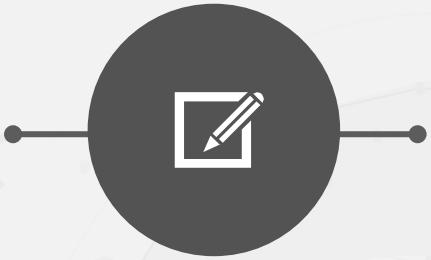
01 现存问题

现存问题

资源要求

对训练数据与算力资源有较高的要求

鲁棒性



逐帧合成语音导致容易出现跳过或重复某些单词的情况

合成速度

自回归模型不可避免地存在合成速度过慢的问题

可控性

几乎无法对合成语音的速度或韵律等属性进行人为控制

章节
part

02 解决思路

解决思路

四点改进

软对齐模型

通过软对齐模型学习<音素或字符序列, 声谱图> 对之间的注意力分布情况，进而获得相应地对齐关系

A

残差卷积块

在不改变编码器-解码器架构的前提下，仅使用简单的残差卷积块搭建编码器与解码器，从而加快训练/推断的速度

C

持续时间预测

通过引入持续时间预测与序列长度规范，将模型由自回归转化为非自回归，从而实现真正的并行化

D

超分辨率重建

通过引入视觉领域的超分辨率重建技术解决将高度压缩的源文本“解压”为语音时可能存在的信息不足的问题

S

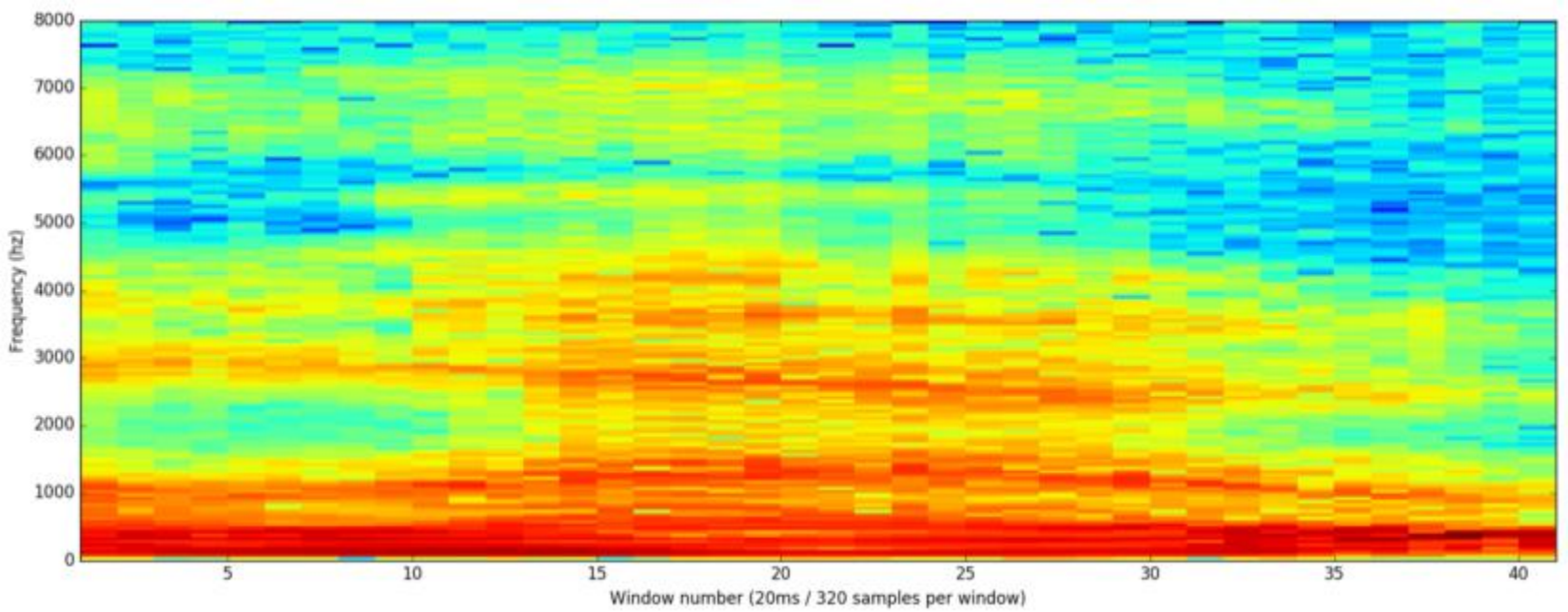
章节
part

03 方法介绍

方法介绍

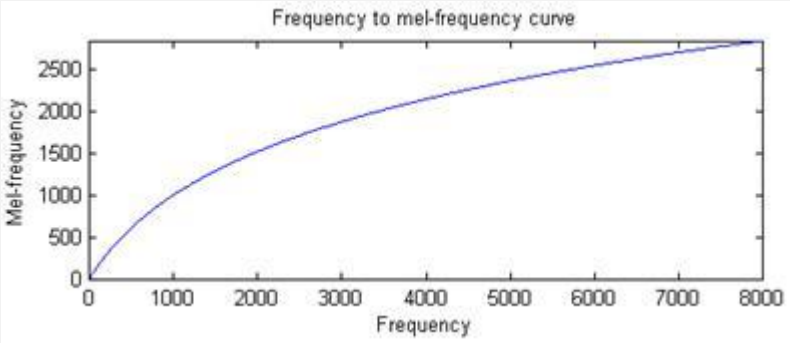
原始波形：从音频文件读出的原始语音信号称为原始波形，是一个一维数组，数组中值的大小通常表示振幅。

线性声谱图：对原始波形进行短时傅里叶变换可以得到线性声谱图。

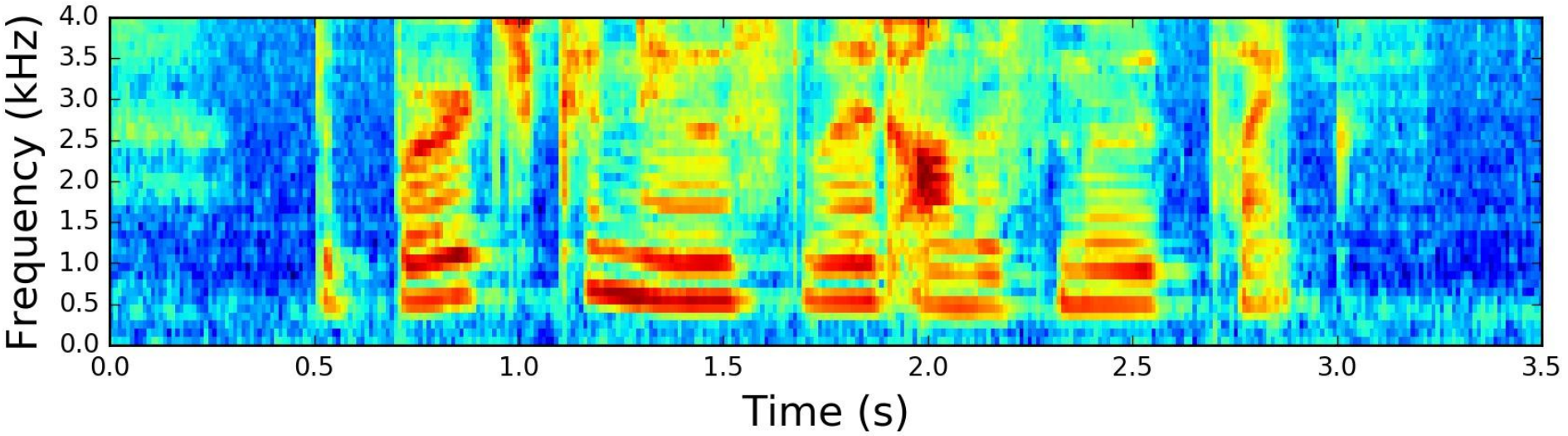


方法介绍

梅尔频率：人耳听到的声音高低和赫兹频率并不是线性关系，梅尔频率更加符合人耳的听觉特性。



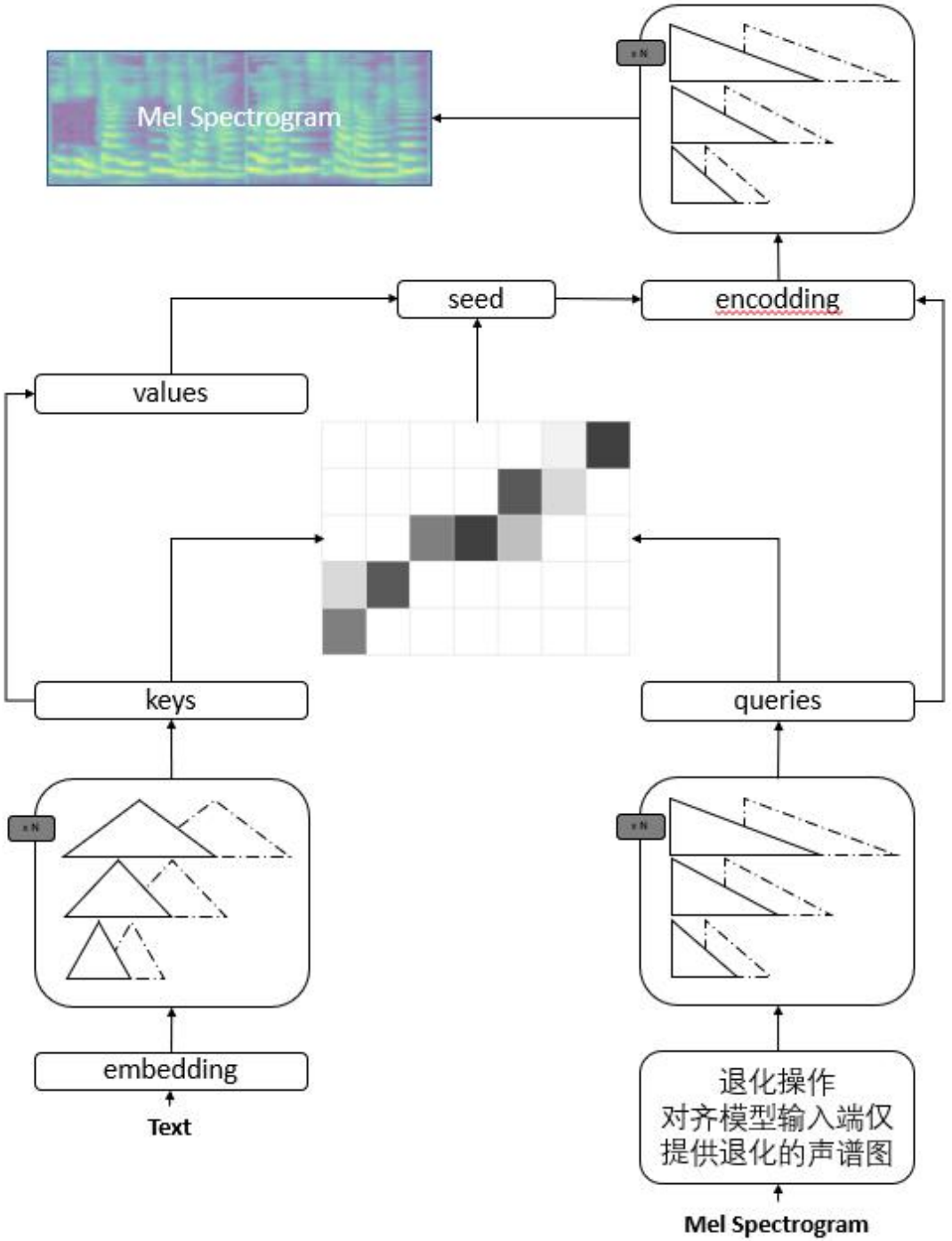
梅尔声谱图：对线性声谱图应用梅尔滤波器组可得梅尔声谱图。



方法介绍

软对齐模型:

- 文本编码器
- 声谱编码器
- 引导注意力
- 声谱解码器



方法介绍

软对齐模型各个模块的具体结构:

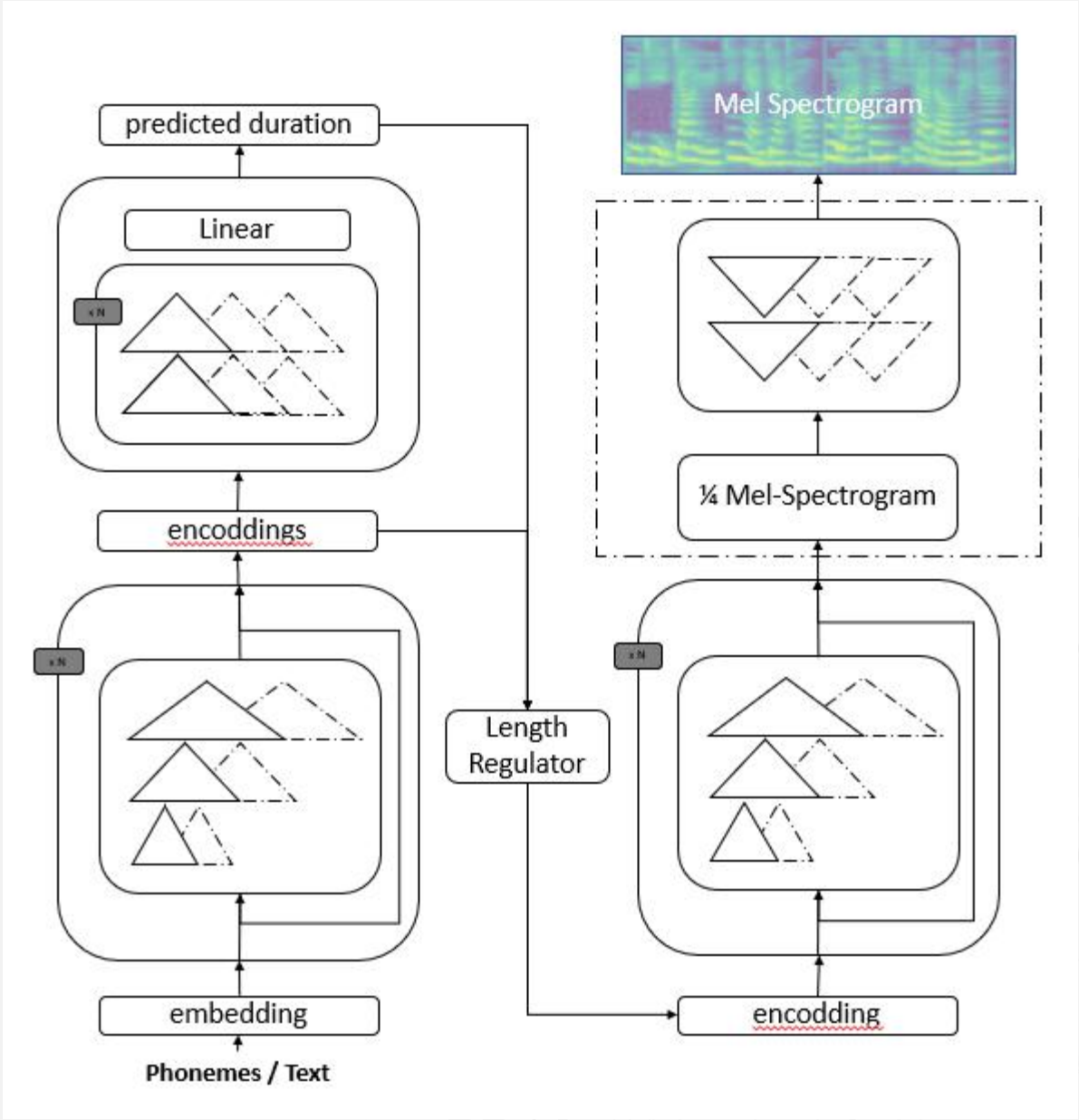
$$
\begin{aligned}
\mathrm{TextEnc}(\cdot) =& \mathrm{C}_{1\times1}^{2d\leftarrow2d} \triangleleft (\mathrm{RC}_{1\times1}^{2d\leftarrow2d})^2 \triangleleft (\mathrm{RC}_{3\times27}^{2d\leftarrow2d} \triangleleft \mathrm{RC}_{3\times9}^{2d\leftarrow2d} \triangleleft \\
& \mathrm{RC}_{3\times3}^{2d\leftarrow2d} \triangleleft \mathrm{RC}_{3\times1}^{2d\leftarrow2d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{2d\leftarrow e} \triangleleft \mathrm{E}^e(\cdot) \\
\mathrm{SpecEnc}(\cdot) =& \mathrm{C}_{1\times1}^{d\leftarrow d} \triangleleft (\mathrm{RC}_{1\times1}^{d\leftarrow d})^2 \triangleleft (\mathrm{RC}_{3\times27}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times9}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times3}^{d\leftarrow d} \triangleleft \\
& \mathrm{RC}_{3\times1}^{d\leftarrow d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow 2d}(\cdot) \\
\mathrm{SpecDec}(\cdot) =& \mathrm{C}_{1\times1}^{d\leftarrow d} \triangleleft (\mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow d})^3 \triangleleft (\mathrm{RC}_{1\times1}^{d\leftarrow d})^2 \triangleleft (\mathrm{RC}_{3\times27}^{d\leftarrow d} \triangleleft \\
& \mathrm{RC}_{3\times9}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times3}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times1}^{d\leftarrow d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow 2d}(\cdot)
\end{aligned}
$$

其中，E^e 表示嵌入维度为 e 的嵌入层，$\mathrm{C}_{k\times\delta}^{o\leftarrow i}$ 表示输入通道数为 i 输出通道数为 o 内核尺寸为 k 空洞因子为 δ 的空洞卷积层，$\mathrm{RC}_{k\times\delta}^{o\leftarrow i}$ 表示输入通道数为 i 输出通道数为 o 内核尺寸为 k 空洞因子为 δ 且带有残差连接的空洞卷积层。

方法介绍

声谱合成模型:

- 并行文本编码器
- 持续时间预测器
- 序列长度调整器
- 并行声谱解码器
- 超分辨率模型



方法介绍

声谱合成模型各个模块的具体结构:

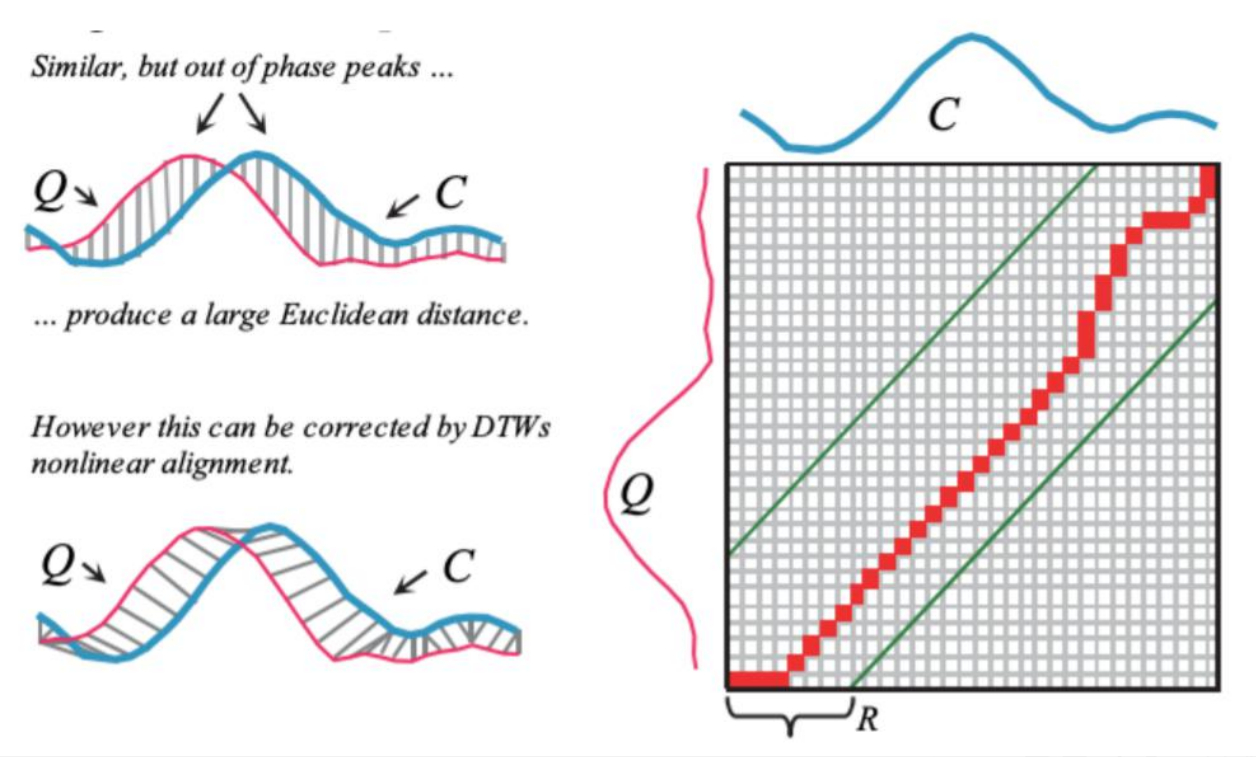
$$
\begin{aligned}
\mathrm{ParaTextEnc}(\cdot) =& \mathrm{C}_{1\times1}^{2d\leftarrow -2d} \triangleleft (\mathrm{RC}_{1\times1}^{2d\leftarrow -2d})^2 \triangleleft (\mathrm{RC}_{3\times27}^{2d\leftarrow 2d} \triangleleft \mathrm{RC}_{3\times9}^{2d\leftarrow -2d} \triangleleft \\
& \mathrm{RC}_{3\times3}^{2d\leftarrow 2d} \triangleleft \mathrm{RC}_{3\times1}^{2d\leftarrow 2d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{2d\leftarrow e} \triangleleft \mathrm{E}^{e}(\cdot) \\
\mathrm{DuraPre}(\cdot) =& \mathrm{C}_{1\times1}^{1\leftarrow d} \triangleleft \mathrm{RC}_{1\times1}^{d\leftarrow d} \triangleleft (\mathrm{RC}_{3\times1}^{d\leftarrow d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow 2d}(\cdot) \\
\mathrm{ParaSpecDec}(\cdot) =& \mathrm{C}_{1\times1}^{d\leftarrow -d} \triangleleft (\mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow -d})^3 \triangleleft (\mathrm{RC}_{1\times1}^{d\leftarrow d})^2 \triangleleft (\mathrm{RC}_{3\times27}^{d\leftarrow d} \triangleleft \\
& \mathrm{RC}_{3\times9}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times3}^{d\leftarrow d} \triangleleft \mathrm{RC}_{3\times1}^{d\leftarrow d})^2 \triangleleft \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{d\leftarrow 2d}(\cdot) \\
\mathrm{SpecSRN}(\cdot) =& \mathrm{C}_{1\times1}^{F\leftarrow c} \triangleleft (\mathrm{RC}_{1\times1}^{c\leftarrow c})^2 \triangleleft (\mathrm{RC}_{3\times3}^{c\leftarrow c} \triangleleft \mathrm{RC}_{3\times1}^{c\leftarrow c} \triangleleft \mathrm{DC}_{2\times1}^{c\times c})^2 \triangleleft \\
& \mathrm{ReLU} \triangleleft \mathrm{C}_{1\times1}^{c\leftarrow F}(\cdot)
\end{aligned}
$$

其中，$\mathrm{C}_{k\times\delta}^{o\leftarrow i}$ 表示输入通道数为 i 输出通道数为 o 内核尺寸为 k 空洞因子为 δ 的空洞卷积层，$\mathrm{RC}_{k\times\delta}^{o\leftarrow i}$ 表示输入通道数为 i 输出通道数为 o 内核尺寸为 k 空洞因子为 δ 且带有残差连接的空洞卷积层，$\mathrm{DC}_{k\times\delta}^{o\leftarrow i}$ 表示输入通道数为 i 输出通道数为 o 内核尺寸为 k 空洞因子为 δ 的空洞反卷积层。

方法介绍

损失函数：软动态时间规整（Soft Dynamic Time Warping，Soft-DTW），其核心在于寻找两个时间序列之间最好的对齐方式。

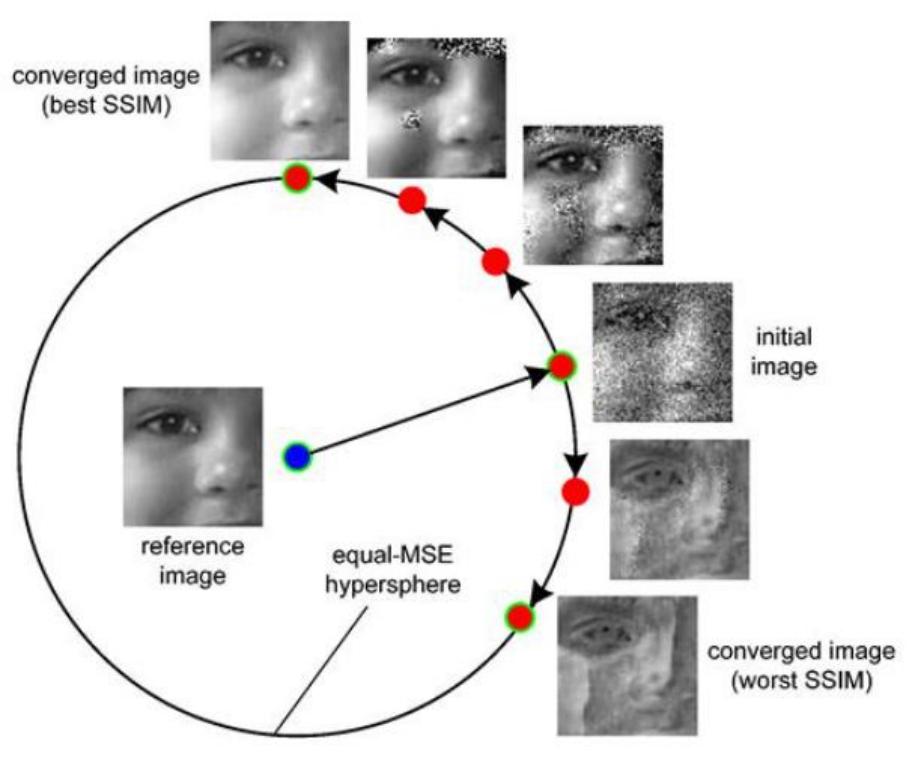
适用于语音这类并不需要完全严格对齐的序列数据。

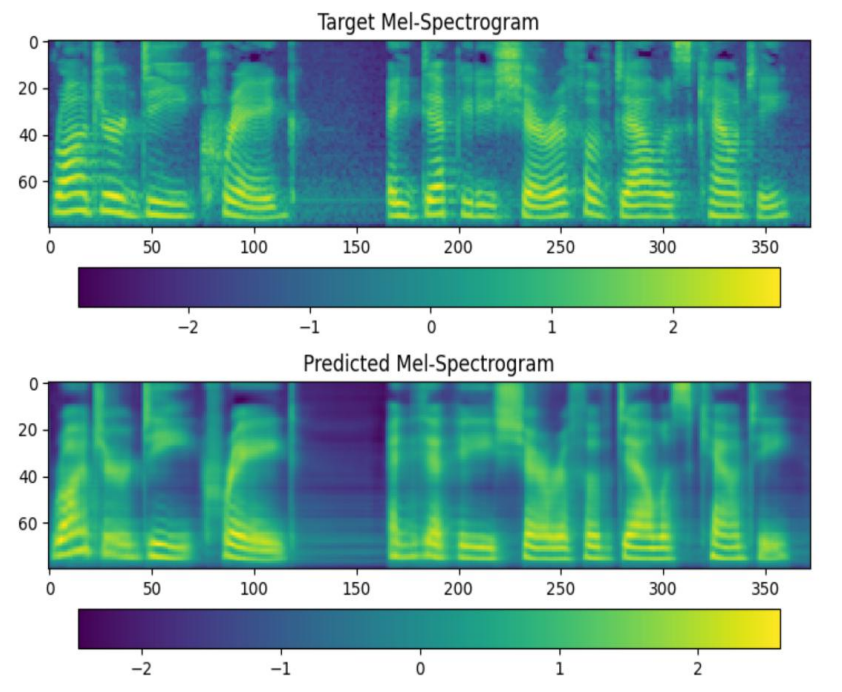


方法介绍

损失函数：结构相似度指数（Structural Similarity Index，SSIM），其计算基于 亮度、对比度与结构 三个维度。

原用于图像领域，应用于声谱图可以使模型更关注整体能量分布而非局部细节。





方法介绍

使用到的几个数据增强方式：

- 随机噪声，即随机地往输入梅尔声谱图中添加少量高斯噪声；
- 随机掩蔽，即随机地将输入梅尔声谱图的若干帧替换成空白帧；
- 随机交换，即随机地将输入梅尔声谱图的若干帧交换位置；
- 随机替换，即随机地将输入梅尔声谱图的若干帧替换成当前模型预测帧；
- 通过 Griffin-Lim 算法将真实声谱图重建为质量较低的语音波形，再重新提取声谱图（退化声谱图）。

Note：上述数据增强方式仅用于软对齐模型的训练过程。

章节
part

04 实验分析

实验分析

评分方式：将 LJSpeech 数据集的 13100 条数据分割为 13000 + 100 两部分，前者用于训练模型，后者保留，用于作为评分时的参考音频；评分调查在 5 名志愿者之中进行，评分等级 0-100 分，取分数均值。

质量评分：在 LJSpeech 数据集上， ParallelTTS 模型的合成质量显著高于 Tacotron 2 模型，当使用 MelGAN 作为声码器时，提升更为明显。

Model	Score	Interval
Tacotron 2 + G&L	42.80	(38, 47)
Tacotron 2 + MelGAN	62.60	(59, 66)
ParallelTTS + G&L	47.80	(43, 51)
ParallelTTS + MelGAN	74.40	(67, 79)

实验分析

训练时间：对于 LJSpeech 数据集，设置批次尺寸为 64，可以在单张 8GB 显存的 GTX 1080 显卡上进行训练。训练约 8 小时（~300 Epochs）即可合成质量较高的语音。

合成时间：以下测试在 Intel Core i7-8550U / NVIDIA GeForce MX150 下进行，每段合成音频在 8.5 秒左右（约 20 词）。

Batch Size	Spec (GPU)	Audio (GPU)	Spec (CPU)	Audio (CPU)
1	0.042	0.218	0.100	2.004
2	0.046	0.453	0.209	3.922
4	0.053	0.863	0.407	7.897
8	0.062	2.386	0.878	14.599

章节
part

05 未来工作

未来工作

接下来还可以继续进行的工作:

- 在更多数据规模较小的小语种上测试模型的有效性
- 语音风格的迁移（进行中）、语音情绪的变化
- …… ……

THANKS

2021-04-12 周志洋