与语速相关的人脸语音动画合成及其评估

周 维 汪增福
(中国科学技术大学自动化系,合肥 230027)

摘 要 为了有效地合成人脸语音动画,根据对唇区肌肉收缩力与语速关系的分析,以及在对皮肤肌肉组织的粘弹性力学进行研究的基础上,提出了一种新的基于不同语速的唇动模型,并将其应用在了汉语人脸语音动画系统中。该模型根据获得的肌肉收缩力与语速的关系,并通过对皮肤肌肉组织的粘弹性分析,首先得到了语速、唇动速度与唇动位移三者之间的关系,并建立了不同语速下的唇动模型;然后通过这个唇动模型合成了不同语速状态下的具有较高自然度和个性化的人脸语音动画;最后,通过设计感知学评估实验,对合成的语音动画的效果和可理解性进行了评估。实验结果表明,该模型可以合成较高可接受性和可理解性的不同语速状态下的人脸语音动画。

关键词 语速 皮肤-肌肉组织 语音动画 感知学评估实验
中图法分类号: TP391.9 **文献标识码**: A **文章编号**: 1006-8961(2009)07-1399-07

Speech Rate Related Facial Animation Synthesis and Evaluation

ZHOU Wei, WANG Zeng-fu
(*Department of Automation, University of Science & Technology of China, Hefei* 230027)

Abstract A novel speech rate related lip movement model is proposed in this paper. The model is based on the research results on the viscoelasticity of skin-muscle tissue and the quantitative relationship between lip muscle force and speech rate. In order to show the validity of the model, we have applied it to our Chinese speech animation system. The experimental results show that our system can synthesize the individualized speech animation with high naturalness at different speech rates. Finally, the perceptual evaluation experiment is designed to evaluate the quality and intelligibility of the synthesized speech animation.

Keywords speech rate, skin-muscle tissue, speech animation, perception evaluation experiment

1 引 言

具有真实感的语音同步人脸动画是当今计算机动画领域的一个热点问题。它在人机交互、电影特效、游戏制作、视频会议、医学辅助治疗和教学辅助领域有着极高的应用价值。

从 20 世纪 90 年代起至今,人脸语音动画理论有了长足的发展。1993 年,Waters 和 Levergood 提出并制作完成了文本驱动的人脸语音动画,即将文本作为系统的输入,并将文本转换成为语音同步的人脸动画[1];1993 年,Cohen 和 Massaro 建立了协同发音模型,构造了连续语流中协同发音现象的模型[2];1997 年,Bregler 等人构建了上下文相关的三音子视素数据库,并且从数据库中选取了与语音信息最相匹配的视素数据,用来生成动画序列[3];2000 年,Kshirsagar 和 Magnenat-Thalmann 提出了语音驱动的人脸语音动画,直接将语音信息转换成人

基金项目:中国科学技术大学/中国科学院自动化研究所智能科学与技术联合实验室开放基金资助项目(JL0602)
收稿日期:2007-03-28;**改回日期**:2008-02-22

第一作者简介:周 维(1981~),男。2007 年获中国科学技术大学博士学位。主要研究方向为图像处理,人脸语音动画以及相关人脸技术研究。E-mail:zhouwei8@mail.ustc.edu.cn

脸动画参数,生成了语音同步的人脸动画[4];2003 年,Song Ming-li 等人将语音和文本同时作为输入,提出了混合驱动的人脸语音动画[5]。

通过参数调节或高质量数据库的建立,上述方法虽然在一定程度上都可以合成比较流畅自然的语音动画,但对于个性化的合成,这些方法往往无能为力。根据 Kuehn、Moll 以及 Ostry、Munhall 等人的研究[6-7],在连续语流中,语速对人说话时的嘴唇运动速度和运动幅度有显著影响。语速变化时,不同人选择不同的唇动策略,即语速增加时,一些人减小了嘴唇幅度,但是却保持了嘴唇的运动速度;一些人虽增加了嘴唇的运动速度,但保持了运动幅度;另一些人对这两种参数都进行了调节[6-7]。为了描述这一现象,在总结前人工作的基础上,本文提出了一种新的适用于不同语速下的嘴唇运动模型,即首先根据测量得到的说话时的不同语速下唇区的 EMG (Electromyography)信号来获得肌肉收缩力与语速的关系;然后,根据人体皮肤肌肉组织的粘弹性来建立嘴唇运动速度和运动幅度与肌肉收缩力之间的关系,并最终得到语速与嘴唇运动之间的关系模型;最后,将这个模型应用到本文的语音动画系统中,即可得到不同语速的人脸语音动画。

2 肌肉收缩力与语速的关系

人体肌肉运动产生肌电效应,它反映了肌肉的活动兴奋程度。通过肌电图可以很方便地探测到肌电信号。肌电信号与肌肉收缩力大小和语速快慢之间有着密切的关系[8-9]。研究表明,肌电信号幅值随肌肉收缩力的增大或者语速的加快而增大。

De Luca 对 3 种肌肉收缩力和 EMG 幅值关系进行了定量分析,并得到了归一化的"肌肉收缩力/EMG 幅值"关系曲线[8]。通过观察该关系曲线发现,可以近似认为 EMG 信号幅值随着肌肉收缩力的增加而线性递增,并且可以得到以下函数表达式:

$$F = aA + b \tag{1}$$

式中,F 为肌肉收缩力,其在测量过程中可看作是恒定的;a,b 为常数,A 为 EMG 信号幅值。

Wohlert 和 Hammen 通过测量 20 个成年人用不同语速阅读一段文字时的嘴唇 EMG 信号幅值,并通过定量分析得到了"EMG 信号幅值/语速"的关系曲线[9]。研究结果表明,语速越快,EMG 信号幅值越大,反之越小。因此可以近似认为随着语速的增加,下嘴唇的 EMG 信号幅值线性递增,而上嘴唇的 EMG 信号幅值则分段线性递增。其函数关系表达式为

$$\begin{cases} A_{\text{upper}} = cR + d, & R_1 \leqslant R < R_2 \\ A_{\text{upper}} = eR + f, & \text{如果 } R_2 \leqslant R \leqslant R_3 \\ A_{\text{lower}} = gR + h, & R_1 \leqslant R \leqslant R_3 \end{cases} \tag{2}$$

其中,A_{upper}为上唇区域的 EMG 信号幅值,A_{lower}为下唇区域的 EMG 信号幅值;c,d,e,f,g,h 为通过曲线拟合得到的常数;R 为语速;3 种阈值语速 R_1,R_2,R_3 可以从"EMG 信号幅值/语速"关系曲线中获得,并对分段线性曲线的不同区间进行了定义。将式(2)代入式(1)即可得到

$$\begin{cases} F_{\text{upper}} = C_1R + D_1 & R_1 \leqslant R < R_2 \\ F_{\text{upper}} = C_2R + D_2 & \text{如果 } R_2 \leqslant R \leqslant R_3 \\ F_{\text{lower}} = C_3R + D_3 & R_1 \leqslant R \leqslant R_3 \end{cases} \tag{3}$$

其中,$C_1 \sim C_3$,$D_1 \sim D_3$ 都为常数,F_{upper}和 F_{lower}分别是上唇区和下唇区在运动时的肌肉收缩力。由此就可以建立以下说话时唇区肌肉收缩力与语速的关系模型:$F = f(R)$。

3 不同语速下的嘴唇运动模型

通过对人体皮肤肌肉组织的生物力学性质进行研究即可知道,人体皮肤肌肉组织的拉压响应具有粘弹性性质[10]。如图 1 所示,许多线性肌肉附着在嘴唇区域,用于将骨骼与皮肤软组织连接起来,并控制着嘴唇的运动。图 2 显示了在嘴唇区域中,皮肤

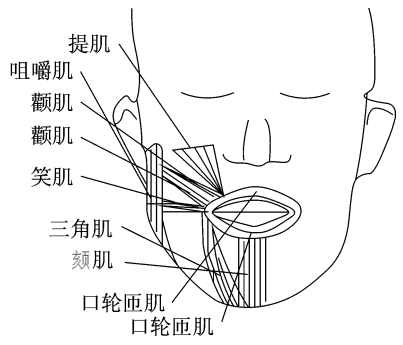

图 1 嘴唇区域中连接皮肤组织和骨骼的肌肉

Fig. 1 Muscles in the lip area connecting the bone and the lip skin tissue

与肌肉相连接点的位置。因此,嘴唇上附着并连接某一线性肌肉的区域可以看成是一个独立的粘弹性系统,该区域内任何一个质点的运动都是在某一个等效的粘弹性系统中完成的(如图 3 所示)。图 3 为粘弹性系统的示意图。

图 2　唇区皮肤与肌肉相连接点的位置

Fig. 2　The points of the skin that adhere to the muscles in the lip skin-muscle tissue area

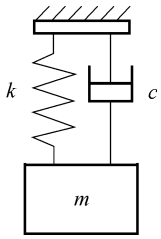

图 3　粘弹性系统模型

Fig. 3　The model of visco-elastic system

在肌肉收缩条件下,肌肉收缩力可看成是对粘弹性系统的一个阶跃激励,因此皮肤上某一质点的运动方程可写为

$$mx'' + cx' + kx = F(t) \tag{4}$$

其中,x 为质点运动位移,$F(t) = Fu(t)$ 为肌肉收缩力,c 为粘性阻尼系数,m 为质量,$u(t)$ 为阶跃函数,F 为收缩力幅值。本文假设皮肤-肌肉组织系统中,在恒定力的作用过程中,弹簧的弹性系数 k 可近似看作是不变的,并且,由于肌肉收缩是一个相对快速的过程,因此为了使问题得到简化,皮肤-肌肉组织可近似看成是一个阻尼比恒定的欠阻尼系统。系统在肌肉收缩力 $F(t)$ 作用下的阶跃响应为

$$x(t) = \frac{F}{k}\left[1 - \frac{e^{-\xi\omega_n t}}{\sqrt{1-\xi^2}}\cos(\omega_d t - \varphi)\right] \tag{5}$$

其中,$\varphi = \mathrm{tg}^{-1}\dfrac{\xi}{\sqrt{1-\xi^2}}$,振动圆频率 $\omega_d = \omega_n\sqrt{1-\xi^2}$,$\omega_n = \sqrt{\dfrac{k}{m}}$。

在欠阻尼条件下,肌肉组织的振动频率为

$$T = 2\pi\sqrt{\frac{m}{k(1-\xi^2)}} \tag{6}$$

其中,ξ 为阻尼比。在阶跃收缩力的作用下,肌肉收缩过程的总时长为

$$t_{con} = \frac{T}{4} = \frac{s}{v} \tag{7}$$

其中,s 为嘴唇皮肤-肌肉组织区域中的附着在皮肤组织上的点在肌肉收缩后相对于肌肉未收缩的自然状态的位移,v 为该点在收缩过程中的平均速度。结合式(6)、式(7),即可得到

$$k = \frac{\pi^2 m v^2}{4(1-\xi^2)s^2} \tag{8}$$

在肌肉运动过程中,某一时刻皮肤肌肉组织的刚度为该时刻肌肉收缩力与肌肉长度变化的比[11],即

$$K = \frac{F(t)}{l} \tag{9}$$

其中,l 为肌肉长度变化,可近似认为与 s 相等。

根据 Zhang 等人的研究成果[12],皮肤肌肉组织的刚度可以表示为

$$K = k(1+s^2)^{\alpha} \tag{10}$$

其中,系数 α 控制着非线性因素调节。函数 K 可以根据不同的 α 值来模拟不同的拉-压关系。

通过上述分析以及第 2 章节中得到的肌肉收缩力与语速的关系表达式可见,当 $\alpha = 0.5$ 时,v 和 s 的求解表达式为

$$v = \sqrt{\frac{4s(1-\xi^2)f(R)}{\pi^2 m\sqrt{1+s^2}}} \tag{11}$$

$$s = \sqrt{\frac{\pi^4 m^2 v^4}{16f^2(R)(1-\xi^2) - \pi^4 m^2 v^4}} \tag{12}$$

由于该模型综合了皮肤-肌肉组织粘弹性模型与“肌肉力/语速”关系,因此可以用来模拟嘴唇皮肤-肌肉组织区域中某点的位移与运动平均速度。同时,由于皮肤组织的粘性,致使影响域内的其余点的位移与速度大小由 s、v 和所在影响域内的位置所决定。

4　实验结果

根据笔者之前的工作,唇区肌肉模型和连续语流中的汉语可视化协同发音模型已经建立[13]。本文提出的唇区肌肉模型是以 Waters 肌肉模型[14]为基础,对唇区肌肉进行了较精确地建模,并通过定义

一些肌肉子模型与肌肉影响域来精确地合成各种口型,而所有口型都可由肌肉参数的调节和各种子模型的组合获得。根据汉语协同发音的规则以及与连续语流中上下文相关的三音子模型,本文构造了基于三音子模型的可视化协同发音模型,并合成了各种可能的音素-视素映射与过渡口型,进而生成了语音动画。

为了合成各种语速下具有个性化唇动语音动画,本文以先前的工作为基础,对连续语流中的动态口型视位计算模型进行了改进,即增加了语速 R 和嘴唇运动速度 v 这两个输入参数,从而得到了各种语速下的不同唇动效果。其中,R 可从语音信息中获得,v 可从视频序列中估计得到。按照如下步骤,本文获得了不同语速的嘴唇运动语音动画。

首先,唇区内某点的运动速度可以通过光流估计法分析口型视频序列来获得[15]。如图 4 所示的例子,图 4(c)表示对图 4(a)向图 4(b)过渡的光流估计,图中覆盖在唇区上的是光流向量场;然后,通过所获得的语速和 EMG 信号幅值,唇肌收缩力可以通过 $F=f(R)$ 得到;最后,由式(12)计算得到肌肉收缩时唇区肌肉上的某点相对于肌肉未收缩时的位移。肌肉影响域内其余点的位移可以通过笔者之前提出的肌肉模型得到。因此,将该模型与已有的语音动画系统相结合,就可以得到不同语速条件下的语音动画。

(a) 与/u/对应的口型　(b) 与/o/对应的口型　(c) 口型过渡的光流估计

图 4　光流估计图像

Fig. 4　Optical flow estimation images

在本文的实验中,说话人分别以正常的、快速的和慢速的 3 种语速阅读一段文字,采用改变唇动幅度并保持唇动速度的唇动策略。正常语速、快语速和慢语速分别为 190,286 和 117 个字/min。以阅读“中国科大”为例,在 3 种语速下所获得的口型视频序列如图 5 ~ 图 7 的上排图像所示,相应的合成口型如图 5 ~ 图 7 的下排图像所示。

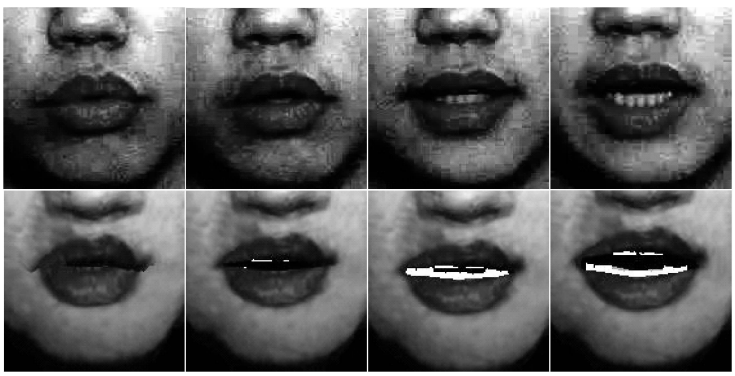

图 5　正常语速的口型序列 (190 个字/min)(上排为实拍视频序列,下排为合成的口型序列)

Fig. 5　Lip shapes sequence at habitual speech rate(190 words/min)

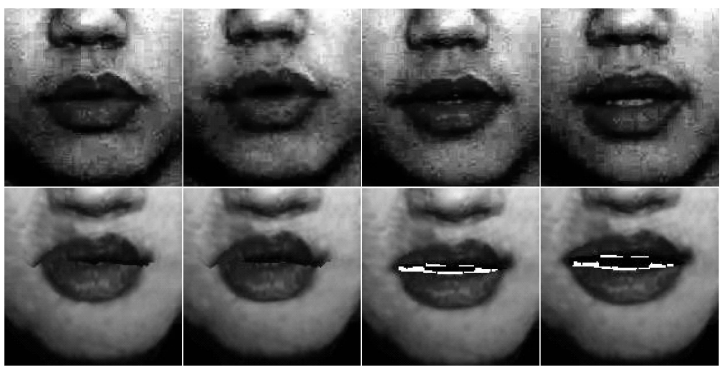

图 6　快语速的口型序列 (286 个字/min)(上排为实拍视频序列,下排为合成的口型序列)

Fig. 6　Lip shapes sequence at fast speech rate (286 words/min)

图 7　慢语速口型序列 (117 个字/min)(上排为实拍视频序列,下排为合成的口型序列)

Fig. 7　Lip shapes sequence at slow speech rate (117 words/min)

5　系统评估

对于已完成的系统,还需要对其进行感知上的评估,以确定系统合成的逼真度和有效性。对合成效果进行评估有客观和主观两种测试方法。其中客观法是运用数学方法或客观特征来对合成效果进行定量评估;主观法是利用人的感官来进行评估。目前,在合成评估中,由于客观法无法完整描述主观感受,因而主观法被广泛认可和接受;然而过分主观的测试则往往因为缺乏客观特征分析,从而使被试无法全面描述其感受。

本系统评估是根据人脸语音动画特点,采用以主观测试为基础,主观客观相结合的测试方法。通常,主观测试需要一组被试者,测试被分为合成效果测试和可理解性测试[16]。这两种测试组成了系统的整体评估测试。

合成效果测试采用 DAM(diagnostic acceptability measure)诊断可接受性方法[16]。在测试中,被试者首先观看一组配合语音的真实嘴唇运动视频,并将合成的相应嘴唇动画与之进行比较;之后,被试者对嘴唇运动各特征的合成效果以及整体效果进行评分,这些运动特征可以很好地帮助被试者完成嘴唇运动逼真度的评估,分数范围为 0 ~ 100;最后根据评分,就可以对合成效果进行评估。

对于可理解性评测,在 Li 等人构建的 CMRT(Chinese modified rhyme test)的基础上[16],再根据嘴唇运动视觉上的表现来设计实验,用于评估人们对于合成嘴唇运动的可理解程度。此实验被命名为 VCMRT(visible Chinese modified rhyme test),其分为单音节测试、双音节测试和多音节测试。实验中选取具有熟练汉语普通话背景的被试者若干人。对于未受过专业唇读训练的被试者来说,测试中音节过多是不现实的,因为被试者在测试中往往只能准确评估单音节或双音节视素的可理解性,而对于过多音节则往往会降低评估的可靠性。

通过视觉混淆树就可以将声母分为以下 7 类[17]:dtnl/bpm/f/zhchshr/jqx/zcs/gkh/(如图 8 所示),zcs 与 jqx 相差 1 个等级,由于它们在视觉上比较相似,因此混淆的概率高;zcs 与 bpm 相差 4 个等级,由于它们视觉上相差较大,故混淆概率低。

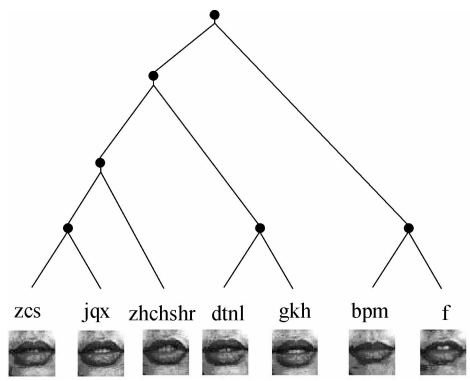

图 8　各声母之间相似度差别示意图
(黑结点联结的两分支之间相差一个等级)

Fig. 8　Differences of visual similarity between initials
(The difference between the two branches which are linked by a blue node is one level)

图 9 为 VCMRT 实验所用的部分音节对,笔者合成了与该音节对对应的口型动画。在实验中,被试者首先通过观察真实口型视频与音节对对应的合成口型动画来选出与真实口型视频对应的合成口型动画,并进行评分。通常按照对口型内容的可理解

zai/zhai	cai/chai	dai/gai	tai/kai
san/shan	nan/han	zan/ran	ban/fan
zhang/dang	chang/gang	zang/nang	cang/kang
sao/bao	zao/pao	zhao/bao	rao/pao
zuan/zun	cang/ceng	song/sen	zen/zeng
qian/qin	xian/xin	jian/juan	jian/jin
zhen/zhan	chang/chong	dian/din	nan/nuan
gan/guan	hun/hang	pen/pian	fen/feng

图 9 VCMRT 实验中部分音节对

Fig. 9 Part of syllable pairs in VCMRT

程度分为:非常好(90 ~ 100 分),不错(80 ~ 89 分),一般(60 ~ 79 分),很难理解(60 分以下)。如果被试者选择正确,则记录相应评分;如果选择错误,则得分为零,然后比较所选答案与正确答案的视觉差别,其中差别大的在计算总评估得分时,给出大权重;而差别小的则给出小权重;若完全一样,权重为 0,则从评分体系中剔除,不参与评分。

本实验选取 10 名具有普通话背景,并且未经唇读训练的学生作为评估被试者。其嘴唇运动合成效果的评估结果与可理解性评估结果如表 1 所示。

表 1 嘴唇运动合成效果评估结果与可理解性评估结果

Tab. 1 Evaluation result of lip movement synthesis quality and intelligibility

语 速	嘴唇合成效果评估(得分)						可理解性评估(得分)			
	嘴唇运动快慢逼真度	口型过渡效果	口型逼真度	嘴唇运动幅度	嘴角自然度	总体效果评分	标准差	单音节对	双音节对	三音节对
正常语速	92.9	92.5	87.6	85.6	85.3	88.8	3.30	89.2	87.4	82.5
快语速	90.2	90.2	86.3	84.8	86.4	87.6	2.21	87.7	81.2	78.8
慢语速	89.8	90.0	87.2	85.5	86.5	88.8	2.06	90.1	88.9	82.6

6 结 论

本文提出了一种新的基于不同语速的唇动模型。该模型先通过对“肌肉收缩力/EMG 幅值”与“EMG 幅值/语速”关系进行分析来得到“肌肉收缩力/语速”关系,并通过对皮肤肌肉组织的粘弹性进行研究来得到基于不同语速的唇动模型,然后再结合笔者之前所提出的肌肉模型与可视化协同发音模型,即可合成不同语速条件下的语音动画,并可对其做出评估。该方法分析了说话时语速与嘴唇运动状态之间的关系,并建立了相应模型。与以前的无语速信息模型相比,该模型可模拟不同的唇动策略,并富于个性化。

同时,本文的研究工作还有一些值得期望的地方。由于该唇动模型是基于单个相互独立的粘弹性系统的,其只关心每个线性肌肉区域的粘弹性,与真实皮肤肌肉组织多自由度弹簧网络模型并不一致,因此可能缺乏一定的细节表现力。今后可以利用弹簧网络来对唇动进行建模,但在复杂的网络面前,这依然有很大的挑战性。另外,除语速之外,音量、音调、情绪状态等也会影响说话时的唇动,对它们进行的建模还需要进一步研究,以进一步完善唇动的自然度与个性化。对系统的感知评估实验的进一步完善化也是今后的一个努力方向。

参考文献(References)

1 Waters K, Levergood T M. DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces [R]. DEC Cambridge Research Laboratory, UK, 1993.

2 Cohen M, Massaro D. Modeling coarticulation in synthetic visual speech[A]. In: Thalmann N M, Thalmann D, editors: Models and Techniques in Computer Animation [M], Tpkyo, Japan: Springer-Verlag, 1993.

3 Bregler C, Covell M, Slaney M. Video Rewrite: driving visual speech with audio[A]. In: Proceedings of SIGGRAPH 97[C], Los Angeles, CA, USA, 1997: 353-360.

4 Kshirsagar S, Magnenat-thalmann N. Lip synchronization using linear predictive analysis [A]. In: Proceedings of IEEE International Conference on Multimedia and Expo[C], New York, USA, 2000: 1077-1080.

5 Song Ming-Li, Chen C, Bu J, *et al*. 3D realistic talking face co-driven by text and speech[A]. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics[C], Washington, DC, USA, 2003, **3**: 2175-2186.

6 Kuehn D P, Moll K L. A cineradiographic study of VC and CV articulatory velocities[J]. Journal of Phonetics, 1976, **4**: 303-320.

7 Ostry D J, Munhall K G. Control of rate and duration of speech movements[J]. Journal of the Acoustical Society of America, 1985, **77**(2): 640-648.

8 De Luca C J. The use of surface electromyography in bio-mechanics [J]. Journal Application Biomechanics, 1997, **13**(2): 135-163.

9 Wohlert A B, Hammen V L. Lip muscle activity related to speech rate and loudness [J]. Journal of Speech, Language, and Hearing Research, 2000, **43**(5): 1229-1239.

10 Fung Y. Biomechanics: Mechanical Properties of Living Tissues [M]. New York, USA: Springer Verlag, 1993.

11 Basmajian J V, De Luca C J. Muscles Alive: Their Functions Revealed by Electro-myography (5th ed.)[M]. Baltimore, Maryland, USA: Williams & Wilkins, 1985.

12 Zhang Y, Prakash E C, Sung E. A new physical model with multilayer architecture for facial expression animation using dynamic adaptive mesh[J]. IEEE Transactions on Visualization and Computer Graphics, 2004, **10**(3): 339-352.

13 Zhou Wei, Wang Zeng-fu. Speech animation based on Chinese mandarin triphone model[A]. In: Proceedings of IEEE International Conference on Computer and Information Science[C], Melbourne, Australia, 2007: 924-929.

14 Waters K. A muscle model for animating three dimensional facial expression[J]. Computer and Graphics, 1987, **21**(4): 17-24.

15 Zhou W, Bao Y, Yu J, *et al*. Improved optical flow method and its application in the study on FAE[J]. Opto-Electronic Engineering, 2006, **33**(2): 9-11. [周维, 鲍远律, 於俊等. 改进的光流法及其在云爆弹研究中的应用[J]. 光电工程, 2006, **33**(2): 9-11.]

16 Li Z, Tan E C, McLoughlin I, Teo T T. Proposal of standards for intelligibility test of Chinese speech[J]. IEE Proceedings Vision, Image & Signal Processing, 2000, **147**(3): 254-260.

17 Wang Zhi-ming, Cai Lian-hong, Ai Hai-zhou. Text-to-visual speech in Chinese based on data-driven approach[J]. Journal of Software, 2005, **16**(6):1054-1063.