

语音活动检测对方言辨识系统的影响研究

张 宁¹, 顾明亮^{1,2}, 朱俊梅², 周 杰¹

(1. 江苏师范大学 物理与电子工程学院, 江苏 徐州 221116;

2. 江苏师范大学 语言科学学院, 江苏 徐州 221116)

摘 要:分别把基于阈值判断和基于统计模型的语音活动检测(VAD)应用于汉语方言辨识系统中,对比了系统识别率及运算时间。其中基于能量、过零率等阈值判断的方法以其算法简单、计算量少的优点在高信噪比噪声环境下取得较好的效果,但在低信噪比噪声环境下准确性及鲁棒性急剧下降。在相同测试环境下,采用统计模型的DD+Hang-over算法取代传统经典阈值算法。实验表明,基于统计模型的算法在高斯混合模型(GMM)系统下运算时间稍长,但抗噪声性能明显优于基于阈值判断的算法,尤其在低信噪比的加性噪声环境下效果更显著。

关键词:方言辨识;语音活动检测;DD+Hang-over 算法

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2012)11-0073-04

Study on Influence of Voice Activity Detection for Dialects Identification System

ZHANG Ning¹, GU Ming-liang^{1,2}, ZHU Jun-mei², ZHOU Jie¹

(1. School of Physics & Electronic Engineering, Jiangsu Normal University, Xuzhou 221116, China;

2. School of Linguistic Science, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: Voice active detection (VAD) based on threshold judgement and statistical-model is used in Chinese dialect identification system respectively. The recognition accuracy and computing time have also be compared. Threshold-based methods such as short-term energy has achieved good results in a high SNR noisy environment with its simple algorithm and less computation. Whereas, its performance declines sharply in the low SNR environment. Statistical model-based method with DD+Hang-over is used instead of the traditional threshold-based methods under the same test environment. Experiments with the system of GMM show that although statistical model-based method with DD+Hang-over takes more time, it has better anti-noise performance, especially in the low SNR environment.

Key words: dialect identification; voice activity detection; DD+Hang-over algorithm

0 引 言

汉语方言辨识是一项根据说话人所说的语音自动判断说话人所属的方言类别或其所在籍贯的技术,它在语音识别、信息检索、口语翻译、军事安全等领域有及其广泛的应用前景。20世纪70年代初,美国德州仪器公司的研究人员首次提出自动语言辨识的研究课题。汉语方言辨识是自动语言辨识的一个重要的研究方向,当前识别系统中比较热点的两个问题是特征的提取和分类器的设计,忽略了预处理中语音活动检测(voice activity detection, VAD)的重要性。语音活动检

测的目的就是从包含语音的一段信号中确定出语音的起始点和终止点,又称端点检测^[1]。有效的端点检测不仅能缩短处理时间,而且能排除无声段冗余信息干扰,进而提高系统的识别率。在文献[2]中,Ramírez等人验证了VAD对语音识别系统的重要性,但至今还没有人讨论VAD对汉语方言识别系统的影响。

目前语音端点检测方法主要分为两类:基于阈值的方法和基于模式识别的方法。其中基于阈值的方法因其简单、快速的优点而得到广泛的使用,其中比较典型的有基于时域的短时能量(Short-term energy)^[3]、短时过零率(Zero cross rate)^[4,5]等方法,基于频域的子带谱熵(Band-partitioning spectral entropy)^[6]等方法,基于倒谱域的倒谱特征^[7,8]等方法。模式分类的方法有基于DD+Hang-over^[9]、基于隐马尔科夫模型^[10]、基于支持向量机^[11]等方法。端点检测的准确性直接影响着后续的工作的有效进行。在方言识别系统的前

收稿日期:2012-03-11;修回日期:2012-06-21

基金项目:国家自然科学基金项目(61040053);江苏省普通高校研究生科研创新计划资助项目(CXZZ11_0903)

作者简介:张 宁(1987-),男,硕士研究生,研究方向为语音信号处理、模式识别;顾明亮,博士,教授,研究领域为语音信号处理、模式识别、机器学习等。

端处理中,文中利用 Sohn 等人提出的 DD+Hang-over 算法对比传统的经典阈值方法。为了检验该算法在汉语方言识别系统应用中的优越性,首先列出各种算法的原理,以及实验所用的语音数据库和系统的各项参数。在预处理中端点检测部分分别采用传统方法和统计模型方法,然后在高斯混合模型(GMM)^[12]系统下测出闽、粤、吴 3 种方言的识别率。实验证明在高斯白噪声的环境下,改进的系统平均识别率更高。

1 语音信号端点检测的原理

1.1 短时能量(Short-term energy, STE)

语音和噪声的主要区别在能量上,语音段的能量比噪声段的大,语音段的能量是噪声能量与语音能量的和。设语音波形的时域信号为 $x(t)$, 经过加窗分帧等处理后得到第 n 帧语音信号为 $x_n(m)$, 则满足下式:

$$x_n(m) = w(m)x(n+m) \quad 0 \leq m < N-1 \quad (1)$$

其中, $n=0, 1T, 2T, \dots$ 为帧号, N 为帧长, T 为帧移长度。 $w(m)$ 为窗函数, 通常选择汉明窗。设第 n 帧语音信号的短时能量用 E_n 表示, 则其计算公式如下:

$$E_n = \sum_{m=0}^{N-1} x_n^2(m) \quad (2)$$

设定适当的阈值就可以将语音段和无声段分开。

1.2 短时过零率(Zero-cross-rate, ZCR)

设一帧信号的短时过零率为 Z_n

$$Z_n = \frac{1}{2} \sum_{m=0}^{N-1} |\text{sgn}[x_n(m)] - \text{sgn}[x_n(m-1)]| \quad (3)$$

式中, $\text{sgn}[\]$ 是符号函数, 即:

$$\text{sgn}[x] = \begin{cases} 1 & (x \geq 0) \\ -1 & (x < 0) \end{cases} \quad (4)$$

短时过零率是信号频率的简单度量, 语音段有较高的过零率, 无声段的过零率相对较低。

1.3 子带谱熵(Band-partitioning spectral entropy, BSE)

对一帧信号求解 FFT, 则每个频率分量归一化谱的概率密度函数为:

$$p_i = \frac{x(f_i)}{\sum_{i=1}^n X(f_i)} \quad (5)$$

其中 p_i 是频率分量 f_i 的概率密度, $X(f_i)$ 表示频率 f_i 的振幅, n 是帧长。因此第 k 帧的子带谱熵定义为:

$$H_k = - \sum_{i=1}^n p_i \log_2(p_i) \quad (6)$$

从信息论角度上讲, 语音信号的谱熵低于噪声信号, 设定适当阈值能够有效区分语音段和静音段, 而且具有一定的稳健性。

1.4 统计模型

Sohn 等人用直接决策(Decision-directed, DD)方法优化的最大似然准则估计未知参数, 使用 Hang-over 方案对语音帧建立一阶隐马尔科夫模型。

假设语音帧可以分解成不相关的两类: $X = N$, $X = N + S$, 其中 S, N, X 分别代表 L 维的语音、噪声、含噪语音的离散傅里叶变换(DFT)系数的矢量集, 由文献[9]可知矢量集之间可以认为是互相独立的, H_0, H_1 概率密度函数为:

$$p(X | H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \quad (7)$$

$$p(X | H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi [\lambda_N(k) + \lambda_S(k)]} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\right\} \quad (8)$$

其中 $\lambda_N(k), \lambda_S(k)$ 分别代表 N, S 的第 k 维频谱分量的方差。第 k 维频谱分量的似然比为:

$$\Lambda_k = \frac{p(X_k | H_1)}{p(X_k | H_0)} = \frac{1}{1 + \xi_k} \exp\left\{\frac{\gamma_k \xi_k}{1 + \xi_k}\right\} \quad (9)$$

其中 $\xi_k = \lambda_S(k)/\lambda_N(k)$, $\gamma_k = |X_k|^2/\lambda_N(k)$ 分别为先验信噪比、后验信噪比。每一维频谱分量得:

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \begin{matrix} > \\ < \end{matrix} \eta \rightarrow X \in \begin{cases} H_1 \\ H_0 \end{cases} \quad (10)$$

对 ξ_k 做最大似然估计(ML)可得到下式:

$$\hat{\xi}_k^{(ML)} = \gamma_k - 1 \quad (11)$$

把(11)式代入(10)式得到:

$$\log \hat{\Lambda}^{(ML)} = \frac{1}{L} \sum_{k=0}^{L-1} \{\gamma_k - \log \gamma_k - 1\} \begin{matrix} > \\ < \end{matrix} \eta \rightarrow X \in \begin{cases} H_1 \\ H_0 \end{cases} \quad (12)$$

由于(12)式的左半部分恒大于零, 由文献[9]可知似然概率将偏向于 H_1 , 为了减少误差, 采用 DD 方法对先验信噪比估计:

$$\hat{\xi}_k^{(DD)}(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_N(k, n-1)} + (1 - \alpha) P[\gamma_k(n) - 1] \quad (13)$$

其中 $\hat{\xi}_k^{(DD)}(n)$ 为第 n 帧的先验信噪比, $\hat{A}_k(n-1)$ 为当前信号前一帧的幅度估计。

传统的 Hang-over 算法通常以适当提高误检率为代价来延迟 H_1 到 H_0 的状态转移, Sohn 等人把语音帧序列状态的转移建立一阶隐马尔可夫模型, 因为当前语音帧的状态依赖于当前帧及前一帧的状态。于是得到如下决策准则:

$$L(n) = \frac{p(X_n | q_n = H_1)}{p(X_n | q_n = H_0)} = \frac{P(H_0)}{P(H_1)} \frac{P(q_n = H_1 | X_n)}{P(q_n = H_0 | X_n)} \begin{matrix} > \\ < \end{matrix} \eta \rightarrow q_n \in \begin{cases} H_1 \\ H_0 \end{cases} \quad (14)$$

其中 $\chi_n = \{X(n), X(n-1), \dots, X(1)\}$ 代表从第一帧到当前帧的观值, q_n 表示第 n 帧的状态。后验概率比记为 $\Gamma(n) = P(q_n = H_1 | \chi_n) / P(q_n = H_0 | \chi_n)$, 最终的统计决策函数简化为:

$$L(n) = [P(H_0) / P(H_1)] \Gamma(n) \tag{15}$$

2 实验比较与分析

2.1 实验一

语音库采用 4 人录音, 2 男 2 女, 分别读取 50 个生活中常用的单词或短语, 语音采用 Wav 格式, 采样频率为 11025Hz, 量化比为 16bits, 并用 cooledit 人工标注每段语音的起点和终点, 作为参考依据。

语音帧长为 256(20ms), 帧移 128(10ms), 程序检测出起点或终点的误差在 200ms 范围内认为正确, 端点检测的正确率 = 正确的样本个数/总的样本数。VAD 运算的复杂度以运行时间为参考依据。实验分别在纯净和不同高斯白噪声环境下进行, 为了描述方便, 基于统计模型的 DD+Hang-over 算法记 Sohn, 实验结果如表 1。

表 1 不同环境下的端点检测结果比较

检测方法	纯净语音	15dB	10dB	5dB	0dB	-5dB	平均耗时
STE	100%	91%	68%	13%	8%	0	2.56s
ZCR	97%	92%	75%	68%	43%	21%	3.45s
BSE	99%	96%	78%	70%	62%	43%	4.62s
Sohn	100%	100%	95%	83%	80%	46%	11.91s

实验一表明, 在背景噪声较小时用 STE 较为有效, 而在背景噪声较大时抗噪声效果不如 ZCR、BSE, Sohn 相对于传统方法有较高的精确度, 特别是在强噪声环境下其抗噪能力明显高于其他方法。缺点是计算相对复杂, 耗时较长。传统的算法最关键的问题是阈值的设定, 阈值的自适应性值得进一步研究。

2.2 实验二

实验二在不同环境下对语音进行训练和测试, 分为纯净语音环境和含噪声语音环境两部分。主要验证不同 VAD 方法在相同环境下对汉语方言识别系统识别率的影响。

2.2.1 语音数据库

语音库包含 3 种有代表性的南方方言: 闽方言、吴方言、粤方言。语音采用 Wav 格式, 采样频率为 11025Hz, 量化比为 16bits。每种方言说话人为 20 人并按男女比例 1:1 选取。语音库为训练集、测试集和开发集。其中训练集约占整个语音时长的 1/2, 主要用于训练 GMM 语言模型; 测试集约占 1/6, 主要用于测试整个系统的性能; 开发集约占 1/3, 主要用于开发系统的性能。训练集中每种方言各有约 40 分钟的训练语料, 以上 3 种语音集互不交叉重叠。

2.2.2 实验系统说明

整个系统原理框图如图 1 所示, 包括训练阶段和测试阶段。训练阶段包括预处理、特征提取、训练各方言的 GMM。其中预处理包括预加重、分帧、加窗、语音活动检测等。本实验采用的预加重滤波器为 $1 - 0.95z^{-1}$, 帧长为 256, 帧移 128, 窗函数为 Hamming 窗。训练提取 SDC 特征^[13], 特征取 20 维。采用 Kmeans 算法对 GMM 初始化, 用 EM 迭代算法训练三种方言的 GMM, 最后利用最大似然准则, 测试语音的特征与训练好的模型分别进行概率打分, 概率最高的模型对应的方言类别即为语音方言类别。

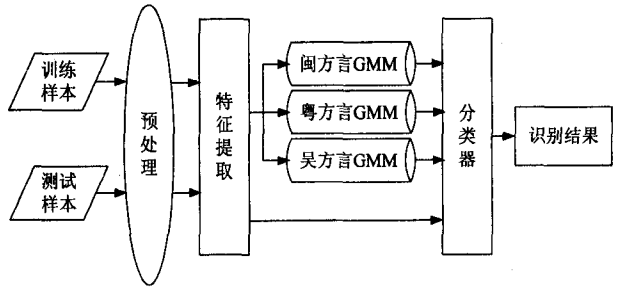


图 1 基于 GMM 的方言识别系统

2.2.3 不同环境下的系统识别结果

为了评估端点检测对系统性能的影响, 先后采用不同端点检测算法应用到不同训练环境中, 进行比较和分析。根据文献[13], 本实验固定 SDC 特征的参数为 10-1-4-2。

首先在纯净语音的环境下对闽、粤、吴 3 种方言分别训练 16, 32, 64 阶高斯混合模型, 各方言训练时间约为 45 分钟。由于实验验证的是预处理部分对整个系统的影响, 所以测试语音均采用 10s 的测试集。系统的识别率如图 2 所示, 可以看出随着 GMM 阶数的增加, 系统的识别率逐渐提高。因为阶数越大对数据的概率密度函数刻画的越详细。由图 2 可见, 相同阶数时采用 Sohn 平均识别最高。

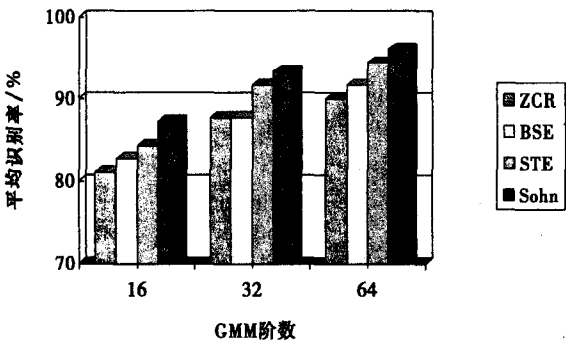


图 2 纯净语音环境下不同阶数的平均识别率

其次, 在噪声环境下训练 32 阶 GMM, 除了端点检测方法不同, 系统其他参数保持不变的前提下, 对 3 种训练语音引入加性高斯白噪声, 测试语音仍采用 10s 的测试集。由于在强噪声环境下系统的识别性能

不够理想,本实验采用信噪比为 10dB ~ 30dB。含不同信噪比的高斯白噪声环境下的识别结果如图 3 所示。可以看出采用 Sohn 方法,系统的平均识别率仍在 90% 左右,鲁棒性相对较强。

从实验二结果可以看出:在噪声环境下,信噪比越小,汉语方言辨识系统的识别率越差。STE 在高信噪比时效果比 ZCR、BSE 好,但抗噪能力不及其他方法。Sohn 的精确度高,鲁棒性强,在信噪比为 10dB 时,系统的识别率可达 89.45%。

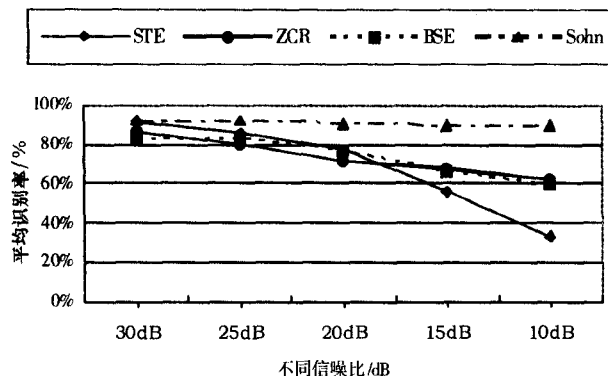


图 3 不同信噪比下系统平均识别率

3 结束语

汉语方言辨识系统中比较重要的两个问题是特征的提取和分类器的设计,但前端的预处理部分不容忽视,准确的端点检测对系统尤为重要,文中实验结果表明基于 DD+Hang-over 的算法比传统方法更精确,尤其在噪声环境下效果更显著。此外,噪声环境下汉语方言辨识系统的识别率明显低于纯净语音,如何在噪声的环境下提高系统性能有待于进一步的研究。

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

(上接第 72 页)

- [4] Levanti A, Giordano F, Tinnirello I. A CAPWAP Architecture for Automatic Frequency Planning in WLAN [C]//Proceedings of IEEE ISCC. [s. l.]: [s. n.], 2007.
- [5] Bernaschi M, Cacace F, Davoli A, et al. A CAPWAP-based solution for frequency planning in large scale networks [J]. Computer Communications, 2011, 34(11): 1283-1293.
- [6] Levanti A, Giordano F, Tinnirello I. A CAPWAP-compliant Solution for Radio Resource Management in Large-scale 802.11 WLAN [C]//Proceedings of IEEE GLOBECOM. [s. l.]: [s. n.], 2007.
- [7] Montemurro M, Stanley D, Calhoun P. CAPWAP Protocol Specification [S]. RFC 5415, 2009.
- [8] 向望, 王志伟, 高传善. 集中式 WLAN 体系结构通信协

和音节分割 [J]. 哈尔滨工程大学学报, 1997, 18(5): 91-95.

- [2] Segura J C, Benítez C, de la Torre A, et al. An Effective OSF-based VAD with Noise Suppression for Robust Speech Recognition [J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(6): 1119-1129.
- [3] 张仁志, 崔慧娟. 基于短时能量的语音端点检测算法研究 [J]. 电声技术, 2005(7): 52-54.
- [4] 韩纪庆, 张磊, 郑铁然. 语音信号处理 [M]. 北京: 清华大学出版社, 2004.
- [5] 张亚歌, 张太猛, 夏川. 一种基于能量聚类分析的句子语音端点检测法 [J]. 计算机技术与发展, 2008, 18(4): 13-15.
- [6] 李晔, 张仁智, 崔慧娟, 等. 低信噪比下基于谱熵的语音端点检测算法 [J]. 清华大学学报: 自然科学版, 2005, 45(10): 1397-1400.
- [7] 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.
- [8] 贾川, 张健, 陈振标, 等. 噪声环境下的端点检测算法研究 [C]//第六届全国人机语音通信学术会议论文集. 出版地不详: 出版者不详, 2001: 441-445.
- [9] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection [J]. IEEE Signal Processing Lett., 1999, 6(1): 1-3.
- [10] 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.
- [11] 董恩清, 赵鹤鸣, 周亚同, 等. 支持向量机在语音激活检测中的应用研究 [J]. 通信学报, 2003, 24(3): 70-77.
- [12] 沈兆勇, 顾明亮. 基于符号化和语言模型方法的汉语方言自动辨识 [J]. 徐州师范大学学报 (自然科学版), 2006, 24(2): 54-57.
- [13] Kohler M A, Kennedy M. Language identification using shifted delta cepstra [C]//Midwest Symposium on Circuits and Systems. [s. l.]: [s. n.], 2003.
- 议 [J]. 计算机工程, 2008, 34(22): 115-117.
- [9] Montemurro M, Stanley D, Calhoun P. Control and Provisioning of Wireless Access Points (CAPWAP) Protocol Binding for IEEE 802.11 [S]. RFC 5416, 2009.
- [10] 李和光. AC-AP 架构中 CAPWAP 协议的研究与开发 [D]. 南京: 东南大学, 2010.
- [11] Bernaschi M, Cacace F, Iannello G, et al. OpenCAPWAP: An open source CAPWAP implementation for the management and configuration of WiFi hot-spots [J]. Computer Networks, 2009, 53(2): 217-230.
- [12] 郭子明. 基于 CAPWAP 协议的无线接入点扩展的设计与实现 [D]. 西安: 西安电子科技大学, 2011.