[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G10L 17/00
//G10L101：04

[12] 发明专利说明书

[21] ZL 专利号 00804893.2

[45] 授权公告日 2004 年 5 月 5 日

[11] 授权公告号 CN 1148720C

[22] 申请日 2000.2.25 [21] 申请号 00804893.2

[30] 优先权

[32] 1999. 3.11 [33] GB [31] 9905627.7

[32] 1999. 7. 2 [33] EP [31] 99305278.6

[86] 国际申请 PCT/GB2000/000660 2000.2.25

[87] 国际公布 WO00/054257 英 2000.9.14

[85] 进入国家阶段日期 2001.9.11

[71] 专利权人 英国电讯有限公司

地址 英国伦敦

[72] 发明人 西蒙·尼古拉斯·唐尼

审查员 杨 叁

[74] 专利代理机构 永新专利商标代理有限公司

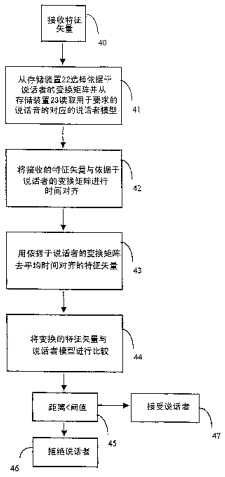
代理人 韩 宏

权利要求书 3 页 说明书 17 页 附图 7 页

[54] 发明名称 说话者识别

[57] 摘要

本发明涉及一种用于说话者识别的方法和设备。 在本发明中，在将从语音导出的特征矢量与一存储的参考模型进行比较之前，通过施加一与说话者相关的变换对这些特征矢量进行处理，该变换匹配于一特定的说话者的发声带的特性。 从具有与该变换所依据的说话者的特性不类似的特性的语音导出的特征通过该变换被严重地变形，而从具有与该变换所依据的说话者的特性类似的特性的语音导出的特征通过该变换所产生的变形非常小。



ISSN1008-4274

1、一种说话者识别方法，包括有步骤：

接收来自一未知的说话者的语音信号；

根据一变换对该接收的语音信号进行变换，该变换是与一特定的说话者相关联的；

将该变换的语音信号与一代表所述特定的说话者的模型进行比较；且

将依据于该未知的说话者是所述的特定说话者的似然性的一参数提供作为输出。

2、根据权利要求 1 的方法，其中该变换步骤包括有子步骤：

检测该接收的语音信号内的一语音开始点和一语音结束点；

生成从该接收的语音信号导出的一特征矢量序列；及

将对应于该检测的开始点和检测的结束点之间的语音信号的该特征矢量序列与用于所述特定说话者的一代表性的特征矢量序列相对准以使在被对准的特征矢量序列中的各特征矢量对应于该代表性的特征矢量序列中的一特征矢量。

3、根据权利要求 2 的方法，其中该变换步骤还包括有子步骤：用该代表性的特征矢量序列中的对应特征矢量对该被对准的特征矢量序列中的各特征矢量进行平均。

4、根据以上任一权利要求的方法，其中该模型是一隐藏的马克夫模型。

5、根据权利要求 4 的方法，其中该模型是一左至右隐藏的马克夫模型。

6、根据权利要求 5 的方法，当权利要求 4 是从属于权利要求 2 或权利要求 3 时，其中该代表性的特征矢量序列包括与隐藏的马克夫模型中的状态数量相同数量的特征矢量。

7、一种用于说话者识别的设备，包括有：

用于接收来自一未知的说话者的语音信号的接收装置；

用于存储多个说话者变换的说话者变换存储装置，各变换与多个说话者中对应的一个相关联；

用于存储多个说话者模型的说话者模型存储装置，各说话者模型与所述多个说话者中对应的一个相关联；

与该接收装置和说话者变换存储装置耦合的变换装置，被配置用于根据一选择的说话者变换对该接收的语音信号进行变换；

耦合至该变换装置和说话者模型存储装置的比较装置，被配置用于将该变换的语音信号与对应的说话者模型进行比较；和

用于提供一指示该未知的说话者是与该选择的说话者变换相关联的说话者的似然性的一信号的输出装置。

8、根据权利要求 7 的设备，其中该变换存储装置存储各所述变换作为一代表性的特征矢量序列；且其中该变换装置包括

一起始点和结束点检测器，用于检测该接收的语音信号内的一语音开始点和一语音结束点；

一特征矢量发生器，用于生成从该输入语音导出的一特征矢量序列；及

一对准装置，用于将对应于该检测的开始点和检测的结束点之间的语音信号的该特征矢量序列与一代表性的特征矢量序列相对准以使在得到的被对准的特征矢量序列中的各特征矢量对应于该代表性的特征矢量序列中的一特征矢量。

9、根据权利要求 8 的设备，其中该变换装置还包括有：平均装置，用于用该代表性的特征矢量序列中的对应特征矢量对该被对准的特征矢量序列中的各特征矢量进行平均。

10、根据权利要求 7 至 9 中任一的设备，其中该说话者模型存储装置被配置用于存储一隐藏的马克夫模型形式的说话者模型。

11、根据权利要求 10 的设备，其中该说话者模型存储装置被配置以存储是一左至右隐藏的马克夫模型形式的说话者模型。

12、根据权利要求 11 的设备，当权利要求 10 从属于权利要求 8 或权利要求 9 时，其中该存储的代表性的特征矢量序列包括与对应的隐藏的马克夫模型中的状态数量相同数量的特征矢量。

说话者识别

技术领域

本发明涉及说话者识别。在说话者识别中，说话者的身份被识别或验证。在说话者识别中，一说话者或者被识别为一组已知说话者之一，或者作为一未知的说话者而被拒绝。在说话者识别中，说话者或者作为具有一声称的身份而被接受或者被拒绝。说话者可例如通过一口令、一个人身份识别号或一卡而输入一要求的身份。

背景技术

通常，对于说话者识别，语音处理目的在于提高对于不同说话者的所说词的影响，而对于语音识别，其中一特定的词（或有时一个短语或者一个音素，或者其他所说的内容）被识别，语音处理目的在于减少对不同说话者的所说词的影响。

输入语音数据（通常是数字形式的）到一前端处理器是共同的，该前端处理器从输入语音数据流导出更紧凑、感性上更明显的称之为输入特征矢量的数据（或有时称之为前端特征矢量）。其中说话者说一对于识别设备和说话者是已知的预定的词（例如在银行中的个人身份识别号），该技术已知为“正文相关（text-dependent）”技术。在说话者识别的一些应用中，使用一种技术，其中该技术不要求语音的内容是预定的，这样的技术别已知为“正文无关（text independent）”技术。

在正文相关技术中，存储的该词的一表示，称之为模板或模型，被预先从一已知是真实的说话者导出。从待被识别的说话者导出的输入特征矢量被与该模板进行比较且两者之间的类似性的测量被与一接受判定的阈值进行比较。可借助于在 Chollet&Gagnoulet 所著的

“On the evaluation of Speech Recognisers and Data Bases using a Reference System（使用参考系统的语音识别器及数据基础的评估）”，1982 IEEE, International Conference on Acoustics（国际声学会），Speech and Signal Processing（语音及信号处理），pp 2026-2029（2026－2029 页）中所述的动态时间扭曲（Dynamic time warping－DTW）来进行该比较。其他比较的手段包括隐藏马克夫模型（Hidden Markov Model－HMM）处理和神经网络。这些技术在 British Telecom Technology Journal, Vol. 6, No.2 April 1988（英国电信技术刊物，第 6 卷，第 2 号，1988 年 4 月）中 105－115 页的由 SJ Cox 所著的“Hidden Markov Models for Automatic Speech Recognition: Theory And Application（用于自动语音识别的隐藏马克夫模型：理论及应用）”；131－139 页的由 McCullogh 等人所著的“Multi-layer perceptrons applied to speech technology（用于语音技术的多层感知器）”和 140－163 页的由 Tattershall 等人所著的“Neural arrays for speech recognition（用于语音识别的神经阵列）”中被进行了描述。

各种类型的特征已被用于或被建议用于语音处理。通常，由于用于语音识别的特征类型倾向于从对于说话者是不敏感的另一词中分辨出一词，而用于说话者识别的特征类型倾向于对于一（些）已知的词而言在若干个说话者之间进行辨别，适用于一种识别的一种特征对于另一种识别可能是不合适的。在 Atal 所著的“Automatic Recognition of Speakers from their voices（从他们的话音自动识别说话者）”，Proc IEEE vol 64 pp 460-475, April 1976（Proc IEEE 第 64 卷 460－475 页，1976 年 4 月）中描述了适用于说话者识别的一些特征。

发明内容

根据本发明，提供有一种说话者识别方法，包括有步骤：接收来

自一未知的说话者的语音信号；根据一变换（transform）对该接收的语音信号进行变换，该变换是与一特定的说话者相关联的；将该变换的语音信号与一代表所述特定的说话者的模型进行比较；且将依据于该未知的说话者是所述的特定说话者的似然性的一参数提供作为输出。

较佳地，该变换步骤包括有子步骤：检测该接收的语音信号内的一语音开始点和一语音结束点；生成从该接收的语音信号导出的一特征矢量序列；及将对应于该检测的开始点和检测的结束点之间的语音信号的该特征矢量序列与用于所述特定说话者的代表性的特征矢量序列相对准以使在被对准的特征矢量序列中的各特征矢量对应于该代表性的特征矢量序列中的一特征矢量。

有利地，该变换步骤还包括有子步骤：平均带有该代表性的特征矢量序列中的对应特征矢量的该被对准的特征矢量序列中的各特征矢量。

较佳地，该模型是一隐藏的马克夫模型且可以是一左至右（left to right）隐藏的马克夫模型。

有利地，该代表性的特征矢量序列包括与隐藏的马克夫模型中的状态数相同数量的特征矢量。

根据本发明的另一方面，提供有一种用于说话者识别的设备，包括有：用于接收来自一未知的说话者的语音信号的接收装置；用于存储多个说话者变换的说话者变换存储装置，各变换与多个说话者中对应的一个相关联；用于存储多个说话者模型的说话者模型存储装置，各说话者模型与所述多个说话者中对应的一个相关联；与该接收装置和说话者变换存储装置耦合的变换装置，被配置用于根据一选择的说话者变换对该接收的语音信号进行变换；耦合至该变换装置和说话者模型存储装置的比较装置，被配置用于将该变换的语音信号与对应的说话者的模型进行比较；和用于提供一指示该未知的说话者是与该选择的说话者变换相关联的说话者的似然性的一信号的输出装置。

较佳地，该变换存储装置存储各所述变换作为一代表性的特征矢量序列；且该变换装置包括一起始点和结束点检测器，用于检测该接收的语音信号内的一语音开始点和一语音结束点；一特征矢量生成器，用于生成从该接收的语音信号导出的一特征矢量序列；及一对准装置，用于将对应于该检测的开始点和检测的结束点之间的语音信号的该特征矢量序列与用于所述特定说话者的代表性的特征矢量序列相对准以使在得到的被对准的特征矢量序列中的各特征矢量对应于该代表性的特征矢量序列中的一特征矢量。

有利地，该变换装置还包括有：平均装置，用于平均带有该代表性的特征矢量序列中的对应特征矢量的该被对准的特征矢量序列中的各特征矢量。

较佳地，该说话者模型存储装置被配置用于存储一隐藏的马克夫模型形式的说话者模型且可被配置以存储是一左至右（left to right）隐藏的马克夫模型形式的说话者模型。

有利地，该存储的代表性的特征矢量序列包括与隐藏的马克夫模型中的状态数相同数量的特征矢量。

众所周知，发音期间的说话者的发声道可被模型化为一时间变化滤波器。在本发明中，在将从语音导出的特征矢量与一存储的参考模型进行比较之前，通过施加与一特定的说话者的发声道的特性匹配的与说话者相关的变换，对这些特征矢量进行处理。从具有与该变换所依据的说话者的特性非常不类似的特性的语音导出的特征通过该变换可被严重地失真，而具有与该变换所依据的说话者的特性类似的特性的语音导出的特征则被失真小得多。这样一与说话者相关的变换可被看作为与常规的匹配的滤波处理（其中使用一匹配的滤波器使滤波的信号不发生失真）类似的一处理。这样被变换的特征因此提供说话者之间的更多辨别。这样变换的特征然后被用于常规的说话者识别比较过程。

附图说明

图 1 示出了结合有一识别处理器的一电信系统；

图 2 示出了结合有一频谱信号抽取器的图 11 中的识别处理器的部分；

图 3 示出了图 2 中的频谱信号抽取器；

图 4a 是说明载荷说话者验证期间图 1 中的识别处理器的操作的流程图；

图 4b 是说明在说话者识别期间图 1 中的识别处理器的操作的流程图；

图 5 示出了两特征矢量 M 和 R 之间的一扭曲函数（warping function）的例子；

图 6 示出了在扭曲期间可被施加的一加权函数的例子；

图 7 是说明在两特征矢量之间的时间正规化距离的计算的流程图；

图 8 是一马克夫模型的例子；

图 9 示出了该转变矩阵和图 8 的马克夫模型的一起始（initialisation）矢量的例子；

图 10 示出了一六状态隐藏的马克夫模型的前向概率的计算；及

图 11 示出了使用韦特比算法计算的一可能状态序列。

具体实施方式

下面参照附图，通过例子对本发明进行描述。

在图 1 中，示出了包括有说话者识别设备的一电信系统，该电信系统包括有一麦克风 1（通常形成电话手机的部分）、一电信网络 2（例如公共交换电信网（PSTN）或数字电信网）、一被连接以接收来自网络 2 的话音信号的识别处理器 3、和一应用设备 4，其被连接至该识别处理器 3 且被配置以从识别处理器 3 接收一话音识别信号，指示一特定说话者的识别或未识别，并响应其而采取行动。例如该应用

设备 4 可以是远程操作的银行终端，用于影响银行交易。在许多情况下，该应用设备 4 将生成对用户的一音频响应，经网络 2 发送给一扬声器 5（通常形成电话手机的部分）。

在操作中，一说话者对麦克风 1 说话且一模拟语音信号被从麦克风 1 发送进网络 2 到识别处理器 3，其中该语音信号被分析且生成指示一特定说话者的识别或未识别的信号并发送给该应用设备 4，应用设备 4 然后在一特定说话者的识别或未识别的情况下采取适当的动作。如果识别处理器正执行说话者识别，则该信号或者指示被识别的说话者或者指示该说话者已被拒绝。如果该识别处理器正执行说话者验证，则该信号指示该说话者是否是所声称的说话者。

该识别处理器需要获取涉及该语音信号与其比较的说话者的身份的数据。该数据获取可由识别处理器在操作的第二模型中执行，其中识别处理器 3 未被连接至应用设备 4，而接收来自麦克风 1 的语音信号以形成用于该说话者的识别数据。然而，获取说话者识别数据的其他方法也是可能的；例如，说话者识别数据可被容纳于由说话者携带的一卡上且可被插入一卡读取器中，从而读取该数据并在传送该语音信号之前，通过网络发送给该识别处理器。

通常，识别处理器 3 不知道自麦克风 1 及通过网络 2 到其所经由的路径；麦克风 1 例如可通过一移动模拟或数字无线电链路被连接至网络 2，或可自另一城市始发。该麦克风可以是多种接收机手机之一的部分。类似地，在网络 2 内，可采取多条传输路径中的任一条，包括无线电链路、模拟及数字路径等。

图 2 示出了识别处理器 3 的部分。一频谱信号抽取器 20 例如从一数字电话网络或者从一模数转换器接收数字语音。从该数字语音导出多个特征矢量，各特征矢量代表多个连续数字样本。例如，这些语音样本可以 8khz 的取样率被接收，且一特征矢量可代表 256 个连续样本的一帧，即 32ms 的语音。

频谱信号抽取器 20 将特征矢量提供给一端点检测器 24，该端点

检测器 24 提供指示该接收的语音的开始点和结束点的输出信号。这些特征矢量在由说话者识别处理器 21 进行处理之前还被存储在帧缓冲器 25 中。

使用一常规的基于能量的端点器（endpointer）提供这些语音的开始和结束点。在一改进的技术中，来自被配置用于识别特定词的一语音识别器的信号可被使用。

说话者识别处理器 21 接收多个特征矢量，其从说话者变换存储装置 22 读取与一特定说话者相关联的与说话者相关的变换矩阵并从一说话者模型存储装置 23 读取与该特定说话者相关联的一参考模型。该说话者识别处理器然后根据所抽取的说话者变换矩阵处理接收的特征矢量，并根据由所抽取的模型代表的说话者和产生由接收的特征矢量代表的语音的与说话者相关的变换的似然性而生成一输出信号。该说话者识别处理器的操作将参照图 4a 和 4b 进行更全面的描述。该说话者识别处理器 21 构成本发明的变换装置、比较装置和输出装置。

现参见图 3，将更详细地描述频谱信号抽取器 20 的操作。一高频加重滤波器 10 以例如 8khz 的取样率接收数字化的语音波形作为一序列 8 位数并执行高频加重滤波处理（例如通过执行 $1-0.95^{-1}$ 滤波器）以增加较高频率的幅度。被滤波的信号的连续样本的一帧通过一窗口处理器 11 被开窗（即这些样本被乘以预定的加权常数），使用例如汉明窗，以减少由这些帧边缘生成的寄生污迹。在一优选实施例中，这些帧被重叠例如 50%，以使在该例中每 16ms 提供一帧。

256 开窗样本的各帧然后由一 MFCC（Mel Frequency Cepstral Coefficient－美频率倒谱系数）发生器 12 处理以生成一 MFCC 特征矢量，该 MFCC 特征矢量包括一组 MFCC 系数（例如 8 个系数）。

该 MFCC 特征矢量是这样被导出的：对一语音信号的各帧执行一频谱变换例如快速傅里叶变换（FFT）以导出一信号频谱；将该频谱的这些项集成为一系列宽带，该些宽带沿频率轴以“美－频率”标度分布；取各带中的幅度的对数；且然后执行进一步的变换（例如离散余

弦变换 DCT）以生成用于该帧的 MFCC 系数组。可发现有用的信息通常被限制在下级系数。该美一频率标度是在 0 和 1khz 之间的一线性频率标度上均匀间隔的、且在 1khz 上的一对数频率标度上均匀间隔的频带。

通过一或多个适当编程的数字信号处理器（DSP）和/或微处理器，可提供高频加重滤波器 10、MFCC 发生器 12、端点检测器 24 和说话者识别处理器 21。帧缓冲器 25、说话者变换存储装置 22 和说话者模型存储装置 23 可被设置在连接至这些处理器装置的读/写存储器装置中。

图 4a 概略地示出了在说话者验证期间说话者识别处理器 21 的操作。在步骤 40，说话者识别处理器接收一特征矢量序列和来自端点检测器 11 的一检测的开始点和一检测的结束点。在步骤 41，对于使用者被声称是该说话者，说话者识别处理器从说话者变换存储装置 22 选择一与说话者相关的变换矩阵并从该说话者模型存储装置 23 读取表示与该代表的特征矩阵相同的说话者的一对应模型。

该与说话者相关的变换矩阵表示用于一特定说话者的一特定词。它包括当由该代表的说话者说出时的该代表的词的一代表性特征矢量序列。该与说话者相关的变换矩阵在这里也被称作为代表性的特征矢量序列。在步骤 42，使用动态时间扭曲（DTW）处理，对应于检测的开始点和检测的结束点之间的语音信号的该接收的特征矢量序列与该与说话者相关的变换矩阵进行时间对准。

现将参照图 5、6 和 7 更加详细地描述在步骤 42 执行的时间对准。

该与说话者相关的变换矩阵包括用于一特定词的一代表性的特征矢量序列。

$$M = m_1, m_2, \ldots m_i \ldots m_I$$

一特征矢量序列

$$R = r_1, r_2, \ldots r_j \ldots r_J$$

被接收。如下所述，该接收的特征矢量序列与该代表性的特征矢量序

列进行时间对准。

参见图5，该代表性序列被沿i轴表示且该接收的序列沿j轴表示。

点序列C=（i,j）表示一"扭曲"函数，其近似地实现从该接收的特征矢量序列的时间轴到该代表性的特征矢量序列的时间轴的映射。

$$F = c(1), c(2),...,c(k),...c(K) \text{ where } c(k) = (r(k),m(k))$$

作为两特征矢量M和R之间的差的测量，使用一距离 $d(c)=d(i,j)=\|m_i-r_j\|$ 在该扭曲函数上这些距离的求和是

$$\sum_{k=1}^{K} d(c(k))$$

其给出了该扭曲函数F如何将一组特征矢量映射到另一组特征矢量上的量度。当F被确定最佳地调节该两特征矢量序列之间的时间差时，该量度达到一最小值。可替换地，可采用一加权函数以使一加权的求和被使用

$$\sum_{k=1}^{K} d(c(k)).\omega(k)$$

且ω(k)被使用以对该距离量度进行加权。加权函数的一个例子是

$$\omega(K)=(i(K)-i(K-1))+(j(K)-j(K-1))$$

其被概略地示出在图6中。

两特征矢量序列之间的时间正规化的距离被定义为

$$D(M,R)=\overset{Min}{\forall F}\left[\frac{\sum_{k=1}^{K}d(c(k)).\omega(k)}{\sum_{k=1}^{K}\omega(k)}\right]$$

如 Sskoe 和 Chiba 所著的“Dynamic Programming Algorihtm Optimisation for Spoken Word Recognition（用于所说的词识别的动态编程算法最优化）”，IEEE Transactions on Acoustics Speech and Signal Processing，vol 26，No.1，February 1978（声学语音和信号处理学报，第 6 卷，第 1 期，1978 年 2 月）中所述的，可对该扭曲函数施加各种不同的约束。计算时间正规化距离连同提供所需的最小值的扭曲函数一起的方程如下：

$$g_1(c(1))=d(c(1),\omega(1))$$
$$g_k(c(k))=\overset{Min}{\forall c(k-1)}\left[g_{k-1}(c(k-1))+d(c(k)).\omega(k)\right]$$

其被称之为“动态编程”方程

该时间正规化距离是

$$D(M,R)=\frac{1}{\sum_{k=1}^{K}\omega(k)}\ gK(c(k)).$$

如果先前示出的加权函数被使用，则该动态编程（DP）方程变为

$$g(i,j)=Min\begin{bmatrix}g(i,j-1)&+d(i,j)\\g(i-1,j-1)&+2d(i,j)\\g(i-1,j)&+d(i,j)\end{bmatrix}$$

及

$$\sum_{K=1}^{k}\omega(K)=I+J$$

在图 7 中示出使用图 6 的加权函数计算该时间正规化距离的流程图。

在步骤 74，i 和 j 被初始化等于 1。在步骤 76，g（1，1）的初始值被设置等于乘以 2 的 $m_1 - r_1$ (d(1,1)) （根据加权函数 w）。然后，在步骤 78，i 被增加 1 且除非在步骤 80，i 大于 1，在步骤 86，该动态编程方程被进行计算。如果 i 大于 1，则在步骤 88，j 被增加且在步骤 96，i 被复位至 1。然后重复步骤 78 和 86 直至最后对于所有的 I 和 J 的值，该动态编程方程已被进行了计算，则在步骤 92，计算了该时间正规化距离。

在一更加有效的算法中，该动态编程方程仅对于在一大小为 r 的限制窗口内的值进行计算，以使

$$j-r\le i\le j+r$$

然后如下通过“退回（backtracking）”来确定该扭曲函数 F：

$$C(K)=(I,J)$$

$$C(k-1)=i,j$$

其中当扭曲函数 F

$$=C(1),C(2),C(3),..C(k)..C(K)$$ 时

$$\left.\begin{matrix}g(i,j-1)\\g(i-1,j-1)\\g(i-1,j)\end{matrix}\right\}$$ 为最小值

其中 $C(k)=(r(k),m(k))$

然后可能确定一“时间对准的”接收的特征矢量序列

$$\omega = \omega_1, \omega_2, \ldots, \omega_l$$

在图5中所示的例子中

$$\begin{aligned} C(1) &= (1,1) \\ C(2) &= (1,2) \\ C(3) &= (2,2) \\ C(4) &= (3,3) \\ C(5) &= (4,3) \end{aligned}$$

即 r_1 被映射至 m_1，r_1 被映射至 m_2，r_2 被映射至 m_2、r_3 被映射至 m_3 等。

可看到在此情况下 r_1 和 r_2 两者都已被映射到 m_2 且对于那个接收的特征矢量应被用于时间对准的特征矢量作出确定。选择接收的特征矢量之一的另一种方法是计算映射到一单个的代表性特征矢量上的接收的特征矢量的平均值。

如果第一个这样的接收的特征矢量被使用，则 $\omega_p = r_q$

其中

$$q = \underset{j(k)}{Min} \forall i(k) = p$$

或者如果最后一个这样的接收的特征矢量被使用，则 $\omega_p = r_s$

其中

$$s = \underset{j(k)}{Max} \forall i(k) = p$$

或者如果使用一平均值

$$\omega_p = Ave\,(r_{j(k)}) \forall i(k) = p$$

这样，在图5所示的例子中，假定第一个这样接收的矢量被使用

$$\begin{aligned} \omega_1 &= r_1 \\ \omega_2 &= r_2 \\ \omega_3 &= r_3 \\ \omega_4 &= r_4 \end{aligned}$$

等。

显然这样一对准处理导致一对准的特征矢量序列，其中该对准的特征矢量序列中的各特征矢量对应于该代表性特征矢量序列中的一特征矢量。

再参见图 4a，在该变换处理的一改进的版本中，在任选的步骤 43 中，各被时间对准的接收的特征矢量还用该与说话者相关的变换矩阵的对应的特征矢量进行平均。如果该时间对准的接收的特征矢量与该与说话者相关的变换矩阵的对应的特征矢量明显不同，则这样一平均步骤将严重地变形改时间对准的接收的特征矢量，而如果这些时间对准的接收的特征矢量类似于改与说话者相关的变换矩阵，则该平均处理将很少地变形该接收的特征矢量矩阵。这些变换的特征将增强在任何随后的比较过程中的辨别。

然后在步骤 44 中，这些变换的特征在一常规的说话者识别比较过程中被使用。在本发明的该实施例中，由一左至右隐藏的马克夫模型提供该说话者模型，且使用韦特比算法进行比较（在后将参照图 8 至 11 进行描述）。在步骤 45，指示该被表示的说话者产生由这些接收的特征矢量代表的语音的似然性的一距离量度被生成且随后与一阈值进行比较。如果其间的差异小于该阈值，在步骤 47，该说话者被接受位对应于该存储的模板；否则在步骤 46，该说话者被拒绝。

现将参照图 8 至 11 对使用隐藏的马克夫模型和韦特比算法模型化语音的原理进行描述。

图 8 示出了一例子 HMM。五个圆圈 100、102、104、106 和 108 表示该 HMM 的状态且在一离散时间瞬间 t，该模型被认为处于这些状态之一且被认为发出一观测值（observation）Ot。在语音或说话者识

别中，各观测值通常对应于一特征矢量。

在瞬间 t＋1，该模型或者移至一新的状态或者呆在相同的状态中且在另一情况下发出另一观测值等等。该发出的观测值仅取决于该模型的状态。在时间 t＋1 占用的状态仅取决于在时间 t 占用的状态（该特性被称之为马克夫特性）。从一状态移至另一状态的概率可被列表在一 N×N 状态转变矩阵（$A=[a_{i,j}]$）中，如图 9 所示。在该矩阵的第 i 行和第 j 列的项是从在时间 t 的状态 Si 移至在时间 t＋1 的状态 S_j 的概率。当从一状态移动的概率是 1.0（如果该模型呆在相同的状态下，则被认为是到其自身的一转变），该矩阵的各行求和至 1.0。在示出的该例子中，该状态转变矩阵仅具有在上三角形中的项，因为该例子是一左至右模型，其中不允许“向后”转变。在一更加通常的 HMM 中，转变可从任何状态到任何其他的状态。还示出一起始矢量 (π)，其第 i 分量是在时间 t=1 占用状态 Si 的概率。

假定 W 个这样的模型存在 M1，...Mw，各表示一特定的说话者且假定来自一未知的说话者的语音信号由一 T 个观测值 O_1，O_2，O_3，...，O_T 的序列表示，则问题是确定哪个模型最有可能已发出了该观测值序列，即确定 k，其中

$$P_k = \max_{i=1,2,3,\ldots,W} \Pr(O \mid M_i).$$

Pr（O | M）被如下地递归地计算：

该前向概率 $\alpha_t(j)$ 被确定是一模型发出该特定的观测值序列 O_1，O_2，O_3，...，O_t 且在时间 t 占用状态 S_j 的概率。

因此

$$\Pr(O \mid M) = \sum_{j=1}^{N} \alpha_T(j)$$

该模型在时间 t＋1 占用状态 Sj 且发出观测值 O_{t+1} 的概率可从在时间 t 的前向概率、状态转变概率（$a_{i,j}$）和状态 S_j 发出观测值 O_{t+1} 的概率 $b_j(O_{t+1})$ 被计算如下

$$\alpha_{t+1}(j)=\left(\sum_{i=1}^{N}\alpha_t(i)a_{i,j}\right)b_j(O_{t+1})$$

图 10 示出了对于一个六状态 HMM 的计算。

通过设置 $\alpha_1(j)=\pi(j)b_j(O_1)$ 来初始化该递归。

上述算法的一个计算上更加有效的变型被称之为是韦特比算法。在替代如上所述的求和前向概率的韦特比算法中，使用前向概率的最大值。

即 $$\phi_{t+1}(j)=\left(\underset{i=1,2,\ldots,N}{Max}\ \phi_t(i)a_{i,j}\right)b_j(O_{t+1})$$

如果要求恢复该最大可能的状态序列，则每次 ϕ_t 被计算 $\psi_t(j)$ 被记录，其中假定在时间 t 是状态 S_j，$\psi_t(j)$ 是在时间 t－1 的最大可能的状态，即最大化上述方程的右手侧的状态。在时间 T 的最大可能状态是对于其 $\phi_T(j)$ 是最大的状态 S_k 且 $\psi_T(k)$ 给出了在时间 T－1 的最大可能状态等等。

图 11 示出了对于十六个帧的观测值（特征矢量）序列及一个五状态左至右隐藏的马克夫模型，使用韦特比算法计算的一可能状态序列。

图 4b 示出了在说话者识别中说话者识别处理器的对应操作；在此情况下，使用多个说话者变换和对应的说话者模型。进而选择各与说话者相关的变换并使用其在步骤 42 时间对准接收的特征矢量。然后在步骤 48，将该时间对准的接收的特征矢量序列与对应的说话者模型进行比较。如先前参照图 4a 所述，在任选的步骤 43，各时间对准的接收的特征矢量还用与说话者相关的变换矩阵的对应的特征矢量被进行平均。然后由于具有指示该已知的说话者对应于该未知的说话者的最大似然性的距离量度，该说话者被识别为已知的说话者。然而，如果在步骤 53，该最小的距离量度大于一阈值，指示没有说话

者具有是该已知说话者的特定的高似然性，则在步骤 54 ，由于对于该系统是未知的，该说话者被拒绝。

历史上，DTW 比较处理相比于 HMM 比较处理，对于说话者识别的效果更佳。将一特征矢量序列与一隐藏的马克夫模型进行比较和使用一动态时间扭曲（DTW）算法将相同序列与一代表性模板进行比较之间的差异在于图形匹配阶段。在 DTW 方案中，一接收的特征矢量可被匹配至两或更多的代表性特征矢量，对应于图 5 中的水平路径。然而，在 HMM 方案中，各接收的特征矢量可仅被匹配至一个状态。它不可能具有图 11 中的一水平路径。将接收的特征矢量序列与与说话者相关的变换矩阵对准，允许将接收的特征矢量映射至 HMM 状态的更多的可能性，且因此可改善基于 HMM 的说话者识别器的性能。

基于 HMM 说话者的识别器和基于 DTW 的说话者识别器之间的另一差异是 DTW 模板是整体地基于一个个体（individual）的语音，而一单个的 HMM 拓扑经常在用一个体的语音训练一组模型之前被定义。在本发明的一改善的实施例中，根据各个体的训练语音，由具有不同数量的状态的 HMM 提供这些说话者模型。例如，用于一特定词的一组特定个体的训练发声中的最小数量的特征矢量可被用于选择用于该特定个体的该特定词的 HMM 的状态数目。在与说话者相关的变换矩阵中的特征的数量可被类似地确定，其中在该代表性特征矢量序列中的特征数量将与隐藏的马克夫模型中的状态数量相同。

已参照 MFCC 对本发明进行了描述，但显然任何适当的频谱表示可以使用。例如，线性预测系数（LPC）倒谱系数、快速傅里叶变换（FFT）倒谱系数、线谱对（LSP）系数等。

尽管已讨论了使用隐藏的马克夫模型的比较处理，本发明同等地适用于采用其他类型的比较处理的说话者识别，例如动态时间扭曲技术或神经网络技术。

本发明采用用于各待被识别的说话者的一与说话者相关的变换。在此所述的本发明的实施例中，借助于用于各词的一代表性特征

矢量序列，提供与说话者相关的变换矩阵。

导出代表性的特征矢量序列的方法是众所周知的，且对于理解本发明，指出各代表性特征矢量序列可通过接收由一说话者对于同一词的多个发声并如上所述地对于各发声导出一组特征矢量的处理而被形成是足够的。这些序列然后被时间对准，例如先前所述，且然后对用于该多个发声的时间对准的特征矢量序列进行平均以导出提供该与说话者相关的变换矩阵的一平均的特征矢量序列。

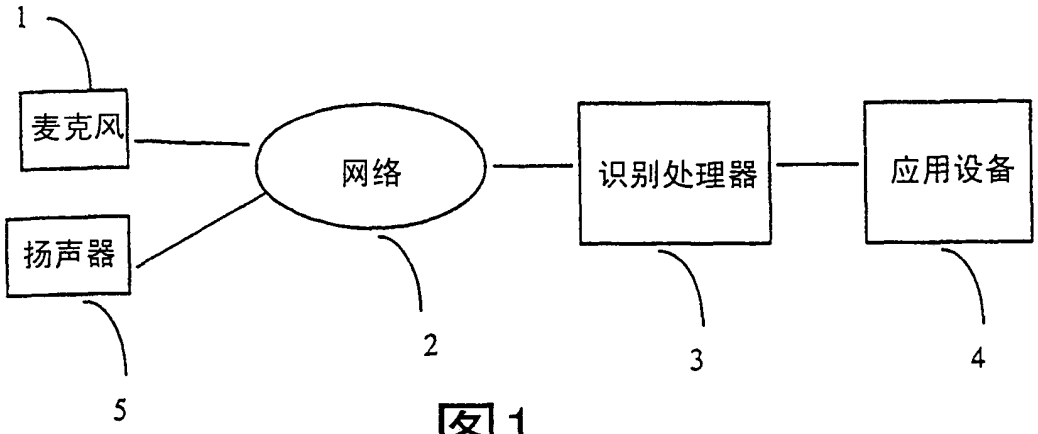

1
麦克风
扬声器
网络
识别处理器
应用设备
2
3
4
5

图1

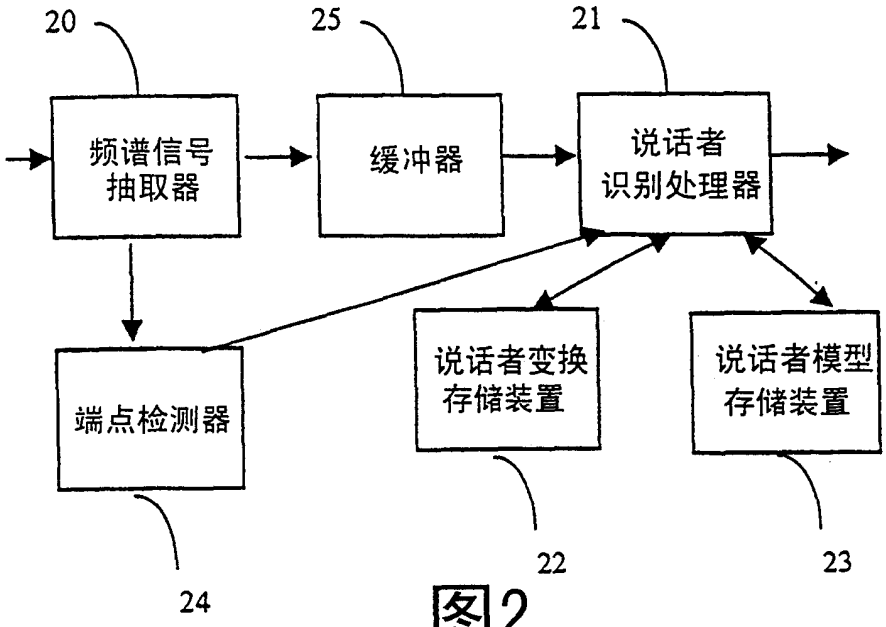

20
25
21
频谱信号
抽取器
缓冲器
说话者
识别处理器
端点检测器
说话者变换
存储装置
说话者模型
存储装置
22
23
24

图2

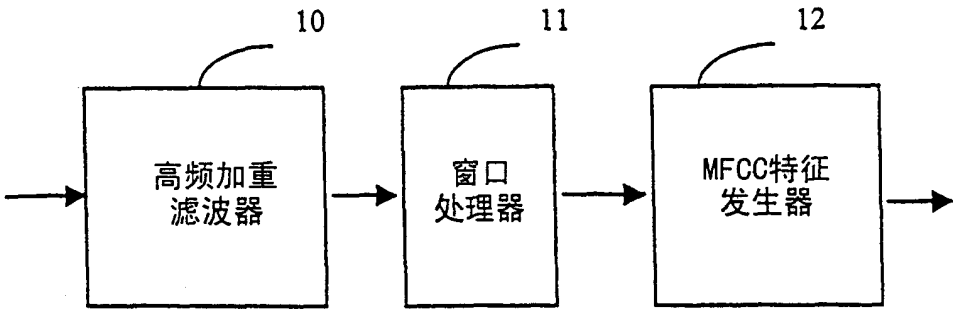

10
11
12
高频加重
滤波器
窗口
处理器
MFCC特征
发生器

图3

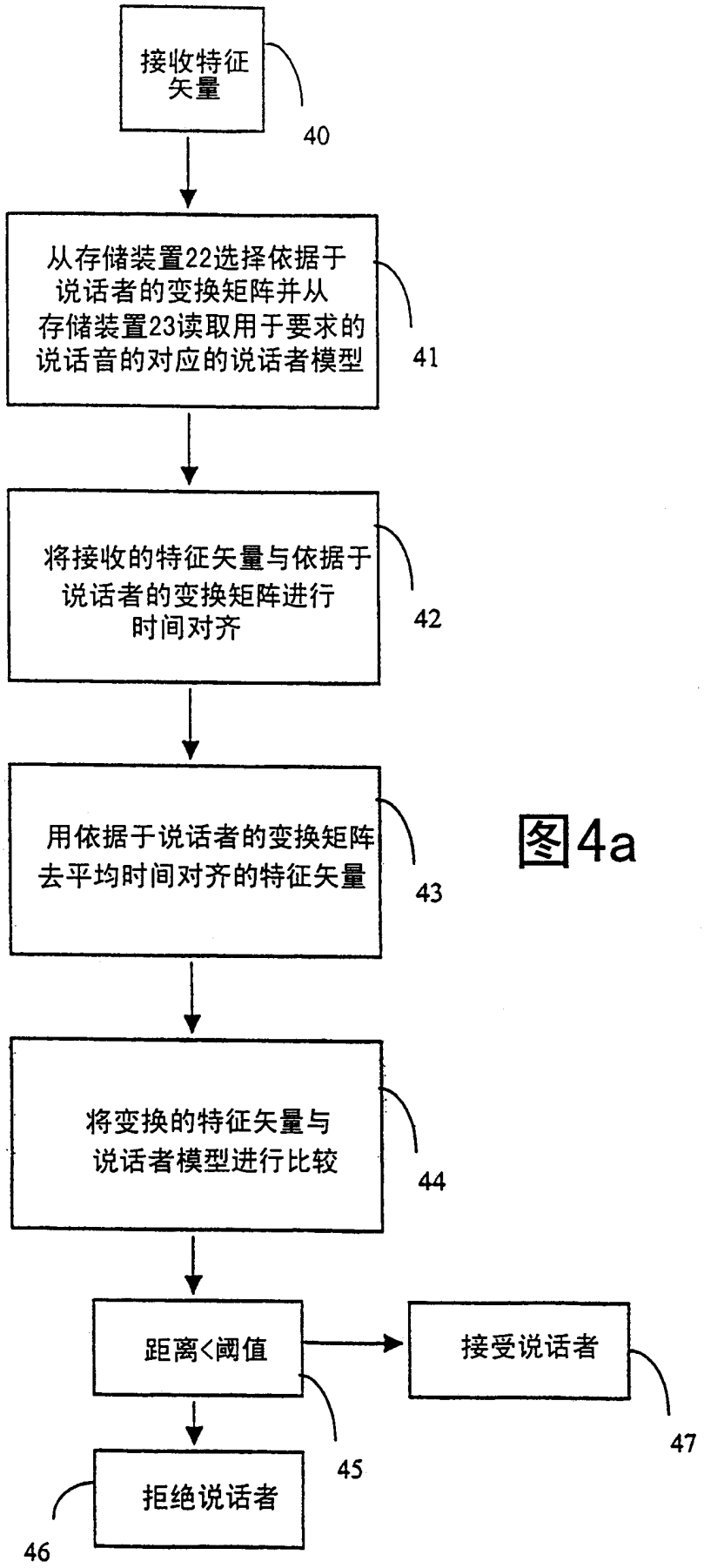

接收特征矢量
40
从存储装置22选择依据于说话者的变换矩阵并从存储装置23读取用于要求的说话音的对应的说话者模型
41
将接收的特征矢量与依据于说话者的变换矩阵进行时间对齐
42
用依据于说话者的变换矩阵去平均时间对齐的特征矢量
43
将变换的特征矢量与说话者模型进行比较
44
距离<阈值
45
接受说话者
47
拒绝说话者
46

图4a

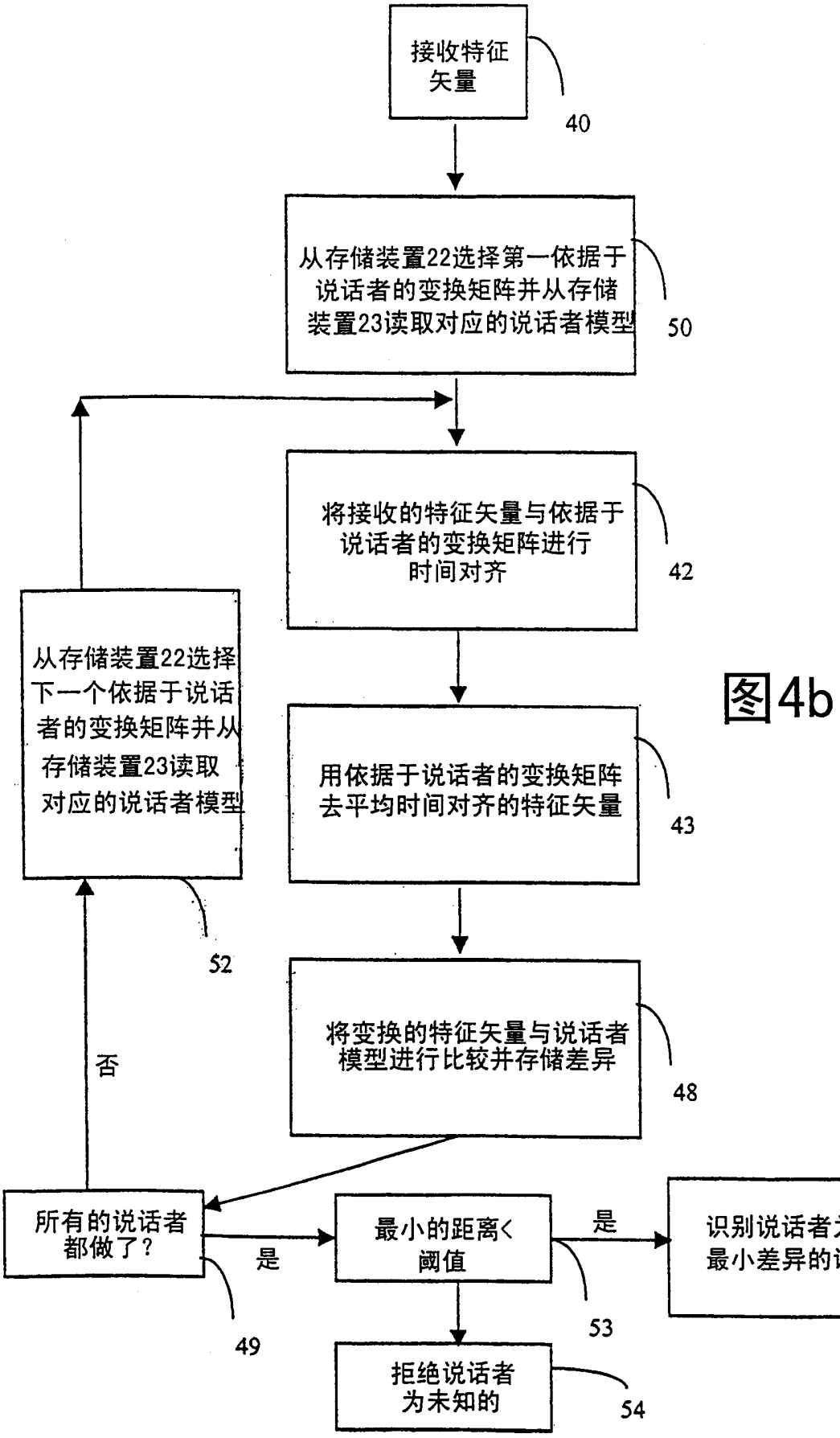

接收特征矢量
40
从存储装置22选择第一依据于说话者的变换矩阵并从存储装置23读取对应的说话者模型
50
将接收的特征矢量与依据于说话者的变换矩阵进行时间对齐
42
用依据于说话者的变换矩阵去平均时间对齐的特征矢量
43
将变换的特征矢量与说话者模型进行比较并存储差异
48
所有的说话者都做了?
49
否
从存储装置22选择下一个依据于说话者的变换矩阵并从存储装置23读取对应的说话者模型
52
是
最小的距离<阈值
53
是
识别说话者为具有最小差异的说话者
51
拒绝说话者为未知的
54

图4b

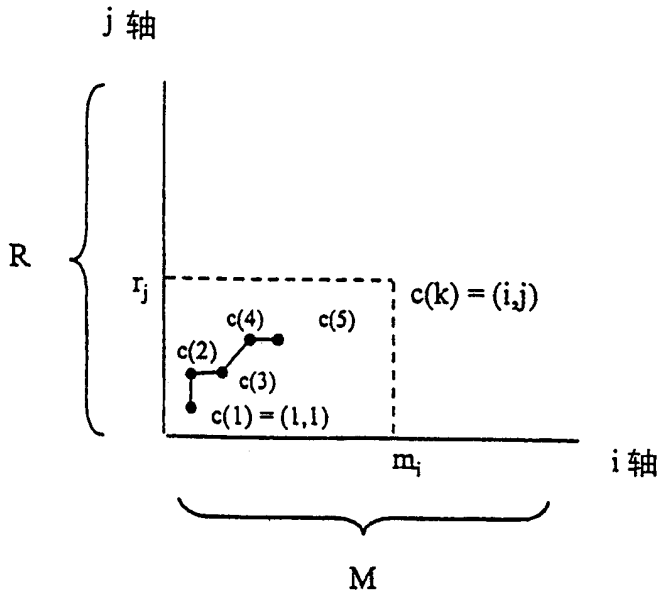


图5

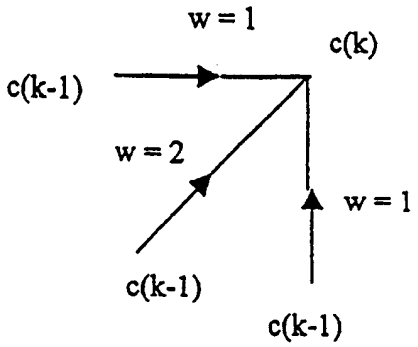


图6

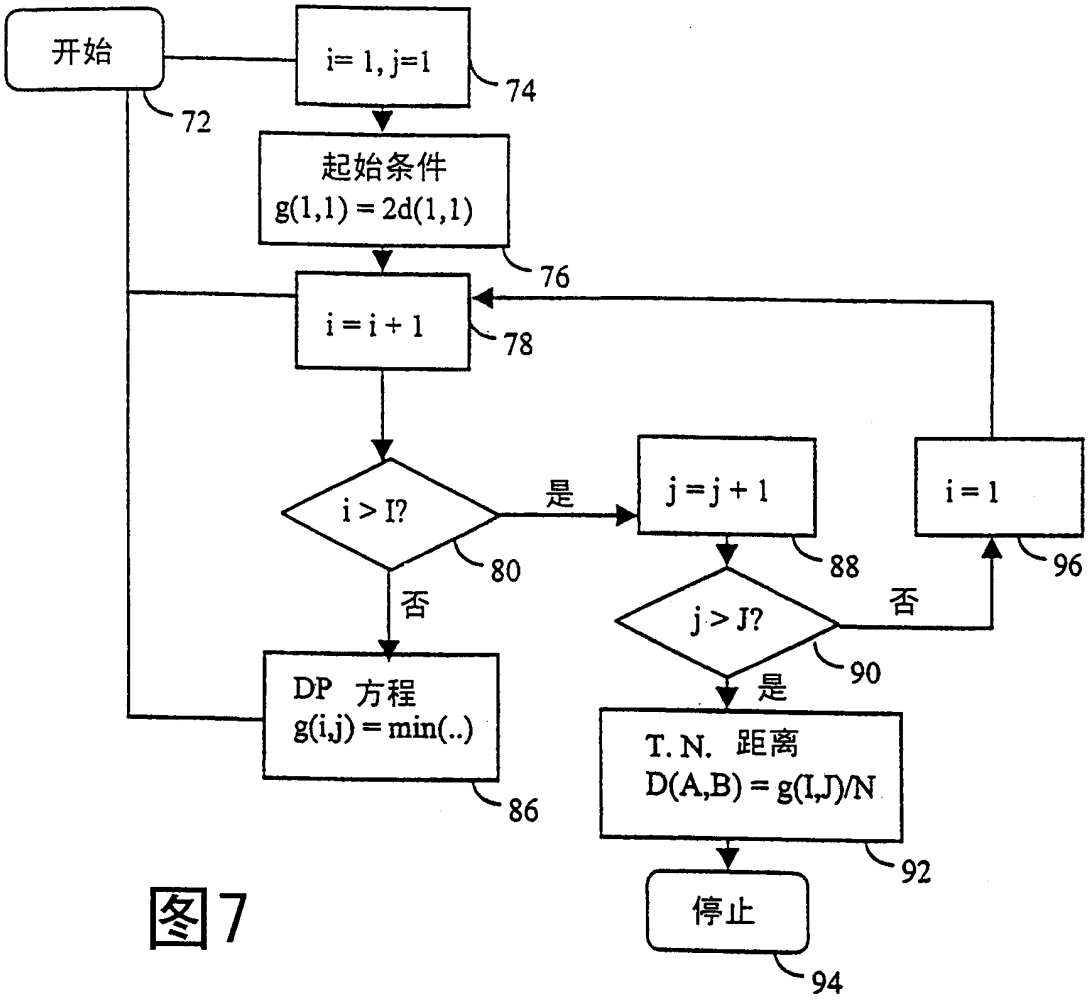

开始
72
i= 1, j=1
74
起始条件
g(1,1) = 2d(1,1)
76
i = i + 1
78
i > I?
80
是
j = j + 1
88
i = 1
96
否
j > J?
90
否
是
DP 方程
g(i,j) = min(..)
86
T. N. 距离
D(A,B) = g(I,J)/N
92
停止
94

图7

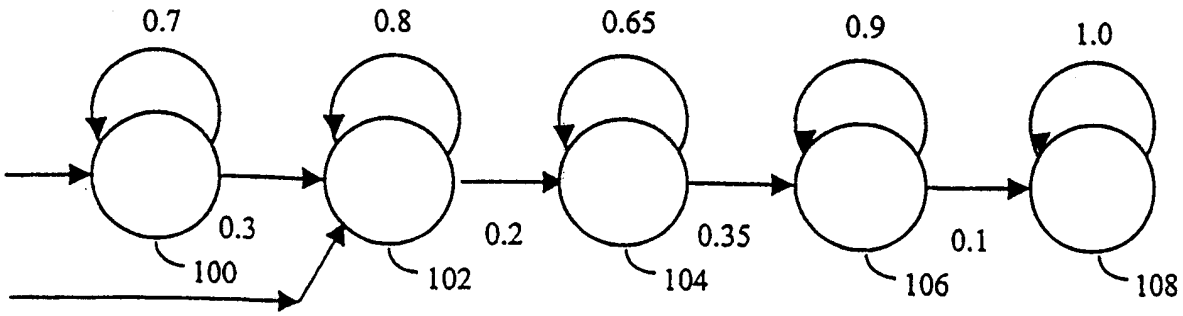

0.7
0.8
0.65
0.9
1.0
0.3
0.2
0.35
0.1
100
102
104
106
108

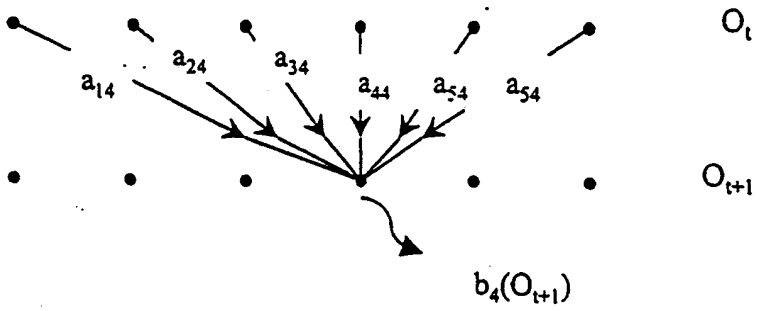
图8

图9

$$\begin{pmatrix} 0.7 & 0.3 & 0 & 0 & 0 \\ 0 & 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.65 & 0.35 & 0 \\ 0 & 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$ 转变矩阵

$$\begin{pmatrix} 0.7 & 0.3 & 0 & 0 & 0 \end{pmatrix}$$ 起始矢量

图10



$$a_{t+1}=\left[\sum_{i=1}^{5} a_t(i)a_{j4}\right]b_4(O_{t+1})$$

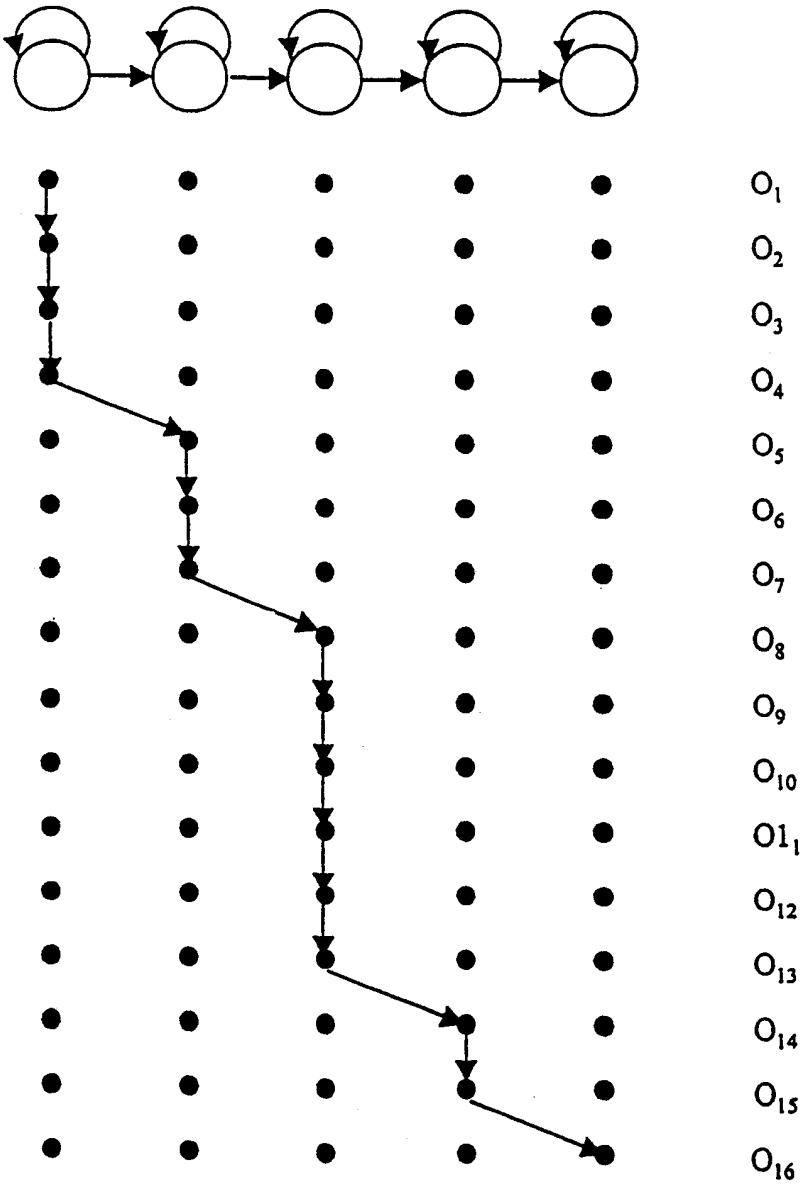


图11