基于内容的音频检索关键技术研究

朱爱红[1] ， 李 连[1,2]

（1. 海军航空工程学院，烟台 264001；2. 北京理工大学，北京 100081）

摘 要：音频是一种重要的媒体，包含丰富的听觉特征。本文根据当前音频检索研究的进展，综述基于内容的音频检索方法，讨论了一些音频检索技术研究中的关键技术：音频特征提取、音频分类、语音识别技术等。最后展望了音频检索技术的发展前景。

关键词：基于内容的音频检索；语音识别；音频分类

引 言

基于人工输入的属性和描述来进行音频检索是一种传统的方法，其主要缺点有：一是当数据量越来越多时，人工注释的工作量加大；二是人对音频的感知有时难以用文字注释表达清楚，人工注释存在不完整性和主观性；三是不能支持实时音频数据流的检索。

为解决上述问题，基于内容的音频检索技术应运而生。基于内容的音频检索就是通过音频特征分析，对不同音频数据赋予不同的语义，使具有相同语义的音频在听觉上保持相似。

目前许多研究机构对音频检索进行了多方面的研究。Muscle Fish 是一个商业化的基于音频感知特征的音频检索引擎。MIT、南加州大学等都展开了音频检索研究，通过哼唱查询、音频分类、结构化音频表示和基于说话人的分割和索引等方面的研究。

根据当前相关的研究进展，本文综述了音频检索方法，讨论了一些音频检索中的关键技术：音频特征提取、音频分类、语音识别技术等。最后展望了音频检索技术的发展前景。

1 基于内容的音频检索

基于内容的音频索引和检索通常采用下面的步骤：

（1）将音频数据分类，通常可分成语音、音乐和噪声等类型。

（2）不同类型的音频数据可以以不同的方式进行处理和索引。如，对语音可运用语音识别技术且可基于识别过的词汇对其进行索引。

（3）查询音频片段要同样地进行分类、处理和索引。

（4）根据查询索引和数据库中音频索引之间的相似性，对音频片段进行检索。

下面根据音频检索过程，着重介绍音频特征提取、音频分类、音频检索方式等音频检索关键技术。

2 音频特征提取

音频分类以一些主观或客观的音频特征为基础。

（1）音频时域特征的提取

可提取的音频时域特征有平均能量、过零率和静音比等。

平均能量说明了音频信号的强度，可用于静音检测，对于一个音频例子，如这个音频例子中的某一短时帧的平均能量低于一个事先设定的阈值，则可判定该短时帧为静音[2]。

过零率指每秒内信号值通过零值的次数，一定程度上，它说明了平均信号频率。一般语音信号由单词构成，单词又由元音和辅音交替的音节组成，辅音信号的过零率低，而元音信号的过零率高。语

音信号开始和结束都大量集中了辅音信号,所以其开始和结束部分的过零率总会有显著升高,利用过零率可判断语音是否开始和结束。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。

静音比表示静音的声音片段的比例。

(2)音频频域特征的提取

傅里叶变换可分解出音频信号的频率成分,可提取的音频频域特征有带宽、频谱中心、谐音、音调等[1]。

带宽说明了声音的频率范围,音乐通常比语音信号具有更高的带宽。

频谱中心也称亮度,是一个声音频谱能量分布的中心点。语音与音乐相比,频谱中心较低。

频率为最低频率的倍数的频谱成分称为谐音。在有谐音的声音中,频谱成分大部分是最低频率的整数倍数,音乐通常比其他声音具有更多的谐音。

音调是听觉分辨声音高低的特性,完全由频率决定,可通过频谱估计。只有阶段性的声音,如那些由音乐设备和语音产生的声音,才会产生一种音调的感觉。可根据音调的级别对声音排序。音调是一个主观特征。

3 音频分类

由于不同的音频类型需要不同的处理和索引检索技术,音频分类这一步很重要。

(1)不同类型声音的主要特征

语音信号的带宽通常在 100~7000Hz 范围内,比音乐的带宽低。由于语音信号主要由低频成分组成,其频谱中心(或亮度)通常比音乐的频谱中心低。而且语音信号中单词和句子之间经常出现停顿,其静音比通常比音乐的静音比高。

语音的特征结构是包括短摩擦周期的连续音节(由辅音引起的),后面跟随着元音的较长间隔。研究证明,在摩擦周期内,平均过零率(ZCR)有明显的提高,因此,与音乐相比,语音在 ZCR 上有着更高的可变性。

音乐的带宽通常在 16~20000Hz 之间,具有较高的频率范围,其频谱中心比语音的要高。与语音相比,除单弦乐器或没有伴奏音乐的歌唱产生的音乐外,其他的音乐具有较低的静音比。音乐在 ZCR 上具有比语音低的可变性,另外音乐具有可抽取的正常的跳动,语音一般没有。

(2)音频分类

根据音频的特征值可将音频分类。常见的分类方法是:首先计算输入音频片段的频谱中心,如果其频谱中心值比预先设定的阈值高,则认为它是音乐;否则它是语音,但由于有的音乐也具有低的频谱中心值,因此它也可能是音乐。其次,计算静音比,如果它的静音比低,则认为它是音乐;否则,认为它是语音或独奏音乐。最后计算平均过零率 ZCR,如果它有着非常高的 ZCR 可变性,则它是语音,否则它是独奏音乐。

在这种分类方法中,特征判定的顺序是非常重要的,通常由计算的复杂性和特征的差别决定的。一般首先判定差别性大、复杂性低的特征,这样可减少一个特殊音频片段将要经历的步骤数,同时也可降低所需的整个计算量。

4 音频检索

将音频分类为语音和音乐后,就可以使用不同的技术对它们进行单独处理。

(1)语音识别和检索

语音索引和检索的基本方法是运用语音识别技术把语音信号转化为文本,然后应用 IR 技术进行索引和检索。除实际的发声词汇(spoken words)外,包含在语音中的其他信息,如发音者的身份和情绪等,都有助语音索引和检索。

①语音识别

一般来说,自动的语音识别(ASR)问题就是一个模式匹配问题。一个 ASR 系统通常包括训练和模式匹配两个阶段[3]。

在训练阶段,ASR 系统收集大量的发音者的语音序列,因为数字信号比模拟语音信号更适合实现语音识别,所以应把这些语音序列转化为数字格式。然后 ASR 系统提取每个语音单位的特征并存放在系统中,通常最小的语音单位为音素。最常用的特征是 MFCC 系数(Mel Frequency Cepstral Coefficients)。最后进行音素模型化,用上面获得的特征矢量、包含所有单词及其可能的发音的字典以及语法使用统计规则产生一个音素模型集合或模板,还可得到一个包括音素模型集合、同义词词库和语法的识别数据库。

在模式匹配(即识别)过程中,ASR 系统用与训练阶段相似的方法对输入语音进行处理,产生特征矢量,找到与输入语音的特征矢量最匹配的特征矢量的单词序列。

目前具有代表性的 ASR 技术有动态时间环绕技术、隐藏马可夫(Markov)模型(HMM)和人工神经网(ANN)模型。其中基于 HMM 的技术最为流行且语音识别性能最好,下面将详细介绍。

每个音素被分解成输入状态、中间状态和输出状态 3 个可听到的状态,每个状态可持续超过一个帧的时间(通常为 10ms)。在训练阶段,使用训练语音数据为每个可能的音素构建 HMM。每个 HMM 都具有以上 3 个状态,并由状态转换概率和符号发生概率来定义。在该环境中,符号是为每个帧计算的特征矢量。由于时间只向前流动,因此一些转换是不允许的。

在训练阶段末期,由不同的发音者、时间变化和周围的声音引起的变化,是每个音素都由捕获不同帧的特征矢量变化的一个 HMM 表示。在语音识别阶段,按照帧的顺序计算每个输入音素的特征矢量。识别问题的目的是去发现哪个音素 HMM 最可能产生输入音素的特征矢量序列。HMM 对应的音素被认为是输入音素,由于一个单词含有大量的音素,因此通常把音素序列放在一起进行识别。要计算 HMM 产生一个给定特征矢量序列的概率有许多算法,如前向算法和 Viterbi 算法等。前向算法用于识别隔离的单词,而 Viterbi 算法用于识别连续的语音。

②发音者识别

虽然语音识别集中于识别语音的内容,但是发音者的识别或口音(voice)识别可设法找到发音者的身份或提取个人语音的有关信息。发音者识别技术非常适用于多媒体信息检索,能改善信息检索性能。

(2)音乐索引和检索

音乐的类型有两种:结构化的(或综合的)音乐和基于样本的音乐。一般说来,音乐索引和检索的有效技术的研发仍处于初期阶段。

①结构化音乐的索引和检索

结构化音乐和声音效果是由一系列指令或算法来表示的。最常见的结构化音乐是 MIDI,它把音乐表示成大量的音符和控制指令。由于结构化音频的简明结构和音符描述的原因,没有必要从音频信号中抽取特征,因此结构化音频更便于检索。

结构化音乐和声音效果非常适合于音频基于精确匹配的查询。用户可指定一个音符序列作为查询,尽管可以找到该音符序列的精确匹配,但是由于相同结构化的声音文件可以由不同的设备以不同的方式进行表现,因此检索结果可能不是用户想要的声音文件,检索准确性能不是很高。

对于结构化音乐和声音效果,由于两个音符序列之间的相似性定义的困难性,基于相似性的检索很复杂。目前一种可行的方法是基于音符序列的音调变化来检索音乐。其基本思想是:查询声音和数据库声音文件中的每个音符(第一个音符除外)都被转换成相对前一个音符的音调变化。音调变化有三种状态:该音符比前一音符高(U)、该音符比前一音符低(D)和该音符与前一音符相同或相似(S)。按这种规则,任意一段旋律可转化为一个包含字母 U、D、S 的符号序列,检索任务也就变成了一个字符串匹配过程。该方法是针对基于样本的声音检索提出的,也同样适用于结构化声音检索,根据音符音阶可较容易地获得音调变化[2]。

②基于样本的音乐的索引和检索

对于基于样本的音乐的索引和检索有两种通用的方法:一是基于抽取的声音特征集合,二是基于音乐音符的音调。

●基于特征集的音乐检索

在这种音乐检索方法中,对每种声音(包括查询)抽取听觉特征集,将其表示成一个矢量。通过计算查询音乐和每个存储音乐片段相应的特征矢量之间的近似度来计算它们的相似性。该方法可应用于一般的声音中,包括音乐、语音和声音效果。

Muscle Fish LLC[4]完成的一项研究工作就是使用该方法的一个较好的实例。在这项研究中,共使用了 5 个音频特征:强度、音调、亮度、带宽和谐音。这些特征随着时间的变化而变化,因此可对每个帧进行计算,然后用统计学中的均值、方差和自动相关 3 个参数来表示每个特征。查询矢量和每个存储的音乐片段的特征矢量之间的欧几里德距离或 Manhattan 距离,可用作它们之间的距离。

●基于音调的音乐检索

该方法与基于音调的结构化音乐检索相似,两者之间的主要区别在于基于音调的音乐检索必须

抽取或估计每个音符的音调。

将一段旋律转化为一系列相对音调转移序列的过程称为音调跟踪。音调跟踪是自动化音乐转录的简化形式,它把音乐声音转化成符号表示。

该方法的基本思想为:由于音乐的每个音符都是由它的音调表示的,因此一个音乐片段或部分可表示成一个序列或音调串。检索是以查询音乐和每个存储音乐片段相应的音调串之间的相似性为基础,音调跟踪和串相似测量是检索过程的关键。

音调通常被定义为声音的基本频率。为了找到每个音符的音调,首先必须把输入音乐分割成单个音符。连续音乐,尤其是鼓乐和歌唱的分割是非常困难的。因此通常假定音乐是以计分的方式存储在数据库中,每个音符的音调是已知的。常用的查询请求形式是哼唱(humming)。为了改善查询请求的音调跟踪性能,通常要在相邻音符之间有一个停顿。

音调表示方法通常有两种。第一种方法,每个音调(第一个除外)都被表示成相对于前一个音符的音调方向(或变化)。音调方向可能是U(上)、D(下)或S(相似的),因此每个音乐片段都可表示成3个符号或字符组成的字符串。

第二种音调表示方法是基于选择的参考音符,把每个音符表示成一个值,该值是由最接近估计音调的标准音调值集合分配的。如果把每个许可值都表示成一个字符,则每个音乐片段都可表示成字符串,但是在这种情况下,许可符号的数量要比用于前一种方法的3个符号数量大。

在把每个音乐片段都表示成一个字符串后,需要进行字符串之间的匹配。考虑到哼唱不很准确且用户不只对一个音乐片段感兴趣而对所有相似的音乐片段都感兴趣,通常使用近似匹配而不是精确匹配。所谓近似匹配问题,就是查询音乐字符串和存储音乐片段的字符串最多可有k个不匹配的字符,变量k是由系统的用户决定的。研究人员已经设计出了几种解决近似字符串匹配问题的算法。

Muscle Fish LLC 的系统和 Waikato 大学[5,6]的系统都具有较好的检索性能,但是性能的好坏依赖于哼唱(humming)输入信号的音调跟踪的准确性,只有当在相邻的音符之间插入一个停顿时才能获得很高的性能。

5 展 望

运用IR技术对使用语音识别方法识别过的单词,进行语音索引和检索相对比较容易,但是没有任何词汇限制的一般主题的语音识别性能仍有待改进。对于音乐检索,主要是基于音频特征矢量匹配和近似音调匹配,这方面研究人员已经做了大量的工作。然而,对于一般情况下如何感知音乐和音频,以及关于音乐片段之间的相似性比较,还有许多工作要做。

在电影和电视节目等许多场合经常把音频和视频一起使用,因此音频检索技术可有助于定位一些具体的视频剪辑,同样视频检索技术也可有助于定位一些音频片段。利用这些关系可开发综合的多媒体数据库管理系统。

为满足大容量数据库和WWW检索的要求,需设计有效的检索界面,可基于语义内容进行检索。并且研究快速的大规模音频库的浏览、检索和提交问题。

结 语

本文介绍了基于内容的音频索引和检索的一些常用技术及相关的问题。基于内容的音频检索是一个涵盖十分广泛的研究领域,与信号处理、人感知心理研究和模式识别等学科紧密相联。为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,我们面临的挑战还很多。

参考文献

[1]朱学芳. 多媒体信息处理与检索技术. 北京:电子工业出版社,2002

[2]庄越挺,潘云鹤,吴飞. 网上多媒体信息分析与检索. 北京:清华大学出版社,2002

[3]李国辉,李恒峰. 基于内容的音频检索:概念和方法.小型微型计算机系统,2000(11)

[4]Erling Wood el. At. Content based Classification,Search and Retrieval of Audio. IEEE Multimedia,1996

[5]McNab R,J.,Smith L.A.,Witten I.H.,Henderson C.L. and Cunningham S.J. Towards the Digital Music Library:tune Retrieval from Acoustic Input. Proc Digital Libraries,1996

[6]McNab R,J.,Smith L.A.,Witten I.H. and Henderson C.L. Tune Retrieval in the Multimedia Library. Submitted to Multimedia Tools and Applications,1996

(收稿日期:2003-06-10)

(英文摘要见第51页)

参考文献

[1]Simon Robinson. C# 高级编程,2002

[2]Drashansky, T. T., Houstis, E. N., Joshi, A. and Rice, J. R. Networked Agents for Scientific Computing. Communications of the ACM, 1999(3)

[3]Drashansky, T. T., Joshi, A. and Rice, J. R. SciAgents——An Agent based Environment for Distributed Cooperative Scientific Computing

[4]Vigna,G. Cryptographic Traces for Mobile Agents. In Mobile Agents and Security.Berlin:Springer-Verlag,1998

(收稿日期:2003-08-29)

A Molile Agent System´s Security Model and Realization based on Microsoft .NET Frameset

XIAO Hai-peng

(China Agriculture Development Bank Guangdong Branch,Guangzhou 510100 China)

Abstract:In this paper, we introduce the security model and application technique of an mobile Agent system which is used in network and based on Microsoft .NET framework.

Key words:Mobile Agent;Security Model

(上接第 40 页)

Research on Content-based Audio Retrieval Key Techniques

ZHU Ai-hong[1] , LI Lian[1,2]

(1. Naval Aeronautical Engineering Institute,Yantai 264001 China;2.Beijing Institute of Technology, Beijing 100081 China)

Abstract:Audio is an important medium,and contains many hearing features. This paper reviews the research works and recent advances in audio retrieval, summarizes Content-based Audio Retrieval method. Some key techniques on Content-based Audio Retrieval(CBAR) are discussed: audio feature extraction, audio classification, speech recognition,and so on. Some future research directions are also given.

Key words:Content-based Audio Retrieval;Speech Recognition;Audio Classification