DOI:10.13274/j.cnki.hdzj.2005.07.002
2005 年第 7 期

中图分类号:TP206.1　　文献标识码:B　　文章编号:1009-2552(2005)07-0005-04

语音信号端点检测方法与展望

杨胜跃，周宴宇，黄深喜

（中南大学信息科学与工程学院，长沙 410075）

摘　要：对语音信号端点检测的主要方法，如基于短时能量的方法、基于 HMM 的方法、基于自相关相似距离的方法等进行了深入研究。分析了这些方法的原理、优点与不足，并进行了比较。针对各种检测方法在应用中，尤其是应用于噪声环境中的局限性，指出了语音信号端点检测领域进一步研究的方向。

关键词：语音信号；端点检测；噪声

A survey of endpoint detection methods for speech signal

YANG Sheng-yue，ZHOU Yan-yu，HUANG Shen-xi

（School of Information Science & Engineering，Center South University，Changsha 410075，China）

Abstract：The main speech signal endpoint detection methods，such as short time energy based scheme，HMM based scheme，related alike scheme and so on are investigated deeply. The analysis about the principle of these methods，their advantages and disadvantages，and the comparison between these methods are given. According to their limitations，especially when applied in noisy environment，the research direction about speech signal endpoint detection is pointed out.

Key words：speech signal；endpoint detection；noise

0　引言

语音是人类最自然的信息载体，理应成为未来人机交互的主要方式。而语音端点检测是语音分析、语音合成和语音识别中的一个重要环节。在实际运用中，通常要求首先对系统的输入信号进行判断，准确地找出语音信号的起始点和终止点。这样才能采集真正的语音数据，减少数据量和运算量，并减少处理时间。

在语音识别中，通常是先根据一定的端点检测算法，对语音信号中的有声片段和无声片段进行分割，而后在针对有声片段，依据语音的某些特征进行识别。研究表明[1]，即使在安静的环境中，语音识别系统一半以上的识别错误来自端点检测器。因此，作为语音识别系统的第一步，端点检测的关键性不容忽视，尤其是噪声环境下语音的端点检测，它的准确性很大程度上直接影响着后续的工作能否有效进行。可以说，语音信号的端点检测至今天为止仍是有待进一步深入的研究课题。

1　语音端点检测主要方法

在很长一段时间里，语音端点检测算法主要是依据语音信号的时域特性[2~3]。其采用的主要参数有短时能量、短时平均过零率等，即通常说的基于能量的端点检测方法。这些算法在实验室环境下具有良好的性能，但在噪声环境下，则无法达到其应有的效果。近年来，随着通信业的迅猛发展，在各行业对通信系统语音质量的客观评价以及语音识别方法等技术实用化的强烈需求下，又出现了很多的语音端点检测算法。它们主要是通过采用各种新的特征参数，以提高算法的抗噪声性能。如基于 1994 年由 Junqua J-C 提出的 TF 参数的语音端点检测[4~5]，还有诸如倒谱系数[6~7]、短时频带方差[8]、自相关相似距离[9]、信息熵[10~11]等也逐渐的被应用到端点检测

收稿日期：2005-04-29
基金项目：国家自然科学基金资助项目(69975003)
作者简介：杨胜跃(1969-)，男，博士，主要研究方向为控制理论、信号处理等。

中。有时，还通过将信号的几种特征组合成为一个新的特征参数来进行端点检测。对语音端点的判决方式也由原来的单一门限[12]、双门限[13]发展到基于模糊理论的判决方式[14]。

1.1 基于短时能量或短时平均幅度的检测方法

语音和噪声的区别主要是在它们的能量上，语音段的能量比噪声段的大，语音段的能量是噪声段能量叠加语音声波能量之和。传统的检测方法认为，如果环境噪声和系统输入噪声比较高，以至能够保证系统的信噪比相当高，那么只要计算输入信号的短时能量或短时平均幅度就能把语音段和噪声背景区分开。但是实际应用中往往难以保证有较高的信噪比，所以仅靠短时能量或短时平均幅度来检测语音段的起止点常会遇到问题。

1.2 基于短时能量和短时平均过零率的检测方法

该方法也称为双门限比较法[13]，它是在短时能量检测方法的基础上，加上短时平均过零率，利用能量和过零率作为特征来进行检测。

对于连续语音信号，可以观察到语音的时域波形通过横轴(时间轴)的情况。在离散时间语音信号情况下，如果相邻的采样具有不同的代数符号就称为发生了过零，由此可以计算得到过零的次数。单位时间内过零的次数称为过零率，而一段长时间内的过零率称为平均过零率。然而语音信号是宽带信号，所以用平均过零率就不太确切，只能运用短时平均过零率。

研究显示[15]，清音的过零率较高，浊音和噪声次之，而且浊音和噪声的过零率相当。因此这种方法对语音信号中的浊音和噪声很难区分，因此在检测时同样会漏掉某些音素[16]。

1.3 基于自相关相似距离的语音信号端点检测方法

这种检测方法将噪声与语音信号都当成随机过程，然后通过其相似性，判断所检测到的信号是噪声或是语音信号。对于广义随机平稳过程 $s(n)$，其自相关函数定义为：

$$R(l)=E[s(n)s(n+l)]=\lim_{n\to\infty}\frac{1}{2N}\sum_{n=-N}^{N}s(n)s(n+l) \tag{1}$$

因为语音信号是非平稳过程，因此通常采用短时分析。短时的语音信号相当于一个平稳的随机过程，将短时语音信号定义为 $s(n)$，其自相关函数为：

$$R_w(l)=\frac{1}{N-l}\sum_{n=0}^{N-l-1}s(n)s(n+l) \tag{2}$$

其方差为：

$$D[R_w(l)]=\frac{1}{N-l}D[s(n)s(n+l)] \tag{3}$$

当 $s(n)$ 是零均值，且方差为 σ^2 的高斯白噪声时，易得：

$$D[R_w(l)]=\frac{1}{N-l}\sigma^4 \tag{4}$$

上式表明：当 $l\leq N$ 时，方差较小；$l\to N$ 时，方差较大。假设两个平稳随机过程 $s_0(n)$ 和 $s(n)$，其短时自相关函数分别为 $R_0(l)$ 和 $R_w(l)$，则定义：

$$\lambda=\min_{\alpha}\left|\frac{\sum_l[R_w(l)-\alpha R_0(l)]^2}{\sum_l R^2(l)}\right| \tag{5}$$

为这两个随机过程的自相关函数相似距离。由(5)式容易得：

$$\alpha=\frac{\sum_l R_w(l)R_0(l)}{\sum_l R_0^2(l)} \tag{6}$$

可以发现：当 $\lambda_{\min}=0$ 时，存在 α，对任意 l 均满足 $R_w(l)=\alpha R_0(l)$，即这两个随机过程具有相同的谱结构，认为它们相似；当 $\lambda_{\max}=1$ 时，这两个信号正好正交，也即 $\alpha=0$，认为它们之间的相似性最弱。

因为 λ 的取值客观反映了信号之间的相似距离，可以把它用于端点检测。设噪声模型的自相关函数为 $R_0(l)$。容易理解，一段实际的语音信号看成两种状态：状态“0”和状态“1”，在状态“0”时无语音信号，$y(n)=w(n)$；状态“1”时有语音信号，$y(n)=s(n)+w(n)$。其中 $w(n)$ 为噪声，$s(n)$ 为语音信号。

在状态“0”时，$R_w(l)=R_N(l)$($R_N(l)$ 是 $w(n)$ 的短时自相关函数)。如果与噪声模型是同种信号，则有 $E[R_w(l)]=\alpha R_0(l)$。此时 λ 最小，表明被测信号与噪声模型最相似。易得：

$$E[\lambda_0]\cong\frac{\sum_l E[R_N(l)-\alpha R_0(l)]^2}{E[\sum_l R_N^2(l)]}=\frac{\sum_l D[R_O(l)]}{\sum_l(\alpha^2R_0^2(l)+D[R_N(l)])} \tag{7}$$

当噪声就是白噪声时：

$$E[\lambda_0]\cong\frac{\sum_l\frac{1}{N-l}}{1+\sum_l\frac{1}{N-l}}$$

其中：$l\in[0,N/2-1]$；$E[\lambda_0]\in[1/3,1/2]$。当噪声为有色喊声时，$E[\lambda_0]$ 更小。

在状态“1”时，$R_w(l)=R_s(l)+R_N(l)$。$R_s(l)$ 由两个部分组成：$R_s'(l)$ 与 $\gamma\cdot R_0(l)$。其中 $R_s'(l)$ 与 $R_0(l)$ 正交，所以：$R_w(l)=R_s'(l)+R_N(l)+\gamma R_0(l)$。因此有：

$$E[\lambda_1] = \left| 1+\frac{\sum_l (\gamma+\alpha)^2 R_0^2(l)}{\sum_l R_s'^2(l)+\sum_l \frac{\alpha^2\sigma^4}{N-l}} \right|^{-1}$$

一般情况下，因 $\sum_l R_s'^2(l) \gg \sum_l \frac{\alpha^2\sigma^4}{N-l}$，且 $\sum_l R_s'^2(l) = \sum_l R_s^2(l) - \sum_l \gamma^2 R_0^2(l)$，所以 $E[\lambda_1]$ 主要取决于：

$$\frac{\sum_l R_N^2(l)}{\sum_l R_s^2(l)} \cdot \frac{(1+\gamma/\alpha)^2}{1-\gamma^2 \sum_l R_0^2(l) / \sum_l R_s^2(l)}$$

经观察可以发现：第一项近似于信噪比(SNR)倒数的平方；第二项则取决于 $R_s(l)$ 与 $R_0(l)$ 的相似程度 α_0，定义：

$$\alpha = \frac{\gamma^2 \sum_l R_0^2(l)}{\sum_l R_s^2(l)} \tag{8}$$

可以发现：当信噪比SNR＞6db时，第一项小于；当 $\alpha = 0.3$，$\gamma/\alpha = 1$ 时，有 $E[\lambda_1] > 0.8$。从以上分析可知，信噪比一定时，λ 主要取决于信号和噪声的相似程度。如果两者完全相似，即相似度为1，那么 $E[\lambda_1] \to E[\lambda_0]$，这时便认为被测信号与噪声模型一致而无分界点。

一般情况下，由于浊音的自相关函数呈明显的周期分布，具有较强的峰值，只要噪声信号不具备同样的周期，则相似度 $\alpha \ll 1$，相似距离 $\lambda \to 1$，则端点特征明显。但由于语音信号开始的功率较小，经实验研究[9]，此方法对清音的检测精度不高，因为清音的自相关函数类似高频信号，无明显周期，对于频谱结构相似的噪声，其相似度 α 较大，所以在信噪比较低的情况下容易造成误判。

1.4 频带方差检测法

语音和噪声的频谱差异很大，在噪声的频谱中，各频带之间变化得很平缓，这与“白噪声”的称谓相符；而语音则是有“色”的，各频带之间变化较激烈。根据这一特征，可以明显的区分语音和噪声[11]。为此，首先定义一个参数来定量地描述这种特征，这个参数称为“频带方差”。由于系统是时变的，所以实际计算的是短时频带方差，它的实质就是计算某一帧信号的各频带能量之间的方差，这种以短时频带方差作为参数检测语音段起止点的方法称为频带方差检测法，以下为具体算法：定义一个矢量

$$X = \{x(w_0), x(w_1), \cdots, x(w_n)\} \tag{9}$$

其中的分量 $x(w_n)$ 定义为 $w(n)$ 的滤波器的输出能量，它可以根据一帧信号通过一带通滤波器来计算，也可以首先计算一帧信号FFT，然后把某几个频带分量组合而得。对数字信号，最低频是0，最高频是 n，其余各中心频率按一定规则从0至 n 递增。定义均值：

$$E = \frac{1}{n}\sum_{i=0}^{1} x(w_i) \tag{10}$$

则频带方差：

$$D = \frac{1}{n}\sum_{i=0}^{n} [x(w_i) - E]^2 \tag{11}$$

设检测门限值为 M，在实际应用中，其取值可以根据实际环境噪声特性来确定，一般取 $M = (3 \sim 5)D_r$（D_r 为背景噪声的频带方差值）。但在实际运用中，会遇到一些脉冲干扰，在这些区域短时频带方差也可能较大，门限值就难以确定。

1.5 其他方法

除了以上几种方法之外，还有短时分形维数的带噪声语音信号端点检测方法[17]；应用倒谱系数作为判决特征的带噪语音端点检测方法[18]，它包括应用倒谱距离测量轨迹和应用循环神经网络的方法。实验发现[18]，倒谱特征参数的语音信号端点检测方法在噪声环境下具有传统的能量方法无法比拟的优越性。基于HMM模型的检测方法[19]也是语音信号端点检测中的重要方法，该方法先用训练的方法生成背景噪声和废料的模型参数，再用Viterbi解码算法对待测信号进行分解，求出语音的哪些语音帧与背景噪声相匹配，哪些与废料相匹配，从而得出端点所在处。实验表明[19]，这种方法的准确率明显高于基于能量的方法。但是HMM的训练环境通常与实际被测信号的语音环境会有很大的差异，即背景噪声模型与实际情况不符合，此时性能会显著下降。因此，必须采用能自适应调节的背景噪声模型[18]，具体实现方法还在研究中。另外，还有采用多层感知机MLP网络实现语音信号端点检测的方法[20]、采用自适应线形神经元网络(ADALINE)的端点检测方法[21]等。

2 各类方法比较

随着越来越多的学者对语音端点检测技术的关注，大量的新的语音端点检测算法相继被提出。通过大量的文献调研与实际研究发现，现有的各种语音信号端点检测技术都存在各自的不足，比如基于自相关相似距离的语音信号端点检测方法，总的来说它与HMM方法的效果大致相同，但是对于结尾的判断却优于HMM模型，这是因为语音大多以浊音结尾，此时自相关法的判断精度较高，但是对于清音开头的语音，尤其是[s]、[ks]、[n]等音节，自相关

算法的检测精度就不高。主要几类方法各自的优点与不足列于表 1。

表 1　各类方法优缺点比较

方　法	优　点	缺　点
短时平均过零率法	较简单	难以识别弱爆破音、摩擦音、末尾的鼻音拖长的元音等
短时能量法或平均幅度法	较简单	弱摩擦音与结尾时的鼻音易和噪声混淆
HMM 法	较准确	需要事先训练
双门限比较法	有效区分语音信号中的浊音和噪声	难以区分浊音和噪声
基于自相关相似距离的方法	对浊音的检测精度较高	对开端的清音检测精度不够
频带方差法	较准确	在脉冲干扰下门限值需要测定

3　展望

提高语音端点检测的精度关键是要提高在噪声环境下语音的端点检测能力，目前已有的各种方法均有其局限性，语音端点检测还有待于进一步深入的研究课题。要做好这方面的工作，可以从两个方面入手：(1)进一步埋头于基础研究，寻找新的特征参数，能够将所有(或更多、更广泛类型)的语音信息与噪声信息很好地区分开来，这方面的工作具有相当的难度与挑战性，不过一旦取得突破，其意义不可估量；(2)各种现有方法的综合运用。如基于自相关相似距离的端点检测法，因为它对清音的检测精确度不高，将它与能量法相结合(对清音检测精确度较高)或许能够在一定程度上改善语音端点的检测精确度。第二类研究实质是基于多种语音信号特征的端点检测方法。

参 考 文 献：

[1] Junqua J C. Robustuess and Cooperative Multimodel Man－machine Communication Applications[M]. Proc. Second Venaco Workshop and ESCA ETRW. 1991.9.

[2] 朱学芳，徐建平. 计算机语音信号处理与语音识别系统[J]. 南京邮电学院学报，1998，18(5－6)：113－119.

[3] He Suning, Yu Juebang. A Novel Chinese Continuous Speech Endpoint Detection Method Based on Time Domain Features of the Word Structure[J]. IEEE Int. Conf. on Commun. Circuits and Systems and West Sino Expositions, 2002.992－996.

[4] 杨崇林，李雪耀，孙羽. 强噪声背景下汉语语音端点检测和音节分割[J]. 哈尔滨工程大学学报，1997，18(5)：28－32.

[5] 李雪耀，林娟，杨崇林. 舰船指挥舱室强噪声环境下语音识别[J]. 船舶工程，1999，(2)：50－53.

[6] 韦晓东，等. 应用倒谱特征的带噪语音端点检测方法[J]. 上海交通大学学报，2000.34(2)：185－188.

[7] 胡光锐，韦晓东. 基于倒谱特征的带噪语音端点检测[J]. 电子学报，2000.28(10)：95－97.

[8] 李祖鹏，姚佩阳. 一种语音段起止端点检测新方法[J]. 电讯技术，2000，(3)：68－70.

[9] 陈斐利，朱杰. 一种新的基于自相关相似距离的语音信号端点检测方法[J]. 上海交通大学学报，1999，33(9)：1097－1099.

[10] 李四根，和应民. 一种基于信息熵的语音端点检测方法[J]. 应用科技，2001，28(3)：13－14.

[11] Abdallah I, Montresor S, Baudry M. Robust Speech/non－speech Detection in Adverse Conditions Using an Entropy Based Estimator [C]. In: International Conference on Digital Signal Processing 1997.757－760.

[12] 何方，朱杰，郁桦，等. 一种语音信号端点检测方法及其在 DSP 上的实现[J]. 微型电脑应用，2002，18(5)：48－50.

[13] 高瑞华，朱君波，王守觉. 一种噪声环境下连续语音识别的快速端点检测算法[J]. 计算技术与自动化，2003.(23)：95－97.

[14] 朱民雄，等. 计算机语音技术[M]. 北京：北京航空航天大学出版社，2002.

[15] Wu Yadong, Li Yan. Robust Speech/non－speech Detection in Adverse Conditions Using the Fuzzy Polarity Correlation Method[J]. IEEE Int. Conf. on Systems, Man and Cybernetics, Nashville, TN, USA, 2000.2935－2939.

[16] Seneff S. Real－time Harmonic Pitch Detector[J]. IEEE Transaction on Acoustics, Speech and Signal Processing, 1978, 26(4)：358－365.

[17] 沈亚强. 低信噪比语音信号端点检测和自适应滤波[J]. 电子测量和仪器学报，2002.

[18] 易克初，田斌，付强. 语音信号处理[M]. 北京：国防工业出版社，2000.

[19] 朱杰，韦晓东. 噪声环境中基于 HMM 模型的语音信号端点检测方法[J]. 上海交通大学学报，1998，32(10)：14－16.

[20] Aini Hussain, Salina Abdul Samad, Liew Ban Fah. Endpoint Detection of Speech Signal Using Neural Network[R]. TENCON 2000. Proceedings, Malaysia, 2000.271－274.

[21] 胡瑞敏，薛东辉，姚天任. 神经网络方法及其在语音识别中的应用[J]. 高技术通讯，1995，(6)：11－15.

责任编辑：杨立民

信息天地

何谓信用经济

信用经济是指信用在生产、分配、交换、消费的四个环节中都发挥着重要的纽带作用，并且信用体系相当完善。在信用经济阶段，经济的发展更多信赖于信用的竞争，信用活动已成为经济各环节中不可或缺的纽带，并进一步成为市场经济健康发展的前提条件及原动力。在信用经济阶段各种经济主体包括政府、企业、个人等都离不开信用活动，信用活动已经成为最基本、最普遍的经济活动。