(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 101156197 B
(45)授权公告日 2011.06.29

(21)申请号 200680010948.9
(22)申请日 2006.04.06
(30)优先权数据
0507148.5 2005.04.08 GB
(85)PCT申请进入国家阶段日
2007.09.29
(86)PCT申请的申请数据
PCT/EP2006/061380 2006.04.06
(87)PCT申请的公布数据
WO2006/108795 EN 2006.10.19
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 R·J·鲍沃特 C·W·小克罗斯
L·R·威尔逊 Y-L·王
(74)专利代理机构 北京市中咨律师事务所 11247
代理人 于静 李峥
(51)Int.Cl.
G10L 15/28(2006.01)
H04M 3/493(2006.01)
(56)对比文件
US 2002/0194388 A1,2002.12.19, 全文.
CN 1601992 A,2005.03.30, 全文.
CN 1329433 A,2002.01.02, 全文.
审查员 万济萍

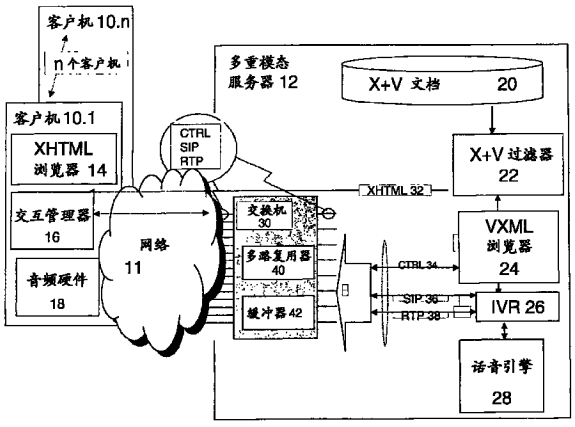
权利要求书 1 页 说明书 5 页 附图 2 页

(54)发明名称

用于多重模态话音和Web服务的方法和装置

(57)摘要

本发明基于:能够定位话音服务器,临时对其进行分配,向其发送你所说的"683航班应何时到达"的音频,将你所说的内容的结果返回到浏览器中,以及对所述话音服务器解除分配以便供与其浏览器交谈的下一人使用。话音信道和交互式话音响应服务器(IVR)端口最初由使用常规音频协议的IVR和交换机建立。该话音信道最初并未连接至客户机。所述交换机在不必与IVR进行更多通信的情况下处理IVR话音信道的分配和解除分配。用户(通常通过按PTT按钮)向客户设备指示他希望在X+V会话期间发起话音交互。这转换为在控制(CTRL)信道上对同步XHTML和VXML表单的请求,实施例将其用作VXML浏览器执行话轮的触发器。多路复用器截获该控制命令并且连接设备与现有的开放但独立的话音端口之间的虚拟话音电路。无需建立RTP信道即可连接虚拟电路。然后,将CTRL信号转发至交互管理器,从而使得会话能够发生。在该会话结束时断开该虚拟电路。



CN 101156197 B

1. 一种用于控制从音频接口到音频处理器的音频连接的方法，其包括：

为音频数据建立交换机与所述音频处理器的处理器链路；

响应于音频接口会话的建立，在所述音频接口会话中为音频数据建立交换机与所述音频接口的接口链路；

响应于指示在所述音频接口会话内音频交互的开始的信号，将所述处理器链路和所述接口链路连接在一起，由此，音频数据可以在所述音频接口与所述音频处理器之间流动；

响应于指示所述音频交互的结束的信号，断开所述处理器链路和所述接口链路之间的连接；以及

响应于所述音频接口会话的结束，拆除所述接口链路。

2. 根据权利要求 1 的方法，其中，所述音频处理器是具有用于语音识别和 / 或文本到语音转换的任选的话音服务器的交互式话音响应系统。

3. 根据权利要求 1 的方法，其中，所述音频接口是具有用于多重模态操作的话音功能性的 XML 浏览器。

4. 根据权利要求 2 的方法，其中，所述音频接口是具有用于多重模态操作的话音功能性的 XML 浏览器。

5. 根据权利要求 1、2、3 或 4 的方法，其进一步包括：如果连接所述接口链路与所述处理器链路存在延迟，则缓冲所述音频数据。

6. 根据权利要求 1、2、3 或 4 的方法，其中，所述建立交换机与所述音频处理器的处理器链路的步骤包括：使用 SIP 协议协商 RTP 连接。

7. 根据权利要求 1、2、3 或 4 的方法，其中，指示音频交互的开始的信号是指示多重模态对话的开始的预先存在的信号。

8. 根据权利要求 1、2、3 或 4 的方法，其中，指示音频交互的结束的信号是指示多重模态对话中字段的同步的预先存在的信号。

9. 一种用于控制从音频接口到音频处理器的音频连接的交换机，其包括：

用于为音频数据建立交换机与所述音频处理器的处理器链路的装置；

用于响应于音频接口会话的建立，在所述音频接口会话中为音频数据建立交换机与所述音频接口的接口链路的装置；

用于响应于指示在所述音频接口会话内音频交互的开始的信号，将所述处理器链路和所述接口链路连接在一起的装置，由此，音频数据可以在所述音频接口与所述音频处理器之间流动；

用于响应于指示所述音频交互的结束的信号，断开所述处理器链路和所述接口链路之间的连接的装置；以及

用于响应于所述音频接口会话的结束，拆除所述接口链路的装置。

用于多重模态话音和 Web 服务的方法和装置

技术领域

[0001] 本发明涉及一种用于多重模态话音和 Web 服务的方法和装置。

背景技术

[0002] 随着设备越来越小，除了键盘和记录笔之外的交互模式成为了必需。特别地，像蜂窝电话和 PDA 这样的小巧的手持设备提供很多功能并且含有足够的处理能力来应对各种任务。如今以及将来的设备将极大地获益于对多重模态访问方法的使用。

[0003] 多信道访问是由多种方法或信道（例如，电话、膝上型计算机或 PDA）来访问企业数据和应用的能力。例如，用户在办公室或在家时可以使用因特网浏览器来访问他或者她在 Web 上的银行账户余额，并且在路上时可以使用话音识别以及文本到语音转换通过非智能电话 (dumb phone) 来访问相同的信息。

[0004] 相比之下，多重模态访问是在相同的交互或会话中组合多种模式或信道的能力。输入的方法包括语音识别、键盘、触摸屏和记录笔。取决于形势和设备，输入模式的组合将使得对小巧设备的使用更加容易。例如，在 PDA 上的 Web 浏览器中，你可以通过敲打或者通过提供话音输入来选择项目。类似地，你可以使用话音或记录笔将信息输入到字段。利用多重模态技术，既可以显示又可以说出设备上的信息。

[0005] 使用 XHTML+Voice（可扩展超文本置标语言 + 话音）的多重模态应用为现今基于 VoiceXML（话音可扩展置标语言）的话音应用以及基于 XHTML 的可视应用提供了向可以服务于这两种环境以及多重模态环境的单个应用的自然的转移途径。通过建立两种信道，一种用于图形用户接口以及另一种用于话音，多重模态应用集成了话音接口和图形用户接口交互。在书写 XHTML+Voice(X+V)Profile 1.2 的同时，于 2004 年 3 月 16 日在 www.voicexml.org 上对其进行了发布。

[0006] 在具有远程话音处理的多重模态浏览器的已知实现中，在客户机与话音服务器之间建立话音信道，并且分配该信道以便在 X+V 会话内的话音交互的持续时间传送话音数据。在该话音交互之后，断开该话音信道，并且继续 X+V 会话。对于 X+V 会话内的每个单独的交互，必须建立新的话音信道，因为这避免了在 X+V 会话处于空闲时消耗服务器上宝贵的话音资源。

[0007] 为每个话音交互建立和关闭话音信道存在这样的缺点，即，由于使用现有协议 (SIP 和 RTP) 来开启和关闭话音信道所花费的时间，增加了每个话音交互的响应时间。所增加的时延是设备与服务器之间可用的网络带宽的直接函数。这导致在窄带网络（例如，慢速因特网连接）上和慢速无线网络上的问题。例如，pre-3G 无线网络上的网络带宽受限。

发明内容

[0008] 根据本发明的第一方面，提供了一种用于控制从音频接口到音频处理器的音频连接的方法，其包括：为音频数据建立与所述音频处理器的处理器链路；响应于音频接口会话的建立，在所述接口会话中为音频数据建立与所述音频接口的接口链路；响应于指示在

所述接口会话内音频交互的开始的信号，将所述处理器链路和所述接口链路连接在一起，由此，音频数据可以在所述音频接口会话与所述音频处理器之间流动；响应于指示所述音频交互的结束的信号，断开所述处理器链路和所述接口链路；以及响应于所述接口会话的结束，拆除所述接口链路。

[0009] 虚拟化在每轮基础上 (on a per turn basis) 使用的 IVR 端口同时保留虚拟化单元与设备之间的固定连接，这无疑是新颖的。在窄带网络上使用 SIP 协议建立 RTP 信道占用了不适当的容量，因为在握手时需要标识会话号和 IVR 端口号。此外，SIP 呼叫用于新的呼叫，并且在建立 RTP 信道之前需要分配资源 (VoiceXML 浏览器、语音引擎等)。新的连接协议使用单个信号触发来获得相同的连接 - 由于在现有会话信道上发送了该信号，因此无需会话号。无需端口号、会话或者话音引擎资源分配，因为已经分配了端口。所需要的唯一的分配是可用的且被配置的 RTP 信道。此外，所述新的连接协议使用现有信号来触发虚拟电路，因此根本不存在建立开销。

[0010] 创建可连接并且可断开的音频信道 (虚拟电路) 的优势在于：1) 当同时的虚拟电路的数目少于或者等于可用的物理 IVR 端口时，系统实质上并未向交互的响应时间附加时延；2) 在客户机以及在 IVR 上无需代码改变；3) 当虚拟电路的数目超过可用的物理 IVR 端口的数目时，系统性能适度下降；以及 4) 由于设备与交换机之间的永久话音电路，任何的用户音频输入均在所述交换机中被获取，并且一旦虚拟电路变得可用，该音频输入就可以在稍后被回放给所述 IVR。

[0011] 对虚拟话音信道与中间控制器的连接和断开给予控制，使得所述客户机和服务器均能够触发连接和断开。

[0012] 所述音频处理器可以是具有用于语音识别和 / 或文本到语音转换的任选的话音服务器的交互式话音响应系统。

[0013] 所述音频接口可以是具有用于多重模态操作的话音功能性的 VoiceXML 浏览器或 XML 浏览器。

[0014] 所述方法进一步包括：如果连接所述接口链路与所述处理器链路存在延迟，则缓冲所述音频数据。

[0015] 所述建立与所述音频处理器的处理器链路的步骤包括：使用 SIP 协议协商 RTP 连接。

[0016] 触发音频交互的开始的信号是指示多重模态对话的开始的预先存在的信号。

[0017] 触发音频交互的结束的信号是指示多重模态对话中字段的同步的预先存在的信号。

附图说明

[0018] 现将参照附图，仅借助于例子来描述本发明的实施例，在附图中：

[0019] 图 1 是本发明的多重模态服务器和 n 个客户机的代表；以及

[0020] 图 2 是本发明的多重模态系统中的一个客户机与多重模态服务器之间的时序图。

具体实施方式

[0021] 优选实施例的系统包括通过网络 11 连接至多重模态服务器 12 的多个客户机 10.1

到 10.n。每个客户机 10 包括 :XHTML 浏览器 14 ;交互管理器 16 ;以及音频硬件 18。服务器 12 包括 :X+V 文档数据库 20 ;XHTML+Voice(X+V) 过滤器 22 ;VoiceXML(VXML) 浏览器 24 ;交互式话音响应系统 (IVR)26 ;话音引擎 28 ;以及交换机 30。

[0022] 网络 11 支持 :XHTML(X) 信道 32 ;CTRL(控制) 信道 34 ;SIP(会话起始协议) 信道 36 ;以及 RTP(实时协议) 信道 38。X 信道 32 传送用于 X+V 交互的 XHTML 应用。在优选的实施例中,CTRL 信道 34 传送用于同步对应的 XHTML 和 VXML 字段的同步信号。X+V 定义了在 XHTML 表单和 VXML 表单中对应字段的概念。当多重模态交互可以发生的时候,发送同步数据事件信号,并且在交互之后发送同步数据信号,以便将一个 XML 字段与对应的 XML 字段同步。RTP 信道 38 传送用于音频的数据。SIP 信道 36 用于建立 RTP 信道。CTRL、RTP 和 SIP 信道并不直接地而是通过交换机 30 将客户机 10 连接至 VXML 浏览器 24/IVR 26。交换机 30 消耗很少的资源,并且因此可以支持大量的并行客户机 10,大于可用的 IVR 容量。

[0023] 关于一个客户机 10 和一个 RTP 信道 38 描述了优选的实施例,但是当存在的客户机多于存在的话音信道时,本发明的优势变得显而易见。

[0024] 客户机 XHTML 浏览器 14 解释通过 X 信道 32 所接收到的 XHTML 文档。XHTML 浏览器 14 是已知的 XHTML 浏览器,其具有附加的功能性以便与 VXML 浏览器进行交互以及与音频硬件进行交互。

[0025] 通过在 CTRL(控制) 信道 34 和 SIP(会话起始协议) 信道 36 上发送和接收控制信息,交互管理器 16 控制 XHTML 浏览器 14 与 VXML 浏览器 24 之间的交互。XHTML 浏览器 14 与 VXML 浏览器 24 之间的交互的重要方面在于,正好在话音交互之前从交互管理器 16 发送的同步事件信号,以及在话音交互之后发送的同步数据信号。同步事件信号触发 VXML 浏览器 24 中的话音对话。同步数据信号在话音交互之后同步对应的字段数据。

[0026] 在交互管理器 16 的控制下,由音频硬件在 RTP(实时协议) 信道 38 上发送和接收音频内容。

[0027] X+V 文档数据库 20 存储 X+V 文档,并且按照请求将其发送至 X+V 过滤器 22。

[0028] 按照来自 XHTML 浏览器 14 的请求,X+V 过滤器 22 从 X+V 文档数据库 20 获得 X+V 文档。将 X+V 文档过滤成 XHTML 组成部分和 VXML 组成部分。将 XHTML 组成部分发送至 XHTML 浏览器 14,并且将 VXML 组成部分发送至 VXML 浏览器 24。XHTML 组成部分含有话音处理器,以便显示对应的 X 和 V 字段以及标记 XHTML 中需要与 VXML 浏览器交互的部分。

[0029] VXML 浏览器 24 是常规的 VXML 浏览器。VXML 浏览器接收使用 X+V 会话的 X+V 文档的 VXML 组成部分来进行话音交互的请求。VXML 浏览器管理该 X+V 会话内的话音交互。对各个话音功能的处理被传递给 IVR。

[0030] 尽管在优选的实施例中,VXML 浏览器 24 和 X+V 过滤器被示于服务器 12 中,然而其还可以在客户机 10 上实现。

[0031] IVR 26 处理话音交互。响应于来自所浏览的 VXML 文档部分的请求以及作为所浏览的 VXML 文档部分的输入而被接收的双音多频信号 (DTMF),可以播放预先记录的提示。IVR 26 还与话音引擎 28 连接。IVR 26 通过交换机 30 连接至 SIP 信道 36 和 RTP 信道 38。

[0032] 话音引擎 28 为 IVR 26 进行语音识别输入和文本到语音转换输出。

[0033] 交换机 30 包括多路复用器 40 和缓冲器 42。多路复用器 40 将 (其自身与大概数千个移动设备客户机之间的) 大量客户机话音链路之一与 (其自身与 IVR 之间的) 少量 IVR

话音链路之一相连。当交换机 30 在控制信道 (CTRL) 上截获同步事件信号时，其将对应的客户机话音链路与 IVR 话音链路相连，以便在客户机与 IVR 端口之间创建虚拟话音电路。

[0034] 一旦创建了虚拟话音电路，同步事件信号便被传递给 VXML 浏览器 24，用于话音交互的处理。VXML 浏览器 24 可以更新 VXML 字段，并且然后指示 IVR 26 播放提示以及在虚拟电路上获得话音输入。虚拟电路仅维持单个话音交互的持续时间 (“话轮 (conversational turn)”)，并且由同步数据信号来告知交互的结束。当触发同步数据信号时，交换机 30 断开虚拟电路。另一客户设备然后可以在断开后立刻重用该虚拟链路的话音资源。出于可扩缩性，可以通过高速网络将交换机连接至 IVR。

[0035] 在优选的实施例中，在 X+V 会话期间的第一话音交互处开启 RTP 信道 38。可选地，交换机可以被配置以在 X+V 会话的开始处，于第一话音交互之前开启 RTP 会话。对于预先开启的 RTP 信道，交换机 30 与 IVR 26 之间的话音信道保持连接。这意味着，当需要建立新的虚拟电路时，其变成在交换机内建立路由的一种简单情况，而无需发生与 IVR 的额外的物理呼叫建立。实际上，这意味着向系统的响应性附加了可忽略的时延。

[0036] 在所有物理 IVR 端口均在使用的情况下 (即，太多的设备正在同时试图进行话音交互)，交换机 30 可以将音频存储在缓冲器 42 中。当稍后 IVR 端口变为可用时，建立虚拟电路并且将所缓冲的音频重放给 IVR，这样就完成了交互 (例如，留下话音消息)。当然，如果这经常发生，那么其意味着系统被 100％地利用，并且需要额外的 IVR 端口。

[0037] 为了说明本实施例的方法，现参照例子以及图 2 的事件图来描述两个典型的话音交互的事件。

[0038] 使用手持 PDA，用户“冲浪”至航班信息网站以查找航班到达的预计时间。GUI (用户图形接口) 显示具有两个字段的查询表单 :航班号和航班日期。用户关注第一字段，并且音频提示被播放“请输入航班号”。用户使用其 PDA 上的键盘输入航班号 (参照下文的“Web 同步”描述了该交互)。用户然后关注下一字段，并且听到音频提示“请输入日期”。这次，用户使用音频硬件，对着麦克风说“今天”。话音数据被处理，并且该天的日期被自动插入到日期字段。在下文的“话音同步”事件序列中描述了该交互。在该例中，填写了两个字段，并且航班信息站点以文本形式或话音形式或者二者来返回预计的到达时间。

[0039] 起初，客户机 10 请求 50 建立话音信道 38 以及在 SIP 信道 36 上发送一次性 SIP 请求给 IVR 26 以建立话音信道 38。然而，在交换机 30 与客户机之间建立了 52 一个客户机链路，并且在交换机与服务器之间建立了 54 至少一个服务器链路，而不是从客户机 20 到服务器的话音信道 38。客户机链路和服务器链路形成了可连接的话音电路，其可以由交换机连接并且然后断开。服务器链路的数目受限于 IVR 上的端口数，并且客户机链路的数目受限于交换机 30 的实现。

[0040] 在图 2 中，Web 同步交互 56 示出了在这样的交互中的事件，即在该交互中，在客户机 XHTML 浏览器中，XHTML 字段被引起关注并且与对应的 VXML 字段同步。由于 XHTML 字段具有对应的 VXML 字段，因此将同步事件信号从客户机发送 58 至服务器。由交换机截获同步事件，该交换机在向 X+V 服务器上的 VXML 浏览器转发同步事件之前首先连接 60 话音电路。当 VXML 浏览器接收到同步事件时，其找到对应的 VXML 组件，并且运行 62 该组件作为话音交互。在该例中，话音交互在新连接的话音信道上发送 64 话音提示。客户机接收话音提示，并且使用音频硬件来播放 66 该话音提示。然而，在该例中，用户选择使用 XHTML 接口

输入响应来响应话音提示，因此，不存在即时 VXML 响应。用户使用图形接口向 XHTML 字段中输入数据。XHTML 字段具有对应的 VXML 字段并且同步尾随其后。交互管理器在 CTRL 信道上向 X+V 服务器发送 68 同步数据信号－这是包括所输入的 XHTML 数据的同步数据信号。交换机在向 VXML 浏览器转发以便向 VXML 字段输入之前截获同步数据信号。由于 CTRL 信号是同步数据信号，因此交换机获知断开 70 虚拟话音电路。VXML 浏览器同步 72 VXML 数据，并且结束 74 话音交互。

[0041] 图 2 中的话音同步交互 76 示出了在这样的交互中的事件，即在该交互中，数据被输入到服务器 VXML 浏览器中的 VXML 字段，并且与客户机浏览器中对应的 XHTML 字段同步。当具有对应的 VXML 字段的 XHTML 字段在图形接口中被引起关注时，将同步事件信号从交互管理器经由交换机发送 78 至服务器。由交换机截获同步事件，该交换机将客户机链路连接 80 至 IVR 链路以形成话音电路。将同步事件转发至 X+V 服务器上的 VXML 浏览器。当 VXML 浏览器接收到该同步事件时，其运行 82 对应的 VXML 交互。在该例中，话音交互通过新连接的话音信道发送 84 话音提示。客户机接收话音提示，并且使用音频硬件播放 86 该话音提示。在该例中，用户通过向音频硬件说出应答（例如，“今天”）来响应该话音提示。通过由交换机所创建的虚拟话音电路将话音数据发送 88 至 IVR。IVR 向话音引擎发送话音数据用于识别 90，并且将作为结果的文本数据输入到 VXML 字段。VXML 字段具有对应的 XHTML 字段，并且沿 CTRL 信道将同步数据信号发送 92 至交互管理器。在将同步数据信号转发至交互管理器之后，交换机断开 94 话音电路。客户机结束 96 交互。X+V 服务器处理字段中的数据并且输出结果。

[0042] 只有在客户设备即将要被切断的时候，才发送 97 话音拆除信号。响应于该信号，交换机拆除 98 客户机链路，并且服务器拆除 99 IVR 链路。

[0043] 总而言之，实施例基于：能够定位话音服务器，临时对其进行分配，向其发送你所说的“今天的 683 航班应何时到达？”的音频，将你所说的内容的结果返回到浏览器中，以及对话音服务器解除分配以便供与其浏览器交谈的下一人使用。话音信道和 IVR 端口最初由使用常规音频协议的 IVR 和交换机建立。话音信道最初并未连接至客户机。交换机在不必与 IVR 连续通信的情况下处理 IVR 话音信道的分配和解除分配。用户（通常通过按 PTT 按钮）向客户设备指示他希望在 X+V 会话期间发起话音交互。这转换为在 CTRL 信道上对同步 XHTML 和 VXML 表单的请求，实施例将其用作 VXML 浏览器执行话轮的触发器。多路复用器截获该控制命令，并且连接设备与现有的开放但独立的 (unattached) 话音端口之间的虚拟话音电路。无需建立 RTP 信道即可连接虚拟电路。然后，将 CTRL 信号转发至交互管理器，从而使得会话能够发生。在会话结束时，断开虚拟电路。

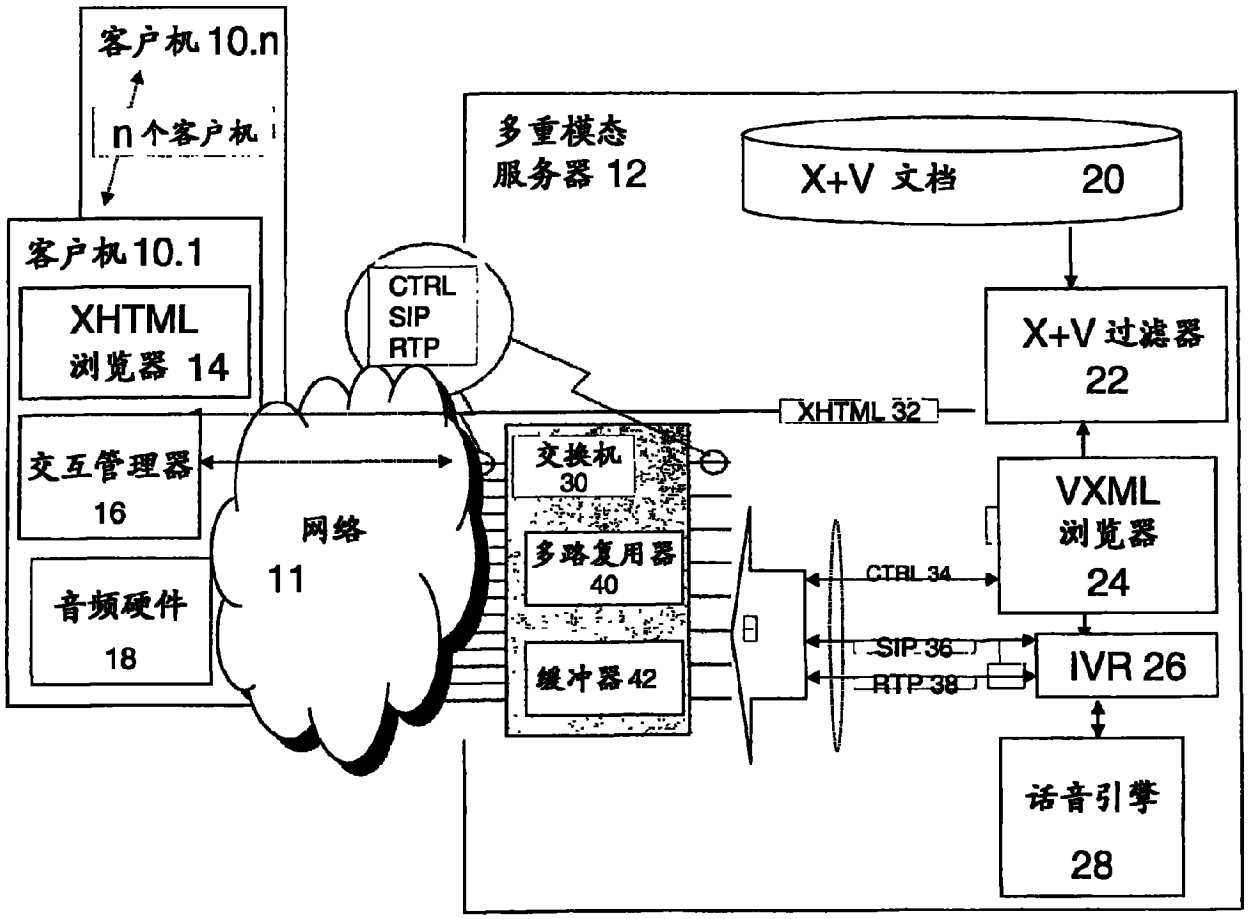

客户机 10.n
n 个客户机
客户机 10.1
XHTML
浏览器 14
交互管理器
16
音频硬件
18
网络
11
CTRL
SIP
RTP
多重模态
服务器 12
X+V 文档 20
X+V 过滤器
22
XHTML 32
交换机
30
多路复用器
40
缓冲器 42
CTRL 34
SIP 36
RTP 38
VXML
浏览器
24
IVR 26
话音引擎
28

图 1

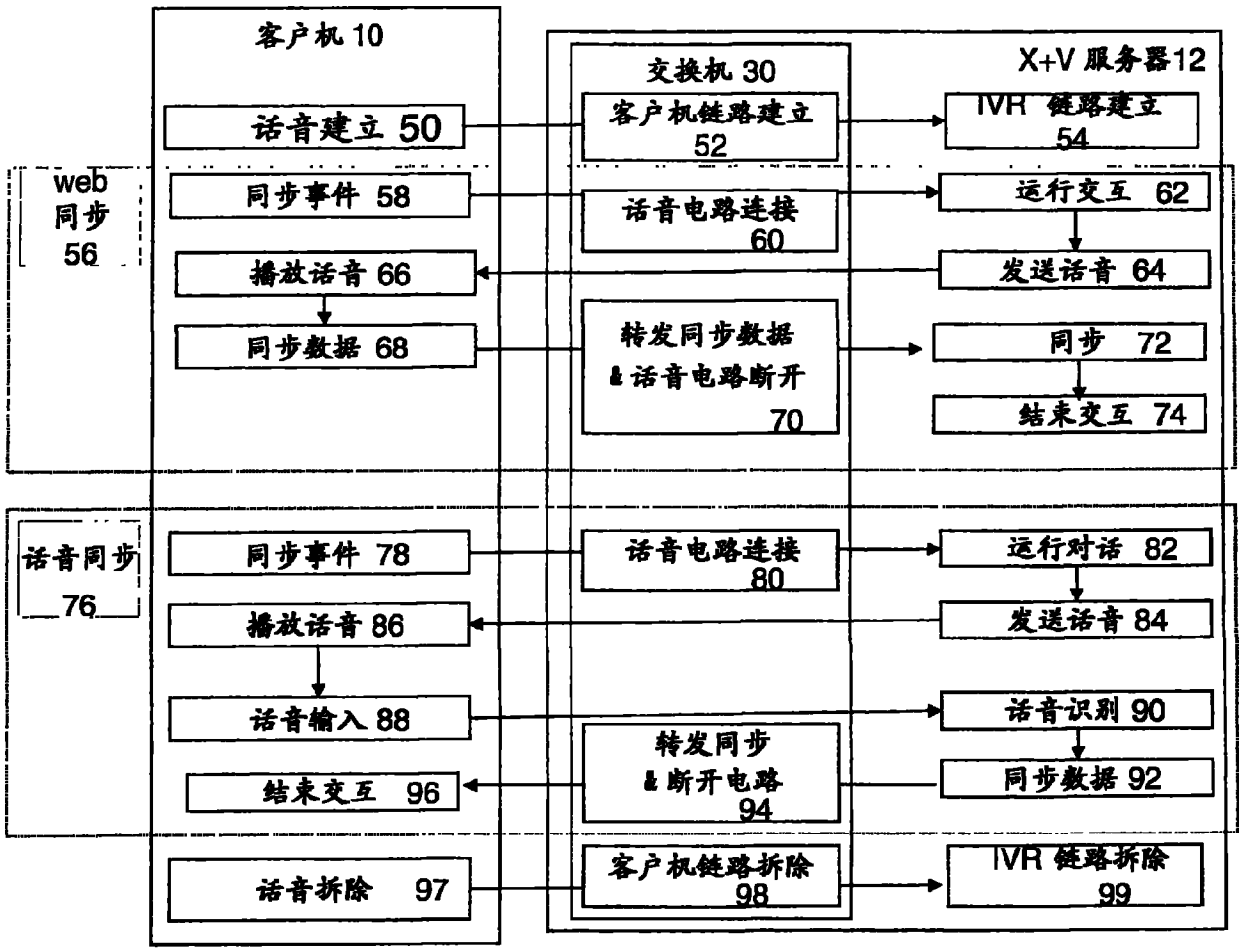

客户机 10
交换机 30
X+V 服务器12
话音建立 50
客户机链路建立 52
IVR 链路建立 54
web 同步 56
同步事件 58
话音电路连接 60
运行交互 62
播放话音 66
发送话音 64
同步数据 68
转发同步数据 & 话音电路断开 70
同步 72
结束交互 74
话音同步 76
同步事件 78
话音电路连接 80
运行对话 82
播放话音 86
发送话音 84
话音输入 88
话音识别 90
转发同步 & 断开电路 94
结束交互 96
同步数据 92
话音拆除 97
客户机链路拆除 98
IVR 链路拆除 99

图 2