[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 11/00 (2006.01)
G10L 11/06 (2006.01)

[12] 发明专利说明书

专利号 ZL 200510128718.X

[45] 授权公告日 2010 年 1 月 27 日

[11] 授权公告号 CN 100585697C

[22] 申请日 2005.11.25
[21] 申请号 200510128718.X
[30] 优先权
[32] 2004.11.25 [33] KR [31] 10-2004-0097650
[73] 专利权人 LG 电子株式会社
地址 韩国首尔
[72] 发明人 金灿佑
[56] 参考文献
US2002/165713A1 2002.11.7
US2004/122667A1 2004.6.24
US6615170B1 2003.9.2

A SEMI-CONTINUOUS STATE TRANSITION PROBABILITYHMM-BASED VOICE ACTIVITY DETECTION. H. OThman. IEEE, Vol. 5. 2004
审查员 时 鹏
[74] 专利代理机构 上海专利商标事务所有限公司
代理人 张政权

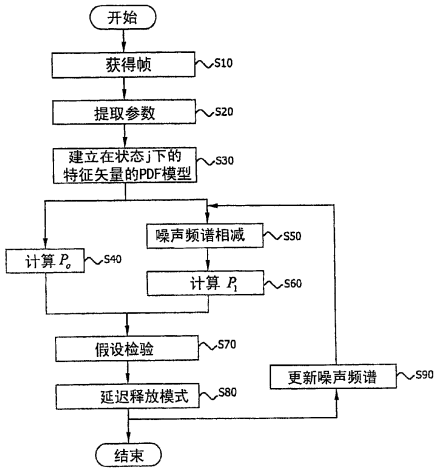
权利要求书 4 页 说明书 6 页 附图 2 页

[54] 发明名称

语音区别方法

[57] 摘要

一种语音区别方法，它包括把输入话音信号划分多个帧；从划分的帧中获得参数；使用获得的参数，为每个帧建立状态 j 的特征矢量的概率密度函数模型；从所建的 PDF 模型和获得的参数获得相应帧将是噪声帧的概率 P_0 以及相应帧将是语音帧的概率 P_1。进一步，使用获得的概率 P_0 和 P_1，执行假设检验，以确定相应的帧是噪声帧还是语音帧。



1．一种语音区别方法，该方法包含：

把输入话音信号划分为多个帧；

从划分的帧中获得参数；

使用获得的参数，为每个帧在状态 j 的特征矢量建立概率密度函数模型；

从所建的 PDF 模型和获得的参数中获得相应帧是噪声帧的每个状态的最大概率 P_0 以及相应帧是为语音帧的每个状态的最大概率 P_1；以及

使用获得的概率 P_0 和 P_1，执行假设检验，以确定相应的帧为噪声帧还是语音帧。

2．如权利要求 1 所述的方法，其特征在于，所述参数包含：

从帧中获得的语音特征矢量 0；

在状态 j 下第 k 个混合物的特征的均值矢量 m_{jk}；

在状态 j 下第 k 个混合物的权值矢量 c_{jk}；

在状态 j 下第 k 个混合物的协方差矩阵 C_{jk}；

一帧将是静音帧或噪声帧的的先验概率 $P(H_0)$；

一帧将是语音帧的先验概率 $P(H_1)$；

假设该帧是噪声帧，当前状态将是噪声帧的第 j 个状态的先验概率 $P(H_{0,j}|H_0)$；以及

假设该帧是语音帧，当前状态将为语音帧的第 j 个状态的先验概率 $P(H_{1,j}|H_1)$。

3．如权利要求 2 所述的方法，其特征在于，基于要求的性能、参数文件的大小以及实验获得的在状态和混合物的数量与所要求性能间的关系确定状态和混合物的数量。

4．如权利要求 1 所述的方法，其特征在于，使用包含收集并记录的实际语音和噪声的数据库来获得所述参数。

5．如权利要求 1 所述的方法，其特征在于，使用高斯混合物、log 凹函数或椭圆对称函数来建立所述概率密度函数的模型。

6．如权利要求 5 所述的方法，其特征在于，使用所述高斯混合物的所述概率密度函数用下列等式表示：

$$b_j(o)=\sum_{k=1}^{N_{mix}} c_{jk} N(o, m_{jk}, C_{jk})$$

。

7．如权利要求 1 所述的方法，其特征在于，由下列等式获得所述帧将是噪声帧的概率 P_0：

$$P_0=\max_j\left(b_j(o)\cdot P(H_{0,j}|H_0)\right)=\max_j\left(\sum_{k=1}^{N_{mix}} c_{jk} N(o, m_{jk}, C_{jk})\cdot P(H_{0,j}|H_0)\right)$$

。

8．如权利要求 1 所述的方法，其特征在于，由下列等式获得所述帧将是语音帧的概率 P_1：

$$P_1=\max_j\left(b_j(o)\cdot P(H_{1,j}|H_1)\right)=\max_j\left(\sum_{k=1}^{N_{mix}} c_{jk} N(o, m_{jk}, C_{jk})\cdot P(H_{1,j}|H_1)\right)$$

。

9．如权利要求 1 所述的方法，其特征在于，使用概率 P_0 和 P_1 以及选择的准则，所述假设检验确定相应的帧是语音帧还是噪声帧。

10．如权利要求 9 所述的方法，其特征在于，所述准则是 MAP（最大后验）准则、最大似然性（ML）极小极大准则、Neman-Pearson 检验、恒定虚警率检验中之一种。

11．如权利要求 10 所述的方法，其特征在于，所述 MAP 准则由下列等式定义：

$$\frac{P_0}{P_1}\begin{array}{c}H_0\\ >\\ <\\ H_1\end{array}\eta \quad , \quad \eta=\frac{P(H_1)}{P(H_0)}$$

。

12．如权利要求 1 所述的方法，其特征在于，所述方法进一步包含：

使用在获得概率 P_1 前先前获得的噪声频谱结果，有选择地在相应的帧上执行噪声频谱相减过程。

13．如权利要求 1 所述的方法，其特征在于，所述方法进一步包含：

在执行假设检验后有选择地应用延迟释放模式。

14．如权利要求 12 所述的方法，其特征在于，

当相应的帧被确定为噪声帧时，用确定的噪声帧的当前噪声频谱来更新噪声频谱相减过程。

15．一种用于区别语音的话音活动检波器，包括：

微处理器，配置成把输入话音信号划分成多个帧；

为这些划分的帧获得参数；

使用获得的参数模拟为每个在状态 j 下建立特征矢量的概率密度函数模型

从所建的 PDF 模型和获得的参数获得相应的帧是噪声帧的每个状态的最大概率 P_0 和相应的帧是语音帧的每个状态的最大概率 P_1；以及

使用获得的概率 P_0 和 P_1 执行假设检验以确定相应的帧是噪声帧还是语音帧。

16. 如权利要求 15 所述的话音活动检波器，其特征在于，所述参数包含：

从帧中获得的语音特征矢量 $\underline{o}$；

在状态 j 下第 k 个混合物的特征的均值矢量 m_{jk}；

在状态 j 下第 k 个混合物的权值矢量 c_{jk}；

在状态 j 下第 k 个混合物的协方差矩阵 C_{jk}；

一帧将是静音帧或噪声帧的的先验概率 $P(H_0)$；

一帧将是语音帧的先验概率 $P(H_1)$；

假设该帧是噪声帧，当前状态将是噪声帧的第 j 个状态的先验概率 $P(H_{0.j}|H_0)$；以及

假设该帧是语音真，当前状态将是语音帧的第 j 个状态的先验概率 $P(H_{1.j}|H_1)$。

17. 如权利要求 15 所述的话音活动检波器，其特征在于，使用所述高斯混合物建立所述概率密度函数模型用下列等式表示：

$$b_j(\underline{o}) = \sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk})$$

。

18. 如权利要求 15 所述的话音活动检波器，其特征在于，由下列等式获得将所述帧是噪声帧的概率 P_0：

$$P_0 = \max_j \left(b_j(\underline{o}) \bullet P(H_{0,j}|H_0)\right) = \max_j \left(\sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk}) \bullet P(H_{0,j}|H_0)\right)$$

。

19. 如权利要求 15 所述的话音活动检波器，其特征在于，由下列等式获得将所述帧是语音帧的概率 P_1：

$$P_1 = \max_j \left(b_j(\underline{o}) \bullet P(H_{1,j}|H_1)\right) = \max_j \left(\sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk}) \bullet P(H_{1,j}|H_1)\right)$$

。

20. 如权利要求 15 所述的话音活动检波器，其特征在于，使用概率 P_0 和

P_1 以及一选择的准则，确定相应的帧为语音帧还是噪声帧。

21. 如权利要求 20 所述的话音活动检波器，其特征在于，所述准则是 MAP（最大后验）准则、最大似然性（ML）极小极大准则、Neman-Pearson 检验、恒定虚警率检验中之一种。

22. 如权利要求 21 所述的话音活动检波器，其特征在于，所述 MAP 准则由下列等式定义：

$$\frac{P_0}{P_1} \underset{H_1}{\overset{H_0}{\gtrless}} \eta \quad , \quad \eta = \frac{P(H_1)}{P(H_0)} 。$$

23. 如权利要求 15 所述的话音活动检波器，其特征在于，

所述微处理器还配置成使用在获得概率 P_1 前先前获得的噪声频谱结果，有选择地在相应的帧上执行噪声频谱相减过程。

24. 如权利要求 23 所述的话音活动检波器，其特征在于，

所述微处理器还配置成当相应的帧被确定为噪声帧时，用确定的噪声帧的当前噪声频谱来更新所述噪声频谱相减过程。

语音区别方法

技术领域

本发明涉及语音检测方法，并且更为具体地，涉及有效地确定包括语音和噪声数据的输入话音信号中的语音和非语音（例如，噪声）部分的语音区别方法。

背景技术

先前的研究指出，两个人之间的一般电话交谈大约包括40%的语音和60%的静音。而且，噪声数据可以比用舒适的噪声生成（CNG）技术的语音数据更低的比特率编码。以不同的编码率对输入话音信号（包括噪声和语音数据）进行编码称为可变速率编码。此外，可变速率语音编码通常用于无线电话通信中。为了有效地完成可变速率的语音编码，用话音活动检波器（VAD）来确定语音部分和噪声部分。

在国际电信联盟（ITU-T）的电信标准部分提出的G.729标准中，可以获得如线谱密度（LSF）、全频带能量（E_f）、低频带能量（E_l）、零点交叉速率（ZC）等的输入信号的参数。也可以获得该信号的频谱失真（△S）。然后，获得的值与先前由实验结果确定的特定常量进行比较，以确定输入的信号的特定部分是语音部分还是噪声部分。

此外，在GSM（全球移动通信系统）网络中，当输入话音信号（包括噪声和语音）时，估计噪声的频谱，使用估计的频谱构造噪声抑制滤波器，且该输入的话音信号穿越噪声抑制滤波器。然后，计算该信号的能量，并把计算出的能量与预设的阈值进行比较，以确定特定部分是语音部分还是噪声部分。

上述方法要求多个不同的参数，并基于先前确定的经验数据，即，过去的数据确定输入信号的该特殊部分是语音部分还是噪声部分。然而，语音的特性对每个特定的人来说是非常不同的。例如，不同年龄的人的语音的特性，无论是男性还是女性等等，会改变语音的特性。因此，因为VAD使用先前确定的经验数据，故VAD不提供最佳的语音分析性能。

改善经验主义方法的另一种语音分析方法使用概率理论来确定输入信号

的特定部分是否为语音部分。然而，这种方法也是有缺点的，因为它不考虑基于任一特定谈话而具有各种频谱的噪声的不同特性。

发明内容

因此，本发明的一个目标是解决上述以及其他问题。

本发明的另一个目标是提供有效确定包括语音和噪声数据的输入话音信号中的语音和噪声部分的语音区别方法。

为了达到根据本发明的目的的这些以及其他优点，作为这里体现并广泛描述的，提供了一种语音区别方法。根据本发明的一个方面的语音检测方法包括把输入话音信号分为多个帧、从分开的帧中获得参数、使用获得的参数为每个帧在状态 j 下的特征矢量建立一概率密度函数模型、从所建的 PDF 模型和获得的参数中获得相应的帧为噪声帧的概率 P_0 和相应的帧为语音帧的概率 P_1。而且，使用获得的概率 P_0 和 P_1 完成假设检验以确定相应的帧是噪声帧还是语音帧。

根据本发明的另一个方面，提供了一种用于执行计算机指令的计算机程序产品，该计算机指令包括配置成把输入话音信号分成多个帧的第一计算机代码、配置成获得为这些分开的帧的参数的第二计算机代码、配置成使用获得的参数为每个在状态 j 的特征矢量建立概率密度函数模型的第三计算机代码、以及配置成从所建的 PDF 模型和获得的参数中获得相应的帧为噪声帧的概率 P_0 和相应的帧为语音帧的概率 P_1 的第四计算机代码。该计算机指令也包括配置成使用获得的概率 P_0 和 P_1 执行假设检验以确定相应的帧是噪声帧还是语音帧的第五计算机代码。

从此后给出的详细描述中，本发明的适用性的又一个范围将变得明显。然而，应该理解，详细描述和特定的例子尽管指出了本发明优选的实施例，但仅是为了说明，因为从这种详细描述中的各种变化和修改都在本发明的精神和范围之内，这对本发明的技术人员来说是显而易见的。

附图说明

从下面给出的详细描述及相应的附图中，本发明将变得更能全面理解。详细描述和相应的附图仅是为了说明，因此并非是本发明的限制，并且其中：

图 1 是显示根据本发明的一个实施例的语音区别方法的流程图；以及

图 2A 和图 2B 是显示完成的试验结果以分别确定许多状态和混和物的图表。

具体实施方式

现在，将对本发明优选的实施例做详细描述，附图示出其例子。

根据本发明的一方面的语音区别方法的算法使用下面两个假设：

H_0：为只包括噪声数据的噪声部分。

H_1：为包括语音和噪声的语音部分。

为了检验以上假设，执行自反（reflexive）算法，将参考图 1 显示的流程图讨论该算法。

参考图 1，输入话音信号被分为多个帧（S10）。在一个例子中，输入话音信号被分为 10 毫秒间隔的帧。进一步，当整个话音信号被分为 10 毫米间隔的帧时，每个帧的值被称为概率过程内的“状态”。

在输入信号被划分为多个帧后，从划分的帧（S20）中获得一组参数。这些参数包括，例如，从相应的帧中获得的语音特征矢量 0；在状态 j 的第 k 个混合物的特征的均值矢量 m_{jk}；在状态 j 的第 k 个混合物的权值矢量 c_{jk}；在状态 j 的第 k 个混合物的协方差矩阵 C_{jk}；一帧将对应于静音帧或噪声帧的先验概率 $P(H_0)$；一帧将对应于语音帧的的先验概率 $P(H_1)$；假设该帧包括静音，当前状态将为静音帧或噪声帧的第 j 个状态的当前状态的先验概率 $P(H_{0,j}|H_1)$；以及假设该语音帧包括语音，当前状态将为语音帧的第 j 个状态的先验概率 $P(H_{1,j}|H_1)$。

可通过训练过程获得上述参数，其中记录实际话音和噪声并将其存储在语音数据库内。由相应的应用、参数文件的大小以及试验获得的许多状态和性能要求间的关系确定要被分配给语音和噪声的状态数量。类似地确定混合物的数量。

例如，图 2A 和图 2B 是示出用于确定状态和混合物数量的试验结果的图表。具体地，图 2A 和图 2B 是分别显示根据状态和混合物的数量的语音区别速率的图表。如图 2A 所示，当状态数量过小或过大时，语音区别率降低。类似地，如图 2B 所示，当混合物的数量过小或过大时，语音区别率降低。因此，

使用试验过程来确定状态和混和物的数量。此外，可以使用各种参数估计技术来确定上述参数，如期望最大值算法（E-M 算法）。

进一步，参考图 1，在步骤（S20）提取参数后，由使用提取的参数的高斯混合物建立状态 j 的特征矢量的概率密度函数（PDF）模型（S30）。也可以使用 log 凹函数或椭圆对称函数来计算 PDF。

L.R.Rabiner 和 B-H.HWANG 所写的“Fundamentals of Speech Recoginition”（Englewood Cliffs, 新泽西. Prentice Hall,1993），以及由 S.E.Levinson、L.R.Rabiner 和 M.M.Sondhi 所写的“An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition（贝尔系统技术.J,1983 年 4 月）”中描述了使用高斯混合物的 PDF 方法，两者因此整体结合与此。因为该方法众所周知，故省略了详细描述。

此外，使用高斯混合物在状态 j 的特征矢量的 PDF 由下列等式表示：

$$b_j(\underline{o}) = \sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk})$$

这里，N 表示采样矢量的总数。

接着，使用计算出的 PDF 和其他参数获得概率 P_0 和 P_1。具体地，从提取的参数中获得对应帧为静音帧或噪声帧的概率 P_0（S40），以及从提取的参数中获得对应帧为语音帧的概率 P_1（S60）。进一步，计算概率 P_0 和 P_1，因为并不知道该帧是语音帧还是噪声帧。

进一步，可使用下列等式计算概率 P_0 和 P_1：

$$P_0 = \max_j (b_j(\underline{o}) \bullet P(H_{0,j} \mid H_0)) = \max_j \left(\sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk}) \bullet P(H_{0,j} \mid H_0) \right)$$

$$P_1 = \max_j (b_j(\underline{o}) \bullet P(H_{1,j} \mid H_1)) = \max_j \left(\sum_{k=1}^{N_{mix}} c_{jk} N(\underline{o}, \underline{m}_{jk}, C_{jk}) \bullet P(H_{1,j} \mid H_1) \right)$$

同样地，如图 1 所示，在计算概率 P_1 之前，在分开的帧上执行噪声频谱相减过程（S50）。相减技术使用先前获得的噪声频谱。

此外，在计算概率 P_0 和 P_1 后，执行假设检验（S70）。使用计算出的概率 P_0 和 P_1 及来自估计统计值标准的特定准则，用该假设检验来确定相应的帧

是噪声帧还是语音帧。例如，该准则可能为由以下等式定义的 MAP（最大后验）准则：

$$\frac{P_0}{P_1} \underset{H_1}{\overset{H_0}{\gtrless}} \eta \quad ，这里，\quad \eta = \frac{P(H_1)}{P(H_0)} 。$$

也可以使用其他准则，如最大似然性（ML）极小极大准则、Neman-Pearson 检验、CFAR（恒定虚警率）（Constant False Alarm Rate）检验等等。

然后，在假设检验后，应用延迟释放模式（Hang Over Scheme）（S80）。使用延迟释放模式来阻止低能量的声音，如“f”、“th”、“h”等等因其他高能量的声音被错误地确定为噪声，以及阻止中止声音，如“k”、“p”、“t”等等（开始为高能量后来为低能量的声音）在用低能量发音时被确定为静音。进一步，如果帧被确定为噪声帧，且该帧在被确定为语音帧的多个帧之间，则延迟释放模式任意决定该静音帧为语音帧，因为当考虑很小的 10 毫秒间隔的帧时，语音不会突然变为静音。

此外，如果应用延迟释放模式后，相应的帧被确定为噪声帧，则为确定的噪声帧计算噪声频谱。因此，根据本发明的一个实施例，可使用计算出的噪声频谱来升级步骤 S50 执行的噪声频谱相减过程（S90）。进一步，可有选择地执行分别在 S80 和 S50 的延迟释放模式和噪声频谱相减过程。即，这一个或两个步骤可省略。

正如迄今为止所述，在根据本发明的实施例的语音区别方法中，分别将语音和噪声（静音）部分作为状态处理，从而适合具有各种频谱的语音或噪声。同样，在数据库内集合的噪声数据上使用训练过程，以提供对不同类型噪声的有效响应。此外，在本发明中，因为由如 E-M 算法的方法可获得随机优化参数，故确定帧为语音帧还是噪声帧的过程得到改善。

进一步，也可通过在话音记录中只记录语音部分而不记录噪声部分，使用本发明来节省存储空间，或者本发明也可被用作有线或无线电话中为可变速率编码器的算法的一部分。

根据本发明的教义，使用传统的通用数字计算机或编程的微处理器可方便地实现本发明，这对本领域的技术人员而言是明显的。熟练的程序员根据本发

明的教义，可轻易地进行适当的软件编码，这对本领域的技术人员而言是明显的。本发明也可准备用由此互联传统计算机电路的适当网络的应用专用集成电路来实现，这对本领域的技术人员而言是明显的。

在通用数字计算机或微处理器上实现的本发明的任何部分包括计算机程序产品，该产品是包括能被用于对计算机编程以执行本发明的过程的指令的存储介质。该存储介质包括但不限于，包括软盘、光盘、CD-ROM、以及磁性光盘、ROM、EEPROM、磁卡或光卡的任何类型的磁盘，或者适于存储电子指令的任何类型的介质。

本发明可以许多形式实现，而不会脱离其精神或基本特性，也应该理解，除非另外指明，上述实施例不作为前面详细描述的限制，但应宽泛地被认为处在附加的权利要求的精神和范围内，并且因此所有的变化和修改都落入权利要求的界限和范围内，或者因此附加的权利要求也意图包含这种界限和范围的等价物。

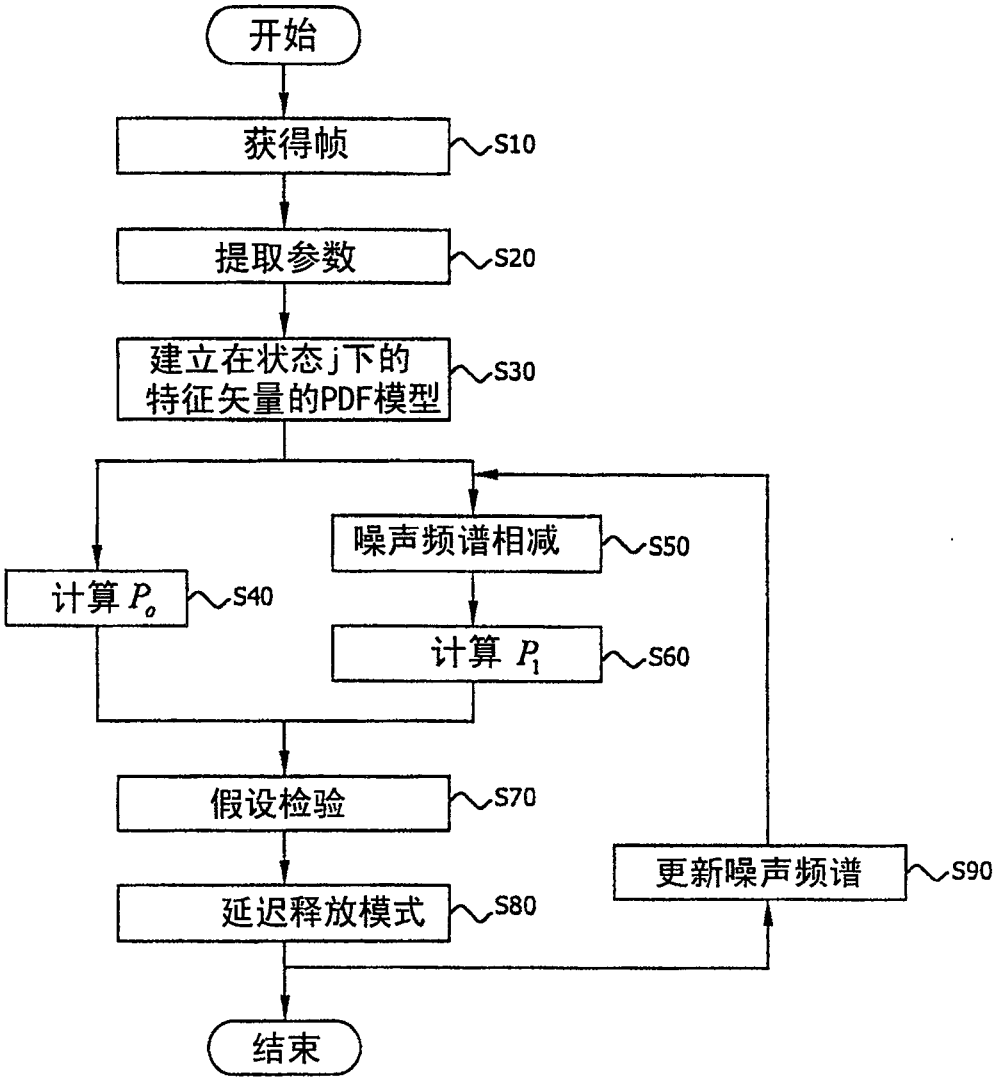

开始
获得帧
S10
提取参数
S20
建立在状态j下的
特征矢量的PDF模型
S30
噪声频谱相减
S50
计算 P_o
S40
计算 P_1
S60
假设检验
S70
更新噪声频谱
S90
延迟释放模式
S80
结束

图 1

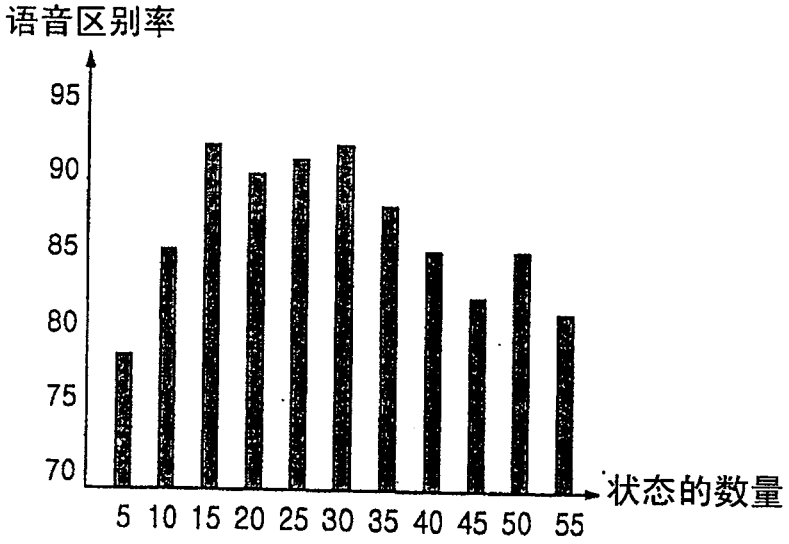

语音区别率
95
90
85
80
75
70
5 10 15 20 25 30 35 40 45 50 55
状态的数量

图 2A

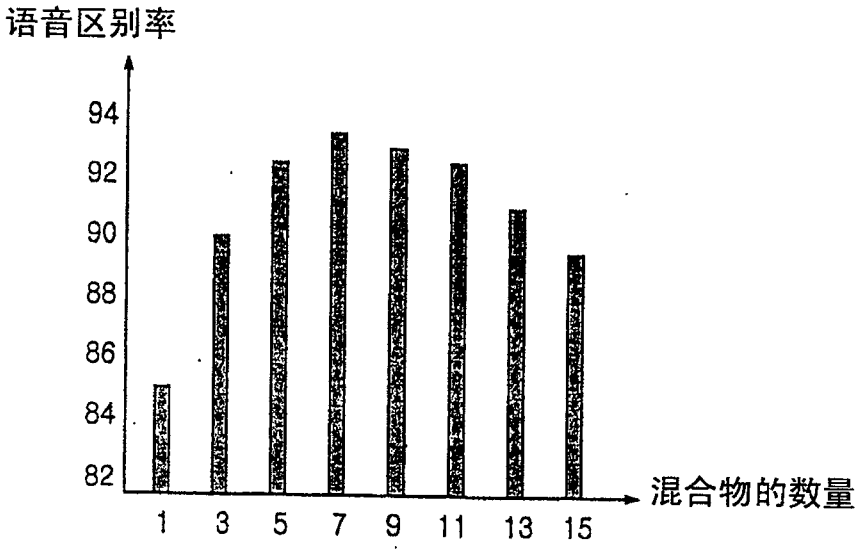

语音区别率
94
92
90
88
86
84
82
1 3 5 7 9 11 13 15
混合物的数量

图 2B