第30卷 第18期 计 算 机 工 程 2004年9月
Vol.30 №18 Computer Engineering September 2004
·人工智能及识别技术· 文章编号：1000—3428(2004)18—0133—02 文献标识码：A 中图分类号：TP391.4

基于高斯混合模型的乐器识别方法

张 奇，苏鸿根
(中国科学院研究生院，北京 100039)

摘 要：传统的乐器识别方法采用的是树型分类方法，这种方法分类过程比较繁琐，而且精度不高。该文把话者识别的方法应用到乐器识别之中，采用模式识别的方法实现对乐器的识别。采用MFCC系数和它的一阶导数作为音品的声学特征，分别对6种管弦乐器建立高斯混合模型。在识别过程中，首先假设各乐器的先验概率相同，根据高斯混合模型得出的后验概率确定待识别乐器所属的种类。实验表明这种识别方法十分有效，取得了较高的识别精度。

关键词：高斯混合模型；乐器识别；话者识别

A Novel Method for Instrument Recognition Based on Gauss Mixture Model

ZHANG Qi, SU Honggen
(Graduate School of Chinese Academy of Sciences, Beijing 100039)

【Abstract】 The traditional instrument recognition method adopts binary-tree classifying method. The process of this method is trivial and inaccurate. This paper applies speaker recognition methods into instrument recognition. The pattern recognition is utilized to implement instrument recognition. The MFCC coefficient and its derivative are taking as the acoustic features. A GMM model is constructed for each instrument set. In the process of recognition, the prior probability is supposed to be the same, the posterior probability is calculated according to GMM, and then the instrument class is determined. The experiment shows that this method is quite efficient and has better precision.

【Key words】 Gauss mixture model; Instrument recognition; Speaker recognition

乐器识别是音频检索的一个重要领域，它既涉及声源的声学属性，也涉及到人耳对音频的感知心理，是深入研究音频检索的基础。乐器识别在科学研究和实际应用中有重要意义。目前人们对于音频有不同的认识，还没有一种成熟的理论用于研究人类如何识别声源。然而声源识别在实际应用中有着重要的意义，例如可以用计算机标注多媒体数据或者转录音乐的演奏信号，形成用于编码或者理论研究的乐谱。在声源识别的基础之上，可以进一步发展理论和模型理解音乐的语义行为。

以往的乐器识别方法大多采用树形分类方法[4,5]，这种方法容易造成错误率的积累，识别的精度不高。乐器识别和话者识别十分相似，乐器识别主要研究乐器的音频特性，而话者识别[1]研究的对象是不同话者的发音特征。尽管两者之间存在差异，但它们要解决的都是声源识别的问题。本文把话者识别的方法应用到乐器识别中，采用模式识别的方法来实现乐器的分类。

1 MFCC声学特征及其提取

MFCC系数是一个感知特征参数[1]。它不同于LPC等通过对人的发声机理的研究而得到的声学特征，Mel倒谱系数MFCC是受人的听觉系统研究成果推动而导出的声学特征。对人的听觉机理的研究发现，当两个频率相近的音调同时发出时，人只能听到一个音调。临界带宽指的就是这样一种令人的主观感觉发生突变的带宽边界，当两个音调的频率差小于临界带宽时，人就会把两个音调听成一个，称之为屏蔽效应。Mel刻度是对这一临界带宽的度量方法之一。大量的实验表明，MFCC系数对于乐器识别十分有效。它反映了音频信号的能量在不同频带的分布。不同乐器的音频信号能量集中在某些特定的频带，例如大提琴音频能量集中在低频部分，然而小号的能量却集中高频部分，小提琴能量集中的频带位于大提琴和小号的频带之间。MFCC系数能够有效地描述乐器的声学特性。

在MFCC[3]系数计算过程中，首先用FFT将时域信号转化成频域，然后对其对数能量谱用依照Mel刻度分布的三角滤波器组进行卷积，最后对各个滤波器的输出构成的向量进行离散余弦变换DCT，取前N个系数。其具体的计算过程如下：

(1)封帧。将音频信号封装成若干个相邻的帧，相邻前后帧之间存在重叠。帧的大小一般为10~20ms之间，重叠大小为帧长的1/4~1/2之间。

(2)加窗。为了减少每一帧信号的头和尾的不连续性，防止信号的扭曲，给每一帧信号与一个特定的窗做卷积，使之在头和尾部分别逐渐增大和减小。

(3)快速傅利叶变换。对每一帧作FFT变换，并求其幅度谱。

(4)Mel刻度转化。心理学研究表明：人耳对音频信号频率内容的感知并非遵循一个线性尺度。因此，对于每一个用Hz单位度量的音调，主观感知的音高是遵循Mel尺度。Mel尺度在1 000Hz以下和Hz频率之间存在着线性的关系，而在1 000Hz以上则和Hz频率的对数存在线性关系。频率为1 000Hz、感知听觉门限40dB以上的音高作为一个参考点，定义为1 000Mel。因此使用以下的近似公式给出Mel和频率Hz刻度之间的关系：

作者简介：张 奇(1972－)，男，硕士生，研究方向为基于MPEG-7的音频内容检索方法；苏鸿根，研究员、教授

收稿日期：2003-07-17 **E-mail**：zq00ff@sina.com

$$mel(f) = 2595 * \log(1 + f/700)$$

一种模拟这种主观频谱的度量方法是使用一组沿着Mel刻度均匀分布的滤波器组，滤波器组有一个三角形的带通频率响应，带宽之间的间隔由一个Mel常数决定，如图1所示。当S(w)作为输入时，滤波器组的输出功率就是Mel谱，Mel谱系数的数目k一般情况下为20。我们可以把Mel封装的滤波器组看作一个频域直方图，直方图中的直方间在频域上存在着重叠。

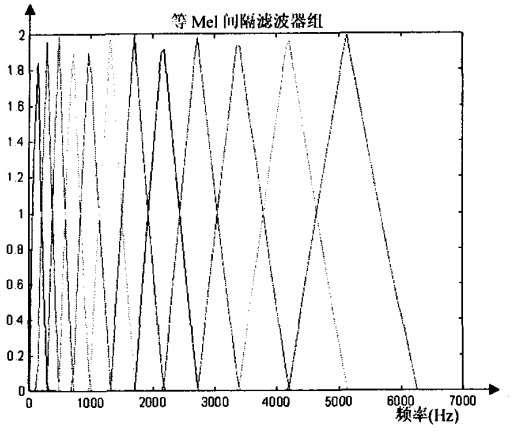

图1 Mel和频率Hz刻度之间关系图

(5)倒普计算。在最后一步，需要把Mel频谱转化到时间域，所得结果称为Mel频率倒谱系数(MFCC)。由于滤波器组输出的Mel功率谱系数是实数，可以采用离散余弦变换对它进行压缩。如果采用 $\tilde{S}_k, k = 1,2,..., K$ 表示Mel功率谱系数，那么采用如下公式计算MFCC系数 $\tilde{C}_n$：

$$\tilde{C}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 1,2,...,K$$

离散余弦变换之中的 $\tilde{C}_0$ 表示输入信号的均值，所描述的音频内容特征极少，一般情况下将其舍去。

2 模式识别的方法

模式识别的目标是把一组称作测试集的模式分成两类或者更多的类别，这种分类方法是通过比较被测试数据和已知类别的相似性完成的。被测试数据称作测试集；已知类别的分类数据称作训练集，它作为一个基础来判断未知类别的数据和哪一类别相似。

这种方法和人们对音频数据的主观分类行为非常类似。人们首先对不同声源发出的声音建立一个主观描述，当听到一个新的、未知类别的声音时，大脑会自动把未知声音同主观描述相比较，看看和哪一类别的数据最为接近，从而形成一个判断。而在计算机识别声音的过程中，提取未知声音的特征向量作为相似性比较的对象，用未知数据和已知类别之间的相似度作为分类依据。

本文采用的分类器为高斯混合模型(GMM)[2]。它广泛应用于模式识别和数据分析等领域。它是一种无导师的学习的方法，模型中的参数用一类训练样本通过最大似然估计方法得到。它的学习方法如下：

一个具有M个混合数的D维GMM，用M个分量的加权和来表示，即

$$P(x \mid \lambda) = \sum_{i=1}^{M} P_i b_i(x)$$

其中x是一个D维的观测矢量；$P_i, i = 1,2,\cdots,M$ 为混合权值，且 $\sum_{i=1}^{M} P_i = 1$，$b_i(x)$ 为D维高斯函数，即

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1}(x - \mu_i)\right\}$$

其中 μ_i 为均值矢量，Σ_i 为协方差矩阵。整个高斯混合模型便可由各均值矢量、协方差矩阵及混合分量的权值来描述。因此，将一个模型 λ 表示为如下三元式

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i = 1,2,\cdots,M$$

设训练特征矢量系列为 $X = \{x_t, t = 1,2,\cdots,T\}$，它对于模型 λ 的似然度表示为 $P(X \mid \lambda) = \prod_{t=1}^{T} P(x_t \mid \lambda)$，训练的目的就是为了找到一组参数 λ，使 $P(X \mid \lambda)$ 最大，即

$$\lambda = \arg \max_{\lambda} P(X \mid \lambda)$$

$P(X \mid \lambda)$ 是 λ 的非线性函数，直接求其最大值是不可能的，这种最大参数估计可以利用EM算法的一种特殊形式，通过迭代得到。

然而采用EM算法进行参数估计时，需要给 λ 设定一个初始值 λ^0，现在尚没有解决这一问题的理论框架。一种方法是从训练数据中任取50个数据，求其均值和方差，作为初始值和方差；另一种方法是采用隐马尔科夫模型(HMM)对训练数据进行分段，分出不同的状态，得到各分量的均值和方差的初值。

在乐器识别的阶段，设S类乐器对应的GMM模型分别为 $\lambda_1, \lambda_2, \cdots, \lambda_s$，目标则是对一个观察序列X，找到使之有最大后验概率的模型所对应的乐器 λ_s，即

$$\hat{s} = \arg \max_{1 \le k \le s} P_\gamma(\lambda_k \mid x) = \arg \max_{1 \le k \le s} \frac{P(x \mid \lambda_k) P_\gamma(\lambda_k)}{P(x)}$$

假定 $P_\gamma(\lambda_k) = 1/s$，即每个乐器出现为等概率，且因 $P(x)$ 对每个乐器都是相同的，上式可简化为

$$\hat{s} = \arg \max_{1 \le k \le s} P(X \mid \lambda)$$

如果使用对数表示的后验概率，乐器识别的任务就是计算：

$$\hat{s} = \arg \max_{1 \le k \le s} \sum_{t=1}^{T} \log P(x_t \mid \lambda_k)$$

3 实验结果

试验采用两类乐器——管乐器和弦乐器，6种乐器——大提琴、小提琴、琵琶、古筝、长笛和小号。每种乐器由80个独奏样本组成，其中60个样本作为训练集，20个样本作为测试集。样本采样率为11 025Hz，样本的长度为3s，格式为WAV文件，帧长为20ms，帧移为10ms。对每一样本采用八维的零阶MFCC系数描述乐器的静态特征，八维的一阶MFCC系数反映乐器的动态特征。高斯混合模型的混合数设为7。获得的实验结果如表1。

(下转第173页)

道路卡口的摄像机将现场的视频采集并传输到监控指挥中心，当有车辆通过道路卡口时，触发系统抓拍车牌图像，并由车牌识别软件对车辆的牌照进行识别处理，其识别结果在问题车辆牌照数据库中加以查询和比对，一旦发现有牌照匹配的问题车辆即报警处理，以便有关警力予以有效地追踪或拦截，其数据流图的主要部分如图4所示。

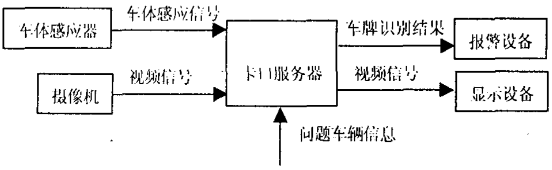

图4 道路卡口子系统数据流图

3.3 接/处警子系统

接/处警子系统主要由固定目标报警、110报警、122交通事故报警和119消防报警几个部分组成。

固定目标报警通过采用技术防范和报警设备完成对重要现场的报警检测和监控，由用户端系统、报警中心系统和联网报警信号传输系统组成，采用的传输方式是公共电话线传输。用户端报警系统由前端探测器、控制主机和报警输出等部分组成，可以通过控制通信主机面板使固定目标报警系统处于布防或撤防状态。当用户端由红外探头检测到物体移动，或烟感探头检测到火警，或发生紧急事件按动紧急按钮发生报警时，除了本地警号发声或闪灯报警外，还会通过主机内置拨号器拨通监控指挥中心接收/处理系统，并通过约定的信号将报警信号向外传输，监控指挥中心就可根据报警类别和报警地点自动查找相关数据库(重点单位数据库、预警方案数据库、地理信息数据库、公安基本信息库、机动车信息、身份证等)，对交通、火警、医疗急救则分别转入122/119/120指挥中心，或进行集中受警处理。

110/122/119报警部分的主要功能包括人工报警、自动报警、GPS卫星定位、视频监控、调度指挥、自动去除误报、自动呼叫分配、自动识别主叫号码、自动获取主叫的相关资料、自动获取主叫所在范围内的警力信息、接/处警同步数字录音、多个接警席位并行工作、分区调度、电子地图显示等，它主要由接警台、监听台、服务器、智能交换平台和电子地图系统组成。接警台和监听台的电话机通过公用电话网与公安局110指挥中心、122交通报警中心和119消防报警中心相连接，110服务器连接到公安指挥中心的局域网上，各种报警电话信息可通过监控指挥中心的110服务器将报警信息送往监控指挥中心大屏显示控制器，在电子地图上显示报警地点和警力分配，并可随时通过公安电话专线对警力发出调遣命令。智能交换平台是接/处警系统的核心，需要完成交换、组网、排队录音、多级接警等功能，实现110系统组网、接警、主叫号提取、错号拦截、被叫锁定、同时振铃、调度指挥、语音信箱、电话会议、汇接等功能。电子地图融地理信息和相关数据于一体，实现运动目标自动定位、动态数据连接、模糊查询、漫游移动、缩放和编辑修改等功能，并提供完备的数据可视化及地理信息分析能力。

4 视频的显示和记录

视频的显示和记录是城市监控系统中非常重要的内容之一。视频的显示是指将摄像机采集的监控现场的视频信号以灵活、适当的方式加以展示，方便监控指挥中心对监控现场进行实时过程监视，为警力的指挥调度提供可视图像证据。根据系统应用规模，可配置若干台彩色纯平收监两用机作为监视器组成电视墙，对视频信号进行一对一或有选择性地控制显示，使用一套大屏幕显示投影墙，作为整个监控指挥中心信息的显示主体。也可使用一体化投影单元拼接成投影墙，用大屏幕综合显示管理软件对显示系统进行管理，主要功能包括用户权限定义、窗口管理、显示模式管理、信号源管理、投影机控制和预案管理。

根据系统的规模，配备若干台多媒体数字硬盘录像机完成视频信息的记录和存储，以便实施视频的检索和回放。多媒体数字硬盘录像机通过网络互联，以实现资源共享。

5 结语

城市监控系统的建立必将大大提高城市的综合管理水平和管理效率，能有效预防、及时发现和制止违法犯罪案件的发生，充分发挥现有警力的作用，维护社会稳定，有效地遏制交通堵塞、火灾等诸多城市问题，为城市经济建设创造良好的环境。笔者于2002年对云南省丽江地区公安局城市监控系统进行了分析和设计，工程项目的建成将有效地实现该地区城市的治安和道路监控，提高公安机关协调处置突发事件的能力。

☆☆

(上接第134页)

表1 MFCC系数(tr60,tt20)

类别	大提琴	长笛	小号	小提琴	琵琶	古筝	合计
正确	16	20	18	20	17	20	111(92.5%)
错误	0	2	0	4	0	3	9(7.5%)

试验表明采用基于高斯混合模型的方法能够有效地识别乐器，试验所得的正确率为 92.5%，这说明了识别算法是有效的。

4 结论

本文把话者识别的方法应用到乐器识别中，采用MFCC系数作为声学特征，用高斯混合模型作为分类器，实现弦乐器和管乐器的识别。这种基于高斯混合模型的方法克服了传统树形分类方法的错误累积缺陷，把分类的准确率提高到90%以上。然而，高斯混合模型是一种基于概率密度参数估计的分类方法，它要求在学习的样本尽可能大的情况下才能准确估计概率密度参数。今后需要进一步改进乐器识别算法的性能，使之在小样本的情况下仍能够准确地识别乐器。

参考文献

1 马继勇.话者识别[博士论文].哈尔滨工业大学,1998
2 易克初.语音信号处理.北京:国防工业出版社,2000
3 Do M N.Digital Signal Processing Mini-project. http://lcavwww.epfl.ch/~minhdo/asr_project/asr_project.html
4 Herrera P,Amatriain X,Batlle E,et al.Towards Instrument Segmentation for Music Content Description:a Critical Review of Instrument Classification Techniques. In Proc. of International Symposium on Music Information Retrieval, 2000
5 Martin K D,Kim Y E.Musical Instrument Identification:A Pattern-recognition Approach.136th Meeting of the Acoustical Society of America, 1998-10-13