第29卷 第17期 计 算 机 工 程 2003年10月
Vol.29 № 17 Computer Engineering October 2003
·人工智能及识别技术· 文章编号：1000—3428(2003)17—0120—02 文献标识码：A 中图分类号：TP311

带噪汉语语音识别的端点检测方法

王 朋，塔维娜，陈树中
（华东师范大学计算机系，上海 200062）

摘 要：在语音识别系统中产生错误识别的原因之一是端点检测有误差，在高信噪比情况下，正确地确定语音的端点并不困难，然而，大多数实际的语音识别系统需工作在低信噪比情况下，一些常规的端点检测方法，例如基于能量的端点检测方法在噪声环境下不能有效地工作。该文利用改进的隐马尔柯夫模型(HMM)进行语音检测以适应噪声的变化，实验结果表明本方法可得到高正确率的带噪语音端点检测。

关键词：语音识别；端点检测；语音检测

Endpoint Detection Method of Noisy Chinese Speech Recognition

WANG Peng, TA Weina, CHEN Shuzhong
(Department of Computer,East China Normal University,Shanghai 200062)

【Abstract】 A major cause of error in automatic speech recognition (ASR) systems is the inaccurate detection of the beginning and ending boundaries of test and reference patterns.Accurate determination of endpoints of speech is not very difficult if the SNR is high.Unfortunately, most practical ASR systems must work with a small SNR, and the conventional speech detection methods based on some simple features such as energy cannot work well in noisy environments. In the paper, modified HMM is used in speech detection to make it adaptive to the change of noise. The experiments show high accurate rates can be obtained.

【Key words】 Speech recongnition; Endpoint detection; Speech detection

在语音识别系统中，语音端点检测的传统方法通常采用语音的短时能量，这些方法在高信噪比(SNR)时具有良好的性能,而在低信噪比时性能很差。本文给出了基于新型HMM的语音检测的改进方法，采用更新噪声模型来改进算法性能。实验结果表明，本文提出的算法具有优越的性能。

1 HMM的修正训练算法

下面对最大分量连续HMM模型(MCHMM)作详尽描述并提出了一个修正的“矫正(CT)训练”算法。

1.1 HMM用于语音识别

定义 $W=\{w_i\}(i\in V)$ 为语音识别系统的语言学识别单元，汉语单音节识别中 w_i 基本上采用音节或声韵母作为单元，V为识别单元的个数。 $Y:O=\{O_t\}_{t=1}^{T_0}$ 为连续特征空间上的语音特征矢量序列, O_t 为K维特征矢量， $O_t=[O_{1t},O_{2t},\cdots,O_{kt}]^T$, $O_t\in R^k$, R^K 为K维欧氏空间, T_0 为语音序列长度。 $\Lambda=\{l_m\}(m\in M)$ 为连续特征空间的一个最佳划分。对应于HMM的每一状态,存在的一个概率分布，并满足

$$\sum_{m=1}^{M}P(l_m/S_i)=1 \quad P(l_m/S_i)\geq 0, \quad i\in N \tag{1}$$

其中， l_m 为划分子空间的标号， l_m 定义在正整数域。当用矢量量化把连续特征序列映射为离散标号序列时，即完成如下的变换过程：

$$O=\{O_t\}_{t=1}^{T_0}\rightarrow B=\{l_{mt}\}_{t=1}^{T_0}$$

在用HMM作语音识别之前，首先用一组训练数据Y估计HMM的参数 $\Omega_V=[\Pi_v,A_v,F_v]$ 。通常采用最大似然训练算法，即

$$\Omega_V=\arg\max_V P(Y/\Omega,w_v) \tag{2}$$

估计过程采用著名的Baum-Welch前后向算法。

语音识别的过程是设计一个最佳解码器：

$$w_v=\arg\max_V P(W/Y) \tag{3}$$

式(3)中的P(W/Y)可以分解为3个部分：

$$P(W/Y)=P(W)P(Y/W)P(Y) \tag{4}$$

式(4)中与声学模型有关的是P(Y/W)，可以采用Viterbi算法得到一个最大似然解码器。

1.2 HMM参数的最大似然估计

HMM模型中与状态相联系的概率函数一般可以表达为

$$f_i(O_t)=\sum_{m=1}^{M}P(O_t/l_m,S_i)P(l_m/S_i) \tag{5}$$

式中 $P(l_m/S_i)$ 为 S_i 状态上的概率函数，它满足式（1）；$P(O_t/l_m,S_i)$ 为任意对数凸函数或椭圆对称函数。此时式(2)中的似然函数是关于模型参数的非凸函数，利用EM算法可以逼近一个局部最优点。利用前后向公式可以导出A, $P(l_m/S_i)$ 即当 $P(O_t/l_m,S_i)$ 为高斯分布时其均值 m_{ij} 和方差 $\sum ij$ 的估计公式。

定义前向概率为

$$a_t(i)=P(O_1,O_2,\cdots,O_t,S_t=S_t/\Omega) \tag{6}$$

后向概率为

$$b_t(i)=P(O_{t+1},O_{t+2},\cdots,O_{T_0}/S_t=S_i,\Omega) \tag{7}$$

初始条件为

$$a_1(i)=p_i,b_{T_0}(i)=1(i=N);b_{T_0}(i)=0 \quad \text{i为其他}$$

定义两个辅助概率函数：

$$g_t(i,j)=P(S_{t-1}=S_i,S_t=S_j/O,\Omega)=a_{t-1}(i)a_{ij}f_i(O_t)b_t(j)/P(O/\Omega) \tag{8}$$

作者简介：王 朋（1972—），男，硕士，研究方向：模式识别，神经网络；塔维娜，硕士生；陈树中，教授、博导

收稿日期：2002-08-30 修回日期：2002-10-30

$$x_t(i,j,k) = P(S_{t-1}=S_t, S_t=S_j, l_{kt}/O,\Omega)$$
$$= a_{t-1}(i)a_{ij}P(l_{kt}/S_j)P(O_t/l_{kt},S_j)b_t(j)/P(O/\Omega) \quad (9)$$

式(9)中，$l_{kt}=T_{VQ}[O_t], T_{VQ}$ 为矢量量化映射变换或聚类算子。

采用EM算法导出如下参数估计公式

$$\hat{a}_{ij} = \sum_{t=1}^{T_0} g_t(i,j) \Big/ \sum_{t=1}^{T_0}\sum_{j=1}^{N} g_t(i,j), i,j \in N \quad (10)$$

$$\hat{P}(l_m/S_i) = \sum_{t=1}^{T_0}\sum_{k=1}^{N} x_t(k,i,m) \Big/ \sum_{t=1}^{T_0}\sum_{j=1}^{N} g_t(i,j), i\in N, m\in M \quad (11)$$

$$\hat{m}_{ij} = \sum_{t=1}^{T_0}\sum_{k=1}^{N} x_t(k,i,j)O_t \Big/ \sum_{t=1}^{T_0}\sum_{k=1}^{N} x_t(k,i,j), i\in N, j\in M \quad (12)$$

$$\hat{\sum}_{ij} = \sum_{t=1}^{T_0}\sum_{k=1}^{N} x_t(k,i,j)(O_t-m_{jk})(O_t-m_{jk})^T \Big/ \sum_{t=1}^{T_0}\sum_{k=1}^{N} x_t(k,i,j), i\in N, j\in M \quad (13)$$

式（10）~（13）为普遍适用的公式，在不同的模型形式下有相应的变化。

1.3 修正训练算法

HMM的经典训练方法是采用最大似然（ML）准则，ML估计在模型假设是正确实施一种最佳的估计器，但由于语音信号并非马尔柯夫源，因此ML训练只能保证训练过程的最佳性而不能一定能保证识别过程也是最佳。本文将Bahl等提出的“矫正训练（CT）算法”推广到非特定人MCHMM识别系统中，取得较好的效果。

对高斯分布的均值矢量 $[m_{ij}]$ 作CT训练，其他参数如A，$[C_{ij}]$ 和 $[\sum_{ij}]$ 均不改变。整个训练的过程如下所述。

(1) 初始化模型

用所述的前后向算法得到非特定人汉语普通话声韵母的MCHMM模型作为CT训练的初始模型记为 $\Omega_1=[A,m,\sum]$。

(2) 训练数据的似然概率打分

对声母和韵母采用不同的处理方法，全部韵母单元视为一类。声母则分成9个子类，即具有相同后续韵头的声母视为一类。CT训练的修改只在同一类单元之内进行。

设发音序列O属于 W_v（W_v 称正确单元），在 W_v 相应的子类内进行似然概率打分，得到似然概率比

$$R_v = \lg P(O/W_v,\Omega_v) - \lg P(O/W_x,\Omega_x) \quad (14)$$

其中 W_x 为子类中的其他单元，称为非正确单元，把属于 W_x 的所有满足：$R_v < d$ 的发音构成修正用数据集合 O_c。

(3) 模型参数的修改

1) 正确单元的修改

用 $\{W_v,\Omega_v\}$ 对 O_c 作Viyerbi分段把每次发音均最佳分割成 N_v 段（N_v 为 Ω_v 的状态数），然后把每段对应的数据进行聚类，求出该段数据的聚类中心 $m_{vci}(i\in N_v)$。用 m_{vci} 对 Ω_v 进行修改

$$\hat{m}_{vim}^{(t-1)} = \hat{m}_{vim}^{(t-1)} + h_1(m_{vci}-\hat{m}_{vim}^{(t-1)}) \qquad i\in N_v, m\in M \quad (15)$$

其中

$$m = \arg\max_k \{C_{vik}\exp[-\frac{1}{2}(m_{rci}-\hat{m}_{vik})^T\sum_{vik}^{-1}(m_{vct}-\hat{m}_{vik})$$

2) 非正确单元的修改

把 O_c 根据识别结果和式(15)的判别结果分成S组，S为 O_c 中所包含的误识单元的个数。在每一组内用 Ω_s 对 O_c 中的相应序列进行分段、聚类，得到相应于每一非正确单元的每一状态的修正矢量 $m_{scj}, j\in N_s, s\in S$，$N_s$ 为 Ω_s 的状态数，那么，对非正确模型修改方法为

$$\hat{m}_{sjn}^{(t)} = \hat{m}_{sjn}^{(t-1)} - h_2(m_{scj}-\hat{m}_{sjn}^{(t-1)}), j\in N_s, n\in M \quad (16)$$

其中

$$n = \arg\max_k \{C_{sjk}\exp[-\frac{1}{2}(m_{scj}-\hat{m}_{sjk})^T\sum_{sjk}^{-1}(m_{scj}-\hat{m}_{sjk})]\}$$

(4) 依据所有训练数据对所有模型作上述识别，修改后即完成一次迭代，统计每次迭代的总误识率，如果连续两次迭单的误识率相同则停止迭代，训练结束，否则重复(2)~(4)步骤。

上述调整中采用的参数最佳值一般通过实验获得，选取：$d=10, h_1=0.1, h_2=0.05$。

2 基于改进的HMM的端点检测

改进的HMM语音端点检测器以及所采用的语法模型见图1、图2。

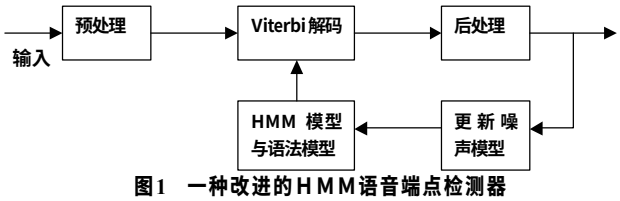

图1 一种改进的HMM语音端点检测器

图2 语法模型

在给定时间T内，根据先前在解码时确定的噪声帧所提供的信息来更新噪声模型，假定噪声HMM模型在M个混合高斯概率密度函数时具有一个状态，更新规则为：(1) 给定时间T，从已经检测白噪声帧计算平均倒谱向量 $\vec{c}$；(2) 在观察概率密度函数的M个混合高斯概率密度函数中找到一个均值 $\vec{m}_i$，其序号为i，和 $\vec{c}$ 具有最小汉明距离的函数；(3) 通过更新第 i个观察概率密度函数 $\vec{m}_i$ 来重新估计噪声模型，即

$$\vec{m}_i = p\vec{m}_i + (1-p)\vec{c}$$

式中p为一调节因子。

3 实验结果

语音端点检测方法是在不同的噪声条件下进行测试的。首先，语音信号经8kHz抽样和16bit量化后，与不同电平的白噪声相混合。在所有实验中，语音信号被分为240采样的帧，相邻帧有50%重叠。每帧采用12阶LPC倒谱系数，对每个语音文本通过手工标号以区分语音与背景噪声，可作为测试端点检测正确率的标准。表1给出实验结果。

(下转第135页)

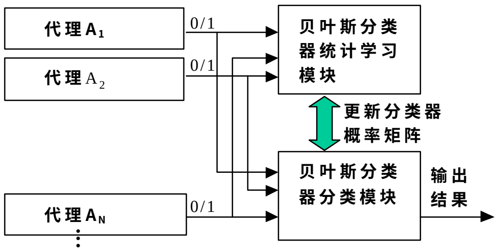

图4 多代理检测结果的数据融合判决

由于每个代理的检测结果只有报警和不报警两种，假设A表示代理发出入侵报警，$\overline{A}$ 表示没有报警，I 和 $\overline{I}$ 分别表示入侵和正常,针对检测结果，有以下情况：

(1)实际系统正常，检测结果正常，该情况对应概率$P(\overline{A}\mid\overline{I})$；

(2)实际系统正常，检测结果异常，该情况对应概率$P(A\mid\overline{I})$；

(3)实际系统异常，检测结果正常，该情况对应概率$P(\overline{A}\mid I)$；

(4)实际系统异常，检测结果异常，该情况对应概率$P(A\mid I)$.

其中(2)、(3)两种情况分别对应虚警和漏警两种误判。在利用贝叶斯分类器进行检测结果的融合时，分类器把各检测结果组成的一维向量$(a_1,a_2,\cdots,a_N)$作为输入，然后分别计算该检测向量出现时，目标系统正常和被入侵的概率$P(\overline{I}\mid a_1,a_2,\cdots,a_N)$和$P(I\mid a_1,a_2,\cdots,a_N)$。

贝叶斯分类算法描述：假设目标函数：$f:X\rightarrow V$，其中X为事例集，V为函数的值域。应用到分类领域，则V就是事例的类别的集合。若每个事例$x\in X$，由它的属性值组成的n维向量$(a_1,a_2,\cdots,a_N)$表示,那么向量x属于类别V的概率为: $P(v_j\mid a_1,a_2,\cdots,a_N)$. 贝叶斯分类器定义目标函数$f$：

$$f(x)=v_{MAP}=\arg\max_{v_j\in V} P(v_j\mid a_1,a_2,\cdots a_N)$$

$$=\arg\max_{v_j\in V}\frac{P(a_1,a_2,\cdots,a_N\mid v_j)P(v_j)}{P(a_1,a_2,\cdots a_N)}$$

$$=\arg\max_{v_j\in V} P(a_1,a_2,\cdots,a_N\mid v_j)$$

为简化计算，假设实例的属性是相互独立的，则有：

$$P(a_1,a_2,\cdots,a_N\mid v_j)=\prod_{i=1}^{N}P(a_i\mid v_j)$$

由此得到简化的贝叶斯分类器：

$$v_{NB}=\arg\max_{v_j\in V} P(v_j)\prod_{i=1}^{N}P(a_i\mid v_j)$$

然后根据分类的结果和设定的阈值来确定目标系统的状态，从以上过程不难看出，分类的有效性取决于提供检测结果的各代理的性能。

该方案在进行结果融合时，考虑了不同代理检测结果的影响，不仅具有结果预测和推理能力，而且利用分类器的自学习功能，有效避免了大量的人工分析工作。

4 结束语

本文通过对多代理的协作问题进行分析以及对网络中海量数据处理的研究表明：一个集成了多个不同的检测子系统的入侵检测系统，通过这些检测子系统的协作可以有效提高整个系统的性能。另外，在面向网络的入侵检测系统中，分布式的代理要进一步加强自身对其收集数据的处理能力，使其输出的数据尽可能精简，从而减少中央数据融合处理代价以及网络传输的开销，这将是进一步工作研究的重点。

参考文献

1 马恒太, 蒋建春, 陈伟锋等. 基于Agent的分布式入侵检测系统模型. 软件学报，2000 ,11 (10)：1312-1319

2 Lee W, Stolfo S J, Mok K W. A Data Mining Framework for Building Intrusion Detection Models. In Proccedings of the 1999 IEEE Sym-posinm on Security and Privacy, 1999

3 Denning D. An Intrusion-detection Model. IEEE Transactions on Software Engineering，1987, 2：222- 232

4 Lee W. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems. The Requirement for the Degree of Doctor, 1999

5 Bass T, Road S. Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems. IRIS National Symposium Dnaft, 1999

6 Bass T. Intrusion Detection Systems & Multisensor Data Fusion. In: Creating Cyberspace Situational Awareness Communications of the ACM to Appear, 2000

☆☆

(上接第78页)

参考文献

1 范玉顺. 企业建模理论与方法学导论. 北京: 清华大学出版社, 2001

2 高洪深. 决策支持系统（DSS） 理论 · 方法 · 案例. 北京:清华大学出版社，2000

3 安徽省水利厅科技处. 安徽水利科技工作文件选编. 安徽省水利厅，2002

4 陈六禹. IDEF 建模分析与设计方法. 北京: 清华大学出版社，1999

5 徐立中. 水利工程信息化与CIMS. 中国电子学会第4届学术年会论文集，2001

☆☆

(上接第121页)

表1 语音端点检测测试实验结果

检测器	正确率	SNR=16db 白噪声	SNR=6db 白噪声	SNR=5db 白噪声
Energy	P(A/S)	0.96	0.73	0.63
	P(A/N)	0.97	0.58	0.50
	P(A)	0.96	0.69	0.57
HMM	P(A/S)	0.97	0.96	0.96
	P(A/N)	0.98	0.75	0.71
	P(A)	0.97	0.89	0.97

注：在表1中,Energy表示基于能量对数的端点检测器，HMM表示基于HMM的端点检测器，P(A/S)表示语音检测的正确率, P(A/N)表示非语音检测的正确率，P(A)表示总的检测正确率。

4 结论

基于改进的HMM的语音端点检测方法在不利的环境下比通常的基于能量的端点检测方法的鲁棒性好。这一特性使其适合实际应用的需要,如在噪声环境下的语音增强与鲁棒语音识别等。