文章编号:2096-1472(2023)12-0059-04
DOI:10.19644/j.cnki.issn2096-1472.2023.012.012

基于 Transformer 的面部动画生成

豆子闻，李文书

(浙江理工大学计算机科学与技术学院，浙江 杭州 310018)
✉ 1055459759@qq.com; charlie@zstu.edu.cn

摘 要:在面部动画生成领域,克服人脸几何形状的复杂性是一项极具挑战性的任务。为了更好地应对这一挑战,文章采用了一种创新的方法,即将经过一维卷积堆叠和自注意力提取后的音频特征作为输入,通过 Transformer 模型从音频信号中生成面部动画。这个过程采用时间自回归模型逐步合成面部运动。使用 BIWI 数据集开展实验证明,该方法成功地将唇部顶点误差率缩小至令人满意的 6.123%,同步率超过 MeshTalk79.64%,这意味该方法在口型同步和面部表情生成方面表现出色,在完成面部动画生成任务中表现出很高的潜力,可为未来相关研究提供方向和参考。

关键词:动画生成;自回归;深度学习;唇形同步;Transformer
中图分类号:TP391 **文献标志码**:A

Facial Animation Generation Based on Transformer

DOU Ziwen, LI Wenshu

(*School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou* 310018, *China*)
✉ 1055459759@qq.com; charlie@zstu.edu.cn

Abstract: In the field of facial animation generation, overcoming the complexity of face geometry has always been a highly challenging task. To better meet this challenge, this paper proposes an innovative approach which uses audio features that have undergone one-dimensional convolutional stacking and self-attention extraction as input, and generates facial animations from audio signals using a Transformer model. During the process, a time auto-regression model is used to gradually synthesize facial movements. Experiments on BIWI dataset show that this method has successfully reduced the lip vertex error rate to a satisfactory 6.123%, with a synchronization rate 79.64% higher than MeshTalk. This means that the proposed method performs well in lip synchronization and facial expression generation, and shows high potential in accomplishing facial animation generation tasks. It provides directions and reference for future related research.

Key words: animation generation; auto-regression; deep learning; lip synchronization; Transformer

0 引言(Introduction)

在过去的几年里,数字人类引起了广泛关注,它们以高度逼真的方式模拟真实人类,现已被应用于各个领域,比如游戏中的虚拟化身、电影中的角色等[1]。VR 设备的普及,使数字人类被更广泛地应用于虚拟现实场景中,这些数字人通过附着于用户各个关节的传感器联合驱动,能够实时模拟现实中真人的动作,但是对于面部的表情,只能通过面捕设备的摄像头捕捉,其不仅操作不便,更会因为遮挡等原因导致无法跟踪。

在过去的研究中,英伟达(NVIDIA)公司发布的唇音同步算法 Audio2Face 基于深度卷积神经网络,主要集中在学习短音频

收稿日期:2023-05-25

窗口的音素级特征，偶尔会导致嘴唇运动不准确[2]。TIAN 等[3]采用两个双向长短时记忆网络(Bidirectional LSTM)，将音频特征作为输入提取高级语义信息，并输出到注意力层学习注意力权重。这种结构使网络能够记住以往的音频特征，并可鉴别对当前动画帧产生影响的音频特征，但是 LSTM 作为顺序模型仍然存在瓶颈，在有效学习音频数据中跨足够长的时间间隔提取相关信息的能力不足。

Transformer 在自然语言处理和计算机视觉任务中都取得了卓越的表现[4]。研究人员在音频特征的提取方面，加入循环卷积和注意力机制，使得输入不再局限于短时特征，并且显著提高了特征精度。受线性偏差注意力的启发，研究人员在查询键注意力评分中添加了时间偏差，并设计了周期性位置编码策略，以提高模型对较长音频序列的泛化能力。在本文研究中，主要关注三维模型上的面部动画，而三维人脸的复现主要分为基于语言的方法和基于学习的方法[5]。

1 相关理论(Related theory)

1.1 唇音同步

基于语言学的方法通常在音素和视觉对应物之间建立一套复杂的映射规则，即视觉语音音素(Visemes)。Visemes 用于表示人类口型和面部表情的视觉表示，它们对应于发音时嘴巴的不同形状，在计算机图形学、动画和虚拟现实领域有着广泛的应用，尤其是在语音同步(Lip-Sync)动画中。

也有一些方法考虑了音素和音素之间的多对多映射关系[6]。例如，基于心理语言学的考虑，在面部动作编码系统的基础上，将嘴巴运动音素纳入嘴唇和绑定后的下巴动画，可以产生良好的联合发音效果。唇音同步的理论基于对音频的分解，从每段音素中提取梅尔频谱，得到独立的音素级特征。

1.2 Transformer

Transformer 基于 encoder-decoder 进行架构，使用了自注意力机制(Self-Attention)捕捉序列中任意两个位置之间的依赖关系，解决了长期依赖问题。通过多头注意力机制(Multi-Head Attention)同时关注不同位置和不同语义的信息[7]。整体由多个 Encoder 和 Decoder 层堆叠在一起，每一层都包含 Self-Attention 和 Feed Forward Neural Network。每个 Encoder 和 Decoder 层都接收整个句子所有的词作为输入，然后为句子中的每个词都做出一个输出。Transformer 的机制使得它相较于 RNN 和 CNN 每层计算复杂度更优，并且可直接计算点乘结果，不用考虑序列的顺序，可以进行并行处理。在自然语言处理(NLP)和计算机视觉(CV)领域都应用广泛。

1.3 Wav2Vec2.0

Wav2Vec2.0 相比 Wav2Vec，使用 Transformer 代替 RNN，同时引用了一个乘积量化的操作，使得语音表示更加紧凑或离散[8]；其通过对比学习进行自监督学习，首先使用一个卷积神经网络将原始音频信号编码成一个连续的隐层表示，其次使用一个量化模块将这个表示转换成一个离散的潜在表示，最后使用一个 Transformer 网络捕捉这个潜在表示的上下文信息。Wav2Vec2.0 在训练过程中会随机地掩盖一些潜在表示，然后让 Transformer 网络预测被掩盖的部分，这样就可以学习到语音信号中有用的结构和模式。

2 本文所提方法(The proposed method)

首先，从原始音频提取音频特征，将音频作为输入，采用预训练的 Wav2Vec2.0 模型对输入的音频进行编码，得到音频嵌入向量，通过一维卷积和注意力机制帮助模型学习更多时序信息和长距离依赖关系。其次，采用一个基于 Transformer 架构的自回归的 Viseme 解码器对音频特征进行解释，生成一组泛化的运动特征 $\hat{v}_{1:T}$。最后，使用一个风格自适应的动作解码器通过结合 Viseme 特征 $\hat{v}_{1:T}$ 和风格嵌入 S_i，从而将这些运动特征映射到人脸网格体的顶点位置上，从而实现在模型上生成人物的面部动作。唇音同步模型流程图如图 1 所示。

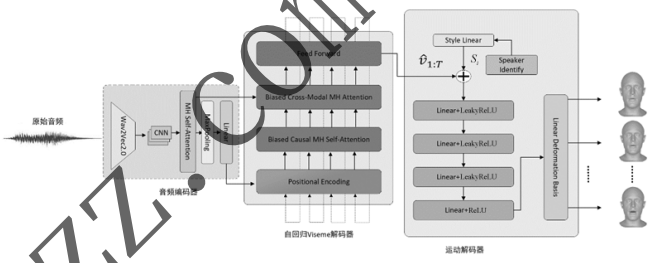

图 1 唇音同步模型流程图

Fig. 1 Flow chart of labial synchronization model

2.1 音频编码器

在编码器的设计方面，研究人员参考了最先进的运动合成模型，使用广义语音模型 Wav2Vec2.0 编码音频输入。编码器以自监督和半监督的方式训练模型，通过使用对比损失预测当前输入语音的近期未来值，使模型能够从大量未标记的数据中学习。

Wav2Vec2.0 接受原始音频信号作为输入，然后在不用手动标注的情况下学习音频特征表示，对原始音频的潜在表示进行建模。Wav2Vec2.0 的输出序列是一组潜在向量，这些向量表示输入音频信号的时间结构特征。每个潜在向量都对应输入音频的一小段时间。在音频特征提取阶段，使用一维卷积处理具有时间顺序的数据，它可以捕捉局部特征并保留输入数据的顺序结构，并且在计算方面更加高效，适用于处理大规模数据。一维卷积层的参数如表 1 所示。

表1 一维卷积层参数

Tab.1 One-dimensional convolution layer parameters

卷积层	卷积核大小	步长	填充	输出通道
Conv1	3	1	1	64
Conv2	5	2	2	128
Conv3	3	1	1	256

为了捕捉序列中的全局依赖关系,将多头自注意力机制应用于输出序列,在 Transformer 编码器层中,输入序列首先经过多头自注意力子层,其次通过位置前馈网络(Position-wise Feed-Forward Network,FFN)。该输出将作为动作编码器中 Biased Cross-Modal MH Attention 层的输入。

经过多头自注意力机制处理后,将序列输入后续的池化层和线性映射层,最终结果作为 Postional Encoding 层的输入。

通过以上方法提取音频的特征并匹配运动的采样频率(本文设置面部运动的帧率为 30 fps,音频为 16 kHz),从而得到 T 个动作帧的音频序列的上下文表示 $\{\hat{a}\}_{t-1}^{T}$。

2.2 自回归解码器

解码器 F_v 将音频的上下文表示作为输入,并以自回归的方式生成风格不可知的 Viseme 特征 V_t,这些视觉特征描述了给定上下文音频和之前的视觉特征,以及嘴唇是如何变形的。与 Faceformer 相比,本文提出了使用经典 Transformer 架构作为 Viseme 解码器,它用来学习从音频特征 $\{\hat{a}\}_{t-1}^{T}$ 到未知的 Viseme 特征 $\{\hat{v}\}_{t-1}^{T}$ 的映射。自回归解码器的公式为

$$\hat{v}_t = F_v + (\theta_v; \hat{v}_{1:t-1}, \hat{a}_{1:T}) \tag{1}$$

其中,θ_v 为 Transformer 的可学习参数。

与产生离散文本的传统神经机器翻译(NMT)架构相比,本文研究的输出表示是一个连续的向量。NMT 模型使用一个开始和结束的标记指示序列的开始和结束。在推理过程中,NMT 模型自回归地生成标记,直到结束。研究人员在输入特征时进行 Linear 操作,在开始处包含输入信息。然而,由于序列长度 T 是由音频输入的长度给出的,所以不适用于结束标记。通常将编码时间添加到序列中的 Viseme 特征中,将时间信息注入序列中。将位置编码的中间表示表述为

$$\hat{h}_t = \hat{v}_t + PE(t) \tag{2}$$

其中,$PE(t)$为正弦编码函数。给定位置编码输入序列 $\hat{h}_t$,使用多头自注意力根据输入的相关性对输入进行加权,从而生成输入的上下文表示。这些上下文表示作为一个跨通道的输入多头注意块也需要来自音频编码器的音频特征 $\hat{a}_{1:T}$ 作为输入。最后一个前馈层将这个音频-动作注意力层的输出映射到 Viseme 嵌入 $\hat{v}_t$。

2.3 动作解码器

动作解码器根据风格不确定的 Viseme 特征 $\hat{v}_{1:T}$ 和风格嵌入 S_i 生成 3D 面部动画 $\hat{y}_{1:T}$。动作解码器由两个组件组成,即风格嵌入层和动作合成块。对于风格未知的 Transformer 的训练和动作解码器的预训练,可以假设训练集的身份为独热码。风格嵌入层将此标识信息作为输入并生成风格嵌入 S_i,它编码了特定身份的动作。

风格嵌入和 Viseme 特征 $\hat{v}_{1:T}$ 拼接并输入动作合成块中。动作合成块由非线性层构成,它将风格感知的 Viseme 特征映射到由线性变形基础定义的动作空间。在训练过程中,通过数据集中的所有身份学习变形基础。变形基础经过微调,可以适应训练不足的身份,网格输出 $\hat{y}_{1:T}$ 通过将估计的顶点变换添加到模板网格体中进行计算。

3 实验与结果(Experiment and result)

3.1 实验数据集

研究人员使用公开的 3D 数据集 BIWI 对本文中的面部动画生成模型进行训练和测试。该数据集提供了英语口语的音频-3D 扫描对。BIWI 包含 40 个独特的句子,该 40 个句子覆盖了常用的发音口型,适用于所有说话者。

BIWI 数据集是一个包含情感语音和相应的密集动态三维人脸几何的语料库。14 名受试者被要求阅读 40 个英语句子,每个句子分别在中性或情绪化的语境中被录下两次。3D 面部几何图形以 25 fps 的速度捕获,每个图形有 23 379 个顶点。每个片段平均时长为 4.67 s。实验中使用了在情感语境中记录句子的子集。具体来说,将数据分为 6 名受试者共说 192 句话的训练集(BIWI-Train),每名受试者说 32 句话,以及两个测试集(BIWI-Test-A 和 BIWI-Test-B)。BIWI-Test-A 包含 6 个可见的被试者共说的 24 句话(每人说 4 句话),BIWI-Test-B 包含 8 个不可见的被试者共说的 32 句话(每人说 4 句话)。

3.2 训练细节

训练 Transformer 编码器、解码器和嵌入块进行跨模态映射。为了从大规模语料库的语音表示学习中受益,使用预先训练好的 Wav2vec2.0 权重初始化 Transformer 编码器。

在编码器的第一阶段,选择 AdamW 作为优化器,其参数如表 2 所示。

表2　AdamW 参数设置

Tab.2 AdamW parameter settings

β_1	β_2	ε
0.9	0.999	1e−8

在第二阶段,用 Adam 优化器训练时间自回归模型,训练时间为 100 个 *epoch*,其他超参数不变。

3.3 评价结果

使用唇形同步度量评估嘴唇运动的质量。所有唇边顶点的最大误差定义为每一帧的唇型误差。误差是通过比较预测和捕获的三维人脸几何数据计算得来的。表 3 统计了使用 MeshTalk[9]、FaceFormer[10] 和本文方法得出的唇形顶点误差比较结果。

表3　唇形顶点误差率比较

Tab.3 Comparison of lip shape error rates

方法	唇部顶点误差率/%
MeshTalk	8.434
FaceFormer	7.242
本文方法	6.123

3.4 相关方法之间的结果比较

人类的感知系统能够理解细微的面部动作和捕捉唇形同步。因此,在语音驱动的面部动画任务中,人类的感知仍然是一个最可靠的度量。本文进行了一项用户调查研究,并与MeshTalk、FaceFormer和Ground Truth(GT)进行感知结果比较。采用A/B(两种方法在各种数据上的比值)测试每个比较,即在逼真的面部动画和口型方面与上述方法的比较。对于BIWI,分别从BIWI-test-B中随机选取30个样本,得到四种比较的结果。为了在说话风格方面达到最大的变化,必须确保抽样结果可以相应地涵盖所有的条件反射风格。因此,本文基于BIWI-test-B创建了120对A和B,共30个样本、4个对照。每组由至少3名不同的参赛者分别评判,最终共收集到372组评价结果。表4为模型在测试集BIWI-Test-B上的用户学习结果的对比,分别比较了同步率和真实值,证明本文所提方法相比MeshTalk和FaceFormer有较显著的提升。

表4 在BIWI-Test-B上的感知评价对比

Tab.4 Comparison of perceptual ratings on BIWI-Test-B

模型对比	同步率/%	真实值/%
本文模型对比MeshTalk	79.64	74.32
本文模型对比FaceFormer	85.62	76.45
本文模型对比GT	22.14	46.05

3.5 结果可视化

为了验证算法的效果,以美国前总统奥巴马的一次演讲Obama Delivers Thanksgiving Greeting中的片段作为音频输入用于合成面部动画帧,从音频到面部动画的生成结果如图2所示。

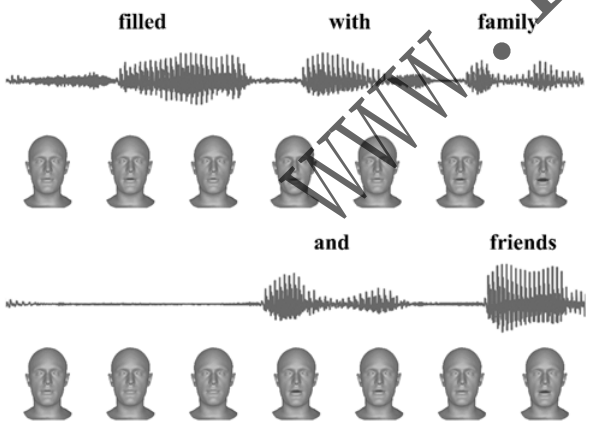

图2 从音频到面部动画的生成结果

Fig. 2 Generating results from audio to facial animation

4 结论(Conclusion)

通过使用一维卷积和自注意力机制,本文所提的方法更好地捕捉到了Wave2Vec2.0输出的特征,帮助研究人员在面部动画合成算法中合成高质量的动画。此外,研究人员展示了在离散空间中将语音驱动的面部动画转换为代码查询任务的优势,即显著地提高了对抗跨模态模糊运动的合成质量。实验结果表明,该方法在实现准确的唇形同步和生动的面部表情方面具有优势。

参考文献(References)

[1] 冯晓宇. 元宇宙背景下虚拟数字人发展面临的挑战及对策建议[J]. 文化产业,2022(36):19-21.

[2] KARRAS T,AILA T,LAINE S,et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion [J]. ACM Transactions on Graphics,36(4):94.

[3] TIAN G Z,YUAN Y,LIU Y. Audio2Face:generating speech/face animation from single audio with attention-based bidirectional LSTM networks[C]//IEEE. Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway:IEEE,2019:366-371.

[4] HAN K,WANG Y H,CHEN H T,et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(1):87-110.

[5] 蔡国鑫. Audio2Face基于音频文件智能生成虚拟角色面部动画[J]. 现代电影技术,2021(9):60-61.

[6] YADAV M,ALAM A. Dynamic time warping (dtw) algorithm in speech:a review[J]. International Journal of Research in Electronics and Computer Engineering,2018,6(1):524-528.

[7] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[DB/OL]. (2017-06-12)[2023-01-04]. https://arxiv.org/abs/1706.03762.

[8] FAN R C,ZHU Y Z,WANG J H,et al. Towards better domain adaptation for self-supervised models:a case study of child ASR[J]. IEEE Journal of Selected Topics in Signal Processing,2022,16(6):1242-1252.

[9] RICHARD A,ZOLLHÖFER M,WEN Y D,et al. MeshTalk:3D face animation from speech using cross-modality disentanglement [C]//IEEE. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE,2021:1153-1162.

[10] FAN Y R,LIN Z J,SAITO J,et al. FaceFormer:speech-driven 3D facial animation with transformers[C]//IEEE. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2022:18749-18758.

作者简介:

豆子闻(1994-),男,硕士生。研究领域:三维重建,虚拟现实。

李文书(1975-),男,博士,教授。研究领域:图像处理,认知建模,虚拟现实,物联网集成开发。