噪声环境中基于 HMM 模型的语音信号端点检测方法*

朱 杰, 韦晓东
(上海交通大学与贝尔实验室通信与网络联合实验室)

摘 要 在噪声环境下如何提高语音信号端点检测的准确性是自动语音识别(ASR)研究中的一个重要课题. 常用的基于短时能量的端点检测方法对于能量较低的音节或在信噪比较低的环境下, 检测性能不够理想. 讨论了一种基于 HMM 模型的语音信号端点检测方法. 先用训练的方法生成背景噪声和废料的模型, 再用 Viterbi 解码算法对待测信号进行处理, 并给出了具体的实现方法. 实验测试结果表明, 基于 HMM 的端点检测方法的检测性能接近于人工检测, 方法是有效的.

关键词 隐马尔可夫模型; 端点检测; 语音识别
中图法分类号 TN 912.34

Speech Signal Endpoint Detection Method Based on HMM in Noise

Zhu Jie, Wei Xiaodong
Shanghai Jiaotong University & Bell Labs Communications and Networks Joint Laboratory, Shanghai, China

Abstract To improve performance of endpoints detection in noisy environments is a significant subject in automatic speech recognition (ASR). The performance of general endpoints detection methods based on short time energy is unsatisfied for lower energy syllable or in the environments with lower signal-to-noise ratio. The endpoints detection method based on HMM is discussed. The background model and garbage model are built with training, then the Viterbi decoding algorithm is used to process the speech signal. the steps of realization are presented in this paper. The results from experiments show that this method is very effective and the performance is approaching to that of manual detection.

Key words hidden Markov model(HMM); endpoint detection; speech recognition

---

语音信号数字处理中的端点检测是语音识别研究中一个十分重要的环节. 准确地从背景噪声中检测出语音信号的起始点和终止点, 可以减少采集的数据量, 删除不含语音成分的背景噪声, 从而可以大大降低语音识别处理中的计算量和处理时间, 提高识别的准确性, 为语音识别系统在实时处理中得以应用创造条件. 随着自动语音识别(ASR)技术在车载电话通信、关键词识别、人员身份确认、连续语音识别等方面的应用逐渐广泛, 在噪声背景下语音信号端点的准确检测已变得十分重要. 一个好的端点检测

收稿日期: 1998-04-04
* 美国贝尔实验室上海分部资助项目
朱 杰: 男, 1963 年生, 副教授. 邮编: 200030

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

器应该具备以下性能[1]: ①在信噪比比较低的环境下(如: 在汽车噪声中, 在麦克风噪声中、在有嘈杂人声的环境中等), 仍应具有端点的准确检测能力. ②对于一些能量较低的爆破音、鼻音, 如: /f/、/th/、/t/、/s/、/sh/等, 不会截去这些音节的有效成分, 否则会对识别结果造成影响. ③能有效地对字间间隙进行光滑, 消除字间间隙对端点检测可能造成的误判.

本文讨论了一种基于 HMM 模型的语音信号端点检测方法, 给出了具体的实现步骤, 并将此方法与基于能量的端点检测方法进行了测试对比实验.

1　端点检测的常用方法

1.1　基于短时能量的方法

有不少端点检测算法是基于信号的短时能量. 先算出背景噪声能量的统计特性, 定出能量门限, 利用能量门限来确定语音信号的起止点. 这种方法在背景噪声幅度保持恒定, 且远低于语音信号幅度时, 并且对孤立字的最小帧数、最大帧数、句子间间隙的最小帧数, 以及人为的突变性音节帧数有充分先验知识的条件下, 可以十分有效地准确检测出语音信号的端点[2]. 尤其在用过零率方法作辅助处理来调整检测后的端点时, 测出的端点位置是比较准确的. 然而, 当该类算法在信噪比较低的情况下, 检测性能开始恶化. 在更恶劣的情况下, 甚至完全不能检测出其端点. 而且, 过零率方法在背景噪声是汽车噪声、麦克风噪声或白噪声时, 噪声的过零率均不相同, 有时与语音某些音节的过零率相重叠, 也很难作为一种辅助的判据.

1.2　模式识别方法

此类算法把语音信号端点检测问题看作是对每帧信号进行分类, 任意时刻的语音特征矢量 O_t 可以看作由一对分布函数 $f_{\theta_i}(O_t)$ 产生的. 即 H_0(语音): O_t——$f(\theta_0)$ 和 H_1(背景): O_t——$f(\theta_1)$. 而 $f(\theta_i)$ 满足 Gaussian 分布, 即 $f(\theta_i)$——$N(m_i, R_i)$. 通过建立相应的检测准则(如 Bayes 准则), 对每帧语音矢量进行划分, 确定其属于 H_0 或 H_1. 一般, 可以建立如下的正交识别准则:

$$\min \hat{d}_n = \min\{(O_t - m_n)^t R_n^{-1}(O_t - m_n)\}$$

根据上式可以识别出该时刻 O_t 是属于 H_0 还是 H_1. 此时一定有 $p_n f(\theta_n) > p_i f(\theta_i)$. 其中: p_i 为先验概率; $n \neq i$; $n = 0, 1$; $i = 0, 1$.

该类算法的优点是考虑到了语音帧之间的相关性及误差的概率最小. 但主要缺点是采用这种算法很难找到能显著区分浊音和清音的语音特征[3]. 必须注意, 由于上述的判据是对每帧语音进行识别, 所以必须用相应的中值滤波来进行平滑.

2　基于 HMM 的端点检测方法

HMM 是语音识别技术中目前应用最广泛的一种模型. 在训练阶段, 训练语音对模型各状态的统计特性进行训练, 得出模型参数. 在测试阶段, 待测语音与训练模型进行匹配, 选择得分最高的作为识别结果.

根据 HMM 的基本处理方法, 尝试把 HMM 方法直接用于语言信号的端点检测, 因为所谓"端点", 无非就是把被测信号看作是有两部分组成: 背景(Background 或 Silence)和废料(Garbage, 在语音处理中, 习惯上把有用或无用的发音统称为"废料"), 而废料就是上述两部分的分界处. 在训练阶段, 分别得出背景噪声和废料的模型参数. 在测试阶段, 用 Viterbi 解码方法在训练模型基础上对被测语音进行分解, 求出语音的哪些帧与背景噪声匹配, 哪些帧与废料匹配, 从而得出端点的所在处.

一个完整的基于 HMM 方法的端点检测系统如图 1 所示.

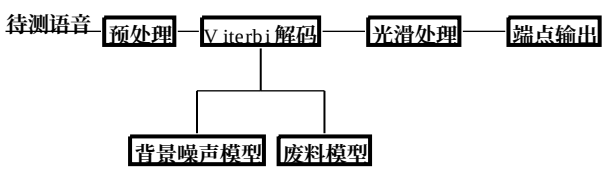

图 1　基于 HMM 方法的端点检测系统

Fig. 1　The endpointing system based on HMM

(1) 为了能有效地采用 HMM 方法进行处理, 须对每帧待测语音进行预处理. 包括: ①预加重处理. 按下式设计一个一阶高通滤波器: $H(z) = 1 - \alpha Z^{-1}$, 其中预加重系数 α 一般选择为 0.95.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

采用预加重的原因是它可以有效地压缩输入语音的动态范围, 使后面的LPC 分析更稳定 同时, 高通滤波器可以有效地滤除输入信号中的直流成分. ②开窗处理 一般采用汉明(Hamming)窗 在本试验中, 数字采样率为 8 kHz, 窗口总长度设计成 30 ms, 即每个窗口有 240 个采样点 窗口每次位移 10 ms, 有 20 ms 的重叠成分. 窗口的重叠起到了平滑特征参数的作用 ③倒谱计算. 对每帧语音求出 p 阶倒谱系数, 分析中取 $p = 12$ ④倒谱加权处理 为了避免倒谱系数数值过小而对识别造成影响, 一般均采用上升正弦函数进行倒滤波处理 ⑤倒谱系数的一阶和二阶导数处理 对每帧倒谱系数求出其一阶和二阶导数系数:

$$\Delta C_i(m) = C\sum_{k=-2}^{2} kC_{i-k}(m), \quad \Delta\Delta C_i(m) = C\sum_{k=-2}^{2} k\Delta C_{i-k}(m)$$

其中: i 为第 i 帧语音信号; m 为第 m 个倒谱系数; C 为常数 ⑥能量及其他特征处理 对每帧语音求出其对数能量, 及能量的一阶和二阶导数 这样, 在本实验中, 通过预处理, 对每帧待测语音共提取出 39 个特征值, 构成一特征矢量

(2) Viterbi 解码 经上述预处理后的语音送入Viterbi 解码器, 采用Baum-Welch 算法, 从Viterbi 解码器的输出端即可得到待测信号的端点

(3) 光滑处理 由于基于HMM 的端点检测方法是对待测语音逐帧进行处理, 对字间间隙比较敏感, 所以, 必须用中值滤波进行平滑处理

3 结果与结论

作者对同一组女生(10 人)语音信号(在麦克风噪声环境下, 信噪比小于 6 dB)分别采用基于能量的端点检测方法和基于HMM 的端点检测方法进行测试, 并对照语音波形用人工检测方法进行测试 测试结果如表 1 所示 实验表明: ① 基于HMM 的端点检测方法检测的准确率明显高于基于能量的方法 在信噪比逐渐降低的情况下, 效果更加明显 ② 基于HMM 的端点检测方法检测低能量的清音或爆破音、鼻音的端点位置时, 性能明显高于基于能量的方法, 很少出现截去音节有效成分的现象

表 1 端点检测结果比较

Tab 1 Comparison of the experiment

序 号	开始帧位置			结束帧位置		
	能量方法	HMM 方法	手工方法	能量方法	HMM 方法	手工方法
1	176	185	187	245	238	236
2	189	197	195	237	230	232
3	153	156	160	212	208	205
4	154	162	164	201	196	196
5	287	296	297	326	320	321
6	179	188	197	241	238	237
7	181	181	184	239	237	238
8	208	206	207	269	268	262
9	202	196	196	245	250	256
10	180	189	191	236	231	234

HMM 的训练环境与实际被测信号的语音环境会有很大差异. 比如: 当训练是在安静的环境下进行, 而实际测试环境是在汽车噪声中进行时, 由于背景噪声模型与实际情况很不相符, 其性能会显著下降 因此, 必须采用能自适应调节的背景噪声模型 其具体的实现方法是正在研究的课题 此外, 特征矢量维数的大小对检测性能的影响也值得研究, 因这关系到在实时处理时该算法的实用性

致谢: 本课题研究中获得的一些结论是作者在美国贝尔实验室(新泽西州)进修期间得到的, 在此期间得到了美国贝尔实验室及其上海分部的大力支持, 并得到了 Sunil K. Gupta 博士和 Frank Soong 博士的指导, 在此深表谢意

参 考 文 献

1 Savoji M H. A robust algorithm for accurate endpointing of speech signals Speech Commun, 1989, 8: 45~60

2 Janqua J-C. A robust algorithm for word boundary detection in the presence of noise IEEE Trans SAP, 1994, 2(3): 406~412

3 Pawate B I (Raj). A new method for segmenting continuous speech. '94 ICASSP, 1994 I-53~I-56

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net