

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G10L 15/14 (2006.01)

G10L 21/00 (2006.01)

G10L 21/02 (2006.01)

G10L 15/00 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200680022639.3

[43] 公开日 2009 年 12 月 16 日

[11] 公开号 CN 101606191A

[22] 申请日 2006.6.13

[21] 申请号 200680022639.3

[30] 优先权

[32] 2005. 6.28 [33] US [31] 11/168,770

[86] 国际申请 PCT/US2006/022863 2006. 6. 13

[87] 国际公布 WO2007/001821 英 2007.1.4

[85] 进入国家阶段日期 2007.12.24

[71] 申请人 微软公司

地址 美国华盛顿州

[72] 发明人 张正友 刘自成 A·艾斯洛

A·塞博拉曼亚 J·G·德罗波

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 陈 斌

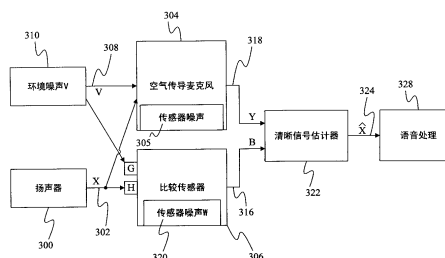
权利要求书 3 页 说明书 14 页 附图 6 页

[54] 发明名称

使用语音状态模型的多传感语音增强

[57] 摘要

一种方法和装置基于一比较传感器信号(316)和空气传导麦克风信号(318)确定语音状态的似然性。该语音状态的似然性连同该比较传感器信号和该空气传导麦克风信号一起被用于为清晰语音信号(324)估计(322)清晰语音值。



1、一种确定对代表经降噪语音信号的一部分的经降噪值的估计的方法，所述方法包括：

使用比较传感器来生成比较传感器信号；

生成空气传导麦克风信号；

使用所述比较传感器信号和所述空气传导麦克风信号来估计语音状态的似然性；以及

使用所述语音状态的似然性来估计所述经降噪值。

2、根据权利要求1所述的方法，其中使用所述语音状态的似然性来估计所述经降噪值包括，使用所述语音状态的似然性和所述经降噪值的期望值来估计所述经降噪值。

3、根据权利要求2所述的方法，其中所述经降噪值的期望值是基于所述比较传感器信号和所述空气传导麦克风信号。

4、根据权利要求1所述的方法，其中估计语音状态的似然性包括，为多个语音状态中的每一个估计单独的似然性，并且其中使用所述语音状态的似然性来估计所述经降噪值包括使用所述多个语音状态中的每个语音状态的似然性。

5、根据权利要求4所述的方法，其中使用所述多个语音状态中的每个语音状态的似然性包括，使用每个语音状态的似然性来对所述经降噪值的各期望值加权。

6、根据权利要求1所述的方法，其中估计语音状态的似然性包括，为一组频率分量中的每一个估计单独的似然性分量并且组合这些单独的似然性分量以形成所述语音状态似然性的估计。

7、根据权利要求1所述的方法，还包括使用语音状态的似然性的估计来确定

一帧所述空气传导麦克风信号是否包含语音。

8、根据权利要求7所述的方法，还包括使用被确定为不包含语音的一帧空气传导麦克风信号来确定噪声源的方差，以及使用该噪声源的方差来估计所述经降噪值。

9、根据权利要求1所述的方法，还包括建模在给定状态下经降噪值的概率来作为具有方差的分布。

10、根据权利要求9所述的方法，还包括估计所述分布的方差为前一帧的经降噪值的估计与当前帧的空气传导麦克风信号的经滤波形式的线性组合。

11、根据权利要求10所述的方法，其中所述空气传导麦克风信号的经滤波形式是通过使用频率相关的滤波器来形成的。

12、根据权利要求10所述的方法，其中所述空气传导麦克风信号的经滤波形式是通过信噪比相关的滤波器来形成的。

13、根据权利要求1所述的方法，还包括通过使用所述经降噪值的估计来形成所述经降噪值的新估计来执行迭代。

14、一种计算机可读介质，具有用于执行步骤的计算机可执行指令，所述步骤包括：

基于比较传感器信号和空气传导麦克风信号确定语音状态的似然性；以及使用所述语音状态的似然性来估计清晰语音值。

15、根据权利要求14所述的计算机可读介质，其中使用所述语音状态的似然性来估计清晰语音值包括，对期望值加权。

16、根据权利要求 14 所述的计算机可读介质，其中使用所述语音状态的似然性来估计清晰语音值包括：

使用所述语音状态的似然性来将一帧信号标识为非语音帧；

使用所述非语音帧来估计噪声的方差；以及

使用所述噪声方法来估计所述清晰语音值。

17、根据权利要求 14 所述的计算机可读介质，其中估计语音状态的似然性包括，为多个频率分量中的每一个估计单独的似然性，并且组合这些单独的似然性以形成所述语音状态的似然性。

18、一种为清晰语音信号标识出清晰语音值的方法，所述方法包括：

形成一模型，在所述模型中所述清晰语音信号取决于语音状态，比较传感器信号取决于所述清晰语音信号，并且空气传导麦克风信号取决于所述清晰语音信号；以及

基于所述模型、所述比较传感器信号的值、以及所述空气传导麦克风信号的值来确定所述清晰语音值的估计。

19、根据权利要求 18 所述的方法，还包括为状态确定似然性，并且其中确定所述清晰语音值的估计还包括使用所述状态的似然性。

20、根据权利要求 18 所述的方法，其中形成所述模型包括形成其中所述比较传感器信号和所述空气传导麦克风信号取决于噪声源的模型。

使用语音状态模型的多传感语音增强

背景技术

语音识别和语音传输中一个常见的问题是由加性噪声对语音信号造成的破坏。尤其是，由于另一扬声器的语音造成的破坏被证明是难于检测和/或纠正的。

最近，已开发出试图通过使用比较传感器——例如骨传导麦克风和空气传导麦克风——来移除噪声的系统。已开发出使用比较传感器信号和空气传导麦克风信号来形成比空气传导麦克风信号具有更少噪声的增强语音信号的各种技术。然而，完美地增强的语音尚未被实现，并且需要在增强的语音信号的形成上的进一步进展。

概要

一种方法和装置基于比较传感器信号和空气传导麦克风信号确定语音状态的似然性。语音状态的似然性被用于为清晰语音信号估计清晰语音值。

附图简要说明

图1是可在其中实践本发明的实施例的一个计算环境的框图。

图2是可在其中实践本发明的实施例的替换计算环境的框图。

图3是本发明的一般的语音处理系统的框图。

图4是在本发明一实施例下的用于增强语音的系统的框图。

图5是在本发明一实施例下的语音增强所基于的模型。

图6是在本发明一实施例下的用于增强语音的流程图。

示例性实施例的具体说明

图1示出了本发明的实施例可以在其上实现的合适的计算环境100的示例。计算环境100仅仅是合适的计算环境的一个例子，并不旨在限制本发明的使用范围或功能集。也不应将计算环境100解释为具有与示例性操作环境100中图解的任一

组件或组件组合的依赖或要求。

本发明的实施例可使用许多其他通用或专用计算系统环境或配置来运行。适合于本发明实施例的公知的计算系统、环境、和/或配置可包括但是不限于个人计算机、服务器计算机、手持或膝上设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络 PC、微型计算机、大型计算机、电话系统、包括任何上述系统或设备的分布式计算环境等。

本发明的实施例可以在由计算机执行的诸如程序模块的计算机可执行指令的一般化上下文中描述。一般来说，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。本发明被设计为在分布式计算环境中实施，在分布式计算环境中任务是由通过通信网络链接的远程处理设备来执行的。在分布式计算环境中，程序模块位于包括记忆存储设备的本地和远程计算机存储介质上。

参见图 1，用于实现本发明的示例系统包括计算机 110 形式的通用计算设备。计算机 110 的组件可以包括，但不限于，处理单元 120、系统存储器 130、和将包括系统存储器在内的各种系统组件耦合到处理单元 120 的系统总线 121。系统总线 121 可以是包括存储器总线或存储器控制器、外围总线和使用各种总线体系结构中的任何一种的局部总线在内的数种类型的总线结构中的任何一种。作为示例，而非限定，这样的体系结构包括工业标准体系结构 (ISA) 总线、微通道体系结构 (MCA) 总线、增强型工业标准结构 (EISA) 总线、视频电子标准协会 (VESA) 局部总线、和也称为 Mezzanine 总线的外围组件互连 (PCI) 总线。

计算机 110 典型地包括各种计算机可读介质。计算机可读介质可以是能被计算机 110 访问的任意可用介质，并且计算机可读介质包括易失性和非易失性介质、可移动和不可移动介质。作为示例，而非限定，计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以任何方法或技术实现的用于存储例如计算机可读指令、数据结构、程序模块或其他数据的信息的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括，但不限于，RAM、ROM、EEPROM、闪存或其他存储器技术，CD-ROM、数字视频光盘 (DVD) 或其他光盘存储，磁卡带、磁带、磁盘存储或其他磁性存储设备，或任何其他可被用于存储所需信息并可被计算机 110 访问的介质。通信介质典型地具体表现为计算机可读指令、数据结

构、程序模块或在例如是载波或其他传输机制的已调制的数据信号中的其他数据，并且包括任何信息传输介质。术语“已调制的数据信号”是指其特征中有一个或多个被以在该信号中编码信息的方式设置或改变的信号。作为示例，而非限定，通信介质包括例如是有线网络或直接有线连接的有线介质、以及例如是声音、RF、红外和其他无线介质的无线介质。任何上述的组合也应该包括在计算机可读介质的范围内。

系统存储器 130 包括例如是只读存储器(ROM)131 和随机存取存储器(RAM) 132 的易失性和/或非易失性存储器形式的计算机存储介质。包含在例如启动期间帮助在计算机 110 内各元件之间传输信息基本例程的基本输入/输出系统 131 (BIOS) 典型的存储于 ROM 131 中。RAM 132 典型地包含可由处理单元 120 立即存取和/或当前正被其操作的数据和/或程序模块。作为示例，而非限定，图 1 示出了操作系统 134、应用程序 135、其他程序模块 136 和程序数据 137。

计算机 110 还可以包括其他可移动/不可移动、易失性/非易失性计算机存储介质。仅作为示例，图 1 示出了读写不可移动、非易失性磁性介质的硬盘驱动器 141、读写可移动、非易失性磁盘 152 的磁盘驱动器 151、以及读写例如是 CD ROM 或其他光学介质的可移动、非易失性光盘的光盘驱动器 155。其他能用于示例操作环境的可移动/不可移动、易失性/非易失性计算机存储介质包括，但不限于，磁卡带、闪存卡、数字多功能盘、数字录像带、固态 RAM、固态 ROM 等。硬盘驱动器 141 典型地通过例如是接口 140 的非易失性存储器接口连接到系统总线 121，并且磁盘驱动器 151 和光盘驱动器 155 典型地通过例如是接口 150 的可移动存储器接口连接到系统总线 121。

上述以及在图 1 中示出的各驱动器和它们相关的计算机存储介质为计算机 110 提供计算机可读指令、数据结构、程序模块和其他数据的存储。例如，在图 1 中，硬盘驱动器 141 被图解为存储了操作系统 144、应用程序 145、其他程序模块 146、以及程序数据 147。注意到这些组件可以和操作系统 134、应用程序 135、其他程序模块 136 和程序数据 137 相同或不同。操作系统 144、应用程序 145、其他程序模块 146 和程序数据 147 在这里给予不同的附图标记以至少说明他们是不同的副本。

用户可以通过例如是键盘 162、麦克风 163 和定点设备 161 的输入设备将命令

和信息输入至计算机 110 中, 定点设备 161 例如是鼠标、轨迹球或触摸板。其他输入设备(未图示)可以包括操纵杆、游戏垫、卫星天碟、扫描仪等。这些及其他的输入设备通常通过耦接到系统总线的用户输入接口 160 连接到处理单元 120, 但也可以通过例如是并行端口、游戏端口或通用串行总线(USB)的其他接口和总线结构来连接。监视器 191 或其他类型的显示设备也经由例如是视频接口 190 的接口连接到系统总线 121。除了监视器, 计算机还可以包括例如是扬声器 197 和打印机 196 的其他的输出设备, 其可以通过输出外围接口 195 连接。

计算机 110 使用到例如是远程计算机 180 的一个或多个远程计算机的逻辑连接在连网环境中运行。远程计算机 180 可以是个人计算机、手持设备、服务器、路由器、网络 PC、对等设备或其他通用网络节点, 且典型地包括以上关于计算机 110 描述的许多或所有的要素。在图 1 中描述的逻辑连接包括局域网(LAN) 171 和广域网(WAN) 173, 但是也可以包括其他网络。这样的网络化环境在办公室、企业范围的计算机网络、内联网、以及因特网中是常见的。

当用在 LAN 网络化环境中时, 计算机 110 通过网络接口或适配器 170 连接到 LAN 171。当用在 WAN 网络化环境中时, 计算机 110 典型地包括调制解调器 172 或其他用于在例如是因特网的 WAN 173 上建立通信的手段。可以是内置的或外置的调制解调器 172 可以经由用户输入接口 160 或其他合适机构连接到系统总线 121。在连网环境中, 以上关于计算机 110 描述的模块或其部分可以存储在远程记忆存储装置中。作为示例, 而非限定, 图 1 示出了远程应用程序 185 驻留在远程计算机 180 上。应知晓所示的网络连接是示例性的, 而且可以使用在各计算机之间建立通信链路的其他手段。

图 2 是移动设备 200 的框图, 其是示例性的计算环境。移动设备 200 包括微处理器 202、存储器 204、输入/输出(I/O) 组件 206 和用于与远程计算机或其他移动设备通信的通信接口 208。在一实施例中, 前面提到的组件在合适总线 210 上耦合以便相互通信。

存储器 204 用电池备份模块(未图示)实现为诸如随机存取存储器等的非易失性电子存储器, 以使存储在存储器 204 中的信息在对移动设备 200 的总电源关闭时不会丢失。存储器 204 的一部分优选地被分配为用于程序执行的可寻址存储器, 而存储器 204 的另一部分优选地被用于存储, 例如模拟硬盘驱动器上的存储。

存储器 204 包括操作系统 212、应用程序 214 以及对象存储 216。在运行期间，操作系统 212 优选地由处理器 202 从存储器 204 执行。操作系统 212 在一优选实施例中是可从微软公司购得的 WINDOWS® CE 操作系统。操作系统 212 优选地为移动设备设计，并且实现数据库特征，该数据库特征能被应用程序 214 通过一组暴露的应用程序编程接口和方法来利用。在对象存储 216 中的对象由应用程序 214 和操作系统 212 至少部分地响应对暴露的应用程序编程接口和方法的调用来维护。

通信接口 208 代表多种允许移动设备 200 发送和接收信息的设备和技术。仅举数例，这些设备包括有线和无线调制解调器、卫星接收机和广播调谐器。移动设备 200 还可以直接连接到计算机以与其交换数据。在此类情形中，通信接口 208 可以是红外收发机或者串行或并行通信连接，所有的这些都能传输流信息。

输入/输出组件 206 包括例如是触敏屏、按钮、滚轮、和麦克风等的各种输入设备，以及包括音频发生器、振动设备和显示器等的输出设备。以上列出的设备是作为示例且不需要都出现在移动设备 200 上。此外，其他输入/输出设备可以被附连至移动设备 200 或在移动设备 200 中找到，这落在本发明的范围内。

图 3 提供了本发明的实施例的基本框图。在图 3 中，扬声器 300 生成语音信号 302 (X)，其被空气传导麦克风 304 和比较传感器 306 检测到。比较传感器的例子包括测量用户喉咙振动的喉部麦克风、定位于或邻近于用户的面部或颅骨（例如是颌骨）或在用户的耳朵中、且感测对应于由用户发出的语音的颅和颌的振动的骨传导传感器。空气传导麦克风 304 是通常用于将音频空气波转化成电信号的类型的麦克风。

空气传导麦克风 304 接收由一个或多个噪声源 310 产生的环境噪声 308 (V) 且产生它自身的传感器噪声 305 (U)。取决于环境噪声的类型和环境噪声的声级，环境噪声 308 也可以被比较传感器 306 检测到。然而，在本发明实施例之下，比较传感器 306 对环境噪声的灵敏度相比空气传导麦克风 304 来说较弱。因此，由比较传感器 306 生成的比较传感器信号 316 (B) 通常比由空气传导麦克风 304 生成的空气传导麦克风信号 318 (Y) 包括更少的噪声。尽管比较传感器 306 对环境噪声灵敏度较低，但是它的确会产生一些传感器噪声 320 (W)。

从扬声器 300 到比较传感器信号 316 的路径能被建模为具有信道响应 H 的信道。从环境噪声 308 到比较传感器信号 316 的路径能被建模为具有信道响应 G 的

信道。

比较传感器信号 316 (B) 和空气传导麦克风信号 318 (Y) 被提供至清晰信号估计器 322, 其估计清晰信号 324。清晰信号估计 324 被提供至语音处理器 328。清晰信号估计 324 或可以是时域信号, 或可以是傅立叶变换矢量。如果清晰信号估计 324 是时域信号, 则语音处理 328 就可以采取监听器、语音编码系统、或语音识别系统的形式。如果清晰信号估计 324 是傅立叶变换矢量, 则语音处理 328 典型地将是语音识别系统, 或包含傅立叶逆变换以将傅立叶变换矢量转换成波形。

在清晰信号估计器 322 内, 比较传感器信号 316 和麦克风信号 318 被转换成用于估计清晰语音的频域。如图 4 所示, 比较传感器信号 316 和空气传导麦克风信号 318 分别被提供至模数转换器 404 和 414 以生成一序列的数字值, 它们分别由帧构建器 406 和 416 编组成数值帧。在一实施例中, 模数转换器 404 和 414 在 16 kHz 和每采样 16 比特下采样模拟信号, 由此创建每秒 32K 字节的语音数据, 且帧构建器 406 和 416 每 10 毫秒创建各自新的一帧, 其包括了相当于 20 毫秒的数据。

由帧构建器 406 和 416 提供的相应各帧数据分别使用快速傅立叶变换 (FFT) 408 和 418 转换至频域。

对应于比较传感器信号和空气传导信号的频域值被提供给清晰信号估计器 420, 其使用这些频域值来估计清晰语音信号 324。

在一些实施例之下, 清晰语音信号 324 使用快速傅立叶逆变换 422 转换回时域。这创建了清晰语音信号 324 的时域形式。

本发明采用包括语音状态以便实现清晰语音的图 3 系统的模型, 以产生增强的语音信号。图 5 提供了该模型的图形表示。

在图 5 的模型中, 清晰语音 500 取决于语音状态 502。空气传导麦克风信号 504 取决于传感器噪声 506、环境噪声 508 和清晰语音信号 500。比较传感器信号 510 取决于传感器噪声 512、通过信号响应 514 的清晰语音信号 500 以及通过信道响应 516 的环境噪声 508。

图 5 的模型在本发明下用于从噪声观察 Y_t 和 B_t 估计清晰语音信号 X_t , 并标识多个语音状态 S_t 的似然性。

在本发明一实施例下, 通过首先在系统模型中对噪声分量假定高斯分布来形成清晰语音信号估计和清晰语音信号估计的状态的似然性。由此:

$$V \sim N(0, g^2 \sigma_v^2) \quad \text{式 1}$$

$$U \sim N(0, \sigma_u^2) \quad \text{式 2}$$

$$W \sim N(0, \sigma_w^2) \quad \text{式 3}$$

其中每个噪声分量被建模为一个各自具有方差 $g^2\sigma_v^2$ 、 σ_u^2 和 σ_w^2 的零均值高斯。 V 是环境噪声、 U 是空气传导麦克风中的传感器噪声，而 W 是在比较传感器中的传感器噪声。在式 1 中， g 是允许环境噪声的方差被调谐的调谐参数。

此外，本发明的这一实施例将给定一状态下清晰语音信号的似然性建模为具有方差 σ_s^2 的零均值高斯以使：

$$X | (S = s) \sim N(0, \sigma_s^2) \quad \text{式 4}$$

在本发明一实施例下，给定状态的先验的概率被假定为均匀概率以使所有的状态相等地似然。具体地，先验概率被定义为：

$$P(s_i) = \frac{1}{N_s} \quad \text{式 5}$$

其中 N_s 是模型中可用语音状态的数目。

在下面为确定清晰语音信号的估计以及语音状态的似然性而对等式的描述中，所有变量在复数频谱域中建模。每个频率分量（Bin）被独立于其他频率分量地来对待。为简化注记，该方法在下面针对单频率分量来描述。本领域技术人员将认识到，计算是对输入信号的频谱形式中的每个频率分量来执行的。对于随时间变化的变量，下标 t 被添加至该变量。

为了从有噪观察 Y_t 和 B_t 估计出清晰语音信号 X_t ，本发明使条件概率 $p(X_t | Y_t, B_t)$ 最大化，其是给定有噪空气传导麦克风信号和有噪比较传感器信号情况下清晰语音信号的概率。由于在该模型下清晰语音信号的估计取决于语音状态 S_t ，因此这个条件概率确定为：

$$p(X_t | Y_t, B_t) = \sum_{s \in \{S\}} p(X_t | Y_t, B_t, S_t = s) p(S_t = s | Y_t, B_t) \quad \text{式 6}$$

其中 $\{S\}$ 指示所有语音状态的集合， $p(X_t | Y_t, B_t, S_t = s)$ 是给定当前有噪观察和语音状态 s 情况下 X_t 的似然性，而 $p(S_t = s | Y_t, B_t)$ 是给定有噪观察情况下语音状态 s 的似然性。在本发明下任意数目的可能的语音状态可被使用，包括对应于浊音、摩擦音、鼻音和后元音的语音状态。在一些实施例中，为例如音素等的语音单元集合中的每一个提供一单独的语音状态。然而，在一个实施例下，只提供两个语音状态，一个

是给语音而一个是给非语音。

在一些实施例下，单状态被用于所有的频率分量。因此，每一帧具有单个语音状态变量。

在式 6 的右边项能被计算为：

$$p(X_t | Y_t, B_t, S_t = s) = \frac{p(X_t, Y_t, B_t, S_t = s)}{p(Y_t, B_t, S_t = s)} \propto p(X_t, Y_t, B_t, S_t = s) \quad \text{式 7}$$

$$p(S_t = s | Y_t, B_t) = \frac{\int p(X_t, Y_t, B_t, S_t = s) dX}{p(Y_t, B_t)} \propto \int p(X_t, Y_t, B_t, S_t = s) dX \quad \text{式 8}$$

其指示给定这些观察情况下清晰语音信号的条件概率能由清晰语音信号、观察和状态的联合概率来估计，并且指示了给定这些观察的情况下状态的条件概率能通过将清晰语音信号、观察以及状态的联合概率在所有可能的清晰语音值上积分来近似。

使用以上在式 1-3 中讨论的噪声分布的高斯假设，清晰语音信号、观察和状态的联合概率可被计算为：

$$p(X_t, S_t, Y_t, B_t) = N(Y_t; X_t, \sigma_u^2 + g^2 \sigma_v^2) p(X_t | S_t) p(S_t) \cdot N\left(G \frac{g^2 \sigma_v^2 (Y_t - X_t)}{\sigma_u^2 + g^2 \sigma_v^2}; B_t - H X_t, \sigma_w^2 + |G|^2 \frac{g^2 \sigma_v^2 \sigma_u^2}{\sigma_u^2 + g^2 \sigma_v^2}\right) \quad \text{式 9}$$

其中 $p(X_t | S_t = s) = N(X_t; 0, \sigma_s^2)$ ， $p(S_t)$ 是由式 5 中的均匀概率分布给定的状态的先验概率， G 是比较传感器环境噪声的信号响应， H 是该比较传感器信号对清晰语音信号的信道响应，而且在垂线标志之间的复数项例如 $|G|$ 指示该复数值的幅值。

比较传感器对背景语音的信号响应 G 是自空气麦克风 Y 的信号以及跨用户没有说话的最后 D 帧的比较传感器 B 的信号估计出的。具体地说， G 被确定为：

$$G = \frac{\sum_{t=1}^D (\sigma_u^2 |B_t|^2 - \sigma_w^2 |Y_t|^2) \pm \sqrt{(\sum_{t=1}^D (\sigma_u^2 |B_t|^2 - \sigma_w^2 |Y_t|^2))^2 + 4\sigma_u^2 \sigma_w^2 \sum_{t=1}^D B_t^* Y_t}}{2\sigma_u^2 \sum_{t=1}^D B_t^* Y_t} \quad \text{式 10}$$

在此 D 是其中用户没有说话但是存在背景语音的帧的数目。这里，我们假定 G 是跨所有时间帧 D 恒定的。在其他实施例中，不是平等地使用所有 D 个帧，我们代之以使用称作“指数式老化”的技术以使最近的帧相比较早前的帧对 G 的估计贡

献更多。

比较传感器的对清晰语音信号的信号响应 H 是自空气麦克风 Y 的信号和跨用户没有说话的最后 T 帧的比较传感器 B 的信号估计出的。具体地说, H 被确定为:

$$H = \frac{\sum_{t=1}^T (g^2 \sigma_v^2 |B_t|^2 - \sigma_w^2 |Y_t|^2) \pm \sqrt{(\sum_{t=1}^T (g^2 \sigma_v^2 |B_t|^2 - \sigma_w^2 |Y_t|^2))^2 + 4g^2 \sigma_v^2 \sigma_w^2 |\sum_{t=1}^T B_t^* Y_t|^2}}{2g^2 \sigma_v^2 \sum_{t=1}^T B_t^* Y_t}$$

式 11

在此 T 是其中用户在说话的帧的数目。这里, 我们假定 H 是跨所有时间帧 T 恒定的。在其他实施例中, 不是平等地使用所有 T 个帧, 我们代之以使用称作“指数式老化”的技术以使最近的帧相比较早前的帧对 H 的估计贡献更多。

状态 $p(S_t=s|Y_t, B_t)$ 的条件似然性使用式 8 的近似以及式 9 的联合概率计算来计算为:

$$p(S_t | Y_t, B_t) \propto \int_x N(Y_t; X_t, \sigma_u^2 + g^2 \sigma_v^2) \cdot N(G \frac{g^2 \sigma_v^2 (Y_t - X_t)}{\sigma_u^2 + g^2 \sigma_v^2}; B_t - HX_t, \sigma_w^2 + |G|^2 \frac{g^2 \sigma_v^2 \sigma_u^2}{\sigma_u^2 + g^2 \sigma_v^2}) \cdot p(X_t | S_t) p(S_t) dX$$

式 12

其可被简化为:

$$p(S_t | Y_t, B_t) \propto N(B_t; \frac{(\sigma_s^2 H + g^2 \sigma_v^2 G) Y_t}{\sigma_s^2 + g^2 \sigma_v^2 + \sigma_u^2}, \sigma_w^2 + |G|^2 \frac{g^2 \sigma_v^2 \sigma_u^2}{\sigma_u^2 + g^2 \sigma_v^2} + |H - G \frac{g^2 \sigma_v^2}{\sigma_u^2 + g^2 \sigma_v^2}|^2 \frac{\sigma_s^2 (\sigma_u^2 + g^2 \sigma_v^2)}{\sigma_s^2 + \sigma_u^2 + g^2 \sigma_v^2}) \cdot N(Y_t; 0, \sigma_s^2 + \sigma_u^2 + g^2 \sigma_v^2) p(S_t)$$

式 13

仔细观察式 13 显示在某些传感建模中第一项是比较传感器信道与空气传导麦克风信道之间的互相关, 而第二项利用状态模型和噪声模型来解释在空气麦克风信道中的观察。第三项仅仅是状态上的先验, 在一实施例下其是均匀分布。

如在式 13 中计算出的给定观察的情况下状态的似然性有两个可能的应用。首先, 它能被用于建立语音状态分类器, 其可用于将观察分类成包括语音或不包括语音以使噪声源的方差能自不包括语音的帧确定。它还能被用于在如下进一步示出地估计清晰语音信号时提供“软”权重。

如上所述，在上面各式中的每个变量是在复数频谱域中的特定频率分量定义的。因此，式 13 的似然性是对应于与特定频率分量相关联的状态。然而，由于每帧仅有单个状态变量，因此通过跨频率分量累计似然性来形成一帧的状态的似然性如下：

$$L(S_t) = \prod_f L(S_t(f)) \quad \text{式 14}$$

其中 $L(S_t(f)) = p(S_t(f) | Y_t(f), B_t(f))$ 式 13 中定义的频率分量 f 的似然性。乘积是在除了 DC 和 Nyquist 频率外的所有频率分量上确定的。注意到如果似然性计算是在对数-似然域中执行的，则在上式中的乘积用总和来替换。

上面的似然性能用于建立语音/非语音分类，基于似然性比率测试：

$$r = \log \frac{L(S_t = \text{语音})}{L(S_t = \text{非语音})} \quad \text{式 15}$$

其中帧在比值 r 大于 0 时被认为包含语音，否则被认为不包含语音。

通过使用语音状态的似然性，能形成对清晰语音信号的估计。在一实施例下，基于以上式 6 通过使用最小均方估计（MMSE）来形成这一估计：

$$\hat{X}_t = E(X_t | Y_t, B_t) = \sum_{s \in \{S\}} p(S_t = s | Y_t, B_t) E(X_t | Y_t, B_t, S_t = s) \quad \text{式 16}$$

其中 $E(X_t | Y_t, B_t)$ 是给定该观察情况下清晰语音信号的期望，而 $E(X_t | Y_t, B_t, S_t = s)$ 是给定该观察和该语音状态的情况下清晰语音信号的期望。

通过使用式 7 和 9，可据之计算期望 $E(X_t | Y_t, B_t, S_t = s)$ 的条件概率 $p(X_t | Y_t, B_t, S_t = s)$ 可被确定为：

$$p(X_t | Y_t, B_t, S_t = s) \propto \frac{N(Y_t; X_t, \sigma_u^2 + g^2 \sigma_v^2) \cdot N(\frac{g^2 \sigma_v^2 G(Y_t - X_t)}{\sigma_u^2 + g^2 \sigma_v^2}; B_t - HX_t, \sigma_w^2 + \frac{g^2 \sigma_v^2 \sigma_u^2 |G|^2}{\sigma_u^2 + g^2 \sigma_v^2}) \cdot N(X_t; 0, \sigma_s^2) p(S_t = s)}{\quad} \quad \text{式 17}$$

这产生期望：

$$E(X_t | Y_t, B_t, S_t = s) = \sigma_s^2 \left(\frac{\sigma_p^2 Y_t + M * ((\sigma_u^2 + g^2 \sigma_v^2) B_t - g^2 \sigma_v^2 G Y_t)}{\sigma_p^2 (\sigma_u^2 + g^2 \sigma_v^2 + \sigma_s^2) + |M|^2 \sigma_s^2 (\sigma_u^2 + g^2 \sigma_v^2)} \right) \quad \text{式 18}$$

其中

$$\sigma_p^2 = \sigma_w^2 + \frac{g^2 \sigma_v^2 \sigma_u^2}{\sigma_u^2 + g^2 \sigma_v^2} |G|^2 \quad \text{式 19}$$

$$M = H - \frac{g^2 \sigma_v^2}{\sigma_u^2 + g^2 \sigma_v^2} G \quad \text{式 20}$$

且 M^* 是 M 的复共轭。

因此，清晰语音信号 X_t 的 MMSE 估计由下式给出：

$$\hat{X}_t = \sum_{s \in \{S\}} \pi_s E(X_t | Y_t, B_t, S_t = s) \quad \text{式 21}$$

其中 π_s 在状态上的后验且其由下式给出：

$$\pi_s = \frac{L(S_t = s)}{\sum_{s \in \{S\}} L(S_t = s)} \quad \text{式 22}$$

其中 $L(S_t=s)$ 由式 14 给出。因此，清晰语音信号的估计是部分基于特定语音状态的相对似然性而且这一相对似然性为清晰语音信号的估计提供了软权重。

在上述的计算中， H 被假定为认为高精度已知。然而，在实践中， H 仅为有限精度已知。在本发明的另一实施例中， H 被建模为高斯随机变量 $N(H; H_0, \sigma_H^2)$ 。在这样的实施例下，上述所有的计算在 H 的所有可能值之上被边缘化。然而，这将使得在数学上难以处理。在一实施例下，使用迭代处理来克服这一难以处理之处。在每次迭代中， H 在式 13 和 20 中用 H_0 来代替，并且 σ_w^2 被用 $\sigma_w^2 + |\hat{X}_t|^2 \sigma_H^2$ 替换，其中 $\hat{X}_t$ 是自前次一迭代确定的清晰语音信号的估计。然后使用式 21 来估计清晰语音信号。然后，清晰语音信号的这一新的估计被设置为 $\hat{X}_t$ 的新值，接着执行下一次迭代。当清晰语音信号的估计变得稳定就结束该迭代。

图 6 提供了使用以上各式来估计清晰语音信号的方法。在步骤 600 中，标识出输入说话中用户没有说话的帧。然后，使用这些帧来确定环境噪声 σ_v^2 的方差、比较传感器噪声 σ_w^2 的方差以及空气传导麦克风噪声 σ_u^2 的方差。

为了标识出用户在其中没有说话的帧，可检查比较传感器信号。由于比较传感器信号将对背景语音产生比对噪声产生的小的多的信号值，因此当比较传感器信号的能量低时，它能初始假定为扬声器不在说话。对应于不包含语音的帧的空气传导麦克风信号和比较传感器信号的值被存储在缓冲中且被用于计算噪声方差如下：

$$\hat{\sigma}_v^2 = \frac{1}{N_v} \sum_{t \in I'} |Y_t|^2 \quad \text{式 23}$$

$$\hat{\sigma}_w^2 = \frac{1}{N_v} \sum_{\text{所有 } t \in V} |B'_t|^2 \quad \text{式 24}$$

其中 N_v 是在说话中被用于形成方差的噪声帧的数目， V 是用户在其中没有说话的噪声帧的集合， B'_t 指已经虑及漏泄之后的比较传感器信号，其计算为：

$$B'_t = B_t - GY_t \quad \text{式 25}$$

其在一些实施例中被替换地计算为：

$$B'_t = \left(1 - \frac{|GY_t|}{|B_t|}\right) B_t \quad \text{式 26}$$

在一些实施例下，基于低能量级在比较传感器信号中标识出非语音帧的技术仅仅在初始训练帧期间执行。在已形成噪声方差的初始值之后，它们可以被用于通过使用式 15 的似然比来确定哪些帧包含语音以及哪些帧不包含语音。

作为可用于增大或者减小估计的方差 σ_v^2 的调谐参数 g 的值在一特定实施例下被设置为 1。这暗示了在噪声估计过程中的完全置信度。在本发明的不同实施例下可以使用不同的 g 值。

空气传导麦克风的噪声方差 σ_u^2 是基于空气传导麦克风相比于比较传感器较不易于受传感器噪声影响的观察来估计的。如此，空气传导麦克风的方差能被计算为：

$$\sigma_u^2 = 1e^{-4} \sigma_w^2 \quad \text{式 27}$$

在步骤 602 中，使用具有时间平滑的噪声抑制滤波器来估计语音方差 σ_s^2 。该抑制滤波器是差谱的一般化。具体地，语音方差计算为：

$$\hat{\sigma}_s^2 = \tau |\hat{X}_{t-1}|^2 + (1-\tau) K_s^2 |Y_t|^2 \quad \text{式 28}$$

其中

$$K_s = \begin{cases} [1 - \alpha Q^{\gamma_1}]^2 & \text{如果 } Q^{\gamma_1} < 1/(\alpha + \beta) \\ [\beta Q^{\gamma_1}]^2 & \text{其他} \end{cases} \quad \text{式 29}$$

连同

$$Q = \frac{\sigma_v}{|Y_t|} \quad \text{式 30}$$

其中 $\hat{X}_{t-1}$ 是来自于前帧的清晰语音估计， τ 是平滑因子，在某些实施例下设置为 .2， α 控制噪声减少的程度以使如果 $\alpha > 1$ ，则以语音失真增大为代价降低更多的噪声，并且 β 给出最小噪声本底并提供增加背景噪声以掩饰察觉到的剩余音乐噪声的手段。在一些实施例下， $\gamma_1 = 2$ 且 $\gamma_2 = 1/2$ 。在一些实施例下，为对纯噪声帧，

将 β 设置为 0.01 以实现 20dB 的噪声减少。

因此，在式 28 中，方差被确定为在前帧的估计的清晰语音信号和经噪声抑制滤波器 K_s 滤波的空气传导麦克风信号的能量加权总和。

在一些实施例下， α 是根据信噪比和屏蔽原理来选择的，该屏蔽原理表明在高语音能量带中的相同量的噪声比在低语音能量带中具有更小的影响，以及在一个频率上高语音能量的存在将减少在邻近频带中的噪声的感知度。在这一实施例下， α 被选择为：

$$\alpha = \begin{cases} \alpha_0(1 - SNR/B) & \text{如果 } SNR < B \\ 0 & \text{其他} \end{cases} \quad \text{式 31}$$

其中 SNR 是以分贝计的信噪比，B 是所需的信噪比等级，在该等级之上不应执行噪声减少，并且 α_0 是在信噪比值为 0 时应该被移除的噪声量。在一些实施例下，B 被设置成等于 20 dB。

通过使用信噪比的定义：

$$SNR = 10 \log\left(\frac{|Y_t|^2 - \sigma_v^2}{\sigma_v^2}\right) \quad \text{式 32}$$

式 29 的噪声抑制滤波器变成：

$$K_s = \begin{cases} [1 - \alpha_0(1 - SNR/B)/(1 + 10^{SNR/10})]^{1/2} & \text{如果 } Q^2 < 1/\alpha + \beta \\ [\beta Q^2]^{1/2} & \text{其他} \end{cases} \quad \text{式 33}$$

这一噪声抑制滤波器为正的信噪比提供弱噪声抑制，而为负的信噪比提供较强的噪声抑制。事实上，对于足够负的信噪比，所有观察到的信号和噪声都将被移除而唯一存在的信号是噪声本底，其由式 33 的噪声抑制滤波器的“其他”分支加回。

在一些实施例下，使 α_0 频率相关以使对不同频率移除不同量的噪声。在一实施例下，通过在 30 Hz 下的 α_0 和 8KHz 下的 α_0 之间使用线性内插来形成这一频率相关：

$$\alpha_0(k) = \alpha_{0\min} + (\alpha_{0\max} - \alpha_{0\min})k/225 \quad \text{式 34}$$

其中 k 是频率分量的计数， $\alpha_{0\min}$ 是在 30Hz 下期望的 α_0 值， $\alpha_{0\max}$ 是在 8KHz 下期望的 α_0 值，并且假定存在 256 个频率分量。

在步骤 602 中确定语音方差后，在步骤 604 中通过使用上述式 13 和 14，利用方差来确定每个语音状态的似然性。然后在步骤 606 中利用语音状态的似然性为当

前帧确定清晰语音估计。如上所述，在高斯分布被用于代表 H 的实施例中，通过在每一次迭代中使用清晰语音信号的最近估计以及通过使用上述各式的变化来容纳 H 的高斯模型，从而来迭代步骤 604 和 606。

尽管本发明已经参考具体实施例描述，但本领域技术人员会认识到在形式和细节上可作出改变而不脱离本发明的精神和范围。

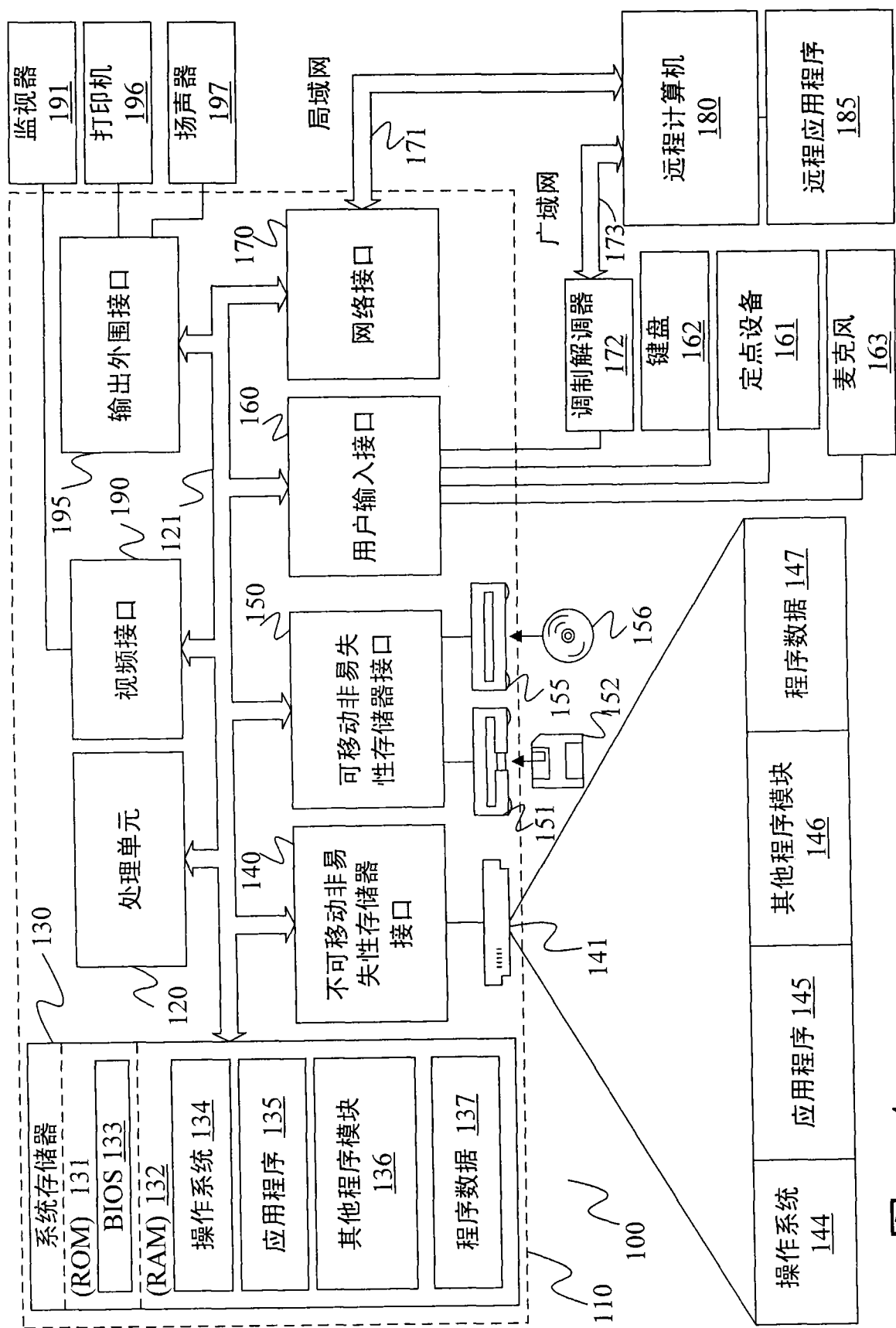


图 1

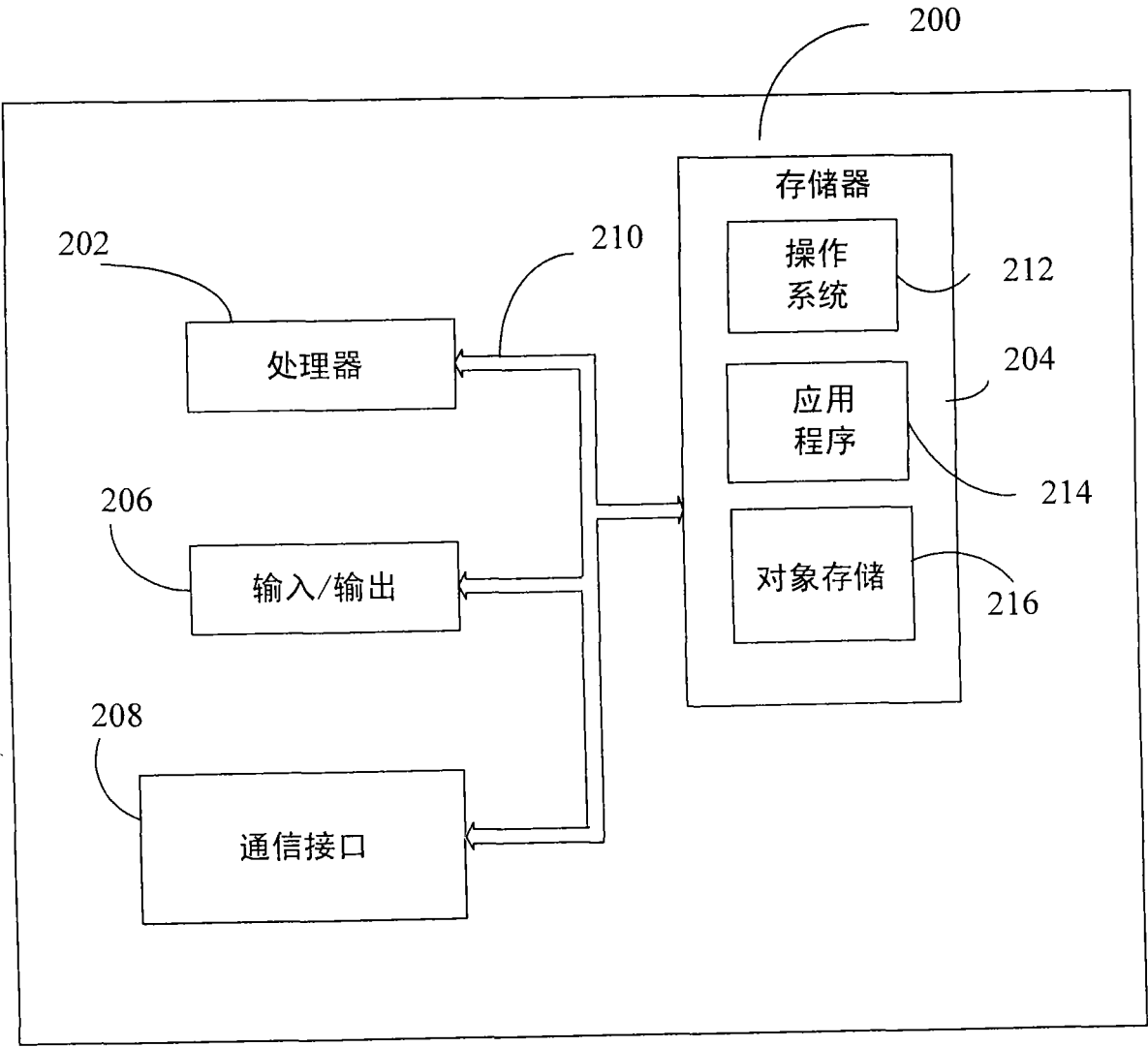


图 2

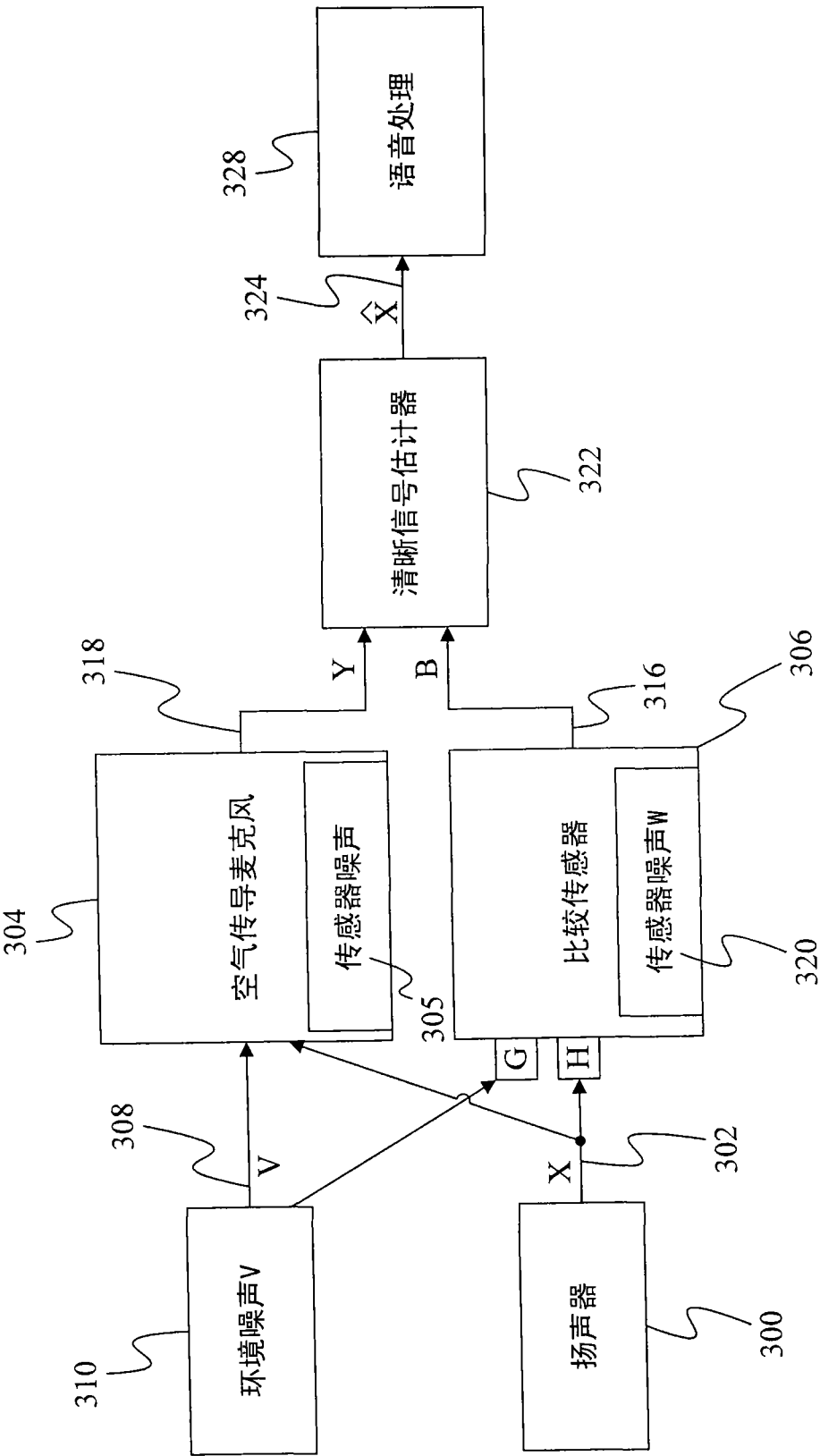


图 3

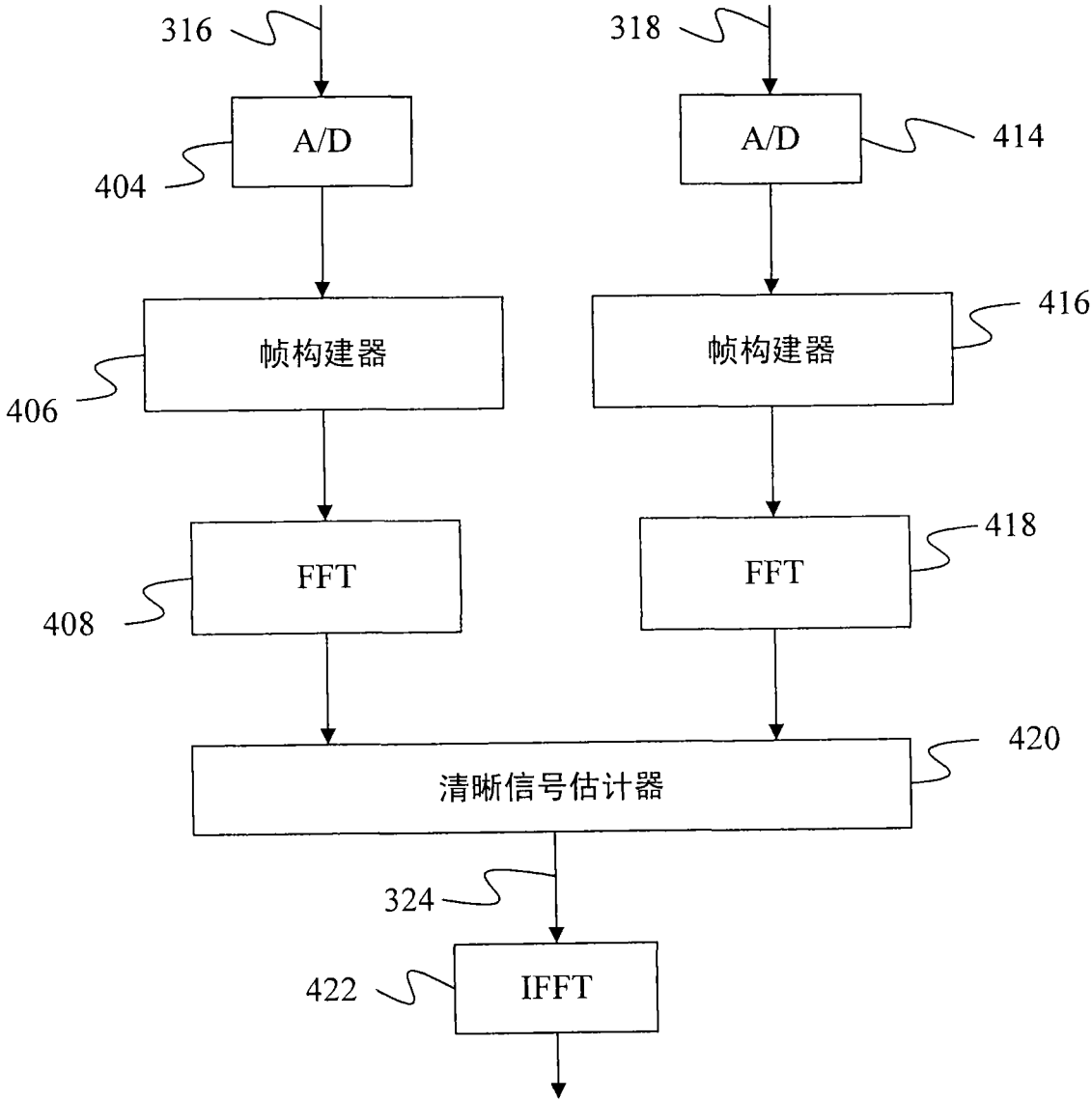


图 4

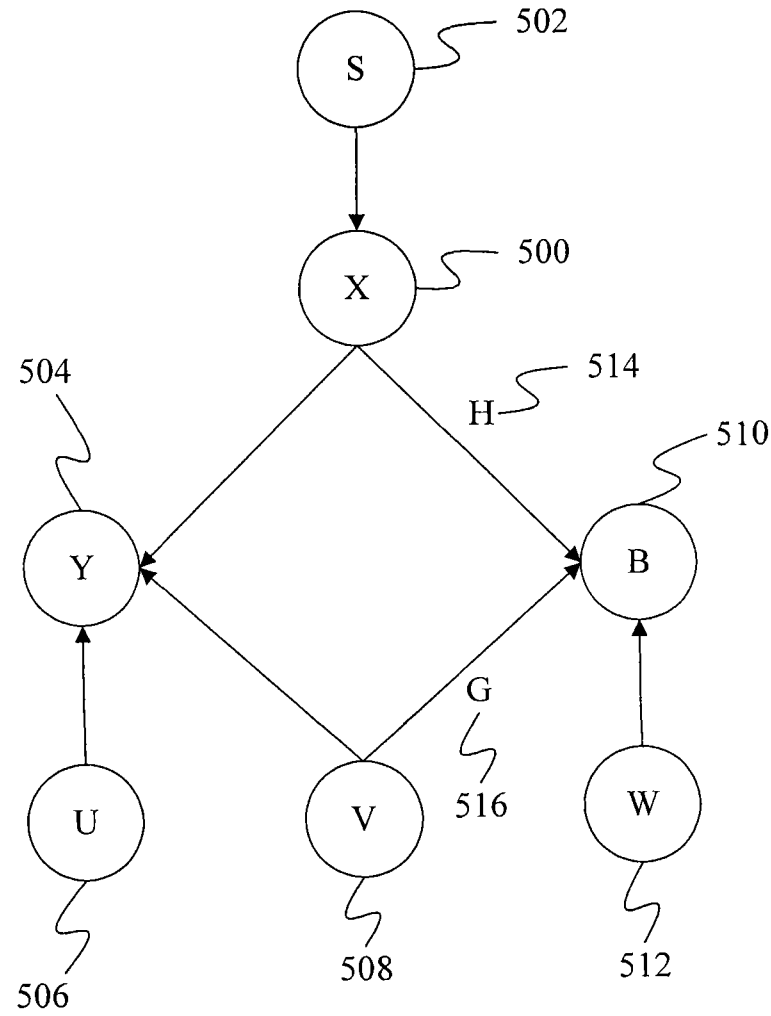
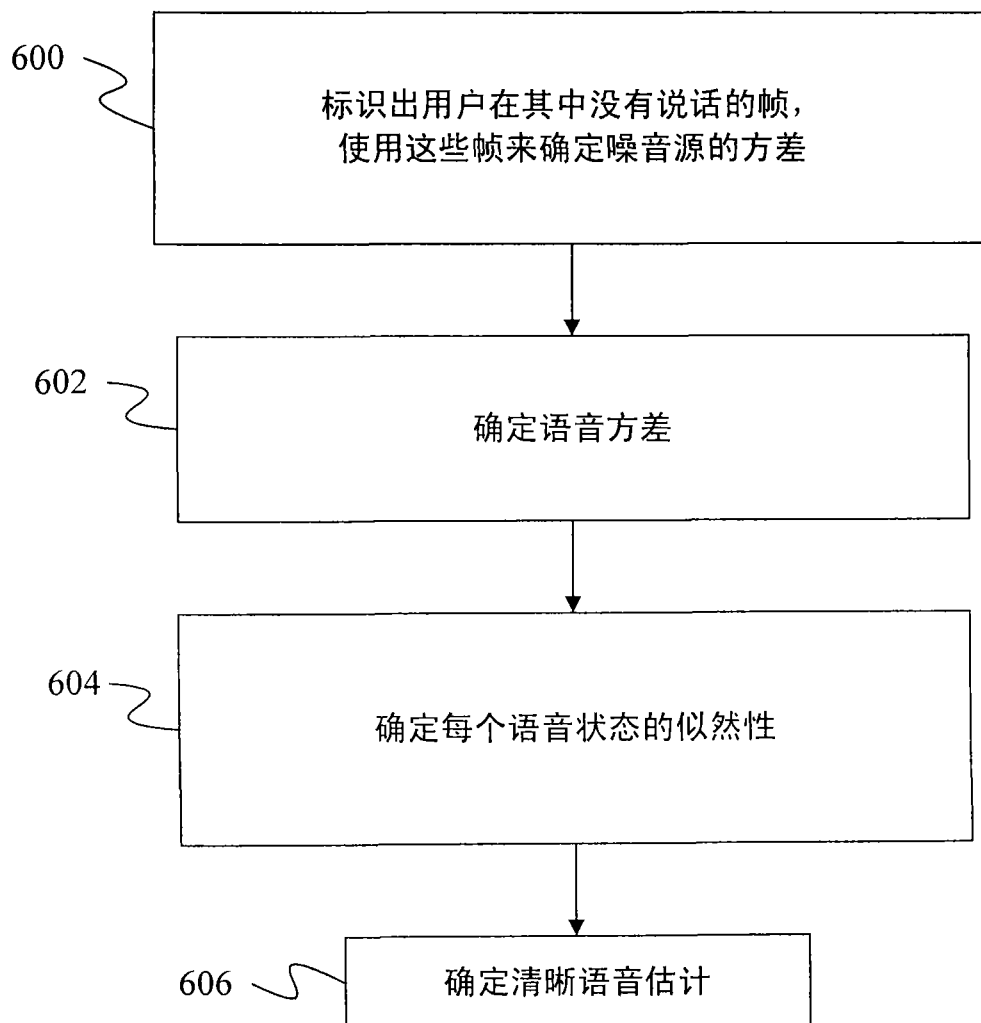


图 5

**图 6**