汉语文本-可视语音转换的研究

王志明　蔡莲红　吴志勇　陶建华

清华大学计算机系(100084)

摘要：本文通过对发音者可见器官动作的研究，从视觉方面抽取汉语发音的 26 个基本口形，并利用 MPEG-4 所规定的面部动画参数(FAP)来描述这些口形，从而获得了符合国际标准的描述汉语发音的视觉参数。另外，我们研究了这些参数在连续语流中的变化及协同发音对口形的影响，基于已有的汉语文语转换系统(Sonic)和二维网格人脸模型(PlaneFace)实现了一个汉语文本-可视语音转换系统(TTVS)。

关键词：视觉语音，面部动画参数(FAP)，文语转换系统(TTS)，文本-可视语音转换系统(TTVS)，协同发音

1.引言：

在多媒体技术迅速发展的今天，多种媒体之间的信息融合越来越受到人们的重视。对视觉语音(Visual Speech)的研究正是这样一种综合考虑语音和发音时各可视部分器官动作的多媒体技术。视觉语音是指人们在用语言交流时所表现出的看得见的各种面部动作，它能在一定程度上传达人们想要表达的思想，帮助人们加深对语言的理解。研究表明，在环境噪声较大或听话者有听力障碍的情况下，如果在给出声音信息的同时能给出一个“讲话的头”(Talking Head)，即可表现说话者面部各器官的动作的头像，则可使信噪比提高约 8-12dB。在人机交互、电子商务等应用环境中，如果人们面对的不是单纯的文本，而是一个会说话的人物形象则使人觉得计算机界面更为友善，方便人们和计算机的交流[1]。因此，如何合成生动逼真的说话者人脸图像已成为近几年来多媒体技术领域研究的一个热点。此外，对视觉语音的研究还可用于视频音频联合编码、自动唇读和帮助聋哑人学习等研究领域[2]。

在对于 TTVS(Text-To-audioVisual Speech)的研究与实现过程中，人们已往主要是集中于图像处理方面。而在建立起语音和口形的对应关系时，一般只是根据主观的判断对某一种语言中各种发音的口形作一个简单的分类，如 Tony Ezzat 将英语发音的口形分为 16 个基本类[3]、Woei-Luen Perng 将汉语发音也分为约 10 个基本类[4]，在实现 TTVS 系统时依靠这些关键帧图像采用图像变形(Warping)技术来生成中间帧图像。虽然也有人作了建立视觉语音数据库方面的工作[5][6][7]，但不同的人对视觉语音的描述各不相同，且对连续语流中协同发音对口形的影响处理较少。本文首先根据汉语发音时发音器官可见部分的动作变化，将汉语发音的可视状态分为 26 个基本类型，并从 MPEG-4 所规定的 FAP(Facial Animation Parameters)参数中选取了 28 个参数来描述这些基本类，然后利用发音时的正面和侧面图像获得相对于各种发音的相关参数。最后，基于已有的 TTS(Text-To-Speech)系统(Sonic)，我们利用这些参数和二维网格模型(PlaneFace)实现了一个汉语 TTVS 系统。

本文的第二部分介绍我们对汉语可视化的基本分类和用以描述这些类的参数，第三部分介绍了如何获取相关参数和 TTVS 系统的实现，第四部分介绍了在连续语流中协同发音的处理，最后是结束语。

2.汉语发音可视化的基本分类及其描述：

汉字发音是由声母和韵母拼接而成，其中声母有 21 个(不计零声母)、韵母 38 个。发音的部位或发音方式影响口形的形状，发音部位和发音方式相同或相似的声韵母的口形相似，如’b’、’p’、’m’的口形非常相似。因此，可以从视觉方面对声母和韵母分类进行研究，我们曾探索过如何客观、合理地从视觉上对汉语的声韵进行分类[8]。

首先，可以用近似地用同一种口形来表示发音部位和发音方式相同或相似的声韵母，也就是说在声韵母到口形的映射中存在着多对一的映射关系。另一方面，也存在着音位到口形的一对多的映射，如许多声母发音时的口形与其后面的韵母有很大的关系，处在不同的韵母前时口形不同，如’du’和’da’中的’d’口形是不同的。基于以上两种情况，我们按照声母发音部位和方式及其后接韵母分属开、齐、合、撮四呼的不同，将声母的发音口形归为 14 类，如表 1 所示。在选取例词时，考虑到了每个声母和韵母本身在汉语中出现的概率，具体数据参见[9]。

表 1　声母口形分类表

声母	b,p,m	f	d,t,n	l	g,k,h	j,q,x	zh,ch,sh,r	z,c,s
开口呼 齐齿呼	爸 (示例字)	发	大	拉	哈	机	沙	杂
合口呼 撮口呼			毒	路	姑	句	书	组

在汉语一个音节的发音中韵母占了大部分的时长，一般来说韵母有比较稳定的典型的口形。汉语中韵母有 39 个，其中ê 很少用，-i(知韵和资韵)的口形可用齐齿呼前的声母口形来表示，我们对剩余的 36 个韵母分为单一口形和复合口形两大类，单一口形的韵母可由 11 个典型的口形来表示，复合口形则由多个单一口形拼接组成。我们对韵母的分类结果如表 2 所示。

表 2　韵母口形分类表

单一口形韵母	a ,ang	阿，盎	ei,en,	诶(ei)，恩
	ai,an	哀，安	er	耳
	ao	奥	i	衣
	o	喔	u	屋
	ou	欧	ü, iong	鱼
	e,eng	鹅，鞥(eng)		
复合口形韵母	ia=i+a，ie=i+e，in=i+ei，ing=i+eng，iao=i+ao，iou=i+ou			
	ian=i+an，iang=i+ang，ua=u+a，uo=u+o，uai=u+ai			
	Uei=u+ei，uan=u+an，un=u+en，uang=u+ang，ueng(ong)=u+eng			
	üan=ü+an，üe=ü+e，ün=ü+ei			

在确定了基本口形之后，需要用一定的参数来描述它。考虑到描述的通用性和灵活性，我们采用 MPEG-4 所定义的 FAP 参数来描述基本口形。这样做不尽有利于在连续语流中对参数的修改，还有利于基于模型的音视频联合编码。MPEG-4 定义 FAP 共有 68 个参数，除去两个高层参数(Viseme, Expression)外，其余 66 个共分为 9 组，分别用以描述人脸的不同部位，如内唇及下腭、外唇、舌头、眼睛、眉毛、鼻子等部位的动作，考虑到人们发音影响到的部位，我们选取了其中 28 个参数来描述汉语中的基本口形，具体参数如表 3 所示。

表 3　用于描述汉语口形的 FAP 参数

FAP#	名称	FAP#	名称
3	open_jaw	44	raise_tongue_tip
4	lower_t_lip	45	thrust_tongue_tip
5	raise_b_midlip	46	raise_tongue
6	stretch_l_cornerlip	47	tongue_roll
7	stretch_r_conerlip	51	lower_t_lip_o
8	lower_t_lip_lm	52	raise_b_midlip_o
9	lower_t_lip_rm	53	stretch_l_cornerlip_o
10	raise_b_lip_lm	54	stretch_r_conerlip_o
11	raise_b_lip_rm	55	lower_t_lip_lm_o
12	raise_l_cornerlip	56	lower_t_lip_rm_o
13	raise_r_cornerlip	57	raise_b_lip_lm_o
14	thrust_jaw	58	raise_b_lip_rm_o
16	push_b_lip	59	raise_l_cornerlip_o
17	push_t_lip	60	raise_r_cornerlip_o

3.汉语 FAP 参数的提取及 TTVS 系统的实现：

为了测定汉语每个基本口形的相关 FAP 参数，我们在对发音人的口形进行录象时，在紧靠面部侧面放置一个与正面成 45 度角的镜子，同步地记录发音时的正面和侧面图

图 1 汉语发音 FAP 参数的测量

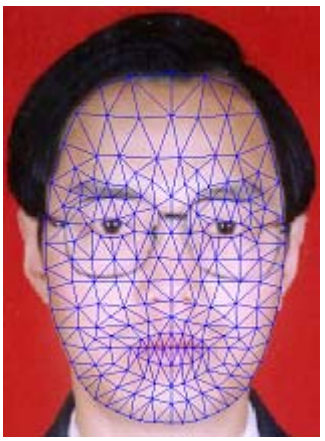

图 2 二维网格人脸模型 PlaneFace

像。然后采用一个自行开发的测量软件 FapMeasure 测出相关的 FAP 参数，如图 1 所示。

必须指出，在对 FAP 测量的过程中存在许多不稳定的因素。如不同人发音时口形的差别、测量时的误差等等。因此在测量过程中，我们首先对多个人录象测量取平均值，然后根据用这些参数合成图像的反馈信息反复调整，以获得较为理想的数值。另外，对某些参数如舌头的运动，无法直接通过录象测得。因此，我根据发音侧面的 X 光录象，并参考发音习惯确定了舌头的运动参数。表 4 是韵母'a'发音时的相关 FAP 参数值。

表 4　韵母'a'口形的 FAP 参数

FAP#	数值	FAP#	数值	FAP#	数值	FAP#	数值
3	510	10	-620	44	-400	54	-45
4	-100	11	-620	45	-100	55	-40
5	-690	12	-160	46	-300	56	-40
6	-30	13	-160	47	0	57	-520
7	-30	14	0	51	-100	58	-520
8	-90	16	-30	52	-510	59	-260
9	-90	17	0	53	-45	60	-260

在对人脸合成时，我们设计了一个由三角形组成的二维网格人脸模型(PlaneFace)，整个模型共包括约 220 个点和 350 个三角形，如图 2 所示。其中的顶点涵盖了 FDP(Facial Definition Parameters) 所定义特征点中与发音部位有关且外部可见的所有点，我们通过 FAP 参数和网格中每个点到各个 FDP 特征点距离等因素的综合考虑，获得在一定的 FAP 参数下网格中每一个点的位置，再通过对相关三角形的变形(Warping)技术来调整面部图像。对于口内的图像，我们采用固定的模型，并根据开口度的大小和上下唇的突出度来调整亮度。

结合到我们已有的 Sonic TTS 系统，由 TTS 系统提供发音的拼音码和每个音节的时长，然后按照固定的帧速率确定每一帧图像的 FAP 参数，从一张说话者的正面照片生成说话者发各种音的口形，图 3 是生成的'zh','a','u'口形。

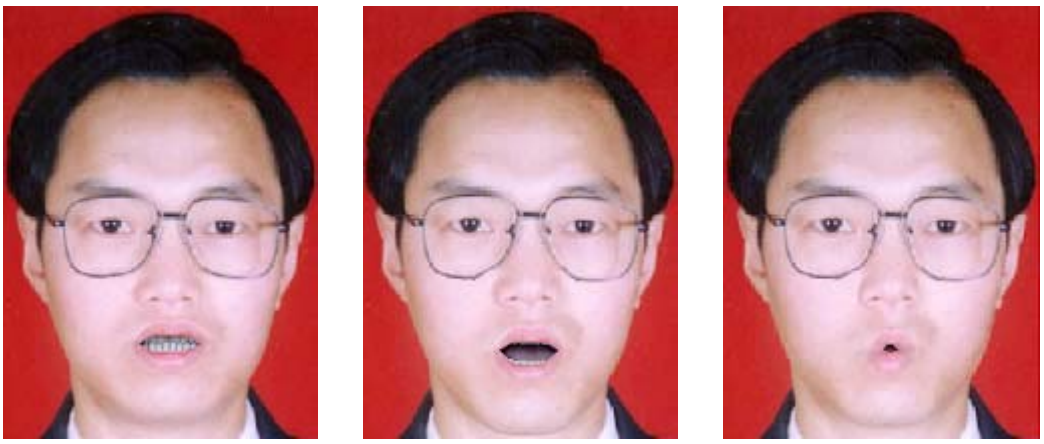

图 3 根据 FAP 参数合成的'zh'、'a'和'u'的口形

4.连续语流中的参数确定及对协同发音的考虑：

对于 FAP 参数，我们只是进行了 26 个典型口形的测量。而在人们说话过程中，口形是连续变化的，所以在合成人脸时我们需要所有中间变化口形的 FAP 参数。另一方面，各个音素单独发音时的口形与其处于连续语句中的口形是有所差别的，因为每个口形要受其前后口形的影响，即协同发音。

4.1 连续语流中的 FAP 参数确定：

当然，对于人们说话时的各种口形，我们可以设计一个发音词汇表包括所有可能的 600 多种各典型之间的互相过渡，对其发音进行录象，再从每个过渡变化的过程中取出若干幅图片测量其 FAP 参数。这样将有数千幅图片，若再考虑到多个发音人的测量，这将是一个很繁重的工作，且测得的数据还需要逐个去校正。即使这样，我们还是无法保证包含了实际说话中可能出现的所有口形。因此，我们采用计算插值的方法来获得不同典型口形之间过渡口形的 FAP 参数。插值的计算方法有许多种，最简单的线性插值，即：

$$Fap(t) = Fap(t_1) + \alpha(t - t_1) \quad \text{.....(1)}$$

其中 t，t_1，t_2 为时间，且 $t_1 \le t \le t_2$，$Fap(t)$ 为 t 时刻的 FAP 参数值，

$\alpha = (Fap(t_2) - Fap(t_1))/(t_2 - t_1)$ 为斜率。

但实验表明 FAP 参数并不是均匀地线性变化的，且线性插值在端点处斜率不连续。Jörn Ostermann 通过对多种方法的比较指出，以三阶 Hermite 函数曲线来计算插值更接近实际情况[10]，当 t1≤t≤t2 时这种方法可简写为：

$$Fap(t) = Fap(t_1) + (3\beta^2 - 2\beta^3)(Fap(t_2) - Fap(t_1)) \quad \text{.....(2)}$$

其中 $\beta = (t - t_1)/(t_2 - t_1)$

上述两种方法都是对所有参数之间的变化采用了相同的曲线来插值，没有考虑到具体 FAP 值对变化曲线的影响，而实际上 FAP 参数的变化曲线同前后特定的发音是有关系的。比如在汉语音节发音中，声母所占时长较短，韵母所占时长较长。当口形从声母向韵过渡时，将很快变化到接近韵母的口形，而从韵母向声母过渡时则较为缓慢。因此，从声母到韵母和从韵母到声母的过渡曲线应是不同的。基于这种考虑，我们采用两个二阶分段连续函数来进行中间口形参数的插值计算，即在 $0 \le \beta \le \beta_0$ 和 $\beta_0 \le \beta \le 1$ 两段分别采用不同的插值函数（β 的定义同(2)式），β_0 为一个可调节的分隔点。

设两段的函数分别为：

$$f_1(\beta) = a\beta^2 + b\beta + c \quad \text{.....(3)}$$

$$f_2(\beta) = m\beta^2 + n\beta + k \quad \text{.....(4)}$$

考虑到连续性，以上两个函数应满足以下几个约束条件：

（1） $f_1(0) = 0$ ；　　（2） $f'_1(0) = 0$ ；

（3） $f'_1(\beta_0) = f'_2(\beta_0)$ ；（4） $f_1(\beta_0) = f_2(\beta_0)$ ；

（5） $f_2(1) = 1$ ；　　（6） $f'_2(1) = 0$ ；

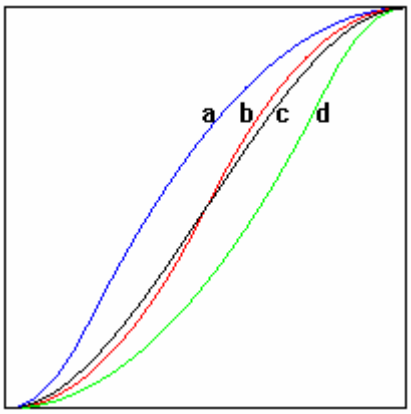

图 4　β_0=0.2,0.5,0.8 与 Hermite 曲线

通过以上约束条件，我们可以求得给定在 β_0 后的两段函数各系数。当 β_0<0.5 时，FAP 参数比(2)式更快地从第一个点值过渡到第二个点值；反之，当 β_0>0.5 时，则 FAP 参数变化比(2)式更慢，β_0=0.5 时与(2)式较为接近。因此，我们在从声母到韵母的过渡中取 β_0<0.5，而在从韵母到声母的过渡中取 β_0>0.5。图 4 是 β_0=0.2（a）、β_0=0.5（b）、β_0=0.8（d）时的曲线与 Hermite 函数曲线(c)的比较。

此外，在连续发音的过程中还有许多特殊情况要考虑，如有的韵母的口形是不表现出来的，像’de’中的’e’、’fu’中的’u’等等。

4.2 协同发音对口形的影响：

协同发音对口形有着明显的影响，我们对’du’和’da’中的’d’取不同的口形就是一种对协同发音的考虑。但仅仅考虑这些区别是远远不够的，在连续语流中，每个发音的口形都要受到其前后发音的影响，可以说每个典型口形都不如其单独发音时那样“ 到位 ”，如我们单独发’a’时的开口高度要大于在’da’中的’a’的开口高度。

我们处理协同发音对口形影响的方法是让每个发音在一定成度上向其前后的口形靠近。在连续发音的语句中，除了韵母分属不同的四呼对其前面的声母影响较大外，一般来说是前音口形对后音口形的影响更为较大，而后音对前音相对较小。因此，我们对后音对前音的影响只考虑了不同韵前声母口形的不同，对前音对后音的影响则是根据每个前音的口形对后音的口形参数作一定的调整，具体处理方法如下：

对后音的所有 FAP 参数值，增加与前音相同参数值的差值的 α 倍，即

$$Fap'(n) = Fap(n) + \alpha(Fap(n-1) - Fap(n)) \quad \text{.....(5)}$$

其中 $Fap(n)$ 为第 n 个关键帧口形的原始 FAP 值，$Fap(n-1)$ 为其前一个关键帧口形的 FAP 值，$Fap'(n)$ 为修改后第 n 个关键帧的 FAP 值，α 为调节比例因子，且 0<α<0.5。

当然，协同发音的影响还有其他许多更复杂的关系，不能一率采用上式修改，如对爆破声母 b,p,m 等的发音，总是要先将嘴合上的，所以对它们的开口度参数如 FAP4#、FAP5#、FAP8#、FAP9#、FAP10#和 FAP11#等参数就不能作修改。

综合考虑上述关键帧口形之间插补帧 FAP 值的计算和对协同发音的考虑，我们实现了 TTVS 系统中的口形连续变化，且可以实现稳定可调的图像帧速率，从而获得了较为逼真的口形合成图像。另外，我们的 TTVS 系统还有以下几个优点：(1) 用照片实现人脸合成，比单纯依靠人造模型合成的人脸更为逼真、生动；(2) 对于系统中人脸模型更换非常方便，只需通过调整二维网格模型中的关键点将相应的模型调整到新的模型上即可；(3) 可针对某个人的发音特点，调整各个口形的 FAP 参数实现具有个性特色的口形模拟。

5.结束语：

基于语音与图像信息之间的内在联系，本文对汉语发音可视化作了初步的探讨。首先对汉语声韵母发音的口形分类和参数化作了研究，然后基于分类和参数化的方法实现了汉语发音口形的标准化和连续语流中的口形模拟，其中主要讨论了连续语流中口形的变化和协同发音对口形的影响。最后，基于我们已有的 Sonic TTS 系统实现了一个 TTVS 系统，增强了合成语音的可理解性和用户界面的友善性。

从我们的工作中来看，对连续语流中口形变化和协同发音对口形影响的处理是模拟口形是否逼真的关键，对这一部分需要更为深入的研究。另外，如何根据 FAP 参数使二维图像中呈现出更好的三维视觉效果也是一个有进一步研究的问题。

参考文献:

[1] Ostermann, J. and Millen, D., Talking heads and synthetic speech: an architecture for supporting electronic commerce, 2000 IEEE International Conference on Multimedia and Expo, Volume: 1 , 2000, P71 –74.

[2] Brooke,N.M., Scott,S.D. and Tomlinson,M.J, Making talking heads and speechreading with computers, IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, 1996 , P2/1-2/6.

[3] Tony Ezzat and Tomaso Poggio, MikeTalk: A Talking Facial Display Based on Morphing Visemes. Appears in Proceedings of the Computer Animation Conference, Philadelphia, Pennsylvania, June, 1998.

[4] Woei-Luen Perng; Yungkang Wu; Ming Ouhyoung, Image Talk: a real time synthetic talking head using one single image with Chinese text-to-speech capability, Sixth Pacific Conference on Computer Graphics and Applications, 1998, P140 -148.

[5] C.C. Chibelushi, F. Deravi and J.S.D. Mason, Survey of Audio Visual Speech Databases, http://mambo.ucsc.edu/psl/ccampbel/survey.txt, 1996.

[6] Tobias Ohman, An audio-visual speech database and automatic measurements of visual speech,

TMH-QPSR 1-2, P61-76, 1998.

[7] Shah,D. and Marshall,S., An image/speech relational database and its application, IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication (Digest No: 1996/213) , 1996 , P 6/1 -6/5.

[8] 王志明、蔡莲红，汉语音节与口形关系的研究，第九届全国多媒体技术学术会议(NCMT'2000)，2000 年 10 月。

[9] 陈永彬、王仁华，《语音信号处理》，中国科技大学出版社，1990，第 54 页。

[10] Jörn Ostermann, Yao Wang Bookmarks for TTS-FBA Synchronization, ISO/IEC JTC1/SC29/WG11, MPEG 98/3823, June 1998.

Study of Text to Visual Speech in Chinese

Wang Zhiming Cai LianHong Wu Zhiyong Tao Jianhua
Dept. of Computer Science and Technology
Tsinghua University, Beijing, China

Abstract: After study the motion of visual organ of the speaker, we divided Chinese phonemes into 26 basic visual classes. We described these basic classes by FAPs defined by MPEG-4, and then we got the universal parameters of Chinese phonemes. We also study the modification of these parameters in successive speech and coarticulation circumstance. Base on our TTS system (Sonic) and image warping technology on our 2-D mesh model (PlaneFace), we realized a TTVS sytem.

Keyword: Visual Speech, Facial Animation Parameter(FAP), Text-To-Speech(TTS), Text-To_audioVisual Speech(TTVS), Coarticulation