计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):44−50 [doi: 10.15888/j.cnki.csa.008776]

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

基于语音驱动的三维人脸动画技术综述①

刘贤梅, 刘　露, 贾　迪, 赵　娅, 田　枫

(东北石油大学 计算机与信息技术学院, 大庆 163318)
通信作者: 刘　露, E-mail: liulu_all@163.com

摘　要: 随着三维数字虚拟人的发展, 语音驱动三维人脸动画技术已经成为虚拟人交互的重要研究热点之一. 其关键技术在于语音-视觉映射模型的建立以及三维人脸动画的合成. 首先分析了音-视素匹配法和音-视觉参数映射两类方法的特点; 之后阐述了目前三维人脸模型的建立方法, 并依据三维人脸模型的表示方法不同, 分析了不同运动控制方法的优缺点; 然后阐述了语音驱动三维人脸动画的主观评价和客观评价方法; 最后总结了语音驱动三维人脸动画技术的未来发展方向.

关键词: 三维人脸动画; 语音; 语音-视觉映射模型; 虚拟人

引用格式: 刘贤梅,刘露,贾迪,赵娅,田枫.基于语音驱动的三维人脸动画技术综述.计算机系统应用,2022,31(10):44–50. http://www.c-s-a.org.cn/1003-3254/8776.html

Overview on Speech-driven 3D Facial Animation Technology

LIU Xian-Mei, LIU Lu, JIA Di, ZHAO Ya, TIAN Feng

(School of Computer & Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: With the development of 3D digital virtual humans, speech-driven 3D facial animation technology has become one of the important research hotspots in virtual human interaction. The key parts of the speech-driven 3D facial animation technology include the construction of a speech-visual mapping model and the synthesis of 3D facial animation. Specifically, the characteristics of phoneme-viseme matching methods and speech-visual parameter mapping methods are described. Next, the current methods of building 3D facial models are expounded, and the advantages and disadvantages of different motion control methods are analyzed according to the different representation methods of 3D facial models. Then, the subjective and objective evaluation methods for speech-driven 3D facial animation are expounded. Finally, the future development directions of speech-driven 3D facial animation technology are summarized.

Key words: 3D facial animation; speech; speech-visual mapping model; virtual human

1　引言

近年来, 三维数字虚拟人正逐渐走入大众视野, 如2021 年登上春晚舞台的虚拟偶像“洛天依”, 央视推出的虚拟主持人“小 C”等. 虽然目前大多三维数字虚拟人模型精美、动作逼真, 但面部动画的合成严重依赖人为设定, 使用动作捕捉设备[1]、三维扫描设备[2]、单摄像头设备[3] 等硬件设备的表演驱动方法, 因设备价格昂贵、获取和处理数据过程复杂、受面部遮挡、光照、姿态的影响较大等原因限制了应用场景. 由于语音获取方便, 受外界影响较小, 因此有学者提出使用语音驱动的方法合成三维人脸动画, 提高用户的体验感及交互的友好性.

人对面部的细微变化敏感, 面部运动与语音不一致, 会使用户产生违和感. 语音驱动三维人脸动画主要涉及语音到视觉的映射和三维人脸动画合成两个关键技术问题. 语音到视觉的映射技术是从语音中预测视

① 基金项目: 黑龙江省自然科学基金 (LH2020F003); 黑龙江省高等教育教学改革重点委托项目 (SJGZ20200037); 黑龙江省教育科学“十四五”规划重点课题 (GJB1421114)
收稿时间: 2021-12-31; 修改时间: 2022-01-28; 采用时间: 2022-04-02; csa 在线出版时间: 2022-07-07

觉信息, 通过寻找语音与视觉信息之间的复杂联系, 建立非线性映射模型, 得到与语音保持同步的嘴部运动信息和面部表情信息. 三维人脸动画合成通过视觉信息使静态人脸模型发生形变, 实现眼睛、眉毛、嘴唇及面部其他部位的运动, 完成声画同步的三维人脸动画. 语音驱动三维人脸动画应用领域广泛, 在服务行业实现虚拟客服、虚拟助手, 提高用户服务体验; 在影视行业实现自动化真实感虚拟角色动画制作, 减少人工成本, 提高生产效率; 在教育行业实现智慧教室, 促进学生个性化学习; 在娱乐行业实现虚拟偶像、游戏制作, 提高玩家趣味性.

本文将从语音-视觉映射、三维人脸动画合成, 以及语音驱动三维人脸动画效果的评价 3 个方面对已有的研究进行阐述, 分析各种方法的优缺点, 对三维人脸动画的未来发展方向做出展望.

2 语音-视觉映射技术

2.1 音-视素匹配

音素是语音中的最小单位, 一个发音动作构成一个音素, 通常使用语音识别技术提取语音中的音素. 视素 (viseme)[4] 起源于视觉 (visual) 和音素 (phoneme) 两个单词, 表示音素对应的面部动作模型.

音-视素匹配分为传统机器学习方法和深度学习方法. 传统机器学习方法方面, Hofer[5] 提出多阶段隐马尔科夫模型 (multi-stream hidden Markov model, MHMM), 通过隐马尔科夫模型 (hidden Markov model, HMM) 根据语音特征流生成相应的视素序列, 并送入基于轨迹的 HMM, 生成平滑的唇部运动轨迹. 深度学习方法方面, Zhou 等人[6] 提出 VisemeNet 模型, 使用三级长短期记忆网络 (long short-term memory, LSTM) 完成音素组的提取、面部标志几何位置的预测、下颚与嘴部的权重预测, 实现语音可视化.

音-视素匹配依赖语音识别技术, 忽略了语音中语气变化、语调顿挫等情感信息, 在虚拟人语音交互时缺乏生动的面部表情.

2.2 音-视觉参数映射

音-视觉参数映射通过建立语音特征和视觉参数序列的映射模型, 完成语音可视化.

2.2.1 语音特征提取

语音特征提取主要分为手工提取方法和深度学习提取方法, 手工提取方法主要提取语音低级描述符 (low level descriptions, LLDs), 采用全局统计的方式 (如方差、极值、极值范围等) 表征语音特征. LLDs 分类如表 1 所示.

表 1 LLDs 分类

特征	具体特征
韵律特征	能量、共振峰、音高、时长、发音、基音频率、过零率
音质特征	相位、频率、声门参数
谱特征	MFCC、LPCC、SDC

Englebienne 等人[7] 使用梅尔倒谱系数 (Mel-frequency cepstral coefficients, MFCC) 提取语音的语义和韵律信息. Xie 等人[8] 在 MFCC 中加入一阶导数和二阶导数, 描述语音的动态信息. Bandela 等人[9] 将 Teager 能量算子和 MFCC 融合形成新的特征, 用于识别语音信号的情绪. 目前常用的 LLDs 提取的开源工具为 Eyben 等人[10,11] 开发的 OpenSMILE 和 OpenEAR, 可批量自动提取包括时长、基频、能量和 MFCC 等常用的声学特征. Ramanarayanan 等人[12] 使用 OpenSMILE 从音频中提取短时特征, 用于识别语音中的副语言信息.

由于手工定义的 LLDs 不能完整描述语音信号, 因此近年来学者尝试使用深度学习的方法从 LLDs 中进一步提取语音高级特征或者直接处理原始语音. 常用的方法有深度神经网络 (deep neural networks, DNN)、卷积神经网络 (convolutional neural networks, CNN)、循环神经网络 (recurrent neural network, RNN) 等. Zhang 等人[13] 设计一个从大量原始数据中学习帧级说话者特征的 DNN 模型, 此模型在短的语音段中获得良好的识别准确率. Mustaqeem 等人[14] 采用 CNN 从语谱图中提取语音特征, 改善 MFCC 对语音高频信息识别准确率不高的问题. Wu 等人[15] 采用两个循环链接的胶囊网络提取特征, 增强语音的时空信息表达能力. Zhao 等人[16] 采用局部特征学习块, 从 MFCC 中提取局部特征, 然后使用 LSTM 进一步提取语音全局的上下文特征.

2.2.2 视觉参数定义

Parke[17] 将视觉参数分为形状参数和表情参数, 形状参数控制个性化人脸细节, 表情参数控制人脸表情.

形状控制参数使用三维坐标点 (x, y, z) 表示. 倪虎[18] 定义 8 个三维特征点表示三维人脸嘴部运动. 文献 [19–21] 使用三维人脸模型中的全部顶点坐标表示面部及嘴部运动.

Blendshape 权重是具有语义信息的表情参数, 可以直接控制嘴角、眉眼等部位运动. Pham 等人[22,23]、Tian 等人[24] 分别采用 46 维和 51 维 blendshape 权重控制 blendshape 三维人脸模型合成三维人脸表情.

视觉参数定义与后续三维人脸模型运动控制方法一一对应. 使用三维坐标点作为视觉参数时, 动画实现效果与定义的三维特征点数量相关, 数量越多, 人脸运动精度越高, 但计算量会增加, 达到一定数量之后难以实现实时计算. 使用 blendshape 权重作为视觉参数时, 三维人脸模型运动控制方法简单、控制数据量较少, 是目前常用的视觉参数.

2.2.3 音-视觉映射模型建立

音视觉映射模型建立分为传统机器学习方法和深度学习方法. 传统机器学习方法主要采用 HMM 和高斯混合模型 (Gaussian mixture model, GMM). Brand[25] 根据 HMM 可以存储上下文信息的能力从语音中获得的信息来预测全脸动画. Xie 等人[26] 在文献 [25] 的基础上提出双层 HMM, 训练多流 HMM 模型建立对应关系. 之后 Xie 等人[27] 引入了耦合 HMM 来解决由协同发音引起的视听活动之间的异步性. HMM 在训练阶段具有较大的计算量, 没有考虑输入语音的个体差异, 且难以对复杂的上下文依赖关系进行建模, 精确度不高. Deena 等人[28] 采用 GMM 实现语音参数与人脸动画的匹配, 对表情动作和语音参数分别建立数据模型, 建立表情与语音的相互联系, 实现语音信息与表情细节的同步. Luo 等人[29] 对传统的 GMM 方法进行改进, 提出基于双高斯混合模型的音频到视觉的转换方法, 解决了视觉参数误差的积累. 但是 GMM 无法改变训练数据的内在结构, 对数据的依赖性较大, 导致了跨数据库的通用性不强.

由于深度学习在建立非线性映射上效果较好, 因此有学者使用该方法建立音-视觉映射模型, Karras 等人[19] 将网络划分为频率分析层、发音分析层、顶点输出层, 使用 LPCC 语音特征点输出视觉参数. 该方法忽略了语音情绪与表情关联的时序性, 难以合成真实的人脸表情. Cudeiro 等人[20] 提出了 Voca 网络, 该网络采用基于 CNN 的编码器-解码器结构, 编码器将语音特征转换为低维嵌入, 使用解码器得到三维顶点位移的高维空间. Richard 等人[21] 提出 MeshTalk 网络, 该网络通过判断面部与音频相关性的强弱, 对人脸上下区域的视觉参数分别建模, 合成带眉眼运动的三维人脸动画. Pham 等人[22] 使用 LSTM 通过分析语音频谱图、MFCC 和色谱图预测三维人脸表情动画参数, 该方法一定程度上解决语音协同发音的现象, 但因语音特征的限制, 对快乐的情绪拟合较差. 之后 Pham 等人[23] 首次将经典的 CRNN (convolutional recurrent neural network) 模型结构应用于端到端音视觉映射模型的建立, 并且该网络模型无需加入额外表征情绪的语音特征, 就可以推断出眉、眼等表征情绪的视觉参数. 网络模型结构如图 1 所示, 使用 CNN 从语谱图中完成语音频域和时域信息的特征提取, 其中, F-Conv1 到 F-Conv5 用于频域特征提取, T-Conv1 到 T-Conv3 用于时域特征提取. 由于语谱图的横纵坐标的物理意义不同, 两个维度包含的信息也不同, 因此使用一维卷积核分别遍历语谱图的横轴和纵轴, 提取不同维度的语音全局特征. 该方法相比二维卷积可以有效地减少计算量, 加速语音提取的过程. 每个卷积层包括卷积、批处理归一化和 ReLU 激活 3 个操作, 使用卷积步长为 2 的方式进行下采样. 然后使用不同的 RNN 接入全连接层 (fully connected layers, FC) 分别建立语音与视觉参数的时序关联性建模, 提高视觉参数精度.

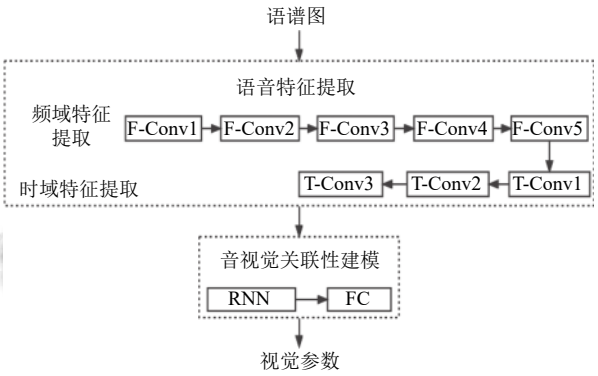

图 1 CRNN 网络模型结构

使用深度学习建立音-视觉映射模型需要三维视听数据集作为支撑. Fanelli 等人[30] 提出 B3D(AC)^2, 该数据集共有 14 名演员、1 109 条语音, 包括消极、悲伤、愤怒、压力、诱惑、恐惧、惊喜、兴奋、自信、快乐、积极, 共计 11 种情绪. 视觉参数采用三维坐标点的形式, 共计 23 370 个顶点. 该数据集的视觉参数仅包含人脸结构, 并不包含头部等运动信息. Pham 等人[23] 提出一种视觉参数为 blendshape 权重的三维视听数据集, 该数据集包括 24 名演员, 每名演员有 60 条语音, 包括自然、平静、快乐、悲伤、愤怒、恐惧、

惊讶和厌恶, 共 8 种情绪, 每种情绪有平缓、强烈两种情况. Cudeiro 等人[20] 提出 VOCASET, 包含 12 个主题和 480 条语音, 视觉参数使用三维坐标点形式, 共计 5 023 个顶点, 包含头部旋转等运动信息. 该数据集仅有中立的可视化语音信息, 不包含其他情绪.

由于三维人脸数据集的构造需要借助三维运动捕捉等硬件设备, 需要耗费大量的人力物力, 导致目前开源数据集较少.

3 三维人脸动画合成技术

3.1 三维人脸模型建立

由于人脸生理结构和几何外观的复杂多样性, 不同肤色、不同性别的人, 其五官比例、面部特征具有极大的差异, 因此建立逼真、自然的三维人脸模型具有较大的难度. 目前建模方式主要有基于三维建模软件的手工建模、基于硬件设备的捕捉建模和基于二维图像的人脸建模.

基于三维建模软件的手工建模主要使用 3DS MAX, MAYA 等商业软件. 此方法建模效果精致、形状可控度高, 但对操作者的专业知识要求较高、建立过程耗时耗力, 效果受人为因素影响较大.

基于硬件设备的捕捉建模主要是通过先进的工业设备 (如三维激光扫描仪、结构光扫描仪), 通过传感器获取人脸面部特征点信息与纹理特征等信息, 然后将获得的信息经过计算机图形学技术恢复三维人脸几何模型. Peszor 等人[31] 首先通过结构光扫描仪获得真实人脸模型, 然后通过修正模型来建立合适的人脸几何模型. Li 等人[32] 采用多个摄像机捕获高质量的三维头部扫描数据. Ye 等人[33] 使用结构光扫描仪构建了 SIAT-3DFE 高精度三维人脸表情数据集. 该类方法虽然可以建立高精度人脸模型, 但其设备价格昂贵、且获取的数据量较大、数据处理较复杂.

基于二维图像的人脸建模使用二维图像结合视觉技术重构面部的三维数据. Jackson 等人[34] 提出 VRN (volumetric regression networks) 端到端的神经网络从单幅图像直接进行三维面部重建. Chen 等人[35] 使用基于条件生成对抗网络的深度面部细节网络, 直接从人脸图像中重建细节丰富的三维人脸. Feng 等人[36] 设计 UV 位置图的二维表示方法, 记录三维形状在 UV 空间中的表示, 然后使用 CNN 从图像中回归. 该类方法获取数据方便、成本低、建模过程自动化, 但重建时可能会因三维人脸形状过度泛化导致人脸个性化信息缺失.

3.2 三维人脸模型运动控制

在建立好三维人脸模型后, 需要控制三维人脸模型运动, 使人脸模型发生形变, 合成三维人脸动画. 依据三维人脸模型表示方法的不同, 三维人脸模型运动控制方法分为参数模型运动控制方法和肌肉模型运动控制方法.

参数模型依据运动方式的不同, 分为多边形形变模型和 blendshape 模型. 多边形形变模型将三维人脸模型用多边形面片表示, 通过控制面片上三维坐标点来实现三维人脸模型运动. Richard 等人[21] 通过控制 5 023 个顶点的多边形形变模型, 实现三维人脸模型运动. 多边形形变模型虽然可以控制高精度的三维人脸模型运动, 但调整参数过程复杂.

Blendshape 模型将人脸表示为一组拓扑结构相同的表情基的线性组合, 包括一个基准三维人脸模型和一系列具有指定人脸动作的表情基, 通过调整不同的表情基权重, 完成三维人脸模型的运动控制. Blendshape 模型如式 (1) 所示:

$$S = B_0 + \sum_{i=1}^{N} (B_i - B_0)e_i \tag{1}$$

其中, N 是表情基个数, e_i是 blendshape 权重, S 是三维人脸模型形变后的状态, B_0是基准三维人脸模型, B_i是第 i 个人脸动作的表情基.

Yu 等人[37] 使用 blendshape 模型对面部表情进行重建与优化, 提高了表情的精准度的同时, 维持了 blendshape 方法的高效性. Alkawaz 等人[38] 使用 blendshape 模型设计一个面部表情动画系统. Wang 等人[3] 使用 RGBD 相机和 blendshape 模型实现了支持表情细节变化的实时面部跟踪系统. blendshape 模型的运动控制操作简单, 但其实现效果依赖表情基精度和完备性.

手工建立 blendshape 表情基的方法耗时耗力, 并且建立的表情基不能重复使用, 因此有学者使用表情迁移自动化建立不同人脸模型的表情基. 表情迁移是将已有角色模型 (源模型) 的人脸表情克隆到新模型 (目标模型) 上. 表情迁移分为标记点迁移方法和深度学习迁移方法. 标记点迁移方法方面, Sumner 等人[39] 使用手工标记的顶点建立源模型到目标模型的相对映射, 通过线性优化函数和映射关系完成表情迁移. 深度学习迁移方法方面, Gao 等人[40] 提出了自动形变两个

不成对形状集 (VAE-CycleGAN) 方法, 使用两个卷积变分自编码器将源模型表情和目标模型映射到潜在空间, 然后使用 GAN 将潜在空间的信息映射到目标模型上, 最后采用相似性约束条件保证迁移表情一致性. Jiang 等人[41] 使用三维顶点形变表示高维模型表情信息, 并使用图卷积网络 (graph convolutional network, GCN) 实现表情迁移. 由于人脸结构空间维度高, 并且人们对表情变化细节极其敏感, 因此保证迁移后表情模型的个性细节特征是该方法的难点.

肌肉模型是通过模拟肌肉底层的位移来控制三维人脸模型运动, 依据解剖学原理将面部肌肉分为线性肌、括约肌和块状肌等. Platt 等人[42] 率先提出该模型, 使用弹簧特性对人脸肌肉建模, 通过肌肉的弹力控制人脸运动. Zhang 等人[43] 采用弹簧-质点模型建立肌肉模型, 模拟人脸皮肤的弹性效果. Yue 等人[44] 建立下巴旋转模型与口部肌肉模型, 然后运用 GFFD (广义自由变形) 面模拟面部皮肤运动, 最后通过融合肌肉模型与皮肤变形实现面部表情的变化. 基于肌肉模型的运动控制法通过对人脸结构进行物理仿真, 可以真实的模拟人脸运动, 但由于人脸肌肉结构复杂, 使用该方法生成动画需要大量的人工交互辅助, 因此不适用普通消费级用户.

4 语音驱动的三维人脸动画效果评价

语音驱动三维人脸动画效果评价包括主观评价和客观评价两种方法. 主观评价通过给出不同分值的动画参考样例, 使用平均分 (mean opinion score, MOS)[45]、诊断可接受性测量 (diagnostic acceptability measure, DAM)[46] 方法进行评价. 评价内容包括合成人脸动画整体的自然度、流畅度, 以及语音与嘴部运动及面部神态的一致性.

客观评价包括合成动画实时性评价、语音-视觉映射精度评价、动画流畅度评价. 在实时性方面, 通过计算语音预处理、语音-视觉映射、三维人脸模型形变渲染的总时间判断合成动画的实时性[23,24]. 在语音-视觉映射精度方面, 通过计算真实值与动画面部关键点的差值判断语音-视觉映射的精度, 计算方法如欧氏距离[20,21]、均方根误差[22,24]、关键点运动轨迹差值评估[18] 等. 在动画流畅度方面, 通过计算当前动画帧面部关键点位置与前后帧的位移判断动画流畅度[23].

5 结论及展望

随着人工智能与虚拟人的不断结合, 使用深度学习方法实现端到端的语音驱动三维人脸动画成为研究的主流方向. 综合国内外对该技术的研究现状, 在未来的发展中仍然有许多挑战, 特别是在数据集、面部表情细节动画、头部运动姿态等方面.

(1) 由于深度学习需要大量数据作为支撑, 数据集的全面性直接影响了语音-视觉映射模型的构建效果, 现有的公开三维视听数据集较少, 且没有统一的构建标准, 因此很难对不同的语音-视觉映射模型进行统一的客观评价.

(2) 人们会通过细微的表情变化揣摩说话时人的情感, 虚拟人的面部微表情可以增强角色的感染力, 因此可以考虑从提高语音情绪细节特征的表达能力入手, 模拟眼角、嘴角、眉毛等面部细节的变化.

(3) 目前语音驱动三维人脸动画的表情合成是基于离散情绪的, 只能刻画有限的几种情绪类型. 但在现实生活中, 人类的情绪是复杂的, 存在悲喜交加、惊喜交集等情况. 因此可以使用语音情绪识别中的连续情感模型, 分析可视化的复合语音情绪, 实现人脸表情的丰富性.

(4) 人们在说话时会产生不同频率的头部运动, 然而语音与头部姿态关联性较弱, 因此可以考虑使用眼动追踪等相关技术实现头部姿态估计, 增强语音动画的真实感.

(5) 由于人脸的结构复杂, 在生成人脸动画时需要复杂的协同控制模拟真实的人脸运动和表情变化, 使用基于三维顶点坐标的形状参数控制多边形形变模型, 虽然可以拟合表情细节运动, 但是难以达到实时的运行效率. 因此可以使用基于 blendshape 权重的表情语义参数控制 blendshape 模型, 合成三维人脸动画, 通过优化表情基中的面部皱纹等个性细节特征, 实现高精度的三维人脸动画.

参考文献

1 Zoss G, Sifakis E, Gross M, *et al*. Data-driven extraction and composition of secondary dynamics in facial performance capture. ACM Transactions on Graphics, 2020, 39(4): 107.

2 Jara-Quito HJ, Guerrero-Vasquez LF, Parra-Luzuriaga KA, *et al*. AVATAR: Human-computer interface for interaction with children using a live animation process based in facial

and body landmarks recognition. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). Greater Noida: IEEE, 2021. 715–720.
3 Wang ZB, Ling JW, Feng CZ, *et al*. Emotion-preserving blendshape update with real-time face tracking. IEEE Transactions on Visualization and Computer Graphics, 2020. 1.
4 Fisher CG. Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 1968, 11(4): 796–804. [doi: 10.1044/jshr.1104.796]
5 Hofer G. Speech-driven animation using multi-modal hidden Markov models [Ph.D. thesis]. Edinburgh: University of Edinburgh, 2009.
6 Zhou Y, Xu Z, Landreth C, *et al*. VisemeNet: Audio-driven animator-centric speech animation. ACM Transactions on Graphics, 2018, 37(4): 161.
7 Englebienne G, Cootes TF, Rattray M. A probabilistic model for generating realistic lip movements from speech. Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2007. 401–408.
8 Xie L, Liu ZQ. Realistic mouth-synching for speech-driven talking face using articulatory modelling. IEEE Transactions on Multimedia, 2007, 9(3): 500–510. [doi: 10.1109/TMM.2006.888009]
9 Bandela SR, Kumar TK. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Delhi: IEEE, 2017. 1–5.
10 Eyben F, Wöllmer M, Schuller B. Opensmile: The Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia. Firenze: ACM, 2010. 1459–1462.
11 Eyben F, Wöllmer M, Schuller S. Open EAR-introducing the Munich open-source emotion and affect recognition toolkit. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam: IEEE, 2009. 1–6.
12 Ramanarayanan V, Pugh R, Qian Y, *et al*. Automatic turn-level language identification for code-switched Spanish-English dialog. 9th International Workshop on Spoken Dialogue System Technology. Singapore: Springer, 2019. 51–61.
13 Zhang M, Chen YX, Li LT, *et al*. Speaker recognition with cough, laugh and "Wei". 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Kuala Lumpur: IEEE, 2017. 497–501.
14 Mustaqeem, Kwon S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, 2020, 20(1): 183.
15 Wu XX, Liu SX, Cao YW, *et al*. Speech emotion recognition using capsule networks. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6695–6699.
16 Zhao JF, Mao X, Chen LJ. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 2019, 47: 312–323. [doi: 10.1016/j.bspc.2018.08.035]
17 Parke FI. Parameterized models for facial animation. IEEE Computer Graphics and Applications, 1982, 2(9): 61–68. [doi: 10.1109/MCG.1982.1674492]
18 倪虎. 基于 Dirichlet 自由变形算法的人脸表情动画技术研究 [硕士学位论文]. 武汉: 武汉理工大学, 2019.
19 Karras T, Aila T, Laine S, *et al*. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics, 2017, 36(4): 94.
20 Cudeiro D, Bolkart T, Laidlaw C, *et al*. Capture, learning, and synthesis of 3D speaking styles. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10093–10103.
21 Richard A, Zollhöfer M, Wen YD, *et al*. MeshTalk: 3D face animation from speech using cross-modality disentanglement. 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1153–1162.
22 Pham HX, Cheung S, Pavlovic V. Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 2328–2336.
23 Pham HX, Wang YT, Pavlovic V. Learning continuous facial actions from speech for real-time animation. IEEE Transactions on Affective Computing, 2020.
24 Tian GZ, Yuan Y, Liu Y. Audio2Face: Generating speech/face animation from single audio with attention-based bidirectional LSTM networks. 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Shanghai: IEEE, 2019. 366–371.
25 Brand M. Voice puppetry. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles: ACM, 1999. 21–28.
26 Xie L, Liu ZQ. Speech animation using coupled hidden

Markov models. 18th International Conference on Pattern Recognition. Hong Kong: IEEE, 2006. 1128–1131.
27 Xie L, Liu ZQ. A coupled HMM approach to video-realistic speech animation. Pattern Recognition, 2007, 40(8): 2325–2340. [doi: 10.1016/j.patcog.2006.12.001]
28 Deena S, Galata A. Speech-driven facial animation using a shared Gaussian process latent variable model. Proceedings of the 5th International Symposium on Visual Computing. Las Vegas: Springer, 2009. 89–100.
29 Luo CW, Yu J, Wang ZF. Synthesizing real-time speech-driven facial animation. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence: IEEE, 2014. 4568–4572.
30 Fanelli G, Gall J, Romsdorfer H, *et al*. A 3-D audio-visual corpus of affective communication. IEEE Transactions on Multimedia, 2010, 12(6): 591–598. [doi: 10.1109/TMM.2010.2052239]
31 Peszor D, Polanski A, Wojciechowski K. Estimation of marker placement based on fiducial points for automatic facial animation. AIP Conference Proceedings, 2015, 1648(1): 660014.
32 Li TY, Bolkart T, Black MJ, *et al*. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, 2017, 36(6): 194.
33 Ye YP, Song Z, Guo JG, *et al*. SIAT-3DFE: A high-resolution 3D facial expression dataset. IEEE Access, 2020, 8: 48205–48211. [doi: 10.1109/ACCESS.2020.2979518]
34 Jackson AS, Bulat A, Argyriou V, *et al*. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1031–1039.
35 Chen AP, Chen Z, Zhang GL, *et al*. Photo-realistic facial details synthesis from single image. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9428–9438.
36 Feng Y, Wu F, Shao XH, *et al*. Joint 3D face reconstruction and dense alignment with position map regression network. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 557–574.
37 Yu H, Liu HH. Regression-based facial expression optimization. IEEE Transactions on Human-Machine Systems, 2014, 44(3): 386–394. [doi: 10.1109/THMS.2014.2313912]
38 Alkawaz MH, Mohamad D, Basori AH, *et al*. Blend shape interpolation and FACS for realistic avatar. 3D Research, 2015, 6(1): 6. [doi: 10.1007/s13319-015-0038-7]
39 Sumner RW, Popović J. Deformation transfer for triangle meshes. ACM Transactions on Graphics, 2004, 23(3): 399–405. [doi: 10.1145/1015706.1015736]
40 Gao L, Yang J, Qiao YL, *et al*. Automatic unpaired shape deformation transfer. ACM Transactions on Graphics, 2018, 37(6): 237.
41 Jiang ZH, Wu QY, Chen KY, *et al*. Disentangled representation learning for 3D face shape. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11949–11958.
42 Platt S, Badler NI. Animating facial expressions. Proceedings of the 8th Annual Conference on Computer Graphics and Interactive Techniques. Dallas: ACM, 1981. 245–252.
43 Zhang Y, Prakash EC, Sung E. Real-time physically-based facial expression animation using mass-spring system. Proceedings. Computer Graphics International 2001. Hong Kong: IEEE, 2001. 347–350.
44 Yue S, Kitajima K. A method of simulating the fusion of speech and facial expression for individuals based on the GFFD method. 2011 IEEE International Conference on Mechatronics and Automation. Beijing: IEEE, 2011. 898–903.
45 肖磊. 语音驱动的高自然度人脸动画 [硕士学位论文]. 合肥: 中国科学技术大学, 2019.
46 周维. 汉语语音同步的真实感三维人脸动画研究 [博士学位论文]. 合肥: 中国科学技术大学, 2008.

(校对责编: 牛欣悦)