汉语语音识别中基频特征的直接声学建模方法

黄 浩[1],哈力旦[2]
HUANG Hao[1],Halidan[2]

1.新疆大学 信息科学与工程学院,乌鲁木齐 830046
2.新疆大学 电气工程学院,乌鲁木齐 830046
1.Department of Information Science and Engineering,Xinjiang University,Urumqi 830046,China
2.Department of Electrical Engineering,Xinjiang University,Urumqi 830046,China
E-mail:hwanghao@gmail.com

HUANG Hao,Halidan.Direct F0 incorporation for acoustic modeling in Mandarin speech recognition.Computer Engineering and Applications,2009,45(30):132-134.

Abstract: Hidden Conditional Random Fields(HCRFs) based acoustic modeling is proposed by directly using discontinuous fundamental frequency (F0) sequences for Mandarin speech recognition.The method is based on the fact that F0 observations are continuous in voiced portion in Mandarin speech and missing in unvoiced portion,and HCRFs are more suitable for integrating such non-uniform features.Tonal syllable classification tasks are carried out on continuous speech database.Results show HCRFs trained on discontinuous F0 are significantly better than those trained on smooth F0 sequences from artificial interpolation.Comparisons with hidden Markov models under various training criteria are also given.

Key words: hidden conditional random fields;Mandarin speech recognition;acoustic modeling

摘 要:提出了隐条件随机场对断续基音频率序列进行直接声学建模的方法,该方法针对汉语语音中基频值在清音段连续,浊音段断续的特点,根据隐条件随机场区别于隐马尔可夫模型的重要特性——无需对观察值采用统一的建模方式,直接对不连续基频值与连续谱特征观察值一起进行声学建模。大词汇语音库上的汉语带调音节分类实验表明,隐条件随机场下对断续基音频率序列的直接建模较使用清音段人工平滑基频特征的识别率有明显的提高, 还给出了与不同区分性准则训练的隐马尔可夫声学模型的实验性能的比较。

关键词:隐条件随机场;汉语语音识别;声学模型
DOI:10.3778/j.issn.1002-8331.2009.30.041 **文章编号:**1002-8331(2009)30-0132-03 **文献标识码:**A **中图分类号:**TN912.34

1 引言

汉语是一种带调语言, 声调在汉语语音中具有重要的意义。在上下文缺失的情况下,声调在汉语中承担着重要的构字辨义的作用。将声调信息应用于汉语普通话的语音识别系统当中,将会有效地提高识别系统的性能。利用声调信息来提高连续语音识别系统性能的方法在总体上可以分为两类:一种是将基音频率(F0)序列与传统谱特征(如美尔频率倒谱系数MFCC)形成同一特征流来进行模型训练和识别, 称为隐式声调建模。另一种是对声调进行独立建模, 然后在语音识别输出的格(lattice)结构的基础上通过在二次解码中加入声调得分来降低误识率,称为显式声调建模。隐式声调建模由于可以不经过二次解码过程直接获得识别输出,是目前汉语语音识别系统中的主要方法。

图 1(a)显示了微软语音库训练语句 00500500.wav(标记文本为星星点点,若隐若现,孩子出生许久还未消退)的基频序列,如图所示,声调只存在于音节的浊音段,F0 值在整个语音段上是不连续的。由于隐马尔可夫模型(Hidden Markov Model, HMM)对观察矢量采用统一的建模方式,因此在传统 HMM 框

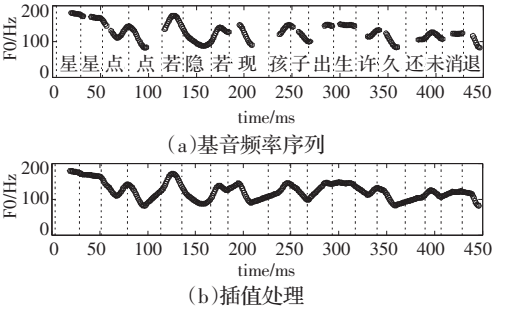

图 1 汉语语音的基音频率序列

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60865001);新疆高校科研计划(Scientific Research Program of the Higher Education Institution of Xinjiang under Grant No.XJEDU2008S15)。

作者简介:黄浩(1976-),男,博士,讲师,主要研究方向:自动语音识别技术、模式识别与机器学习;哈力旦(1959-),女,教授,主要研究方向:智能信号处理、数字图像处理。

收稿日期:2009-06-25 **修回日期:**2009-08-24

架下,原始声调特征难以直接与谱特征一起建模。在连续语音识别中,常采用的处理方法是利用插值的方法对基频不连续区域进行插值(图 1(b))以便于 HMM 进行处理。这种插值产生的 F0 值对识别声调没有贡献,甚至可能带来建模的误差。

条件随机场(Conditional Random Fields,CRFs)[1]是近年来在自然语言处理领域成功应用的数学模型。在自然语言处理应用中,训练数据中观察序列的每个观察值事先经过人工或者自动的标记。而语音识别中,标注语料往往只给出语音数据的模型,对观察值属于哪个状态则是未知的,因此文献[2]提出一种状态隐藏的随机场模型,称为隐条件随机场(Hidden Conditional Random Fields,HCRFs),该模型去除了条件随机场需要观察值需明确标注的局限,在语音识别、图像识别任务中获得了应用。条件随机场(隐条件随机场)较 HMM 的一个优势就是无需对观察值采用统一的处理方式,其特征可以取自观察值的任意函数,因而对观察值缺失的处理十分方便。在自然语言处理系统中,条件随机场多数特征都属于离散特征。这种能力用于处理基频值不连续的现象十分方便。论文将采用具有离散特征的隐条件随机场进行隐式声调建模来解决汉语声调基频观察值不连续的问题。

2　隐条件随机场

2.1　条件随机场

条件随机场的基本定义为:若 o 是一个值可以被观察的输入随机变量集合,s 是一个值能够被模型预测的输出随机变量的集合,这些输出随机变量之间通过指示依赖关系的无向边所连接,CRFs 将输出随机变量值的条件概率定义为与无向图中各个团的势函数的乘积成正比,每个势函数具有如下形式:

$$\exp\left(\sum_j \lambda_j f_j^t(s_{t-1},s_t,o,t)+\sum_k \mu_k f_k^s(s_t,o,t)\right) \tag{1}$$

其中 $f_j^t(s_{t-1},s_t,o,t)$ 表示在整个观察序列中在第 t 时刻和第 $t-1$ 时刻所处状态 s_{t-1}、s_t 发生转移时观察序列 o 的函数,称为转移特征;$f_k^s(s_t,o,t)$ 是 t 时刻处于状态 s_t 时观察序列 o 函数,称为状态特征;λ_j 和 μ_k 是需要从训练数据中估计的模型参数。为叙述简洁,通常将状态特征和转移特征统一表示为 $f(s_i,o,t)$,则线性链的条件随机场定义状态序列的条件概率为:

$$p(s|o)=\frac{1}{Z(o)}\exp\left(\sum_j \lambda_j f_j(o,s)\right) \tag{2}$$

其中归一化因子 $Z(o)=\sum_{s\in H}\exp\sum_k \lambda_k f_k(o,s)$ 是所有状态路径 s 下势函数乘积之和。

2.2　隐条件随机场

HCRFs 同样对条件似然度进行建模,但作为 CRFs 的扩展,HCRFs 在公式(2)中分子部分的观察值所属的具体状态是不可知的,其数学表达为:

$$p(s|o)=\frac{1}{Z(o)}\sum_{s\in h}\exp\left(\sum_j \lambda_j f_j(s,o)\right) \tag{3}$$

其中 h 表示了正确的模型中所有可能的状态路径。将其与公式(2)中 CRFs 条件概率的定义比较,HCRFs 分子部分考虑到所有可能路径的概率之和。两者之间的关系是:当分子部分仅考虑到一条正确的路径时,则 HCRFs 转化为 CRFs。

3　特征选择与参数更新

3.1　基本特征

在 CRFs 中,特征函数的选取不仅决定了对观察值的利用方式,而且特征函数由于可以将前后时刻的观察值作为变量,因而也决定了隐条件随机场的模型结构。首先按照文献[2]选取特征保证 HCRF 与 HMM 有相同的模型结构进行连续 F0 观察序列下的两种模型性能的比较。选择观察值的常数项(零次项)、一次项、二次项使得 HCRFs 模型对观察值的利用方式上与 HMM 相同并与具有相同的模型结构,这些特征称为基本特征,数学表示为:

$$\begin{aligned}
f_{ss'}^{(Tr)}&=\sum_1^T \delta(s_{t-1}=s)\delta(s_t=s') && \forall s,s'\\
f_s^{(Occ)}&=\sum_1^T \delta(s_t=s) && \forall s\\
f_s^{(M1)}&=\sum_1^T \delta(s_t=s)o_t && \forall s\\
f_{ss'}^{(M2)}&=\sum_1^T \delta(s_t=s)o_t^2 && \forall s
\end{aligned} \tag{4}$$

其中 $\delta(s_t=s')$ 当 $s_t=s'$ 时为 1 反之为 0。$f_{ss'}^{(Tr)}$ 为转移特征,用于累加状态转移 s,s' 在路径 s 中发生的次数。从 HMM 的角度来观察,转移特征可看作是 HMM 中的转移概率。$f_s^{(Occ)}$ 为发生概率特征,用于累加状态 s 的发生次数。$f_s^{(M1)}$ 以及 $f_s^{(M2)}$ 为和特征以及平方和特征,用于累计观察值以及观察值平方对齐至状态 s 的累加。

3.2　离散基频特征

不失一般性,可以将公式(4)中特征分为两类:谱观察值的零次、一次、二次表示为 o_t^S,基频观察系数的零次、一次二次项表示为 o_t^T。在使用人工平滑基频观察值时,(o_t^S,o_t^T) 对应参数为 $(\lambda_s^S,\lambda_s^T)$,在状态 s 的概率可表示为 $\exp(\lambda_s^S o_t^S,\lambda_s^T o_t^T)$。如前所述,清音段人工平滑的基频特征在清音段会给概率带来影响,甚至可能带来建模的误差。因此在直接使用断续基频值作为特征时,采用通过加入清浊音指示函数 $voiced(t)$ 来消除伪基频值带来的影响。当前帧为清音帧时,$voiced(t)=0$;当前帧为浊音帧时,$voiced(t)=1$。这时的离散特征为 $(o_t^S,voiced(t)o_t^T)$,t 时刻的状态概率可表示为 $\exp(\lambda_s^S o_t^S+\lambda_s^T voiced(t)o_t^T)$。这样在清音帧时将不会有 F0 相关的观察值对概率值有影响,将这类特征选择方法称为第一类离散特征。

通常来说 $(\lambda_s^S,\lambda_s^T)$ 的大小决定了状态 s 中谱特征流和基频观察值两个流作用程度的大小。采用上述离散特征时,由于谱特征在清浊音段都存在,采用相同的系数不能体现谱观察值在清浊音段的不同作用程度。因此可以将谱特征再按照清浊音分开:$(voiced(t)o_t^S,unvoiced(t)o_t^S,voiced(t)o_t^T)$,对应参数 $(\lambda_s^{SV},\lambda_s^{SU},\lambda_s^T)$ 分别为谱特征在浊音段的参数、谱特征流在清音段的参数和声调特征观察值在清音段的参数,这种特征的选择在浊音段对谱特征给予了不同的特征参数,考虑了谱特征存在的两种发音状态;而在浊音段谱特征和声调特征的参数则考虑了同一状态下两种观察值的作用程度,将这种特征选择称为第二类离散特征。

3.3　参数的优化更新

基于 LBFGS 的梯度更新是条件随机场的参数优化常使用

的方法[3],LBFGS 方法通过利用前几次迭代过程中的一阶导数来近似二阶导数,避免了对二阶 Hessian 矩阵及其逆阵的直接计算,从而使得大规模的非线性优化成为可能。而在该文中将采用文献[2]使用的随机梯度下降法(Stochastic Gradient Descent,SGD)作为 HCRFs 参数的优化方法。由于语音识别的观察值以帧为单位,训练观察值数量巨大,利用 SGD 方法的优点更加体现在训练速度上,该方法只需 5 次左右迭代就能够达到最佳的识别结果。采用 LBFGS 或者 SGD 进行参数更新时,都需要计算目标函数对参数的导数,将目标函数定为对公式(3)对数条件概率加上用于减少过训练的高斯先验:

$$L=\log\sum_{s\in h}\exp\left(\sum_j\lambda_j f_j(s,o)\right)-\log Z(o)-\sum_j\frac{\lambda_j^2}{2\sigma^2} \quad (5)$$

其中 σ^2 是人工选择的高斯先验的方差,由此计算目标函数对参数的偏导数为:

$$\frac{\partial L}{\partial\lambda_j}=\sum_{s\in h}(p(s|o)c_j(s,o))-\sum_{s\in H}(p(s|o)c_j(s,o))-\frac{\lambda_j}{\sigma^2} \quad (6)$$

第一项为所有路径正确路径 h 中经过路径 s 中特征 f_j 发生的次数 $c_j(s,o)$ 与通过路径后验概率 $p(s|o)$ 的乘积之和,第二项为所有路径 H 中经过路径 s 中特征 f_j 发生的次数 $c_j(s,o)$ 与通过路径后验概率 $p(s|o)$ 的乘积之和。随机梯度下降的更新公式为:

$$\lambda_k^{(n+1)}=\lambda_k^{(n)}+\eta^{(n)}\frac{\partial L}{\partial\lambda_k^{(n)}} \quad (7)$$

式中 $\eta^{(n)}$ 是对于样本 $o^{(n)}$ 时的学习速度,根据经验选取,文中采用固定值 $\eta^{(n)}$ 取得较好的结果。$o^{(n)}$ 可以是从训练集中随机选取的样本,而且同一个训练样本也可以处理多次。在该文实验中将采取顺序计算全部 N=454 291 个训练样本,处理所有样本之后利用处理第 n 个样本之后得到的参数进行平均得到本次迭代的参数值来保证参数优化时的稳定性。

4 实验与结果

4.1 数据库与实验配置

将使用带调音节分类实验来验证提出方法对汉语语音声学建模的效果。实验在微软亚洲研究院大词汇量连续语音库[4]基础上进行。训练语料包含 100 个男性发声的 31.5 小时的 19 688 条语句,测试语料包含另外 25 个男性发声的 0.74 小时 500 条语句,共计 9 570 个带调音节。语音数据采样率为 16 bit/16 kHz。

先利用微软最大似然估计(Maximum Likelihood Estimation,MLE)训练的上下文相关 3 状态 8 高斯 HMM 模型进行 Viterbi 对齐来获得每个音节的起始和结束时间。这时每个音节的边界已知,可以将带调音节识别简单看作是一种孤立词识别。对于 MSR 汉语语音语料库,训练集中 454 291 个带调音节分属 1 287 个不同的音节。谱观察向量采用 39 维,包括使用了倒谱均值归一化的 12 阶美尔频率倒谱系数(MFCC)、归一化对数能量及其一阶、二阶导数。基频观察向量采用基于语句归一化的 F0 及其一阶导数 ΔF0。利用 Praat 语音分析工具箱[5]在提取基音频率时加入平滑选项来平滑语句中清音段的不连续基频值。

4.2 结果与分析

首先给出仅使用 39 维 MFCC 观察值的 HMM 声学模型对测试集的带调音节分类结果,HMM 的最大似然训练采用 HTK 工具箱[6]的 HERest 命令进行 Baum-Welch 参数重估直至似然度不再增加为止,不同高斯数下带调音节分类错误率(Tonal Syllable Classification Error,TSCER)为 70.7%~58.0%。对于谱特征 MFCC+平滑 F0 的 HMM 基础上的声学模型,训练准则仍然使用最大似然,当加入基于 F0 的声调特征后,TSCER 降低至 64.3%~49.8%,这里的误识率下降原因是来自于声调相关观察序列的加入。实验还考察了 HMM 对不连续的 F0 观察序列的建模能力(对断续 F0 部分补零),当采用断续 F0 时错误率为 65.4%~50.4%,分类错误率较采用平滑 F0 有所增大,这表明常规结构的 HMM 不适合对这种断续特征序列的建模。

表 1 HMM 带调音节分类错误率 (%)

实验配置	Mix1	Mix2	Mix4	Mix8
最大似然 MFCC	70.7	66.3	62.4	58.0
最大似然 MFCC+平滑 F0	64.3	59.7	53.6	49.8
最大似然 MFCC+断续 F0	65.4	60.3	54.0	50.4
条件最大似然 MFCC+平滑 F0	54.2	49.5	46.0	43.4

对于区分性训练下的 HMM 声学模型下的带调音节分类结果。与最大似然准则的结果差别在于训练方法不同,区分性训练准则采用条件最大似然(Conditional Maximum Likelihood,CML)训练方法。参数更新采用 EBW 算法进行优化[7]。区分性训练利用在最大似然训练的 HMM 单音子模型每个训练音节产生 N 最佳列表作为正确音节的竞争假设。将 N 取值为 20 作为识别率和训练速度的折中。从结果看出,条件最大似然结果较最大似然训练的 HMM 误识率下降为 54.2%~43.4%,这里的误识率的下降来自于基于 CML 准则的区分性训练方法。

表 2 HCRFs 带调音节分类错误率 (%)

实验配置	Mix1	Mix2	Mix4	Mix8
平滑 F0,无先验	53.6	48.0	45.6	43.0
平滑 F0,先验	52.0	46.6	43.9	41.7
断续 F0	49.7	44.8	42.3	40.2
断续 F0 离散特征 II	48.8	44.1	41.8	39.7

接下来考察基于 HCRFs 的声学模型的分类结果。模型参数初始化自最大似然训练的基于平滑 F0 观察序列的 HMM 声学模型参数(表 1 中最大似然 MFCC+平滑 F0 的实验配置)。表 2 先给出了基于平滑 F0 观察序列的隐条件随机场模型,采用无参数平滑,在带调音节分类实验的训练过程中出现了参数优化过程中随着训练迭代次数识别率提高之后反而下降的现象,这时误识率为 53.6%~43.0%,较采用同样特征的区分性训练的 HMM 误识率平均相对下降仅为 1.4%。而加入平滑时分类错误率为 52.0%~41.7%,较区分性训练的 HMM 平均相对误识率下降 4.4%。这表明在同样的线性一阶链式结构、相同平滑 F0 观察序列以及同样的 CML 区分性训练准则下,随机梯度下降法训练的 HCRFs 要优于 EBW 训练的 HMM 的声学模型。

再考察使用离散观察序列的 HCRFs,对于第一类离散特征,误识率为 49.7%~40.2%,较采用平滑 F0 误识率下降 2.3%~1.5%,这表明隐条件随机场非常适合建模这种观察序列有缺失的情况,直接采用离散基频特征能够带来误识率的降低。对于第二类离散特征,分类误识率进一步下降 0.9%~0.5%,这说明条件随机场在加入谱特征在不同发音状态下的相关参数将会较直接利用离散 F0 具有更好的识别结果。而比较隐条件随机场的最优结果 48.8%~39.7%,较 HMM 的最优结果(54.2%~43.4%)误识别率平均下降 9.2%。

(下转 164 页)