

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G10L 21/06 (2006.01)

G10L 15/00 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200510121718.7

[43] 公开日 2007 年 7 月 4 日

[11] 公开号 CN 1991981A

[22] 申请日 2005.12.29

[21] 申请号 200510121718.7

[71] 申请人 摩托罗拉公司

地址 美国伊利诺斯

[72] 发明人 祖漪清 黄建成 王开志

[74] 专利代理机构 中国国际贸易促进委员会专利商
标事务所

代理人 吴丽丽

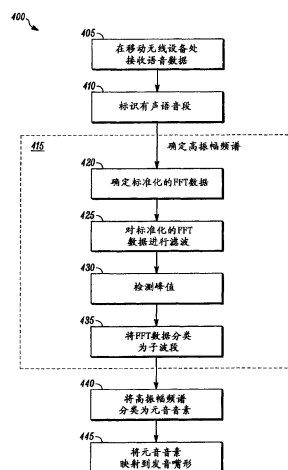
权利要求书 2 页 说明书 8 页 附图 4 页

[54] 发明名称

对语音数据进行分类的方法

[57] 摘要

一种用于对实时语音数据进行分类的计算上不密集的方法可用于改善虚拟角色的动画化。该方法包含标识语音数据的有声语音段(步骤 410)。然后通过对有声语音段的高振幅成分执行频谱分析来确定高振幅频谱(步骤 415)。之后将该高振幅频谱分类为元音音素,该元音音素是从简化的元音集合中选择出来的(步骤 440)。



1、一种对语音数据进行分类的方法，包含：

标识语音数据的有声语音段；

通过该有声语音段的高振幅成分执行频谱分析，确定高振幅频谱；和
将该高振幅频谱分类为元音音素，其中该元音音素是从简化的元音集合中选择出来的。

2、根据权利要求1中的方法，其中该简化的元音集合仅包含主元音音素/a/、/i/和/u/。

3、根据权利要求2中的方法，其中主元音音素/a/包含从由/ax/、/aa/、/ae/、/ao/、/aw/、/er/、/ay/、/eh/和/ey/构成的组中选出的英语音素；主元音音素/i/包含从由/ih/和/iy/构成的组中选出的英语音素；并且主元音音素/u/包含从由/ow/、/oy/、/uh/和/uw/构成的组中选出的英语音素。

4、根据权利要求1中的方法，进一步包含：将元音音素映射到发音嘴形以使虚拟角色动画化。

5、根据权利要求1中的方法，进一步包含：在移动无线通信设备处接收语音数据。

6、根据权利要求5中的方法，其中当接收到语音数据时，实时地对有声语音段的高振幅成分进行分类。

7、根据权利要求1中的方法，其中确定高振幅频谱的步骤包括：

通过根据振幅而使有声语音段的高振幅成分的快速傅立叶变换、即FFT标准化，确定已标准化了的快速傅立叶变换、即FFT数据；

对已标准化了的FFT数据进行滤波，以创建已标准化和滤波了的FFT数据；

检测已标准化和滤波了的FFT数据中的峰值；和

基于所检测到的峰值而将已标准化和滤波了的FFT数据分类为子波段。

8、根据权利要求7中的方法，其中检测已标准化和滤波了的FFT数据中的峰值的步骤包含：对峰值的数目进行计数、测量峰值分布和测量峰值能量。

9、根据权利要求7中的方法，其中子波段被索引为从0到3，并且简化的元音集合包含以下主元音音素：

/a/, 其中高振幅频谱的能量集中在子波段1或2中;

/i/, 其中高振幅频谱的能量集中在子波段0和2中; 与

/u/, 其中高振幅频谱的能量集中在子波段0中。

10、根据权利要求1中的方法, 其中有声语音段的高振幅成分包括语音数据的关键帧。

对语音数据进行分类的方法

技术领域

本发明通常涉及语音识别。尤其是，虽然并非专门地，但是本发明涉及对语音数据进行分析 and 分类以有助于使虚拟角色（avatars）动画化。

背景技术

语音识别是一种将例如在扩音器处所接收到的声信号转换为例如音素、单词、以及句子之类的语言成分的处理。语音识别对许多功能来说是有用的，这些功能包含将口语转换成书面文本的听写以及利用口头命令来对软件应用程序进行控制的计算机控制。

进一步涌现出的语音识别技术的应用是控制计算机生成的虚拟角色（avatars）。根据印度教的神话，虚拟角色（avatars）是起到与人类的居间人这样的作用的天神的化身。在电子通信的虚拟世界中，虚拟角色是人或各类生物的像卡通那样的“二维”或“三维”图形表示。作为一个“正在说话的头”，虚拟角色通过将表示通信的虚拟图像发送给接收者而能够让例如语音呼叫或电子邮件之类的电子通信变得生动。例如，利用语音合成技术而通过虚拟角色将电子邮件的文本“讲给”接收者。此外，利用正在说话的虚拟角色，可以将仅将声数据从主叫方传送到被叫方的传统电话呼叫转化为准视频会议呼叫。对于参加者来说，这种准视频会议呼叫比传统的仅有音频的会议呼叫更加有趣并且信息更加丰富，但是需要比实际视频数据传输少得多的带宽。

使用虚拟角色的准视频会议采用语音识别技术以标识出所接收到的音频数据中的语言成分。例如，在移动电话屏幕上所显示的虚拟角色能够实时地使主叫方的话音动画化。当主叫方的话音通过电话扬声器扩散时，电话中的语音识别软件标识主叫方的话音中的语言成分，并且将该语言成分映射成虚拟角色的嘴部的图形表示的变化。实时地利用主叫方的话音，虚拟角色（avatars）因此看起来就像是正在讲话的电话用户。

发明内容

根据一个方面，本发明是一种用于对语音数据进行分类的方法。该方法包含标识语音数据的有声（voiced）语音段。然后，通过该有声语音段的高振幅成分执行频谱分析而确定高振幅频谱。之后将该高振幅频谱分类为元音音素，其中该元音音素是从简化的元音集合中选择出来的。

因此，使用本发明，利用实时语音数据可改善对虚拟角色的动画化。本发明的方法与大多数的传统语音识别方法相比具有更少的计算密集度，这可使得本发明的方法被更快地执行，同时使用较少的处理器资源。

附图说明

为了能够容易地理解本发明并将本发明投入实际使用，现在参考通过参考附图所说明的示例性实施例，其中在各个图中相似的参考数字表示相同的或功能上相似的元件。根据本发明，附图与下面的详细说明一起包含在说明书中并且构成了说明书的一部分，并且用来进一步说明实施例和解释各种原理和优点，其中：

图 1 给出了用于执行本发明的方法的以无线电话形式的移动设备的示意图；

图 2 是说明了根据本发明实施例的例如在移动设备处所接收和处理的语音数据的图表和相关的频谱图；

图 3 给出了根据本发明实施例的语音分类和嘴部运动映射处理的功能部件的模块图；并且

图 4 是说明了根据本发明实施例的对语音数据进行分类的方法的一般流程图。

熟练的技术人员应该理解的是，为了简单和清楚起见，对附图中的元件进行了举例说明，并且其不一定是按比例绘制的。例如，附图中一些元件的尺寸相对于其他元件而言被放大了，以有助于提高对本发明的实施例的理解。

具体实施方式

在详细描述根据本发明的实施例之前，应该注意到，该实施例主要在于与用于对语音数据进行分类的方法相关的方法步骤和装置部件的组合。因此，在

附图中在适当的地方用惯用的符号表示了装置部件和方法步骤，其中仅示出了与理解本发明的实施例有关的那些特定细节，以便不会因对于受益于在此的说明书的本领域普通技术人员来说很容易明白的细节而模糊了本公开。

在本文中，诸如左和右、第一和第二等这样的关系术语仅仅被用来区分一个实体或动作与另一个实体或动作，而不一定要求或暗示这种实体或动作之间的这种任何实际的关系或顺序。术语“包括”，“包含”或它的任何其他变化用于覆盖非排他性的包含，从而使得包括一系列要素的处理、方法、制品、或者装置不仅包括这些要素，而且还包括其他没有明确列出的或是为这种处理、方法、制品、或装置所固有的要素。由“包括一个……”限定的要素在不受到更多限制的情况下不排除在处理、方法、制品、或者装置中还存在其他相同的要素。

参见图 1，该示意图说明了用于执行本发明的方法的、以无线电话 100 的形式的移动设备。该电话 100 包含与处理器 103 耦合进行通信的射频通信单元 102。该电话 100 还具有与处理器 103 耦合进行通信的小键盘 106 和显示屏 105。对于本领域技术人员来说，显然屏幕 105 可以是触摸屏从而使小键盘 106 是可选择的。

该处理器 103 包含编码器/解码器 111 和相关的用于存储下述数据的代码只读存储器 (ROM) 112，所述数据用于对可以由无线电话 100 发送或接收的语音或其它信号进行编码和解码。处理器 103 还包含通过公共数据和地址总线 117 而与编码器/解码器 111 耦合的微处理器 113、字符只读存储器 (ROM) 114、随机存取存储器 (RAM) 104、静态可编程存储器 116 和 SIM 接口 118。静态可编程存储器 116 和可操作地与 SIM 接口 118 耦合的 SIM 中的每一个都尤其可以存储选定的到来文本消息以及电话号码数据库 TND (电话簿)，该电话号码数据库 TND 包括用于电话号码的号码字段以及用于与姓名字段中的号码之一相关联的标识符的姓名字段。例如，电话号码数据库 TND 中的一个条目可能是在姓名字段中具有相关标识符 “Steven C ! at work” 的 9199911111 (输入到号码字段中)。

微处理器 113 具有用于与小键盘 106 和屏幕 105 以及警报器 115 相耦合的端口，其中警报器 115 一般包括有警报扬声器、振动电动机、以及相关的驱动器。此外，微处理器 113 具有用于与麦克风 135 和通信扬声器 140 相耦合的端口。字符只读存储器 114 存储用于对可以由通信单元 102 所接收的文本消息进

行解码或编码的代码。在这个实施例中，该字符只读存储器 114 还存储微处理器 113 的操作代码 (OC) 以及用于执行与无线电话 100 相关联的功能的代码。

射频通信单元 102 是具有公共天线 107 的组合的接收机和反射机。该通信单元 102 具有通过射频放大器 109 而与天线 107 相耦合的收发信机 108。该收发信机 108 还与组合的调制器/解调器 110 相耦合，该调制器/解调器 110 使通信单元 102 与处理器 103 相耦合。

参见图 2，根据本发明的实施例，图表 200 和相关的频谱图 205-n 描述了例如在无线电话 100 处所接收和处理的语音数据。图表 200 将语音数据描绘为相对于时间的声音振幅。本领域技术人员会将三种主要峰值波形包络 210-n 识别为有声 (voiced) 语音，并且将峰值波形包络 210-n 之间的相对低振幅的间隔识别为无声 (unvoiced) 语音。

传统的语音识别处理解决了标识音素这样的复杂的技术问题，音素是用于造词的最小元音声音单位。语音识别通常是需要对语音数据进行计算上密集的分析的统计处理。这种分析包含：识别诸如背景噪声和传感器引起的噪声之类的声音变化性；以及识别诸如单个音素中的声音差异之类的语音可变性。

根据一个实施例，本发明是一种计算密集度要显著小于传统语音识别方法的方法，该方法用于对语音数据进行分类以使虚拟角色 (avatars) 的嘴部特征的动画更加可信且看起来更真实。例如，虚拟角色可以显示在电话 100 的屏幕 105 上，并且看上去像是正在实时说着主叫方的那些由收发信机 108 接收并在通信扬声器 140 上放大的话语。下面将详细描述这样一种方法。

首先，通过标识语音数据的有声语音段而对例如在图 200 中所描述的语音数据进行滤波。通过使用诸如能量分析和过零率分析之类的本领域所熟知的各种技术来执行对有声语音段的标识。语音数据的高能成分通常与有声的声音有关，同时低能量语音数据至中能量语音数据通常与无声的声音有关。语音数据的极低能量成分通常与静寂或背景噪声有关。

过零率是对语音数据的频率内容进行的简单测量。语音数据的低频成分通常与有声的语音有关，而语音数据的高频成分通常与无声的语音有关。

在标识有声的语音段后，为每一段确定一个高振幅频谱。因此，对每一段而言，通过根据振幅来使每一有声语音段的高振幅成分的快速傅立叶变换 (FFT) 标准化，来确定标准化的快速傅立叶变换 (FFT) 数据。例如，在图

表 200 中, 将每一段的高振幅成分标识为“关键帧”, 该“关键帧”包含每一段之内的峰值振幅。该关键帧一般具有固定的时间窗(大约 30ms), 并且抽样的次数会依据该语音数据的抽样率而改变。例如, 一个典型的关键帧可以包含以 8 kHz 抽样率的长度 $L = 256$ 的抽样, 或是以 16 kHz 抽样率的 $L = 512$ 的抽样。

此后对标准化的 FFT 数据进行滤波, 以便使该数据中的峰值更加明显。例如, 可以应用具有阈值设定为 0.1 的高通滤波器, 它将 FFT 数据中的所有低于阈值设定的值设置为零。

然后由一个或更多峰值检测器来处理已被标准化和滤波了的 FFT 数据。该峰值检测器对诸如峰值数目、峰值分布、以及峰值能量这样的各种峰值属性进行检测。此后利用来自峰值检测器的数据, 将可能表示主要元音声音的高振幅频谱的已被标准化和滤波了的 FFT 数据划分为子波段。例如, 根据本发明的一个实施例, 使用被索引为从 0 到 3 的四个子波段。如果高振幅频谱的能量集中在子波段 1 或 2 中, 那么该频谱很有可能被分类为与主元音音素/a/相对应。如果该高振幅频谱的能量集中在子波段 0 和 2 中, 则该频谱很有可能被分类为与主元音音素/i/相对应。最后, 如果该高振幅频谱的能量集中在子波段 0 中, 则该频谱很有可能被分类为与主元音音素/u/相对应。图 2 描述了与峰值波形包络 210-1 以及主元音音素/a/相对应的标准化频谱 205-1; 与峰值波形包络 210-2 以及主元音音素/i/相对应的标准化频谱 205-2; 以及与峰值波形包络 210-3 以及主元音音素/u/相对应的标准化频谱 205-3。

根据本发明的一个实施例, 利用已分类的频谱来使虚拟角色 (avatars) 的特征动画化, 以便创建该虚拟角色实际上正在“说出”语音数据这样的印象。这种动画是通过将已分类的频谱映射到不连续的嘴部运动来执行的。如本领域所熟知的那样, 由虚拟角色使用一系列发音嘴形 (viseme) 来使不连续的嘴部运动重复, 所述一系列发音嘴形 (viseme) 实质上是已映射到可视域中的基本语音单位。每一个发音嘴形 (viseme) 表示一个静态的、在视觉上形成对比的嘴部形状, 它通常与当一个人发出特定音素时所使用的嘴部形状相对应。

本发明能够通过使用下述事实有效地执行这种音素到发音嘴形的映射, 所述事实就是语言中的音素数目远大于相应发音嘴形的数目。更进一步地, 将上述主元音音素/a/、/i/和/u/的每一个映射到三个完全不同的发音嘴形之一。通过

仅使用这三个不同的发音嘴形——其与从闭合到张开并且然后再到闭合位置运动的嘴的图像帧有关——像卡通片一样，可以创建可信的嘴部运动。由于在语音数据中仅识别三个主要元音音素，因此本发明的实施例中的语音识别明显比现有技术中的语音识别具有更少的处理器密集度。例如，根据本发明的一个实施例，利用如下述表1中所示的/a/、/i/和/u/的三个主要元音音素，可使英语中的各种元音音素分组成为简化的元音集合。

表1 - 英语中的简化的元音集合

/a/	ax, aa, ae, ao, aw, er, ay, eh, ey
/i/	ih, iy
/u/	ow, oy, uh, uw

利用诸如根据语音能量进行嘴部宽度映射或根据语音数据的频谱结构进行嘴部形状映射之类的技术，根据本发明的教义进行分类的语音数据可以用于控制虚拟角色（avatars）的嘴部和嘴唇图形的动作。例如，嘴部宽度映射与在峰值波形包络 210-n 期间嘴部的张开和闭合有关。考虑将被编号为从 0 到 i-1 的 i 个图像用于描述峰值波形包络 210-n。嘴部宽度映射首先将该峰值波形包络 210-n 的开始无声段设置为零，由此表示闭合的嘴部。然后根据在每一相应帧中的语音能量，将在峰值波形包络 210-n 中剩余的数据帧映射到该图像 1 到 i-1 上。最后，为了使所感知到的虚拟角色的嘴部和嘴唇动作看上去更自然，执行对嘴部和嘴唇图形的后处理以提供图像之间的平滑转换。

参见图 3，示意性框图 300 描述了根据本发明实施例的语音分类和嘴部运动映射处理的功能部件。可将处理分为三个主要功能块：关键帧标识块 305、简化的元音集合分类块 310、以及动画合成块 315。

在关键帧标识块 305 中，在能量分析块 320 和过零率块 325 中并行地接收并处理输入的语音数据。将来自能量分析块 320 和过零率块 325 的数据提供给用于分离有声和无声语音数据的有声/无声检测块 330。然后，在有声包络产生器块 335 中对来自有声/无声检测块 330 的数据进行处理，该有声包络产生器块 335 用于标识语音数据中的有声语音段。如图所示，在上述关键帧标识块 305 中，在有声包络产生器块 335 中还使用来自能量分析块 320 的原始数据。

在简化的元音集合分类块 310 中，向关键帧频谱分析块 340 提供有声语音段，该关键帧频谱分析块 340 对有声语音段的高振幅成分执行频谱分析并确定高振幅频谱。接下来，在分类块 345 中，将高振幅频谱分类为主元音音素/a/、/i/或/u/。

最后，在动画合成块 315 中，将从分类块 350 输出的主元音音素映射到用于使虚拟角色动画化的发音嘴形(visemes)。该动画合成块 315 从动画素材数据库 355 中检索这样的信息，该信息包括例如发音嘴形定义以及与当在音素之间进行转换时所使用的传统嘴部张开和闭合位置有关的信息。该动画合成块 315 的最后输出因此是与语音同步的动画。

参见图 4，一般流程图描述了根据本发明实施例的用于对语音数据进行分类的方法 400。首先，在步骤 405，在例如无线电话 100 之类的移动无线通信设备处接收语音数据。在步骤 410，标识该语音数据的有声语音段。接下来，在步骤 415，通过对有声语音段的高振幅频谱成分执行频谱分析来确定高振幅频谱。步骤 415 可以包含以下子步骤：在步骤 420，通过根据振幅而使有声语音段的高振幅成分的 FFT 标准化来确定已标准化的 FFT 数据。在步骤 425，对标准化的 FFT 数据进行滤波，以创建已标准化和滤波了的 FFT 数据。之后，在步骤 430，检测在已标准化和滤波了的数据中的峰值。在步骤 435，基于诸如峰值总数、峰值分布和峰值能量之类的所检测峰值的属性来将已标准化和滤波了的 FFT 数据分类为子波段。然后，在步骤 440，将高振幅频谱分类为例如上述主元音音素/a/、/i/或/u/之一这样的元音音素。最后，在步骤 445，将主要元音音素映射到例如与虚拟角色的嘴部位置相关联的发音嘴形。

本发明的优点因此包含利用实时的语音数据而改善了虚拟角色的动画化。此外，本发明的方法与大多数的传统语音识别方法相比具有更少的计算密集度，这使得本发明的方法能够被更快地执行，同时使用较少的处理器资源。因此，本发明的实施例特别适合于具有有限的处理器和存储器资源的移动通信设备。

以上的详细描述仅提供了示例性实施例，而不意味着限定了本发明的范围、可适用性、以及结构。相反，上述示例性实施例的详细说明向本领域技术人员提供了允许实现本发明的示例性实施例的描述。应当理解的是，在不脱离所附权利要求中所阐明的本发明的精神和范围的情况下，可对元件以及步骤的功能和设置进行各种改变。应该理解的是，这里所描述的本发明的实施例可以包括

一个或多个传统的处理器和独特的所存储的程序指令，所述程序指令用于对一个或多个处理器进行控制，以结合某些非处理器电路而实现如这里所描述的对语音数据进行分类的一些功能、大多数功能、或者所有功能。该非处理器电路可以包含但不局限于：无线电接收机、无线电反射机、信号驱动器、时钟电路、电源电路和用户输入装置。同样地，可将这些功能解释为用于对语音数据进行分类的方法步骤。或者，一部分或全部功能可以由不具有存储的程序指令的状态机实现，或者用一个或多个专用集成电路（ASIC）实现，在所述专用集成电路中每个功能或某些功能的某些组合可以作为定制的逻辑实现。当然，可以使用两种方法的组合。因此，这里已对用于这些功能的方法和装置进行了描述。此外，可以预料到的是，虽然由于例如可用时间、当前的技术、以及经济方面的考虑而促发了可能的显著效果以及许多设计选择，但是普通技术人员在受到这里所公开的概念和原则的指导时，可利用最少的试验而很容易地产生这种软件指令和程序以及 IC。

在上述的说明书中，已经描述了本发明的具体实施例。然而，本领域普通技术人员应该知道的是，在不脱离权利要求中所阐明的本发明的精神和范围的情况下可做出各种修改和变化。因此，说明书和附图被认为是说明性的而不是限制性的，并且意图使所有这种修改包含在本发明的范围之内。益处、优点、问题的解决方案、以及可产生任何益处、优点或解决方案或使其变得更加显著的任何要素都不应当被认为是任何一个权利要求或是所有权利要求的关键的、必需的、或者基本的特征或要素。本发明仅仅由所附权利要求来定义，其中包括在该申请的待审期间所做出的任何修改以及这些权利要求的等效内容。

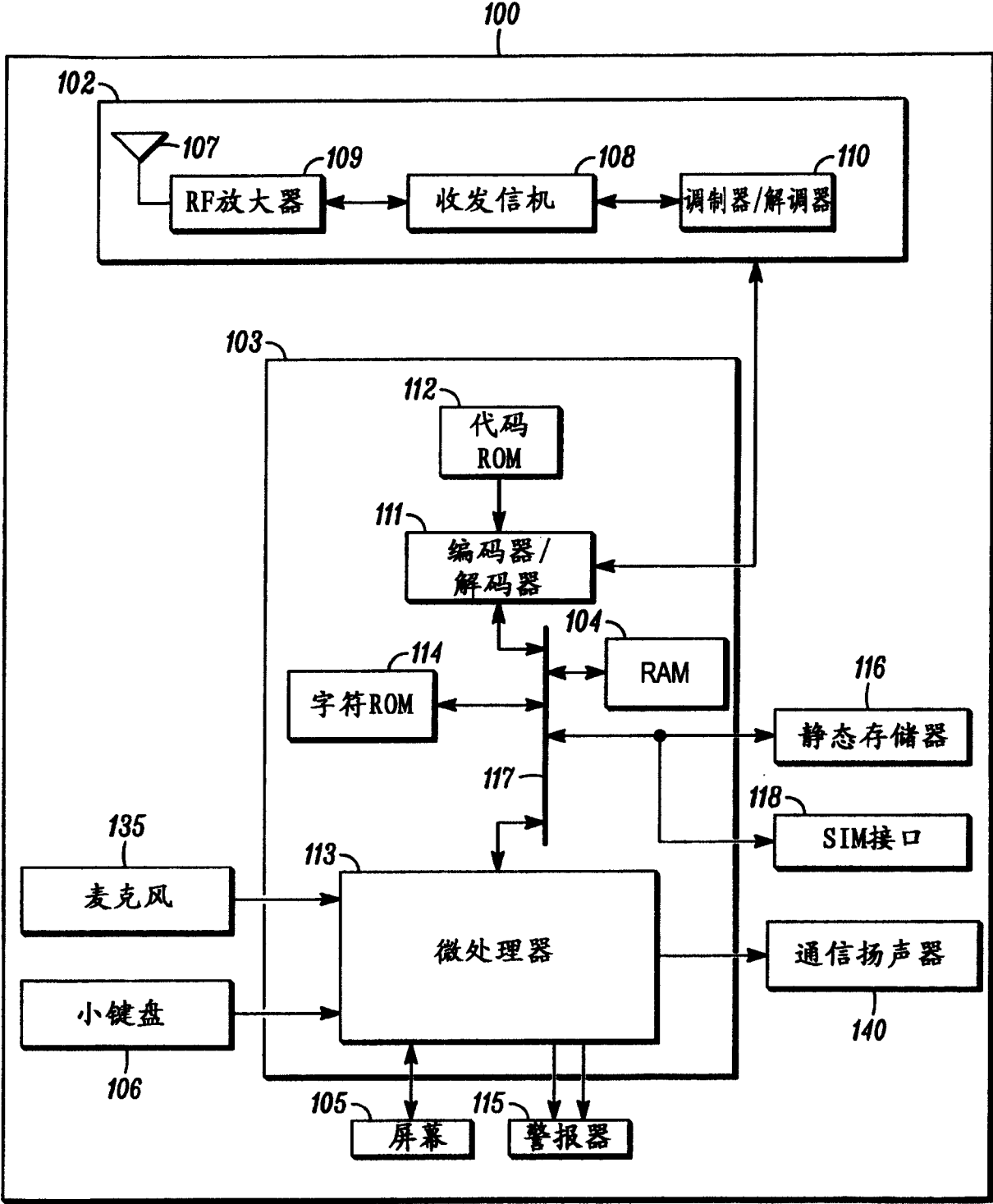


图1

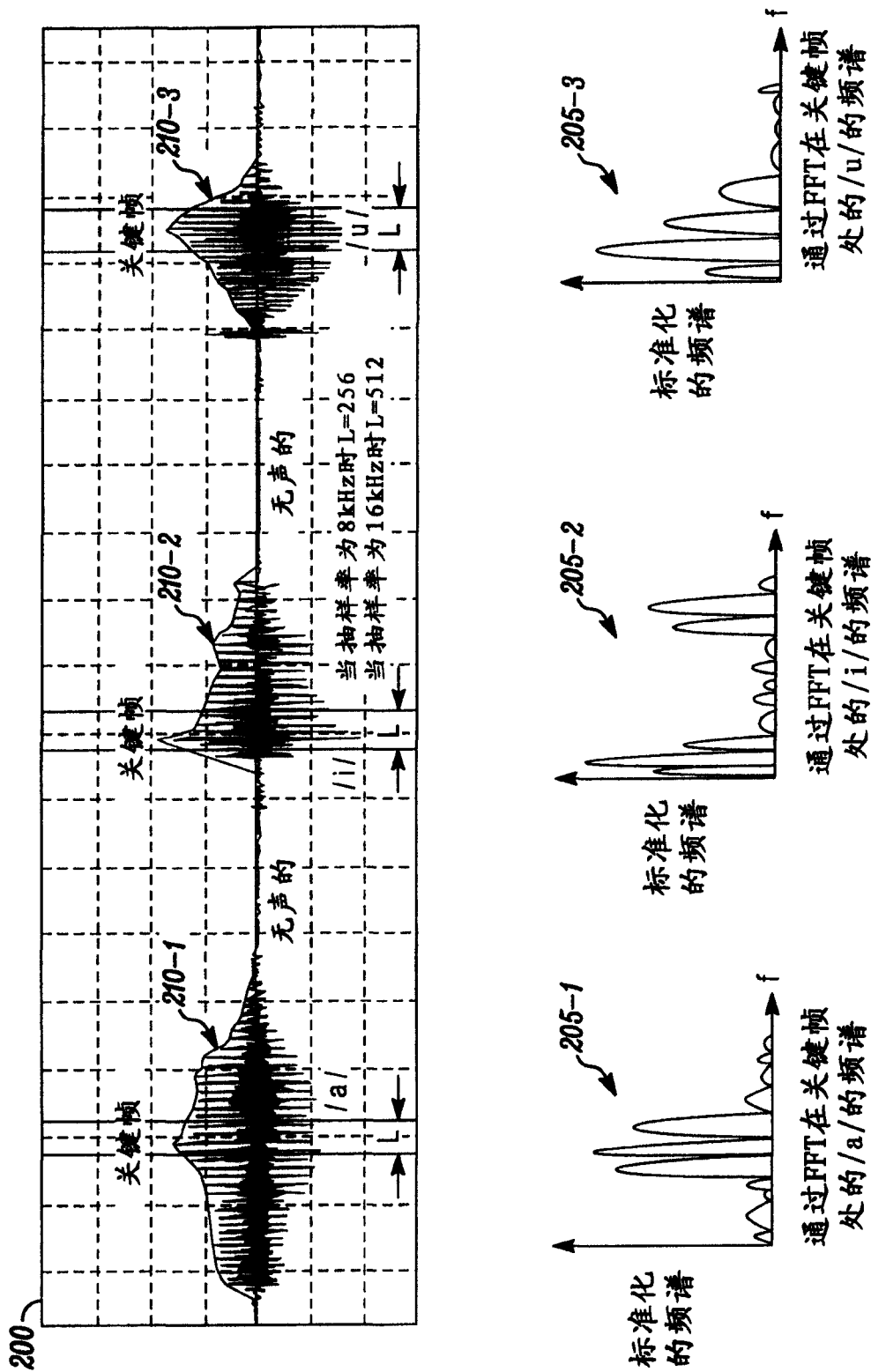


图2

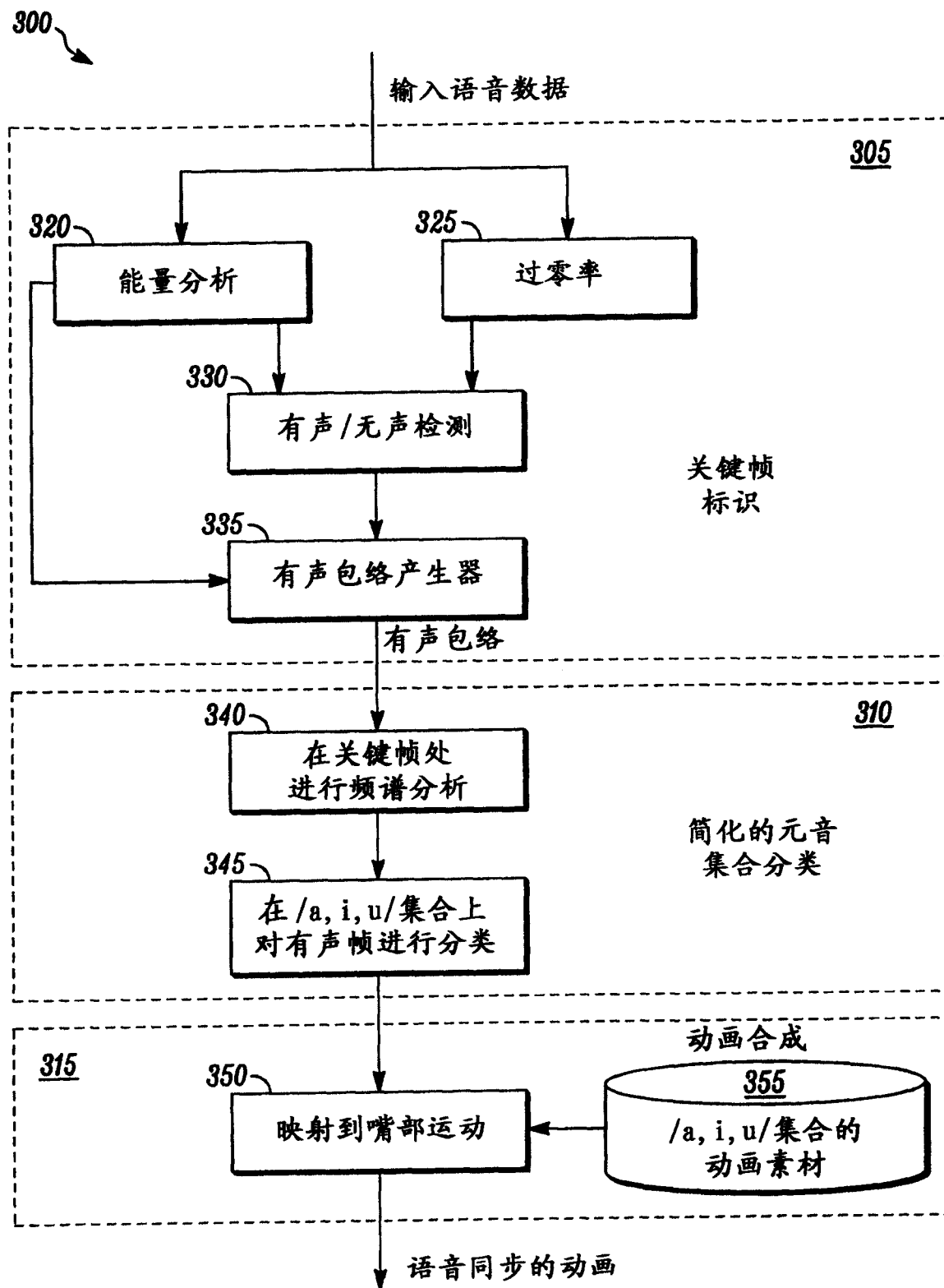


图 3

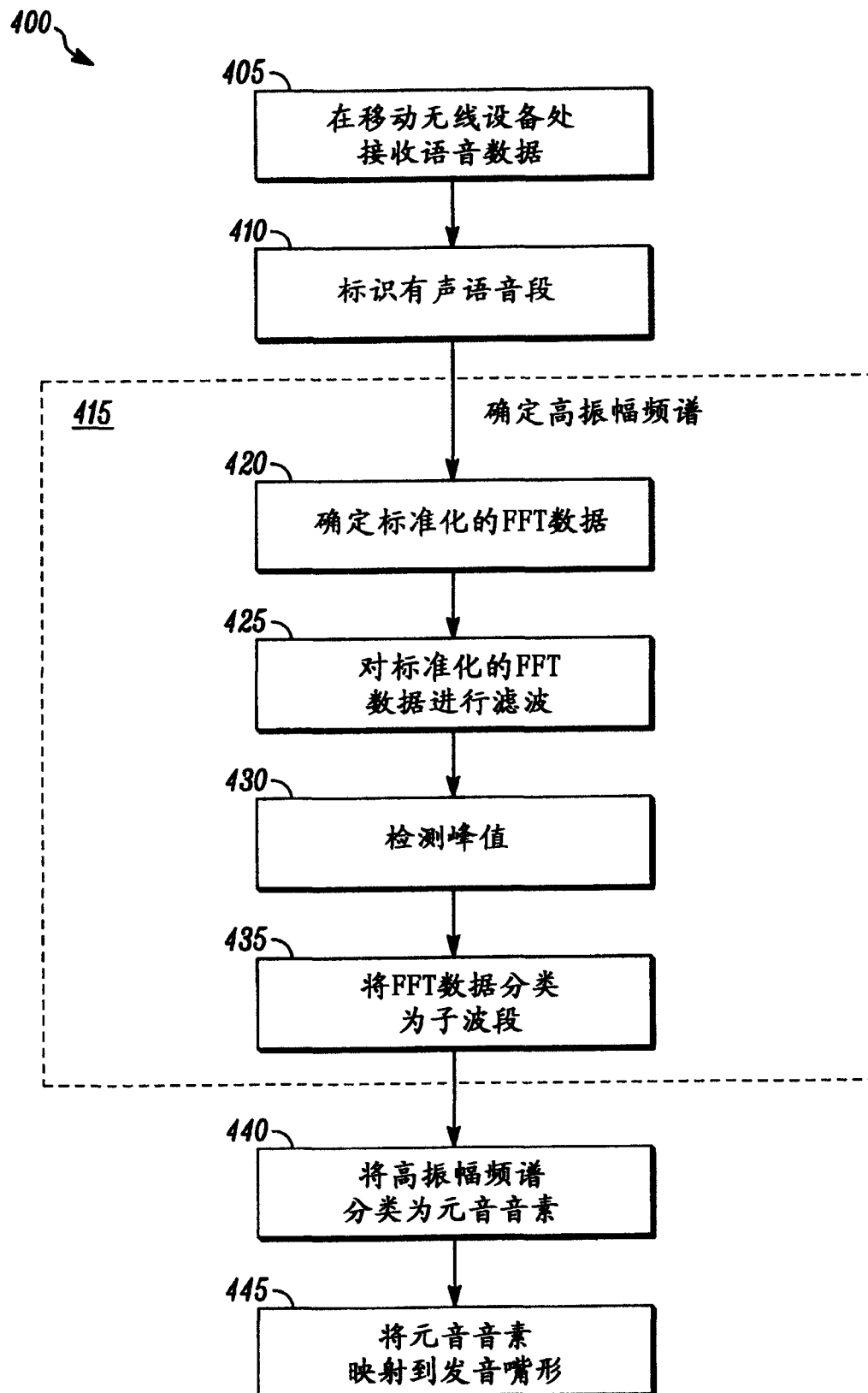


图 4