[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G10L 15/02
G10L 15/00
G10L 11/02

[12] 发明专利申请公开说明书

[21] 申请号 200410095656.2

[43] 公开日 2005 年 6 月 8 日

[11] 公开号 CN 1624765A

[22] 申请日 2004.11.26
[21] 申请号 200410095656.2
[30] 优先权
[32] 2003.11.26 [33] US [31] 10/723,995
[71] 申请人 微软公司
地址 美国华盛顿州
[72] 发明人 A·阿塞罗 H·阿蒂亚斯
L·J·李 邓 立

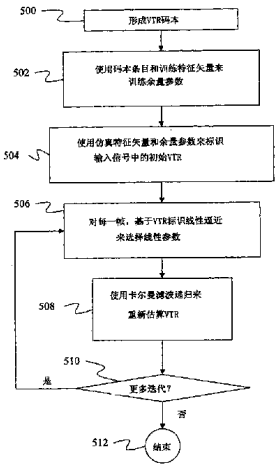
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 谢喜堂

权利要求书 3 页 说明书 12 页 附图 6 页

[54] 发明名称 使用分段线性逼近的连续值声道共振跟踪方法和装置

[57] 摘要

一种方法和装置跟踪语音信号中的共振分量，包括频率和带宽。 通过定义对过去的声道共振矢量线性、且预测当前声道共振矢量的状态方程式来跟踪这些分量。 也定义对当前声道共振矢量为线性的、且预测观测矢量的至少一个分量的观测方程式。 状态方程式、观测方程式和观测矢量序列用于使用卡尔曼滤波器算法来标识声道共振矢量序列。在一个实施例中，基于对非线性函数的分段线性逼近来定义观测方程式。 基于预定义的区域来选择线性逼近的参数，这些区域根据声道共振矢量的粗略估算来确定。



ISSN1008-4274

1．一种跟踪语音信号中的声道共振频率的方法，其特征在于，它包括：

定义对过去的声道共振矢量为线性的、且预测当前声道共振矢量的一状态方程式；

定义对当前声道共振矢量为线性的、且预测观测矢量的至少一个分量的一观测方程式；以及

使用所述状态方程式、所述观测方程式和所述观测矢量序列来标识一声道共振矢量序列，每一声道共振矢量包括至少一个声道共振频率。

2．如权利要求 1 所述的方法，其特征在于，使用所述状态方程式、所述观测方程式和所述观测矢量序列来标识声道共振矢量序列包括向一卡尔曼滤波器应用所述状态方程式、所述观测方程式和所述观测矢量序列。

3．如权利要求 1 所述的方法，其特征在于，标识声道共振矢量包括根据一组连续值标识声道共振矢量。

4．如权利要求 1 所述的方法，其特征在于，定义所述观测方程式包括定义对所述声道共振矢量非线性的函数的线性逼近。

5．如权利要求 4 所述的方法，其特征在于，定义所述观测方程式还包括定义对两个函数的乘积的线性逼近，该两个函数的每一个对所述声道共振矢量都为非线性。

6．如权利要求 5 所述的方法，其特征在于，对所述声道共振矢量非线性的所述函数的其中之一是一个对所述声道共振矢量的带宽分量非线性的指数函数。

7．如权利要求 5 所述的方法，其特征在于，对所述声道共振矢量非线性的所述函数的其中之一是对所述声道共振矢量的频率分量非线性的正弦函数。

8．如权利要求 4 所述的方法，其特征在于，定义线性逼近包括从共同形成对所述非线性函数的分段线性逼近的一组线性逼近中选择一线性逼近。

9．如权利要求 4 所述的方法，其特征在于，定义线性逼近包括基于声道共振矢量的估算来计算所述非线性函数的值以生成一非线性函数值，并使用所述非线性函数值来选择所述线性逼近的参数。

10．如权利要求 9 所述的方法，其特征在于，定义线性逼近还包括使用所述非线性函数值以从共同形成对所述非线性函数的分段线性逼近的一组线性逼近中

选择一线性逼近。

11．如权利要求 1 所述的方法，其特征在于，它还包括：

使用所标识的声道共振矢量来重定义所述观测方程式；以及

使用所述重定义的观测方程式、所述状态方程式和所述观测矢量来标识一声道共振矢量的新序列。

12．如权利要求 11 所述的方法，其特征在于，重定义所述观测方程式包括使用一已标识的声道共振矢量来选择对声道共振矢量非线性的函数的至少一个线性逼近的参数。

13．如权利要求 12 所述的方法，其特征在于，使用已标识的声道共振矢量来选择参数包括使用所述声道共振矢量来计算所述非线性函数的值以生成一非线性函数值、及使用所述非线性函数值来选择至少一个线性逼近的参数。

14．一种具有计算机可执行指令的计算机可读媒质，其特征在于，所述指令执行以下步骤：

使用至少一个声道共振分量的估算来选择对所述声道共振分量非线性的函数的线性逼近；

使用所述线性逼近来定义一观测方程式；以及

使用所述观测方程式和至少一个观测矢量来重新估算所述声道共振分量。

15．如权利要求 14 所述的计算机可读媒质，其特征在于，选择线性逼近包括从形成所述非线性函数的分段线性逼近的一组线性逼近中选择一线性逼近。

16．如权利要求 14 所述的计算机可读媒质，其特征在于，选择线性逼近包括向所述非线性函数应用所述声道共振分量以形成一函数值、及基于所述函数值选择所述线性逼近。

17．如权利要求 14 所述的计算机可读媒质，其特征在于，重新估算所述声道共振分量的值还包括使用对所述声道共振分量线性的一状态方程式。

18．如权利要求 17 所述的计算机可读媒质，其特征在于，重新估算所述声道共振分量的值还包括向一卡尔曼滤波器应用所述状态方程式、所述观测方程式和所述至少一个观测矢量。

19．如权利要求 14 所述的计算机可读媒质，其特征在于，它还包括选择对所述声道共振分量非线性的第二函数的第二线性逼近、及使用所述第二线性逼近来定义所述观测方程式。

20．如权利要求 14 所述的计算机可读媒质，其特征在于，所述非线性函数包括一指数函数。

21．如权利要求 14 所述的计算机可读媒质，其特征在于，所述非线性函数包括一正弦函数。

22．如权利要求 14 所述的计算机可读媒质，其特征在于，所述声道共振分量是连续值。

使用分段线性逼近的连续值声道共振跟踪方法和装置

技术领域

本发明涉及语音识别系统，尤其涉及利用语音中的声道共振的语音识别系统。

背景技术

在人类语音中，大量的信息包含在语音信号的前三个或前四个共振频率内。特别地，当说话者发出元音时，这些共振的频率（对较小的范围，为带宽）指示正在说出哪一元音。

这一共振频率和带宽通常被总称为共振峰（formant）。在通常为有声的响音语音中，可发现共振峰为语音的频率表示中的谱突起。然而，在非响音语音中，不能直接找到共振峰为谱突起。为此，术语“共振峰”有时被解释为仅应用于语音的响音部分。为避免混淆，某些研究人员使用词组“声道共振”来指出现在响音和非响应语音中的共振峰。在两种情况下，共振仅指声道共振的口腔道部分。

为检测共振峰，现有技术的系统分析语音信号帧的频谱内容。由于共振峰可以是任何频率，因此现有技术试图在标识最可能的共振峰值之前限制搜索空间。在某些现有技术系统中，可能的共振峰的搜索空间通过标识帧的频谱内容中的峰值来减小。通常，这通过使用线性预测编码（LPC）来完成，LPC 试图找出表示语音信号帧的频谱内容的多项式。该多项式的每一根值表示信号中的一个可能的共振频率，并由此表示可能的公正共振峰。由此，使用 LPC，搜索空间被减小至形成 LPC 多项式根的那些频率。

在现有技术的其它共振峰跟踪系统中，通过将帧的频谱内容与一组在其中由专家标识了共振峰的频谱模板进行比较来减小搜索空间。然后选择最接近的“n”个模板，并将它们用于计算该帧的共振峰。由此，这些系统将搜索空间减小至与最接近的模板相关联的那些共振峰。

由本发明的相同的发明人开发的现有技术的一种系统使用了对输入信号的每一帧都相同的一致搜索空间。搜索空间中的每一组共振峰被映射到一特征矢量。每一特征矢量然后被应用到一模型以确定哪一组共振峰是最可能的。

该系统能够较好地工作，然而它需要很大的计算量，因为它通常使用梅尔频率（Mel-Frenquency）倒谱系数频率矢量，这需要将一组频率应用到基于要映射的共振峰组中的所有共振峰的复杂滤波器，随后执行加窗步骤和离散余弦变换步骤，以将共振峰映射到特征矢量。这一计算在运行时执行太耗时，由此所有共振峰组都必须在运行之前映射，并且映射的特征矢量必须被储存在一个大表中。这并不理想，因为它需要充足的存储器来储存所有映射的特征矢量。

在由本发明的发明人开发的另一系统中，一组离散声道共振矢量被储存在码本中。每一离散矢量被转化成一仿真特征矢量，将该仿真特征矢量与输入特征矢量相比较，以确定哪一离散矢量能最好地表示输入语音信号。该系统并不理想，因为它不确定声道共振矢量的连续值，而是选择离散的声道共振码字的其中之一。

发明内容

一种方法和装置跟踪语音信号中的声道共振分量。通过定义对过去的声道共振矢量为线性、且预测当前的声道共振矢量的状态方程式来跟踪该分量。也定义对当前声道共振矢量为线性、且预测观测矢量的至少一个分量的观测方程式。状态方程式、观测方程式和一系列观测矢量用于标识一系列声道共振矢量。在一个实施例中，基于对非线性函数的线性逼近来定义观测方程式。基于声道共振矢量的估算来选择该线性逼近的参数。

附图说明

图 1 是可在其中实践本发明的实施例的通用计算环境的框图。

图 2 是语音信号的幅度频谱曲线图。

图 3 所示是对指数函数的分段线性逼近的曲线图。

图 4 所示是对正弦函数的分段线性逼近的曲线图。

图 5 是本发明的方法的流程图。

图 6 是用于训练余量模型的训练系统的框图。

图 7 是本发明的一个实施例中共振峰跟踪系统的框图。

具体实施方式

图 1 示出了适合在其中实现本发明的计算系统环境 100 的一个示例。计算系

统环境 100 仅为合适的计算环境的一个示例，并非暗示对本发明的使用范围或功能的局限。也不应将计算环境 100 解释为对示例性操作环境 100 中示出的任一组件或其组合具有依赖或需求。

本发明可以使用众多其它通用或专用计算系统环境或配置来操作。适合使用本发明的众所周知的计算系统、环境和/或配置的示例包括但不限于：个人计算机、服务器计算机、手持式或膝上设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费者电子设备、网络 PC、小型机、大型机、电话系统、包括任一上述系统或设备的分布式计算环境等等。

本发明可在诸如由计算机执行的程序模块等计算机可执行指令的一般上下文环境中描述。一般而言，程序模块包括例程、程序、对象、组件、数据结构等等，执行特定的任务或实现特定的抽象数据类型。本发明被设计成在分布式计算环境中实践，其中，任务由通过通信网络连接的远程处理设备来执行。在分布式计算环境中，程序模块可以位于本地和远程计算机存储媒质中，包括存储器存储设备。

参考图 1，用于实现本发明的示例系统包括以计算机 110 形式的通用计算装置。计算机 110 的组件可包括但不限于，处理单元 120、系统存储器 130 以及将包括系统存储器的各类系统组件耦合至处理单元 120 的系统总线 121。系统总线 121 可以是若干种总线结构类型的任一种，包括存储器总线或存储器控制器、外围总线以及使用各类总线体系结构的局部总线。作为示例而非局限，这类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 ISA（EISA）总线、视频电子技术标准协会（VESA）局部总线以及外围部件互连（PCI）总线，也称为 Mezzanine 总线。

计算机 110 通常包括各种计算机可读媒质。计算机可读媒质可以是可由计算机 110 访问的任一可用媒质，包括易失和非易失媒质、可移动和不可移动媒质。作为示例而非局限，计算机可读媒质包括计算机存储媒质和通信媒质。计算机存储媒质包括以用于储存诸如计算机可读指令、数据结构、程序模块或其它数据等信息的任一方法或技术实现的易失和非易失，可移动和不可移动媒质。计算机存储媒质包括但不限于，RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字多功能盘（DVD）或其它光盘存储、磁盒、磁带、磁盘存储或其它磁存储设备、或可以用来储存所期望的信息并可由计算机 110 访问的任一其它媒质。通信媒质通常在诸如载波或其它传输机制的已调制数据信号中包含计算机可读指令、数据结

构、程序模块或其它数据，并包括任一信息传送媒质。术语“已调制数据信号”指以对信号中的信息进行编码的方式设置或改变其一个或多个特征的信号。作为示例而非局限，通信媒质包括有线媒质，如有线网络或直接连线连接，以及无线媒质，如声学、RF、红外和其它无线媒质。上述任一的组合也应当包括在计算机可读媒质的范围之内。

系统存储器 130 包括以易失和/或非易失存储器形式的计算机存储媒质，如只读存储器（ROM）131 和随机存取存储器（RAM）132。基本输入/输出系统 133（BIOS）包括如在启动时帮助在计算机 110 内的元件之间传输信息的基本例程，通常储存在 ROM 131 中。RAM 132 通常包含处理单元 120 立即可访问或者当前正在操作的数据和/或程序模块。作为示例而非局限，图 1 示出了操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137。

计算机 110 也可包括其它可移动/不可移动、易失/非易失计算机存储媒质。仅作示例，图 1 示出了对不可移动、非易失磁媒质进行读写的硬盘驱动器 141、对可移动、非易失磁盘 152 进行读写的磁盘驱动器 151 以及对可移动、非易失光盘 156，如 CD ROM 或其它光媒质进行读写的光盘驱动器 155。可以在示例性操作环境中使用的其它可移动/不可移动、易失/非易失计算机存储媒质包括但不限于，磁带盒、闪存卡、数字多功能盘、数字视频带、固态 RAM、固态 ROM 等等。硬盘驱动器 141 通常通过不可移动存储器接口，如接口 140 连接到系统总线 121，磁盘驱动器 151 和光盘驱动器 155 通常通过可移动存储器接口，如接口 150 连接到系统总线 121。

图 1 讨论并示出的驱动器及其关联的计算机存储媒质为计算机 110 提供了计算机可读指令、数据结构、程序模块和其它数据的存储。例如，在图 1 中，示出硬盘驱动器 141 储存操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147。注意，这些组件可以与操作系统 134、应用程序 135、其它程序模块 136 和程序数据 137 相同，也可以与它们不同。这里对操作系统 144、应用程序 145、其它程序模块 146 和程序数据 147 给予不同的标号来说明至少它们是不同的副本。

用户可以通过输入设备，如键盘 162、麦克风 163 和定位设备 161（如鼠标、跟踪球或触摸板）向计算机 110 输入命令和信息。其它输入设备（未示出）可包括操纵杆、游戏垫、圆盘式卫星天线、扫描仪等等。这些和其它输入设备通常通过耦合至系统总线的用户输入接口 160 连接至处理单元 120，但是也可以通过其它接口

和总线结构连接，如并行端口、游戏端口或通用串行总线（USB）。监视器 191 或其它类型的显示设备也通过接口，如视频接口 190 连接至系统总线 121。除监视器之外，计算机也可包括其它外围输出设备，如扬声器 197 和打印机 196，通过输出外围接口 195 连接。

计算机 110 可以在使用到一个或多个远程计算机，如远程计算机 180 的逻辑连接的网络化环境中操作。远程计算机 180 可以是个人计算机、手持式设备、服务器、路由器、网络 PC、对等设备或其它公用网络节点，并通常包括许多或所有上述与计算机 110 相关的元件。图 1 描述的逻辑连接包括局域网（LAN）171 和广域网（WAN）173，但也可包括其它网络。这类网络环境常见于办公室、企业范围计算机网络、内联网以及因特网。

当在 LAN 网络环境中使用时，计算机 10 通过网络接口或适配器 170 连接至 LAN 171。当在 WAN 网络环境中使用时，计算机 110 通常包括调制解调器 172 或其它装置，用于通过 WAN 173，如因特网建立通信。调制解调器 172 可以是内置或外置的，通过用户输入接口 160 或其它合适的机制连接至系统总线 121。在网络化环境中，描述的与计算机 110 相关的程序模块或其部分可储存在远程存储器存储设备中。作为示例而非局限，图 1 示出远程应用程序 185 驻留在远程计算机 180 上。可以理解，示出的网络连接是示例性的，也可以使用在计算机之间建立通信链路的其它装置。

图 2 是人类语音的一个片段的频谱曲线图。在图 2 中，频率沿水平轴 200 示出，频率分量的幅度沿垂直轴 202 示出。图 2 的曲线图示出了响音人类语音包含的共振或共振峰，如第一共振峰 204、第二共振峰 206、第三共振峰 208 和第四共振峰 210。每一共振峰由其中心频率 F 与其带宽 B 描述。

本发明提供了在响音和非响音语音中，跨共振峰频率和带宽的连续范围标识语音信号中的共振峰频率和带宽的方法。由此，本发明能够跟踪声道共振频率和带宽。

为完成这一过程，本发明将隐含的声道共振频率和带宽模型化为一列隐含的状态，其每一个都产生一观测。在一个具体的实施例中，隐含的声道共振频率和带宽使用以下状态方程式 1 和观测方程式 2 来模型化：

$$x_t = \Phi x_{t-1} + (I - \Phi)T + w_t \qquad \text{公式 1}$$

$$o_t = C(x_t) + v_t \qquad \text{公式 2}$$

其中，x_t是 t 时刻的隐含声道共振矢量，它由$x_t=\{f_1,b_1,f_2,b_2,f_3,b_3,f_4,b_4\}$构成，$x_{t-1}$是前一时刻 t-1 的隐含声道共振矢量，Φ是系统矩阵，I 是单位矩阵，T 是声道共振频率和带宽的目标矢量，w_t是状态方程式中的噪声，o_t是已观测矢量，$C(x_t)$是从隐含声道共振矢量到观测矢量的映射方程，v_t是观测中的噪声。在一个实施例中，Φ是对角矩阵，其每一元素具有根据经验所确定的 0.7 和 0.9 之间的值，T 是矢量，在一个实施例中，它的值为：

$(500\ 1500\ 2500\ 3500\ 200\ 300\ 400\ 400)^T$

在本实施例中，噪声参数 w_t和 v_t的值由具有零平均值矢量和对角协方差矩阵的随机高斯样值来确定。本实施例中，这些矩阵的对角元素的值对 w_t在 10 和 30,000 之间，对 v_t在 0.8 和 78 之间。

在一个实施例中，已观测的矢量是线性预测编码倒谱（LPC 倒谱）矢量，该矢量的每一分量表示一 LPC 阶。结果，可由解析非线性函数来精确地确定映射函数 $C(x_t)$。帧 t 的矢量值函数 $C(x_t)$的第 n 个分量为：

$$C_n(x_t)=\sum_{k=1}^{K}\frac{2}{n}e^{-\pi n\frac{b_k(t)}{f_s}}\cos(2\pi n\frac{f_k(t)}{f_s}) \qquad \text{公式 3}$$

其中，$C_n(x_t)$是第N阶LPC倒谱特征矢量中的第n个元素，K是声道共振（VTR）频率的数量，$f_k(t)$是帧 t 的第 k 个 VTR 频率，$b_k(t)$是帧 t 的第 k 个 VTR 带宽，f_s是采样频率，在许多实施例中为 8kHz，在其它实施例中为 16kHz。C_0元素被设为等于 log$G$，其中 G 是增益。

为从一列观测矢量标识一列隐含声道共振矢量，本发明使用卡尔曼（Kalman）滤波器。卡尔曼滤波器提供了一种递归技术，它可确定由公式 1 和 2 表示的线性动态系统中的连续值隐含声道共振矢量的最佳估算。这一卡尔曼滤波器在本领域中是众所周知的。

卡尔曼滤波器需要公式 1 和 2 的右侧对隐含声道共振矢量为线性。然而，公式 3 的映射函数对声道共振矢量是非线性的。为解决该问题，本发明使用了分段线性逼近来替代公式 3 中的指数和余弦项。在一个实施例中，指数项由 5 个线性段来表示，余弦项由 10 个线性段来表示。

图 3 示出了对公式 3 中的指数项的分段线性逼近。指数的值沿垂直轴 300 示出，第 k 个 VTR 带宽的带宽 b_k 的值沿水平轴 302 示出。在图 3 中，使用 5 个线段 304、306、308、310 和 312 来近似指数曲线 314。下表提供了每一线段所覆盖的指数值的范围。

线段	指数值的范围
304	0-100 Hz
306	100-200 Hz
308	200-300 Hz
310	300-400 Hz
312	400-500 Hz

表 1

图 4 示出了对公式 3 中的余弦项的分段线性逼近的示例。余弦函数的值沿垂直轴 400 示出，第 k 个 VTR 频率的频率 f_k 的值沿水平轴 402 示出。在图 4 中，示出了余弦函数的单个周期，然而，本领域的技术人员将认识到，可对余弦函数的每一周期使用同一分段线性逼近。在图 4 的实施例中，余弦函数 424 由 10 个线段 404、406、408、410、412、414、416、418、420 和 422 来近似。下表 2 提供了由每一线段覆盖的余弦值的不均匀范围，假定完整的周期覆盖了从 0 Hz 到 8000 Hz 的频率范围。

线段	余弦值范围
404	0-500 Hz
406	500-1000 Hz
408	1000-3000 Hz
410	3000-3500 Hz
412	3500-4000 Hz
414	4000-4500 Hz
416	4500-5000 Hz
418	5000-7000 Hz
420	7000-7500 Hz
422	7500-8000 Hz

表 2

使用这些线性逼近，公式 3 可重写为：

$$C_n(x_t)=\sum_{k=1}^{K}\frac{2}{n}(\alpha_{kx}x_t+\beta_{kx})(\gamma_{kx}x_t+\delta_{kx}) \qquad \text{公式 4}$$

其中，α_{kx} 是近似指数项的线段的斜率，β_{kx} 是其截矩，γ_{kx} 是近似余弦项的线段的斜率，δ_{kx} 是其截矩。注意，这四项都依赖于 x_t，因为用于近似非线性函数的线段

是基于由依照表 1 和 2 的 x_t 的值来确定的区域上选择的。

公式 4 中的映射函数的形式在 x_t 中仍非线性，这是由于二次项的存在。在本发明的一个实施例中，忽略该项的递增部分，由此获得从 x_t 到 $C_n(x_t)$ 的线性方程式。

在该形式中，只要参数基于表 1 和 2 中例示的范围是固定的，则可直接应用卡尔曼滤波器以从一列已观测的 LPC 特征矢量 $o_{1:T}$ 来获取一列连续值状态 $x_{1:T}$。

图 5 提供了一种一般的方法的流程图，该方法选择线性逼近，并在卡尔曼滤波器中使用该近似以使用公式 1、2 和 4 来标识一列连续值的状态，同时忽略公式 4 中二次项的递增部分。图 6 和 7 提供了图 5 的方法中使用的组件的框图。

在图 5 的步骤 500，通过量化可能的声道共振（VTR）频率和带宽形成一组量化值，然后对量化值的不同组合形成条目，来构造储存在一表中的 VTR 码本。由此，所得的码本包含作为 VTR 频率和带宽的条目。例如，如果码本包含四个 VTR 的条目，码本中第 i 个条目 x[i]为矢量[F_{1i}, B_{1i}, F_{2i}, B_{2i}, F_{3i}, B_{3i}, F_{4i}, B_{4i}]，其中，F_{1i}、F_{2i}、F_{3i} 和 F_{4i} 是第一、第二、第三和第四 VTR 的频率，B_{1i}、B_{2i}、B_{3i} 和 B_{4i} 是第一、第二、第三和第四 VTR 的带宽。在以下的讨论中，码本的索引 i 可与储存在该索引上的值 x[i]交换使用。当下文单独使用索引时，它意味着表示储存在该索引上的值。

在一个实施例中，依照下表 3 中的条目量化共振峰和带宽，其中 Min(Hz)是以赫兹表示的频率或带宽的最小值，Max(Hz)是以赫兹表示的最大值，“Num.Quant.”是量化状态数。对于频率和带宽，最小值和最大值之间的范围由量化状态数来划分，以在每一量化状态之间提供分隔。例如，对于表 3 中的带宽 B_1，260Hz 的范围由 5 个量化状态均匀地划分，使得每一状态按照 65Hz 与其它状态分隔（即，40、105、170、235、300）。

	Min(Hz)	Max(Hz)	Num.Quant.
F_1	200	900	20
F_2	600	2800	20
F_3	1400	3800	20
F_4	1700	5000	20
B_1	40	300	5
B_2	60	300	5
B_3	60	500	5

B_4	100	700	5

表 3

表 3 中的量化状态数可生成总共 1 亿个以上不同的 VTR 组。然而，由于约束 $F_1<F_2<F_3<F_4$，实际上码本中 VTR 的组较少。

在形成了码本之后，在步骤 502，码本中的条目用于训练描述剩余随机变量的参数。剩余随机变量是一组观测训练特征矢量和一组仿真特征矢量之差。以公式表示：

$$v_t = o_t - S(x_t[i]) \qquad \text{公式 5}$$

其中，v_t 是余量，o_t 是 t 时刻的已观测训练特征矢量，$S(x_t[i])$ 是仿真特征矢量。

如图 6 所示，当向 LPC 倒谱计算器 602 应用 VTR 码本 600 中的一组 VTR $x_t[i]$ 需要时，构造仿真矢量 $S(x_t[i])$，它执行以下计算：

$$S_n(x_t[i]) = \sum_{k=1}^{K} \frac{2}{n} e^{-\pi n \frac{b_k[i]}{f_t}} \cos(2\pi n \frac{f_k[i]}{f_s}) \qquad \text{公式 6}$$

其中，$S_n(x_t[i])$ 是 n 阶 LPC 倒谱特征矢量中的第 n 个元素，K 是 VTR 的数量，f_k 是第 k 个 VTR 频率，b_k 是第 k 个 VTR 带宽，f_s 是采样频率，在许多实施例中为 8kHz。S_0 元素被设为等于 $\log G$，其中，G 是增益。

为产生用于训练余量模型的已观察训练特征矢量 o_t，人类说话者 612 生成由麦克风 616 检测的声学信号，麦克风 616 也检测附加噪声 614。麦克风 616 将声学信号转化成提供给模一数（A/D）转化器 618 的模拟电信号。模拟信号由 A/D 转化器 618 以采样频率 f_s 来采样，并将所得的样值转化成数字值。在一个实施例中，A/D 转化器 618 以 8kHz 和每样值 16 比特对模拟信号进行采样，由此创建了每秒 16 千字节的语音数据。在其它实施例中，A/D 转化器 618 以 16kHz 对模拟信号进行采样。数字样值被提供给帧构造器 620，它将样值组合成帧。在一个实施例中，帧构造器 620 每隔 10 毫秒创建包含 25 毫秒数据的新帧。

数据帧被提供给 LPC 倒谱特征提取器 622，它使用快速傅立叶变换（FFT）将信号变换到频域，然后使用 LPC 系数系统 626 标识表示语音信号帧的频谱内容的多项式。使用递归 628 将 LPC 系数转化成 LPC 倒谱系数。递归 628 的输出是表示训练语音信号的一组训练特征矢量 630。仿真特征矢量 610 和训练特征矢量 630 被提供给余量训练器 632，它训练余量 v_t 的参数。

在一个实施例中，v_t 是具有平均值 h 和精度 D 的单个高斯型，其中，h 是对特征矢量的每一分量具有单独的平均值的矢量，D 是对特征矢量的每一分量具有单

独的值的对角精度矩阵。

在本发明的一个实施例中，使用期望值最大化（EM）算法来训练这些参数。在该算法的 E 步骤，确定后验概率 $\gamma_t(i)=p(x_t[i]|o_1^N)$。在一个实施例中，该后验概率使用后向递归来确定，定义如下：

$$\gamma_t(i)=\frac{\rho_t(i)\sigma_t(i)}{\sum_i \rho_t(i)\sigma_t(i)} \qquad \text{公式 7}$$

其中，$\rho_t(i)$ 和 $\sigma_t(i)$ 被递归地定义为：

$$\rho_t(i)=\sum_j \rho_{t-1}(j)p(x_t[i]|x_{t-1}[j])p(o_t|x_t[i]=x[i]) \qquad \text{公式 8}$$

$$\sigma_t(i)=\sum_j \sigma_{t+1}(j)p(x_t[i]|x_{t+1}[j])p(o_t|x_t[i]=x[i]) \qquad \text{公式 9}$$

在本发明的一个方面，使用上述公式 1 来确定转移概率 $p(x_t[i]|x_{t-1}[j])$ 和 $p(x_t[i]|x_{t+1}[j])$，此处为方便起见，使用码本索引表示法来重复该公式：

$$x_t[i]=\Phi x_{t-1}[i]+(I-\Phi)T+w_t \qquad \text{公式 10}$$

其中，$x_t[i]$是帧 t 的 VTR 的值，$x_{t-1}[j]$是前一帧 t-1 的 VTR 的值，Φ 是速率，T 是与帧 t 相关联的 VTR 的目标，w_t 是帧 t 的噪声，在一个实施例中假定噪声为具有精度矩阵 B 的零均值高斯型。

使用这一动态模型，转移概率可被描述为高斯函数：

$$p(x_t[i]|x_{t-1}[j])=N(x_t[i];\Phi x_{t-1}[i]+(I-\Phi)T,B) \qquad \text{公式 11}$$

$$p(x_t[i]|x_{t+1}[j])=N(x_{t+1}[i];\Phi x_t[i]+(I-\Phi)T,B) \qquad \text{公式 12}$$

可选地，可通过令概率仅取决于当前观测矢量而非矢量序列来估算后验概率 $\gamma_t(i)=p(x_t[i]|o_1^N)$，使得后验概率变为：

$$\gamma_t(i)\approx p(x_t[i]|o_t) \qquad \text{公式 13}$$

它可被计算如下：

$$p(x_t[i]|o_t)=\frac{N(o_t;S(x_t[i])+\hat{h},\hat{D})}{\sum_{i=1}^{I} N(o_t;S(x_t[i])+\hat{h},\hat{D})} \qquad \text{公式 14}$$

其中，$\hat{h}$ 是余量的平均值，$\hat{D}$ 是余量的精度，余量是根据 EM 算法的前一次迭代确定的，或者如果是第一次迭代，则是最初设定的。在执行了 E 步骤来标识后验概率 $\gamma_t(i)=p(x_t[i]|o_1^N)$ 之后，执行 M 步骤，使用以下公式来确定余量的方差 D^{-1}（精度矩阵的逆）的平均值 h 和每一对角元素 d^{-1}：

$$\hat{h}=\frac{\sum_{t=1}^{N}\sum_{i=1}^{I}\gamma_t(i)\{o_t-S(x_t[i])\}}{N} \qquad \text{公式 15}$$

$$\hat{d}^{-1}=\frac{\sum_{t=1}^{N}\sum_{i-1}^{I}\gamma_t(i)\{o_t-S(x_t[i])-\hat{h}\}^2}{N} \qquad \text{公式 16}$$

其中，N是训练话语中的帧的数量，I是VTR的量化组合的数量，o_t是t时刻的已观测特征矢量，$S(x_t[i])$是VTR $x_t[i]$的仿真特征矢量。

余量训练器632通过重复E步骤和M步骤来多次更新平均值和方差，每次都使用前一次迭代的平均值和方差。在平均值和方差达到稳定值之后，它们被作为余量参数634储存。

一旦构造了余量参数634，它们可在图5的步骤504中用于标识输入的语音信号中的VTR矢量。图7示出了用于标识VTR矢量的系统的框图。

在图7中，语音信号由说话者712生成。语音信号和附加噪声714由麦克风716、A/D转化器718、帧构造器720和特征提取器722转化成特征矢量流710，特征提取器包括FFT 724、LPC系统716和递归728。注意，麦克风716、A/D转化器718、帧构造器720和特征提取器722以与图6的麦克风616、A/D转化器618、帧构造器620和特征提取器622相同的方式操作。

特征矢量流730连同余量参数634和仿真特征矢量610一起提供给VTR跟踪器732。VTR跟踪器732使用动态编程来标识一列最可能的VTR矢量734。特别地，它使用维特比（Viterbi）解码算法，其中，网格图中的每一节点具有下列公式的最优部分得分：

$$\delta_t(i)=\max_{x[i]_1^{t-1}}\prod_{\tau=1}^{t-1}p(o_\tau\mid x_\tau[i])p(o_t\mid x_t[i]=x[i])$$
$$\times p(x[i]_1)\prod_{\tau=2}^{t-1}p(x_\tau[i]\mid x_{\tau-1}[i])p(x_\tau[i]=x[i]\mid x_{t-1}[i])$$

公式 17

基于最优原理，t+1处理阶段的最优部分似然性可使用以下维特比递归来计算：

$$\delta_{t+1}(i)=\max_{i'}\delta_t(i')p(x_{t+1}[i]=x[i]\mid x_t[i]=x[i'])p(o_{t+1}\mid x_{t+1}[i]=x[i]) \qquad \text{公式 18}$$

在公式18中，“转移”概率$p(x_{t+1}[i]=x[i]\mid x_t[i]=x[i'])$使用上文的状态方程式10来计算，以生成高斯分布：

$$p(x_{t+1}[i]=x[i]\mid x_t[i]=x[i'])=N(x_{t+1}[i];\Phi x_t[i']+(I-\Phi)T,B) \qquad \text{公式 19}$$

其中，$\Phi x_t[i]+(I-\Phi)T$是该分布的平均值，B是该分布的精度。

公式18的观测概率$p(o_{t+1}[i]=x[i])$被作为高斯型处理，并根据观测方程式5和余量参数h和D来计算，使得：

$$p(o_{t+1}\mid x_{t+1}[i]=x[i])=N(o_{t+1};S(x_{t+1}[i]+h,D) \qquad \text{公式 20}$$

公式 20 中最优量化索引 i'的后向跟踪提供了初始 VTR 序列 734。

为减少必须执行的计算数量，可执行修剪（pruning）束搜索来替代严格的维特比搜索。在一个实施例中，在对每一帧仅标识一个索引时，使用修剪的极端形式。

在步骤 504 标识了初始 VTR 序列 734 之后，将初始 VTR 序列提供给线性参数估算器 736，它选择用于上述步骤 506 处的公式 4 的线性逼近的参数。具体地，对于每一帧，该帧的初始 VTR 矢量用于确定对每一声道共振索引 k 和每一 LPC 阶 n 的线性参数α_{kx}、β_{kx}、γ_{kx}和δ_{kx}的值。

在一个实施例中，通过向指数项$e^{-\pi n\frac{b_k}{f_s}}$应用初始 VTR 矢量的带宽 b_k 并计算该指数的值来对 LPC 阶 n 确定线性参数α_{kx}和β_{kx}的值。然后选择图 3 中跨越该指数值的线段，由此选择定义线段的线性参数α_{kx}和β_{kx}。注意，这些参数的每一个是对除与带宽 b_k 相关联的矢量分量之外的每一矢量分量具有零值的矢量。

在一个实施例中，通过向余弦项$\cos(2\pi n\frac{f_k}{f_s})$应用初始 VTR 矢量的频率 f_k 并计算该余弦的值来对 LPC 阶 n 确定线性参数γ_{kx}和δ_{kx}的值。然后选择图 4 中跨越该余弦值的线段，由此选择了定义线段的线性参数γ_{kx}和δ_{kx}。注意，这些参数的每一个是对除与频率f_k相关联的矢量分量之外的每一矢量分量具有零值的矢量。

在步骤 508，将每一帧的线性参数应用到公式 4。忽略公式 4 中二次项的递增部分，公式 4 在公式 2 中使用。然后将公式 1 和 2 提供给卡尔曼滤波器 738，它对每一帧重新估算 VTR 矢量。在步骤 510，过程确定是否存在更多迭代要执行。如果存在更多迭代，则过程返回到步骤 506，根据新 VTR 矢量重新估算线性参数。然后将新线性参数应用到公式 2 到公式 4，并且在步骤 508 在卡尔曼滤波器 738 中使用公式 1 和 2 来重新估算 VTR 矢量。重复步骤 506、508 和 510，直到在步骤 510 确定不需要更多的迭代。在这一点上，过程在步骤 512 结束，VTR 矢量 734 的最后一次估算用作输入信号的声道共振频率和带宽序列。

注意，卡尔曼滤波器 738 提供了声道共振矢量的连续值。由此，所得的声道共振频率和带宽的序列不限于 VTR 码本 600 中找到的离散值。

尽管参考具体实施例描述了本发明，然而本领域的技术人员将认识到，可在不脱离本发明的精神和范围的情况下在形式和细节上作出改变。

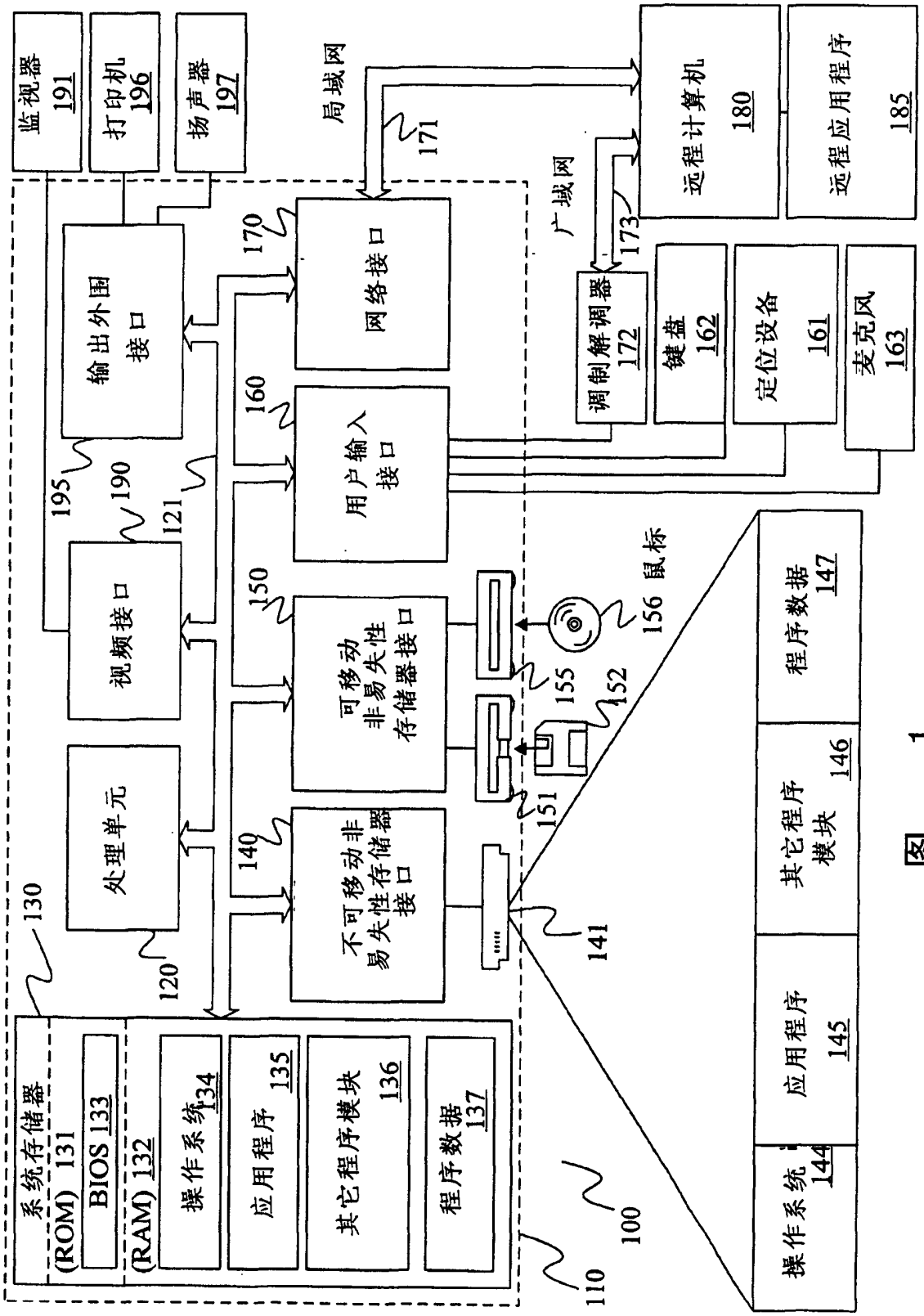

监视器 191
打印机 196
扬声器 197
局域网
171
远程计算机 180
远程应用程序 185
广域网
173
输出外围接口
195
190
121
网络接口
170
160
用户输入接口
调制解调器 172
键盘 162
定位设备 161
麦克风 163
视频接口
150
可移动非易失性存储器接口
处理单元
120
140
不可移动非易失性存储器接口
130
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
110
100
151
155
152
156 鼠标
141
操作系统 144
应用程序 145
其它程序模块 146
程序数据 147

图 1

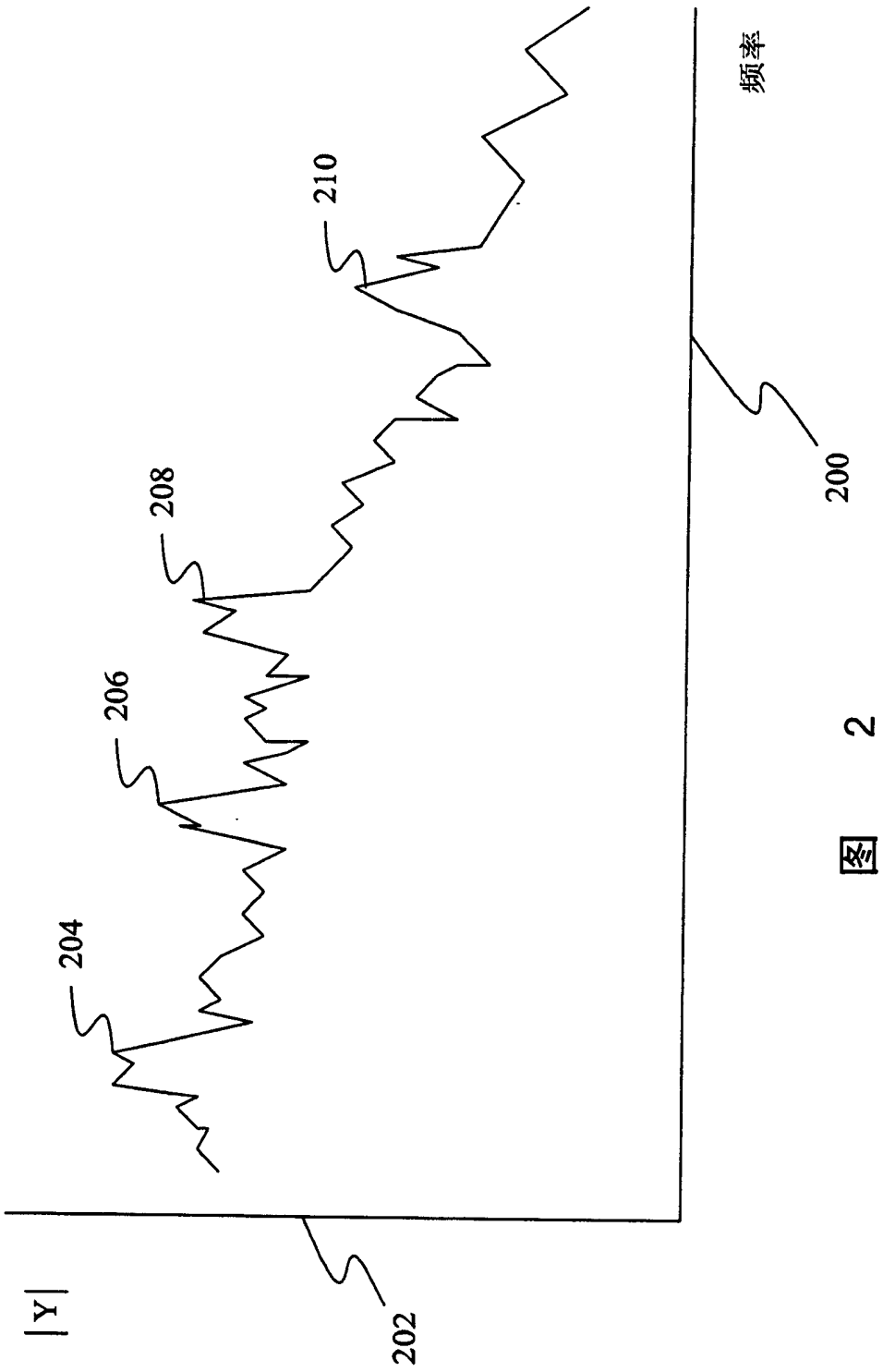


图 2

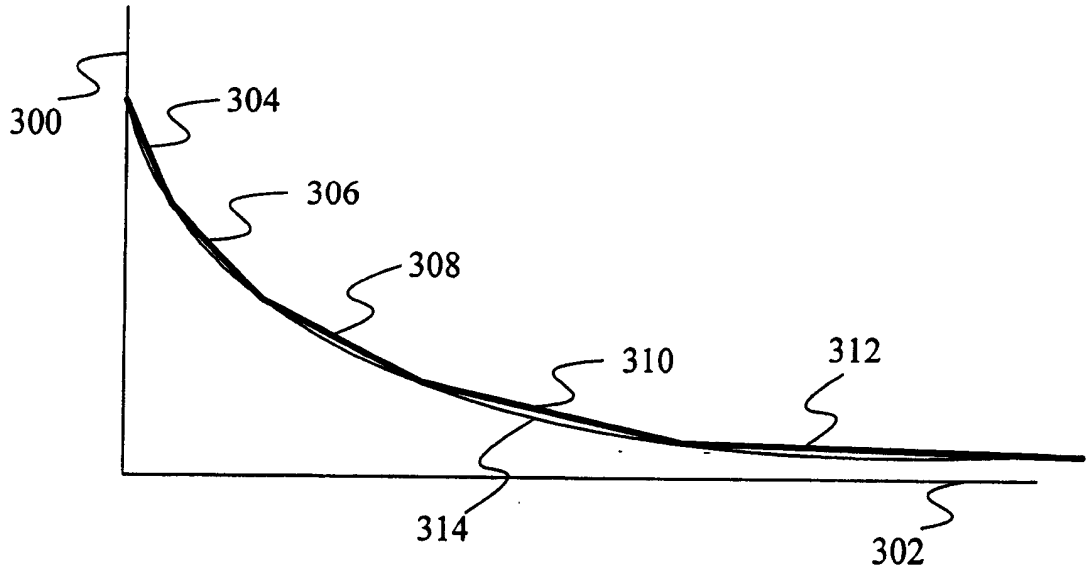

300
304
306
308
310
312
314
302

图 3

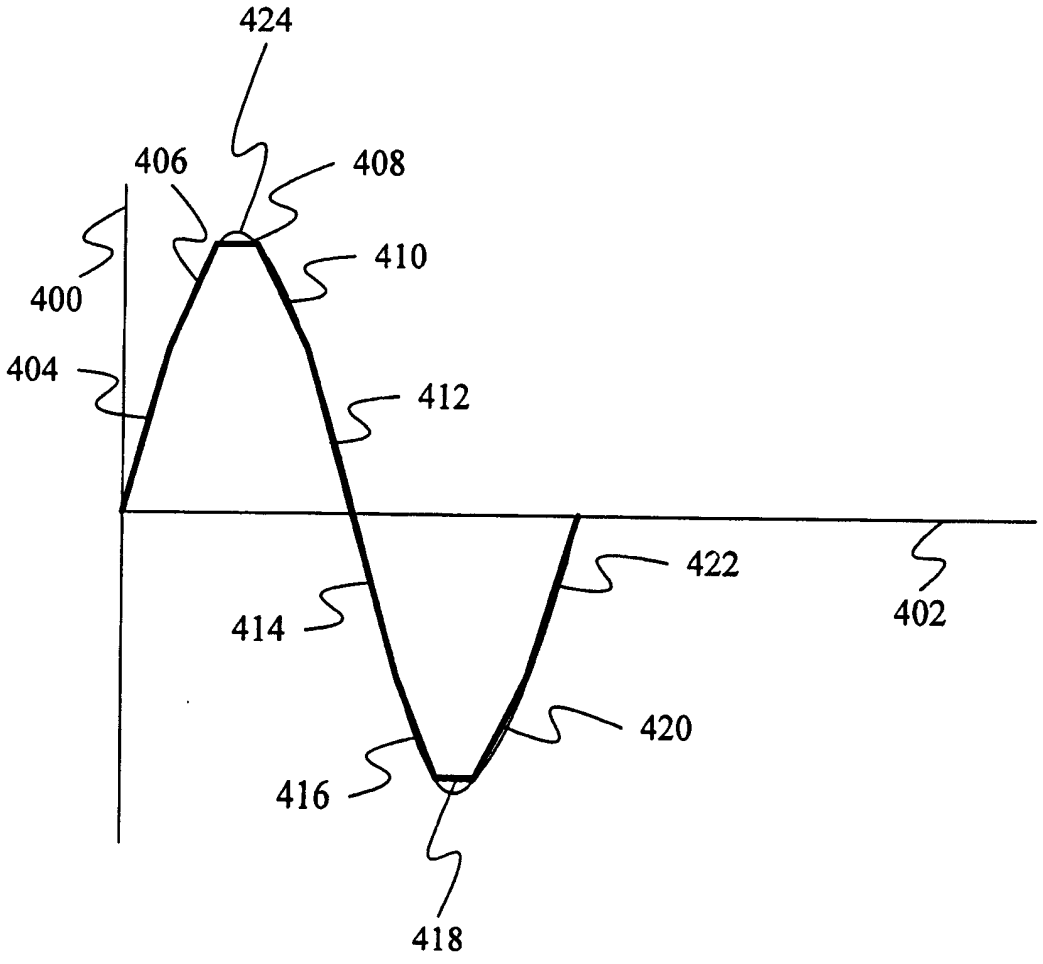

424
406
408
400
410
404
412
422
402
414
420
416
418

图 4

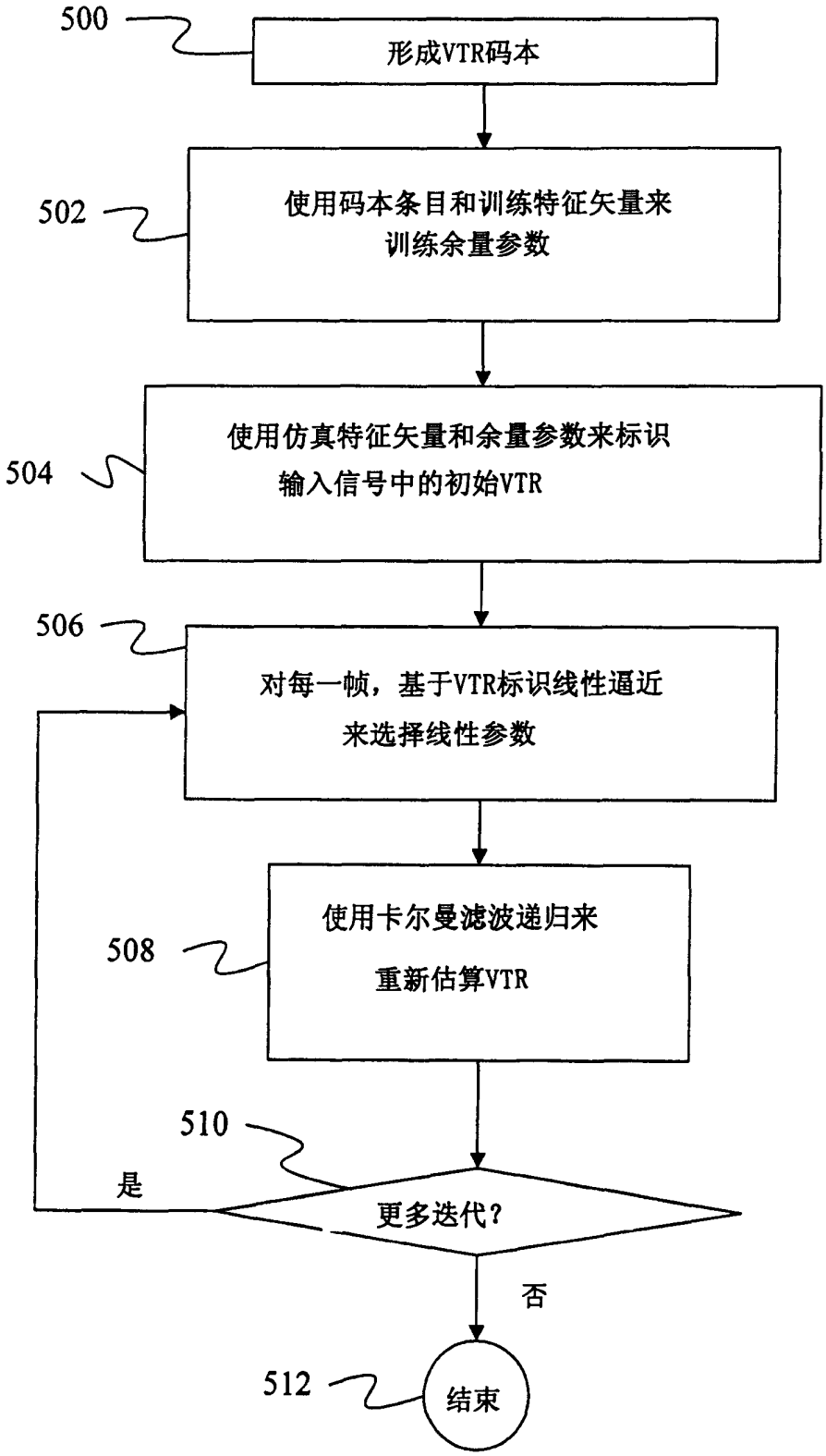

500
形成VTR码本
502
使用码本条目和训练特征矢量来
训练余量参数
504
使用仿真特征矢量和余量参数来标识
输入信号中的初始VTR
506
对每一帧，基于VTR标识线性逼近
来选择线性参数
508
使用卡尔曼滤波递归来
重新估算VTR
510
是
更多迭代？
否
512
结束

图 5

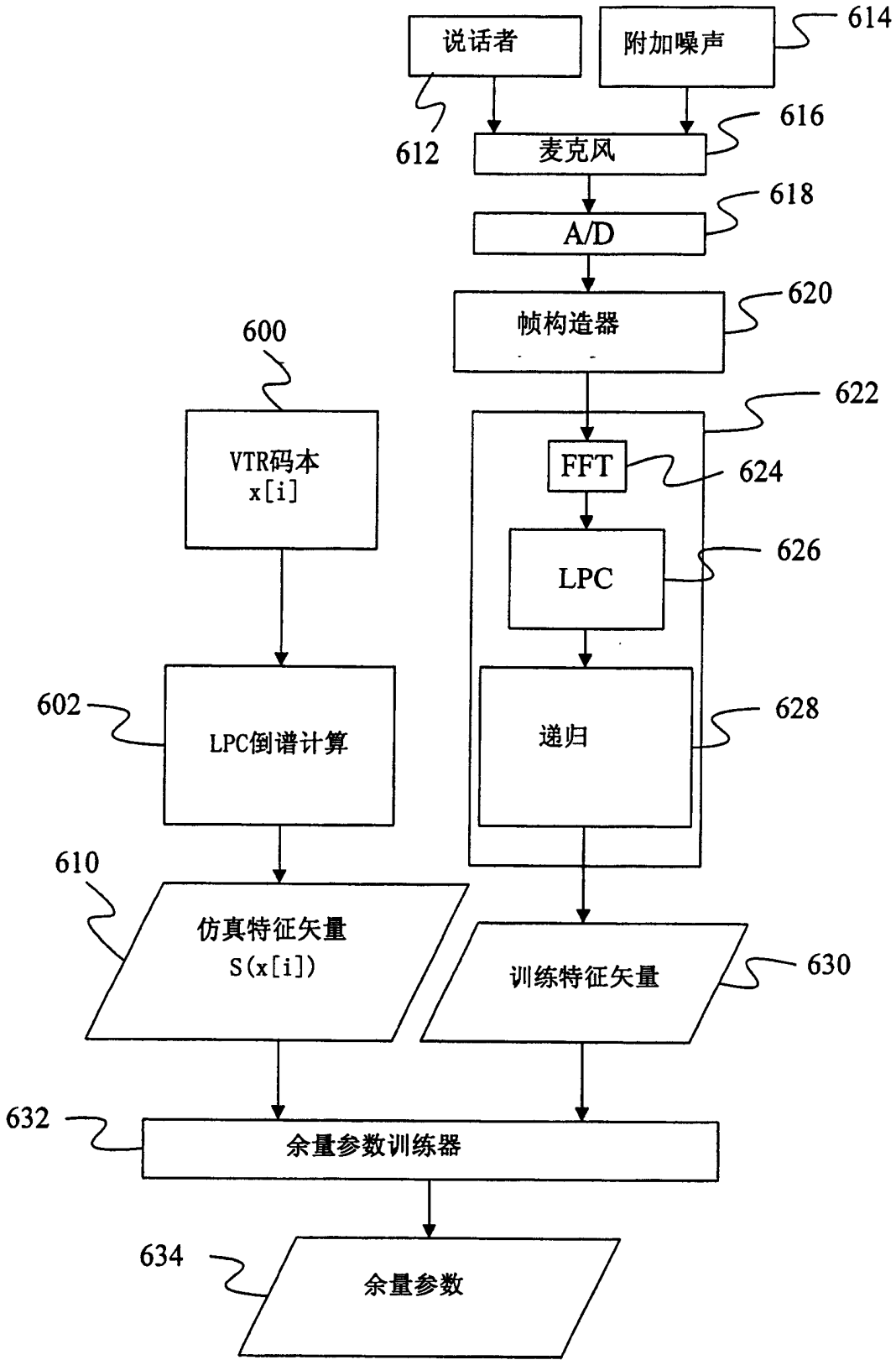

说话者
附加噪声
614
612
麦克风
616
A/D
618
帧构造器
620
600
622
VTR码本
x[i]
FFT
624
LPC
626
602
LPC倒谱计算
递归
628
610
仿真特征矢量
S(x[i])
训练特征矢量
630
632
余量参数训练器
634
余量参数

图 6

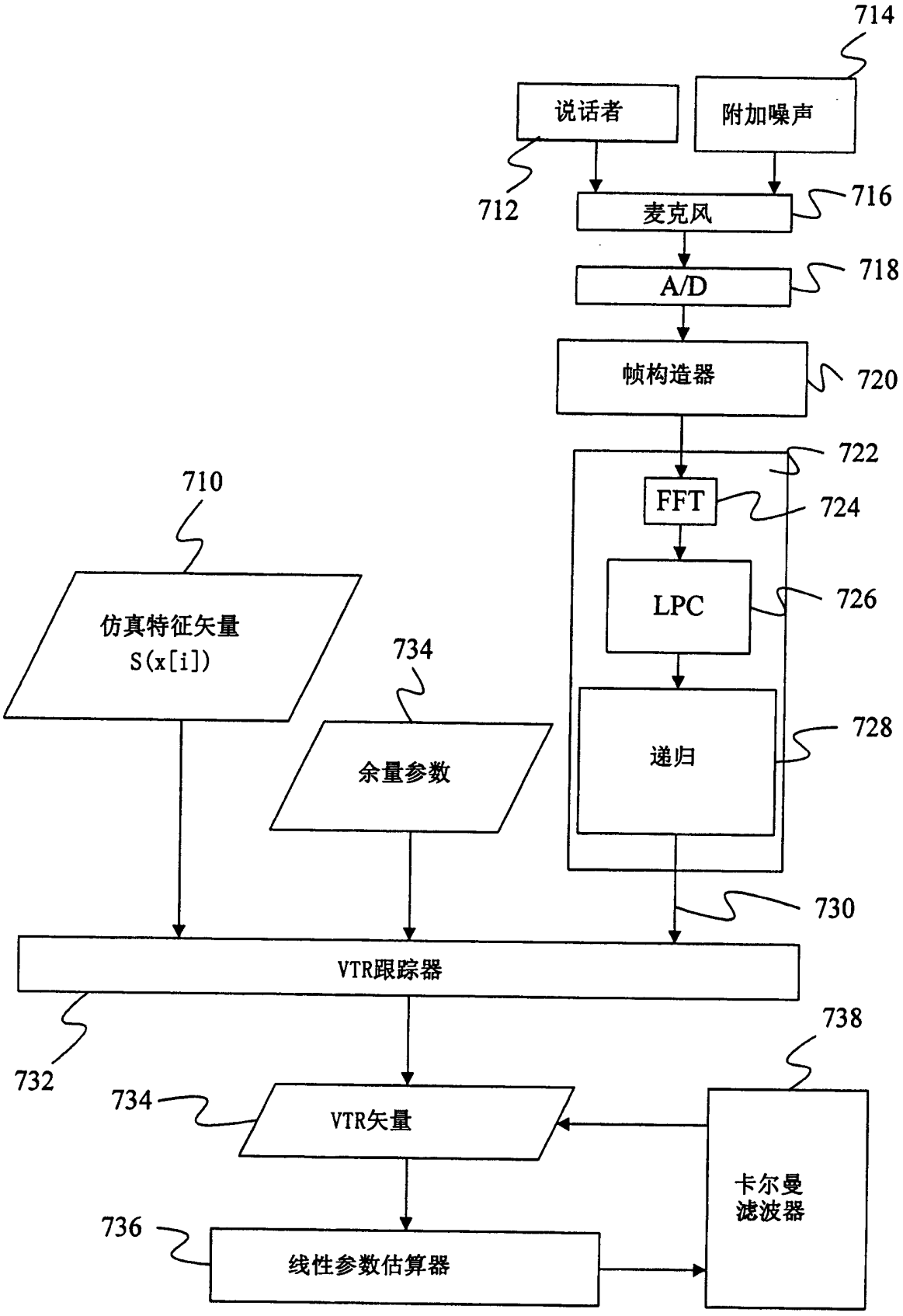

714
说话者
附加噪声
712
麦克风
716
A/D
718
帧构造器
720
722
FFT
724
710
LPC
726
仿真特征矢量
S(x[i])
734
余量参数
递归
728
730
VTR跟踪器
732
738
734
VTR矢量
卡尔曼
滤波器
736
线性参数估算器

图 7