ISSN 1000-0054 清华大学学报(自然科学版) 2004 年 第 44 卷 第 1 期 16/36
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2004, Vol 44, No. 1 61-64

汉语连续语音识别中上下文相关的声韵母建模

李 净[1], 郑 方[1,2], 张继勇[1], 吴文虎[1]

(1. 清华大学 计算机科学与技术系, 智能技术与系统国家重点实验室, 北京 100084; 2. 北京得意音通技术有限公司, 北京 100085)

摘 要: 声学建模是汉语连续语音识别中的关键步骤之一。根据汉语语音的特点, 采用扩展声韵母(XIF)作为识别基元, 并针对XIF基元设计了相应的问题集, 利用基于决策树的状态共享策略建立上下文相关声韵模型(Tri-XIF)。将Tri-XIF模型与上下文相关音素模型(Tri-phone)、上下文无关音节模型进行了对比。提出了几种方法用于改善标注、改进问题集和降低模型规模。实验结果表明, Tri-XIF模型与Tri-phone模型、音节模型相比, 识别性能有了很大提高, 其音节误识率分别降低了24.53%和41.65%。采用了所提出的优化策略后, 模型规模降低20%以上, 而性能下降很少。

关键词: 语音识别; 决策树; 上下文相关; 声韵母

中图分类号: TN 912.34 **文献标识码**: A

文章编号: 1000-0054(2004)01-0061-04

Context dependent initial/final acoustic modeling for continuous Chinese speech recognition

LI Jing[1], ZHENG Fang[1,2], ZHANG Jiyong[1], WU Wenhu[1]

(1. State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Beijing d-Ear Technologies, Beijing 100085, China)

Abstract: Acoustic modeling is very important for continuous Chinese speech recognition. The extended Initial/Final (XIF) set chosen as the basic speech recognition unit set to analyze the Chinese language characteristics outperformed the standard IF set. Decision tree-based state tying technology was used to construct the context dependent Initial/Final acoustic model (Tri-XIF model), with an appropriate question set design based on Chinese linguistic knowledge. Methods were developed to optimize the Tri-XIF modeling, including transcription refinement, question set extension, and model size reduction. Tests show that the Tri-XIF modeling is much better than either Tri-phone modeling or syllable modeling, with the syllable error rate reduced by 24.53% relative to the Tri-phone model and 41.65% relative to syllable model. More than 20% model size reduction was obtained with little performance deterioration using the methods in the Tri-XIF model.

Key words: speech recognition; decision tree; context dependent; initial/final

近年来, 对语音识别的研究重心已经从小词表、孤立词的研究逐步转向大词表、连续语音识别, 以及基于连续语音的各种应用。

音节、音素基元均被广泛应用于汉语连续语音识别, 并取得了很好的效果。虽然声韵母基元被认为很好地反映了汉语的特点, 也有一些研究选用标准声韵母作为识别基元, 但对其研究还不够深入细致。本文采用扩展的声韵母为识别基元, 研究如何利用基于决策树的状态共享策略进行了上下文相关建模, 并与音节、音素模型进行对比。

1 识别基元的选择

识别基元的选择原则可以是基于语音学知识的, 也可以是基于数据驱动方式的。在汉语连续语音识别中, 常用的基元包括: 词(word)、音节(syllable)、声韵母(initial/final)和音素(phone)等。

汉语约有400个无调音节和1 300多个有调音节[1]。在进行上下文无关的声学建模时, 选用音节作为基元可以取得比较好的性能。但在连续语音识别中, 音节间的协同发音现象比较严重, 选用音节基元来描述这种现象是十分困难的。

汉语有大约35个音素。音素基元在英语连续语音识别中得到了广泛的应用, 并取得了很好的识别性能[2,3]。对于汉语, 音素也是一个很好的选择。但音素并没有反映出汉语语音的特点, 而且, 相对于声韵母, 音素显得更加不稳定, 这就给标注带来了困难, 进而影响声学建模。

声韵结构是汉语音节特有的结构, 使用声韵母作为识别基元具有以下优点:

1) 声韵结构是汉语的独特音节结构;

2) 上下文关系比较确定(详见第2节);

收稿日期: 2003-01-13

作者简介: 李净(1975-), 男(汉), 河北, 博士研究生。

通讯联系人: 吴文虎, 教授, E-mail: WuWH@tsinghua.edu.cn

3) 有许多相关语音学知识可以应用;

4) 基元数目和语音段长度比较恰当。

2 扩展的声韵母基元定义

表1给出了扩展的声韵母基元定义,一共有27个声母和38个韵母。

表1 扩展的声韵母基元列表

声母基元(27)	韵母基元(38)
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r, -a, -o, -e, -I, -u, -v	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn

与标准的声韵母定义相比,增加了6个零声母{-a, -o, -e, -I, -u, -v},这样,每个音节都是由两部分组成的,分别对应其声母部分和韵母部分。

当使用标准的声韵母基元集合时,有一些音节只有韵母部分,而没有声母部分。所以,当考虑上下文相关信息时,这些韵母既可以搭配声母,又可以搭配韵母,因此,上下文相关声韵母基元数目会很大,超过10万个。而使用扩展的声韵母基元集合时,上下文关系比较确定,基元数目减少为约3万多个,有效地缓解了数据稀疏问题。

同时,使用扩展的声韵母基元也有效地减少识别中的插入错误,其性能也优于标准声韵母基元。因此,本文采用扩展的声韵母作为基元。

3 基于决策树的状态共享策略

在连续语音中,协同发音现象是十分严重的,因此,建立上下文相关模型来刻画协同发音现象是非常必要的。为解决上下文相关建模时的数据稀疏问题,本文采用基于决策树的状态共享策略。

基于决策树的状态共享策略已经广泛地应用于改善大词表连续语音识别系统的声学模型性能[4,5]。决策树是一个二叉树,每个结点都绑定着一个"Yes/No"问题,所有允许进入根结点的HMM状态要回答结点上绑定的问题,根据回答的结果选择进入左枝还是右枝。最后,每个进入根结点的HMM状态都会根据对一系列结点问题的回答进入设定的一个叶子结点。进入同一个叶子结点的HMM状态会被认为是相似的,其参数将被共享起来。它是一种结合了基于数据驱动方法和基于知识方法的方法。与基于数据驱动方法相比,它能够对训练数据稀少的基元和没有训练样本的基元给出适当的参数估计。与基于知识的方法相比,它能够弥补专家知识不足等缺陷。本文使用隐式Markov模型(HMM)描述声学模型。

3.1 问题集的设计

问题集就是供决策树构造使用的问题的集合。结点分裂时选中的那个问题,就与此结点绑定,从而决定哪些基元的哪些状态被共享起来。问题集的好坏会影响到上下文相关模型的性能。

本文中使用的问题集是基于语音学知识的[6]。根据这些先验知识,中心基元的上下文被划分为若干类,每一类作为一个问题。本文针对音素和声韵基元,设计了各自的问题集。

以我们提出的Tri-XIF基元为例,作为问题的声母基元类有(共22个):

响音(Sonorant) {m, n, l}
塞音(Stop) {b, d, g, p, t, k}
唇音(Labial) {b, p, m, f}
塞擦音(Affricate) {z, zh, j, c, ch, q}
⋮

作为问题的韵母基元类有:(共39个)
前高(HighFront) {i, u, v}
开口n(Open-n) {an, en}
开口ng(Open-ng) {ang, eng}
⋮

3.2 决策树的构造

首先将所有可能共享的HMM状态放入一个状态共享池(State Pool)中,然后根据一定的分裂准则进行逐级分裂,当满足停止分裂准则时,分裂过程停止。如图1所示。这是对中心基元为an的扩展声韵母基元的各状态建立决策树的示例。

本文中使用的决策树是基元相关且状态相关的。所使用的分裂准则是最大似然准则,即选择结点

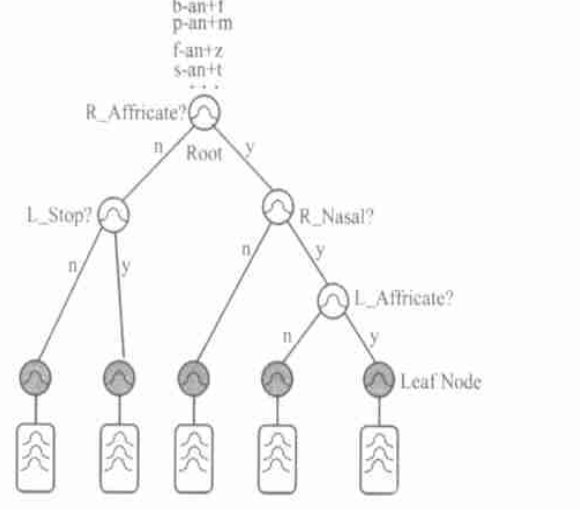

图1 决策树结构图

分裂后似然分增加最大的问题作为本结点绑定的问题。决策树的停止分裂采用阈值进行控制。当分裂后的结点中训练样本数目少于一定数量时, 或者, 当本结点分裂后对数似然分数的增加小于一定的阈值时, 停止分裂。

4 模型训练

对于音节模型, 每个基元使用自左往右的可单步跳转的 6 状态 HMM 来描述, 即每个状态只能驻留或跳转到相邻的下一个状态。对于音素和声韵基元, 使用 3 状态 HMM 来描述。

由于计算量的问题, 一般在利用决策树进行状态共享时, 都使用单混合的初始模型, 而单混合的模型描述能力是非常有限的, 因此, 本文采用混合分裂的方式来增加混合数目, 而最终的混合数目是根据模型的性能和规模综合决定的, 且不同的基元可以使用不同的混合数目。本文采用的是中心分裂准则, 由单混合依次分裂为 2 个、4 个、8 个混合, 并最终使用 8 混合的模型。

5 模型优化策略

为了进一步改进模型, 本文采用了 3 种策略: 1) 采用强制对准解码来重新获得基元的时间切分点, 以改善标注。2) 对问题集进行扩充, 加入双向问题集, 使左右两边的相关性在进行状态共享时同时被考虑进来。3) 考虑到中间状态受两边相关性影响相对较小, 所以通过调整参数, 增加模型中间状态的共享程度, 以减少模型规模。

6 实验结果

6.1 实验条件

本文中使用的数据库是“863 数据库”中的男声数据库[7]。数据库中的句子是用略带口音的普通话读出的。数据库中共有 1 560 个不同的句子, 被划分为 3 组, 分别称为 A、B 和 C 组。库中共有 80 个人的语音数据。标注信息的获得是利用手工标注和机器切分相结合的方法得到的。

本文中从数据库中选取 70 人的数据作为训练集合, 剩余 10 人数据作为测试集合。所以, 测试集合中的说话人都不在训练集合中。

实验中使用 42 维的 (mel-frequency cepstrum coefficients, MFCC) 作为特征参数, 包含能量参数, 以及一阶差分和二阶差分参数, 并且利用倒谱均值归一化方法 (CMN) 来对特征进行归一化, 归一化窗宽定义为 1 s。实验中没有使用语言模型。

本文使用 HTKv3.0 工具进行模型训练[8]。测试结果用连续识别的音节正确率来评价。

6.2 声学模型性能比较

表 2 列出了音节、音素、声韵模型的音节正确率。其中 Phone, XIF, Syllable 分别表示上下文无关音素、声韵和音节模型, Triphone 和 Tri-XIF 为上下文相关音素、声韵模型。

从表 2 可以看出, 对于上下文无关模型, 音节模型的识别率远高于音素和声韵基元模型, 这是因为音节模型使用了更多的参数来描述模型, 音节内部的相关性已经得到了很好的描述。同时, 扩展的声韵母基元的性能也明显优于音素基元。

对于上下文相关模型, 其性能要远远好于对应的上下文无关模型, 也明显优于无关模型中性能最好的音节模型。对于 8 混合模型, 声韵、音素模型与音节模型相比, 其音节误识率分别降低了 41.65% 和 22.68%。这主要是由于引入了上下文相关建模和状态共享策略的缘故。同时, 声韵母与音素模型相比, 音节误识率降低了 24.53%。此时, 声韵模型参数数目多于音素模型, 但规模相当。

表 2 声学模型音节正确率

%

模 型	1 混合	2 混合	4 混合	8 混合
Phone	29.73	38.62	44.41	49.55
XIF	41.60	49.56	55.91	60.11
Syllalbe	58.88	64.46	69.61	73.06
Triphone	70.47	74.45	77.93	79.17
Tri-XIF	76.96	79.67	82.72	84.28

表 3 模型规模

模 型	状态数
Syllable	2 412
Triphone	10 851
Tri-XIF	11 708

对于本文采用的优化策略 1 改善标注, 2 优化问题集, 通过实验发现, 在状态中的混合数目较少时, 这两种策略都可以将识别率提高 1%～2%, 甚至更高。但当混和数增加后, 这两种策略影响变得很小了, 只有少量提高。这主要是因为增加混合数以后, 模型的描述能力已经大大增加, 初始标注和问题集带来的影响已大大减小了。采用优化策略 3, 即加强中间状态的共享程度而得到的模型后, 可以使模型规模降低 20% 以上, 而音节识别率只降低约 0.6%, 说明这种策略还是比较有效的。

7 总 结

本文根据汉语语音的特点,选用扩展声韵母作为识别基元,并利用汉语语音学知识设计了适当的问题集,利用基于决策树的状态共享策略训练上下文相关模型。通过与音节、音素基元的对比,结论如下:

1) 连续语音中协同发音现象严重,进行相关性建模是很有必要的;

2) 所采用的基于决策树的状态共享策略充分利用了汉语语音学知识,并与数据驱动的方式结合,使上下文相关声韵母建模得以实现,并取得了很好的性能,是一种很好的上下文建模方法;

3) 上下文相关声韵母(Tri-XIF)基元是几种基元中的最佳选择,其性能明显优于音节、音素基元。同时,扩展的声韵母与标准声韵母集合相比,大大减少了上下文数目,降低了模型规模;

4) 提出的改进策略在一定程度上优化了模型。

参考文献 (References)

[1] 郑方, 牟晓隆, 徐明星, 等. 汉语语音听写机技术的研究与实现[J]. 软件学报, 1999, **10**(4): 436-444.
ZHENG Fang, MOU Xiaolong, XU Mingxing, et al. Studies and implementation of the techniques for Chinese dictation machines [J]. *J Software*, 1999, **10**(4): 436-444. (in Chinese)

[2] Lee C-H, Rabiner L, Pieraccini R, et al. Acoustic modeling for large vocabulary speech recognition [J]. Computer Speech and Language, 1990, **4**(2): 127-165.

[3] Young S J, Woodland P C. Tree-based state tying for high accuracy acoustic modeling [A]. Proc ARPA Human Language Tech Workshop [C]. Plainsboro, NJ: Morgan Kaufmann Publisher, 1994, 307-312.

[4] Reichl W, Chou W. Decision trees state tying based on segmental clustering for acoustic modeling [A]. Proc Int Conf Acoustics, Speech, Signal Processing'98 [C]. Seattle, Washington: IEEE Press, 1998. 801-804.

[5] Reichl W, Chou W. Robust decision tree state tying for continuous speech recognition [J]. *IEEE Trans Speech and Audio Proc*, 2000, **8**(5): 555-566.

[6] 曹剑芬. 现代语音基础知识[M]. 北京: 人民教育出版社, 1990.
CAO Jianfen. Fundamentals of Modern Chinese Phonetics [M]. Beijing: People's Education Press, 1990. (in Chinese)

[7] ZHENG Fang, SONG Zhanjiang, XU Mingxing. EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine [A]. EuroSpeech '99 [C]. Budapest, Hungary: ISCA, 1999, 819-822.

[8] Yong S, Kershaw D, Odell J, et al. The HTK Book [EB/OL]. http://htk.eng.cam.ac.uk, 2002.

(上接第 60 页)

基于OpenSSL 0.9.7beta4, 实现了TLS的改进的记录协议,并在排除网络和磁盘的干扰下与原协议对比进行了运算性能测试。表1给出了测试结果。

表1 TLS协议与改进协议的运算性能对比

数据	t/ms		节省时间比率/%
	原协议	本文协议	
发送	36.3	32.2	11.0
接受	32.4	32.0	1.0

(CipherSuite 采用 TLS-RSA-WITH-3DES-EDE-CBC-SHA)

其中,记录协议的发送或接收数据的时间分别指记录协议发送或接收 1~40 kB 的数据,并对数据进行签名和验证,所消耗时间的算术平均值。

由上表可知,与原协议相比, TLS 改进协议的运算性能并未降低,反而有一定的提升。

4 结 论

本文提出了一种TLS协议的改进方案,该方案在保持TLS协议原有的安全性的基础上,在TLS协议内部引进了数字签名及验证机制,并且支持数字签名和验证的动态协商。这不仅为TLS协议添加了"抗抵赖"的安全特性,同时在性能上并未增加额外的负担,从而进一步增加了TLS协议的安全性和实用性。

参考文献 (References)

[1] Dierks T, Allen C. RFC 2246, The TLS Protocol, Version 1 [S]. 1999.

[2] 任江, 袁宏春. 对SSL协议及其安全性分析[J]. 电子科技大学学报, 1998, **27**(4): 416-420.
REN Jiang, YUAN Hongchun. Analysis of security SSL [J]. *J University of Electronic Sci and Tech of China*, 1998, **27**(4): 416-420. (in Chinese)

[3] 宋志敏, 王卫京, 南相浩. SSLV3.0及其安全性分析[J]. 计算机工程与应用, 2000, (10): 145-149.
SONG Zhimin, WANG Weijing, NAN Xianghao. SSL V3.0 and the analysis of its security [J]. *Computer Eng and Appl*, 2000, (10): 145-149. (in Chinese)

[4] Rescorla E. SSL and TLS: Designing and Building Secure Systems [M]. USA: Addison-Wesley, 2000.

[5] 候小梅, 莫鸿强, 毛宗源. 基于SSL协议的电子商务解决方案[J]. 计算机工程与应用, 2001, (8): 35-37.
HOU Xiaomei, MO Hongqiang, MAO Zongyuan. E-commerce solution based on SSL protocol [J]. *Computer Eng and Appli*, 2001, (8): 35-37. (in Chinese)

[6] Blake-Wilson S, Nystrom M, Hopwood D, et al. TLS Extensions. Internet-Draft: draft-ietf-tls-extensions-05.txt. [EB/OL]. http://www.ietf.org/internet-drafts/draft-ietf-tls-extensions-05.txt, 2002.