DOI:CNKI:50-1165/N.20120704.1115.013

智能保健监测系统中音频信号的分类算法研究*

李玲俐
(广东司法警官职业学院 信息管理系,广州 510520)

摘要:针对 Mel 频率倒谱系数(MFCCs)信息在区分音频信号时的局限性,提出一种基于不同特征提取技术的两级分类策略,对智能保健监测系统的 9 种音频信号进行分类。分类的第一级采用 MFCCs 及其变化率(ΔMFCCs)作为隐马尔可夫模型(HMM)的输入。在第二级,将不同频段的功率谱密度的一阶差分均值和标准差作为分类的特征。实验结果表明,功率谱密度的一阶差分包含了 MFCCs 所不含有的重要分类信息,该方法使得实时保健监测系统的平均分类准确度高达 97.37%,具有较好的鲁棒性和分类准确性。

关键词:音频信号;Mel 频率倒谱系数;特征提取;隐马尔可夫模型;分类
中图分类号:TP391.42 **文献标志码**:A **文章编号**:1672-6693(2012)04-0073-04

智能保健监测系统主要用于家庭以照顾和监测高龄人群,具有成本低、复杂度小、高效并且能够充分保护隐私等优点。在智能保健监测系统中,音频信号的特征提取和分类决定了系统的性能,是研究者关注的重点。然而,现有的音频分类文献中,对非语音信号的分类方法研究较少,而是借用成熟的语音识别系统[1-4]来进行分类。例如,文献[1,4]采用 3 种经典的特征提取技术,包括 Mel 频率倒谱系数(Mel frequency cepstrum coefficient,MFCCs)、连续小波变换(Continuous wavelet transform,CWT)和短时傅立叶变换(Short-time Fourier transform ,STFT)来提取环境声音的特征,然后根据提取的特征,使用动态时间归整(Dynamic time wrapping,DTW)和学习矢量量化(Learning vector quantization,LVQ)两种分类技术来实现分类。实验仿真结果表明,MFCCs 和 DTW 相结合能达到最佳的分类性能。对于咳嗽声信号的识别[2],Samantha J. Barry 等人将提取的特征谱系数作为输入,采用概率神经网络(Probabilistic neural network,PNN)分类器进行分类,获得了良好的效果。也有研究人员提出应该先识别语音信号。例如,Abu-EI-Quran 等人[5]提出先根据音调比率参数区分语音和声学声音,对于非语音音频片段的识别,采用带 MFCC 和 ΔMFCC 特征的时延神经网络进一步进行分类。

由文献[1-5]可知,MFCCs 特征包含了大量可用于声音分类的基本信息,然而,要从一个复杂的随机过程(如咳嗽、语音等)中提取所有可能的信息并进行分类,仍缺少系统的分类方法。本文提出一种算法简单、易于实现的两级分类策略来识别智能保健监测系统中 9 种感兴趣的声音,并对实际环境中采集的音频信号来进行分类。

1 实验系统

智能音频监测系统包括以下信号处理步骤:音频信号的采集与检测、预处理、特征提取和分类,如图 1 所示。

1.1 数据预处理

文献[6]提出音频数据预处理包括去除采样的音频信号的噪音,和将分段后的声音信号加载到计算机中两个步骤。本文首先阐述去除噪音和音频信号分段的过程,然后进行特征提取和分类。实时应用系统中,将语音、咳嗽、清嗓、杯碟碰撞声、口哨声、开门和关门、电话铃声、沸水声和冲厕水 9 种声音信号作为测试样本。所有音频的采样频率均为 8 kHz。

1.1.1 去除噪音 通常情况下,采样音频信号的低频噪音犹如直流浮动或扰动,如图 2 所示。因此,可以先用高通滤波器来消除这些低频噪音。文献[6]中提到,设计的高通滤波器转换函数为 $H(z)=1-k*z^{-1}$,其中参数 k 设置为 0.937 5。从图 2 可以看

* **收稿日期**:2012-03-23 **网络出版时间**:2012-07-04 11:15:00
资助项目:广东省自然科学基金(No.101754539192000000)
作者简介:李玲俐,女,讲师,硕士,研究方向为数据挖掘与模式识别。
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20120704.1115.201204.73_013.html

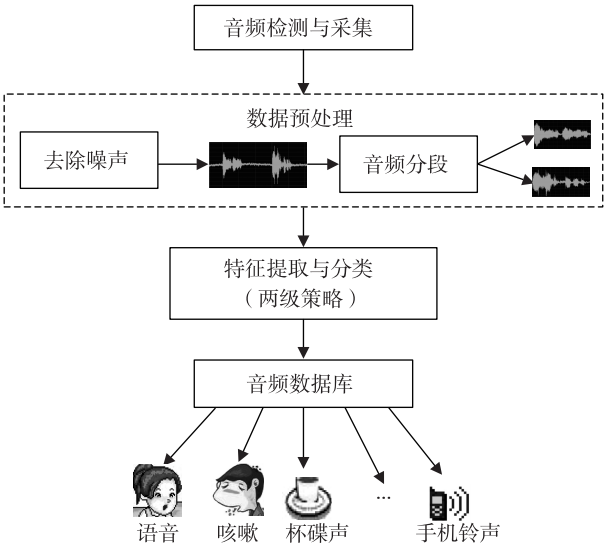
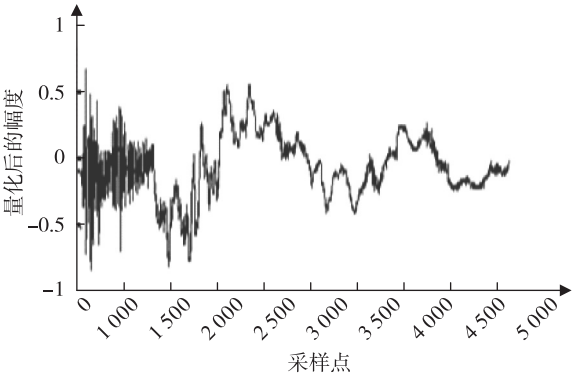
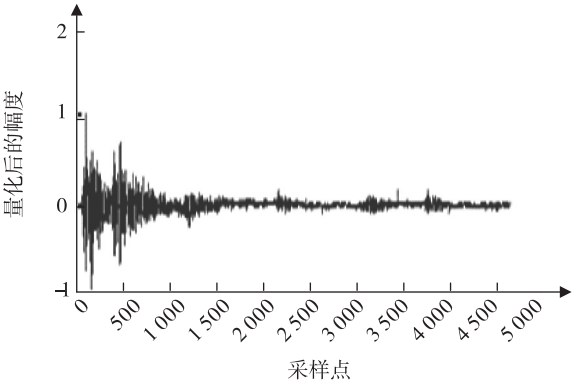


图1　智能音频监测系统中信号的处理步骤

出,采样的原始波形在低频部分浮动。信号经过高通滤波器后,不但调整了浮动的信号,而且处理后的音频信号在低音情况下听起来更清晰。



(a)原始采样数据



(b)高通滤波后的数据

图2　调整声波信号的直流浮动

1.1.2 音频分段　通过采样,每个声音片段包括目标声音片段和一些接近静音的背景声音。通过对声音分段来获得目标声音片段[3]。声音的分段通过分析音频信号的能量和过零率来实现。声音的能量可以通过公式(1)计算。

$$E(t)=(1-\alpha)E(t-1)+\alpha x(t)^2,\alpha\in[0,1) \quad (1)$$

其中,$E(t)$表示在t时刻信号$x(t)$的能量,α是调节的参数,本文设置其值为0.99。如果$\max(E(T))$表示一个信号的最大能量,T_1用作一个尺度参数,那么每个单一的目标声音的能量阈值可以设置为$T_1 * \max(E(T))$。为了获得过零率,需统计信号的变化情况,并计算两个相邻采样间的距离。当两个相邻采样的信号正负情况互不相同,同时两者之间的距离超过阈值T_2时,统计该次过零率。对于语音信号,T_2是经验设定值。对于本文的音频信号,T_2为自适应阈值。目标声音片段的确定不但取决于能量,同时也取决于过零率,过零率的阈值设置为T_3。

1.2 特征提取

音频信号的分类中,MFCCs、STFT和DWT是非常经典的特征提取技术[2-3,7]。文献[7]对这3项技术进行了比较,并指出MFCCs能实现更好分类效果。因此,本文将使用MFCCs作为特征提取技术之一。然而,当所研究的对象包含更多音频信号时,一些能将某些声音进行分类的重要信息并没有被包含在MFCCs所提取的特征中。因此,本文提出通过提取更多的非MFCCs特征的思想,从而更好地完成分类过程。

1.2.1 Mel频率倒谱系数(MFCCs)　MFCCs计算包含两个过程。首先将信号分离为帧,对每一帧做加窗做傅立叶变换,然后将Mel频率滤波器组应用到每个加窗帧,再做傅里叶逆变换,将所获得的倒谱系数转换成单一的特征向量。本文实验将信号分隔帧长为256、步长为84的片段,其重叠长度是172。为了更好地进行比较,本文对MFCCs中不同的特征及其变化(ΔMFCCs)做了实验。

1.2.2 功率谱密度(PSD)函数　功率谱密度(Power spectral density,PSD)函数显示了不同频率下功率的强弱。通过计算PSD,可以知道哪些信号频率的功率强、哪些频率的功率弱。PSD可以通过计算自相关函数,然后做傅里叶变换得到。对于一个给定的观察信号$x(t)$,其PSD函数$p(w)$如公式(2)所示。

$$p(w)=\int_{-\infty}^{+\infty}E\{x(\tau)*x(\tau+t)\}e^{-j2\pi wt}dt \quad (2)$$

其中$E\{x(t)x(t+\tau)\}$为信号$x(t)$的自相关函数。

如果在一个特定的频率范围内对PSD函数进行积分运算,可以得到该频率的功率。随着频率变化,功率也将发生变化,其变化关系$\Delta p(w,t)$如公式(3)所示。

$$\Delta p(w,t)=p(w,t)-p(w,t-1),\ t=1,2,\cdots,T \tag{3}$$

其中,$p(w,t)$表示 t 时刻信号的短时 PSD,$p(w,t-1)$表示 $t-1$ 时刻的短时 PSD,T 是信号的持续时间。对于一些音频信号,功率的分布情况会显示一些重要的特征,如过频的最大功率、共振频率等。此外,通过计算功率密度的一阶或二阶差分,功率变化的特征将更加明显。这些特征将在策略的第二级中被用来分类语音、沸水声和冲厕水声等。

1.3 两级分类策略

本文提出的方法是通过建立一个两级分类框架,在不同级别,用不同的特征提取和分类方法对不同的声音进行分类。第一级主要是提取 MFCCs 和 ΔMFCCs 特征,并将隐马尔可夫模型(Hidden Marcov model,HMM)用作分类器。HMM 通过对大量音频数据进行数据统计,建立识别统计模型,然后从待识别音频中提取特征,与这些模型匹配,通过比较匹配概率以获得分类结果和稳健的统计模型,能够适应实际音频中的各种突发情况[8]。借助于 MFCCs 和 ΔMFCCs,音频信号,如电话铃声、开门和关门、口哨声、语音和沸水声等目标类都很容易被识别。而其他如咳嗽,清嗓的音频信号往往被错误地和语音归为一类,冲厕水声则容易与沸水声相互混淆。因此仅仅依靠 MFCCs 和 ΔMFCCs 特征是很难区分这些信号的,本文提出在第二级策略中通过各信号独有的特征来区分。由于在不同的频段,不同的声音表现出的不同特征,例如,咳嗽、清嗓和语音在第一级分类的基础上,它们在不同频段的一阶差分的平均值和标准偏差将很大不同,因此可以通过计算各信号 PSD 函数的一阶差分(公式(3))来对各种信号进行识别。

2 实验设置与结果

本实验共采集了 1 186 个声音样本,其中包含 211 个语音采样,89 个咳嗽采样,52 个清嗓采样,67 个杯盘碰撞声采样,217 个口哨声采样,147 个开门和关门声音采样,55 个电话铃声采样,183 个沸水声采样,165 个冲厕水声采样。抽取 2/3 的采样作为训练集,1/3 的数据作为测试集。本文将用不同的特征方法和分类器相结合来比较它们的性能。

先对支持向量机(Support vector machine,SVM)与 HMM 进行比较。对于 SVM,MFCCs 特征的均值和标准差作为输入属性。当使用 SVM 作为分类器,对于不同的特征数,使用 24 个 MFCCs 特征获得的平均分类准确度与使用 36 个 MFCCs 特征获得的平均分类准确度几乎相同,但比使用 12 个 MFCCs 特征获得的平均分类准确度高一点。当使用 HMM 用作分类器,设置每个阶段的两个状态和两个概率密度函数。表 1 所示的仿真实验和实时测试结果表明,HMM 的性能比使用 SVM 的更好一点。在 HMM 分类模型中,变换概率矩阵被初始化为[0.5, 0.5, 0, 1]。

表 1　不同特征组合的不同分类技术的分类结果　%

方法	仿真系统下的平均分类准确度	实时环境下的平均分类准确度
MFCCs+SVM	95.21	75.00
MFCCs+HMM	95.30	75.00
Abu's method[5]	非常低	—
本文的方法	95.80	90.37

正如表 1 所示,仿真试验中,虽然这些方法的平均准确度相差不大,但在实时系统测试中,前两种方法的分类性能迅速降低。像咳嗽这样的一些音频信号往往被错误地认为是语音,清嗓声往往被错划为咳嗽。这主要是由于 MFCCs 的低鲁棒性而导致对一些声音,如咳嗽、清嗓的声音分类不够准确。本文提出的方法将弥补这一缺陷。

文献[5]中,Abu-EI-Quran 等人建议首先从将语音与其他非语音类声音区分开来,然后再对其他声音进行分类。本实验也对这个方法进行了测试。除了 Abu-EI-Quran 的基音检测技术[5],本文还对其他两个音高检测技术,即信号的逆频转换(Signal inverse frequency transformation, SIFT)算法和自相关的检测算法[6]进行比较,发现语音的音高范围非常接近咳嗽、清嗓的声音,仅仅通过基音检测方法[6]要完全区分语音与其他声音是比较困难的。本文实验也表明该分类方法[5]的分类准确度是非常低的。

两级分类策略中,在第一级,MFCCs 可以先分类出大部分声音。在第二级,由于不同频段的 PSD 的一阶差分不同,以此来区分其他声音。如表 2 所示,当只使用 MFCCs 特征,可以很准确地区分诸如电话铃声、开门和关门、口哨声、语音和沸水声。因此,在第一级,这 5 种音频信号被选为 5 个不同的目标类。由于在这一级分类中,咳嗽、清嗓声有时错误地认为是语音,冲厕水有时错误地认为是沸水声,于是将咳嗽、清嗓声归为语音类一组,冲厕水与沸水声归为一组。在第二级,需要将咳嗽、清嗓声与语音区分开来,将沸水声和冲厕水声也区分开来。从表 2 的分类结果所示,在第一级,仿真和实时实验中,使用 MFCCs 和 HMM 分类,语音、电话铃声、口哨声、沸水声均能达到很好的识别率。在第二级,第一级

容易混淆的声音也被区分开来,表明其具有良好的分类准确性。

表 2 仿真与实时环境下二级策略分类法的平均分类准确度 %

方法	采样	仿真系统下的平均分类准确度	实时环境下的平均分类准确度
第一级	语音	99.21	95.20
	杯碟碰撞声	91.49	90.00
	口哨声	100	98.60
	开门和关门	96.17	95.74
	电话铃声	100	100
	沸水声	100	96.60
第二级	咳嗽	97.02	95.80
	清嗓	95.32	94.89
	冲厕水	94.89	91.49

从表 1 和表 2 的仿真和实时实验结果表明,PSD 的一阶差分包含一些 MFCCs 所没有包含的重要分类信息。在不同的频段,PSD 的一阶差分的平均值和标准差功率下降,表明两级的分类框架,在实际应用中具有较强的鲁棒性。

3 结束语

本文对智能医疗保健监测系统中的音频信号分类技术展开研究。首先分析讨论了一些常用的特征提取和分类技术,鉴于常用的方法在实际应用中鲁棒性差而难于提供稳定的性能,提出一种音频信号的两级分类策略。实时环境中进行的测试结果表明,该智能保健监测系统采用的两级分类方法具有结构简单、高分类精度和鲁棒性强的特点,适合于实际应用。

参考文献:

[1] Cowling M. Non-speech environmental sound classification systems for autonomous surveillance[D]. Brisbane, Queensland: Griffith University, 2004 .

[2] Barry S J, Dane A D, Morice A H, et al. The automatic recognition and counting of cough[J]. Cough (London, England) , 2006, 2(9): 8-14.

[3] Kimber D, Wilcox L. Acoustic segmentation for audio browersers[C]//Proceedings of Interface Conf, Australia: IEEE_CS, 1996.

[4] Cowling M, Sitte R. Comparison of techniques for environmental sound recognition [J]. Pattern Recognition Letters, 2003, 24(15): 2895-2907 .

[5] Abu-ei-quran A R, Goubran R A, Chan A D C. Security monitoring using microphone arrays and audio classification[J]. IEEE Trans Instrumental and Measurements, 2006, 55(4): 1025-1032.

[6] Jang R. Audio signal processing and recognition [J/OL]//(2012-05-01). http://neural. cs. nthu. edu. tw/jang/ books/audio Signal Processing.

[7] Zhou N, Ser W, Yu Z L, et al. Enhanced class-dependent classification of audio signals[C]// 2009 WRI World Congress on Computer Science and Information Engineering, LOS Angeles, California USA: IEEE _ CS, 2009, 7: 100-104.

[8] 于晓明,柏松. 基于前向-后向 HMM 的连续语音识别系统的研究[J]. 计算机工程与应用, 2009, 30(18): 4339-4341.

Audio Signal Classification Algorithm for a Smart Health-Care Monitoring System

LI Ling-li

(Dept. of Information Management, Guangdong Justice Police Vocational College, Guangzhou 510520, China)

Abstract: Aiming at the deficiency of Mel-frequency cepstral coefficients (MFCCs) in discriminating acoustic signals, a two-level classification strategy based on different feature extraction techniques was proposed to classify nine audio signals in a smart health-care monitoring system. In the first level, the MFCCs and its variants (ΔMFCCs) are used as the inputs of the hidden Markov model (HMM) for classification. Then, the mean and standard deviation of the first-order difference of power spectral density over different frequency bands are calculated as features for further classification in the second step. Experiment results in real-time health monitoring system reveal that the first-order derivatives of power spectral density contain some important information which is not included in MFCCs. The approach in this paper shows better robustness and high classification accuracy whose average is as high as 97.37%.

Key words: audio signal; MFCC; feature selection; hidden Markov models (HMM); classification

(责任编辑 游中胜)