一种基于三维模型和照片的合成“说话头”

赖 伟 孙 岭 王仁华

(中国科学技术大学电子工程与信息科学系,合肥 230027)

摘 要 视觉语音的研究已经成为人机交互技术中一个非常活跃的领域,在语音的相关视觉信息当中,最主要的是说话人的口型乃至整个头部的图像,即“说话头”(talking head)。为了合成具有真实感的三维“说话头”模型,提出了一种基于三维模型和真人照片来合成真实“说话头”的方法,即在一个中性的三维人头部模型的基础上,从任意人的正面和侧面两张照片当中,通过提取脸形和五官位置等特征参数来校正模型,并且从照片中提取皮肤和头发等纹理,使得合成的模型能在较大程度上贴近真人。该方法综合了基于三维模型和基于图像库的建模方法,因此同时具有两者的优点,即既能够灵活控制表情和口型,又可自由旋转,不仅可实时合成,而且合成效果接近真人,自然度高。已将此模型应用于视觉语音合成系统,并获得了满意的效果。

关键词 说话头 视觉语音合成 三维模型 人脸动画

中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2004)07-0886-07

A Talking Head Synthesis System Based on 3D-Model and Photo

LAI Wei, SUN Ling, WANG Ren-hua

(*The University of Science and Technology of China, Department of Electronics Engineering and Information Science, Hefei* 230027)

Abstract Recently, research on Visual Speech attracts more and more attention. It has become a very active research field of the Human-Machine Interface. The chief information relative to speech is lip motion, face, and even the whole head, which is called “Talking Head”. To synthesis a lifelike three-dimension (3D) talking head model, a novel method is proposed in this paper, which is based on an individual independent 3D-model and photos of human face. At first, the features of face shape and the position of facial organs are extracted from a front-face and a side-face photo to revise the 3D-model and make it adapt the real person. Then, the textures of the skin and hair are picked from the photos and pasted on the revised 3D-model to make it looks like the person. This method integrates the techniques of 3D-model based modeling and photo lib based modeling, and has both of their advantages: the model has strong flexibility of synthesizing lip motions and expressions, can be rotated freely, can be synthesized in real-time, and can achieve a highly natural, lifelike 3D talking head visual effect. Then, the model is applied in a visual Text-to-Speech (TTS) talking head synthesis system, and gets a satisfying result.

Keywords talking head, visual text-to-speech, 3D model, face animation

1 引 言

近年来,对于视觉语音的研究得到人们越来越多的重视,该技术已经成为人机交互技术中一个非常活跃的领域。在语音识别中加入视觉信息,可以增加识别的鲁棒性和提高识别率。其用于网络虚拟主持人可以提高网络的人性化,而在电话通话中增加视觉信息,则会大大方便听力有障碍的人使用[1]。

在语音的相关视觉信息当中,最主要的是说话人的口型乃至整个头部的图像,即“说话头”。目前,其合成主要有基于模型的和基于图像库的两大类方法。

其中,基于图像库的方法是通过录制大量真人说话的图像来组建一个库,合成时,再由库中挑选单元拼接而成,该方法的优点是合成效果好,贴近真

收稿日期:2003-08-15;改回日期:2004-03-12

人,缺点是数据库非常庞大,不仅因其运算复杂度高而难以达到实时合成,而且其合成模型也缺乏自由变换的能力;而基于模型的方法则是先建立一个类似真人头部的三维模型,然后根据说话人的脸部特点,通过调节模型参数来实现不同表情和口型的显示,它的优点是模型的调节非常灵活,还可以有旋转等变换,且复杂度低,可以实时合成[2,3],但是这种方法一个比较普遍的问题是,模型的外形比较卡通化,与真人的效果相差太远。

本文提出了一种基于三维模型和真人照片来合成“说话头”的方法,即在保留模型方法的优点的前提下,从一个中性的人头部模型着手,首先利用任意人的正面和侧面照片,从中提取出特征参数来校正模型,以使模型在外形上贴近真人;然后从照片中提取纹理,使得合成的模型在皮肤、头发、五官等各部分都能在较大程度上贴近真人,而且能够很灵活地控制表情和口型,以实现高自然度的三维“说话头”。本文提出的这种合成方法,由于综合了模型方法和图像库方法的优点,因此既有比较高的仿真自然度,又保证了较高的灵活性和较低的复杂度,还可以实时实现。

文中还阐述了此模型在视觉语音合成等视觉语音合成系统中的应用,包括如何提取自然的口型和表情,如何进行语音和口型/表情同步的合成。

2　三维模型及其校准

本文提出的“说话头”模型由两个部分组成,其一部分是“骨架”(图 1(a)),就是人头部外表面的三维模型,它由若干顶点和三角形组成(本文系统中采用的模型包含有 1703 个三角形,1309 个顶点);另一部分就是“纹理”,也就是覆盖于模型表面的外壳,比如头发部分是黑色,嘴唇部分是红褐色(现图上为深灰色)的纹理(如图 1(b)所示)[4]。

(a)“骨架”　　(b)“纹理”　　(c) 模型的三维坐标方向示意

图 1　“说话头”模型组成

初始的模型只是粗略的与人的头部相似,但当需要制作和某个特定的人相像的模型时,就需要根据此人的特征来定制,也就是要用特征参数来调整校准原始的模型。这些参数的来源,就是此人的正面和侧面两张照片[5]。

校准的方法是先定位出图像中的特征点,再根据这些特征点的位置来修正对应模型上关键点的坐标。由于模型校准的精确程度对模型的最终合成效果有很大的影响,因此可以采用半自动的手工辅助调整方法来精确校准。

需要调整的参数主要是以下两类,一是头形、脸形;二是眼、眉、鼻、耳和嘴巴这些五官的位置。校准时,首先将正面及侧面两张照片缩放到同样的大小,作为背景图片,然后将与照片同姿势的原始模型的正面和侧面分别放置于其上,再参考背景图像来调节模型上轮廓一圈的顶点坐标(图 2 上白色点),其邻近的顶点的坐标也要以一定的权重随之改变,然后保存新的模型数据(如图 2 所示)。

图 2　模型的校准

接着是调整五官的位置,同样,以嘴部为例进行说明(如图 3 所示),嘴唇内部的顶点(最内层的圆点),包括牙齿和舌头处的顶点都以一定权重一起移动,而嘴唇外部相邻的顶点,则分别以较低权重(中间一层的方形点)和更低权重(最外层的三角形点)相应移动。

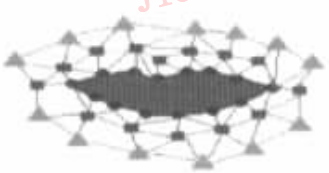

图 3　模型调整示意

最后保存调整完毕的模型数据,并将其作为适合此人的模型“骨架”。

3 照片纹理的提取及贴图

为了使合成的模型像真人，除了“骨架”之外，还需要真实感的纹理，其中包括皮肤、毛发、五官，这些数据将从真人的正面和侧面两张照片中提取。

3.1 贴图数据的获取

合成三维人头像模型时，为了使模型上每个顶点所对应的贴图坐标都能找到相应的图像数据，需要一张展开铺平的人脸图像，包括四面和头顶等所有方向，但是实际上有用的和被关注的主要是人的正面和侧面，而后脑和头顶等地方一般很少被注意，合成模型一般也不会显示到这些部分，因此这些部分只需要相对粗糙的图像即可，即可以先从正面和侧面的图片中提取出那一部分，然后采用重复延伸的方式来显示即可。

为了提取人脸正面和侧面的纹理，需要先在模型上为正面和侧面划定一个分界线，并且假定人脸的左右两面是完全对称的；然后在模型正面的部分，就以 x 和 y 两个方向（坐标方向如图 1(c)所示）作为贴图坐标来从正面照片提取贴图纹理数据，而在模型侧面的部分则以 z 和 y 两个方向作为贴图坐标来从侧面照片提取贴图纹理数据[2]。

由于眼睛、睫毛、牙齿、舌头等处的贴图很难从照片当中提取，因此可采用从一些通用的眼球、牙齿的图片当中进行提取的方法。

3.2 正侧面边界反差的消除

由于为一个人制作模型时，使用的照片很可能受到拍摄环境的影响，其正面和侧面照片的光照条件不是完全一样，这样就造成了模型在分界处两侧的皮肤纹理的色调和亮度会有不一致的情况，因而在相互对比时，因正侧面分界处看起来反差很大而严重影响了模型的真实度（如图 4(a)所示）。

其解决的方法是，对边界处的顶点以及附近的顶点，再设置一个调节色，用于控制在该顶点处实际显示的纹理色彩。作此调整之后，就可使边界处顶点对正面和侧面显示的纹理的色彩达到一致，即达到了消除边界处反差的效果（图 4(c)左部）。

(a) 边界反差 (b) “假耳” (c) “说话头”模型最终合成效果

图 4 模型边界反差消除及“假耳”处理

3.3 “假耳”的处理

提取侧面贴图时，信息来自侧面的照片，而耳朵后面的皮肤和头发是被耳朵挡住的，如果仅仅取侧面的 z 和 y 坐标作为贴图坐标参考，则将导致耳朵部分的图像被重复使用，即会在耳朵后面又形成一个“假耳”（如图 4(b)所示）。

由于头像模型的后部很少显示出来，一般是处于隐藏状态，因此无需额外的图像信息来精确弥补被耳朵遮挡的部分，可采用“假耳”以外正确的皮肤和头发的纹理来插值模拟即可，即将落在“假耳”内的所有三角形的顶点的贴图坐标取为外部任意一个相邻顶点的贴图坐标，就可以消除“假耳”，这样就达到了可接受的模型合成效果（如图 4(c)右部耳朵后部位所示）。

经过上面处理，最终模型的合成效果如图 4(c)所示。

4 利用 FAP 参数驱动模型的口型和表情

4.1 人脸模型的参数定义

在 MPEG4 标准中，有面部定义参数（Facial Definition Parameter，缩写为 FDP）和面部动作参数 Facial Animation Parameter，缩写为 FAP）两组参数用来描述脸部的表情，感情和发音，和对脸部进行编码。面部定义参数（FDP）是用来描述通用人脸模型对应特定人脸的特征点，包括嘴角、嘴巴轮廓、

眉心、眼角等等。面部动作参数(FAP)是用于描述脸部动作和脸部肌肉运动的一组相关参数。每个 FAP 参数值代表一个特征点或一个区域偏离自然位置的程度。图 5 是 MPEG4 中定义的面部定义参数(FDP)特征点。这些参数经过归一化之后就确定了一个模型,而与模型的大小和形状无关[6,7]。

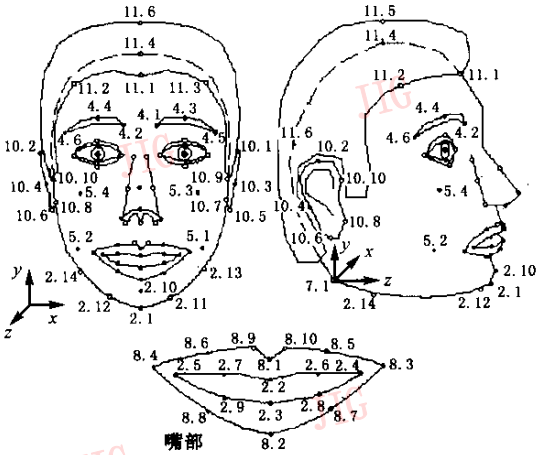

图 5　FDP 特征点

4.2　FAP 参数和模型顶点坐标变动的关系

本文采用的这个模型框架使用了 68 个 FAP 参数,其中 f_1 和 f_2 是矢量,属于高层 FAP 参数,将在下一节说明。余下的 66 个 FAP 参数 $f_3 \sim f_{68}$ 是标量,属于低层 FAP 参数,每个都对应人脸模型的一种动作,其参数的数值表示动作的强度,而这种动作的具体反映就是一个或一些顶点的坐标位置偏离其初始位置,比如 f_{20} 是控制左眼眨眼的参数,其数值用来控制模型左眼上下眼睑一些顶点运动做上下运动的距离。

低层 FAP 参数对模型运动的解释是随模型而不同的,即不同结构的模型,不仅其顶点数目、分布、序号对应部位不同,而且其顶点运动的计算公式也不同[8]。

4.3　表情(f_1)和口型(f_2)参数

在 68 个 FAP 参数当中,第 1 个(f_1)和第 2 个(f_2)是高层参数,分别用来表示表情(expression)和口型(visem),f_1 和 f_2 都是 66 维的 FAP 矢量,其分量就是 66 个低层参数 $f_3 \sim f_{68}$。

本文定义了 6 种基本表情:高兴(joy)、悲伤(sadness)、愤怒(anger)、恐惧(fear)、厌恶(disgust)、惊讶(surprise),以 1～6 编号。每种表情都对应于由一组低层参数 $f_3 \sim f_{68}$ 构成的一个 FAP 矢量,6 种基本表情就对应 6 个矢量,分别记为 $\boldsymbol{E}_1 \sim \boldsymbol{E}_6$(其设定将在 5.2 节中说明),再定义一个全零的 FAP 矢量 $\boldsymbol{E}_0$,表示无表情。其余复杂表情可以认为是这些基本表情中某两个表情的线性组合,因此,表情对应的 FAP 矢量 $\boldsymbol{f}_1$,可由两个基本表情对应的 FAP 矢量 $\boldsymbol{e}_1$、$\boldsymbol{e}_2$、和各自的强度值 s_1、s_2 来组合得到。$\boldsymbol{e}_1$、$\boldsymbol{e}_2$ 取值为 $\boldsymbol{E}_0 \sim \boldsymbol{E}_6$ 中的一个,s_1、s_2 取值范围为 0～100,它们关系如下:

$$\boldsymbol{f}_1 = (\boldsymbol{e}_1 \times s_1 + \boldsymbol{e}_2 \times s_2)/100 \qquad (1)$$

若 $\boldsymbol{e}_1 = \boldsymbol{E}_1$,$\boldsymbol{e}_2 = \boldsymbol{E}_6$,$s_1 = s_2 = 50$,则表示一半喜(joy)一半惊(surprise)(如图 6(a)所示)。

若 $\boldsymbol{e}_1 = \boldsymbol{E}_3$,$\boldsymbol{e}_2 = \boldsymbol{E}_0$,$s_1 = 100$,$s_2 = 0$,则表示愤怒(anger)(如图 6(b)所示)。

(a) 表情:惊喜　(b) 表情:发怒　(c) 口型:“ao”　(d) 口型:“f”

图 6　表情和口型合成效果示例

在中文语音合成系统当中,汉语拼音被分成了 15 大类,并将发音相近的声母/韵母分为同一类,复合韵母看成是单韵母、韵尾(n 和 ng)的组合,加上静音(闭口)一共是 16 种口型(见表 1)[7]。口型和基本表情一样,16 种基本口型也分别对应于由一组低层参数 $f_3 \sim f_{68}$ 构成的一个 66 维的 FAP 矢量,与这些基本口型对应的矢量,分别记为 $\boldsymbol{V}_0 \sim \boldsymbol{V}_{15}$(其设定将在 5.2 节中说明;表示静音的矢量 $\boldsymbol{V}_0$ 和 $\boldsymbol{E}_0$ 一样,

都是一个全零的 FAP 矢量)。在说话的过程中,由于任何任一时刻的口型可以认为是 16 种基本口型中的一个,或者其中两个的过渡,因此,与口型对应的 FAP 矢量 f_2 可以由两个基本口型对应的 FAP 矢量 v_1、v_2 和过渡系数 b 来组合得到。v_1、v_2 在 $V_0 \sim V_{15}$ 中取值,b 的取值在 0～100,有

$$f_2 = (v_1 \times b + v_2 \times (100 - b))/100 \quad (2)$$

例如:$v_1 = V_1$, $v_2 = V_2$, $b = 75$,则表示在韵母"*ao*"之间的一个口型(如图 6(*c*)所示)。

$v_1 = V_8$, $v_2 = V_0$, $b = 100$,则表示声母"f"的口型(如图 6(d)所示)。

表 1 汉语的口型分类

口型编号	发音
0	"sil"(静音)
1	a
2	o
3	e
4	i
5	u
6	v
7	b p m
8	f
9	d t n l
10	g k h
11	j q x
12	zh ch sh r
13	z c s
14	-n
15	-ng

5 "说话头"合成系统的实现

5.1 合成系统的架构

整个合成器由文本与处理器、文语转换(text-to-speech,简称 TTS)引擎、口型和表情 FAP 表、FAP 参数序列重采样和混合、动画生成、声音动画同步播放等几个模块组成(如图 7 所示)。

5.2 人的自然口型和表情的提取

对一个"说话头"合成系统,合成的最基本单元就是口型(f_2)和表情(f_1)。

对于表情,其中与 6 种基本表情对应的 FAP 参数 $E_1 \sim E_6$,可以通过人工主观调整的方法来设置获取,即可通过调整各 FAP 参数值来使模型显示的效果在视觉上贴近真实的表情,在达到了某一基本表情的标准之后,还需将其 $f_3 \sim f_{68}$ 参数值保存,以作为与这个基本表情对应的 FAP 矢量,由此得到 6 种与基本表情对应的 FAP 矢量 $E_1 \sim E_6$。

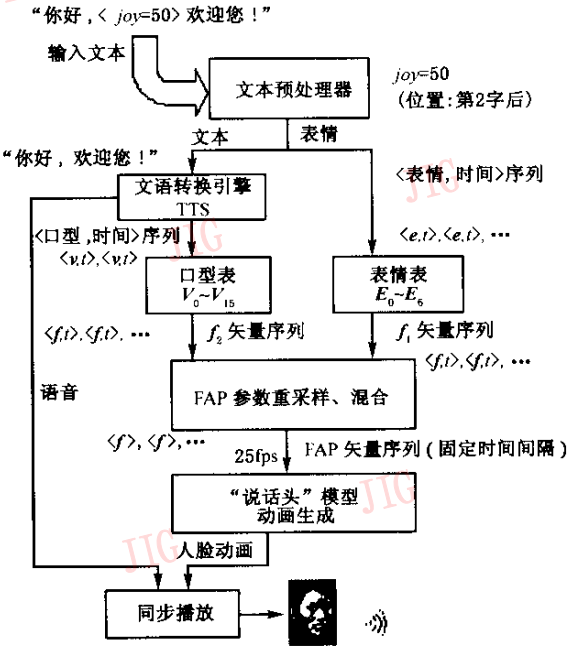

图 7 合成系统结构

对于 16 种基本口型,其 FAP 矢量 $V_0 \sim V_{15}$ 的定义则是由大量的数据统计而来,即先设计涵盖所有 16 种基本口型的句子和短语,然后录制一个人朗读这些句子的面部录像,同时从录像视频当中提取出基本口型的图片,再用检测程序检测出嘴部的高度、宽度、嘴角位置和嘴唇形状等参数,最后采用了建立并且训练隐马尔可夫模型(HMM)的方法,将各个口型映射到各自的一组 FAP 参数值上[9],这样就得到了与 16 个基本口型对应的 FAP 矢量 $V_0 \sim V_{15}$。

5.3 声音和动画的同步合成

整个"说话头"合成系统的前端输入是带有表情标签的文本,如:"你好,〈joy=50〉欢迎您!"。系统处理时,首先经过简单的文本预处理分离出表情,然后将纯文本送入文语转换(TTS)引擎,便得到合成的语音波形,同时,从文语转换(TTS)引擎中获取这段文本的拼音信息(包括声母/韵母)及其所持续的时间,其中韵母将拆分成单韵母和韵尾的组合,并将持续时间均分,如"iang"被分成"i","a","-ng",其持续时间都是"iang"的三分之一;接下来,声韵母通过查表得到 1～15 种基本口型。由此即可获得一个〈口型:时间〉的序列,再将前面分离出的表情,加上由其

通过文本当中的位置得到的时间信息来得到〈表情:时间〉的序列,如上面的例子,在第 2 个字结束这个时间上得到 $e_1=E_1(joy)$、$e_2=E_0$、$s_1=50$、$s_2=0$;最后,通过基本口型 FAP 矢量表 $V_0\sim V_{15}$和基本表情 FAP 矢量表 $E_0\sim E_6$,分别由式(2)和式(1)计算出与口型对应的 f_2 序列和与表情对应的 f_1 序列。

在合成动画时,通常固定一个播放动画时的帧速率,比如 25fps,这样两帧的时间间隔就是 40ms。在 FAP 参数重采样和混合模块中,则先对表情和口型序列在时间轴上做一个线性插值曲线,再按照 40ms 的间隔取值,并且把与口型和表情对应的 FAP 矢量 f_2 和 f_1 按照 40ms 的时间关系叠加在一起,就得到了合成这句话(包含口型和表情)需要播放的 FAP 矢量序列,即每帧都有 $f_3\sim f_{68}$这一组低层 FAP 参数值,其过程如图 8 所示,黑方块表示口型 f_2 某低层 FAP 参数分量 $f_i(i=3\sim 68)$的采样值,黑圆圈表示表情 f_1 的分量 f_i 的采样值,最终输出的 f_i 值为两者的和(用黑三角表示)。如此生成的低层 FAP 参数序列,就可以在时间上和语音同步吻合;然后动画生成器根据每帧的低层 FAP 参数来生成人脸动画,最后由同步播放模块播放语音和动画,以构成整个合成系统的输出。

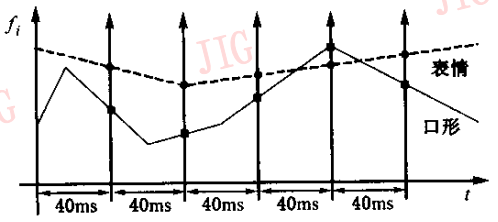

图 8　口型、表情的同步叠加

为了使合成的效果更加生动,本系统还为模型添加了一些定时的和随机的表情扰动,比如一段时间之后眨眼一次,并且在长时间没有合成语音的时候,随机做个脸部和嘴部的细微动作,这样就使得合成系统的表现更加人性化,更加逼真。

6　结　论

本文提出的"说话头"模型,以一个三维模型和人的正面侧面两张照片为基础,通过调整模型和提取照片纹理贴图等方法,首先使模型达到逼近真人的效果;而且由于该模型是基于三维模型的,因此它具有灵活度大的特点,它不仅可以随意变换生成口型和表情,还可以实时合成。可见该模型综合了基于图像库方法点和基于三维模型方法两者的优点。

本文使用面部动作参数(FAP)参数来驱动人脸模型的方案,其中包括表示表情(f_1)和口型(f_2)的两个高层 FAP 参数,它们都是由低层 FAP 参数 $f_3\sim f_{68}$组成的 FAP 矢量。

这种新的"说话头"模型用于视觉语音合成系统(Visual TTS),不仅使系统能够同步播放合成的语音和包括口型和表情的人脸动画,而且有比较逼真自然的性能。

目前,笔者正在模型的自然形变、生成更复杂精细的脸部动作等方面做进一步的研究,以便使"说话头"模型更加地自然和人性化。

参 考 文 献

1 Sumedha Kshirsagar, Nadia Magnenat-Thalmann. Virtual humans personified[A]. In: Proceedings of Autonomous Agents and Multi-agent Systems (AAMAS) Conference [C], Bologna, Italy, July, 2002: 356～357.

2 A. Roy Chowdhury, S. Krishnamurthy, T. Vo, *et al*. 3D face rconstruction from video using a generic model [A]. In: Proceedings of The IEEE International Conference on Multimedia & Expo(ICME)[C], Lausanne, Switzerland, August, 2002.

3 CHENG Chia-Ming, LAI Shang-hong. An integrated approach to 3D face model reconstruction from video [A]. In: Proceedings of IEEE International Conference of Computer Vision (ICCV) Workshop on RATFG-RTS'01 [C], Vancouver, Canada, August, 2001: 16～22.

4 Žiga Kranjec, Franc Solina. Building animated 3D face models from range data[A]. In: Proceedings of Electrical and Computer Science Conference ERK 2000 [C], Portoroz, Slovenia, September 2000,**B**: 193～196.

5 Volker Blanz, Thomas Vetter. A morphable model for the synthesis of 3D faces[A]. In: Proceedings of SIGGRAPH '99 [C], Los Angeles, CA, USA, August 1999: 187～194.

6 Sumedha Kshirsagar, Nadia Magnenat-Thalmann. Viseme space for realistic speech animation[A]. In: Proceedings of Audio-Visual Speech Processing (AVSP) 2001[C], Aalborg, Denmark, September 2001: 30～35.

7 Sumedha Kshirsagar, Marc Escher, Gael Sannier, *et al*. Multimodal animation system based on the MPEG-4 standard [A]. In: Proceedings of Multimedia Modeling '99[C], Ottawa, Canada, October 1999: 215～232.

8 WANG Zhi-ming, CAI Lian-hong, AI Hai-zhou. A dynamic viseme model for personalizing a talking head [A]. In: Proceedings of the International Conference on Signal Processing (ICSP) 2002[C], Beijing, China, August 2002: 29～34.

9 SUN Ling, LAI Wei, WANG Ren-hua. Face synthesis driven by audio speech input based on HMMs [A]. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP 2002) [C], Taipei, Taiwan of China, August 2002.

赖 伟 1982 年生，中国科学技术大学电子工程与信息科学系在读博士生，2000 年和 2002 年先后获中国科学技术大学电子工程学士学位和硕士学位。现主要研究领域为多模态智能人机接口、多媒体内容的分析处理。
E-mail：laiwei@ustc.edu

孙 岭 1977 年生，2000 年获中国科学技术大学电子工程学士学位，2003 年获中国科学技术大学信息与通信工程学科工学硕士学位，现主要研究领域为嵌入式系统。

王仁华 1943 年生，中国科学技术大学教授，博士生导师，主要研究领域为信号与信息处理、多媒体通信。