信息科学技术学术著作丛书

基于内容的音频检索技术

周明全　耿国华　王小凤　李　鹏　著

科学出版社

信息科学技术学术著作丛书

基于内容的音频检索技术

周明全　耿国华　王小凤　李　鹏　著

科学出版社

北京

内 容 简 介

本书系统介绍基于内容的音频检索技术的相关理论、方法和框架。全书共十章内容，主要涉及音频认知机理、音频信号特征表示和基于内容的音频处理分类技术；音频检索的特点、检索处理框架、声学特征级和语义特征级的音频检索技术；音乐表示、基于内容的音乐检索技术、框架和算法、基于语音识别和情感的音乐检索；音频处理技术的进展。重点介绍音乐特征库构建、基于旋律的音乐哼唱检索和乐理支持下的音乐检索方法。

本书适合从事信号处理、音频识别、基于内容检索交叉研究和应用的专业技术人员阅读，也可作为本科生、研究生的教学及参考用书。

图书在版编目(CIP)数据

基于内容的音频检索技术/周明全等著. —北京：科学出版社，2014
(信息科学技术学术著作丛书)
ISBN 978-7-03-041662-0

Ⅰ. 基… Ⅱ. 周… Ⅲ. 音频信号处理-情报检索 Ⅳ. TN912.3

中国版本图书馆 CIP 数据核字(2014)第 185681 号

责任编辑：裴育 陈琪杰 杨向萍 / 责任校对：张小霞
责任印制：肖 兴 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京源海印刷责任有限公司印刷

科学出版社发行 各地新华书店经销

*

2014 年 8 月第 一 版 开本：B5(720×1000)
2014 年 8 月第一次印刷 印张：13 3/4
字数：277 000

定价：70.00 元

(如有印装质量问题，我社负责调换)

《信息科学技术学术著作丛书》序

21世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术，数据知识化和基于知识处理的未来信息服务业，低成本信息化和信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前沿交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性；体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国工程院士
原中国科学院计算技术研究所所长
李国杰

序

20世纪70年代，我国发射了第一颗人造地球卫星——东方红一号。发射这颗卫星的重要任务之一就是向太空播放《东方红》乐曲。当电子频率构成的乐音传到人们的耳边时，我们第一次接触了这种有特色的声音，知道了有一种奇妙的数字音乐是用电子脉冲合成的。后来，我们上大学，学的是计算机专业，就和这种数字音乐结识了。在那个年代，我们搞计算机和数字设备的研制，只要涉及研制计算机，成功的重要标志就是计算机唱出了“东方红”。我们研制成功的数字控制机床、针织提花控制计算机都需要在唱出“东方红”时候，方可去报喜，大家再庆功。

改革开放以后，技术发展了，数字音乐表现的内容越来越广泛，表现形式越来越多样化。数字音乐已成为了时代的主流，即使过去的传统唱片、录音磁带，也需要把它们数字化，以便传承与保护。随着互联网的兴起，向生活领域不断拓展，数字音乐发生了巨大的变化。从磁带、光盘这种传统的方式，发展成为网上音乐的时尚。

《中国互联网络发展状况统计报告》显示，截至2013年12月底，我国网民规模达到6.18亿，互联网普及率为45.8%。移动互联网规模更是发展迅速，我国基于移动互联网的数字娱乐发展是极有潜力的。移动互联网正在改变用户的娱乐习惯，手机和各种移动终端成为娱乐的智能设备，使娱乐的方式和消费的成本大大降低。据统计，截至2013年12月底，我国有8亿的移动互联网网民。如果其中80%的网民有听歌的需求，每天人均听半小时，那么每年播放次数、播放时长、商业消费又多么巨大，由此也将带来非常大的社会变化和商业机会。

基于Internet传播方式的发展和众多用户的增加，使人们对“音乐服务”提出了前所未有的要求。无处不在的音乐服务和自然方便的检索方式成为需要突破的两个关键技术。

云计算为音乐的无处不在提供了强有力的技术支持。音乐云计算被认为是云计算领域中最有前景的应用方向之一。云计算由一系列可以动态升级和被虚拟化的资源组成，这些资源被所有云计算的用户共享，并且可以方便地通过网络访问，用户无需掌握云计算的技术，只需要按照个人需要购买云计算的资源。音乐云计算就是将音乐资源放在云端服务器，用户可通过网络按需采用。当前的云音乐服务主要有两种模式，一种是云端音乐存储服务，一种是云端音乐订阅服务。在音乐云计算的上网环境下，音乐爱好者可以通过从云端获取内容，而不必在网络中寻找、下载歌曲，再整理到移动终端上进行收听，因为云端的服务器已经帮助用户完

成了。音乐云计算实现了移动音乐应用的更大智能化，极大程度地发挥了网络互动性特色，将这些音乐以便捷、有效的方式与朋友共享。

云中漫步音乐的另一重要问题是对于云中音乐的挑选。也就是我们常提到的音乐信息检索方式，成为音乐服务中的又一关键技术。

目前多媒体数据库管理方法一般是人工进行分类和检索，不但费力费时，而且对描述音乐来说高度主观、不准确，甚至存在误导。音乐信息检索是从音乐资源中找到满足用户所需信息的匹配、定位过程。传统的基于文本描述的音乐检索技术已经无法满足大量音频数据的检索需要，基于音频内容的检索技术就是为了解决这个问题，包括分类和查询，即利用音乐本身的特征对其进行自动分类，取代手工的文本描述，用哼唱的方法、弹奏的方式进行查询。

基于内容的音频检索是继基于内容的图像检索之后发展起来的一个新兴研究方向，是指通过音频特征分析，对不同音频数据赋以不同的语义，使具有相同语义的音频在听觉上保持相似。基于内容的音乐检索是根据音乐的内容特征进行检索，也就是根据音乐的旋律、节奏等特征进行检索。基于内容的音乐检索在音乐数据库管理、Internet 音乐检索及生活娱乐方面都具有非常重要的意义。现在每年仅中国就有上千张新音乐专辑出现，对于音乐数据库的管理，简单依据人工标注分类已远远不够，这就需要对音乐依照基于音乐内容的分类管理。面对大量涌现的新音乐和海量的经典音乐，对于喜爱音乐的人来说，通过他们熟悉的音乐旋律特征查找音乐是一种更受欢迎的方法。同时，作曲家和音乐家则需要通过音乐的旋律等特征进行音乐的查找和比对，以解决音乐的著作权和版权问题。因此，基于内容的音乐检索技术和检索系统研究具有广泛的理论研究价值和实际应用价值。

检索是人类重要而基本的动作之一。我们有一个模式，按照它去查找符合的结果，这就是检索。如果我们按照原有的语义名查找，那就是传统的查找方法；如果根据已有的内容进行物体多维度的描述，根据物体内容特征查找，那就是基于内容的检索技术。在过去的 20 年里，我们一直在进行基于内容的检索技术研究，从基于图像内容的检索，到基于音频内容的音乐检索、基于三维模型的实体检索、基于视频内容的检索技术，这些研究都不同程度的推动了学术进步和应用的发展。

本书是我和同事们在此方面研究工作的总结，虽颇费时日，但往往浅尝辄止。管见所及，难免有谬误之处，还往智者不吝指正，共同促进这一技术的更好发展。

检索技术随着科学技术的发展和人们需求的增长，方法和技术创新刚刚开始。

路漫漫其修远兮，吾将上下而求索。

周明全

前　言

音频是多媒体信息的重要类型。视频节目中含有大量的音频内容，在歌曲、会议录音、语音通信中都以纯音频形式存在。传统的关键词检索方式已经不能适应大量音频信息多类型的应用需求，迫切需要对海量音频数据进行有效的组织管理和使用。基于内容的音频检索技术应运而生，涉及认知科学、人工智能、模式识别、音频处理和信息检索等多个领域。根据给定的音频片段内容，驱动对大量音频文件的查询过程，可以有效地支持音频审查、音频监听、歌曲检索及数字音频产品的版权保护等应用，具有广阔的前景。

基于内容的音频检索技术已经得到广泛关注，成为研究和应用的热点，国内外学者对此进行了广泛的研究，并已研制应用系统，取得了重要的研究成果与进展。本书作者的研究团队从1998年开始，率先开展了有关音频信号处理、自动音频分类和音乐音频检索的相关研究，先后得到国家自然科学基金等多个项目的支持，培养了多名硕士、博士研究生，发表了40余篇学术论文，设计并实现了多个原型系统。

本书汇聚了该领域的经典成果与前沿研究，介绍了基于内容的音频检索技术分支和国际上相关研究的新成果，给出了音频检索体系与算法上的研究与实例。内容力求体现本领域研究前沿，阐述本领域的理论、技术、方法与发展趋势，并将团队十余年的研究成果贯穿其中，总结了基于内容的音频检索、音乐特征库和基于内容音乐检索等关键技术，分析介绍了适应不同需求的音频检索系统。

内容包括四部分。第一部分概述(第一章)，概括介绍基于内容的音频检索的意义、内容、发展，以及基于内容的音乐检索的研究进展。第二部分基于内容的音频检索关键技术(二～四章)，包括音频信号的特征表示、处理，音频分类技术，音频检索框架和检索算法。第三部分基于内容的音乐检索关键技术(五～九章)，包括音乐检索框架、特征表示、主旋律识别及分割、数据库构建、基于旋律的音乐哼唱检索技术、乐理支持下的音乐检索方法、基于语音识别和情感的音乐检索技术。第四部分音频处理技术的进展(第十章)，分析介绍音频识别、音乐可视化、自动作曲和歌声合成等技术的研究发展。

周明全教授和耿国华教授负责全书的总体安排和定稿。第一章由周明全教授编写，第二章由耿国华教授编写，第六章由王小凤副教授编写，第七章由李鹏博士编写，第三、四、五、八、九、十章由周明全、耿国华、王小凤和李鹏共同编写。王学松、刘晓宁、郭红波等参与了本书的资料整理和书稿校对工作。在本书的写作过程

中，我们先后参阅了很多相关研究文献，在此一并表示感谢！

基于内容的音频检索研究还处于不断发展的过程之中，相关理论和技术尚待进一步探讨，加之作者的实践经验和理论水平所限，书中不足之处在所难免，敬请读者批评指正。

作　者

2013 年 12 月

目　录

第一章　绪　论

随着现代信息技术，特别是多媒体技术和网络技术的迅速发展，多媒体信息的数据量急剧增多，但由于缺乏有效的多媒体检索技术，人们难以充分有效地利用这些海量资源。例如，人们知道巨大的网络信息海洋中有自己需要的歌曲和电影，但却不知道它们到底在哪里。因此，如何在浩如烟海的数据中快速准确地挑选出感兴趣的信息，对于充分利用不断积累的信息资源具有极其重要的意义。

音频是一类重要的多媒体数据，包含大量信息，如何从众多音频资料中检索出需要的信息是一个迫切需要解决的问题，具有非常重要的研究价值。

1.1　基于内容的音频检索技术概述

对声音进行数字化处理和保存得到的结果称为音频。音频媒体是除视觉媒体外最重要的媒体，占总信息量的20%左右。音频信息按内容可以分成语音类和非语音类。语音是人类发出的含语义内容的声音，含有词字、语法等语素，是一种高度抽象的概念交流媒体。非语音包括音乐、音效、非规则声音等，其中音乐是人声和(或)乐器等声响配合构成的一种声音，具有节奏、旋律或和声等语义要素。音效是由声音所制造的效果，是指为增进场面的真实感、气氛或戏剧信息，而加于声带上的杂音或声音。非规则声音则是指没有规律的声音。我们能够听见的音频频率范围是20Hz～20kHz，其中语音大约分布在300Hz～4kHz之内，而音乐和其他自然声响是全范围分布的。

从大量音频文件中查找想要的音频片段就是音频检索，目前音频检索主要分为基于文本关键词的检索和基于音频内容的检索。基于文本关键词的检索主要是采用文件名、文件大小和文件属性等已知的或人工标注的信息进行检索，目前已经发展得非常成熟，已经熟悉的如Google、Baidu和Yahoo等搜索引擎采用的就是这种技术。由于已知的属性和标注的信息有限，不能表示音频所有内容，因此基于内容的音频检索(content based audio retrieval，CBAR)研究应运而生。

基于内容的音频检索是指通过音频特征分析，对不同音频数据赋以不同的语义，使具有相同语义的音频在听觉上保持相似。它主要是研究如何利用音频的幅度、频谱等物理特征，响度、音高、音色等听觉特征，词字、旋律等语义特征实现基于内容的音频信息检索。它涉及多方面领域的知识，包括数字信号、模式识别、统计学习、神经网络和语音识别等。

1.1.1 基于内容的音频检索意义

目前，互联网上主要的音频信息有语音、音乐和结合语音音乐的音频文件等。对于语音，人们有时想找讲述特定内容或某个特定人的讲话部分。对于音乐，人们总是想找到自己喜欢的旋律或情感的音乐。对于语音和音乐的结合体，如广播等音频数据，其中包含了广告、天气预报、主持人主题新闻和新闻详细报告等不同部分。这些部分往往是混合在一起的，不同的人对这些不同部分偏好不同，如果能够将音频按类别分类，可以满足人们对广播新闻进行不同层次需要的检索。同时，像图像和视频一样，人们对相似音频例子的检索需求也很大，总是想从互联网中找到自己需要的音频例子。例如，有些人想找相似的“枪声”，有些人想找相似的“鼓掌声”等。基于内容的音频检索为这些音频媒体检索需求提供了一个新思路，是一种更智能的检索方式。

基于内容的音频检索技术有着广泛的应用前景。

① 它是音频信息搜索引擎的关键技术，用户可通过该技术快速获取所需的信息资源，还可以根据音频信息的内容实现更加灵活的信息搜索策略。

② 它的实现可对音视频点播和网上电视节目等媒体中的音频信息进行实时检索、审查和有效监控，可应用于市场调查、网络管理、信息安全等诸多领域。

③ 它可用于监听，如用声音辅助监测犯罪事件和在医院里监视小孩喊叫、心脏跳动等。陈斯中等[1]将音频多普勒信号的多种参数综合起来用于对孕妇脐动脉血流的诊断，力图准确地判别出胎儿生长发育中存在的异常。

④ 它可用于各种数字音频产品的版权保护，如音乐的版权保护，即搜索未经授权的使用等。

⑤ 它在音频信息分类与统计技术的研究中扮演重要的角色。例如，广播电视新闻节目、学术会议的录音报告、数字图书馆等内容中包含着大量的语音、音乐等信息，使用音频信息检索技术可以有效地对这些信息进行分类、统计与检索，更好地利用这些资源。

1.1.2 基于内容的音频检索概述

基于内容的音频检索需要经过特征提取、音频分割、音频识别、音频分类和索引检索等步骤。它是继基于内容的图像检索之后发展起来的一个新兴研究方向，近年来，已成为国内外研究的热点问题之一，引起了各国众多研究机构和学者的广泛重视。所以，音频信息检索技术已经成为信息检索技术的研究重点之一。

从整体上看，音频内容可划分成三个等级，即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，如图 1.1 所示。在物理样本级，音频内容是以媒体流的形式存在，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。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，可以分为物理特征和感知特征。物理特征包括音频的基频、幅度和共振峰结构等。感知特征表达用户对音频的感知，如音调、响度和音色等。感知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，是音频内容和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。

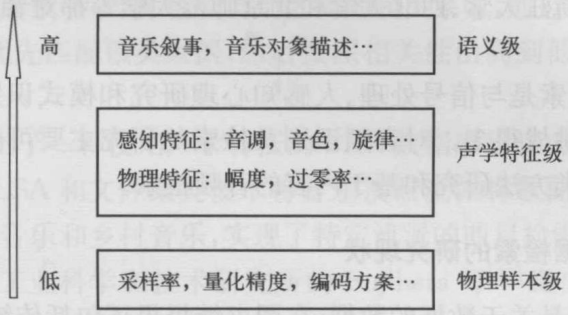

图 1.1 音频内容的抽象层次

在大量、形式多样的音频数据中，如何能够自动、准确和快速地查找到感兴趣的内容，实现基于内容的音频信息检索，是一个既迫切又具有挑战性的研究课题。由于起步晚、数据复杂、研究难度大等原因，音频信息检索技术和文本检索技术相比仍存在很大差距，还有大量问题亟须解决。

1.2 基于内容的音频检索发展

基于内容的音频信息检索技术的研究工作是从 20 世纪 90 年代中后期开始的[2,3]。近年来，它已成为国内外研究的热点问题之一，引起了众多研究机构和学者的广泛重视，如卡内基梅隆大学、马里兰大学、麻省理工学院、康奈尔大学、南加州大学，以及剑桥大学等都对音频信息检索做了大量的研究工作，取得了许多研究成果。这个研究领域中比较重要的期刊和会议包括 IEEE Transaction on Speech and Audio Processing，IEEE Transaction on Pattern Analysis and Machine Intelligence，IEEE Transaction on Multimedia，IEEE Transaction on Signal Processing，IEEE International Conference on Acoustics、Speech and Signal Processing (ICASSP)，IEEE International Conference on Multimedia and Expo(ICME)和 International Symposium on Music Information Retrieval (ISMIR)等。

国外研究机构对音频检索进行了多方面的研究。Muscle Fish[4]是一个商业化的基于音频感知特征的音频检索引擎。卡内基梅隆大学的 Informedia 项目[5]结合语音识别、视频分析和文本检索技术支持视频广播的检索。剑桥大学的

VMR(视频邮件检索)小组利用基于网格的词组发现技术检索视频邮件中的消息。马里兰大学的 Voice Graph[6,7]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。Speech Skimmer[8]是一种音频交互的接口,以层次结构构造出音频文档的"鱼饵"视图。

国内在这方面的研究也很多,李国辉等[9]开发的一套基于内容的音频信息检索与分类系统——ARS 系统。中国科学院声学研究所、上海交通大学、北京大学、微软亚洲研究院、浙江大学、西北大学和北京师范大学等都对音频中的音乐检索做了大量的研究。

音频处理和检索是与信号处理、人感知心理研究和模式识别等学科相关的研究领域,其面临的挑战很多。对音频识别及检索的研究主要可分为传统的元数据检索方式、音频分类方法研究和基于内容的音频检索。

1.2.1 基于元数据检索的研究现状

元数据被认为是关于数据的数据,在图书情报界还包括传统的机读目录格式。音乐元数据方案是从数字音乐信息的外部特征入手的方案。目前国际上出现过多个音乐元数据的研究机构和相关项目,包括国际音乐元数据计划工作小组、MusicBrainz 元数据计划、美国弗吉尼亚大学音乐表示文献类型定义(document type definetion,DTD)和 Musicat DTD、北京大学中文元数据标准框架等。

Pinto 等[10]详细描述了 IEEE PAR1599 (MX)定义的一种采用 XML 实现对音乐、音频、视频进行建模表示的标准。将元数据嵌入相关模型,通过一种新颖的音乐信息检索对象中的结构层次表达音乐和音频的语义信息。

采用元数据形式的音频和音乐信息检索主要依靠关键字符或对文件的外部标注实现,也可通过导航形式实现对特定类型音乐文件的检索。目前基于元数据形式的音频、音乐检索仍然是主流检索形式,但多样化的检索需求和超大规模音乐数据库的增长迫切需要一种具有更高自动化程度和智能程度的自然检索方式,因此基于内容的音乐检索研究逐步展开。

1.2.2 音频分类研究现状

常规的音频分类方法往往通过外部标注实现,特别是音乐文件分类,主要通过乐曲风格分类、艺术家分类、专辑分类等音乐分类方式。对于飞速增长的海量音乐数据,这些分类方式的局限性逐渐凸显出来,已经越来越难以满足用户检索的需求。音乐索引和检索需要新的技术手段来满足用户需求。

基于内容的音频分类主要是针对音频媒体库,采用分析音频文件的声学特征方法对其划定类别。不同的研究课题对音频分类体系的设计也各不相同,目前互联网上主要的音频信息有音乐、语音、结合音乐和语音的音频文件等。对于音乐,