计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
45(11):1882-1891,2008

基于动态贝叶斯网络的连续语音识别框架及其 Token 传递模型

苗夺谦　王睿智　冉　巍
(同济大学嵌入式系统与服务计算教育部重点实验室　上海　201804)
(同济大学计算机科学与技术系　上海　201804)
(wei_ran@mail.tongji.edu.cn)

A Dynamic Bayesian Network Based Framework for Continuous Speech Recognition and its Token Passing Model

Miao Duoqian, Wang Ruizhi, and Ran Wei
(*Ministry of Education Key Laboratory of Embedded Systems and Service Computing, Tongji University, Shanghai* 201804)
(*Department of Computer Science and Technology, Tongji University, Shanghai* 201804)

Abstract　Recently, dynamic Bayesian network (DBN) based speech recognition has aroused an increasing interest, because of its interpretability, factorization and extensibility, which hidden Markov models (HMMs) lack. Although a huge success of the introduction of DBNs into speech recognition in many areas and DBNs has been presented with promising potential to overcome inherent limitations of HMMs in speech recognition, previous work on DBN based speech recognition mainly focuses on isolated word speech recognition, and the frameworks and recognition algorithms for DBN based continuous speech recognition are not as mature and flexible as those for HMM based one. This paper is trying to address the problems of flexibility and extensibility in DBN based continuous speech recognition. To achieve this purpose, the token passing model, which works very well to address the above problems for HMM based continuous speech recognition, is adapted for DBN based one, and a general framework based on it is proposed. In this framework, the advantages of both token passing model and DBN are combined. A novel recognition algorithm independent of the upper layer language model is proposed under this framework, and a toolkit DTK for building DBN based speech recognition under this framework is developed.

Key words　speech recognition; pattern recognition; dynamic Bayesian networks; token passing model; stochastic model

摘　要　近年来,由于动态贝叶斯网络(DBN)相对于传统的隐马尔可夫模型(HMM)更具可解释性、可分解性以及可扩展性,基于 DBN 的语音识别引起学者们越来越多的关注.但是,目前关于基于 DBN 的语音识别的研究主要集中在孤立语音识别上,连续语音识别的框架和识别算法还远没有 HMM 成熟和灵活.为了解决基于 DBN 的连续语音识别的灵活性和可扩展性,将在基于 HMM 的连续语音识别中很好地解决了上述问题的 Token 传递模型加以修改,使之适用于 DBN.在该模型基础上,为基于 DBN 的连续语音识别提出了一个基本框架,并在此框架下提出了一个新的独立于上层语言模型的识别算法.还介绍了作者开发的一套基于该框架的可用于连续语音识别及其他时序系统的工具包 DTK.

关键词　语音识别;模式识别;动态贝叶斯网络;Token 传递模型;随机模型

中图法分类号　TP391.4

收稿日期:2007-12-13;**修回日期**:2008-01-22
基金项目:国家自然科学基金项目(60775036,60475019);高等学校博士学科点专项科研基金项目(20060247039)

大部分成熟的连续语音识别系统[1],都是选择某个随机模型对一组声学单元(acoustic unit)建模,描述一些特定的声学特征.连续语音识别的任务就是寻找到一个能最好的匹配语音观察样本的声学单元序列,而不需要事先知道每个声学单元的边界.

隐马尔可夫模型[2](hidden Markov model,HMM)是较常用的描述语音识别中的声学特征的随机模型.一个 HMM 工作在这样一个假设下,观察样本序列是由一组隐含的有限状态序列在某个随机分布下产生的,隐含状态的转换则满足一个一阶马尔可夫链.

Token 传递模型由 Young 等人[3]提出,用来为基于 HMM 的连续语音识别提供一个简单而强大的抽象模型.在这个模型中,识别过程被看作一个在某个转换网络中传递 Token 的过程.因此,在该模型下,基于 HMM 的连续语音识别中的语言模型和声学模型被分离开,可以很容易地改变词法和语法结构,而不影响底层的声学模式匹配算法.

动态贝叶斯网络[4](dynamic Bayesian network,DBN)是一个强大而灵活的随机模型,用于对离散随机过程的表示和推理.事实上,一个 HMM 可以描述为 DBN 的一个特例.DBN 的可解释性、可分解性及可扩展性正是 HMM 缺少的,因此引起了学者们越来越多的关注.

Zweig 等人[5-6]首次提出利用 DBN 来对语音识别建模,用来克服 HMM 的限制.之后,不少文献讨论和分析了用 DBN 建模的声学特征[7-10]和网络结构[11-12]问题.Bilmes[13]综述了一系列基于 DBN 的语音识别技术,并开发了一个通用的工具包 GMTK[14],用于构建基于 DBN 的语音识别和其他时序系统.

虽然 DBN 在语音识别的许多领域取得巨大的成功,但目前 DBN 主要用在孤立语音识别上.为了将 DBN 用于连续语音识别,研究者提出了一些网络结构和框架,但它们这些方法大致分两类:第 1 类方法用 DBN 对每个声学单元独立建模,然后用一个特定的语言模型将它们连接起来组成一个大的 DBN[12,15].这类方法要求每个声学单元的 DBN 具有相同的结构或者可以扩展成为相同的结构,而且一般只是对基本的 HMM 等价 DBN 的一个特定扩展.第 2 类方法是将声学模型和语言模型当作一个整体建模[16].这类方法的扩展性较差,很难做到只改变声学模型或者语言模型其中之一而不影响对方,且不能表示无法用 DBN 建模的语言模型.由此可见,现存的基于 DBN 的连续语音识别框架和识别算法在灵活性和可扩展性上还不如 HMM.

本文的目标是解决基于 DBN 的连续语音识别中的灵活性和可扩展性问题.鉴于 Token 传递模型在基于 HMM 的连续语音识别中能很好地实现这个目标,所以本文对它加以修改,使之能适用于 DBN,并在此基础上提出了一个基于 DBN 的连续语音识别基本框架,在该框架下提出了一个新的独立于上层语言模型的基于 DBN 的 Token 传递算法.

1 隐马尔可夫模型和 Token 传递模型

1.1 隐马尔可夫模型

在基于 HMM 的语音识别中,一个声学单元用一个 HMM 来建模.一个声学单元可以是一个单词、词根、音节或者音素等,取决于具体的应用.不失一般性,在本文中用单词模型或者单词作为声学单元的同义词.

一个 HMM 工作在这样一个假设下,观察样本序列是由一组隐含的有限状态序列在某个随机分布下产生的,隐含状态的转换则满足一个一阶马尔可夫链.用 s_t^w 表示单词 w 在时刻 t 的状态.单词 w 的 HMM 可以表示为$<A^w, b^w>$,其中 A^w 表示马尔可夫链的转移矩阵,元素 $a_{i,j}^w = P(s_t^w = j \mid s_{t-1}^w = i)$,$b^w$ 是一个生成概率分布向量,$b_i^w = P(o_t \mid s_t^w = i)$.

令 $S^w(t) = \{s_1^w, s_2^w, \cdots, s_t^w\}$ 表示单词 w 从时刻 1 到 t 的任意一个隐含状态序列,$O = \{o_1, o_2, \cdots\}$ 表示一个观察样本序列,$\phi_j^w(t)$ 表示单词 w 相对于观察向量 o_1 到 o_t 且在时刻 t 处于状态 j 的部分最大似然,即 $\phi_j^w(t) = \max_{S^w(t-1)} P(O_{1:t}, S^w(t-1), s_t^w = j)$.部分最大似然 $\phi_j^w(t)$ 可以按式(1)递归地求出:

$$\phi_j^w(t) = \max_i(\phi_i^w(t-1)\, a_{i,j}^w b_j^w(o_t)), \quad (1)$$

其中 $\phi_1^w(0) = 1$,$\phi_j^w(0) = 0$ 对 $1 < j \leq N^w$,N^w 表示单词 w 的状态数.而单词 w 的最大似然则为 $\phi_{N^w}^w(T)$.

1.2 Token 传递模型

Token 传递模型[3]提供了一个简单而强大的抽象模型,成功地被应用于 HTK[2]中.在该模型中,识别过程被看作一个在某个转换网络中传递 Token 的过程,语言模型与底层声学模型相互独立.

下面简单介绍用 HMM 作为其底层声学模型的 Token 传递模型.本文对 Young 等人[3]给出的 Token 传递模型在描述上作了一些修改,使之也能适用于用 DBN 建模的底层声学模型,而不需要改变

上层语言模型算法.

单词 w 在时刻 t 的状态 j 所持有的 Token 包含了(还有一些其他信息)它的部分似然 $\phi_j^w(t)$,可以传递到时刻 $t+1$ 的其他状态.上一小节中描述的最大似然的递归过程可以描述为一个 Token 传递的过程.

令 $Tok_j^w(t)$ 表示单词 w 在时刻 t 的状态 j 所持有的 Token,用 $\phi(tok)$ 表示一个 Token tok 的部分似然,用 tok_zero 表示一个部分似然为 0 的 Token,即 $\phi(tok\_zero)=0$. 基于 HMM 的单词模型 Token 传递算法如算法 1 所示.

算法 1. 基于 HMM 的单词模型 Token 传递算法.

Token *step_word_model*(Token *tok*, int *t*, int *w*)

1) IF $\phi(tok) > \phi(Tok_0^w(t-1))$ THEN

$Tok_0^w(t-1) := tok$;

2) 单词 w 在时刻 t 的所有状态都持有 Token *tok_zero*;

3) FOR $i := 1$ TO N^w DO

FOR $j := 1$ TO N^w DO

$tok' := Tok_i^w(t-1)$;

$\phi(tok') := \phi(tok') a_{i,j}^w b_j^w(o_t)$;

IF $\phi(tok') > \phi(Tok_j^w(t))$ THEN

$Tok_j^w(t) := tok'$;

END

END

4) RETURN $Tok_N^w(t)$.

算法 1 中函数 *step_word_model* 实现了一个单词模型的匹配算法,可以视作一个抽象模式匹配器,控制 Token 传递流在单词模型内部传递.另一方面,单词模型之间的 Token 传递则由在高层语言模型控制. *step_word_model* 的定义起到一个语言模型和单词模型之间的接口作用,如图 1 所示:

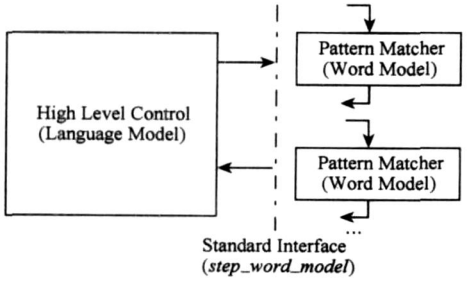

Fig. 1 Separation of language model and word models.

图 1 语法模型和单词模型的分离

我们以 Bi-gram 语言模型为例来阐述语言模型如何利用标准接口 *step_word_model* 来控制单词模型之间的 Token 传递,如算法 2 所示.在 Bi-gram 语言模型中,某个单词出现的概率只和它前面一个单词相关,即它给出了某个单词出现相对于前一个单词的条件概率 $P(w_i|w_{i-1})$.

算法 2. Bi-gram 语言模型的 Token 传递算法.

1) 除每个单词在时刻 0 的状态 1 外,所有状态都持有 Token *tok_zero*;

2) 每个单词 w 在时刻 0 的状态 1 都持有一个部分似然为 1 的 Token,即 $\phi_1^w(0)=1$;

3) 令 *toks_in* 和 *toks_out* 为一个 Token 数组,且对每个单词 w 初始化 $\phi(tok\_in_w)=0$;

4) FOR $t := 1$ TO T DO

FOR EACH word w

$toks\_out_w = step\_word\_model(toks\_in_w, t, w)$;

END

FOR $i := 1$ TO N^w DO

$k := argmax_j \phi(tok\_out_j) P(i|j)$;

$toks\_in_i := toks\_out_k$;

$\phi(toks\_in_i) := \phi(tok\_out_k) P(i|k)$;

END

END

事实上,算法 2 只能求出所有可能的单词序列的最大似然,但是,我们的任务是找到一个似然最大的单词序列.为了满足这样的要求,Young 等人[3]对 Token 的概念作了扩展.除部分似然外,一个 Token 还包含一个路径标识符(path identifier),表示某个单词模型在某个时刻具有最大部分似然的单词序列及其边界.

由于单词模型和语言模型之间的标准接口 *step_word_model* 的存在,其他语言模型下的语音识别也很容易得到,而不影响底层单词模型.其他语言模型,如 *Tri-gram* 和上下文无关文法,在文献[3]中有详细阐述.

2 动态贝叶斯网络及其推理算法

一个贝叶斯网络[17](BN)是一个有向无环图(DAG),其中结点表示随机变量,弧表示两个随机变量之间存在因果影响关系.一个动态贝叶斯网络[4](DBN)包含一个 BN 的实例随时间重复多次,同时包含跨越时间的因果依赖关系.

事实上,给定一个时间范围,一个 DBN 总能展开成为一个普通 BN,任意 BN 的推理算法[18-19]对

DBN 都适用.但是为了利用 DBN 的特殊性质,一些针对 DBN 的特殊算法被提出来.

接口算法(interface algorithm)[20]利用了向前接口(forward interface)的概念,它 d-separate 了过去和未来.因为在该算法中,每一步概率传递都被限制在一个"$1\frac{1}{2}$-slice BN",这种简化正好适合于本文的 Token 传递模型,所以本文提出的基于 DBN 的识别算法以接口算法为基础.

时刻 t 的向前接口用 I_t 表示,是至少有一个子结点在时刻 $t+1$ 的随机变量的集合,即 $I_t=\{x|\ x\in X_t,\ (\exists y)\ x\in parent(y)\ \wedge y\in X_{t+1}\}$,其中 X_t 表示时刻 t 的所有随机变量的集合.可以证明[20],I_t d-separate 时刻 t 之前的随机变量和时刻 t 之后的随机变量.

一个在时刻 t 的"$1\frac{1}{2}$-slice BN"是一个由 $H_t=I_{t-1}\cup X_t$ 生成的子 BN.我们可以为每个 H_t 构造一个联合树(junction tree) J_t,而且必须保证 I_{t-1} 和 I_t 各自包含在一个团(clique)里,因为我们需要求出 $P(I_{t-1})$ 和 $P(I_t)$.这可通过在构造 J_t 之前,在正图(moral graph)中对 I_{t-1} 中的每两个结点之间都增加一条边,也对 I_t 作类似操作.

Murphy[20]提出接口算法的概率传递过程需要进行两遍,向前向后各一遍.但是如果只关心 $P(I_t|O_{1:t})$,向前传递已经足够了,不需要进行向后传递.

令 D_t 和 C_t 分别表示 J_t 中包含 I_{t-1} 和 I_t 的团.给定先验概率 $P(I_{t-1}|O_{1:t-1})$,我们先初始化所有的团和隔集(separator)的势(potential)都为 1,将先验概率乘到 D_t 的势上.然后我们以 C_t 为根对 J_t 调用 *collect-evidence* 过程.这时,C_t 的势就表示 $P(C_t|O_{1:t})$.边缘化(marginalize) $P(C_t|O_{1:t})$,我们就可以得到 $P(I_t|O_{1:t})$.

构造一棵联合树的方法及 *collect-evidence* 过程与普通 BN 推理算法中一致,在文献[21]中有详细介绍.

在语音识别中,我们更关心 $\max_{X_{1:t}\setminus I_t} P(X_{1:t}|O_{1:t})$ 而不是 $P(I_t)=\sum_{X_{1:t}\setminus I_t} P(X_{1:t}|O_{1:t})$,其中 $X_{1:t}$ 表示从时刻 0 到 t 的所有隐含变量的集合.广义分配率[22](generalized distributive law, GDL)把 BN 推理抽象为一个"marginalizing a product function"(MPF)问题.GDL[22]证明了 $\max_{X_{1:t}\setminus I_t} P(X_{1:t}|O_{1:t})$ 可以由计算 $\sum_{X_{1:t}\setminus I_t} P(X_{1:t}|O_{1:t})$ 同样的方法得到,只是将上述接口算法和 *collect-evidence* 过程中的 marginalize 操作符中的求和操作替换为求最大值操作.

3 一个基于 DBN 的连续语音识别框架

3.1 框架设计目标

本文框架设计的总体目标是继承目前成熟的基于 HMM 的框架的大部分优点(主要集中在 HTK[2]的特点),并将 DBN 的可解释性、可分解性及可扩展性引入该框架.具体的框架设计目标如下,

1) 支持语言模型和声学模型的分离.每个声学单元用一个 DBN 独立地建模,不需要考虑上层的语言和语法限制.一组声学单元能够在不同的语言模型下交互工作,而不需要改变各个 DBN 的结构和参数.换句话说就是支持 Token 传递模型.

2) 对各个声学单元支持异构的 DBN.各个声学单元的 DBN 结构可以完全不同,描述的声学特征也可以完全不同.例如,基波频率(pitch)只与浊音相关,与清音无关[5].总之,我们不对每个 DBN 的结构和参数作过多假设,只要它们满足一个最小接口使其能在各个声学单元之间通信即可.

3) 支持高效的、可扩展的识别算法.能够容易地对标准的识别算法进行扩展,如 N-best 识别,beam 剪枝[2]等.此外,支持高效的训练算法.

3.2 基本框架的总体结构

本文对 GMTK[14]中给出的模板稍作修改来定义一个 DBN.一个 DBN 模板包含 3 个部分:主干(chunk)、前缀(prologue)、后缀(epilogue).主干部分表示一个 DBN 的重复部分,可以被重复地展开(unroll)直至相对于一个发声序列足够长.前缀和后缀分别在一个展开网络的开始和结束部分.在 GMTK 中,模板中的每个部分都可以跨越几个时间帧,为了简化,在本文中我们只关心单帧的情况,单帧可以很容易地扩展到多帧情况.此外,我们限制前缀和后缀不包括任何观察变量,且时刻 t 的任何一个变量的父结点只能在时刻 t 或者 $t-1$.

怎样将各个声学模型的 DBN 连接在一起,且独立于上层语言模型,这是本文的首要问题.在回顾了基于 HMM 的语音识别的 Token 传递模型后,我们有这样一个结论:单词模型和语言模型之间的标准接口 *step_word_model* 接受一个 Token 作为其输入,表示某个单词序列与时刻 0 到 $t-1$ 的观察

样本序列最佳匹配,并且单词 w 从此刻开始. 该接口输出一个 Token,表示某个单词序列与时刻 0 到 t 的观察样本序列最佳匹配,并且单词 w 从此刻结束. 算法 1 中描述的针对 HMM 的接口实现表明,输入 Token 只可能直接传递给 HMM 的状态 1,而状态 1 表示一个单词模型的开始. 输出 Token 则是由状态 N 传递而来,而状态 N 表示一个单词模型的结束.

基于以上结论,如果我们要针对 DBN 实现接口 *step_word_model*,需要两个实体来分别表示一个单词模型的开始和结束. 所以,在本文提出的框架中,要求在每个单词 w 的 DBN 模板的前缀必须包含两个二值随机变量 b^w 和 e^w,分别用来表示一个单词模型是否开始和结束. 变量 b^w 是一个没有父结点的根变量,而变量 e^w 则没有子结点.

图 2 描述了在该框架下,基本的 HMM 等价 DBN. 图 2(a)是 DBN 定义模板,而 2(b)则是展开了的贝叶斯网络. 变量之间的条件概率定义如下: $P(s^w(1)=1|b^w)=1$, $P(s^w(t)=j|s^w(t)=i)=a^w_{i,j}$, $P(o_t|s^w(t)=i)=b^w_i$, $P(e^w=1|s^w(T)=N^w)=1$, $P(e^w=0|s^w(T)\neq N^w)=1$, 其中 $a^w_{i,j}$, b^w_i 如第 1 节中的定义.

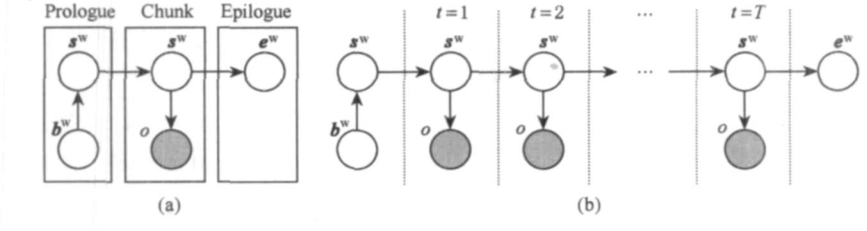

Fig. 2 The baseline HMM-equivalent DBN. (a) The DBN definition template and (b) The unrolled network.

图 2 基本的 HMM 等价 DBN. (a) DBN 定义模板;(b) 展开网络

几个在该框架下的示例 DBN 定义模板如图 3 所示. 图 3(a)表示了一个如文献[5,7]中所述的带一个辅助变量的 DBN. 图 3(c)表示了一个如文献[16]中所述的多流(multi-stream)半同步的 DBN. 图 3(b)表示了一个如文献[16]中所述的显式转换(transition-explicit) DBN,声学单元之间的转换是可观察的.

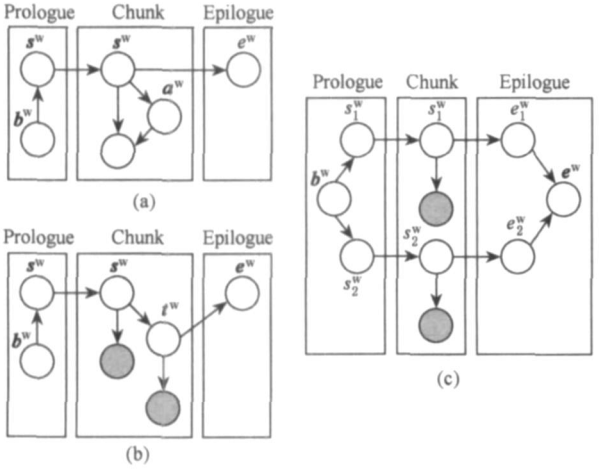

Fig. 3 Example DBN definition templates. (a) A DBN with an auxiliary variable; (b) Transition-explicit DBN; and (c) Multi-stream semi-synchronous DBN.

图 3 示例 DBN 定义模板. (a) 带一个辅助变量的 DBN;(b) 显式转换 DBN;(c) 多流半同步的 DBN

每个声学单元的 DBN 不一定相同,甚至可能完全不一样,但只要每个 DBN 都分别在前缀和后缀中有两个二值变量 b^w 和 e^w 来分别表示这个声学单元是否开始和结束,而且满足上面所述的最低要求,下一节中详细讨论的针对基于 DBN 的连续语音识别的 Token 传递模型就能在不同的语言模型下高效地识别,而不用改变每个 DBN.

4 针对 DBN 的 Token 传递模型

4.1 寻找最佳匹配单词序列

在上一节中,本文提出了一个基于 DBN 的连续语音识别基本框架. 该框架需要完成的任务是找到一个单词序列最佳匹配语音观察序列. 一个单词序列 S 用一个二元组数组表示, $S=\{\langle w_i, t_i\rangle | i=1, 2, \cdots\}$,其中 $t_1=1$ 且 $t_{i-1}<t_i$ 对所有 $2\leqslant i\leqslant |S|$,表示该序列中第 i 个单词从时刻 t_i 开始. 如果 $t_{|S|}\leqslant T$,我们就说单词序列 S 在时刻 T 有效.

给定一个在时刻 T 有效单词序列 $S=\{\langle w_i, t_i\rangle | i=1, 2, \cdots\}$,从时刻 1 到 T 的连接 BN 可以通过如下方法:先对每个单词 w_i 从时刻 t_i 到 $t_{i+1}-1$ 展开(假设 $t_{|S|+1}=T+1$),然后通过添加一个从 $e^{w_{i-1}}(t_i-1)$ 到 $b^{w_i}(t_i)$ 箭头,把各个展开的 BN 连接到一起,并令 $P(b^{w_i}(t_i)=1|e^{w_{i-1}}(t_i-1)=1)=1$. 由于单词

序列的限制,$e^{w_{i-1}}(t_i - 1)$ 现在都是确定的,即对所有 $1 \leq i < |S|$, $P(e^{w_{i-1}}(t_i - 1) = 1) = 1$. 如图 4 所示:

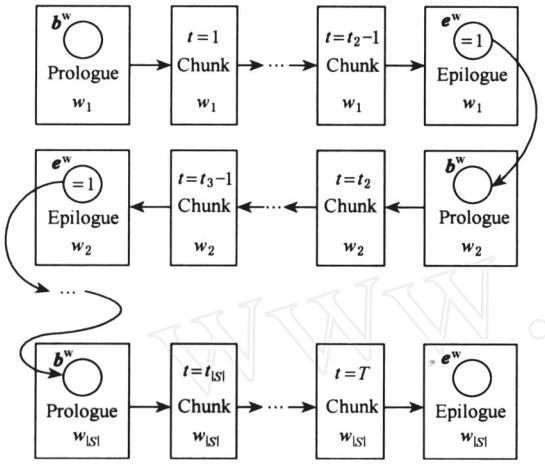

Fig. 4 A connected BN constructed by a word sequence S.

图 4 由单词序列 S 构造出的连接 BN

令 $BN(S, T)$ 表示由一个在时刻 T 有效单词序列构造出的连接 BN, $var(S, T)$ 表示这个 BN 中所有隐含变量的集合. 相对于观察序列 $O_{1:T}$ 最佳匹配的单词序列是指使 S 的似然达到最大的单词序列,即

$$S^* = \arg\max_S \max_{var(S,T)} P(e^w(T) = 1, var(S, T), O_{1:T}) P(S). \quad (2)$$

理想地,如果我们能枚举出所有可能单词序列,我们总能得到最佳匹配的一个序列. 但是单词序列的状态空间的大小已经超出我们的处理范围.

在任意时刻 T, 所有有效单词序列可以分为 W 组,其中 W 为单词模型的个数. 令 $\mathcal{S}^w(T)$ 表示最后一个单词为单词 w 的单词序列组,即任意 $S \in \mathcal{S}^w(T)$ 都有 $w_{|S|} = w$. 每个单词 w 的 $\mathcal{S}^w(T)$ 都可以再被分为两个子组 $\mathcal{S}_1^w(T)$ 和 $\mathcal{S}_2^w(T)$, 对任意 $S \in \mathcal{S}_1^w(T)$ 都有 $t_{|S|} = T$, 而对任意 $S \in \mathcal{S}_2^w(T)$ 都有 $t_{|S|} < T$.

对任意单词序列 $S \in \mathcal{S}_1^w(T)$, S 在时刻 $T-1$ 无效. S 的前序列,用 $prev(S)$ 表示,是 S 的前 $|S| - 1$ 个元素,即 $prev(S) = \{<w_i, t_i> | i = 1, 2, \cdots, |S| - 1\}$. 显然, $prev(S)$ 在时刻 $T-1$ 有效. 所以, $BN(S, T)$ 可以在 $BN(prev(S), T-1)$ 的基础上通过增加一条从 $e^{w_{|S|-1}}(T-1)$ 指向 $b^w(T)$ 的箭头,设置 $P(b^w(T) = 1 | e^{w_{|S|-1}}(T-1) = 1) = 1$ 把单词 w 的前缀添加进来, 然后再展开单词 w 的主体部分一次, 如图 5(a) 所示.

对任意单词序列 $S \in \mathcal{S}_2^w(T)$, S 在时刻 $T-1$ 有效. 所以 $BN(S, T)$ 只需要在 $BN(prev(S), T-1)$ 的基础上再展开一次单词 w 的主干部分即可,如图 5(b) 所示. $e^w(T-1) = 0$ 表示单词 w 在时刻 $T-1$ 还没有结束.

令 $\mathcal{S}$ 表示一个在时刻 T 有效的单词序列组. 如果所有 $\mathcal{S}$ 中的单词序列所构造的 BN 在时刻 T 有相同的部分网络结构, $X(T)$ 是一个包含在这部分公共结构中的变量集,那么 $X(T)$ 在当前时刻相对于组 $\mathcal{S}$ 的似然定义为

$$\phi(X(T), \mathcal{S}) = \max_{S \in \mathcal{S}} \mathrm{var}_{var(S,T) \setminus X(T)} P(var(S, T), O_{1:T}) P(S), \quad (3)$$

而在上一时刻相对于组 $\mathcal{S}$ 的似然定义为

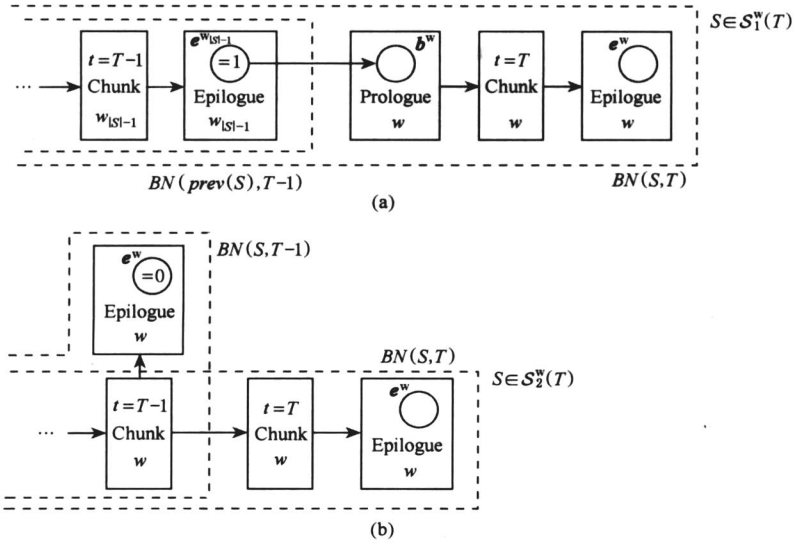

Fig. 5 Constructing a BN recursively from a word sequence $S \in \mathcal{S}^w(T)$. (a) $S \in \mathcal{S}_1^w(T)$ and (b) $S \in \mathcal{S}_2^w(T)$.

图 5 从一个单词序列 $S \in \mathcal{S}^w(T)$ 递归构造一个 BN. (a) $S \in \mathcal{S}_1^w(T)$, (b) $S \in \mathcal{S}_2^w(T)$

$$\phi^1(X(T), \mathcal{S}) = \max_{S\in\mathcal{S}} \max_{var(prev(S),T-1)} P(var(prev(S),T-1), X(T), O_{1:T-1})P(S). \quad (4)$$

寻找一个最佳匹配单词序列(似然最大的序列)的问题,就被转化为寻找一个似然为如下值的单词序列:

$$\max_w \phi(e^w(T)=1, S\in\mathcal{S}^w(T)). \quad (5)$$

如图 5(b)所示,所有由$\mathcal{S}_{2^w}(T)$中的单词序列构造出的 BN 在时刻 $T-1$ 和 T 都具有相同结构. 如第 2 节中所述,单词 w 在时刻 $T-1$ 的主干部分的向前接口(记作 $I_c^w(T-1)$)在所有由$\mathcal{S}_{2^w}(T)$中的单词序列构造出的 BN 中 d-separate 时刻 T 之前的和时刻 T 时的部分结构.

我们构造一个由 $I_c^w(T-1)$和单词 w 在时刻 T 时的主干和后缀部分组成的 BN,在构造联合树之前,在正图中对 $I_c^w(T-1)$中的每两个结点之间都增加一条边,也对 $I_c^w(T)$作类似操作,然后为这个 BN 构造一棵联合树. 如果对 $I_c^w(T-1)$的势初始化为如式(3)所定义的 $\phi(I_c^w(T-1), \mathcal{S}_{2^w}(T-1))$,在执行如第 2 节中所述的以 $I_c^w(T)$为根的 *collect-evidence* 操作后,$I_c^w(T)$的势此时就表示 $\phi(I_c^w(T), \mathcal{S}_{2^w}(T))$.

如图 5(a)所示,所有由$\mathcal{S}_{1^w}(T)$中的单词序列构造出的 BN 在时刻 T 都具有相同结构,$b^w(T)$ d-separate 时刻 T 之前的和时刻 T 时的部分结构. 如第 2 节中所述,单词 w 在时刻 T 的前缀部分的向前接口(记作 $I_p^w(T)$) d-separate 它之前和它之后的部分结构.

我们先构造一个只包含单词 w 在时刻 T 时的前缀部分组成的 BN,在构造联合树之前,在正图中对 $I_p^w(T-1)$中的每两个结点之间都增加一条边,然后再为这个 BN 构造一棵联合树. 如果对 $b^w(T)$的势初始化为如式(4)所定义的 $\phi^1(b^w(T), \mathcal{S}_{1^w}(T))$,在执行如第 2 节中所述的以 $I_p^w(T)$为根的 *collect-evidence* 操作后,$I_p^w(T)$的势此时就表示 $\phi^1(I_p^w(T), \mathcal{S}_{1^w}(T))$.

我们再构造一个由 $I_p^w(T)$和单词 w 在时刻 T 时的主干和后缀部分组成的 BN. 如第 3.2 节中所述,单词 w 的前缀和主干部分有相同的向前接口 $I_p^w(T)$和 $I_c^w(T-1)$. 所以我们可以通过与$\mathcal{S}_{2^w}(T)$类似的前向传递过程,由 $I_p^w(T)$的势得到 $I_c^w(T)$的势表示 $\phi(I_c^w(T), \mathcal{S}_{1^w}(T))$.

现在还只剩下一个问题,那就是如何得到 $\phi^1(b^w(T), \mathcal{S}_{1^w}(T))$. 如我们之前的定义,对任意 $S\in\mathcal{S}_{2^w}(T)$, $P(b^w(T)=1 \mid e^{w_{|S|-1}}(T-1)=1)=1$. 所以,我们容易求得

$$\max_{var(prev(S),T-1)} P(var(prev(S),T-1), b^w(T), O_{1:T-1}), \quad (6)$$

这正是式(4)右部分的第 1 项. 但是,我们如何得到 $P(S)$,它是由语言模型决定的. 为了简化,我们在这里只考虑 Bi-gram 语言模型,即 $P(S\mid prev(S)) = P(w^{|S|}\mid w^{|S|-1})$. 对其他语言模型的扩展,我们在后面讨论.

假设对每个单词 w',$\phi(e^{w'}(T-1)=1, \mathcal{S}_{1^{w'}}(T))$都已知,

$$\begin{aligned}&\phi^1(b^w(T)=1, \mathcal{S}_{1^w}(T)) = \\ &\max_{S\in\mathcal{S}} \max_{var(prev(S),T-1)} P(b^w(T)=1, var(prev(S),T-1), O_{1:T-1})P(S) = \\ &\max_{S\in\mathcal{S}} \max_{var(prev(S),T-1)} P(e^{w_{|S|-1}}(T-1)=1, var(prev(S),T-1), \\ &O_{1:T-1})P(prev(S))P(S\mid prev(S)) = \\ &\max_{S\in\mathcal{S}} \max_{var(prev(S),T-1)} P(e^{w_{|S|-1}}(T-1)=1, var(prev(S),T-1), \\ &O_{1:T-1})P(prev(S))P(w\mid w_{|S|-1}) = \\ &\max_w \phi(e^{w'}(T-1)=1, \mathcal{S}_{1^{w'}}(T))P(w\mid w')\end{aligned} \quad (7)$$

且 $\phi^1(b^w(T)=0, \mathcal{S}_{1^w}(T))=0$.

事实上,$\mathcal{S}^w(T)$的两个子组可以统一处理. 令 $I^w(T-1)=I_c^w(T-1)\cup e^w(T-1)$. 如果我们令 $I^w(T-1)$的势表示式(3)所定义的 $\phi(I^w(T-1), \mathcal{S}_w(T-1))$,我们可以很容易得到

$$\phi(e^w(T-1)=1, \mathcal{S}_{1^w}(T)) = \phi(e^w(T-1)=1, \mathcal{S}_w(T-1)), \quad (8)$$

及

$$\phi(I_c^w(T-1), \mathcal{S}_{2^w}(T)) = \phi(\{I_c^w(T-1), e^w(T-1)=0\}, \mathcal{S}_w(T-1)). \quad (9)$$

这时,我们就可以通过式(7)由 $\phi(e^w(T-1)=1, \mathcal{S}_{1^w}(T))$得到 $\phi^1(b^w(T), \mathcal{S}_{1^w}(T))$,然后按照上面所述的前向传递过程,求得 $\phi^1(I_p^w(T), \mathcal{S}_{1^w}(T))$. 因为单词 w 的前缀和主干部分有相同的向前接口 $I_p^w(T)$和 $I_c^w(T-1)$,我们可以定义一个虚接口 $I_v^w(T)$来同时表示两种向前接口. $I_v^w(T)$的势定义为

$$\phi(I_v^w(T)=i, \mathcal{S}_w(T)) = \max\{\phi(I_c^w(T-1)=i, \mathcal{S}_{2^w}(T)), \phi(I_p^w(T)=i, \mathcal{S}_{1^w}(T))\}. \quad (10)$$

如图 5 所示,所有由$\mathcal{S}_w(T)$中的单词序列构造出的 BN 在时刻 T 时 $I_v^w(T)$之后都具有相同结构,

所以,我们可以构造一个由 $I_v^w(T)$ 和单词 w 在时刻 T 时的主干和后缀部分组成的 BN,然后通过与 $\mathcal{S}_2^w(T)$ 中类似的前向传递过程,由 $I_v^w(T)$ 的势得到 $I^w(T)$ 的势表示 $\phi(I^w(T),\mathcal{S}^w(T))$.

上述过程就是由时刻 $T-1$ 时所有单词 w 的 $\phi(I^{w\prime}(T-1),\mathcal{S}_w'(T-1))$ 递归地求出时刻 T 时所有单词 w 的 $\phi(I^w(T),\mathcal{S}_w(T))$. 然后,最佳匹配序列的最大似然可以由式(5)和式(8)得到,而这个单词序列本身则可以由回溯求出.

4.2 基于 DBN 的单词模型 Token 传递算法

上述步骤可以像基于 HMM 的连续语音识别一样,抽象为一个 Token 传递的过程. 对一个随机变量集 $X(t)$,它的每一个实例(instantiation) $X(t)=x$ 都可以持有一个 Token,记作 $tok(X(t)=x)$. Token 分为两类,基本 Token 和扩展 Token. $X(t)=x$ 持有的如果是基本 Token,那么它只有一个字段,记作 $\Phi(X(t)=x)$,表示如式(3)和式(4)定义的似然 $\phi(X(t),\mathcal{S})$ 或者 $\phi^{-1}(X(t),\mathcal{S})$. 扩展 Token 除了上述字段外,还包括其他信息,如路径标识符等,与 HMM 中的 Token 类似. 一个随机变量集的所有实例要么全部持有基本 Token,要么全部持有扩展 Token. 一个随机变量集的实例持有基本 Token 还是扩展 Token 将在本文后面讨论.

为了设计基于 DBN 的单词模型的 Token 传递算法,本文重新定义了贝叶斯网络推理的 2 个基本操作,边缘化(marginalization)和乘法(multiplication),具体如算法 3 和算法 4 所示. 令 X 和 Y 为 2 个随机变量集,且 $Y\subseteq X$.

算法 3. $marginalize(X,Y)$.

设置 Y 的所有实例持有一个 $\Phi(Y=y)=0$ 的 Token.

FOR EACH 实例 $X=x$

令实例 $Y=y$ 为与 $X=x$ 相容的实例;

IF $\Phi(X=x)>\Phi(Y=y)$ THEN 将 $tok(X=x)$ 复制到 $Y=y$ 替换原先的 Token;

END

算法 4. $multiply(Y,X)$.

FOR EACH 实例 $X=x$

令实例 $Y=y$ 为与 $X=x$ 相容的实例;

$tmp:=\Phi(Y=y)$;

IF $Y=y$ 持有的是扩展 Token, THEN 将 $tok(Y=y)$ 复制到 $X=x$ 替换掉原先的 Token;

$\Phi(X=x):=\Phi(X=x)\cdot tmp$;

END

以上述两个基本操作为基础,本文提出基于 DBN 的单词模型 Token 传递算法(详见算法 5),其基本思想如下:首先,我们对每个单词 w 在时刻 t 构造两个 BN,第 1 个 BN 由单词 w 在时刻 t 的前缀部分组成,记作 $B_1^w(t)$,第 2 个 BN 由单词 w 的主干部分在时刻 $t-1$ 的向前结构及时刻 t 的主干和后缀部分组成,记作 $B_2^w(t)$. 在为每个 BN 构造联合树之前,我们在 $B_1^w(t)$ 的正图中对 $I_p^w(t)$ 中的每两个结点之间都增加一条边,对 $B_2^w(t)$ 的正图中 $I_c^w(t-1)$ 和 $I^w=I_c^w(t)\cup e^w(t)$ 作类似操作. 然后,对每个 BN 像普通 BN 一样构造一棵联合树.

在 $B_1^w(t)$ 的联合树中,令包含 $b^w(t)$ 的团记作 $B^w(t)$,包含 $I_p^w(t)$ 的团记作 $C^w(t)$. 在 $B_2^w(t)$ 的联合树中,令包含 $I_c^w(t-1)$ 的团记作 $D^w(t)$,包含 $I^w(t)$ 的团记作 $E^w(t)$. 在 $B_1^w(t)$ 的联合树中所有在从 $B^w(t)$ 到 $C^w(t)$ 的路径上的团和隔集,以及在 $B_2^w(t)$ 的联合树中所有在从 $D^w(t)$ 到 $E^w(t)$ 的路径上的团和隔集,都被称作关键变量集. 我们定义关键变量集的实例都持有扩展 Token,其他变量集的实例都持有基本 Token.

如第 1.2 节中所述,Token 传递模型的最大优点就是让单词模型和语言模型相互独立. 只要它们之间的标准接口 $step\_word\_model$ 的定义保持不变,就能只改变底层单词模型,不改变上层语言模型.

算法 5 是本文设计的基于 DBN 的单词模型的 Token 传递算法,该算法实现了这个标准接口.

算法 5. 基于 DBN 的单词模型 Token 传递算法.

Token $step\_word\_model$(Token tok, int t, int w)

1) 令 $b^w(t)=1$ 持有 Token tok, 令 $b^w(t)=0$ 持有 tok_zero;

2) $Multiply(b^w(t),B^w(t))$;

3) 对 $B_1^w(t)$ 的联合树调用以 $C^w(t)$ 为根的 $collect\text{-}evidence$ 操作;

4) $marginalize(C^w(t),I_p^w(t))$;

5) 对 $I_p^w(t)$ 每个实例, IF $\Phi(I_p^w(t)=i)>\Phi(I_c^w(t-1)=i)$ THEN 将 $tok(I_p^w(t)=i)$ 复制到 $I_c^w(t-1)=i$,替换掉原先的 Token;

6) $multiply(I_c^w(t-1),D^w(t))$;

7) 对 $B_2^w(t)$ 的联合树调用以 $E^w(t)$ 为根的 $collect\text{-}evidence$ 操作;

8) $marginalize(E^w(t),I^w(t))$;

9) FOR EACH 实例 $I^w(t)=i$

IF $I^w(t) = i$ 可推出 $e^w(t) = 0$ THEN 将 $tok(I^w(t) = i)$ 传递给与 $I^w(t) = i$ 相容的 $I_c^w(t) = j$;

ELSE IF $\Phi(I^w(t) = i) > \Phi(e^w(t) = 1)$ THEN 将 $tok(I^w(t) = i)$ 复制到 $e^w(t) = 1$, 替换掉原先的 Token;

END

10) RETURN $tok(e^w(t) = 1)$.

式(7)所描述步骤在 Token 传递模型中由一个特殊语言模型(Bi-gram 语言模型)来控制. 因为算法 5 对接口 *step_word_model* 的定义没有改变, 所以算法 2 描述的 Bi-gram 语言模型的 Token 传递算法也适用于基于 DBN 的单词模型. 同样, Young 等人[3]描述的其他语言模型, 如 Tri-gram 和上下文无关文法语言模型, 也对算法 5 描述的基于 DBN 的单词模型适用.

对算法 5 的扩展, 如 N-best 识别和 beam 剪枝等, 能够容易的实现. 本文提出的框架下的训练算法也很直接. 因为事先已经知道每段语音序列所对应的单词序列, 所以由这个单词序列就可以构造一个连接 BN(如图 4 所示). 那么, 普通的 BN 训练算法就可以很容易地应用到本文的框架中.

5 一个针对本文提出框架的工具包(DTK)

根据本文提出的基于 DBN 的连续语音识别框架和相应识别算法, 我们开发了一个可用于连续语音识别及其他时序系统工具包 DTK.

在 DTK 中, 声学模型(DBN)和语言模型之间的接口按照上一节所提出的算法 5 实现. DTK 内置了 3 种语言模型的实现, 即 Bi-gram, Tri-gram 和上下文无关文法. 它还提供了程序接口以供用户实现其他的语言模型.

除此之外, DTK 也支持 N-best 识别和 beam 剪枝. 每个 DBN 参数都可以由数据训练得到. 但是, 和 HTK[2]一样, DTK 并不支持结构学习.

6 结 论

虽然将 DBN 引入语音识别中在许多领域都取得了巨大的成功, 但是现有的基于 DBN 的连续语音识别的框架和识别算法都还远不如 HMM 成熟和灵活. 主要的不足表现在, 现有的研究要么要求所有声学模型的 DBN 都有(或者可以扩展成为)相同的结构, 通常只是对基本的 HMM 等价 DBN 的一个特定扩展, 要么把声学模型和语言模型统一建模, 这就很难只改变声学模型或者语言模型而不影响对方, 而且不能表示无法用 DBN 建模的语言模型.

为了克服这些不足, 本文修改了声学模型和语言模型相互独立的 Token 传递模型, 使之适用于基于 DBN 的连续语音识别. 本文在此基础上提出了一个连续语音识别的基本框架, 继承了目前成熟的基于 HMM 框架的大部分优点, 并将 DBN 的可解释性、可分解性及可扩展性引入该框架. 在这个框架下, 本文提出了一个新的独立于上层语言模型的基于 DBN 的 Token 传递算法. 最后, 介绍了作者开发的一个基于本文所提框架的可用于连续语音识别及其他时序系统的工具包.

本文提出的框架在灵活性和可扩展性方面相对于现有研究来说是较高的, 但仍然存在一些为了简化而作的限制. 我们未来研究工作的重点是扩展该框架和识别算法, 尽量减少这些限制, 使其更具灵活性和可扩展性.

参 考 文 献

[1] Young S. A review of large-vocabulary continuous speech recognition [J]. IEEE Signal Processing Magizine, 1996, 13(5): 45-56

[2] Young S, Evermann G, Gales M, *et al*. The HTK book (for HTK version 3.3) [EB/OL]. (2007-07-10) [2005-07-25]. http://htk.eng.cam.ac.uk/

[3] Young S, Russell N, Thornton J. Token passing: A simple conceptual model for connected speech recognition systems, CUED/F-INFENG/TR38 [R]. Cambridge: Engineering Department Cambridge University, 1989

[4] Dean T, Kanazawa K. Probabilistic temporal reasoning [C]//Proc of AAAI88. Menlo Park, CA: AAAI Press, 1988: 524-528

[5] Zweig G. Speech recognition with dynamic Bayesian networks [D]. Berkeley: U C Berkeley, 1998

[6] Zweig G, Russell S. Speech recognition with dynamic Bayesian networks [C]//Proc of AAAI98. Menlo Park, CA: AAAI Press, 1998

[7] Stephenson T, Doss M, Bourlard H. Speech recognition with auxiliary information [J]. IEEE Trans on Speech and Audio Processing, 2004, 12(3): 189-203

[8] Nefian A, Liang L, Pi X, *et al*. Dynamic Bayesian networks for audio-visual speech recognition [J]. EURASIP Journal on Applied Signal Processing, 2002(11): 1274-1288

[9] Livescu K, Glass J, Bilmes J. Hidden feature models for speech recognition using dynamic Bayesian networks [C] // Proc of Eurospeech'03. Bonn: ISCA Archire, 2003: 2529-2532

[10] Wu Zhiyong, Cai Lianhong. Audio-visual bimodal speaker identification using dynamic Bayesian networks [J]. Journal of Computer Research and Development, 2006, 43(3): 470-475 (in Chinese)
(吴志勇, 蔡莲红. 基于动态贝叶斯网络的音视频双模态说话人识别[J]. 计算机研究与发展, 2006, 43(3): 470-475)

[11] Bilmes J. Dynamic Bayesian Multinets [C] //Proc of the 16th Conf on Uncertainty in Artificial Intelligence. San Francisco: Margan Kaufmann, 2000: 38-45

[12] Daoudi K, Fohr D, Antoine C. Dynamic Bayesian networks for multi-band automatic speech recognition [J]. Computer Speech and Language, 2003, 17(12): 263-285

[13] Bilmes J. Graphical models and automatic speech recogniton [C] //Mathematical Foundations of Speech and Language Processing. New York: Springer, 2003

[14] Bilmes J, Zweig G. The graphical models toolkit: An open source software system for speech and time-series processing [C] //Proc of IEEE Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2002: 3916-3919

[15] Deviren M, Daoudi K. Continuous speech recognition using dynamic Bayesian networks: A fast decoding algorithm [C] // Electronic Proc of the 1st European Workshop on Probabilistic Graphical Models. Cuence, Spain: [s. n.], 2002

[16] Bilmes J, Bartels C. Graphical model architectures for speech recognitin [J]. IEEE Signal Processing Magazine, 2005, 22(9): 89-100

[17] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference [M]. San Francisco: Morgan Kaufmann, 1988

[18] Lauritzen S, Spiegelhalter D. Local computation with probabilities on graphical structures and their application to expert systems [J]. Journal of Royal Statistics B, 1988, 50(2): 157-224

[19] Jensen F, Lauritzen S, Olesen K. Bayesian updating in causal probabilistic networks by local computations [J]. Computer Statist Quart, 1990, 5(4): 269-292

[20] Murphy K. Dynamic Bayesian networks: Representation, inference and learning [D]. Berkeley: U C Berkeley, 2002

[21] Huang C, Darwiche A. Inference in belief networks: A procedural guide [J]. International Journal of Approximate Reasoning, 1996, 15(3): 225-263

[22] Aji S, McEliece R. The generalized distributive law [J]. IEEE Trans on Information Theory, 2000, 46(2): 325-343

Miao Duoqian, born in 1964. Professor and Ph. D. supervisor. Senior member of China Computer Federation. His main research interests include machine learning, rough set theory, granular computing, Web intelligence, data mining, and artificial intelligence.

苗夺谦,1964 年生,教授,博士生导师,中国计算机学会高级会员,主要研究方向为机器学习、粗糙集、粒度计算、Web 智能、数据挖掘、人工智能等.

Wang Ruizhi, born in 1968. Ph. D. candidate and lecturer. Her main research interests include biclustering, rough set theory, and machine learning.

王睿智,1968 年生,博士研究生,讲师,主要研究方向为异质聚类、粗糙集、机器学习等.

Ran Wei, born in 1983. Master candidate. His main research interests include pattern recognition, artificial intelligence, speech processing, *etc.*

冉　巍,1983 年生,硕士研究生,主要研究方向为模式识别、人工智能、语音处理等.

Research Background

Although the introduction of dynamic Bayesian networks into speech recognition in many areas is a huge success, the frameworks and recognition algorithms for DBN-based continuous speech recognition are not as mature and flexible as those for HMM-based one. The flexibility and extensibility of DBN-based continuous speech recognition have not been addressed well in previous work yet. After reviewing the HMM-based token passing model and current DBN-based speech recognition algorithms, we are trying to combine the advantages of both of them. The general framework and its token passing model are proposed in this paper to achieve this goal. This work is supported by the National Natural Foundation of China (No. 60775036 and 60475019) and the Research Fund for the Doctoral Program of Higher Education (No. 20060247039).