计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):236−243 [doi: 10.15888/j.cnki.csa.009875] [CSTR: 32024.14.csa.009875]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于注意力机制的端到端语音合成模型①

耿　盈, 朱欣娟

(西安工程大学 计算机科学学院, 西安 710600)
通信作者: 朱欣娟, E-mail: zhuxinjuan@xpu.edu.cn

摘　要: 随着语音合成应用场景不断扩展, 对多人多情感语音合成的需求越来越大. 在实际应用中经常需要合成具有特定风格的语音信号. 为此提出一种基于注意力机制的端到端语音合成模型. 首先设计了说话人编码模块, 通过注意力机制提取语音信号中说话者的特征表示, 结合数据集中性别、年龄等特征标签构建说话人特征库; 其次设计风格编码模块, 通过注意力机制为不同梅尔特征通道赋予不同关注程度和权重; 然后使用 K 近邻构建虚拟说话人特征, 实现在不提供说话人及真实数据的情境下, 灵活组合不同说话人特征和风格特征, 搭配合成出具有特定特征风格的声音. 实验结果表明, 该模型对比 SV2TTS 模型有较快的训练速度, 能够合成具有特定风格的高质量的语音.

关键词: 语音合成; 说话人编码器; 语音风格; 注意力机制

引用格式: 耿盈,朱欣娟.基于注意力机制的端到端语音合成模型.计算机系统应用,2025,34(7):236–243. http://www.c-s-a.org.cn/1003-3254/9875.html

End-to-end Speech Synthesis Model Based on Attention Mechanism

GENG Ying, ZHU Xin-Juan

(College of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: With the continuous expansion of speech synthesis application scenarios, the demand for multi-speaker and multi-emotion speech synthesis is increasing. In practical applications, there is often a need to synthesize speech signals with specific styles. To address this, an end-to-end speech synthesis model based on the attention mechanism is proposed. First, a speaker encoding module is designed to extract speaker feature representations from speech signals using the attention mechanism, combined with dataset features such as gender and age labels to construct a speaker feature database. Second, a style encoding module is designed to assign different levels of attention and weights to different Mel feature channels using the attention mechanism. Then, virtual speaker features are constructed using K-nearest neighbors, allowing for the flexible combination of different speaker and style features to synthesize voice with specific characteristic styles, even without requiring real speaker data. Experimental results show that this model has a faster training speed compared to the SV2TTS model and can synthesize high-quality speech with specific styles.

Key words: speech synthesis; speaker encoder; speech style; attention mechanism

语音合成, 又称文语转换, 是服务于语音交互、信息播报、有声朗读等任务的核心技术[1]. 基于神经网络的语音合成方法已经成为当前最流行的方法之一, 使用深度神经网络来学习输入语音与目标风格之间的映射关系, 目前在语音合成和语音转换等任务中取得了很好的效果[2–4].

Tacotron2[5]是一种经典的端到端语音合成模型, 首先输入文本序列, 使用 LSTM 对整个文本序列进行

① 基金项目: 陕西省重点研发计划 (2024GX-YBXM-548)
收稿时间: 2024-10-23; 修改时间: 2024-11-29; 采用时间: 2025-01-07; csa 在线出版时间: 2025-05-29
CNKI 网络首发时间: 2025-05-29

逐步处理, 获得包含上下文信息的文本表示. 同时也需要对输入文本对应的音频文件进行预处理得到参考音频的梅尔谱及声学特征, 使用位置敏感注意力将输入文本序列与输出声学特征之间进行关联, 提高语音合成的准确性和流畅度. 最后阶段使用 WaveNet[6]神经网络模型来将声学特征转换为高质量、逼真的语音波形. 但是传统的端到端的语音合成模型在多人多风格的应用场景中适应性较差.

为实现更富有情感风格变化的语音合成, 使用变分自编码器[7–9]建模语音中的表达特征, 变分自编码器能够有效捕捉和重建各种表达, 使生成的语音更具自然性和表现力. 语音合成中使用层次化 Conformer 结构[10,11], 在编码阶段将输入的文本和语音信息通过层次化网络进行处理, 生成多层次的表达特征, 使得模型能够更好地捕捉并合成多样的说话风格. 传统的语音合成模型通常难以处理精细韵律和层次化特征, 导致生成语音的质量受限. Sun 等人[12]提出了完全层次化的韵律建模方法, 通过多个层次来捕捉不同层面的韵律信息, 建模韵律特征, 并通过层次化网络来融合这些信息, 模型不仅关注宏观的韵律特征, 还深入微观细节, 如单个音节的韵律调整, 提高模型的可解释性和生成语音的质量. 在语音合成中, 保持目标说话人的特有风格同时迁移语音特征是一个具有挑战性的任务[13,14]. Pan 等人[15]引入“prosody bottleneck”在神经语音合成中实现跨说话人风格迁移, 引入一个瓶颈层来专门编码和解码韵律特征来实现压缩和抽象韵律信息, 避免了传统模型中韵律特征的过度复杂化, 尽管模型在风格迁移上表现优异, 但对于复杂的瓶颈机制和风格迁移过程的具体操作, 模型的解释性较低. 低延迟和高质量的语音生成, 在实时语音合成应用中也是关键要求, 为了减少延迟, 提出了一种神经增量 TTS 方法[16,17], 在生成当前音频片段的同时, 开始生成下一个片段. Ren 等人[18]提出了一种基于 Transformer 架构的非自回归语音合成模型“FastSpeech”. 非自回归的 TTS 模型[19–21]能够并行生成语音, Transformer[22,23]的自注意力机制能并行处理整个输入序列, 传统的 RNN 或 LSTM 通常依赖于逐步处理, 因此速度较慢, 并行处理能够显著提高文本到语音的生成速度. Ellinas 等人[24]提出了一种流式语音合成架构, 其模型设计更关注实时性, 通过减少处理时间来实现低延迟, 保持延迟与句子长度无关, 无论输入句子多长, 语音合成的延迟都保持稳定. 为了解决 FastSpeech 在韵律建模和音质上的不足, Ren 等人[25]在韵律建模和音素对齐技术方面进行了改进, 提出了“FastSpeech 2”模型, 在提升语音合成质量和自然性方面做出了显著进步.

随着神经网络模块设计的不断复杂化, 以及其层化处理结构的引入, 模型的计算复杂性显著增加, 这对系统的计算能力和效率提出了更高的要求. 而大多数语音合成模型是使用具有表达力的、单一风格的语料库进行训练的, 为了训练具有良好神经表达能力的语音合成模型, 需要采集大量具有多种风格的语音数据. 发音人完成不同风格的录制既昂贵又耗时, 并且当前语音风格迁移技术仍然局限在训练数据中存在的风格和说话人[26]. 随着语音合成应用场景地不断扩展, 对多人多情感语音合成的需求越来越大. 然而, 也必须注意到由于语音合成可以模拟个人的声音, 可能会面临非法身份验证、诈骗或未经用户授权的私人数据泄露的风险[27,28].

为了降低对高质量训练数据的需求, 降低模型的复杂程度, 本文提出基于注意力机制的端到端语音合成模型. 本文的主要工作与创新包括: (1) 设计说话人编码模块, 对单一风格的语料库使用注意力机制进行说话人特征提取, 结合数据集中的性别、年龄等特征标签构建说话人特征库; (2) 设计风格编码模块对高质量、具有表现力的数据使用注意力机制进行说话人风格提取, 作为语音合成中的风格条件参与语音合成; (3) 为了实现虚拟说话人的语音合成, 在提取的说话人特征库中, 选择多个样本生成一个新的样本, 使用 K 近邻筛选一定数量的真实样本, 微调新样本的参数. 实现在不提供说话人及真实数据的情境下, 灵活组合不同说话人特征和风格特征, 搭配合成出具有特定特征风格的声音.

1 语音合成模型

本文的语音合成模型基于编码器-解码器结构, 使用 Tacotron2 作为基础模型, 其整体结构如图 1 所示.

预训练阶段的目标是通过大规模语音数据的学习, 使得模型能够理解并生成具有目标特征的语音. 说话人的语音信号经过编码转化为一个固定长度的向量, 这个向量包含了说话人独特的声音特征和一定语音风格信息, 图 1 中展示了设计的一个语音风格编码模块和一个说话人编码模块, 分别完成说话人特征与说话

人的语音信息解耦、说话人语音风格与语音信息解耦. 这两个模块都使用残差注意力机制[29], 该机制能够动态调整音频特征的权重, 并且由于各模块任务的差异, 注意力机制在通道间和空间位置上的重要性会有所不同, 从而使得网络能够专注于任务相关的特征. 模块的独立性使得其能分别训练, 从而有效学习具有较高区分性的风格和说话人嵌入, 进一步提高特征提取和训练的效率及准确性.

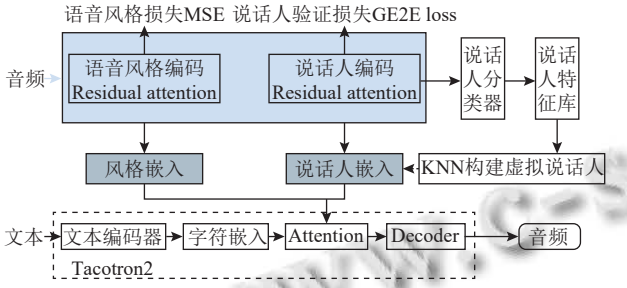

图 1 基于注意力机制的语音合成模型框架

推理阶段的目标是利用经过训练的模型生成自然且具有目标说话人特征的语音. 在这一阶段, 模型接收输入文本特征和目标说话人的音频特征以及风格特征, 经过 Tacotron2 的解码器后生成对应的梅尔频谱图. 在目标说话人语音合成任务中, 增加使用构建虚拟说话人特征代替目标音频, 并结合采用残差注意力机制训练的说话人编码网络, 能够更有效地捕捉目标说话人的语音特征, 同时提升模型对未见过的说话人的泛化能力, 从而实现高质量的语音合成.

1.1 说话人编码模块

编码器包含两个核心部分, 说话人编码模块和语音风格编码模块. 说话人编码模块从语音信号中提取与语音内容和背景噪声无关的说话人特征. 首先, 对任意长度的音频文件处理得到梅尔频谱图, 并将梅尔频谱图切分成长度相同的若干数量梅尔频谱帧. 在说话人编码模块中加入注意力机制, 借助层级结构有效地提取梅尔频谱帧中说话人语音相关的特征, 经过注意力机制调整后, 这些特征与原始特征相加. 然后梅尔频谱帧被传递到一个由 3 层 256 个单元的 LSTM 堆叠组成的网络中, 每层后面跟随一个 256 维的投影层. 对最后一帧的顶层输出进行 L2 归一化处理, 得到一个 256 维度的嵌入向量, 作为说话人的特征.

说话人编码模块的训练目标是优化说话人验证损失, 使得来自同一说话人的特征嵌入具有高余弦相似度, 而来自不同说话人的特征嵌入在嵌入空间中相距较远[30]. 训练过程如图 2 所示, 编码器提取说话人特征传递给 GE2E[31], GE2E 对每个特征计算其与同类样本与其他类样本的相似度. 损失梯度通过反向传播到编码器输出层, 通过链式法则在编码器中传递, 调整网络和注意力权重, 优化说话人的区分性. 假设有 N 个说话人, 每个说话人有 M 条语音, 用 x_{ij} 表示一条语音片段的向量, 通过注意力神经网络的特征提取结果为 $f(x_{ij};\omega)$, 对该结果做一次 L2 归一化. 对于每个说话人, 将其 M 条语音的特征向量进行平均, 得到说话人的特征向量 c_k, 同时避免在计算 loss 时遇到目标本身, 特征向量计算公式为:

$$c_k = \frac{1}{N-1}\sum_{\substack{i=1\\ i\neq j}}^{N} x_{ki} \tag{1}$$

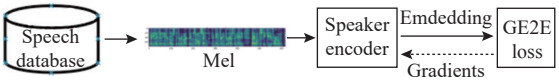

图 2 说话人编码模块训练

用 Softmax 函数来优化模型, 提高正确类别的得分同时对所有类别进行归一化, 帮助模型在训练过程中进行有效地学习以便更好地区分不同的说话人. 计算一个相似矩阵 $S_{ji,k}$, 表示说话人 k 对样本 j 的得分, 得到 Softmax 损失函数, 其表达式为:

$$L(x_{ji}) = -S_{ji,j} + \log\sum_{k=1}^{N}\exp(S_{ji,k}) \tag{2}$$

为确保同一说话人的嵌入向量彼此接近, 不同说话人的嵌入向量彼此远离, 使用 Contrastive 损失函数. 属于同一说话人的嵌入向量 e_{ji} 的输出分数 $\sigma(S_{ji,j})$ 尽可能高, 不属于同一说话人的输出分数尽可能低, 损失函数的表达式如式 (3):

$$L(e_{ji}) = 1 - \sigma(S_{ji,j}) + \max_{\substack{1\leqslant k<N\\ k\neq j}}\sigma(S_{ji,j}) \tag{3}$$

1.2 语音风格编码模块

语音风格编码模块使用富有风格信息的高质量音频数据进行风格训练, 完成语音风格特征提取. 为获得更高质量和更多细节, 使用 80 通道对数梅尔频谱图作为输入. 在图 3 语音风格编码模块设计中加入注意力机制, 首先将梅尔频谱图经过一层卷积处理得到特征图 p, 分别经过主干分支和掩码分支. 主干分支通过全

局注意力获得特征图 t 的全局信息, 使用全局平均池化计算得到. 掩码分支用来增强重要特征并抑制不重要特征, 通过最大池化降低特征图的空间维度以提取更深层次特征, 再通过线性插值将特征图恢复至原始大小.

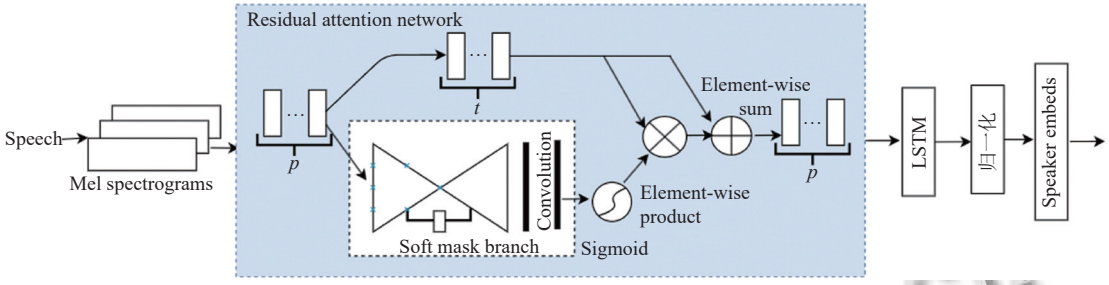

图 3 语音风格编码模块

使用 Sigmoid 层将输出归一化到[0, 1]的范围, 在掩码分支生成一个掩码 m:

$$m = \sigma(W_m \times p + b_m) \tag{4}$$

其中, σ 是 Sigmoid 激活函数, W_m 是权重, b_m 是偏置, 每个值在 0–1 之间, 0 表示该像素不重要或被忽略, 1 表示该像素重要并应被考虑. 将掩码 m 与特征图 p 进行逐元素相乘, 以增强重要特征. 注意力模块的输出 H 可以表示为:

$$H = \sum_i \sum_c M(x_i)T(x_i) \tag{5}$$

其中, i 遍历所有空间位置, $c \in \{1, \cdots, C\}$是通道的索引. 再将注意力模块的输出的结果传递到 LSTM 层堆叠组成的网络中, LSTM 可以有效地捕捉长期依赖关系处理序列数据, 再将注意力模块的输出结果传递到 LSTM 层堆叠组成的网络中. 通过对比生成语音的特征F_i与目标风格特征G_i之间的差异, 计算说话人风格损失, 进而优化风格编码的效果. 说话人风格损失可表示为式(6), 其中, *Gram* 表示特征图矩阵.

$$L(x_i) = \sum_i \|Gram(F_i) - Gram(G_i)\|^2 \tag{6}$$

1.3 虚拟说话人

说话人编码模块关注计算并提取说话人特征生成说话人嵌入, 通过优化训练集中每个说话人的嵌入向量, 构建一个通用的嵌入表示, 从而为每个说话人创建一个独特的特征表示, 能够唯一标识该说话者. 将说话人编码模块提取的特征构建说话人特征库, 并根据数据集中提供的说话人性别和年龄分类, 每个说话人选择多个语音的平均值作为一个说话的人的特征. 首先, 在已分类的说话人特征中, 选择多个样本生成一个新的说话人特征, 作为虚拟说话人的初始特征. 再使用 K 近邻从真实说话人特征中筛选一定数量的样本, 来微调虚拟说话人初始特征的参数. 实现在不提供说话人及真实数据的情况下, 灵活组合不同说话人特征和风格特征, 合成出具有特定特征风格的声音, 同时在语音合成的应用中避免泄露用户的私人数据.

1.4 解码器

与 Tacotron2 解码器的原理相同, 文本编码器和注意力机制协同工作获得上下文信息的文本表示, 并动态地调整输入文本序列与输出声学特征之间的关联, 生成预测的梅尔频谱帧. 通过改进位置敏感的注意力机制使得模型可以接收说话人嵌入和语音风格嵌入控制信息的输入, 位置敏感的注意力机制也有助于将文本与生成的语音风格特征对齐. 预训练的说话人嵌入被连接到每个解码器时间步的隐层表示中, 使得生成的梅尔频谱与目标说话人的特征相匹配. 每一步预测梅尔频谱帧先经过 2 层 LSTM 处理序列任务, 再经过一层线性投影映射, 作为解码器的输出, 同时作为解码器的输入. 作为解码器的输入的预测梅尔频谱帧会经过一个预处理网络, 再结合注意力的输出传递给 2 层 LSTM.

2 实验

2.1 数据集

本文选取了 LibriTTS[32]多说话人数据集和游戏音频, LibriSpeech 数据集是从 LibriVox 项目中提取的有声书音频, 包含来自 1 172 位说话者的 436 h 语音的大量的英语语音样本, 采样率为 16 kHz. LibriSpeech 数据集被划分为训练集、验证集和测试集, 说话者集合在训练、验证和测试集之间完全不重叠, 这种分割方式

有助于评估模型的泛化能力. 游戏音频从游戏官方网站提供的语音对话中构建数据集.

LibriSpeech 数据集和游戏音频的语音质量相对较高, 但许多录音包含明显的环境和静态背景噪声, 实验中使用简单的谱减法去噪程序对目标谱图进行了预处理. 数据集中的转录文本没有标点符号, 通过强对齐音频与转录文本, 使用自动语音识别 (automatic speech recognition, ASR)[33]模型将数据重新分段为更短的语句, 并在静音处进行分割, 将中位持续时间从 14 s 减少至 5 s. 游戏音频进行预处理去除背景噪音, 使用 ASR 同样将音频数据重新分段为更短的语句, 并在静音处进行分割.

2.2 实验设置

实验使用显卡为 NVIDIA GeForce RTX 3090, 操作系统为 Ubuntu 18.04, 深度学习框架为 PyTorch, 训练过程包括对说话人编码网络的说话人验证训练、语音风格编码模块风格重建训练、结合虚拟说话人特征的语音合成训练. 将音频的采样率设为 16 000 Hz, 并修剪静音的片段. 采用 Hann 窗函数, 帧长为 50 ms, 帧移为 12.5 ms, Mel 滤波器的数量为 80, 预加重滤波器系数为 0.97.

使用 Tacotron2 作为基础模型并将训练集划分为预训练数据部分和微调数据部分, 分别用于预训练阶段和微调阶段. 在预训练阶段, 选择 LibriSpeech 验证集和构建的虚拟说话人作为未见说话人, 训练集作为可见说话人数据集. 在微调阶段, 冻结了在大数据集上预训练好模型参数, 使其作为一个特征提取器从输入音频文件与文本中提取音频特征和上下文信息, 从 LibriSpeech 验证集随机选择 20 个句子, 构成微调数据集来微调模型. 模型采用的批量大小为 64, 训练 150k 步, decoder 每一步输出 2 个预测梅尔帧, Dropout 比率 0.5, 学习率初始化为 0.001, 模型评估的步数间隔为 500 验证编码器与解码器步骤之间的对齐结果. 使用 Adam 作为优化器, 优化器参数为:

$$\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1\mathrm{E}-8$$

2.3 评估

本文采用主观评价和客观评价两种评价方式. 评估语音合成的主观听力测试在两个维度上的表现: 合成语音的自然性和与目标说话者真实语音以及风格方面的相似性. 自然性使用平均意见评分 (MOS) 评估, 相似性使用相似度平均意见评分 (SMOS), 评分范围为 1–5, 1 表示差, 5 表示优秀, 增量为 0.5, 结果如表 1 所示. 假定有 N 个评分人, 有 M 条语音片段, 表示某个评分人对一条语音片段的评分, 均值公式为:

$$x = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij} \tag{7}$$

表 1 语音质量主观评价

System	可见说话人		未见说话人	
	MOS	SMOS	MOS	SMOS
Ground truth	4.52	—	4.52	—
Tacotron2	4.15	3.91	3.78	3.34
SV2TTS	4.23	3.97	3.85	3.36
本文方法	4.28	4.01	3.92	3.42

本文评估不同模型在训练 200 轮和 1 000 轮生成的 10 个句子 MOS 和 SMOS 得分, 排除其他干扰因素, 统一输入具有不同长度的 10 个文本内容, 结果见表 2.

表 2 训练过程语音质量主观评价

Epoch	Method	MOS	SMOS
—	Ground truth	4.52	—
200	Tacotron2	2.56	1.12
	SV2TTS	2.57	1.19
	本文方法	2.61	1.23
1 000	Tacotron2	3.16	1.17
	SV2TTS	3.33	1.24
	本文方法	3.41	1.32

本文客观评价通过对比 SV2TTS (speaker verification to multispeaker text-to-speech synthesis)[34]模型. 使用目标语音 A 与生成语音 B 的说话人特征向量余弦相似度, 作为说话人相似度评价指标, 计算公式为:

$$\cos(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i \cdot B_i}{\|A_i\|\|B_i\|} \tag{8}$$

其中, n 的含义是一条目标语音对应一条生成语音, 有 n 组; A_i 表示目标语音说话人特征, B_i 表示生成语音说话人特征; 余弦值越接近 1, 表示两个向量夹角越趋近 0°, 向量越相似; 余弦值越接近 0, 则表示夹角越趋近 90°, 向量越不相似. 通过计算原始语音与生成语音在基频和时长上的均方根误差, 作为语音合成质量评价指标, 均方根误差公式:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{9}$$

其中, n 为目标语音与生成语音的组数, y_i 代表生成语音, x_i 代表目标语音结果. 语音质量客观评价结果如表 3 所示. 由式 (9) 可知得分越低, 生成语音和目标语音越接近, 语音合成的效果越好.

表 3 语音质量客观评价

System	相似度	无风格		有风格	
		基频 (Hz)	时长 (s)	基频 (Hz)	时长 (s)
Tacotron2	0.799	28.2	0.362	84.0	0.921
SV2TTS	0.837	26.2	0.308	78.5	0.763
本文方法	0.845	22.3	0.256	44.8	0.472

为了观测说话人嵌入的相似性和分布关系, 使用 UMAP 算法对说话人嵌入向量进行处理, 得到一个二维空间投影, 其中横纵坐标表示点在新特征空间中的位置. 我们使用不同颜色代表不同的说话人, 随机选取 10 个说话人的若干条语音样本, 每个语音样本被标识为一个带有颜色的点. 通过这种方式实现了说话人嵌入的相似性和分布关系的可视化 (见图 4). 可以观察到相同说话人的语音样本在空间中聚集在一起, 不同说话人的语音样本则分布在空间的不同区域, 彼此之间保持较远的距离, 进一步证明了说话人分类训练的有效性. 同样, 通过 UMAP 可视化构建的虚拟说话人特征见图 5, 其中, Average_data 为所构建的虚拟特征, 6407_mean.npy 为数据集中对编号 6407 的说话人提取的特征文件. 将使用 K 近邻筛选出的说话人特征标识为一个带有颜色的点, 可以观察出所构建的虚拟特征是区别真实说话人特征的.

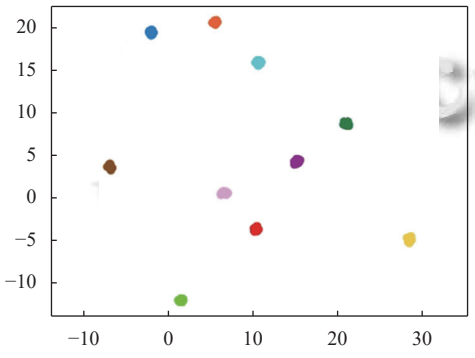

图 4 说话人特征聚类

在说话人验证任务中, 希望在一个合适的阈值下同时最小化假正例率和假负例率, EER 可以衡量假正例率和假负例率的平衡. 因此使用 EER 作为评估说话人验证性能的重要指标, 在保证系统较高准确性的同时, 避免过多的误判和漏判, EER 越低表示模型的性能越好. 说话人编码模块训练结果的对比见图 6, 可以观察出在进行说话人验证任务时本文模型收敛速度快于 SV2TTS. 本文模型训练损失对比结果图 7, 结合实验结果可以得出在引入残差注意力模块的迭代训练时模型合成的音频效果优于 Tacotron2 和 SV2TTS 模型.

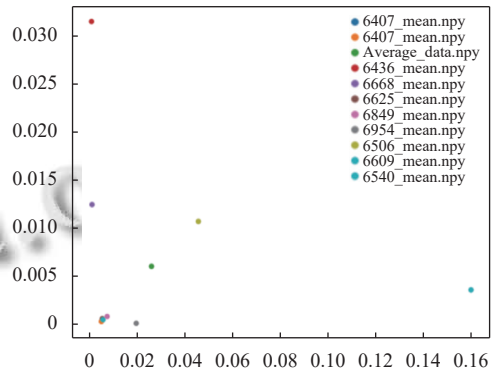

图 5 虚拟说话人特征空间投影

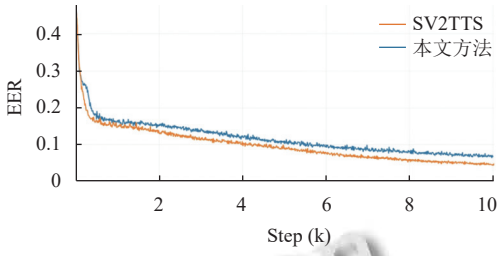

图 6 说话人验证训练

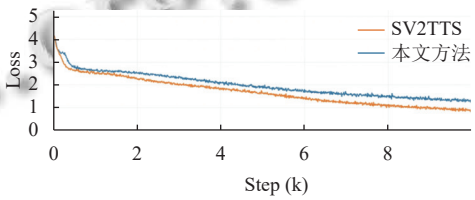

图 7 模型训练损失

人工模型评估成本较高, 使用编码器步骤与解码器步骤之间的对齐结果对模型进行评估如图 8, 线段连续表明模型可以处理输入与输出之间的关系, 并在生成过程中保持了一定的连贯性和一致性. 不同颜色代表对齐状态, 黄色代表需要重点关注的部分, 图 8 表明模型能够有效地捕捉到关键特征. 梅尔声谱图纵坐标是梅尔频率轴, 横坐标是时间轴, 通过颜色强度来表示每个时间点上的梅尔频率分布能量的大小, 颜色强度越亮 (如黄色), 表示该频率在某一时刻的能量越强, 颜色越暗 (如蓝色), 表示能量较弱. 目标声谱图 (target

Mel-spectrogram) 通过真实的音频信号即原始语音生成的梅尔频谱图, 预测声谱图 (predicted Mel-spectrogram) 是模型输出的梅尔声谱图, 目标声谱图与预测声谱图的对比见图 9. 预测声谱图通过模型训练生成, 整体结构清晰, 在保持重要细节同时, 整体看起来更加紧凑; 在时域和频域的过渡上较为平滑, 有助于合成语音更加自然且连续.

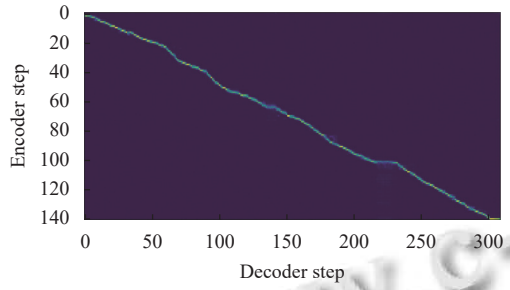

图 8 对齐关系

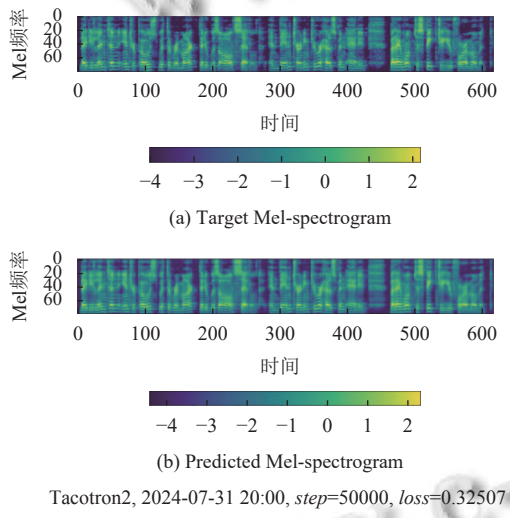

图 9 真实声谱图与预测声谱图

3 结论与展望

本文设计了一个基于注意力机制的语音合成模型. 首先, 设计了独立的说话人编码模块和语音风格编码模块, 使用残差注意力模块让两个模块专注不同的任务, 在说话人模块区分不同说话人特征, 在语音风格模块完成风格提取. 然后对 Tacotron2 的注意力进行修改, 使得该模型能够接收说话人特征的嵌入和风格信息的控制, 并使用 K 近邻方法构建新的说话人特征, 可以在语音合成的应用中避免涉及未经授权的用户私人数据. 对比 Tacotron2 和 SV2TTS, 本文模型有较快的训练速度, 在 LibriTTS 数据集上的实验说明本文方法能够合成高质量的语音. 在未来的工作中需要考虑: 改进当前模型, 使其可以更好地契合中文语音合成; 优化编码器结构, 提升梅尔频谱图到高保真音频波形的合成能力; 构建更高质量虚拟说话人特征.

参考文献

1 豆子闻, 李文书. 基于神经网络和虚幻引擎的数字人客服系统. 软件工程, 2023, 26(10): 49–52. [doi: 10.19644/j.cnki.issn2096-1472.2023.010.011]

2 Tobing PL, Wu YC, Hayashi T, *et al*. Voice conversion with CycleRNN-based spectral mapping and finely tuned WaveNet vocoder. IEEE Access, 2019, 7: 171114–171125. [doi: 10.1109/ACCESS.2019.2955978]

3 An XC, Soong FK, Xie L. Disentangling style and speaker attributes for TTS style transfer. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 646–658. [doi: 10.1109/TASLP.2022.3145297]

4 Choi BJ, Jeong M, Lee JY, *et al*. SNAC: Speaker-normalized affine coupling layer in flow-based architecture for zero-shot multi-speaker text-to-speech. IEEE Signal Processing Letters, 2022, 29: 2502–2506. [doi: 10.1109/LSP.2022.3226655]

5 Shen J, Pang RM, Weiss RJ, *et al*. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 4779–4783. [doi: 10.1109/ICASSP.2018.8461368]

6 Oord A, Dieleman S, Zen H, *et al*. Wavenet: A generative model for raw audio. arXiv:1609.03499, 2016.

7 Akuzawa K, Iwasawa Y, Matsuo Y. Expressive speech synthesis via modeling expressions with variational autoencoder. arXiv:1804.02135, 2018.

8 Peng KN, Ping W, Song Z, *et al*. Non-autoregressive neural text-to-speech. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 703.

9 蒿晓阳, 张鹏远. 使用变分自编码器的自回归多说话人中文语音合成. 声学学报, 2022, 47(3): 405–416. [doi: 10.15949/j.cnki.0371-0025.2022.03.004]

10 An XC, Wang YX, Yang S, *et al*. Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis. Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Singapore: IEEE, 2020. 184–191. [doi: 10.1109/ASRU46091.2019.9003859]

11 Peng ZL, Guo ZH, Huang W, *et al*. Conformer: Local features coupling global representations for recognition and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9454–9468. [doi: 10.1109/TPAMI.2023.3243048]

12 Sun GZ, Zhang Y, Weiss RJ, *et al*. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 6264–6268. [doi: 10.1109/ICASSP40776.2020.9053520]

13 尚影, 韩超, 吴克伟. 基于分离对比学习的个性化语音合成. 计算机工程与应用, 2023, 59(22): 158–165. [doi: 10.3778/j.issn.1002-8331.2306-0127]

14 Li T, Yang S, Xue LM, *et al*. Controllable emotion transfer for end-to-end speech synthesis. Proceedings of the 12th International Symposium on Chinese Spoken Language Processing. Hong Kong: IEEE, 2021. 1–5. [doi: 10.1109/ISCSLP49672.2021.9362069]

15 Pan S, He L. Cross-speaker style transfer with prosody bottleneck in neural speech synthesis. arXiv:2107.12562, 2021.

16 Ma MB, Zheng BG, Liu KB, *et al*. Incremental text-to-speech synthesis with prefix-to-prefix framework. arXiv: 1911.02750, 2019.

17 Stephenson B, Besacier L, Girin L, *et al*. What the future brings: Investigating the impact of lookahead for incremental neural TTS. arXiv:2009.02035, 2020.

18 Ren Y, Ruan YJ, Tan X, *et al*. FastSpeech: Fast, robust and controllable text to speech. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., 2019. 285.

19 蔡玥清. 基于 Transformer 的非自回归中文语音合成方法研究 [硕士学位论文]. 武汉: 武汉理工大学, 2021. [doi: 10.27381/d.cnki.gwlgu.2021.001246]

20 何挺. 基于深度学习的端到端汉语语音合成研究 [硕士学位论文]. 杭州: 浙江大学, 2021. [doi: 10.27461/d.cnki.gzjdx.2021.000318]

21 王志超, 吴浩, 李栋, 等. 基于非自回归模型中文语音合成系统研究与实现. 计算机与数字工程, 2023, 51(2): 325–330, 335. [doi: 10.3969/j.issn.1672-9722.2023.02.010]

22 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

23 Li NH, Liu SJ, Liu YQ, *et al*. Neural speech synthesis with transformer network. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019. 6706–6713. [doi: 10.1609/aaai.v33i01.33016706]

24 Ellinas N, Vamvoukakis G, Markopoulos K, *et al*. High quality streaming speech synthesis with low, sentence-length-independent latency. arXiv:2111.09052, 2021.

25 Ren Y, Hu CX, Tan X, *et al*. FastSpeech 2: Fast and high-quality end-to-end text to speech. arXiv:1904.02882, 2019.

26 Whitehill M, Ma S, McDuff D, *et al*. Multi-reference neural TTS stylization with adversarial cycle consistency. arXiv: 1910.11958, 2019.

27 张振国. 面向个性化隐私保护的声纹生成方法研究 [硕士学位论文]. 广州: 广州大学, 2023. [doi: 10.27040/d.cnki.ggzdu.2023.000638]

28 蒲治北. 面向神经网络合成语音的检测技术研究与实现 [硕士学位论文]. 成都: 电子科技大学, 2023. [doi: 10.27005/d.cnki.gdzku.2023.005401]

29 Wang F, Jiang MQ, Qian C, *et al*. Residual attention network for image classification. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6450–6458. [doi: 10.1109/CVPR.2017.683]

30 Shang C, Wu QB, Meng FM, *et al*. Instance segmentation by learning deep feature in embedding space. Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019. 2444–2448. [doi: 10.1109/ICIP.2019.8803021]

31 Wan L, Wang Q, Papir A, *et al*. Generalized end-to-end loss for speaker verification. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary: IEEE, 2018. 4879–4883. [doi: 10.1109/ICASSP.2018.8462665]

32 Zen H, Dang V, Clark R, *et al*. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. arXiv:1904.02882, 2019.

33 Li JY. Recent advances in end-to-end automatic speech recognition. arXiv:2111.01690, 2021.

34 Jia Y, Zhang Y, Weiss RJ, *et al*. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 4485–4495.

(校对责编: 王欣欣)