文章编号: 1006-2467(2000) 02-0185-04

应用倒谱特征的带噪语音端点检测方法

韦 晓东[1], 胡 光锐[2], 任晓林[2]
(1. 上海交通大学与贝尔实验室通信和网络联合实验室; 2. 上海交通大学 电子工程系, 上海 200030)

摘 要: 传统的语音端点检测方法以信号的短时能量、过零率等简单特征作为判决特征参数. 这些方法在实际应用中, 尤其当信号信噪比比较低时, 无法满足系统的需要. 文中利用语音信号的倒谱特征作为判决抽样信号帧是否为语音信号的依据, 并提出了倒谱距离测量法和循环神经网络法. 通过对宽带噪声-白噪声干扰情况和一种特殊噪声——汽车噪声情况的实验, 发现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, 更适合于实际应用.

关键词: 端点检测; 倒谱距离; 神经网络
中图分类号: TN 912.34 文献标识码: A

Endpoint Detection of Noisy Speech by the Use of Cepstrum

WEI Xiao-dong[1], *HU Guang-rui*[2], *REN Xiao-lin*[2]
1. Shanghai Jiaotong Univ.-Bell Labs Comm. and Network Joint Lab. Shanghai 200030, China;
2. Dept. of Electronic Eng. Shanghai Jiaotong Univ., Shanghai 200030

Abstract: Most practical automatic speech recognition(ASR) systems must work with a small signal-noise ratio(SNR), and the conventional speech detection methods based on some simple features such as energy cannot work well in the noisy environments. In this paper, cepstrum was used as the feature to detect the voice activity. Two algorithms for endpoint detection of noisy speech signal were proposed. The first one takes the cepstral distances as the decision thresholds instead of short-time energy. The second approach takes advantages of recurrent neural networks. The experiments show that the high accurate rates can be obtained in the noisy cases.

Key words: endpoint detection; cepstral distance; neural network

语音信号端点检测的目的是从连续采样得到的数字信号中检测出语音信号段和噪声信号段. 准确的语音端点检测不仅提高了系统处理效率, 同时也能够提高系统的识别率[1,2]. 传统的语音端点检测方法在噪声环境下性能下降, 它们在信噪比较低的情况下性能很差, 有时甚至无法工作. 本文针对传统方法通常采用能量等一些简单特征的缺陷, 提出了应用倒谱系数作为判决特征的带噪语音端点检测方法. 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. 通过对带噪语音的实验比较, 证实了基于倒谱特征的带噪语音端点检测法的有效性.

1 语音信号端点检测的基本步骤和常用方法

语音信号端点检测的一般过程为:

(1) 采样信号被分割成相邻有重叠的信号帧;

(2) 对每一帧信号, 选取并计算一种特征向量;

(3) 根据对应于信号的特征向量序列, 利用一定的判决准则, 得到语音段和非语音段的判决;

(4) 对第(3) 步的判决结果做后处理得到语音的起始点和终止点, 即得到语音端点检测的结果. 后

收稿日期: 1999-05-10
基金项目: 贝尔实验室(中国) 上海分部资助项目
作者简介: 韦晓东(1970~), 男, 博士生.

处理过程是为了避免把人在发声过程中出现的自然停顿当作背景噪声. 这些停顿包括句子内的停顿, 还包括一个词甚至一个音素内的发音间歇.

常用的语音信号端点检测方法是能量判决法. 在这种方法中, 每一帧信号的短时能量或者短时能量的对数被作为判决特征. 它用门限判决的方法来判断每一帧信号是语音还是背景噪声, 其最大优点是它非常简单. 它在低噪声情况下具有非常好的性能, 比如在信噪比大于 20 dB 时, 检测准确性接近 100%. 但是, 实际的语音识别系统可能应用于不同的环境. 在低信噪比情况下, 基于能量的判决方法已经无法取得满意的检测结果, 甚至无法工作. 这是因为适当的噪声能量门限非常难以估计. 针对这些缺陷, 一些其他的语音信号特征, 比如过零率、基音周期被加入和短时能量一起作为判决特征. 但是这一类方法在低信噪比情况下仍然无法得到满意的检测结果. 因为, 不同噪声类型, 如办公室噪声、汽车噪声的过零率区别很大, 所以很难通过经验值得到合适的门限.

上述方法最根本的问题就是判决门限往往通过经验值来确定, 而门限值对整个端点检测的影响极大.

2 基于倒谱特征的带噪语音信号端点检测

在目前大多数语音识别系统中, 倒谱系数如线性预测倒谱系数(LPC-CEP) 或 Mel 刻度倒谱系数(MFCC) 被选为代表语音信号的特征参数, 因为倒谱被认为是语音信号一种较好的时频表示. 本文用倒谱向量取代这些简单特征参数作为语音端点检测的判决特征向量.

2.1 倒谱距离测量法

信号倒谱可以看成是信号能量谱密度函数 $S(\omega)$ 的对数的傅里叶级数展开[2], 即

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{-\mathrm{j}n\omega} \tag{1}$$

式中, c_n 为倒谱系数, $c_n = c_{-n}$ 是实数, 且

$$c_0 = \int_{-\pi}^{\pi} \log S(\omega) \frac{\mathrm{d}\omega}{2\pi} \tag{2}$$

对一对谱密度函数 $S(\omega)$ 和 $S'(\omega)$, 应用 Parsavel 定理可用倒谱距离来表示对数谱的均方距离:

$$d^2_{\mathrm{cep}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|\,\mathrm{d}\omega = \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2 \tag{3}$$

式中, c_n 和 c'_n 为对应于谱密度函数 $S(\omega)$ 和 $S'(\omega)$ 的倒谱系数. 对数谱的均方距离可以表示两个信号谱的区别, 故它可以作为一个判决参数. 本文用倒谱距离测量的方法来判定各个信号帧是语音帧还是噪声帧. 此方法就是倒谱距离测量法. 实际上基于能量的检测方法可看成是倒谱距离测量法的特例, 因为 c_0 含有信号的能量信息.

倒谱距离测量法根据每个信号帧与噪声帧的倒谱距离的轨迹进行检测. 它也采用门限判决的方法, 只是同能量方法相比, 门限值是倒谱距离门限而不是短时能量门限. 这种方法的过程如下:

(1) 假定抽样信号的前几帧信号是背景噪声. 利用这些帧的倒谱系数的平均作为背景噪声倒谱系数的估计值. 用向量 $\boldsymbol{C}$ 表示.

(2) 计算每帧信号的倒谱系数, 然后计算每帧信号的倒谱系数与噪声倒谱系数估计值的倒谱距离. 式(3) 可近似为[2]

$$d'_{\mathrm{cep}} = 4.342\,9 \sqrt{(c_0 - c'_0)^2 + 2\sum_{n=1}^{P}(c_n - c'_n)^2} \tag{4}$$

式中: c_n 为对应于 $\boldsymbol{C}$ 的倒谱系数; P 为倒谱系数的阶数.

(3) 由步骤(2) 计算的各帧倒谱距离得到倒谱距离轨迹. 然后利用近似能量方法中门限判决的方法检测语音段和噪声段.

(4) 为使背景噪声倒谱系数的估计值 $\boldsymbol{C}$ 能够适应噪声的变化, 采用一个平滑处理过程. 背景噪声倒谱估计 $\boldsymbol{C}$ 利用已经检测过的上一信号帧倒谱向量, 按照

$$\boldsymbol{C} = a\boldsymbol{C} + (1 - a)\boldsymbol{C}_t \tag{5}$$

规则更新. 式中: t 为上一信号帧帧号; $\boldsymbol{C}_t$ 为倒谱向量; a 为一个时间调整因子, 并且这一帧被认为是非语音帧.

(5) 后处理得到语音的起始点和终止点. 后处理可以通过中值滤波实现.

图 1 所示为一个采样信号的倒谱距离轨迹与短时能量轮廓的比较图. 一段录有三个词发音的信号(见图 1(a)) 被人为加入白噪声形成带噪语音信号(见图 1(b)). 从图 1(b) 中可以看出, 原始语音信号几乎完全被噪声信号掩蔽. 其信噪比小于 - 10 dB. 在如此恶劣的情况下, 根据图 1(c) 倒谱距离的轨迹, 仍然可以找到词与词、词与背景噪声之间的界线(尽管图中所示的情况存在一些误差). 而利用图 1(d) 中的短时能量轮廓就无法达到这一目的.

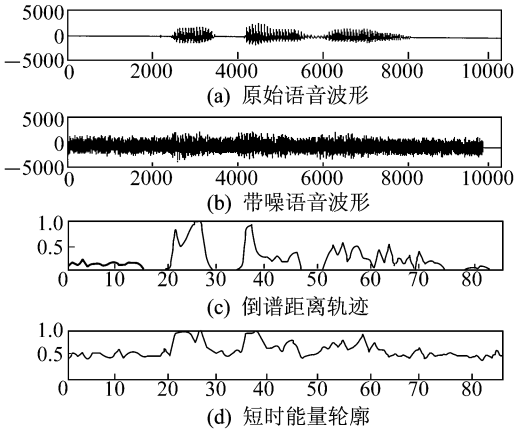

图 1　倒谱距离轨迹与短时能量轮廓比较图

Fig. 1　Trajectory comparison between cepstral distance and energy

这种算法实质上仍然依靠门限判决. 在一些特殊情况下, 如信噪比非常低时, 语音信号本身严重的谱失真会给门限的估计带来困难. 另外一些非平稳噪声, 如开关门的声音, 它们与一些语音信号的倒谱距离非常小, 所以很难区分当前是语音还是噪声. 为克服门限判决法固有的缺陷, 本文提出了应用神经网络的模式判别方法.

2. 2　应用循环神经网络的语音端点检测方法

近年来, 循环神经网络(Recurrent Neural Network, RNN) 在一些语音识别系统中被用作声音模型. 这些方法主要利用了 RNN 能够并行计算子词(如音素) 出现概率, 并且能表示语音上下相关信息的特性[1, 3]. 本文提出一种用 RNN 进行带噪语音信号端点检测的算法, 选用三层 RNN(见图 2), 且隐层的输出全部反馈到输入层. 这一个三层 RNN 是一个动态系统[4], 它可以表示输入信号的前后相关信息. 语音信号被认为是一种上下相关的动态信号, 可以用这种神经网络来表示语音信号特征, 并用来判决语音信号的模式. 文中用 RNN 进行语音信号的端点检测, 实际上是利用了 RNN 的模式分类能力. 输入的采样语音信号被分成三种模式: 浊音(U)、清音(V) 和背景噪声(N). 故采用的 RNN 输出层含有三个输出节点, 它们分别对应于三种模式, 见图 2 中的标号. 这里把浊音和清音分开是因为清音信号有时与噪声有类似的统计特性. 为减少判决时出现的混淆, 提高系统性能, 本文把清音单独作为一个模式.

图 2 中输出层的三个输出对应于三个不同的模式, 其输出为 $y_i(t)(i=1,2,3)$; 隐层输出为 $s_j(t)$; 输入为 $x_k(t)$. 它们之间的关系如下:

$$y_i(t) = f\left(\sum_{j=1}^{N_h} w_{ij} s_j(t) + \theta\right) \qquad i = 1, 2, 3$$

$$s_j(t) = f\left(\sum_{k=1}^{N_i} \psi_{jk} x_k(t) + \sum_{l=1}^{N_h} \psi'_{jl} s_l(t-1) + \theta_j\right)$$

其中: $j = 1, 2, \cdots, N_h$; $f(\cdot)$ 为 Sigmoid 函数[5]; w_{ij} 为从隐层到输出层的连接权值; ψ_{jk} 为从输入层到输入节点到隐层的连接权值; ψ'_{jl} 为输入层反馈节点到隐层的连接权值; θ 和 θ_j 分别为隐层和输出层门限阈值; N_i 和 N_h 分别为输入层输入节点数和隐层节点数; 输入层总节点数为 $N_i + N_h$.

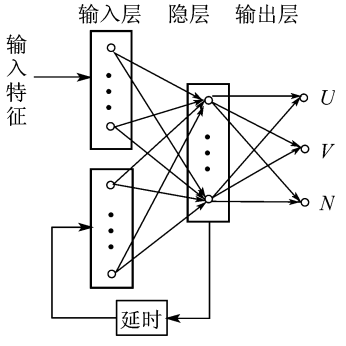

图 2　循环神经网络的结构

Fig. 2　Structure of recurrent neural network

如前所述, 在噪声环境下, 短时能量、过零率和其他简单特征不适于判决信号是语音还是非语音. 本文仍用信号的倒谱特征向量作为循环神经网络的输入特征. 采样信号为经过信号预处理后得到的特征向量序列, 即倒谱向量序列. 信号预处理过程类似前面所述的信号处理模块. 包括信号分帧、预加重、加窗、倒谱计算和特征向量提取等. 这里用于表示每一帧信号的特征参数向量由 12 阶倒谱系数、12 阶倒谱系数的导数和短时能量的导数组成, 共 25 维. 故采用 $N_i = 25$.

采样信号的特征参数序列输入后, 循环神经网络的输出层输出相应的序列. 本算法通过 3 个输出节点输出序列在每个时刻的择大判决来判断当前帧信号属于哪种模式. 择大判决的结果输出是一个判决结果序列. 通过对判决结果序列的后处理, 即选择第一个判决为清音或浊音的帧为发音的开始, 最后一个被判为非噪声的帧作为发音的结束. 为避免出现把清音信号截去的现象, 作者在最后的端点检测结果前后加入一些 "保护" 帧.

神经网络连接权值的训练采用方向传播算法(即 BP 算法)[4, 5]. 这里, 输入的学习样本就是语音信号的特征向量序列, 教师样本就是特征向量对应的三种模式: 浊音、清音和噪声. 为使神经网络的权

值和阈值对各种噪声环境具有适应能力, 本文采用了一种称为免疫训练的训练过程. 这一过程就是首先用干净的语音库训练神经网络的权值集. 然后逐渐用具有不同噪声电平的带噪语音库对训练好的权值集进行更新. 这里用于训练的语音库信噪比逐渐降低, 即信噪比从20、15、10、5 dB 到0 dB. 最后得到的权值集对信号中的噪声具有一定的 '免疫 '能力, 也就是说, 神经网络的权值对输入信号的信噪比有一定的不敏感性. 这里利用了神经网络具有记忆能力的特点.

3 实验结果与结论

测试语音信号在安静环境下录制, 以 8 kHz 抽样, 16 位脉码调制(PCM) 量化, 并人为地以不同程度加入白噪声和汽车噪声形成带噪语音. 实验中语音信号被分为 30 ms 的帧, 相邻帧有 2/ 3 重叠. 倒谱系数采用 12 阶 LPC 倒谱系数. 各段采样语音信号通过人耳区分的手工标号作为测试各种语音端点检测结果的标准. 实验中比较各种方法性能的测度主要包括: ① $p(A|S)$ —— 当信号帧为语音, 并正确检测为语音的准确比率; ② $p(B|N)$ —— 当信号帧为噪声, 并正确检测为噪声的准确比率; ③ $p(A)$ —— 整体检测准确率, $p(A) = p(A|S)p(S) + p(B|N)p(N)$. 其中 $p(S)$ 和 $p(N)$ 分别是在用于测试的语音库中语音和噪声所出现的概率. 测试了应用倒谱特征系数的两种算法和传统的能量法在白噪声和汽车噪声情况下的性能. 选来做测试的带噪语音文件每组(每种信噪比情况) 300 个, 包括男声和女声发音. 两组带有汽车噪声的语音信号平均信噪比分别为9 dB和 5 dB. 它们分别模拟在汽车中速和高速行驶时的环境下录制的语音. 各种方法的检测实验结果见表 1. 表中: Energy 为对数短时能量判决法; CDM 为倒谱距离测量法; RNN 为用应用循环神经网络的算法; WN 表示白噪声; CN 表示汽车噪声. 对于语音识别系统应用而言, 一个语音端点检测方法的最重要性能指标就是语音检测准确率 $p(A|S)$. 本文提出的两种方法的语音检测准确率 $p(A|S)$ 在信号信噪比只有几个 dB 的恶劣情况下, 检测准确率在 90% 以上. 但为了避免在系统特征提取中出现丢弃语音信号特征的情况, 需要在后处理过程中在端点检测结果的前后加入一定的保护帧.

表 1 噪声情况下各种语音端点检测方法实验结果比较

Tab. 1 Experimental results of endpoint detection of noisy speech

方法	测度	WN/dB			CN/dB	
		15	5	0	9	5
Energy	$p(A\|S)$	0.98	0.76	0.64	0.86	0.70
	$p(B\|N)$	0.99	0.60	0.51	0.79	0.50
	$p(A)$	0.98	0.71	0.59	0.84	0.61
CDM	$p(A\|S)$	0.99	0.96	0.92	0.93	0.92
	$p(B\|N)$	0.99	0.80	0.70	0.83	0.75
	$p(A)$	0.99	0.90	0.81	0.90	0.86
RNN	$p(A\|S)$	0.98	0.95	0.94	0.96	0.91
	$p(B\|N)$	0.98	0.81	0.74	0.85	0.78
	$p(A)$	0.98	0.90	0.88	0.92	0.89

通过对宽带噪声 —— 白噪声干扰情况和一种特殊噪声 —— 汽车噪声情况的实验, 作者发现基于倒谱特征参数的方法具有更高的准确度. 这表明, 倒谱特征参数比短时能量等其他参数对语音发生环境的适应力强, 更适合于实际的语音处理系统.

参考文献:

[1] Lee C H, Soong F K, Paliwal K K. Automatic speech and speaker recognition-advanced topics [M]. Boston: Kluwer Academic Publishers, 1996.

[2] Rabiner L R, Juang B H. Fundamentals of speech recognition [M]. New Jersey, USA: Murray Hill, 1993.

[3] Gerven S, Xie Fei. A comparative study of speech detection methods [A]. EUROSPEECH'97 [C], Greece, 1997.

[4] Chen S H, Liao Y F, Chiang S M, *et al*. An RNN-based preclassification method for fast continuous Mandarin speech recognition [J]. IEEE Trans Speech and Audio Processing. 1998, 6(1): 86～90.

[5] Morgan D P, Scofield C L. Neural networks for speech processing [M]. Boston: Kluwer Academic Publishers, 1991.