[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04N 5/265 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200510093269.X

[43] 公开日 2006 年 2 月 8 日

[11] 公开号 CN 1731833A

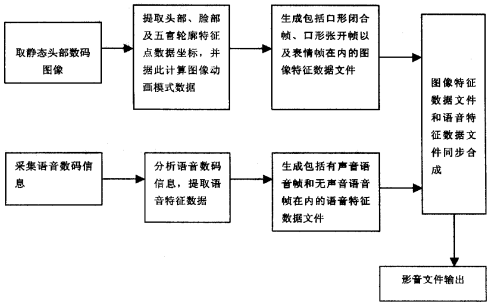
[22] 申请日 2005.8.23
[21] 申请号 200510093269.X
[71] 申请人 孙 丹
地址 100044 北京市西直门北大街 41 号天兆家园 4C501
共同申请人 王维国
[72] 发明人 孙 丹 王维国
[74] 专利代理机构 北京连和连知识产权代理有限公司
代理人 王 昕
权利要求书 3 页 说明书 11 页 附图 6 页

[54] 发明名称

语音驱动头部图像合成影音文件的方法

[57] 摘要

本发明涉及一种语音驱动头部图像合成影音文件的方法。它包括取静态头部数码图像；提取头部、脸部及五官轮廓特征点数据坐标，并计算图像动画模式数据；生成包括口形闭合帧、口形张开帧、表情帧以及三种类型图像帧的各自数目在内的图像特征数据文件；采集语音数码信息；分析、提取语音特征数据，生成包括有声音语音帧和无声音语音帧在内的语音特征数据文件；将图像特征数据文件和语音特征数据文件同步合成，使得有声音语音帧对应口形张开帧，无声音语音帧对应口形闭合帧，然后输出。本发明具有图像逼真、可实现人脸及动物脸部表情的动画，制作方便简单，便于推广使用等特点。



1. 一种语音驱动头部图像合成影音文件的方法，其特征是包括下列步骤：

步骤 1、取静态头部数码图像；

步骤 2、提取头部、脸部及五官轮廓特征点数据坐标，并计算图像动画模式数据，用以确定口形闭合帧帧数、口形张开帧帧数以及表情帧帧数；

步骤 3、生成包括口形闭合帧、口形张开帧、表情帧以及三种类型图像帧的各自数目在内的图像特征数据文件；

步骤 4、采集语音数码信息；

步骤 5、分析语音数码信息，提取语音特征数据；

步骤 6、生成包括有声音语音帧和无声音语音帧在内的语音特征数据文件；

步骤 7、上述图像特征数据文件和语音特征数据文件同步合成，使得有声音语音帧对应口形张开帧，无声音语音帧对应口形闭合帧，然后输出。

2、根据权利要求 1 所述的方法，其特征是步骤 2 进一步包括下列步骤：

步骤 21、根据头部、脸部及五官轮廓特征点数据坐标计算脸部动画模式数据，进而获取典型的头部、脸部及五官动画模式数据，将所获取的整个头部动画模式数据按照一定角度进行偏转，用以生成图像特征数据文件中摇头的表情帧。

3、根据权利要求 2 所述的方法，其特征是步骤 2 还进一步包括下列步骤：

步骤 22、取五官轮廓特征点数据坐标中眼睛特征点数据，用眼睛上下

边缘图像覆盖眼睛中间图像，用以实现眨眼睛的表情帧。

4、据权利要求 3 所述的方法，其特征是步骤 2 还进一步包括下列步骤：

步骤 23、取五官轮廓特征点数据坐标中嘴部特征数据，将嘴部轮廓特征点数据坐标上下左右向中间部位移动或向四周扩张，用以实现口形张开帧和口形闭合帧。

5、根据权利要求 1 所述的方法，其特征是步骤 5 中的分析语音数码信息，提取语音特征数据进一步包括下列步骤：

步骤 51、读取语音数码信息，判断语音数码信息文件是否终止，如是，结束本程序，并生成语音特征数据文件；如否，则执行步骤 52；

步骤 52、通过分析大量语音数码信息文件，进行有声音语音帧和无声音语音帧的过零率计算，用以确定有声阈值；

步骤 53、进行噪音修正以及过零有效修正；

步骤 54、过零率判断，用以确定有声音语音帧和无声音语音帧是否在有声阈值内，如是，输出为“1”，如否，输出为“0”；

步骤 55、将输出的“0”或“1”表示的无声音语音帧和有声音语音帧分别放入缓存区并重新开始读取语音数码信息。

6、根据权利要求 5 所述的方法，其特征是上述有声阈设为 100~728。

7、根据权利要求 6 所述的方法，其特征是上述有声阈值中女声和童声的有声阈值上限设为 690~725。

8、根据权利要求 6 所述的方法，其特征是上述有声阈值中男声的有声阈值上限设为 710~730。

9、根据权利要求 1 所述的方法，其特征是步骤 7 所述图像特征数据文件和语音特征数据文件同步合成，进一步包括下列步骤：

步骤 71、读取图像数据特征文件中口形张开帧数或口形闭合帧数；

步骤 72、读取语音特征数据文件中无声音语音帧或有声音语音帧；

步骤 73、判断语音特征数据是否有变化，如否，重新读取语音特征数据；如是，当语音特征数据变为无声音语音帧（即数值为 0 时），则进行步骤 74；如语音特征数据变为有声音语音帧（即数值为 1 时），则进行步骤 75；

步骤 74、当语音特征数据文件中连续的无声音语音帧的帧数小于 4 时，按照误差处理；当语音特征数据文件中连续的无声音语音帧的帧数大于图像数据特征文件中表情帧数+（20±5）时，在口形闭合帧中合理插入表情帧；当语音特征数据文件中连续的无声音语音帧的帧数为上述以外的值时，直接对应图像数据特征文件的口形闭合帧。

步骤 75、当语音特征数据文件中连续的有声音语音帧的帧数小于 4 时，按照误差处理；当语音特征数据文件中连续的有声音语音帧的帧数大于或等于 4 并且小于或等于两倍的口形张开帧数时，选取部分口形张开帧与该有声音语音帧对应；当语音特征数据文件中连续的有声音语音帧的帧数大于两倍的口形张开帧数时，按帧数循环输出口形张开的图像序列，如果是最后一帧图像，则采用闭合口形帧图像。

语音驱动头部图像合成影音文件的方法

技术领域

本发明涉及一种语音与图像合成影音文件的方法，特别是涉及一种语音驱动头部图像合成影音文件的方法。

背景技术

语音信号和视觉信号是人类进行学习和交流的重要手段，而综合利用语音信号和视觉信号使人们对信息的理解和掌握更加快捷、方便。语音驱动图像正是这样一种综合考虑声音和图像的多媒体技术。语音驱动图像是指用语音来驱动人们在用语言交流时所表达出的口形变化以及面部表情和动作，它能在一定程度上传达人们想要表达的意思，并能帮助人们加深对语言的理解。在计算机的人机交互的过程中或者在第三代移动通信应用中，如果人们面对的是一个会说话的人物形象，则使人觉得界面更为友善，方便人们的交流。中国专利申请 02140286 中，公开了一种“基于统计与规则结合的语音驱动人脸动画方法”，该方法是：预先得到人脸的动态音频和视频，运用统计和自学习的方法分析出人脸的运动参数并建立人脸运动模型，然后对语音和人脸特征点之间的关联模式进行统计学习。当给定新语音，利用学习到的模型以及一些规则，可以得到与该语音对应的人脸特征点运动参数，驱动人脸动画模型。这种方法有三个方面的局限性：一是必须预先得到人脸的视频（即动态图象），也就是说无法根据单一的静态图像进行处理；二是需要进行统计分析和建模，即要建立一个庞大的数据库，投入人力物力大，制作周期长，不便于普遍推广；三是对于动物脸的动态数据采集困难，其动画效果难以实现。

发明内容

为克服上述已有技术中存在的不足，本发明的目的是提供一种对单一的静态图像进行处理、即通过语音与图像合成技术使该静态图像中的人物或动物具有面目表情的简单易行的方法。

为实现上述目的，本发明提出一种语音驱动头部图像合成影音文件的方法包括下列步骤：

步骤 1、取静态头部数码图像；

步骤 2、提取头部、脸部及五官轮廓特征点数据坐标，并计算图像动画模式数据，用以确定口形闭合帧帧数、口形张开帧帧数以及表情帧帧数；

步骤 3、生成包括口形闭合帧、口形张开帧、表情帧以及三种类型图像帧的各自数目在内的图像特征数据文件；

步骤 4、采集语音数码信息；

步骤 5、分析语音数码信息，提取语音特征数据；

步骤 6、生成包括有声音语音帧和无声音语音帧在内的语音特征数据文件；

步骤 7、所述图像特征数据文件和语音特征数据文件同步合成，使得有声音语音帧对应口形张开帧，无声音语音帧对应口形闭合帧，然后输出。

上述步骤 2 进一步包括下列步骤：

步骤 21、根据头部、脸部及五官轮廓特征点数据坐标计算脸部动画模式数据，进而获取典型的头部、脸部及五官动画模式数据，将所获取的整个头部动画模式数据按照一定角度进行偏转，用以生成图像特征数据文件中摇头的表情帧。

步骤 22、取五官轮廓特征点数据坐标中眼睛特征点数据，用眼睛上下边缘图像覆盖眼睛中间图像，用以生成图像特征数据文件中眨眼睛的表情帧。

步骤 23、取五官轮廓特征点数据坐标中嘴部特征数据，将嘴部轮廓特征点数据坐标上下左右向中间部位移动或向四周扩张，用以生成图像特征数据文件中口形张开帧和口形闭合帧。

上述步骤 5 中进一步包括下列步骤：

步骤 51、读取语音数码信息，判断语音数码信息文件是否终止，如是，结束本程序，并生成语音特征数据文件；如否，则执行步骤 52；

步骤 52、通过分析大量语音数码信息文件，进行有声音语音帧和无声音语音帧的过零率计算，用以确定有声阈值；

步骤 53、进行噪音修正以及过零有效修正；

步骤 54、过零率判断：设定有声阈值为 100~728,其中女声和童声有声阈值上限为 690~725 之间，男声有声阈值上限为 710~730 之间。用以确定有声音语音帧和无声音语音帧是否在有声阈值内，如是，输出为“1”，如否，输出为“0”；

步骤 55、将输出的“0”或“1”表示的无声音语音帧和有声音语音帧分别放入缓存区并重新开始读取语音数码信息。

上述步骤 7 进一步包括下列步骤：

步骤 71、读取图像数据特征文件中口形张开帧数或口形闭合帧数；

步骤 72、读取语音特征数据文件中无声音语音帧或有声音语音帧；

步骤 73、判断语音特征数据是否有变化，如否，重新读取语音特征数据；如是，则当语音特征数据变为无声音语音帧（即数值为 0 时），则进行步骤 74；如语音特征数据变为有声音语音帧（即数值为 1 时），则进行步骤 75；

步骤 74、当语音特征数据文件中连续的无声音语音帧的帧数小于 4 时，按照误差处理；当语音特征数据文件中连续的无声音语音帧的帧数大于图像数据特征文件中表情帧数+（20±5）时，在口形闭合帧中合理插入表情帧；当语音特征数据文件中连续的无声音语音帧的帧数为上述以外的值时，直接对应图像数据特征文件的口形闭合帧。

步骤 75、当语音特征数据文件中连续的有声音语音帧的帧数小于 4 时，按照误差处理；当语音特征数据文件中连续的有声音语音帧的帧数大于或等于 4 并且小或等于两倍的口形张开帧数时，选取部分口形张开帧与

该有声音语音帧对应；当语音特征数据文件中连续的有声音语音帧的帧数大于两倍的口形张开帧数时，按帧数循环输出口形张开的图像序列，如果是最后一帧图像，则采用闭合口形帧图像。

为让本发明之上述和其它目的、特征和优点能更明显易懂，下文特举较佳实施例，并配合附图，作详细说明如下。

附图说明

图 1 为本发明所述方法的示意图；

图 2 为一种提取头部轮廓特征点数据坐标方法示意图；

图 3 为一种提取脸部轮廓特征点数据坐标方法示意图；

图 4 为一种提取五官轮廓特征点数据坐标方法示意图；

图 5 为另一种提取五官轮廓特征点数据坐标方法示意图；

图 6 为生成图像特征数据文件流程图；

图 7 为生成语音特征数据文件流程图；

图 8 为图像特征数据文件与语音特征数据文件合成影音文件方法的流程图。

具体实施方式

图1为本发明所述方法的示意图，在实际运用中，本发明可采取如下步骤：

(1) 获取静态头部数码图像，例如人物头像或动物头像：图像可以是通过数码相机、扫描仪等方式获取的照片、图片，分辨率最好在800x600以上，图像清晰，以正面图像为好、头部突出，表情自然，格式可以为BMP、JPG、GIF等。BMP，JPG、GIF都是计算机和数码像机上常用和通用的图像存储格式。

(2) 图像处理，利用图像跟踪技术，对图像进行预处理，选取图像轮廓

特征点数据坐标：头部图像的处理和勾画方法如图2所示，例如选取头部轮廓图像中4个特征点数据坐标。脸部图像的处理和勾画方法如图3所示，可通过人工勾画或计算机边缘处理和边缘识别的方法勾勒出头脸部，其余部分作为背景，例如选取头脸部轮廓图像中4个特征点数据坐标；五官轮廓图像的处理和勾画方法如图4、图5所示，例如可选取五官轮廓图像中2~6个特征点数据坐标。

（3）图像特征提取，即提取上述图像轮廓特征点数据坐标，用以生成图像特征数据文件，其提取流程如图6所示。利用人类视觉特点以及图像处理中的检测技术,可以将图像中的头部图像轮廓特征点数据坐标值以及脸部图像轮廓特征点数据坐标从图像中提取出来。具体方法如下：

——颜色分离：通常得到的图片均为彩色图片，即每个点包含RGB三种颜色，我们在处理时需要把它转化成YUV彩色空间，转化公式是：

$$Y = (0.257 * R) + (0.504 * G) + (0.098 * B) + 16$$

$$U = -(0.148 * R) - (0.291 * G) + (0.439 * B) + 128$$

$$V = (0.439 * R) - (0.368 * G) - (0.071 * B) + 128$$

Y 表示亮度，U 和 V 表示色度和饱和度，在处理时，我们仅处理亮度信息。

——去噪处理：是通过滤波等手段进行平滑处理以去除噪声，

通常采用中值滤波，这是一种最基本的图像处理算法，可在任何一本图像处理的书中得到。

——微分运算：是利用微分算子进行图像边缘检测，可采用的算子有 Laplace（拉普拉兹） 算子、Sobel（索贝尔）算子等，拉普拉兹算子是 2 阶微分算子，也就是说，相当于求取 2 次微分，它的精度还算比较高，但对噪声过于敏感(有噪声的情况下效果很差)是它的重大缺点，所以这种算子并不是特别常用。索贝尔算子是最常用的算子之一(它是一种一阶算子)，方法简单效果也不错，但提取出的边缘比较粗，要进行细化处理。在这里我们选用效果比较理想的 Sobel 算子，模板大小为 3X3，该算子可在任何一本图像处理的书中得到。

——二值化处理：即是对处理后的图像进行阈值操作，可先进行直方图分析，找到分界阈值，然后对图像中象素值大于阈值的象素取值为 1，否则为 0。

——参数计算：主要计算图像中各线段或图形的长度、面积、重心等参数，计算方法是进行象素的累加。这样就得到了头像中的各图形（如头部、脸部、眼睛、嘴巴等）的特征数据。

（4）图像动画模式：根据图像特征点坐标数据可以计算出脸部动画参数数据，从而获取典型的头部及脸部动画模式：摇头即将整个头部（第3步已经获得了头部特征和其他各有关部位特征）按照一定的角度进行偏转；眨眼即在眼睛特征范围内，用眼睛上下边缘图像覆盖眼睛中间图像（一般认为未经任何处理的头部图像的眼睛是睁开的），嘴巴张开和闭合的实现可通过嘴巴特征中上下侧向中间部位移动来实现。并计算图像动画模式数据，用以确定口形闭合帧帧数、口形张开帧帧数以及表情帧帧数；生成包括口形闭合帧、口形张开帧、表情帧以及三种类型图像帧的各自数目在内的图像特征数据文件；

（5）采集语音数码信息：语音可以通过录音设备或文本语音的转换技术获得，比如说语音格式可采用为WAVE，也可以是PCM (Pulse Code Modulation脉冲编码调制)，AAC (Advanced Audio Coding高级音频编码) ,MP3, AMR (Adaptive Multi-Rate，适应多比率)等。

（6）分析语音数码信息，提取语音特征数据：语音数码信息分析的方法包括时域分析和频域分析等，主要通过线性预测、过零率分析、傅立叶变换、小波变换、时频分析等技术对语音数码信息进行分析。

（7）语音特征数据提取，其步骤如图 7 所示，通过对语音数据进行分析，可以获得语音特征参数：如能量、基频、功率谱等。

语音特征数据提取与选择是语音识别的一个重要环节。语音特征数据提取主要解决时域语音信号的数字表示问题，提取与选择的好坏直接影响到最后影音同步的效果。

语音信号的特征主要有时域和频域两种。时域特征如短时平均能量、短时平均过零率、共振峰、基音周期等；频域特征有傅立叶频谱等。现在还有结合时间和频率的特征，即时频谱，充分利用了语音信号的时序信息。

每一帧信号所对应的时域参数有，第每一帧信号所对应的时域参数有，第t 帧语音的短时平均能量为

$$Eng(t)=\frac{1}{N}\sqrt{\sum_{n=0}^{N-1}S_t^2(n)} \tag{1}$$

或

$$Eng(t)=\frac{1}{N}\sum_{n=0}^{N-1}|S_t(n)| \tag{2}$$

其中 N 为分析窗的宽度，St(n)表示第 t 帧中第 n 个点的信号样值。短时平均过零率（Zero-Crossing-Rate，以下简称 ZCT）为

$$ZCT(t)=\sum_{n=0}^{N-1}\frac{1}{2}[Sgn(S_t(n)\cdot S_t(n-1))+1] \tag{3}$$

其中符号函数定义为

$$\begin{cases}\mathrm{Sgn}(x)=1, & x>0\\ \mathrm{Sgn}(x)=0, & x<0\end{cases} \tag{4}$$

目前时域参数（能量和过零率）多用在语音的端点检测，判断语音的开始与结束上。而能量的使用多利用其对数值或把能量的包络作为参数使用。在我们的语音识别方法中，以短时过零率这个时域参数来作为例子。首先对大量语音文件进行人工分析，结合语音波形文件，分别统计有声语音帧和无声语音帧的 ZCT 数值，可以知道当该语音帧有声音时，ZCT 往往会处于某个范围内，称之为有声阈值，无声语音帧则反之。流程通过分析每个语音帧 ZCT 值是否位于有声阈值内，自动判断是否有声。

短时过零率由于各人特点和说话环境的不同会由不小的变化，因此需要在应用的时候加入辅助判断，提高其精确性。由于时域参数不同于频域

参数，不能直接判断出噪音，所以识别程序中还需要增加有除噪环节和过零准确性修正环节。通过辅助的修正判断提高了程序的准确性。

有声阈值的确定，是通过统计得出的。针对不同环境，不同年龄，不同性别的发音人，要提出一个通用的阈值，并不容易。以 PCM 格式的语音文件处理为例，对大量 PCM 音频文件的分析结果表明，声音越高，有声部分的 ZCT 值就越低。男声的有声阈值较大，跨度也大，女声和童声的有声阈值较小，跨度也小。在增加了两个修正环节，对 PCM 波形进行滤波以后，将有声阈值定在 100～728。其中下限的设置影响不大，设为 100 是为了能够将无关信号造成的 ZCT 无序变动的影响消除。而女声和童声的 ZCT 有声阈值上限在 690～725 之间，无声区在 725 以上，若某发音人的 ZCT 有声阈值上限为 710，一般而言，往往很少有无声段 ZCT 统计值位于 710～725 这一个区段内。意即将有声阈值上限设为 725，对女声和童声基本是没有什么问题。而男声的 ZCT 有声阈值上限往往在 710～730 之间。这样男声的有声段和女声的无声段就有一个重叠，即 725～730 这一段。这个重叠区会对最后的判断结果造成误差。好在这个区域很窄。因此将有声阈值上限设一个居中值 728，判断有一定误差，经过测试在 4% 以内。是可以接受的。效果测试还证明取上限值为 728 的效果要稍好于取 727 为阈值上线。

除噪修正：前面提到，采用时域信号进行识别，并不能从分析结果中直接获得音频特征的全部具体信息，需要对其特征进行辨识。

无效音判断：PCM 是将声音采样值转化为二进制数据进行存储的，存在一个有效范围。以 16-bit 采样而言，其采样值在 0~65535 之间。在语音识别程序中，当读取的采样值>=65485 或<=50 时，均视为无效值，若这个采样点与前点或后点形成过零，同样视为无效。

噪音修正：对于有时某些接近于静音的音频段，由于录音设备或是周围环境的干扰，都可能造成采样点在 0 点附近浮动的，在此设置了两个环节处理这样，一是若过零量太小，视为过零无效，再有若过零存在，但是

前后两个采样值距离太近，视为过零无效。这样可以有效提高了语音识别的准确性。

（8）语音特征序列：根据语音特征对语音数据进行重新分类，生成包括有声音语音帧和无声音语音帧在内的语音特征数据文件；从而形成新的语音特征数据序列。

（9）图像特征数据文件和语音特征数据文件同步合成，使得有声音语音帧对应口形张开帧，无声音语音帧对应口形闭合帧。根据分析得到的语音特征文件和图像特征文件，通过音视频合成算法得到一个与语音特征数据文件相对应的图像序列，保证在有声音语音帧对应口形张开的图像帧，无声音的语音帧对应口形闭合的图像帧。对于没有声音的语音帧用口形闭合的图像与之对应，口形张开的图像帧选择口形连续变化的一系列图像帧，将这些图像帧合理安排对应一段有声音的语音信号段，这样在连续播放图像序列的时候可以保证口形变化的连续性和平滑性；同时考虑到语音特征的自动判断必然存在一定的误差，没有声音的地方可能误判为有声音，因此需要加入误差判断机制，即找到一个合适的阈值，如果有声音的语音段大于这个值才按有声音进行图像的对应，否则，认为是判断误差，按没有声音处理。如果图像特征序列存在表情帧，则在连续多帧语音数据文件没有声音的时候，适当地插入表情帧，这样可以使形象更加逼真。

根据输入的语音特征分析文件以及图像特征文件，经过分析处理得到与输入语音特征相对应的一个新的图像序列，将原始语音和该图像序列合成后可以得到一个口形与声音对应的影音文件。

如图 8 所示，将生成的语音特征数据文件及图像特征数据文件作为输入。语音特征数据文件仅由 0 和 1 组成，其中可以“0”代表无声音语音帧，“1”代表有声音语音帧。图像特征数据文件由三个部分组成，口形闭合帧，口形张开帧以及表情帧，并且三种类型图像帧的各自数目存放在文件的开头。其中表情帧可包括摇头的表情帧，眨眼睛的表情帧等。因此图像特征文件的数据格式为：口形闭合帧数，表情帧帧数，口形张开帧数，口形闭合图像帧数据，表情帧图像数据，口形张开图像帧数据。本发明最重要的

一点在于要实现口形与声音的对应，即在有声音的语音帧对应口形张开的图像，在没有声音的语音帧对应口形闭合的图像，而实现难点在于如何保持口形变化的连续性，从而达到一个比较好的合成效果。因此，在读取语音特征分析结果的同时，需要计算连续 1 或 0 的个数，即在某一段连续有声音语音帧或连续无声音语音帧的帧数。

当出现某一段连续无声音语音帧时，可以分为三种情况来处理：

- 连续无声音语音帧的帧数<4 时：此种情况下说明没有声音的语音段小于 0.3s，而实际语音中没有声音的时间段明显长于 0.3s，因此，认为此段分析结果为误差，按有声音处理。考虑到前一段语音同样为有声音，如“0”代表无声音语音帧，“1”代表有声音语音帧，所以从前一段语音开始，重新计算 1 的个数，并重新安排对应图像帧的输出。
- 连续无声音语音帧的帧数>表情帧+（20±5）时，若表情帧不等于零(存在表情帧)时：此种情况下将表情帧循环安插在无声音语音帧的语音段内。
- 连续无声音语音帧的帧数为其它数值时：此种情况下，语音帧全部对应口形闭合的图像。

当出现某一段连续无声音语音帧时，同样可以分为三种情况来处理：

- 连续为 1 的语音帧的帧数小于 4 时，同样认为是分析误差，按 0 处理，对应口形闭合的图像帧。
- 连续为 1 的语音帧的帧数大于或等于 4 且小于两倍的口形连续变化的图像帧数，此种情况下选取口形张开的图像中的一部分来和语音帧对应。
- 连续为 1 的语音帧的帧数大或等于两倍的口形连续变化的图像帧数时，此种情况下，按帧数循环输出口形张开的图像序列，考虑到序列最后一帧图像的口形不一定与口形闭合图像过渡平滑，因此可以采用闭合口形。

（10）MPEG4 （Moving Picture Expert Group）压缩：在合成好的影音文件中视频可以是YUV格式，音频可以是WAVE格式，采用MPEG4压缩技术对影音文件进行压缩以降低其对存储介质的需求。MPEG4是目前通用的计算机及数码设备采用的视频存储格式。

（11）输出：为满足在3G中应用的需求，可以对MPEG4压缩后的文件按照3GPP标准格式进行封装。

本发明具有影音效果逼真、制作方便简单，便于推广使用等特点。

虽然本发明已以较佳实施例披露如上，然其并非用以限定本发明，任何所属技术领域的技术人员，在不脱离本发明之精神和范围内，当可作些许之更动与改进，因此本发明之保护范围当视权利要求所界定者为准。

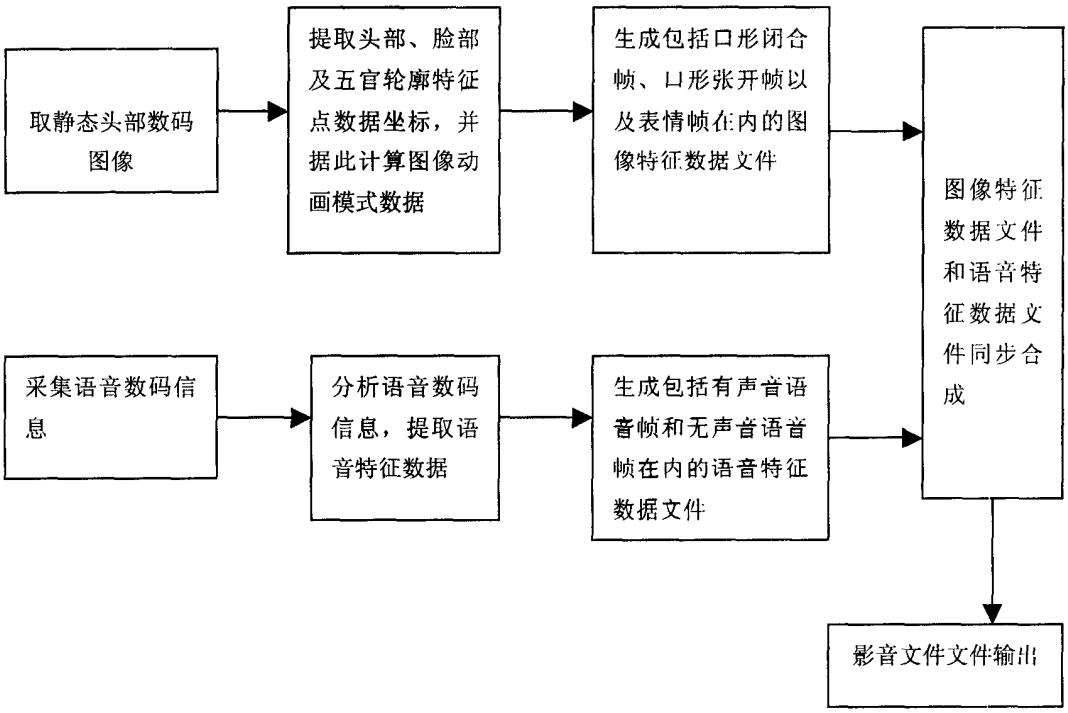

取静态头部数码图像
提取头部、脸部及五官轮廓特征点数据坐标，并据此计算图像动画模式数据
生成包括口形闭合帧、口形张开帧以及表情帧在内的图像特征数据文件
图像特征数据文件和语音特征数据文件同步合成
采集语音数码信息
分析语音数码信息，提取语音特征数据
生成包括有声音语音帧和无声音语音帧在内的语音特征数据文件
影音文件文件输出

图 1

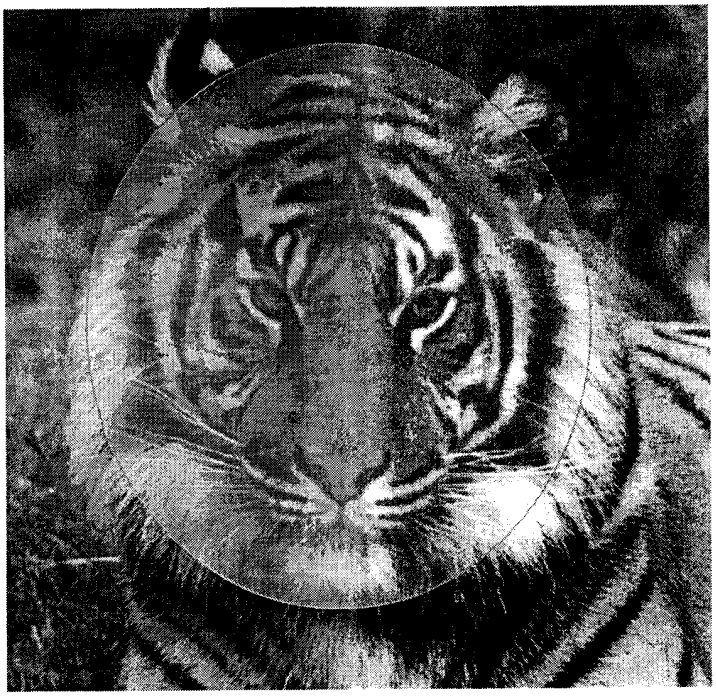


图 2

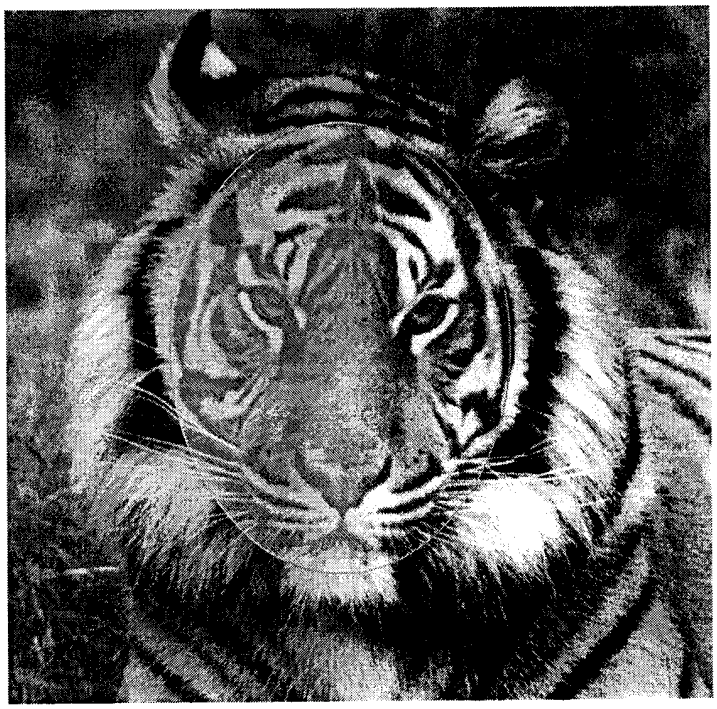


图 3

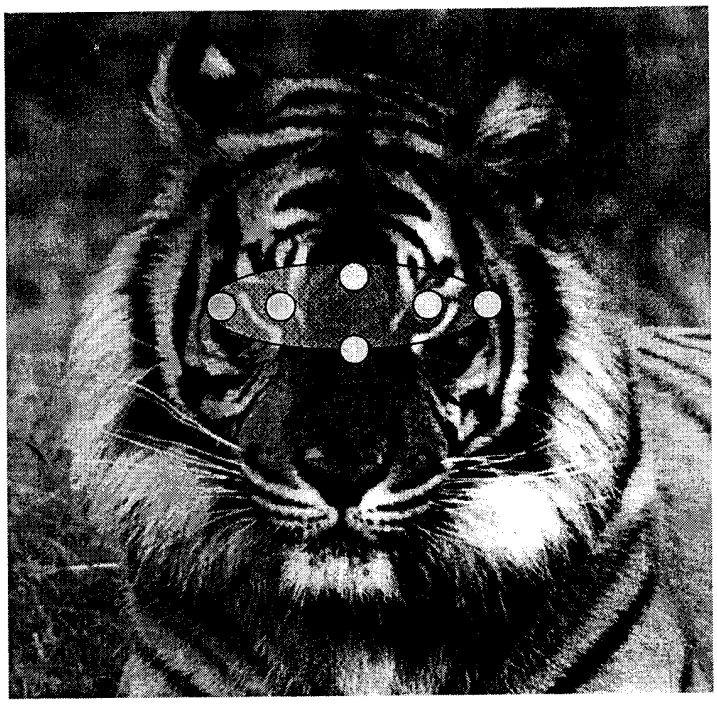


图 4

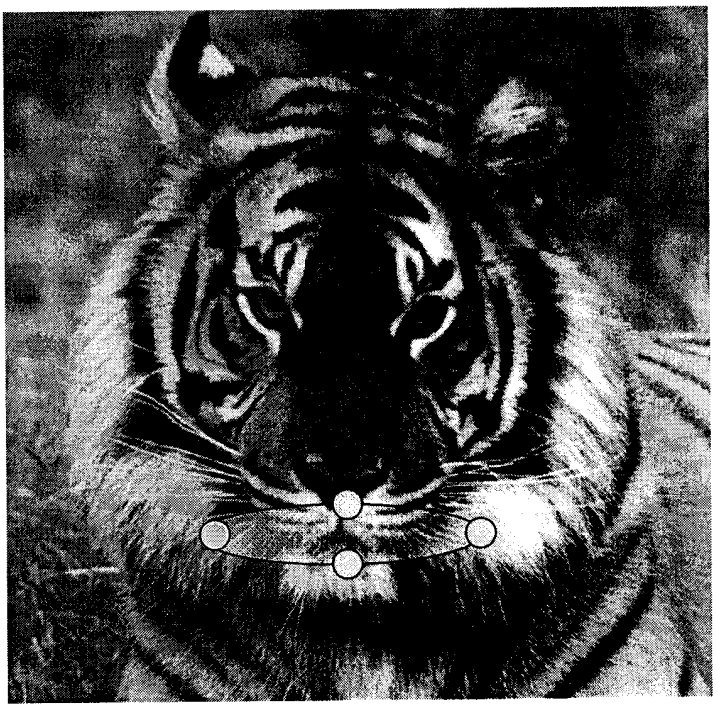


图 5

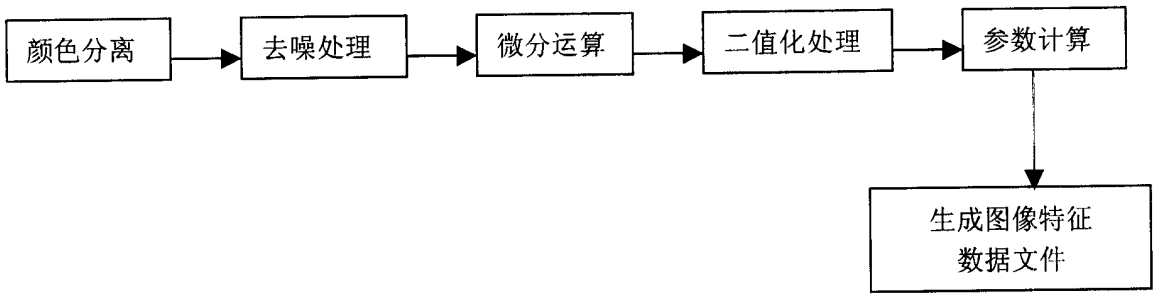


图6

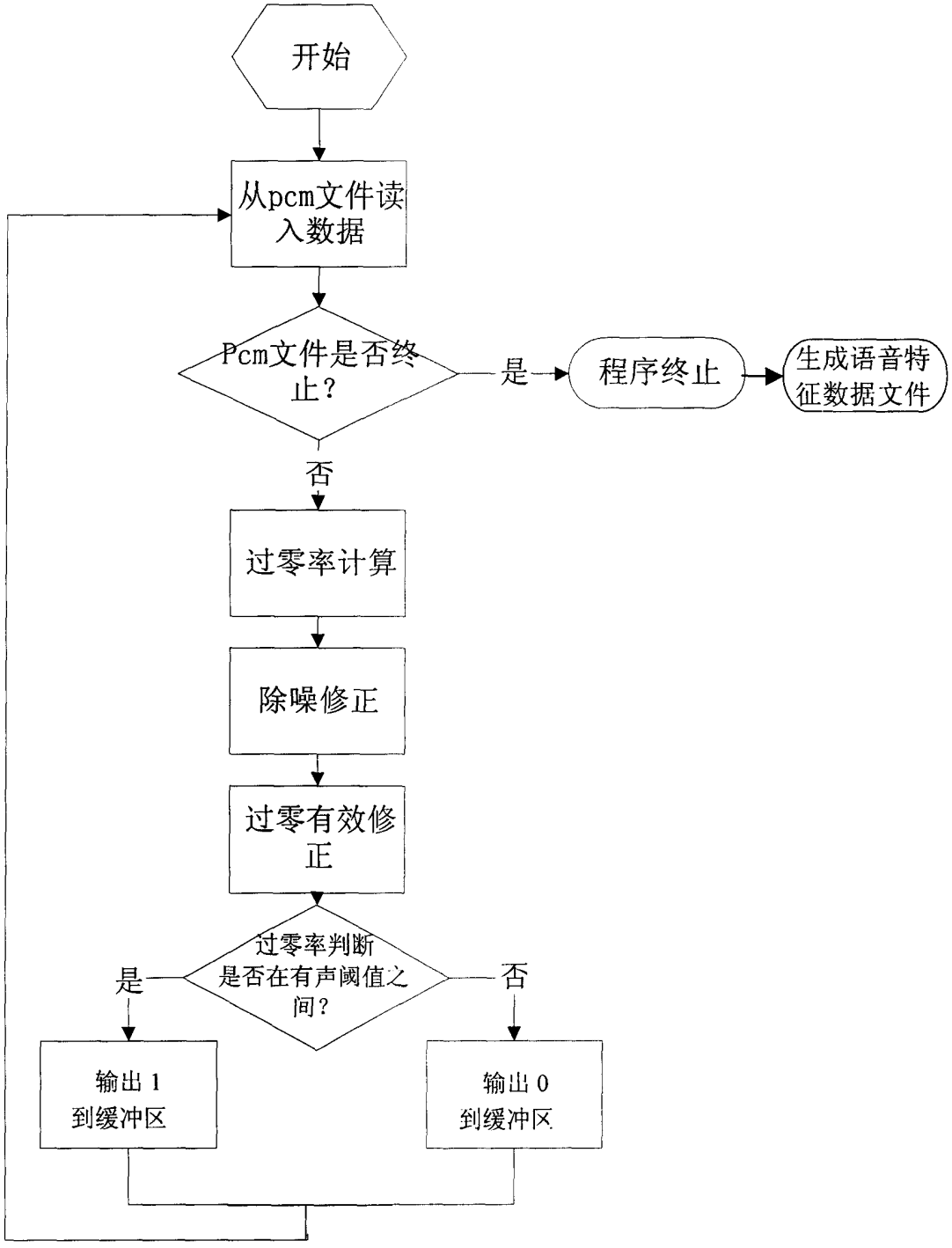

开始
从pcm文件读入数据
Pcm文件是否终止?
是
程序终止
生成语音特征数据文件
否
过零率计算
除噪修正
过零有效修正
过零率判断是否在有声阈值之间?
是
否
输出 1 到缓冲区
输出 0 到缓冲区

图 7

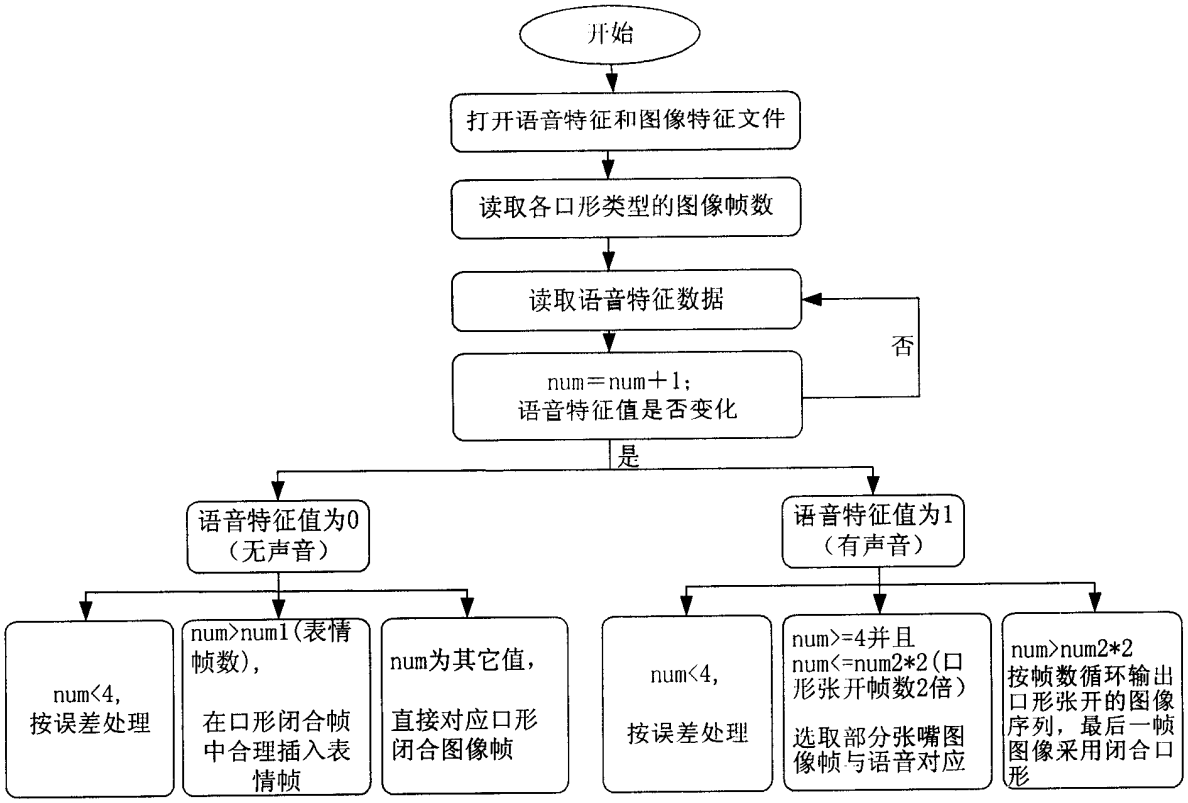

开始
打开语音特征和图像特征文件
读取各口形类型的图像帧数
读取语音特征数据
否
num＝num＋1；
语音特征值是否变化
是
语音特征值为0
（无声音）
语音特征值为1
（有声音）
num<4,
按误差处理
num>num1(表情帧数),
在口形闭合帧中合理插入表情帧
num为其它值，
直接对应口形闭合图像帧
num<4,
按误差处理
num>=4并且num<=num2*2(口形张开帧数2倍)
选取部分张嘴图像帧与语音对应
num>num2*2
按帧数循环输出口形张开的图像序列，最后一帧图像采用闭合口形

图8