



(12) 发明专利

(10) 授权公告号 CN 101390155 B

(45) 授权公告日 2012.08.15

(21) 申请号 200780006100.3

(22) 申请日 2007.02.06

(30) 优先权数据

11/358,001 2006.02.21 US

(85) PCT申请进入国家阶段日

2008.08.20

(86) PCT申请的申请数据

PCT/US2007/061707 2007.02.06

(87) PCT申请的公布数据

W02007/098316 EN 2007.08.30

(73) 专利权人 索尼电脑娱乐公司

地址 日本东京都

(72) 发明人 R·陈

(74) 专利代理机构 中国专利代理(香港)有限公

司 72001

代理人 柯广华 张志醒

(51) Int. Cl.

G10L 15/06 (2006.01)

(56) 对比文件

CN 1585967 A, 2005.02.23,

CN 1441948 A, 2003.09.10,

US 6442519 B1, 2002.08.27,

EP 1607939 A1, 2005.12.21,

Cherif, A.. Pitch and formants
extraction algorithm. 《Electronics,
Circuits and Systems, 2000. ICECS 2000. The
7th IEEE International Conference on》. 2000,
第1卷 595-598.

W. H. Abdulla, N. K. Kasabov. Improving
speech recognition performance through
gender separation. 《Proceedings of the
Fifth Biannual Conference on Artificial
Neural Networks and Expert Systems
(ANNES2001)》. 2001, 218-222.

审查员 贾杨

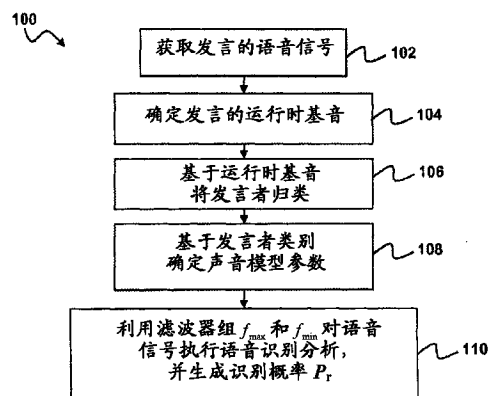
权利要求书 2 页 说明书 7 页 附图 1 页

(54) 发明名称

发言者适配的语音识别和利用基音的注册

(57) 摘要

公开语音识别方法和系统。获取发言者的发言的语音信号。根据发言的语音信号确定运行时基音。基于运行时基音将发言者归类,并基于发言者的类别调整一个或多个声音模型参数。可在所述识别期间的任何时间瞬间进行参数调整。然后基于声音模型来执行对发言的语音识别分析。



1. 一种用于语音识别的方法,所述方法包括:

获取发言者的发言的语音信号;

根据所述发言的语音信号确定运行时基音;

基于所述运行时基音将所述发言者归类;

基于所述发言者的类别调整执行所述语音识别分析时所用的滤波器组的最大频率 $f_{\max}$ 和最小频率 $f_{\min}$;以及

基于所述最大频率 $f_{\max}$ 和最小频率 $f_{\min}$ 来执行对所述发言的语音识别分析。

2. 如权利要求 1 所述的方法,其中确定所述运行时基音包括通过如下公式确定时间 t 处的移动平均基音 $p_{\text{avg}}(t)$:

$$p_{\text{avg}}(t) = \frac{1}{NP} \sum_i p(t_i)$$

行的。

3. 如权利要求 2 所述的方法,其中所述基音 $p(t_i)$ 中的每一个基音的概率 $\text{prob}(\text{pitch})$ 都高于预定阈值,其中 $\text{prob}(\text{pitch}) = \frac{\text{correlation}(\frac{1}{\text{pitch}})}{\text{correlation}(0)}$, 其中

$\text{correlation}(t) = \sum_i \text{signal}(t+i)\text{signal}(i)$ 是分析语音信号的相关性。

4. 如权利要求 1 所述的方法,其中确定所述运行时基音包括如下类型的计算:

$p_{\text{run}}(t) = c \cdot p_{\text{run}}(t-1) + (1-c) \cdot p(t)$, 其中 c 是 0 与 1 之间的常数, $p_{\text{run}}(t)$ 和 $p_{\text{run}}(t-1)$ 分别是时间 t 和 $t-1$ 处的运行时基音值, $p(t)$ 是时间 t 处的当前基音值;其中在 $p(t)$ 的基音概率高于预定阈值时使用该公式。

5. 如权利要求 1 所述的方法,其中将所述发言者归类包括确定所述发言者的年龄和/或性别。

6. 如权利要求 5 所述的方法,其中确定所述发言者的年龄和/或性别包括:确定所述运行时基音是否落在一个范围内,其中所述范围取决于发言者的年龄和/或性别。

7. 如权利要求 5 所述的方法,其中确定所述发言者的年龄和/或性别包括:根据所述基音确定所述发言者是男性、女性还是小孩发言者。

8. 如权利要求 1 所述的方法,还包括:存储所述发言者类别和/或基于所述发言者的类别的所述最大频率 $f_{\max}$ 和最小频率 $f_{\min}$, 并将所述发言者的所述发言者类别和/或基于所述发言者的类别的所述最大频率 $f_{\max}$ 和最小频率 $f_{\min}$ 与特定发言者关联。

9. 如权利要求 8 所述的方法,还包括:在对发言者进行后续语音识别分析期间,使用所存储的发言者类别和/或基于所述发言者的类别的所述最大频率 $f_{\max}$ 和最小频率 $f_{\min}$ 。

10. 一种语音识别系统,包括:

适于获取语音信号的接口;

耦合到所述接口的一个或多个处理器,其中所述一个或多个处理器被配置成:

获取发言者的发言的语音信号;

根据所述发言的语音信号确定运行时基音;

基于所述运行时基音将所述发言者归类;

基于所述发言者的类别调整执行所述语音识别分析时所用的滤波器组的最大频率 $f_{\max}$ 和最小频率 $f_{\min}$; 以及

基于所述最大频率 $f_{\max}$ 和最小频率 $f_{\min}$ 来执行对所述发言的语音识别分析。

发言者适配的语音识别和利用基音的注册

[0001] 相关申请的交叉引用

[0002] 本申请涉及与本申请同一日提交的 Ruxin Chen 的共同转让的美国专利申请 11/358, 272, 标题为“具有并行性别和年龄规范化的语音识别”(VOICE RECOGNITION WITH PARALLEL GENDER AND AGE NORMALIZATION), 其全部公开通过引用结合于本文。

技术领域

[0003] 本申请涉及语音识别, 并且更具体来说涉及适于基于基音的发言者的语音识别系统。

背景技术

[0004] 语音和语言识别技术使得配备声音输入源(例如麦克风)的计算机和其他电子装置能够解释人的语言, 例如用于抄录或作为一种与计算机交互的备选方法。正在开发语言识别软件, 以在消费电子装置(例如移动电话、游戏平台、个人计算机和个人数字助理)中使用。在典型的语言识别算法中, 表示人语言的时域信号被分成多个时间窗口, 每个窗口通过例如快速傅立叶变换(FFT)转换到频域信号。然后通过采用频谱域信号的对数并且再执行另一个 FFT 来压缩此频域或频谱域信号。根据压缩的频谱(称为倒谱), 可以使用统计模型来确定该信号所表示的语言内的音位和上下文。倒谱(cepstrum)可以视为有关语言信号内的不同频带中的速率变化的信息。对于语言识别应用, 常常首先使用 Mel 频带来变换频谱。其结果称为 Mel 频率倒谱系数或 MFCC。可以根据如下公式将以赫兹为单位的频率 f (每秒的周期数) 转换成 mel 频率 m : $m = (1127.01048\text{Hz}) \log_e(1+f/700)$ 。相似地, 可以使用 $f = (700\text{Hz}) (e^{m/1127.01048} - 1)$ 将 mel 频率 m 转换成以赫兹为单位的频率 f 。

[0005] 在语音识别中, 常常使用一组三角形滤波器函数来对频谱滤波。该滤波器函数将频谱分成一组部分重叠的频带, 该组频带位于最小频率 $f_{\min}$ 与最大频率 $f_{\max}$ 之间。每个滤波器函数以感兴趣的频率范围内的特定频率为中心。当转换到 mel 频率标度时, 每个滤波器函数可以表示为一组 mel 滤波器组, 其中每个 mel 滤波器组 MFB_i 由如下公式给出:

[0006]
$$\text{MFB}_i = \left(\frac{mf - mf_{\min}}{mf_{\max} - mf_{\min}} \right)_i$$
 其中索引 i 指代滤波器组的编号, 以及 $mf_{\min}$ 和 $mf_{\max}$ 是对应于 $f_{\min}$ 和 $f_{\max}$ 的 mel 频率。

[0007] $f_{\min}$ 和 $f_{\max}$ 的选择决定语音识别算法所使用的滤波器组。通常, $f_{\min}$ 和 $f_{\max}$ 由正在使用的语音识别模型来固定。语音识别的一个问题在于, 不同发言者具有不同的声道长度并产生具有相应不同频率范围的语音信号。为了对此进行补偿, 语音识别系统可以在滤波之前执行语音信号的声道规范化。举例来说, 规范化可以使用这种类型的函数:

[0008]
$$f' = f + \frac{1}{\pi \arctan \alpha \left(\frac{\sin(2\pi f)}{1 - \alpha \cos(2\pi f)} \right)}$$

[0009] 其中 f 是规范化的频率, 以及 α 是调整规范化函数的曲率的参数。

[0010] 具有 N 个不同 mel 频带的语言信号的分量可以表示为具有 N 个分量的矢量 A。矢量 A 的每个分量是语言信号的 mel 频率系数。矢量 A 的规范化通常包括如下类型的矩阵变换：

[0011] $F' = [M] \cdot F + B$, 其中 [M] 是由如下公式给出的 $N \times N$ 矩阵：

$$[0012] \quad [M] = \begin{bmatrix} M_{11} & M_{12} & \dots & M_{1N} \\ M_{21} & M_{22} & \dots & M_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{N1} & M_{N2} & \dots & M_{NN} \end{bmatrix}$$

[0013] 以及 B 是如下公式给出的偏置矢量：

$$[0014] \quad B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{bmatrix},$$

[0015] F' 和 F 是如下形式的矢量：

$$[0016] \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_N \end{bmatrix}, F' = \begin{bmatrix} F'_1 \\ F'_2 \\ \vdots \\ F'_N \end{bmatrix}$$

[0017] 其中离线计算矩阵系数 M_{ij} 和矢量分量 B_i , 以使 HMM 系统中观察的语言序列的概率最大。通常, 对于给定的帧和给定的特征 F' , 按如下高斯函数计算观察的概率：

$Gaussian_k(F'_1 \dots F'_N) = \frac{1}{\sqrt{\sigma_k}} \exp\left(-\sum_i \frac{(F'_i - \mu_k)^2}{2 \cdot \sigma_k^2}\right)$ 规范化的矢量 F' 的每个分量是规范化的语言信号的 mel 频率分量。

[0018] 众所周知, 男性和女性发言者产生由不同 mel 频率系数 (MFCC) 表征的语音信号。在现有技术中, 语音识别系统已经使用训练以在发言者是男性还是女性之间进行鉴别, 并基于发言者是男性还是女性来调整语音识别中所用的声音模型。通常, 通过让多个 (例如 10 个) 男性发言者和相等数量的女性发言者说相同的词以产生语音样本来训练声音模型。将基于语音样本的特征分析与用于语音识别的超模型组合在一起。

[0019] 上面的规范化的主要缺点在于, 矢量 F 可能具有多达 40 个分量。因此, 矩阵 [M] 可能具有多达 1600 个系数。这种大量系数的计算可能花费太长时间来调整语音识别算法。

[0020] 而且, 因为现有技术的语音识别系统和方法使用 $f_{\min}$ 、 $f_{\max}$ 、 $mf_{\min}$ 和 $mf_{\max}$ 的固定值来进行滤波和规范化, 所以它们并未充分地考虑到发言者之间声道长度方面的变化。因此, 语音识别精确度可能差于最优情况。因此, 需要一种克服此类缺点的语音识别系统和方法。

发明内容

[0021] 通过本发明涉及语音识别方法和系统的实施例克服了与现有技术关联的缺点。根据本发明的实施例, 获取发言者的发言的语音信号。根据发言的语音信号确定运行时基音。基于运行时基音将发言者归类, 并基于发言者的类别调整一个或多个声音模型参数。然后基于声音模型参数来执行对发言的语音识别分析。

附图说明

[0022] 通过参考结合附图的下文详细描述可以容易地理解本发明的技术, 其中:

[0023] 图 1 是图示根据本发明实施例的语音识别算法的流程图。

[0024] 图 2 是图示根据本发明实施例的语音识别系统的框图。

具体实施方式

[0025] 虽然出于说明的目的, 下文详细描述包含许多特定细节, 但是本领域技术人员将认识到对下文细节的许多变化和备选方案均在本发明的范围内。因此, 下文描述的本发明的实施例是在不失一般性且不对要求权利的本发明施加任何限制的前提下提出的。

[0026] 根据本发明实施例, 语音识别方法 100 可以如图 1A 所示的进行。在 102, 获取来自发言者的发言的语音信号。可以采用任何常规方式来获取语音信号, 例如使用麦克风, 并使用将语音信号制作成数字格式的波形数字转换器。可以通过以大于工作特征分析频率的采样频率对语音信号进行过采样 (over-sampling) 来获取语音信号。具体来说, 采样频率可以大于训练时间语言采样率。通过非限制性实例, 如果通过 12 千赫的工作特征分析频率来表征语音信号, 则可以采用例如 16-22 千赫的采样频率来对该信号采样。

[0027] 在 104, 确定发言的运行时基音值 p_{run} 。有许多方法来确定运行时基音 p_{run} 。例如, p_{run} 可以是按如下公式在包括时间 t 的给定时间窗口上计算的移动平均基音 $p_{avg}(t)$:

[0028] (公式 1)
$$p_{avg}(t) = \frac{1}{NP} \sum_i p(t_i),$$

[0029] 其中求和是对在基音概率高于预定阈值的时间窗口期间在时间 $t_i = \{t-(NP-1), t-(NP-2), \dots, t\}$ 处所取的 NP 个基音测量值进行的。计算基音概率的一种简单方式是

[0030]
$$prob(pitch) = \frac{correlation(\frac{1}{pitch})}{correlation(0)},$$
 其中 $correlation(t) = \sum_i signal(t+i)signal(i)$ 是分析语言信号的相

关性。或者, 可以通过例如如下公式将运行时基音 p_{run} 与当前基音相关:

[0031] [0031] (公式 2) $p_{run}(t) = c \cdot p_{run}(t-1) + (1-c) \cdot p(t), \text{ for } t > 0$

[0032] 以及 $p_{run}(0) = p(0), \text{ for } t = 0$

[0033] [0033] 其中 c 是 0 与 1 之间的常数, $p(t)$ 是时间 t 处的基音的当前值。常数 c 的值与窗口大小相关。例如, $c = 0$ 的值对应于无窗口 (其中 $p_{run}(t) = p(t)$), $c = 1$ 的值对应于无穷窗口 (其中 $p_{run}(t) = p_{run}(t-1)$)。注意, 对于 $t > 0$ 的值, t 之前的时间的基音值构成运行时基音 $p_{run}(t)$ 的值。这可以在 $c = 0.6$ 的情况中以数值示例说明。在此情况中, 公式 2 得到:

[0034] $p_{run}(0) = p(0)$

[0035] [0035] $p_{run}(1) = 0.6 \cdot p_{run}(0) + (1-c) \cdot p(1) = 0.6 \cdot p(0) + 0.4 \cdot p(1)$

[0036] $p_{run}(2) = 0.6 \cdot p_{run}(1) + (1-c) \cdot p(2) = 0.6 \cdot (0.6 \cdot p(0) + 0.4 \cdot p(1)) + 0.4 \cdot p(2)$

[0037] [0037] 在本发明的一些实施例中, 如果基音概率高于某个阈值 (例如高于约 0.4), 则可以根据公式 2 来计算 $p_{run}(t)$ 。

[0038] 举例来说, 图 1A 的 106 处执行的发言者归类可以基于发言者的年龄和 / 或性别来进行。例如, 根据训练数据, 可以确定男性、女性和小孩发言者的平均基音落在不同的范围内。可以由语音信号的当前基音所落在的基音范围来将发言者归类。举例来说, 成年男性

发言者具有约 120Hz 与约 160Hz 之间的平均基音,成年女性发言者具有约 180Hz 与约 220Hz 之间的平均基音,以及小孩发言者具有大于约 220Hz 的平均基音。如果当前基音是 190Hz,则发言者被归类为女性发言者。在这些情况的任何一种中,可以将发言者的平均基音作为矢量 F 中的特征来包括。

[0039] 一旦将发言者归类,则可以按 108 所指示的相应地选择声音模型的参数。然后在 110,在语音识别分析中使用这些参数。参数的选择取决于语音识别分析中所用的声音模型的类型。例如,语音识别分析可以使用一组滤波器函数来对语音信号滤波。滤波器函数(例如,三角形滤波器函数)将频谱分成一组部分重叠的频带。每个语音识别分析使用由不同的最大频率 $f_{\max}$ 和不同的最小频率 $f_{\min}$ 定义的滤波器组。 $f_{\max}$ 和 $f_{\min}$ 可以是赫兹标度的频率或 mel 标度的基音。最大频率 $f_{\max}$ 指滤波器组的频率范围的上限,而最小频率 $f_{\min}$ 指滤波器组的频率范围的下限。参数 $f_{\min}$ 和 $f_{\max}$ 的值可以在语音识别分析期间在任何时间瞬间(例如语音识别分析期间的任何时间窗口)动态地进行调整。语音识别分析产生一个或多个语言单元的识别的识别概率 P_r 。语言单元可以是短语、词或词的子单元(例如音位)。

[0040] 举例来说,一旦发言者被归类为男人、女人或小孩,则可以相应地选择用于对发言进行语音识别分析的值 $f_{\min}$ 和 $f_{\max}$ 。例如,如果假定发言者是男人,则 $f_{\max}$ 可以约为 70Hz 而 $f_{\min}$ 可以约为 3800Hz。如果假定发言者是女人,则 $f_{\max}$ 可以约为 70Hz 而 $f_{\min}$ 可以约为 4200Hz。如果假定发言者是小孩,则 $f_{\max}$ 可以约为 90Hz 而 $f_{\min}$ 可以约为 4400Hz。

[0041] 在 110,基于调整的模型参数从对发言的语音分析得到识别概率 P_r 。举例来说且不失一般性地,语音识别分析可以使用隐藏的马尔可夫模型(HMM)来确定给定语音信号中的语言单元。语言单元可以是词、双词组合或子词单元(例如音位等)。HMM 可以由如下项来表征:

[0042] L ,表示系统的多个可能状态;

[0043] M ,表示系统中存在的高斯的总数;

[0044] N ,表示给定时间处的不同可观察特征的数量;这些特征可以是语言信号的频谱上的(即频域)或时间上的(时域)特征;

[0045] $A = \{a_{ij}\}$,状态过渡概率分布,其中每个 a_{ij} 表示如果系统最初在时间 t 处于第 i 状态,则系统将在时间 $t+1$ 过渡到第 j 状态的概率;

[0046] $B = \{b_j(k)\}$,第 j 个状态的观察特征概率分布,其中每个 $b_j(k)$ 表示当系统处于第 j 个状态时第 k 个特征的观察值的概率分布;以及

[0047] $\pi = \{\pi_i\}$,最初状态分布,其中每个分量 π_i 表示系统将在某个最初时间处于第 i 个状态的概率。

[0048] 隐藏的马尔可夫模型可以应用于语音信号以解决一个或多个基本问题,包括:(1)从语音信号获得给定观察序列的概率;(2)给定观察序列,什么对应状态序列最好地解释该观察序列;以及(3)如何调整该组模型参数 A 、 B 、 π 以使得给定观察序列的概率最大化。

[0049] 将 HMM 应用于语言识别由例如 Lawrence Rabiner 在 1989 年 2 月 IEEE 学报第 2 期卷 77 的“语言识别中隐藏的马尔可夫模型和选择的应用的教程”(“A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”in Proceedings of the IEEE, Vol. 77, No. 2, February 1989,其通过引用基于所有目的结合于本文)中进行了详细描述。

[0050] 在 110 处实现的语音识别分析可以通过公知为音位 (phoneme) 的多个可识别模式来表征语言。这些音位的每一个可以分成多个部分,例如开始、中间和结尾部分。注意,中间部分通常是最稳定的,因为开始部分常常受到前一个音位的影响,而结尾部分受到后一个音位的影响。不同部分的音位由频域特征来表征,这些特征可以通过信号的适当统计分析来识别。统计模型常常使用高斯概率分布函数来预测特征的每个不同状态的概率,所述特征构成与不同音位的不同部分对应的信号部分。一个 HMM 状态可以包含一个或多个高斯。给定可能状态的特定高斯,例如第 k 个高斯可以由一组 N 个平均值 μ_{ki} 和方差 σ_{ki}^2 表示。在典型的语言识别算法中,确定给定时间窗口的哪个高斯是最大的一个。根据最大高斯,可以推断时间窗口的最大可能音位。

[0051] 举例来说,在 110 的语音识别分析可以分析时域信号以获得 N 个不同可观察信号特征 $x_0, \dots, x_n$, 其中 $n = N-1$ 。系统的观察的特征可以表示为具有分量 $x_0, \dots, x_n$ 的矢量。这些分量可以是给定观察的语言信号的频谱上、倒谱上或时间上的特征。

[0052] 举例来说且不作为本发明实施的限制,分量 $x_0, \dots, x_n$ 可以是在 102 处获得的语音信号的 mel 频率倒谱系数 (MFCC)。倒谱是如同作为信号来取分贝频谱的傅立叶变换 (FT) 的结果。时域语言信号的倒谱可以在言语上定义为该时域信号的傅立叶变换的对数 (具有去包裹的相位 (unwrapped phase)) 的傅立叶变换。时域信号 $S(t)$ 的倒谱可以在数学上表示为 $FT(\log(FT(S(t)) + j2\pi q))$, 其中 q 是将复对数函数的角度或虚数部分正确去包裹所需的整数。算法上为:倒谱可以由如下操作序列来生成:信号 $\rightarrow$ FT $\rightarrow$ 对数 $\rightarrow$ 相位去包裹 $\rightarrow$ FT $\rightarrow$ 倒谱。

[0053] 有复倒谱和实倒谱。实倒谱使用为实值定义的对数函数,同时复倒谱使用为复值定义的复对数函数。复倒谱具有有关初始频谱的量值和相位的信息,从而能够重构信号。实倒谱仅使用频谱的量值的信息。举例说明且不失一般性地,在 110 处实现的语音识别分析可以使用实倒谱。

[0054] 分量 $x_0, \dots, x_n$ 的组合的某些模式对应于语言单元 (例如词或短语) 或子单元 (例如音节、音位或词的其他子单元)。每个单元或子单元可以视为系统的状态。系统的给定高斯 (第 k 个高斯) 的概率密度函数 $f_k(x_0 \dots x_n)$ 可以是任何类型的概率密度函数,例如具有如下形式的高斯函数:

$$[0055] \quad f_k(x_0 \dots x_n) = \frac{1}{\sqrt{\delta_k}} \exp\left(-\sum_i \frac{(x_i - \mu_{ki})^2}{2 \cdot \sigma_{ki}^2}\right) \quad (1)$$

$$[0056] \quad \delta_k = \prod_i (2\pi \cdot \sigma_{ki}^2) \quad i=1 \dots N, k=1 \dots M.$$

[0057] 在上面的公式中,“i”是特征的索引,以及“k”是高斯的索引。在公式 (1) 中,下标 k 是高斯函数的索引。可能有数百至数十万个高斯为语言识别算法所使用。数量 μ_{ki} 是系统的第 k 个高斯的特征 x_i 的平均值。数量 σ_{ki}^2 是第 k 个高斯中的 x_i 的方差。可以将一个或多个高斯与一个或多个不同的状态关联。例如,可以有 L 个不同的状态,其包含系统中的总数 M 个高斯。数量 μ_{ki} 是在训练数据的所有时间窗口上属于 $f_k(x_0 \dots x_N)$ 的 x_i 的所有测量值的平均值,而 σ_{ki}^2 是用于计算 μ_{ki} 的对应测量值的方差。

[0058] 公式 (1) 可以计算每个高斯的概率以得到对应的识别概率 P_r 。根据具有最大概率的高斯,可以构建该特定时间窗口的最可能的状态、词、音位、字符等。注意,也可以使用给

定时间窗口的最可能的状态来帮助确定较早或较晚时间窗口的最可能的状态,因为它们可以确定其中发生该状态的上下文。

[0059] 根据本发明的实施例,按如上所述来工作的图 1A 或 1B 中所示类型的识别方法(例如,语音识别方法)可以作为信号处理设备 200 的一部分来实现,如图 2 所示。系统 200 可以包括处理器 201 和存储器 202(例如, RAM、DRAM、ROM 等)。此外,如果要想实现并行处理,则信号处理设备 200 可以具有多个处理器 201。存储器 202 包括按如上所述来配置的数据和代码。确切地说,存储器包括表示信号特征 204 的数据和概率函数 206,每个概率函数 206 可以包括代码、数据或代码与数据的某种组合。

[0060] 设备 200 还可以包括公知的支持功能 210,例如输入/输出(I/O)部件 211、电源(P/S)212、时钟(CLK)213 和高速缓存 214。设备 200 可以可选地包括用于存储程序和/或数据的海量存储装置 215,例如磁盘驱动器、CD-ROM 驱动器、磁带驱动器等。控制器还可以可选地包括用于帮助控制器 200 与用户之间交互的显示单元 216 和用户接口单元 218。显示单元 216 可以采用显示文本、数字、图形符号或图像的阴极射线管(CRT)或平板屏幕的形式。用户接口 218 可以包括键盘、鼠标、游戏杆、光笔或其他装置。此外,用户接口 218 可以包括用于对待分析信号进行直接捕获的麦克风、视频摄像机或其他信号换能装置。处理器 201、存储器 202 和系统 200 的其他组件可以经由图 2 所示的系统总线 220 彼此交换信号(例如代码指令和数据)。麦克风 222 可以通过 I/O 功能 211 耦合到设备 200。

[0061] 如本文所使用的,术语 I/O 一般指将数据传输到系统 200 或从系统 200 传输数据以及将数据传输到外设或从外设传输数据的任何程序、操作或装置。每次传输是来自一个装置的输出和对另一个装置的输入。外设包括只输入装置,例如键盘和鼠标;只输出装置,例如打印机;以及例如可以用作输入和输出装置的可写 CD-ROM 的装置。术语“外设”包括外部装置,例如鼠标、键盘、打印机、监视器、麦克风、摄像头、外部 Zip 驱动器或扫描仪,以及内部装置,例如 CD-ROM 驱动器、CD-R 驱动器或内部调制解调器或如闪速存储器读取器/写入器、硬盘的其他外设。

[0062] 处理器 201 可以执行程序 204 的程序代码指令中的对信号数据 206 和/或概率的信号识别,程序 204 由存储器 202 存储并检索并由处理器模块 201 执行。程序 203 的代码部分可以遵循多种不同编程语言(例如汇编、C++、JAVA 或多种其他语言)的任何一种。处理器模块 201 形成在执行例如程序代码 204 的程序时成为专用计算机的通用计算机。虽然本文中将程序代码 204 描述为以软件形式实现并在通用计算机上执行,但是本领域技术人员将认识到,作为备选方式,该任务管理方法可以使用硬件、例如专用集成电路(ASIC)或其他硬件电路来实现。同样地,应该理解本发明实施例可以全部或部分地以软件、硬件或二者的某种组合来实现。

[0063] 在一个实施例中,其中程序代码 204 可以包括一组处理器可读指令,该组处理器可读指令实现与图 1A 的方法 100 或图 1B 的方法 110 具有共有特征的方法。程序 204 一般可以包括一个或多个指令,该一个或多个指令指示处理器 201 获取发言者发言的语音信号;根据发言的语音信号确定运行时基音(runtime pitch);基于运行时基音将发言者归类;基于发言者的类别调整一个或多个声音模型参数;以及基于声音模型参数来执行对发言的语音识别分析。

[0064] 举例来说,程序 204 可以是较大的整体程序(例如计算机游戏的程序)的一部分。

在本发明的某些实施例中，程序代码 204 可以在提供语言样本的初始化阶段（例如在游戏开始时）提示发言者说某个词或短语（例如发言者的姓名）。根据此样本，程序 204 可以参考图 1 所描述的进行，以查找该发言者的最佳参数（例如 $f_{\min}$ 和 $f_{\max}$ ），并使用这些参数来运行 110 的语音识别。可以在程序得出结果之后保存这些参数，并在发言者再次使用该程序时使用。

[0065] 本发明的实施例提供更鲁棒且更精确的语言识别。在采用声音模型参数选择的语言识别的一个示例中，对单个女性发言者使用基于基音的发言者归类产生了 94.8% 的词精确度。不采用声音模型参数选择的常规语言识别算法，对于同一个女性发言者使用基于基音的发言者归类仅达到 86.3% 的词精确度。

[0066] 虽然上文是对本发明优选实施例的完整描述，但是还可以使用多种备选方案、修改和等效物。因此，本发明的范围应该不是参照上文描述来确定的，而是应该参照所附权利要求连同其等效物的完整范围来确定。可以将本文描述的任何特征（无论是否是优选的）与本文描述的任何其他特征（无论是否是优选的）组合。在下文的权利要求中，除非另行明确指示，否则不定冠词“一个”指该冠词后面为一个或多个项目的数量。所附权利要求不应解释为包括部件加功能的限制，除非在给定的权利要求中使用短语“用于... 的部件”来明确指出此类限制。

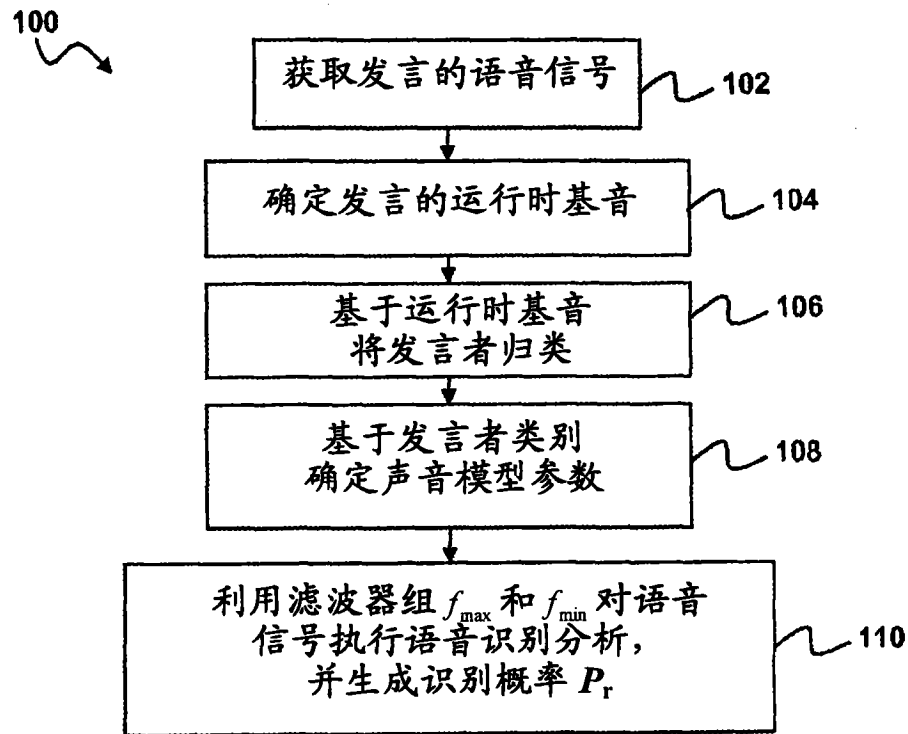


图 1

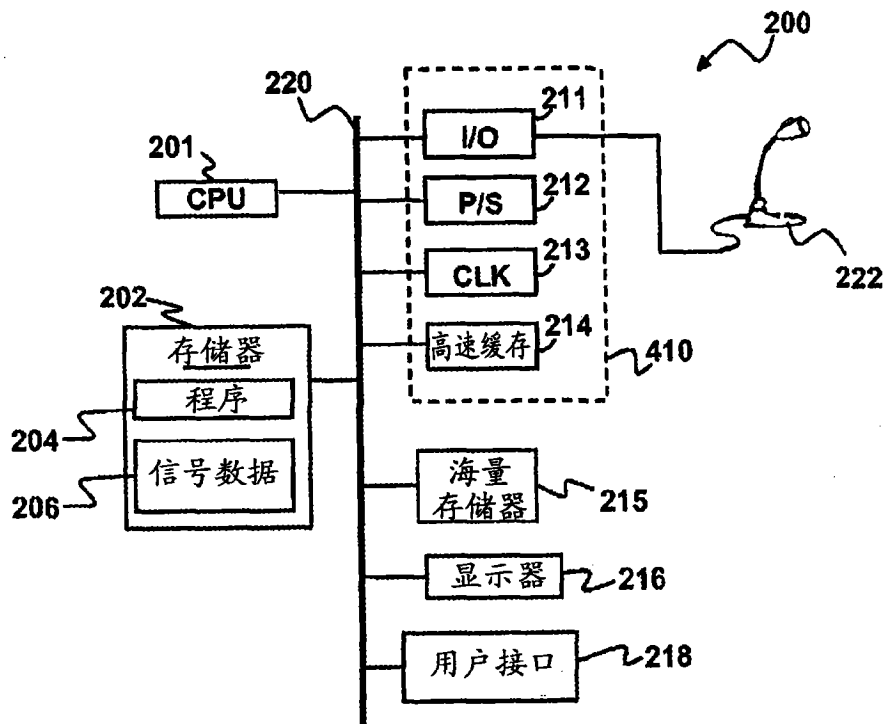


图 2