第18卷 第8期
2008年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.18 No.8
Aug. 2008

音频检索技术研究

李 晨，周明全
（西北大学 信息科学与技术学院，陕西 西安 710127）

摘 要：结合音频检索发展现状，描述了当前相关研究的进展，介绍了现在最常用到的音频检索方法，讨论了与音频检索相关的关键技术：音频特征提取、音频分割和分类。基于内容的音乐检索研究是一种涉及音乐理论、信号处理、模式识别等相关领域的综合学科研究，其在音乐数据库管理、Internet音乐检索以及生活娱乐等方面都具有非常重要的意义。分析并总结出音乐内容及其检索的概念，给出音乐检索的系统结构，综述了基于内容的音乐检索方法，最后指出了音频检索发展的前景。

关键词：基于内容的音频检索；特征提取；音频分割；基于内容的音乐检索
中图分类号：TP391.3 **文献标识码**：A **文章编号**：1673－629X(2008)08－0215－04

Research on Technology of Audio Retrieval

LI Chen, ZHOU Ming-quan
(College of Information Science and Technology, Northwest University, Xi'an 710127, China)

Abstract: Traditional simple method of retrieval is no longer adequate for multimedia data, since the digitized representation of images, video, or data do not convey the reality of these mediaitems. The research of content－based musical retrieval is composed of the study of musical theoretics, signal processing and pattern recognition. And it is significant for the management of the music digital library, the music search in the Internet and the daily amusement. Therefore, content－based retrieval for audio data is realized taking such intrinsic features of multimedia data into account. Surveys recent studies on content－based retrieval for music databases from the point of view of some fundamental issues.

Key words: content－based audio retrieval; feature extraction; audio segmentation; content－based musical retrieval

0 引 言

音频检索与图像检索、视频检索并列为当今基于内容检索研究的热点，而目前基于内容的多媒体信息检索技术研究成果主要集中在图像和视频方面。随着人们每次能够处理的音频信息量越来越大、音频信息的种类越来越多，其中占所有总信息量的20%左右，要从这海量的音频信息中迅速、有效地检索出所需要的音频信息就变得越来越重要。

国内外研究机构对音频检索进行了多方面的研究。例如，美国的Muscle Fish是一个商业化的基于音频感知特征的音频检索引擎，马里兰(Maryland)大学的Voice Graph结合基于内容和基于说话人的查询，检索已知的说话人和词语，并设计了一种音频图示查询接口。另外，国内的ARS系统[1]是基于内容的音频信息检索与分类系统。ARS系统建立了一个原始音频库，并收录了包括语音、音乐、动物声、笑声、电话铃声等十几类近300个音频文件，文件格式为“wav”格式。在实际的检索过程中，采用音调、音强、亮度、带宽、过零率等5个特征，并用基于欧氏距离的聚类算法将所有文件聚为50类，形成一个聚类参数库。聚类后，该系统对原始音频库进行特征处理，建立了音频特征库，进而形成一个音频数据库。实际检索就是对该音频数据库的检索，检索方式有三种，分别为基本属性检索、特征值检索、示例检索。

1 音频以及音频内容

音频是多媒体中的一种重要媒体，包含丰富的听觉特征。人耳能够听见的音频频率范围是20Hz～60Hz，其中语音频率大约分布在300Hz～4000Hz之间，而音乐和其他自然声响则是全范围分布。语言是人类进行思想、观点和情感交流最自然便捷的交互方式，音乐又是人们日常生活中形影不离的朋友，所以人

收稿日期：2007－11－06
基金项目：国家自然科学基金(60673100)
作者简介：李 晨(1982－)，女，陕西西安人，硕士研究生，研究方向为图形图像处理与多媒体应用；周明全，博士，博士生导师，研究方向为图形图像处理与多媒体应用。

们希望能通过这些自然的听觉特征来检索声音信息。

音频的内容从整体上来看可以划分成三个等级：最底层的物理样本级、中间层的声学特征级和最高层的语义级。

在物理样本级，音频内容是以媒体流的形式存在，其中包含原始音频数据和注册数据(如采样频率、量化精度和压缩编码方法等)。用户通过音频录放软件如CoolEdit等只能以时间刻度来检索和浏览音频内容。

中间层是声学特征级。声学特征是从音频数据中自动抽取的，它可以分为物理特征(Physical Feature)和感觉特征(Perceptual Feature)，前者包括音频的基频、幅度和共振峰结构等，后者表达用户对音频的感知，例如音调、响度和音色等，感觉特征一般都与某些物理特征之间存在一定的联系。

最高层是语义级，它是音频内容、音频对象的概念描述。具体来说，在这个级别上，音频的内容可以是语音识别、辨别后的结果(文本)、音乐旋律和叙事说明等。

2 音频检索系统通用流程

在音频检索中，需要经过特征提取、音频分割、音频识别分类和索引检索这几个关键步骤，见图1。

原始音频流⇒特征提取⇒音频分割⇒音频识别⇒音频检索

图1 音频检索流程图

迄今为止，音频检索的方法有很多种。人是通过听觉特征来感知声音的，所以人们希望能通过这些自然的听觉特征来检索声音信息。在这里介绍一种很有效的技术——基于内容的音频检索技术。基于内容的音频检索(Content-Based Retrieval, CBR)，就是指从媒体数据总体取出特定的信息线索，建立音频数据表示方法和数据模型，采用有效和可靠的查询处理算法，使得用户可以在智能化的查询接口的辅助下，从大量存储数据库中的媒体进行查找，检索出具有相似特征的媒体数据出来。

基于人工输入的属性和描述来进行音频检索是首先想到的方法。该方法的主要缺点反映在以下几个方面：当数据量越来越多时，人工的注释强度加大；人对音频的感知，如音乐的旋律、音调、音质等，难以用文字注释表达清楚。这些正是基于内容的音频检索需要研究和解决的问题。

基于内容的音频检索技术最关心的是声学特征级和语义级的音频检索。在这两个层次上，用户可以提交某一概念或按照特定的声学特征进行查询。基于内容的方法从新的角度来管理多媒体信息，基于内容的检索就是从多媒体数据中提出特定的信息线索，然后根据这些线索从大量存储在数据库中的多媒体数据里进行查找，检索出具有相似特征的多媒体数据。最简单的基于内容的音频检索使用查询和存储的音频片段之间的样本到样本之间的比较。基于内容的检索不同于图像处理、模式识别、图像理解、语音识别等，但要以这些技术作为其重要的基础。

音频信息的检索技术具有广泛的应用前景。Niessen[2]根据音频的相似性匹配技术来辨别病变的心音，从而可以及早发现心脏的病变情况。Foote[3~5]利用Mel频率倒谱系数(Mel Frequency Cepstrum Coeficient，简称MFCC)作为音频特征，采用一种应用动态规划技术(Dynamic Programming，简称DP)的树形结构分类器，对音频信号进行分类与检索处理。为了度量音频信号之间的相似性，建立一个代价模型对不相似的音频点加以惩罚，较低的代价表示检索到相似的音频片段。但是由于MFCC特征不能很好地反映声音的音色、音品和音质等属性，因此该方法对音乐和多种声音混合而成的环境音区分效果不理想。

抽取音频特征是进行音频检索的首要任务。特征提取指的是寻找原始音频信号表达形式，提取能代表原始信号的数据。根据音频信号的短时平稳特性，可以固定长度的音频帧为单位，统计帧内音频信号的各项属性以表征该帧音频信号，即音频特征。为了改善音频分类与检索的准确性、速度等性能，选择能够有效捕获音频信号能量谱特性短时变化的音频特征是非常重要的。首先对音频数据进行加窗处理形成帧，加窗大小在几到几十微秒，相邻帧之间一般有30%～50%的叠加。然后对每一帧作离散傅里叶变换(DFT)，实际上常用快速傅里叶变换(FFT)，得到傅里叶系数$F(w)$和频域能量E。最后应用不同算法计算相应的帧特征，而后对帧特征计算其标准偏差、数学期望和方差，把帧特征推广成片段特征。常见特征有：

1)短时平均能量(STE)：指在一个短时音频窗口内采样的点信号所聚集的平均能量。短时平均能量可以较好地表示音频信号幅度随时间的变化。应用短时平均能量特征的主要原因可概括为如下三点：

(1)对于纯语音信号，短时平均能量能够较好地区分语音中的清音成分与浊音成分，因为清音成分的短时平均能量通常明显地小于浊音成分的短时平均能量；

(2)当音频信号的信噪比较高时，短时平均能量可以有效地区分其中的静音部分；

(3)短时平均能量随时间的变化，可以反映音频的节奏、周期等属性。

2)短时过零率：指在一个短时帧内，离散采样信号

值由正到负和由负到正变化的次数。短时平均过零率是另一种区分纯语音信号中清音成分与浊音成分的有效度量方法,因为清音成分通常比浊音成分具有更高的过零率,这也导致过零率幅度变化比较明显;另外,大多数音乐信号集中在低频部分,其过零率不表现出突然升高或降落的起伏特性,所以有时也可用过零率来区分语音和音乐两种不同音频信号。

3)线性预测系数(Line prediction coefficients, LPC):在一个短时帧内,用有限个参数的数学模型来近似表示音频采样序列 $x(n)$,这些参数就成为 $x(n)$ 的重要特征,叫做线性预测系数。无论在音频压缩编码还是在音频信息检索方面均有极广的应用。

4)基于Mel频率的倒谱系数(MFCCs)。这是音频数据经过Z-变换和对数处理后得出的结果。一般对每帧数据取12个系数就可以很好地表现每帧的特征。其处理过程见图2。

特征提取完成后,所有这些提取出来的特征被用来表征音频数据流,在后续处理时被用到。由于音频信息是时间序列数据流,好比不能对100MB大小的纯文本信息直接分析而要把他分成不同主题子段一样,也不能对持续时间很长的音频直接处理,而是在其特征发生突变的地方进行分割,把连续多媒体数据流分成不同长度的数据片段,这是音频分割需要完成的任务,然后对分割好的数据片段进行处理。

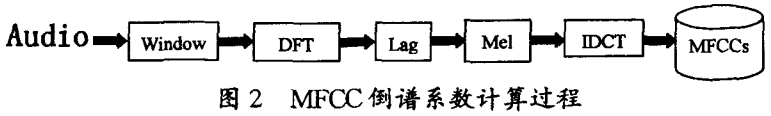

图2　MFCC倒谱系数计算过程

音频数据流分割基本是根据所提取的音频低层物理特征完成的,所分割出来音频数据只是些物理单元,需要对这些物理单元进行识别分类,将它们归属成事先定义好的不同语义类,这由音频识别分类这一步完成。音频物理单元被分割出来再进行细分,如分成静音、音乐和语音、环境音等,也可以进行某事件或某人物的精细分类,如"爆炸"事件、"演讲"事件等。下一步工作就是检索工作要完成的任务,即把分割出来的音频物理单元识别分类成哪些语义类。

最后要做的就是音频检索的最后一步,对识别出来的语义类建立索引,进行检索。建立索引可以有三个途径:

(1)用文字形成的抽象概念描述这些类别,这样用户必须通过文字查询音频数据。

(2)用音频特征建立索引,查询时用户提交的是对特征的描述,如对音频能量描述的"音调"。

(3)提交一个音频例子,提取这个音频例子的特征,按照前面介绍的音频例子识别方法判断这个音频例子属于哪一类,然后把识别出的这类所包含的若干个样本按序返回给用户,这是基于例子的音频检索。

3　音乐检索

一般情况而言,音乐检索系统可分成4个模块:检索界面、音调跟踪、特征音乐数据库生成和检索引擎(见图3)。

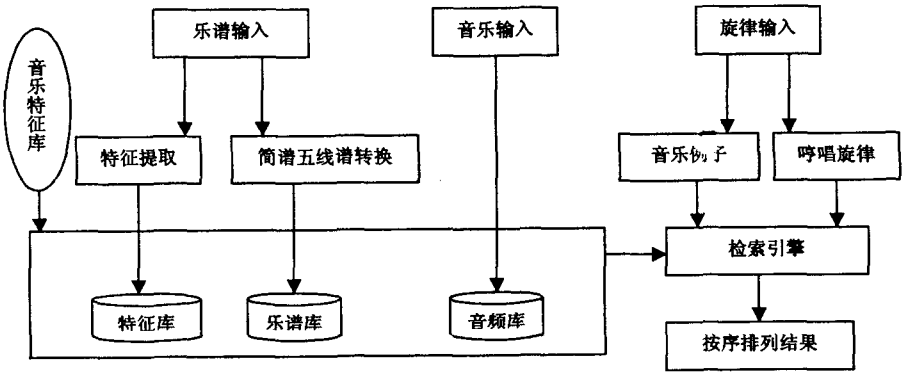

图3　音乐检索系统

3.1　绝对音高序列

绝对音高序列包含了旋律的准确音高,其优点是可对音乐旋律进行完全精确的检索。为了理解上的方便和简化计算机的处理,可采用128个梯度来表达从最低到最高的音高范围,使用的音高符号从低到高依次为:C、Db、D、Eb、E、F、Gb、G、Ab、A、Bb、B,八度音阶从0到10,这样整个音高范围就可以用从C0到G10的符号来表示。如民族音乐《二泉映月》,其旋律开头部分的绝对音高序列可表示为E5Gb5D5D5E5Gb5A5B5A5。

尽管通过绝对音高序列可以非常精确地检索到相关的乐曲旋律,但也有其不足之处:首先,它要求检索者非常准确地把握此旋律的音调和音高,这种方式对于音乐知识并不丰富的一般检索者来说是比较困难的。其次,有些音乐的旋律中存在大量即兴性、不确定性因素。比如一些民族音乐的音调是不固定的,不同的演奏者、不同的乐器在演奏时其调号可能会发生改变,尽管此时音乐的旋律还是一样的,但由于其音调不同,所以基于绝对音高的音高序列就是完全不同的,这将导致检索的失误。

3.2　相对音高序列

针对绝对音高序列旋律轮廓的不足,一种改进的

使用相对音高序列的旋律轮廓在实际中应用得更为广泛,这就是采用音频的三步轮廓表示法。对任意一个旋律中的音符,它都有以下三种状态:"U"该音符比前一音符音调高,"D"该音符比前一音符音调低和"S"该音符与前一音符音调相等。按这种规则,任意一段旋律可转化为一个包含字母 U、D、S 的字符序列。

把音乐和歌曲转化为由 U 、D、S 三个字符组成的字符串表示,装入音频特征库。把检索请求表示成三步轮廓形式后,就可对音频特征库进行检索了。该方法的优点还在对于非音乐专业人员,即使其给定的拟声查询不是很准确,相对音序列旋律轮廓可以有效地解决绝对音高序列旋律轮廓的不足。

特征音乐库中装入的是原始音乐(歌曲)和它们的曲调特征,检索引擎将检索请求的曲调与存储在数据库中的音乐进行特征匹配,得到与之匹配的音乐(歌曲),最后将它们按照匹配程度大小反馈给用户。

3.3 基于内容的音乐检索

现阶段在音乐检索中广为流行的是基于内容的音乐检索方法。众所周知,网络海量的音乐数据通过人工实现分类和标注逐渐变得不切实际,需要自动分类和检索技术的帮助;同时人们在查询数据时往往并不知道音频数据的名称、作者等版权信息,仅仅知道一些内容片断,这也要求基于内容的音频检索技术帮助查询。因此,基于内容的音乐检索在音乐数据库管理、Internet 音乐检索以及生活娱乐方面都具有非常重要的意义,该领域具有很强的实用意义和研究价值。目前该领域中,音频的分类技术(如将音频文件分为"音乐"、"语音"、"噪音"等)得到了较快的发展,而由于音乐自身的特征表达和模式匹配问题,基于内容的音乐检索技术则发展缓慢。

音乐与人的听觉感知密切相关,它更多地表达了一种感情,一种很难量化的情绪,音乐的这种特性决定了在音频的分类检索技术中所用到的物理特征对音乐分析并不适用,基于内容的音乐检索是根据音乐的内容特征来进行检索,也就是根据音乐的旋律、节奏等音乐特征进行检索。基于内容的音乐信息检索技术是一个涉及交叉学科的研究方向,涉及到的学科包括:计算机科学、信息检索、音乐学、音频技术、数字信号处理和认知科学等。

基于内容的音乐检索涉及音乐旋律的表达、音乐旋律的特征提取、用户查询构造、音乐旋律匹配以及音乐数据库构造等很多方面的问题,这些问题的解决是建立一个完整、有效的音乐检索系统的关键。基于内容的音乐检索通常采用下面通用的步骤:

1) 音乐旋律的表达.即音频信号的预处理;

2) 通过对音乐旋律的特征提取,形成查询索引;

3) 对音乐数据库中的音乐建立音频索引;

4) 用户查询构造;

5) 根据查询索引和数据库中音频索引之间的相似性,对音乐片段进行检索。

基于内容的音乐检索主要是基于音频特征矢量匹配和近似音调匹配。计算机对信息的表达归根结底是一种状态表达,要将听觉感知的信息借助计算机进行存储与检索,这当中进行的转换难以避免实际信息的失真。对于音频检索来说,由于感官上与表达上的不一致性大大增加了检索的处理难度。因此,基于内容的音频检索只能是一种相似性检索,而无法实现传统的精确匹配检索。虽然研究人员已在基于内容的音乐检索技术方面做了大量的研究,但是为了满足大容量数据库和 WWW 检索的要求还有许多工作要做。

4 展 望

由于基于内容的音频检索技术不够成熟,该研究领域还有许多急待解决的问题,如:高层概念和低层特征的关联,以实现媒体语义的计算机自动抽取;Web 上基于内容的音频检索,需解决快速、大规模音频库的浏览、检索和连续音频媒体提交等;用户的音频查询接口,需要一种友善易用的用户接口来提交音频查询,使用户在主动的交互过程中表达对音频媒体语义的感知,调整查询参数,最终获得满意的查询结果。总之,为使计算机能像人那样对音频语义实现自动理解,需要做的工作还很多。

WWW 上基于内容的音频检索问题,需要研究快速的大规模音频库的浏览、检索和提交;长音频的浏览,即结构化表示音频流,并设计出新形式的内容浏览界面;长音频的检索,研究通用的基于片段级的内容检索,在时间轨迹上匹配一组特征,这需要研究模糊的匹配方法;继续研究有效的听觉解析特征,以支持通用和专用的音频检索问题;用户的音频查询接口和检索引擎;音频索引问题,以满足大容量数据库和网络检索的要求。

5 结束语

音频检索是一个非常宽泛的研究领域,要想使计算机能像人那样对音频语义实现自动理解,并根据语义高级内容进行音频检索,将要有很长的路要走。

参考文献:

[1] 李国辉,李恒峰.基于内容的音频检索:概念和方法[J].小

(下转第 222 页)

口接收到数据,就会产生一个串口接收数据缓冲区中有字符的消息事件,刚才添加的函数就会执行,在 OnComm()函数加入相应的处理代码就能实现自己想要的功能了。具体实现代码如下:

```
void CCMyCommDlg::OnComm()
{
VARIANT variant_inp;
COleSafeArray safearray_inp;
LONG len,k;
BYTE rxdata[2048]; //设置 BYTE 数组 An 8-bit integerthat is not signed.
CString strtemp;
if(bReceive)
{
if (m_ ctrlComm. GetCommEvent()==2) //事件值为 2 表示接收缓冲区内有字符
{ variant_inp = m_ ctrlComm. GetInput(); //读缓冲区
safearray_inp = variant_inp; //VARIANT 型变量转换为 ColeSafeArray 型变量
len=safearray_inp. GetOneDimSize(); //得到有效数据长度
for(k=0;k<len;k++)
safearray_inp. GetElement(&k,rxdata+k);//转换为 BYTE 型数组
for(k=0;k<len;k++) //将数组转换为 Cstring 型变量
{
BYTE bt= *(char *)(rxdata+k); //字符型
strtemp. Format("%c",bt);
m_edit_receive += strtemp; //加入接收编辑框对应字符串
}
}
UpdateData(FALSE);
}
```

4 结束语

所介绍的利用 MSComm 控件基于 VC++6.0 文档视图体系结构开发的 CPM2A * 系列 PLC 系统监控软件已经成功地运用在 CPM2AE PLC 上。在 PLC 运行长达 8 小时的监控运行中稳定可靠。PC 机定时读取 PLC 36 个输入端口、24 个输出端口的工作状态,并实时显示。图 3 是笔者开发的 PC 机与 OMRON PLC 之间的串行通信软件界面。

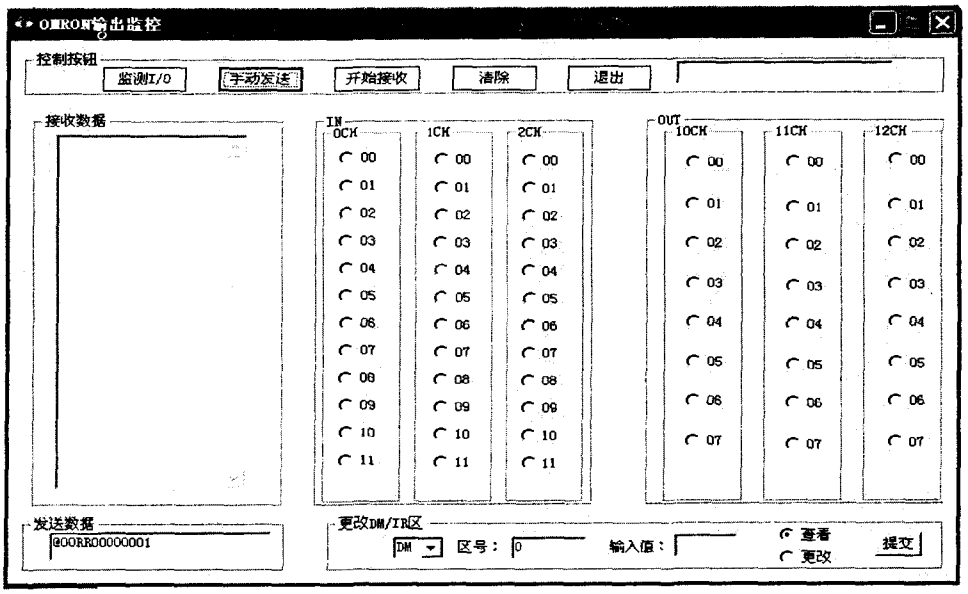

图 1 PC 机与 OMRON PLC 间的串行通信软件界面

参考文献:

[1] 朱善君. 可编程控制系统原理、应用和维护[M]. 北京: 清华大学出版社,1992:15-34.

[2] OMRON 公司. C200HX/C200HG/C200HE 编程手册[M]. [s.l.]:OMRON 公司,1997:75-96.

[3] 李东晓,李晓明,李留振,等. PC 与 PLC 实时通信的 VisualC++6.0 实现[J]. 计算机应用研究,2002(1):111-113.

[4] 何华东,赵喜荣,王程远,等. PLC 与上位计算机的串行通信程序设[J]. 机电工程,2002,19(2):24-26.

[5] Nelson M. 串口通讯开发指南[M]. 北京:中国水利水电出版社,1999:20-45.

(上接第 218 页)

型微型计算机系统,2000(11):1173-1177.

[2] Nilssen E J. Feature Extraction and Classification of Earth Sounds[D]. Diploma Thesis in Applied Physics. Institute of Mathematical and Physical Sciences, University of Thromso, 1996.

[3] Foote J. Content-based Retrieval of Music and Audio[J]. SPIE,1997,3229:138-147.

[4] Foote J. A Similarity Measure for Automatic Audio Classification[C]//AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora. USA:Stanford University,1997:1-7.

[5] Foote J. ARTHUR: Retrieving Orchestral Music by Long-term Structure[C]//Proceedings of the International Symposium on Music Information Retrieval. [s.l.]:[s.n.],2000:1-7.