基于 GMM的声音活动检测方法

陈奇川 蔡 骏 林 茜
(厦门大学计算机科学系 福建 厦门 361005)

摘 要 为了提高声音活动检测的鲁棒性,提出了一种基于 GMM模型的声音活动检测方法。此方法在频谱特征空间中建立背景噪音和语音的 GMM模型,然后采用模型匹配的方法对被测信号进行区分。此方法自适应更新 GMM模型的参数,使之可以适应环境的变化。实验结果显示该方法在噪音环境中比传统的声音活动检测方法具有更高的准确率。

关键词 声音活动检测 特征空间 GMM模型 参数自适应 语音识别

A VOICE ACTIVITY DETECTION METHOD BASED ON GMM

Chen Qichuan Cai Jun Lin Qian
(Department of Computer Science, Xiamen University, Xiamen 361005, Fujian, China)

Abstract To improve the robustness of the voice activity detection (VAD), a GMM-based approach for VAD has been proposed in this paper. With this method, two GMMs are constructed to model the noise and the speech respectively in spectrum feature space, and the signal frames to be detected are discriminated in the way of GMMs matching. This method is designed to self-adapt the GMM parameters updating to accommodate environmental variation. Experimental results show that the proposed method generally performs better in accuracy than traditional VAD approaches in noisy environments.

Keywords VAD Feature space GMMs Parameter self-adapting Speech recognition

0 引 言

随着多媒体通信和语音识别技术应用的深入发展,声音活动检测技术 VAD(voice activity detection)受到广泛的关注[1]。VAD技术经过多年发展,已经有很多成熟的算法投入到实际应用中,常用的算法有短时能量检测[2]、过零率检测[3]和高阶统计分析[4]等,近几年又发展出一些新的方法,如自适应能量检测[5]、相关性检测[6]、概率检测[7]和基于 HMM模型检测[8]等。传统的基于短时能量方法虽然在高信噪比的环境下具有良好的效果,且具有计算复杂度低、容易实现的优点,但是在低信噪比的环境下效果并不理想。如何设计一种在噪音环境中具有良好性能的 VAD方法是当前的一个研究热点[9]。

VAD的目的在于区分语音和环境噪音,之所以可以区分是因为它们具有各自不同的特征,如能量、自相关系数等。因此,如果能够建立模型来描述语音和环境噪音在特征空间中的分布,那么就可以用模型匹配的方法将它们区分开来。基于这样的一种思想,本文提出了一种在噪音环境中的 VAD方法——基于 GMM模型的声音活动检测方法。该方法假设语音和背景噪音在特定的特征空间中符合高斯混合分布,在特征空间中分别建立它们的 GMM模型,然后用模型匹配的方法在被测信号中检测出有效的语音段。

1 基于 GMM的声音活动检测方法

假设背景噪音(非语音)和语音在特定的特征空间中符合高斯混合分布,其 GMM模型分别为 H_0和 H_1,被测的输入信号帧为 D维的特征矢量 X_t,问题就是求解 $P(H_0|X_t)$ 与 $P(H_1|X_t)$ 的大小关系,根据贝叶斯法则:

$$P(H_0|X_t) = P(X_t|H_0)P(H_0)/P(X_t) \quad (1)$$

$$P(H_1|X_t) = P(X_t|H_1)P(H_1)/P(X_t) \quad (2)$$

其中,$P(H_0)$ 表示背景噪音出现的先验概率,$P(H_1)$ 表示语音出现的先验概率。由于分母相同,故公式(1)、(2)可简化为:

$$P(H_0|X_t) \propto P(X_t|H_0)P(H_0) \quad (3)$$

$$P(H_1|X_t) \propto P(X_t|H_1)P(H_1) \quad (4)$$

假设语音的 GMM模型由 M个 D维的高斯分量组成,可以用 M个高斯分量加权和的形式来表示:

$$P(X_t|H_1) = \sum_{i=1}^{M} W_i P(X_t|\lambda) \quad (5)$$

其中,W_i($i=1, 2, \cdots, M$)为混合权值,相当于第 i个高斯分量出现的概率,满足 $\sum_{i=1}^{M} W_i = 1$。$\lambda_i(\mu_i, \Sigma_i)$ 表示 GMM模型的第 i个高斯分量,μ_i为均值矢量,Σ_i为协方差矩阵。由于背景噪音可以认为是平稳的,所以,背景噪音的模型 H_0采用单高斯概率模型 $\lambda_0(\mu_0, \Sigma_0)$,即:

$$P(X_t|H_0) = P(X_t|\lambda_0) \quad (6)$$

高斯概率密度函数为:

收稿日期:2008-04-11。福建省自然科学基金项目(2006J0043)。
陈奇川,硕士生,主研领域:语音信号前端处理,语音识别。

$$P(X_t|\lambda_i)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(X_t-\mu_i)^T\Sigma_i^{-1}(X_t-\mu_i)\right\} \quad (7)$$

在实际应用中, 为了提高计算效率, 通常采用协方差矩阵的对角阵:

$$P(X_t|\lambda)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{D}\frac{(X_{tk}-\mu_{ik})^2}{\sigma_k^2}\right\} \quad (8)$$

式中 σ_k^2表示特征矢量第 k维的方差。

2 特征空间与特征提取

语音的频谱分布具有不均匀性的特点, 以汉语的数码语音频谱特征[10]为例: 元音的低频(0.1kHz~0.4kHz)和中频(0.64kHz~2.8kHz)能量较高; 浊辅音的低频能量较高, 中频能量较低; 清辅音的高频(3.5kHz以上)能量较高。而噪音相对稳定, 其频谱分布比较均匀。以下面的一段含噪语音的语谱图为例。

图 1显示语音在频谱上的分布是不均匀的, 背景噪音的频谱分布是比较均匀的。受此启发, 我们选择以频谱为特征空间。

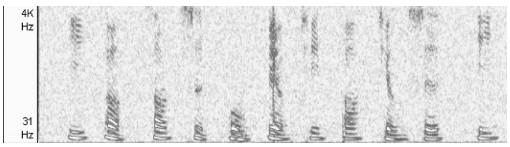

图 1　含噪语音的语谱图

如上所述, 可以把语音的频谱分为三段: 低频、中频和高频。假设语音在每个频带上的分布近似于高斯混合分布, 就可以用高斯混合模型来表示语音。由于背景噪音相对平稳, 只用单高斯概率模型即可。由于语音和噪音频谱均方差的差异, 在特征矢量中我们加入了频谱均方差, 由此, 我们构作的 4维的特征矢量包含如下分量: 低频能量, 中频能量, 高频能量, 频谱均方差, 记作 $[L, M, H, \Sigma]$。三个频带能量的提取可以采用三角滤波器组来得到, 其频谱特性如图 2所示。

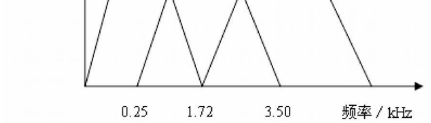

图 2　三角滤波器组

3 GMM模型的建立与应用

3.1 初始化噪音模型

一般认为, 录音信号段开始的 200ms是没有语音的, 所以可以用开始的 200ms的噪音来初始化噪音模型 $\lambda_0(\mu_0, \Sigma_0)$, 其中, μ_0是均值矢量, Σ_0是协方差对角阵, 计算公式如下:

$$\mu_0=\frac{1}{N}\sum_{t=1}^{N}X_t \quad (9)$$

$$\Sigma_{0k}=\frac{1}{N}\sum_{t=1}^{N}(X_{tk}-\mu_{0k})^2 \quad (10)$$

其中 N表示初始 200ms的特征矢量个数, k表示特征矢量的第 k维。

3.2 无噪语音模型

无噪语音模型可以从无噪语音中获得, 先采用 K-Means聚类算法从一定量的无噪语音中得到 GMM模型初始值, 然后再采用 EM算法调整 GMM模型的参数, 得到的每个高斯分量为 $\lambda(\mu, \Sigma)$, 相应的混合权重为 W_i。

3.3 含噪语音模型

在实际运用中, 由于噪音的存在, 被测的输入信号是夹杂了噪音的语音, 所以, 应该把上面的无噪语音模型和噪音模型进行混合, 得到含噪语音模型, 即:

$$\lambda_0+\lambda_i\rightarrow\overline{\lambda}_i \quad (11)$$

3.4 声音活动检测的判定规则

根据公式(3)、(4), 在计算 $P(H_0|X_t)$与 $P(H_1|X_t)$时, 首先应知道噪音和语音的先验概率 $P(H_0)$和 $P(H_1)$, 由于语音信号具有前后连续性的特点, 假设当前输入帧 X_t是语音或噪音的先验概率只和前帧有关, 令前帧 X_{t-1}是语音的概率是 $P_{1(t-1)}$, 那么, 当前帧是语音的先验概率 $P(H_{1(t)})\approx P_{1(t-1)}$, 同理, $P(H_{0(t)})\approx P_{0(t-1)}$, 则:

$$P(H_0|X)\approx P(X_t|H_0)P_{0(t-1)} \quad (12)$$

$$P(H_1|X)\approx P(X_t|H_1)P_{1(t-1)} \quad (13)$$

经归一化处理, $P_{0(t)}$和 $P_{1(t)}$的算式如下:

$$P_{0(t)}=P(H_0|X)/[P(H_0|X)+P(H_1|X)] \quad (14)$$

$$P_{1(t)}=P(H_1|X)/[P(H_0|X)+P(H_1|X)] \quad (15)$$

这里, $P_{0(t)}$和 $P_{1(t)}$满足 $P_{0(t)}+P_{1(t)}=1$。

假设平滑系数 a, 经平滑处理后的概率计算公式如下:

$$P_{0(t)}=aP_{0(t-1)}+(1-a)P_{0(t)} \quad (16)$$

$$P_{1(t)}=aP_{1(t-1)}+(1-a)P_{1(t)} \quad (17)$$

根据实验观察, a的数值不宜取得太大, 否则会出现语音起点和落点判断滞后的问题, 一般取值为 $0.3<a<0.5$效果比较理想。

在经过平滑处理后, 则可根据 $P_{1(t)}$的值来判定这帧信号是语音还是噪音: 当 $P_{1(t)}>\theta$时, 判定为语音; 否则, 判定为噪音, 其中 θ为判定阈值, 一般取值为 0.5。

3.5 模型参数的自适应更新

GMM模型参数重估的常用方法是 EM算法, EM算法的思想是使密度函数最大化来确定参数的值, 即所谓极大似然估计。我们借鉴了 EM算法的思想, 使高斯模型参数的更新步长与当前输入帧在该高斯模型下的概率密度成正比, 从而得到含噪语音及背景噪音的高斯模型参数的更新步长的推导公式:

$$S_{i(t)}=\beta P_{1(t)}\frac{W_iP(X_t|\overline{\lambda}_i)}{\sum_{k=1}^{M}W_kP(X_t|\overline{\lambda}_k)} \quad (18)$$

$$S_{0(t)}=\beta P_{0(t)} \quad (19)$$

上面两式中的 β表示更新速率系数, 表示高斯模型参数的更新快慢, 实验证明, β的取值在 $0.075<\beta<0.15$范围比较合适。最后, 得到高斯参数的更新公式如下:

$$\mu_{i(t)}=(1-S_{i(t)})\mu_{i(t-1)}+S_{i(t)}X_t \quad (20)$$

$$\Sigma_{ik(t)}=(1-S_{i(t)})\Sigma_{ik(t-1)}+S_{i(t)}(X_k-u_{ik(t-1)})^2 \quad (21)$$

3.6 基于 GMM模型声音活动检测的实现

根据以上关于模型和参数自适应更新过程的描述, 可以得到基于 GMM模型的声音活动检测方法的流程图, 如图 3所示。

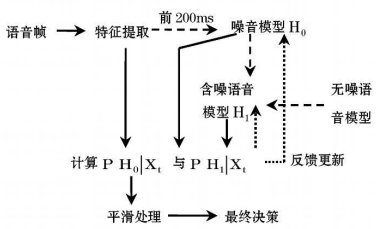

图 3 基于 GMM模型的活动音检测方法流程图

我们按照图 3的流程, 采用 4维特征矢量 $\{L, M, H, \Sigma\}$、8高斯分量的 GMM模型对语音进行建模, 模型参数更新速率系数 $\beta = 0.1$, 平滑系数 $a = 0.4$, 判定阈值 $\theta = 0.5$。下面示例中的语音数据来源于男性口音, 采用 8kHz采样, 16Bit量化, 帧长为 20ms,帧移为 10ms,采样后的语音数据与白噪音混合。下图是信噪比为 6db的测试数据的波形及检测结果。

图 4中的方框是采用本文所述方法检测出来的语音, 图 5中的曲线是计算出的语音概率值 $P_{1(t)}$, 两图的横坐标尺度一致且对齐。

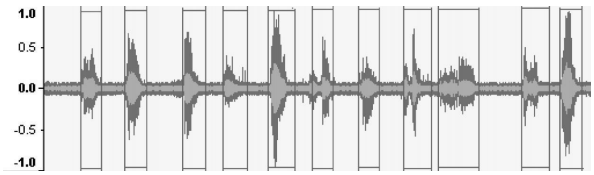

图 4 含噪语音波形及检测结果

图 5 计算输出的语音概率值 $P_{1(t)}$

4 实验结果

我们采用美式英语语料库 TIMIT[11]进行测试。TIMIT的测试语料集共有 3696个录音句子, 由 462个说话人每人 8句构成, 采用 16kHz采样, 16Bit量化。我们通过把原始语音数据与不同电平的白噪音混合得到不同信噪比的测试数据。

在实验中我们共采用了三种方法进行对比, 分别是: 短时能量, 短时过零率, 和本文提出的基于 GMM模型的声音活动检测方法。在本文方法的实验中, 仍然采用 4维特征矢量 $\{L, M, H, \Sigma\}$、8高斯分量的 GMM模型, 模型参数更新速率系数 $\beta = 0.1$, 平滑系数 $a = 0.4$, 判定阈值 $\theta = 0.5$, 帧长为 20ms,帧移为 10ms。实验在不同信噪比测试数据下测试, 实验结果通过与手工切分的语音端点进行比较得出最终的准确率。

图 6是 TIMIT语料库中一段原始语音波形, 图 7是与之对应的含噪语音波形及采用基于 GMM模型声音活动检测方法的检测结果, 其中方框表示检测出的语音。

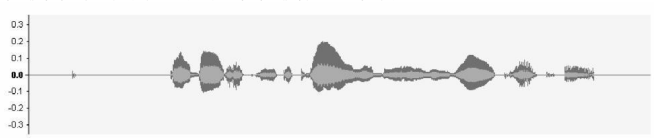

图 6 原始语音波形

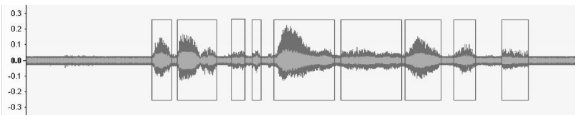

图 7 含噪语音波形及检测结果

表 1是三种声音活动检测方法在不同的信噪比测试数据下的最终检测结果。

表 1 活动音检测结果(准确率)

方法	15db 白噪音	10db 白噪音	5db 白噪音	0db 白噪音
短时能量	98%	91%	79%	66%
过零率	98%	92%	81%	68%
GMM	98%	95%	86%	76%

表 1显示: 在不同信噪比的测试数据下, 基于 GMM的声音活动检测方法比基于短时能量和过零率方法具有更高的准确率; 此外, 随着信噪比的下降, 短时能量和过零率方法的准确率均有较大幅度的下降, 而基于 GMM的声音活动检测方法的准确率下降幅度要明显小于前两种方法, 说明此方法具有更好的鲁棒性。

5 总 结

实验结果表明, 基于 GMM模型的活动音检测方法在噪音环境中具有较好的准确率和鲁棒性。同时采用自适应更新模型参数的方法, 使之可以适应不同的语音及环境噪音, 在嘈杂的环境中具有较好的使用价值。在后续的研究中我们将会研究提高高斯分量的个数及特征矢量的维数, 以进一步提高该方法在极低的信噪比环境下的准确率。由于此方法的构建过程具有通用性, 语音与噪音的特征空间选择并不局限于频谱, 因此, 我们也将尝试寻找其它的特征空间。

参 考 文 献

[1] Hersent O, Petit J P, Gurle D. Beyond VoIP Protocols: Understanding Voice Technology and Networking Techniques for IP Telephony[M]. Chichester: John Wiley & Sons Ltd, 2005: 91-100.

[2] Dong E, Liu G, Zhou Y, et al. Voice activity detection based on short-time energy and noise spectrum adaptation[J]. In: Proc of the 6th International Conference on Signal Processing, 2002, 1: 464-467.

[3] Sangwan A, Chiranth M C, Jamadagni H S, et al. VAD techniques for real-time speech transmission on the Internet[J]. In: Proc of the 5th IEEE International Conference on High Speed Networks and Multimedia Communications, 2002: 46-50.

[4] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(3): 217-231.

[5] Venkatesha P R, Sangwan A, Jamadagni H S, et al. Comparison of voice activity detection algorithms for VoIP[J]. In: Proc of the 7th International Symposium on Computers and Communications, 2002: 530-535.

[6] Sangwan A, Jamadagni H S, Chiranth M C, et al. Second and third order adaptable threshold for VAD in VoIP[J]. In: Proc of the 6th International Conference on Signal Processing, 2002, 2: 1693-1696.

(下转第 75页)

台是 CUP为奔腾 2 8GHZ, 内存为 512M的安装有 matlab和 VC 6 0的计算机。下面是采用该方法检测阴影的实验数据及结果。

图 3为阳光强度中等的室外行人, 图 5为强阳光下公路上的行车, 四个目标样本都在地面上产生了阴影。先用高斯背景估计估计背景, 用背景差检测运动目标区域, 包括人体区域和阴影区域。然后统计每个目标区域的像素亮度下降比率直方图及累积亮度下降比率直方图。根据公式(9—12)计算伽玛分布的形状参数和尺度参数, 并根据公式(7)求伽玛分布, 根据 2 4节所示的算法求出阴影亮度下降比率分布区, 根据公式(13)分割出候选阴影。最后用规一化 rgb彩色模型及连通域分析检测出真实阴影。

图 3　室外行人

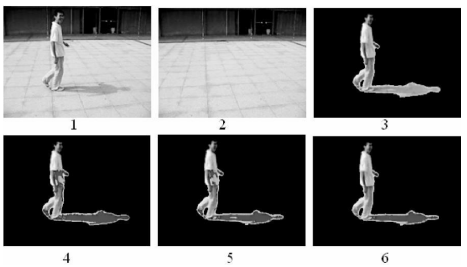

图 4　室外行人的阴影检测结果, 1是原始视频, 2是高斯估计的背景, 3是检测的带有阴影的目标, 4 5分别是 Gauss分布模型和伽玛分布模型检测的阴影, 6是 5处理后的结果

由图 4可看出在我们的参数模型下几乎把所有的投影阴影都包含在内, 由于目标亮度下降比率与投影阴影亮度下降比率具有很大的相似性, 所以用此模型检测的结果会包含少量的自阴影。但本模型可以把候选阴影缩小在一个比较小的范围, 后期借助基于规一化 rgb纹理一致性等的阴影检测方法可以得到很好的真实阴影。由图 4的 4 5也是看出在伽玛分布假设下检测的结果略优于高斯分布假设下检测的结果。

图 5和图 6是强阳光下的公路行车阴影检测实验结果。

图 5　公路行车

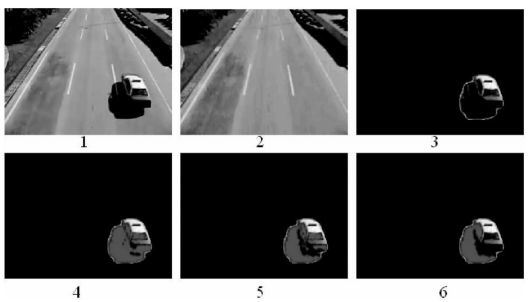

图 6　公路行车的阴影检测结果

通过多组实验验证了本文所提的基于亮度下降比率直方图的参数估计在阴影检测中的有效性和鲁棒性, 它在各种光照强度下都可以比较准确地自适应估计出阴影的亮度比率范围, 可较准确地确定候选阴影区域。

4　结束语

本文分析了运动阴影与背景的光照变化具有强相关性, 而目标与背景之间光照变化具有弱相关性的特点, 利用多个目标的亮度下降比率直方图进行统计分析, 并用伽玛分布拟合累积亮度下降比率直方图, 得到在该光照模型下的投影区域亮度下降比率的统计模型参数。实验结果表明本文方法得到的统计模型参数比较稳定, 可以鲁棒地检测出阴影区域, 同时可以较好地保持目标原有的特征。

参　考　文　献

[1] Cucchiara R, Grana C, Piccardi M, et al. Improving shadow suppression in moving object detection with HSV color information[J]. IEEE Proc ITSC'01, Oakland, USA, 2001: 334-339.

[2] Tattersall S, Dawson-Howe K. Adaptive shadow identification through automatic parameter estimation in video sequences[J]. Proc Irish Machine Vision and Image Processing Conference, Coleraine, Ireland, 2003: 57-64.

[3] Salvador E, Cavallaro A, Ebrahimi T. Shadow identification and classification using invariant color models[J]. Proc of IEEE Proceedings IEEE ICASSP'01, 2001(3): 1545-1548.

[4] Jun-Wei Hsieh, Shih-Hao Yu, Yung-Sheng Chen, Wen-Fong Hu. A Shadow Elimination Method for Vehicle Analysis[J]. Proceedings of the Pattern Recognition, 17th International Conference on ICPR, 2004, 4 (04): 372-375.

[5] Pang C C C, Lam W W L, Yung N H C. A novel method for resolving vehicle occlusion in a monocular traffic-image sequence[J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(3): 129-141.

[6] Fatih Porikli, Jay Thornton. Shadow Flow: A Recursive Method to Learn Moving Cast Shadows[J]. Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005, 1(1): 891-898.

[7] Stauder J, Mech R, Ostermann J. Detection of moving cast shadows for object segmentation[J]. IEEE Transactions on Multimedia, 1999, 1 (1): 65-76.

[8] Baisheng Chen and Duansheng Chen. Shadow Detection Based on rgb Color Model[J]. Intelligent Computing in Signal Processing and Pattern Recognition, ICIC2006, LNCIS 345, 2006: 1068-1074.

(上接第 62页)

[7] Joon-Hyuk Chang, Nam Soo Kim. Voice activity detection based on complex Laplacian model[J]. Electronics Letters, 2003, 39(7): 632-634.

[8] Othman H, Aboulnasr T. A semi-continuous state transition probability HMM-based voice activity detection[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2007(1): 2-2.

[9] Kojoe Agyei-Kodie. Development of Voiced Activity Detection (VAD) Algorithm that is Robust at Low Signal-to-Noise Ratios[J]. December 2003.

[10] 李虎生. 汉语数码串语音识别及说话人自适应[D]. 北京: 清华大学电子工程系, 2000.

[11] TIMIT Acoustic-Phonetic Continuous Speech Corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.