一种噪声环境下的实时语音端点检测算法

徐大为　吴　边　赵建伟　刘重庆
（上海交通大学图像处理与模式识别研究所，上海 200030）
E-mail：xu_dawei@sina.com

摘　要　语音识别中的端点检测要求对噪声有很强的鲁棒性。该文提出一种方法，综合采用了语音信号中的4个相互之间独立性强的特征-短时能量、倒谱距离、能量谱方差和能量-熵特征，有效地改进传统的基于单一语音特征方法的缺陷，在动态变化的噪声环境中，大大提高了端点检测对噪声的鲁棒性；为了克服分类回归树（CART）决策法的过度复杂性，引入一种新的5状态自动机进行快速决策，以保证算法的实时性能，并且能够提高端点检测的可靠性。通过各种实际噪声环境的测试，实验表明这一算法可以显著提高在低信噪比、噪声动态变化的各种环境下的端点检测性能。

关键词　端点检测　倒谱距离　能量-熵特征　5状态自动机

文章编号 1002-8331-(2003)01-0115-03　**文献标识码** A　**中图分类号** TP301.6；TP391.42

A Robust Algorithm for Real-time Endpoint Detection in Noisy Environments

Xu Dawei　Wu Bian　Zhao Jianwei　Liu Chongqing
(Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030)

Abstract: In speech recognition, the endpoint detection must be robust to noise. A method is presented in this paper, four speech features, e.g. short-time energy, cepstral distance, energy variance and energy-entropy, are taken into consideration. Because of the high independence of those four features, this method can adapt to various environments. The described algorithm not only uses four features but also introduces a 5-states automation decision logic to increase the robustness in both low SNR and various noisy environments. The performance of the algorithm is evaluated by experiments in various noisy environments, and the performance of this endpoint detection algorithm is greatly improved. At the same time, the proposed algorithm has a low complexity and is very suitable for real time mobile conditions.

Keywords: Endpoint detection, Cepstral distance, Energy-entropy feature, 5-states automation

1　简介

端点作为语音分割的重要特征，在很大程度上影响语音识别的性能，因此在有背景噪声的环境下，自动语音识别系统（ASR）需要对端点进行精确的检测。如何在噪声环境下设计一种鲁棒的端点检测算法还是一个非常棘手的问题。Savioji[2]认为，一种理想的端点检测算法应当具有以下几个特征：可靠性、鲁棒性、精确性、自适应性、简单性、实时性和对噪声特征无需先验知识。在所有的这些特征中，鲁棒性是最难达到的要求。

传统的端点检测算法通常只依赖于一个特征，例如信号能量，过零率，持续时间以及线性预测能量误差。尽管这些方法通过获取语音信号的一维特征可以降低算法的复杂度，但是却对各种噪声失去了抵制力。通常，基于能量检测方法认为加性噪声是最常见的噪声分布，而实际情况却并非如此，所以只利用一个语音特征很难处理各种各样的噪声情况。Shin[4]提出利用多个特征进行端点检测，并采用“分类和回归树”（CART）来综合各个特征进行决策。但是CART给算法带来了成指数形式增长的计算复杂度，根本无法实时实现，而且这种方法采用特征的都是由短时能量推导出的特征，所以特征之间存在较大的冗余性，在某些噪声条件下也不能取得好的效果。

该文提出的方法综合运用语音信号的4个独立性强的特征，目的在于利用不同特征对各种噪声的适应特性，提高在各种环境下对噪声的鲁棒性。算法中设计了一种5状态的自动机推理决策，使算法复杂度大大下降，适合实时计算。为了改进对各种噪声的鲁棒性，采用了语音的4个特征：短时能量、能量谱方差、倒谱距离和熵。对语音特征的选择原则是：被选择的特征应当可以从不同的方面反映噪声和语音信号之间的差别。尽管某个特征可以在某些特定环境下可以作为端点检测最有效的手段，但是这个特征并不总是能够保证在各种环境下有效。

分析所选择的这4个语音特征，可以得到以下简单的结论。短时能量是最有效的端点检测手段，被广泛采用；能量谱的方差反映了噪声信号和语音信号之间的能量谱的差别；倒谱距离是能量谱的傅立叶变换系数，是一种较为理想的分类特征；熵是从信息论中引用的一个概念，表示信息的有序程度，对于噪声而言，其有序程度要远低于语音信号的有序程度。这些特征相互之间的冗余度极大地降低，在不同的噪声环境中体现出各自的优势；端点检测通过这4种特征适当地综合加权得到。最后，采用一种5状态的自动机进行推理决策最终的端点检测结果。

2　四特征端点检测算法

基金项目：国家863计划资助项目（编号：1863-306-ZD13-05-61）
作者简介：徐大为，博士生，从事语音与视频信号处理与计算机视觉研究。吴边，博士生。赵建伟，博士生。刘重庆，博导。

为了提高端点检测的鲁棒性,选择了 4 个语音特征。这些特征为 :短时能量,倒谱距离,能量谱方差,以及能量-熵特征[1]。首先检测所有基于这 4 个特征的端点,然后采用一种 5 状态的自动逻辑推理机[3]作为最终的可靠性决策机制。当数字语音信号输入之后,倒谱系数,短时能量的均值以及熵作为环境参数事先计算出来,作为以后要进一步计算 4 个特征的参数,输入信号的开始几帧通常被认为是噪声。

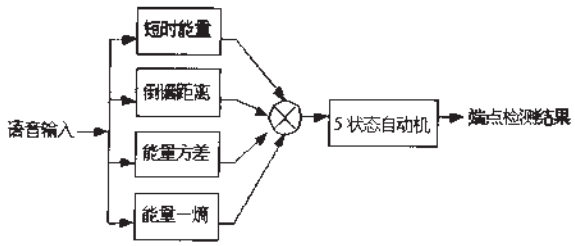


图 1　四特征端点检测算法示意

对采样率为 8KHz 的语音信号在 10ms(80 个采样点)的时间间隔内,完全可以认为是稳定的。各帧的窗口是互相重叠的,这样帧与帧之间特征就具有平滑的过渡。对于每一帧,要提取的特征是 :每一帧的短时能量,能量谱的方差,该帧与噪声帧的倒谱距离,以及能量-熵特征。如果当前帧和相邻帧的这 4 个特征的 2 个以上特征发生剧烈的变化时,通过设置的这 4 个特征的阈值进行判断。给定每一个特征的加权系数并计算最终的阈值,由此就可以确定当前观测值是否为可能的语音帧。然后采用一个 5 状态的自动推理机,推理机的状态转移是由各特征的阈值决定,输入语音信号的开始和结束最终是由状态推理机来确定。

2.1　选择端点检测的 4 个特征

2.1.1　特征分析

为了改进端点检测的性能,人们对各种噪声环境下的语音信号特征提取进行了非常多的研究,短时能量在端点检测中应用最为广泛。短时能量的计算相对于提取其他特征而言,要简单快捷得多。语音信号叠加了噪声之后的能量要比纯粹的噪声信号的能量更大,因此语音信号的能量轮廓包络线是表征采样信号中出现语音信号的最显著的特征。但是,在低信噪比条件下,比如环境中有开、关门的声音,咳嗽声,机器震动声,短时能量特征就无法将期望的语音与背景噪声区分开来。要达到对噪声的鲁棒性,短时能量特征需要和其它特征相互结合使用。用短时能量法对信噪比低的语音,图 2(b),很难检测出某些部分的端点,图 2(c)。

通过观察包含有语音信号的帧,可以看出,在不同的频带之间存在很大的差别。由于存在共振峰,该帧的能量将集中到某个频段上,因此在白噪声的环境下,能量的方差特征可以发挥一定作用,图 2(d),但是检测的效果还不能令人满意。

观测值与最小噪声帧之间的倒谱距离[5]是另外一个可以从每一帧中提取的特征。这个特征在低信噪比条件下,具有比短时能量特征更好的鲁棒性。图 2(e)显示(倒谱图形显示经过最大值归一化处理),在低信噪比环境下,采用倒谱距离可以找出信号与噪声之间的边界, 但是在某些部分端点检测不是很准确 :相比之下,基于短时能量特征的检测性能大大下降。

能量-熵特征是能量与熵的结合。熵是信息的期望值,在信息编码中广泛采用。实际上,语音的熵和非语音的熵相差非常大[6]。因为从熵的角度看来,有序信号具有小的熵值,而无序信号具有大的熵值, 因此可以用这个特征进行正确的端点检测。在某些场合下,用基于熵的方法要比纯粹基于能量的方法可靠得多,尤其是在具有机械噪声的非稳定噪声的环境中。图 2(f)表示能量-熵特征的检测结果,显然,它具有比其它特征更好的噪声鲁棒性。

2.1.2　特征计算

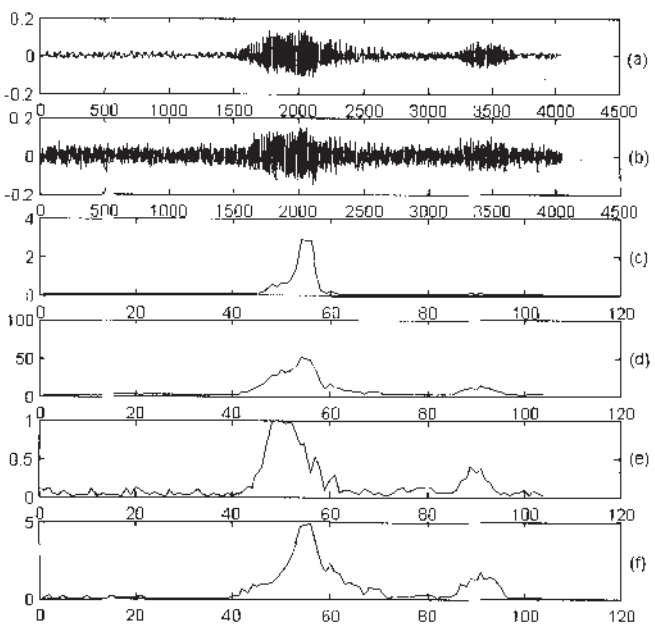
对每一帧,计算 256 点的 FFT,可以得到能量谱 EngChart[0],EngChart[1],...EngChart[127],以及短时能量 E。倒谱距离 C 和能量谱的方差 D 可以由下面(1)~(3)求出。

$$E=\sum_{i=0}^{127} EngChart[i] \tag{1}$$

$$D=\sum_{i=0}^{127}\left(EngChart[i]-\frac{E}{128}\right)^2 \tag{2}$$

$$C=4.3429\sqrt{(c_0-c'_0)^2+2\sum_{n=1}^{P}(c_n-c'_n)^2} \tag{3}$$

其中,$C_n=\sum_{i=0}^{127} EngChart\ [i]e^{-j\frac{2\pi}{128}in}$,$c'_n$ 为噪声帧的倒谱系数的平均值。



(a)原始语音信号;(b)加入噪声的信号;(c)短时能量方法;
(d)能量方差方法;(e)倒谱距离方法;(f)能量-熵方法

图 2　各种端点检测方法的性能比较

计算能量-熵特征。将 0~4000Hz 的全频段划分为 16 个频带,每 8 点(250Hz)构成一个频带。计算每一个频带的能量 $S(f)$,那么 $S(f)$ 就是频带中的总能量。$P(i)$ 是相应的概率密度函数。

$$S(f_i)=\sum EngChart[j] \tag{4}$$

$$P(f_i)=\frac{S(f_i)}{\sum_{k=0}^{8} S(f_k)} \tag{5}$$

对 $P(i)$ 和 $S(f)$ 进一步改进。因为语音信号的频谱一般分布在 250~3500Hz,因此不属于这个频率范围内的 $S(f)$,将它设为零。另外,如果某一个频带的能量超过总能量的 90%,为了消除集中在一些特殊频率的噪声,可以限定 $P_i<0.9$,即采用下述约

束关系：

$S(f_i)=0$，if $f_i \le 250\text{Hz}$ or $f_i \ge 3500\text{Hz}$

$P_i=0$，if $P_i \ge 0.9$

经过上述改进，第 i 帧的负熵 H_i 定义为：

$$H_i=\sum_{j=0}^{7} P_j \log P_j \qquad (6)$$

于是可以获得 EE 特征：

$$M_i=|(E_i-C_E)(H_i-C_H)| \qquad (7)$$

C_E 表示平均能量，C_H 表示噪声帧的熵。

2.2 5 状态推理机决策

在基于阈值的端点检测系统中结合一种自适应的 5 状态自动机推理。5 种状态的定义为：①无语音；②语音假设；③语音；④爆破音/无语音；⑤可能的语音连续。状态之间的转移是由语音的持续时间约束和一个标志来控制，控制标志可以从每一帧中的特征提取出来。在上面提到的 5 个状态，其中有三个状态是新引入的状态：语音假设；爆破音/无语音；可能的语音连续。在突发噪声信号影响下，某些特征的计算值增加剧烈，语音假设状态可以防止自动机转移到语音状态。但是在特征值很大，并且停留在此状态持续时间很长，自动机就会转移到语音状态。

在算法中，为了精确地检测出语音的起点和终点，对自动机进行了改进。设定特征的阈值时，对每一个特征选择两个 T_u（上限）和 T_l（下限）。自动机的每个不同状态，用不同的值来确定状态标志 $flag$。如果当前状态为①无语音和④爆破音/安静，T_u 就作为实际的阈值来确定状态标志；如果当前状态为②语音假设，③语音，⑤可能的语音连续，于是用 T_l 作为确定状态标志的实际阈值。

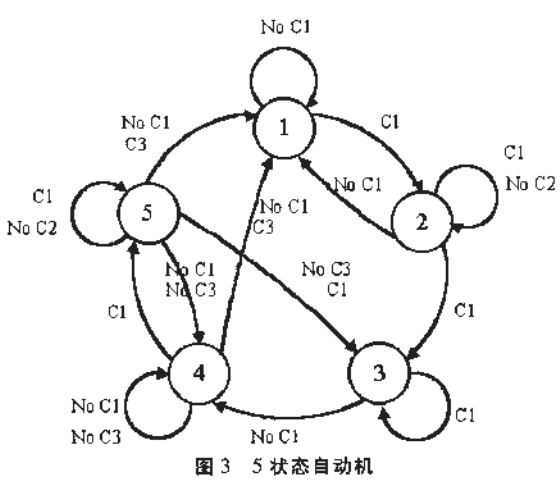


图 3 5 状态自动机

3 实验

通过实验，对算法端点检测的性能进行了评价。评价的测试数据是按照 8KHz 采样率，16Bit 存储的各种环境噪声下的语音信号，环境包括：车站、商场、街道和办公室。在每一种环境下，采集的数据来自两个不同的麦克风，其中一个靠近嘴唇，具有较高的信噪比；另一个远离嘴唇，具有较低的信噪比。一共采用了 5 组数据进行评价。每一组数据有 200~700 句话，每句话包含 2~6 个单词不等。

算法性能评估。为了评价该文算法，与通常的端点检测算法进行性能比较，经典的算法大多采用基于单一的短时能量特征，或者短时能量特征与倒谱距离特征的结合。结果表明，采用倒谱距离信息与短时能量两个特征在大多数情况下可以改善端点检测的准确率；而采用该文提出的 4 个特征（短时能量，倒谱距离，熵和能量谱方差）之后，端点检测准确率获得极大的提高，相对于短时能量法平均提高了 13.9%，相对于能量-倒谱方法平均提高了 9.9%。

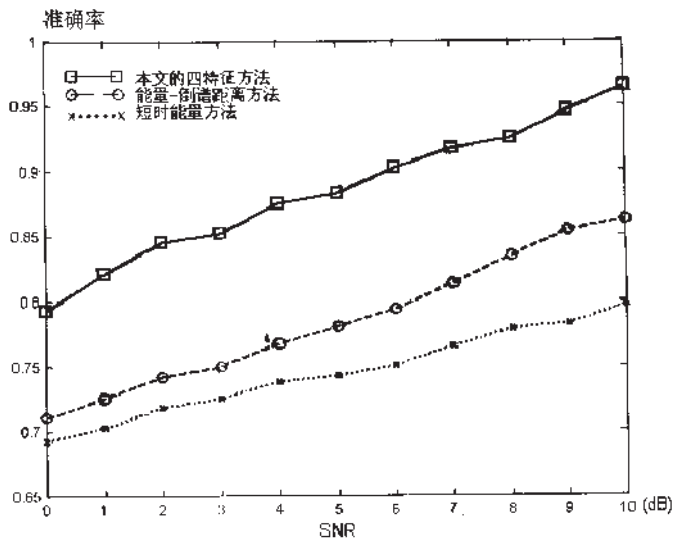


图 4 3 种端点检测方法性能比较

4 结论

该文的端点检测方法采用 4 个独立性强的语音信号特征（特征-短时能量、倒谱距离、能量谱方差和能量-熵特征），在各种噪声环境下进行实时端点检测。每一种特征具有对某种噪声的适应能力，利用每个信号特征检测出的端点，通过一个 5 状态的自动机进行推理决策，可以有效克服单个特征检测的不准确性，提高对各种环境噪声的抵制能力。采用相互独立性强的多个特征可以有效地提高算法对各种噪声环境的适应能力。由于算法的复杂度相对较低，因此非常适合实时应用场合。

（收稿日期：2001 年 12 月）

参考文献

1.L S Huang，C H Yang.A Novel Approach to Robust Speech Endpoint Detection in Car Environments[C].In：ICASSP'00，vol.3，2000：1751~1754

2.M H Savoji.A Robust Algorithm for Accurate Endpointing of Speech[J].Speech Communication，1989：45~60

3.A Martin，D Charlet.Robust Speech/non-speech Detection Using LDA Applied To MFCC[C].In：ICASSP'01，vol.1，237~240

4.W H Shin.Speech/non-speech Classification Using Multiple Features for Robust Endpoint Detection[C].In：ICASSP'00，vol.3，1399~1402

5.Haigh J A，Mason J S.Robust Voice activity Detection Using Cepstral Features[C].In：IEEE Region 10 Conference on Computer，Communication，Control and Power Engineering，1993：vol.3，321~324

6.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，1997：2：757~760