

基于动态贝叶斯网络的汉语方言辨识

周杰¹, 顾明亮^{1,2}, 张宁¹, 杨帆¹

(1. 徐州师范大学 物理与电子工程学院, 江苏 徐州 221116;

2. 徐州师范大学 语言科学研究所, 江苏 徐州 221116)

摘要: 方言的差异性在语音层面上反映在时间序列结构的不同。传统的语音建模方法只能建立稳定的时间序列结构, 而方言语音是典型的动态时变时间序列结构。为了更好地提取方言时间序列结构, 文中采用动态贝叶斯网络(DBN)进行建模分析, 并对DBN的构建方法进行了研究, 这种结构与常用于语音识别中的隐马尔可夫模型的不同之处在于它揭示多个时间片内的节点之间的影响。文中探索了不同结构和参数对识别效果的影响。文中的研究表明动态贝叶斯网络对汉语方言的识别比传统方法要好, 识别率达到了98.9%。

关键词: 动态贝叶斯网络; 汉语方言辨识; 联合树算法

中图分类号: TP391.4

文献标识码: A

文章编号: 1673-629X(2012)11-0179-04

Chinese Dialect Identification Based on DBN

ZHOU Jie¹, GU Ming-liang^{1,2}, ZHANG Ning¹, YANG Fan¹

(1. School of Physics & Electronic Engineering, Xuzhou Normal University, Xuzhou 221116, China;

2. School of Linguistic Science, Xuzhou Normal University, Xuzhou 221116, China)

Abstract: The differentiation of Chinese dialect is the different time series in the phonetic. Traditional speech modeling methods can only establish time series, but the dialect speech is typical time-varying series. It chose dynamic Bayesian networks to model the speech in order to extract the time series structure of dialect speech. It also studied the method to model the DBN structure and the influence of the model complexity on recognition rate. The structures of this paper is more complex than the HMM because these structures notice the influence of the nodes in more than two time series. Experiments show that the DBN method is an excellent method with high rate 98.9%.

Key words: dynamic Bayesian networks (DBN); Chinese dialect identification; junction tree algorithm

0 引言

汉语方言辨识比一般的语种识别更复杂,这是由于方言间的差异性比语种间的差异性更小造成的。目前的汉语方言辨识系统大致由两部分组成,前段是音素识别或类音素识别,它将输入语音转换成一串音素或类音素符号,后段是利用各类方言特有的音素约束关系或类音素间的关联统计信息进行识别。其中音素或类音素识别是系统优劣的关键,常用的音素识别方法是HMM方法^[1,2],类音素识别则多用GMM方法^[3,4]。前者往往需要大量的标注语音,这对于方言辨识具有相当的困难,即使可行,标注成本也太高,有些方言还很难找到合适的语言专家。后者虽然可以不需要标注语音信号,但精度相对要低一些,尤其是GMM

对语音特征只具有空间统计特点,它难于描述语音的时间变化特性。如何在描述特征空间特性的基础上,同时刻画语音的时间变化特性,仍是人们讨论的热点。DBN是目前较为成功的一种统计推理方法,它能够根据给定的数据自动生成推理网络结构及网络参数,语音的时间演化过程正好与DBN有某种相似性。但要想将其引进方言辨识系统中,必须解决以下几个问题,系统参数如何确定、系统结构如何确定、样本数对模型训练的影响等。

1 用于汉语方言辨识的DBN结构的构建

动态贝叶斯网络^[5,6]由若干片静态贝叶斯网络串接而成,静态贝叶斯网络由顶点和有向边组成有向无环图,其中,顶点表示变量,有向边表示变量之间的因果关系。根据语音信号的产生原理,语音信号可以看做两个随机过程联合产生的。其中一个随机过程表示发音状态的演变过程,它是无法观察得到的,属于隐过程;另一个随机过程是观察得到的语音信号,它受控于

收稿日期:2012-03-24;修回日期:2012-06-27

基金项目:国家自然科学基金(61040053)

作者简介:周杰(1988-),男,硕士研究生,研究方向为语音信号处理、模式识别;顾明亮,博士,教授,研究方向为语音信号处理、模式识别、机器学习等。

语音的发音状态。为了清晰地表示两者的关联关系,可用 DBN 网络来表示,其中每一片网络由两个节点组成,见图 1。

图中,在第 t 个时间片内:上层节点表示状态变量,用 $\vec{S}_t = [S_{t1}, S_{t2}, \dots, S_{tN}]^T$ 来表示, $t=1, 2, \dots, T$, T 为语音信号的总帧数(片数), N 是语音信号的状态数;下层节点表示观察矢量,它由语音分帧后提取的特征矢量组成,设第 t 帧特征矢量为 $\vec{X}_t = [x_{t1}, x_{t2}, \dots, x_{tD}]^T$, $t=1, 2, \dots, T$, T 为总的帧数(片数), D 为特征矢量的维数。上层节点决定下层节点的观察值。考虑到音联关系,下层节点一般受同一片内的上层节点或相邻片内的上层节点影响,用有向箭头表示。上层节点之间有跳转存在,但跳转规定发生在相邻片内,且只能向前跳转。

为了表示状态之间和观察变量之间的相互影响关系,引入三个参数 (k, p, f) 来描述。其中第一个参数 k 表示上层节点(状态节点)受第 $t-k, t-k-1, \dots, t-1$ 个时间片内的上层节点的影响; p 和 f 表示第 t 个片内的下层节点(观察节点)受第 $t-p, t-p-1, \dots, t+f$ 个时间片内上层节点的影响。实验中可以调节 (k, p, f) 的值来改变网络结构,这样就可以探究不同的网络结构对方言识别效果的影响。

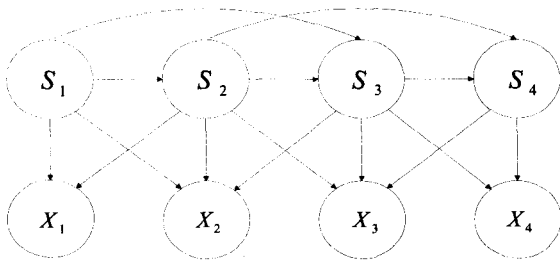


图 1 方言识别中的 DBN 片段结构图
(k, p, f) = (2, 1, 1)

2 用于方言识别的 DBN 网络推理算法

2.1 基本识别原理

设训练得到的 N 个方言的 DBN 网络结构和参数记为: $G^i = \{V^i, E^i\}$, $\theta^i = \{A^i, B^i, \lambda^i\}$, $i=1, 2, \dots, N$ 。待测试方言的语音片段,经特征提取后,得到的特征矢量为: $O = \{\vec{o}_1, \vec{o}_2, \dots, \vec{o}_T\}$, 其中 \vec{o}_t 是第 t 帧的特征矢量。识别时首先将特征矢量输入第 i 个 DBN 模型^[7], 然后计算其输出概率: $P_i(O) = \sum_k P(O, S^{(k)} | G^i, \theta^i)$, 其中, $S^{(k)} = \{s_1^k, s_2^k, \dots, s_T^k\}$ 表示第 k 种可能的状态序列, $s_t^k \in S$ 是第 k 种状态序列中第 t 帧的状态值。最后,根据选择最大的 $P_i(O)$ 所对应的 DBN 方言模型,作为识别结果。

2.2 DBN 的推理算法

计算 DBN 网络模型输出概率可利用 Spiegelhalter 和 Lauritzen 提出的联合树算法^[8,9], 该算法首先将网络简化成树,然后根据所得的树计算概率。

在树状图中对任何一个节点 X_i , 图形都可以分解成三个部分(图 2), 分别为: e_i^0 , e_i^- 和 e_i^+ 。其中, e_i^0 是当前节点 X_i 的观测值, e_i^- 是以 X_i 为父节点的各节点的值, e_i^+ 是以 X_i 为子节点的值。设 $CON(e_i^0)$ 为 e_i^0 所有可能的取值, 若它为观测节点, $CON(e_i^0)$ 就只包含观测值。注意到:

$$P(e, X_i = j) = P(e_i^0, e_i^-, e_i^+, X_i = j) = P(e_i^+, X_i = j) P(e_i^-, e_i^0 | X_i = j, e_i^+) = P(e_i^+, X_i = j) P(e_i^-, e_i^0 | X_i = j) \quad (1)$$

如果 $X_i = j$ 不在 e_i^0 中, $P(e_i^-, e_i^0 | X_i = j)$ 就等于 0。下面的两个变量在推导过程中是关键, 它们将用于计算每个 X_i 。

$$\lambda_j^i = P(e_i^-, e_i^0 | X_i = j) \quad (2)$$

$$\pi_j^i = P(e_i^+, X_i = j) \quad (3)$$

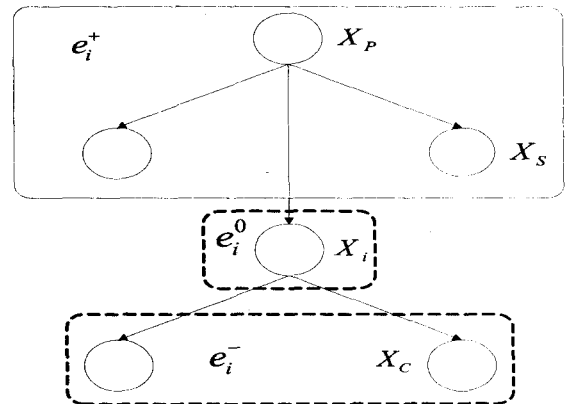


图 2 树状图

这两个变量的定义可以导出:

$$P(O) = \sum_j \lambda_j^i \pi_j^i \quad (4)$$

$$P(X_i = j | O) = \lambda_j^i \pi_j^i / \sum_j \lambda_j^i \pi_j^i \quad (5)$$

这样就可以通过计算 λ 和 π 来对一个网络进行推导, 计算所需要的联合概率和边缘概率。这里的 λ 和 π 类似于 HMM 的 α 和 β 。这两个变量可以通过如下方法计算:

λ 的计算:

如果 X_i 没有子节点

$$\lambda_j^i = 1 \quad (6)$$

如果 X_i 有子节点

$$\lambda_j^i = \prod_{c \in Ch(X_i)} \sum_f \lambda_c^i P(X_c = f | X_i = j) \quad (7)$$

这里 $Ch(X_i)$ 表示 X_i 子节点集, 可以看出要先计算子节点的 λ 。

π 的计算:

如果 X_i 没有父节点,

$$\pi_j^i = P(X_i = j) \tag{8}$$

如果 X_i 有父节点,

$$\pi_j^i = \sum_p P(X_i = j | X_p = v) * \prod_{s \in sib(X_i)} \sum_f \lambda_f^s P(X_s = f | X_p = l) \tag{9}$$

其中 X_p 是 X_i 的父节点,而 $sib(X_i)$ 是 X_i 的兄弟节点集。

3 网络参数的学习

贝叶斯网络的参数学习是为了使网络更好地描述特征变量的数学分布。Heckman 在文献[10]中介绍了贝叶斯网络的一般学习方法。文中结合方言语音特点,将观测节点看作是连续节点,用高斯混合分布来表示其条件概率分布,状态节点看成离散节点,节点维数取 32,即每个状态节点都是一个 32 维的矢量。因为当前的状态对当前的观察语音帧的影响最大,所以当前节点的权重最大,取 0.8,前后两个节点的权重设为 0.1。

网络观测的对数似然:

$$L = \sum_m \log P(D_m) = \sum_h \log \sum_h P(H = h, D = D_m) \tag{10}$$

这里的 H 表示隐节点, O 表示观测节点, D_m 是观测值。EM 算法的基本思想是利用 Jensen 不等式,对凹函数 f 有:

$$f(\sum_j \lambda_j y_j) \geq \sum_j \lambda_j f(y_j) \tag{11}$$

其中, $\sum_j \lambda_j = 1$ 。即 f 的平均大于平均的 f 。由于 \log 函数是典型的凹函数,所以根据公式(11)就转化为,

$$\begin{aligned} L &= \sum_m \log \sum_h P_\theta(H = h, O_m) \\ &= \sum_m \log \sum_h q(h | O_m) \frac{P_\theta(H = h, O_m)}{q(h | O_m)} \\ &\geq \sum_m \sum_h q(h | O_m) \log \frac{P_\theta(H = h, O_m)}{q(h | O_m)} \\ &= \sum_m \sum_h q(h | O_m) \log P_\theta(H = h, O_m) \\ &\quad - \sum_m \sum_h q(h | O_m) \log q(h | O_m) \end{aligned} \tag{12}$$

这里 q 是满足 $\sum_h q(h | O_m) = 1$,且 $0 \leq q(h | O_m) \leq 1$ 的任意函数。最大化下界,对于 q ,就是:

$$q(h | O_m) = P_\theta(h | O_m) \tag{13}$$

这就是 Expectation 过程。

最大化下界,对于 θ 相当于最大化期望 \log 似然度,

$$Q(\theta' | \theta) = \sum_m \sum_h P(h | O_m, \theta) \log P(h, O_m | \theta') \tag{14}$$

选择合适的 θ' ,使得 $Q(\theta' | \theta) > Q(\theta | \theta)$,那么它可以保证使得 $P(D | \theta') > P(D | \theta)$,也就是说增大期望 \log 似然度,相当于增大实际的似然度。这是因为,等式 $q(h | O_m) = P_\theta(h | O_m)$ 保证了下界已经和实际似然度的弧线吻合,所以提高下界,相当于提高实际似然度。

4 实验及结果

4.1 语料和特征提取

汉语方言数据库^[11]是我校语言所(江苏省语言学重点学科单位)方言学专家建立的 HFY 方言语音数据库,该数据库包括普通话、闽方言、粤方言和吴方言。发音者以高校学生为主,也包含了一部分中、老年人。每段录音长度一般为 20 分钟到半个小时。三种方言被切分为 5s,10s,15s 的短时语音各 60 段作为测试语音,训练语音的录音长度为 10 分钟左右。语音信号以 11 KHz 进行采样,16 bit 量化。上述三个语音集语音互不交叉重叠。

文中所用的方言特征主要是 SDC 特征^[12],该特征提取前需要进行一定的预处理,表 1 给出了语音预处理和特征提取所用的参数。

表 1 实验所用参数

参数	预加重	汉明窗	去除静音	窗长	步进	MFCC	SDC
值	0.97	使用	使用	32ms	16ms	10+10	10-1-4-3

4.2 实验比较

为了考察 DBN 在方言识别中的效果,首先对 DBN 网络的参数进行了详细的讨论,然后与 HMM 和 GMM 识别系统进行了比较。这里 HMM 看做为 DBN 的特例,GMM 的参数选择为混合阶数为 32。

4.2.1 DBN 结构复杂度对识别率的影响

探究结构的复杂程度对识别结果的影响,取不同的 (k, p, f) 分别进行实验。实验的结果如图 3 所示。

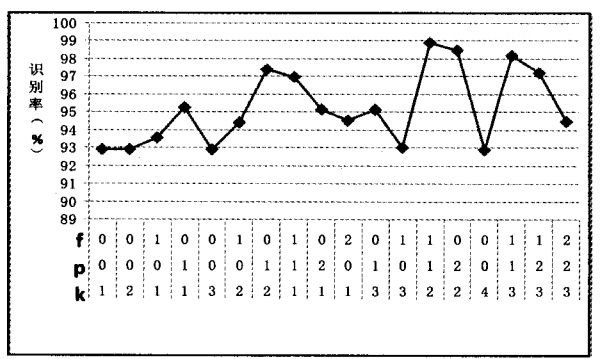


图 3 不同结构复杂度的比较

对结构复杂程度的实验可以看出 $p=0, f=0$ 时, k 的变化对实验结果几乎没有影响。相同复杂程度 $(1, 0, 1), (1, 1, 0)$ 和 $(2, 0, 1), (2, 1, 0)$ 时 p 大的有更高的识别率, 这可能是因为语音信号先前状态对后续的影响比较大。随着结构复杂程度的提高在 $(2, 1, 1)$ 时识别率达到了最高, 随后随着结构复杂度的变高识别率反而下降。说明了语音信号的状态变化与紧靠它之前的一个状态关系比较大, 与更前面的状态几乎没有关系。

4.2.2 HMM 和 GMM 的比较

HMM 可以看成是 (k, p, f) 取 $(1, 0, 0)$ 的 DBN, 用 HMM 的结构进行实验, 实验结果如图 4, 然后将用 GMM 进行汉语方言识别的方法与 DBN 的进行比较, 通过用不同的方法计算它们在 5s, 10s, 15s 不同时间的平均识别率, 实验结果如图 5。

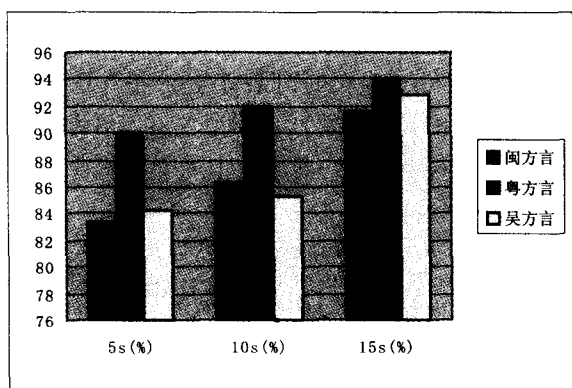


图4 不同时长下的识别率

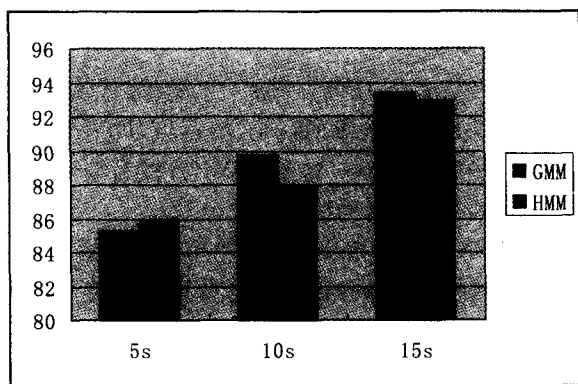


图5 GMM 和 DBN 的识别率比较

从图中可以看出, 随着测试语音长度的增加, 两种识别方法的识别率均有所增加, 但显然 GMM 更适用于长时语音的方言识别, 而 HMM 更适用于短时语音的方言识别。从 GMM 和 HMM 对比的结果中可以看

出随着测试语音时间的增加, 识别率明显升高, 这是因为 GMM 能更好地模拟出统计性, 长时间的语音表现出更强的统计性。而且 HMM 的识别率和 GMM 的识别率相差不大。这是因为 HMM 本身就可以转化为 GMM。

5 结束语

文中利用动态贝叶斯网络的方法对汉语方言进行识别, 比较 GMM 和 HMM 对汉语方言的识别率发现相差不大。同时文中研究了 DBN 结构的复杂度对识别率的影响。实验表明 $(2, 1, 1)$ 时识别率比较理想。如何从数据中学习出动态贝叶斯网络的结构将是今后的工作的主要方向。

参考文献:

- [1] 包亚萍, 郑 骏, 武晓光. 基于 HMM 和遗传神经网络的语音识别系统[J]. 计算机工程与科学, 2011, 33(4): 139-144.
- [2] 韩 普, 姜 杰. HMM 在自然语音处理领域中的应用研究[J]. 计算机技术与发展, 2010, 20(2): 245-248.
- [3] 顾明亮, 夏玉果, 张长水. 基于支撑矢量机的汉语方言辨识[J]. 计算机工程与应用, 2007, 43(29): 210-213.
- [4] 顾明亮, 马 勇. 基于高斯混合模型的汉语方言辨识系统[J]. 计算机工程与应用, 2007, 43(3): 204-206.
- [5] Murphy K. Dynamic Bayesian Networks: Representation, Inference and Learning[D]. USA: U. C. Berkeley, 2002.
- [6] 肖秦琨, 高 嵩, 高晓光. 动态贝叶斯网络推理学习理论及应用[M]. 北京: 国防工业出版社, 2007.
- [7] Deviren M, Daoudi K. Language Modeling Using Dynamic Bayesian Networks[C]//Proceedings of the LREC. Lisbon: [s. n.], 2004.
- [8] Cowell R G, Dawid A P, Lauritzen S L, et al. Probabilistic Networks and Expert Systems[M]. [s. l.]: Springer, 1999.
- [9] 周本达, 王 浩, 姚宏亮. 一又二分之一片联合树算法在动态贝叶斯网络中的应用[J]. 计算机工程与应用, 2005(14): 81-84.
- [10] Heckerman D. A tutorial on learning with Bayesian networks[R]. Redmond, Washington: Microsoft Research, 1995.
- [11] 高 原, 顾明亮, 孙 平, 等. 多用途汉语方言语音数据库的设计[J]. 计算机工程与应用, 2012, 48(5): 118-120.
- [12] Allen F, Ambikairajah E, Epps J. Warped Magnitude and Phase-based Features for Language Identification[C]//ICASSP. [s. l.]: [s. n.], 2006.