(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104541324 B
(45)授权公告日 2019.09.13

(21)申请号 201380031695.3
(22)申请日 2013.06.26
(65)同一申请的已公布的文献号
申请公布号 CN 104541324 A
(43)申请公布日 2015.04.22
(30)优先权数据
P.403724 2013.05.01 PL
(85)PCT国际申请进入国家阶段日
2014.12.17
(86)PCT国际申请的申请数据
PCT/EP2013/063330 2013.06.26
(87)PCT国际申请的公布数据
WO2014/177232 EN 2014.11.06
(73)专利权人 克拉科夫大学
地址 波兰克拉科夫
(72)发明人 巴尔托什·焦尔科
托马什·贾奇克
(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 曾贤伟 周捷
(51)Int.Cl.
G10L 15/14(2006.01)
(56)对比文件
US 7346510 B2,2008.03.18,
CN 102411931 A,2012.04.11,
US 6292776 B1,2001.09.18,
US 6542866 B1,2003.04.01,
US 2003/0212552 A1,2003.11.13,
Todd A.Stephenson等."Automatic Speech Recognition using Dynamic Bayesian Networks with both Acoustic and Articulatory Variables".《International Conference on Spoken Language Processing,2000》.2000,第2卷951-954. (续)
审查员 可杨

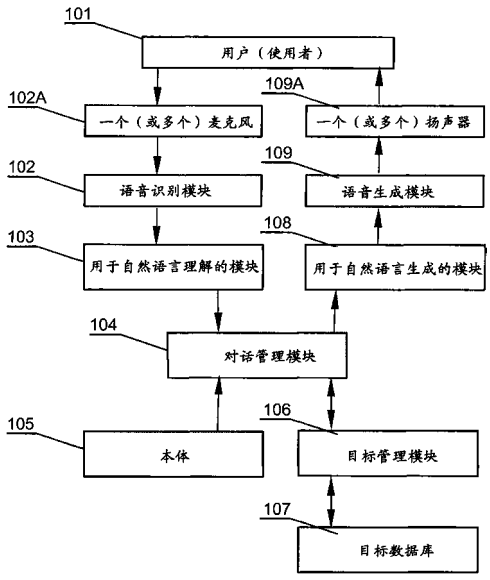
权利要求书1页 说明书8页 附图4页

(54)发明名称

一种使用动态贝叶斯网络模型的语音识别系统和方法

(57)摘要

一种用于语音识别的计算机实现的方法,包括以下步骤:通过输入设备(102A)的方式,记录(201)表示语音的电信号,并将该信号转换为频域或时-频域(202),基于动态贝叶斯网络在分析模块中分析信号(205),被配置为基于观察到的信号特征(OA,OV)生成单词(W)的假设和它们的概率,基于特定单词(W)假设和它们的概率,识别(209)出表示语音的电信号所对应的文本。该方法的特征在于,将观察到的信号特征(308-312)输入到分析模块(205)中,其中,所述观察到的信号特征是在至少两条并行信号处理线(204a,204b,204c,204d,201a)上,为频域或时-频域(202)中信号而确定的,其中每条线上的时间片段都不同,以及,在分析模块(205)中对至少两个不同的时间片段分析观察到的信号特征(308-312)之间的关系。



CN 104541324 B

[转续页]

[接上页]

(56)对比文件

Timothy J.Hazen."Visual model structures and synchrony constraints for audio-visual speech recognition".《IEEE Transactions on Audio Speech & Language Processing》.2006,第14卷(第3期),1082-1089.

Xavier Domont等."Hierarchical spectro-temporal features for robust speech recognition".《IEEE International Conference on Acoustics,2008》.2008,4417-4420.

John N.Gowdy等."DBN based multi-stream models for audio-visual speech recognition".《IEEE International Conference on Acoustics,2004》.2004,993-996.

Kate Saenko and Karen Livescu."An asynchronous DBN for audio-visual speech recognition".《2006 IEEE Spoken Language Technology Workshop》.2006,154-157.

1.一种用于语音识别的计算机实现的方法,包括以下步骤:

-通过输入设备(102A),记录(201)表示语音的电信号,并将该信号转换为频域或时-频域(202),

-基于动态贝叶斯网络在分析模块中分析信号,所述分析模块被配置为基于观察到的声学信号特征或声学及视觉信号特征生成单词的假设和它们的概率,

-基于特定单词假设和它们的概率,识别(209)出表示语音的电信号所对应的文本,

该方法的特征在于,

-在识别过程期间内,将观察到的声学信号特征或声学及视觉信号特征(308-312)输入到分析模块(205)中,该观察到的信号特征(308-312)是对于在至少两条并行信号处理线路(204a,204b,204c,204d,201a)上的频域或时-频域(202)中的信号而确定的,在所述至少两条并行信号处理线路中的至少两条信号处理线路被配置为确定声学信号特征,并且至少一条线路被配置为接收以与其他线路不同方式分段的信号,以使得所述至少一条线路的至少一些帧能够在所述动态贝叶斯网络的不同状态之间被公用,

-以及在分析模块(205)中对至少两个不同的时间片段分析所观察到的信号特征(308-312)之间的关系。

2.如权利要求1所述的方法,其中时间片段具有预定的持续时长。

3.如权利要求1或2所述的方法,其中时间片段取决于语音片段的内容。

4.如权利要求3所述的方法,所述语音片段的内容包括音素、音节和单词。

5.如权利要求1所述的方法,特征在于,在分析模块(205)中,定义描述模型的变量之间的确定性和概率性关系,而至少对观察到的信号特征与当前状态的关联定义概率性关系。

6.如权利要求1所述的方法,其特征在于同时分析不同的观察到的信号特征。

7.一种用于语音识别的、由计算机实现的系统,包括:

-用于记录代表语音的电信号的输入设备(102A),

-用于将表示语音的所记录的电信号转换为频域或时-频域的模块(202),

-基于动态贝叶斯网络的分析模块(205),被配置为分析表示语音的信号,并且被配置为基于观察到的声学信号特征或声学及视觉信号特征生成单词的假设和它们的概率,

-用于基于已定义的单词的假设与它们的概率来识别表示语音的电信号所对应的文本的模块(209),

该系统的特征在于进一步包括:

-至少两个信号参数化模块(204a,204b,204c,204d,201a),用于在识别过程期间内,通过分析模块(205)确定在至少两条并行信号处理线路上的至少两个观察到的声学信号特征或声学及视觉信号特征(308-312),在所述至少两条并行信号处理线路中的至少两条信号处理线路被配置为确定声学信号特征,并且至少一条线路被配置为接收以与其他线路不同方式分段的信号,以使得所述至少一条线路的至少一些帧能够在所述动态贝叶斯网络的不同状态之间被公用,

-其中分析模块(205)被配置为对至少两个不同的时间片段分析所观察到的信号特征(308-312)之间的相关性。

8.一计算机可读介质,存储计算机可执行指令,当在计算机上执行时,所述指令执行根据权利要求1-6的任何一个的计算机实现的方法的所有步骤。

一种使用动态贝叶斯网络模型的语音识别系统和方法

技术领域

[0001] 本发明的目标是实现一种使用贝叶斯网络的语音识别系统和方法。特别地，涉及一种自动语音识别系统，其可以在用于广告和信息意图的对话系统中应用。对话系统的实施可以采用报亭或货摊的形式，其与顾客或观众开始一对话，并且将呈现适当的多媒体内容。

背景技术

[0002] 语音识别系统在日常生活中变得越来越常见。比如，它们可以被用于信息电话中心，比如为公共交通所用。然而，这些系统仍然经常依赖于键盘和文本作为输入信息源，而不是使用语音作为输入信息源而运行。

[0003] 已知各种类型的计算机化的交互报亭被用于与用户进行对话。比如，美国专利US6256046公开了一种在计算机化的报亭内的有源公共用户交互接口，其通过处理视觉数据、通过使用动作和色彩分析以检测表示用户出现的环境中的改变来感知用户。交互空间被定义，系统记录其环境的初始模型，该环境随着时间更新，以反映出不活动对象的添加或减去，并且补偿光的改变。该系统研发了针对移动对象的模型，因此当他们在交互空间的附近移动时，该系统能够跟踪用户。一立体摄像系统进一步增强了该系统感知位置和移动的能力。该报亭呈现出音频和视频的反馈来反映其“看到”了什么。

[0004] 美国专利申请US20080204450公开了一种用于提供虚拟世界的系统、方法和程序产品，其中主动提供的广告被嵌入在自动虚拟角色中。所提供的系统包括：用于将广告虚拟角色引入虚拟世界的注册系统；用于定向用户虚拟角色以实现广告虚拟角色所传递的广告内容的定向系统；用于定义广告虚拟角色如何在虚拟世界中移动的移动系统；以及用于定义广告虚拟角色如何将广告内容传递给用户虚拟角色的广告传递系统。

[0005] 诸如上述的已知的对话系统的缺陷包括，在与用户进行复杂对话时缺乏足够的语音识别能力。

[0006] 美国专利US7203368公开了一种模式识别程序，其使用HMM(隐马尔科夫模型)和CHMM(耦合隐马尔科夫模型)形成了分级的统计模型。分级的统计模型支持具有多个超节点的父层和具有与每一个父层的超节点相关联的多个节点的子层。经过训练之后，分级统计模型使用从数据集中提取的观察矢量来寻找基本的最优状态序列片段。对该过程进行改进是很有利的。

[0007] 一个比基于HMM的方案少一些限制的、更加通用的方案，是将贝叶斯网络用于语音识别。使用贝叶斯网络的方案包括动态贝叶斯网络(DBN)，已经在以下出版物中被公开：

[0008] -M.Wester，J.Frankel，以及S.King所著的：″Asynchronous articulatory feature recognition using dynamic Bayesian networks″(Proceedings of IEICI Beyond HMM Workshop，2004)(“使用动态贝叶斯网络的异步分节特征识别”，公开于2004年HMM研讨会的IEICI会议录)；

[0009] -J.A.Bilmes和C.Bartels所著的″Graphical model architectures for speech

recognition″,IEEE Signal Processing Magazine,vol.22,pp.89-100,2005(“用于语音识别的图形模型构造”,公开于IEEE信号处理杂志,2005年,vol.22,pp.89-100);

[0010] -J.Frankel,M.Wester和S.King所著的″Articulator/feature recognition using dynamic Bayesian networks″,Computer speech and Language,vol.21,no.4,pp.620-640,October 2007(“使用动态贝叶斯网络的发音器/特征识别”,公开于2007年10月,计算机语音和语言vol.21,no.4,pp.620-640)。

[0011] 使用贝叶斯网络的语音识别方法依据特征矢量对声音时长进行建模。在DBN中,使用表示声音的变量替换表示时长的变量已经变得可能。然而,所有的现有技术的方案都在预定的时间范围内进行语音分析。

[0012] 考虑到之前的现有技术,有必要设计和实现一种允许提高人类和机器之间的对话效率的语音识别系统和方法。

发明内容

[0013] 本发明的目的在于提供一种用于自动语音识别的计算机实现的方法,包括以下步骤:通过输入设备记录表示语音的电信号,并将该信号转换至频域或时-频域,基于DBN在模块分析中分析信号,被配置为基于观察到的信号特征(OA,OV)生成单词(W)的假设和它们的概率,基于特定单词(W)假设和它们的概率识别出表示语音的电信号所对应的文本。该方法的特征在于,将观察到的信号特征输入到分析模块中,该观察到的信号是对于多个时间段、在至少两条并行信号处理线上的频域或时-频域中为信号而确定的,其中在每条线上的时间片段都不同,并且,在分析模块中对至少两个不同的时间片段分析观察到的信号特征之间的关系。

[0014] 优选地,时间片段具有预定的时长。

[0015] 优选地,时间片段取决于语音片段的内容,比如音素(phonemes)、音节(syllables)、单词(words)。

[0016] 优选地,该方法进一步包括在分析模块定义描述模型的变量之间的确定性和概率性关系,而概率性关系至少被定义用于将观察到的信号特征与当前状态(Sti)进行关联。

[0017] 优选地,该方法进一步包括同时分析不同的观察到的信号特征(OA,OV)。

[0018] 本发明的另一个目标是实现用于语音识别的、计算机实现的系统,包括用于将代表语音的电信号进行记录的输入设备,用于将表示语音的记录的电信号转换为频域或时-频域的模块,基于DBN的分析模块,被配置为分析表示语音的信号,并且,被配置为基于观察到的信号特征(OA,OV)生成单词(W)的假设和它们的概率,用于基于已定义的单词(W)的假设与它们的概率识别表示语音的电信号所对应的文本的模块。该系统进一步包括至少两个信号参数化模块,用于为分析模块在至少两条并行信号处理线上为每条线上不同的时间片段确定至少两个观察到的信号特征,其中分析模块被配置为分析在至少两个不同的时间片段上,观察到的信号特征之间的相关性。

[0019] 本发明的目标是还提供一种计算机程序,包括当所述程序在计算机上运行时,用于执行根据本发明的计算机实现的方法的所有步骤的程序代码装置,还有存储计算机可执行指令的计算机可读介质,当在计算机上执行该指令时,该指令执行根据本发明的计算机实现的方法的所有步骤。

[0020] 附图简要说明

[0021] 已经在附图中的示例性实施例中公开了本发明的目标，其中：

[0022] 附图1示出了依据本发明的系统的方框图；

[0023] 附图2示出了自动语音识别过程的方框图；

[0024] 附图3示出了在不同长度的并行时间周期内使用DBN对语音进行模型化；

[0025] 附图4描述了使用与附图3中示出的DBN相似的DBN进行单词序列解码的例子(为了示例性目的，已经被简化的版本)。

具体实施方式

[0026] 附图1示出了依据本发明的系统的方框图。该系统可以被用于交互性广告或其它提供信息的对话系统中。对话尽可能地接近现实中的对话。由于使用诸如模式识别、语义分析的技术，使用语音合成所伴随的本体认知和自然语言生成，这种假设可以被实现。

[0027] 在可能使用本发明的对话系统中，可以包括高质量显示器或图像投影器。在优选的实施例中，对话系统还配备有用户出现检测器，或者在更加高级的例子中，配备有诸如生物检测器、面部识别模块等的用户特征检测器。对话系统还可以包括方向性麦克风，以能够更加有效地采集语音。

[0028] 输出信息被调整为对话的情景并且根据用户爱好而被确定。对话系统优选地还输出与用户对话的视觉虚拟角色或用户的图像。

[0029] 采用语音识别的对话系统与用户或多个用户101交互地交谈。用户101通过说话将问题输入至声音输入模块，比如输入至麦克风102A中。被麦克风记录的声音通过语音识别模块102进行处理，之后传送至用于识别自然语言的模块103。

[0030] 用于理解的模块103负责在预期的响应情景中解释用户101的状态识别假设，其采用使得它们可以被机器所理解并且可以容易地并且快速地被处理的方式。比如，若在旅游信息现场实施该系统，基于伴随着语音假设概率的语音假设列表的用于理解的模块103需要确定说话者是否在寻找一个具体的地点，如果是的话，这个地点是什么，或者服务、公共交通运行的时间信息等。在最简单的版本中，为了实现该意图，该模块使用关键词，但是它们也可以使用基于句法模型(例如，句子解析器)和/或语义(比如，单词网或语音HMM)的更加高级的方案，这些内容在D.Jurafsky，J.H.Martin，″Speech and Language Processing″，Second Edition，Pearson Education，Prentice Hall，2009中被公开。

[0031] 被自然语言理解模块103处理之后，句子或句子假设被输入至对话管理模块104(比如，诸如在D.Jurafsky，J.H.Martin，″speech and Language Processing″，Second Edition，Pearson Education，Prentice Hall，2009中所描述的)，其通过适当地询问本体模块105，与目标管理模块106和目标数据库107合作，决定将要对用户询问呈现的响应。

[0032] 本体模块105包括关于世界的有规则的知识，比如，在某些特定类型中可以购买哪些产品、人们将与所选择的产品一起购买哪种产品等等的信息。本体模块可以包括来自社会服务的额外的不同类型的数据，例如，核实正在与用户进行对话的朋友是否在该用户访问的城市中等等。本体模块还可以包括任何其它的实际知识，使用一种计算机或其它机器可以处理它的方式将其系统化。

[0033] 目标管理模块106被用于在计算机中实现商业、广告、谈判等的已知规则，其用于

指引专业人员(比如,商业雇员),其职责被依据本发明的系统所执行。

[0034] 确定响应的内容之后,自然语言中的响应在生成自然语言的模块108中产生,紧接着被送至语音生成模块109中。通过安装在系统内的扬声器或其它输出设备109A将以语音形式生成的响应输出至用户101。

[0035] 本发明中所使用的关键元件是使用贝叶斯网络进行分析的计算机实现的模块。贝叶斯网络能够对复杂的现象进行模型化,其中分开的元素可以彼此依赖。将基本模型产生为有向的、非循环的图形,其中节点表示模型的分开的元素(随机变量),而边表示这些元素之间的相关性。

[0036] 附加地,边被分配有概率值,该概率值指定在另一个事件假定了一特定值的情况下,发生多个事件中的一个事件。通过使用贝叶斯法则,复杂的条件概率可以在贝叶斯网络的特定路径中计算。这些概率可以被用于推断网络中的各个元素将要采用的值。

[0037] 每一个网络变量必须有条件地独立于不和它连接的其它变量。使用这种方式生成的图形可以被解释为事件的紧密表示(compact representation),考虑到图形的节点之间的条件独立性,这些事件和假定发生的累积概率。

[0038] DBN可被用于语音识别。之后,节点不仅仅表示单个随机变量,还表示变量序列。这些可以被解释成时间序列,其允许根据时间的经过来对语音模型化。因此,多个连续的观察状态给出对于到达最终状态的明确路径的判断。

[0039] 使用标准贝叶斯网络基于声音的期望时长,依赖于发音者特征的矢量。网络具有用于每个特征的单个离散变量和用于声音时长的单个连续变量。网络描述了特征之间的关系。表示特征的节点的值依赖于输入到网络中的值,并且可选地,依赖于其它特征。表示时长的节点值是隐层(如在HMM中一样),仅仅直接依赖于从其它节点接收的值。

[0040] 引入DBN允许用代表声音的变量代替代表时长的变量。具有特征之间的关系的整个网络使用这样的方式被复制,这种方式为:多个网络中的一个网络表示在时刻t-1处所分析的信号,并且下一个网络表示在时刻t所分析的信号。这两个网络通过边相连,这些边具有可能随着时间变化的状态之间的转移的概率值。

[0041] 应注意到,本发明并不只限制于具有两个子网络的实施例。可以有更多的子网络,每一个子网络用于后续的时刻。典型地,可以有数百或数千个网络。可以将这种结构多次复制至多个后续的时刻。附加地,在某些情况下,这种局部贝叶斯网络结构可以在离散时间之间修改其本身。

[0042] DBN模型还可以被用于将来自不同源的信号的相关信息联系起来,比如声学特征和视觉特征(比如嘴唇运动)。这种类型的系统对于在恶劣的声学条件的地方的应用特别有用。在例如街道、飞机场、工厂等地方,低信噪比(SNR)值使得仅仅使用源自声学路径的信息极大地降低所获取的结果的质量。增加从对相同类型的噪声不敏感的其它信号类型获取的信息,就会消除这些困难,并且使得在这些地方也可以使用语音识别系统。

[0043] 发明人已经注意到,当用于语音分析时,与HMM方法相比,贝叶斯网络具有更少的限制条件。

[0044] 附图2示出了语音识别过程的方框图。下面的描述也会参考附图3中的某些特征,附图3示出了在不同长度的时间周期上使用DBN对语音进行模型化。

[0045] 此处通过这样的方法使用DBN对语音进行模型化,即分开的观察表示不同的时

长——如图3中所示。这些不同的时长可以是预定长度的片段，比如5ms，20ms，60ms，这依赖于语音片段的内容，比如音素、音节、单词或者两种类型的结合，比如，5ms，20ms，音素，单词。

[0046] 所呈现的方法允许提取不同的信息类型，并且对所获取的特征进行直接融合，这是由于使用DBN模型对状态概率进行了估计(附图3中的St1至St6)。

[0047] 基于那些描述模型的变量之间的两种类型的关系，在DBN中进行推断：确定性关系(在图3中用直箭头表示)和概率性关系(在图3中用波浪形箭头表示)。

[0048] 基于那些公知事实定义确定性关系，比如，当分析一个给定单词Wti时，已知第一个音素Pti的类型和位置Wps。之后，通过知晓是否已经发生了从该音素到下一个音素的转移Ptr，可以确定当前音素在单词中的位置：若还没有出现音素转移，那么t+1时刻的Wps等于t时刻的Wps，若已经观察到了前述的转移，那么，t+1时刻的Wps等于Wps+1。

[0049] 可以使用相似的方式获取到从一个单词到另一个单词的转移Wtr有关的信息。出现从用音标标出的单词的最后一个音素进行的转移，暗示着有必要对另一个单词Wti进行分析。

[0050] 另一种类型的关系是概率性关系。为了基于其之间存在概率性关系的变量进行推断，有必要确定限定发生这些事件的概率(概率密度函数--PDF)的函数。该类型的关系用于将信号的观察到的特征和当前状态Sti进行关联。优选的PDF函数是高斯混合模型-GMM。

[0051] 某些关系可以既是确定性也是概率性的，比如连续单词Wti。一旦没有发生从一个单词到另一个单词的转移，该关系就是确定性的——该单词与t-1时刻的单词一样。一旦发生转移，使用从语言模型中得到的知识，以概率性方式确定下一个单词Wti+1。

[0052] 基于声学特征的观察值在DBN中进行推断。然而，任何观察都可能受到测量误差(s)的影响。引入属于相同组(图3中的OA11，OA23和OA33或者OV11和OV23)的、相关的、随时间变化的观察量之间的概率关系同样可以减小这种误差。

[0053] 状态Ｓti和之前的状态Sti-1用于估计观察量是说出给定音素(附图3中的Pt1至Pt6)的结果的概率。

[0054] 给定音素的出现还和转移状态Ptr在一定概率上相关。音素Pti、音素转移率Ptr、音素在单词Wps中的位置和从单词Wtr进行的转移率允许估计假设记录的声音中包含单词W的正确性。

[0055] 语音具有某些频率特征和能量特征在短时期内几乎保持恒定的特性。然而，在长时期内它们显著变化。然而，并没有定义第一和第二种情况出现时的特殊时刻，因而，使用DBN模型非常有利。不同片段中的观察量之间的关系可以呈现，也可以不呈现。

[0056] 比如，对于具有四个时间长度的配置的变形而言，它们假定5ms、20ms、音素和单词用于并行分析。可能存在不同的模型配置，例如，存在全部四个范围之间的关系，还可能仅仅在5ms和20ms层与音素层之间存在关系，仅仅在20ms层和音素层之间存在关系，或者仅仅在音素层和单词层之间存在关系。

[0057] 此外，每一个范围可以具有一些观察量类型，其与不同种类的语音特征相关。比如，它们中的一个可以是频率特征矢量，另一个可以是能量，在一个可以是视觉特征矢量。这些同样可以是相同种类的声学特征，但是通过不同的方法获得(比如WFT(小波傅里叶变换)、MFCC(梅尔倒谱系数))，还可以是在不同的时间范围内使用同样的方法所获取的声学

特征,比如在20ms的滑动窗中、或是在50ms的滑动窗中,两者均是每隔10ms提取一次。

[0058] 另外,某些范围可以仅仅在特别类型的特征分析中出现,并且,在其它类型中并不出现(附图3——声学特征1观察量(308)持续60ms、声学特征2观察量(310)持续20ms、以及视觉特征1观察量(309)持续30ms)。

[0059] 可能存在在分析过程中使用(也是同时使用)的更多类型的描述信号的特征,比如,声音的有声/无声描述、共振峰频率或基音频率。

[0060] 附图2中公开的方法开始于步骤201,获取语音信号。下一个步骤202通过诸如WFT的方法或使用短时傅里叶变换(STFT)进行的时频变换将信号处理成频率域。可以应用其它变换以允许对包括在不同时刻上的不同频率子带内的信息(比如信号能量)进行定量描述。

[0061] 接着,在步骤203,时-频谱被分割成恒定的帧,比如5ms、20ms、60ms等等,或者依据预定的算法进行分割,诸如,在例如如下中所公开的:

[0062] -P.Cardinal,G.Boulianne和M.Comeau所著的"Segmentation of recordings based on partial transcriptions",Proceedings of Interspeech,pp.3345-3348,2005;或者

[0063] -K.Demuynck和T.Laureys所著的"A comparison of different approaches to automatic speech segmentation",Proceedings ofthe 5th International Conference on Text,Speech and Dialogue,pp.277-284,2002;或者

[0064] -Subramanya,J.Bilmes和C.P.Chen所著的"Focused word segmentation for ASR",Proceedings of Interspeech 2005,PP.393-396,2005。

[0065] 分割模块(203)将频谱分析过程分成多个线程,独立地对其进行参数化。

[0066] 线程数目可以不是前面所述的4个。附图2中的例子使用四个分开的线程,其具有5ms的帧——204a、20ms的帧——204b、音素——204c和单词——204d,而在方框204a至204d中从每个线程中提取的特征,表示了在特定时刻的语音。这些参数化方框使用诸如MFCC、感知线性预测(PLP)的处理算法,或者是其它的,比如:

[0067] -H.Misra,S.Ikbal,H.Bourlard和H.Hermansky所著的"Spectral entropy based feature for robust ASR",Proceedings of ICASSP,pp.1-193-196,2004;和/或

[0068] -L.Deng,J.Wu,J.Droppo和A.Acero所著的"Analysis and comparison of two speech feature extraction/compensation algorithms",IEEE Sigral Processing Letters,vol.12,no.6,pp.477-480,2005;和/或

[0069] -D.Zhu and K.K.Paliwal所著的"Product of power spectrum and group delay function for speech recognition",Proceedings of ICASSP,pp.1-125-128,2004.

[0070] 从模块204a-204d获取的特征与诸如信号能量和视觉特征矢量的观察量201a一起被传至DBN205。DBN模型,使用对BN的近似推测的其自己的嵌入的算法(比如变分消息传播(variational message passing)、期望传播(Expectation Propagation)和/或吉布斯采样(Gibbs Sampling)),利用语音识别中所使用的动态编程算法(比如维特比解码和/或Baum-Welch),并基于词典206和语言模型207的内容(比如单词的二元语法)来确定单词假设并且计算它们的概率。在大多数情况下,这些假设会部分重叠,因为DBN可以在相同的时间长度上呈现不同的假设。在进一步的语言模型208中(优选地,比DBN中使用的第一个语言

模型更加高级)可以继续处理这些假设,以便获取识别的语音文本209。

[0071] 附图3公开了示例性DBN结构。术语W301表示单词,Wtr302表示单词转移率,Wps303表示在特定单词中的音素的位置,Ptr304表示音素转移率,Pt 305表示音素,Spt 306表示之前的状态,S 307表示状态,OA1 308表示在60ms的时间窗中,观察到的第一种类型的声学特征,OV1 309表示在30ms时间窗中观察到的第一种类型的视觉特征,OA2 310表示在20ms时间窗中观察到的第二种类型的声学特征,OA3 311表示在10ms时间窗中观察到的第三种类型的声学特征,同时,OV2 312表示在10ms时间窗中观察到的第二种类型的视觉特征。

[0072] 箭头表示变量之间的相关度(依赖度),如前面所描述的一样。通过条件概率分布(CPD)定义转移率,其是基于训练数据,在贝叶斯网络的训练过程中被计算出的。

[0073] 附图4描述了使用附图3中所示的DBN用于解码单词序列的一个例子。与图3中的区别在于,为了实现语音识别,对于不同长度的两帧,这里使用了一种类型的信号声学特征。网络呈现出解码短语:“Cat is black”的过程——语音转换是:/kæt iz blæk /。音素状态依赖于两种类型的观察量O1和O2。时刻t处的前一状态306是时刻t-1处的状态307的完美复制。依赖于在单词303中的当前位置、音素转移至另一个音素304的出现率、音素的状态306和前一状态307,对单词301的后续音素进行分析。若转移概率值大于0.5,则出现音素转移。附图3中的贝叶斯网络的分开的节点的符号被这些状态的值取代。对于302和304而言,它们是这样的值:T(真)/F(假)分别表示后续单词或后续音素之间出现转移或未出现转移。对于音素在单词303中的位置而言,它是当前所分析的音素的索引(1-3指代单词‘cat’、1-2指代单词‘is’、1-4指代单词‘black’)。只有在时刻t-1的前述时刻中,获取到音素304的转移值为‘T’时,才会出现音素索引的变化。此外,只有当在特定单词中的上一个索引的音素转移304的时刻所获得的单词转移302发生的情况下,单词301才会变化。在这种情况下,作为使用语言模型的结果,连续的单词之间的关系从确定性改变成概率性。在附图上部的表中示出二元语法语言模型(应用多对单词的模型)的示例性值。此外,还公开了示例性的语言模型中初始单词概率值。通过同时处理不同时长内的片段和各种类型的特征所达到的技术效果是提高了语音识别质量,因为在一种类型的时间片段下会更好地识别以各种方式说出的一种类型的音素,并且其它的需要不同类型的片段,但是,为每一种音素类型确定合适的分析时间窗是复杂的。此外,考虑到在更局部的时间片段上精确地提取信息,某些特征表示稳定特性,同时其它的特征需要更全局的时间片段。使用如图3所示的结构,可以一次提取两种类型的特征。在传统的系统中,只是通过局部特征或只是通过全局特征来提取所用的一些信息。此外,比如,视觉特征可以具有与与声学特征不同的持续时间,即,比如,对要说出声音的嘴唇的观察量可以持续长于或短于特定声音的时间。

[0074] 本领域技术人员能够容易地认识到,前述的语音识别方法可以被一个或多个计算机程序执行或控制。这些计算机程序典型地使用诸如个人电脑、个人数字助理、移动电话、数字电视的接收器和解码器、信息亭或类似的计算设备中的计算资源来被执行。应用被存储于非易失性存储器中,比如闪存,或者被存储于易失性存储器中,比如RAM,并且通过处理器执行应用。这些存储器是用于存储计算机程序的示例性存储介质,所述计算机程序包括执行依据本文所公开的技术概念的计算机执行的方法的所有步骤的计算机指令。

[0075] 尽管已经参考特定的优选实施例限定、描述和描写了本文公开的发明,在前述具体方式中的这些参考和实施的例子并不对本发明作任何限制。然而,在不违背技术概念的

更广范围的条件下，很显然可以对本发明做出各种修改和改变。所公开的优选实施例仅仅是示例性的，并非本文公开的技术概念的全部范围。

[0076] 因此，保护范围不被限定为说明书中描述的优选实施例所，而仅仅通过随后的权利要求所限定。

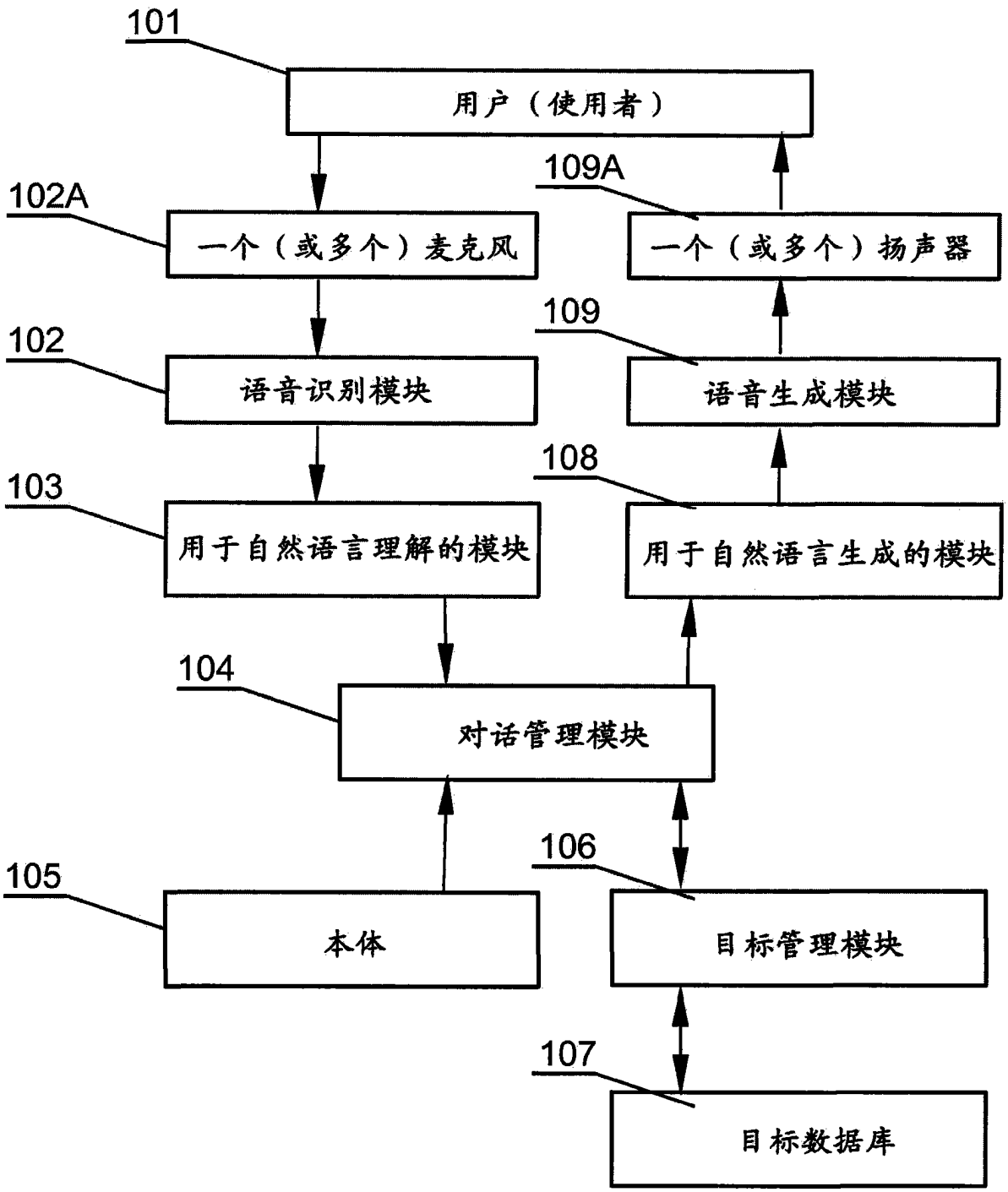

101
用户（使用者）
102A
一个（或多个）麦克风
109A
一个（或多个）扬声器
102
语音识别模块
109
语音生成模块
103
用于自然语言理解的模块
108
用于自然语言生成的模块
104
对话管理模块
105
本体
106
目标管理模块
107
目标数据库

图1

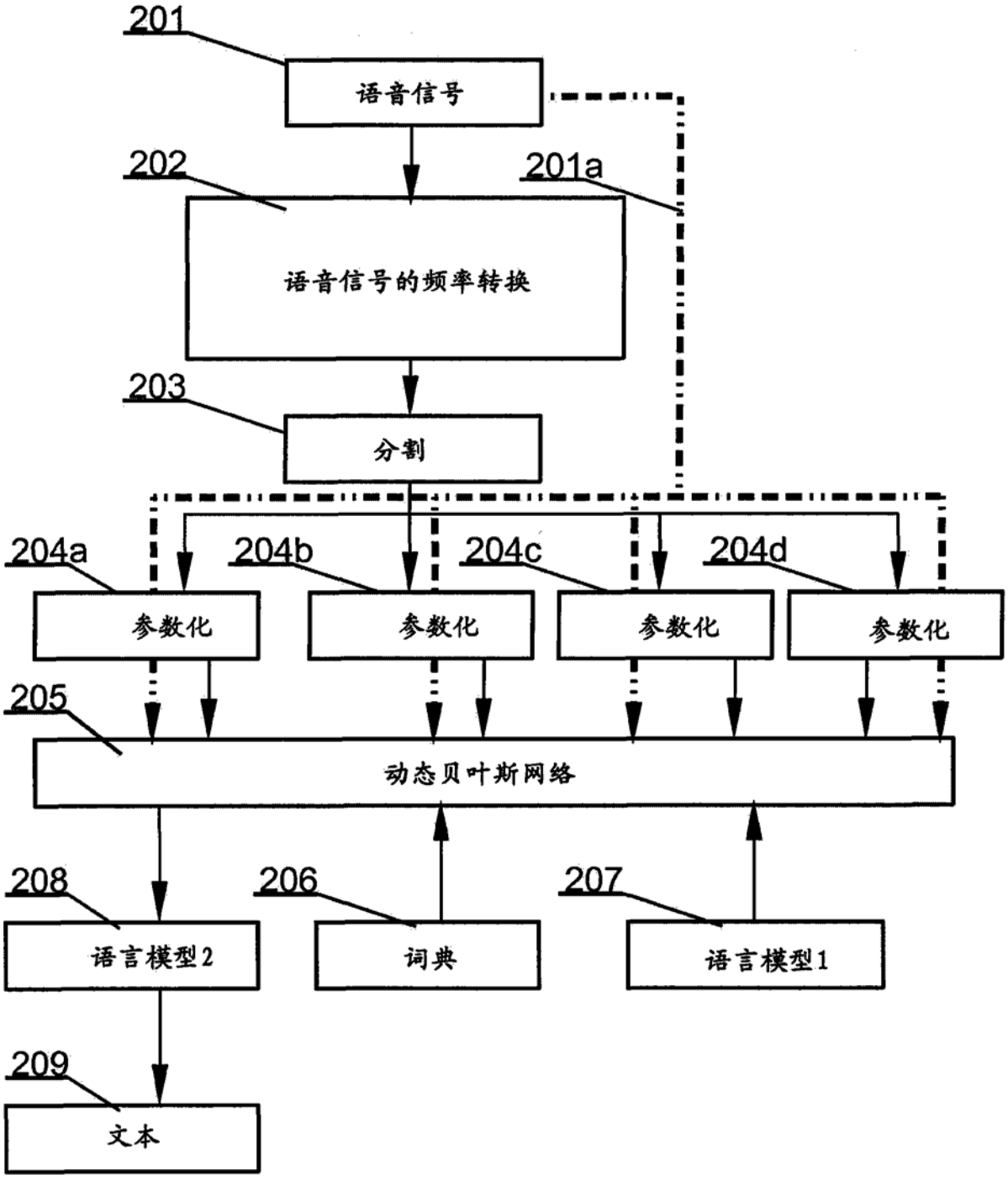

201
语音信号
202
201a
语音信号的频率转换
203
分割
204a
204b
204c
204d
参数化
参数化
参数化
参数化
205
动态贝叶斯网络
208
206
207
语言模型2
词典
语言模型1
209
文本

图2

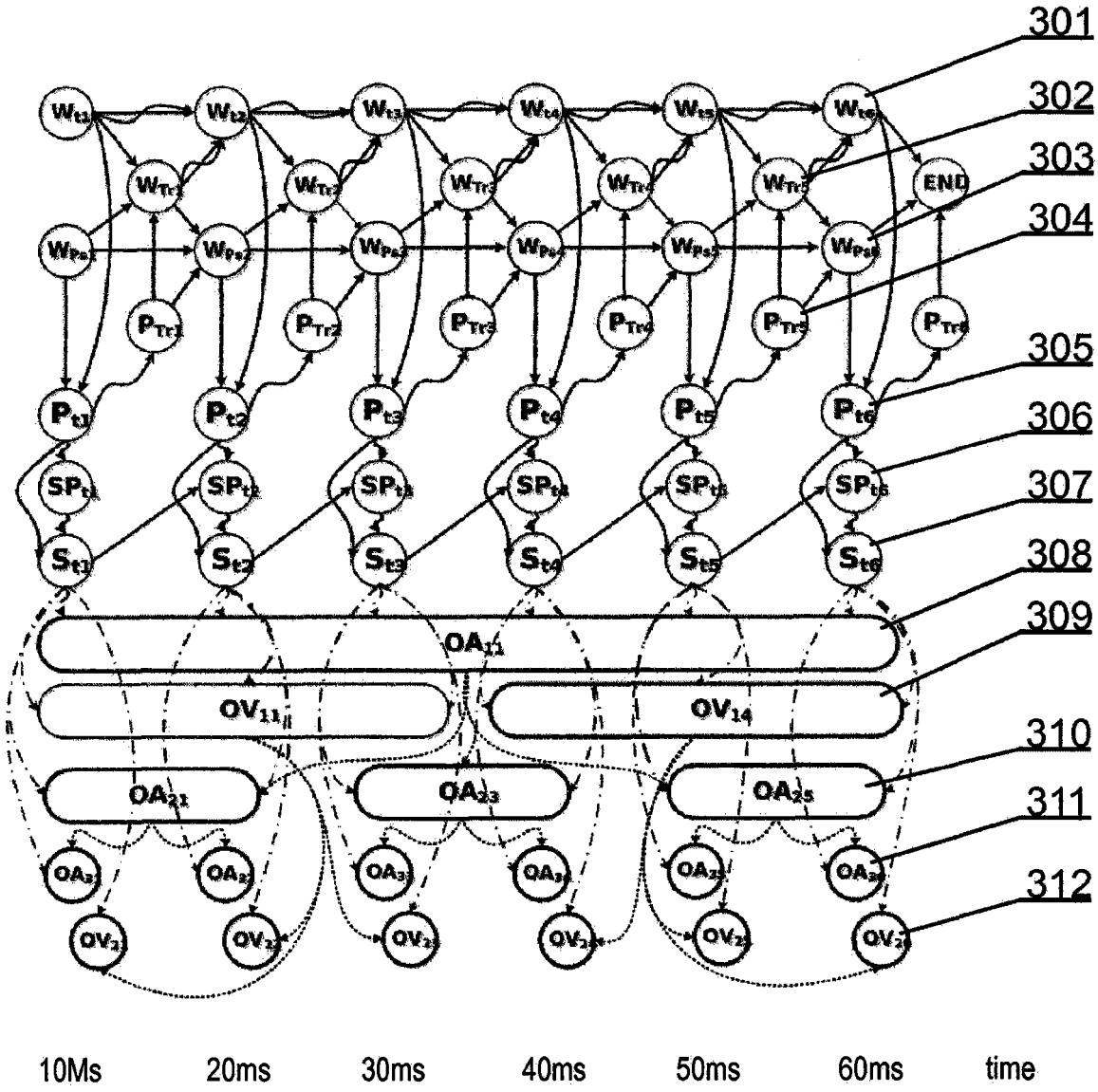

301
302
303
304
305
306
307
308
309
310
311
312
END
10Ms
20ms
30ms
40ms
50ms
60ms
time

图3

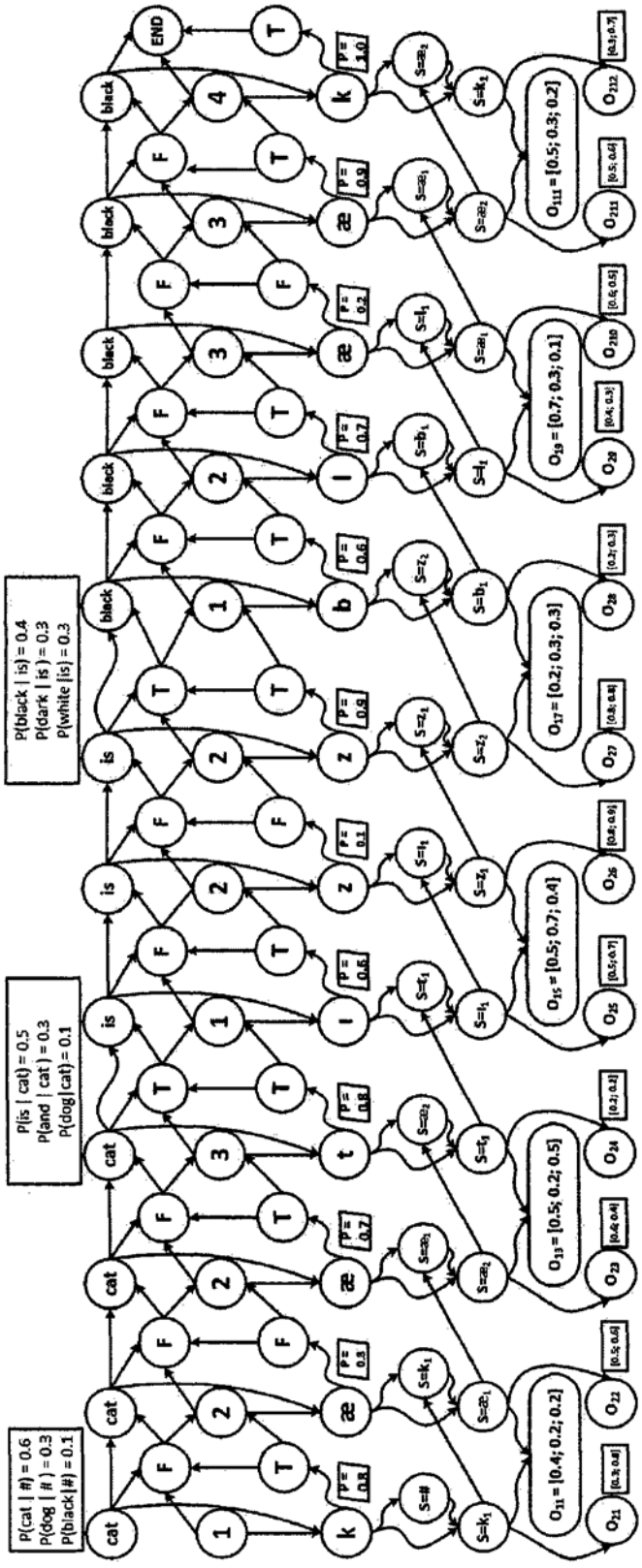

P(cat | #) = 0.6
P(dog | #) = 0.3
P(black |#) = 0.1
P(is | cat) = 0.5
P(and | cat) = 0.3
P(dog|cat) = 0.1
P(black | is) = 0.4
P(dark | is) = 0.3
P(white |is) = 0.3
cat
is
black
END
O11 = [0.4; 0.2; 0.2]
O13 = [0.5; 0.2; 0.5]
O15 = [0.5; 0.7; 0.4]
O17 = [0.2; 0.3; 0.3]
O19 = [0.7; 0.3; 0.1]
O111 = [0.5; 0.3; 0.2]

图4