[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G10L 15/00
G10L 15/14

[12]发明专利申请公开说明书

[21]申请号 00134269.X

[43]公开日 2001年6月6日

[11]公开号 CN 1298172A

[22]申请日 2000.11.29 [21]申请号 00134269.X

[30]优先权

[32]1999.11.29 [33]US [31]09/450,392

[71]申请人 松下电器产业株式会社

地址 日本大阪府

[72]发明人 罗兰·奎恩 马托·坎特里尼
让-克劳德·詹卡

[74]专利代理机构 中国国际贸易促进委员会专利商标事务所

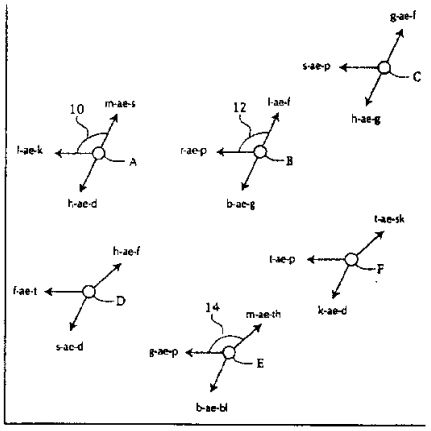
代理人 于 静

权利要求书2页 说明书14页 附图页数7页

[54]发明名称 用于中等或大词汇量语音识别的上下文相关声模型

[57]摘要

在训练时使用降低维数本征语音分析技术来为异音素构造上下文相关声模型。在运行时,对于新说话者的语音还使用本征语音技术。该技术去掉各个说话者 特异性以产生用途广并且是健壮的异音异模型。在一实施例中,使用本征语音 技术来标识每个说话者的质心,然后从识别等式中将其减去。在另一实施例中 ,使用最大似然估计技术来构造通用决策树框架,在构造说话者空间的本征语 音表示时可以在所有的说话者之间共享该框架。



ISSN1008-4274

权 利 要 求 书

1. 一种为自动语音识别建立上下文相关模型的方法，包括:

产生本征空间表示训练说话者组;

为至少一个训练说话者提供一组声数据，在所述本征空间中表示所述声数据以便为所述训练说话者确定至少一个异音素质心;

从所述声数据中减去所述质心以便为所述训练说话者产生说话者调节的声数据;

使用所述说话者调节的声数据为不同的异音素增长其叶节点包含上下文相关模型的至少一决策树。

2. 权利要求 1 的方法，进一步包括使用多个训练说话者的一组声数据，为所述多个训练说话者的每一个产生所述说话者调节声数据。

3. 权利要求 1 的方法，其中通过根据来自所述训练说话者组的语音构造超向量并对所述超向量进行维数降低以定义覆盖所述训练说话者组的降低维数空间来产生所述本征空间。

4. 一种使用权利要求 1 的上下文相关模型来进行语音识别的方法，包括:

提供来自新的说话者的语音;

使用所述本征空间确定新的说话者的至少一个新说话者质心并从所述说话者的所述语音数据中减去所述说话者质心以产生说话者调节数据; 以及

对于采用所述上下文相关模型的语音识别器应用所述说话者调节数据。

5. 一种使用权利要求 1 的上下文相关模型进行语音识别的方法，包括:

提供来自新的说话者的语音;

使用所述本征空间确定新的说话者的至少一个新说话者质心并将所述新说话者质心加到所述上下文相关模型中去以产生新的说话者调节上下文相关模型; 以及

对于采用所述新说话者调节上下文相关模型的语音识别器应用所述语音数据。

6. 一种用于训练自动语音识别的上下文相关模型的方法，包括:

构造具有用于存储上下文相关异音素模型的叶节点的“是”-“否”问题的决策树框架;

为多个训练说话者训练一组说话者相关模型，并且使用所述决策树框架为所述训练说话者构造多个决策树，在各个决策树的叶节点中为每个训练说话者存储说话者相关模型;

使用所述一组决策树产生之后将通过维数降低进行变换的超向量来构造本征空间。

7. 一种构造为自动语音识别存储上下文相关模型的决策树的方法，包括:

提供“是-否”问题池以标识声单元的不同上下文;

提供测试说话者数据的语料库;

对于多个由所述语料库表示的测试说话者以及所述问题池中的多个问题，迭代地执行步骤（a）至步骤（e）

（a）从所述问题池中选择一问题;

（b）使用来自所述测试说话者的第一说话者的说话者数据对于选择的问题构造第一“是”模型和第一“否”模型;

（c）为所述第一“是”模型和所述第一“否”模型计算概率得分的第一乘积;

（d）使用来自所述测试说话者的第二说话者的说话者数据对于选择的问题构造第二“是”模型和第二“否”模型;

（d）为所述第二“是”模型和第二“否”模型计算概率得分的第二乘积;

（e）通过计算包括第一和第二乘积的全局积对于所述选择的问题计算全局得分;

增长具有用从问题池中选择的不同问题繁殖的节点的决策树以便在每个节点使用具有最高全局得分的问题。

说　　明　　书

用于中等或大词汇量语音识别的上下文相关声模型

小词汇量语音识别系统将需要识别的小词汇量中的词作为其基本单元。例如，用于识别英文字母的系统通常有26个模型，每个字母一个模型。该方法对于中等和大词汇量语音识别系统来说是不切合实际的。这些大的系统通常将语言的音素或音节作为其基本单元。如果一系统对于语言的每一音素都有一模型（例如隐藏马尔科夫模型），则将该系统称作带有"与上下文无关"的声模型的系统。

如果一系统对于一给定的音素，依据周围音素的识别来使用不同的模型，则将该系统称为应用"上下文相关"声模型。异音素（allophone）是由其上下文定义的音素的专门版本。例如，在"t"之前发音"ae"的所有情况（例如在"bat"，"fat"等中）定义了异音素"ae"。

对于大多数语言，音素的发声对其前后音素的依赖性很大：例如，前边带有"y"的"eh"（例如在"yes"中）完全不同于前边带有"s"的"eh"（例如"set"）。于是，对于一个中等或较大词汇量的系统，上下文相关声模型的性能要优于上下文无关模型。今天，大多数实际应用的中等或较大词汇识别系统都采用上下文相关声模型。

当今许多上下文相关识别系统应用决策树聚类来定义上下文相关的、与说话者无关的声模型。树增长算法寻找关于感兴趣的音素周围的音素的问题并从声音上将感兴趣的音素的不相似样例分开。结果产生一个"是－否"问题的决策树，用于选择能够最佳识别给定的异音素的声模型。通常，"是－否"问题与异音素如何出现在上下文中（例如谁是它的相邻音素）相关。

传统的决策树对于每个音素定义了包含根节点和中间节点（例如根节点的子、孙节点）中的是/否问题的二叉树。端节点，或叶子节点包括为音素的特定异音素设计的声模型。于是，在使用时，识别系统

遍历该树，根据考虑的音素的上下文分枝“是”或“否”，直到标识出包含可应用模型的叶节点。于是标识出的模型用于识别。

不幸的是，传统的异音素模型可能出错。我们认为这是因为当前的方法没有考虑每个训练说话者的特定特异性。当前的方法假设如果使用大的训练讲话者库，可以平均掉各个说话者的特异性。然而，在实际中，我们发现这种假设并不是总成立。传统的基于决策树的异音素模型在新的说话者的语音刚好与该组训练的说话者的语音相似时能工作得很好。然后，在新的说话者的语音在该组训练的说话者语音域外，则传统的技术则失败。

本发明通过降低说话者空间维数估计技术来解决前述问题，该技术能够快速地标识各个说话者的特异性的从识别等式中除去该特异性来产生可应用于各种情况并且具有健壮性的异音素模型。该降低说话者空间维数估计技术可以在降低了维数的空间（我们称为本征语音空间或本征空间）中应用。本征语音技术的一个重要优点就是速度。当新的说话者使用识别器时，他或她的语音被迅速地放入或投影到从训练的一组说话者中得出的本征空间中。甚至能够使用新的说话者的非常快速的发音就将该新的说话者放入本征空间中。在本征空间中，异音素可以通过诸如在说话者空间中说话者的位置等非相干因素的最小影响来表示。

通过以下参照附图的说明可以更完整地理解本发明，和其目的以及优点。在以下说明中，给出两个基本实施例。对于本领域技术人员来说可以作各种修改和变型。

图1是用于描述在理解一组说话者的质心和相关的异音素向量对于不同的说话者是不同时使用的说话者空间的示意图；

图2是一称为本征质心加δ树实施例的第一优选实施例的方框图；

图3是利用由图2所示的实施例得出的δ决策树的语音识别器的一个实施例；

图4是利用由图2所示的实施例得出的δ决策树的语音识别器的另一实施例；

图5说明如何使用由图2所示的实施例产生的说话者调节数据来构造δ树;

图6示出在相应于图5的δ树的声空间中对说话者调节数据分组的过程;

图7示出包括关于本征空间维数问题的示例的δ决策树;以及

图8示出了本发明的第二实施例,用于对于每个说话者存在一个较完整数据的情况。

本发明的技术可应用于各种不同的语音识别问题。该项技术最适合应用于中等或大的词汇量方面的应用,其中不易由其自己的模型来表示每个完全词。在此描述本发明的两个实施例。可以理解本发明的原理可以扩展到其它实施例。

该第一实施例对于每个训练说话者已经提供了中等数量的训练数据(例如,每个说话者的训练数据为20-30分钟)的情况下是优化的。对于这样大小的训练数据量,预期将有足够的语言声样例用于为每个说话者构造合理的较佳上下文无关、说话者有关模型。如果希望,可以使用说话者自适应技术产生足够的数据以训练上下文无关模型。尽管不必具有每个说话者的全部异音素的样例全集,但数据在某方面应反映数据中每个音素的最重要异音素(即,已由至少一小部分说话者发声了几次的异音素)。

该实施例的识别系统基于异音素的上下文(例如基于其相邻音素)应用决策树来为每个异音素标识合适的模型。然而,与基于决策树的传统建模系统不同,该实施例在构造决策树时使用说话者调节的训练数据。事实上,说话者调节过程去掉每个训练说话者的语音的特异性以产生较佳的异音素模型。然后,在使用识别系统时,对新的说话者的语音进行类似的调整,从而可以访问说话者调节异音素模型以进行高质量、上下文有关的识别。

本实施例的识别系统的一个重要组成部分是本征语音技术,通过使用该技术可以快速地分析训练说话者的语音以及新的说话者的语音,以提取各个说话者的特异性。以下详细讨论的本征语音技术定义

了减少了维数的本征空间，该本征空间集合地表示了训练说话者组。当在识别过程中一个新的说话者讲话时，快速地将他或她的语音放入或投影到本征空间以快速地确定该说话者的语音质心如何落入与训练说话者相关的说话者空间。

如以下详细解释的，通过平均每个说话者如何发音该系统的音素来定义新的说话者的质心（以及每个训练说话者的质心）。为了方便，人们可以将质心向量考虑成由给定说话者上下文无关模型中每个音素HMM的每个状态中的并置高斯均值向量构成。然而，"质心"概念是一标量并且与对每个训练说话者可得到多少数据有关。例如，如果存在足够训练数据来为每个说话者训练较富足的说话者相关模型（例如双音素(diphone)模型），则每个训练说话者的质心是来自该说话者相关双音素模型的并置高斯均值。当然，也可以实现诸如三音素模型等其他模型。

图1通过图示六个不同的训练说话者A－F在不同的上下文中可能怎样对音素"ae"发声来描述质心概念。图1示出说话者空间，为了方便将其图示为二维空间，其中每个说话者的质心位于该二维空间中用于该说话者的异音素向量的中心。于是，在图1中，说话者A的质心位于从说话者A发音的下列词"mass"、"lack"以及"had"导出的各个异音素向量的起点。于是说话者A的质心包括粗略表示该说话者"平均"音素"ae"的信息。

通过比较，说话者B的质心在说话者空间中位于说话者A的右边。通过下述发音："laugh"、"rap"以及"bag"产生说话者B的质心。如图所示，其他说话者C－F位于说话者空间的其他区域中。请注意，每个说话者具有表示为始发于质心的向量（图1中表示的三个异音素向量）的一组异音素。如图所示，这些向量定义了同经常粗略比较的不同说话者之间的角关系。将说话者A的角10和说话者B的角12进行比较。然而，因为各个说话者的质心与其他说话者不同心，从而产生的说话者A和B的异音素不同。本发明通过去除由不同质心位置表征的说话者相关特异性来处理该问题。

尽管在说话者中通常可以对异音素向量之间的角关系进行比较，但这并不是说向量是相同的。确实，向量长度对于不同说话者可能是不同的。男性说话者和女性说话者很容易具有彼此不同的异音素向量长度。此外，对于说话者不同方言，也存在不同角关系。就此来说，比较说话者 E 的角 14 和说话者 A 的角 10。这样的角度差例如反映了说话者 A 讲的是美国北方方言，而说话者 E 讲的是美国南方方言。

除了这些向量长度和角度差别之外，质心位置的差异表示了重要的传统上下文相关识别器没有解决的说话者相关的产物。正如以下更完整地说明的那样，本发明提供了一种方便补偿质心位置差异以及其他向量长度和角度差异的机制。

图 2 示出了第一优选实施例，我们称作本征质心加δ树实施例。具体地说，图 2 示出了用于训练识别器将使用的δ树的优选步骤。图 3 和图 4 示出了对于由新的说话者提供的语音使用该识别器的实施例。

参照图 2，通过如在 16 处描述的，提供来自多个训练说话者的声数据来增长本实施例使用的δ决策树。将来自每个训练说话者的声数据投影或放置到本征空间 18。在本发明的优选实施例中，可以将本征空间截断以减小其尺寸和计算复杂性。在此我们称减小尺寸的本征空间为 K－空间。

由步骤 20－26 描述了用于产生本征空间 18 的过程。该过程使用训练说话者声数据 16 产生用于每个训练说话者的说话者相关（SD）模型，如步骤 20 所示。然后在步骤 22 将这些模型向量化。在该优选实施例中，通过并置每个说话者的语音模型的参数来将说话者相关模型向量化。通常使用隐藏马尔科夫模型为每个说话者产生一个超向量，该向量包括对应于该说话者的隐藏马尔科夫模型的至少一部分参数的有序列表（通常为浮点数）。可以按任何方便的顺序来组织参数。该顺序并不重要，然而，一旦采纳一种顺序，就应适用于所有训练说话者。

接着在步骤 24 对超向量进行维数降低步骤以定义本征空间。可以通过将原高维超向量降低成基向量的任何线性变换来进行维数的降低。维数降低技术的非穷尽列表包括：主要分量分析（PCA）、独立向

量分析（ICA）、线性鉴别分析（LDA）、因子分析（FA）和奇异值分解（SVD）。

在步骤 24 产生的基向量定义了由本征向量覆盖的本征空间。维数降低为每个训练说话者产生了一个本征向量。于是，如果存在 n 个训练说话者，则维数降低步骤 24 产生 n 个本征向量。这些本征向量定义了本征语音空间或本征空间。

构成本征空间的本征向量每一个都表示一个不同说话者可能是不同的维数。在原始训练集中的每个超向量可以表示了这些本征向量的线性组合。这些本征向量按它们在建模数据中的重要性来排序：第一个本征向量比第二个重要，第二个比第三个重要，等等。

尽管在步骤 24 产生最大 n 个本征向量，实际中，有可能丢弃一些本征向量，只保持前 K 个本征向量。于是，可选地在步骤 26 从 n 个本征向量中提取 K 个以包括降低参数的本征空间或 K 空间。可以丢弃高次序的本征向量，因为它们通常包括的信息对于鉴别说话者来说不重要。将本征语音空间降至低于训练说话者的总数有助于消除原始训练数据中的噪声，还提供了在有限的存储器和处理器资源情况下构造实际系统时有用的本征数据压缩。

在构造了本征空间 18 之后，将各个训练说话者的声数据投影或放到本征空间，如 28 所示。每个说话者数据在本征空间（K－空间）中的位置表示了每个说话者的质心或平均音素发音。如图 1 所示，预计这些质心对于不同说话者是不同的。使用本征空间技术确定说话者音素质心的最大优点是速度。

当前用于将每个说话者数据放入本征空间的优选技术涉及我们称为最大似然性估计技术（MLED）。在实际中，最大似然技术将选择本征空间的超向量，它与说话者输入的语音最一致，而不管实际可获得多少语音。

为了说明，假设说话者是亚拉巴马的青年女性。在接收到来自该说话者发出的一些音节时，最大似然技术将选择表示与该说话者的亚拉巴马女性音调一致的所有音素（甚至那些在输入语音中不表示的）

的本征空间中的一个点。

最大似然技术应用表示对隐藏马尔科夫模型的预定集合产生观测数据的概率的概率函数 Q。如果函数不仅包括概率项 P，而且还包括该项的对数，log P，则更容易处理该概率函数 Q。然后通过相对于每个本征值，单独对概率函数取导，使概率函数极大化。例如，如果本征空间是 100 维，该系统计算概率函数 Q 的 100 个导数，将每个设置成零，并对各个本征值 W 求解。

如此得到的集 Ws 表示了识别本征空间中对应于最大似然度点的点所需的本征值。于是 Ws 集包括了本征空间中的最大似然向量。然后可以使用该最大似然向量来构造对应于本征空间中的优化点的超向量。

在本发明最大似然框架情况下，我们希望相对于给定模型使观测 O 的似然最大。这可以通过迭代使以下的辅助函数 Q 最大来完成

$$Q\left(\lambda,\hat{\lambda}\right)=\sum_{\theta\in states} P\left(O,\theta\mid\lambda\right)\log\left[P\left(O,\theta\mid\hat{\lambda}\right)\right]$$

其中λ表模型，$\hat{\lambda}$是估计模型。

作为初步估计，我们仅想对于均值进行最大化。在此情况下，由一组 HMM 给出概率 P，我们得到：

$$Q(\lambda,\hat{\lambda})=const\ -\frac{1}{2}P(O\mid\lambda)\sum_{\substack{states\\ in\lambda}}^{S_\lambda}\sum_{\substack{mixt\\ gauss\\ inS}}^{M_s}\sum_{\substack{time\\ t}}^{T}\left\{\gamma_m{}^{(s)}(t)[n\log(2\pi)+\log\mid C_m{}^{(s)}\mid+h(o_t,m,s)]\right\}$$

其中：

$$h(o_t,m,s)=(o_t-\hat{\mu}_m{}^{(s)})^T C_m{}^{(s)-1}(o_t-\hat{\mu}_m{}^{(s)})$$

并且：

Q_t：是时间 t 的特征向量

$C_m^{(s)-1}$是状态 s 的混合高斯 m 的协方差倒数

$\hat{\mu}_m^{(s)}$是状态 s、混合分量 m 近似采纳均值

$\gamma_m^{(s)}(t)$是 P（使用混合高斯 m|λ,Q_t）

假设对于新的说话者的 HMM，高斯均值在本征空间中。该均值超向量$\bar{\mu}_j$覆盖该空间，j = 1...E

$$\bar{\mu}_j = \begin{bmatrix} \bar{\mu}_1^{(1)}(j) \\ \bar{\mu}_2^{(1)}(j) \\ \cdot \\ \cdot \\ \bar{\mu}_m^{(s)}(j) \\ \bar{\mu}_{Ms_\lambda}^{(S_\lambda)}(j) \end{bmatrix}$$

其中$\bar{\mu}_m^{(s)}(j)$表示在本征空间（本征模型）j 的状态 s 中对于混合高斯 m 的均值向量。

然后我们需要:

$$\hat{\mu} = \sum_{j=1}^{E} w_j \bar{\mu}_j$$

$\bar{u}_j$是正交的，并且 Wj 是我们的说话者模型的本征值。在此我们假设可以将任何新的说话者模型化为我们的观测的说话者数据库的线性组合。然后:

$$\hat{\mu}_m^{(s)} = \sum_{j=1}^{E} w_j \bar{\mu}_m^{(s)}(j)$$

s 在λ的状态中，m 在 M 的混合高斯中。

由于我们需要使 Q 最大化，我们只需设置:

$$\frac{\partial Q}{\partial w_e} = 0, \qquad e=1..E.$$

（注意，因为本征向量是正交的$\frac{\partial w_i}{\partial w_j} = 0 \quad i \neq j...$）

于是:

$$\frac{\partial Q}{\partial w_e} = 0 = \sum_{\substack{states \\ in\lambda}}^{S_\lambda} \sum_{\substack{mixt \\ gauss \\ inS}}^{M_s} \sum_{\substack{time \\ t}}^{T} \{\frac{\partial}{\partial w_e} \gamma_m^{(s)}(t) h(o_t, s)\}, \quad e=1...E.$$

计算上述的导数，我们使:

$$0 = \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \{-\overline{\mu}_m^{(s)T}(e) C_m^{(s)-1} o_t + \sum_{j=1}^{E} w_j \overline{\mu}_m^{(s)T}(j) C_m^{(s)-1} \overline{\mu}_m^{(s)}(e)\}$$

从其中我们可以发现线性等式

$$\sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \overline{\mu}_m^{(s)T}(e) C_m^{(s)-1} o_t = \sum_s \sum_m \sum_t \gamma_m^{(s)}(t) \sum_{j=1}^{E} w_j \overline{\mu}_m^{(s)T}(j) C_m^{(s)-1} \overline{\mu}_m^{(s)}(e),$$
$$e=1..E.$$

一旦已确定每个说话者的质心，则在步骤 30 将它们减去以产生说话者调节的声数据。参照图 1。该质心减去过程将移动说话者空间中的所有说话者，这样它们质心将彼此一致。事实上，这去掉了说话者的特异性，只剩下异音素相关数据。

在以此方式处理了所有训练说话者数据后，在步骤 32 使用产生的说话者调节训练数据以如在 34 图示的那样增长δ决策树。对于每个音素按此方式增长决策树。在 34 描述的是音素 "ae"。每个树包括一根节点 36，包含关于音素的上下文的问题（即：关于音素的相邻或其他上下文信息的问题）。可以用 "是" 或 "否" 回答根节点问题，从而向一对子节点向左或右分枝。子节点可以包含附加问题，如在 38 处描述的，或语音模型，如在 40 处描述的。请注意，所有叶节点（节点 40、42、和 44）包含语音模型。选择这些模型作为最适合识别一特定异音素的模型。于是，在叶节点的语音模型是上下文相关的。

在已经生成了δ决策树之后，如在图 1 描述的，可以使用系统识别新的说话者的语音。下面结合图 3 和 4 描述两个识别器的实施例。两个识别器实施例的区别主要在于是在上下文相关识别之前从声数据中减去新的说话者质心（图 3），还是在上下文相关识别之前将质心信

息加到上下文相关模型中去。

参见图 3，新的说话者 50 提供一发音，送到所示的几个处理模块。将发音送到说话者无关识别器 52，简单地启动 MLED 过程。

在将新的说话者发音送到上下文相关识别器 60 之前，从说话者声数据中减去新的说话者的质心信息。这可以通过计算新的说话者在本征空间（K－空间）的位置（如在 62 所示）从而确定新的说话者的质心（如在 64 所示）来完成。最好使用先前描述的 MLED 技术计算新的说话者在 K－空间的位置。

已经确定了新说话者的质心后，从新的说话者声数据中减去质心数据（如在 66 所示）。这产生说话者调节声数据 68，然后将其送到上下文相关识别器 60。

在图 4 示出的另一实施例以类似的方式工作。将新的说话者发音送到说话者无关识别器 52，如前所述，以启动 MLED 过程。当然，如果在特定实施例中不使用 MLED 过程，可能不需要说话者无关识别器。

同时，将新的说话者发音放入本征空间（如步骤 62）以确定新的说话者的质心（在步骤 64）。然后将质心信息加到上下文相关模型中（步骤 72）从而产生一组说话者调节的上下文相关的模型 74。然后上下文相关识别器 60 使用这些说话者调节模型产生识别器输出 70。以下表 1 示出了三个说话者的例示数据项如何通过减去质心被进行说话者调节。表中的所有数据项是音素“ae”的发音（在各种上下文中）。然后图 5 示出如何使用该说话者调节数据构造δ树。图 6 示出说话者调节数据在声空间的分组。在图 6 中 +1 表示下一音素，摩擦音是音素{f，h，s，th，…}的集合，辅音是{b，d，g…}。

TABLE I

Spkr1: centroid = (2,3)					
	"half" = >	<h *ae f>	(3, 4)	- (2,3)	= (1,1)
	"sad" = >	<s *ae d>	(2, 2)	- (2,3)	= (0,-1)
	"fat" = >	<f *ae t>	(1.5, 3)	- (2,3)	= (-0.5, 0)
Spkr2: centroid = (7,7)					
	"math" = >	< m *ae th>	(8,8)	- (7,7)	= (1,1)
	"babble" = >	< b *ae b l>	(7,6)	- (7,7)	= (0,-1)
	"gap" = >	< g *ae p>	(6.5, 7)	- (7,7)	= (-0.5,0)
Spkr3: centroid = (10,2)					
	"task" = >	< t *ae s k>	(11,3)	- (10,2)	= (1,1)
	"cad" = >	< k *ae d>	(10,1)	- (10,2)	= (0,-1)
	"tap" = >	< t *ae p>	(9.5,2)	- (10,2)	= (-0.5,0)

如果需要，可以在说话者调节过程中使用标准差以及均值。例如可以应用一单位方差条件来实现（如在倒频谱标准化中）。在说话者相关质心训练之后，提供给 MLED 的超向量将包含标准差和均值。对于每个训练数据项，在从其中减去音素状态质心后，该数据项将通过被质心标准差除而进一步被调整。这将导致由树更准确地集中异音素数据。在使用该技术时，会在运行时存在一些计算成本，因为对于输入帧的说话者调节会更复杂。

如前面指出的，共同发声可能受到说话者类型的影响，导致异音素向量的方向不同。这描述在图 1 中，其中依据说话者来自北方或南方，偏置向量的角关系不同。通过包括关于本征维数的决策树问题，可以考虑该现象。图 7 示出示例性δ决策树，在确定对特定的异音素应用哪个模型时包括关于本征维数的问题。在图 7 中，问题 80 和 82 是本征维数问题。该问题询问是否特定本征维数（在此情况中维数为 3）大于零。如果大于零，还可以询问有关本征维数的其他问题。

下面结合图 8 说明本发明的其他实施例。该实施例适用于对于每个说话者存在足够的数据以合理准确地训练说话者相关模型的应用。在该实施例中，不必要确定每个说话者的质心。

然而，为了应用本征语音技术，有必要具有一组超向量（来自于每个训练说话者相关模型）。这些超向量必须具有相同的维数并且在相同的方面对准，即指标 i 必须在所有的说话者相关模型中表示相同的参数。

于是，为了对于各说话者共用的给定音素增长良好的上下文相关异音素树，该实施例集中各讲话者数据，但保持跟踪哪一数据项来自哪一讲话者。于是将选择一问题的最大似然估计（MLE）标准扩展成计算每个测试问题的全部得分，同时分别估计和保持各个说话者的得分。图 8 描述了该技术。

参见图 8，通过提供问题池 100 来增长决策树。这些问题由树增长算法各个测试以确定哪些问题最佳地定义了异音素树的结构。

通过迭代技术检验问题池，一次一个问题。于是，图 8 的系统包括用于从池 100 中选择问题以对其测试的迭代器 102。当前测试的问题描述在 104。

回想一下，每个测试问题可能以某种方式与出现特定音素的上下文相关。于是，测试问题例如可能是给定的音素是否由摩擦音领先。树增长算法对于每个音素增长各个树，以根节问题开始并领先于附加节点，如需的话直至该音素的异音素由树结构较佳地表示。如图 8 所示，继续根节点问题和任何其他中间节点问题的选择。

通过假设已为树的那个节点选择了当前估计的问题（问题 104），选择测试问题的过程工作。通过测试问题 104 估计来自训练说话者 106 的说话者数据，从而将语音数据分成两部分：一部分对测试问题回答“是”，一部分对测试问题回答“否”。然后使用测试说话者数据构造语音模型。具体地说，对于每个说话者构造“是”模型 106 和“否”模型 108。这不同于集中所有说话者数据，并且对于给定问题，从集中的数据中训练一个“是”模型和一个“否”模型的传统过程。通过

对测试问题回答“是”的所有语音数据样例上训练声特征并且在对测试问题回答“否”的数据上类似地训练其他组声特征来训练模型。

在为每个说话者都已生成“是”模型106和“否”模型108之后，系统计算给定“是”模型106的所有“是”数据的概率得分，以及给定“否”模型108的所有“否”数据的概率得分。一个高概率得分意味着构造的模型能够很好地识别其训练数据部分。低概率得分意味着尽管使用训练数据能产生最佳模型，但该模型就此音素不能很好地进行识别。

确定概率得分以为测试问题104计算整个得分。计算过程如图8所示按以下所述地继续。首先，为第一训练说话者（说话者A）计算对于“是”模型和“否”模型的各自概率得分。这些概率得分被乘在一起给出累积得分以表示模型对说话者A的工作情况。这表示在112处。然后对于其他训练说话者进行相同的过程，如在114和116所示，一次一个说话者。最后，当对所有训练说话者都进行以上过程，通过将各个讲话者的积相乘计算全局得分。将在步骤112、114和116确定积相乘在118得出对该测试问题的全局得分。

在已产生第一测试问题的全局得分，迭代器102存储全局得分结果并从问题池100中取出第二个问题，以相同方式测试。当问题池中的问题已被全部测试完，为决策树的该节点选择给出最佳全局得分的问题。

在如上所述地已确定了决策树的根节点之后，迭代器102可以继续确定中间节点在异音素识别中是否产生足够的改进以有理由对树增加附加节点。最后，当以此方式增长了该树时，叶节点包含最佳识别特定音素的异音素的模型。

在通过前边的过程标识了决策树结构之后，现在应用本征语音技术。如果每一叶节点单一高斯就足以表示异音素，则使用共享树结构来训练异音素说话者相关模型以获得一组超向量，之后将超向量用于通过维数降低来构造本征空间。对于现在完成的训练，在线步骤是本征语音素数的简单MLED估计。多高斯模型略复杂，因为必须解决对

准问题。即，尽管已知说话者相关模型 1 的叶节点 N 和说话者相关模型 2 的叶节点 N 表示相同的异音素，不能保证说话者相关模型 1 的叶 N 的高斯 i 与说话者相关模型 2 的叶 N 的高斯 i 表示相同的现象。解决该问题的一个方法是为说话者相关模型的每个叶节点找出质心，然后由说话者调节达到所有叶子的数据。然后对于全部说话者相关模型，集中对于给定叶子的数据并且计算共享δ高斯。在运行时，MLED 将产生所有叶子质点的估计，然后在对δ高斯估计它之前从新的说话者数据中减去它。

从前边的描述中可知，本发明较佳地利用本征语音训练技术为中等和大词汇量语音识别系统开发上下文相关声模型。尽管参照优选实施例描述本发明，应理解，在不背离所附权利要求书限定的本发明精神实质和范围情况下，本发明可扩展到其他实施例中。

说 明 书 附 图

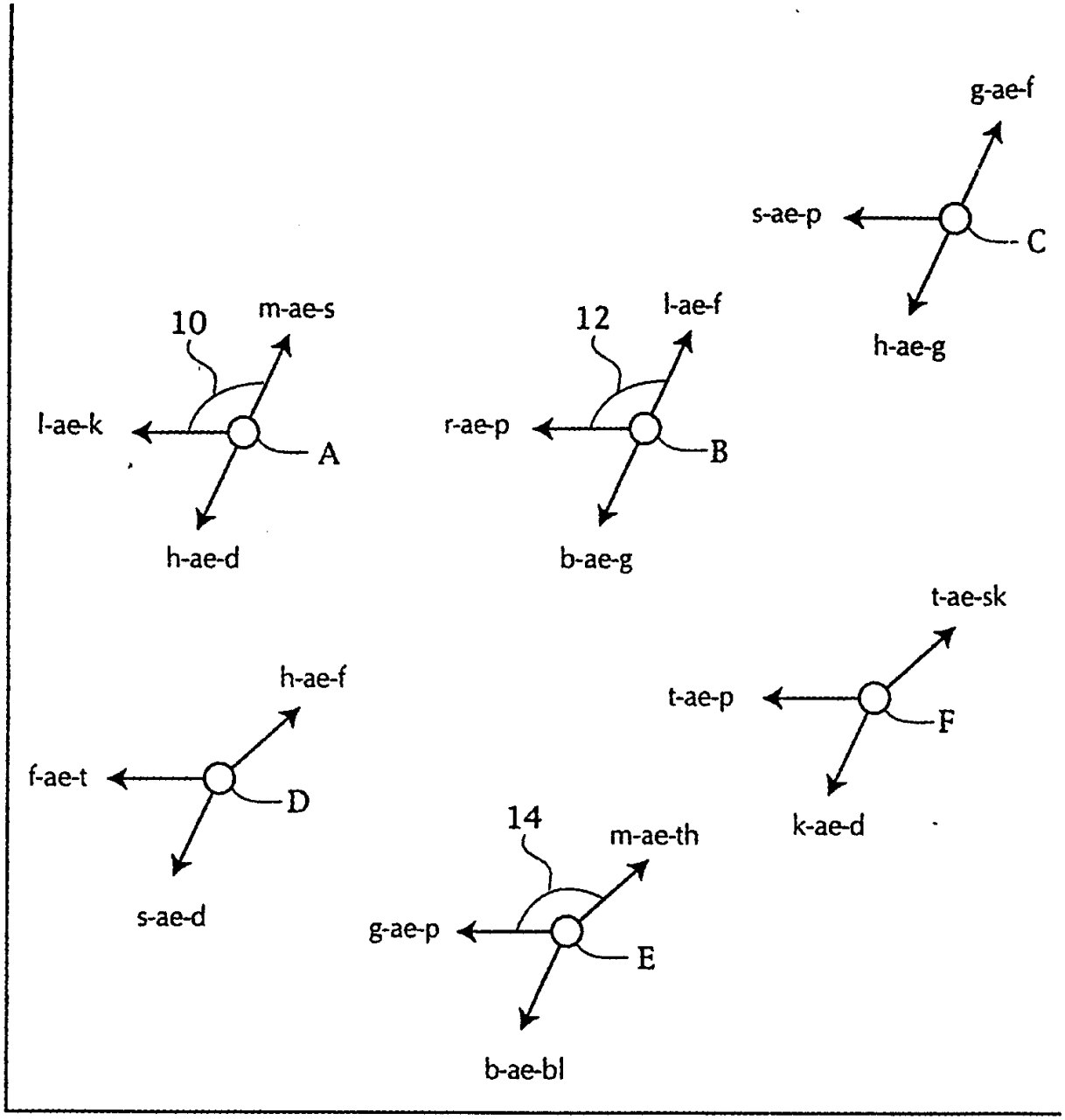

g-ae-f
s-ae-p
C
h-ae-g
10
m-ae-s
l-ae-k
A
h-ae-d
12
l-ae-f
r-ae-p
B
b-ae-g
t-ae-sk
t-ae-p
F
k-ae-d
h-ae-f
f-ae-t
D
s-ae-d
14
m-ae-th
g-ae-p
E
b-ae-bl

图1

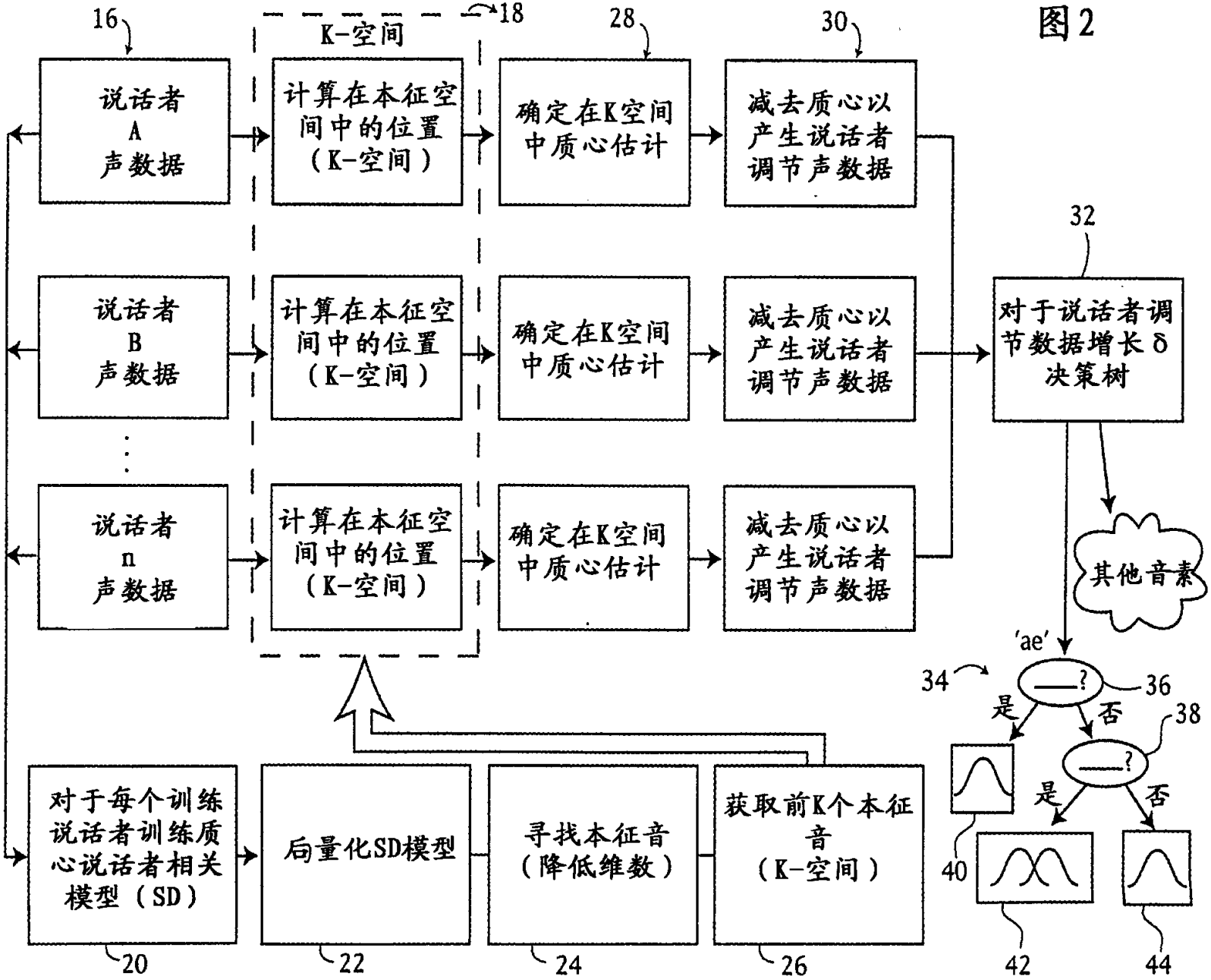
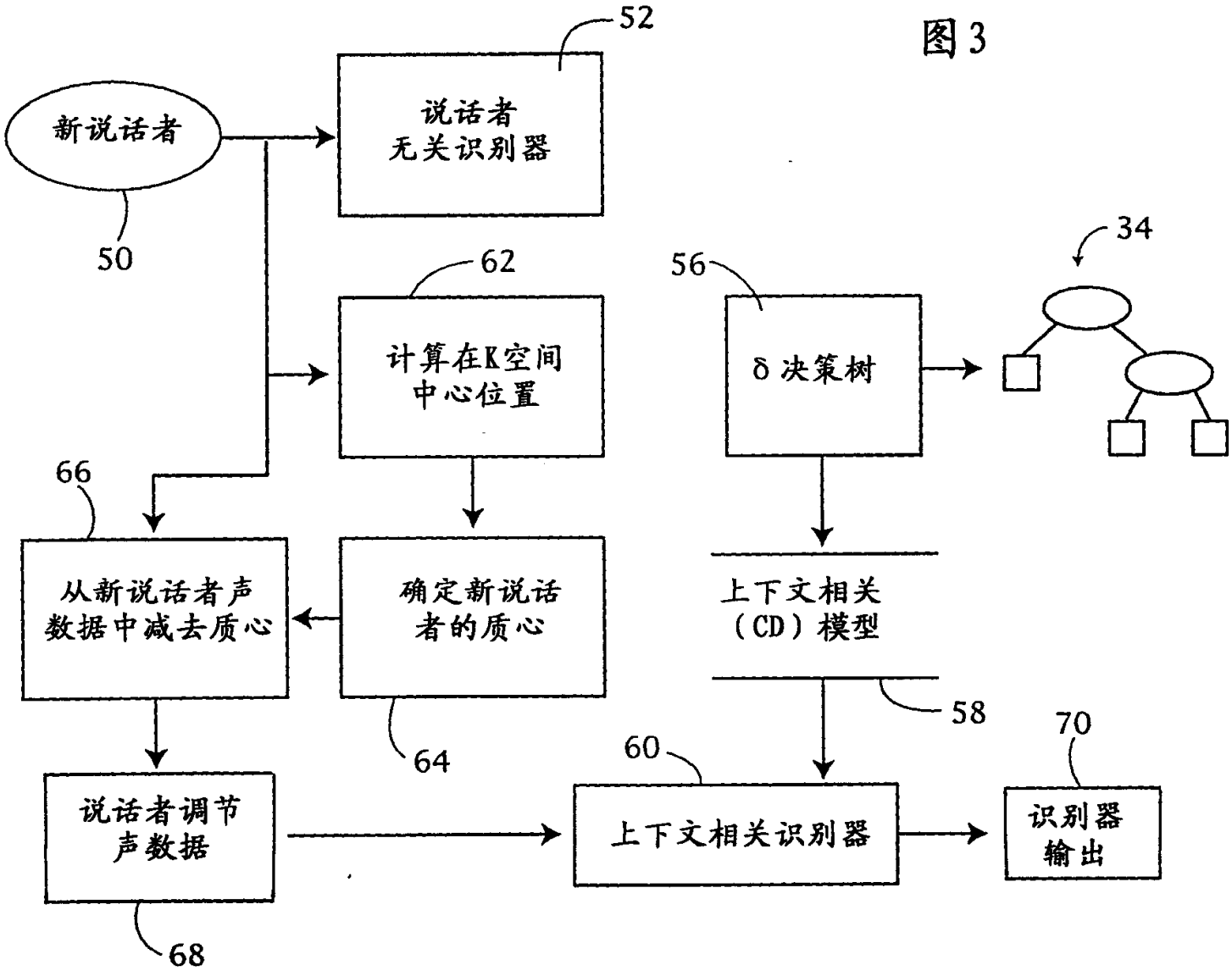

16
K-空间
18
28
30
说话者
A
声数据
计算在本征空间中的位置（K-空间）
确定在K空间中质心估计
减去质心以产生说话者调节声数据
说话者
B
声数据
计算在本征空间中的位置（K-空间）
确定在K空间中质心估计
减去质心以产生说话者调节声数据
说话者
n
声数据
计算在本征空间中的位置（K-空间）
确定在K空间中质心估计
减去质心以产生说话者调节声数据
32
对于说话者调节数据增长δ决策树
其他音素
'ae'
34
36
是
否
38
是
否
40
42
44
对于每个训练说话者训练质心说话者相关模型（SD）
后量化SD模型
寻找本征音（降低维数）
获取前K个本征音（K-空间）
20
22
24
26

图2

图 3


新说话者
50
说话者
无关识别器
52
计算在K空间
中心位置
62
δ 决策树
56
34
从新说话者声
数据中减去质心
66
确定新说话
者的质心
64
上下文相关
（CD）模型
58
说话者调节
声数据
68
上下文相关识别器
60
识别器
输出
70

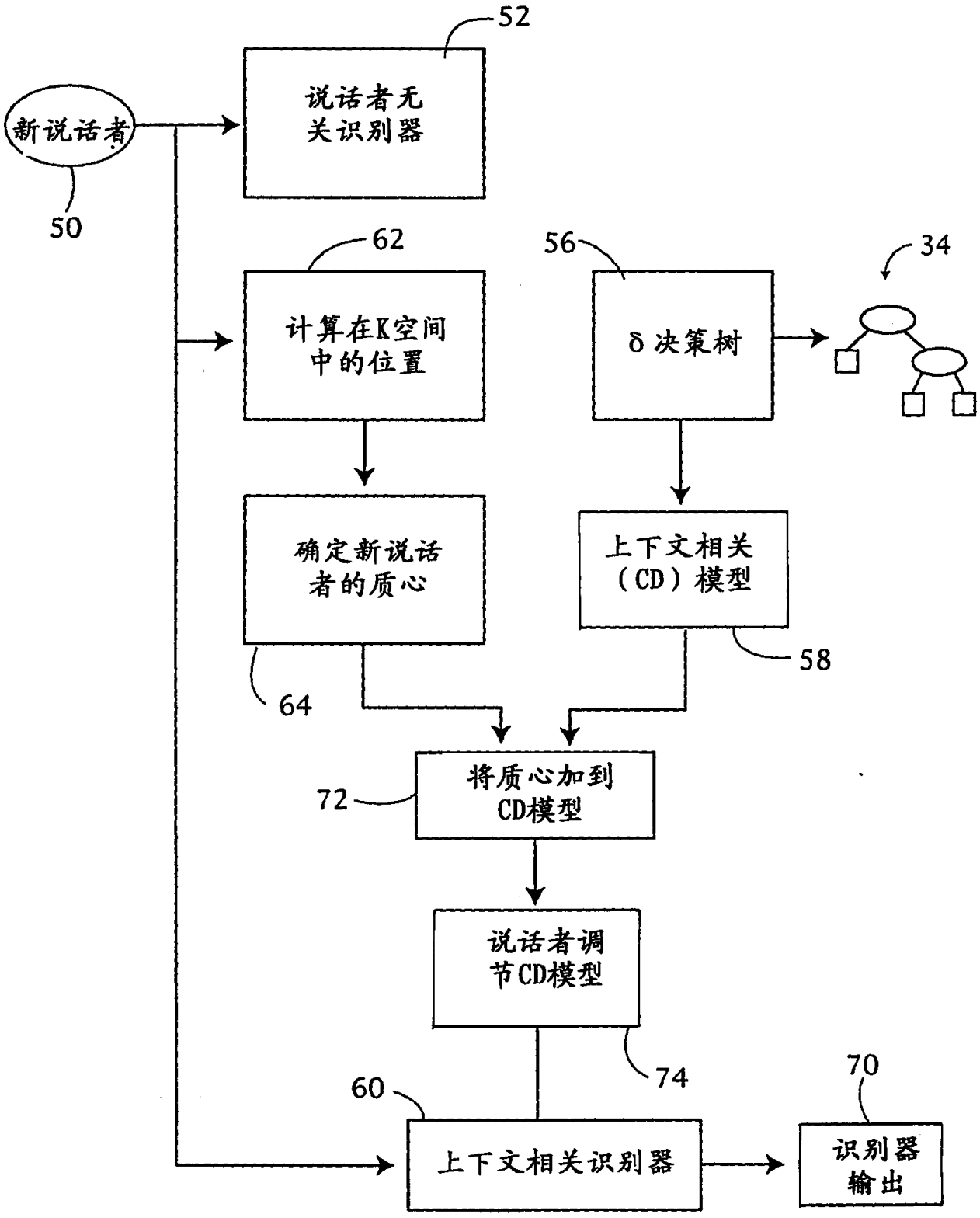

新说话者
50
52
说话者无关识别器
62
计算在K空间中的位置
56
δ决策树
34
确定新说话者的质心
64
上下文相关（CD）模型
58
72
将质心加到CD模型
说话者调节CD模型
74
60
上下文相关识别器
70
识别器输出

图4

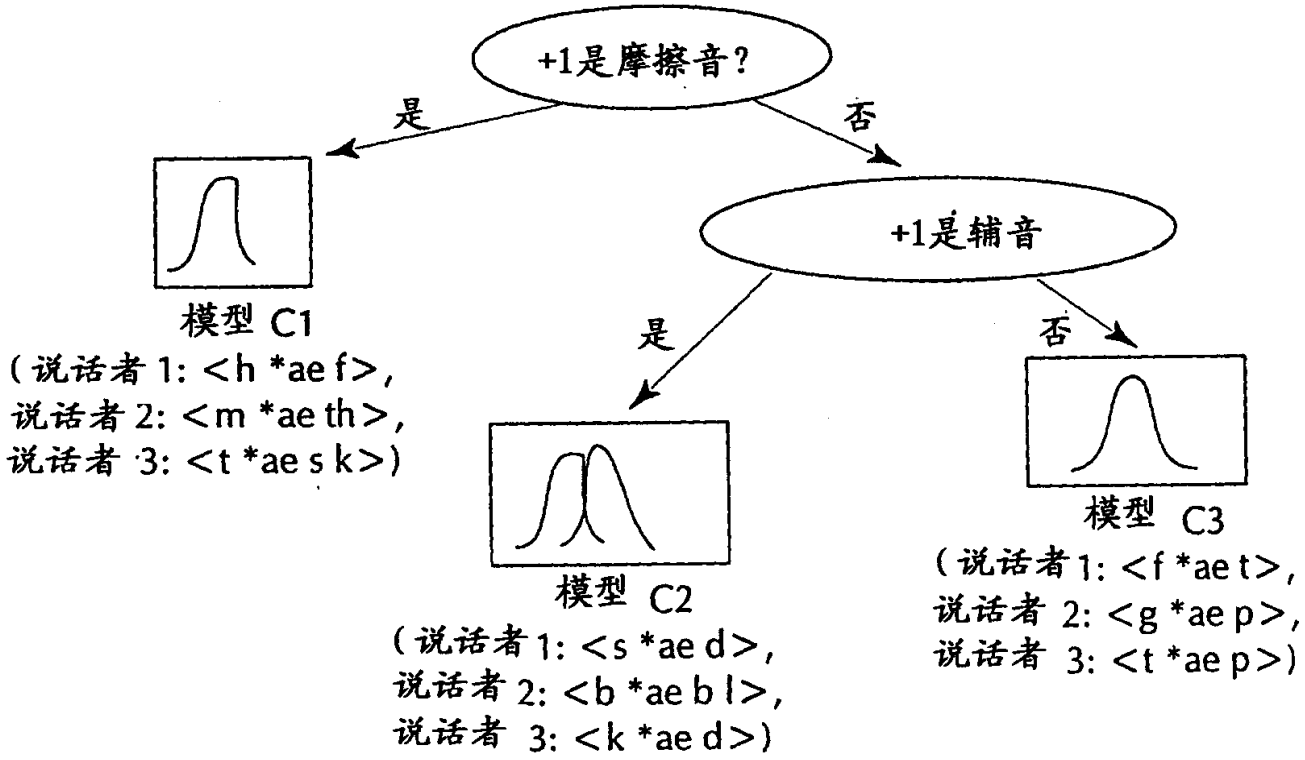

+1是摩擦音?
是
否
+1是辅音
是
否
模型 C1
(说话者 1: <h *ae f>,
说话者 2: <m *ae th>,
说话者 3: <t *ae s k>)
模型 C2
(说话者 1: <s *ae d>,
说话者 2: <b *ae b l>,
说话者 3: <k *ae d>)
模型 C3
(说话者 1: <f *ae t>,
说话者 2: <g *ae p>,
说话者 3: <t *ae p>)

图5

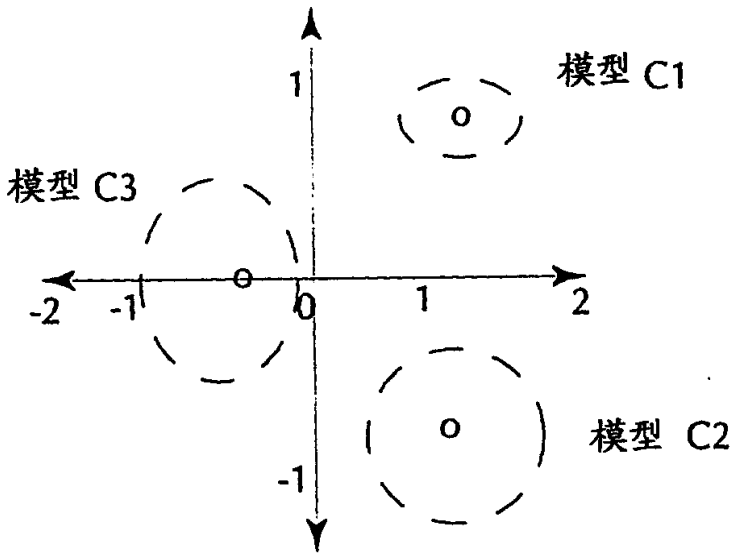

模型 C1
模型 C3
模型 C2
1
-1
-2
0
1
2

图6

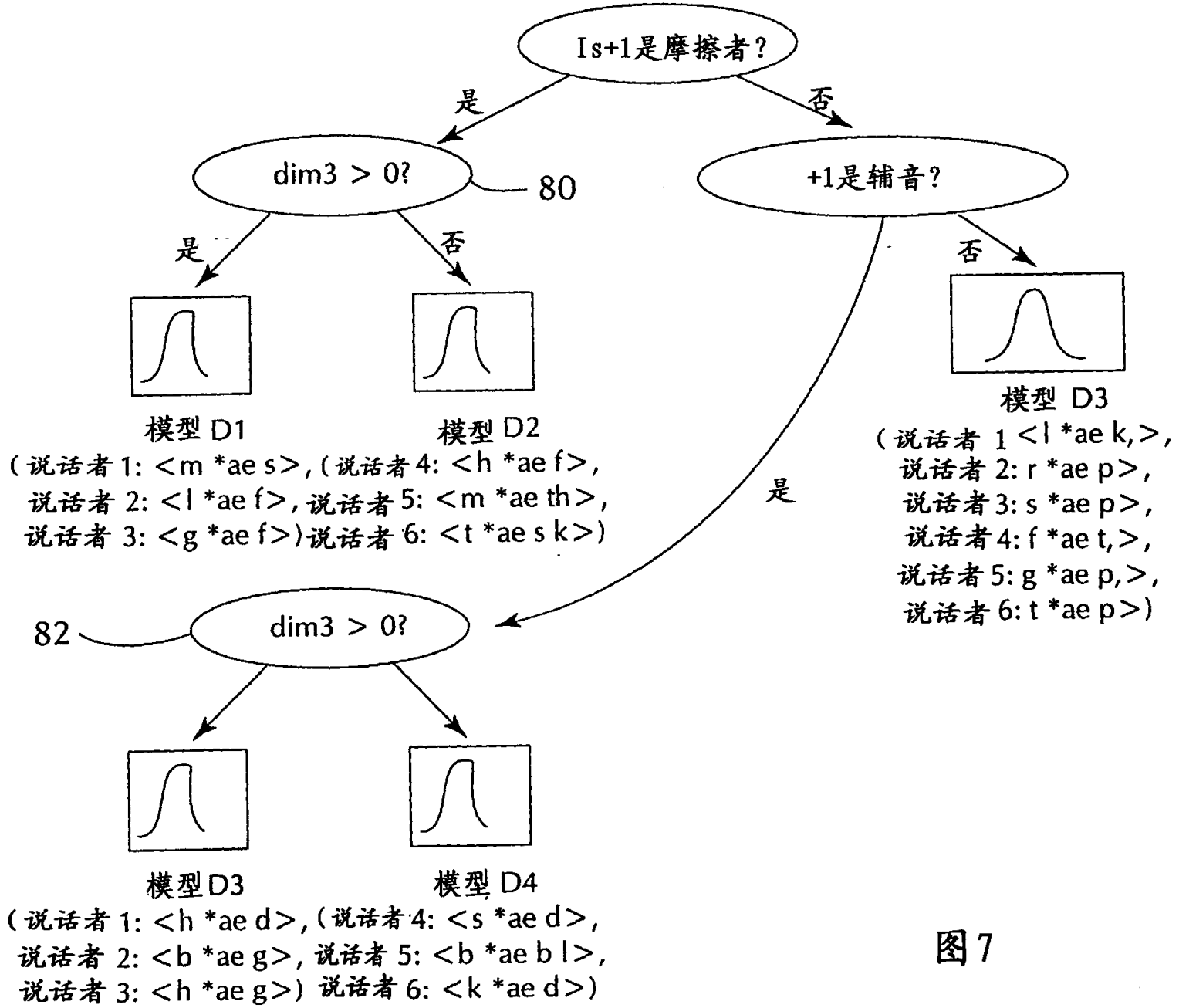

Is+1是摩擦者?
是
否
dim3 > 0?
80
+1是辅音?
是
否
否
模型 D1
(说话者 1: <m *ae s>, (说话者 4: <h *ae f>,
说话者 2: <l *ae f>, 说话者 5: <m *ae th>,
说话者 3: <g *ae f>) 说话者 6: <t *ae s k>)
模型 D2
模型 D3
(说话者 1 <l *ae k,>,
说话者 2: r *ae p>,
说话者 3: s *ae p>,
说话者 4: f *ae t,>,
说话者 5: g *ae p,>,
说话者 6: t *ae p>)
是
82
dim3 > 0?
模型D3
(说话者 1: <h *ae d>, (说话者 4: <s *ae d>,
说话者 2: <b *ae g>, 说话者 5: <b *ae b l>,
说话者 3: <h *ae g>) 说话者 6: <k *ae d>)
模型 D4

图7

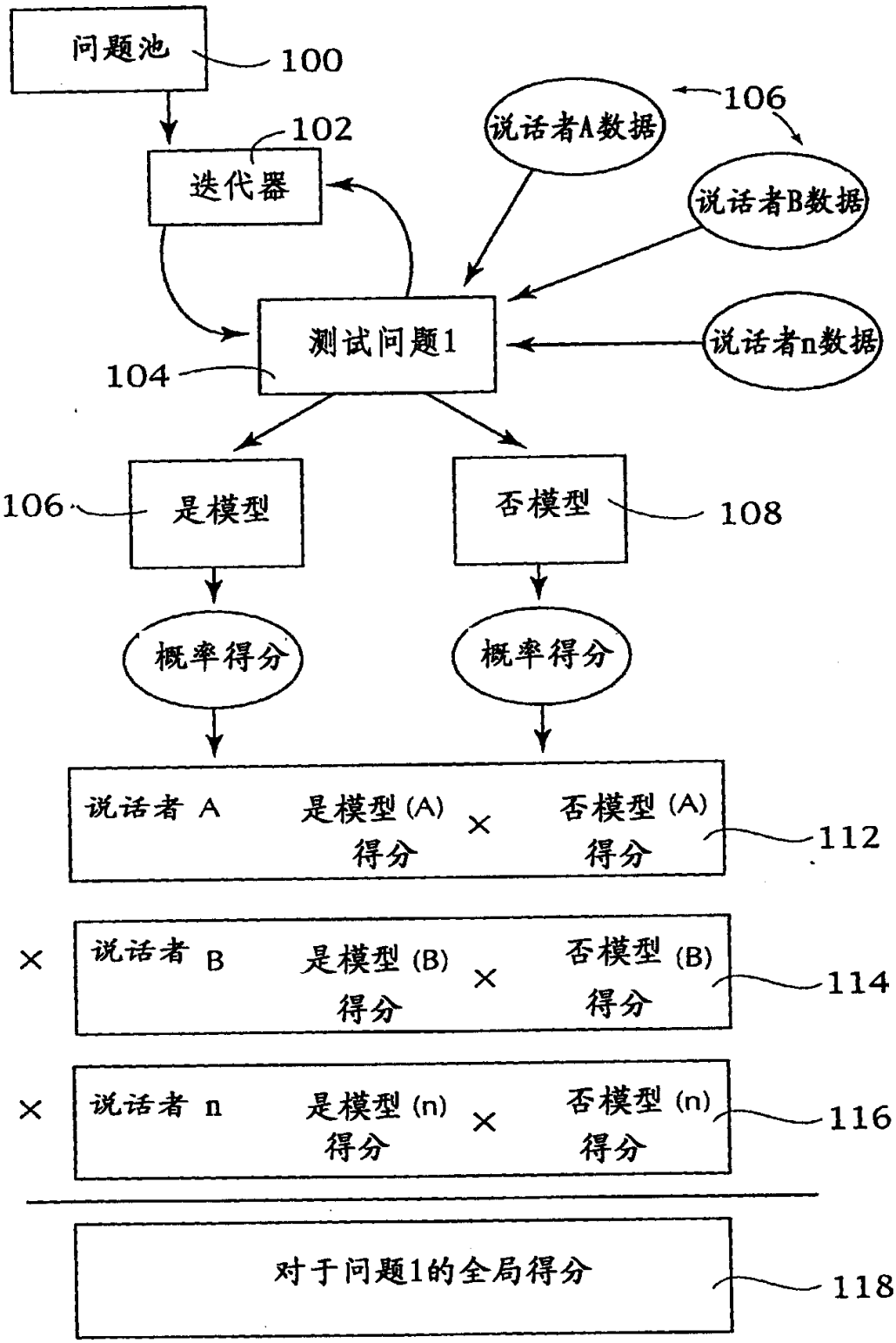

问题池
100
迭代器
102
说话者A数据
106
说话者B数据
测试问题1
104
说话者n数据
106
是模型
否模型
108
概率得分
概率得分
说话者 A
是模型 (A) 得分
×
否模型 (A) 得分
112
×
说话者 B
是模型 (B) 得分
×
否模型 (B) 得分
114
×
说话者 n
是模型 (n) 得分
×
否模型 (n) 得分
116
对于问题1的全局得分
118

图8