



[12] 发 明 专 利 说 明 书

[21] ZL 专利号 00803548.2

[45] 授权公告日 2004 年 9 月 22 日

[11] 授权公告号 CN 1168070C

[22] 申请日 2000.2.4 [21] 申请号 00803548.2

[30] 优先权

[32] 1999. 2. 8 [33] US [31] 09/246,413

[86] 国际申请 PCT/US2000/002904 2000.2.4

[87] 国际公布 WO2000/046794 英 2000.8.10

[85] 进入国家阶段日期 2001.8.8

[71] 专利权人 高通股份有限公司

地址 美国加利福尼亚州

[72] 发明人 张承纯

审查员 胡 婧

[74] 专利代理机构 上海专利商标事务所

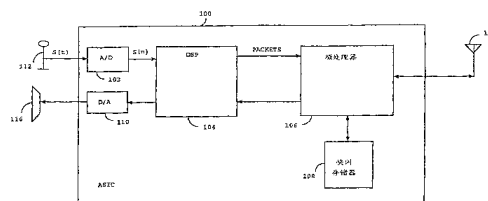
代理人 吴蓉军

权利要求书 2 页 说明书 8 页 附图 3 页

[54] 发明名称 分布式语音识别系统

[57] 摘要

一种分布式语音识别系统，包括数字信号处理器(DSP)(104)、非易失存储媒体(108)和微处理器(106)。DSP(104)配置成从数字化输入话音采样中提取参数并向微处理器(106)提供所提取的参数。非易失存储媒体包括话音模板的数据库。将微处理器配置成读取非易失媒体(108)的内容、将参数与内容相比较并根据比较结果选择话音模板。非易失存储媒体可以是快闪存储器。DSP(104)可以是声码器。如果DSP(104)是声码器，那么参数可以是由声码器产生的诊断数据。分布式语音识别系统可位于用途特定集成电路(ASIC)上。



1. 一种分布式语音识别系统，其特征在于，包括：

数字信号处理器，用于接收数字化话音采样，并从中提取多个参数，所述数字信号处理器包括一个声码器，用于产生诊断数据和声码器分组，其中所述诊断数据包含非量化参数，所述声码器分组是经量化的参数；

存储媒体，它包括多个话音模板；和

处理器，它与所述存储媒体和所述数字信号处理器耦合，用于接收来自所述数字信号处理器的所述诊断数据和所述声码器分组，由所述诊断数据或所述声码器分组获得所述多个参数，并将所述多个参数与所述多个话音模板相比较。

2. 如权利要求 1 所述的系统，其特征在于，所述处理器还被配置成记录所述声码器分组，并根据所述多个参数与所述多个话音模板的比较，从所述多个话音模板中选择一个模板。

3. 如权利要求 1 所述的系统，其特征在于，所述数字信号处理器、所述存储媒体和所述处理器位于一特定用途集成电路上。

4. 如权利要求 1 所述的系统，其特征在于，所述多个参数构成由所述声码器产生的至少一部分诊断数据。

5. 如权利要求 4 所述的系统，其特征在于，所述声码器、所述存储媒体和所述处理器位于一特定用途集成电路上。

6. 一种用于分布语音识别处理的方法，其特征在于，包括下述步骤：

在数字信号处理器中，用声码器从多个数字化话音采样中提取多个参数，所述声码器被配置成用于产生诊断数据和声码器分组，其中所述诊断数据包含非量化参数，所述声码器分组是经量化的参数；

向微处理器提供所述诊断数据和所述声码器分组；和

在微处理器中，从所述诊断数据或所述声码器分组获得所述多个参数，并将所述多个参数与多个话音模板相比较。

7. 如权利要求 6 所述的方法，其特征在于，还包括下述步骤：将所述声码器分组记录在微处理器中；并且在微处理器中，根据所述比较步骤的结果，从所述话音模板中选择一个模板。

8. 如权利要求 6 所述的方法，其特征在于，所述多个参数构成由声码器

产生的至少一部分诊断数据。

9. 一种分布式语音识别系统，其特征在于，包括：

用于从多个数字化话音采样中提取多个参数的装置，所述提取装置包括用于产生声码器分组和诊断数据的装置，其中所述诊断数据包含非量化参数，所述声码器分组是经量化的参数；

用于存储多个话音模板的装置；和

用于接收来自所述提取装置的所述声码器分组和所述诊断数据，由所述诊断数据或所述声码器分组获得所述多个参数，并将所述多个参数与所述多个话音模板相比较的装置。

10. 如权利要求 9 所述的系统，其特征在于，所述接收与比较装置包括用于记录所述声码器分组，并根据所述多个参数与所述多个话音模板的比较，从所述多个话音模板中选择一个模板的装置。

11. 如权利要求 9 所述的系统，其特征在于，所述提取装置、所述存储装置和所述接收与比较装置位于一特定用途集成电路上。

12. 如权利要求 9 所述的系统，其特征在于，所述多个参数构成由用于产生声码器分组和诊断数据的装置产生的至少一部分诊断数据。

13. 如权利要求 9 所述的系统，其特征在于，所述存储装置包括一快闪存储器。

14. 如权利要求 9 所述的系统，其特征在于，所述声码器、所述存储装置和所述接收与比较装置位于一特定用途集成电路上。

分布式语音识别系统

发明领域

本发明一般涉及通信领域，特别涉及语音识别系统。

发明背景

语音识别(VR)代表赋予机器模拟智能以识别用户或用户发声的命令并促进人机接口的最重要技术之一。VR 还代表人们话音理解的关键技术。采用从声学语音信号中恢复语言消息的技术的系统被称为语音识别器。语音识别器一般包括声处理器，它提取获得入局原始话音的 VR 所需的一系列带信息特征或矢量；字解码器，它解码特征或矢量系列以提供有意义和所需的输出格式，诸如与输入说话(utterance)相对应的一系列语言字。为了增加给定系统的性能，需要进行训练来将系统设置有效参数。换句话说，系统在它能够起到最佳作用之前需要学习。

声处理器代表在语音识别器中的前端话音分析子系统。响应于输入话音信号，声处理器提供一种表示时间变化话音信号的特征的适当表示法。声处理器应丢弃无关信息，诸如背景噪声、信道失真、扬声器特征和说话方式。有效的声处理为语音识别器提供增强的声区别(acoustic discrimination)能力。为此，要分析的有用特征是短时间谱包络。表示短时间谱包络的两种常用频谱分析技术是线性预测编码(LPC)和基于滤波器组的频谱建模(filter-bank-based spectral modeling)。在美国专利号 5,414,796(已转让给本发明的受让人并作为参考资料在此引入)和 L.B.Rabiner 与 R.W.Schafer 所著的“话音信号的数字处理”(Digital Processing of Speech Signals 396-453(1978)，作为参考资料在此引入)中描述了示例 LPC 技术。

处于安全原因，VR(一般称为话音识别)的使用变得越来越重要。例如，VR 可用来代替在无线电话键盘上人工按键的工作。这在用户在开车时要打电话时特别有用。当使用没有 VR 的电话时，司机必需从方向盘上腾出一只手并在按键拨号的同时要看着键盘。这种动作增加了交通事故的可能性。话音使能的电话(即，为话音识别设计的电话)允许司机打电话，同时继续看路。此外，

车辆上的免提配套系统允许司机在呼叫启动期间将两只手放在方向盘上。

话音识别装置分为依赖扬声器或不依赖扬声器装置。不依赖扬声器装置能够接受来自任何用户的语音命令。训练更加普遍的依赖扬声器的装置来识别来自特定用户的命令。依赖扬声器的 VR 装置一般在两个阶段操作, 即, 训练阶段和识别阶段。在训练阶段, VR 系统提醒用户说在系统词汇表中的每个字一遍或两遍, 从而系统能够对于这些特定字或短语学习用户的话音特征。另一方面, 对于表示语音的(phonetic)VR 装置, 通过阅读覆盖语言中的所有音素的一篇或多篇短文可以完成训练。对于免提汽车配套设备的示例词汇表可包括键盘上的数字; 关键字“呼叫”、“发送”、“拨号”、“删除”、“清除”、“添加”、“删除”、“历史”、“程序”、“是”和“否”; 和预定数量的一般称为共同工作者、朋友或家庭成员的名字。一旦完成训练, 用户就能够通过说出训练的关键字在识别阶段启动呼叫。例如, 如果名字“John”是一个经训练的名字, 那么用户能够通过说出短语“呼叫 John”来启动对 John 的呼叫。VR 系统识别字“呼叫(Call)”和“John”, 而且能够拨出用户此前作为 John 的电话号码输入的号码。

传统的 VR 装置一般用数字信号处理器(DSP)或微处理器来分析入局语音采样、提取相关参数、解码参数并将解码的参数与存储的一组字或 VR 模板(template)相比较, 所述模板包括 VR 装置的词汇表。把词汇表存储在非易失存储器中, 诸如, 快闪存储器。在具有 DSP 和微处理器(诸如, 数字蜂窝电话)的传统 VR 系统中, 非易失存储器一般可被微处理器访问, 但是 DSP 不能访问。在这样的系统中, 如果完全在微处理器中执行 VR, 那么微处理器通常缺乏计算能力, 从而以合理的等待时间传递识别结果。另一方面, 如果完全在 DSP 中执行 VR, 那么微处理器需要读取快闪存储器并将读取的内容传给 DSP, 因为 DSP 的单片存储器(on-chip memory)的尺寸相对较小不足以保持大 VR 模板。由于在 DSP 和微处理器之间的接口的典型低带宽限制了可在给定时间内在两个装置之间传递的数据量, 因此这是冗长的过程。于是, 需要 VR 装置有效地将 DSP 的计算能力与微处理器的存储器容量相结合。

发明内容

本发明提供一种将 DSP 的计算能力与微处理器的存储器容量有效地结合的 VR 装置。因此, 在本发明的一个方面, 分布式语音识别系统最好包括数字

信号处理器，配置成接收数字化语音采样并从中提取多个参数；存储媒体，包括多个语音模板；和耦合到存储媒体和数字信号处理器的处理器，所述处理器被配置成接收来自数字信号处理器的多个参数并将多个参数与多个语音模板相比较。

在本发明的另一个方面，分布式语音识别处理的方法最好包括步骤：在数字信号处理器中，从多个数字化语音采样中提取多个参数；将该多个参数提供给微处理器；以及在微处理器中，将该多个参数与多个语音模板进行比较。

在本发明的另一个方面，分布式语音识别系统最好包括从多个数字化语音采样中提取多个参数的装置；永久存储多个语音模板的装置；和接收来自所述提取装置的多个参数并将所述多个参数与所述多个语音模板相比较的装置。

附图简述

图 1 是传统语音识别系统的方框图；

图 2 是分布式语音识别系统的方框图；

图 3 是将声码器用于前端处理的分布式语音识别系统的方框图。

较佳实施例的详细描述

如图 1 所示，传统的语音识别系统 10 包括模拟-数字变换器(A/D)12、声处理器 14、VR 模板数据库 16、模式比较逻辑 18 和判定逻辑 20。例如，VR 系统 10 可位于无线电话或免提汽车配置设备中。

当 VR 系统 10 处于语音识别阶段，人(未图示)说出一个字或短语，产生语音信号。用传统的变换器(未图示)将语音信号转换成电信号 $s(t)$ 。向 A/D12 提供语音信号 $s(t)$ ，其中根据已知采样方法(例如，脉冲编码调制(PCM))将语音信号 $s(t)$ 转换成数字化语音采样 $s(n)$ 。

向声处理器 14 提供语音采样 $s(n)$ 用于参数判定。声处理器 14 产生一组参数，对输入语音信号 $s(t)$ 的特征建模。根据多种已知的话音参数判定技术(例如，包括话音编码器编码)并运用基于快速傅利叶变换(FFT)的倒谱系数(如上述美国专利 5,414,796 所述)，可以判定这些参数。可将声处理器 14 作为数字信号处理器实施。DSP 可包括话音编码器。另一方面，可将声处理器 14 作为话音编码器来实施。

在训练 VR 系统 10 期间，还执行参数判定，其中将一组用于 VR 系统 10

的所有词汇表字的模板路由到 VR 模板数据库 16 以永久地存储在其中。VR 模板数据库 16 最好以任一种传统的非易失存储媒体形式来实施，诸如，快闪存储器。这允许当关断到 VR 系统 10 的电源时，模板保留在 VR 模板数据库 16 中。

向模式比较逻辑 18 提供这组参数。模式比较逻辑 18 最好检测说话的起点和终点、计算动态声特征(acoustic feature)(诸如，时间导数、第二时间导数，等等)、通过选择相关帧压缩声特征并量化静态和动态声特征。例如，在 Lawrence Rabiner & Biing-Hwang Juang 所著的话音识别基础(Fundamentals of Speech Recognition(1993)，作为参考资料在此引入)中描述了各种已知的端点判定方法、动态声特征推导、模式压缩和模式量化。模式比较逻辑 18 将所得参数组与存储在 VR 模板数据库 16 中的所有模板相比较。向判定逻辑 20 提供在参数组和存储在 VR 模板数据库 16 中的所有模板之间的比较结果或者距离。判定逻辑 20 从 VR 模板数据库 16 中选择最接近匹配参数组的模板。另一方面，判定逻辑 20 可用传统的“N-最佳(N-best)”选择算法，它选择在预定匹配门限内的 N 个最接近匹配。于是，询问人趋于选择哪个。判定逻辑 20 的输出是关于说出词汇表中的哪个字的判定。

模式比较逻辑 18 和判定逻辑 20 最好作为微处理器来实施。例如，VR 系统 10 可以是特定用途集成电路(ASIC)。VR 系统 10 的识别精度是对 VR 系统 10 能够多正确地识别所说的词汇表中的字或短语的测量标准。例如，识别精度 95%表示 VR 系统 10 在 100 次中能够正确地识别词汇表中的字 95 次。

在如图 2 所示的一个实施例中，分布式 VR 系统位于 ASCII 100 中。例如，ASCII 100 可位于无线电话内。ASIC 100 包括 A/D102、DSP104、处理器 106、非易失存储媒体 108 和数字-模拟变换器(D/A)110。处理器 106 最好是微处理器 106，但是另一方面，处理器 106 可以是任何传统的处理器、控制器或状态机。非易失存储器媒体 108 最好是快闪存储器 108，但是另一方面，非易失存储器媒体 108 可以是任何形式的传统非易失、可写存储器。DSP104 包括话音编码器(未图示)和话音解码器(也未图示)。DSP 104 可以执行几种功能，包括，声编码、VR 前端处理 (即，话音分析)背景噪声抑制、和声回波消除。另一方面，DSP104 可以是传统声码器，诸如，8 千比特/秒(kbps)码激励线性预测 (CELP) 编码器、13kbps CELP 编码器 (如上述美国专利第 5414796 号中所述)、或码分多址 (CDMA) 加强可变速率编码(EVRC)编码器。

麦克风 112 耦合到 ASIC100 作为到 A/D102 的输入。说话者(未图示)对着麦克风 112 说话, 它向 A/D102 提供模拟话音信号 $s(t)$ 。A/D102 根据已知的方法(诸如, PCM)将模拟话音信号 $s(t)$ 转换成数字模板 $s(n)$ 。将 A/D102 耦合到 DSP104。向 DSP104 提供数字化话音采样 $s(n)$ 用于前端处理。

在 DSP104 中, 对数字化话音采样 $s(n)$ 执行傅利叶变换, 将时域采样 $s(n)$ 转换成频域值。通过使信号经过多个带通滤波器(未图示)将频域分成几个相邻频带宽。带通滤波器最好是有限脉冲响应(FIR)滤波器。将滤波器抽头设为以每个连续滤波器通过更宽频带宽, 其中在连续滤波器之间的宽度变化以对数增加。所有滤波器通过的频率范围覆盖了人语音范围。在较低端, 通过的带宽相对窄, 同时在较高端, 通过带宽相对宽。这有助于人们耳朵更加精确地察觉声音, 因为人的耳朵在较低频率下能够更好地区分。对数频率标度可以是 Bark 标度, 或者另一方面, 该标度可以是 Mel 标度。已知这些标度, 而且在上述 Lawrence Rabiner & Biing-Hwang Juang 所著的话音识别基础(Fundamentals of Speech Recognition 78-79 (1993))中描述。

在一个实施例中, 采用 13 个带通滤波器。下列等式描述滤波器组:

$$b_1 = C,$$

$$b_i = b_{i-1}, \quad 2 \leq i \leq 13, \quad \text{和}$$

$$f_i = f_1 + \sum_{j=i-1}^{j=1} b_j + (b_i - b_1)/2$$

其中, C 和 f_1 分布是任意设定带宽和第一滤波器的中心频率, 而 α 是对数生长矢量 (logarithmic growth factor)。

对于每帧, DSP104 都根据传统的特征提取方法来提取频带功率值矢量(a vector of band power values), 这在上述美国专利号 5,414,796 (Rabiner & Schafer) 和 Rabiner & Juang 69-140) 的文中所述。在一个实施例中, 每帧提取 13 个频带功率值的矢量, 其中帧为 24 毫秒长并 50%重叠。对频带功率值进行分组并将其作为数据分组向微处理器 106 提供。

微处理器 106 访问快闪存储器 108 以读取快闪存储器 108 中所存的模板词汇总表内容。如果在 VR 解码器(未图示)中要求端点, 那么微处理器 106 根据接收到的数据分组检测说话的起点和终点。然后, 微处理器 106 计算动态声特征(诸如, 时间导数, 第二时间导数, 等等), 通过选择相关帧来压缩声特征并量化静态和动态声特征。例如, 在上述 Lawrence Rabiner & Biing-Hwang Juang

所著的话音识别基础(Fundamentals of Speech Recognition(1993))中描述了各种已知的端点检测、动态声特征导数、模式压缩和模式量化方法。

微处理器 106 将说话与快闪存储器 108 的内容相比较。微处理器 106 选择与数据分组最接近相似的所存字。在上述 Rabiner & Juang (141-241) 的文中描述了多个模式识别技术。于是,微处理器 106 执行与所存字相关的命令,诸如,通过耦合到微处理器 106 的天线与基站(未图示)进行通信来启动到特定人的电话呼叫。

当开始对话,由 DSP104 对话务信道话音分组编码,路由到微处理器 106 并由调制电路(未图示)调制以在空中传输。在用微处理器 106 中的编码逻辑(未图示)调制之前也可以对分组进行编码。然后,通过天线 114 在空中发送分组。将从对话的另一侧接收到的话音分组从天线 114 路由到解调电路(未图示)。用微处理器 106 中的解码逻辑对解调分组进行解码。于是,将分组从微处理器 106 路由到 DSP104,该矢量未量化所述分组。于是,向 D/A110 提供分组以转换成模拟话音信号。向扩音器 116 提供模拟话音信号,所述扩音器产生说话者可听到的字。

在一个实施例中,ASIC200 包括声码器 202 以对 VR 执行前端处理。声码器 202 耦合到分组缓冲器 204。分组缓冲器 204 耦合到处理器 206。处理器 206 耦合到非易失存储媒体 208。例如,ASIC200 可位于无线电话听筒中。声码器 202 是可变速 13kbps CELP 声码器,如在上述美国专利号 5,414,796 中所述,但是同样可以是 8kbps CELP 声码器或 CDMA EVRC 声码器。处理器 206 最好是微处理器,但是也可以是任何传统形式的处理器、控制器或状态机。非易失存储媒体 208 最好与快闪存储器一起实施,但是另一方面,可以与任何传统形式的可写、非易失存储器一起实施。分组缓冲器 204 是传统缓冲器元件。

向声码器 202 提供数字化话音采样 $s(n)$ 。声码器 202 用在上述美国专利 5,414,796 中所述的已知声编码技术对话音采样 $s(n)$ 进行编码。在所述的特定实施例中,声码器 202 每帧产生 128 个字节,其中每帧有 20 毫秒长。向分组缓冲器 204 提供每帧,即 128 个字节。微处理器 206 每帧从分组缓冲器 204 读取 128 个字节(即一个分组)。如果在 VR 解码器(未图示)中需要端点,那么微处理器 206 根据接收到的数据分组检测说话的起点和终点。然后,处理器 206 计算动态声特征(诸如时间导数、第二时间导数等等);通过选择相关帧来压缩声特征;并且量化静止和动态的声特征。在上述 Lawrence Rabiner & Biing-Hwang

Juang 所著的话音识别基础(Fundamentals of Speech Recognition(1993))中描述了端点检测、动态声特征导数、模式压缩和模式量化的多种已知方法。微处理器 206 访问快闪存储器 208、将存储器的内容与从分组缓冲器 204 中读取的每个分组相比较并根据已知的 (在 Rabiner & Juang 141-241 的文中所述) 模式识别技术进行字判定。

在所述的特定实施例中, 如果声码器 202 以全速编码, 那么产生每帧 35 个字节, 而 128 字节中的剩余部分可提供给声码器诊断数据。如果声码器 202 以半速编码, 那么产生每帧 17 个字节, 而 128 字节中的剩余部分可提供给声码器诊断数据。如果声码器 202 以四分之一速率编码, 那么产生每帧 7 个字节, 其中对于声码器诊断数据可获得 128 个字节的剩余部分。如果声码器 202 以八分之一速率编码, 那么产生每帧 3 个字节, 其中对于声码器诊断数据可获得 128 个字节的剩余部分。有利的是, 在话音识别期间, 声码器 202 以半速编码。因此, 对于声码器诊断数据可获得 111 个字节。

在所述的特定实施例中, 诊断数据需要 51 个字节, 而与声码器 202 的编码速率无关。如果声码器 202 以全速编码, 那么对于声码器诊断数据需要 40 个附加字节。如果编码速率是半速, 那么对于声码器诊断数据只需要 16 个附加字节。如果编码速率是四分之一速率, 那么对于声码器诊断数据只需要 5 个附加字节。而且, 如果编码是八分之一速率, 那么对于声码器诊断数据, 只需要一个附加字节。

虽然声码器分组构成量化参数(例如, 在半速下, 17 个声码器字节是量化的), 但是声码器诊断数据字节包含非量化的(unquantized)参数(例如, 在半速下, 声码器诊断数据高达 $128-17=111$ 字节是非量化的)。此外, 虽然声码器诊断数据没有包括所有声码器数据, 但是声码器诊断数据包括了与 VR 相关的所有声码器数据。因此, 声码器诊断数据可用于 VR 目的。这提供了显著的好处, 如下所述。

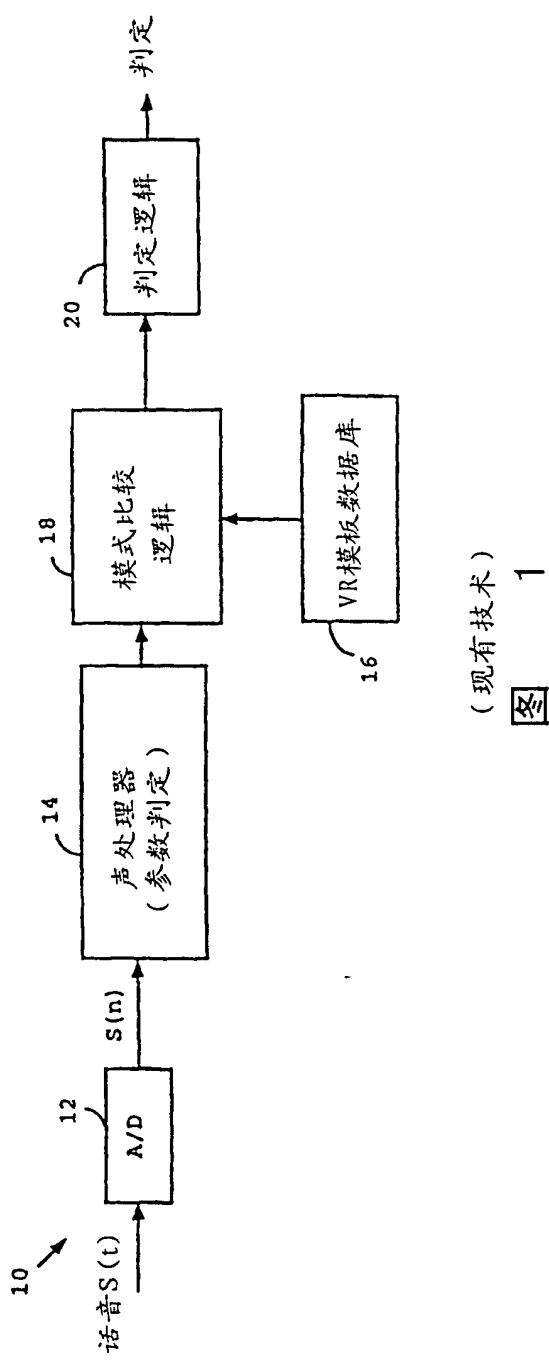
从声码器分组获得相关的 VR 参数需要所述分组首先是解量化的。例如, 在声码器 202 中, 线性预测编码(LPC)系数被转换成线状谱对(LSP)频率。每帧 10 个 LSP 频率是被量化分组成常规声码器分组的矢量。在微处理器 206 中, 以帧为基础对接收到的声码器分组进行解封装, 并对它们解量化, 从而获得解量化的 LSP 频率。然后, 将解量化 LSP 频率转换成 LPC 系数。然后, 将 LPC 系数转换成倒谱系数, 它们是对数幅度频谱(magnitude spectrum)的傅利叶变化

表示的系数。通过对数字化语音采样 $s(n)$ 的傅利叶变化系数的对数求反傅利叶变换，获得倒谱系数。在上述美国专利 5,414,796 中详细描述将 LSP 频率转换成 LPC 系数再转换成倒谱系数的技术以及将 LPC 系数转换成 LSP 系数再转换成声码器分组的技术。

以上述方法使来自声码器分组的数据矢量解量化或“解压缩”会导致计算量很大，需要大量处理容量。此外，需要大的 LSP 矢量量化(VQ)电码本(codebook)，而微处理器 206 必需将它永久存储在快闪存储器 208 中。在上述实施例中，由于将声码器诊断数据用于 VR 而不是解量化声码器分组用于 VR，所以有利地利用了必需的计算能力和存储容量。

于是，已经描述了一种新颖的经改进的分布式语音识别系统。熟悉本技术领域的人员应理解可以通过数字信号处理器(DSP)、特定用途集成电路(ASIC)、分立门电路或晶体管逻辑、分立硬件元件(诸如，寄存器和 FIFO)、执行一组固件指令的处理器或任何传统可编程软件模块和处理器，来实施或执行结合这里所述的实施例描述的各种说明性逻辑块和算法步骤。处理器最好是微处理器，但是另一方面，处理器可以是任何传统的处理器、控制器、微处理器或状态机。软件模块可以位于 RAM 存储器、快闪存储器、寄存器或在现有技术中已知的任何形式的可写存储媒体。熟悉本技术领域的人员应理解，最好用电压、电流、电磁波、磁场或微粒、光场或微粒或它们的任意组合来表示在上述说明中提到的数据、指令、命令、信息、信号、位、码元和码片。

已经示出和描述了本发明的较佳实施例。对于熟悉本技术领域的人员而言，可对这里所述的实施例进行各种变换而不偏离本发明的构思和范围是显而易见的。因此，应由所附权利要求来限定本发明。



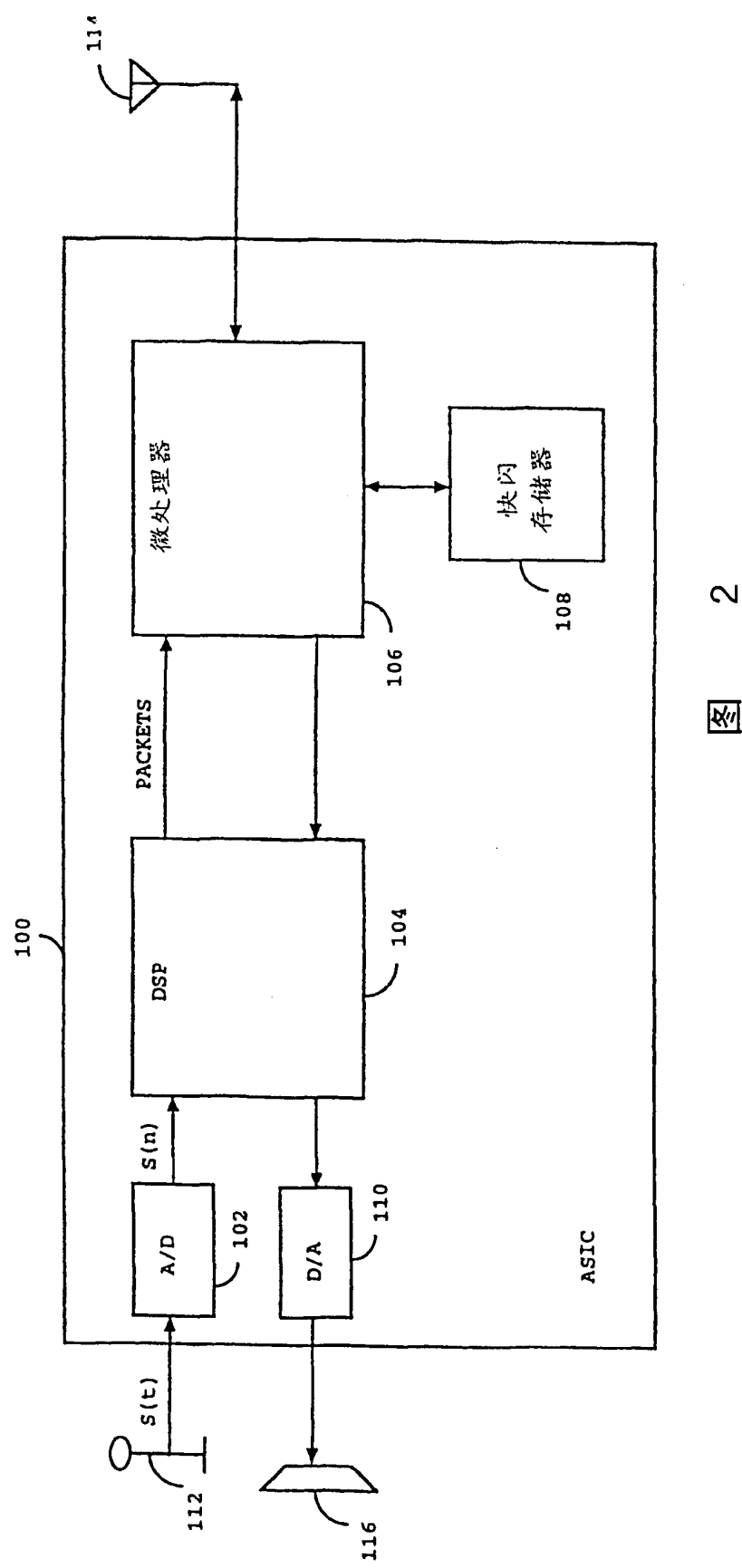


图 2

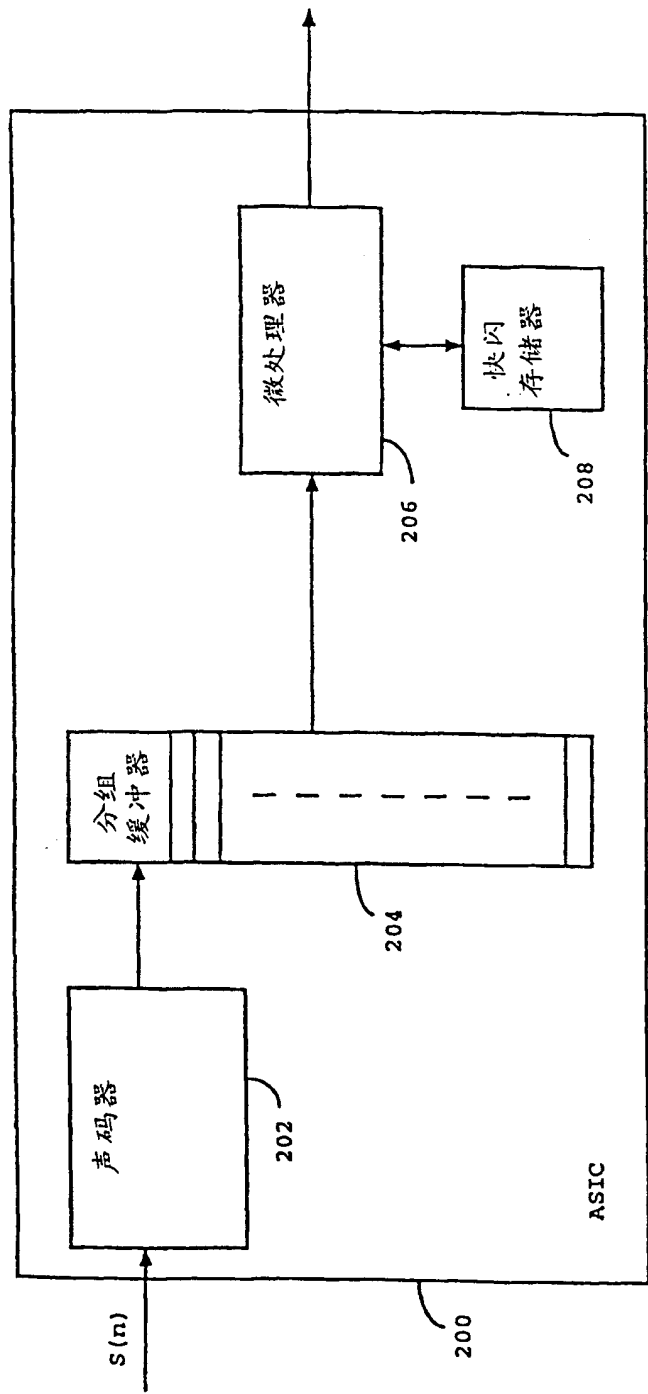


图 3