第29卷 第3期 计 算 机 工 程 2003年3月
Vol.29 № 3 Computer Engineering March 2003
· 多媒体技术及应用 · 文章编号：1000—3428(2003)03—0120—02 文献标识码：A 中图分类号： TP391.42

一种语音端点检测方法的探究

刘庆升，徐霄鹏，黄文浩
(中国科学技术大学精密仪器系，合肥 230027)

摘 要：研究了一种以过零率ZCR和能量E为特征的语音端点检测方法。在进行大量实验的基础上，经过分析，对该方法提出了几点改进。

关键词：端点检测；过零率（ZCR）；能量（E）；幅度(M)

Research on a Speech Endpoint Detection Method

LIU Qingsheng，XU Xiaopeng，HUANG Wenhao
(University of Science and Technology of China , Hefei 230027)

【Abstract】A key problem in speech recognition is how accurately speech must be detected so as to provide the best speech patterns for recognition. In this paper, a classic method of a speech endpoint is discussed, and based on a mass of experiments and analysis, some improvements are proposed.

【Key words】Endpoint detection; Zero cross ratio(ZCR); Energy(E); Magnitude(M)

语音端点检测就是检测语音信号的起点和终点，因此也叫起止点识别[1]。它是语音处理技术中的一个重要方面，其目标是要在一段输入信号中将语音信号同其它信号(如背景噪声)分离开来。在语音识别中，一个关键问题就是如何将语音信号精确地检测出来，为获得准确的识别提供前提[2]。

本文作者在进行语音识别算法的研究过程中在对一种经典的端点检测方法—Lawrence Rabiner提出的以过零率ZCR和能量E为特征的起止点检测方法[1]进行研究之后，针对具体的应用提出了几点改进，并且达到了较好的效果。

1 原理

以过零率ZCR和能量E[1,3,4]为特征的起止点算法的根据是背景噪声与语音的短时段ZCR及E特征从统计看都有相当的区别[3]。这里的E特征指的是能量类特征[1]，用到的是该类特征中的短时段平均幅度M特征。

过零率ZCR的定义为：在统计的短时段中，信号波形穿越零电平的次数。

记x(n)为离散语音信号时间序列；ω(n)为时窗函数（其有效长度为N），可以为矩形窗，如Hamming窗等。则ZCR和M的计算式分别为：

$$ZCR = \sum_{m=-\infty}^{+\infty} \left| sign[x(m)] - sign[x(m-1)] \right| \bullet w(n-m)$$

$$M = \sum_{m=-\infty}^{+\infty} \left| x(m) \right| \bullet w(n-m)$$

其中sign[x(n)]=1(当x(n)≥0)或sign[x(n)]= -1(当x(n)<0)；ω(n)=1/2N（0≤n≤N-1）或ω(n)=0（n为其它）。

该方法的要点为：由于采集声音信号的最初的短时段为无语音段，仅有均匀分布的背景噪声信号。这样就可以用已知为"静态"的最初几帧（一般取10帧）信号计算其过零率阈值IZCT及能量阈值ITL（低能量阈）和ITU（高能量阈）。

IZCT的具体计算式为：

$$IZCT = \min[IF, \overline{IZC} + 2 \cdot \sigma_{ZC}] \quad (1)$$

其中IF为固定值，一般取25；$\overline{IZC}$、σ_{ZC}分别为根据所取最初10帧样值算得的过零率的"均值"和"标准差"。

计算ITL和ITU时，先算出最初10帧信号每帧的平均幅值M，最大者记为IMX，最小者记为IMN。然后令：

$I_1 = 0.03 \times (IMX - IMN) + IMN$

$I_2 = 4 \times IMN$

最后按下式计算出ITL和ITU：

$$ITL = \min(I_1, I_2) \quad (2)$$

$$ITU = 5 \times ITL \quad (3)$$

接下来就可以用过零率阈值IZCT及能量阈值ITL（低能量阈）和ITU（高能量阈）来进行起点及止点的判别。

先根据ITL、ITU算得一初始起点N_1。方法为从第11帧开始，逐次比较每帧的平均幅度，N_1为平均幅度超过ITL的第一帧的帧号。但若后续帧的平均幅度在尚未超过ITU之前又降到ITL之下，则原N_1不作为初始起点，改记下一个平均幅度超过了ITL的帧的帧号为N_1，依此类推，在找到第一个平均幅度超过ITU的帧时停止比较。

N_1只是根据能量信息找到的起点，还未必是语音的精确起点。这是由于语音的起始段往往存在着能量很弱的清辅音(如[f]、[s]等)，仅依靠能量很难把它们和无声区分开。但研究发现它们的过零率明显高于无声段，因此可以利用过零率这个参数来精确判断清辅音与无声区二者的分界点[4,5]。图1显示的是语音"升高"的波形、平均幅值和过零率。从图中可以十分明显地看到清辅音能量弱而过零率高的特点。

当N_1确定后，从N_1帧向N_1-25帧搜索，依次比较各帧的过零率，若有3帧以上的ZCR≥IZCT，则将起点N_1定为满足ZCR≥IZCT的最前帧的帧号，否则即以N_1为起点。这种起点检测法也称双门限前端检测算法[4]。

语音结束点N_2的检测方法与检测起点相同，从后向前搜索，找一第一个平均幅度低于ITL、且其前向帧的平均幅度

作者简介：刘庆升（1975~），男，硕士生，研究方向为声控产品及语音识别芯片开发；徐霄鹏，硕士；黄文浩，教授、博导
收稿日期：2002-02-22

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在超出ITU前没有下降到ILT以下的帧的帧号，记为N_2，随后根据过零率向N_2+25帧搜索，若有3帧以上的ZCR≥IZCT，则将结束点N_2定为满足ZCR≥IZCT的最后帧的帧号，否则即以N_2作为结束点。

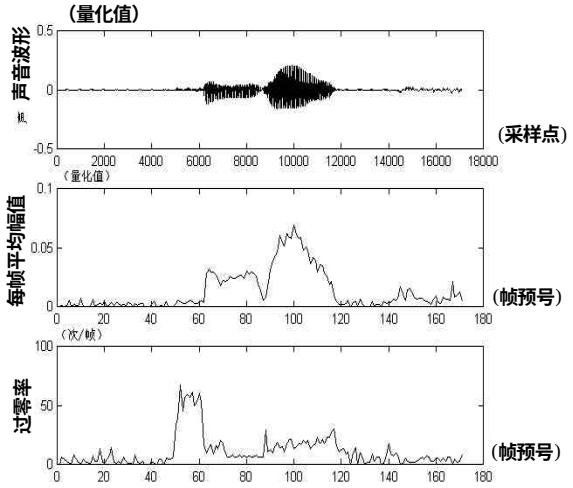

图1 语音"升高"的波形、平均幅值和过零率

2 几点改进

按文献中提出的上述方法完成识别程序后，实验中发现端点检测的效果很不理想，能够较理想地找出起止点的语音大约只有80%左右。在大量的观察、统计和实验的基础上经过分析，现提出了如下的几点改进方案。

(1) 在得到IZCT后加一个处理步骤，令

$$IZCT = \max[IZCT', 15] \quad (4)$$

式中IZCT′为式(1)输出的IZCT。这是因为：起始噪音段的 $\overline{IZC} + 2s_{IZC}$ 有可能为0或十分接近0的数值，这样根据式(1)计算得到的IZCT可能为0或1这样很小的数，此时如果后续噪音帧中过零率略有增大，则很容易使系统过度敏感从而导致起点识别不准。而根据我们的观察和统计，语音起始段如为能量很小的清辅音，其过零率数值在连续10余帧一般都会超过20，而噪音则只会在偶尔的情况下有极少数帧的过零率超过20且不连续。

(2) 将高、低能量阈值的计算改为如下形式：

$$ITL = \max[3.1 \times IMA, a] \quad (5)$$

$$ITU = 2.5 \times ITL \quad (6)$$

式(5)中的IMA是前10帧的平均幅值；a是根据大量语音的能量通过统计得出的数据，在我们的系统中所用的a大致等于语音平均能量E的1/9。E的计算式为：

$$E = \sum_{m=-\infty}^{+\infty} [x(m) \bullet w(n-m)]^2$$

因为在能量阈值ITL和ITU的计算中，式(2)所用到的I_1、I_2均由起始各帧平均幅值中最大和最小值IMX和IMN决定。经过对实际采样数据的观察可以看到，当在正常室内环境下，前10帧背景噪音平均幅值中最小的往往十分接近于0，而某些帧的波形中有时会有毛刺存在，导致该帧的平均幅值很大，这样，IMX和IMN常常可能差距很大，甚至可以达到好几个数量级，在此情况下，由于I_1很大而I_2很小，用式(2)得到的ITL将完全由IMN决定。该理论的提出者的本意应该是忽略突发毛刺的影响，但它存在一个明显缺陷就是如果10帧中只要有1帧的平均幅值很小，则其余9帧数据将变得毫无意义。此时得到的低能量阈值ITL极小，从而导致系统过度敏感，十分容易将噪声误判为语音，使系统识别率因起点识别不准而大受影响。

经过对大量语音数据的统计分析和实验，作了以上修改。这样既利用了噪声的平均能量，又可以防止得到的ITL过低导致的问题。式(5)中的系数3.1和式(6)中的系数2.5均为通过大量统计和实验得到的经验值。

(3) 在结束点的检测上，将对过零率的检测忽略掉，并将高低能量阈值合并为一个阈值，该阈值取为ITL的1.2倍。这样改动的依据是：辅音(Consonant)能量弱，过零率高，元音(Vowel)过零率低而能量高，在其它语种中，存在C-V、V-C、C-V-C等多种结构，而在汉语中的字只有C-V结构或V结构，均以元音结尾，在结尾再次进行过零率检测将毫无益处，甚至适得其反。而汉语语音信息最强的部分为前面和中间的部分，在结尾部分往往只是信息弱时间长的拖尾音，将过多的拖尾音取入识别匹配所用的特征序列对识别无甚贡献，甚至可能有害。而这些拖尾音的能量通常逐渐减弱，因此适当提高低能量阈ITL将有助于截断过多的拖尾音。

3 实验及结论

图2为改进前后的端点检测算法对图1所示语音"升高"的端点检测结果的比较。

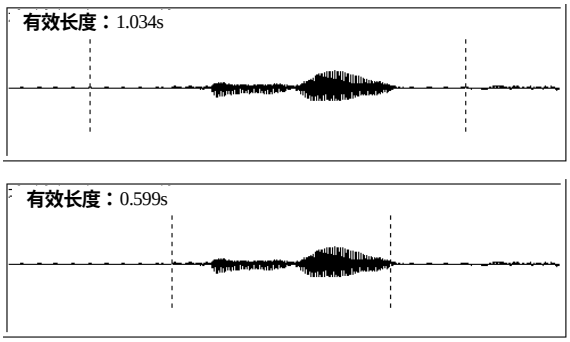

图2 算法改进前后语音"升高"的端点检测结果

图2中竖虚线内为检测到的有效语音段。从中可以清楚地看到，算法改进前检测的起点比实际起点靠前，而结束点又比实际止点靠后。这样不仅造成语音的帧数增加、计算量加大，还可能会导致识别错误。改进后的起止点识别比较准确，这样就不仅缩短了识别时间还提高了识别率。

上文提到的实验，都是基于如下实验环境及实验对象。

实验环境：PIII750，128MB内存，YAMAHA724声卡和廉价微型麦克风。计算所使用的软件为Windows 98操作系统和VC++6.0。

实验对象：14个两至三字词（打开、关闭、制冷、取暖、升温、降温、强风、弱风、温度一、温度二、温度三、温度四、温度五、温度六），由两名男性青年发音作为测试集。其中一名男青年在噪音（说话声、音乐声、较大的空调噪声混合）大、中、小情况下对各语音各发12遍，另一名男青年在噪音适中情况下对各语音发12遍，每个词的前两次发音作为训练音，后10次发音作为被识别音。被识别音按上述顺序分别构成第1、2、3、4测试集，各由140个语音构成。其中1、4测试集以各种不同规律进行发音（轻、重、缓、急、先轻后重、先重后轻、先缓后急，先急后缓），2、3测试集发音相对一致性较好。 （下转第138页）

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

7 结束语

国内电子商务应用还刚刚起步，有关概念和技术还在发展，一些法规和制度正在建立，但这不应成为观望的理由。谁先把先进的技术用于自己的商务中，谁就掌握了主动权。

该系统于2000年6月投入运行，已经稳定运行近两年的时间，其间曾根据公司的发展需要作过几次小的改进，对部分功能，如库房管理、页面美观等作了进一步完善，使其更加方便实用。由于该系统能实时提供总公司/代理商的产品销售信息，使公司高层得以及时掌握市场动态，及时调整公司产品的研发方向和生产状况，从而为公司带来了巨大的经济效益。

参考文献

1方美琪. 电子商务概论北京: 清华大学出版社, 1999

2 陈启申. 制造资源计划基础.北京: 企业管理出版社, 1997

3 Domino Design Component2.0 Tutorial. Lotus Corp, 1999

4 Lotus Domino Release 5.0: A Developer's Handbook. IBM Corp., 1999

☆☆

（上接第57页）

```
            <complexType>
              <attribute name="name" type="string">
            <complexType>
          </element>
        </sequence>
        <attribute name="name" type="string">
      <complexType>
    </element>
    <element name="link" type="semanticLink"/>
  </sequence>
 <complexType>
</element>
```

3 基于XKR的分布式知识库

3.1 XKR知识库的组织

以上我们通过XKR用一种统一的形式描述了常用的5种知识表示方法。由于不同类型的知识有了统一的形式，因此就可以把从多种不同知识源获得的知识融合在一个知识库中。用XKR所组织起来的知识库既不同于一般的文件系统；也不同于关系型数据库系统。XML天生是与分布式的Internet/Intranet融于一体的。所以基于XKR的知识库是分布式的知识库。XKR知识库由多个知识文档构成。不同的文档可能处在同一节点机上，也可能在不同的节点机上，甚至一篇知识文档的不同部分可能位于不同的节点机上。各个文档之间可以通过XLink链接在一起或者使用include/import包含在一起。对知识文档的访问则通过HTTP协议完成。

由于XKR知识库分布在Internet上，因此并不存在一个全局的管理系统（就像关系型数据库的DBMS一样）。各个节点机各自管理各自的知识文档。但是每个节点机所提供的知识文档是透明的。如果由于某种原因某个节点失效了，那么该节点上的所有知识也相应失效，所有与该节点上的知识相关联的知识也失效。但是一个节点的失效并不会导致整个知识库的崩溃，只是导致知识库的部分知识失效。

3.2 XKR知识库的特点

分布式的知识库结构十分有利于知识库扩充。一方面基于XKR的知识库能充分吸收已有的各种知识库资源。通过XKR可以直接把已有的知识形式转换为XML知识。可以简单地把两个XKR知识库链接在一起构成容量更大的知识库。这样做避免了知识库重复建设，提高知识库效能，节省开发时间；另一方面用XKR可以表述数据挖掘所得到的最终模式。这样就能够利用数据挖掘来大大提升知识系统的学习能力。

但是XKR知识库由于缺乏全局知识管理系统，因此存在很多知识冗余。对某一条具体的知识检索效率比不上关系型数据库系统。因此把XKR知识库和文本挖掘（Text Mining）、信息检索（Information Retrieval）结合起来是提高XKR知识库应用性能的一条路径。

4 结论

利用XKR可以把规则、框架、过程、表格、语义网络等等多种知识表示方法统一在一个形式描述语言之中。在此基础上我们能够融合不同类型的知识，并进一步构造出分布式知识库。XKR的表示能力比BNF更强，它不仅可以描述知识结构与属性，还可以描述知识或者属性之间的联系。XKR技术在国家“十五”计划863-11主题下“虚拟作物生长系统的可视化技术”项目中得到应用。通过实践发现基于XKR的分布式知识库便于知识融合、扩充，但是需要和文本挖掘以及信息检索结合起来提高性能。

参考文献

1 王永庆. 人工智能原理与方法. 西安：西安交通大学出版社，1998

2 何新贵. 模糊知识处理的理论与技术. 北京：国防工业出版社，1994

3 王国胤. Rough集理论与知识获取. 西安：西安交通大学出版社，2001

4 Han Jiawei, Kamber M. 数据挖掘——概念与技术. 北京：高等教育出版社，2001

5 XML1.0（第二版）建议书. http://www.w3.org/TR/REC-xml

6 XML Schema建议书. http://www.w3.org/TR/xmlschema-0, http://www.w3.org/TR/xmlschema-1，http://www.w3.org/TR/ xmlschema-2

7 XLink1.0版建议书. http://www.w3.org/TR/xlink/

☆☆

（上接第121页）

经过改进后的端点检测算法，在上述实验环境下采用相同的语音识别算法（提取端点检测后的有效语音段的LPC系数，然后用DTW（动态时间归正）法进行匹配），其识别的准确率由原来的80%左右上升至95%左右。

参考文献

1陈尚勤, 罗承烈, 杨　雪 .近代语音识别. 成都：电子科技大学出版　社, 1991

2 Rabiner L, Juang B H. Fundamentals of Speech Recognition. Prentice - Hall International, Inc,1993

3 杨行峻, 迟惠生. 语音信号数字处理. 北京: 电子工业出版社, 1995

4 易克初, 田　斌, 付　强. 语音信号处理. 北京: 国防工业出版社, 2000

5 何　强, 张歆奕, 张有为. 基于定点ＤＳＰ的实时语音命令识别模块.电子技术应用，2000,26(7):51

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net