(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101197868 B
(45)授权公告日 2012.04.04

(21)申请号 200710186930.0
(22)申请日 2007.11.15
(30)优先权数据
11/567,235 2006.12.06 US
(73)专利权人 纽昂斯通讯公司
地址 美国马萨诸塞州
(72)发明人 小查尔斯·W.·克罗斯
苏恩索恩·阿蒂瓦尼查亚丰
杰拉尔德·M.·麦科布
(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 11038
代理人 李颖
(51)Int.Cl.
H04M 1/27(2006.01)
H04M 3/493(2006.01)
(56)对比文件
US 2005172232 A1,2005.08.04, 全文.
US 2001013001 A1,2001.08.09, 全文.
WO 200219097 A1,2002.03.07, 全文.
审查员 韩峥

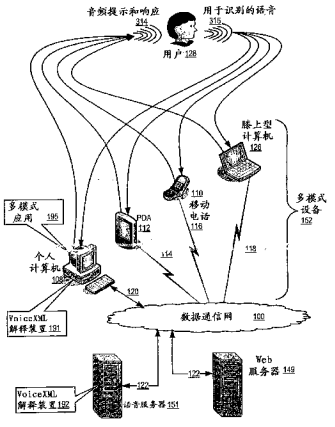
权利要求书 2 页 说明书 17 页 附图 6 页

(54)发明名称

在web页框架中启用语法的方法和系统

(57)摘要

在web页框架中启用语法,包括:在多模式设备上的多模式应用中接收框架集文档,其中该框架集文档包括定义web页框架的标记;多模式应用获取显示在每个web页框架中内容文档,其中该内容文档包括可导航标记元素;多模式应用为每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段,包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记;以及多模式应用启用所有产生的用于语音识别的语法。



CN 101197868 B

1. 一种在 web 页框架中启用语法的方法，所述方法包括：

在多模式设备上的多模式应用中接收框架集文档，所述框架集文档包括定义 web 页框架的标记；

由所述多模式应用获取显示在每个 web 页框架中的内容文档，所述内容文档包括可导航标记元素；

由所述多模式应用为每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个语音识别语法中插入识别当语音识别语法中的词匹配时待显示的内容的标记和识别将显示所述内容的框架的标记；

由所述多模式应用动态产生规定语音识别语法的标记语言片段，并向自动话音标记语言解释装置提供所述标记语言片段，以启用所有产生的用于语音识别的语法；

由所述多模式应用向自动话音标记语言解释装置提供来自用户的用于识别的语音；

由带有所述多模式应用产生的语音识别语法的所述自动话音标记语言解释装置对至少部分用于识别的语音进行匹配；以及

将指示代表匹配语音的指令的事件从所述自动话音标记语言解释装置返回至多模式应用。

2. 根据权利要求 1 所述的方法，其中：

所述 web 页框架按照一个或多个框架集组织在分层结构中，所述分层结构由最顶层的框架和一个或多个子框架表征；而且

获取所述内容文档进一步包括为所述最顶层的框架和每个子框架反复获取显示于每个框架中的独立的内容文档。

3. 根据权利要求 1 所述的方法，其中：

所述多模式设备进一步包括自动话音标记语言解释装置；并且

由所述多模式应用启用所有产生的语音识别语法进一步包括通过从所述多模式应用到自动话音标记语言解释装置的一个或多个应用编程接口 (API) 调用向所述自动话音标记语言解释装置提供规定语音识别语法的标记语言片段。

4. 根据权利要求 1 所述的方法，其中：

所述多模式设备耦接于话音服务器以用于数据通信，所述话音服务器包括自动话音标记语言解释装置；并且

由所述多模式应用启用所有产生的语音识别语法进一步包括通过从所述多模式设备到话音服务器上的自动话音标记语言解释装置的一个或多个数据通信协议消息向所述自动话音标记语言解释装置提供规定语音识别语法的标记语言片段。

5. 一种在 web 页框架中启用语法的系统，所述系统包括：

在多模式设备上的多模式应用中接收框架集文档的部件，所述框架集文档包括定义 web 页框架的标记；

由所述多模式应用获取显示于每个 web 页框架中的内容文档的部件，所述内容文档包括可导航标记元素；

由所述多模式应用为每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段的部件，其中在每个语音识别语法中插入识别当语音识别语法中的词匹配时待显示的内容的标记和识别将显示所述内容的框架的标记；以及

由所述多模式应用动态产生规定语音识别语法的标记语言片段，并向自动话音标记语言解释装置提供所述标记语言片段，以启用所有产生的用于语音识别的语法的部件；

由所述多模式应用向自动话音标记语言解释装置提供来自用户的用于识别的语音的部件；

由带有所述多模式应用产生的语音识别语法的所述自动话音标记语言解释装置对至少部分用于识别的语音进行匹配的部件；以及

将指示代表匹配语音的指令的事件从所述自动话音标记语言解释装置返回至多模式应用的部件。

6. 根据权利要求 5 所述的系统，其中：

所述 web 页框架按照一个或多个框架集组织在分层结构中，所述分层结构以最顶层的框架和一个或多个子框架表征；而且

获取所述内容文档进一步包括为所述最顶层的框架和每个子框架反复获取显示于每个框架中的独立的内容文档。

7. 根据权利要求 5 所述的系统，其中：

所述多模式设备进一步包括自动话音标记语言解释装置；并且

由所述多模式应用启用所有产生的语音识别语法进一步包括通过从所述多模式应用到自动话音标记语言解释装置的一个或多个应用编程接口 API 调用向所述自动话音标记语言解释装置提供规定语音识别语法的标记语言片段。

8. 根据权利要求 5 所述的系统，其中：

所述多模式设备耦接于话音服务器以用于数据通信，所述话音服务器包括自动话音标记语言解释装置；并且

由所述多模式应用启用所有产生的语音识别语法进一步包括通过从所述多模式设备到话音服务器上的自动话音标记语言解释装置的一个或多个数据通信协议消息向所述自动话音标记语言解释装置提供规定语音识别语法的标记语言片段。

在 web 页框架中启用语法的方法和系统

技术领域

[0001] 本发明的领域涉及数据处理，或者，更具体地，涉及用于在 web 页框架中启用语法的方法、装置和产品。

背景技术

[0002] 由于小型设备已经日益变小，与通过键盘或者指示笔运行于小型设备上的应用的用户交互变得越来越受限和麻烦。特别地，类似移动电话和 PDA 的小型手持设备通过诸如多模式接入等其他方式提供许多功能并包含充分的处理能力来支持用户交互。支持多模式接入的设备将多个用户输入方式或者通道组合在同一个交互中，允许一个用户通过多个输入模式或者通道同时与该设备上的应用交互。输入的方法包括语音识别、键盘、触摸屏、指示笔、鼠标、手写以及其他。多模式输入往往会使得小型设备的使用更加容易。

[0003] 多模式应用往往运行于提供多模式 web 页以在多模式浏览器上显示的服务器。作为本说明书中所使用的术语，“多模式浏览器”通常意味着能够接收多模式输入并且以多模式输出与用户交互的 web 浏览器。典型地，多模式浏览器展现了用 XHTML+Voice（“X+V”）编写的 web 页。X+V 提供了使用户能够通过除了诸如键盘敲击和鼠标指针动作等传统输入方式以外的口语对话与通常运行于服务器上的多模式应用交互的标记语言。X+V 通过将 XHTML（可扩展超文本标记语言）和 VoiceXML 所支持的语音识别词汇表结合起来的方式为标准 web 内容增加了口语交互。对于可视化的标记，X+V 包括 XHTML 标准。对于话音标记，X+V 包括 VoiceXML 的子集。

[0004] 目前，轻量级话音解决方案需要开发人员建立语法和词典对自动语音识别 (automatic speech recognition，ASR) 引擎所必须识别的词的可能数量加以限制——作为提高准确度的手段。典型地，一些普及的设备已经由于设备的形状因数的缘故限制了交互和输入模态，信息站设备也已经通过设计限制了交互和输入模态。在这两种情况下，实施与说话者无关的话音识别的使用来增强用户体验以及与设备的交互。与说话者无关的识别的当前技术允许写下一些复杂的话音应用，只要每个可能的话音命令都有与之相关联的有限的词汇表。例如，如果用户被提示说出城市的名称，则系统就可以相当自信地识别出所说的城市名称。

[0005] 话音交互的特性与 X+V 相结合，从而可以直接用于 X+V 内容之中。X+V 包括支持语音合成、语音对话、命令和控制以及语音语法的话音模块。话音处理装置可以附着于 X+V 元素并对具体事件做出响应。对 VoiceXML 元素与相应的可视化接口元素进行同步，X+V 采用 XML Events 事件（本文档中通常称为“事件”）。对 X+V 的详细说明可以从网页 http://www.voicexml.org 的 VoiceXML 论坛上获得。对 XHTML 和 XML Events 的详细说明可以从网址为 http://www.w3.org/MrakUp 的万维网联盟的 HTML 的主页上获得。对 VoiceXML 的详细说明可以从网址为 http://www.w3.org/voice 的万维网联盟的 Voice Browser Activity 上获得。

[0006] 多模式应用可以跨越多个 XHTML web 页。这些 web 页中的一个可以规定多个框

架，其中每个框架包含其自身的 XHTML 页面。对于 HTML 框架的概述， 可参见万维网联盟的网站 http://www.w3.org/TR/html401/present/frames.html。框架允许作者呈现浏览器同时显示的多个视图或者子窗口。一个常见的用途是将应用的导航作为独立的子窗口分离。当另一个子窗口中的内容更新时，导航子窗口并不改变。为了规定多个框架，在包括包含 <frameset> 标记元素的应用的文档中，有一个被称为“框架集文档”的顶级 XHTML 文档。一个或多个 <frame> 元素像 <frameset> 的子代一样被配置为框架集文档中的标记。每个框架都有一个名称以便多个 XHTML 文档可以作为新内容放置于其内部。每个框架都可以在识别文档的标记中通过其名称被目标定位从而显示在由框架定义的子窗口中。XHTML 文档内的 <link> 和 <anchor> 元素规定哪一个框架将通过“目标”标记属性装载引用的 XHTML 文档。如果缺少“目标”属性，则默认当前框架为目标。如果用户通过图形用户界面 (GUI) 利用鼠标单击激活框架中的超链接，则只有目标框架随着新内容被更新。

[0007] 然而，在现有技术中，只有当前受到关注的框架将启用语音识别语法。由于用户可以同时看见浏览器显示的所有框架，所以用户希望启用针对所有框架的语法。针对超链接的框架通过 GUI，而不是通过话音启用。

[0008] 另外，当话音用于激活超链接时，没有框架目标定位。与用户的言语匹配时激活启用话音的超链接的语法可源于链接的属性、标题属性、名称属性、另外的属性或者源于链接标记中开始标签和结束标签之间的文本。但是当用户说出超链接的标题且该链接被激活时，整个页面，而不是目标框架将随着新内容被更新。包括其导航框架的所有应用的框架将由单一的新页面代替。定义在框架集文档中的框架结构会被破坏，应用就变成了单框架应用。

发明内容

[0009] 本发明试图通过同时话音启用所有显示框架中的超链接并设置每个超链接的目标、以便更新的内容出现在适当框架中的方法、系统和产品来克服在多模式浏览器的 web 页框架中启用语法的技术现状的局限性。所述在 web 页框架中启用语法的方法、装置和计算机程序产品包括：在多模式设备上的多模式应用中接收框架集文档，其中该框架集文档包括定义 web 页框架的标记；由多模式应用获取显示在每个 web 页框架中的内容文档，其中该内容文档包括可导航标记元素；由多模式应用针对每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示内容的标记和识别该内容将显示于何处的框架的标记；以及由多模式应用启用所有产生的用于语音识别的语法。

[0010] 本发明前述以及其他目的、特征和优势将通过以下对如附图所示的本发明示范性实施例的更为具体的描述变得显而易见，附图中相同的参考数字通常代表本发明示范性实施例的相同部分。

附图说明

[0011] 图 1 示出了根据本发明的实施例在 web 页框架中启用语法的示范性系统的网络图。

[0012] 图 2 示出了根据本发明的实施例在 web 页框架中启用语法的包括用作多模式设备

的计算机实例的自动计算机器的框图。

[0013] 图 3 示出了根据本发明的实施例在 web 页框架中启用语法的包括用作话音服务器的计算机实例的自动计算机器的框图。

[0014] 图 4 示出了根据本发明的实施例在 web 页框架中启用语法的示范性装置的功能框图。

[0015] 图 5 示出了根据本发明的实施例在 web 页框架中启用语法的另一个示范性装置的功能框图。

[0016] 图 6 示出了根据本发明的实施例在 web 页框架中启用语法的示范性方法的流程图。

具体实施方式

[0017] 根据本发明的具体实施例，下面将结合从图 1 开始的附图对用于在 web 页框架中启用语法的示范性方法、装置和产品进行描述。图 1 示出了根据本发明的实施例在 web 页框架中启用语法的示范性系统的网络图。根据本发明的实施例，图 1 的系统通常通过：在多模式设备 (152) 上的多模式应用 (195) 中接收框架集文档，其中该框架集文档包括定义 web 页框架的标记；由多模式应用获取显示在每个 web 页框架中的内容文档，其中该内容文档包括可导航标记元素；由多模式应用针对每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段 (segment of markup)，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别该内容将显示于何处的框架的标记；以及由多模式应用启用所有产生的用于语音识别的语法，从而在 web 页框架中启用语法。典型地，图 1 中系统的工作还包括由多模式应用向自动话音标记语言解释装置 (interpreter) 提供来自用户的用于识别的语音；由带有启用语法的自动话音标记语言解释装置对至少部分用于识别的语音进行匹配；以及将指示代表匹配语音的指令的事件从自动话音标记语言解释装置返回至多模式应用。

[0018] 根据本发明的实施例，多模式应用 (195) 是能够将多模式设备作为支持在 web 页框架中启用语法的装置来操作的计算机程序指令的模块。多模式设备 (152) 为自动设备，即在能够接收来自用户的语音输入、将语音数字化并且向自动话音标记语言解释装置提供数字化语音和话音识别语法的自动计算机器或者在自动设备上运行的计算机程序。多模式设备可以和例如膝上型计算机上话音启用的浏览器、电话听筒上的话音浏览器、与个人计算机上的 Java 一同执行的在线游戏以及和本领域的技术人员可能想到的其他硬件和软件的组合一起实现。图 1 的系统包括几个实例多模式设备：

[0019] ●为了数据通信通过有线连接 (120) 耦接于数据通信网 (100) 的个人计算机 (108)，

[0020] ●为了数据通信通过无线连接 (114) 耦接于数据通信网 (100) 的个人数字助理 (PDA)(108)，

[0021] ●为了数据通信通过无线连接 (116) 耦接于数据通信网 (100) 的移动电话 (110)，以及

[0022] ●为了数据通信通过无线连接 (118) 耦接于数据通信网 (100) 的膝上型计算机 (126)。

[0023] 图 1 系统中的每个实例多模式设备 (152) 都包括麦克风、音频放大器、数模转换器以及能够从用户 (128) 接受用于识别的语音 (315)、将语音数字化并且向自动话音标记语言解释装置提供数字化语音和话音识别语法的多模式应用。可以根据工业标准的编解码器，包括但不局限于那些同样用于分布式语音识别的编解码器对语音进行数字化。用于对语音进行“编码 / 解码”的方法称为“编解码器”。欧洲电信标准协会 (ETSI) 提供了几种可用于 DSR 中的对语音进行编码的编解码器，包括，例如 ETSI ES 201 108 DSR 前端编解码器、ETSIES 202 050 高级 DSR 前端编解码器、ETSI ES 202 211 扩展 DSR 前端编解码器以及 ETSI ES 202 212 扩展高级 DSR 前端编解码器。在诸如标题为

[0024] RTP Payload Format for European TelecommunicationsStandards Institute(ETSI)European Standard ES 201 108Distributed Speech Recognition Encoding

[0025] 的 RFC3557 和标题为

[0026] RTP Payload Formats for European TelecommunicationsStandards Institute(ETSI)European Standard ES 202 050, ES 202211, and ES 202 212 Distributed Speech Recognition Encoding

[0027] 的因特网草案的标准中，IETF 为不同的编解码器提供了标准的 RTP 净荷格式。因此，值得注意的是本发明中没有关于编解码器、净荷格式或者分组结构的限制。根据本发明的实施例，可以通过包括例如：

[0028] ● AMR(自适应多速率语音编码器)

[0029] ● ARDOR(自适应速率失真优化声音编码器)

[0030] ●杜比数码 (A/52, AC3)

[0031] ● DTS(DTS 相干声学)

[0032] ● MP1(MPEG 音频层 -1)

[0033] ● MP2(MPEG 音频层 -2) 层 2 音频编解码器 (MPEG-1, MPEG-2 和非 ISO MPEG-2.5)

[0034] ● MP3(MPEG 音频层 -3) 层 3 音频编解码器 (MPEG-1, MPEG-2 和非 ISO MPEG-2.5)

[0035] ●感知音频编码

[0036] ● FS-1015(LPC-10),

[0037] ● FS-1016(CELP),

[0038] ● G.726(A DPCM),

[0039] ● G.728(LD-CELP)

[0040] ● G.729(CS-ACELP)

[0041] ● GSM

[0042] ● HILN(MPEG-4 参数音频编码) 以及

[0043] ●本领域的技术人员可能想到的其他

[0044] 任何编解码器对用于在 web 页框架中启用语法的语音进行编码。

[0045] 图 1 系统中的每个实例多模式设备 (152) 可以包括自动话音标记语言解释装置。自动话音标记语言解释装置 (191) 可以本地安装于多模式设备本身，或者自动话音标记语言解释装置 (192) 可以跨过数据通信网 (100) 相对于该多模式设备远程安装在话音服务器 (151) 中。当多模式设备包括自动话音标记语言解释装置时，可以通过借助从多模式应用到

自动话音标记语言解释装置的一个或多个应用编程接口 (API) 调用向该自动话音标记语言解释装置提供语法完成启用产生的语法。当自动话音标记语言解释装置位于话音服务器时，该多模式设备可以为了数据通信耦接于话音服务器，可以通过借助从多模式应用到话音服务器上的自动话音标记语言解释装置的一个或多个通信协议消息向自动话音标记语言解释装置提供语法完成启用产生的语法。

[0046] 根据本发明的实施例，图 1 系统中的每个实例多模式设备 (152) 都被配置并编程为能够通过：在多模式设备 (152) 上的多模式应用 (195) 中接收框架集文档，其中该框架集文档包括定义 web 页框架的标记；由多模式应用获取显示在每个 web 页框架中的内容文档，其中该内容文档包括可导航标记元素；由多模式应用针对每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记；以及由多模式应用启用 (enable) 所有产生的用于语音识别的语法，在 web 页框架中启用语法。

[0047] 对这四个实例多模式设备 (152) 的描述仅仅用于解释本发明，而并非对本发明加以限制。根据本发明的实施例，任何能够接受来自用户的语音、向自动话音标记语言解释装置提供数字化的语音并且接收和播放语音提示和响应的自动计算机器都可以改进为用于在 web 页框架中启用语法的多模式设备。

[0048] 图 1 的系统还包括通过有线连接 (122) 连接于数据通信网 (100) 的话音服务器 (151)。话音服务器 (151) 是运行例如，诸如 VoiceXML 解释装置等自动话音标记语言解释装置的计算机，自动话音标记语言解释装置通过接受带有话音识别语法的语音识别请求并返回可能包括表示识别的语音或事件的文本以由多模式客户应用处理的响应，来为多模式设备提供话音识别服务。话音服务器 (151) 还向多模式客户端应用，例如诸如 X+V 应用或者 Java 话音应用中的用户输入提供用于话音提示和话音响应 (314) 的文本到语音 (TTS) 转换。

[0049] 图 1 的系统包括为了数据通信连接多模式设备 (152) 和话音服务器 (151) 的数据通信网 (100)。根据本发明的实施例，用于在 web 页框架中启用语法的数据通信网是由多个为了带有分组交换协议的数据通信而连接的起数据通信路由器作用的计算机组成的数据通信网。这种数据通信网可以通过光连接、有线连接或者无线连接的方式实现。这种数据通信网可以包括企业内部互联网、因特网、局域数据通信网 (LAN) 和广域数据通信网 (WAN)。这种数据通信网可以实现，例如：

[0050] ●具有 Ethernet™ 协议或者无线 Ethernet™ 协议的链路层，

[0051] ●具有因特网协议 (IP) 的数据通信网络层，

[0052] ●具有传输控制协议 (TCP) 或者用户数据报协议 (UDP) 的传输层，

[0053] ●具有超文本传输协议 (HTTP)、会话初始协议 (SIP)、实时协议 (RTP)、分布式多模式同步协议 (DMSP)、无线接入协议 (WAP)、手持设备传输协议 (HDTP)、被称为 H.323 的 ITU 协议的应用层，以及

[0054] ●本领域的技术人员所能想到的其他协议。

[0055] 组成图 1 所示的示范性系统的话音服务器 (151)、多模式设备 (152) 和数据通信网 (100) 的排列仅仅是为了解释本发明，而并非对本发明加以限制。根据本发明的不同实施例，可用于在 web 页框架中启用语法的数据处理系统可以包括图 1 中未示出而本领域的

技术人员可能想到的额外的服务器、路由器、其他设备和对等体系结构。这种数据处理系统中的数据通信网可以支持除上面所提及的那些协议之外的许多数据通信协议。可以在除图1所示的那些硬件平台之外的多种硬件平台上实现本发明的不同实施例。

[0056] 术语“标记”用于本文指的是HTML、XHTML、XML、X+V、VoiceXML等标记语言中的标记元素和标记属性。web页框架是定义了多个用于内容显示的视图、窗口或子窗口的标记，例如，XHTML<frame>元素。术语“框架”(frame)既用来指定义视图的标记又用来指视图本身。多个视图为设计者提供了使特定信息可视的途径，而其他视图可以被滚动或替换。例如，在同一个窗口中，一个框架可能显示静态横幅，第二个框架可能显示导航菜单，而第三个框架可能显示能够通过第二个框架中的导航滚动或者替换的主文档。

[0057] 框架集文档是描述框架布局的标记文档，例如诸如X+V文档。框架集文档具有与没有框架的HTML文档不同的标记。标准的HMTL、XHTML或者X+V文档有一个<head>部分和一个<body>。框架集文档具有<head>和取代了<body>的<frameset>。标记文档的<frameset>部分规定了计算机显示屏上视图的布局。框架中的待显示内容不包括在框架集文档中框架被定义的同一文档里。这些内容在另一个文档，“内容文档”中，典型地，该文档远程存储在web服务器上，而往往不是向多模式设备提供框架集文档的同一web服务器上。内容文档的位置在框架标记，“scr”属性中规定。典型地，每个内容文档实际上都是web页本身，典型地，HTML、XHTML、XML或者X+V文档还包含诸如链接<link>元素和锚<a>元素等可导航标记元素。

[0058] 语法是向自动话音标记语言解释装置传递可被识别的词和词的顺序的标记。根据本发明的实施例，用于在web页框架中启用语法的语法可以以任何ASR引擎所支持的任何格式表示，包括以例如Java语音语法格式(JSGF)、W3C语音识别语法规范(SRGS)的格式、源于IETF RFC2234的增强型Backus-Naur格式(ABNF)、以W3C的随机语言模型(N-Gram)规范中描述的随机语法的形式以及本领域技术人员可能想到的其他语法格式来表示。典型地，语法如同对话的元素，例如诸如VoiceXML<menu>或者X+V<form>一样工作。语法的定义可以在对话(dialog)中内嵌表示。或者语法可以在独立的语法文档中外部实现并在对话内通过URL引用。这里是用JSFG表示语法的实例：

[0059] <grammar scope =”dialog”><! [CDATA[

[0060] #JSGF V 1.0;

[0061] grammar command;

[0062] <command> = [remind me to]call|phone|telephone<name>

[0063] <when>;

[0064] <name> = bob|martha|joe|pete|chris|john|artoush;

[0065] <when> = today|this afternoon|tomorrow|next week;

[0066]]]>

[0067] </grammar>

[0068] 在本实例中，标记元素<command>、<name>和<when>是语法的规则。规则是规则名称和向自动话音标记语言解释装置建议当前哪些词可以被识别的规则扩展的组合。在本实例中，扩展包括联合(conjunction)和析取(disjunction)，垂直条“|”表示“或”。自动话音标记语言解释装置依次对规则进行处理，首先是<command>，其次是<name>，再次是

<when>。<command>规则匹配“call”或“phone”或“telephone”加上,即结合从<name>规则和<when>规则返回的任何东西。<name>规则匹配“bob”或“martha”或“joe”或“pete”或“chris”或“john”或“artoush”,<when>规则匹配“today”或“this afternoon”或“tomorrow”或“next week”。命令语法总体上匹配类似这些的言语,例如:

[0069] ●“phone bob next week,”

[0070] ●“telephone martha this afternoon,”

[0071] ●“remind me to call chris tomorrow,”以及

[0072] ●“remind me to phone pete today.”

[0073] 图1的系统包括采用诸如HTTP等请求/响应协议向多模式设备(152)提供web页、常规web页和框架集文档的web服务器(149)。可以通过在HTTP消息中接收诸如本实例框架集文档的框架集文档来完成在多模式设备(152)的多模式应用(195)中框架集文档的接收,其中框架集文档包括定义web页框架的标记:

[0074] <! DOCTYPE HTML PUBLIC “-//W3C//DTD HTML 4.01

[0075] Frameset//EN”

[0076] “http://www.w3.org/TR/html4/frameset.dtd”>

[0077] <HTML>

[0078] <HEAD>

[0079] <TITLE>A frameset document</TITLE>

[0080] </HEAD>

[0081] <FRAMESET id=“frameset1”cols=“33%,33%,33%”>

[0082] <FRAMESET id=“frameset2”rows=“*,200”>

[0083] <FRAME id=“frame1”scr=“contents_of_frame1.html”>

[0084] <FRAME id=“frame2”scr=“contents_of_frame2.gif”>

[0085] </FRAMESET>

[0086] <FRAME id=“frame3”scr=“contents_of_frame3.html”>

[0087] <FRAME id=“frame4”ser=“contents_of_frame4.html”>

[0088] </FRAMESET>

[0089] 该框架集文档定义了通过框架集“frameset1”和“frameset2”组织在分层结构中的四个框架。Frameset2嵌套在frameset1中,创建了frame3和frame4在顶层而frame1和frame2在下层的分层结构。每个框架中待显示的内容文档在src属性中被识别为名为“contents_of_frame1.html”、“contents_of_frame3.html”和“contents_of_frame4.html”的三个HTML文档以及一幅图像,名为“contents_of_frame2.gif”的可交换图形格式(GIF)文档。每个src值,即每个内容文档名称实际上都是相对的统一资源定位符(URL),它除了提供内容文档的名称以外,还规定了该内容文档在信息空间中的位置(在本实例中,相对于被视为基准位置的//www.w3.org/TR/html4/)。

[0090] 本实例中的每个HTML内容文档都可以包含可导航标记元素、链接元素和锚元素。GIF文档可以不包含导航元素。通过借助HTTP从//www.w3.org/TR/html4/检索被识别的内容文档,可由多模式应用获得显示在每个web页框架(此处为frame1到frame4)中的内容文档。然后,多模式应用通常将每个内容文档显示在其被称为内容文档的“目标框架”的

指定框架中。

[0091] 多模式应用为每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记。在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记可以通过在每个文档中扫描可导航标记元素、链接元素和锚元素来完成（每个元素都具有规定为另一个内容文档提供位置的 URL 并且将“href”值、URL 写入语法的“href”属性）。当自动话音标记语言解释装置将词与来自用户的用于识别的语音匹配时，则语法中的该词“匹配”。在每个这种语法中插入识别将显示该内容的框架的标记可通过在语法中插入该内容文档的目标框架的框架标识，“id”属性值来完成。这样，下面来自内容文档的实例锚元素：

[0092] Pizza

[0093] Demo

[0094] 就为话音激活该锚元素所表示的超链接产生下列语法：

[0095] $grammar = Pizza Demo {$.link =”pizza/pizza.html”;

[0096] $.target =”contentFrame”}

[0097] 根据本发明的实施例，多模式应用为由框架集文档中的文档定位的每个内容文档中的每个导航元素创建语法。然后，多模式应用可以通过动态产生规定语法的标记语言片段并向自动话音标记语言解释装置提供该标记语言片段来启用所有产生的用于语音识别的语法。动态产生规定语法的标记语言片段意味着将每个产生的语法放置在当这一语法中的词由自动话音标记语言解释装置匹配时向多模式应用返回事件的标记段中。

[0098] 这样，多模式应用可以利用应用编程接口 (API) 调用或者数据通信协议中的消息为自动话音标记语言解释装置提供包含 <link> 元素的标记段，例如诸如 VoiceXML 段。当链接语法被匹配时，解释结果作为事件被提交回应用。以下是包括产生的语法和事件的 VoiceXML 链接元素的实例：

[0099] <vxml:link

[0100] eventexpr =”application.lastresult$.interpretation.c3n”>

[0101] <vxml:grammar><! [CDATA[

[0102] #JSGF V1.0;

[0103] $grammar = Pizza Demo{$.link =”pizza/pizza.html”;

[0104] $.target =“contentFrame”}

[0105]]]>

[0106] </vxml:grammar>

[0107] <catch event =”command link”>

[0108] <value

[0109] expr =”window.c3nEvent(application.lastresult$.interpretation.c

[0110] 3n)”/>

[0111] </catch>

[0112] <vxml:link>

[0113] 当 VoiceXML 解释装置与用户的言语匹配时，其语义解释功能构造事件串。事件是

与内容文档中的元素变得关联（以其为目标）的特定异步发生（如元素表示上的鼠标单击、元素的语法中词的匹配、元素的属性值中的算术错误或者众多其他可能性中的任何一种）的表示。多模式应用的一般行为是当事件发生时，通过将其传递至 DOM 文档树来将其分派到事件发生处的元素（称为其目标）。动作是对事件进行响应的某种方式；处理装置 (handler) 是针对这种动作的某种规范，例如采用脚本或者某种其他方式。监听器是这种处理程序到以文档中某个元素为目标的事件的绑定。在本实例中，事件是锚元素所代表的超链接的话音激活，处理程序是 <catch> 元素，而监听器是由多模式应用中的 <form> 元素所规定的对话。

[0114] 包括该 Pizza Demo 实例里 <vxml:link> 的“eventexpr”属性中的事件串导致了语义解释功能将该事件串作为调用 Pizza Demo 锚元素所代表的超链接的事件提交 (raise)。<vxml:link> 也包括处理由语义解释功能产生的事件的 <catch> 元素。在 catch 元素内，文档对象模型 (DOM) 功能“window.c3nEvent()”被执行，并经过事件串。

[0115] 多模式应用为来自由目标框架引用的内容文档中可导航标记元素的 <vxml:link> 元素产生标记。多模式应用将 <vxml:link> 和 <catch> 添加至带有语法的标记段并将完整的标记段提供给 VoiceXML 解释装置。 现在如果用户发出“Pizza Demo”，则包含“application.lastresult$.interpretation.c3n”的 <vxml:link> 的事件表达属性解析到串“link.pizza/pizza.html.contentFram”。该事件被 <vxml:link> 抛出并由 <vxml:link> 中的 <catch> 处理程序捕获。捕获处理程序中被调用的 DOM API 根据由包含在 <vxml:link> 元素中的语法所建立的事件分层结构对该事件串进行解释。以“command.”开始的串可以解释为菜单命令，而以“link.”开始的串可以解释为内容导航。该 Pizza Demo 是内容导航的实例。

[0116] 根据本发明的实施例，在 web 页框架中启用语法通常通过一个或多个多模式设备，即自动计算机器或者计算机实现。例如，在图 1 的系统中，所有的多模式设备至少在某种程度实现为计算机。因此，为了进一步解释本发明，图 2 示出了根据本发明的实施例在 web 页框架中启用语法的包括用作多模式设备 (152) 的计算机实例的自动计算机器的框图。图 2 的多模式设备 (152) 包括至少一个计算机处理器 (156) 或“CPU”以及通过高速存储器总线 (166) 和总线适配器 (158) 连接于处理器 (156) 和多模式设备其他部件的随机存取存储器 (168)(RAM)。

[0117] 根据本发明的实施例，存储在 RAM(168) 中的有多模式应用 (195)，能够将多模式设备作为支持在 web 页框架中启用语法的装置来操作的计算机程序指令的模块。根据本发明的实施例，本实例中的多模式应用 (195) 被编程为通过：在多模式设备 (152) 上接收框架集文档，其中该框架集文档包括定义 web 页框架的标记；由多模式应用获取显示在每个 web 页框架中的内容文档，其中该内容文档包括可导航标记元素；由多模式应用针对每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记；以及由多模式应用启用所有产生的用于语音识别的语法，来在 web 页框架中启用语法。本实例中的多模式应用 (195) 被编程为向自动话音标记语言解释装置提供来自用户的用于识别的语音。在本实例中，自动话音标记语言解释装置表示为 VoiceXML 解释装置 (192)。当自动话音标记语言解释装置将用户话音中的一个或多个词与启用的语法匹配时，多模式应用从

解释装置接受并处理指示代表匹配语音的指令的事件。自动话音标记语言解释装置 (192) 包括语法 (104)，该语法如上所述依次包括定义当前针对识别启用了哪些词和词的顺序的规则。

[0118] 典型地，多模式应用 (195) 是提供语音接口的用户级、多模式、客户端的计算机程序，其中通过所述语音接口，用户可以通过麦克风 (176) 提供用于识别的口述语音，通过音频放大器 (195) 和声卡 (174) 的编码器 / 解码器 (编解码器) (183) 将语音数字化，并且将用于识别的数字化话音提供给此处表示为 VoiceXML 解释装置的自动话音标记语言解释装置 (192)。多模式应用可以是其本身处理语法并直接通过 API 为 ASR 引擎 (150) 提供语法和用于识别的数字化语音的 Java 话音应用。或者多模式应用可以是运行于浏览器或者微浏览器内将 VoiceXML 语法通过 API 调用直接传递给嵌入式 VoiceXML 解释装置 (192) 处理的 X+V 应用。嵌入式 VoiceXML 解释装置 (192) 可以直接通过 API 调用向嵌入式 ASR 引擎 (150) 依次发出语音识别请求。多模式应用 (195) 还通过到嵌入式 TTS 引擎 (194) 的 API 调用，向例如诸如 X+V 应用或者 Java 话音应用等多模式应用中的用户输入提供用于话音提示和话音响应的 TTS 转换。本实例中的多模式应用 (195) 不通过网络将用于识别的话音发送给话音服务器识别，本实例中的多模式应用 (195) 不通过网络从话音服务器接收 TTS 提示和响应。本实例中所有的语法处理、话音识别和文本到语音转换都在多模式设备本身中以嵌入式的方式完成。

[0119] 在本实例中，同样存储在 RAM 中的 ASR 引擎 (150) 是用于完成自动语音识别的计算机程序指令的模块。根据本发明的实施例，可以改进为用于在 web 页框架中启用语法的嵌入式 ASR 引擎的实例是 IBM 的 Embedded ViaVoice Enterprise，一种也包括嵌入式 TTS 引擎的 ASR 产品。存储在 RAM(168) 中的还有嵌入式 TTS 引擎 (194)，是将文本作为输入接受并且将相同文本以数字编码语音的形式返回的计算机程序指令的模块，可用于为多模式系统的用户提供作为提示和响应的语音。

[0120] 存储在 RAM(168) 中的还有操作系统 (154)。根据本发明的实施例，可用于话音服务器中的操作系统包括 $Unix_{TM}$、$Linux_{TM}$、Microsoft NT_{TM}、AIX_{TM}、IBM's $i5/OS_{TM}$ 以及本领域的技术人员可能想到的其他操作系统。图 3 的实例中，操作系统 (154)、多模式应用 (195)、VoiceXML 解释装置 (192)、ASR 引擎 (150)、JVM(102) 和 TTS 引擎 (194) 都显示为在 RAM(168) 中，但是典型地，这种软件的许多组件也存储在非易失存储器中，例如，在磁盘驱动器 (170) 上。

[0121] 图 2 的多模式设备 (152) 包括总线适配器 (158)，包含针对高速总线、前端总线 (162)、视频总线 (164) 和存储器总线 (166) 以及针对较慢扩展总线 (160) 的驱动电子技术的计算机硬件部件。根据本发明的实施例，可用于多模式设备的总线适配器的实例包括 Intel 北桥 (Northbridge)、Intel 存储器控制器集线器、Intel 南桥和 Intel I/O 控制器集线器。根据本发明的实施例，可用于多模式设备的扩展总线的实例包括工业标准体系结构 (ISA) 总线和外设部件互联 (PCI) 总线。

[0122] 图 2 的多模式设备 (152) 包括通过扩展总线 (160) 和总线适配器 (150) 耦接于处理器 (156) 和多模式设备 (152) 的其他部件的磁盘驱动适配器 (172)。磁盘驱动适配器 (172) 以磁盘驱动器 (170) 的形式将非易失数据存储器连接至多模式设备 (152)。可用于多模式设备的磁盘驱动适配器包括集成驱动电子技术 (IDE) 适配器、小型计算机系统接口

(SCSI) 适配器和本领域的技术人员可能想到的其他适配器。另外，非易失计算机存储器可以针对多模式设备实现为光盘驱动器、电可擦除可编程只读存储空间（所谓的“EEPROM”或者“Flash”存储器）、RAM 驱动器以及本领域的技术人员可能想到的其他存储器等等。

[0123] 图 2 的实例多模式设备包括一个或者多个输入 / 输出 (I/O) 适配器 (178)。多模式设备中的 I/O 适配器通过例如，用于控制到诸如计算机显示屏等显示设备以及来自诸如键盘和鼠标等用户输入设备 (181) 的用户输入的软件驱动程序和计算机硬件实现面向用户的输入 / 输出。图 2 的多模式设备包括视频适配器 (209)，它是为了向诸如显示屏和计算机监视器等显示设备 (180) 进行图形输入而专门设计的 I/O 适配器的实例。视频适配器 (209) 通过高速视频总线 (164)、总线适配器 (158) 和同样为高速总线的前端总线 (162) 连接于处理器 (156)。

[0124] 图 2 的多模式设备还包括声卡 (174)，它是为了从麦克风 (176) 接受模拟音频信号并将该模拟音频信号转换为数字格式以便由编解码器 (183) 做进一步处理而专门设计的 I/O 适配器的实例。声卡 (174) 通过扩展总线 (160)、总线适配器 (158) 和前端总线 (162) 连接于处理器 (156)。

[0125] 图 2 的示范性多模式设备 (152) 包括用于与其他计算机 (182) 进行数据通信以及与数据通信网 (100) 进行数据通信的通信适配器 (167)。这种数据通信可以通过串行地通过 RS-232 连接、通过诸如通用串行总线 (USB) 等外部总线、通过诸如 IP 数据通信网等数据通信网以及本领域的技术人员可能想到的其他途径完成。通信适配器实现硬件级的数据通信，通过该数据通信，一台计算机直接或通过数据通信网将数据通信发送给另一台计算机。根据本发明的实施例，可用于在 web 页框架中启用语法的通信适配器的实例包括用于有线拨号通信的调制解调器、用于有线数据通信网通信的 Ethernet(IEEE802.3) 适配器和用于无线数据通信网通信的 802.11b 适配器。

[0126] 根据本发明的实施例，某些实施例中在 web 页框架中启用语法可以通过提供语音识别的一个或者多个话音服务器、计算机（即自动计算机器）来实现。因此，为了进一步解释本发明，图 3 示出了根据本发明的实施例在 web 页框架中启用语法的包括用作话音服务器的计算机实例的自动计算机器的框图。图 3 的话音服务器 (151) 包括至少一个计算机处理器 (156) 或者 CPU 以及通过高速存储器总线 (166) 和总线适配器 (158) 连接于处理器 (156) 和话音服务器的其他部件的随机存取存储器 (168)(RAM)。

[0127] 存储在 RAM(168) 中的有多模式服务器应用 (188)，能够操作系统中话音服务器的计算机程序指令的模块，该系统被配置为完成从多模式客户机设备接收语法和用于识别的数字化语音、将语法和数字化语音传递给自动话音标记语言解释装置进行处理、并且将响应从自动话音标记语言解释装置返回至多模式设备所需的数据通信。这种响应可以包括表示被识别语音的文本、用作对话中变量值的文本以及事件（即作为来自语义解释的脚本的串表示的事件文本）。多模式服务器应用 (188) 还包括为多模式应用（例如，诸如 X+V 应用或者 Java 语音应用）中的用户输入提供用于话音提示和话音响应的文本到语音 (TTS) 转换的计算机程序指令。

[0128] 多模式服务器应用 (188) 可以用 Java、C++ 或者其他语言实现为通过向来自 X+V 客户机的 HTTP 请求提供响应支持 X+V 的 web 服务器。对于另一个实例，多模式服务器应用 (188) 可以实现为运行于 Java 虚拟机 (102) 并通过向运行于多模式设备的来自 Java 客户

机应用的 HTTP 请求提供响应支持 Java 话音框架的 Java 服务器。支持在 web 页框架中启用语法的多模式服务器应用还可以以本领域的技术人员可能想到的其他途径实现，而且所有的这些途径都在本发明的范围之内。

[0129] 图 3 的实例中设置于 RAM 的还有 ASR 引擎 (150)。ASR 引擎 (150) 是利用能够由 ASR 引擎识别的词的 ASR 词典 (106) 完成语音识别的计算机程序指令的模块。词典 (106) 是文本形式的词和表示每个词发音的音素的关联。在完成自动语音识别的过程中，ASR 引擎以至少一个数字化词的形式从自动话音标记语言解释装置接收用于识别的语音，利用该数字化词的频率分量派生语音特征矢量 (Speech Feature Vector，SFV)，再利用该 SFV 从语言特定的声学模型（未示出）推断该词的音素。举例来说，语言特定的声学模型是将 SFV 与表示具体语言中所有词的所有发音的音素关联到该做法是实际可行的程度上的数据结构、表或者数据库。然后 ASR 引擎利用该音素查找词典中的词。如果找到该词，则将该词的文本作为被识别的语音返回给自动话音标记语言解释装置。然后，自动话音标记语言解释装置可以确定该被识别的语音是否与启用的语法中的词相匹配。

[0130] 存储在 RAM 中的还有例如此处表示为 VoiceXML 解释装置 (192) 的自动话音标记语言解释装置，处理 VoiceXML 语法的计算机程序指令的模块。到 VoiceXML 解释装置 (192) 的 VoiceXML 输入可以来源于远程运行于多模式设备的 VoiceXML 客户机，来源于远程运行于多模式设备的 X+V 多模式客户机应用，或者来源于远程运行于多模式设备的 Java 客户机应用。在本实例中，VoiceXML 解释装置 (192) 解释并执行通过多模式服务器应用 (188) 从远程多媒体客户机软件接收并提供给 VoiceXML 解释装置 (192) 的 VoiceXML 段。VoiceXML 解释装置 (192) 包括语法 (104)，该语法如上所述依次包括定义当前针对识别启用了哪些词和词的顺序的规则。存储在 RAM(168) 中的还有文本到语音 (TTS) 引擎 (194)，将文本作为输入接受并以数字编码语音的形式返回相同文本的计算机程序指令的模块，可用于向多模式系统的用户提供作为提示和响应的语音。

[0131] 存储在 RAM(168) 中的还有操作系统 (154)。根据本发明的实施例，可用于话音服务器的操作系统包括 $Unix_{TM}$、$Linux_{TM}$、Microsoft NT_{TM}、AIX_{TM}、IBM's $i5/OS_{TM}$ 以及本领域的技术人员可能想到的其他操作系统。图 3 的实例中，操作系统 (154)、多模式服务器应用 (188)、VoiceXML 解释装置 (192)、ASR 引擎 (150)、JVM(102) 和 TTS 引擎 (194) 都显示为在 RAM(168) 中，但是典型地，这种软件的许多组件也存储在非易失存储器中，例如，在磁盘驱动器 (170) 上。

[0132] 图 3 的话音服务器 (151) 包括总线适配器 (158)，包含针对高速总线、前端总线 (162)、视频总线 (164) 和存储器总线 (166) 的驱动电子技术以及针对较慢扩展总线 (160) 的驱动电子技术的计算机硬件部件。根据本发明的实施例，可用于话音服务器的总线适配器的实例包括 Intel 北桥、Intel 存储器控制器集线器、Intel 南桥和 IntelI/O 控制器集线器。根据本发明的实施例，可用于话音服务器的扩展总线的实例包括工业标准体系结构 (ISA) 总线和外设部件互联 (PCI) 总线。

[0133] 图 3 的话音服务器 (151) 包括通过扩展总线 (160) 和总线适配器 (158) 耦接于处理器 (156) 和话音服务器 (151) 的其他部件的磁盘驱动适配器 (172)。磁盘驱动适配器 (172) 以磁盘驱动器 (170) 的形式将非易失数据存储器连接于话音服务器 (151)。可用于话音服务器的磁盘驱动适配器包括集成驱动电子技术 (IDE) 适配器、小型计算机系统接口

(SCSI) 适配器和本领域的技术人员可能想到的其他适配器。另外，非易失计算机存储器可以针对话音服务器实现为光盘驱动器、电可擦除可编程只读存储空间（所谓的“EEPROM”或者“Flash”存储器）、RAM 驱动器以及本领域的技术人员可能想到的其他存储器等等。

[0134] 图 3 的实例话音服务器包括一个或者多个输入 / 输出 (I/O) 适配器 (178)。话音服务器中的 I/O 适配器通过例如，用于控制到诸如计算机显示屏等显示设备以及来自诸如键盘和鼠标等用户输入设备 (181) 的用户输入的软件驱动程序和计算机硬件实现面向用户的输入 / 输出。图 3 的话音服务器包括视频适配器 (209)，它是为了向诸如显示屏和计算机监视器等显示设备 (180) 进行图形输入而专门设计的 I/O 适配器的实例。视频适配器 (209) 通过高速视频总线 (164)、总线适配器 (158) 和同样为高速总线的前端总线 (162) 连接于处理器 (156)。

[0135] 图 3 的示范性话音服务器 (151) 包括用于与其他计算机 (182) 进行数据通信以及与数据通信网 (100) 进行数据通信的通信适配器 (167)。这种数据通信可以通过串行地通过 RS-232 连接、通过诸如通用串行总线 (USB) 等外部总线、通过诸如 IP 数据通信网等数据通信网以及本领域的技术人员可能想到的其他途径完成。通信适配器实现硬件级的数据通信，通过该数据通信，一台计算机直接或通过数据通信网将数据通信发送给另一台计算机。根据本发明的实施例，可用于在 web 页框架中启用语法的通信适配器的实例包括用于有线拨号通信的调制解调器、用于有线数据通信网通信的 Ethernet(IEEE802.3) 适配器和用于无线数据通信网通信的 802.11b 适配器。

[0136] 为了进一步解释本发明，图 4 示出了根据本发明的实施例在 web 页框架中启用语法的示范性装置的功能框图。在图 4 的实例中，只有多模式设备 (152) 和用户 (128)，没有网络，没有 VOIP 连接，也没有包含远程 ASR 引擎的话音服务器。根据本发明的实施例，所有在 web 页框架中启用语法所需的部件都要安装或者嵌入于多模式设备本身，膝上型计算机、PDA、蜂窝电话等等。

[0137] 图 4 的装置与图 2 的系统以相似的方式工作。根据本发明的实施例，多模式应用 (195) 是能够将多模式设备作为在 web 页框架中启用语法的装置操作的计算机程序指令的模块。在本实例中，根据本发明的实施例，本实例中的多模式应用 (195) 也配置为通过：在多模式设备上接收框架集文档，其中该框架集文档包括定义 web 页框架的标记；由多模式应用获取显示在每个 web 页框架中的内容文档，其中该内容文档包括可导航标记元素；由多模式应用针对每个内容文档中的每个可导航标记元素产生定义语音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记；以及由多模式应用启用所有产生的用于语音识别的语法在 web 页框架中启用语法。本实例中的多模式应用 (195) 编程为向自动话音标记语言解释装置提供来自用户的用于识别的语音。在本实例中，自动话音标记语言解释装置表示为 VoiceXML 解释装置 (192)。多模式应用 (195) 接受来自用户的用于识别的语音，并通过 API(175) 将该用于识别的语音发送给 VoiceXML 解释装置 (192)。当借助启用的语法通过自动话音标记语言解释装置匹配用户语音中的一个或多个词时，多模式应用从解释装置接受并处理指示代表该匹配语音的指令的事件。VoiceXML 解释装置 (192) 包括语法 (104)，该语法如上所述依次包括定义当前针对识别启用了哪些词和词的顺序的规则。

[0138] 多模式应用 (195) 是提供语音接口的用户级、多模式、客户端的计算机程序，通过

该语音接口，用户可以通过麦克风 (176) 提供用于识别的口述话音，通过音频放大器和编解码器将语音数字化，并且将用于识别的数字化语音提供给嵌入式 ASR 引擎 (150)。多模式设备应用可以是其本身处理语法并直接通过 API(179) 为嵌入式 ASR 引擎 (150) 提供语法和用于识别的数字化语音的 Java 话音应用。或者多模式应用可以是运行于浏览器或者微浏览器内将 VoiceXML 语法通过 API(175) 直接传递给嵌入式 VoiceXML 解释装置 (192) 处理的 X+V 应用。嵌入式 VoiceXML 解释装置 (192) 可以转而通过 API(179) 向嵌入式 ASR 引擎 (150) 发出语音识别请求。多模式设备应用 (195) 还通过到嵌入式 TTS 引擎 (194)API 调用，向例如诸如 X+V 应用或者 Java 话音应用等多模式应用中的用户输入提供用于话音提示和话音响应的 TTS 转换。本实例中的多模式设备应用 (195) 不通过网络将用于识别的话音发送给话音服务器识别，本实例中的多模式设备应用 (195) 不通过网络从话音服务器接收 TTS 提示和响应。所有的语法处理、话音识别和文本到语音转换都在多模式设备本身中以嵌入式方式完成。

[0139] 为了进一步解释本发明，图 5 示出了根据本发明的实施例在 web 页框架中启用语法的另一个示范性装置的功能框图。图 5 的实例包括为了数据通信由 VOIP 连接 (216) 通过数据通信网 (100) 连接的多模式设备 (152) 和话音服务器 (151)。多模式应用 (195) 在多模式设备 (152) 上运行，而多模式服务器应用 (188) 在话音服务器 (151) 上运行。话音服务器 (151) 上还安装有带有 ASR 词典 (106) 的 ASR 引擎 (150)、JVM(102) 以及带有启用语法的 VoiceXML 解释装置 (192) 。

[0140] 代表“Voice Over Internet Protocol”的 VOIP 是用于在基于 IP 的数据通信网上对语音进行路由的一般术语。语音数据流过通用分组交换数据通信网，而不是传统的专用电路交换话音传输线。用于在 IP 数据通信网上携带话音信号的协议通常称为“Voice over IP”或者“VOIP”协议。可以在任何 IP 数据通信网，包括缺少到因特网其余部分的连接的数据通信网，例如在专用建筑物范围的局域数据通信网或者“LAN”上部署 VOIP 业务。

[0141] 许多协议用于实现 VOIP。两类最为普遍的 VOIP 是通过 IETF 的会话初始协议 (SIP) 和被称为“H.323”的 ITU 协议实现的。SIP 客户机采用 TCP 和 UDP 端口 5060 连接于 SIP 服务器。SIP 本身用于建立和拆除用于语音传输的呼叫。然后，带有 SIP 的 VOIP 采用 RTP 来传送实际的编码语音。类似地，H.323 是来自国际电信联盟标准部门的保护性建议，以便在任何分组交换数据通信网上提供视听通信会话。

[0142] 图 5 的装置和上述图 3 的系统以相似的方式工作。多模式应用 (195) 将语音接口呈现给用户 (128)，将启用的语法发送给话音服务器，提供音频提示和响应 (314) 并且接受来自用户 (128) 的用于识别的语音 (315)。多模式应用 (195) 根据某种编解码器对用于识别的语音数字化，根据 VOIP 协议将该语音打包在识别请求消息中，并且通过网络 (100) 上的 VOIP 连接 (216) 将该语音发送给话音服务器 (151)。多模式服务器应用 (188) 通过接受用于语音识别的请求（包括启用的语法和数字化语音）并返回语音识别结果（包括识别语音的文本、用作对话中的变量值的文本和作为来自语义解释的脚本的串表示的文本）为多模式设备提供话音识别服务。多模式服务器应用 (188) 包括向例如诸如 X+V 应用或 Java 语音应用等多模式应用中的用户输入提供用于话音提示和话音响应的文本到语音 (TTS) 转换的计算机程序指令。

[0143] 多模式服务器应用 (188) 接收语法和来自用户的用于识别的语音，并且将该语法

和语音传递给 VoiceXML 解释装置 (192)。VoiceXML 解释装置利用 ASR 引擎 (150) 识别单独的词并且确定词或者词的顺序是否被语法所匹配。ASR 引擎从 VoiceXML 解释装置接收用于识别的数字化语音，利用数字化语音的频率分量派生 SFV，利用该 SFV 从语言特定的声学模型（未示出）推断该词的音素，并且利用所述音素在词典 (106) 中查找该语音。

[0144] 为了进一步解释本发明，图 6 示出了根据本发明的实施例在 web 页框架中启用语法的示范性方法的流程图。图 6 的方法包括在多模式设备的多模式应用中接收 (302) 框架集文档。典型地，通过响应于数据通信协议请求消息（例如诸如返回框架集文档的 HTTP 请求）接收 web 页来完成对框架集文档的接收。该框架集文档包括定义 web 页框架的标记。以下是根据两个框架集将三个框架组织在分层结构中的框架集文档的实例：

[0145] <！ DOCTYPE HTML PUBLIC “//W3C//DTD HTML 4.01

[0146] Frameset//EN”

[0147] “http://www.w3.org/TR/html4/frameset.dtd”>

[0148] <HTML>

[0149] <HEAD>

[0150] <TITLE>A simple frameset document</TITLE>

[0151] </HEAD>

[0152] <FRAMESET id =“frameset1”cols =“20%,80%”>

[0153] <FRAMESET id =“frameset2”rows =“100,200”>

[0154] <FRAME id = “frame1” src =

[0155] “contents_of_frame1.html”>

[0156] <FRAME id = “frame2” src =

[0157] “contents_of_frame2.gif”>

[0158] </FRAMESET>

[0159] <FRAME id =“frame3”src =“contents_of_frame3.html”>

[0160] </FRAMESET>

[0161] </HTML>

[0162] 图 6 的方法还包括由多模式应用获取 (304) 显示在每个 web 页框架中的内容文档。典型地，所述内容文档是包括诸如 XHTML 链接元素和锚元素等可导航标记元素的 web 页。本实例中的内容文档是框架集文档内框架定义中的“scr”URL 值所规定的内容文档。在本实例中，内容文档被 URL 识别为 contents_of_frame1.html、contents_of_frame2.gif 和 contents_of_frame3.html。

[0163] 在本实例中，根据两个框架集将 web 页框架组织在分层结构中，而且该分层结构以最顶层的框架 frame3 以及两个子框架 frame1 和 frame2 为特征。因此，在本实例中，可以通过为最顶层的框架和每个子框架反复获取显示在每个框架中的独立的内容文档来完成对至少两个内容文档的获取。

[0164] 图 6 的方法还包括由多模式应用为每个内容文档中的每个可导航标记元素产生 (306) 定义话音识别语法的标记段，包括在每个这种语法中插入识别当语法中的词匹配时待显示的内容的标记和识别将显示该内容的框架的标记。识别当语法中的词匹配时待显示的内容的标记可以从内容文档内可导航标记元素中的“href”属性获得。识别该内容将显

示于何处的框架的标记可以从框架集文档中针对内容的目标文档的“id”属性获得。

[0165] 图 6 的方法还包括由多模式应用启用 (308) 所有产生的用于语音识别的语法。启用产生的语法可进一步地通过动态产生规定语法的标记语言片段并向自动话音标记语言解释装置提供该标记语言片段来完成。在图 6 的方法中，多模式设备可以包括自动话音标记语言解释装置，启用产生的语法可以通过借助从多模式应用到自动话音标记语言解释装置的一个或多个应用编程接口 (API) 调用向自动话音标记语言解释装置提供语法来完成。在图 6 的方法中，可选择地，多模式设备可以为了数据通信耦接于话音服务器；该话音服务器可以包括自动话音标记语言解释装置；启用所有产生的语法可以通过借助从多模式设备到话音服务器上的自动话音标记语言解释装置的一个或多个数据通信协议消息向自动话音标记语言解释装置提供语法来完成。

[0166] 图 6 的方法还包括由多模式设备向自动话音标记语言解释装置提供 (310) 来自用户的用于识别的话音。即多模式设备从麦克风获得作为模拟音频信号的用户语音并根据编解码器将该语音数字化。然后，通过 API 调用 (如果该解释装置在多模式设备上)，或通过数据通信协议消息 (如果该解释装置在网络话音服务器上)，多模式应用将数字化的语音提供给自动话音标记语言解释装置。

[0167] 图 6 的方法还包括由带有启用语法的自动话音标记语言解释装置匹配 (312) 用于识别的至少部分语音。解释装置接收数字化的语音，将其传递给 ASR 引擎并接收响应中的文本词。然后解释装置确定该文本词的任何一个是否在值和顺序上与启用的语法中的词相匹配。

[0168] 图 6 的方法还包括将指示代表匹配语音的指令的事件从自动话音标记语言解释装置返回 (314) 至多模式应用。如果解释装置将词或者词的顺序与启用的语法匹配，则解释装置将事件返回至多模式应用中的事件监听器。如果解释装置在带有多模式应用的多模式设备上，则将该事件从 API 调用返回至与该事件所定向到的元素相对应的 DOM 目标。如果解释装置在网络话音服务器上，则该事件在送往相应的 DOM 目标之前首先传递回数据通信协议消息中的多模式设备。

[0169] 鉴于本文档中前面所提出的解释，读者将认识到根据本发明的实施例在 web 页框架中启用语法提供了如下的好处 :

[0170] ●启用了将对显示中所有框架话音启用 (voice-enable) 内容导航的语法，以及

[0171] ●当话音用于激活超链接时，对特定目标框架进行定位。

[0172] 此处用于在 web 页框架中启用语法的全功能计算机系统的上下文中大量描述了本发明的示范性实施例。然而，熟悉本技术的读者将认识到为了用于任何合适的数据处理系统，本发明也可以在设置于信号承载介质上的计算机程序产品内具体化。这种信号承载介质可以是传输介质或者是针对机器可读信息的可记录介质，包括磁介质、光介质或者其他合适的介质。可记录介质的实例包括硬件驱动器中的磁盘或磁碟、用于光驱动器的紧致磁盘、磁带以及本领域的技术人员可能想到的其他介质。传输介质的实例包括用于话音通信的电话数据通信网和数据通信网，例如诸如 Ethernets™ 和通过因特网协议通信的数据通信网以及万维网。对本技术熟悉的人们将立刻认识到任何具有合适的编程手段的计算机系统都能够如同程序产品中所体现的那样执行本发明方法的步骤。对本技术熟悉的人们将立刻认识到尽管本说明书中所描述的某些示范性实施例是面向安装的软件并在计算机硬

件上执行的，然而，作为固件或者硬件实现的可选择的实施例也在本发明的范围之内。

[0173] 从前面的描述可以理解，可以对本发明的不同实施例进行修改和改变而不背离本发明真正的精神。本说明书中的描述的目的仅仅在于解释本发明而不是对其加以限制。本发明的范围仅仅通过以下权利要求书的语言来限制。

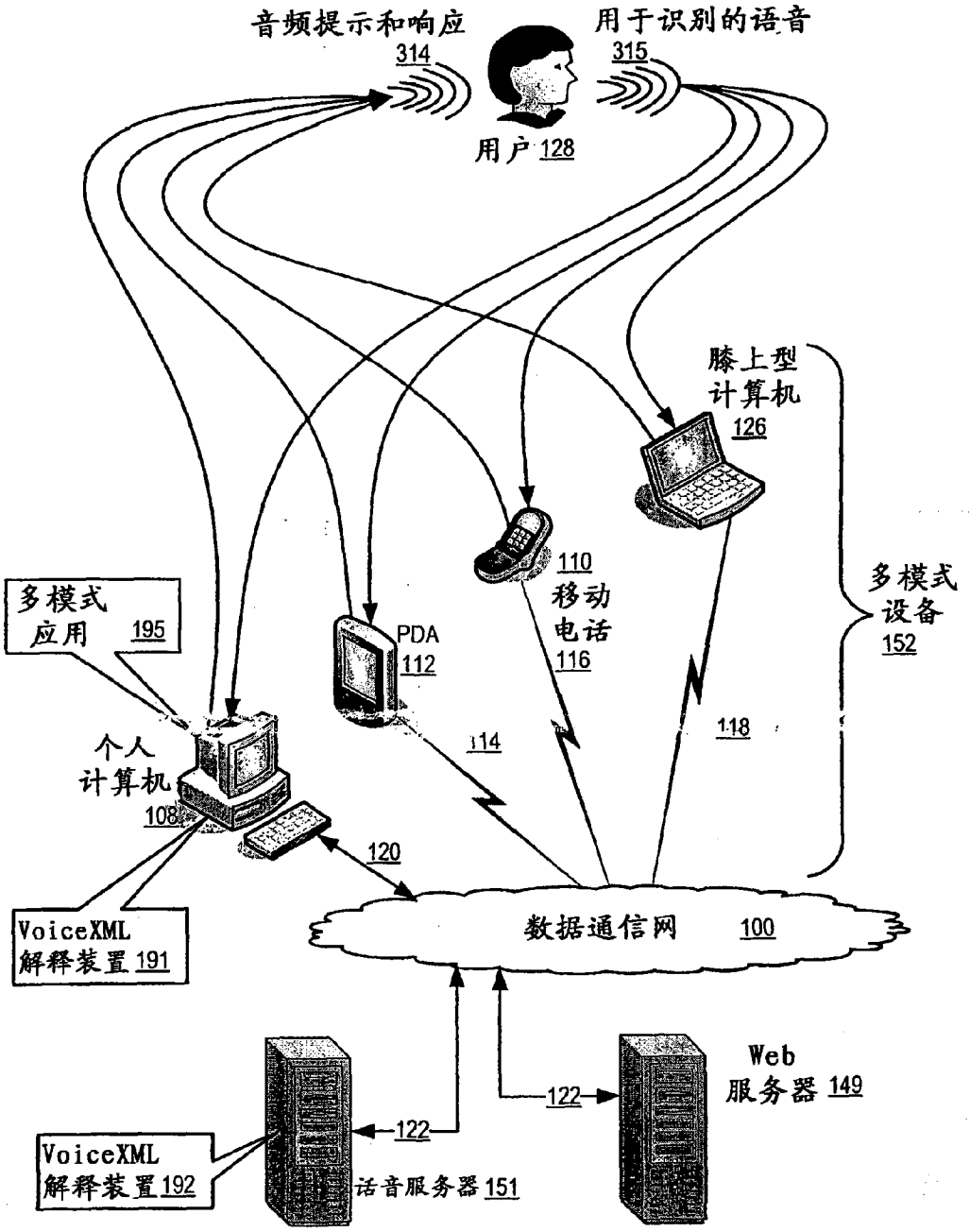

音频提示和响应
314
用于识别的语音
315
用户 128
膝上型
计算机
126
多模式
设备
152
110
移动
电话
116
多模式
应用 195
PDA
112
个人
计算机
108
114
118
120
VoiceXML
解释装置 191
数据通信网 100
Web
服务器 149
122
122
VoiceXML
解释装置 192
话音服务器 151

图1

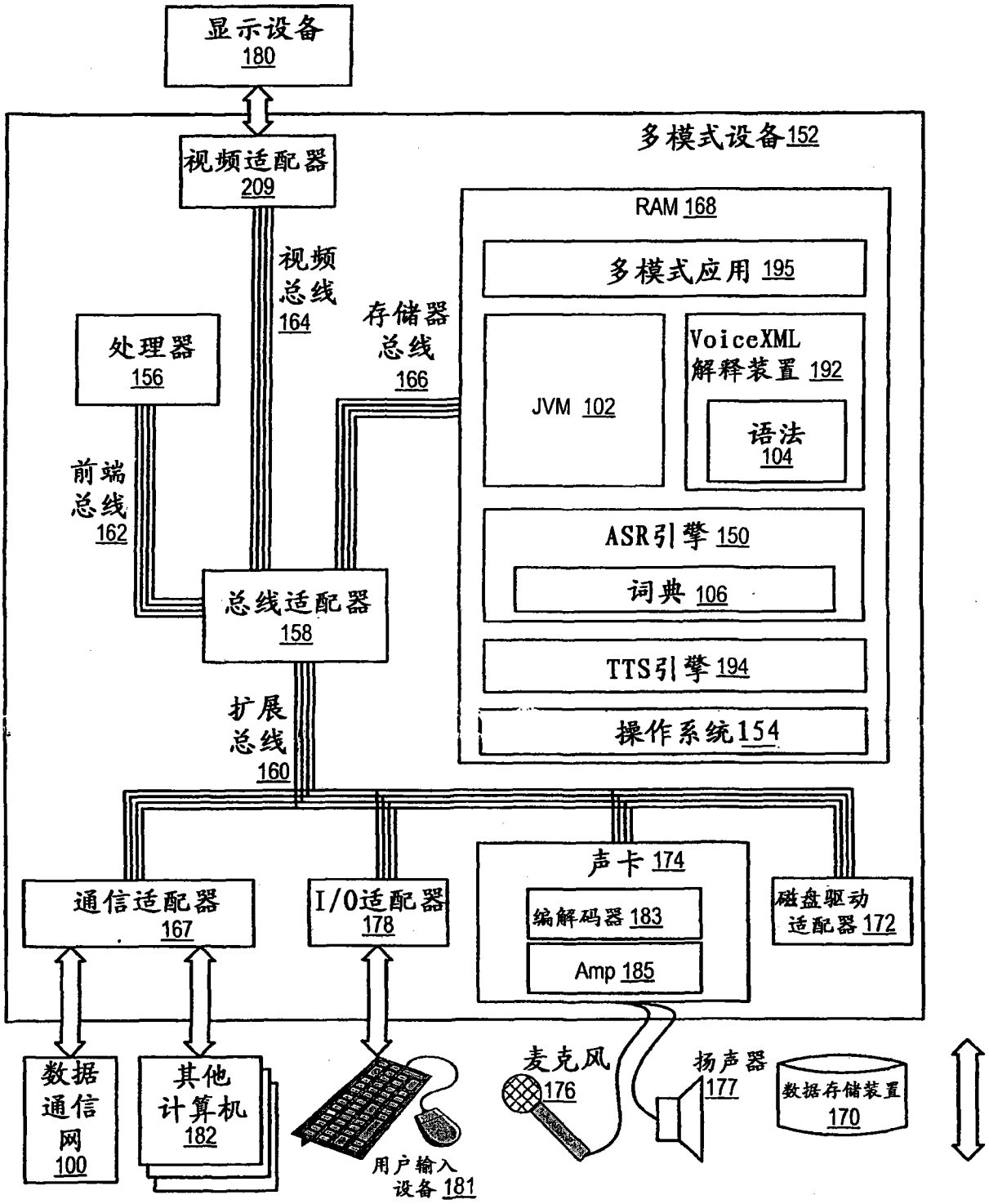

显示设备
180
视频适配器
209
多模式设备152
RAM 168
多模式应用 195
视频
总线
164
存储器
总线
166
处理器
156
JVM 102
VoiceXML
解释装置 192
语法
104
前端
总线
162
ASR引擎 150
词典 106
总线适配器
158
TTS引擎 194
操作系统154
扩展
总线
160
声卡 174
通信适配器
167
I/O适配器
178
编解码器 183
磁盘驱动
适配器 172
Amp 185
数据
通信
网
100
其他
计算机
182
麦克风
176
扬声器
177
数据存储装置
170
用户输入
设备 181

图2

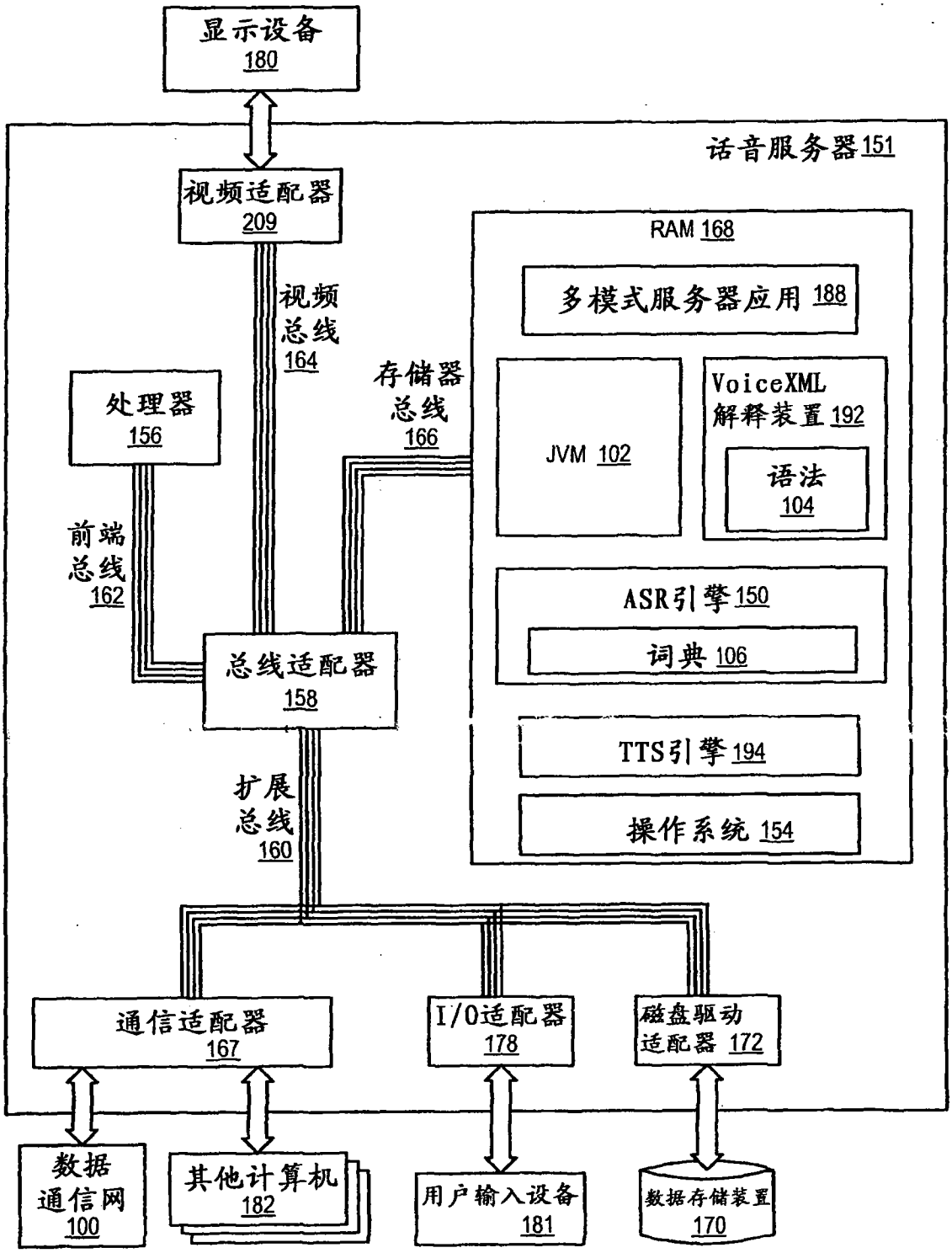

显示设备
180
话音服务器151
视频适配器
209
RAM 168
多模式服务器应用 188
视频
总线
164
存储器
总线
166
处理器
156
JVM 102
VoiceXML
解释装置 192
语法
104
前端
总线
162
ASR引擎 150
词典 106
总线适配器
158
TTS引擎 194
扩展
总线
160
操作系统 154
通信适配器
167
I/O适配器
178
磁盘驱动
适配器 172
数据
通信网
100
其他计算机
182
用户输入设备
181
数据存储装置
170

图 3

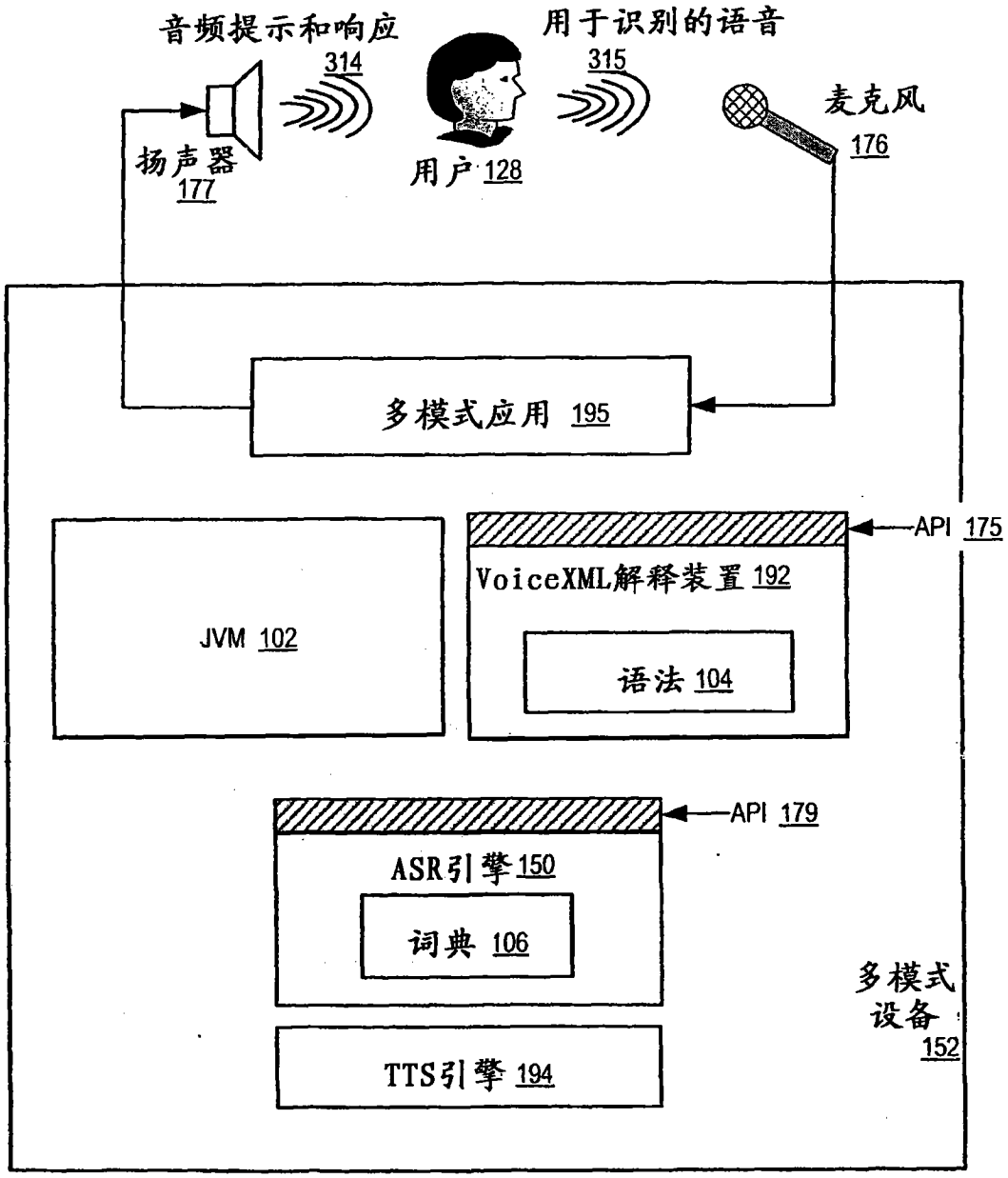

音频提示和响应
314
用于识别的语音
315
麦克风
176
扬声器
177
用户 128
多模式应用 195
API 175
VoiceXML解释装置 192
JVM 102
语法 104
API 179
ASR引擎 150
词典 106
多模式
设备
152
TTS引擎 194

图 4

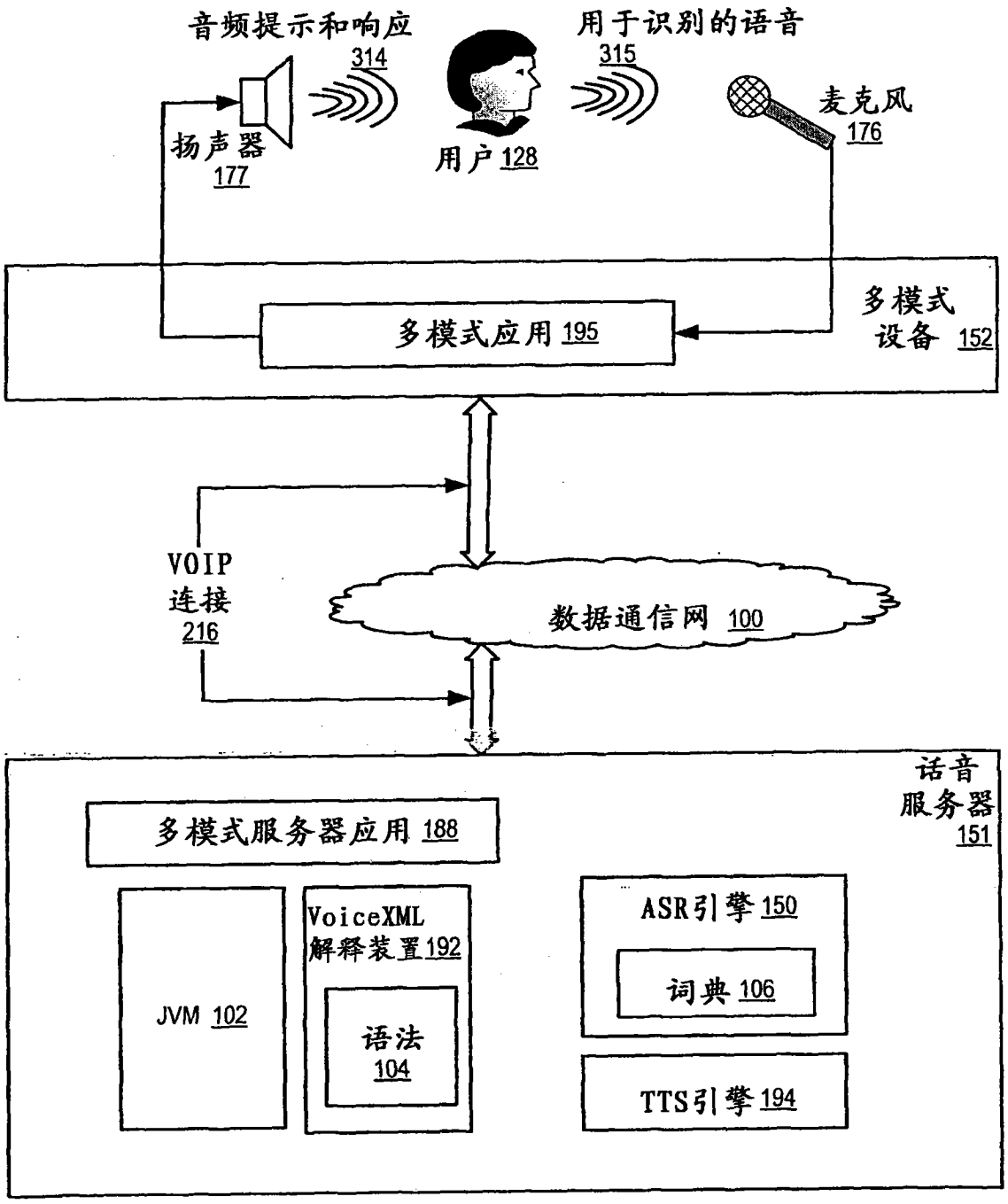

音频提示和响应
314
用于识别的语音
315
麦克风
176
扬声器
177
用户 128
多模式
设备 152
多模式应用 195
VOIP
连接
216
数据通信网 100
话音
服务器
151
多模式服务器应用 188
JVM 102
VoiceXML
解释装置 192
语法
104
ASR引擎 150
词典 106
TTS引擎 194

图 5

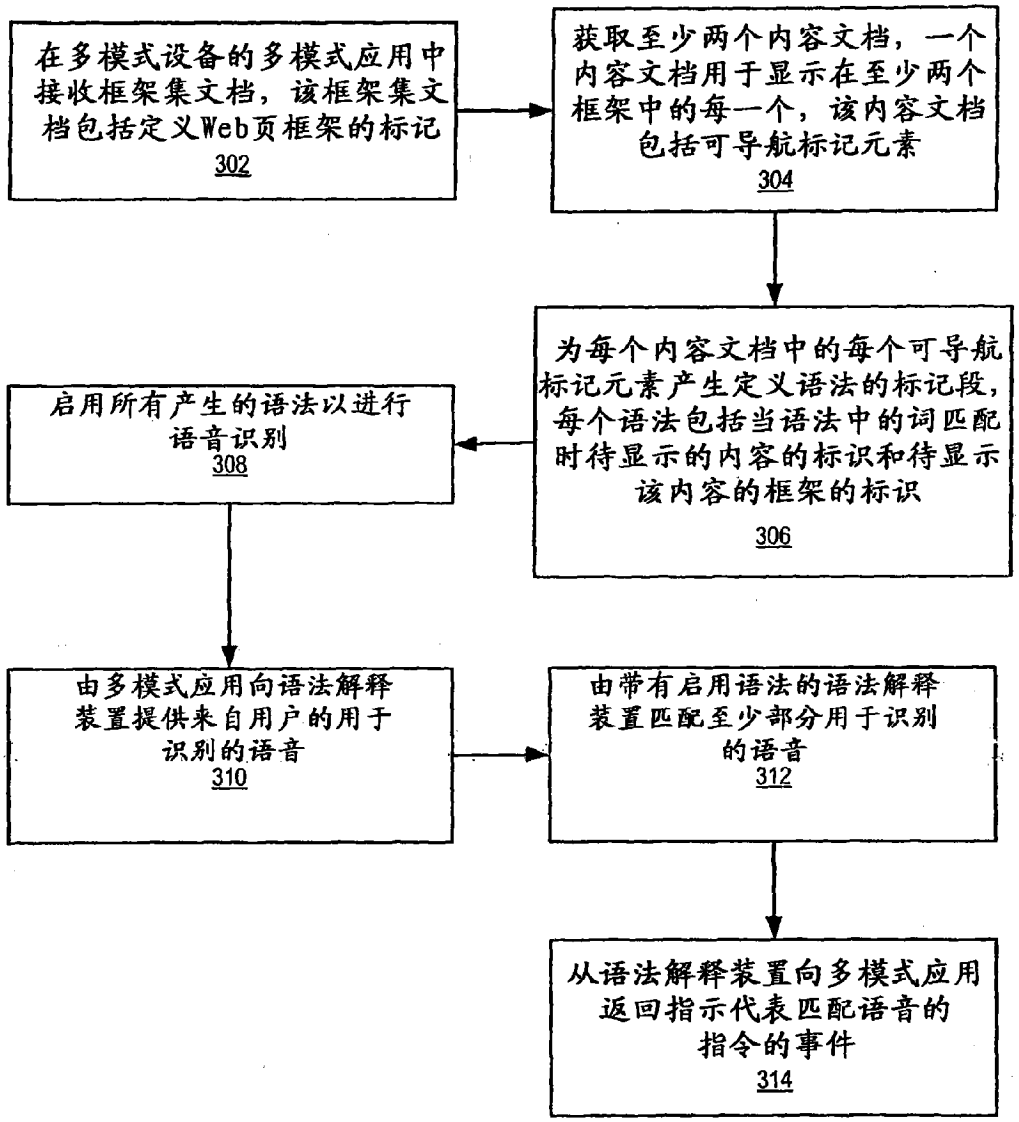

在多模式设备的多模式应用中接收框架集文档，该框架集文档包括定义Web页框架的标记
302
获取至少两个内容文档，一个内容文档用于显示在至少两个框架中的每一个，该内容文档包括可导航标记元素
304
为每个内容文档中的每个可导航标记元素产生定义语法的标记段，每个语法包括当语法中的词匹配时待显示的内容的标识和待显示该内容的框架的标识
306
启用所有产生的语法以进行语音识别
308
由多模式应用向语法解释装置提供来自用户的用于识别的语音
310
由带有启用语法的语法解释装置匹配至少部分用于识别的语音
312
从语法解释装置向多模式应用返回指示代表匹配语音的指令的事件
314

图6