汉语三维发音动作合成和动态模拟

郑红娜[1,2] 朱 云[1] 王 岚[1] 陈 辉[3]

[1]（中国科学院深圳先进技术研究院集成所环绕智能实验室 深圳 518055）
[2]（太原理工大学信息工程学院 太原 030024）
[3]（中国科学院软件研究所 北京 100080）

摘 要 本文以帮助聋儿言语康复为出发点，从聋儿音频发音数据中获得了聋儿易错发音文本以及聋儿易混淆发音文本对。设计了一个数据驱动的3D说话人头发音系统，该系统以EMA AG500设备采集的发音动作为驱动数据，逼真模拟了汉语的发音，从而可使聋儿观察到说话人嘴唇及舌头的运动情况，辅助聋儿发音训练，纠正易错发音。最后对系统的性能进行了人工评测，结果表明：3D说话人头发音系统可以有效地模拟说话人发音时口腔内外器官的发音动作。此外，本文还用基于音素的CM协同发音模型合成的方法，合成了聋儿易错发音文本的发音动动作，并用RMS度量了合成发音动作与真实发音动作的误差，得到了均值为1.25 mm的RMS误差值。

关键词 聋儿易错发音文本；3D说话人头；CM协同发音模型；电磁发音动作采集仪（EMA）；Dirichlet Free-Form Deformation(DFFD)算法

Chinese 3D Ariticulatory Movement Synthesis and Animation

ZHENG Hong-na[1,2] ZHU Yun[1] WANG Lan[1] CHEN Hui[3]

[1](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)
[2](*College of Information Engineering, Taiyuan University of Technology, Taiyuan* 030024, *China*)
[3](*Institute of Software Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract In order to help the hearing loss children, we obtained hearing loss children's fallible pronunciation texts and the confusing pronunciation text pairs form a good deal of hearing loss children's audio pronunciation data. We designed a data-driven 3D talking head articulatory animation system, it was driven by the articulatory movements which were collected from a device called Electro-magnetic articulography (EMA) AG500, the system simulated Chinese articulation realistically. In that way, the hearing loss children can observe the speaker's lips and tongue's motions during the speaker pronouncing, which could help the hearing loss children train pronunciation and correct the fallible pronunciations. Finally, a perception test was applied to evaluate the system's performance. The results showed that the 3D talking head system can animate both internal and external articulatory motions effectively. A modified CM model based synthesis method was used to generate the articulatory movements. The root mean square between the real articulatory movements and synthetic articulatory movements was used to measure the synthesis method, and an average value of RMS is 1.25 mm.

Keywords hearing loss children's fallible pronunciation texts; 3D talking head; CM co-articulation model; electro-magnetic articulography (EMA); Dirichlet Free-Form Deformation (DFFD) algorithm

1 引 言

长期以来，政府和社会各界都十分关注聋儿的康复。学者们也进行了大量有关聋儿康复的语音可视化方面的研究。早在上世纪，就有研究详细阐述了聋儿在语言学习过程中及语言感知过程中，视觉信息输入的重要性[1,2]。有早期的研究将唇部和面部信息应用

基金项目：国家自然科学基金项目(NSFC61135003，NSFC90920002)、中国科学院知识创新工程项目(KJCXZ-YW-617)。
作者简介：郑红娜，硕士研究生，研究方向为语音信号处理，E-mail：girlzhhna@126.com；朱云，硕士研究生，研究方向为语音可视化；王岚，博士，研究方向为构建先进智能信息系统；陈辉，博士，研究方向为人机交互技术。

到聋儿言语康复中，获得了良好的效果[15]。

国内外有关语音可视化的研究大多着眼于以下两个方面：

(1)建立符合动力学特性的2D或3D说话人头发音系统模型，这些模型大多是基于生理特性的，旨在重构包括说话人的面部、唇部、舌头、牙齿、软腭、下巴等器官的三维模型[8]。近年来，学者们已实现了基于生理结构的虚拟3D说话人头模型，可逼真展示通常不易观察到的舌头、软腭等发音器官运动的发音动作[2-5]。目前，已成功模拟了法语和英语的发音动作[9]。本文构建了一个模拟汉语发音动作的虚拟3D说话人头模型，并逼真地模拟了汉语三维发音动作，直观呈现了说话人发音时唇部、舌头、软腭、牙齿等器官的标准普通话发音动作。

(2)模型的驱动。驱动数据的来源有三种：①实时MRI采集。这种方法的采样率偏低，因而不适于采集连续变化的发音动作；②x-ray采集。这种方法有害于人体健康，尤其不适合对同一测试者采集大量的数据；③利用电磁发音动作采集仪（Electro-magnetic articulography，EMA)采集驱动数据[6]。该设备可捕获高精度的语音信息，且不损害人体健康，是捕获复杂多变发音动作的理想方式[6]。本文用德国CARSTENS股份有限公司生产的电磁发音动作采集仪EMA AG500采集真实说话人上下唇和舌头等发音器官的发音动作数据，并基于汉语音素发音动作数据的位移矢量，运用Cohen和Massaro提出的Cohen-Massaro协同发音模型（CM协同发音模型）的合成及平滑技术[10]合成了聋儿易错发音文本的连续发音动作，进而驱动和控制3D说话人头模型。精确模拟了汉语的发音动作，并直观展示了标准汉语发音时，口腔内外各个发音器官的发音动作，使聋儿在听觉（通过佩戴助听器或植入人工耳蜗实现）和视觉信息的双重刺激和指导下进行发音训练。此模型不仅可帮助聋儿进行言语康复，还可用于帮助国外的汉语学习者更容易地区分易混淆音素之间发音动作的区别，练习准确发音等[9]。

最后，我们对3D说话人头模拟的发音动作效果做了人工评测，实验结果说明：可透明的3D说话人头逼真地模拟了真人的发音动作。通过视觉和听觉信息的融合可以很好地帮助聋儿纠正错误发音，提高他们的发音准确率，有效指导聋儿进行言语康复。

2 聋儿易混淆发音分析

为了观察聋儿的发音特点，获得聋儿易混淆的音素级别上的发音文本。我们进行了多次聋儿发音的音频和视频数据采集工作。数据采集对象是3岁到8岁的20名言语障碍儿童。用于聋儿发音数据采集的语料，参照了《听力障碍儿童评测标准内容》和《汉语普通话发音标准》的相关语料片段，并力求做到各个音素出现的频率相等，内容包括：汉语拼音的36个韵母和21个声母、单音素和音节对比对（如表2的“发音文本对”所示）以及常见词汇，如“阿姨、弟弟、洗衣机等”和句子。其中聋儿发音数据采集过程主要是通过拼音及汉字、图片等多重信息的提示，引导聋儿发音并录制聋儿的发音数据，最后将音频信息保存为16 kHz、16 bit的单声道wav文件。

我们邀请语言学专业人士对采集的聋儿发音数据进行标注，发现聋儿的发音与正常儿童的发音相比，存在以下不足：首先，聋儿在说话时呼吸短促，不会把握适度的音量，不善于在言语过程中换气和停顿，缺乏流畅感；其次，不能准确发卷舌音，即没有正确把握汉语音素中的4个卷舌音（/zh/、/ch/、/sh/、/r/）的发音技巧；最后，阴平、阳平、上声、去声和清音这五个声调发音也不准确。注：本文中出现的/zh/、/a/、/chī/等均表示汉语拼音。

本文旨在通过直观展示正确发音动作来纠正音素级别的聋儿的易错发音，例如音节/zhi/错读为/zi/，音节/shi/错读为/zhi/等。我们对采集到的聋儿发音数据标注进行总结和归纳，获得了聋儿的易错的发音文本，如表3所示，并把聋儿混淆频率最高的发音文本对成对列出，组成了聋儿易混淆发音文本对，例如/a/和/o/、/zi/和/zhi/等，如表2中所示。总结聋儿易错的发音文本，主要包括：韵母易错发音文本、声母易错发音文本和卷舌音易错发音文本。

3 合成发音文本的运动轨迹

3.1 三维发音动作数据采集和分析

3.1.1 汉语三维发音动作数据采集

为了获得控制和驱动三维说话人模型的数据，本文选择电磁发音动作采集仪（Electro Magnetic Articulography，EMA）AG500采集真实说话人舌头和唇部一定离散点处的发音动作数据。为了获得标准的发音动作，实验者请了一位有多年汉语教学经验的老师。

数据采集的过程如下：在说话人的舌头和唇部等离散点处粘贴EMA的微型传感器，如图1所示：采用10个传感器分别置于说话人的面部、唇部及舌头的几个离散点处。其中一个传感器置于鼻梁H_1，两个位于耳后$H_{2,3}$，H_1、H_2和H_3共同作为参考传感器，用于校准数据。唇部的四个传感器分别置于上唇L_2、下唇L_3、左嘴角L_4和右嘴角L_1四个点处，舌头上的三个传感器分别置于舌头的三个不同位置，即舌根T_3、舌体T_2和舌尖T_1，如图1左图所示。图1的右图展示了标有传感器离散位置的3D说话人头模型。在说话人阅读语料时，使用电磁发音动作采集仪EMA AG500，以200帧每秒的采样率录制发音动作，并录制与其同步的语音波形。

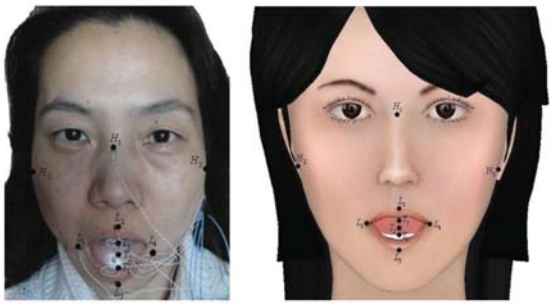

图1　EMA传感器在说话人和3D说话人头模型上的离散数据采集点

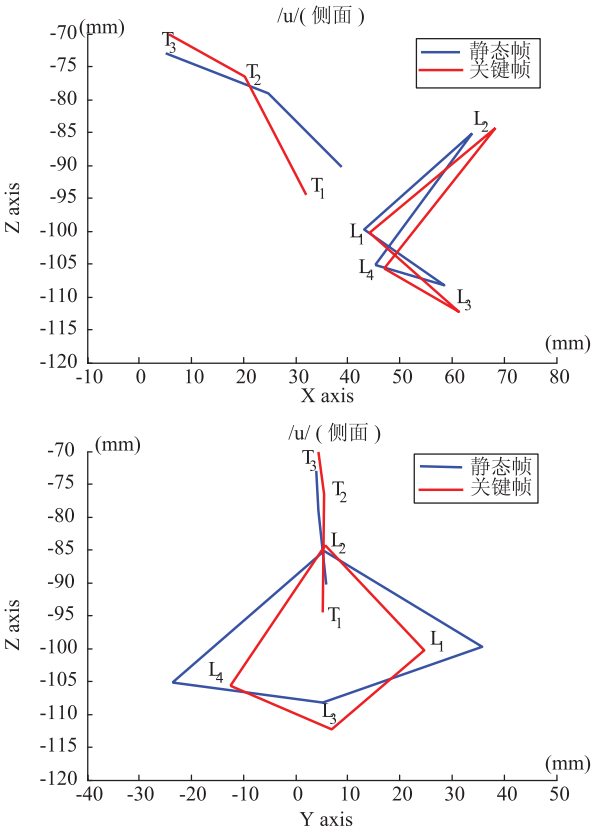

图2　音素/u/静态帧和关键帧的侧面（上图）和正面（下图）视图

数据采集结束后，要对数据进行处理。由于采集的数据是头部运动和发音动作的混合运动，此时要发挥参考传感器的作用（即仅仅记录单一的头部运动数据），应用自动校准程序进行头部校准，去掉头部动作带来的影响，获得所需发音器官的发音动作。

3.1.2　汉语三维发音动作数据分析

采集的发音数据是200帧/秒的离散值，每个音素发音是由若干帧组成的，包括静态帧和关键帧等。其中，静态帧指舌头和上下唇都处于放松状态的一帧，是从不发音的起始静音状态中选取出来的。关键帧指能反映音素的发音特点的一帧，音素的关键帧常选取为与静态帧之间的欧式距离最大的一帧。

图2描绘了发汉语拼音/u/时，上下唇和舌头的静态帧（虚线）和关键帧（实线）的侧面视图（上图）和正面视图（下图）。从图中可观察出，发/u/时，上下唇拢成圆形，向前突出，且舌头后缩。这些特点同时也符合发汉语音素/u/的标准发音动作，即双唇拢圆，留一小孔，舌头后缩，使舌根接近软腭[12]。

3.2　合成三维发音运动轨迹

基于汉语单音素的三维发音动作可以合成聋儿易错发音文本的发音动作。由于协同发音在连续语流中会对嘴唇的发音动作产生很大的影响，所以直接连接的方法并不可取。CM协同发音模型是一个符合实际发音过程的有效数学模型，能准确反映说话人上下唇的变化情况，可以达到高质量的可视语音合成效果[13]。本文采用CM协同发音模型的方法[10]来合成聋儿易错音节的发音动作。利用CM协同发音模型合成连续发音动作包括以下两个主要步骤：

步骤一：合成。用单音素的位移矢量（即音素的关键帧与静态帧的差值）作为合成基本单位。设待合成的聋儿易错发音文本包括若干个音素，将第p个音素的位移矢量定义为R_P, 该音素的幅度α_P由以下公式定义：

$$\alpha_p=\begin{cases}\dfrac{max_p\left|R_p\right|-\left|R_p\right|}{max_p\left|R_p\right|}, & if\ \left|R_p\right|\neq max_p\left|R_p\right| \\ 1.0, & if\ \left|R_p\right|=max_p\left|R_p\right|\end{cases} \tag{1}$$

接下来，使用一个指数函数（也称主导函数）来近似单个发音动作。第p个音素的主导函数$D_p(\tau)$定义为：

$$D_p(\tau)=\begin{cases}\alpha_p.e^{-\theta_d|\tau|^c}, & if\ \ \tau\le 0 \\ \alpha_p.e^{-\theta_g|\tau|^c}, & if\ \ \tau>0\end{cases} \qquad \tau=t_{pk}-t \tag{2}$$

上式中，c是一个常数，试验中取2；θ_d和θ_g是两个常量，分别代表指数函数的增长速度和下降速度，由以下公式决定：

$$\begin{cases} \alpha_p \cdot e^{-\theta_d |t_{pk}-t_{pe}|} = \varepsilon \\ \alpha_p \cdot e^{-\theta_d |t_{pk}-t_{ps}|} = \varepsilon \end{cases} \quad (3)$$

其中，第p个音素的起始帧、关键帧和结束帧时刻分别为t_{ps}、t_{pk}、t_{pe}。由于单个音素的发音时长较其在连续语言中的发音长度要短一些，因此单音素的时长是通过自动语音识别系统进行强制对齐得到的。公式（3）中ε是一个常数，指两个指数函数的重叠部分，也叫局部最小值，试验中取0.22。

步骤二：平滑。使用论文[11]中提出的混合函数，即加权的高斯函数，来平滑前后两个相邻音素的发音动作：

$$\hat{F}_w(t) = \sum_{p=1}^{N} R_p \cdot D_p(t - t_p) \quad (4)$$

合成试验中使用了唇部四个点和舌头三个点的数据，分别合成了聋儿易错发音文本在舌尖、舌体、舌后各点的上下方向（Z轴）和前后方向（X轴）的运动轨迹，以及上唇点、下唇点、左嘴角点、右嘴角的前后方向和上下方向的运动轨迹。试验合成了表1中列出的聋儿易错发音文本的发音动作。

以下唇的Z轴方向为例，图3上图画出了音节/xī/所包含的两个单音素/x/和/ī/在下唇Z轴方向上的主导函数，下图描绘了音节/xī/的合成发音轨迹（下图中的虚线）及EMA录制的真实发音轨迹（下图中的实线）。

试验用合成发音曲线与真实发音曲线的均方根误差RMS来客观地度量合成发音轨迹与真实发音轨迹之间的误差，RMS的计算方法如下：

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(F_i - E_i)^2}{n}} \quad (5)$$

公式中，n为EMA录制的发音文本的帧数，F_i和E_i分别表示合成的与EMA实际采集的发音动作的坐标值。

各聋儿易错发音文本的RMS值，如表1所示，其RMS误差分布在0.27 mm到2.70 mm的范围内，均值为1.25 mm。

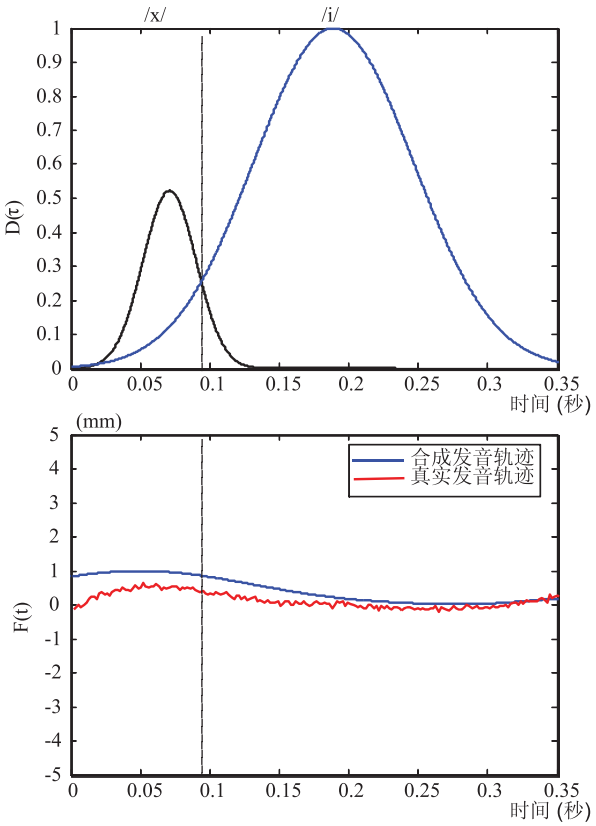

图3 汉语音素/xi/下唇Z轴的主导函数（上图）、合成及真实发音曲线（下图）

最后，用已合成的三维发音动作数据驱动3D说话人头发音模型，以直观地显示口腔内外聋儿易错发音文本的发音动作。

表1 合成发音文本的RMS值（mm）

文本	RMS	文本	RMS	文本	RMS	文本	RMS
/gē/	0.92	/nē/	0.67	/lē/	2.69	/rī/	1.11
/dē/	2.26	/qī/	0.41	/zī/	0.86	/zhī/	1.34
/hē/	2.70	/xī/	0.27	/cī/	0.95	/chī/	0.39
				/sī/	0.72	/shī/	0.33

4 3D说话人头模拟发音运动

3D说话人头模型的结构包括：小舌、舌头、上下唇、上下牙、下颚的三维模型。而关于3D说话人模型的变形算法有很多，其中FFD(Free-Form Deformation)算法是计算机图形界最有名的自由变形算法。DFFD（Dirichlet Free-Form Deformation）算法[15]是FFD的扩展算法，通过引用新的坐标系统，即自然邻居坐标：Sibson坐标（也称希普森坐标）可

以把控制点设置在网格的任何位置，同时控制网格的形状也可以是任意的。DFFD算法更适合于复杂多变的几何变形。因此，本文采用DFFD算法为变形算法，使虚拟3D说话人头产生了符合人体生理特性的发音动作。

5　人工评测

我们设计了两组人工评测的实验。

第一组人工评测实验的目的是：测试3D说话人头模型模拟的三维发音动作在区分聋儿易混淆发音文本对上的性能。实验选取了10对聋儿易混淆发音文本对（即表2所示）作为待辨别的易混淆发音文本。

试验方法如下：假设要从3D说话人头模型模拟的发音动作视频中辨别汉语发音文本对/cī/和/chī/，先随机播放/cī/或/chī/其中一个文本对应的3D说话人模型的发音动作视频，再播放另一个文本对应的3D说话人模型的发音动作视频。在告知测试者所观察到的两个视频对应的文本是/cī/和/chī/的前提下，让测试者对前后两个视频对应的文本（/cī/、/chī/）进行区分，并记录辨认结果。图4展现了3D说话人模型所模拟的三维发音动作视频的截图，其中上面两图分别对应文本/cī/和/chī/发音动作的正面视频截图，此时模型的驱动数据是关键帧位移矢量。与其相对应的下面两图分别对应文本/cī/和/chī/的侧面视频发音动作视频截图。

实验请了10名不同的具有良好普通话水平的学生作为测试者，分别辨别了每一对聋儿易混淆发音文本对应模型的发音动作。实验的辨别结果如表2所示，列出了每对聋儿易混淆发音文本对的辨别率。待识别的10对发音文本中，能正确辨别的平均辨别率为90%。而正确辨别率较低的几对（如/lē/和/nē/等）不能正确辨别的原因主要是：这些文本对在唇部和舌头这两个发音部位的发音动作差别微妙，并无明显发音动作区别，所以影响了发音文本对的正确辨别率。

第二组人工评测实验的目的是：测试3D说话人头模型模拟聋儿发音易错文本三维发音运动的效果。评测方法如下：为了将3D说话人模型的发音动作与真人的发音动作相比较，首先，播放聋儿易错发音文本的真人发音动作的音视频；其次，播放合成的该发音文本对应的3D说话人头模型的发音动作视频。如图4的视频截图所示：分别从正面（左上图）和侧面（左下图）直观地展现了音节/cī/的唇部和舌头的发音动作。实验请了8名具有良好语言学素养的学生作为评价者对3D说话人模型的发音动作效果进行评价并打分，分数分为1到5共五个等级，分别代表较差、一般、良好、较好、很好五个等级。各个聋儿发音易错发音文本所得分值的平均值如表3所示，3D说话人头模型模拟的发音动作的得分总平均值约为4.2分。评价者一致认为：3D说话人头所模拟的发音文

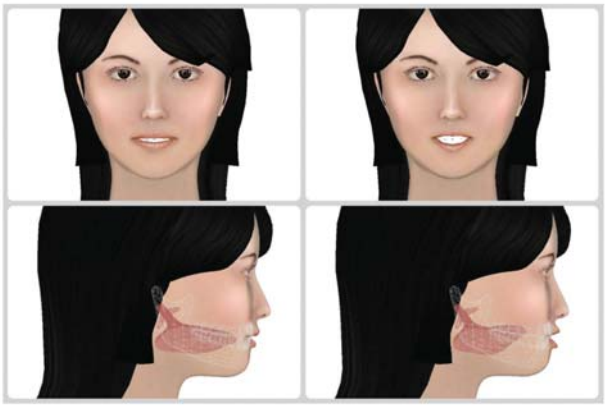

图4　音节/cī/（左侧上下图）和/chī/（右侧上下图）3D说话人头模型的发音动作视频截图

表2　合成发音文本的RMS值（mm）

发音文本对	辨别率	发音文本对	辨别率	发音文本对	辨别率
/u/vs./ü/	90%	/gē/vs./dē/	85%	/rī/vs./chī/	85%
/a/vs./o/	100%	/ē/vs./nē/	80%	/zī/vs./zhī/	100%
/e/vs./i/	90%	/xī/vs./qī/	80%	/cī/vs./chī/	100%
				/sī/vs./shī/	90%

表3　聋儿易错发音文本及其3D模型发音效果评分

发音文本	平均分值	发音文本	平均分值	发音文本	平均分值	发音文本	平均分值
/a/	4	/ü/	4	/qī/	4	/sī/	4
/o/	3	/gē/	4	/xī/	4	/rī/	4
/e/	4	/dē/	4	/hē/	3	/zhī/	4
/i/	4	/lē/	3	/zī/	4	/chī/	4
/u/	4	/nē/	3	/cī/	4	/shī/	4

本的发音动作可以很好地模拟发音器官的三维发音运动。

6 结 论

本文用一个数据驱动的3D说话人头模型，逼真地模拟了聋儿易错发音文本各发音器官的发音动作，并运用CM协同发音模型合成方法将单音素的发音动作合成了发音文本的发音动作。对模型性能进行的人工评测结果表明：3D说话人头可以逼真地模拟人发音的三维发音运动。这一技术的实现有着重大的现实意义：通过3D说话人头模型，聋儿不但可以直观地看到说话人发音时舌头和嘴唇的发音动作，而且可以比较易混淆发音文本对的发音动作从而纠正不准确的发音。同时也为人们对汉语及其他语言的研究提供了一条新的途径。

参 考 文 献

[1] Baker A E. The development of phonology in the blind child [J]. Hearing by Eye: the Psychology of Lip Reading, 1987: 145-161.

[2] Sumby W H, Pollack I. Visual contribution to speech intelligibility in noise [J]. Acoustical Society of America, 1954, 26: 212-215.

[3] Stoel-Gammon C. Prelinguistic vocalizations of hearing-impaired and normally hearing subjects [J]. A Comparison of Consonantal Inventories Speech and Hearing Disorders, 1988, 53: 302-315.

[4] Mulford R. First Words of the Blind Child, the Child's Development of a Linguistic Vocabulary [M]. New-York: Academic Press, 1988: 293-338.

[5] Benoît C, Le Goff B. Audio-visual speech synthesis from French text: eight years of models, designs and evaluation at the ICP [J]. Speech Communication, 1988, 26: 117-129.

[6] Pfitzinger H. Concatenative speech synthesis with articulatory kinematics obtained via three-dimensional electro-magnetic articulography [J]. Fortschritte der Akustik, 2005, 31(2): 769-770.

[7] Heracleous P, Hagita N. Automatic recognition of speech without any audio information [C] // International Conference on Acoustics, Speech, and Signal Processing. 2011.

[8] Serrurier A, Badin P. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on mri and ct data [J]. Acoustic Society of American, 2008, 123(4): 2335-2355.

[9] Pierre B, Frédéric E. An audiovisual talking head for augmented speech generation: models and animations based on a real speaker's articulatory data [J]. Computer Science, 2008, 5098/2008: 132-143.

[10] Cohen M, Massaro D W. Modeling coarticulation in synthetic visual speech [J]. Models Technique in Computer Animation, 1993: 139-156.

[11] Wang L, Chen H, Li S, et al. Phoneme-level articulatory animation in pronunciation training [J]. Speech Communication, 2012, 54: 845-856.

[12] 黄伯荣, 廖序东. 现代汉语 [M]. 北京:高等教育出版社, 2002.

[13] 王志明, 蔡莲红. 动态视为模型及其参数估计 [J]. 软件学报, 2003, 14(03): 461-466.

[14] Thomas W S, Scott R P. Free-form deformation of solid geometric models [J]. ACM Computer Graphics, 1986, 20(4): 151-160.

[15] Moccozet L, Thalmann N M. Dirichlet free-form deformations and their application to hand simulation [J]. Computer Animation, 1997, 97: 93-102.

[16] Watson D F. Computing the n2 dimention delaunay tessellation with application to voronoi polytopes [J]. Computer Journal, 1981, 24(2): 167-172.