第33卷 第18期 计 算 机 工 程 2007年9月
Vol.33 No.18 Computer Engineering September 2007
·多媒体技术及应用· 文章编号：1000—3428(2007)18—0239—03 文献标识码：A 中图分类号：TP37

语音驱动人脸唇形动画的实现

林爱华[1]，张文俊[2]，王毅敏[2]，赵光俊[1]
(1. 上海大学通信与信息工程学院，上海 200072；2. 上海大学影视艺术技术学院，上海 200072)

摘 要：提出了一种实现语音直接驱动人脸唇形动画的新方法。结合人脸唇部运动机理，建立了与唇部运动相关肌肉拉伸和下颌转动的唇形物理模型，对输入的语音信号进行分析和提取其与唇部运动相关的特征参数，并直接将其映射到唇形物理模型的控制参数上，驱动唇形动画变形，实现输入语音和唇形动画的实时同步。仿真实验结果表明，该方法有效实现了语音和唇形的实时同步，唇形动画效果更接近自然，真实感更强。且该唇形物理模型独立于人脸几何模型，可广泛应用于各类人脸唇形动画的语音驱动，具有良好的普适性和可扩展性。
关键词：唇形动画；唇形物理模型；语音唇形同步

Realization of Speech-driven Facial Lip Animation

LIN Ai-hua[1], ZHANG Wen-jun[2], WANG Yi-min[2], ZHAO Guang-jun[1]
(1. School of Communication & Information Engineering, Shanghai University, Shanghai 200072;
2. School of Film Video Art & Technology, Shanghai University, Shanghai 200072)

【Abstract】 A new approach of speech-driven facial lip synchronized animation is presented. Based on the analyzing mechanism of the facial lip motion, the physical model of muscles and mandible corresponding to lip motion is established, and the input speech is analyzed and parameterized into the feature coefficients, and that are mapped to the control coefficients of the lip model, then the lip animation is driven directly by input speech, and its real-time synchronization is realized. The results of simulations show that the input speech and facial lip animation is synchronized more precisely and efficiently in this way, and the effect realities of facial lip animation during speech are greatly improved, and due to the lip model is independent to the geometric facial model, it is also suitable for speech driven lips in the field of various facial animations.
【Key words】 lip animation; physical lip model; speech-lip synchronization

1 概述

近年来，人脸动画技术在数字娱乐、人机交互以及远程会议等方面得到了广泛的应用，实现语音与唇动面部表情之间的同步是人脸动画的关键之一。目前就语音与人脸唇形动画的同步驱动的研究与实现方法可分为两类：一类是基于文本输入，通过分析给定文本的音素确定口型和表情，生成人脸动画，再由文本合成与唇型同步的语音，使人们在听计算机说话的同时可以看到一个说话的人头[1]。但是这种合成的声音听起来不太真实自然；另一类是基于语音信号输入，它又分为通过语音识别和不通过语音识别两种不同方法，前者通过语音识别出语音信号的音素序列(phoneme)，再将其转换为按时间排列的视素序列(viseme)，按照某种规则驱动与其基本视素对应的唇形。此方法直接、有效，但忽视了动态因素和同步问题，很难协调潜在的语音段落与唇形运动之间一致性[2,3]。后者不通过语音识别，直接将语音特征参数映射到人脸动画参数上，如通过聚类和机器学习的方法学习语音信号和唇动面部表情之间的同步关系，再利用神经网络直接将语音特征参数映射到MPEG-4 定义的人脸运动参数(FAP)模式中[4]。这种方法不仅可以回避语音识别遇到的问题，同时又能与真人发声有效地实现同步，增强真实感和逼真度。语音直接驱动三维计算机虚拟动画角色的动作，是计算机动画设计与制作技术的进步，可以极大地拓展计算机动画技术的应用空间。

本文在分析人脸唇部运动机理的基础上，提出了一种语音驱动人脸唇形动画的新方法。本方法不仅能大大简化动画制作中烦琐的声像同步过程，而且能为实时交互式计算机动画应用提供一条新的互动途径。

2 人脸唇部运动机理

从解剖学角度，人脸面部组织可精细划分为6层包括表皮层、真皮层、皮下脂肪层、筋膜层、肌肉层和骨骼层，而表皮层、真皮层、皮下脂肪层和筋膜层统称为皮肤层，所以人脸面部组织可粗略分为3层：皮肤层，肌肉层和骨骼层。骨骼的大小和形状决定了人脸的基本轮廓，人脸除下颌骨可转动外，其它骨骼基本上不动，下颌骨的转动可影响嘴唇的张闭；脸部肌肉层富有良好的弹性，它的丰满程度决定了人脸的外部特征，同时肌肉的拉伸可以产生各种脸部表情；而皮肤层仅仅体现面部的色泽和衰老程度，它对人脸的几何模型变化影响不大。

人在说话时，是由下颌的转动、舌头的自由活动、唇的展开或收缩以及控制气流进入鼻腔的软腭，使口腔形成不同功能的共鸣器，气流通过时发出不同声音。人脸的唇部是一个复杂的非刚体模型，其形变过程由分布在面部的肌肉组织的收缩以及人体下颌骨的运动来控制，图1给出了与唇部运动相关的面部肌肉示意图[5]。考虑到说话时，影响唇形变化

基金项目：上海市教育委员会科学发展基金资助项目(04AB45)
作者简介：林爱华(1981－)，女，硕士研究生，主研方向：图形图像处理，人脸动画等；张文俊，教授、博士生导师；王毅敏，讲师；赵光俊，硕士研究生
收稿日期：2006-10-15 **E-mail：**aiwa11077@163.com

的主要是口轮匝肌、笑肌、提上唇肌、降下唇肌、主颊骨肌以及降口角肌，因此，在仿真实验中主要是为表 1 所示的肌肉层建立一个简单又能模拟其形变过程的物理模型。表 1 列出了与唇部运动相关的主要肌肉及其功能。

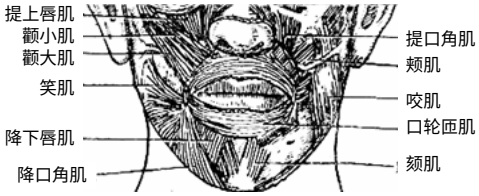

图 1　唇部运动相关的面部肌肉分布图

表 1　唇部运动相关的主要肌肉模型及其功能

模拟肌肉	外形动作
括约肌(轮匝肌)	嘴唇变圆，嘴唇宽度变窄，嘴唇突出
笑肌(左/右线性肌)	嘴唇宽度增加
提上唇肌(左/右线性肌)	上嘴唇抬高
降下唇肌(左/右线性肌)	下嘴唇下降
主颊骨肌(左/右线性肌)	嘴角抬高
降口角肌(左/右线性肌)	嘴角降低
下颌转动肌	嘴巴开合

3　唇形运动变形的物理模型建立

3.1　肌肉模型

本文采用的肌肉模型是线性肌和括约肌两种伪肌肉模型[6]。线性肌通常是用一个向量表示，也称向量肌，如图 2 中的v_1v_2，其影响区域是一个扇形，向量的两端点归类为附属点和插入点，附属点等价于连接在骨头上的真实肌肉(v_1)，其位置固定不变，插入点等价于连接在皮肤上的真实肌肉(v_2, p)，可沿向量方向移动。当线性肌收缩时，便带动肌肉影响区(扇形区)内的网格结点向附属点移动，从而使人脸发生变形。式(1)给出在以附属点为中心的三角扇形区域内任意网格结点(p)的位移方程。

$$p' = p + kR\frac{p - v_1}{D}\cos\theta \tag{1}$$

其中，$D = |p - v_1|$；$R = \begin{cases} \cos(\frac{1-D}{R_1}), & \text{当}p\text{在扇形}v_1R_1R_3\text{内} \\ \cos(\frac{D-R_1}{R_2-R_1}), & \text{当}p\text{在扇形}R_1R_2R_3R_4\text{内} \end{cases}$；$p'$表示 p 点的新位置；k 是肌肉收缩系数，反映肌肉的弹性状况。

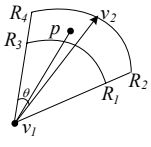

图 2　线性肌示意图

括约肌也称轮匝肌，主要用来控制人脸嘴唇部分的动作，其影响区用一个椭圆表示，如图 3 所示。位于影响区内的网格结点，在括约肌发生挤压时均向椭圆中心移动，同时稍微向前凸起。而椭圆体的中心不受肌肉收缩影响，以防止顶点堆积在椭圆中心。肌肉挤压后其影响区内的任意一点 p 的新位置计算如下：

$$p' = p + kR\frac{p - c}{D} \tag{2}$$

其中，$D = |p - c|$；$R = \begin{cases} 1 - \frac{\sqrt{l_y^2 p_x^2 + l_x^2 p_y^2}}{l_x l_y} \cdot \frac{D}{l_x}, & \text{当} D > l_y \\ 0, & \text{当} D \le l_y \end{cases}$；$k$ 是肌肉挤压系数，反映肌肉的弹性状况。

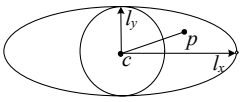

图 3　括约肌示意图

3.2　下颌的转动

人脸的颌骨可分为上颌和下颌，上颌基本保持不动，下颌会因说话、咀嚼等动作发生转动。如果下颌没有动作嘴唇基本不会有动作，故本文仅对下颌的张合进行模拟。如图 4 示，点P代表上下颌的支点；点C代表嘴唇的原始位置；τ_1和τ_2分别表示上下唇随着下颌的张合产生的位置偏移量，其与下颌转动之间的关系如下[7]：

$$\tau_1 = Dis(P,C).\tan(\theta_1)\ ;\tau_2 = Dis(P,C).\tan(\theta_2) \tag{3}$$

其中，$Dis(P,C)$ 表示点 P 和点 C 之间的距离。

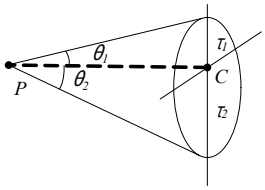

图 4　下颌示意图

4　语音数据处理与系统实现

音频特征参数表征原始音频信息。在研究音频参数对唇部运动控制的规律基础上，选定可以影响角色唇形动作的参数，以保证角色在声音的驱动下做出符合要求的动作。比如，可以通过音强来控制张嘴大小，通过语速来调整张嘴过程的快慢等。提取语音特征参数的方法有多种，如 LPC、MFCC、PLP 等。

本文不采用传统的语素级别(音节、声韵母等)的语音特征提取，直接从原始语音信号中提取出LPC倒谱及短时能量参数组成特征矢量，避免了由于人的口音、方言等因素造成的误差，增强了系统的适用性。LPC倒谱是线性预测参数LPC在倒谱域中的表示，常用LPCC表示。LPCC参数对元音有较好的描述能力，所以更适合对唇形的控制。LPC的倒谱系数$C_{LPCC}(n)$和LPC系数$C_{LPC}(n)$之间的关系如下[7]：

$$C_{LPCC}(n) = C_{LPC}(n) + \sum_{k=1}^{n-1}\frac{n-k}{n}C_{LPCC}(n-k)C_{LPC}(k) \tag{4}$$

本文构建的系统流程如图 5 所示。首先对输入的语音信号(.WAV 文件或麦克风)进行分析，提取出与唇动相关的参数 LPCC 和短时能量参数，然后将这些特征参数映射到相应的物理模型控制参数上，利用肌肉的拉伸和下颌的转动来驱动嘴唇运动，从而实现语音驱动唇形同步动画。

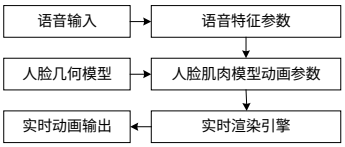

图 5　系统流程

由于语音信号是非平稳过程，而人的发音器官的肌肉运动速度较慢，因此语音信号可以认为是局部平稳或短时平稳，常用分段或分帧来处理，约为 33~100 帧/s，可采用移动的有限长度的窗口进行加权的方法实现分帧。

从麦克风输入的模拟语音信号经过 11 025Hz采样和 8bit

量化，变成数字信号，加哈明窗分帧，每帧取 256 个样本点，再提取出每帧的短时能量系数和LPC系数。第m帧的p阶LPC系数用 $a_m=[a_{1m},a_{2m},\ldots,a_{pm}]$ 表示，短时能量用l_m表示，然后，利用式 (1) 求得第 m 帧的 LPC 倒谱特征参数 $c_m=[c_{1m},c_{2m},\ldots,a_{pm}]$。

语音参数和脸部肌肉参数之间的关系可以用函数 $V_m=f(c_m,l_m)$ 表示。V 为基本唇形单位，是以肌肉模型弹性系数构成的向量，用式 $V=(m_1,m_2,\ldots,m_n,\theta)$ 表示，其中，m_i 代表第 i 块肌肉模型的收缩度；n 代表肌肉模型的数量；θ 代表下颌的转动角度。比如，不说话时，$\boldsymbol{V}=(0,0,\ldots,0,0)$。

5 仿真结果与结论

本文采用的人脸模型是由三角形组成的三维网格模型，整个模型共包括约 256 个点和 882 个三角形，如图 6 所示。系统采用一般的人脸网格模型，可通过 3DMAX、MAYA 等三维软件设计导出。

图 6　一般人脸网格模型

图 7　发“e”音时的唇形

本系统以 VC++作为开发平台，利用 OpenGL 实现。在显示真实感图形和生成动画图形时，采用了深度缓存和双缓存技术。输入“e”音的语音信号，即可获得嘴唇和下颌微张，宽度变宽的动画结果，其部分典型分解画面如图 7 示。

本文以真人语音输入实现了声音与人脸唇形动画的同步。在一般人脸网格模型上建立了与唇部运动相关肌肉的物理模型，通过分析和利用输入语音特征参数与唇形变形物理模型控制参数之间的关联性，建立了两者之间的映射关系，实现了语音和嘴唇同步动画。仿真实验结果表明，用语音驱动唇形运动变形模型的方法所生成的人脸唇形动画，能保证输入语音与唇形变形动画之间较精确的实时同步，且仿真效果也更趋自然和真实。

参考文献

1 Gudukbay U B, Ozguc U B. Realistic Speech Animation of Synthetic Faces[C]//Proc. of Conference on Computer Animation. 1998: 111-118.

2 Shan S, Gao W, Yan J, et al. Individual 3D Face Synthesis Based on Orthogonal Photos and Speech-driven Facial Animation[C]//Proc. of IEEE International Conference on Image Processing. 2000: 238-241.

3 Verma A. Using Viseme Based Acoustic Model for Speech Driven Lip System[C]//Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing. 2003.

4 陈益强, 高　文, 王兆其, 等. 基于机器学习的语音驱动人脸动画方法[J]. 软件学报, 2003, 14(2): 215-221.

5 郑　放. 人体头颈部解剖图谱[M]. 杭州: 浙江科学技术出版社, 1985.

6 Waters K. A Muscle Model for Animating Three-dimensional Facial Expression[J]. Computer Graphics, 1987, 21(4): 17-24.

7 王炳锡. 实用语音识别基础[M]. 北京: 国防工业出版社, 2004.

(上接第 229 页)

在 2 个像素值之内时，改进 ASM 的定位图像个数占的比例大；平均误差点在 3 个像素或更多时，传统 ASM 定位图像个数占的比例较大，表明改进 ASM 算法比传统 ASM 算法的定位精度高。

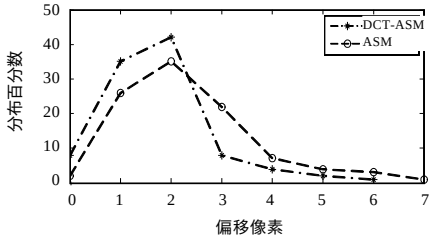

图 7　ASM 和改进 ASM 平均点误差结果比较

本文分别按照传统 ASM 方法和本文改进的 ASM 方法对这些图像做了训练和测试，改进的方法能够快速收敛到脸的轮廓，平均迭代次数不超过 5 次。而传统的方法一般要经过 8 次~10 次的迭代才可以收敛，导致处理每幅图像的时间增加。本文算法是在 CPU 为 2.0GHz、内存为 512MB 的计算机上使用 Matlab 编程实现的，改进 ASM 方法处理每幅图像的时间平均为 0.9s~2.6s，传统 ASM 方法为 1.2s~3.4s。

通过以上的评价准则、图像定位比较，认为本文改进的 ASM 方法是有效的。

4 结束语

DCT 特征在保留更多有效信息的同时，消除了不必要的冗余并采用八方向搜索匹配，大大提高了 ASM 的定位速度。实验证明本文的算法比传统方法有很大改善。

ASM 方法本身的特点使得定位与初始位置有很大关系，如果初始位置与定位目标相差很远往往导致收敛效果差，可以考虑结合面部部件的位置信息和图像边缘信息使图像初定位到人脸附近位置。

参考文献

1 Kass M, Witkin A, Terzopoulos D. Snake: Active Contour Models [C]//Proc. of the 1st International Conference on Computer Vision. 1987.

2 Kirby M, Sirovich L. Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(1): 103-108.

3 Cootes T F, Taylor C J. Locating Faces Using Statistical Feature Detectors[C]//Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition. 1996.

4 Cootes T F, Edwards G J, Taylor C J, et al. Active Appearance Models[C]//Proc. of European Conf. on Computer Vision. 1998.

5 刘　洵, 张大力. 基于 ASM 的 CT 图像序列标记点定位方法研究[J]. 计算机工程与应用, 2005, 41(13): 180-182, 213.

6 胡永利, 王　巍. 基于 ASM 模型的骨龄评价系统研究[J]. 中国图象图形学报, 2003, 8(1): 33-40.

7 Froba B, Kastner T, Zink W, et al. Real-time Active Shape Models for Face Segmentation[C]//Proc. of International Conference on Image Processing. 2001.

8 Li Yong, Zhang Changshui, Lv Xiaoguang. Face Contour Extraction with Active Shape Models Embedded Knowledge[C]//Proc. of the 5th International Conference on Signal Processing. 2000.