噪声评估在端点检测中的应用

张亚歌,张太镒,夏 川
(西安交通大学 电子与信息工程学院,陕西 西安 710049)

摘 要:端点检测是语音识别中非常重要的部分,其准确性直接影响语音识别系统的识别率。传统端点检测方法预设经验门限对语音的短时特征进行判决,因为预设门限难以适应不同环境,其准确度和噪声鲁棒性较差。为了改善上述缺点,提出噪声评估的概念,对环境噪声的短时能量与短时过零率等短时特征进行分析,得到了更能表征环境噪声的门限。噪声评估结合传统的双门限法用于端点检测过程,解决了经验门限对不同环境适应性不强的问题。实验表明,噪声评估增加了端点检测的准确度和噪声鲁棒性。

关键词:端点检测;噪声评估;短时特征
中图分类号:TP391.42 **文献标识码**:A **文章编号**:1673-629X(2010)09-0177-04

Application of Noise Evaluation in Endpoint Detection

ZHANG Ya-ge, ZHANG Tai-yi, XIA Chuan
(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Endpoint detection is an important part in speech recognition and its accuracy affects the recognition rate of the whole recognition system directly. The traditional endpoint detection method uses empirical thresholds which are preset to judge the short-time characteristics of speech, preset thresholds are not accurate and noisy robust because they couldn't adapt to the environment. Therefore the notion of noise evaluation is proposed in this paper, evaluate the short-time characteristics of environmental noise such as short-time energy and short-time zero-cross-rate to get the thresholds that can better describe the environmental noise. Combining the noise evaluation with traditional two-threshold method in endpoint detection, the inelasticity of empirical thresholds is improved. Experimental results show that noise evaluation increases the accuracy and noisy robust of endpoint detection

Key words: endpoint detection; noise evaluation; short-time characteristics

0 引 言

在语音识别技术中,端点检测属于前端处理部分,它是指从含有语音和噪声的声音信号中把语音和非语音信号时段区分开来的技术,后续处理就可以只对语音信号段进行处理。端点检测的精确度作为影响识别性能最大的因素在语音识别技术中有着重要的地位。有研究表明,即使在安静环境下,语音识别系统一半以上的识别错误来自于端点检测部分[1]。端点检测通常依据的语音特征有短时能量和过零率[2]、LPC参数[3,4]、频谱熵[5~7]、倒谱特征[8]、TF参数[1]、分形特征[9]以及几种参数相结合[8,10~12]。由于频域参数的计算量太大,通常使用时域参数短时能量和过零率进行端点检测。

基于短时能量和过零率的双门限法是普遍使用的一种端点检测方法[13],通过预定经验门限对短时能量和过零率进行划分以检测出语音部分的端点。然而其准确度和噪声鲁棒性并不理想。文中在考虑实际系统的情况下,提出噪声评估的概念,通过事先对环境噪声进行分析得到合适的特征判决门限。大量实验表明,噪声评估提高了端点检测的准确度和噪声鲁棒性。

1 传统端点检测算法

1.1 短时能量

短时能量是语音信号经过分帧、加窗等前端处理后帧内信号的能量之和。它是描述语音信号幅度的时域特征量。设 $S(n)$ 为一帧语音信号,帧长为 N,则该帧的短时能量为

$$E = \sum_{n=0}^{N-1} S^2(n) \tag{1}$$

语音信号中浊音的短时能量较大,清音和噪声部分的短时能量相对较小。故使用短时能量可以有效地

收稿日期:2010-01-07;**修回日期**:2010-04-03
作者简介:张亚歌(1984-),男,河南郑州人,硕士研究生,研究方向为语音信号处理、语音识别;张太镒,博士生导师,研究方向为新一代移动通信技术、软件无线电,数字音视频技术。

区分出浊音段。

1.2 短时过零率

短时过零率是语音信号经过前端处理后帧内信号通过零值的次数，它简单地表示了语音信号的频域特性。语音帧 $S(n)$ 的短时过零率定义为：

$$Z = \sum_{n=1}^{N-1} \frac{1}{2} \mid \mathrm{sgn}(S(n)) - \mathrm{sgn}(S(n-1)) \mid \quad (2)$$

式中 $\mathrm{sgn}(S(n)) = \begin{cases} 1, & S(n) \geqslant 0 \\ -1, & S(n) < 0 \end{cases}$ 为符号函数。

清音和噪声部分比浊音有更大的过零率，因此短时过零率对于检测清音有重要作用。为了区分噪声和清音，通常将过零率改变为过 m 值率：

$$Z_m = \sum_{n=1}^{N-1} \frac{1}{2} \mid \mathrm{sgn}(S(n) - m) - \mathrm{sgn}(S(n-1) - m) \mid \quad (3)$$

式中 m 为预设固定数值，当噪声信号不能通过 m 值而清音信号能通过时就能较好地区分开清音和噪声。

1.3 双门限端点检测

传统的端点检测采用双门限法对短时能量和过零率进行判决以得到语音部分。语音信号由清音、浊音和噪声三部分组成，为短时能量确定高低两个门限，短时能量大于高门限的帧可以确定为语音帧；处于高低门限之间的可能是语音信号也可能是由噪声引起，用过零率进行辅助判定；短时能量低于低门限且过零率也很低的划归为噪声。对一段被分为 N 帧的语音信号进行端点检测的流程如下：

Step1：计算每帧的短时能量 $E(n)$ 和过零率 $Z(N)$，$1 \leqslant n \leqslant N$。

Step2：确定短时能量的高低阈值 E_H、E_L 和过零率阈值 Z_{th}。

Step3：寻找高短时能量段。段起点为第 H_1 帧，H_1 满足

$$H_1 = \arg\min_n(E(n) > E_H), 1 \leqslant n \leqslant N$$

段终点为第 H_2 帧，H_2 满足

$$H_2 = \arg\max_n(E(n) > E_H), 1 \leqslant n \leqslant N$$

式中 $\arg\min_n(E(n) > E_H)$ 表示满足条件 $E(n) > E_H$ 的最小 n 值。

Step4：根据低能量门限 E_L 和过零率阈值 Z_{th} 确定整个语音段，起点为第 L_1 帧，L_1 满足

$$L_1 = \arg\min_n(E(n) > E_L \text{ 或 } Z(n) > Z_{th})$$

终点为第 L_1 帧，L_2 满足

$$L_2 = \arg\max_n(E(n) > E_L \text{ 或 } Z(n) > Z_{th})$$

Step5：段长检验。如果语音段 $L_2 - L_1$ 长度合适则认为成功的检测出语音段，将语音段交付下一环节，否则若 $L_2 - L_1$ 过大或过小，则认为检测失败。

1.4 门限设置

端点检测中门限一般根据经验设置，通常有以下几种设置或者其组合：

1：最大短时能量乘以一个系数，在存在突发性大能量噪声时不可靠。

2：最小短时能量乘以一个系数，在噪声较大时变得不可靠。

3：短时能量的中值乘以一个系数，在噪声帧或语音帧占半数以上的情况下不可靠。

门限的设置是端点检测的一个难点，为了克服传统设置门限方法的缺点，提出了噪声评估的概念。

2 噪声评估

所谓噪声评估，是在进行语音识别前对系统所在环境噪声状况进行估计。在环境噪声基本平稳的条件下，通过对环境噪声进行录音并分析表征环境噪声的各项参数以指导端点检测过程。文中对环境噪声的短时能量、短时过零率和信号幅度进行了分析，得到了短时能量、过零率的自适应门限。

2.1 短时能量

一段实验室噪声的短时能量、短时能量分布及对数能量分布柱形图如图 1 所示。

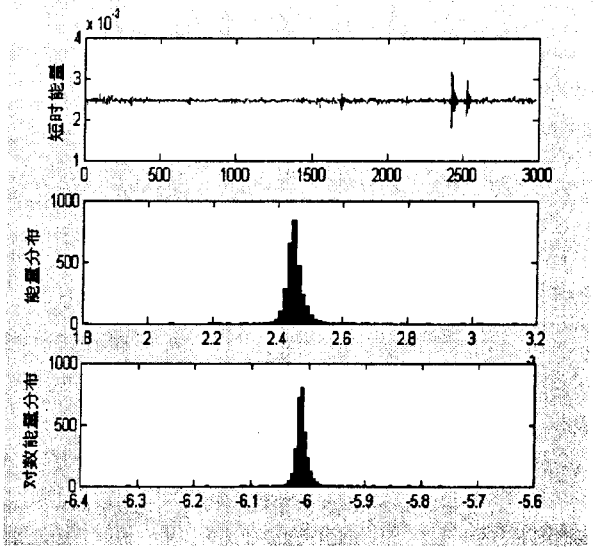

图 1 短时能量、能量分布、对数能量分布图

从图中可以看出，短时能量除了偶然的能量变化外大部分都集中在一个特定值附近，方差处于一个很低的水平。因为短时能量总是大于零的，其取值范围随环境信噪比的变化而变化。可以认为短时能量大致符合对数正态分布，即短时能量的对数符合正态分布。

根据正态分布的性质，变量 X 服从均值为 μ 方差为 σ^2 的正态分布，记作 $X \sim N(\mu, \sigma^2)$，则有：

$$P(x < \mu + 6\sigma) > 99.99\% \quad (4)$$

$$P(x < \mu + 3\sigma) > 99.87\% \quad (5)$$

即随机变量 X 99.87% 以上的取值不会超过其均值以上 3σ 水平,99.99% 以上的取值不超过均值以上 6σ 水平。该结论广泛地运用于工业控制领域。

充足的短时能量数据使得我们可以用样本均值、样本标准差描述其分布,由此设置短时能量高门限和低门限:

$$E_H = \text{mean}(E) + 6 * \text{std}(E) \quad (6)$$

$$E_L = \text{mean}(E) + 3 * \text{std}(E) \quad (7)$$

根据上述正态分布性质,这两个门限很好地给出了噪声短时能量的取值上限,采用这两个数值作为端点检测的依据可以最大限度地保证端点检测的准确性,实验结果也很好地证明了这一结论。

2.2 短时过零率

为了较大程度地除去噪声干扰,根据式(3)进行过 m 值率计算。选取 $m = \text{mean}(s) + 3 * \text{std}(s)$ (8)

一段实验室噪声的短时过零率及分布如图 2 所示。

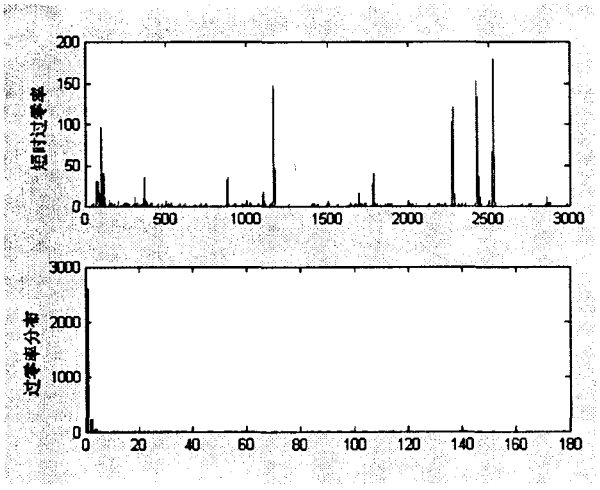

图 2　短时过零率及其分布图

从图中看出,过 m 值率集中在零值附近,零星出现的高值对整个系统影响不大,采用门限

$$Z_{th} = \text{mean}(\text{zcr}) + 3 * \text{std}(\text{zcr}) \quad (9)$$

大量实验证明,该门限在端点检测过程中取得了很好的效果。

2.3 评估时间

提出了噪声评估的概念之后,对环境噪声进行多长时间的录音才能有效地得到端点检测过程中所需的门限,这就引出了评估时间的概念。语音识别系统的实用性通常不容许评估时间过长,而太短的评估时间又无法保证门限的可靠性。经过多次实验,不同评估时间下得到的门限参数如表 1 所示。

表 1　不同评估时长下得到的门限

评估时长	高能量门限	低能量门限	过零率门限
5s	0.0031	0.0028	5
10s	0.0030	0.0028	7
30s	0.0028	0.0027	6
60s	0.0028	0.0027	8
100s	0.0028	0.0027	9

从表中可以看出,最短 30s 的评估时长已经可以得到相对稳定的参数,为了获得更稳定可靠的参数,本实验中取评估时长 60s。在系统启动或环境改变时进行长为 60s 的录音,既得到了噪声参数又不影响识别过程。

2.4 噪声评估端点检测流程

含噪声评估的端点检测流程为:

Step1:对环境噪声进行录音。

Step2:对噪声进行分帧加窗等前端处理,并计算短时能量和短时过零率。

Step3:利用式(6)、(7)、(9)得到短时能量的高低门限和过零率门限。

Step4:利用评估过程得到的门限进行传统的双门限端点检测。

3 实验结果

实验在不同信噪比的噪声环境下进行,将传统端点检测与噪声评估端点检测进行对比。实验中采用普通实验室噪声,对汉字数字 0～9 的共 500 组发音在 10～45dB 的信噪比下进行端点检测,采样率为 22.05kHz,帧长 20ms,帧间重叠 10ms。手工标定语音段的起点和终点作为准确端点值 $d_{\text{准确值}}$,定义检测结果 $d_{\text{检测值}}$ 偏离准确端点的绝对值 $D = |d_{\text{检测值}} - d_{\text{准确值}}|$ 为检测误差,以检测误差作为衡量检测结果优劣的参数。经过实验对误差 D 进行统计,传统方法误差均值 $\text{mean}(D) = 2445$,标准差 $\text{std}(D) = 1142$,噪声评估法误差均值 $\text{mean}(D) = 1749$,标准差 $\text{std}(D) = 891$。可以看出,噪声评估法的检测误差更小且更集中,得到的结果更接近准确端点。

实验结果表明,噪声评估端点检测在检测的准确度和噪声鲁棒性方面优于传统的端点检测方法,以存在清音、浊音、鼻音和爆破音,具有代表性的汉字“三”和“八”为例,首先对发音进行手工标定得到准确端点,然后在不同信噪比下用两种方法进行检测并计算检测误差,图 3 是在信噪比为 45dB 时对汉字“三”“八”的手工标定,“三”得到的起点 3250,终点 14100;“八”的起点 7150,终点 14200。

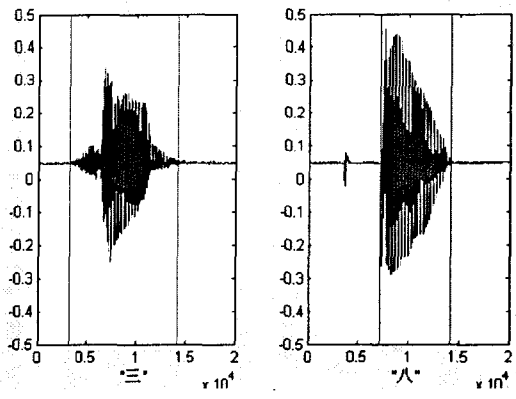

图 3 汉字"三"和"八"手工标定端点示意图

表 2 和表 3 分别是对汉字发音"三"和"八"在信噪比为 45dB 到 10dB 时两种检测方法的检测误差。因为汉字"三"以清音开始,鼻音结束,两者检测难度较大,而汉字"八"以爆破音开始,浊音结束,两者因为短时能量较高而容易检测,故表 2 中检测误差均值和标准差均大于表 3。

表 2 汉字发音"三"不同信噪比下检测误差

信噪比	传统方法		噪声评估	
	起点误差	终点误差	起点误差	终点误差
45dB	65	1503	65	177
40dB	286	1945	65	177
35dB	286	2608	286	1061
30dB	507	2829	286	1945
25dB	728	2829	507	2608
20dB	949	3050	728	2829
15dB	949	3050	728	2829
10dB	1170	3050	949	3050
	均值:1612		标准差:1133	
	均值:1143		标准差:1112	

表 3 汉字发音"八"不同信噪比下检测误差

信噪比	传统方法		噪声评估	
	起点误差	终点误差	起点误差	终点误差
45dB	133	56	88	56
40dB	133	56	133	56
35dB	133	277	133	277
30dB	133	498	133	498
25dB	133	719	133	498
20dB	133	1161	133	719
15dB	68	1603	133	940
10dB	检测不出	检测不出	133	1603
	均值:374		标准差:473	
	均值:280		标准差:276	

从表 2 和表 3 中可以看出,随着信噪比的降低,检测误差变大,与实际情况相符,即噪声越大,越难准确地检测出语音端点,但噪声评估法始终比传统方法更接近准确端点。实验中多次发现,在低信噪比的情况下,传统方法甚至无法检测出语音段。这也验证了结论:噪声评估端点检测法与传统方法相比在检测效果上有着更好的表现。

4 结束语

提出了噪声评估的概念,并对噪声的短时能量和短时过零率进行统计分析,得到了自适应的门限,并将其应用于端点检测过程。实验表明,噪声评估端点检测与传统的检测方法相比,有着更好的准确度和噪声鲁棒性。文中提出的噪声评估,可以应用于实际语音识别系统的端点检测过程,对于不同环境下语音识别系统的构建,有着重要的实用价值。

参考文献:

[1] 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):91-95.

[2] 江官星,王建英.一种改进的检测语音端点的方法[J].微计算机信息,2006,22(5):138-139.

[3] 王亚涛,朴春俊,权花紫.强噪音情况下的多种端点检测方法研究[J].信息技术,2005(2):34-36.

[4] Rabiner L R,Sambur M R. Voiced-unvoiced-silence detection using the Itakura LPC distance measure[C]//In Proc. ICASSP.[s.l.]:[s.n.],1977:323-326.

[5] 韩 韬,王 玲,刘 辉.一种应用于语音识别的端点检测改进方法[J].微电子学与计算机,2008,25(5):146-149.

[6] 白顺先.基于信息熵的语音端点检测方法的研究[J].微计算机信息,2009,25(11):196-198.

[7] 郭丽惠,何 昕,张亚昕,等.基于顺序统计滤波的实时语音端点检测算法[J].自动化学报,2008,34(4):419-425.

[8] 李洪波,于洪志.噪声环境下语音识别的端点检测技术[J].西北民族大学学报,2007,28(1):44-47.

[9] 李 凯,徐强樯,左万利.基于分形特征变化的语音端点检测技术研究[J].小型微型计算机系统,2007,28(8):1523-1526.

[10] Gu Lingyun,Zahorian S A. A New Robust Algorithm for Isolated Endpoint Detection[C]//In Proc. IEEE ICASSP.[s.l.]:[s.n.],2002:4161-4164.

[11] 刘华平,李 昕,徐柏龄,等.语音信号端点检测方法综述及展望[J].计算机应用研究,2008,25(8):2278-2282.

[12] Wu G D,Lin C T. Word boundary detection with mel-scale frequency bank in noise environment[J]. IEEE Trans. Speech Aduio Process,2000,8(3),541-554.

[13] Rabiner L,Juang B H. Fundamentals of Speech Recognition [M].[s.l.]:Prentice Hall,1993.