动态视位模型及其参数估计*

王志明+, 蔡莲红

(清华大学 计算机科学与技术系,北京 100084)

A Dynamic Viseme Model and Parameter Estimation

WANG Zhi-Ming+, CAI Lian-Hong

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: 86-10-62771587, E-mail: wzm00@mails.tsinghua.edu.cn

http://www.cs.tsinghua.edu.cn

Received 2001-09-17; Accepted 2002-02-26

Wang ZM, Cai LH. A dynamic viseme model and parameter estimation. *Journal of Software*, 2003,14(3): 461~466.

Abstract: Visual information can improve speech perception. But how to synthesis the realistic mouth shape is a complex problem. After studying the rule of lip movement in speaking, a dominance blending dynamic viseme model for visual speech synthesis is proposed in this paper. Furthermore, considering the characteristic of Chinese speech, a systemic learning method is given to learn the model parameters from training data, which is more reliable than desire parameters according to subjective experience. Experimental results show that the dynamic viseme model and learning method are effective.

Key words: visual speech; viseme; static viseme; dynamic viseme; co-articulation

摘 要: 视觉信息可以加强人们对语音的理解,但如何在可视语音合成中生成逼真自然的口形是个复杂的问题.在深入地研究了人们说话过程中口形变化的规律后,提出了一个基于控制函数混合的动态语音视位模型.并针对汉语发音的特点给出了一种系统的从训练数据学习模型参数的方法,这比依靠主观经验人为指定模型参数更为可靠.实验结果表明,视位模型和通过训练数据学习得到的模型参数可以有效地描述汉语发音过程中口形的变化过程.

关键词: 可视语音;视位;静态视位;动态视位;协同发音

中图法分类号: TP391 **文献标识码**: A

视频信息和音频信息是人们感知外界最主要的两个信息来源,视觉信息和听觉信息的结合比任何单一信息能传达更多的信息.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,这也正是可视语音合成的意义所在.但另一方面,“McGurk 效应”又指出:当人们面对互相冲突的音频和视频信息刺激时,人们所理解的信息既不符合音频也不符合视频.比如,在给出/ba/的声音的同时,给出一个/ga/的口形,

* Supported by the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20010003049 (国家教育部博士点基金)

第一作者简介: 王志明(1968－),男,山西运城人,博士生,工程师,主要研究领域为语音可视化,多模态语音合成.

则人们会理解为/da/.这说明,如果在可视语音(visual speech)合成时合成的口形错误或不准确,将会对人们理解语音产生误导.因此,我们必须准确地了解人们说话时口形的变化规律,建立有效的可视语音描述模型,以达到高质量的可视语音合成.

许多学者在可视语音合成的过程中提出了各种各样处理口形变化的方法,包括建立参数控制的协同发音模型[1,2]、建立基于统计模型(如隐马尔可夫模型)的协同发音模型[3,4]和无参数的数据驱动方法[5~7]等等.参数控制模型是人们最早研究也是研究最多的一种处理连续语流中协同发音对口形影响的方法,如早期的前向预测模型(look-ahead model)、时间锁定模型(time_locked model)、混合模型(hybrid mode)以及后来加洲大学 Cohen 和 Massaro 提出的基于指数控制函数的协同发音模型[1].但如何使所建立的数学模型更符合实际发音过程中口形的规律,针对某一模型如何得到模型中的各个参数值仍没有一个公认的解决方法,还有待于进一步深入地研究.

1 语音视位与协同发音

随着可视语音的发展,MPEG-4 标准定义了语音视位(viseme),用来描述人们发某一音位时对应的可视发音器官所处的物理状态[8].现在的国际标准仅定义了静态视位(static viseme),但人们发某一音位时口形处在一个动态的、连续变化的过程中,仅用一幅静态图像来描述往往是不够的.我们将一个音位发音时口形从产生到消失的完整变化过程称为动态视位(dynamic viseme),它可以为人们理解语音提供更多的信息.

在连续语流中,协同发音会对口形产生很大的影响.可视语音中的协同发音是指连续说话过程中每个音位的发音口形受到其周围音位口形的影响,同时也在影响着周围其他音位的口形.这种相互作用的结果使得连续语流中每个音位的口形都不同于其单独发音时的口形,最终形成一个复杂的口形变化过程.

一般来说,元音的口形对辅音影响较大,在汉语中主要表现为韵母的口形对声母影响较大,如声母‘d’在/du/中和/de/中的口形差别很大,主要是受到其后韵母‘u’和‘e’的影响;但也有一些声母的口形几乎不受周围其他音位的口形的影响,如‘b’,‘p’,‘m’,人们在发这几个音时上下唇必须合上.也正因为它们受周围环境影响较小,所以人们在感觉语音与口形的同步时主要依靠对这些音的口形的辨识.

以往各种各样的可视语音模型主要着眼于协同发音,即各口形之间的相互影响.我们认为,应首先将每个音位的发音口形视为一个完整的过程,再来考虑相互影响,即应该首先建立动态视位模型.

2 控制函数混合动态语音视位模型(DB-DVM)

在建立动态语音视位模型时,我们应考虑以下几点:(1) 模型可以模拟出人们发某一音位时完整的口形参数变化曲线;(2) 模型可以适应不同语速的口形变化;(3) 可以方便地应用于可视语音合成.

通过对大量发音录像的观察、跟踪处理,并总结在可视语音合成过程中的经验,我们在 Cohen&Massaro 的协同发音模型的基础上,发展并提出一种对口形变化描述更为完整的语音视位模型,即基于指数控制函数的动态语音视位模型(dominance blending dynamic viseme model,简称 DB-DVM).每个动态语音视位的某一口形参数由一个基本控制函数与前后两个无声模型的控制函数来决定,整个动态视位的参数变化过程由静态视位参数值及自然无声状态下的参数按这 3 个控制函数加权形成.每个语音视位参数的基本控制函数随时间按指数衰减,可表示如下:设 D_{sp} 表示第 s 个视位的第 p 个口形参数的控制函数值,则有

$$D_{sp}=\alpha_{sp}\mathrm{e}^{-\theta_{\leftarrow sp}|\tau|^{c}},\quad \text{if } \tau\geq 0, \tag{1}$$

$$D_{sp}=\alpha_{sp}\mathrm{e}^{-\theta_{\rightarrow sp}|\tau|^{c}},\quad \text{if } \tau< 0. \tag{2}$$

其中 α_{sp}, $\theta_{\leftarrow sp}$ 和 $\theta_{\rightarrow sp}$ 是与 s 和 p 相关的正系数,α_{sp} 表示峰值处的控制函数值,此值越大,表示该视位的口形在连续语流中越不易受到其周围口形的影响;$\theta_{\leftarrow sp}$ 和 $\theta_{\rightarrow sp}$ 分别代表了向前和向后控制函数衰减的快慢程度,值越大则衰减得越快,表示该音位的口形消失得越快;c 为一常数,$\tau=t_{cs}+t_{osp}-t$,t_{cs} 为第 s 个音位的语音时间中心,t_{osp} 为从语音中心到 p 参数峰值点的距离,t 为当前时刻,即τ代表了当前时间到控制函数中心时刻的距离(ms).

无声模型用于表示无声到有声和有声到无声的口形转变.从无声到有声(左无声模型)的控制函数可表示如下:

$$D_{lp}=\alpha_{lp}\mathrm{e}^{\mathrm{sgn}(\tau)\theta_{lp}|\tau|^{c}}. \tag{3}$$

$\tau=t_s+t_l-t$,其中 t_s 表示后续语音起始时刻,t_l 表示从语音起始时刻到左无声模型中心的时间;从有声到无声(右无声模型)的控制函数为

$$D_{rp}=\alpha_{rp}\mathrm{e}^{-\mathrm{sgn}(\tau)\theta_{rp}|\tau|^{c}}. \tag{4}$$

$\tau=t_e+t_r-t$,其中 t_e 表示其前一语音结束时刻,t_r 表示从语音结束时刻到右无声模型中心的时间.

最终任意时刻的动态视位参数由这三者按其控制函数值加权构成:

$$F_{sp}(t)=\frac{D_{sp}(t)\times T_{sp}+(D_{lp}(t)+D_{rp}(t))\times T_{0p}}{D_{sp}(t)+D_{lp}(t)+D_{rp}(t)}. \tag{5}$$

其中 T_{0p} 为自然状态下 p 参数值.控制函数及参数变化过程如图 1 所示,图中上边显示各个控制函数值随时间变化的过程,下窗口显示在各个控制函数作用下的视位参数随时间变化的过程.无声模型中的控制函数权值随时间单调变化,使口形最终回到自然状态.

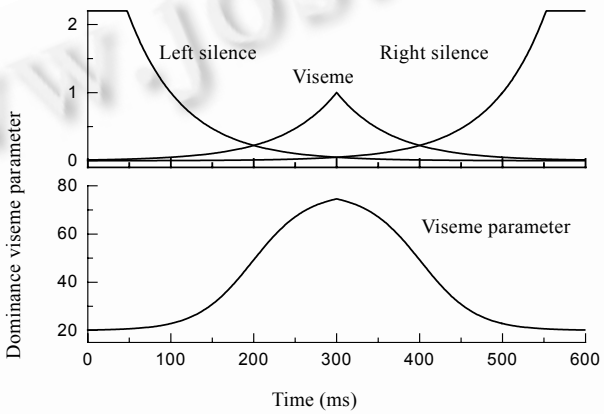


Fig.1 Domiance and viseme parameter of dynamic viseme model

图1 动态视位模型的控制函数和视位参数

从图中可以看出,模型包括了视位参数从有到无的整个变化过程,并可根据语音的时长调节参数的变化过程.

3 连续语流中的参数生成

在连续语流中,各个视位的口形受到其周围视位的影响,最终的视位参数可以由各个视位参数按基本控制函数值加权得到,即

$$F_p(t)=\left(\sum_{s=1}^{N}(D_{sp}(t)\times T_{sp})\right)\Big/\left(\sum_{s=1}^{N}D_{sp}(t)\right), \tag{6}$$

其中的s包含了无声模型,且相邻视位之间的无声模型由左、右无声模型相交构成,随时间间隔的不同,无声模型作用的时间和幅度也不相同.当两个音位的发音相连时,相连处的无声模型自动消失.从而可以模拟人们说话过程中不同停顿时长情况下口形变化的过程,如图2所示.

对于一些特殊视位参数,如受周围环境影响较小、对同步感明显的部分音位,如‘b’,‘p’,‘m’的上下唇高度参数,我们定义一种特殊的混合规则:在这些音位中心值的某一固定时间范围内,它们的控制函数值与所有其他音位控制函数值总和的比例不断增大,直到中心处使它们的权值远大于所有其他音位控制函数值总和.这样就保证了在任何情况下发这些音时其上下唇都会处于闭合状态.可用如下公式表示:

$$D_{sp} = D_{sp0} + \left(C\sum_{k \neq p} D_{kp} - D_{sp0} \right)\left(1 - \frac{|\tau|}{T}\right), \tag{7}$$

式中T为一固定时间值,D_{sp0}为距音位中心时间为T时的控制函数值,C为一常数,表示在中心时刻,控制函数值与其他所有控制函数值之和的比例,一般应大于20.

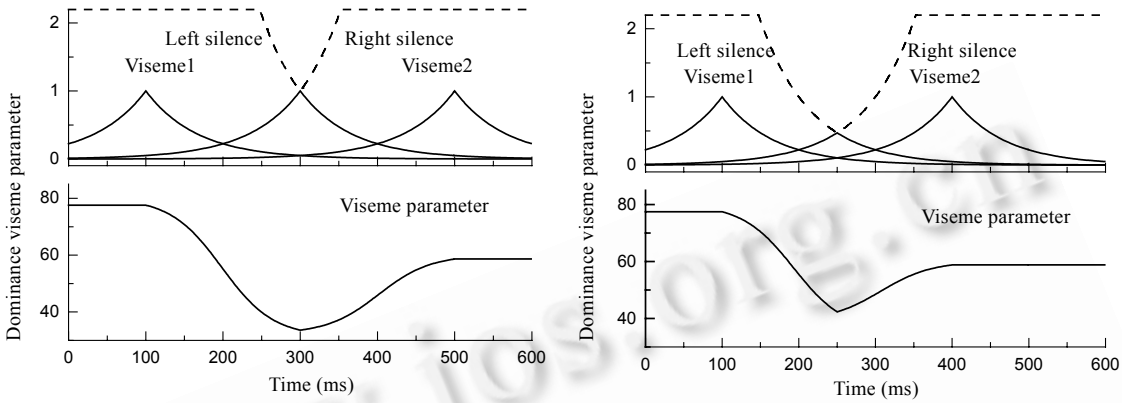


Fig.2 Domiance and viseme parameter of different pause in speech

图 2 语音中不同停顿时间的控制函数和视位参数

这样,由各个动态视位模型和左、右无声模型,我们就可以生成连续说话过程中各个口形参数的变化曲线,从而模拟人们说话时复杂多变的口形变化.

4 模型参数的估计

根据我们上面所述的动态视位模型可以模拟出各种各样的口形参数变化曲线,但要在可视语音合成的过程中使用它,首先需要确定模型中的参数.从式(1~4)可以看出,对每个视位的某一个口形参数,有5个模型参数需要确定,即 α_{sp} (中心处控制函数值)、$\theta_{\leftarrow sp}$ (前向衰减系数)、$\theta_{\rightarrow sp}$ (后向衰减系数)、t_{osp} (控制函数中心与语音中心的时间距离)和 T_{sp}(口形参数值);对于如'b','p','m'的特殊音位的某些参数需要7个参数(外加 C 和 T);对于左、右无声模型共有 7 个参数,α_{lp}, α_{rp}, θ_{lp}, θ_{rp}, t_l, t_r 和 T_{0p};另外还有一个各视位共同的指数参数 c.可以看出,模型中待定参数较多,自由度大,其参数的估计本身就是一个复杂的问题.

估计模型参数的具体方法不外乎两类:一是依靠发音规则和主观经验知识人为地设定,二是动用机器学习的方法从实际数据中获取.文献[9,10]在可视语音合成中都利用了 Cohen&Massaro 模型,其中文献[9]通过观察人的说话手动调整模型参数,文献[10]提到了从实际数据中获取参数,但并没有详述其方法.要合理地从实际数据中估计模型参数,必须首先确定一个适当的学习策略.我们针对汉语可视语音合成的特点提出了一种分组、分阶段、分步骤学习的方法.首先将汉语发音的口形进行分类,并分为声母、单一口形韵母、复合口形韵母 3 组.对单一口形韵母、声韵母组成的音节分别进行发音录像,采用变形模板方法跟踪口形变化,得到实际发音过程中的口形参数,同时计算语音的短时能量、过零率等参数,以便自动切分.

参数估计的第 1 阶段包括两个步骤:第 1 步,先根据实际数据设定一个初始的无声模型参数,利用它对每个韵母视位的口形参数变化模型进行学习,以模型产生的口形参数与实测口形参数的误差平方和作为相似度准则或能量函数,利用梯度下降法找到误差最小的模型参数值.第 2 步,在对所有单韵母视位学习完一遍之后,再调整无声模型的参数及指数系数 c.同样,在一个大的循环中利用梯度下降法找到误差最小的无声模型参数值及系数 c.在第 2 阶段,利用已有的无声模型参数和韵母模型参数值来学习声母视位的模型参数,整个过程如图 3 所示.图 3 中下面的虚线框代表与学习韵母视位模型参数相似的学习过程,图中作了简化.

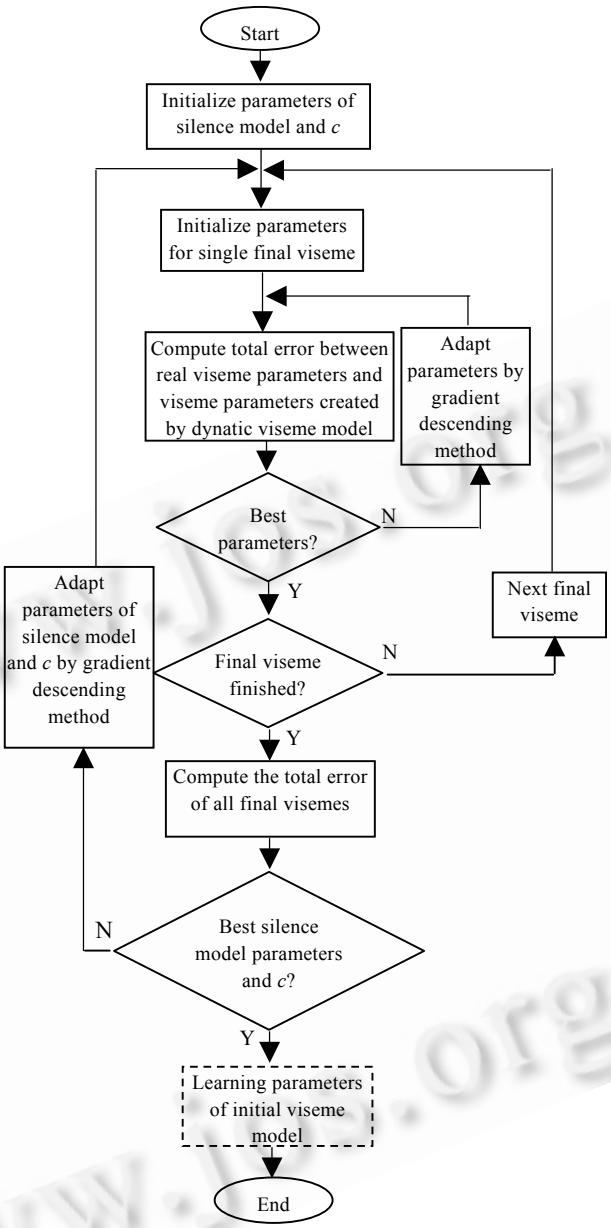


Fig.3 Flow chart of learning parameters for dynamic viseme model

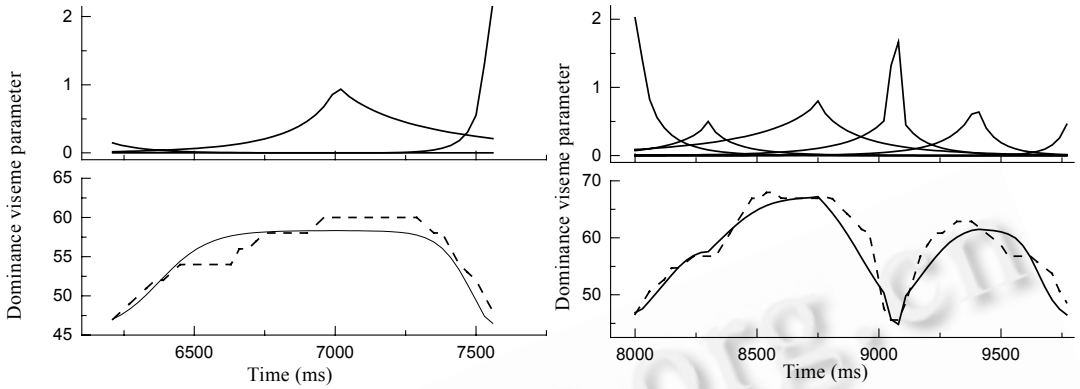
图 3 动态视位模型参数学习流程图

5 实验结果

图 4 是我们在学习汉语动态视位模型参数及合成连续语流中视位参数时所得到的实验结果,图中上窗口所示为各个视位的基本控制函数值随时间变化的过程,下窗口为在各个控制函数作用下的开口高度参数随时间变化的过程.图 4(a)是由学习得到的韵母/e/视位模型和左、右无声模型的控制函数曲线以及由此模型合成的参数值(下窗口中的实线)与实测的参数值对比(下窗口中的虚线).

图 4(b)是汉语“一安培”发音时各视位的控制函数值(上窗口)及利用这些控制函数合成的开口高度参数变化曲线(下窗口中实线)与实测参数(下窗口中虚线)的对比.

从图中可以看出,模型合成的参数与实测所得口形参数吻合得较好.说明我们提出的动态视位模型可以有效地描述汉语发音时的口形参数变化规律,同时也说明我们提出的参数学习方法是可行的.



(a) Viseme /e/ (b) Sentence 'yi an pei'

(a) 发/e/音 (b) 发'一安培'音

Fig.4 Mouth height created dynamic viseme model (solid) and tracking result (dotted)

图4 动态语音视位模型生成的开口高度值(实)与实际跟踪值(虚)

6 结束语

本文通过对人们发音过程中口形变化规律的研究和总结可视语音合成中口形合成的经验,提出了一个动态语音视位模型,它可以模拟人们说话时各种各样的口形参数变化曲线.在此基础上,针对汉语发音的特点,我们给出一种行之有效的模型参数估计方法.

在参数的学习过程中,我们以平误差之和作为相似度的衡量准则,也可以选取其他相似度的衡量准则,如最大绝对误差、交叉相关性等等.哪一种相似度的衡量准则更接近于人的主观感觉,有待于进一步的研究.另外,图像跟踪的精度对学习结果也有一定的影响,但精确地跟踪口形变化是一个复杂的问题,如何准确地跟踪口形的三维变化过程也是我们下一步要研究的问题之一.

References:

[1] Cohen MM, Massaro DW. Modeling coarticulation in synthetic visual speech. In: Thalmann NM, Thalmann D, eds. Models Techniques in Computer Animation. Tokyo: Springer-Verlag, 1993. 139~156.

[2] Reveret L, Bailly G, Badin P. Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In: Yuan Bao-Zong, Huang Tai-Yi, Tang Xiao-Fang, eds. Proceedings of the 6th International Conference on Spoken Language Processing (Ⅱ). Beijing: China Military Friendship Publish, 2000. 755~758.

[3] Brooke NM, Scott SD. Computer graphics animations of talking faces based on stochastic models. In: International Symposium on Speech, Image Processing and Neural Networks. 1994. 73~76.

[4] Masuko T, Kobayashi T, Tamura M. Text-to-Visual speech synthesis based on parameter generation from HMM. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (Ⅵ). 1998. 3745~3748.

[5] Bregler C, Covell M, Slaney M. Video rewrite: driving visual speech with audio. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graphics. 1997. 353~360.

[6] Cosatto E, Potamianos G, Graf HP. Audio-Visual unit selection for the synthesis of photo-realistic talking-heads. In: IEEE International Conference on Multimedia and Expo (Ⅱ). 2000. 619~622.

[7] Steve M, Andrew B. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In: Yuan BZ, Huang TY, Tang XF, eds. Proceedings of the 6th International Conference on Spoken Language Processing (Ⅱ). Beijing: China Military Friendship Publish, 2000. 759~762.

[8] 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).

[9] Zhong J, Olive J. Cloning synthetic talking heads. In: Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis. 1998. 26~29.

[10] Le Goff B, Benoit C. A text-to-audiovisual-speech synthesizer for French. In: Proceedings of the 4th International Conference on Spoken Language Processing (Ⅳ). 1996. 2163~2166.