MASTER DISSERTATION XIDIAN UNIVERSITY

代　号　10701　　学　号　0522121173

分类号　TP391

西安電子科技大學

硕士学位论文

题（中、英文）目　基于内容的音频检索的关键技术研究

Study on Key Techniques of

Content-Based Audio Retrieval (CBAR)

作 者 姓 名　潘文娟　指导教师姓名、职务　刘志镜 教授

学 科 门 类　工学　学科、专业　计算机应用技术

提交论文日期　二〇〇八年一月

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

创新性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其他教育机构的学位或证书而使用过的材料。与我一同工作过的同志对本研究所做的任何贡献已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：潘文娟　　日期 2008.3.5

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密的论文在解密后遵守此规定）

本人签名：潘文娟　　日期 2008.3.5

导师签名：[illegible]　　日期 08.3.5

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

摘要

近年来，随着多媒体技术以及网络技术的迅速发展，网络上的信息资源日益丰富，信息检索技术产生了深刻的变革。人们已经不再满足于传统的基于文本的检索，而是需要一种能对视频、图像、音频等各种媒体进行快速检索的引擎。基于内容的音频检索技术（Content-Based Audio Retrieval：CBAR）应运而生。它从音频中直接提取语义线索，根据语义线索进行检索，从而把检索过程与媒体的语义直接联系在一起，使检索工作更加有效，适应性更强。

本文首先介绍了 CBAR 的技术背景和发展过程；然后阐述了一个成功有效的 CBAR 应用的多种关键技术并提出了对现有音频分割分类方法的改进算法；同时给出了检索实验结果及分析；最后指出了系统的不足和未来的发展研究方向。

稳健有效的音频分割与分类是系统检索的前提条件。传统的基于特征阈值的分割与分类方法采用相对简单的特征和前人的经验值，处理的分类问题比较单一。同时，特征阈值的选取也比较困难。本文采用基于高斯模型的分割算法，并给出一种新特征 Mel-ICA，改进了该算法。该方法不需要采集样本，根据特征变化点来进行分割，取得了良好的分割结果。本文还给出一种基于阈值和模型的组合分类方法，结合了这两种方法的优点，同时采用小波变换和傅里叶变换提取音频特征，提高了分类的准确度。

音频的特征和描述是系统的关键。本文结合采用时域、频域和时频域分析方法，从不同角度刻画音频信号的实质，构成音频信号的描述算子。音频检索采用示例音频查询方式（QBE），先使用最小生成树（MST）聚类方法形成关键帧，然后对同类型帧进行匹配比较，减少了计算的强度，大大提高了检索效率。实验结果表明本文方法能够较好地进行音频检索，取得了较好的性能。最后总结了本文的工作，并提出进一步的研究探索方向。

关键词：基于内容的音频检索　特征提取　音频分割与分类　示例音频检索

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Abstract

With growing development of multimedia and network technique, information is becoming rich day-by-day, and information retrieval technique is under grand change. People are not contented with traditional text-based retrieval, instead, they need a retrieval engine which supports the fast retrieval of multimedia data, such as video, image, and audio. Content-based audio retrieval (CBAR) is presented then for audios, which extracts semantic clues from media and retrieval media based on these semantic clues. In this way, the retrieval process is directly related with the semantic of media, and the flexibility and efficiency of retrieval is enhanced.

In this paper, firstly, the developing process of CBAR is introduced, secondly, secondly, some key technology for a successful CBAR is described and an improved audio segmentation algorithm is presented, and then the results of retrieval experiments and analysis are given. Lastly, shortcomings and the direct of study in the future are put forward.

Effective audio segmentation and classification are the preconditions of CBAR. The traditional threshold-based methods mostly adopt relatively simple feature and experienced values. The classification is single and the feature threshold selection is very difficult. Therefore, we adopt the improved Gaussian-based segmentation algorithm and present a new feature named Mel-ICA to improve it. This method does not need samples. The segmentation is implemented according to feature change-points, and good results are achieved. Besides, a classification approach based on threshold and models is proposed, which combines the advantages of these two methods. And the wave and Fourier transform are used to extract features, with the accuracy of classification increased greatly.

Feature description is the key to the system. This study integrates the time domain, frequency domain and time-frequency domain analysis methods, to depict the essential of audio signal, and constitute the description operator. Query-by-example (QBE) is adopted for audio retrieval. Firstly, the (MST) is used for clustering to form key-frames, and the frames with same type are compared to get the similarity. As a result, the calculation complexity is decreased and the retrieval efficiency is increased greatly. The experiments can be concluded that our method can retrieve at the level of objects and achieve good performance. Finally, the paper summarizes the study and declares the direction of further research and exploration.

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Keywords: Content-based audio retrieval Feature extraction
Audio segmentation and classification Query by example

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目录

第一章 绪论 1
1.1 引言 1
1.2 CBAR 的发展和研究现状 2
1.3 本论文研究的内容和方法 3
1.4 本文内容安排 4
第二章 音频数据分析 7
2.1 主流音频文件格式介绍 8
2.1.1 声音文件 8
2.1.2 MIDI 9
2.1.3 模块文件 10
2.2 音频处理技术与特征提取 10
2.2.1 时域特征 11
2.2.2 频域特征 13
2.2.3 时频特征 14
2.3 本章小结 17
第三章 音频分割与识别分类 19
3.1 基于特征阈值的音频分割与分类算法 19
3.1.1 音频分层分割与分类算法 19
3.1.2 双模式的分割与分类算法 20
3.2 基于模型的音频分割与分类 26
3.2.1 基于 HMM 的说话者分割与分类 26
3.2.2 基于高斯模型的音频分割算法 28
3.2.3 音频分割算法改进及实验结果 33
3.3 一种基于特征阈值和模型的组合分类方法 36
3.3.1 基于特征阈值的初始分类 36
3.3.2 采用 SVM 对语音进一步分类 37
3.3.3 实验结果与分析 41
3.4 本章小结 42
第四章 基于内容的音频检索技术研究 43
4.1 音频特征的相似度模型 44
4.1.1 闵氏距离 44
4.1.2 马氏距离 45
4.1.3 余弦距离 45
4.1.4 非几何的相似度方法 45
4.2 哼唱音乐检索 46
4.2.1 音调跟踪 46
4.2.2 检索引擎 47
4.3 示例音频检索 47
4.3.1 基于分类模型的检索算法 48
4.3.2 基于音频模板的算法实现 48
4.4 本章小结 54
第五章 CBAR 系统设计与实验分析 55

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5.1 系统设计概要.....55
5.2 CBAR 系统模块与库结构.....55
5.3 系统开发平台和界面.....57
5.4 实验结果分析.....58
5.5 本章小结.....59
第六章 总结与展望.....61
致 谢.....63
参考文献.....65
在读期间发表论文.....69

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第一章 绪论

1.1 引言

随着信息化社会的到来，人们越来越多地接触到大量的多媒体信息。多媒体信息是人们接触最广泛的一种信息资源，它以文字、图像、声音和视频等各种形式存在，音频信息是这些信息中重要的一种。近 10 年来互联网的应用和发展，促进了多媒体信息的数据量急剧增长。信息量的增长使人们对多媒体信息检索工具和系统的依赖日益加强。而目前在庞大的多媒体数据中检索出自己需要的内容有一定的难度。对音频文件来说，现在通常是把它作为一种不透明的数据集合来处理，仅涉及到它的文件名、文件格式和采样率。如在网络中检索一首歌曲，主要还是以歌曲的名字、歌曲存取格式等来检索。这种基于人工输入属性或描述的方法的缺点是：数据量越来越大，从而人工注释工作量也随之加大；音频感知难以用文字注释表达清楚。如果能开发出一种先进的检索技术，根据音频的内容特征进行有效的组织，给用户提供直观的操作接口，无疑会极大的节省人们的音频查找时间，使得可以快速的定位到自己真正需要的音频。这种迫切的需求推动着研究者们投身于音频检索技术的研究当中。

基于内容的音频检索（Content-Based Audio Retrieval：CBAR）是数据库、多媒体技术前沿的研究方向之一，从上世纪 90 年代起开始成为一个较活跃的研究领域。所谓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。基于内容的音频检索是一个较新的研究方向，由于原始音频数据除了含有采样频率、量化精度、编码方法等有限的注册信息外，本身仅仅是一种非语义符号表示和非结构化的二进制流，缺乏内容语义的描述和结构化的组织，因而音频检索受到极大的限制。相对于日益成熟的基于内容的图像与视频检索，音频检索相对滞后。但它在相当多的领域中具有极大的应用价值，例如，远程教学、卫生医疗、数字图书馆、环境监测、新闻节目检索和娱乐节目的编辑和制作等。这些应用需求推动着基于内容的图像检索技术的研究工作不断深入。

在基于内容的音频检索的研究中，涉及广泛的学科包括：数据库、语音识别、信息检索、音频分析、信号处理、心理声学、机器学习等。基于内容的音频检索综合利用了数据库和计算机听觉研究领域中各方面的技术，同时对这些技术的研究和发展也起到了一定的推动作用。由于基于内容的检索有着广泛的应用前景和市场前景，因而也引起了国际标准化组织的关注。随着多媒体内容描述的国际标准化，音频内容的描述也将随之标准化，音频内容描述及查询语言将成为研究的热点，基于内容的音频检索将朝商业化方向迈进。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.2 CBAR 的发展和研究现状

音频处理是一个涵义甚广的概念，包括音频数字信号处理、心理声学、语音识别、计算机音乐和多媒体数据库。音频处理已有很长的历史，并且取得了一定的成果。不过对于音频的自动分类与检索，研究的并不多。然而实际上，基于音频属性的分类与检索在很多音频和数据库应用系统中有着广泛的应用。

音频处理从 70 年代就已开始了，不过研究重点在于语音识别、说话者鉴别等技术。比如在语音识别方面，IBM 的 ViaVoice[1]已趋于成熟，剑桥大学的 VMK(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，卡内基梅隆大学的 Informedia 项目[2]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，Maryland 大学的 VoiceGraph[3]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，SpeechSkimmer 是一种音频交互的接口，它以层次结构构造出音频文档的“鱼饵”视图。这些都是很出色的音频处理系统，但对于基于内容的音频分类和检索技术的研究还不很多。显然这样的发展是不均衡的。只有在基于音频物理特征的检索技术方面有所突破，才有可能在更高层次的基于知识辅助的音频检索方面做出更深入的研究。

近年来，已有一些公司和研究机构开始进行基于内容的音频信息检索方面的研究，其中美国的 Muscle Fish 公司较早推出了较为完整的原型系统。Muscle Fish 公司先对带标识的数据进行加窗处理，对每帧数据提取音调、响度、亮度、带宽属性，而后对属性序列计算其均值、方差和自相关值，加上能量共 13 个特征，则此 13 维特征即为音频数据的特征矢量，检索时采用马氏距离，比较样本特征矢量与库中数据的特征矢量，从而输出检索结果。Web 上的演示使用了 400 个声音文件，包括动物声、机器声、乐器声、语音和其它自然声。New Zealand 研究音乐曲调和旋律的检索。Jonathan Foote[4, 5]开发了一种基于量化树的方法，它提取音频数据的倒频谱特征 MFCC，并借鉴了语音分析中的方法，利用音频数据的频谱表示并构造一个量化树，最后的特征是一种量化柄的直方图。南加州大学的 Tong Zhang& C.C. Jay Kuo[6, 7, 8]开发了一种启发式的音频数据分割方法，它可以将一段长时的音频数据分段，每一段属于不同的类别，包括静音、语音、音乐、歌曲、带音乐背景的语音和带音乐背景的环境音等。

另外，MIT, Cornell 大学、南加州大学、澳大利亚 Wollongong 大学、欧洲 ELIROMAEDIA 和 Eurocom 的语音和音频处理小组等研究机构分别开展了用子词方法进行语音检索[9]，通过哼唱查询、音频分类、结构音频表示和基于说话人的分割和索引等方面的研究。

相比之下，国内对基于内容的音频检索研究起步比较晚，但已引起广泛的关

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

注和重视，并已有一些研究单位相继展开了相关方面的研究，开发了一些实验系统。浙江大学人工智能研究所对基于内容的音频检索、广播新闻分割等领域进行了深入的研究、在国内处于领先地位，清华大学计算机科学与语音实验室在语音方面的研究，国防科技大学多媒体数据库检索系统方面展开研究，南京大学等也开始了这方面的研究工作。从目前研究和应用的现状来看，基于内容的音频检索技术还处于起步阶段，虽然已开始了基于内容的音频分类与检索方面的研究，但完全自动化和智能化的要求还没有达到。针对海量数据的特点如何快速地进行音频的检索以及如何引入相关性反馈更好的满足用户的检索需求的问题还需要解决。要推出真正实用的基于内容的音频管理与检索系统还有很长的路要走。因此还需要做很多的理论和实践工作，以实现完善实用的基于内容的音频检索。

1.3 本论文研究的内容和方法

基于内容的音频检索技术的检索技术仍处于起步阶段，还有很多技术难题没有解决，目前还没有适合实际应用的完善的音频管理与检索系统。本文的研究主要针对多媒体数据库中对音频信息的分类和检索要求，在分析研究了现有的音频处理和检索方法的基础上，提出了一套基于内容的音频分割分类与检索方法，并设计实现了音频检索的原型系统，如图 1.1 所示。研究和解决的问题主要有：

1. 音频的有效分割；
2. 音频准确分类；
3. 音频检索算法以及相似测度；
4. 音频的查询接口。

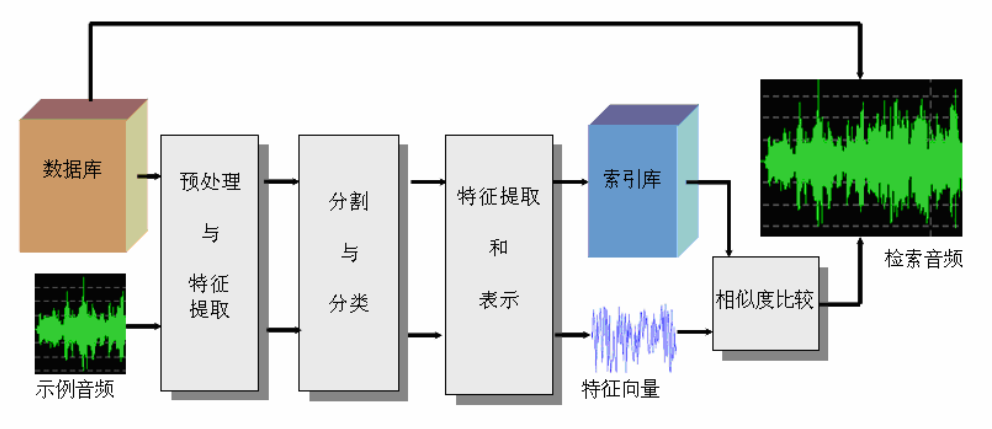


图 1.1 基于内容的音频检索框架图

针对第一个问题，准确稳健的音频分割是音频检索的基础。而用于分割的音频特征的提取又是基础的基础。本文提出一个新的特征——基于 ICA 变换的 mel-ICA，替代用于传统音频分割的特征美尔倒谱系数（MFCCs），能更准确地确

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

定音频变化点。本文采用的是改进的基于高斯模型的音频分割算法[10]，分割效果比采用特征 MFCCs 时有了很大的提高。具体介绍见第三章。

针对第二个问题，音频检索的效率在很大程度上依赖于音频分类，音频的准确分类能提高检索的效率与准确率。而采用的分类特征又直接决定音频分类的效果。音频特征主要有两种，一是感性特征如音高，响度，能量等，一是非感性属性或称物理特性，如美尔倒谱系数。本文综合采用时域、频域和时频的分析方法，提取这两种特征用于分类，更准确更全面地描述了音频的听觉特征。同时提出一种基于特征阈值和模型的分类方法，综合了基于特征阈值方法和基于模型方法的优点，提高了分类的速度和准确率。具体在第三章介绍。

针对第三个问题，音频是一种时序媒体，所以对音频提取的音频特征形成不定长的矢量序列，如何定义矢量序列的距离一直是音频检索算法的重点。本文采用的方法是对此矢量序列进行归一化操作，将其归一化为一个 N 维矢量，维数由系统决定。音频数据的距离归结为此 N 维矢量的距离。本文采用一种基于最小生成树 MST 对音频帧进行聚类的检索算法[11]，减少索引库的存储量，并对相同类型帧进行匹配，大大提高了检索的效率。

针对第四个问题，由于音频与听觉有关，使其难以在计算机上直观的显示。音频对象的特殊性使得常规的基于文本的查询界面无法满足用户的查询要求，音频数据的组织模式和内容描述方法的变化对查询界面和用户接口提出了新的要求，需要引入基于内容的音频检索和查询接口技术。本文分析了基于内容的查询方法和检索中的交互问题，主要采用示例检索（Query By Example：QBE）的形式。

1.4 本文内容安排

本论文第二章分析了音频媒体的特性与分类，介绍了目前主流的几种音频格式，并研究了音频的特征提取技术，包括时域分析、频域分析和时频分析，另外还分析了音频特征如短时平均能量、Mel 倒谱系数等的特性、应用与提取方法。

第三章介绍了基于内容的音频分割与识别分类方法。主要从两个角度展开了音频分割分类识别算法分析研究：基于模型的音频分割和识别算法和基于特征阈值的分割和分类算法。介绍了两种基于模型的分割和识别算法在不同情况下的应用，一是隐马尔可夫模型在说话者分割和识别中的应用[12]，另外一个是基于高斯模型的多变化点音频分割算法[10]，同时给出一种新特征，改进该算法并给出实验结果分析。还介绍了两种基于特征阈值的方法，即传统的非压缩格式的音频分层分割算法[13]和双模式分割算法[14]。最后本文给出了一种基于特征阈值和学习模型的组合分类算法，以一种紧密的方式将特征和模型结合起来对音频进行高效的分类，并通过实验证明该算法获得了较高的音频分类准确率和效率。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第四章分析基于内容的音频检索技术。首先介绍了音频特征的相似度模型及相似度距离计算的几种方法。然后介绍现有的两种主要查询方式，即示例音频检索和哼唱检索，主要介绍了这两种不同查询方式所采用的检索技术。其中详细介绍了两种示例音频检索算法：基于分类模型的检索和基于音频模板的检索算法。还介绍了哼唱”检索在音乐库中搜索最相似音频时采用的匹配方法与技术。

第五章是本文的具体实现工作，应用上述算法，我们实现了一个音频信息分类与检索系统－－CBAR 系统。详细介绍了 CBAR 系统的设计实现、体系结构、功能体现方面，并给出了检索界面和测试结果。在本章的结尾，对 CBAR 系统做了总结评价并指出了有待完善的地方。

在第六章中，总结了以前所作的工作并提出了今后基于内容的音频检索的未来和挑战。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二章 音频数据分析

音频是多媒体中的一种重要媒体，人耳能够听见的音频频率范围是 60Hz～20kHz，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。人耳听到的音频是连续的模拟信号，而计算机只能处理数字化信息，所以模拟音频信号要经过抽样后变成计算机处理的采样离散点，音频信号数字化时采样率必须高于信号带宽的 2 倍才能正确恢复信号。

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别和最高层的语义级，如图 2.1 所示。

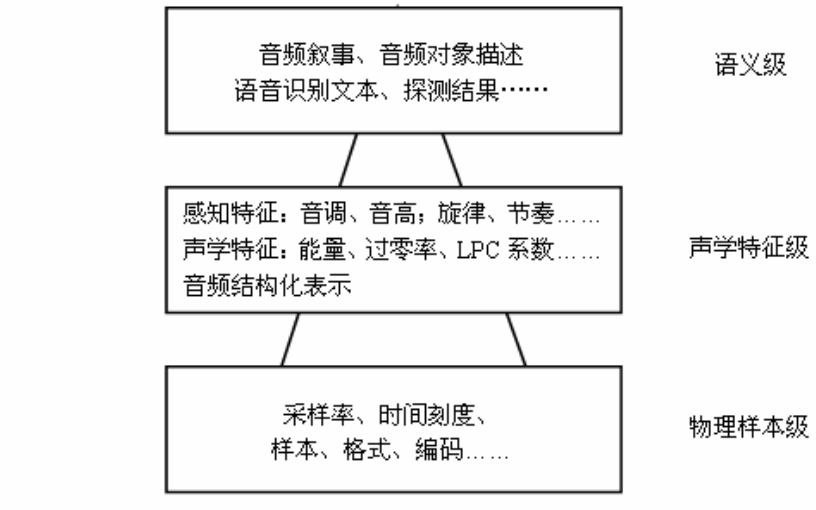


图 2.1 音频内容分层描述模型

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和感觉特征。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。

基于内容的音频检索技术最关心的是声学特征级别和语义级别的音频检索。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。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些步骤。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.1 主流音频文件格式介绍

音频文件通常分为声音文件、MIDI 文件和模块文件三类。前者指的是通过声音录制设备制成的原始文件，直接记录了真实声音的二进制采样数据，其种类很多，文件很大；MIDI 是一种音乐演奏指令序列，相当于乐谱，可以利用声音输出设备或计算机相连的电子乐器进行演奏，由于不包含声音数据，其文件尺寸较小，而后者则是一种记录方式。

2.1.1 声音文件

数字音频是将真实的数字信号保存起来，然后通过声卡恢复成声音。存储声音信息所产生的声音文件是相当庞大的，绝大多数声音文件采用了不同的音频压缩算法，在基本保持声音质量不变的情况下尽可能缩小文件。数字音乐制作过程中编码形式的迥异，造成了格式大相径庭，由此造就了一个庞大的数字音乐家族，现将常见的几种格式进行简要介绍。

WAV 是一种古老的音频文件格式，由微软开发，符合 RIFF 的规范[15]。所有的 WAV 都有一个文件头，这个文件头音频流的编码参数。WAV 对音频流的编码没有硬性规定，除 PCM 外，几乎所有支持 ACM 规范的编码都可以为 WAV 的音频流进行编码。由于 Windows 本身的影响力，这个格式已经成为了事实上的通用音频格式。在 Windows 平台下，基于 PCM 编码的 WAV 是被支持得最好的音频格式，所有音频软件都能够完美支持，由于本身可以达到比较高的音质要求，因此，WAV 也是音乐编辑创作的首选格式，适合保存音乐素材。WAV 作为一种中介的格式，常常使用在其他编码的互相转换之中，例如 MP3 转换成 WMA。通常我们使用 WAV 格式都是用来保存一些没有压缩的音频，因此它的文件很庞大，一般都在几 MB 以上。也正因为没有采用压缩技术，WAV 文件中声音的采样数据很容易被读出来，便于做其它处理[16]。例如：画出声音的信号波形、做出频谱等。现在的应用程序几乎都支持 WAV 文件格式，也有专门软件可以完成从 WAV 文件格式向其它文件格式的转换，因此 WAV 文件在目前仍然有着相当广泛的应用价值。

上面介绍了 WAV 格式文件，下面介绍一种目前在互联网上应用非常广泛的音频文件格式 MPEG。MPEG 是运动图像专家组（Moving Picture Experts Group）缩写，为 MPEG 多媒体压缩标准。这里的音频文件格式指的是 MPEG 音频层。

MPEG 是一种有损的，非平衡编码[17]。有损意味着为达到低比特率，采用了基于听觉心理的压缩模式，人耳最不敏感的一些伴音信息将丢失；非平衡编码意味着其压缩编码过程比解码过程慢的多。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在 MPEG 对任何类型音频编码时，首先通过 32 个过滤器组将原始音频流转换成对应频谱分量，同时运用心理生理学模型来控制每一子带的位分配，通过对各个子带编码来实现原始信号编码。由于 MPEG 编码是非平衡编码，因此相对于复杂而又耗时的编码过程，其解码过程是十分简单的：各子带的序列按照分配的信息被重建,然后各子带的信号通过一个合成过滤器组生成32个连续的16位PCM格式的声音信号。详细算法可参见 ISO MPEG 的相关标准。

MP3 是 MPEG Audio Layer3 的简写，是 20 世纪 90 年代开发成功并得到 Fraunhofer IIS 大力支持的一种常用于播放器的有损压缩编码格式。它是利用人耳的掩蔽效应对声音进行压缩，使文件在较低的比特率下，尽可能地保持了原有的音质，是目前最为流行的压缩方式，也是现在网上收集音乐的最主要的方式，大多数播放器都支持这一文件格式。MP3 格式的声音文件的压缩比达 10∶1-12∶1。不过 MP3 对音频信号采用的是有损压缩方式，为了降低声音失真度，采取了“感官编码技术”，即编码时先对音频文件进行频谱分析，然后用过滤器滤掉噪音电平，接着通过量化的方式将剩下的每一位打散排列，最后形成具有较高压缩比的 MP3 文件，并使压缩后的文件在回放时能够达到比较接近原音源的声音效果。在不小于 128kbps 传输率下，基本保持了原有音质，正是这一特性，使得 MP3 相关产品保持着长盛不衰，而由 MP3 格式衍生出来的其他格式文件也很多，基本目的都是在保持原有音质的情况下，降低文件的传输率。

2.1.2 MIDI

MIDI 是乐器的数字化接口（Musical Instrument Digital Interface）的缩写，是电脑多媒体技术在音频领域中的又一应用。整个 MIDI 系统包括合成器、电脑音乐软件、音源、电脑、MIDI 连线、调音台、数码录音机等周边设备。电脑可以将来源于键盘乐器的声音信息转化为数字信息存入电脑。它规定了不同厂家的电子乐器与计算机连接的电缆和硬件及设备之间数据传输的协议，协议中规定采用数字方式对乐器演奏出来的每个音符作为一个记录，然后播放时采用以下两种方式对记录进行合成：1）PM 调频合成：PM 合成是通过多个频率的声音很和来模拟乐器的声音；2）波表合成：波表合成是将乐器的声音样本存储在声卡的波形表中，播放时再从波形表中取出。

MIDI 分为普通 MIDI1 和 MIDl2 两个版本，不管哪个版本，其声音质量比真实乐器演奏出来的声音都要差很多。MIDI 文件包含 MID 和 RMI 两种格式，还有 XMI 格式，可用于为不同乐器创造数字声音，模拟一些常见乐器。在 MIDI 文件中，只包含产生某种声音的指令，包括使用什么设备的音色、声音的强弱、声音持续多长时间等，计算机将指令发送给声卡，按照指令声音被合成出来。在

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

重放时可以有不同的效果，取决于音乐合成器的质量。相对于保存真实采样数据的声音文件，显得更加紧凑，其文件通常比声音文件小的多。

2.1.3 模块文件

模块（Module）格式是存在了很长时间的声音记录方式，同时具有 MIDI 和数字音频的共同特性。既包含如何演奏乐器的指令，又保存了数字声音信号的采样数字，其声音回放质量对音频硬件的依赖性较小，在不同的机器上可以获得基本相似的声音回放质量。根据不同的编码方式有 MOD、XM、MTM、KAR、IT 等多种不同格式。

2.2 音频处理技术与特征提取

音频信号携带各种信息。在不同应用场合下，人们感兴趣的信息也不同。比如对于语音来说，判断一端语音是否为语音，只需提取人类语音信号的一般特征就足够了；而为了区分是清音还是浊音，就应该了解其能量谱分布和基音频率。

为了满足音频管理和检索的需要，需要提取音频的低层特征来表现音频。音频特征分为两种：一是听觉感知特征如音调，响度等；一是非感知特征或称物理特性，如线性预测系数，MFCCs 等。不同的特征表达音频的不同方面，适用不同应用范围。

音频特征提取的手段是数字信号处理技术，通常它可以分为时域分析、频域分析及时频分析三种，其中时域分析方法主要针对音频信号的波形，频域分析的方法主要是涉及某种形式的音频频谱表示，时频分析是结合时间域和频率域对音频进行表示和分析。以时域度量音频信号的例子有能量和自相关函数等。频域分析方法主要有傅立叶分析（FFT），以频域度量音频信号的特征主要有带宽、频率中心和 MFCCs 等。时频分析方法主要有短时傅立叶变换以及小波变换方法等。这三类特征空间从不同角度刻画音频信号的实质，构成音频的描述算子。

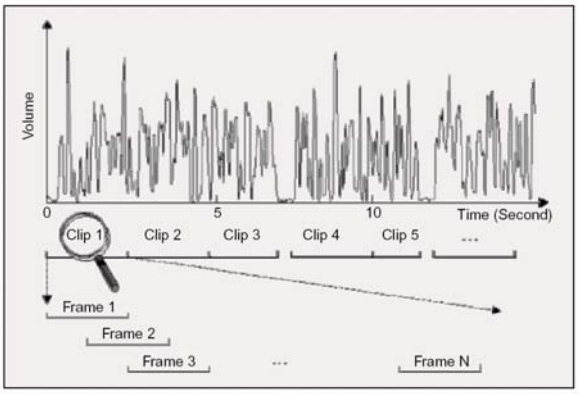


图 2.2 音频信号的帧和段

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

音频信号本质上是非稳定的，也就是说，相隔很短时间音频信号就会发生明显的变化。但是如果在音频处理时对每个采样点都进行处理，又是不现实的。比较常用的方法是使用“帧”，一帧大概持续 10-40 毫秒，包含上百个采样点。音频帧的概念来自于传统的语音信号处理，其前提假设是音频信号在帧内是平稳的，即音频信号的短时平稳性。这样在音频处理时就可以保证基于帧提取的各种内容特征基本稳定。在语音信号处理中，已经发现基于短时信号的分析是最合适的。但另一方面，对于反映音频信号语义内容（如音乐、掌声、爆炸等）的特征来说，基于一个较长时间段（几秒到几十秒）的分析也是必要的[18]。在这里，我们把这种较长的时间段称为“段”。一个音频段包含若干帧，段的特征通常表示帧特征在段内的统计变化，如均值和方差等。如图 2.2 所示，帧和段都可采用前后交叠的方式布局。

2.2.1 时域特征

现假设连续音频信号 x 经过采样后，得到 K 个采样点 $x(n)(1\le n\le K)$。在音频时域特征提取中，认为每个采样点 $x(n)(1\le n\le K)$ 包含了这一时刻音频信号的所有信息，所以直接由 $x(n)(1\le n\le K)$ 提取音频特征，而不需要对 $x(n)(1\le n\le K)$ 做任何进一步处理。

采用这种处理方法，将 $x(n)(1\le n\le K)$ 序列看成个二维数轴，横坐标表示时间（其长度为 K），纵坐标表示 $x(n)(1\le n\le K)$ 的值。下面分别介绍音频信号的短时平均能量，过零率和线性预测系数等时域特征。

1．短时平均能量（STE）

STE 指的是在一个短时音频窗口内采样点信号所聚集的平均能量。假定每个短时帧的大小为 N， $x(n)$ 为用奈奎斯特频率采样后的离散音频信号。对于第 m 个短时帧，短时平均能量可以使用下面的公式计算：

$$E(m)=\frac{1}{N}\sum_{m}(x(n)w(n-m))^2 \qquad \text{式(2-1)}$$

其中 $x(n)$ 表示离散时间音频信号，$w(n)$ 表示长度为 N 即含 N 个采样点的窗函数。

STE 可以较好地表示音频信号幅度随时间的变化。应用 STE 特征的主要原因可概括为如下三点：

1）对于纯语音信号，STE 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 STE 通常明显小于浊音成分的 STE；

2）当音频信号的信噪比较高时，STE 可以有效地区分其中的静音部分；

3）STE 随时间的变化，可以反映音频的节奏、周期等属性。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

STE 可以直接应用到静音检测中。如果音频中的某一帧的平均能量低于一个事先设定的阈值，则判定帧为静音，否则为非静音。如果音频中的静音帧数目超过了一定比例，则把这个音频归为静音音频。

2. 过零率（ZCR）

ZCR 描述过零的速度，是信号频率量的一个简单的度量，它指的是在一个帧内，离散采样信号由正到负和由负到正的变化次数。ZCR 能够大概反映信号在帧内的平均频率。对于音频信号流 x 中第 m 帧，其 ZCR 计算如下：

$$ZCR(m)=\frac{1}{2}\sum_{m}\left|\mathrm{sgn}[x(n)]-\mathrm{sgn}[x(n-1)]\right|w(n-m) \qquad \text{式(2-2)}$$

其中，$x(n)$ 表示第 m 个帧信号中第 n 个采样信号值，$w(n)$ 是长度为 N 的窗口函数。当 $x(n)\geq 0$ 时，$sign[x(n)]=1$；否则 $\mathrm{si}\,gn[x(n)]=0$。

比较而言，语音信号比较规范，一般是由几个单词构成，每个单词又由元音和辅音交替的音节组成。语音产生模型指出，由于声道阻碍较大，所以辅音的能量集中在 3KHz 以下，所带能量较小；相反，由于受声道阻碍较小，元音所带能量较大。这样，对于语音信号，在波形上表现为较短时间内的低能量辅音信号总是后继一个较长时间高能量元音信号。相应的，辅音信号的 ZCR 低，而元音信号的 ZCR 就高[19]。

语音信号开始和结束都大量集中了辅音信号，所以在语音信号中，其开始和结束部分的ZCR总会有显著升高，所以利用ZCR可以去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 ZCR 不表现出突然升高或坠落的跌宕特性，所以有时也用 ZCR 来区分语音和音乐信号。

3. 线性预测系数

对于采样后得到的信号序列$\{s_i\}$，人们总想用一个模型来模拟它的产生。如果用有限个参数的线性模型来近似表示音频序列$\{s_i\}$，这些参数就可以成为描述序列的重要特征，称为线性预测系数。在该线性预测模型下，对下一个样本的预测可以表示之前样本的加权和：

$$\hat{s}_n=\sum_{i=1}^{p}a_i s_{n-i} \qquad \text{式(2-3)}$$

其中$\{a_i\}$为线性预测系数。在实际应用中，主要是对帧内的采样序列建立一个最佳预测模型。一般采用最小均方差的方法。帧内的预测误差定义为

$$e=\sum_{k}(s_k-\sum_{i=1}^{p}a_i s_{k-i})^2 \qquad \text{式(2-4)}$$

假设预测误差取最小值，求解得到最佳预测模型的 p 个参数$\{a_i\}$作为帧的 LPC 特征。

线性预测模型最大的优点是容易计算，模型求解为线性问题。缺点是由于音

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

频信号本质上是一个非线性过程，使用线性模型只是近似计算，精度不高。近来人们越来越重视音频信号非线性分析方法的研究，但是由于非线性模型普遍受到计算复杂、稳定差等问题的困扰，目前还没有一种令人接受的非线性分析方法。

2.2.2 频域特征

音频理论指出：每一个音频信号是由不同时刻、不同频率和不同能量幅度的声波组成的，人们之所以感受到音频信号，是因为人耳这个滤波器在不同时候感受到了不同频率带上不同能量信号的结果。音频是不同频率在不同时刻所附带的不同能量形成的。每个时刻的采样信号 $x(n)(1 \le n \le K)$，只代表部分信息，音频信号的其他信息，需要经过频域分析才能得到。

我们通常采用傅立叶变换将原始音频信号从时域转换到频域，再把音频信号用具有不同频率和幅度的谐波构造出来，然后对这些谐波进行特征系数提取。音频信号频域特征有多种，常用的有频率中心、带宽和 Mel 频率倒谱系数等。

傅立叶变换（Fourier transform）是法国科学家 Joseph Fourier 提出的一种数学方法，可将时空信号变换成频率信号，得到了工程技术领域的广泛应用。

原始的音频数据为时空信号，在时空上有最大分辨率，并可利用时空上的相关性进行数据压缩。Fourier 变换可将时空域中的音频信号映射到频率域来研究，更符合人类听觉特征，也可以利用信号在频率域中的冗余进行数据压缩。Fourier 变换所得的频率信号，在频率域上有最大分辨率，但其本身并不包含时空定位信息。

设一维和二维的时空信号分别为：$f(t),\ t \in (-\infty, +\infty)$（如图 2.3 所示）和 $f(x,y),\ x,y \in (-\infty, +\infty)$。那么 Fourier 变换后的频率信号分别为：$F(w) = \int_{-\infty}^{\infty} f(t)e^{-jwt}dt$（如图 2.4 所示）和 $F(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-j(ux+vy)}dxdy$。

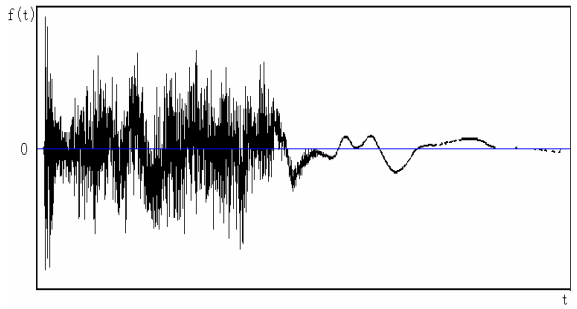


图 2.3 音频信号的时间波形图

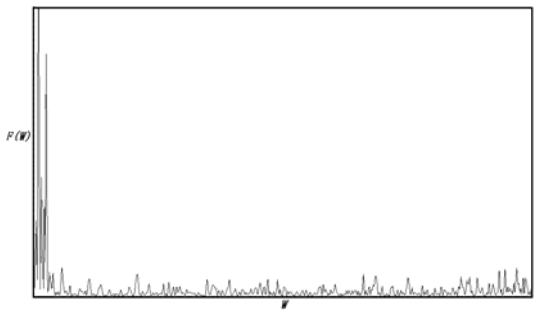


图 2.4 音频信号的频率图

1．频率中心和带宽

频率中心 Centroid 表示频谱分布，是傅立叶变换的频率中心。在压缩域中，它被认为是绝对频谱的平衡频率，其计算公式为：

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\omega_c = \int_0^{\omega} u\left|F(u)\right|^2 du \Big/ \int_0^{\omega} \left|F(u)\right|^2 du \qquad \text{式(2-5)}$$

带宽是频谱成分和频率中心的平方差的能量权重平均值的平方根，如下式：

$$B = \sqrt{\int_0^{\omega} (u-\omega_c)^2 \left|F(u)\right|^2 du \Big/ \int_0^{\omega} \left|F(u)\right|^2 du} \qquad \text{式(2-6)}$$

2．Mel 倒谱系数（MFCC）

Mel 倒谱系数模拟了人耳的听觉特性，在语音识别实际应用中取得了较高的识别率。MFCC 在一定程度上模拟了人耳对语音的处理特点，应用了人耳听觉感知方面的研究成果，在有信道噪声和频谱失真的情况下具有较好的稳健性。

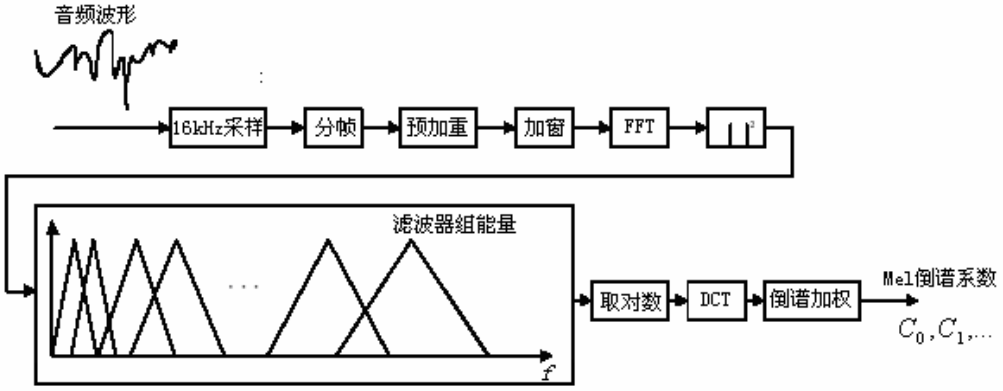


图 2.5 MFCC 的提取过程

MFCC 的典型提取过程如图 2.5 所示。通常要对帧内信号进行预加重以提升高频，然后进行加窗处理以避免短时语音段边缘的影响，常用的窗函数有汉明窗、矩形窗等。然后对处理后的采样信号进行快速傅里叶变换，得到音频帧在每个频率上的大小。为了表达人耳的感知特性，需要把一般频率上的能量映射到更加符合人类听觉的 Mel 频谱上，通过一组三角滤波器实现，它们在 Mel 频谱上是等间隔的。每个滤波器的输出就是该频率段上的能量系数，称为子带能量。为了更加有效的表示，还需要对能量系数取对数，并进行离散余弦变换，最后得到的系数就是 MFCC 特征。

2.2.3 时频特征

在信号处理中，传统的傅立叶分析对确定性和平衡性消耗的分析发挥了重要作用。但是，在现实生活中，某些信号具有很强的时变性，如在某一段短时间内呈现出周期信号的特征，而在另外一些时间段却呈现出噪声特性。对于这些时变剧烈的音频信号，仅仅在频谱空间上面进行的傅立叶分析有一定的局限性。这就需要对信号进行时频分析。

鉴于傅立叶变换不含时空定位信息，Dennis Gabor 于 1946 年提出短时傅立叶变换，可用于时频分析，但窗口大小是固定的。1984 年， Jean Morlet 和 A.Grossman

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

又提出了具有可变窗口的自适应时频分析方法——小波变换[20]。近年来，小波变换已成为了对信号进行时频分析的一个重要工具，其应用领域也越来越广。

1．短时傅立叶变换（STFT）

为了弥补傅立叶变换不能时空定位的不足，可以采用短时傅立叶变换，来对时空信号进行分段或分块的时空-频谱分析，即时频分析。

短时傅立叶变换公式为：

$$F_g(\tau,w)=\int_{-\infty}^{\infty} f(t)\overline{g(t-\tau)}e^{-jwt}dt \qquad \text{式(2-7)}$$

其中，g 为窗口函数。常用的窗口函数有矩形窗、海明窗、高斯窗等。

虽然短时傅立叶变换能部分解决傅立叶变换时空定位问题，但由于窗口的大小是固定的，对频率波动不大的平稳信号还可以，但对音频、图像等突变定信号就成问题了。本来对高频信号应该用较小窗口，以提高分析精度；而对低频信号应该用较大窗口，以避免丢失低频信息；而短时傅立叶变换则不论频率的高低，都统一用同样宽度的窗口来进行变换，所以分析结果的精度不够或效果不好。迫切需要一种更好的时频分析方法。

2．小波变换

近二十年来发展起来的小波分析是一种时频分析方法，具有多分辨分析功能，被誉为数学显微镜。与幂级数、三角级数或傅立叶级数等一样，小波分析研究用一组简单函数来表示任意函数。如三角级数/傅立叶级数可表示为：

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{l}+b_n\sin\frac{n\pi x}{l}\right)=\sum_{n=-\infty}^{\infty}c_n e^{j\frac{n\pi}{l}x} \qquad \text{式(2-8)}$$

其中 $c_n=\frac{1}{2}(a_n-jb_n),\ c_{-n}=\frac{1}{2}(a_n+jb_n),\ e^{j\theta}=\cos\theta+j\sin\theta,\ j=\sqrt{-1}$。

被表示的函数的全体构成一个函数空间，而表示这些函数的函数族{x^n}与{sin nx，cos nx}等则为函数空间的基底。函数展开式中的系数为该函数在函数空间中相对于此基底的坐标，对应于函数空间的一个点。这相当于将函数从原来的域变到新的域，如三角级数将时空域的函数变换到频率域。

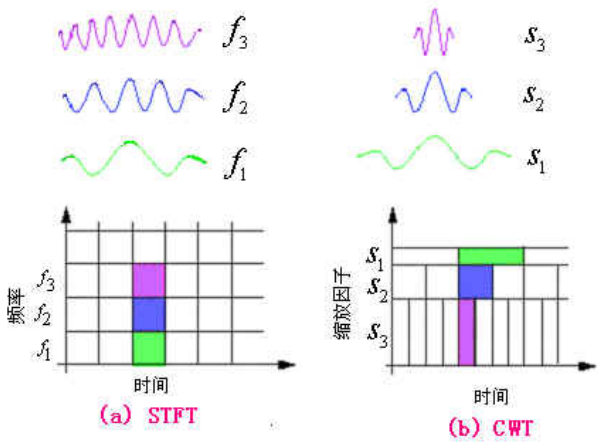


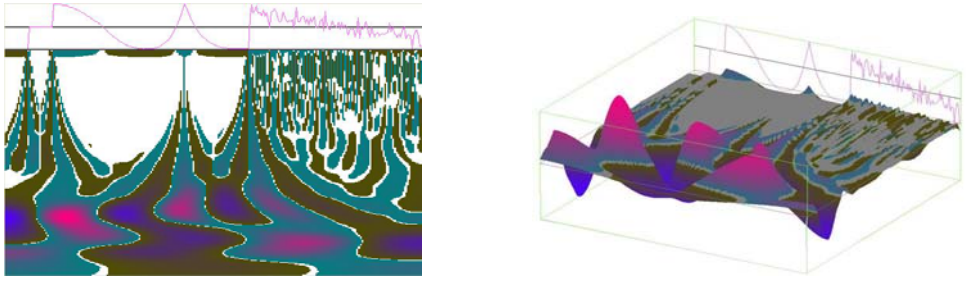
图 2.6 FFT 与小波变换的时频特征

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

如图 2.6 所示，小波变换的特点有：

- 时频局域性、多分辨分析、数学显微镜
- 自适应窗口滤波：低频宽、高频窄
- 适用于去噪、滤波、边缘检测等

与傅立叶变换不同，小波变换的结果有两个参数，多了一个可以表示时空位置信息的平移因子，所以其图示为一个二维曲面。图 2.7（a）和（b）分别为 Mallat 构造的一组典型数据的曲线及其连续小波变换曲面的二维和三维图示：



（a） （b）

图 2.7 Mallat 数据及其连续小波变换的二维和三维图示

小波变换与傅立叶变换的变换核不同：傅立叶变换的变换核为固定的虚指数函数 e^{-jwx}，而小波变换的为任意母小波 $\psi(x)$。前者是固定的，而后者是可选的。实际上，母小波有无穷多种，只要满足一定条件即可。常见的小波函数有：

■ Haar 小波（Alfred Haar，1910 年）：$\psi(x)=\begin{cases}1, & 0\le x<0.5\\ -1, & 0.5\le x<1\\ 0, & \text{其他}\end{cases}$，见图 2.8。

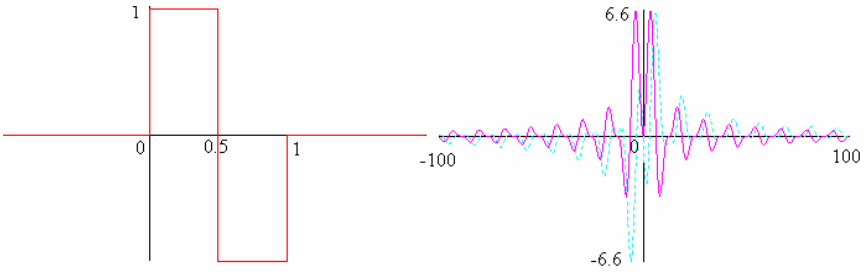


图 2.8 Haar 小波函数及其 Fourier 变换

■ Morlet 小波（Jean Morlet，1984 年）：$\psi(x)=e^{jCx}\cdot e^{-\frac{x^2}{2}}, C\ge 5$，见图 2.9。

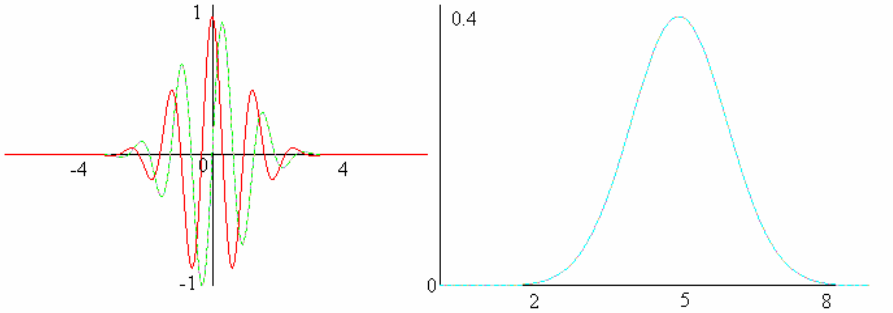


图 2.9 Morlet 小波函数（C=5）及其 Fourier 变换

除 Haar 小波外，其他几种小波都不是初等函数，有的小波函数是用导数/积

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分或微分方程/积分方程来定义，或用其傅立叶变换定义，有的甚至没有解析表达式，而只是一些数字解，很多小波为复函数，所以不太直观。

3．小波特征系数提取

作为音频信号特征的一种补充，小波系数在音频检索中已有成功应用，如使用音频小波系数特征进行音乐和音频例子检索。实际中，小波系数都是音频帧中的采样信号经过小波变换后得到的。

语音信号频率范围在 8000Hz 以内，汇集了大部分能量；音乐信号在 16kHz 频率范围以内，能量分布比较平均；而对于爆炸和钟声等环境背景音信号，其能量集中在高频部分而在低频部分能量很低。小波变换等效的频域表示是：

$$W_f(a,b) = < f, \psi_{a,b} > = |a|^{-1/2} \Big/ 2\pi \int_R F(\omega)\psi(a\omega)e^{j\omega^* \pi} d\omega \qquad \text{式(2-9)}$$

如果$\psi(\omega)$是幅频特性比较集中的带通函数，则小波变换便具有表征分析信号$F(\omega)$频域上局部性质的能力。从频域上看，用不同的尺度做小波变换相当于用一组带通滤波器对信号进行处理。使用幅频特性比较集中的带通函数设计低通和高通滤波器，得到音乐、语音和环境背景音在这些频率子带上的能量，由于音乐、语音和环境背景音在不同频率带上集中的能量多少很不相同，就可以使用这些不同频率带上的能量来粗略地分出它们。

2.3 本章小结

本章研究了音频文件格式与音频信号特征提取与表达。详细介绍了现今流行的四种音频文件格式 WAV、MP3、AAV 和 MIDI 等，并分别从时域、频域和时频域介绍常用的音频特征的提取与表达方法。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第三章 音频分割与识别分类

和图像不同，音频是连续的时间序列信号，具有时间上的结构和语义内容。一段连续的音频序列可能由不同类型的音频单元构成，如语音、音乐、环境背景音等。因此在进行内容分析时，需要对音频序列进行分割，然后对这些得到音频片断进行识别，把相似的片断归为一类。

对音频的分割可以有两类方法。第一类方法利用不同类型音频信号转换时某些听觉特征会发生较大变化的现象，在特征发生突变的地方对音频序列进行分割，再进行后续处理。这种方法需要预先确定不同特征之间变化的阈值，根据阈值判断是否应该分割。它的计算比较简单，缺点是特征阈值的选取比较困难，主要依靠前人的经验值，并且对于不同的应用，阈值也是不确定的。第二类方法是基于模型的。它通过训练模型去模拟某类音频的动态变化，然后根据这个模型的变化确定最佳的分割。这其中，目前应用比较成功的是隐马尔可夫模型（Hidden Markov Model：HMM）和高斯模型（Gaussian Model）。

对分割后的音频片断进行识别属于模式识别的问题。可以采用各种模式分类的方法进行识别，较常用的如 k-近邻法和支持向量机（Support Vector Machine：SVM）等。另外，HMM 也可用于识别分类。还可以直接采用特征阈值对音频进行分类。在下面的章节中将分别具体介绍 HMM、高斯模型、SVM 在音频分割和分类中的应用，以及如何采用特征阈值对音频进行分割和分类。

3.1 基于特征阈值的音频分割与分类算法

自动的音频分割和分类是互相依赖的。好的分割需要好的分类，好的分类依赖于好的分割结果。下面介绍的两种算法都是在分割基础上进行分类的，而分割又是根据分类结果进行的。传统的基于规则的分类算法是根据一种或者几种音频特征及其阈值来判定音频所属的类别，主要有音频分层分割分类算法和双模式的分割分类算法等。

3.1.1 音频分层分割与分类算法

音频分层分割算法[13]是根据音频特征分割音频的算法，当一种音频转换成另外一种音频时，主要几个特征会发生变换，每次选取一个发生变换最大的音频特征，从粗到细，逐步将音频分割成不同的音频例子。

当音频信号在不同种类音频之间转换时，不同特征之间所存在的差异是很不

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

同的[13]。比如，当从C_i，这个音频类别转换到C_j这个音频类别时，特征f_i和f_j之间的差别较大，所以需要比较f_i和f_j之间是否发生了很大变换，不需要比较其他特征之间是否发生变换，如果f_i和f_j之间发生了很大变换，则对音频在发生变化处进行切分；而当从C_i这个音频类别转换到C_k这个音频类别时，特征f_i和f_k之间的差别较大，特征f_i和f_j之间的差别不很明显。所以只需要比较f_i和f_k之间的差别是否较大，不需要比较其他听觉特征之间的差异，如果f_i和f_k之间的差别较大，则对音频流进行切分。

这样，在进行音频切分时，要分层次考虑不同音频特征之间的差异，从而划分出不同的音频信号。当然，在实际应用中，当音频发生转换时，一个特征之间的差异不是很明显，几个音频特征之间的差异累加起来就比较明显了，所以，在分割中，也常常使用组合音频特征实现分割。

实现音频信号流分层分割，关键是找到能够明显区分不同类别音频信号的特征或特征组合，然后通过比较特征之间的差异是否超过了一定阈值，将连续音频信号流分割出实现预定的音频例子。

提取每个音频帧的静音比率（Silence Ratio）、频率质心（Spectrum Centroid）、谐波（Harmonic）和音调（Pitch）四个特征。语音、音乐和噪音由于发音等的不同，分别具有不同的特征，我们可以比较前后相邻若干个帧某个或某些特征是否发生了明显变换，将得到的特征变换值与给定阈值做比较，逐步对连续音频信号流进行切分，分别得到语音（Speech）、音乐（Music）和噪音（Noise）三类音频例子[56]。

这种算法只是完成对音频的粗分，但是这种算法不能对音频进行进一步细分。比如，粗分只能够把枪声，钟声，警笛声，瀑布声，拳击声，尖叫声，鸟叫声，下雨声，鼓掌声和笑声等都笼统归为噪音，不能识别出到底是哪种噪音；还有，虽然这种方法可以把语音和音乐分割开来，但识别不出来到底是谁在说话。分割出来的音频单元需要进一步精细分类，这是模式识别需要完成的任务。

3.1.2 双模式的分割与分类算法

上节详细介绍了传统的非压缩格式的音频分割方法，而通常音频数据库包含了多种格式（压缩/非压缩）、编码方式（MP3、ADPCM 等）、编码参数（采样频率、位速等）的音频文件，因此需要一种综合的音频分割方法，既可以处理压缩音频也可以处理非压缩格式的音频文件。本节介绍一个双模式音频分割分类系统[14]。它有两种工作模式，位流模式主要用于处理压缩格式的音频文件如 MP3、AAC，直接使用位流信息来进行分类和分割，计算速度较快；一般模式用于处理非压缩格式的音频文件如 PCM，从 PCM 样本中提取瞬时的和频谱信息然后对其进行分

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

析，准确率较高。最后将音频分为四大类型：语音、音乐、模糊和静音，其中模糊是混合类型，包括带语音的音乐和各种环境噪声等。该双模式结构支持大规模多媒体数据库的基于音频的检索，图 3.1 为系统框架图：

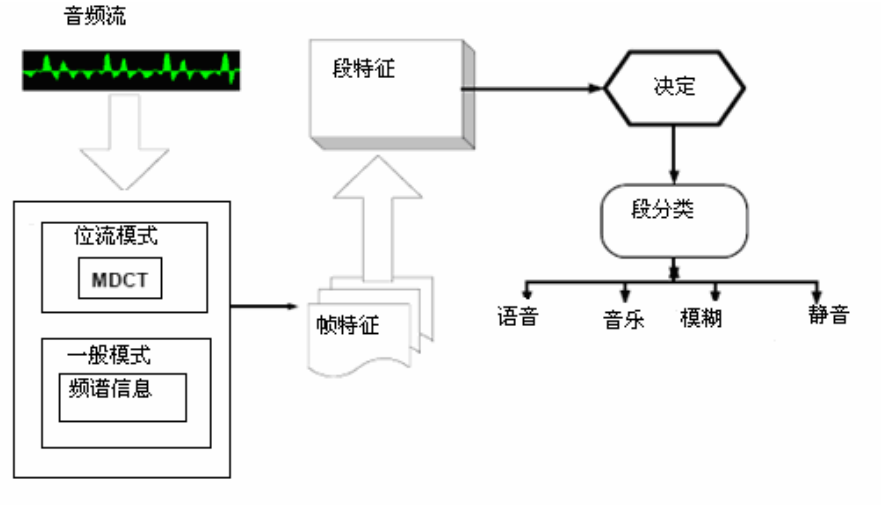


图 3.1 一种通用的音频分段和分类框架

双模式分割系统主要包括三大模板：频谱模板的形成、特征提取以及分割分类算法实现。其中频谱模板的形成是将非压缩格式和压缩格式统一起来，形成标准的频谱模板以提取用于后续分割和分类的特征。特征提取包括帧特征和段特征提取，分别用于帧和段分类。分割和分类最主要的方法是基于特征域的感知建模，根据特征的特性和感知规则对特征空间进行划分。

3.1.2.1 算法原理

双模式的分割分类的原理是基于特征域的感知建模，主要在三个重要的段特征上进行：转换率、基音频率和子带频谱中心段特征。根据段特征的特性，模型提供特征域的基于感知规则的划分，如图 3.2 所示。

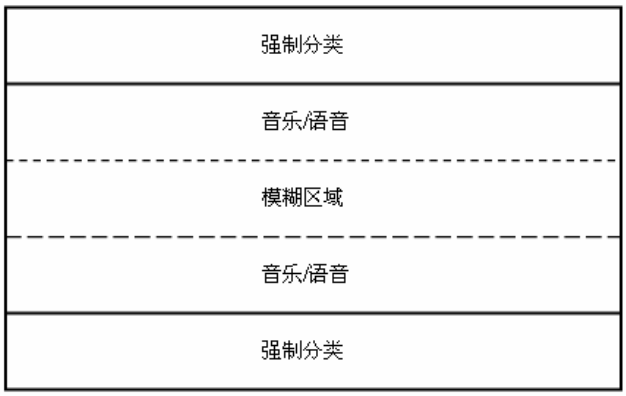


图 3.2 特征域的感知建模

如果某个特征值是一个极限值，可以确定内容的类型，那么就会强制给定该内容的类型，而忽略其它的特征。重要特征的强制类型分别定义如下：

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

- 当转换率在15%以上时，强制划分为语音。
- 当基音频率段特征大于2KHz的时候，被强制划分为音乐。
- 子带频谱中心段特征有两个强制分类的区域：当子带频谱中心平均值超过2KHz时为音乐，当子带频谱中心段特征值超过1200Hz时为语音。

强制分类下面的区域，自然分类会将它分为纯类型如语音或音乐。对于所有的段特征，有来自纯类型的典型值，但是仅靠这个特征值又还远不能确定最后类型。最后，还有一个模糊区域，特征值在该区域不再可靠。另外，在分类真正发生的区域间存有模糊区域可防止大部分关键错误，这种错误可能会由于噪音从一个纯类型区域跳到另一区域而发生。这种噪音例子或异常事件可以放在模糊区进行处理，因此关键性错误可被转换为非关键性错误，便于基于内容的索引和检索。

3.1.2.2 音频特征与提取

在音频特征提取之前，首先要对非压缩和压缩格式音频形成一个标准的频谱模板。在位流模式下，频谱模板 $SPEQ(w,f)$ 是从编码后的位流 MP3/AAC 中获得的，为解码后的 MDCT 系数以及相应的频率值 $FL(f)$，其中 w 代表窗口数目，f 表示频率索引。而在一般模式下，$SPEQ(w,f)$ 是从 PCM 样本的频谱中获得的，为对 PCM 样本进行傅里叶变换后的系数及其相应频率值 $FL(f)$。

频谱模板形成后，首先提取帧特征，然后提取用于段分类的段特征。频谱模板为 $SPEQ(w,f)$ 和 $FL(f)$，这里的 $SPEQ$ 为 $MDCT$ 或 $PSPQ$，取决于当前工作模式。提取的帧特征有帧能量、子带能量比、基音频率和子带频谱中心四个。

1）帧能量（Total Frame Energy：TFE）用来侦测静音帧，可表示为：

$$TFE_j = \sqrt{\sum_{w}^{NoW}\sum_{f}^{NoF}(SPEQ_j(w,f))^2} \qquad \text{式(3-1)}$$

2）子带能量比（Band Energy Ratio：BER）是由一个频率切断的两个频率区域的总能量之比。给定一个截断频率 $f_c\left(f_c \le f_{BW}\right)$，令 $f\langle f_c\rangle$ 为线频率索引，这里 $FL\left(f\langle f_c\rangle\right) \le f_c \prec FL\left(f\langle f_c\rangle+1\right)$，帧 j 的 BER 为：

$$BER_j(f_c) = \sqrt{\sum_{w}^{NoW}\sum_{f=0}^{f\langle f_c\rangle}\left(SPEQ_j(w,f)\right)^2} \Bigg/ \sqrt{\sum_{w}^{NoW}\sum_{f=f\langle f_c\rangle}^{NoF}\left(SPEQ_j(w,f)\right)^2} \qquad \text{式(3-2)}$$

3）基音频率（Fundamental Frequency ： FF）。采用自适应峰值检测算法来检测是否有足够多的峰值，而这些峰值是某个确定频率（某个侯选 FF 值）的整数倍。

4）子带频率中心（Subband Centroid：SC）表示频谱分布，在压缩域中，它被认为是绝对频谱的平衡频率值。使用频率模板数组，$SC(f_{SC})$ 的计算公式为：

$$f_{SC} = \sum_{w}^{NoW}\sum_{f}^{NoF}(SPEQ(w,f)\times FL(f)) \Bigg/ \sum_{w}^{NoW}\sum_{f}^{NoF}SPEQ(w,f) \qquad \text{式(3-3)}$$

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

段特征从帧特征中提取，用于段分类。段指时间窗口，在一个音频片段中具有一定持续时间。段分为静音段和非静音段。对于非静音段，要继续对其进行分类。对非静音段提取主要子带能量比、转换率、基音频率和子带频谱中心段特征。

1）主要子带能量比（Dominant Band Energy Ratio：DBER）：段中比例大的类型为段的类型。用来进行初步分类，合并具有相同类型的段。

2）转换率（Transition Rate：TR）：通常用停顿率来对语音/音乐进行分类，但快速语音的停顿率也比较低。因此采用基于转换的转换率作为段特征，这里的转换指的是由静音帧转到非静音帧或由非静音帧转到静音帧。TR 可表示为：

$$TR(S) = NoF + \sum_{i}^{NoF} TP^{i} \Big/ 2NoF \qquad \text{式(3-4)}$$

其中 NoF 为段 S 的帧数目，i 为帧索引，TP^{i} 为转换处罚因子，如表 3.1 所示：

表 3.1 转换处罚表

变化：$fr^{i} \rightarrow fr^{i+1}$			TP^{i}
静音	⟶	非静音	+1
非静音	⟶	静音	+1
静音	⟶	静音	+1
非静音	⟶	非静音	-1

3）基音频率段特征（Fundamental Frequency Segment Feature：FFSF）：语音的 FF 值以及平均 FF 值通常比较低，而音乐的 FF 值通常较高[21]。但实验证明，一些谐和的女声语音有偏差。因此，仅当帧的相邻帧均为谐和时，其 FF 值才被加入计算，否则扔掉该帧。基于一定条件的 FFSF 平均值可表示为：

$$FFSF(S) = \sum_{i}^{NoF} \begin{pmatrix} FF_i & if & FF_j \neq 0 & \forall j \in NN(i) \\ 0 & & & otherwise \end{pmatrix} \Big/ NoF \qquad \text{式(3-5)}$$

其中 FF_i 为段 S 的第 i 个帧，j 为第 i 帧的最近邻居集 $NN(i)$ 中的帧的索引。

4）子带质心段特征（Subband Centroid Segment Feature：SCSF）：语音的 SC 标准偏差比较大，而音乐较小[22]。SC 的标准偏差通过使用当地窗口的平均和段中的窗口的 SC 的标准偏差来计算，可表示为：

$$\sigma^{SC}(S) = \sqrt{\sum_{j}^{NoF} (SC_j - \mu_j^{SC})^2 \Big/ NoF}\ , \quad \mu_i^{SC} = \sum_{j \in W_i}^{NoW_i} SC_j \Big/ NoW \qquad \text{式(3-6)}$$

其中 μ_i^{SC} 是第 i 帧在有 NoW 个帧的窗口 W_i 中的窗口 SC 平均值，$\sigma^{SC}(S)$ 是有 NoF 个帧的段 S 的 SCSF。

3.1.2.3 分割分类算法的实现步骤

该算法不需要任何先验知识和监督机制，以一种反复的方式进行，先是帧分类和初始分割，然后执行反复步骤直到形成全面分割，从而最后获得每段的成功分类。图 3.3 图示了音频分类和分割的反复方法。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

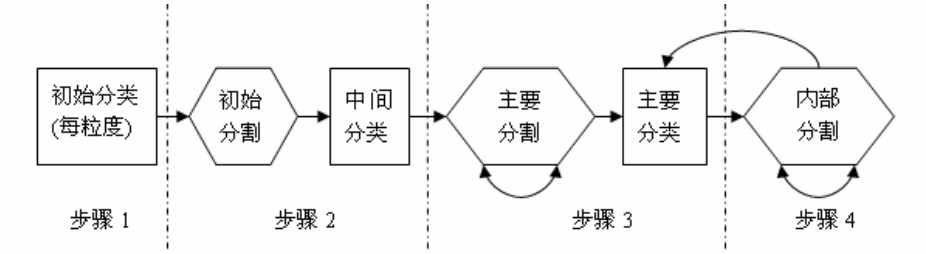


图 3.3 双模式音频分割分类算法的实现步骤

步骤 1：分别计算整个音频片段的最大、最小以及平均帧能量。测试音频片段是否为静音片断，若是，则不需要进一步地处理。一旦检测到非静音帧的存在，继续判断其为静音帧或非静音帧。对于非静音帧，计算其以 500Hz 为截断频率的 BER。若 BER 值大于阈值，那么判定为音乐，否则为语音。

步骤 2：主要是将步骤 1 中的静音帧合并成静音段。如果有足够多的静音帧合并起来的持续时间超过经验值，那么可将该段视为静音段，而介于静音段之间的部分为非静音段。然后采用段特征 DBER 和 TR 来对这些段进行分类。

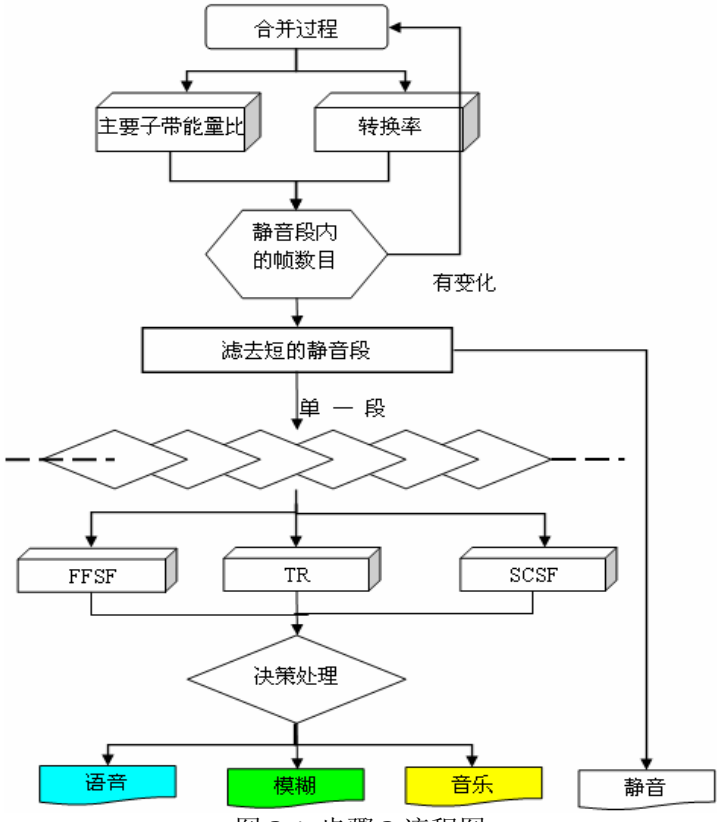


图 3.4 步骤 3 流程图

步骤 3：是主要的分割分类步骤，如图 3.4 所示。若满足条件则消除静音段并合并它相邻的两个非静音段：它的长度低于阈值并且通过 DBER 和 TR 分类后的相邻非静音段的类型是匹配的。合并一些非静音段后，重新提取那些新的非静音段的特征 DBER 和 TR，然后重新分类。对非静音段的新分类可能会产生需要进一步消除静音段的类型，因此进行反复循环以消除全部可能存在的短静音段。

循环完成后，有可能仍然存在还未被合并到相邻段中的短非静音段，强行将

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其合并到相邻的非静音段中。消除这些短的非静音段后所形成的非静音段都包含了某个简单类型，提取主要段特征：TR、FFSF、SCSF。

由于特征空间中的感知建模，所有的段特征可能会映射到强制分类区域，而不考虑通常决策过程。对于不存在强制分类情况的段，分别采用这三个段特征对其进行分类，并应用多数原则确定其最终类型。也就是说如果有两个段特征判为某一类型，就将该段视为该类型。如果不能达成一致，就将该段视为模糊段。

步骤 4：主要进行段内分析并执行一些处理以改进全面的分割机制。在步骤 3 中完成最后的分割分类后，特别长的非静音段可能仍然需要分割，因为它们可能含有两个或更多的不同类型的子段，而这些子段之间没有任何的静音部分。

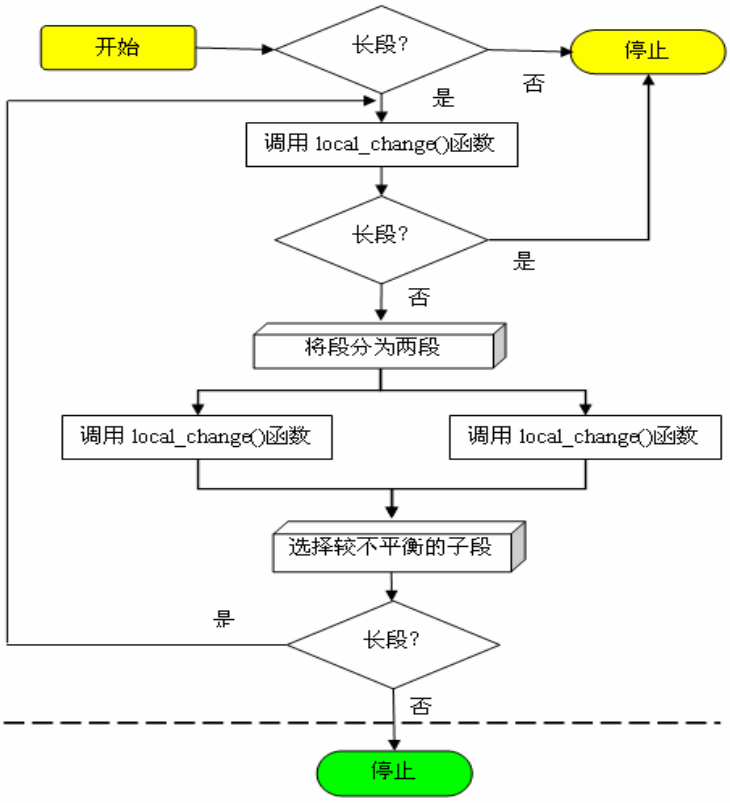


图 3.5 步骤 4 中二元内部分割流程图

首先，检测非静音段的持续时间是否超过一个给定阈值，然后将该段分为两个子段并测试它们的 SC 段特征值是否有很大的不同。如果不是，保留原来的父段，并停止处理。否则，对这两个子段执行相同操作并查找相对不太平衡的那个子段，即具有较高 SC 值的子段。执行该循环直到子段足够小，然后跳出循环。这样就确定了子段边界，然后重新执行步骤 3。若步骤 3 并没有给侦测到的可能的子段分配别的类型，则保留原来的父段的类型不变化。图 3.5 示例了该算法，其中函数 local_change（）执行基于 SC 的分类并返回左右子段的绝对 SC 差值。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.2 基于模型的音频分割与分类

基于模型的分割分类方法具有坚实的理论基础、简单的实现机制等特点，因而为目前的大多数音频分类系统所采用。这种方法要求事先给出一批带有类别标记的训练样本，通过有指导的学习训练来生成分类器，进而对测试样本集合的待分类样本进行测试以衡量其分类性能。比较典型的基于模型的音频分类算法包括HMM、神经网络和 SVM 等。下面分别对基于 HMM 和高斯模型的分割分类算法进行详细介绍。

3.2.1 基于 HMM 的说话者分割与分类

起源于上世纪 60 年代后期的隐马尔可夫模型属于信号统计理论模型[23, 24]，能够很好的处理随机时序过程的识别与预测，在多媒体处理如语音、音乐、视频以及实时监控中得到了很好应用。

一个 HMM 由下面五组参数构成：

1）状态总数 N。设全部的状态为 $S=\{S_1,\cdots,S_N\}$；

2）各个状态的观察事件总数 M。设全部的观测事件为 $V=\{v_1,\cdots,v_M\}$，每个观测事件就是一个特征向量；

3）状态转移矩阵 $A=\{a_{ij}\}$；

4）观测事件对应状态的概率分布 $B=\{b_j(k)\}$，其中 $b_j(k)=P(v_k \mid q_t=S_j)$；

5）起始状态概率 $\pi=\{\pi_i\}$，$\pi_i=P(q_1=S_i)$，$1\le i\le N$。

方便其见，通常用三元组 $\lambda=(A,B,\pi)$ 来表示一个 HMM。HMM 有三个基本问题，它们的解决对于实际应用有十分重要的意义：

问题一（评价 Evaluation）：已知观测序列 $O=O_1\cdots O_T$ 和模型 $\lambda=(A,B,\pi)$，求在该模型下观测序列 O 发生的概率 $P(O\mid\lambda)$；

问题二（解码 Decoding）：已知观测序列 $O=O_1\cdots O_T$ 和模型 $\lambda=(A,B,\pi)$，求该观测序列对应的最优状态序列 $Q=q_1\cdots q_T$；

问题三（学习 Learning）：已知观测序列 $O=O_1\cdots O_T$ 和模型 $\lambda=(A,B,\pi)$，如何调整模型 λ 的参数最大化 $P(O\mid\lambda)$。

其中，问题一和问题二可用于分类决策。不同之处是，在问题一中每一类对应一个模型，而在问题二中每一类对应一个状态。问题一的求解可以用于对给定音频片断的识别分类。问题二的求解则能应用 HMM 给出音频数据流的最佳状态序列，实现统一的音频分割和识别，每个状态对应一定的音频类别，状态之间的转换就是音频分割的边界。此外，无论问题一还是问题二，都需要先对 HMM 进行训练和学习。这个主要是通过问题三的求解来实现的。下面简单介绍一下三个

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

问题求解的算法。

评价问题求解最直接的方法就是枚举所有可能的状态序列，再累计在所有状态序列下出现观测序列O的概率，但是这样计算效率不高，需要大约$2T\cdot N^T$次运算。常用的更有效的算法（大约N^2T次运算）是前（后）向算法。和问题一不同，问题二没有唯一确定的解，这主要是因为最优状态序列的标准可以不同。依据不同的最优标准，就有不同的解。较常用的是最大化$P(Q|O,\lambda)$，即得到一条最优的状态转移路径。对此，一般采用一种基于动态规划的算法——Viterbi 算法来求解。对于学习问题求解，事实上，给定训练的观测序列，我们没有一个最优的解析方法来估计模型的参数。我们只能通过一个递归过程，来调整$\lambda=(A,B,\pi)$的参数使得$P(O|\lambda)$局部最大。常用算法包括 Segmental k-Means 算法和 EM（Expectation-Modification）或称为 Baum-Welch 算法。两者的区别在于前者仅考虑最优状态路径，而后者考虑所有可能的路径。

上面简单地介绍了 HMM 的基本概念和原理，下面介绍 HMM 在说话者分割与分类中的应用。

如果一段语音是由几个不同人的语音构成，那么可以使用一个训练好的模板，将这个连续语音信号分割开来，使每一分割出来的音频单元里只包含一个话者语音。

音频信号本质上是时序变化数据，对于同类音频信号（如某个人的语音），它在不同时刻出现时，虽然会发生变化，但是可以使用一种方法去模拟它的动态过程。这样，当它下一次出现时，看它是否吻合所训练好的模板，如果匹配，则把它与其它类别音频分割开来。对于若干个人，提取其语音特征 MFCCs，然后分别为他们建立模板，用这些建立的模板去模拟不同人的语音动态变化过程。这样，根据得到的模板，就可以把它们对应的语音信号分割开来。

假设需要将一段音频信号中n个人的语音分割出来，那么需要训练n个隐马尔可夫链$\lambda_i(1\le i\le n)$。每个训练好的隐马尔可夫链λ_i表示识别一种时间序列模式的参数模型，即每个训练好的隐马尔可夫链对应一个话者。

对于一个要识别未知音频时间序列o，要判断这个音频时间序列属于哪个说话者，就是计算o属于哪个λ_i的概率最大，然后把概率最大所属的λ_i作为识别结果。

分割流程图如图 3.6 所示，具体算法如下：采集n个人讲话的样本，训练一个带有n个状态的隐马尔可夫链，这里每个状态代表一个人讲话的语音。譬如，第一状态代表某个人，第二个状态代表另外某个人。那么对含有这n个人的语音数据流，可以通过训练好的隐马尔可夫链计算出最佳状态序列，由于每个对应一个讲话人，在得到的最佳状态序列中，如果从第一个状态转换到第二个状态，意味着对应语音信号一个话者转换成了另外一个话者，于是，将语音信号从转换处分割开来。最后，语音音频自动被分割成了若干音频单元，每个单元仅包含一个话者语音。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

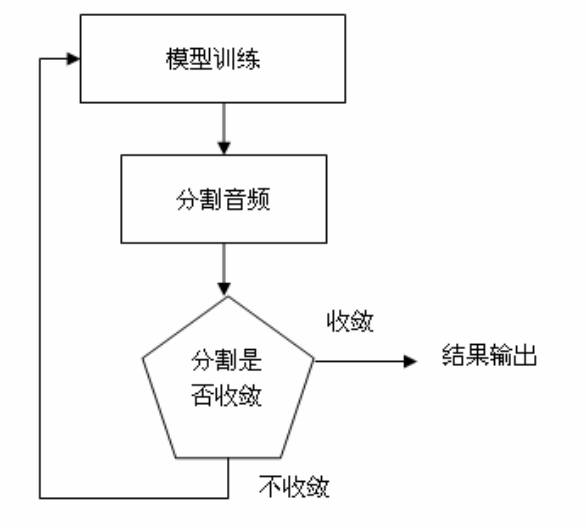


图 3.6 基于 HMM 的说话者音频分割流程图

通过上面的说话者分割过程，我们已经将整个音频分为若干个段，每个段包含单一的内容，即只包含一个说话者的声音，下面要进行的是对这些段进行识别分类。已有实验证明，HMM 用于说话者识别也取得了很好的效果。

得到不同说话者的隐马尔可夫链模板 $\lambda_i(1 \le i \le k)$ 后，给生成的隐马尔可夫链模板赋以相应的语义信息，如说话者 A、说话者 B 等。

对于一个要识别的音频例子，提取其每一帧的特征 MFCC，并得到每个音频帧的特征向量 $O=\{o_1,o_2,...o_T\}$。采用前向算法，计算每个隐马尔可夫链可夫模板 $\lambda_i(1 \le i \le k)$ 对 $O=\{o_1,o_2,...o_T\}$ 的最大似然值 $P_i(X \mid \lambda_i)$。令 $j=\arg\max\{P_i,1 \le i \le k\}$，则这个音频例子被识别为隐马尔可夫链 j 所代表的语义。

同时设置一个最小阈值，如果对任意 $\lambda_i(1 \le i \le k)$，$P_i(X \mid \lambda_i)$ 都小于该阈值，则认为该 $O=\{o_1,o_2,...o_T\}$ 不属于任何一种已知音频模型。

3.2.2 基于高斯模型的音频分割算法

基于 HMM 的音频流分割去掉了阈值的干扰，但是生成这个模板需要采集样本，进行反复训练。在训练隐马尔可夫链时，隐马尔可夫链的结构是需要自己设定的，如隐马尔可夫链的状态数目、每个状态所对应的高斯分布数目，以及状态之间是“自遍历”还是“从左到右”的结构，即需要选择一个结构。但是，并不知道怎样的一个隐马尔可夫链结构才能达到最优结果，所以只能凭借经验。另外，在训练隐马尔可夫链的过程中，并不知道到底多少个训练样本是适合的，于是就会普遍认为，样本越多越好。其实，并非样本越多越好，而是“好样本”越多越好。但是没有一个合适方法去判断哪些是“好样本”，哪些是“坏样本”，只能靠自己的主观感受决定。下面介绍一种基于高斯模型的多变化点音频分割方法[10]，

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

它不需要采集样本，根据特征变化点来进行分割，取得了很好的分割结果。

多变化点音频自动分割方法的框架如图 3.7 所示。一般说来，普通音频数据中的变化点包括前景和背景变化。例如，一个新闻广播可能包含说话者、背景音乐/噪音/语音、演播室/室外环境等等之间的变化。忽略音频类型，假定不同子段中的特征在特征空间中的平均矢量和协方差矩阵中包含重大的变化。采用一个独立且同一地分布多元高斯模型来模拟每个不同类型的子段。对于一个分析帧，提取它的 MFCC 并将其用来估计高斯模型的平均矢量和协方差矩阵。另外，采用多变化点高斯模型来检测音频信号特征空间中的特征变化，并采用最小描述长度 Minimum Description Length（MDL）标准[28]来确定何时停止分割程序，其中 MDL 确定具有最小复杂度的音频描述。还采用了分层二元分割程序[29]来搜索最好的分割方案。下面将详细介绍基于 MDL 的高斯模型的音频分割算法。

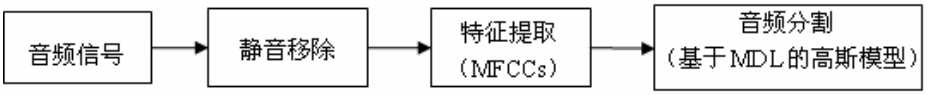


图 3.7 多变化点音频分割框架图

3.2.2.1 静音移除

一般的音频信号，尤其是语音信号都会包含不同长度的静音段。静音段的存在会提高错误警报的总体可能性，因此先采用静音去除程序以移除静音段并消除由静音段造成的错误警报影响。

由于音频源中的环境噪音和声道噪音的影响，静音帧的能量不再是零。估计一个背景能量阈值并将其用来从音频信号中移除静音。首先，所有帧按照汉明能量进行升序排列以形成一个帧序列 $(f^1, f^2, ..., f^{N_r}, f^n)$。如果帧的汉明能量比 (3-7) 式定义的能量阈值低，那么就定义其为静音帧。

$$Threshold = \frac{1}{N_r} \cdot \sum_{i=1}^{N_r} E(f^i) \qquad 式(3\text{-}7)$$

其中 $E(f^i)$ 表示帧 f^i 的汉明能量。N_r 表示汉明能量最高百分之 r 的帧的数目。然后将所有连续的检测到的静音帧合并形成静音段，并移除所有音频中的静音段。

3.2.2.2 特征提取

对于每一非静音帧，提取 12 个 MFCC 系数和对数能量，这些系数与它们的第一时间派生一起用来形成 26 维特征向量。采用基于帧的特征 MFCCs 的原因有两个：首先，根据前面对音频分割的研究，MFCC 在音频分割中得到了广泛地应用[25, 26]并且取得很高的性能。其次，如果变化点在帧层次上被检测到并且鉴别出了用于音频分割的说话者和环境，语音识别可以被用来获得更为精确的分割，从而

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

提高性能。

3.2.2.3 基于高斯模型的音频分割

提取帧的特征 MFCCs 后，相应的特征向量序列表示为 $y=(y_1,y_2,...,y_n)$。设 y 是一系列的独立分布的 d 维高斯随机向量，并且用 b 个分界点 $c=(c_1,c_2,...,c_b)$ 组成 $b+1$ 个子段。假定子段 j 中的数据经过一个独立且同分布的多元高斯模型后，它的平均向量为 μ_j，协方差矩阵为 $\sum_j$。一般说来，对于 y 存在两种段假设：

1）没有变化点，即 $b=0$；

2）$b+1$ 个子段，b 个变化点即 $c=(c_1,c_2,...,c_b)$，并且 $\mu_1\neq\mu_2\neq...\neq\mu_b\neq\mu_{b+1}$，$\sum_1\neq\sum_2\neq...\neq\sum_b\neq\sum_{b+1}$（所有的 $\mu,\sum$ 都是未知的）；

其中 b 表示变化点的未知数目，而 $c=(c_1,c_2,...,c_b)$ 是一系列未知变化点的位置。从概念上来说，在假设 H_b 下测试所有潜在的变化点位置并且确定接受哪个假设后，可以找到最好的一系列变化点。但是，直接估计要求列举 2^n 个可能的组合，计算复杂度以指数增长。在下面设计了一个用于搜索最好变化点序列的分层二元分割程序。

在假设 H_0 下，y 的概率密度函数可以容易地给出为：

$$p(y;\mu,\Sigma)=(2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\times\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right\} \quad \text{式(3-8)}$$

在另一假设下，特征向量序列 y 被分成 $b+1$ 个均一的子段。因为特征向量 y_i 是从 d 维多元高斯分布得来的，所以 $b+1$ 个子段可被假定为在统计上是独立的，并且 y 在假设 H_b 下的概率密度函数为：

$$p(y\left|c,b;\mu,\Sigma\right.)=(2\pi)^{-\frac{nd}{2}}\prod_{j=1}^{b+1}\left(\left|\Sigma_j\right|^{-\frac{m_j}{2}}\times\exp\left|\left(-\frac{1}{2}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\mu_j)'\Sigma_j^{-1}(y_i-\mu_j\right)\right|\right) \quad \text{式(3-9)}$$

其中 m_j 表示第 j 个子段中的特征向量数目；μ_j 和 $\sum_j$ 分别为第 j 个子段的未知平均向量和协方差矩阵，c_0+1 等于 y_1 并且 c_{b+1} 等于 y_n。y 在假设 H_b 下的对数概率密度函数为：

$$\log p(y\left|c,b;\mu,\Sigma\right.)=-\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\Sigma_j\right|-\frac{1}{2}\sum_{j=1}^{b+1}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\mu_j)'\Sigma_j^{-1}(y_i-\mu_j) \text{式(3-10)}$$

因为与在假设 H_b 下特征向量序列 y 相关的被定义为 $\theta=\left((\mu_1,\Sigma_1),(\mu_2,\Sigma_2),...,(\mu_{b+1},\Sigma_{b+1})\right)$ 的所有参数是未知的，所以使用最大似然估计来得到参数 $\hat{\theta}=\left(\left(\hat{\mu}_1,\hat{\Sigma}_1\right),\left(\hat{\mu}_2,\hat{\Sigma}_2\right),...,\left(\hat{\mu}_{b+1},\hat{\Sigma}_{b+1}\right)\right)$，它最大化了对数概率密度函数，其中

$$\hat{\mu}_j=\mu_j=\frac{1}{m_j}\sum_{i=c_{j-1}+1}^{c_j}y_i \quad , \quad \hat{\Sigma}_j=\frac{1}{m_j}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\hat{\mu}_j)(y_i-\hat{\mu}_j)' \quad \text{式(3-11)}$$

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

式(3-11)中的估计参数$\hat{\theta}$代替式(3-10)中的参数θ。在假设H_b下用于y的音频分割的合成多变化点高斯模型如式(3-12)所示。

$$\begin{aligned}\log p(\mathrm{y}|\mathrm{c},b;\hat{\theta}) &= -\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\hat{\Sigma}_j\right|-\frac{1}{2}\sum_{j=1}^{b+1}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\hat{\mu}_j)'\hat{\Sigma}_j^{-1}(y_i-\hat{\mu}_j)\\ &= -\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\hat{\Sigma}_j\right|-\frac{1}{2}\sum_{j=1}^{b+1}\mathrm{tr}\left(\hat{\Sigma}_j^{-1}\sum_{i=c_{j-1}+1}^{c_j}(y_i-\hat{\mu}_j)'\hat{\Sigma}_j^{-1}(y_i-\hat{\mu}_j)\right)\\ &= -\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\hat{\Sigma}_j\right|-\frac{1}{2}\sum_{j=1}^{b+1}\mathrm{tr}\left(\hat{\Sigma}_j^{-1}m_j\hat{\Sigma}_j\right)\\ &= -\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\hat{\Sigma}_j\right|-\frac{1}{2}\sum_{j=1}^{b+1}m_jd\\ &= -\frac{nd}{2}\log(2\pi)-\frac{1}{2}\sum_{j=1}^{b+1}m_j\log\left|\hat{\Sigma}_j\right|-\frac{nd}{2}\end{aligned} \quad \text{式(3-12)}$$

算法采用最小描述长度 MDL 原则[27]进行音频分割，它将分割程序处理为编码程序并且最小化音频信号的描述长度。MDL 标准使得最佳分割的每一子段都能获得最合适的模型类型，并且最小化音频信号的全部描述长度。

应用 MDL 规则的第一步是估计随机复杂度，即有一系列变化点$c=(c_1,c_2,...,c_b)$的整个音频信号描述的平均代码长度。需要编码进音频信号y的必要位数目包括两部分：描述每一子段的特征向量所需的位数目和对一系列变化点$c=(c_1,c_2,...,c_b)$进行编码所需的位数目。

对具有m_j个特征向量的第j个子段的特征向量进行编码需要对这些参数进行编码：多元高斯概率密度函数$\mathrm{p}(\mathrm{y}_j;\theta_j)$的参数$\theta_j=(\mu_j,\Sigma_j)$和特征向量$\mathrm{y}_j$。与模型参数$\theta_j=(\mu_j,\Sigma_j)$相关的代码长度为$(d+(d\cdot|(d+1)|/2))\log_2\sqrt{m_j}$[28]。多元高斯的概率密度函数$\mathrm{p}(\mathrm{y}_j;\theta_j)$，包括平均向量和协方差矩阵在内的参数维数为$d+(d\cdot|(d+1)|/2)$。另外，第$j$个子段的特征向量的熵代码的平均代码长度可表示为：

$$-\log_2\mathrm{p}(\mathrm{y}_j;\theta_j)=-\sum_{y\in\mathrm{y}_j}\log_2\mathrm{p}(y;\theta_j) \quad \text{式(3-13)}$$

接下来，必须确定变化点序列的代码长度。因为y包含n个特征向量和b个变化点，对变化点位置进行编码的代码长度为$DL_{change-points}=b\bullet\log_2 n$。应用 MDL 规则可以得到最小的音频分割方案：

$$\begin{aligned}(\hat{c},\hat{b}) &= \arg\min_{\{c,b\}}DL(\mathrm{y}|c,b;\hat{\theta})\\ &= \arg\min_{\{c,b\}}\left\{\sum_{j=1}^{b+1}-\log_2 L(\mathrm{y}_j;\hat{\theta}_j)+DL_{\theta_j}\right\}+DL_{\text{change-points}}\\ &= \arg\min_{\{c,b\}}\left\{\frac{nd}{2}\log_2(2\pi)+\frac{nd}{2}+\frac{1}{2}\sum_{j=1}^{b+1}\times\left(m_j\log_2\left|\hat{\Sigma}_j\right|+(d+\frac{d\cdot|(d+1)|}{2})\times\log_2 m_j\right)\right\}+b\bullet\log_2 n\end{aligned} \quad \text{式(3-14)}$$

其中$DL(\mathrm{y}|c,b;\hat{\theta})$表示y的描述长度，y中的分割条件为$(c,b)$且其估计模型参

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

数为 $\hat{\theta}$。$\sum_{j=1}^{b+1}-\log_2 L(\mathrm{y}_j,\hat{\theta}_j)$ 表示所有子段的特征向量的平均代码长度。基本上，所有可能的 (c,b) 都要被测试，并且使用式(3-14)来确定最佳 (c,b)，这样复杂度太高，下面介绍分层二元分割方法[29]以获得相似的性能且能降低复杂度。

分割程序在每一搜索步骤中重复地检测一个变化点，并且将变化点排序为从第一个到第 b 个。然后在整个音频数据的基础上寻找每个二元分割最为突然的变化点。由于变化点的数目比音频样本数目少得多，即 $b \prec\prec n$，这里对于长音频信号采用一种“自顶向下”的分层次分割策略。这种方法有两个好处：一是节省时间，另外由于每个检测到的变化点都分配了优先权，如果发生了“超分割”，那么就将那些具有低优先权的变化点抛弃掉。设一系列的 d 维高斯特征向量为 $\mathrm{y}=(y_1,y_2,...,y_n)$，分层二元分割程序的实现步骤如下：

1）计算 $DL(\mathrm{y}\left|c,b=0;\hat{\theta}\right.)$ 结果显示条件为 y 中不存在变化点（即 $b=0$）。

2）得到

$$\hat{c}^{b=1}=\arg\min_{d<c^{b=1}<n-d} DL(\mathrm{y}\left|c^{b=1},b=1;\hat{\theta}\right.) \qquad \text{式(3-15)}$$

假设 y 包含一个变化点。通过扫描所有可能位置以获得最佳变化点 $\hat{c}^{b=1}$ 来确定整个描述长度 $DL(\mathrm{y}\left|c^{b=1},b=1;\hat{\theta}\right.)$。可能位置范围为 d 和 $n-d$ 之间以获得与 y 相关的最大似然估计值。

3）得到

$$\hat{c}^{b=2}=\arg\min_{\substack{d<c^{b=2}<\hat{c}^{b=1}-d,\\ \hat{c}^{b=1}+d<c^{b=2}<n-d}} DL(\mathrm{y}\left|(c^{b=2},\hat{c}^{b=1}),b=2;\hat{\theta}\right.) \qquad \text{式(3-16)}$$

在得到第一个变化点 $\hat{c}^{b=1}$ 的情况下确定第二个分界点 $\hat{c}^{b=2}$。

4）重复直到

$$\frac{DL(\mathrm{y}\left|\hat{\theta},(\hat{c}^{b=k+1},...,\hat{c}^{b=1}),b=k+1\right.)}{DL(\mathrm{y}\left|\hat{\theta},(\hat{c}^{b=k},...,\hat{c}^{b=1}),b=k\right.)}>\lambda \qquad \text{式(3-17)}$$

式(3-17)中的不等式被用来确定何时停止程序以获得最后的分界点数目和位置 $(\hat{c},\hat{b})$。在这个式子中，λ 是一个收敛参数，而且通常被设为 1。如果不等式成立，那么就确定了变化点数目以及它们的相应位置 $((\hat{c}^{b=k},...,\hat{c}^{b=1}),\hat{b}=k)$。

潜在的变化点的范围介于为 d 和 $n-d$ 之间以获得假设 H_b 下 y 的最大似然估计。如果 $(DL(\mathrm{y}\left|c^{b=1},b=1;\hat{\theta}\right.)/DL(\mathrm{y}\left|c,b=0;\hat{\theta}\right.))>\lambda$，那么 y 中不存在变化点。给检测到的变化点赋于优先权，$\hat{c}^{b=1}$ 是 y 中最为突然的变化点位置。

从概念上来说，基于 MDL 的高斯模型的收敛参数 λ 可以被调整来改变音频分割的分辨率。获得的变化点数目随着 λ 的值而增加。例如，采用收敛参数 λ 生成了一系列的变化点 $\hat{c}=(\hat{c}^{b=1},\hat{c}^{b=2},...,\hat{c}^{b=k})$，若 $\lambda'>\lambda$，获得一系列的变化点 $\hat{c}'=(\hat{c}^{b=1},\hat{c}^{b=2},...,\hat{c}^{b=k},\hat{c}^{b=k+1},...,\hat{c}^{b=k+l})$。$\hat{c}'$ 中的前 k 个变化点与 $\hat{c}$ 中的相同，后面的 l 个变化点在 MFCC 特征空间中变化很小。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

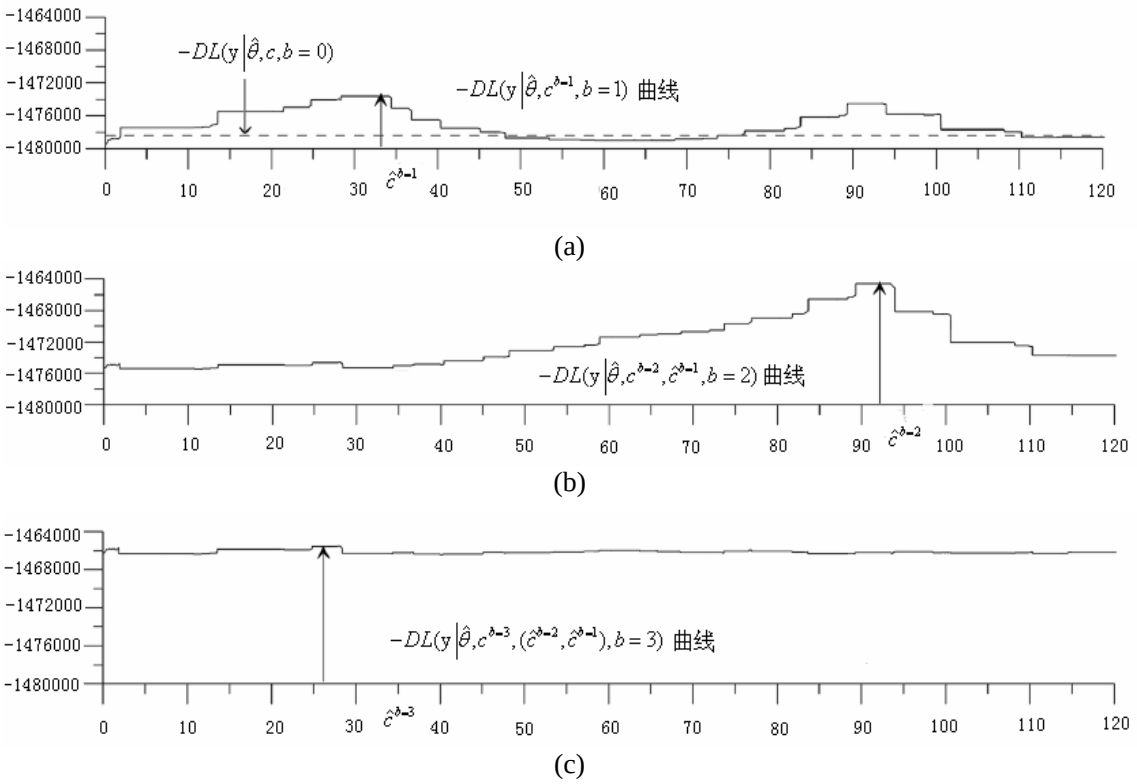


图 3.8 含有 2 个变化点的音频的分割结果

图 3.8 给出了一个使用基于 MDL 高斯模型的两个变化点和分层次的分割程序的例子。为了清楚地示例分割程序，使用了描述长度的一个负的表达式 $-DL(y|c,b;\hat{\theta})$。图 3.8（a）显示了扫描了所有可能位置后的描述长度曲线，并且在 33.3 分处获得了第一个分界点 $\hat{c}^{b=1}$，它的值最大。类似地，图 3.8（b）显示了 92.4 分处的第二个变化点 $\hat{c}^{b=2}$，图 3.8（c）显示了 26.1 分处的第三个变化点 $\hat{c}^{b=3}$。由于 $\hat{c}^{b=2}$ 和 $\hat{c}^{b=3}$ 的描述长度满足式(3-17)中的不等式，所以变化点的最佳数目为 2，因此扔掉第三个变化点 $\hat{c}^{b=3}$。

3.2.3 音频分割算法改进及实验结果

音频中的语音\音乐分割的目的不同于用于自动语音识别的分割，因此，成功用于 ASR 的特征、处理方法和建模概念并不一定适用于分割。用于语音识别的特征最小化说话者和声学环境间的差别，并最大化音素空间的差别。但是，在音频流的语音\音乐分割中，最小化音素方差以产生包含单一声学事件更为可取。用于 ASR 的传统 MFCC 特征对于语音\音乐分割可能并不一定有效。另外，近来的研究通过使用来自人类听觉系统的音频特征提高了音频系统的性能[30]。通过使用线性变换技术如 PCA、ICA 或因子分析来将特征变换到一个更合适的空间中去，传统的基于 MFCC 的音频系统取得了更好的性能。特别地，独立成分分析（Independent Component Analysis）ICA 更适用于这种变换，因为在理论上它可以最小化变换成

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分的统计依赖性。因此，针对上面介绍的基于高斯模型的音频分割算法，本文给出一个新的基于 ICA 变换的特征 Mel-ICA[57]。

ICA 是一个将观察到的多维矢量分解为互相独立的源成分的统计变换。它的基础是在二十世纪八十年代后期由 Jutten 和 Herault 奠定的[31]。ICA 变换发现使用高级统计可以将线性和非正交源信号在统计上尽可能地独立。基本 ICA 定义可以表示为： x=A·s, s=W·x 。

其中 $x=(x_1,x_2,...,x_m)^T$ 为 m 维的平均观测向量， $s=(s_1,s_2,...,s_n)^T$ 为 n 维的源向量。混合矩阵 A 为 $m\times n$ 维矩阵，它的行元素为线性结合源成分以形成观察数据元素的系数，W 为 $n\times m$ 维的非混合矩阵。问题在于估计这个转换，以使得源成分像具有稀疏或超高斯分布一样的统计独立于其它成分。假定 $m=n$ ，其中的一个转换矩阵可以简单地计算为另一个的逆。研究了很多无监督的学习技术来估测上述转换矩阵，实验使用固定点 FastICA 来训练 ICA 转换矩阵。

令 $X(m)$表示音频流频谱能量， W_k 表示关键波段滤波器， $S[k]$ 表示第 k 关键波段的能量， M 表示美尔频率域中的关键波段数目，范围为 20 到 24。那么：

$$S[k]=\sum_{j=0}^{F/2-1}W_k(j)\cdot X(j),k=1,...,M \qquad \text{式(3-18)}$$

选取独立成分的数目为 M 。获得 ICA 变换基数的步骤如下：

1）对能量对数进行处理，获得每帧的向量 s ；

2）设置一个变换基数计数器 i ，初始值为 1，并随机初始化变换基数 w_i ；

3）设 $g(\cdot)=\tanh(\cdot)$ 并计算不偏离的 w_i ： $w_i=E\left[sg(w_i^Ts)\right]-E\left[g'(w_i^Ts)\right]w_i$ ；

4）进行正交化： $w_i=w_i-\sum_{j=1}^{i-1}(w_i^Tw_j)w_j$

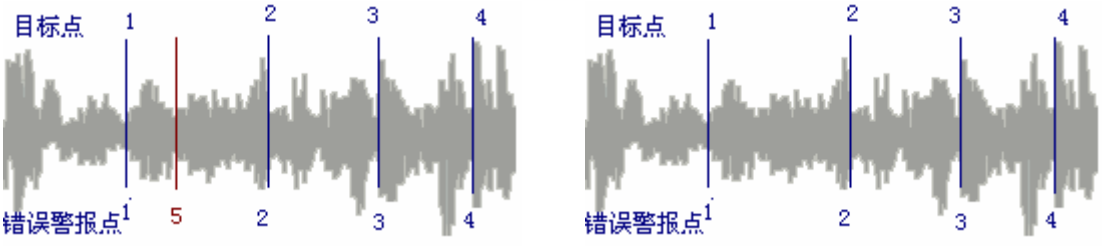
5）对 w_i 进行标准化： $w_i=\frac{w_i}{\|w_i\|}$

6）如果 w_i 不是收敛的，转步骤 5）；

7）计数器 i 加 1。如果 $i<M$ ，转步骤 3），否则结束。

得到基数 w_i （ $i=1,...,M$ ）后， M 维的 Mel-ICA 特征可以通过下式获得：

$$f_{ICA}=\sum_{k=1}^{M}\log(S[k])w_k \qquad \text{式(3-19)}$$



（a） （b）

图 3.9 采用不同音频特征的变化点检测

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

实验证明 Mel-ICA 比原算法采用的 MFCCs 的效果更好，减少了错误警报点，如图 3.9 所示，其中图（a）和（b）分别为采用 MFCCs 和 Mel-ICA 的变化点检测结果，（a）中有一个错误警报点。

为了评估算法的分割性能，采用 I 类错误和 II 类错误的检测错误权衡曲线，它们在前面的研究中已得到了广泛地应用，I 类错误是指误报率（FAR），II 类错误指漏检率（MDR）。图 3.10 给出了一个音频分割中的 MDR、FAR 的表示。

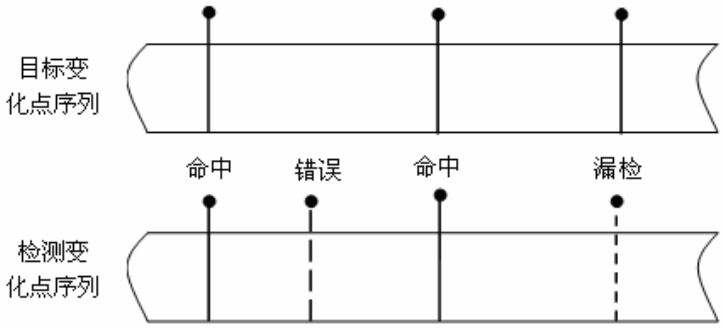


图 3.10 FAR、MDR 示意图

对于所有分割实验，MDR 和 FAR 定义如下：

$$MDR=\frac{\text{漏检的变化点数目}}{\text{目标变化点的数目}}\times 100\% \qquad \text{式(3-20)}$$

$$FAR=\frac{\text{错误警报变化点数目}}{\text{检测到的变化点数目}}\times 100\% \qquad \text{式(3-21)}$$

实验数据为来自 CCTV-1 的新闻广播，音频文件以 16kHz 进行采样，每一样本为 16 位。实验目的在于比较提出的新的特征 Mel-ICA 和 MFCCs 对于分割性能的影响。实验结果如图 3. 11 所示，结果表明给出的新特征比 MFCCs 更适合基于高斯模型的分割算法，MDR 和 FAR 都有所下降，其中式(3-17)中的收敛参数 λ 的范围为 0.7 到 1.2，并且每一步间隔为 0.05。

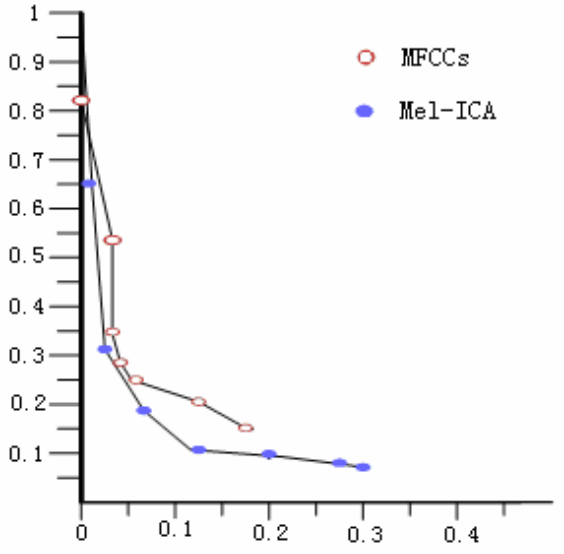


图 3.11 采用不同音频特征的音频分割结果

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.3 一种基于特征阈值和模型的组合分类方法

音频自动分类与分割是提取音频中结构化信息和语义内容的重要手段，是音频和视频内容理解、分析与检索的基础，上面介绍了几种音频分割分类方法，大概可以分为两大类：传统的基于特征阈值的分割分类算法，即根据一种或者几种音频特征及其阈值判定音频所属的类别；基于学习模型的分割分类算法，如 HMM、高斯模型等。

这些现有的音频分割分类方法主要存在这几个方面的问题。第一，这些研究大多采用相对简单的特征和前人的经验值，同时处理的分类问题也比较单一，通常只是对语音和音乐进行分类。在简单的分类中分类精度比较满意，但如果分类对象增加，比如加入环境音、非纯语音等或者取较小的窗口，则只使用简单特征进行分类，精度很低，并且基于傅立叶变换提取的特征的可靠性不高。第二，传统的基于规则的分类算法中，决策规则和分类顺序并不一定是最优的，并且上层的决策错误会积累到下一层而形成“雪球”效应。第三，基于学习模型的分类很大程度上依赖于样本的选取。

因此在这里，我们给出一种结合特征阈值和学习模型的方法，同时结合小波变换提取音频特征，提高了分类的准确度。对于语音/非语音区分，非语音中的音乐/环境噪音的分类，采用基于特征阈值的方法。因为对于语音和音乐，它们在某些特征方面表现特别明显，前面有很多的研究并且都很成熟。而对于具体的语音分类，如将其分为纯语音/带背景的语音等，采用基于模型的方法来进行识别分类。这里采用 SVM[32, 33, 34, 35]，因为 SVM 在分类准确率、计算时间和参数设置稳定性方面表现出比其它传统的非参数分类器高的性能（如 RBF 神经网络、最近邻居（NN）、和 k-NN 分类器[36]），同时它也比基于特征选择程序的组合的传统格调识别方法和传统分类器高效[37, 38, 39]。下面将详细介绍音频特征提取、SVM 的理论基础以及如何使用 SVM 进行音频分类。

3.3.1 基于特征阈值的初始分类

采用分层的基于阈值的分类器[40]将音频分为语音、音乐和环境声音。先采用 LSP－VQ 精练机制，基于高过零率（high zero-crossing rate ratio：HZCRR）、低短时能量比（low short-time energy ratio：LSTER）、频谱流量（spectral flux：SF）三个特征进行语音/非语音分类，然后采用一个使用带周期（band periodicity：BP）、噪音帧比（noise frame ratio：NFR）和频谱流量（spectral flux：SF）三个特征的基于阈值的分类器将非语音区分为音乐和环境声音，如图 3.12 所示。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

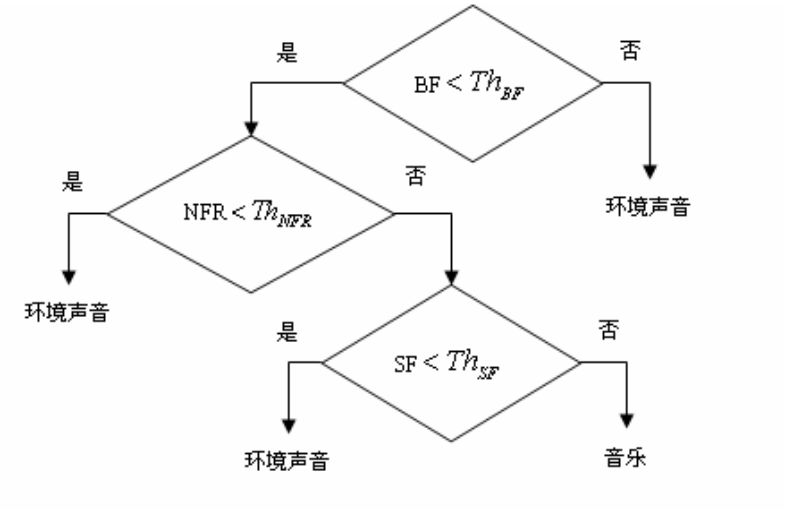


图 3.12 音乐/环境声音分类过程

3.3.2 采用 SVM 对语音进一步分类

将音频初步分为语音、音乐和环境声音后，我们采用 SVM 对语音进一步分类，将其分为纯语音、带音乐的语音和带噪音的语音。首先，提取非静音帧的特征。在进行特征提取前，先将音频信号从时域转换到频域。傅立叶变换是一种最为常用的方法。在前面的许多研究中[41, 42]，小波变换是另外一个不错的选择。因此，本文使用傅立叶变换和小波变换来更好地提取合适的音频特征。下面简单介绍一下小波变换。

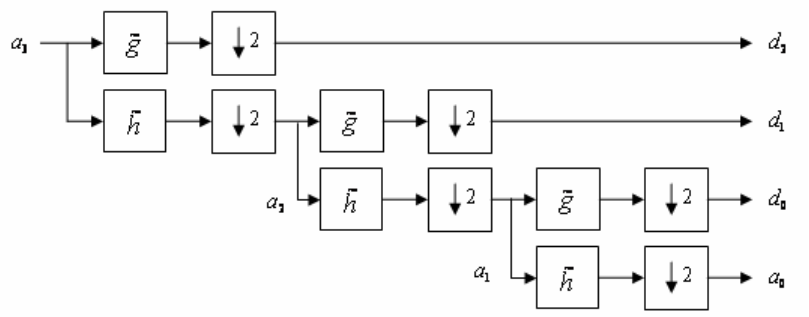


图 3.13 三层小波变换

这里讨论的小波变换是通过一个滤波器组结构来实现的。Mallat 提出的快速离散算法如图 3.13 所示，其中 $\tilde{h}(n)$ 和 $\tilde{g}(n)$ 分别为低通和高通滤波器，符号↓2表示进行 2 的下采样。令 $\{a_3(n)\}_{n\in Z}$ 为分析滤波器的输入，分析滤波器组的输出可由下式得到：

$$a_i(k)=\sum_n \tilde{h}(n-2k)a_{i+1}(n) \qquad \text{式(3-22)}$$

$$d_i(k)=\sum_n \tilde{g}(n-2k)a_{i+1}(n) \qquad \text{式(3-23)}$$

其中 $a_i(k)$ 和 $d_i(k)$ 分别为 $a_{i+1}(n)$ 的小波分解的近似系数和详细系数。在文中，

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

小波变换的计算通过使用 Daubechies 提出的长度为 8 的直交小波来实现。

下面介绍用于 SVM 分类的特征的提取过程。如表 3.2 所示，总共有$6+L$个特征，它们来自每个非静音帧s_i^h的滤波系数和快速傅立叶变换（FFT）系数$F(u)$。$6+L$个特征包括感知特征和L个频率倒谱系数（FCC）。详细的特征提取过程如下所述。

表 3.2 提取特征表

特征		变换类型	特征数目
感知特征	子带能量P_j	小波变换	3
	音调频率f_p	小波变换	1
	响度w_c	傅里叶变换	1
	带宽B	傅里叶变换	1
频率倒谱系数（FCC）c_n		傅里叶变换	L

1）子带能量P_j：计算小波域中的三个子带的能量。令w为采样频率的一半。子带间隔为$[0,w/8]$、$[w/8,w/4]$和$[w/4,w/2]$，分别相应于一个给定音频声音$a_3(k)$的近似和详细系数$a_0(k)$、$d_0(k)$和$d_1(k)$。子带能量计算公式为$P_j=\sum_k z_j^2(k)$，其中$z_j(k)$为子带j的相应近似或详细系数。

2）音调频率f_p：使用一种噪音鲁棒的基于小波的音调检测方法来提取音调频率[43]。该音调检测方法的第一步是应用带有混淆现象补偿的小波变换来将输入声音分解为三个子带，如图 3.13 所示。然后，采用一个改进的空间相关性函数，它由前一步得到的近似信号决定。实验证明该算法优于其它时域、频域和小波域的音调检测算法。

3）响度w_c：响度是傅立叶变换的频率中心，计算公式为：

$$\omega_c=\int_0^{\omega}u\left|F(u)\right|^2du\Big/\int_0^{\omega}\left|F(u)\right|^2du \qquad \text{式(3-24)}$$

4）带宽B：它是频谱成分和频率中心的平方差的能量权重平均值的平方根，如：

$$B=\sqrt{\int_0^{\omega}(u-\omega_c)^2\left|F(u)\right|^2du\Big/\int_0^{\omega}\left|F(u)\right|^2du} \qquad \text{式(3-25)}$$

5）频率倒谱系数（FCC）c_n：L个系数计算公式为：

$$c_n=\sqrt{2/256}\sum_{u=0}^{255}\left(\log_{10}F(u)\right)\cos\left(n(u-0.5)\pi/256\right),\ n=1,2,...,L \qquad \text{式(3-26)}$$

特征提取后，计算$6+L$个特征的平均值和标准偏差，形成$(12+2L)$维特征向量。另外，增加音调比率（音调帧数/总帧数）和静音比（静音帧/总帧数），形成$(14+2L)$维特征向量。

SVM 理论是一个新的统计技术，近些年来对该课题的研究非常热门，可被用作一种音频分类工具。SVM 采用一个已知的核心函数来定义一个超平面，以划分给定点到两个预定义类中。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

令 $x_i \in X \subseteq R^n$ 和 $y_i \in Y = \{1,-1\}$ 分别为输入矢量和目标变量，其中 R^n 表示 n 维的真实空间。设训练集 $S = \{(x_1, y_1),...,(x_l, y_l)\}_{i=1}^{l} \subseteq (X \times Y)^l$， $X \times X$ 上的一个核心函数 $K(x_i, x_j) = \langle \theta(x_i), \theta(x_j) \rangle$，其中 $\langle \cdot, \cdot \rangle$ 表示内部结果，θ 将输入空间 X 映射以另一个高维特征空间 F 。选择合适的 θ，给出的非线性的可分样本 S 可在 F 中被线性分隔，如图 3.14 所示。

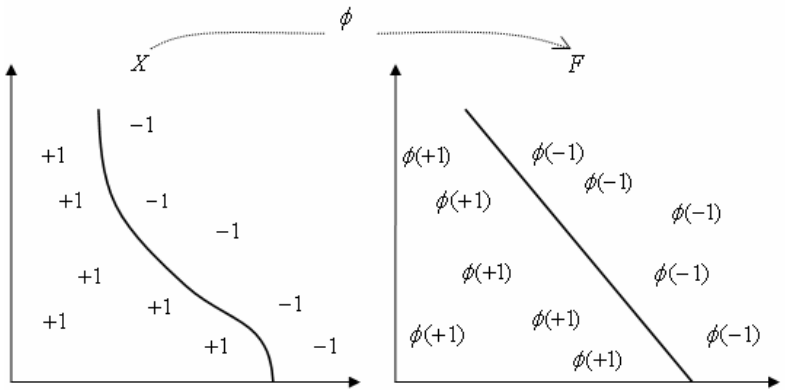


图 3.14 特征映射可以简化分类任务

许多超平面都可以达到上面划分的目的，但是我们的目标是找到最大化边缘的超平面，即从该超平面到每个点的距离最小。超平面可表示为 $(w,b) \in R^n \times R$，它包括了所有满足 $\langle w, x \rangle + b = 0$ 的 x 。问题可以表示为：

$$\begin{cases} \text{最小化} & \frac{1}{2}\|w\|^2 \\ \text{约束} & y_i(\langle w, x_i \rangle + b) \geq 1. \end{cases} \qquad \text{式(3-27)}$$

SVM 的最优化问题是通过拉格朗日函数的鞍点解决的。设 C 为拉格朗日乘数 α_i 的上边界，式(3-27)可表示为：

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \qquad \text{式(3-28)}$$

假设 α_i^* 最大化式(3-28)，那么对于任何满足 $0 < \alpha_i^* < C$ 的 k，解为 $w = \sum_{i=1}^{l} \alpha_i^* y_i x_i$ 和 $b = y_k - \sum_{i=1}^{l} \alpha_i^* y_i \langle x_i, x_k \rangle$ 。最佳辨别函数可表示为 $H(x) = \langle w, x \rangle + b = \sum_{i=1}^{l} \alpha_i^* y_i \langle x_i, x \rangle + b$，相应地，决策函数 $f(x) = sign(H(x))$ 为形成最佳分开超平面的分类函数。

有三种用于非线性特征映射的常用核心函数。1）ERBF $K(x, \bar{x}) = \exp(-|x - \bar{x}/2\sigma^2)$；2）高斯函数 $K(x, \bar{x}) = \exp(-|x - \bar{x}|^2 / 2\sigma^2)$，其中 σ^2 为高斯函数的方差；3）多项式函数 $K(x, \bar{x}) = (\langle x, \bar{x} \rangle + 1)^d$，其中参数 d 多项式的次数。

典型的 SVM 是一个两类分类器，该分类器将所有的训练集分为两类，即正类和负类，可以通过增加一些策略以达到多情况分类目的。常用的有以下 4 种机制，它们的训练和测试复杂度见表 3.3。

1）一对一：在一对类中分类。

2）一对所有：在每一类和所有其它剩下的类中进行分类。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3）上－下二叉树：初始组包括所有类。采用递归过程来区分和减少一个大组的类至一个更小组的类直到测试模式被分配了一个最终类型。

4）底－顶二叉树：在对类中执行一个递归的比较过程。保留与测试模式有更短距离的类以用于进一步比较，直到测试模式被分配了一个最终类型。

其中，底－顶二叉树机制要优于其它三种机制，这是因为其它机制：1）需要附加的决策规则；2）有更高的训练复杂度，见表 3.3；3）在训练策略中使用来自多类的混合特征，这将导致训练数据中的不调和，因此会降低分类准确率[33, 44, 45]。

表 3.3 四种机制的比较

过程 \ 机制	一对一	一对所有	上－下二叉树	底－顶二叉树
总训练复杂度	$\frac{m(m-1)}{2}C(k,k)$	$mC(k,(m-1)k)$	$\sum_{i=1}^{\log_2 m} 2^{i-1}C(\frac{mk}{2^i},\frac{mk}{2^i})$	$\frac{m(m-1)}{2}C(k,k)$
总测试次数	$\frac{m(m-1)}{2}$	$m-1$	$\log_2 m$	$m-1$
结果组合	得票数/统计量		无要求	

* m 种类型中的每个类型均包含 k 成员，其中 m 为 2 的幂。

* $C(X,Y)$ 表示正类中有 X 个向量，负类中有 Y 个向量时的训练复杂度。

此外，通常每个训练和测试音频文件只生成一个特征向量来训练或测试支持向量机，即基于文件的分类。计算所有非静音帧的特征的平均值和标准偏差，并使用这些数据表示音频文件的特征集[21, 32]。

本文提出一种基于帧和底－顶二叉树的多类支持向量机作为音频分类器[58]，可以有效的将语音音频分为纯语音、带音乐的语音和带噪音的语音三种类型。对于一个音频段，将信号划分为许多独立的帧。然后经过特征提取程序将每个帧转化为一个特征向量。

设有一个 N_F 帧的音频文件 $\mathrm{s}^{(k)}$，$k=1,...,N_F$，其类型为 C_n，$n\in\{1,2,...,N\}$。基于帧的底－顶二叉树多类 SVM 的音频分类步骤如下。

1）设置当前具有最短距离的类型 C_m 为 0，并初始化类 C_m 的累积距离 $score(C_m|s^k)$ 为一大数。

2）设置类型计数器 i 为 1。

3）根据下式计算类型 C_i 的累积距离：

$$score(C_i|s^{(k)})=\sum_{j=1}^{N_F}H(\mathrm{wx}^{(j)}+b)-\sum_{j=1}^{N_F}H(-(\mathrm{wx}^{(j)}+b)) \qquad \text{式(3-29)}$$

如果 $score(C_i|s^k)<score(C_m|s^k)$，那么更新 C_m 为 C_i。在式(3-31)中，$H(\cdot)$ 是海维赛德步骤函数。

4）计数器 i 自增 1。 若 $i<N$，转 Step 3。否则，返回 C_m 为最合适类型。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.3.3 实验结果与分析

在本文的实验中，使用了一个来源于 CCTV—1 新闻节目，广告节目和 CD 音乐的音频数据库。该数据库共有 560 个声音文件，带背景噪音的语音 60 个，噪音 46 个，音乐 255 个，纯语音 134 个，带背景音乐的语音 65 个。数据库中的音频文件的采样率为 16kHz，每一样本为 16 位。每个类型文件中的一半用来训练支持向量机，而剩下的用来测试。帧大小设为 512 个样本（31 毫秒），且相邻帧的重叠区域为 50%。如果帧能量小于某个经验阈值，那么标记其为静音帧。提取每个非静音帧的音频特征。

通常，采用准确率来衡量音频分类结果，它被定义为正确分类文件数与测试文件总数之比。先采用基于特征阈值的方法，对语音、音乐和噪音进行分类。实验结果如表 3.4 所示。

表 3.4 音频片段粗分结果

音频类型	准确率
语音	98.1%
音乐	96.8%
噪音	94.3%

对于多类 SVM 的分类，我们采用放射基础核心函数 $K(x,y)=\exp(-\gamma\|x-y\|^2)$，其中 $\gamma=1$。参数 C 确定极限最大化与训练错误最小化之间的平衡，将其设为 1。实验对音频分类方法和音频特征都进行了比较。在音频特征的比较中，对 MFCCs 的特征集的分类结果如表 3.5 所示。本文采用的特征集的分类结果见表 3.6。对于特征 MFCCs 和本文提出的特征集，实验通过采用传统的基于文件的 SVM 分类器[32]和文中提出的基于帧和底－顶二叉树的多类 SVM 对音频文件分别进行分类。结果表明采用文中提出的多类 SVM，使用特征 MFCCs 和本文特征集的分类准确率分别增加了 10.7%、11.0%。同时，与 MFCCs 相比，在基于文件的 SVM 分类器和提出的多类 SVMs 中，本文采用的特征集的准确率都高于 MFCCs。

表 3.5 采用 MFCCs 特征的分类结果

特征集和分类器	MFCCs 和基于文件的分类器		MFCCs 和文中提出的分类器	
命中/未命中	命中	未命中	命中	未命中
音乐	210	45	233	22
噪音	38	8	42	4
纯语音	108	26	120	14
带背景噪音的语音	48	12	53	7
带背景音乐的语音	52	13	57	8
准确率	81.5%		90.2%	

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 3.6 采用本文提出的特征集的分类结果

特征集和分类器	本文特征集和基于文件的分类器		本文特征集和文中提出的分类器	
命中/未命中	命中	未命中	命中	未命中
音乐	222	33	245	10
噪音	40	6	44	2
纯语音	113	21	127	7
带背景噪音的语音	50	10	56	4
带背景音乐的语音	55	10	61	4
准确率	85.8%		95.2%	

从表 3.4 和表 3.6 可以看出，对于语音、音乐和噪声分类，基于特征阈值的方法可以获得与 SVM 分类器更好的性能，这是因为语音和音乐的特征特别明显，并且前人对这方面的研究非常成熟。因此，本文采用基于特征阈值的方法对语音、音乐和噪音进行分类，然后采用多类 SVM 对语音进行具体分类，将其分为纯语音、带背景音乐的语音和带背景噪音的语音。

3.4 本章小结

音频分割和分类方法主要有两种，基于特征阈值的方法和基于模型的方法。其中基于 HMM 的音频分割与分类通过训练样本，对未知样本进行分割和识别分类。训练得到的模型即使对样本数据取得了很高的正确识别率，也不能保证训练好的模型对实际未知数据也产生很高的正确识别率。本文还介绍两种基于特征阈值的音频分割方法，其中音频分层分割方法过于依赖单一特征及前人的经验，分割效果不是特别理想。双模式的音频分割分类算法提供两种处理模式，既可应用于非压缩格式文件也可用于压缩格式文件的分割分类，但是分类类型过于简单且依然受限于前人的经验。本文采用改进的基于高斯模型的多变化点分割方法，不需要对样本进行训练，也无需依赖于前人的阈值经验，取得了很好的分割效果。同时给出了一种结合基于特征阈值和模型的分类方法，将两种方法的优点结合起来，加快了分类的速度，提高了分类准确率。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第四章 基于内容的音频检索技术研究

前面分别介绍了音频特征提取、连续音频流分割和分类等内容，本章将研究分析基于内容的音频检索技术。

与基于内容的图像/视频检索一样，目前比较成熟的基于内容的音频检索是基于听觉内容相似的音频检索。如果要求计算机对大量音频数据自动处理，像人脑对音频理解和分析那样，先形成语义，然后方便人们去检索，这个目标的实现还有相当大的挑战。主要原因在于，计算机自动对音频信号的处理与人脑的处理有一个极大的不同：在人脑形成语义前，所处理的信息直接来自大脑皮层感知器官的输出，而计算机所分析的音频信号，没有经过任何“相似”感知器官的处理。多媒体检索领域将这种现象称为“低级听觉特征与高级语义之间存在鸿沟”。

即使这样，基于听觉内容的音频检索，还是向基于语义的音频检索迈出了坚实的一步，因为任何语义都一定有存在形式。对于音频所蕴涵的高级语义，听觉特征至少表达了部分音频所蕴涵的高级语义，这也是人们逐渐研究基于听觉的音频检索技术的原因所在。

目前 Internet 上主要的音频信息有音乐、语音和广播等。对于音乐，人们总是想从 Internet 上找到自己喜欢旋律的音乐。对于广播等音频数据，由于广播中包含了广告、天气预报、主持人主题新闻和新闻详细报告等不同部分，而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，如果能够对分成如上几个部分，可以很方便人们对广播新闻不同层次的需要。最后，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 Internet 中找到自己需要的音频例子。如，有些人想找相似的“枪声”，有些人想找相似的“鼓掌声”等。而在实现相似风格“歌曲”和相似音频检索的时候，人们提交检索信息主要有三种方式。

当然，最直接的是提交一个语义描述，如“爵士音乐”和“爆炸声”等这样的文字后，然后把蕴涵了这些语义标注音频例子或歌曲寻找出来，反馈给用户。但是，要自动完成这样的任务，是相当困难的。因为在前面介绍过，音频低级听觉特征和其蕴涵的高级语义之间存在很大的鸿沟，不可能自动从“歌曲”或“音频例子”中获取完整语义。如果实在要完成这样的检索任务，一般是对每个收集了相似音频例子或歌曲的音频库进行手工语义标注，识别之后，基于标注信息完成检索。而人为手工标注因为人主观感知不一致，很难取得一个公正的语义标注。

二是提交一个示例音频，提取这个音频的特征，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样本按序返回给用户，或者根据它跟数据库中各个音频的相似度距离，返回若干最为相似的音频，这是示例音频检索。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第三种是使用“哼”作为输入。比如，用户 “哼”一段想寻找的音乐的旋律，然后基于用户 “哼”出来的音乐，去寻找与之相似风格和旋律的歌曲，反馈给用户，这种方式叫做哼唱检索。这其实也是一种示例音频检索方法，不过其示例音频是靠“哼”出来的。

第一种查询方式叫基于语义描述的音频查询方式，由于对一段音频例子可以有不同的语义描述，如何处理不同语义描述其内涵的一致性以及是否存在语义描述不一致的问题，是前一种检索方式面临的挑战。后两种是基于（听觉内容）的音频例子检索（Audio Retrieval by Clip）。下面将详细介绍哼唱检索和示例音频检索方式所采用的关键技术。

4.1 音频特征的相似度模型

在基于文本的检索方法采用的是文本的精确匹配，而基于内容的音频检索则通过计算查询音频和候选音频之间在听觉特征上的相似度匹配进行。因此，定义一个合适的听觉特征相似度度量方法对检索的效果有很大的影响。由于在第二章中介绍的听觉特征大都可以表示成向量的形式，常用的相似度方法都是向量空间模型（vector space model），即将听觉特征看作是向量空间中的点，通过计算两个点之间的接近程度来衡量音频特征间的相似度。

常用的相似度量的距离有闵氏距离、余弦距离、以及马氏距离等。世界上第一个基于内容的音频检索查询商业软件 Muscle Fish[46]中就使用马氏距离作为相似性度量方法。微软研究院的 Lu 等在文献[47]中提出了一种用于音频特征距离度量的最近特征线方法，该方法实际是对解析几何中点到线最短距离的应用；而 Compaq 研究院的 Beth Logan 等[48]则将图像处理中的 EMD（Earth Mover Distance）方法应用于音乐相似性度量，该方法计算较复杂，要经过大量的协方差计算，并且仍然缺乏对语义理解的足够支持。下面分别对这些距离公式进行简单介绍。

4.1.1 闵氏距离

闵氏距离是若干经典距离的通式，如棋盘格距离（$\lambda = 1$）、欧式距离（$\lambda = 2$）等。欧式距离是特征空间距离度量的最自然的方法，也是在音频特征距离度量中应用最多的方法。其特点在于自然、直观、易懂，而缺陷在于没有考虑方差的因素，尤其在进行多元分析时与期望往往存在较大的偏离；另一个缺陷是其距离与分量的单位有关，距离需经过规范化处理才能作为相似性适用。两个音频文件 X 和 Y 之间的闵氏距离可以表示为：

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$d(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^{\lambda}\right)^{\frac{1}{\lambda}} \quad \text{式(4-1)}$$

4.1.2 马氏距离

如果特征向量的各个分量间具有相关性或者具有不同的权重，可以采用马氏距离（Mahalanobis distance）来计算特征之间的相似度。马氏距离的表达为：

$$d(X,Y)=\sqrt{(X-Y)^T C^{-1}(X-Y)} \quad \text{式(4-2)}$$

其中C是特征向量的协方差矩阵。当特征向量的各分量间没有相关性，马氏距离还可以进一步简化，因为这时只需要计算每个分量的方差C_i。简化后的马氏距离如下所示：

$$D=\sum_{i=1}^{n}\frac{(X_i-Y_i)^2}{C_i} \quad \text{式(4-3)}$$

对某个听觉特征选择一种合适的相似度衡量方法是获取满意的检索效率的重要保证。然而，更为重要和困难的是确定不同特征之间或是同一特征的不同分量之间的权重。 马氏距离考虑了样本的统计特性，排除了样本间的相关影响。关键是协方差的计算，尤其是当数据维数较大时，协方差的计算相当耗时。

4.1.3 余弦距离

余弦距离消除了分量单位的影响，并且其计算结果可以很方便地转换为相似性，如相关分析的输出甚至可直接用作相似性，相关又可分为自相关和互相关。两音频文件间的余弦距离可表示为：

$$d(X,Y)=\frac{\sum_{i=1}^{n}X(i)Y(i)}{\sqrt{\sum_{i=1}^{n}X(i)^2\times\sum_{i=1}^{n}Y(i)^2}} \quad \text{式(4-4)}$$

4.1.4 非几何的相似度方法

上述的各种方法都是基于向量空间模型的，采用几何距离作为相似度度量。这样的距离函数通常要满足距离公理的自相似性、最小性、对称性和三角不等性等条件。然而，早在 1950 年，Attneave 用几何距离对一组四边形的感知相似性进行了实验，发现距离度量方法和人对相似性的感知判断之间存在一定差距。Tversky 指出相似性的最小性原理在一些识别中并不一定成立。同时，对于相似对称性原

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

则，在一定情况下存在着方向性[49]。对相似三角不等性也同样存在着一些争议。1977 年，Tversky 提出了著名的特征对比模型[50]。与几何距离不同，该模型不把每个实体看作特征空间中的一个点，而每个实体用一个特征集表示。

4.2 哼唱音乐检索

一般情况而言，音乐检索系统可分成四个模块：检索界面、音调跟踪、特征音乐数据库生成和检索引擎（如图 4.1 所示） 。

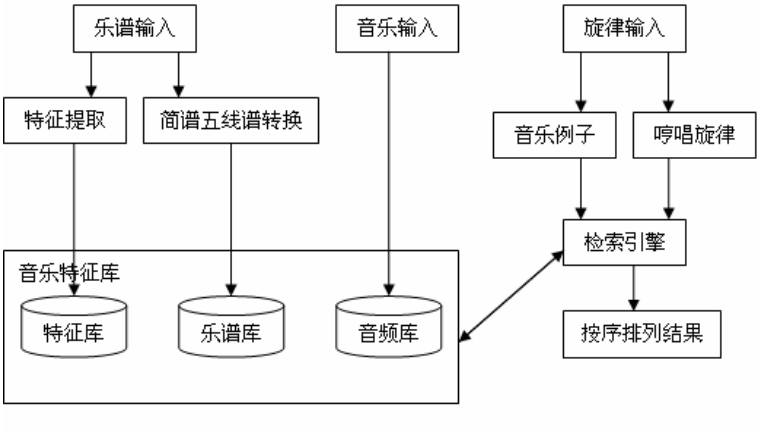


图 4.1 音乐检索系统

音乐检索有两种方式：一是基于音乐或歌曲例子的，这种方式我们将在后面的章节中介绍；还有一种是基于“哼”的检索方式。在这种方法中，检索请求是根据用户用嘴哼出来的曲调，去检索与之相似的音乐或歌曲。当然，在进行检索前，要把用户哼的曲调数字化成音频例子。

不论是基于音乐例子，还是基于“哼”，这两种检索方式的目的都是想找到相似曲调的音乐，而曲调强弱剧烈变化性质要通过对音频信号动态分析得到，这是音调跟踪要完成的任务。特征音乐库存中装入的是原始音乐和它们的曲调特征；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，最后将它们按照匹配程度大小反馈给用户。下面着重介绍音调跟踪和检索引擎部分。

4.2.1 音调跟踪

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。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。把用 U、D 和 S 表示原

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

始音频信号，叫做音频的三步轮廓表示。

从一段旋律中分离出音调并跟踪它们的相对变化，是实现上述转移的关键。对音调的一个直接定义是最接近该音符的某个频率。这个定义与实际生活中反映到人脑的音调有一定的出入，由于音调是定义在人的感知上的一个特征，至今还没有一个统一完整的物理模型能够描述它。

目前主要有三种方法用于提取音频信号中的音调，它们分别是自相关法、最大似然法和谱分析法。自相关是一种经典的音调跟踪方法。它从波形中分离出它的局部最大峰值，继而对峰值进行跟踪，得到它们的出现频率，并将该频率定义为该信号的音调；最大似然法是对自相关的一种改进，既提高准确度，又能避免走样。但由于耗时过多，实践中一般不予采用[51]；谱分析同样也是一种非常经典的音调提取算法[52]，但后来发现其结果并不精确。还可以采用 Malcolm Slaney 提出的模型来提取音频信号中的音调[53]，将 Lyon 耳蜗模型用来计算音频相关图，相关图中最大能量所位于的频率位置作为音调。

4.2.2 检索引擎

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。

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。

4.3 示例音频检索

这一节中主要介绍如何基于用户提交的示例音频，得到相似音频。示例音频检索按照音频例子表示方法的不同，可以分为两种：1）将某类音频用一个模板表示出来，对于用户提交查询的音频例子，先使用模板去进行匹配，判断其属于模板，然后将这类模板对应的音频例子按序反馈给用户；2）对每个音频例子建立模板。这里，所谓为每个音频建立模板，就是如何寻找一种良好的方式去表征音频。音频是用听觉特征表示的，提取每个音频帧的听觉特征，所有短时音频帧的听觉特征就构成了这个音频例子的表示方式。在下面将介绍，采用最小生成树（MST）

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

聚类算法得到每个音频的关键帧，使用关键帧的特征表示每个音频并进行索引，然后进行相似匹配的检索算法[11]。

4.3.1 基于分类模型的检索算法

基于分类模型的示例音频检索过程如下：首先将相似的音频组成一个个音频数据库，然后使用同一音频类别的数据训练生成分类模板，用这个模板来代表此类音频例子。假设有 *auCorpus* 个不同类别的音频数据库，对于每个音频数据库，训练了 *auCorpus* 个隐马尔可夫链 $Hmm_i(1 \le i \le auCorpus)$ 来代表这类音频数据库。

一旦用户提交了需要检索的音频例子 X，首先提取这个音频的每个短时迭加音频帧的特征向量 $X = \{x_1, x_2, ..., x_T\}$。然后采用前向算法，计算每个隐马尔可夫链语义模板 $Hmm_i(1 \le i \le auCorpus)$ 相对于 X 的最大似然值 $P_i(X \mid Hmm_i)$。令 $j = \arg\max\{P_i, 1 \le i \le auCorpus\}$，则这个音频例子属于隐马尔可夫链 j 所代表音频类别库，这个音频类别库中所有的音频例子和 X 是相似的。

有时候，并不需要将属于隐马尔可夫链 j 所代表音频类别库中的所有音频例子返回给用户，只需要将最相似的若干音频例子返回就可以。如果隐马尔可夫链 j 所代表音频类别库中总共包含 Num_j 个音频例子 $auClip_k(1 \le k \le Num_j)$，返回 $P(auClip_k \mid Hmm_j)$ 值最大的前面若干个音频例子给用户，并且将这些返回的音频例子按照 $P(auClip_k \mid Hmm_j)$ 值进行排序。

在基于模型的示例音频检索中，除可以使用隐马尔可夫链外，还可采用其他任何分类模型，如支持向量机等。

4.3.2 基于音频模板的算法实现

基于音频模板的算法指的是对每个音频都建立模板，检索时对于用户提交查询的音频例子，先计算其模板，并与数据库中音频模板进行匹配，将最为匹配的一个或多个音频例子反馈给用户。下面介绍一种通用的音频索引和检索方法，它根据音频类型（语音、音乐等）、音频内容（说话者、主题和环境等）以及声音感知来进行音频索引，而这个声音感知与人类听觉感知机制非常接近。

首先提取音频帧的特征，即分别提取每个不同类型段的特征矢量并进行存储和检索。这对基于内容的检索来说更有意义，因为它便于那些类型匹配的段内的帧之间的相似比较，避免了潜在的相似匹配错误并极大地减少了索引时间尤其是检索时间。然后在音频检索中，采用基于分类的索引机制，取得了复杂度低且鲁棒性好的查询结果。提取特征的 *AFeX* 模块在计算每个子特征的相似距离时，融合音频特征集和联合子特征，并且在相似距离计算过程中，采用处罚机制处罚不完

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

全匹配的片段。例如，查询一个包含语音和音乐的片段时，所有缺少已存在类型的片段会被处罚，如只包含音乐的片段，且该处罚与查询片段中缺少的类型（这里指语音）的覆盖范围有关。这相当于给那些具有完全类型匹配的片段赋以优先权，能获得更为可靠的检索结果。此外，还应用了规格化机制，使子特征的相似距离独立于音频帧时间，因为子特征中存在着不同的音频帧时间并会改变类型中的帧数目。这种独立性使得无需处理子特征融合问题，对音频的每帧进行规格化。

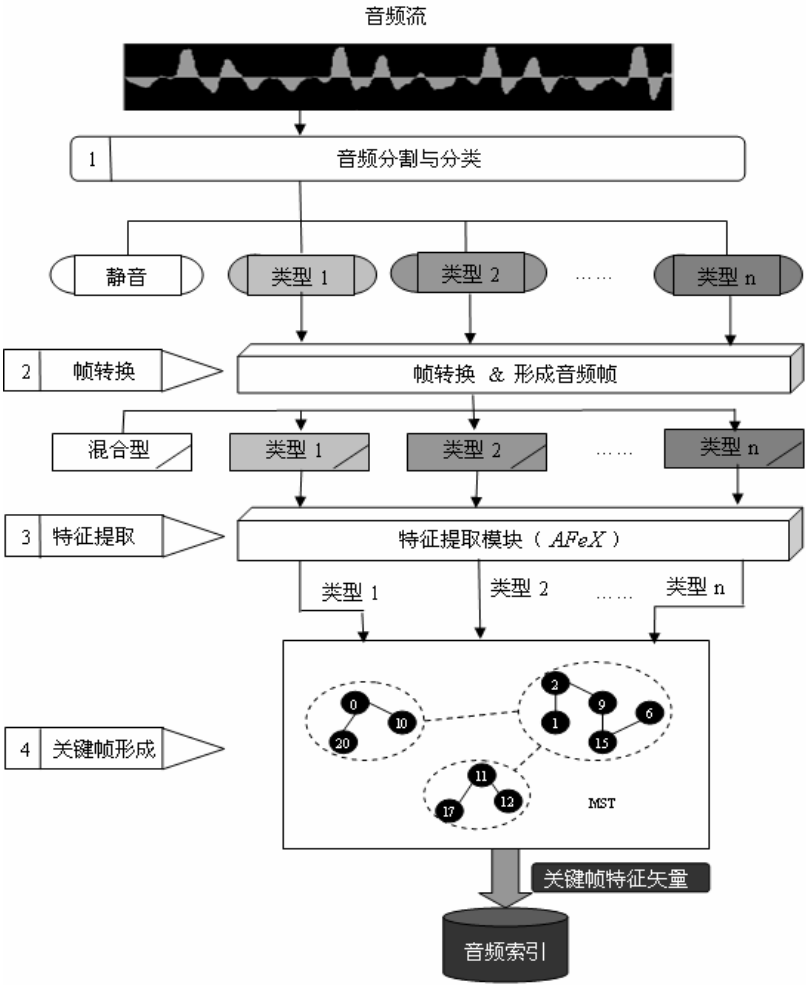


图 4.2 音频索引流程图

如图 4.2 所示，对音频数据库中的每个文件进行索引通过四个步骤完成。首先进行音频流的分类和分割，其结果是将整个音频片段分割为几种类型，除静音帧外的其它类型的音频帧用于索引，因为静音不携带任何的音频内容信息。第二步是进行帧转换，因为分类和分割中的帧的持续时间与稍后的 *AFeX* 模块用到的帧的持续时间不同。边界帧，即那些含有不止一种类型的帧，被扔掉而不用于索引，因为它们的内容混杂而不能提供干净的内容信息。音频特征提取模块 *AFeX* 对剩下的其它类型的帧（在它们相应的段中）进行特征提取，并采用最小生成树（MST）聚类方法[54]对它们相应的特征矢量进行聚类，然后存入描述符文件中形成索引。音频分割与分类已经在前面的章节中介绍过，下面详细介绍后三个步骤。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.2.1 形成音频帧

由于分割和分类都是以帧为单位进行处理的，因此需要转换以获得用于索引的通用音频帧。先将整个音频文件划分为用户音频帧或模型定义的音频帧，每个音频帧都有一个由前面步骤所得到的类型。有可能存在包含两个类型的边界帧，定义其为混合帧，扔掉这些边界帧而不用于索引，因为它们包含的内容不单一。

4.3.2.2 提取帧特征 MFCC

MFCC 已广泛应用于语音和说话者识别系统[55]，因为它们提供了频域的不相关且面向听觉的观测矢量，非常适合人类听觉感知系统。因此我们将 MFCC 用于音频检索的相似度测量，以获得与通常人类听觉感知标准相近的测量，如通过分类的含有高级别内容区分的“听起来像”。

AFeX 模块提取音频帧的特征 MFCC 通过几个步骤实现。首先，对音频帧进行汉明窗处理来增强语音中的元音和音乐中的发音的谐音的谐和特征。此外，汉明窗还可以减少在成帧过程中引入的不连续性和边缘影响，尤其是在算法中，窗口影响比较大。汉明窗是一个向前移动的余弦波的一半，如下面公式所示：

$$w(k) = 0.54 - 0.46\cos(2\pi \frac{k-1}{N-1}) \qquad \text{式(4-5)}$$

其中 N 为窗口的大小，它等于音频帧的大小（PCM 样本的个数）。

为了实现时间域上的滤波，音频帧不是重叠的，以使得它的大小为 2 的 n 次幂，然后进行 FFT 变换，与滤波器组进行简单的相乘以变换到频率域。美尔刻度滤波器组是一系列的滤波组，它们有规律分布在美尔频率域中的中心频率。美尔频率可表示为：

$$mel(f) = m_f = 1127\ln(1+\frac{f}{700}),\quad f = 700(e^{m_f/1127}-1) \qquad \text{式(4-6)}$$

通常低频率区的分辨率较高，而高频率的较低，这与人类的听觉特征相符。滤波后，计算每一子带的能量，并对子带能量值进行倒谱变换。倒谱变换是滤波组幅度对数的离散余弦变换：

$$c_i = \left(\frac{2}{P}\right)^{1/2} \sum_{j=1}^{P} \log m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \qquad \text{式(4-7)}$$

其中 $0 \prec i \le P$，P 为滤波器组的数目。c_i 的子集用作音频帧的特征矢量。另外，*AFeX* 模块应该提供独立于以下变化的通用特征矢量：采样率、音频声道数（单声道/立体声）和音量大小。因此在计算每个滤波器组能量时进行规格化，以中和采样率和音量变化的影响。令 f_S 为采样率，根据奈奎斯特理论，信号的带宽为 $f_{BW} = f_S/2$。每个 FFT 频率线的频率分辨率（Δf）可表示为：

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\Delta f = \frac{f_{BW}}{N_{FL}} = \frac{f_S}{2N_{FL}} \qquad \text{式(4-8)}$$

其中 N_{FL} 为该频率带宽内的频率线数目，$N_{FL} = 2^{\lfloor \log_2(Tf_S) \rfloor}$，其中 $\lfloor . \rfloor$ 表示取下整，T 为音频帧时间。音频帧的 PCM 样本数目为 $N = Tf_S$。经过不同采样频率采样后的音频的各个子带的能量是不同的，因为帧的样本数目不同，故通过一个系数 λ 来对子带能量进行规格化，其中 $\lambda \sim N$。

声音音量（V）可以近似地表示为音频帧的绝对平均能量，如 $V \cong \sum_i^N |x_i| / N$。类似地，具有不同音量的音频会导致子带能量不同，因此能量也用 λ 来进行规格化，$\lambda \sim N$。总规格化可表示为：$\lambda \sim \lambda_V \lambda_f \sim VN \rightarrow \lambda \sim \sum_i^N |x_i|$。

在计算每个滤波的子带能量过程中，能量被 λ 除以避免音量和采样率对倒谱系数计算的影响。滤波的中心频率是均一地分布在美尔刻度上的，可设 f_{CF}^i 为第 i 滤波的中心频率，那么该滤波的中心频率可以通过下式得到：

$$mel(f_{CF}^i) = \frac{imel(f_{BW})}{P} \qquad \text{式(4-9)}$$

由此可见，中心频率也依赖于采样率（$f_{BW} = f_S/2$）。这将导致不同采样率的音频有不同中心频率的滤波，特征向量（MFCC）完全不相关，因为它们是直接来自于每个滤波的子带能量值的。为了固定滤波的位置，使用一个固定截断频率，它与音频文件的最大采样频率相关。若音频的最小和最大采样频率为 16 和 44.1kHz，那么有：

$$mel(f_{CF}^i) = \frac{imel(f_{FCO})}{P}, \qquad f_{FCO} \geq 22050 \qquad \text{式(4-10)}$$

根据上面的公式设定中心频率，可以保证所有的音频采用相同的滤波组。不过，只有采样率为 44.1kHz 的音频才使用所有的滤波（假定 $f_{FCO} = 22050$ Hz），其它采样率低的音频的最大子带能量值（m_j 其中 $j \succ M$）被自动地设置为 0，因为这些值并不在音频信号的带宽内。这将会在计算 MFCC 时产生错误的结果，因为后面的步骤就是对子带能量值进行 DCT 变换。为了防止这些错误的发生，只采用那些常见于所有可能采样率的子带能量值。采用音频的最低（$f_S = 16kHz \Rightarrow f_{BW} = 8kHz$）和最高 $f_S = 44.1kHz \Rightarrow f_{BW} = 22.050kHz$ 采样频率来获得最小可能值 M。根据式(5-6)，有 mel(8000)=2840.03，mel(22050)=3923.35，M 的开始范围为：$M \leq 0.7238P$。因此，采用一个包含 P 个子带滤波的滤波器组，并用其中的 M 来计算 MFCC。这种方法仅使用 MFCC 的常见范围值，消除了由于数据库中音频的采样率不同而带来的影响。

最后，将倒谱变换系数（c_i）的统计值用作音频文件的索引。去掉特征矢量的第一个系数，因为它是帧能量的噪音计算而不含可靠信息，剩下的 P（$c_i \forall 1 \prec i \leq M$）上的 $M-1$ 个系数形成 MFCC 特征矢量。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.2.3 最小生成树（MST）聚类提取关键帧

音频帧的个数与音频片断的持续时间是成正比的，当 *AFeX* 模块执行后，可能会产生很多特征向量，而它们中的大部分可能都是冗余的，因为事实上音频文件中声音的重复率很高，并且有很多时候声音都是相似的。为了在可接受时间内获得高效的音频检索，这里仅仅存储那些来自不同声音帧的特征向量。这与视频特征提取机制的情形是相似的，仅存储关键帧的视频特征向量用于索引。但不同之处在于：视频的关键帧在提取特征以前就确定了的，而音频没有物理的帧结构，先要按一定持续时间成帧，获得每帧的特征后才能进行关键帧分析。

为了有效地提取关键帧，首先聚类那些有相似声音的音频帧（因此具有相似的特征矢量），并且每一聚类的一帧或更多被选为关键帧。图 4.3 给出了一个关键帧提取例子。

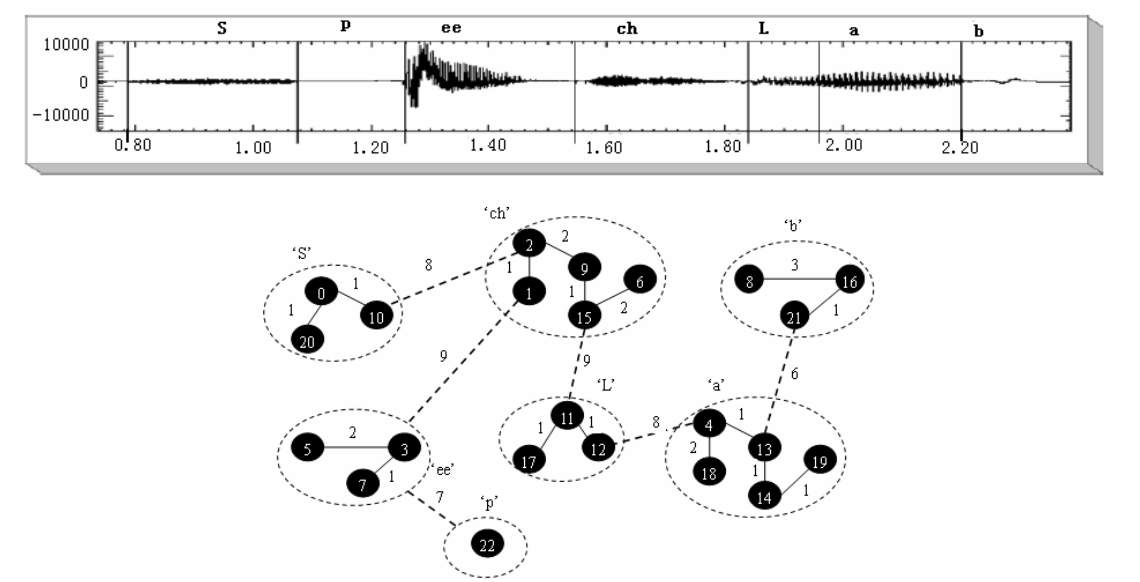


图 4.3 聚类机制示例图

问题在于如何确定提取音频的聚类数目。实际上这个数目是随着音频内容变化的。例如，独白语音的关键帧数比电影少。定义 *KF* 率为关键帧数与音频内的有效帧的总数目之比。设置 *KF* 率后，聚类的数目就可以很容易地计算出来，并且与音频时间成正比。长音频含有相似声音的可能性较大，尤其是当大部分内容是语音时，相似声音（元音和不发音部分）会不断地重复。*KF* 率可以通过一个经验的关键帧形成模型动态设定。

设定关键帧的数目后，采用 MST 方法对音频帧进行聚类。MST 中的每一结点是一个独特音频帧的特征矢量，并计算结点间的距离。形成 MST 后，断开最长的 $KFno-1$ 条枝，得到 *KFno* 个聚类，选择其中一帧（如第一帧）作为关键帧，其特征矢量被用作索引。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.3.2.4 QBE 音频检索机制

在基于音频的索引机制中，使用段的类型信息来提高效率，只与匹配的音频帧，即具有相同类型的帧，通过相似度距离进行比较，如图 4.4 所示。

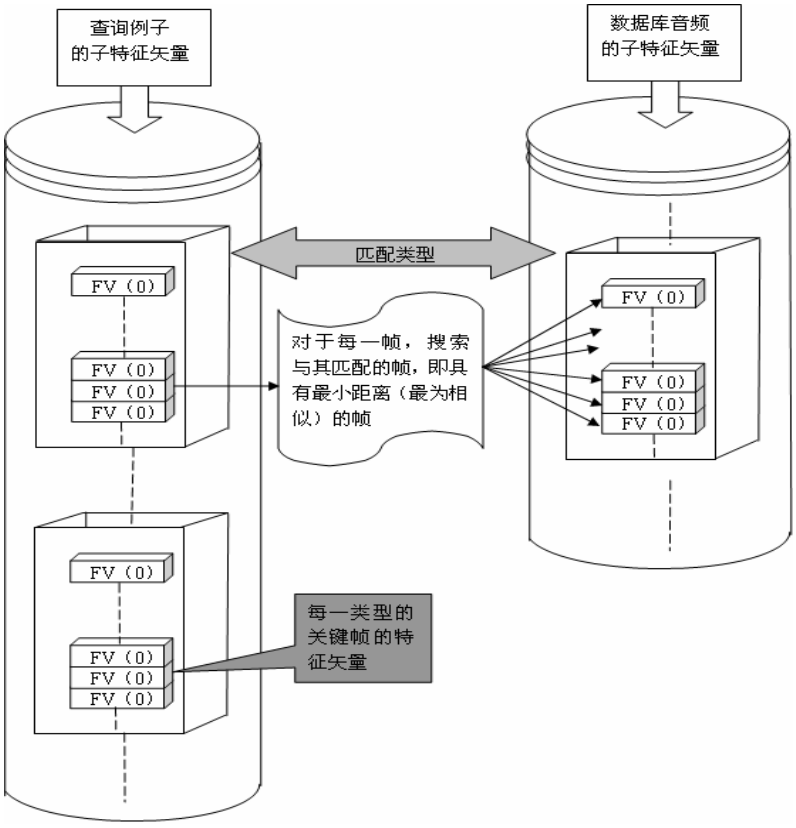


图 4.4 基于类型的音频检索中的帧距离计算

为了实现基于音频的查询，从多媒体数据库选择一个音频文件作为查询例子，在数据库中搜索跟它最相似的音频，并且至少提取了该数据库的一个音频特征。设数据库中的特征集数为 NoS，每一特征集中子特征的数目为 $NoS(s)$，$0 \le s \prec NoS$。令相似距离函数为 $SD(x(s,f),y(s,f))$，其中 x 和 y 分别是特征索引为 s、子特征索引为 f 的特征矢量。设 i 为类型为 C_q 的查询例子的音频帧索引，相似距离仅需计算该帧的一个子特征矢量（如 $QFV_i^{C_q}(s,f)$）和数据库中音频文件（设索引为 c）中的具有相同类型的音频帧（设索引为 j）之间的距离。对于音频文件 c 中所有类型为 $C_q(\forall j \Rightarrow j \in C_q)$ 的音频帧，只有与查询例子中音频帧 i 有最短距离 $D_i(s,f)$ 的那个音频帧，被用来计算这两段音频的总体子特征 (s,f) 的相似距离 $D(s,f)$。这样，查询例子中的帧只与数据库音频中的具有相同类型的帧进行比较。这种内部搜索增加了必要的检索鲁棒性，解决了音频片段中内容变化和持续时间不确定等问题。图 4.4 给出了子特征相似距离计算过程中的类型匹配和最小距离搜索机制。另外，在 $D(s,f)$ 的计算过程中，应用两个因子以获得无偏差且鲁棒性好的结果。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1）处罚函数：若在音频c中没有找到类型为C_q的音频帧，则在计算D（s，f）时应用处罚函数。设$N_Q(s,f)$为查询例子中的有效帧数目，$N_Q^{\varnothing}(s,f)$为那些由于没有在音频c中找到匹配类型而未用于总体子特征相似距离计算的音频帧的数目。令$N_Q^{\Theta}(s,f)$为剩下的音频帧数目，即那些用于总体子特征相似距离计算的音频帧数目。这里$N_Q(s,f)=N_Q^{\varnothing}(s,f)+N_Q^{\Theta}(s,f)$，类型不匹配的处罚函数可表示为：$P_Q^C(s,f)=1+N_Q^{\varnothing}(s,f)/N_Q(s,f)$。如果查询例子的所有类型都与数据库中音频$c$的类型相匹配，则$N_Q^{\varnothing}(s,f)=0\Rightarrow P_Q^C(s,f)=1$，实际上未对$D(s,f)$的计算进行任何的处罚。

2）标准化：由于音频帧的持续时间可能存在变化，一定类型的帧数目也会变化，从而导致子特征的相似距离计算（依赖于帧的数目）出现偏差。为了防止这种情况的发生，通过与子特征相关的总帧数$N_Q(s,f)$来对$D(s,f)$进行标准化。矢量被标准化后，计算查询例子与数据库中音频c的总体查询相似距离QD_c时，可以设置子特征距离的权重。通过实验得到特征集s中的子特征f的权重$W(s,f)$，以找到数据库中现有音频特征的最优并入方案。QD_c的计算如下：

$$D_i(s,f)=\begin{Bmatrix}\min\left[\mathrm{SD}(\mathrm{QFV}_i^{C_q}(s,f),\mathrm{DFV}_j^{C_q}(s,f))\right]_{j\in C_i} & \text{if} \quad j\in C_q \\ 0 & \text{if} \quad j\notin C_q\end{Bmatrix} \quad \text{式(4-11)}$$

$$D(s,f)=\frac{P_Q^C(s,f)}{N_Q(s,f)}\sum_q\sum_{i\in C_q}D_i(s,f)，\quad \mathrm{QD}_c=\sum_s^{\mathrm{NoS}}\sum_f^{\mathrm{NoF}(s)}W(s,f)D(s,f) \quad \text{式(4-12)}$$

另外，QD_c的计算只有当查询例子与数据库中音频c之间至少存在一种匹配类型时才是有效的。如果不存在匹配的类型，将QD_c置为∞，视其为相似度最低的音频，将其置为查询检索队列的末端。

4.4 本章小结

本章对基于内容的音频检索技术进行了分析，介绍了示例检索和“哼唱”检索两种查询方式所采用的音频检索技术。示例检索中采用的基于分类模型的检索算法只能对类型进行匹配，不能较精确地定位相似文件，而“哼唱”检索应用范围太狭窄，基本上只适用于MIDI格式的音乐文件。本文采用的是基于音频模板的示例检索算法，能较精确地定位相似音频文件，并且采用MST聚类方法提取音频的关键帧，检索时仅在同类型的帧中搜索最为匹配的帧，大大提高了检索效率，并在很大程度上降低了索引的存储量。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第五章 CBAR 系统设计与实验分析

5.1 系统设计概要

从大规模的音频数据库角度考虑，音频检索不是简单的相似音频的查找，它包括音频处理、音频分割与分类、特征库生成、音频匹配查询、过滤索引等许多过程。

图 5.1 给出 CBAR 系统的系统结构。图的左边是原始音频数据的预处理，包括一般信息提取、特征提取、音频分割和分类等。右边是用户查询，包括用户查询接口，检索引擎和索引过滤机制。在图的下端是元数据库、特征库和音频媒体数据库。音频处理首先要提取其属性，包括一般属性，如文件名、类型编码格式等信息，以及音频数据的音频特征。在提取时先将数据分成帧，提取每帧的 ICA-mel 特征用于音频分割，然后提取带宽和音调频率等特征，并根据这些特征采用分类器对音频数据分类。最后提取用于检索的特征集，并将其保存在特征数据库中。

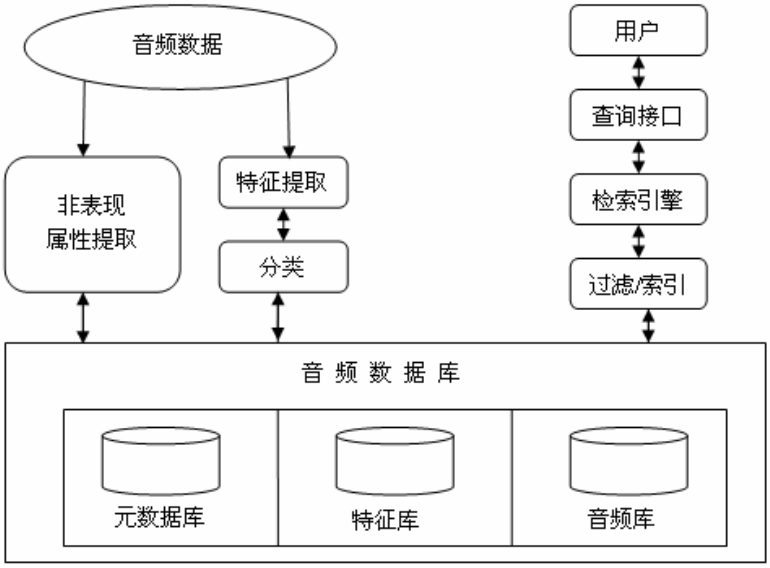


图 5.1 CBAR 系统的系统结构

用户通过用户查询接口检索相似的音频文件，或查询音频数据的类别。检索时首先通过过滤/索引机制缩小检索范围，然后使用上一章中介绍的基于音频模板的 QBE 检索方法，计算示例音频与数据库中音频文件的距离，比较其相似度，并排序输出查询结果。

5.2 CBAR 系统模块与库结构

CBAR 系统主要处理音频媒体的查询和检索。它既可作为独立的音频数据库

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

管理系统，也可支持或嵌入到军用多媒体数据库，或其它多媒体信息系统中等，以提供基于内容的音频信息分类与查询。CBAR 系统由二个子系统构成，即数据库生成子系统和数据库查询子系统，每个子系统由相应的功能模块和部件组成。下面分别详细介绍这两个子系统。

其中数据库生成子系统用于处理音频库输入、元数据库的生成和特征库的生成、插入和删除，主要包含：非表现属性信息提取模块、特征提取模块，分割模块、分类模块和数据库模块，如图 5.2 所示。

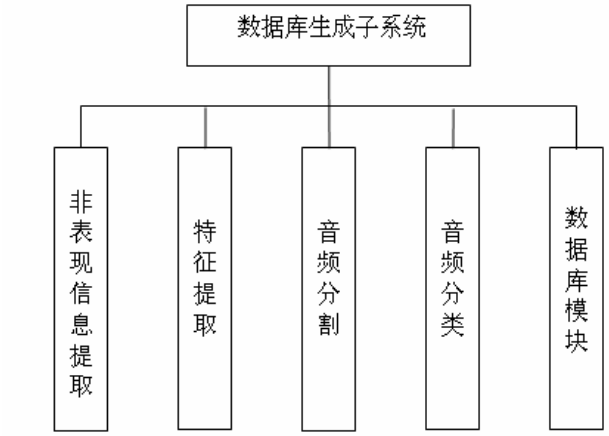


图 5.2 数据库生成子系统的功能模块图

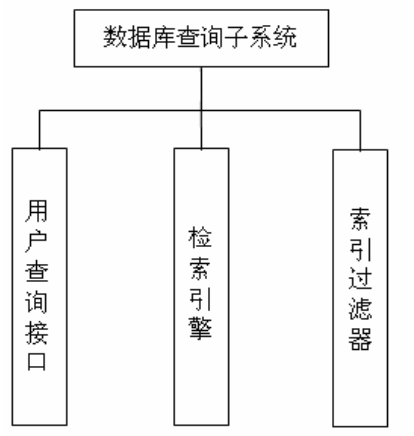


图 5.3 数据库查询子系统的功能模块

1．非表现属性信息提取

用于处理用户的常规查询，本模块提取两大属性：常规文件属性，包括全文件名、文件大小，编辑时间；音频编码属性，包括编码格式、播放时间、声道数、取样率、取样位数。

2．特征提取

这是系统的核心之一。对于数据库的生成和插入，每加入一段音频数据，就要提取其音频特征，分析各个特征的值，这样才能对其进行分割分类，然后插入到数据库中。在音频检索中也常用到特征提取，比如给定示例样本，要提取其特征，才能进行相似度的计算。本模块根据系统的需要，采用短时音频处理技术，将音频数据统一为采样率 11.025kHz，取样位数为 16 比特的单声道数据，每帧 256 个采样点，这样既能同时满足语音和音乐的检索需要，又提高了计算速度。CBAR 系统共提取三种类型的特征：用于音频分割的 ICA-mel 特征，用于分类的带宽和响度等特征以及用于检索的 MFCCs 特征。这几个特征便于用户理解和掌握，也满足系统实际的检索需要和应用。

3．音频分割

由于音频文件通常含有混合内容，如语音和音乐等，因此有必要对其进行分割，使分割后的段含单一内容。采用改进的基于高斯模型的音频分割算法，将音频流分成包含单一内容的段，提高了分类和检索的效率。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4．音频分类

对音频数据进行自动分类是 CBAR 系统的功能之一。对音频分割得到的段进行分类，使每个单独段均含有唯一的类型。分类特征提取后，先采用特征阈值的方法进行语音/音乐/噪声分类，若为语音，再比较查询样本与 SVM 形成的类模板的距离，距离最小的类就认为是查询样本所在的类。

5．数据库

数据库由音频库、特征库和元数据库组成。下面分别介绍各个库的组成：

1）媒体库包含原始音频数据，以文件形式存放。

2）元数据库包含是分类和检索处理所需的必要信息，最主要的部分是音频分类的个数，各类的名字、类模板和其它信息。另外还有音频处理所需的参数和阈值。

3）特征库是基于内容检索的核心，包含对音频数据提取的用于分割、分类以及检索的特征和预处理提取的一般属性信息。

数据库查询子系统以友好的用户查询界面处理音频数据的分类与检索，包含：用户查询接口，检索引擎，和索引过滤器，如图 5.3 所示。

1．用户查询和浏览接口

采用示例查询（QBE）作为音频信息的查询方式，提供友好的查询和浏览接口，可以对音频进行播放，方便用户的查询。

2．检索引擎

检索引擎通过计算音频模板间的距离比较示例样本与数据库数据的相似度，距离公式采用欧氏距离函数，通过大量的检索和分类实验表明，该系统具有较高的检索效率和正确率。

3．索引/过滤器

检索引擎通过索引/过滤模块来达到快速搜索的目的，从而可以应用到数据库中的大型多媒体数据集中。过滤器作用于全部数据，利用 MST 聚类来提取音频关键帧从而进行索引，降低了存储量，同时提高了检索效率。

5.3 系统开发平台和界面

本文的实验系统采用基于 Windows XP 操作系统，在 VC6.0 平台上开发，使用 C++语言开发底层模块，应用模块和界面；系统采用面向对象系统设计方法（OOD）。该实验系统的原型系统界面（GUI）如图 5.4 所示。

在音频检索系统中，左边是查询音频显示区，用户提供的查询音频的波形就出现在波形显示框中；界面右边的结果显示区，按相似度从上到下顺序显示了相似的音频检索结果的，并以数值形式反映音频之间的相似程度，为研究算法提供

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

实验数据，另外在上方还可播放选中的检索结果音频并显示其波形。

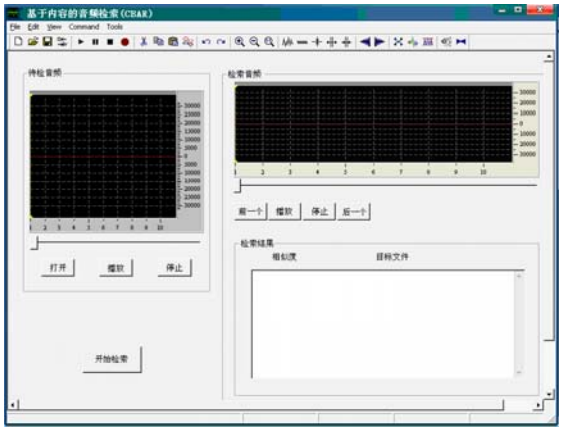


图 5.4 实验原型系统截图

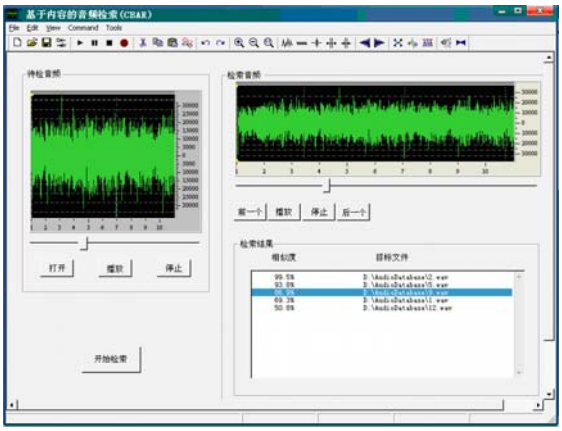


图 5.5 音频检索运行界面

友好完善的用户界面不但是系统功能的完美体现，也是系统实用的基础。设计不好的界面难以为用户理解和掌握，难以使用。对于基于内容的音频分类和检索，更是如此。CBAR 系统提供了友好的用户界面，方便用户的查询、检验。图 5.5 给出了 CBAR 系统音频检索时的界面。

5.4 实验结果分析

测试数据来源于 CCTV—1 新闻节目和 CD 音乐。截取新闻广播节目的纯语音片断以及同时包含语音和音乐的片断，并将每个音频片断转换成通用的 WAV 格式。测试数据库共有 300 个声音文件，每个文件的持续时间为 40 秒。其中纯语音 120 个，音乐 100 个，语音和音乐混合片断 80 个。

利用一段未知的音频片断来对数据库中音频进行对比及检索，便可找到库中是否存在与之相似的音频文件，并显示相似度及文件。

本方法的有效性与选取的检索特征的有效性有很大的关系。由于 MFCC 提供了频域的不相关且面向听觉的观测矢量，非常适合人类听觉感知系统，因此我们采用 MFCC 作为检索特征，以获得与通常人类听觉感知标准相近的测量，如通过分类的含有高级别内容区分的“听起来像”。实验检索的有效性用查准率（Precision）和查全率（Recall）来评估，分别定义如下：

$$\text{查全率} = \frac{\text{检索出的音频中相关音频的数目}}{\text{所有相关的音频数目}} \qquad \text{式(5-1)}$$

$$\text{查准率} = \frac{\text{检索出的音频中相关音频的数目}}{\text{检索出的音频数目}} \qquad \text{式(5-2)}$$

查全率反映系统检索相关音频的能力，而查准率则反映系统拒绝无关音频的能力。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5.1 音频检索结果

音频类型	查全率	查准率
语音	95.6%	96.2%
音乐	83.4%	82.5%
语音和音乐	89.3%	90.1%

实验结果如表 5.1 所示。实验表明，语音的查全率和查准率较高，而音乐较低，这是因为采用的检索特征是 MFCC，目前为止，还没有找到适合音乐检索的特征，这也是以后研究的一个重要方向。

5.5 本章小结

CBAR 系统是基于内容的音频分类与检索的系统，可作为独立的系统，而且可以嵌入到其它多媒体系统中去，以提供满足基于内容的音频分类与检索服务。另外其检索引擎可用于 Web 上，实现网络上的基于内容的音频分类与检索。它能很好的完成音频自动分类和基于内容的检索。实际检索数据表明：

1．系统查询界面友好，可满足用户的查询需求。

2．音频数据的自动分类准确率很高，尤其对语音数据可达到较高的准确率。可用于区分语音，音乐文件等，可达到实用的要求。

3. 对于音频的检索具有较高的效率和准确性，查询结果与人们的感觉很相近。

当然，由于时间和人力有限，CBAR 系统只是一个初步的基于内容的音频检索系统，还有许多功能有待完善，才能达到真正的实用的要求：

1．系统目前对语音的检索准确率较高，但对音乐较低，合适的音乐检索特征还有待研究。

2．CBAR 系统应用了相似查找功能，但还没有找到较好的音频数据的索引结构，很显然对于海量的音频数据库检索这是远远不够的，因此需要研究满足基于音频检索要求的索引结构。

3．系统现只支持示例音频查询，为满足不同用户的需求，查询方式应该多样化，增加“哼唱”查询方式以及提交语义描述的查询等。

4．作为网络上是检索引擎，目前只初步实现了计算机间的引擎调用。与其它系统的实际连接及网络上的通讯传输等问题还没有作深入研究。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第六章 总结与展望

基于内容的音频检索是一个新兴的研究领域，在国内外仍处于研究、探索阶段。只有在基于音频物理特征的识别技术方面有所突破，才可能在更高层次的基于知识辅助的音频检索方面做出更深入地研究。本文主要从以下几个方面对基于内容的音频检索进行了研究：

1. 音频信号主要的时域、频域及时频域特征提取；并对音频信号的时域、频域及时频域特征的应用情况进行了分析。
2. 研究了音频信号的分割和识别算法，并已发表论文三篇，给出的算法已得到同行专家的初步认可。音频信号的分割部分主要研究了基于高斯模型的音频变化点分割方法，并给出一种新特征改进该算法，提高了分割的准确率。音频信号的识别部分主要研究基于特征阈值和基于学习模型的分类算法，并给出一种基于阈值和模型的组合方法，对单一的音频信号进行识别。
3. 从音频查询方式的角度研究了不同的检索方法。现有的查询方式主要有哼唱音乐检索和示例音频检索两种，其中主要研究了示例音频检索采用的基于模板的检索算法。采用 MST 聚类方法提取音频的关键帧，降低了存储量，同时大大提高了检索效率。
4. 设计并实现了一个完整的音频分割和分类系统 CBAR，既可作为独立的系统，也可嵌入到其它系统中，以满足基于内容的音频分类与检索的需要。

由于原始音频数据除了含有采样频率、量化精度、编码方法等有限的注册信息外，本身仅仅是一种非语义符号表示和非结构化的二进制流，缺乏内容语义的描述和结构阻的组织，因而音频检索受到极大的限制。相对于日益成熟的图像与视频检索，音频检索相对滞后。尽管国内外的知识界正在为建立一个相对完备的音频检索研究机理提供普遍指导意义的理论基础。但是，就目前而言，这项工作还处于研究的初级阶段，音频检索技术还并不成熟，它还面临着很多的挑战：

1．内容描述标准

MPE G- 7 是正在制定的多媒体内容描述，其目标就是制定一组标准的“描述子”及其“描述模式”。内容描述与媒体内容结合，使用户能够快速准确地进行检索。MPEG-7 还制定标准的描述定义语言 DDL.这种自描述模式独立于平台、厂商和任何应用，方便多媒体内容的分布处理，同时有利于内容的交换和重用。在丰富而且标准的内容描述模式的支持下，应用将可以支持用户化的多视图。由此可见，内容的标准化，将极大地促进基于内容检索的广泛应用，同时也有利于其它的多媒体应用，如多媒体编辑和处理、过滤代理、超媒体浏览、媒体交互等。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

MPEG-7 的范围不包括特征提取和检索引擎，目的是留有竞争的余地。因此在特征及其提取、查询接口、检索引擎、索引等方面。

2．需要更加方便、有效的检索界面

涉及到用户对内容的感知表达、交互方式的设计、用户如何形成并提交查询等方面。现代多媒体信息系统的一个重要特征就是信息获取过程的可交互性，人在系统中是主动的。除了提供示例和描绘查询基本接口之外，用户的查询接口应提供丰富的交互能力，使用户在主动的交互过程中表达对媒体语义的感知，调整查询参数及其组合，最终获得满意的查询结果。用户的查询接口应该是直观易用的，底层的特征选择对用户是透明的。这里涉及到如何转换用户的查询表达到可以执行检索的特征矢量，如何从交互过程中获取用户的内容感知以便选择合适的检索特征等问题。

3．Internet 多媒体内容检索

从未来应用的需求来看，对 Internet 上音频信息的检索是未来对多媒体的基于内容的组织和处理所面临的又一问题。首先是网络上音频内容的组织和分类，由于网络信息众多且繁杂，对网络信息的检索必须依赖于有效的组织和分类方式。另外，需要研究快速高效的索引技术支持大型 Web 上的多媒体信息的检索。由于 Internet 上的用户是来自不同层次的任何人，其查询标准和理解方式不同，因此需要提供智能化的操作界面。

4．音频相关反馈

由于距离函数计算出来的两个音频之间的相似度是几何意义上的相似度，非主观感知相似。因此有必要引入音频相关反馈概念，用户可以根据检索结果，结合自己的意愿去调整音频特征的权值，来定位检索的侧重点。另外，也可让用户根据自己的感觉对检索出来的每个音频的满意程度作出一个评价然后根据这个评价判断出用户对特征的倾向程度，从而自动调整特征的权值，重新进行检索，使检索的结果符合自己的主观感知。这可以从基于听觉内容的检索向基于语义内容的检索迈进一步。目前国际上已有学者开始这方面的研究。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致 谢

时光如梭，转眼我的硕士学习生活已接近尾声，过去的这段时光，成了我此生难忘的记忆。在导师刘志镜教授的关怀和悉心指导下，我顺利地完成了毕业课题及论文的撰写，刘老师渊博的学识和丰富的实践经验、严谨的治学作风、勤奋的工作态度和谦逊的为人给我留下了深刻的印象，将永远是我学习的榜样，从刘老师身上我不仅学到了如何做学问，还学到了如何做人。近三年来，刘老师给我创造了很多的锻炼机会，并精心指导，倾注了大量的心血，在此，谨对刘老师致以最诚挚的谢意。

感谢实验室的所有同学，他们在我的论文写作过程中提出了许多宝贵的意见和建议，感谢那些曾经帮助过我的同学，另外师弟师妹们的积极提问也是给予我的另一种意义上的帮助。

最后，非常感谢我的家人，他们为我的成长付出了无尽的心血，在我的求学生涯中总是默默地关心我和鼓励我，支持我在成长的道路上有勇气不断跋涉前进。

经过三年的学习生活，相信我获得的不只是一个学位，而是这一生的难得的宝贵财富和美好回忆。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献

[1] 朱学芳. 多媒体信息处理与检索技术. 电子工业出版社，2002. 11。
[2] 蒋丹宁，蔡莲红. 音频分类与音频分类的研究. 第九届全国多媒体技术学术会议论文集. P142-145。
[3] 庞渤，蔡莲红. 基于音频内容检索技术的研究. http://hcsi .cs .tsinghua.edu.cn/paper/paper, 1997。
[4] J.H.Foote. A Similarity Measure for Automatic Audio Classification. In Proc. AAAⅠ 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora. Stanford, March 1997.
[5] Jonathan Foote. Content based retrieval of music and audio. Multimedia Storage and Archiving SystemsⅡ, Proc. of SPIE, 1997. Vol.3229, pp.138-147.
[6] Tong Zhang and C.C Jay Kuo. Heuristic Approach For Generic Audio Data Segmentation and Annotation. Proc ACM'99. pp.67-76.
[7] Tong Zhang and C.C Jay Kuo. Content-Based Classification and Retrieval of Audio. Proceedings of SPIE's Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII. SanDiego, July, 1998.
[8] Tong Zhang and C.-C. Jay Kuo. Hierarchical System for Content-Based Audio Classification and Retrieval. Proceedings of SPIE's Conference on Multimedia Storage and Archiving SystemsⅢ. Boston, 1999, SPIE Vol.3527, p398-409.
[9] Foote J .T. Content-Based Retrieval of Music and Audio. In: Proceedings of SPIF, C.J.kuoctal, editor Multimedia Storage and Archiving Systems 1l， 1997. 32(29).138-147.
[10] Chung-Hsien Wu, Chia-Hsin Hsieh. Multiple Change-Point Audio Segmentation and Classification Using an MDL-Based Gaussian Model. Journal, IEEE Trans. Audio, Speech, and Language Processing, March 2006. Vol.14, No.2.
[11] S. Kiranyaz, M. Gabbouj. A Generic Content-Based Audio Indexing and Retrieval Framework. IEE Proceedings Vision, Image and Signal Processing, 2006. Vol.153, pp.285-297.
[12] Wilcox, L D, F R Chen, D Kimber and V Balasubramanian. Segmentation of Speech Using Speaker Identification. Proc Int Conf Acoustics, Speech and Signal Processing, Adelaide, Australia, April 1994.
[13] 郝杰, 李星. 汉语连续语音识别中关键词可信度的贝叶斯估计. 声学学报，2002. 27(5):393~397。
[14] Kiranyaz, S., Qureshi, A.F., and Gabbouj, M. A generic audio classification and

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

egmentation approach for multimedia indexing and retrieval. Proc. Eur. Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, EWIMT 2004, London, UK, 25–26 November 2004. pp. 55–62.
[15] 齐俊英.基于内容的音频检索技术的研究.辽宁工程技术大学硕士论文，2005年6月。
[16] 贺前华，陆以勤，韦岗.一种新的HMM训练方法.电子学报,Sep，2000. Vol 28 No 9: p56-58。
[17] 卢坚，陈毅松，孙止兴，张福炎.语音/音乐白动分类中的特征分析.计算机辅助设计与图形报，2002.第14卷第3期。
[18] KeChen, Ting-Yao and Hong-Jiang zhan. on the Use of Nearest Feature Line for Speaker Identification. 微软亚洲研究院论文集[多媒体计算组 2002].
[19] 韩纪庆,张磊，郑铁然.语音信号处理.北京&清华大学出版社，2004.01。
[20] C. S. Burrus, R. A. Gopinath, and H. Guo. Introduction to Wavelets and Wavelet Transforms. Journal, Englewood Cliffs, NJ: Prentice-Hall, 1998.
[21] C. C. Lin, S. H. Chen, and T. K. Truong et al. Audio classification and categorization based on wavelets and support vector machine. Journal, IEEE Transactions on Speech and Audio Processing, Sept.2005. vol.13, no.5, pp.644-651.
[22] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification search and retrieval of audio. Journal, IEEE Multimedia Magazine, July 1996. vol.3, pp.27–36.
[23] L Rabiner. A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proceedings of the IEEE, Feburuary 1989. 77, No.2 pp.257-285.
[24] L Rabiner and B Juang. Fundamentals of speech recognition. Pretice Hall, 1993.
[25] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus. Large vocabulary continuous speech recognition of broadcast news—The Philips/RWTH approach. Speech Commun, 2002. vol.37, pp.109–131.
[26] J. Rissanen. Stochastic Complexity in Statistical Inquiry. River Edge. NJ: World Scientific, 1989.
[27] M. Cettolo and M. Federico. Model selection criteria for acoustic segmentation. Proc. ISCA ITRWASR'00 Automatic Speech Recognition, Paris, France, 2000. pp.221–227.
[28] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. IEEE Trans. Inform. Theory, 1998. vol.44, no.6, pp.2743–2760.
[29] Z. Liu, Y. Wang, and T. Chen. A robust audio classification and segmentation method. Proc. 9th ACM Int. Conf. Multimedia, 2001. pp.203–221.
[30] Potamitis I., Fakotakis N., Kokkinakis G.. Spectral and cepstral projection bases

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

constructed by independent component analysis. International Conference on Spoken Language Processing, ICSLP2000. vol.3: 63-66.

[31] Jutten C., Herault J.. Blind separation of sources. Part 1: An adaptive algorithm based on neuromimetic architecture. Signal Processing, 1991. vol.24: 1-10.

[32] G. Guo and S. Z. Li. Content-based audio classification and retrieval by support vector machines. IEEE Trans. Neural Networks, Jan.2003. vol.14, no.1, pp.209–215.

[33] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens., Aug.2004. vol.42, no.8, pp.1778–1790.

[34] P. Clarkson and P. J. Moreno. On the use of support vector machinesfor phonetic classification. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process, Mar.1999. vol.2, pp.585–588.

[35] V. N. Vapnik. Statistical Learning Theory. New York: Wiley, 1998.

[36] F. Schwenker. Hierarchical support vector machines for multi-class pattern recognition. Proc. IEEE Fourth Int. Conf. Knowledge-Based Intelligent Eng. Syst. Allied Technologies, Sep.2000. vol.2, pp.561–565.

[37] S. Z. Li. Content-based audio classification and retrieval using the nearest feature line method. IEEE Trans. Speech Audio Process, Sep.2000. vol.8, no.5, pp.619–625.

[38] T. Zhang and C.-C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. IEEE Trans. Speech Audio Process, May 2001. vol.9, no.4, pp.441–457.

[39] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Process, Oct.2002. vol. 10, no.7, pp.504–516.

[40] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Process, 2002. vol.10, pp.504–516.

[41] S. Mallat. A Wavelet Tour of Signal Processing. New York: Academic, 1998.

[42] C. S. Burrus, R. A. Gopinath, and H. Guo. Introduction to Wavelets and Wavelet Transforms. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[43] S.-H. Chen and J.-F. Wang. Noise-robust pitch detection method using wavelet transform with aliasing compensation. Proc. Inst. Elect. Eng. Vision, Image Signal Process, Dec.2002. vol.149, no.6, pp.327–334.

[44] F. Schwenker. Hierarchical support vector machines for multi-class pattern recognition. Proc. IEEE Fourth Int. Conf. Knowledge-Based Intelligent Eng. Syst. Allied Technologies, Sep.2000. vol.2, pp.561–565.

[45] T. Luo, K. Kramer, and D. B. Goldgof et al. Recognizing plankton images from

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

the shadow image particle profiling evaluation recorder. IEEE Trans. Syst., Man Cybern.—B: Cybern, Aug.2004. vol.34, no.4, pp.1753–1762.

［46］Erling W, ThomB, Douglas K, et al. Content-based classification, search, and retrieval of audio. IEEE Multimedia, 1996. 3(3) :27～36.

［47］Lu L, Jiang H, Zhang H J. A robust audio classification and segmentation method. In : roc of the 9th ACM International Conference on Multimedia. Ottawa : ACM, 2001. 203～211.

［48］Beth L, riel S. A content-based music similarity function. CRL2001/ 2 , 2001.

［49］D.R.Xu. Research on the imagery generation in Design. Ph.D dissertation, Zhejiang University, Hangzhou, 1995.

［50］Tversky A. Feature of similarity. Psychological Review. 1977. 84 (4): 327~352.

［51］James D Wise, James R Caprio and Thomas W Parks. Maximum likelihood pitch estimation. IEEE Trans Acoustics, Speech, Signal Processing, October 1976. 24(5):418-423.

［52］A V Oppenheim. A Speech analysis-synthesis system based on homomorphic filtering. J A coustical Society of Americs, February 1969. 45, 458-465.

［53］Malcolm Slaney and Richard F Lyon. A Perceptual Pitch Detector. Albuquerque NM: in the Proceedings of the 1990 International Conference on Acoustic Speech and Signal Processing, IEEE 1990. pp.357-360.

［54］Graham, R.L., and Hell, O. On the history of the minimum spanning tree problem. Ann. Hist. Comput., 1985. 7, pp.43–57.

［55］Rabiner, L.R., and Juang, B.H. Fundamental of speech recognition, Prentice-Hall, 1993.

［57］邬显康. 基于内容的音频检索技术研究与系统实现. 西安电子科技大学硕士论文，2007 年 3 月。

［57］Wenjuan Pan, Yong Yao, and Zhijing Liu. An Unsupervised Audio Segmentation and Classification Approach. Proceedings of 2007 Fuzzy System And Knowledge Discovery(FSKD2007).

［58］Wenjuan Pan, Yong Yao, and Zhijing Liu. Audio Classification in A Weighted SVM. Proceedings of 2007 International Symposium on Communications and Information Technologies (ISCIT2007).

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在读期间发表论文

[1] 潘文娟，姚勇，刘志镜. An Unsupervised Audio Segmentation and Classification Approach. Proceedings of 2007 Fuzzy System And Knowledge Discovery(FSKD2007). Vol. 3, pp. 303-306(Cited by EI).

[2] 潘文娟，姚勇，刘志镜. Audio Classification in a Weighted SVM. Proceedings of 2007 International Symposium on Communications and Information Technologies (ISCIT2007). 已录用待刊(Cited by EI).

[3] 潘文娟，姚勇，刘志镜. An Automatic Approach towards Audio Segmentation and Classification. Proceedings of 2007 International Symposium on Intelligence Computation and Applications (ISICA'07, Cited by ISTP).

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

西安電子科技大學

地址：西安市太白南路2号

邮编：710071

网址：www.xidian.edu.cn

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.ne