汉语语音视位的研究[1]

王志明　蔡莲红

清华大学计算机系(100084)

摘要：MPEG-4 首次作为国际标准正式定义了视位（Viseme）的概念，它是指与某一音位相对应的嘴、舌头、下腭等可视发音器官所处的状态。本文通过对汉语发音时各可见部分器官动作和发音规则的研究，将汉语发音分为 28 个基本静态视位。利用语音信息从 AVI 文件中自动抽取出这些基本视位图像，从 MPEG-4 所规定的 68 个面部动画参数（FAP）中提取出 28 个来描述这些口形，并实现了部分 FAP 参数的自动测量。最后，我们给出一个视位研究应用的实例。

关键词：视位，面部动画参数，文本-语音转换系统，文本-可视语音转换系统

1.引言：

人类对语言的理解是多模态的，即人们在相互交谈时，不仅听声音，而且用眼睛去观察说话人的面部表情。人们说话时复杂多变的面部表情不仅可以传达丰富的感情，而且可以增强对语言的理解。有的声音在听觉上是很容易混淆(如/bi/和/di/)，但因为它们在发音时口形有较大的差别，如果观察说话者的口形就很容易把它们区分开来。因此，人们在许多方面研究如何利用这种多媒体之间的交互作用。如依靠人工合成的虚拟人脸去提高人们在环境噪声较大的情况下对语音的理解；利用视觉信息辅助的双模态语音识别提高语音识别的识别率；利用人脸表情与语音的关系提高多媒体数据的压缩率，等等。

随着人们研究的不断深入和许多实际应用的驱动，新的国际标准 MPEG-4 提出了视位（Viseme）的概念，它由英文的 Visual 和 Phoneme 两词拼接而成。MPEG-4 对视位的定义是：Viseme is the physical (visual) configuration of the mouth, tongue and jaw that is visually correlated with the speech sound corresponding to a phoneme，即视位是指与某一音位相对应的嘴、舌头、下腭等可视发音器官所处的状态[1]。现在的 MPEG 标准仅定义了静态视位（Static Viseme），但同时也指出不排除将来定义其它类型的视位。为叙述方便，我们将嘴、舌头、下腭等可视发音器官所处的状态简称为口形。

音位是与某一特定的语言密切相关的，因而视位也是与语言相关的。虽然 MPEG-4 把国际音标的发音分为 15 个静态视位，但考虑到各种语言的发音特点和不同的音位组成，各国学者对不同语言的发音口形作了很多研究，现今多限于静态视位。如 Bothe 将德语发音口形分为 12 个静态视位[2]、Le Goff 将法语发音口形分为 19 个静态视位[3]、Ezzat 将英语发音口形分为 16 个静态视位[4]、Lande 将意大利语发音口形分为 23 个静态视位[5]等等。也有人提出了一些动态视位的雏形，如用[2]中用 0-4 帧图片表示一个音位的口形，[6]和[7]用小段原始图像序列合成新的图像序列，但还没有正式提出动态视位(Dynamic Viseme)这个概念，更没有上升到参数化和模型化的高度。国内对于汉语视位的研究相对较少，晏洁将汉语分为 6 个基本口形[8]，实现文本驱动的唇形合成，但分类过于简单；钟晓等人将汉语口形分为个 12 类(包括 10 个基本口形和两个过渡口形)[9]，研究基本口形的识别，但分类时显然没有考虑到汉语的发音方式和舌位等口内器官的差别。

本文首先在分析了汉语发音拼音结构的特点和发音口形后，将汉语发音口形分为 28 个基本的静态视位(第 2 部分)。对于视位的量化描述，我们采用了 MPEG-4 定义的面部动画参数 FAP(Facial Animation Parameter)；对静态视位原始数据的获取，采用了一种基于语音信息指导的自动视位抽取方法，并利用图像跟踪技术实现了部分参数值的自动提取（第 3 部分)。我们讨论了连续语流中视位变体的问题，给出了一个视位研究应用的实例（第 4 部分)，最后是我们工作的总结(第 5 部分)。

[1]本文受到国家教育部高等学校博士学科点专项科研基金(20010003049)资助。

2.汉语静态视位

每种语言均具有特定的音位集和特有的发音特点，所以不同语言的视位并不能完全共用。如汉语的/a/可能在不同的环境中对应不同的国际音标[A](啊)和[a](派)，而 MPEG-4 所定义的 15 个静态视位中没有与汉语中/o/口形对应的视位。根据汉语发音的特点和音位组成，我们对汉语的发音口形建立了一个基本静态视位集。

从音位的角度来考虑，汉语音位有 32 个，包括 22 个辅音音位和 10 个元音音位；从汉语发音的基本组成单位来考虑，可以分为 21 个声母和 38 个韵母，其中韵母又可分为单韵母和复合韵母。针对汉语的发音习惯，我们从声韵母的角度来研究口形的分类。每个声母对应一个辅音音位，但音位到口形的映射中存在着多对一的关系，如/b/、/p/、/m/ 音位的发音口形非常相似。另一方面，由于协同发音的影响，同一个声母在不同的拼音组合中口形可能发生变化，如/du/和/da/中的/d/口形不同，因此音位到口形的映射也不是一一对应。按发音部位，声母可分为双唇音、唇齿音、舌尖中音、舌尖边音、舌根音、舌面音、舌尖后音、舌尖前音。我们研究了声母发音口形与发音部位的关系，分析了大量有关的发音图像数据，对声母的分类如表 1 所示。

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

汉语韵母中的ê 很少用，对剩余的 38 个韵母可分为单韵母和复合韵母，单韵母的发音口形比较稳定，彼此之间有较大的差别，所以每个单韵母的发音口形可以作为一个静态视位，包括/a/、/o/、/e/、/i/、/u/、/ü/、/-i/(知)、/-i/(资)和/er/。其中/o/的口形是一个特例，指得出/o/音发出后稳定的口形，唇形成圆形，开口度适中。实际发音我们发/o/音的口形是先作出/u/的口形，再过渡到/o/的口形。

复合韵母由多个音位组成，口形由一个口形过渡到另一个口形。但有些复合韵母组合较为紧密，口形难以再拆分为多个口形的组合，如汉语中前响二合元音韵母的口形接近第一个元音的口形，可以作为单一口形来考虑，包括/ai/、/ei/、/ao/和/ou/。另外，汉语鼻韵尾-n 和-ng 处在开口呼音位后时对口形的影响较小，-n 较-ng 舌位靠前、开口幅度略小，从而使/an/与/ai/相似、/en/与/ei/相似，也可以把它们作为单一口形来考虑。通过以上的分析，我们将韵母的基本口形分为 13 个静态视位，其它复合韵母的口形则由多个单韵母视位组合而成，如表 2 所示。

表 2　韵母口形分类表

单一口形韵母	a ,ang	阿，盎	er	耳
	ai,an	哀，安	i	衣
	ao	奥	u	屋
	o	喔	ü	鱼
	ou	欧	-i(知韵)	知
	e,eng	鹅，鞥(eng)	-i(资韵)	资
	ei,en,	诶(ei)，恩		
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,iong=ü+eng			

加上无声自然状态的口形(以 NA 表示)，我们将汉语总共分成 28 个基本静态视位。

3.静态视位的定量描述

以往语音学对发音口形的描述只是定性的，如圆唇、扁唇、开口的大小，舌位的高低等等。为了能更好的从客观上对汉语的视位给以度量，需要一个对视位定量的描述，现在许多应用领域也需要有对视位进行客观上的定量的度量，如虚拟人脸合成、机器自动唇读等等。在对视位的量化度量中，我们选择了 MPEG-4 所定义的 FAP(Facial Animation Parameter)参数。

MPEG-4 是经各国专家总结多年的研究成果制定的，它对人脸的模型和参数化描述作出了详细的定义。MPEG-4 定义了一组面部定义参数 FDP(Facial Definition Parameter)，用来定义一个人脸模型，包括 84 个特征点，对其中部分特征点位移的定义形成了一组 FAP 参数。图 1 是嘴唇和下腭区的特征点，实心点是影响 FAP 参数的点。FAP 共有 68 个参数，包括 2 个高级参数和 66 个低级参数，其中 28 个与发音有直接的关系，如表 3 所示[2]。

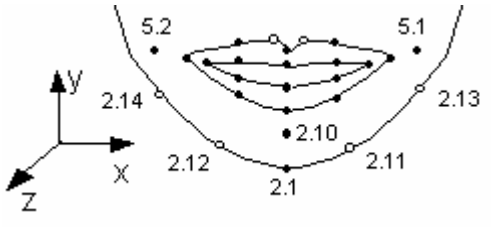

图 1 嘴唇和下腭区的 FDP 特征点

表 3 用于描述汉语口形的 FAP 参数

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

在这些参数中，FAP3 和 FAP14 定义了下腭的上下和前后移动量；FAP4，FAP5 和 FAP8~13 定义了内唇边缘 8 个点的上下移动量；FAP6、FAP7 和 FAP53、FAP54 分别定义了内外唇角的水平方向的位移；FAP51，FAP52 和 FAP55~60 定义了外唇边缘 8 个点的上下移动量；FAP16 和 FAP17 分别定义了下唇和上唇的突出度；定义了下腭的上下和前后移动量；FAP44~47 定义了舌尖和舌体的上下、前后位移及舌头弯成‘U’形的成度。

为了获得静态视位参数的量化值，有人利用从镜子中观察人的发音动作来调节其人脸模型，并以此获得视位的参数值[10]。但这种方法受主观因素的影响较大。我们首先设计发音文本进行录像，再从图像序列中提取相应的视位图片，进而测量其相应的 FAP 参数。

对于静态视位图像的抽取，多是采用手工从图像序列中提取相应的图片。但对手工选择视位的准确定与特定人的理解有一定的关系，不同的人可能选不同的图像作为某一音位的视位。另一方面，当所要选取的视位量较大时，这也是一个很繁重枯燥的工作。Jie Yang 等人提出一种自动寻找最佳视位的方法[11]，先利用语音识别确定每个音位的时间段及图像的起始帧和终止帧，而后以每个图像帧作为参考来作线性插值得到某个音位的其它帧图像，并将合成后图像与原始图像比较，计算总的像素值平方误差和，取误差和最小的一帧图像作为相应音位的视位。但在 Jie Yang 等人的方法中，首先其所依据的线性假设和其逐像素的评价标准有待进一步的讨论，其次对于汉语声母视位的选取显然是不合适。

通过对汉语发音特点的研究，我们提出一种快速有效的视位图片自动抽取方法。汉语韵发音在整个音节中所占时间较长，口形在发音的中间呈现出稳定的口形。图 2 中是发单韵母时语音短时能量与开口高度的关系。图中实线为语音的短时能量，虚线为外唇高度。从图中可以看出，在语音的中间时刻附近口形比较稳定，所以我们抽取韵母视位所选取的时间坐标为切分出的语音时长的中央，如图中竖线所示。

当发一个包括声母和韵母的汉语音节时，其声母的口形在声音发出前已形成，发音时向韵母的口形过渡，所以应取语音出现之时的口形，如图 3 中竖线所示。

对于视位参数值的测量，我们采用类似于 Rui Wang 所介绍的唇形跟踪技术，用一个由两条四次曲线组成的变形模板来表示外唇边缘轮廓，两条抛物线来表示内唇轮廓。对嘴唇附近的唇区和面部图像在 rgb 空间进行聚类，进而计算使其 Fisher 投影向量，以 Fisher 变换后的标量梯度作为变形模板搜索的能量函数，实现了嘴唇外轮廓的跟踪[12]。因为内唇的准确跟踪较为困难，对于内唇的高度和宽度近似为外唇参数的线性组合。另外，录像时在人脸上加上一些标志点，作为参考点和某些参数的测量点；面部某些 FDP 特征点在人脸上并没有明显的特征，加上标志点有利于准确跟踪和测量。通过对内外唇轮廓和这些标志点的跟踪，我们可以计算出外唇区和下腭部分的上下和左右位移，从而计算出 21 个与内外唇轮廓和下腭位移相关二维 FAP 参数值。图 4 是对唇形跟踪的一些实验结果，包括正面图中的内、外轮廓，侧面图中的上下唇突出度、下腭突出度，以及下腭张开度。图中两幅图片分别对应于自动抽取的视位/a/和/b/的口形。

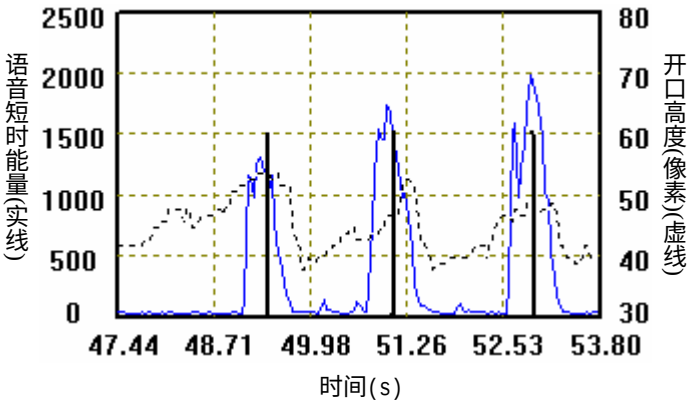

图 2 韵母视位抽取时刻示意图

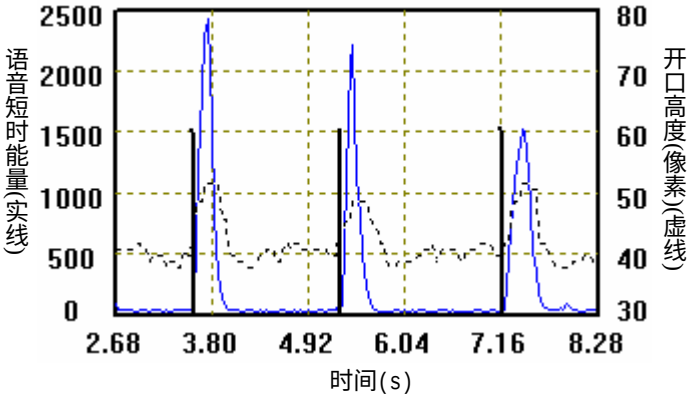

图 3 声母视位抽取时刻示意图

为了测量 FAP 中的三维参数值，如上下唇和下腭的突出度，我们在对发音人的口形进行录像时，在紧靠面部侧面放置一个镜子，同步地记录发音时的正面和侧面图像，再利用鼠标手动测出相关的 FAP 参数。由于实验条件的限制，我们现在可以测量 FAP 参数仅限于外观看得见的 24 个参数，也就是说，不包括与舌头运动有关的四个参数(FAP44~47)。对于这些参数，我们根据发音侧面的 X 光录像，并参考发音习惯来确定，如何获取这些参数的准确值有待进一步的研究。

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像测量取平均值，然后根据用这些参数合成图像的反馈信息反复调整，以获得汉语静态视位较为理想的 FAP 参数值。

图 4 口形自动跟踪的结果及自动抽取的静态视位/b/和/a/

4.视位变体及视位的一项应用

我们知道，同一个音位在不同的环境中或由不同的人发音时不尽相同，这就是音位变体。同理，人们说话时所表现的视位在不同环境下也是不相同的，我们称之为视位变体。这主要来自协同发音对口 形的影响，如果不对它进行有效的处理，将会严重影响合成口形的质量。对于连续语流中协同发音的影响，我们定义了一个协同发音影响因子，用来控制每个视位受其前后视位影响的程度，通过这个因子调整每个视位在实际应用环境中的参数值，以获得不同的视位变体。

在人们正常发音的过程中，口形在不停的变化，是一个动态的过程。要合成动态连续的说话者口形，仅有静态视位是不够的，必须得到各个静态视位之间过渡帧口形。为此，我们定义了一个二阶分段值函数来计算视位间过渡帧口形的参数，通过对分段函数切换点位置的调整可以实现不同速率的参数过渡，适应汉语声母到韵母和韵母到声母的口形过渡。

作为对汉语视位研究的一个直接应用，我们制作了一个汉语文本-可视语音合成(Text-to-visual speech)系统。合成中汉语拼音序列及每个音节的时长等信息来源于我们已有的汉语文本语音合成(Text-to-speech)系统[13]。

基于我们建立的汉语基本静态视位参数库和已有的汉语 TTS 系统，并经过到对视位变体和中间口形插值的处理，我们实现了一个帧速率可调的汉语 TTVS 系统。

5.总结：

现在有越来越多的环境要求了解语音与图像之间的关系，对人们说话口形的准确的、定量的描述是一个非常复杂的问题。我们首先通过对汉语发音规则和发音特点的研究，得出了一组基本的汉语静态视位集。进而从客观性和国际通用性出发，选择了 MPEG-4 定义的 FAP 参数作为对视位的定量描述参数。对于视位的获取，结合汉语的发音特点，提出一种基于语音信息的自动提取方法，并已利用图像跟踪技术实现了视位部分参数值的自动测量。最后，我们给出一个视位研究应用的实际例子，并讨论了视位变体的问题。

参考文献:

[1] International standard, Information technology-Coding of audio-visual objects-Part 2: Visual; Admendment 1: Visual extensions, ISO/IEC 14496-2: 1999/Amd.1:2000(E).

[2] Bothe, H.H. and Wieden, E.A., A neurofuzzy approach for modeling lips movements, IEEE World Congress on Fuzzy Systems, 1994. Proceedings of the Third IEEE Conference on Computational Intelligence, Vol1, P234 –237, 1994.

[3] Le Goff, B. and Benoit, C., A text-to-audiovisual-speech synthesizer for French, Proceedings., Fourth International Conference on Spoken Language, ICSLP 96, Vol4, P2163 –2166, 1996.

[4] 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.

[5] Lande, C. and Francini, G., An MPEG-4 facial animation system driven by synthetic speech, Multimedia Modeling (MMM '98), P203 –212, 1998.

[6] Bregler. C, Covell. M. and Slaney. M., Video Rewrite: Driving visual speech with audio, Proc DIGGRAPH97, p353-360, ACM SIGGRAPH, July 1997.

[7] Cosatto, E., Potamianos, G. and Graf, H.P., Audio-visual unit selection for the synthesis of photo-realistic talking-heads, 2000 IEEE International Conference on Multimedia and Expo, ICME 2000, Vol2, P619 –622.

[8] 晏洁，文本驱动的唇动合成系统，计算机工程与设计，第 19 卷，第 1 期，第 31-34 页，1998 年 2 月。

[9] 钟晓、周昌乐、俞瑞钊，一种面向汉语语音识别的口形形状识别方法，软件学报，第 10 卷，第 2 期，第 205-209 页，1999 年 2 月。

[10] Olives, J.-L., Sams, M., Kulju, J., Seppala, O., Karjalainen, M., Altosaar, T., Lemmetty, S., Toyra, K. and Vainio, M., Towards a high quality Finnish talking head, 1999 IEEE 3rd Workshop on Multimedia Signal Processing, P433 –437, 1999.

[11] Yang, J., Xiao, J. and Ritter, M., Automatic selection of visemes for image-based visual speech synthesis, 2000 IEEE International Conference on Multimedia and Expo, ICME 2000, Vol2, P1081 –1084.

[12] Rui Wang, Wen Gao and Jiyong Ma, An approach of robust and fast locating lip motion, ICMI2000.

[13] 王志明、蔡莲红等，汉语文本-可视语音转换的研究，小型微型计算机系统(已录用)。

Study of Chinese Viseme

Wang Zhiming, Cai LianHong

Dep. of Computer Sci. and Tech. of Tsinghua University, Beijing, China

Abstract: MPEG-4 gives the definition of viseme as the physical (visual) configuration of the mouth, tongue and jaw that is visually correlated with the speech sound corresponding to a phoneme. Based on the study of the visual articulators movement in uttering Chinese speech and of the pronunciation rules, we define 28 basic static visemes of Chinese. We describe these visemes in term of 28 of the total of 68 MPEG-4 FAPs, extract these visemes automatically from AVI files based on speech information, and measured partial FAP values by automatically tracking the mouth contour and some marked points. Finally, we give an example of usage of these viseme.

Keyword: Viseme, Facial animation parameter, Text-to-speech, Text-to-visual speech