基于动态贝叶斯网络的大词汇量连续语音识别和音素切分研究*

吕国云[1]，蒋冬梅[1]，张艳宁[1]，赵荣椿[1]，Hichem Sahli[2]

(1. 西北工业大学 计算机学院，陕西 西安 710072；2. 布鲁塞尔自由大学 电子与信息处理系，比利时 布鲁塞尔B-1050)

摘 要：提出一个新颖的单流多状态动态贝叶斯网络(Single stream Multi-states Dynamic Bayesian Network，SM-DBN)模型，以实现大词汇量连续语音识别和音素切分。该模型在Bilmes等人提出的单流动态贝叶斯网络(Single stream Dynamic Bayesian Network，Phone-shared，SS-DBN-P)模型(识别基元为词)基础上，增加了一个隐含的状态节点层，每个词由它的对应音素组成，而音素采用固定个数的状态描述，状态和观测向量直接连接。它的识别基元为音素，描述了音素的动态发音变化过程。大词汇量语音识别的实验结果表明：在纯净语音环境下，SM-DBN模型的识别率比HMM和SS-DBN-P模型的识别率分别提高了13.01%和35.2%，而音频流的音素切分正确率则分别提高了10%和44%。

关 键 词：动态贝叶斯网络音视频语音识别音素切分

中图分类号：TP391.42 **文献标识码**：A **文章编号**：1000-2758(2008)02-0173-06

近年来，采用动态贝叶斯网络(Dynamic Bayesian Network，DBN)进行语音识别成为一个研究热点[1]。最初的一些动态贝叶斯网络结构是用来模拟标准的隐马尔可夫模型(HMM)及其它的扩展模型[2,3]如factorial HMM，couple HMM等等。最近，Bilmes等人提出一个单流的DBN模型用于连续语音识别[4,5]，这个模型显式地描述了词、音素(整词状态)、观测向量以及他们之间的条件概率分布。每个词由它的对应组成音素构成，而每个音素和观察向量联系并采用高斯混合模型来描述，对于每个词，采用它的组成音素以及音素之间的状态转移概率描述词的动态发音过程，也可以称为共享音素(Phone-share)的单流DBN(SS-DBN-P)模型。同时，针对连续语音识别，Bilmes还构建了一个整词(Whole-Word)结构的单流DBN(SS-DBN-WW)模型，在这个模型中，没有体现音素节点，每个词采用了固定个数的整词状态来描述。报告采用一个图模型工具包(GMTK)对数字连接词数据库进行了语音识别实验[6,7]，和HMM比较，更好的识别结果被得到。

然而上述2个模型在本质上是一个词模型，描述了词的动态变化过程，没有描述音素的动态变化过程，而且SS-DBN-WW模型中没有音素节点，不能进行音素切分，因此这两个模型仅仅适合于小词汇量的语音识别任务。

在上述模型的基础上，本文提出一个新颖的单流多状态(Single-stream Multi-states)DBN模型，简写为SM-DBN模型，这个模型引入了状态节点，每个词由它的对应组成音素构成，每个音素采用固定个数的状态来描述，而状态和状态之间的转移概率关系反映了音素的动态变化过程，每个状态节点都采用高斯混合模型来描述。模型可以输出词识别序列和带时间边界的音素序列。

1 SM-DBN模型介绍

动态贝叶斯网络(DBN)是贝叶斯网络(BN)随时间变化的一个动态扩展，它是由一系列变量节点

* 收稿日期：2007-03-07 基金项目：中国科技部与比利时国际合作项目(No.[2004]487)资助

作者简介：吕国云(1975-)，西北工业大学博士生，主要从事音视频语音信号处理。

和表示节点之间概率关系的有向连接弧组成。从语音识别的角度来说,HMM 模型是DBN 模型的一个特例,而DBN 是HMM 的一般化,它具备灵活的结构和扩展性能,以显式的方式描述了词、音素、状态、观测向量等之间关系。图1描述了SS-DBN-P 模型的结构,它由头(Prologue),中间层(Chunk)和尾(Epilogue)组成,Prologue 是模型的初始化部分,Chunk 可以随着时间 t 不断地扩展和延伸,而Epilogue 表示DBN 模型的结束,从语音识别的角度,表示一句话的结束。带阴影的节点表示可观测节点,而没有阴影的节点表示隐含节点。对于图中节点之间的条件连接概率,实线连接表示确定性的概率,虚线连接表示需要学习的条件概率。图2中描述了几个模型的节点的层次结构关系,图2(a)和图2(b)的本质是一个词模型,语音识别的基元是词,描述了发音时词的动态变化过程。图2(a)是SS-DBN-WW 模型,每个词由固定个数的整词状态组成;图2(b)对应SS-DBN-P 识别模型,有词、音素、观测向量3层节点分布,每个词由它的组成音素构成,音素被所有词共享,每个音素和观测向量直接联系,没有反映音素的动态变化过程,而词的动态过程由组成音素节点及他们之间的转移概率来描述。应该说,SS-DBN-P 模型是词模型的另一种表达形式,采用了共享音素来代替整词状态。

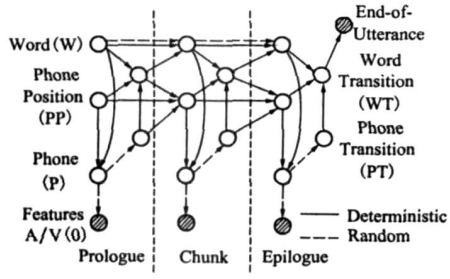

图1 SS-DBN-P 识别模型

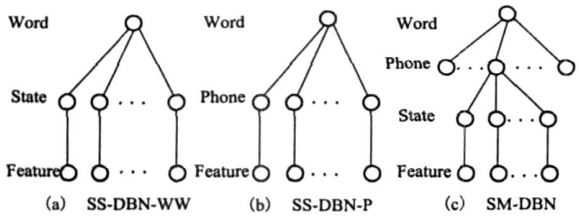

图2 各个节点之间的层次关系描述

借鉴用于连续语音识别的HMM 思想,在本文中,采用音素作为识别基元,用于对大词汇量的连续语音识别和音素切分,基于SS-DBN-P 模型,增加了状态节点,提出了一个新颖的单流多状态DBN (SM-DBN)模型,它的节点的层次关系见图2(c),识别模型结构如图3所示,在这个模型中,词由它的对应组成音素组成,每个音素由固定个数的状态构成,状态节点及其状态之间的转移概率描述了发音时音素的动态变化过程,每个状态和特征观察向量相联系,采用高斯混合模型来描述,这个模型适合于完成大词汇量的语音识别任务。

图3中,括号内为节点变量的简称,W 为词节点,WT 为词转移概率,P 为音素,PP 为音素在词中的位置,PT 为音素转移概率,SP 表示状态在音素中的位置,S 是状态节点,ST 表示状态转移概率,O 为音频特征,End-of-Utterance 表示DBN 模型的结束。假设DBN 模型共有 T 帧,t 表示是第 t 个时间帧,$t \in [1 \cdots T]$,那么识别模型的联合概率分布为:

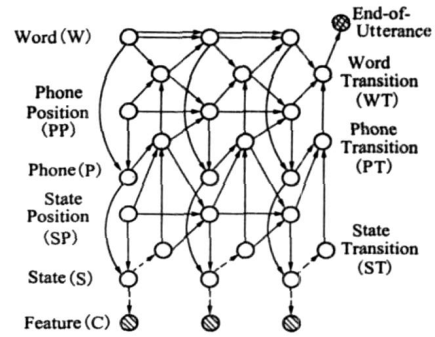

图3 SM-DBN 语音识别模型

$$
\begin{aligned}
&P(W_{1:T}, WT_{1:T}, WP_{1:T}, PT_{1:T}, P_{1:T}, PP_{1:T}, \\
&\quad ST_{1:T}, S_{1:T}, O_{1:T}) \\
&= \prod_{t=1}^{T} P(O_t \mid S_t) P(ST_t \mid S_t) \cdot \\
&\quad P(S_t \mid SP_t, P_t) P(SP_t \mid ST_{t-1}, SP_{t-1}, PT_{t-1}) \times \\
&\quad P(P_t \mid PP_t, W_t) P(PT_t \mid P_t, SP_t, ST_t) \cdot \\
&\quad P(PP_t \mid PT_{t-1}, PP_{t-1}, WT_{t-1}) \times \\
&\quad P(WT_t \mid W_t, PP_t, PT_t) P(W_t \mid W_{t-1}, WT_{t-1}) \quad (1)
\end{aligned}
$$

为更好地理解模型,下面详细描述了主要节点变量及其它们的条件概率分布(Conditional Probability Distribution, CPD)关系。

(1) 特征向量节点(O):类似于HMM 中的特征观测向量,采用连续高斯混合模型来描述如下

$$b_{S_t}(O_t) = f(O_t \mid S_t) = \sum_{k=1}^{M} \omega_{S_t k} N(O_t, \mu_{S_t k}, \sigma_{S_t k}) \tag{2}$$

$\omega_{S_t k}$ 为权值, $\sum_k \omega_{S_t k} = 1$, $N(O_t, \mu_{S_t k}, \sigma_{S_t k})$ 为高斯模型, $\mu_{S_t k}$ 和 $\sigma_{S_t k}$ 分别为均值和协方差。

(2) 状态转移概率(ST), 表示由当前状态转移到一个状态的概率, 类似于 HMM 中的状态转移概率。

(3) 状态节点(S): 它是 SP 和音素 P 的确定性函数, 在这个函数中, 确定了音素和状态之间的详细关系, 如果给出了音素和状态在音素中的位置, 那么状态就可以得到。公式表示如下

$$p(S_t = j \mid P_t = i, SP_t = m) = \begin{cases} 1 & \text{if } j \text{ is the } m\text{-th state of the phone } i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

(4) 状态在音素中的位置节点(SP): 在初始帧, SP_1 为 0; 在其他帧, 当有音素转移概率发生时, 表示一个音素的结束, 状态位置 SP_t 值也复位为 0, 没有音素的转移时, SP_t 的值由状态转移概率确定, 公式表示为

$$p(SP_t = j \mid SP_{t-1} = i, PT_{t-1} = m, ST_{t-1} = n) = \begin{cases} 1 & m = 1, j = 0 \\ 1 & m = 0, n = 1, j = i+1 \\ 1 & m = 0, n = 0, j = i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

(5) 音素转移概率 (PT): 本文中, 每个音素采用了 4 个状态来表示, 对于给定的音素, 仅仅当本状态在音素中的位置(SP) 为本音素的最后一个状态, 而且有状态转移概率发生时那么有音素转移发生, 公式表示为

$$p(PT_t = j \mid P_t = a, SP_t = b, ST_t = m) = \begin{cases} 1 & j = 1, m = 1, b = \text{laststate}(a) \\ 1 & j = 0, m = 1, b = \sim \text{laststate}(a) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

(6) 音素节点(P): 是它的父节点 PP 和 W 的确定性函数, 在这个函数中, 确定了词和音素之间的构成关系, 如果给出了词和音素在词中的位置, 那么音素就可以得到。公式表示为

$$p(P_t = j \mid W_t = i, PP_t = m) = \begin{cases} 1 & \text{if } j \text{ is the } m\text{-th of the word } i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

(7) 音素在词中的位置节点 (PP): 在初始帧, PP_1 为 0; 在其他帧, 当有词转移概率发生时, 表示一个词的结束, 音素位置 PP_t 值也复位为 0, 没有词的转移时, PP_t 的值由音素转移概率确定, 公式表示如下

$$p(PP_t = j \mid PP_{t-1} = i, WT_{t-1} = m, PT_{t-1} = n) = \begin{cases} 1 & m = 1, j = 0 \\ 1 & m = 0, n = 1, j = i+1 \\ 1 & m = 0, n = 0, j = i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

(8) 词转移概率节点WT: 由于每个词的音素构成不同, 模型需要分别处理, 相当于在模型中嵌入了一个词典, 对于给定的词, 首先判断是否有音素转移概率发生, 如果有, 那么判断是否音素在词中的位置为本词的最后一个音素, 如果是, 即有词转移发生; 否则, 没有词转移发生, 公式表示如下

$$p(WT_t = j \mid W_t = a, PP_t = b, PT_t = m) = \begin{cases} 1 & j = 1, m = 1, b = \text{lastphone}(a) \\ 1 & j = 0, m = 1, b = \sim \text{lastphone}(a) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

式中, $b = \text{lastphone}(a)$ 表示当前音素b 是词a 最后一个音素。

(9) 词节点W: 在识别模型中, 本文采用了二元文法模型, 当没有词转移概率发生时, 词保持不变; 当有词转移概率发生时, 转移到下一个词的概率按照二元文法模型得到, 可以采用下面的公式表示

$$p(W_t = j \mid W_{t-1} = i, WT_t = m) = \begin{cases} \text{bigram}(i, j) & \text{if } m = 1 \\ 1 & \text{if } m = 0 \quad i = j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

式中, $\text{bigram}(i, j)$ 表示由词 i 转移到 j 的概率。

而SS-DBN-P 模型的CPD 表示和SM-DBN 类同, 在确定了上述条件概率关系的基础上, 便可以进行推理和学习, 本文采用了图模型工具GMTK 来实现文中3 个单流DBN 模型的推理、学习和训练[5], 训练方法采用了广义的EM 算法(GEM), 最后采用GMTK 输出词识别序列, 音素识别序列以及音素的时间边界切分。

2　实验和结果评价

为了测试模型, 采用了 2 个数据库, 数字连接词(小词汇量)数据库和连续语音数据库(大词汇量数

据库)来对2个模型进行语音识别和音素切分实验,GMTK和HTK工具包分别用来实现本文中所有DBN模型和HMM模型。

2.1 数据库描述和特征提取

音视频数据库采用中国-比利时听视觉信号处理联合实验室录制的数字连接词音视频数据库和大词汇量音视频数据库。数字语音数据库中有数字0～10,加上停顿SP和静音SIL共13个词,涉及到22个音素(phone),数据库的脚本按照Aurora 2.0数据库的句子顺序录制。本文采用100句纯净的音频数据作为训练数据,另外100句以及相应的加噪语音的语音数据作为测试数据。而对于大词汇量音视频数据库,数据库的脚本生成是采用TMIT数据库生成,为了和将来的双流音视频模型比较,采用了500句纯净语音作为训练数据,另取100句并加各种信噪比的语音作为测试数据,500句训练集包括了1 692个词,74个音素。可以模拟大词汇量的数据库对模型进行测试。

实验中,采用HTK提取音频数据的13维MFCC特征和能量特征,加上一阶和二阶差分数据,即MFCC_D_A,共有42维特征。

2.2 基于数字连接词数据库实验结果及分析

2.2.1 词识别结果

针对数字语音数据库,SS-DBN-WW模型,SS-DBN-P模型以及SM-DBN模型的词识别结果见表1,对于SM-DBN模型,每个音素采用了4个状态来描述,作为比较,基于状态捆绑的三音素HMM的识别结果(采用HTK实现)的识别结果也在表1中列出。可以得到下面结论。

(1) SS-DBN-WW模型在不同的噪声下都有最高的识别率,而相同条件下,SS-DBN-P模型的识别率稍低于SS-DBN-WW模型的识别率,可能的原因是词-整词状态的组成结构优于词-对应音素的组成结构,音素是共享的,而整词状态属于词单独拥有的。

(2) SS-DBN-WW模型和SS-DBN-P模型的性能优于SM-DBN模型,平均的识别率(0～30 db下)分别高18.31%和16.59%,说明在小词汇量的语音识别中,选择词作为基元,由于训练充分,且没有更多的中间参数,识别率优越于选择音素作为基元。而在理论上,对于大词汇量数据库,由于缺少更多的训练数据,选用更小的基元:音素,来进行建模。

(3) 当语音信号的信噪比比较高时(大于15 dB时),SS-DBN-WW和SS-DBN-P模型的识别率略低于基于三音素HMM的识别率,其最可能的原因是此DBN模型采用了单音素模型,而HMM采用的是三音素模型;而当信噪比比较低时(小于或等于15 dB时),SS-DBN-WW和SS-DBN-P模型的识别率比HMM的识别率分别高14.78%和12.56%,具有对噪声更强的稳健性和鲁棒性。

2.2.2 音素时间切分结果

对于SS-DBN-P模型和SM-DBN模型,除了给出词一级的识别结果外,还可以进一步给出音素级的识别和时间切分结果。本文提出一个客观评价标准,首先采用三音素HMM模型训练所有纯净语音数据,然后把音素识别结果作为参考的音素序列。然后将不同信噪比下DBN模型的音素切分结果和音素参考序列进行了比较。对于HMM和DBN的音素识别结果,逐帧比较其音素,假如某句话有C帧,如果DBN模型和参考音素序列的音素相同,则累计积分加一分,否则,积分不变,音素相同的帧数假设为A帧,那么就认为本句话的DBN音素切分和HMM音素切分的相似度为

$$P = A/C \tag{10}$$

表1 数据连接词数据库:DBN和HMM词识别率的对比

系统	词识别率/%							
	0 dB	5 dB	10 dB	15 dB	20 dB	30 dB	Clean	0～30 dB
SS-DBN-WW(MFCC_D_A)	48.65	67.32	72.34	78.10	83.46	97.21	99.11	74.51
SS-DBN-P(MFCC_D_A)	42.94	66.10	71.75	77.97	81.36	96.61	97.74	72.79
SM-DBN(MFCC_D_A)	19.6	28.7	46.41	64.71	81.7	96.08	97.04	56.2
HMM(MFCC_D_A)	30.21	41.0	62.67	74.62	85.67	98.04	98.79	65.36

对于N个样本,通过对每个样本的相似度进行平均得到总的音素切分相似度,按照这个评价

准则, 得到各个信噪比下的SS-DBN-P 模型和SM-　　DBN 模型的音素识别结果, 见表2。

表2　数据连接词数据库: 各个信噪比下的音素时间切分的评价结果

系　统	音素切分正确性/%						
	0 dB	5 dB	10 dB	15 dB	20 dB	30 dB	Clean
SS-DBN-P(MFCC-D-A)	33.1	41.3	45.5	53.2	60.2	79.6	81.5
SM-DBN(MFCC-D-A)	30.3	33.7	40.6	45.5	56.3	69.7	86.3

而由表2可以得知, 在如此严格的判断标准下, 纯净语音下SS-DBN-P 模型和SM-DBN 模型的音素切分结果比较令人满意, 分别达到81.5%和86.3%的相似度。而在主观上和原始wav 文件的比较表明, 本文设计的词-音素-状态的组成结构模型是合理的。

2.3　基于大词汇量数据库的实验结果及分析

2.3.1　词识别结果

采用500句纯净的语音数据训练模型, 然后采用另外100句进行识别, 得到的SS-DBN-P 模型, SM-DBN 模型, 三音素状态捆绑HMM 模型识别率见表3。

表3　大词汇量数据库: DBN 和HMM 词识别率的对比

系　统	词识别率/%						
	0 dB	5 dB	10 dB	15 dB	20 dB	30 dB	Clean
SS-DBN-P(MFCC-D-A)	2.39	5.61	9.07	14.80	17.06	22.79	27.57
SM-DBN(MFCC-D-A)	2.51	5.13	9.11	16.47	29.24	50.48	62.77
HMM(MFCC-D-A)	0.72	1.07	3.46	14.32	27.21	44.87	49.76

结果表明: SM-DBN 模型性能明显优于其他两个模型, 纯净语音下的识别率分别比SS-DBN-P 模型和HMM 分别高35.2%和13.01%, 原因是很明显的, SS-DBN-P 模型本质上是词模型, 在大词汇量数据库中不能得到很好的训练, 而SM-DBN 模型属于音素模型, 模型得到了更为充足的训练。而比较HMM 模型, 由于DBN 模型显式地、更为合理地表达了语音的变化规律, 因此在同等情况下, 明显优于HMM 模型。

2.3.2　音素时间切分结果

对于大词汇量的音素切分结果评估, 依旧采用在数字连接库中提出的评价准则, 根据识别结果, 本文仅仅评估纯净语音下的音素切分结果, 参考序列采用手工标注的音素参考序列, 得到SS-DBN-P 模型、SM-DBN 模型、三音素的HMM 的切分结果分别为26%、70%和60%, 很明显, , 比较起SS-DBN-P 模型与单音素的HMM, SM-DBN 模型明显具备最好的音素切分性能。

3　结论和展望

基于SS-DBN-P 模型, 本文提出一种新颖的SM-DBN 模型, 它模拟了一个连续语音识别的HMM。采用一个小词汇量数据库和一个大词汇量数据库进行了语音识别和音素时间切分实验。结果表明: SM-DBN 模型本质上属于音素模型, 适合进行大词汇量的语音识别, 并且取得了较好的识别率, SS-DBN-P 模型和SS-DBN-WW 模型本质上是词模型, 更适合于小词汇量的语音识别, 而且DBN 模型客观上更好地反映了语音的变化规律, 比HMM 模型有更好的噪声鲁棒性。同时, 通过在模型中描述每个词和它的具体音素的对应关系, 音素识别结果和时间切分序列被得到。

在下一步工作中将扩展SM-DBN 模型, 建立听视觉异步的多流DBN 语音识别和音素切分模型, 实现对大词汇量数据库的听视觉语音识别和听觉流和视频流的时间异步切分, 得到听觉流和视觉流之间的异步关系。

参考文献:

[1] Zweig G. Speech Recognition with Dynamic Bayesian Networks [Ph. D. Thesis]. University of California, Berkeley, 1998

[2] Zweig G. Bayesian Network Structures and Inference Techniques for Automatic Speech Recognition, Computer Speech and Language, 2003, 17: 173～193

[3] Murphy K. Dynamic Bayesian Networks: Representation, Inference and Learning [Ph. D. Thesis], University of Cali-fornia Berkeley, 2002 http://www.cs.ubc.ca/～murphyk/Thesis/thesis.pdf

[4] Bilmes J, Zweig G, Richardson T, et al. Discriminatively Structured Graphical Models for Speech Recognition: JHU-WS-2001 Final Workshop Report, Johns Hopkins Univ, Baltimore, MD, Tech Rep CLSP, 2001, http://www.clsp.jhu.edu/ws2001/groups/gmsr/GMRO-final-rpt.pdf

[5] Bilmes J and Bartels C. Graphical Model Architectures for Speech Recognition. IEEE Signal Processing Magazine, 2005, 22(5): 89～100

[6] Bilmes J, Zweig G. The Graphical Models Toolkit: An Open Source Software System For Speech And Time-Series Processing, Proceedings of the IEEE International Conf on Acoustic Speech and Signal Processing (ICASSP). 2002, 4: 3916～3919

[7] Bilmes J. GMTK: The Graphical Models Toolkit. http://ssli.ee.washington.edu/～bilmes/gmtk/

A Novel SM-DBN Model for Large-Vocabulary Continuous Speech Recognition and Phone Segmentation

Lu Guoyun[1], Jiang Dongmei[1], Zhang Yanning[1],
Zhao Rongchun[1], Hichem Sahli[2]

(1. Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
2. Vrije Universiteit Brussel, Department ETRO, Pleinlaan 2, 1050 Brussel)

Abstract: A novel SM-DBN (Single-stream Multi-state Dynamic Bayesian Network) model is proposed. It is an augmentation of the Single Stream DBN Phone-shared (SS-DBN-P) model proposed by Bilmes et al[4] whose basic recognition units are words, to which we add an extra level of hidden nodes-states, resulting in the SM-DBN model. In our model, a word is composed of its corresponding phones, a phone is composed of a fixed number of states, and a state is associated with the observation features. Essentially, it is a phone model whose basic recognition units are phones. We perform the recognition and segmentation experiments with both continuous digital speech database and large-vocabulary speech database, with the experimental results given in Tables 1 through 3 in the full paper. The experimental results on large-vocabulary and clean speech environment show preliminarily that the speech recognition rate of SM-DBN model is 13.01% and 35% higher than those of the HMM (Hidden Markov Model) and the SS-DBN-P model respectively, and that its phone segmentation accuracy is respectively 10% and 44% higher than the other two models.

Key words: single-stream multi-state dynamic Bayesian network (SM-DBN), continuous speech recognition, phone segmentation