一种噪声环境下连续语音识别的快速端点检测算法

崔冬青　李治柱　吴亚栋
(上海交通大学计算机系,上海 200030)
E-mail:cdongq@sjtu.edu.cn

摘　要　根据汉语语音的特点,该算法利用幅度及功率谱对语音端点进行检测,有效地消除了背景噪声及 DC 分量的干扰。算法采用实际语音采样进行分析,试验结果表明此算法不仅能有效地标识出语音的起始及终止点,并且还具有相当高的运算效率。

关键词　功率谱　端点检测　DC 分量

文章编号 1002-8331-(2003)23-0095-03　**文献标识码** A　**中图分类号** TP391;TP301.6

A Fast Endpoint Detection Algorithm for Continuous Speech Recognition in Noisy Background

Cui Dongqing　Li Zhizhu　Wu Yadong
(Computer Science Department,Shanghai Jiaotong University,Shanghai 200030)

Abstract: According to the character of Chinese speech,this algorithm effectively elimilates the deviation caused by background noise and DC-offset in locating the endpoints of speeches in continuous speech recognition by utilizing amplitude and power spectra.Some real audio inputs are taken as examples to evaluate the performance of this algorithm.The result shows that this Algorithm can effectively mark the endpoints of each audio input.

Keywords: Power spectra,Endpoint detection,DC-Offset

1　引言

实时检测语音的起始点与终点是连续语音识别的一个重要环节,它既可以增加语音识别的准确率,也可以节省许多额外的计算开销。目前已经有不少端点检测方法,大多数算法都是基于能量、短时过零率或 LPC[1-3],在噪声环境下它们的准确率比较低。相反采用声学参数的端点检测算法[4]虽然有比较高的准确率,但是计算复杂度相当高,并不太适合实时系统。对于以上两个问题,结合 DC 分量及背景噪声等问题,将提出一个高效、准确的算法。

2　端点检测算法

整个算法由四个部分组成;首先对输入语音信号进行 DC 分量和背景噪音的估值,为参考点判决阈值的选取做准备。接着检出语音信号大致的起止点作为实际起止点的参考点。然后利用汉语语音功率谱的特点,检测出实际的语音起止点。最后对该算法进行一定的修正。

2.1　DC 分量和背景噪声的估值

DC 分量是由硬件产生的,使样本点的均值偏离中心点,端点检测时必须除去采样语音信号的 DC 分量,通常是取所有采样数据的平均值(均值法)作为 DC 分量或者采用预加重的方法来消去 DC 分量:

$$S_n'=S_n-S_{n-1}\quad n=1,2,...N \tag{1}$$

式中 S_n 是量化的语音信号。但是这两种方法都不太适合连续语音识别(CSR),因为前者需要所有数据的平均值,而 CSR 不同于孤立词的识别,不可能事先得到所有的语音数据,而后者在有噪声的情况下会对波形造成相当大的影响,无法准确地利用能量或幅度信息标识出语音的端点。因此,下面将提出一种采用自适应算法的除去 DC 分量的方法。

首先设初始的 DC 分量的值为 0,对于每一个新输入的采样数据 S_n 而言,新的 DC 分量的值由下式所决定:

$$Offset(k)=\alpha*Offset(k-1)+(1-\alpha)*S_n \tag{2}$$

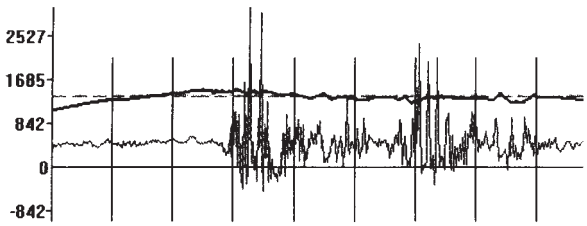
其中 α 是一个固定常数,在实验中取定它为 0.999。随着采样数据的增加,DC 分量的值将会逐渐收敛于某个特定的值。从实验结果来看,如果将 DC 分量的初始值设为第一帧的均值,可大大提高其收敛速度。同时,该方法较之于均值法所不同的是,如果在识别的过程中发生硬件或信号的故障,那么 DC 分量也会做相应的调整,而不是局限于某一固定的值。图 1 为语音'美好前程'的实验结果。其中实线为 offset 曲线,虚线为均值曲线。

对于每个采样的语音数据,将它减去 DC 分量后所得到的值就是实际的采样数据,即:

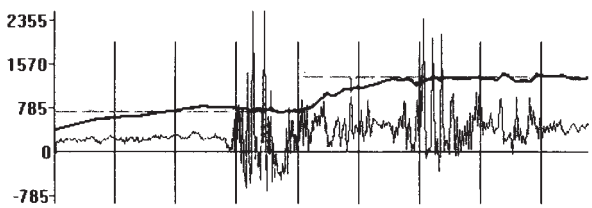
$$S_n'=S_n-Offset(k) \tag{3}$$

作者简介:崔冬青,硕士研究生,计算机应用技术专业。李治柱(1965-),教授,硕士生导师,中国计算机学会微机专业委员会委员,香港国际教育交流中心研究员,主要研究方向为中文信息处理、计算机网络、多媒体技术、语音识别和办公自动化。吴亚栋,工学博士,副教授,主要研究方向为语音信息处理及其应用。

对于减去 DC 分量后的采样语音，实验中得到结果显示它们的均值大致为 0 左右，因此这种算法能相当好地除去硬件对语音分析所造成的影响。



(a)DC 分量不变（放大 3 倍），offset[0]取第一帧的均值



(b)DC 分量发生变化（放大 3 倍），offset[0]取第一帧的均值

图 1　语音‘美好前程’的实验结果

对于输入的采样信号，取它初始阶段的 10 帧无实际语音信号的帧来估计背景噪声能量，首先利用公式(4)来计算这些帧的能量：

$$E_k=\sum_{n=(k-1)W_L+1}^{kW_L}(s_n')^2 \tag{4}$$

式中 $k=1,2...10$ 为帧的序号，W_L 是每帧的采样点数量，实验中取定它为 256（32 毫秒，8kHz）。对于计算得到的 E_k，设其最小值为 $E_{\min}$，最大值为 $E_{\max}$，然后计算这两者的平均值：

$$E_{\text{mid}}=0.5\ (E_{\max}+E_{\min}) \tag{5}$$

基于 E_{mid}，将 E_1 至 E_{10} 分为两部分，大于 E_{mid} 的分为一组，计算其平均能量 E_{big}，小于 E_{mid} 的分为一组，计算其平均能量 E_{small}。最后，噪声能量 E_N 由下面的公式所确定：

$$E_N=\begin{cases}0.5\ (E_{\text{big}}+E_{\text{small}}), & \dfrac{E_{\text{big}}}{E_{\text{small}}}\le 2\\ 0.95E_{\text{small}}+0.05E_{\text{big}}, & \dfrac{E_{\text{big}}}{E_{\text{small}}}>2\end{cases} \tag{6}$$

对于计算所得到的噪声能量 E_N，它必须处于一段阈值范围之内，否则将认为该语音采样的背景噪声过于强烈而不予接受，或者没有语音输入而终止程序。

2.2　参考端点的检测

对于除去 DC 分量后的采样信号，可以利用其波形幅度划出语音的参考起止点，对于汉语语音而言，参考起止点实际上是输入语音的第一个韵母的起点和最后一个韵母的终止点。

参考端点的检测算法是：对输入语音信号，从左到右计算每帧内样本幅值大于阈值 T_A 的样本总数，如果该数目大于阈值 V_1，则该帧即为参考起始点 t_{F3}：

$$t_{F3}=\underset{i}{\operatorname{argmin}}\ \left(\sum_{n=i-W_L+1}^{i}u\ (s_n'-T_A)>V_1\right),W_L\le i\le N \tag{7}$$

同理参考终止点 t_{B3}：

$$t_{B3}=\underset{i}{\operatorname{argmax}}\ \left(\sum_{n=i+1}^{i+W_L}u\ (s_n'-T_A)>V_2\right),0\le i\le N-W_L \tag{8}$$

其中 $u(x)$ 是个阶梯函数：

$$u(x)=\begin{cases}1, & x\ge 0\\ 0, & x<0\end{cases} \tag{9}$$

考虑到噪声同实际语音之间存在着一定的能量差距，因此如果用信噪比来表示这个差距的话，T_A 的计算公式可以由下式表示：

$$T_A=C\sqrt{\frac{E_N}{W_L}} \tag{10}$$

上式中 C 为经验常数，通过实验取定 C 为 8。对于 V，通过实验得到当 V_1 等于 3，V_2 等于 15 时能很好地对语音的端点实施检测。图 2 显示了汉语语音‘琼州海峡’在强噪音情况下通过上述算法所得到的参考起始及终止位置。

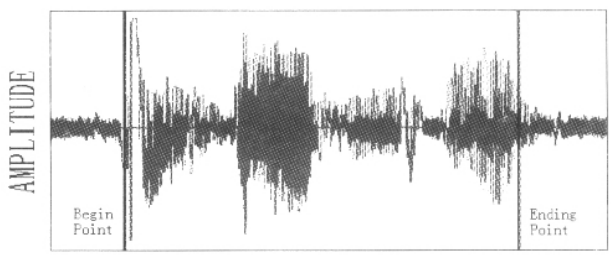


图 2　在强噪声情况下语音‘琼州海峡’的参考起始（Begin Point）及终止（Ending Point）位置

另外，考虑到在无背景噪声（S/N>60DB）的情况下，由于 E_N 的值近似于 0，因此 T_A 也近似于 0。在这种情况下，一旦在语音的尾部跟随了一个很小的随机噪声，那么由于连续幅度小于 T 的帧数小于 V_2，就会造成语音终止端点的后移。为避免这种情况，可以将(10)式改为：

$$T_A=\max\left(C\sqrt{\frac{E_N}{W_L}},D\right) \tag{11}$$

上式中的 D 为一个常数，在实验中取为 800。

对于计算所得的 t_{F3} 和 t_{B3}，必须满足一定的条件即：

$$t_{B3}-t_{F3}\ge t_{\min} \tag{12}$$

其中，$t_{\min}$=160，即对应最小的浊音的长度（20 毫秒）。否则，算法认为没有语音输入而终止。

2.3　实际端点的检测

得到了参考语音端点后，就可以通过对端点邻近帧采样数据的分析来得到实际的语音端点。下面先讨论实际的语音起始点的定位。

由于汉语中韵母的幅度和声母的幅度往往相差很多，因此在有强噪声的情况下上面的算法所标识出的起始端点往往包含了全部韵母的部分而没有包含全部的声母部分，为了得到确切的语音起始端点，必须从参考起始点 t_{F3} 向前搜索，以得到声母的起始位置。统计表明[5]，汉语声母最长可达 200ms 以上，因此，至少应从参考语音起始位置向前追溯 7 帧以上（8K 采样率，每帧 256 个采样点）就可以得到实际的语音起始点。实验中实际取为 20 帧。

由于声母的波形幅度一般都比较小，在有较强噪声的情况下，短时过零率和自相关系数会在很大程度上受噪声的影响，无法准确利用这些信息来定位具体的起始点，而采用 FFT 进行语音端点的判断虽然也要受到噪声的影响，但是通过汉语语音的频谱特性可以去除这一影响。通过对汉语语音的分析可以知道声母和韵母的大部分能量都集中在某些频率段上，因此如

果能通过对这些特定频率上的能量进行分析就可以判断该帧是否属于实际语音。

在追溯的过程中首先必须采用减谱型法[6]来大致地消除噪声：假设噪声中的语音信号是由噪声信号和语音信号线性叠加而成的，因此如果通过采样初始阶段只有噪声的那些帧计算得到噪声的平均功率谱 $P_n(\omega)$，那就可以把计算所得到的含噪声的语音信号的功率谱 $P_x(\omega)$ 减去 $P_n(\omega)$ 来得到实际的语音信号的功率谱 $P_s(\omega)$。考虑到噪声的随机性，在减去 $P_n(\omega)$ 后仍会有些较大的功率谱分量的剩余部分，因此在实际操作中的计算公式为：

$$\begin{cases} P_s(\omega)=P_x(\omega)-aP_n(\omega), & P_x(\omega)\geq aP_n(\omega) \\ P_s(\omega)=0, & P_x(\omega)<aP_n(\omega) \end{cases} \quad (3)$$

通过实验，上式中的 a 取定为 3。虽然用公式（3）求得的语音信号的功率谱仍然在某些频率范围内包含有一些噪声部分，但是由于噪声往往是随机的，因此在不同的帧内这些部分所对应的频率也是不同的。相反，韵母或声母所对应的高能量部分的频率往往是相同的，利用这个特性，就可以把语音信号（可能仍含有噪声）同噪声信号（不含语音信号）区分开来，从而找到真正的端点。由于事先并不知道频率的范围，因此必须通过对参考起始点附近的帧进行分析来得到其功率谱大于 0 的频率范围。首先对参考起始点及参考起始点前的连续 2 帧进行分析，分别得到经过减谱处理以后功率谱仍大于 0 的各自的频率范围，从而得到这三帧相应的频率范围的交集，交集数设为 n，如果在某个交集内这 3 帧的功率谱都大于 0，那么就继续往前寻找，直至到达回溯极限（10 帧），否则，停止回溯，认为已经找到实际语音的起始端点了。考虑到白噪声的能量分布往往是在整个频率范围内随机分布的，因此在相邻三帧中在同一频率范围都满足功率谱 $P_s(\omega)$ 大于 0 的概率是相当小的，所以上面的算法是比较符合实际情况的。

2.4 算法修正

但是，上面的算法中还存在如图 3 的问题：以“蛇”的拼音/she/为例，声母/sh/的能量主要集中在 3kHz 附近，而韵母/e/的能量主要集中在低频区间。由于噪声的影响，在用幅度划分语音端点的时候很有可能划分在声母韵母之间。

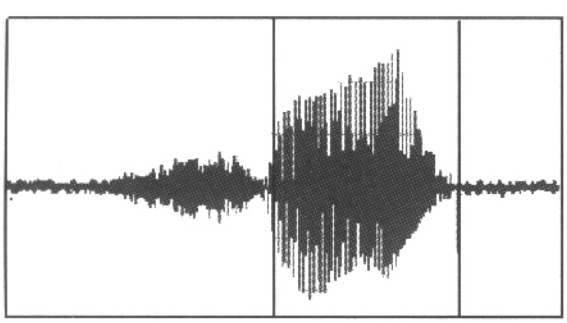


图 3 语音参考起始端点划分在韵母而非声母上

因此，通过上面的回溯算法得到的语音起始端点可能是韵母/e/的起始点而非声母/sh/的起始点。但是，通过对语谱图的分析可以知道，声母和韵母之间一般存在着能量的过渡带，受声母/sh/的影响，在靠近声母处韵母/e/在 3kHz 左右能量也比较高，这就意味着二者之间肯定存在交集，因此如果将上面的算法做如下的调整，就可以解决这个问题：如果在回溯时某一帧信号的语音功率谱 $P_s(\omega)$ 在已求得的频率范围内为 0，而如果该帧的上一帧与下一帧语音功率谱同该帧一样在另一个频率范围内（交集）大于 0，那么就用此频率范围代替原来的频率范围，并以此频率范围为基础做相同的回溯，直至到达回溯极限或找到功率谱为 0 的帧并且不能再找到新的频率范围为止。

对于语音实际终止端点的估计类似于起始点的估计，只需要向后分析一定的帧数就可以了。考虑到汉语语音是以韵母作为结束的，而韵母的幅度一般都比较大，因此向后判断的帧数取 7 左右就可以了。对于语音“琼州海峡”通过该算法修正后的端点可参见图 4。

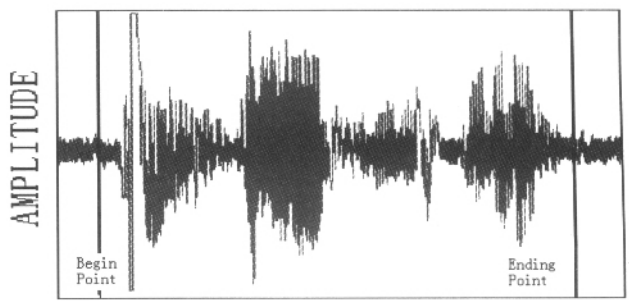


图 4 修正后的语音“琼州海峡”在背景噪声下的端点划分

同图 2 相比，图 4 在划分语音的起始点与终止点上更加精确。

3 实验结果

为了评估算法的准确性，在不同的噪声环境下实验分别对 100 个随机的语音采样进行了分析，如果以端点划分的误差≤3 帧作为正确的评估标准的话，可得到表 1 所示的结果。

表 1

测试环境	语音起始点的划分		语音终止点的划分	
	误差≤3 帧	误差>3 帧	误差≤3 帧	误差>3 帧
无噪声环境（S/N≈60DB）	100	0	100	0
低噪声环境（S/N≈40DB）	99	1	99	1
强噪声环境（S/N≈25DB）	98	2	99	1

从这些统计结果可以看到即使在很强噪声的情况下，算法也能很好地确定出语音的实际起始与终止点。

4 结论

同利用短时过零率、波形能量及自相关系数的语音端点检测算法相比，该算法受噪声的影响比较小，而且准确性也更加高。

另外，从算法的复杂性来说，该文算法的算法复杂性要高于仅利用短时过零率和波形能量的算法。但是同利用声学参数的算法相比，该文算法的开销要小很多。因此该文算法还是相当适用于实时系统的。

最后，同大多数端点检测算法所不同的是该文算法针对的是连续语音的端点检测而非孤立词的端点检测，因此在消除 DC 分量时所采用的自适应的算法还是相当必要的，同时这种自适应的算法也适用于其他实时信号处理。此外，由于连续语音识别过程中背景噪声可能不是恒定的，因此在估计噪声的能量的时候也可以采用一些自适应的方法作为改进。

（收稿日期：2002 年 7 月）

（下转 138 页）

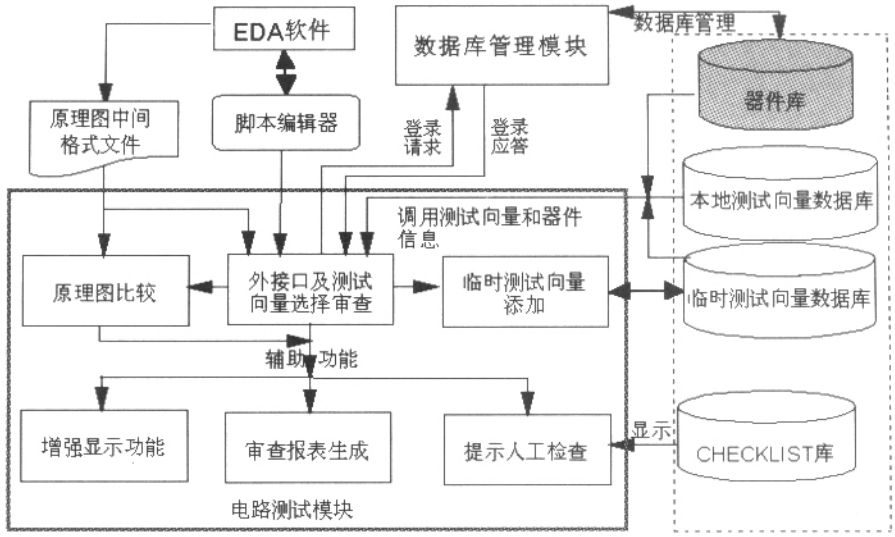


图 5　原理图级电路设计正确性自动审查模块结构框图

现原理图的自动审查，为了使用户更方便地使用原理图级电路自动审查工具，笔者在平台中添加了一些增强显示的辅助功能——显示器件信息、查找器件或网络、关联器件和串通显示。显示器件信息是对选中的器件显示器件的属性、管脚等信息。查找器件或网络是根据用户输入的器件或网络的名称（或属性），查找出器件或网络所在的原理图，并能够反标器件或网络。关联器件的功能是：选定或输入对象（器件或网络），显示关联器件。串通显示的功能是：选中或输入网络的标识，能够高亮显示该网络信号线路径，将串通电阻视为直接连通的信号线。并且在进行关联器件和串通处理时，用户可以配置不需要关联或串通的网络信息。以上几项辅助功能支持 EDA 软件设计原理图的分页分层的设计风格。

该原理图级电路设计正确性自动审查工具的应用客户端作为组件嵌入到 EDA 软件中，从而成为 EDA 软件的菜单进行调用。该平台的测试向量管理模块作为单独的组件提供。该平台采取图形化界面，为用户提供友好的使用环境。

4　原理图级电路设计正确性自动审查工具的可扩展性

目前，笔者开发的原理图级电路设计正确性自动审查工具主要实现用 EDA 软件设计的原理图级电路的自动审查。在电子电路的设计过程中，不仅有原理图的设计，还有 PCB 版图的设计，集成电路版图的设计等。笔者开发的测试平台不能只局限于原理图的自动审查，还应该可以支持其他电子电路设计的审查。笔者目前开发的原理图自动审查工具只是对 EDA 软件开发的初步尝试，对 PCB 版图、集成电路版图设计审查的支持是人们继续开发审查平台的方向。

5　结论

原理图自动审查工具，增加内置测试向量库，将初期设计工程师的设计经验集成在测试向量库中，依据测试向量实现原理图级电路设计正确性的自动测试，实现电子电路设计中设计与测试的融合，提高电子设计的正确性、可靠性，从而提高电子设计的效率。

原理图自动审查工具的开发，为测试平台对各种电子电路设计的正确性测试奠定了基础。（收稿日期：2002 年 8 月）

参考文献

1.Steve Broderick.增进自动测试设备的应用[J].电子工程专辑，1998-09

2.ANSI/EIA-548-1988.Electronic Design Interchange Format，Version 200[S]

3.Cutkosky M R.PACT：An experiment in integrating concurrent engineering systems[J].IEEE Computer，1993，26（1）：28~37

4.Knister M，Praksh A.DistEdit：A distributed toolkit for supporting multiple group editors[C].In：Proceedings of ACM CSCW'90，Los Angeles，CA，1990：343~355

5.Sessions R.Microsoft's Vision for Distributed Objects.COM and DCOM.New York：John Wiley & Sons，Inc，1998

6.Rodden T et al.Distributed system support for computer supported cooperative work[J].Computer Communication，1992，15（8）：527~537

7.潘爱民.COM 原理与应用[M].北京：清华大学出版社，1999

8.M L Maher，J H Rutherford.A model for synchronous collaborative design using CAD and database management[J].Research in Engineering Design，1997，9（1）：95~98

9.李玉山，来新泉等.电子系统及专业集成电路 CAD 技术[M].西安：西安电子科技大学出版社，1997

（止接 97 页）

参考文献

1.L R Rabiner，M R Sambur.An Algorithm for Determining the Endpoints of Isolated Utterances[J].Bell System Tech J，1975，54：297~315

2.L F Lamel，L R Rabiner et al.An Improved Endpoint for Isolated Word Recognition[J].IEEE ASSP，1981，37：777~785

3.Evangelos S Dermatas，Nicos D Fakotakis et al.Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment[J].ICASSP，1990：733~736

4.L R Rabiner，C E Schmidt et al.Evaluation of a Statistical Approach to Voiced-UnVoiced-Silence Analysis for Telephone-Quality Speech[J].Bell System Tech J，1977：455~487.

5.陈永彬，王仁华.语音信号处理[M].合肥：中国科学技术大学出版社，1990

6.J S Lim，A V Oppenheim.Enhancement and bandwidth compression of noisy speech[C].In：Proc IEEE，1979，67：1586~1604