



[12] 发 明 专 利 说 明 书

[21] ZL 专利号 00122702.5

[45] 授权公告日 2004 年 7 月 14 日

[11] 授权公告号 CN 1157710C

[22] 申请日 2000.8.8 [21] 申请号 00122702.5

[30] 优先权

[32] 1999. 8. 10 [33] US [31] 09/371,400

[71] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 第米垂·卡那威斯基

斯第芬·荷曼·梅斯

杰弗里·斯考特·琐伦森

审查员 刘红梅

[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

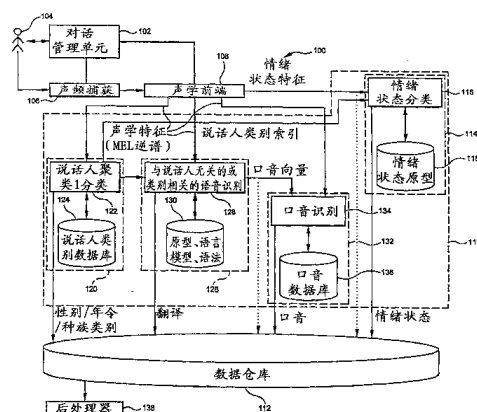
代理人 朱海波

权利要求书 7 页 说明书 16 页 附图 6 页

[54] 发明名称 会话数据开采

[57] 摘要

一种用于收集和话音系统用户的话音相关的数据的方法包括：和用户进行对话，捕获并数字化用户的语音波形，从数字化语音波形中提取至少一个声学特征以及在数据仓库中以利用随后的数据开采的形式存储与声学特征对应的属性数据以及标识标记。在该方法的一种替代形式中，在带有或不带有存储用于随后的数据开采的数据下，可利用用户属性实时修改话音系统的行为。



1. 一种用于在数据仓库中收集和话音系统用户的话音相关的数据的方法，所述方法包括步骤：

(a)通过话务员和话音操纵机器系统中的至少一个进行与话音系统用户的会话；

(b)捕获与所述会话期间由该话音系统用户说出的发音相关的语音波形；

(c)数字化所述语音波形以提供数字化语音波形；

(d)从所述数字化语音波形中提取至少一个和至少一个用户属性相关的声学特征，所述至少一个用户属性包括下述中的至少一个：

(d-1) 用户的性别；

(d-2) 用户的年龄；

(d-3) 用户的口音；

(d-4) 用户的母语；

(d-5) 用户的方言；

(d-6) 用户的社会经济类别；

(d-7) 用户的教育水平；以及

(d-8) 用户的情绪状态；以及

(e) 在数据仓库中以利于随后对其进行数据开采的形式存储和所述声学特征对应的属性数据以及至少一个标识标记，其中所述声学特征和所述至少一个用户属性相关。

2. 权利要求1的方法，还包括下述附加的步骤：

(f) 为多次与其它用户的其它会话重复步骤(a)-(e)，以提供包括属性数据和标识标记的存储数据的集合；以及

(g) 开采存储数据集合，以提供用于修改话音系统的基础商用逻辑的信息。

3. 权利要求1的方法，其中步骤(e)包括存储包含时间印记的至少一个标识标记。

4. 权利要求1的方法, 其中步骤(d)包括提取基频、基频中的偏差、平均运行音调、运行音调方差、音调跳动、运行能量方差、语音速率以及摆振中的至少一个, 以作为和用户的情绪状态相关的至少一个情绪状态特征。

5. 权利要求4的方法, 还包括归一化所述至少一个情绪状态特征的附加步骤。

6. 权利要求1的方法, 还包括处理所述至少一个声学特征以确定所述至少一个用户属性的附加步骤, 其中步骤(e)中的所述属性数据至少包括所述用户属性的一个值。

7. 权利要求6的方法, 还包括响应在数据仓库中存储附加的属性自动改进所述处理步骤的附加步骤。

8. 权利要求1的方法, 其中步骤(e)包括把所述属性数据存储为至少一个实质上原始的声学特征。

9. 权利要求1的方法, 其中步骤(d)包括至少提取 MEL 逆谱, 其中还包括基于所述 MEL 逆谱识别用户的语音、翻译所述语音以及从词选择和词汇集中之一检查所述语音以确定用户的教育水平、用户的社会经济类别及用户的方言中的至少一个附加步骤。

10. 一种用于把话音系统响应加工成从声学上确定的话音系统用户的状态的方法, 所述方法包括步骤:

(a)通过话音系统和该话音系统用户进行会话;

(b)捕获与所述会话期间由该话音系统用户说出的发音相关的语音波形;

(c)数字化所述语音波形以提供数字化语音波形;

(d)从所述数字化语音波形中提取至少一个和至少一个用户属性相关的声学特征, 所述至少一个用户属性包括下述中的至少一个:

(d-1) 用户的性别;

(d-2) 用户的年龄;

(d-3) 用户的口音;

(d-4) 用户的母语;

- (d-5) 用户的方言;
- (d-6) 用户的社会经济类别;
- (d-7) 用户的教育水平; 以及
- (d-8) 用户的情绪状态; 以及
- (e) 根据所述至少一个用户属性修改该话音系统的行为。

11. 权利要求 10 的方法, 其中步骤(e)中的所述修改至少包括下述之一:

实时改变话音系统的商用逻辑; 以及
和不带有所述修改的话音系统的期待响应相对比, 实时地修改话音系统响应。

12. 权利要求 10 的方法, 其中步骤(d)包括提取基频、基频中的偏差、平均运行音调、运行音调方差、音调跳动、运行能量方差、语音速率以及摆振中的至少一个, 以作为和用户的情绪状态相关的至少一个情绪状态的特征。

13. 权利要求 12 的方法, 还包括归一化所述至少一个情绪状态特征的附加步骤。

14. 权利要求 10 的方法, 其中步骤(d)包括至少提取 MEL 逆谱, 其中还包括基于所述 MEL 逆谱识别用户的语音、翻译所述语音以及从词选择和词汇集中之一检查所述语音以确定用户的教育水平、用户的社会经济类别及用户的方言中的至少一个附加步骤。

15. 权利要求 12 的方法, 还包括检查所述至少一个情绪状态特征以判定该用户是否处于快活情绪状态的附加步骤;

其中步骤(e)包括响应于所述快活情绪状态向该用户至少提出一种产品和一种服务中的至少一种。

16. 权利要求 15 的方法, 还包括下述附加步骤:
确定至少一个和情绪状态不同的用户属性; 以及
响应于所述至少一个和情绪状态不同的用户属性而修改所述一种产品和一种服务中的至少一种。

17. 权利要求 12 的方法, 还包括检查所述至少一个情绪状态特征以

判定该用户是否处于快活情绪状态的附加步骤;

其中步骤(e)包括响应所述快活情绪状态至少对该用户进行营销研究。

18. 权利要求 17 的方法, 还包括下述附加步骤:

确定至少一个不同于情绪状态的用户属性; 以及

响应于所述至少一个不同于情绪状态的用户属性而修改所述市场研究。

19. 权利要求 12 的方法, 其中所述话音系统实质上是自动交互语音响应 (IVR) 系统, 该方法还包括检查所述至少一个情绪状态特征以判定用户是否处于厌恶、轻视、害怕和生气的情绪状态中的至少一种状态下的附加步骤;

其中步骤(e)包括响应于所述厌恶、轻视、害怕和生气情绪状态中的至少一种状态至少把所述用户从所述 IVR 转到话务员上。

20. 权利要求 12 的方法, 其中所述话音系统是混合式交互语音响应 (IVR) 系统, 该方法还包括检查所述至少一个情绪状态特征以判定用户是否处于厌恶、轻视、害怕和生气情绪状态中的至少一种状态下的附加步骤;

其中步骤(e)包括响应于所述厌恶、轻视、害怕和生气情绪状态中的至少一种状态至少把所述用户从低层话务员转到高层管理人员。

21. 权利要求 12 的方法, 其中所述话音系统实质上是自动交互语音响应 (IVR) 系统, 该方法还包括检查所述至少一个情绪状态特征以判定用户是否处于迷惑情绪状态下的附加步骤;

其中步骤(e)包括响应于所述迷惑情绪状态把所述用户从所述 IVR 转到话务员。

22. 权利要求 10 的方法, 还包括附加步骤:

(f) 在数据仓库中以利于随后对其进行数据开采的形式存储和所述声学特征对应的属性数据以及至少一个标识标记, 其中所述声学特征和所述至少一个用户属性相关。

23. 一种用于收集和用户的话音相关的数据的设备, 所述设备包括:

- (a) 一个和用户进行会话的对话管理单元;
 - (b) 一个和所述对话管理单元连接并捕获与会话期间由该用户说出的发音相关的语音波形的声频捕获模块。
 - (c) 一个声学前端, 其和所述声频捕获模块连接并配置成:
 - 接收并数字化语音波形以提供数字化语音波形; 以及
 - 从数字化语音波形中提取至少一个和至少一个用户属性相关的声学特征, 所述至少一个用户属性包括下述中的至少一个:
 - (c-1) 用户的性别;
 - (c-2) 用户的年龄;
 - (c-3) 用户的口音;
 - (c-4) 用户的母语;
 - (c-5) 用户的方言;
 - (c-6) 用户的社会经济类别;
 - (c-7) 用户的教育水平; 以及
 - (c-8) 用户的情绪状态;
 - (d) 一个和所述声学前端连接并分析所述至少一个声学特征以确定所述至少一个用户属性的处理模块; 以及
 - (e) 一个和所述处理模块连接并以利于随后对其进行数据开采的形式存储所述至少一个用户属性以及至少一个标识标记的数据仓库。
24. 权利要求 23 的设备, 其中所述声频捕获模块由模数转换器板、交互话音响应 (IVR) 系统和麦克风中之一构成。
25. 权利要求 23 的设备, 其中所述对话管理单元由电话交互话音响应 (IVR) 系统构成。
26. 权利要求 25 的设备, 其中所述处理模块由所述 IVR 的处理器部分构成。
27. 权利要求 23 的设备, 其中所述处理模块由带有适当软件的独立通用计算机构成。
28. 权利要求 23 的设备, 其中所述处理模块由专用电路构成。
29. 权利要求 23 的设备, 其中所述处理模块包括至少一个情绪状态

分类器。

30. 权利要求 29 的设备，其中所述处理模块还包括至少：

一个说话人聚类器及分类器；

一个语音识别器；以及

一个口音识别器。

31. 权利要求 30 的设备，还包括一个和所述数据仓库连接并配置成译出用户的发音和在其上进行关键词确定的后处理器。

32. 一种用于和用户交互的可实时修改的话音系统，所述系统包括：

(a) 一个和用户进行会话的对话管理单元；

(b) 一个和所述对话管理单元连接并捕获与会话期间由该用户说出的发音相关的语音波形的声频捕获模块。

(c) 一个声学前端，其和所述声频捕获模块连接并配置成：

接收并数字化语音波形以提供数字化语音波形；以及

从数字化语音波形提取至少一个和至少一个用户属性相关的声学特征，所述至少一个用户属性包括下述中的至少一个：

(c-1) 用户的性别；

(c-2) 用户的年龄；

(c-3) 用户的口音；

(c-4) 用户的母语；

(c-5) 用户的方言；

(c-6) 用户的社会经济类别；

(c-7) 用户的教育水平；以及

(c-8) 用户的情绪状态；以及

(d) 一个和所述声学前端连接并分析所述至少一个声学特征以确定所述至少一个用户属性的处理模块；

其中所述处理模块根据所述至少一个用户属性修改该语音系统的行为。

33. 权利要求 32 的系统，其中所述处理模块通过提示话务员至少部分地修改该语音系统的行为。

34. 权利要求 32 的系统，其中所述处理模块由交互语音响应（IVR）系统的处理器部分构成，并且其中所述处理器模块通过修改 IVR 的商用逻辑至少部分地修改该语音系统的行为。

35. 权利要求 32 的系统，其中所述声频捕获模块由模数转换器板、交互语音响应（IVR）系统和麦克风中之一构成。

36. 权利要求 32 的系统，其中所述对话管理单元由电话交互语音响应（IVR）系统构成。

37. 权利要求 36 的系统，其中所述处理模块由所述 IVR 的处理器部分构成。

38. 权利要求 32 的系统，其中所述处理模块由带有适当软件的独立通用计算机构成。

39. 权利要求 32 的系统，其中所述处理模块由专用电路构成。

40. 权利要求 32 的系统，其中所述处理模块包括至少一个情绪状态分类器。

41. 权利要求 40 的系统，其中所述处理模块至少还包括：

一个说话人聚类器及分类器，

一个语音识别器；以及

一个口音识别器。

42. 权利要求 41 的系统，还包括一个和所述数据仓库连接并配置成译出用户的发音和在其上进行关键词确定的后处理器。

会话数据开采

本发明涉及面向语音的系统，具体地涉及一种便利数据开采的面向声学的方法和设备以及一种用于加工语音系统的响应成为语音系统用户的声学鉴定状态的面向声学的方法和设备。

数据开采是近来日益普及的一种跨学科领域。它指的是使用各种以无监督方式或极小监督方式从数据中提取信息的方法。“无监督”指的是其中不存在事先标注的技术；允许其自己推导出各个类。声音是簇集的，并由人明白推导出那些类。在交易、风险和欺诈管理中使用数据开采。

在数据开采(data mining)领域中，普遍认同更多数据更好。从而，从事数据开采的公司频繁地编辑或获取用户数据库。这些数据库可基于邮购史、过去顾客史、信用史等。预计顾客的电子商务以及其互联网习性很快地也会提供一种顾客数据库基础。可以从对交易或事件的人工或自动编码处理中得到所存储信息的特性。交易的一个例子可以是某给定人在某些条件下在某给定价格下购买某给定产品，或者，某给定人对某邮寄目录作出响应。事件的一个例子可包括某人在某日出交通事件，或者某家庭在上个月搬入。

传统上把数据开采所运算的数据存储在数据仓库里。一旦确定出商务对象，检查该数据仓库以选择有关特征，评估数据的品质，并把数据转换到适用于预定分析的分析模型中。接着可施加诸如预测建模、数据库分段、连接分析和偏差检测之类的技术，以便输出目标、进行预报和检查。在确认后，可以导出结果模型。

当前，常常在电话上通过话务员或交互式语音响应(IVR)系统进行各种各样的交易。大家知道，作为这种交易中的通信方式的话音携带着有关用户的各种属性的信息，诸如性别、年龄、母语、口音、方言、社会经济条件、教育程度和情绪状态。这些参数中的一个或多个对于参与数据开采的各个人可能是有用的。目前，这些交易中所含有的数据上

所发现的不可多得的内容或者全都丢失给数据开采者，或者为了有效地应用必须进行人工检索。

从而，现有技术需要一种用于收集数据仓库中的和话音系统用户的话音相关的数据的方法，其可有效地和自动地使用利用话音系统（例如电话、商亭等）的交易中所得到的数据。还要求该方法实现成实时的、带有或不带有数据仓库存储的，以允许由诸如交互式话音响应系统等的话音系统进行“飞行中”的修改。

解决现有技术中所确定的需求的本发明提供一种方法，用于收集数据仓库中和话音系统用户的话音相关的数据。该方法包括步骤：和话音系统用户进行会话，捕获语音波形，数字化该语音波形，从数字化的语音波形中提取至少一个声学特征，以及在数据仓库中存储和声学特征对应的属性数据。可通过话务员和话音操纵机器系统中的至少一种进行和话音系统用户的会话。所捕获的语音波形是和会话期间话音系统用户说出的发音相关的。语音波形的数字化提供数字化语音波形。从数字化波形中提取至少一个声学特征，并且和至少一个诸如用户性别、年龄、口音、母语、方言、社会经济类别、教育程度和情绪状态的用户属性相关。存储在数据仓库里的该属性数据对应于声学特征，该声学特征和至少一个用户属性相关并且和至少一个标识标记一起存储。在数据仓库中以便利随后的数据开采的形式存储该数据。

本发明还提供一种把话音系统响应加工成话音系统用户的声学鉴定状态的方法。该方法包括通过该话音系统和话音系统用户进行对话的步骤。该方法还包括如前面所讨论的捕获语音波形并数字化该语音波形的步骤。此外，还如上述那样，该方法包括从数字化语音波形提取声学特征的步骤。最后，该方法包括根据至少一个用户属性修改该话音系统的行为的步骤，其中所述至少一个用户属性和至少一个声学特征相关。

本发明还包括一种可由机器读的程序存储部件，其有形地嵌有可由该机器执行的用于实现上述方法之一的指令程序。

本发明还提供一种设备，用于收集和用户的话音相关的数据。该设备包括一个对话管理单元，一个声频捕获模块，一个声学前端，一个处

理模块以及一个数据仓库。该对话管理单元进行与用户的会话。该声频捕获模块和该对话管理单元相连接，并捕获与会话期间由用户说出的发音相关的语音波形。

该声学前端和该声频捕获模块相连接，并配置成接收及数字化该语音波形以提供数字化语音波形，而且配置成从数字化语音波形中提取至少一个和至少一个用户属性相关的声学特征。该至少一个用户属性可包括上面对各方法讨论的至少一个用户属性。

该处理模块和该声学前端相连并分析至少一个声学特征以确定至少一个用户属性。该数据仓库和该处理模块相连，并以随后的数据开采的格式存储该至少一个用户属性。

本发明还提供一种用于和用户交互的可实时修改的话音系统。该系统包括一个上述讨论类型的对话管理单元，一个上述讨论类型的声频捕获模块以及一个上述讨论类型的声学前端。此外，该话音系统还包括一个上述讨论类型的处理模块。该处理模型配置成根据该至少一个用户属性来修改该话音系统的行为。

为了更好地理解本发明以及它的其它和更多的优点，请在结合各附图下参阅下述说明，而在附属权利要求书中指出本发明的范围。

图 1 是依据本发明的用于收集和用户的话音相关的数据的设备图；

图 2 是依据本发明的用于和用户交互的可实时修改的话音系统图；

图 3 是依据本发明的用于在数据仓库中收集和话音系统用户的话音相关的数据的方法的流程图；

图 4 描述图 3 中所示方法的某些细节，它们也可应用于图 5；

图 5 是依据本发明的用于把话音系统响应加工成话音系统用户的声学鉴定状态的方法的流程图；以及

图 6 描述图 5 方法的某些细节。

现参照图 1，其描述依据本发明的用于收集和用户的话音相关的数据的设备。该设备概括地用 100 标示。该设备包括一个与用户 104 进行会话的对话管理单元 102。设备 100 还包括一个和对话管理单元 102 连接的并在会话期间捕获与用户 104 说出的发音相关的语音波形的声频捕获模块 106。如本文中所使用的那样，应把会话广义地理解成包括第一

人和/或者第二人、或者一个机器或者第二人与机器的组合之间的任何交互，该会话包括至少一些语音。

设备 100 还包括一个声学前端 108，前端 108 和声频捕获模块 106 相连并被配置成接收及数字化该语音波形，从而提供数字化语音波形。此外，声学前端 108 还配置成从该数字化语音波形提取至少一个和用户 104 的至少一个用户属性相关的声学特征。该至少一个用户属性可包括下述中的至少一个：用户的性别、用户的年龄、用户的口音、用户的母语、用户的方言、用户的社会经济类别、用户的教育水平以及用户的情绪状态。对话管理单元 102 可使用从声学前端 108 得到的诸如 MEL 逆谱的声学特征，从而若需要时二者之间可以有直接连接。

设备 100 还包括一个和声学前端 108 连接的处理模块 110，其分析该至少一个声学特征以确定至少一个用户属性。此外，设备 100 还包括一个和处理模块 110 连接的数据仓库 112，其在随后的数据开采的形式下存储该至少一个用户属性以及至少一个标识标记。在本文的其它处讨论标识标记。

用户的性别可以通过对用户语音音调分类或者通过对特征组的简单聚类得到确定。在后一种方法中，建立和一种给定性别的大量说话人相关的话音声纹，然后借助这二组模型进行说话人的分类。以类似于性别分类的方式通过年龄组的分类可以确定用户的年龄。尽管可靠性有限，可以以这种方式区分各种年龄类别，例如，儿童、青少年、成人和老年人。

从声学特征中确定口音在技术上是周知的。例如，Lincoln 等于 1998 年国际口语处理年会[以下称为 ICSLP'98]，Sidney, Australia 提出的论文“A Comparison of Two Unsupervised Approaches to Accent Identification”叙述了各种有用的技术。可以以基本上等同于口音分类的方式确定用户的母语。可以增添说话人的母语的有关元数据以定义每种口音/母语模型。

即，在创造用于各种母语的模型时，可以使用一个或多个以该语言作为他们的母语的说话人。在 ICSLP'98 上由 Matrouf 等提出的论文“Language Identification Incorporating Lexical Information”中讨论了

各种语言识别的技术。

可以从口音以及某给定方言中的专用的关键字或习语的使用确定用户的方言。例如，在法语中，为数字 90 选用“nonante”而不是选用“Quatre Vingt Dix”可以标明该说话人为比利时或瑞士血统，而不是法国人或加拿大人。进而，随后为数字 80 选用“qutre-vingt”而不选用“octante”或“Huitante”可标明该人是比利时人而不是瑞士人，在美式英语中，选用“grocery sack”而不选用“grocery bag”可标识某个人来自于中西部而不是来自于中大西洋部。美语中中西部和中大西洋部不同的另一个例子是，在中西部对软饮料选用“pop”，而在中大西洋部为该相应的软饮料选用“soda”。在国际环境下，使用“holiday”以代替“vacation”可标明某人为英国而不是美国血统。本段中所讨论的操作可利用后面讨论的语音识别器 126 实现。

用户的社会经济分类可包括诸如用户的种族背景、用户的人种背景及用户的经济类型等因素，例如，兰领、白领—中产阶级或富人。可以通过训练时刻带注释的口音和方言以及通过检查用户用词选择对此作出确定，尽管其可靠性只是中等的，确信这些技术可充分了解用户背景，从而对数据开采是有用的。

以类似于社会经济分类的方式，可通过词的选择以及口音确定用户的教育等级；同样只能期待局部可靠性，但对于数据开采目的是足够的。

从声学特征中确定用户的情绪状态在技术上是周知的。可以识别的情绪类型包括：激怒、忍怒、惊慌、恐惧、焦急、悲哀、得意、失望、高兴、感兴趣、厌烦、害羞、轻视、慌乱、厌恶和傲慢。下述在 ICSLP'98 上提交的论文都描述了从有关的声学特征中确定情绪状态的示例方法：Pereira 和 Watson 的“Some Acoustic Characteristics of Emotion”，Amir 和 Ron 的“Towards an Automatic Classification of Emotions in Speech”，以及 Whiteside 的“Simulated Emotions: An Acoustic study of Voice and Perturbation Measures.”

声频捕获模块 106 可包括例如模数转换器板、交互话音响应系统和麦克风中的至少一个。对话管理单元 102 可包括一个电话交互话音响应系统，例如用于实现声频捕获的同一系统。备择地，对话管理单元可以

简单地是一个对话务员的声学接口。对话管理单元 102 可包括自然语言理解 (NLO)、自然语言生成 (NLG)、有限状态语法 (FSG) 和/或用于代替或辅助话务员的对用户进行机器提示的文本至语音的合成 (TTS)。处理模块 110 可在 IVR 的处理器部分中实现, 或者可由一个单独的带有适当软件的用户计算机实现。另外, 可以利用诸如专用集成电路 (ASIC) 的专用电路实现处理模块, 或者用采用分离元件的或采用分离及集成元件组合的专用电路实现。

处理模块 110 可包括一个情绪状态分类器 114。分类器 114 可进而包括一个情绪状态分类模块 116 以及一个情绪状态原型数据库 118。

处理模块 110 还包括一个说话人聚类器及分类器 120。部件 120 还可包括一个说话人聚类及分类模块 122 和一个说话人类别数据库 124。

处理模块 110 还可包括一个语音识别器 126, 后者进而可包括一个语音识别模块 128 和一个语音原型、语言模型及语法数据库 130。语音识别器 126 可以是对话管理单元 102 的一部分, 或者, 例如是处理模块的实现内的一个独立部件。此外, 处理模块 110 可包括一个口音识别器 132, 后者进而包括一个口音识别模块 134 和一个口音数据库 136。

处理模块 110 可包括部件 114、120、126 和 132 中的任一部件; 可包括所有这些部件或它们的任何组合。

设备 100 还可包括一个后处理器 138, 其和数据仓库 112 连接并且配置成译出用户的发音并对用户发音进行关键字识别。尽管在图 1 中是按一个独立物示出的, 后处理器可以是处理模块 110 的一部分或者是处理模块 110 的任何子部件。例如, 它可实现成是语音识别器 126 的一部分。可以按专用电路或带有适当软件模块组的通用计算机上的一部分, 把后处理器 138 实现成 IVR 的处理器部分。后处理器 138 可使用语音识别器 126。后处理器 138 还可包括一个语义模块 (未示出) 以解释短语的含义。语音识别器 126 可使用该语义模块以指示某列表中某些译码候选物是无意义的并应丢弃/或用有意义的候选物代替。

声学前端 108 如技术上周知可以典型地是一个八维加能量前端。但应理解, 可采用 13、24 或任何其它数量的维数。例如可以在带有 10ms 重叠的 25ms 的帧组上计算 MEL 逆谱以及三角 (deleta) 参数和双三角

参数，即，一次和二次微商。如图 1 中所示，可把这些声学特征提供给说话人聚类器及分类器 120、语音识别器 126 和口音识别器 132。

可以由声学前端 108 提取其它类型的声学特征。可把它们标示成情绪状态特征，例如平均运行音调 (pitch)、运行音调方差、音调跳动、运行能量方差、语音速率、摆振 (shimmer)、基频以及基频中的偏差。音调跳动指的是音调的一阶导数中符号的改变次数。摆振是能量跳动。可把这些特征从声学前端 108 提供到情绪状态分类器 114。上述包括 MEL 逆谱和情绪状态特征在内的声学特征可看成是原始的即未处理的特征。

可通过 IVR 等译出用户询问。可首先例如在说话人聚类器和分类器 120 中，通过和文本无关的说话人分类系统处理语音特征。这允许根据其语音的声学类似性对说话人分类。在 1996 年 2 月 2 日提交的美国专利申请 S. N. 60/011, 058, 1997 年 1 月 28 日提交的美国专利申请 S. N. 08/787, 031 (现为 1999 年 4 月 20 日颁布的美国专利 5, 895, 447), 1997 年 1 月 28 日提交的美国专利申请 S. N. 08/788, 471 以及 1997 年 1 月 28 日提交的美国专利申请 S. N. 08/787, 029 中公开这样的系统的实现和使用，所有这些专利申请都共同转让给国际商用机器公司，这些公开都特意与本文相结合供用于各种参考。对说话人的分类可以是受监视的或不受监视的。在受监视的情况下，根据外部信息事先决定类别。典型地，这种分类可区分男与女、成人与儿童、母语说话人与非母语说话人等等。该分类处理的标示构成处理过的特征。该处理的结果可提供给情绪状态分类器 114，并且可与训练期间用于根据对某给定类观测到的均值把情绪状态特征归一化成中性情绪状态。该归一化情绪状态特征由输出情绪状态估计的情绪状态分类器 114 使用。该输出也被当成是处理过的特征的一部分。概言之，情绪状态分类器 114 可根据语音聚类器及分类器 120 生成的每个类归一化情绪状态特征。可按如下归一化一个特征。令 X_0 为额定频率。并令 X_i 为测出频率。归一化特征则由 X_i 减 X_0 给出。该量可为正、为负或为零，通常是无量纲的。

语音识别器 126 可改录来自用户的询问。这可以是与说话人无关的或者与类别相关的大词汇量连续语音识别，或者，系统可简单到是一个

关键字识别器以检测出（例如）危害等。这类系统是技术中周知的。输出可以是完整的句子，但也可得到精细颗粒；例如，各识别出的词的时间校准。该带时间印记的改录也可当作一部分处理过的特征，在后面根据本发明的方法会进一步作出说明。这样，可以译出并存储事务的每个阶段中的会话。如图 1 中所示，适当的数据从说话人聚类器及分类器 120 传送到情绪状态分类器 114 和语音识别器 126。如所提及，有可能利用用户 104 的输入语音进行口音、方言及语言识别。可根据具有要识别的不同口音的几个说话人的语音训练连续语音识别器。每个训练说话人还带有一个口音向量，该向量的各个维代表与每个 lefeme 的每个状态有关的最有可能的混合构件。可以根据这些口音向量之间的距离对说话人聚类，并且例如可通过说话人成员组的口音标识各聚类器。可以通过从用户的语音中提取口音向量并进行分类实现口音识别。如所述，可以根据用户 104 使用的词汇和词序列估计方言、社会经济类别等。通过专家语言知识可以编辑要检测的适当关键字、句子或语法错误。口音、社会经济背景、性别、年龄等是一部分处理过的特征。如图 1 中所示，任何用实箭头指示的处理过的特征可存储在数据仓库 112 中。此外，在数据仓库 112 中还可存储用虚线指示的原始特征。

在完成事务后，可在数据仓库 112 中存储任何处理过的或原始的特征，并把它们和其它已收集的数据关联起来。然后施加传统的数据开采技术。这样的技术是已知的，例如在 Alex Berson 和 Stephen J. Smith 所著的由 McGraw Hill 出版社于 1977 年出版的“Data Warehousing, Data Mining and OAP”一书中以及在 Cabena 等著的由 Prentice Hall 出版社于 1998 年出版的“Discovering Date Mining”一书中所描述的。对于给定的商用目的，例如目标营销（target marketing），通过应用适当的开采方法可自动地得到预测模型或分类器。数据仓库 112 中存储的所有数据可以以一种便利随后的数据开采的格式存储。业内人士知道用于要被开采的数据的各种适当格式，例如上述二本参考书籍中所描述。商用目的例如可包括检查对提出购买某给定产品或服务敏感的用户，检查对自动系统存在问题从而应转到话务员的用户，以及检查对服务生气并应转到上级主管的用户。用户 104 可以是使用设备 100 的某商业顾客，

或者可以是其它类型机构，例如非赢利机构、政府部门等的客户。

通过各种模型可以提取各种特征并且可动态地回送决策。后面会对此进一步讨论。

现参照图 2，其描述依据本发明的并概括地用 200 表示的一个用于和用户交互的可实时修改的话音系统。图 2 中类似于图 1 中的部件的各部件用递增了 100 的相似参考数表示。系统 200 可包括一个与前面的讨论类似的对话管理单元 202。具体地，如图 2 中所建议，单元 202 可以是一个话务员或管理人员、IVR 或者语音用户接口（VUI）。系统 200 还可包括一个和上述讨论类似的音频捕获模块 206 以及一个亦和上述讨论类似的声学前端 208。和设备 100 一样，单元 202 在需要时可和声学前端 208 直接连接，以便使用由前端 208 确定的 MEL 逆谱或其它声学特征。此外，系统 200 包括一个与上述讨论类似的但具有一些现要说明的附加特性的处理模块 210。处理模块 210 可包括一个实行用户 204 的动态分类的动态分类模块 240。从而，处理模块 210 配置成根据至少一个用户属性修改话音系统 200 的行为，所述至少一个用户属性是根据从用户的语音中提取的至少一个声学特征确定的。系统 200 还包括一个商用逻辑单元 242，后者和对话管理单元 202、动态分类模块 240 连接并供选择地和声学前端 208 连接。该商用逻辑单元可按 IVR 或 VUI 的处理部分实现，可以是适当编程的通用计算机的一部分，或者可以是专用电路。目前认为处理模块 110、210（包括模块 240）最好按通用计算机实现并且商用逻辑 242 最好在交互话音响应系统中的处理器部分中实现。如粗线 244 所建议那样，动态分类模块可配置成向商用逻辑单元 242 和对话管理单元 202 提供反馈，该反馈可以是实时反馈。

可以如所示选用地设置数据仓库 212 和后处理器 238，它们可按上面对数据收集设备 100 讨论的那样运行。然而，要强调的是，在本发明的可实时修改的话音系统 200 中，数据仓库是选用的，并且若需要该系统在可局限于由线 244 指示的对部件 240、242、202 所讨论的实时反馈上。

通过如和对话管理单元 202 连接的反馈线 244 所示那样提示话务员，处理模块 210 可以修改，至少部分地，系统 200 的行为。例如，当

检测到用户 204 的生气情绪状态时可以警告话务员并提醒对用户 204 使用缓和的词组, 或把该用户转到更高层的管理人员。另外, 处理模块 210 可修改系统 200 的商用逻辑 242。例如当处理模块 210 和商用逻辑单元 242 都是 IVR 系统的一部分时, 这是可做到的。后面会进一步讨论商用逻辑的修改例子, 其中可包括根据系统 200 检测到的用户属性修改对用户 204 的营销报价。

如所述, 处理模块 210 及其各子部件以和图 1 的处理模块 110 基本相同的方式运行。然而要注意到由图 2 中的虚线和箭头示出的语音识别模块 228 的输出对商用逻辑 242 的反馈的选项。

应注意在本申请中, 包括说明书和各附图中, 术语“心情”被认为是等同于术语“情绪状态”。

现转到图 3, 其中描述用于在数据仓库中收集和话音系统用户的话音相关的数据的一种方法的流程图。在框 302 处启动后, 该方法包括按照框 304 经过话务员和话音操纵机器系统中的至少一种和该话音系统的用户进行会话的步骤。按照框 306, 该方法还包括捕获语音波形, 该波形是和会话期间话音系统用户说出的发音相关的。按照框 308, 该方法还包括数字化该语音波形以提供数字化语音波形的步骤。按照框 310, 该方法还包括从数字化语音波形提供至少一个和至少一个用户属性相关的声学特征的步骤。该至少一个声学特征可以是上面讨论的任何特征, 例如 MEL 逆谱或任一种情绪状态特征。用户属性可包括上面讨论的任何用户属性, 即, 性别、年龄、口音和其余上述属性。最后, 按照框 316 该方法可包括在数据仓库中以利于随后的数据开采的形式存储和声学特征对应的属性数据以及至少一个标识标记的步骤, 其中所述声学特征和至少一个用户属性相关。可以使用任何所需类型的标识标记; 该术语是广泛理解的。例如, 该标识标记可以是时间印记, 其对应于某给定时间进行的某会话的各种特性, 从而标识该给定的事务; 该标识标记可以是标识号或姓名等, 以标识该用户; 或者, 该标识标记可以是任何其它与用于数据开采过程中的属性数据相关的信息条目。

如判定框 320 所示, 可以为多次附加的会话重复框 304、306、308、310 和 316 中的上述各步骤, 以提供包括属性数据和标识标记的存储数

据的集合。这可重复多次直至对数据开采存在充分的数据。接着，如框 322 中所示，可开采该存储数据集合以提供所需的信息，例如用于修改该话音系统的基础商用逻辑的信息。

如所述，按照框 316，存储步骤可包括在其中所存储的至少一个标识标记是时间印记。收集的数据越多，可建立的各种模型越好。数据收集可以注释，这可能通过使用已经训练过的用于识别各条目的现有分类器组或者仅通过估计所需条目的翻译器的注解。也可使用这二种技术的组合。上面讨论的多次附加会话最好是和多个不同的用户进行的，从而可具有来自大量说话人的数据。

按照框 310 的提取步骤可包括提取基频、基频中的偏差、平均运行音调、运行音调方差、音调跳动、运行能量方差、语音速率以及摆振中的至少一个，以作为和用户的情绪状态相关的至少一个情绪状态特征。

按照框 312，可以归一化提取出的特征，当这些特征是情绪状态的指示时相信这是特别有用的。前面已根据本发明的设备对此做了讨论。

按照框 314，方法 300 还可包括处理至少一个声学特征以确定至少一个用户属性的附加步骤。在该情况下，得到处理过的各特征，及属性数据可以是属性值本身，例如，情绪状态的某个值。这不同于只存储原始数据的方法，在该情况下属性数据可简单地是各原始特征，即，上面讨论的 MEL 逆谱或情绪状态特征。从而，概言之，在框 316 中可存储原始声学特征（例如，波形、MEL 逆谱、情绪状态特征）和处理过的声学特征（例如情绪状态的值(高兴、悲伤、迷惑)、会话翻译）中的一种或者原始及处理过的声学特征二者。

参照框 318，在按照框 314 执行处理步骤中所使用的处理模块可在每次把附加属性存储到数据仓库中时自动地得到改进。即，可以利用每条新数据改进上面根据本发明的设备所讨论的聚类、分类和识别功能。

现参照图 4，其中示出一些选用的子步骤，最好把它们和图 3 中示出连带在一起地执行。具体地，若需要，图 3 中的框 310 可包括至少提取 MEL 逆谱，如图 4 中的框 310' 内所示。在这种情况下，该方法还可包括下述步骤：根据 MEL 逆谱识别用户的语音（框 314A），翻译语音（框 314B）以及确定语音（框 314C）。可以为词的选择和词汇集中的

至少一个确定语音，以便确定用户的教育水平、用户的社会经济类别及用户的方言之中的至少一个。根据需要还可确定和词的选择以及词汇集相关的其它用户属性。在另一种意义上，可把步骤 314A、314B 和 314C 当成是图 3 的处理框 314 的子步骤。

回到图 3，该流程的结束可由框 324 表示。

现参照图 5，其中描述代表一种依据本发明的用于把话音系统响应加工成话音系统用户的声学确定状态的方法的流程图 400。在从框 402 开始后，按照框 404，该方法包括通过该话音系统和话音系统用户进行会话的步骤。按照框 406，该方法还包括捕获与会话期间该话音系统用户说的发音相关的语音波形的步骤。此外，按照框 408，该方法包括数字化语音波形的步骤以提供数字化语音波形的步骤。而且，按照框 410，该方法包括从该数字化语音波形提取至少一个和至少一个用户属性相关的声学特征的步骤。该至少一个用户属性可包括上面讨论的任何用户属性。可以理解框 402 - 410 类似于图 3 中的框 302 - 310。

最后，按照框 415，该方法可包括根据该至少一个用户属性修改该话音系统的行为。与不带有修改的话音系统的预定响应相比，话音系统行为的修改可包括实时改变话音系统的商用逻辑和实时修改话音系统响应中的至少一种。应参照上面对本发明的设备的讨论。例如，话音系统响应的实时修改可把烦恼用户转给话务员。

按照框 410 的提取步骤可包括提取任何上述的情绪状态特征，或者提取前面讨论的任何其它特征。按照框 412，该方法可选用地包括归一化声学特征的附加步骤，具体地当声学特征是情绪状态特征的情况下。依据框 416，该方法可选用地包括在数据仓库中存储对应于声学特征的属性数据以及至少一个标识标记的附加步骤，其中所述声学特征和至少一个用户属性相关。该存储可以用有利于随后的数据开采的形式，并且可包括原始状态或处理过状态中的一种状态。该步骤基本上类似于上面由流程图 300 所代表的方法中的有关讨论。应理解，按照框 414，可利用一个处理模块处理特征以确定所需的属性。在该情况下，属性数据可以是属性本身；当未进行处理时，属性数据可以是原始声学特征。尽管图 5 中描述的方法可限制或修改话音系统的行为，若需要也可进行框

418 的改进步骤、由判定框 420 控制的重复以及数据开采步骤（即，如图 3 中描述的方法那样）。框 424 表示该方法的结束。

如流程图 300 所代表的方法那样，流程图 400 所代表的方法可根据对用户语音的翻译确定某些用户属性。从而，在框 410 的提取步骤中，可包括至少 MEL 逆谱的提取。重新参照图 4，这是在框 410' 中实现的。其它的步骤可包括：基于 MEL 逆谱对用户语音的识别（框 414A）；对语音的翻译（框 414B）；以及为词的选择和词汇集中的至少一个检查语音（框 414C），从而确定用户的教育水平、用户的社会经济类别和用户的方言中的至少一种。如前面那样，可确定其它与词选择的词汇集相关的用户属性。

现参照图 6，其中描述和流程图 400 的方法的某些方面相关的某些细节。具体地，在依据流程图 400 的方法的某些实施例中，按照图 6 中的框 414D，处理步骤 414 可包括检查情绪状态特征以确定用户的情绪状态。此外，按照图 6 中的框 415A，行为修改框 415 可包括采取行动以响应于先前确定的情绪状态。这样，可以检查情绪状态特征以判定用户是否为快活（即高兴）的情绪状态还是处于厌恶、轻视、害怕和生气的情绪状态中的至少一种状态下。当发现用户处于快活情绪状态时，如框 415A 中采取的动作那样，可向他或她提出至少一种产品或服务。替代地，当发现用户处于快活情绪状态时，可对该用户进行营销研究，以作为框 415A 中采取的动作。

仍参照图 6，在利用情绪状态特征判定情绪状态的情况下，按照框 426 可以确定不同于情绪状态特征的特征以判定情绪状态之外的属性，并且接着按照框 428 在框 415A 中采取的动作可改变成响应于与情绪状态不同的属性。例如，当向快活的用户提出产品和服务中的至少一个时，可根据至少一个不同于情绪状态的用户属性而修改提出的产品或服务。替代地，当该快活的用户受到营销研究时，可修改该营销研究以响应于至少一个不同于情绪状态的用户属性。例如，假定要向一个愉快的用户提出产品和服务中的一种，可检查语言类型以判定用户来自鲈鱼很流行的美国南部的农区地区，此外若需要还可检查音调以判断定其为男性。然后可对该人提议提供钓鲈鱼的设备及录像带。或者，假定把要进

行营销研究的快活的对象判定成来自富有的城市地区并受过良好教育的中年女性。该营销研究可修改成向她询问其对高级化妆品、时装购买爱好或时髦度假胜地。

如所述，可以检查情绪状态特征以判定用户是否处于厌恶、轻视、害怕和生气的情绪状态之中的一种状态下。若利用 IVR 系统进行该方法，并且检测出这样的情绪状态，则框 415A 应构造成响应于用户的这种被检测出的情绪状态把该用户从 IVR 切换到话务员上。替代地，在采用混合式交互话音响应系统的情况下，若检测出类似的情绪状态，在框 415A 中采取的行动应把该用户从低层话务员切换到高层管理人员上以响应于用户的这种情绪状态。

另外，可以检查情绪状态特征以判定用户是否处于迷惑情绪状态。这可以利用技术上已知的方法，例如上面讨论的 ICSLP'98 中叙述的方法来达到。例如可通过推迟对问题的回答、结结巴巴、重复、不说话等表现出迷惑。这样，语音识别和翻译是有价值的。当检测出迷惑情绪状态时，框 415A 中采取的行动应响应于该迷惑情绪状态把用户从基本上自动的 IVR 系统转到话务员上。

本发明还包括一种可由机器读的程序存储部件，其有形地嵌有可由该机器执行的指令程序，用于实现本文中公开的任何方法的各方法步骤或者这些方法的任何步骤子集。例如，在由通用计算机或 IVR 系统的处理器部分常规地执行各方法步骤的某子集的场所，可在软盘、CD-ROM 等上写入适当的程序指令。在流程图 300 中示出的方法里，这些方法步骤应包括读出对应于语音波形的数字数据，其中所述语音波形和在话音系统用户与话务员及话音操纵机器系统之中的至少一个会话期间由该话音系统用户说出的发音相关。用于附加步骤的程序指令应包括实现框 310 和 316 中所描述的任务或者所需的任何其它框的任务的指令。

类似地，对于流程图 400 中描述的方法，要通过程序指令执行的第一步骤应包括读出和语音波形对应的数字数据，其中所述语音波形和在话音系统用户与话务员及话音操纵机器系统之中的至少一个会话期间由该话音系统用户说出的发音相关。在该程序指令组中带有附加方法步骤应该例如是框 410 和框 415 中的方法步骤，如前面所述，或者，实

际上是本文中讨论的任何其他方法步骤。

应该理解，可以通过本发明的模型提取特征并动态地回送决策。除了已经叙述的这些例子之外，当诸如顾客的用户害怕地说话时，话务员可出于各种原因，例如保证交易不是强迫的，可以倾听通话。此外，可以检测出用户（或者，操作员）的气愤，除了修改自动式或混合式 IVR 系统的响应之外，其还可用于质量控制，例如用作估评和训练顾客服务机构的手段。

本发明可扩充到声信息之外的别的信息。例如，可包括伴有或不伴有声频数据的视频信息。从而，要求进行会话的方法步骤应替代地进行视频交易。视频信息可帮助识别或分类用户属性。可以自然地通过可视电话、商亭中的摄象机、计算机上的摄象机等收集这样的数据。诸如微笑、大笑、哭等的属性和情绪状态可得到识别。此外，可以标记和某些可视觉确定的用户属性或情绪状态相对应的话音部分。这应能建立一个训练数据库，该数据库对于建立用于只通过声学数据识别用户属性的各种自动技术应是有用的。从而，可以在只从视觉确定的用户属性、只从声学确定的用户属性或二者上进行数据开采。

可以根据共同的人类经验从外观上确定用户属性，即，红脸意味愤怒或窘迫，微笑意味高兴或快活的心情，泪水意味着悲伤。此外，可以和视频及声学数据一起得到适当的生物统计数据。另外，同时可得到多于一个人的数据。例如可同时监视父母和孩子，或者，也可同时监视寻找房子或汽车的一对夫妇。可以检测到对一大堆食品高兴的孩子，同时还检测出对这样的选择不悦的父母。当太太选择购买昂贵的珠宝时，丈夫可能是生气的，而妻子是高兴的。替代地，当丈夫的选择是购买昂贵的一组高尔夫球棍时，丈夫是高兴的而他的妻子不高兴。

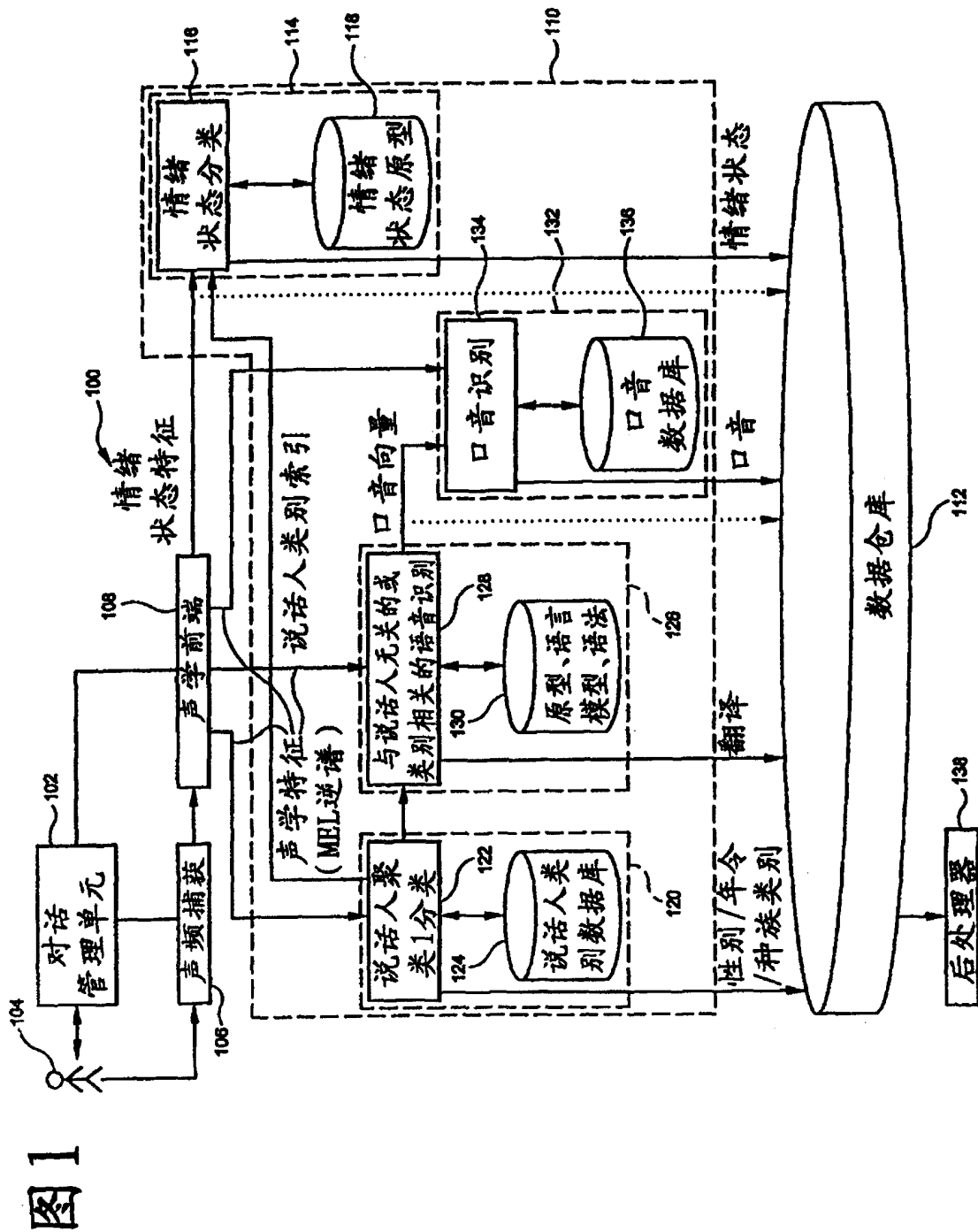
如所述，可以使用时间印记，作为一种和用户属性数据一起存储的标记。这允许研究在一天的不同时刻人们如何响应，或者观察人们在人生的不同阶段的演变，例如从儿童长大成青少年然后是成人，或者人们变老时对成年时的兴趣的改变。还可跟踪和绘出亲属间的相似性。另外，可被跟踪的一个用户属性是疲劳。这样的系统例如可安装在汽车、火车、飞机或长途载重车里以监视操作员的疲劳，并提示操作员换班并休息，

或者大声放音乐以使操作员保持清醒。Zadrozny 和 Kanevsky 的标题为“防止睡觉的汽车上对话系统”的共同转让的美国专利申请 09/078,807 (1998 年 5 月 4 日提交) 特意作为本文的参考资料。

应注意本文中讨论的话音系统可包括对计算机等通话的电话系统、公用电话亭。术语“声学特征”是广义理解的，并且如所讨论那样可包括原始特征或处理过的特征，或者二者。例如，当声学特征是 MEL 逆谱时，一些处理过的特征可包括关键词、句子部分等。某些关键字例如可以是不能接收的脏词，应把它们去掉，并导致召来管理人员，或导致对雇员的处分。还应强调，在用于进行话音系统的实时修改的设备和方法中，在数据仓库里存储带有标记的属性是供选用的，不是必须进行的。

当训练模型时，话务员在对不同的用户属性作出受教育的猜测时可注释数据。替代地，可以利用一组已训练过的现有分类器自动地完成注释。也可采用这两种技术的组合。所存储的标记除了本文中讨论的时间印记和其它项目之外可包括交易事件或结果，或者任何其它有用信息。流程图 400 描述的方法也可用于借助话务员的现场会话，以人工地改变由话务员使用的商用逻辑，或者当注意生气和其它不希望的情况时自动地召来管理人员。

尽管说明了目前认为是本发明的各种优选实施例，业内人士可意识到可在不背离本发明的实质的情况下对本发明作出各种改变和修改，并且本发明的范围内的所有这样的改变和修改是受到权利保护的。



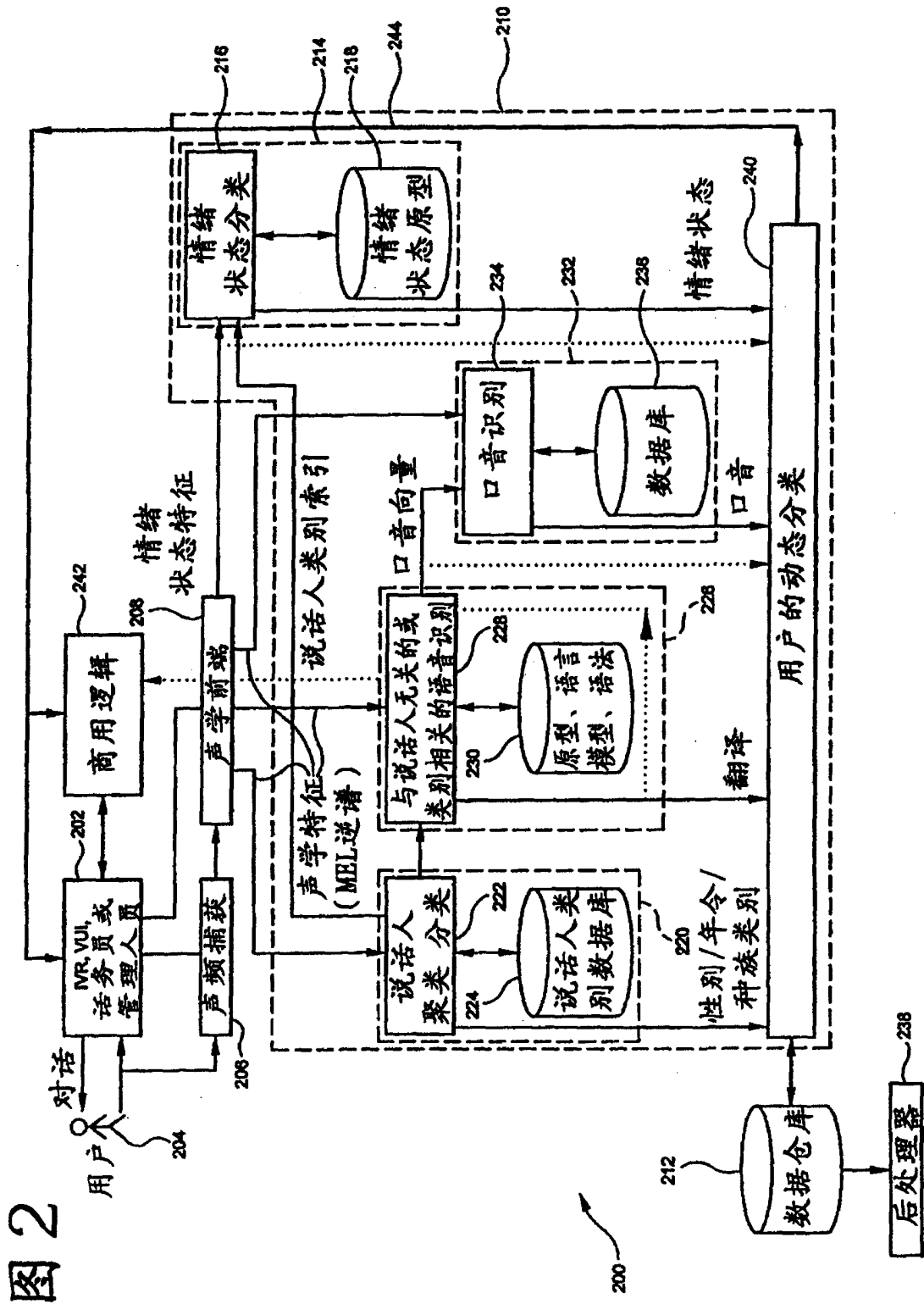


图 3

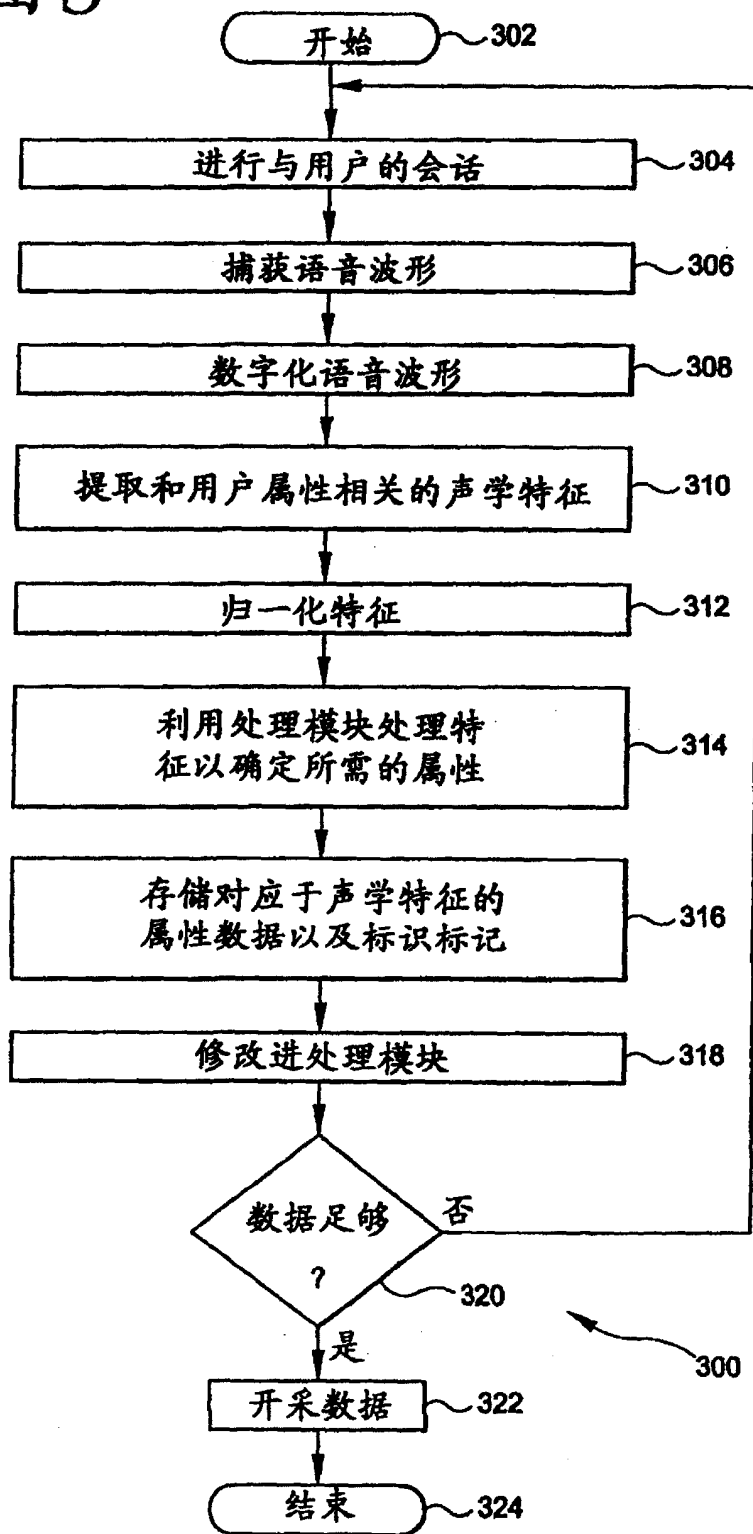


图4

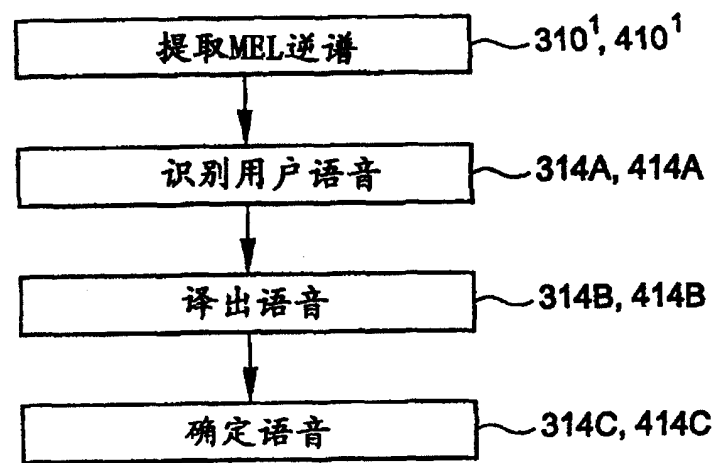


图5

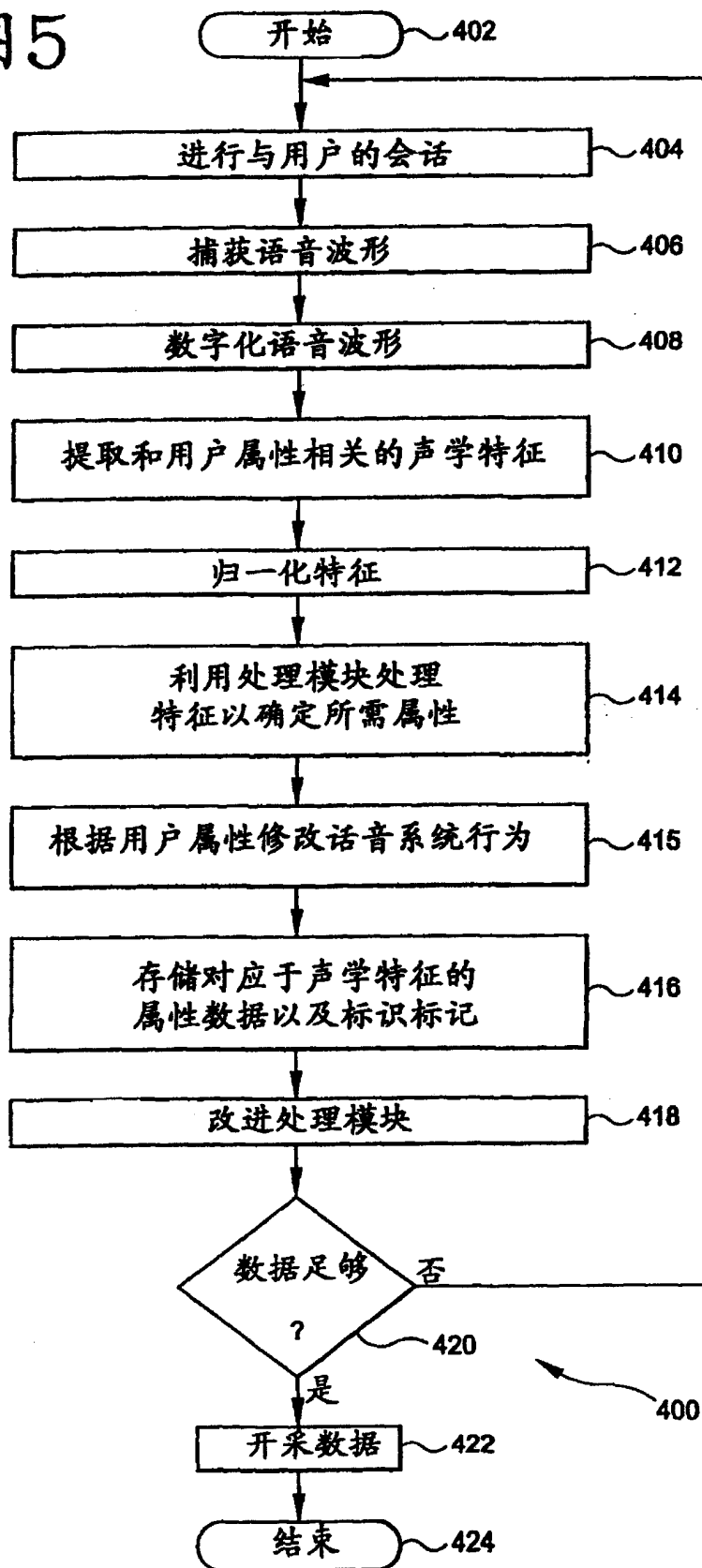


图 6

