分类号 TP391 密级 公开

重庆邮电大学硕士学位论文

论文题目 基于小波变换的音频特征提取与分类研究

英文题目 Research of Audio Feature Extraction and Classification Based on Wavelet Transform

硕士研究生 邢峰

指导教师 郑继明 副教授

学科专业 计算机应用技术

论文提交日期 2007年5月 论文答辩日期 2007.6.2

论文评阅人 汪继锋 教授 重庆邮电大学

黄正洪 教授 重庆工商大学

答辩委员会主席 王国胤 教授 重庆邮电大学

2007年 5 月 30 日

摘 要

视频、图像和音频等多媒体数据已经成为信息处理领域的主要信息媒体，其中音频占有很重要地地位。传统的基于文本的检索存在主观性和不完整性等缺点，为此基于内容的音频检索成为未来必然的研究和应用方向。音频的特征提取与分类是音频检索的基础。如何基于不同的规则提取更加有效的特征以及如何根据提取的音频特征进行更有效的分类是本文的主要研究工作。

本文针对基于小波变换的音频特征提取和分类的关键技术展开分析，主要集中在以下两个方面：（1）音频信号特征提取与分析。对不同变换域的特征进行表征，包括时域特征、频域特征以及时频域特征。主要是研究小波变换域的特征提取与特征描述，提取的特征包括质心、带宽、过零率、小波子带能量、基音频率等。基于不同的时间长度上的音频特征提取，主要包括基于短时音频帧的特征提取和基于音频片段的特征提取，其中基于音频片段的特征有相当一部分是在短时音频帧特征的基础上得到的，如质心、带宽等就是对每一帧的质心带宽求均值得到的；静音比和零过零率比则是在短时帧特征的基础上通过求比运算得到的，当然也有基于整个音频片段的特征，如小波子带能量、近似子带过零率周期等。与传统的特征提取相比较，基于小波变换的特征提取能够减少运算量，节省时间。（2）音频分类方法的研究。典型的音频分类算法有很多，包括神经网络法，隐马尔可夫模型法、支持向量机法、最近特征线法等。这些方法各有优劣，也有不同的适用性，本文主要研究隐马尔可夫模型方法和支持向量机方法在音频分类中的应用，并把两种分类算法结合起来设计新的分类算法，在隐马尔可夫模型训练中充分应用时间序列的优势，使用短时音频帧特征进行训练，得到样本在每个 HMM 模型下的概率，在 SVM 训练中则使用基于片段的音频特征与 HMM 概率特征进行训练，从而把音频分为纯语音、音乐、带背景音乐的语音和环境音四种类型，达到了比较好的分类效果。

关键词：小波变换，特征提取，隐马尔可夫模型，支持向量机，分类精度

Abstract

Audio data take an important part of the multimedia application. Content-based audio retrieval becomes the main aspect of research and application in future because traditional text-based audio retrieval has some disadvantages such as subjectivity and imperfection. Audio feature extraction and classification are bases of audio retrieval. The main research work of this dissertation includes the follow two aspects.

One is audio feature extraction and analysis. Features can be extracted from different transform domains such as time domain, frequency domain and time-frequency domain. Audio features include frame features whose time interval are several milliseconds and clip features whose time interval are one or two seconds. Some of the clip features are statistical of frame features, such as centroid and bandwidth, and some are the quotient of frame features, such as silence ratio. There are also some features such as wavelet sub-band energy and zero-crossing-rate ratio extracted from the whole audio clips. Compared to traditional feature extraction, wavelet method can save time.

The second is the research of audio classification. There are many typical algorithms of audio classification, such as Neural Network (NN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Nearest Feature Line (NFL) and so on. Different algorithms have different advantages and explicabilities. Both HMM and SVM are discussed in this dissertation. The powerful discriminative ability of SVM is combined with the temporal modeling ability of HMM. The frame features based on wavelet transform are used to train HMMs. Then use the HMMs to compute the probability of every train samples. Then the clip features and probability got from the HMMs are used as the input of SVM to train SVMs. Audios can be classified into pure speech, music, speech with music and environment sounds. The result shows that this HMM-SVM algorithm can get a better classification result.

Key words: Wavelet transform, Feature extraction, Hidden Markov Model, Support Vector Machine, Classification accuracy

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，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__重庆邮电大学__或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：孙峰　　签字日期：2007 年 6 月 8 日

学位论文版权使用授权书

本学位论文作者完全了解__重庆邮电大学__有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权__重庆邮电大学__可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：孙峰　　导师签名：郑建明

签字日期：2007 年 6 月 8 日　　签字日期：2007 年 6 月 8 日

第一章 绪论

1.1 研究背景

随着计算机数据处理能力的提高和多媒体编码技术的进步，多媒体逐渐成为人们经常使用的信息载体，于是查找相似的视频和音频成为人们的需要，导致了基于内容的视频、音频检索的兴起。在早期多媒体检索中，所提取的多媒体特征多是从视频（图像）中得到的视觉特征。近来，在多媒体分析和检索中，音频也蕴涵了丰富语义，如在不看电视（视频图像流），甚至不听电视中的语音信号的前提下，只听它的非语音信号，多数人就可以清楚的区分电视节目中的新闻报道和广告场景等，从而区分“鼓掌”、“广告节目”和“观众欢呼”等不同语义内容，因为这些节目中的非语音信号特征区域较大。因此，提取音频特征去识别不同多媒体场景也受到了与视频特征同样的重视。并且，由于很多视频分析是基于图像像素点处理的，计算量很大，而音频信号是一维的，可以减少计算量。

音频的非语义符号表示和无结构化组织的特点阻碍了音频应用的发展，因此，如何提取音频中的结构化信息和内容语义，使得无序的音频数据变得有序，是解决问题的关键。音频的分类，作为提取音频内容语义和结构的重要手段之一，其研究日益引起人们的重视。音频分类本质上是一个模式识别过程，包括特征抽取与分类两个基本过程。

音频是一个非平稳的随机过程，其特性是随时间变化的，但这种变化是很缓慢的。一般在对音频信号的特征进行抽取时，频域特征是基于傅立叶变换域的，有时候为了更好地反映音频信号的非平稳性，采用短时傅立叶变换的方法提取频域特征。但傅立叶变换只能给出信号变换的谱信息，而不能给出信号时域的局部信息，所以提取非平稳信号的特征时，傅立叶变换具有一定的局限性。小波分析方法是一种窗口大小固定但其形状可改变，时间窗和频率窗都可改变的时频局部化分析方法，即在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率。正是这种特性，使小波变换具有对信号的自适应性。音频信号是一种频率随时间改变而改变的振动波形信号，属于非平稳信号，因此需要从音频信号中同时获得时间和频率信息。小波变换能

够同时提取时域和频域的信息，因此可以作为傅立叶变换的一种替代方法，并且能够克服傅立叶变换的局限性。

音频分类是基于内容音频检索的关键技术之一，对分类算法的研究也很多，主要有神经元网络[1]、自组织映射聚类算法[2]、简单决策树分层分类算法[3,4]、最近特征线[5]、隐马尔可夫模型[6]、支持向量机[7]等。这些方法各有优劣，研究两种分类算法的融合，充分利用每一种分类算法的优势克服其劣势，这在音频分类算法的研究中具有很重要的意义。

1.2 音频分类与检索技术的研究现状

国内外的研究机构对音频检索进行了多方面的研究。Muscle Fish[8]是一个商业化的基于音频感知特征的音频检索引擎。Muscle Fish 公司先将带标识的数据加窗处理，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，把这 13 维特征作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，从而检索出结果。另外，MIT、Cornell 大学、南加州大学、澳大利亚 Wollongong 大学、欧洲 EUROMAEDIA 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、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。国内的一些研究单位已相继开展了基于内容的音频分类与检索的研究，并开发了一些实验系统。主要有浙江大学人工智能研究所对基于内容的音频检索、广播新闻分割与分类等领域的研究，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的研究；南京大学也开始了这方面的研究。

对音频进行处理前，通常要进行预处理，将音频流切分成时间长度较短的音频片段，一般是 1-2s，所谓的音频分类就是指对这些音频片段进行识别的过程。从本质上讲音频分类是一个模式识别过程，包括特征抽取和分类两个基本过程。音频分类技术的研究涉及到许多相关技术，包括人耳听觉特征、信号与系统、数字信号处理、语音信号处理、模式识别、机器学习、人工智能、数据挖掘等。目前的研究重点主要包括音频特征分析与抽取，分类器的设计与实现。

1.2.1 音频特征分析与提取

音频特征分析与提取是音频分类的基础，如果选取的特征不好，那么即使设计再好的分类算法，最后的分类效果也是不理想的。因此，选取的特征应该能够充分表示音频频域和时域的重要分类特性，对环境的改变也要具有较好的鲁棒性和一般性。音频分类和检索中常用到的特征是 Mel 倒谱系数（MFCC）特征，该特征比较符合人类的听觉特性，在音频分类与检索中得到了广泛的应用。在音频特征分析领域常用的特征还包括：短时能量、短时过零率、音调、带宽、质心、静音比等。

当然，很多研究人员也致力于新的特征提取方面的研究，使他们提取出的特征能够应用到特定的领域中去。微软亚洲研究院的 Hao Jiang，Lie Lu 等人[3]为了提高环境音的识别精度，提出了噪音率和带周期等新的音频特征，取得了不错的效果。在文献[9]中，提出了一些新的特征诸如谱流量、波段周期和噪音帧比，使用基于核心的支持向量机，通过非线性分类提高了分类的速率，把音频分为静音、音乐、背景音、纯语音和非纯语音，分类的准确率约为 90 %。随着音频分类与检索技术研究的发展，音频分类更加细化，如何抽取能够准确表征音频类别的特征是特征分析和提取的重点。

传统的音频特征提取包括两个方面。其一是基于不同变换域提取的音频特征，主要有音频时域特征提取，提取的特征一般有短时平均能量、过零率、线性预测系数等；音频频域特征提取，对音频进行傅立叶变换，提取的特征一般有质心、带宽、LPC 倒谱系数、MFCC 倒谱系数等。另一方面是基于不同时间长度的音频特征提取，主要有基于音频帧的特征提取和基于音频片段的特征提取。

传统的傅立叶变换方法在信号处理中一直占统治地位，这是因为傅立叶变换对信号的确定性和平稳性的分析具有较强的优势。但是在现实生活中，某些信号具有很强的时变特性，如在某一段短时间内呈现出周期信号的特性，而在另外一些时间段内却呈现出噪声特性。对于时变剧烈的音频信号，仅仅在频谱空间上进行傅立叶变换具有一定的局限性。

小波变换方法是一种广泛应用的信号处理方法，它也成功的应用在语音和音频的特征提取中。同传统的傅立叶变换方法相比，小波变换方法具有下面的优势：小波变换具有恒 Q 性质，在信号的高频部分，可以取得较好的时间分辨率，在时间的低频部分，可以取得较好的频率分辨率。一些

研究表明小波变换域提取的基频和子带能量特征能够有效的改善音频分类和识别的性能，同时也论证了小波变换域特征能够减少特征向量的维数，降低计算复杂度[7]。

1.2.2 音频分类

音频分类的早期研究工作以文献[1,2]为代表。文献[1]训练一种神经元网络直接将声音类别映射到所标注的文本；文献[2]使用自组织映射聚类算法对具有相似感觉特征的声音进行聚类。真正意义上的基于内容的音频分类工作是由美国 Muscle Fish[8]公司完成的，他们详细分析了音频的区别特征，并根据最近邻准则和 Mahalanobis 距离设计音频分类器。

音频分类算法主要有以下几种：

1） 层次分类算法。

文献[3,4]对音频分类的研究采用的是基于简单决策树的层次分类算法，该算法的主要思想是：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，逐步将音频分成不同种类的例子。但层次分类方法存在以下缺点：(1) 只能刻画音频的均值和方差等静态统计特性，而音频信号特征通常具有时间统计特性。(2) 决策规则和搜索顺序并不一定是最优的。(3) 上层的决策错误会积累到下一层而形成“雪球”效应。(4) 分类误差大，且需要人的先验知识和试验分析，比如阈值的确定。

因此基于简单决策树的层次分类方法精度较低，只适用于区别明显的简单的音频分类，难于满足复杂的、多特征的音频分类应用。但是由于这种分类器简单，容易实现，在大部分传统音频分类工作中有广泛的应用。J.T,Foote[10]采用的一种有监督的贪心算法构造分类决策树就是其中的代表；国防科技大学多媒体实验室开发的音频分类系统[11]也是建立在基于规则的层次分类器基础上的。

2） 模板分类算法

国防科技大学多媒体实验室的李恒峰、李国辉等人开发的基于内容的音频分类与检索系统是基于模板匹配的方法的[12]。模板匹配的思想是为每一类音频建立一个模板，然后计算实际音频的特征向量，用特征向量匹配模板向量来确定实际的音频应该属于哪一类音频。向量匹配采用的是计算向量之间的距离，常用的距离有欧几里得距离、马氏距离、相对熵距离、相关距离等。

3） 基于统计学习的分类算法

早期的基于统计学习算法的音频分类的研究主要集中在神经网络算法的应用中。Feiten.B，Frank.R[1]等训练一种神经元网络直接将声音类别映射到所标注的文本。Zhu Liu[13]等根据音频特征为每类音频训练简单的多层感知机，并根据 OCON（One-Class-One-Network）的结构实现它们的连接，进行天气预报、新闻、广告、足球和篮球等电视节目的视频场景分类。虽然神经网络（Neural Network）方法在其它分类问题中应用非常广，但它在音频分类中的应用却不多，原因是神经网络中所需的很多参数都是人工凭经验选定的，会产生过量匹配和陷入局部最小，而且具有时序功能的神经网络如 TDNN,RTRLN 和 BPTTN 等，其拓扑结构比较复杂，训练和分类的计算量都相当大[14]。

近年来，随着人工智能，机器学习领域的快速发展，越来越多的研究者将隐马尔可夫模型（HMM）、支持向量机（SVM）等统计学习模型应用到了音频分类研究中[6,7,9]。HMM 在本质上是一种双随机过程的有限状态自动机，它具有刻画信号的时间随机统计特性的能力，可以克服基于简单决策树的层次分类方法的缺点。SVM 是 Vapnik 等[15]人提出的以结构风险最小化原理为基础的一种分类方法，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳的折中，具有较好的泛化能力。吴飞等[16]提出了增量学习支持向量机训练算法，从大训练样本库中发现“好样本”，清除冗余样本。

基于统计学习算法的音频分类方法是音频分类研究的重点，它为自动和自助学习分类的实现提供了一种行之有效的途径，是目前和未来的主要研究方向。

统计学习算法中的 HMM 和 SVM 应用在音频分类中各有优劣，如何结合两种算法的优势，而克服各自的缺点设计一种新的有效的分类算法，很值得研究。Jianjun Ye 等[17]首先用基于 HMM 的分类器进行分类，接下来使用 SVM 解决那些用 HMM 不能分类的不确定的部分。其中用到了动态时间卷积核函数(DTWK)，并把欧几里德距离应用到高斯核函数中：$K(X,Y)=\exp\{-\lambda D(X,Y)\}$。这种方法提高了手势语言的识别精度。LIU Jiang-hua 等[18]综合应用 SVM 的判别力与 HMM 的短时性能。SVM 的概率输出代替 HMM 中的高斯混合输出。用小波变换方法提取观测向量，减少了数据的维数并且提高了鲁棒性。实验结果表明，SVM 与 HMM 的混合算法同样能够提高识别精度。在文献[19]中，首先为每一类训练对应的 HMM，然后计算每一观察序列在各个 HMM 中的输出概率，经规格化处理后作为

SVM 的训练样本，实现音频的分类。该方法能够提高音频的分类精度，并且具有开放性，当需要增加新的类别时，只需要单独训练该新类别的 HMM，并利用 SVM 增量学习算法比较容易训练得到新的分类器。目前对于 HMM/SVM 相结合的算法应用到音频分类领域的研究还比较少，需要做更进一步的研究。

1.3 论文的研究内容与结构

本文的研究主要涉及到以下两个方面：一方面是特征的提取与分析，主要有基于不同变换域的特征提取与分析和基于不同时间长度的特征提取与分析。对于不同变换域的特征提取的研究主要是基于小波变换域的特征提取，提取的特征包括质心、带宽、小波子带能量、过零率、静音比等；另外还对小波变换域中基音频率的提取算法进行研究。基于不同时间长度的音频特征，是把一些常见和常用的特征按照提取时基于时间片段的长度分为帧特征与段特征。另一方面是对音频分类算法的研究，主要研究是基于统计学习理论的 HMM 和 SVM 算法，并根据 HMM 的时间规整能力和 SVM 的泛化能力，把短时音频特征和音频片段特征相结合，设计实现一种基于 HMM 和 SVM 的新的音频分类算法。

本文共分五章，各章的内容安排如下：

第一章介绍了音频检索与分类出现的背景，音频的特征分析与提取的研究现状以及现有音频分类算法的研究现状。

第二章介绍了音频分类中的小波理论，包括小波分析、小波变换、多分辨分析以及小波变换在音频分类识别中的优势。

第三章对音频分类中应用到的特征进行分析和提取，主要是基于小波变换的音频特征分析与提取，对在小波域和傅立叶变换域的相同特征进行比较；同时也简单介绍了基于时域和傅立叶变换频域的音频特征。另外还基于不同时间长度把音频特征进行分类，分为短时音频帧特征和音频片段特征。

第四章主要介绍两种音频分类的算法 HMM 和 SVM 的原理和方法，并在两种算法的基础上设计一种基于 HMM/SVM 的分类算法，利用提取到的短时帧特征训练 HMM，并随时调整 HMM 的参数，对每一个音频计算其在 HMM 下的 log 概率作为新的特征，和音频的其他片段特征一起作为 SVM 的输入向量，训练新的 SVM 分类器，从而实现了音频的分类。

第五章总结本文所做工作，并探讨进一步的研究方向。

第二章 音频分类中的小波理论

2.1 小波分析

小波变换是本世纪 80 年代发展起来的一个强有力的信号分析工具，小波变换克服了傅立叶变换和窗口傅立叶变换的缺点，具有窗口自适应性，能对信号作不同尺度的分析，通过变换突出信号的某些方面的特征，所以在信号处理中得到了广泛应用。

小波变换的概念是由法国地质学家 J.Morlet 和 A.Grossmann 首先提出的，并用于分析处理地质数据，引进了以他们的名字命名的时间-尺度小波即 Grossmann-Morlet 小波[20]；1986 年著名数学家 Y.Meyer 构造出一个具有一定衰减性质的光滑函数，这个函数的二进尺度伸缩和二进整数倍平移产生的函数系构成著名的函数空间 $L^2(R)$标准正交基[21]，并与 S.Mallat 合作提出了多分辨分析的概念，这标志着小波分析蓬勃发展起来，其中比利时女作家 I.Daubechies 撰写的《小波十讲（Ten Lectures on Wavelet）》对小波的普及起了重要的推动作用。小波变换是一个时间和频率的局域变换，能有效地从信号中提取信息，通过伸缩和平移等运算功能对函数或信号进行多尺度细化分析，解决傅立叶变换不能解决的许多困难问题，享有“数学显微镜”的美称。

小波分析的应用是与小波分析理论的研究紧密地结合在一起的，现在，它已经在科技信息产业领域取得了令人瞩目的成就。电子信息技术是六大高新技术中重要的一个领域，它的重要方面是图像和信号处理。信号处理已经成为当代科学技术工作的重要部分，信号处理的目的就是准确的分析、诊断、编码压缩和量化、快速传递或存储、精确地重构。从数学的角度来看，信号与图像处理可以统一看作是信号处理（图像可以看作是二维信号），在小波分析的许多应用中，都可以归结为信号处理问题。对于性质随时间是稳定不变的信号，处理的理想工具仍然是傅立叶分析，但是在实际应用中，绝大部分信号都是非稳定的，而特别适用于非稳定信号的工具就是小波分析。

事实上，小波分析的应用领域十分广泛，它包括数学领域的许多学科；信号分析、图像处理；量子力学、理论物理；军事电子对抗与武器的智能化；计算机分类与识别；音乐与语言的人工合成；医学成像与诊断；地震

勘探数据处理；大型机械的故障诊断等方面。

小波分析用于信号与图像压缩是小波分析应用的一个重要方面。它的特点是压缩比高、压缩速度快、压缩后能保持信号与图像的特征不变，且在传递中可以抗干扰。

小波在信号分析中的应用也十分广泛。它可以用于边界的处理与滤波、时频分析、信噪分离与提取弱信号、求分形指数、信号的识别与诊断以及多尺度边缘检测等。

在工程技术等方面的应用。包括计算机视觉、计算机图形学、曲线设计、湍流、远程宇宙的研究与生物医学方面。

2.2 小波变换

从数学上讲，小波变换是一种数学工具，它把数据、函数或算子分割成不同频率的成分，然后去研究对应尺度的成分；技术上讲，它是一种变换方法；同时也是一种可伸缩可拓展的思想。

给定一个基本函数 $\psi(t)$，令

$$\psi_{a,b}(t)=\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{2.1}$$

式中 a,b 均为常数，且 $a>0$。显然，$\psi_{a,b}(t)$ 是基本函数 $\psi(t)$ 先作平移再作伸缩以后得到的。若 a,b 不断地变化，我们可得到一族函数 $\psi_{a,b}(t)$。给定平方可积的信号 $x(t)$，即 $x(t)\in L^2(R)$，则 $x(t)$ 的小波变换（Wavelet Transform, WT）定义为：

$$\begin{aligned} WT_x(a,b)&=\frac{1}{\sqrt{a}}\int x(t)\psi^*(\frac{t-b}{a})dt \\ &=\int x(t)\psi^*_{a,b}(t)dt=<x(t),\psi_{a,b}(t)> \end{aligned} \tag{2.2}$$

式中 a,b 和 t 均是连续变量，因此该式又称为连续小波变换（CWT），式中及以后各式中的积分都是从 $-\infty$ 到 $+\infty$。信号 $x(t)$ 的小波变换 $WT_x(a,b)$ 是 a 和 b 的函数，a 是尺度因子，b 是平移因子。$\psi(t)$ 称为基本小波或母小波。$\psi_{a,b}(t)$ 是母小波经平移和伸缩所产生的一族函数，称为小波基函数或简称小波基。这样小波变换又可解释为信号 $x(t)$ 和一族小波基的内积。

在实际应用中，特别是在计算机上有效地实现小波变换时，信号总是要取成离散的，因此研究 a、b 及 t 都是离散情况下的小波变换，进一步发展快速小波变换算法具有重要的意义。

令 $a=a_0^j, j\in Z$，可以实现对尺度因子 a 的离散化。若 $j=0$，则

$\psi_{j,b}(t)=\psi(t-b)$。欲对平移因子b离散化，最简单的方法是将b均匀抽样，如令$b=kb_0$，b_0的选择应保证能由$WT_x(a,b)$来恢复出$x(t)$。当$j\neq 0$时，将a由$a=a_0^{j-1}$变成a_0^j时，即是将a扩大了a_0倍，这时小波$\psi_{j,k}(t)$的中心频率比$\psi_{j-1,k}(t)$的中心频率下降了a_0倍，带宽也下降了a_0倍。因此，这时对b抽样的间隔也可相应的扩大a_0倍。由此可以看出，当尺度因子a分别取$a_0,a_0^1,a_0^2,\cdots,$时，对b的抽样间隔可以取$a_0b_0,a_0^1b_0,a_0^2b_0,\cdots,$这样，对$a$和$b$离散化后的结果是：

$$\begin{aligned}&\psi_{j,k}(t)=a_0^{-j/2}\psi[a_0^{-j}(t-ka_0^jb_0)]\\&=a_0^{-j/2}\psi(a_0^{-j}t-kb_0)\end{aligned}\qquad j,k\in Z \tag{2.3}$$

对给定的信号$x(t)$，连续小波变换可变成如下离散小波变换，即

$$\begin{aligned}&WT_x(j,k)=\int x(t)\psi_{j,k}(t)dt\\&=\int x(t)a_0^{-j/2}\psi(a_0^{-j}t-kb_0)dt\end{aligned} \tag{2.4}$$

此式称为“离散小波变换（Discrete Wavelet Transform， DWT）”。式中t仍是连续变量。取$a_0=2$，$b_0=1$即尺度轴取以 2 为底的对数坐标，称为二进小波变换，表示为

$$\psi_{j,k}(t)=2^{-\frac{j}{2}}\psi(2^{-k}t-k) \tag{2.5}$$

二进小波介于连续小波和离散小波之间，它只是对尺度参量进行离散化，在时间域上的平移量仍保持连续变化，因此二进小波具有连续小波变换的时移共变性，这是它较离散小波变换所具有的连续的独特优点。而且二进小波容易用计算机实现，并且有相应的快速算法，所以在工程和科技研究中得到了广泛应用。

2.3 多分辨分析（Multi-resolution Analysis）和塔式算法

多分辨分析又称多尺度分析是建立在函数空间概念上的理论。它是在$L^2(R)$函数空间内将函数$f(x)$描述为一系列近似函数的极限。每个近似都是函数$f(x)$的平滑版本，而且具有越来越细的近似函数，这些近似都是在不同尺度上得到的，因此称之为多分辨分析。

Mallat 给出了多分辨分析的数学定义：

设$\{V_j\},j\in Z$是$L^2(R)$空间中的一系列闭合子空间，如果它们满足下面性质，则说明$\{V_j\},j\in Z$是一个多分辨近似。这些性质是：

单调性（包容性）：$\forall j$，则有$\cdots V_0\supset V_1\supset V_2\cdots V_j\supset V_{j+1}\supset\cdots$；

逼近性：$\lim_{j\to\infty} V_j = Closure(\bigcup_{j=-\infty}^{\infty} V_j) = L^2(R), \lim_{j\to\infty} V_j = \bigcap_{j=-\infty}^{\infty} V_j = \{0\}$；

伸缩性：$\forall j \in Z$，若 $x(t)\in V_j$，则 $x(\frac{t}{2})\in V_{j+1}$；

平移不变性：$\forall (j,k)\in Z^2$，若 $x(t)\in V_j$，则 $x(t-2^j k)\in V_j$；

Riesz 基存在性：存在一个基本函数 $\theta(t)$，使得 $\{\theta(t-k)\}$，$k\in Z$ 是 V_0 中的 Riesz 基。

Mallat 在用于图像分解的金字塔算法（Pyramidal algorithm）的启发下，结合多分辨分析，提出了信号的塔式多分辨分解与综合算法，通常称为 Mallat 算法。

设 $f(t)\in L^2(R)$，并假定已得到 $f(t)$ 在 2^{-j} 分辨率下 $A_j f\in V_j$，$\{V_j\}_{j\in Z}$ 构成 $L^2(R)$ 的多分辨分析，从而有 $V_j = V_{j+1}\oplus W_{j+1}$，即：

$$A_j f = A_{j+1} f + D_{j+1} f \tag{2.6}$$

其中 $A_j f = \sum_{-\infty}^{\infty} C_{j,k}\phi_{j,k}(t)$，$D_j f = \sum_{-\infty}^{\infty} D_{j,k}\psi_{j,k}(t)$，于是(2.6)变为：

$$\sum_{-\infty}^{\infty} C_{j,k}\phi_{j,k}(t) = \sum_{-\infty}^{\infty} C_{j+1,k}\phi_{j+1,k}(t) + \sum_{-\infty}^{\infty} D_{j+1,k}\psi_{j+1,k}(t) \tag{2.7}$$

由尺度函数的双尺度方程可得：$\phi_{j+1,k}(t) = \sum_{k=-\infty}^{\infty} h(k-2m)\phi_{j,k}(t)$，利用尺度函数的正交性，有：$<\phi_{j+1,m},\phi_{j,k}>= h(k-2m)$，同理由小波函数的双尺度方程得 $<\psi_{j+1,m},\psi_{j,k}>= g(k-2m)$。

于是可得

$$C_{j+1,m} = \sum_{k=-\infty}^{\infty} C_{j,k} h^*(k-2m) \tag{2.8}$$

$$D_{j+1,m} = \sum_{k=-\infty}^{\infty} D_{j,k} g^*(k-2m) \tag{2.9}$$

$$C_{j,k} = \sum_{m=-\infty}^{\infty} h(k-2m)C_{j+1,m} + \sum_{m=-\infty}^{\infty} g(k-2m)C_{j+1,m} \tag{2.10}$$

其中 $h(k-2m)$ 和 $g(k-2m)$ 分别是低通滤波器和高通滤波器，分别记做 H,G，那么上面(2.8)- (2.10)式子变为：

$$C_{j+1} = H^* C_j \tag{2.11}$$

$$D_{j+1} = G^* D_j \tag{2.12}$$

$$C_j = HC_{j+1} + GD_{j+1} \tag{2.13}$$

其中 H^*,G^* 分别是 H,G 得共轭转置矩阵。这就是 Mallat 塔式算法。对

于信号 $C_0(k)$， Mallat 多分辨分解和重构的过程如图 2.1，2.2 所示。

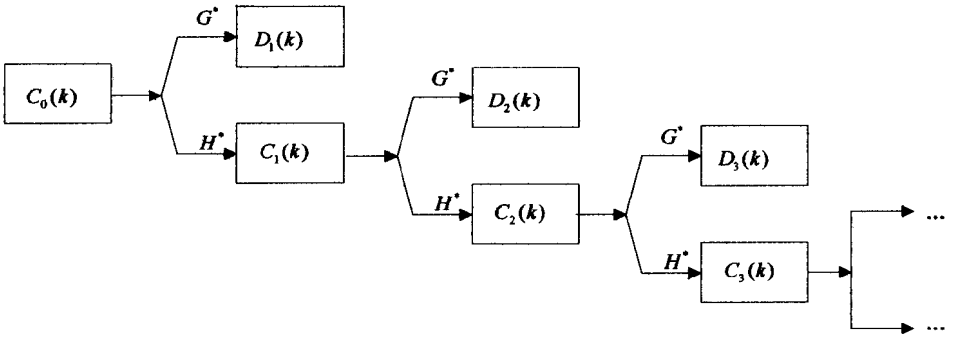


图 2.1 信号多分辨分析的 Mallat 分解示意图

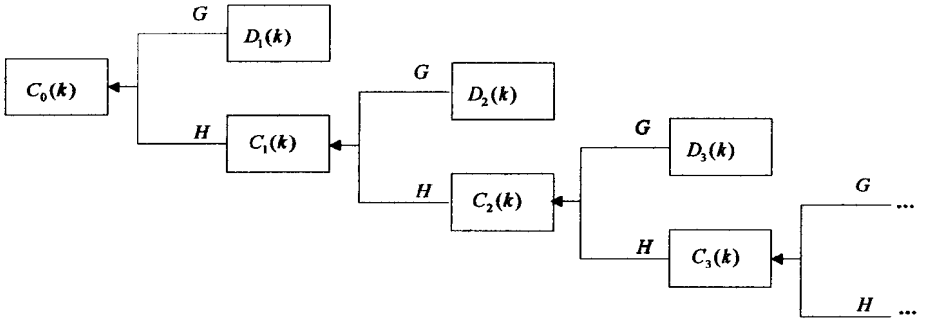


图 2.2 信号多分辨分析的 Mallat 重构示意图

音频信号的感知是一个复杂的过程，除了要利用一些先验知识来指导外，重要的就是要根据音频信号本身所包含的信息对音频进行研究。事实表明，小波的计算特性与人耳的感知过程具有相似性，那么就可以利用小波变换理论，在小波域提取特征作为音频分类和识别的基础。同时在特征提取之前，也可以用小波进行处理，选取音频信号的有用信息，并抑制无关信息对分类和识别产生的干扰。

2.4 小波变换在音频分类识别中的应用

小波分析的应用是与小波分析的理论研究紧密地结合在一起的。小波分析在对数尺度上可将信号分解为相同宽度的频率通道组，这种特性十分接近人耳对声音信号的感知。人的发音过程中，由于声门的瞬间闭合，声道被强烈激励，表现在语音信号波形上就是瞬间突变，小波变换为检测这种状态提供了有力的数学工具。

基于小波计算特性与人耳感知过程的相似性，我们利用小波多尺度性质，在提取分类识别特征之前，用小波进行预处理，选取语音信号的有用信息，并且抑制无关信息对识别所产生的干扰，从而有效地提高了系统的分类准确率与识别率。

2.5 小结

本章主要介绍了音频分类中用到的小波理论的基础知识，包括小波分析的基本原理、离散小波变换以及小波变换的多分辨分析和塔式算法等。同时也说明了小波变换应用在音频分类中的可行性及优势。本章是小波变换域音频特征提取和分析的理论基础。

第三章 音频特征提取与分析

音频特征提取与分析是音频分类和检索的前提和基础，所选取的特征应该能够充分表示音频频域和时域的重要分类特性，对环境的改变具有鲁棒性[9]和一般性。从信号时域或频域处理方式上，可以将音频特征分为时域特征、频域特征和时频特征；从信号是否短时平稳方面考虑，可以将音频特征分为短时音频帧特征和音频例子特征，其中音频例子特征是短时音频帧特征的统计特征，像均值、方差、比值等。本章简要介绍了音频特征提取的相关技术，重点是小波变换域的音频特征提取与分析，同时也分析了不同时间长度上的音频特征。

人耳听到的音频是连续模拟信号，而计算机只能处理数字化的信息，所以模拟连续音频信号要经过离散化即抽样后变成计算机处理的采样离散点。奈魁斯特采样频率指出，音频信号数字化时的采样率必须高于信号带宽的 2 倍，才能正确恢复信号。

3.1 音频信号的短时分析

音频信号是一个非平稳信号，其特征是随时间变化的，但这种变化很缓慢。常用的信号处理方法比如傅立叶变换，自相关算法等都是针对平稳信号的。考虑到音频的形成过程是与发声器官的运动密切相关的，这种物理运动比起音频振动速度来讲要缓慢的多，因此音频信号常常可假定为短时平稳的，即在 10-20ms 这样的时间段内其频谱特性和某些物理特征参数可近似地看作是不变的。鉴于此，可以将音频信号分成一些相继的短段进行处理，这就是短时处理技术。基于上述特点，我们常常以 10-20ms 步长为语音信号分帧。

特征提取前，需要对原始音频数据做预处理：首先对原始音频信号做预加重处理，减少尖锐噪声的影响，提升高频信号，设 $x(n)$ 为原始信号，处理后信号 $y(n)=x(n)-$参数$*x(n-1)$，其中，参数通常取 0.98 或 0.97。短时分析将音频分成一段一段来处理，每一段称为一“帧”，为保证信号的平滑性，相邻的帧之间一般都有重叠部分，即帧移，通常取 0～1/2 帧长。为了减小音频帧的截断效应，需要加窗处理，即 $s_w(n)=s(n)w(n)$，$w(n)$ 是窗函数。通常选用 Hamming 窗，其数学表示为：

$$w(n)=0.54-0.46\cos(2\pi\frac{n}{N-1}),0\le n\le N-1.$$

文中所用到音频信号的采样频率都是 22050Hz，把音频信号分成短时音频帧，每帧包含 256 个采样点，相邻帧之间有 80 个采样点的叠加。在 matlab 中分帧程序是 enframe.m，对信号 x 进行分帧处理的调用语句是 enframe（x，256，80）。

3.2 基于小波变换的特征描述与分析

小波分析方法是一种窗口大小固定但其形状可改变，时间窗和频率窗都可以改变的时频局部化分析方法。同傅立叶变换相比，能够提供可变的时频窗，这样更符合人耳的分辨特性。目前，许多学者都致力于研究小波变换域中的特征提取[22,23,24,25]。文献[22]比较了小波变换方法提取的特征和传统方法提取的特征的分类性能，应用 DB7 小波进行 6 层变换，提取的特征包括质心、带宽、子带能量、子带方差和过零率等，实验结果表明从小波域中提取的特征能够在提高分类准确率的同时节省了计算时间。而在文献[7]中，则综合应用小波域中提取的子带能量和基音频率特征与傅立叶变换域中得到的特征质心、带宽、频率谱系数等特征，用支持向量机的分类方法，完成音频信号的分类，达到了较高的分类准确度。小波变换域中的某些特征能够更好的反映音频信号的局部特性，不同音频之间的区别有时候就在于局部的差别，因此可以在小波变换域中提取分析更加有效的能够区分不同音频的特征。

根据前面小波变换理论，我们知道，对信号 $f(x)$ 进行小波分解相当于经过一系列高通滤波器（对应 G）和低通滤波器（对应 H）的滤波，得到各个尺度的细节信号及尺度信号。信号经过一次小波变换（$j=1$）相当于通过一个高通滤波器和一个低通滤波器，得到高通部分（细节部分），其频率范围为 $f/2\sim f$，低通部分（近似部分），其频率范围是 $0\sim f/2$；同理，$j=2$ 时所得到的细节的频率范围在 $f/4\sim f/2$，近似部分的频率范围在 $0\sim f/4$，当 $j\to\infty$ 时，细节部分和近似部分都接近于 0。

把小波滤波器分离出来的各尺度下的剩余信号和原始信号，分别提取带通滤波器族特征进行实验，结果表明，音频信号的一次小波变换后信号中携带的信息并没有消失，并且比单纯用原始信号效果更好，这是因为小波变换的带通滤波器对无用信号进行了衰减和部分滤出，非常类似于人耳的感知过程。而第二、三、四层小波变换近似信号的识别率逐渐降低，是

因为随着小波滤除器中心频率降低，带宽减小，一些有用成分被逐渐滤掉，丢失很多信息。可以看出，小波变换后提取的特征对音频分类和识别的效果更好。文献[24,25]得出这样的结论：对音频信号进行3层小波分解后提取的特征，得到的分类识别性能较好。因此下面对小波特征的提取都是基于3层小波分解后的。

3.2.1 Daubechies小波基的选取

对音频信号进行小波变换处理提取特征时，一般都采用Daubechies小波，像文献[22,23]。Daubechies小波有以下特点：

(1) 时域上为有限支撑；

(2) 高阶原点矩 $\int t^p \psi(t)dt = 0$， $p = 0 \sim N$， N越大$\psi(t)$长度越长；

(3) 频域上$\psi(\omega)$在$\omega = 0$处有N阶零点；

(4) $\psi(t)$与其整数位移正交归一，即 $\int \psi(t)\psi(t-k)dt = \delta(k)$；

(5) $\psi(t)$可由尺度函数$\varphi(t)$求出。

(6) Daubechies小波基容易计算机实现。

3.2.2 质心和带宽

质心是度量音频亮度的指标，在傅立叶变换域中的定义为 $FC = \int_0^{\bar{w}} |F(w)|^2 \, dw / E$，其中$F(w)$表示信号$f(t)$的傅立叶变换，$\bar{w}$是采样频率的一半，$E$是频域能量，$E = \int_0^{\bar{w}} |F(w)|^2 \, dw$。一般的，音乐的频率中心比语音的要高，语音的频率中心相对较低。

带宽是衡量音频频域范围的指标，根据质心的表述可以定义为 $BW = \sqrt{\int_0^{\bar{w}} (w - FC)^2 |F(w)|^2 \, dw / E}$，一般的，语音的带宽范围在0.3kHz~3.4kHz左右，而音乐的带宽范围比较宽。

取一组样本数据，包括语音和音乐两种，首先对音频信号进行分帧处理，然后用三种方法分别计算每一帧音频的质心和带宽。

方法一：首先对每一帧音频信号进行快速傅立叶变换，然后按照上面

的定义提取每一帧音频的质心和带宽，可以得到图 3.1 所示的结果。从图 3.1 中可以看出来，语音和音乐的质心与带宽在频域中的区分并不是十分明显。

方法二：基于小波变换和傅立叶变换的质心和带宽的计算。首先对音频信号进行一次多尺度小波分解，采用 DB4 小波，进行 3 尺度分解，然后把分解后得到的近似向量 CA_3 作为下面处理的新的信号，记为 $f(t)$。对 $f(t)$ 进行分帧，然后对得到的每一帧进行快速傅立叶变换，再按照上面定义的质心和带宽进行计算。可以得到图 3.2 所示的结果。从图 3.2 中可以看出，音乐与语音的质心和带宽区分相当明显，这就说明对小波变换后的近似信号进行的质心带宽特征比直接对信号计算质心带宽特征要好的多，并且计算量减少了一大半，大大降低了程序运行的时间。

方法三：对音频信号也进行一次多尺度小波分解，同样采用 DB4 小波，进行 3 尺度分解，然后将分解得到的近似向量 CA_3 作为下面处理的新的信号 $f(t)$。对 $f(t)$ 进行分帧处理。类似于傅立叶变换域中质心和带宽的定义，用小波系数重新定义质心和带宽如下：

$$FC' = \frac{\sum_{i=1}^{N} i\,|w_i|^2}{\sum_{i=1}^{N} |w_i|^2}, \quad BW' = \sqrt{\frac{\sum_{i=1}^{N} (i-FC)^2\,|w_i|^2}{\sum_{i=1}^{N} |w_i|^2}} \tag{3.1}$$

式中 w_i 表示第 i 个小波系数，N 表示小波系数的总数。

分别计算每一帧音频的质心和带宽，可以得到如图 3.3 所示的结果。从图 3.3 中，可以看出来，用小波系数定义的质心和带宽同样能够明显的区分语音和音乐，带宽特征对语音音乐的区分更加明显，一般音乐的带宽比语音的带宽要高。

通过比较上面三种方法，我们发现，基于傅立叶变换提取的质心和带宽特征对语音和音乐的区分不是特别明显，虽然一般的音乐质心和带宽要比语音的高，但是有的时候语音的质心和带宽也会比较大，从图 3.1 就可以看出来。并且第一种方法是对整个音频信号进行分帧处理，如果分成 10ms 的短时音频帧，对于 2s 的音频则需要分几百帧，对每一帧都提取质心和带宽，计算量非常大，要花费很多计算时间，实验证明方法一提取特征要花费的时间达到几秒钟，这在实际应用中显得很不合适。方法二和方法三经过了一次小波变换，滤除了部分细节信号，得到了保留大部分信息的近似信号，也能够反映语音与音乐的大部分信息，减少了计算量，节省了计算时间。实验表明，方法二、方法三提取的质心带宽特征能更好的区

分语音和音乐。

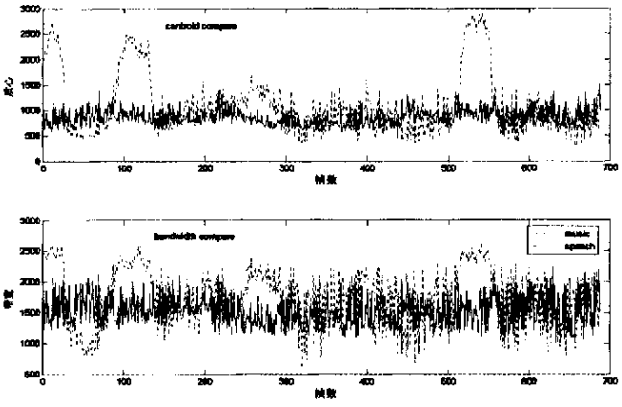


图 3.1 基于傅立叶变换的语音和音乐的质心、带宽比较

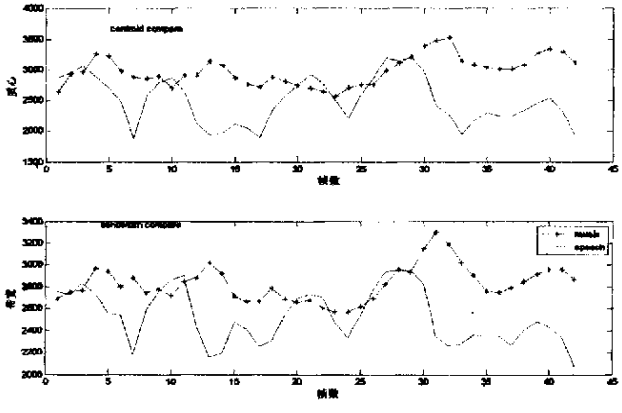


图 3.2 近似子带基于傅立叶变换的语音和音乐的质心、带宽比较

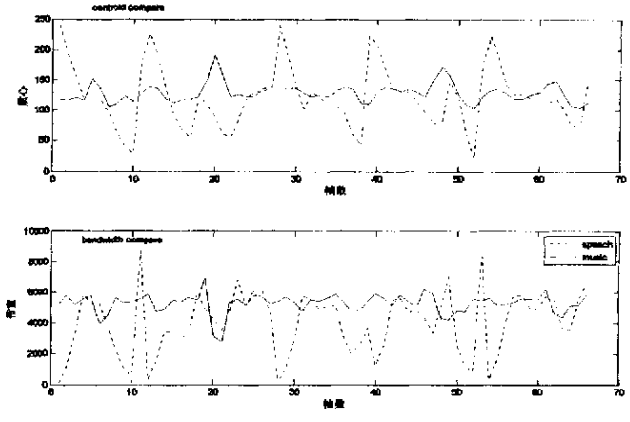


图 3.3 基于小波变换的语音和音乐的质心、带宽比较

通过做类似的实验，可以发现用方法二和方法三提取的质心、带宽特征也能够很好区分其他不同种类的音频。环境音的质心和带宽要比音乐的质心和带宽高，有背景音乐的语音的质心和带宽比纯语音的要高，但是对音乐与带背景音乐的语音的区分不是很明显。因此，质心和带宽特征可以作为分类的主要依据特征，尤其是带宽特征，但不是唯一的特征，需要研究其他的特征。

3.2.3 子带能量

将频域划分成几个子带区间，假设为 4 个，在傅立叶变换域内分别记为 $[0\sim\overline{w}/16]$，$[\overline{w}/16\sim\overline{w}/8]$，$[\overline{w}/8\sim\overline{w}/4]$和$[\overline{w}/4\sim\overline{w}]$，并计算各自的子带能量 SE_0，SE_1，SE_2，SE_3，$SE_i=\int|F(w)|^2\,dw$，其中积分限是每个区间的端点值。不同类型的音频，其能量在各个子带区间的分布有所不同。音乐的频域能量在各个子带区间的分布比较均匀，而在语音中，能量则主要集中在第 0 个子带，约占 80%以上。

在小波变换域中定义小波子带能量则相对比较简单，只需要作一次多尺度小波变换即可。假设仍然要得到四个子带，则只需要作尺度为 3 的小波变换。小波域中子带能量的定义如下：

$$SE_i=\sum_{k=1}^{N_i}\left|w_i^k\right|^2. \tag{3.2}$$

式中，w_i^k表示第i个子带的第k个小波系数，N_i表示第i个子带的小波系数的总数。根据小波域中子带能量的计算，我们可以看出小波子带能量可以明显的区分不同种类的音频。由子带能量可以延伸出各子带能量占总能量的比，定义为 $ER_i=\dfrac{SE_i}{\sum_{i=0}^{N}SE_i}$，$N$为小波子带的总数，这样可以把能量值规范到（0，1）之间，用来简化计算。

选择一组纯语音和音乐信号作为样本，分别作尺度为 3 的小波变换，采用 DB4 小波基，则由小波变换的多分辨分解得到四个子带，近似子带 ca_3 与细节子带 cd_3，cd_2，cd_1，对每个子带信号进行分帧处理，按照前面小波域中能量的定义，分别计算各个子带中每一帧的能量，得到的结果如图 3.4 所示。

从图 3.4 中可以看出，语音信号的能量大部分集中在近似子带，即 ca_3

表示的子带中，而音乐的能量分布则比较分散，在 ca_3，cd_3 中都有大部分的能量分布，在 cd_2，cd_1 表示的子带中也有部分能量分布，并且在这些子带中的能量都比语音的能量大。由小波多分辨分解的理论知，ca_3 表示的是信号的低频部分，cd_3、cd_2、cd_1 则分别表示信号的高频部分，他们所表示的频率带是逐渐增加的，因此可以得出这样的结论：语音信号的能量主要集中在小波子带的低频部分，而音乐信号的能量则在低频与高频小波子带中都有分布，但也主要是分布在低频和频率较低的高频部分，其他高频部分也有少部分的能量分布，但不是主要部分，音乐信号在高频子带的能量要比语音信号在这些子带的能量大。

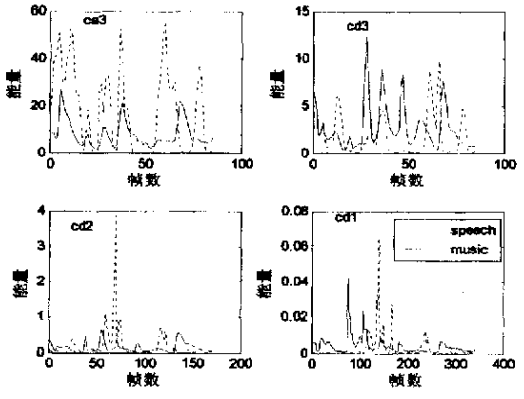


图 3.4 语音音乐在各小波子带能量分布的比较

3.2.4 过零率，零过零率比，过零率周期

“过零率（Zero-crossing Rate）”指在一个短时帧内，离散采样信号值由正到负和由负到正变化的次数，这个量大概能够反映信号在一个短时帧内的平均频率。对于音频信号流 x 中第 m 帧，其过零率计算如下：

$$Z_m = \frac{1}{2}\sum_m \left|sign[x(n)] - sign[x(n-1)]\right| w(n-m) \tag{3.3}$$

其中 $x(n)$ 表示第 m 个短时帧信号中第 n 个采样信号值，$w(n)$ 是长度为 N 的窗口函数。$sign[x(n)]$ 是符号函数，当 $x(n)\geq 0$ 时，$sign[x(n)]=1$，否则为 0。

小波子带过零率是在一个小波子带中，信号频率在一段时间内的改变次数，定义如下：

$$wzr = \frac{1}{N_k - 1}\sum_{i=1}^{N_k-1} \left| sgn[x(i)_k] - sgn[x(i-1)_k] \right| \tag{3.4}$$

式中 $\begin{cases}\mathrm{sgn}[x(i)_k]=1, 当x(i)_k \geq 0时 \\ \mathrm{sgn}[x(i)_k]=-1 当x(i)_k < 0时\end{cases}$，$x(i)_k$ 是子带 k 的第 i 个小波系数，N_k 是子带 k 的小波系数的总数。

在 matlab 中，小波域定义的过零率的计算代码如下：

```
y=enframe(x,256,80);
zcr=zeros(size(y,1),1);
delta=0.02;
for i=1:size(y,1)
    x=y(i,:);
    for j=1:length(x)-1
        if x(j)*x(j+1)<0 & abs(x(j)-x(j+1))>delta
            zcr(i)=zcr(i)+1;
        end
    end
end
```

其中 x 表示待处理的信号，设置了门限 delta＝0.02。这是个经验值，可以进行细微的调整。对于整个音频片段，可以得到如图 3.5 所示的纯语音和音乐过零率波形图。

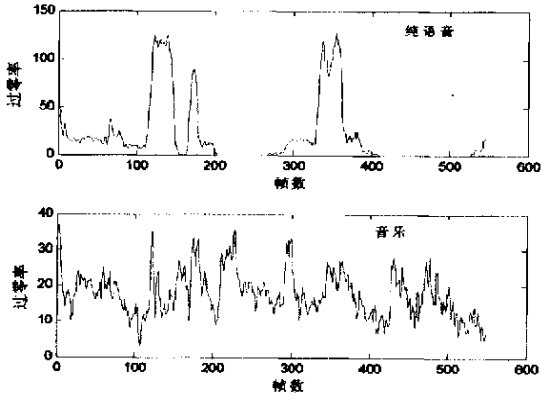


图 3.5 纯语音和音乐的过零率比较

从图中可以看出语音信号的过零率具有明显的峰值，同时有许多过零率为 0 的音频帧，而音乐的过零率分布比较均匀，过零率为零的帧数比较少。根据过零率的这一特点可以定义一个新的特征：零过零率比。如果某一帧的过零率为零，则称该帧为零过零率帧，在一个音频片段中，零过零率帧的数目与总音频帧数目的比值称为零过零率比，数学表达式为：

$$零过零率比=\frac{片段中零过零率帧的数目}{片段中帧总数} \tag{3.5}$$

对纯语音、音乐、带背景音乐的语音和环境音等音频数据分别取其中的 10 个作为样本音频，分别计算零过零率特征，可以得到如图 3.6 所示的结果。

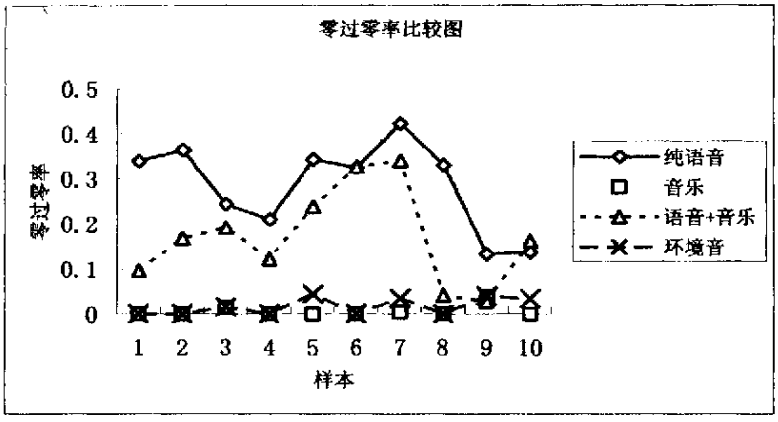


图 3.6 四类音频的零过零率比

从图中可以看出纯语音的零过零率帧在整个音频片段中占的比值较大，带背景音乐的语音零过零率比次之，音乐和环境音的零过零率比较小，在零附近。环境音的零过零率比相对还要比音乐的稍大。零过零率比可以很好的区分纯语音、带背景音乐的语音和音乐，但对音乐和环境音的区分不是很明显。

如果提取小波变换后的近似子带的过零率，选一组语音与音乐作为样本，可以得到图 3.7 所示的结果。

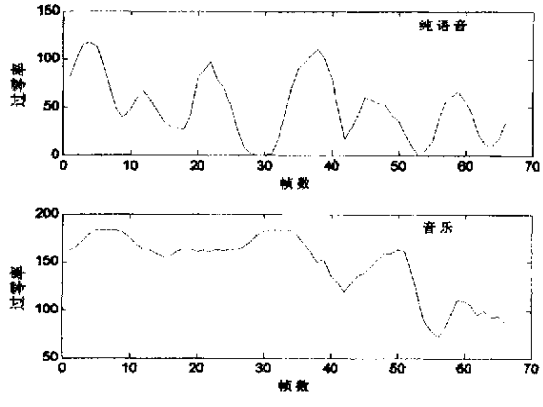


图 3.7 语音和音乐小波近似子带的过零率比较

从图 3.7 中可以看出，经过一次小波变换后得到的近似子带的过零率，语音信号具有明显的周期性，而音乐信号的周期性则很不明显，基于此，可以在小波子带中定义一个新的音频特征：过零率周期。把在小波近似子带中得到的过零率作为待处理的新的信号，然后计算该信号的自相关函数，把自相关函数的第二个峰值作为过零率周期。用到的自相关函数如下：

$$\widehat{R}_w(k)=\sum_{n=0}^{N-1}s_w(n)s_w(n+k),\quad 0\le k\le K. \tag{3.6}$$

3.2.5 静音比

静音比定义为静音帧的数量占所有音频帧数量的比率。如果帧的能量小于某一阈值，那么就把该帧看作是静音帧。静音比的定义为静音比$=\frac{\text{片段中静音帧的数目}}{\text{片段中帧总数}}$。类似地，可以基于小波系数定义小波系数静音比。如果小波系数小于某一阈值，把该系数看作是静音系数，小波系数静音比定义为小波子带中静音系数占所有小波系数的比例，数学表示如下：

$$sr=\frac{N_j}{N_k}. \tag{3.7}$$

式中 N_j 表示小波子带中静音系数的数目，N_k 表示所有小波系数的数目。

3.2.6 基音频率

基音频率，是衡量音调高低的单位，一般采用中心削波短时自相关函数的波峰检测算法计算。采用的中心削波函数是：

$$y(n)=C(n)=\begin{cases}x(n)-T, & x(n)>T\\ x(n)+T, & x(n)<-T\\ 0, & |x(n)|\le T\end{cases} \tag{3.8}$$

一般削波电平 T 取本帧音频最大幅度的 60～70%。将削波后的序列 $y(n)$ 用短时自相关函数估计基音周期，再对基音周期取倒数即为基音频率。

对于小波变换域的音频基音频率的提取方法是利用小波的多分辨分析，对音频信号进行分解，并利用其低频小波系数对音频信号的低频部分进行重构，与传统的 FFT 算法相结合，对基音频率进行估计。具体的步骤

如下：

对音频信号进行小波多分辨分解，采用 DB6 小波，由于音乐和环境音的基音频率比较高，所以分解到 3 级即可。

根据分解得到的各级系数，提取反映信号低频特征的小波系数，即 ca_3。利用低频小波系数对音频信号进行重构，这样能在基音频率提取过程中有效地抑制共振峰的影响，同时小波变换可以有效地去除噪声的影响，使得提取的基音频率的鲁棒性好。

利用 FFT 算法对重构后的音频信号进行基音估计。采用基于短时自相关函数的基音周期估计算法，计算每一帧的基音周期，然后对基音周期求倒数可得到基音频率。自相关函数采用的是修正的短时自相关函数：

$$\hat{R}_w(k)=\sum_{n=0}^{N-1}s_w(n)s_w'(n+k),\quad 0\le k\le K. \tag{3.9}$$

$$s_w'(n)=s(n)w'(n).$$

$$w'(n)=1,\quad 0\le n\le N-1+K.$$

上面介绍了基于小波变换的音频特征的提取，同时还可以基于时域和频域进行特征提取。

基于时域的音频特征提取：在音频时域特征提取中，认为每个采样点 $x(n)$ 包含了这一时刻音频信号的所有信息，所以直接由 $x(n)$ 提取音频特征，而不需要对 $x(n)$ 做任何进一步的处理。提取的特征一般包括：短时平均能量、过零率、线性预测系数等。

由于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，人耳之所以能感受到音频信号，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能量信号的结果。因此，基于频域的特征提取是必要的。傅立叶分析在音频信号的频域分析中起到很重要的作用，因此一般采用傅立叶变换来获取音频信号的频域特征。也可以应用短时傅立叶变换来获取音频信号的频域特征。在傅立叶变换域中提取的特征主要包括：熵、能量比、质心、带宽、静音比等。

3.3 基于不同时间长度的音频特征提取

一般而言，音频流特征提取可以基于两种不同的时间长度：音频帧(audio frame)和音频例子（audio clip）。使用音频帧长度来提取特征的思想来自语音信号处理理论，其前提假设是语音信号在短时刻内（如几十毫秒）是稳定的，因此在稳定短时刻内提取的特征被发现十分适宜。基于音频例

子长度提取特征考虑的是任何音频语义总是要持续一定长的时刻，如爆炸声和掌声等会持续几秒。如果在音频语义持续时间内提取特征会更好反映音频所蕴涵语义。

3.3.1 基于帧（frame）的音频特征

帧是我们处理音频信号的最小单位，计算出每一帧的特征值，然后在此基础上计算出片段层次上的特征值。通常在帧的层次上有以下几种典型的特征：

MFCC（Mel-frequency cepstral coefficients）系数，MFCC 是建立在傅立叶和倒谱分析的基础上的：对短时音频帧中的$[K/M]$个采样点进行傅立叶变换，得到这个短时音频帧在每个频率上的能量大小。如果音频信号的采样频率为 25kHZ，那么由采样定理知，音频帧的最大频率为 12.5kHz。也就是说，短时音频帧在 0 到 12.5kHz 的频率上具有能量，只是每个时刻在不同频率上所带能量大小不同而已。利用人耳的感知特性，把 0kHz-12.5kHz 的频率带划分成若干个子带。在整个频率带划分为频率子带时，可以采取线性划分和非线性划分两种方式。如果要将整个频率带线性划分成若干个子带，每个子带的宽度可以取为：$Mel(f)=2595\log_{10}(1+\frac{f}{500})$；非线性划分中每个频率子带的划分就比较复杂了。无论是线性划分还是非线性划分，如果整个频率带被划分成n个子带，分别计算这n个子带上的总能量，就构成了这个短时音频帧的n个 MFCC 系数。如果对提取出来的 Mel 系数再计算其对应的倒谱系数，就是 Mel 倒谱系数。在 Mel 标度频率域提出来的倒谱参数，非常符合人耳的听觉特性，它被广泛应用于语音识别和说话者识别中，并且实验结果证明它比其它特征如 LPC 性能更好。

基于音频帧地特征还包括基音频率、短时帧频域能量、过零率、频率中心（质心）、带宽等。

3.3.2 基于片段（clip）的音频特征

根据上面计算的帧层次上的基本特征，可以计算片段层次上的特征。

带宽均值和方差，片段中各帧的带宽均值和方差可以作为整个音频片段的均值和方差。

过零率均值和方差，片段中各帧的过零率的均值和方差作为整个音频

片段的均值和方差。

静音比例，如果一帧的频域能量小于阈值，则认为该帧为静音帧，否则为非静音帧，定义为：静音比$=\frac{\text{片段中静音帧的数目}}{\text{片段中帧总数}}$。

子带能量，将频域划分成不同的子带，计算各子带的能量如下：$D=\frac{1}{E}\int_{L_j}^{H_j}|F(\omega)|^2\,d\omega$，$H_j,L_j$为子带的上下边界频率。

子带能量比，定义为各子带能量与频域总能量的比值。

零过零率比：如果某一帧的过零率为零，则认为该帧为零过零率帧，音频片段中具有零过零率的帧数与整个片段中所有帧数的比即是零过零率比。

在文献[12]中还定义了谱通量（SF），高过零率比（HZCR），低短时能量值比率（LSTER）等基于片段的音频特征。

3.4 小结

有效的特征是音频分类和检索的基础。本章介绍了基于不同变换域的音频特征提取，重点是基于小波变换域的音频特征提取，并简单提了一下基于时域和傅立叶变换频域的音频特征提取。基于小波变换提取的特征包括质心、带宽、过零率、静音比、基音频率以及由过零率得到的新的特征零过零率和小波近似子带过零率周期等，这些音频特征能够有效的区分不同种类的音频，并且特征提取的计算比较简单，所花费的时间较少。另外还介绍了基于不同时间长度的音频特征的提取，包括基于毫秒级的短时音频帧的特征和基于秒级的音频片段特征。短时音频帧特征反映音频信号的短时平稳性，音频片段特征则反映音频信号的整体特性，这些特征是下面分类的基础。

第四章 基于 HMM-SVM 的音频分类算法研究

4.1 隐马尔可夫模型（HMM）的基本理论

隐马尔可夫模型起源于 60 年代后期，属于信号统计理论模型，能够很好地处理随机时序数据识别与预测，在语音处理的各个领域中得到广泛的应用。

Markov 链是 Markov 随机过程的特殊情况，即 Markov 链是状态和时间参数都是离散的 Markov 过程。Markov 链的定义为：

随机序列 X_n，在任一时刻 n，它可以处在状态 $S_1,S_2,\cdots,S_N$，N 为状态数目，且它在 $m+k$ 时刻所处的状态 q_{m+k} 的概率只与它在 m 时刻的状态 q_m 有关，而与 m 时刻以前它所处的状态无关，该性质称为随机过程的“马尔可夫性”，即有：$P(X_{m+k}=q_{m+k}\mid X_m=q_m,X_{m-1}=q_{m-1},\cdots,X_1=q_1)=P(X_{m+k}=q_{m+k}\mid X_m=q_m)$，其中 $q_1,q_2,\cdots,q_{m+k}\in(S_1,S_2,\cdots,S_N)$，则称 X_n 为 Markov 链，$P_{ij}(m,m+k)=P(q_{m+k}=S_j\mid q_m=S_i)$，（$1\le i,j\le N$）为 K 步转移概率，并且满足下面所规定的约束条件：

$$P_{ij}\ge 0,\quad \forall i,j \tag{4.1}$$

$$\sum_{j=1}^{n}P_{ij}=1,\ \forall i \tag{4.2}$$

这种随机过程又叫做可观测马尔可夫过程。隐马尔可夫模型（HMM）是在 Markov 链的基础之上发展起来的。在一些实际问题中存在着比 Markov 链更为复杂的模型，其中观察到的事件并不是与状态一一对应，而是通过一组概率分布相联系，这样的模型就称为 HMM。HMM 本质上是一种双重随机过程的有限状态自动机，其中的双重随机过程之一是满足 Markov 分布的状态转换 Markov 链，这是基本的随机过程，它描述状态的转移；另一个随机过程描述状态和观察值之间的统计关系。由于不像 Markov 链模型中的观察值和状态是一一对应的，观察者只能看到观察值而不能直接看到状态，所以只能通过一个随机过程去感知状态的存在及其特性。因此称为“隐”马尔可夫模型（HMM）。

一个 HMM 可以由下列参数描述：

状态总数 N，这表示一个 HMM 所包含的状态总数，即状态的集合为 $S=\{S_1,S_2,\cdots,S_N\}$。在 HMM 中虽然每个状态被“隐藏”起来了，但是在很多实际应用中，状态是有明确意义的。如在音频分类识别中，隐马尔可夫模型中的每个状态对应于一种音频的类型，如果状态发生了改变，则意味着音频的类型也发生了改变。在隐马尔可夫模型中，状态和状态之间通过转移概率被连接起来，状态与状态的之间的连接方式有“自遍历”连接，即每个状态都可以直接到达模型中的任何状态（包括自身），还有一种常见的连接方式是“从左向右”，即每个状态只能达到自身和它相邻的下一个状态。我们可以用 q_t（ $q_t \in (S_1,S_2,\cdots,S_N)$）表示 t 时刻所处的状态。

每个状态对应的观测事件数 M，这是每个状态所可能发生的观测事件的总数。记 M 个观察值为 $V_1,\cdots,V_M$，则 t 时刻观察到的观察值为 O_t（$O_t \in V_1,\cdots,V_M$）。在音频分类与识别中，观测事件是从每个短时音频帧中提取的特征，其特征数目就是观测事件的维数。

状态间的转移概率矩阵 $A=\{a_{i,j}\}$，其中 $a_{i,j}=P[q_{t+1}=S_j \mid q_t=S_i]$，$1\le i,j\le N$，注意 $a_{i,j}$ 要满足前面 4.1-4.2 式的约束条件。

观察事件对应状态的概率分布，每一状态都有一个相关的概率输出函数，用于估计观测值在该状态下的输出概率，$b_j(O_k)=P[O_k \mid q_t=S_j]$ 表示时刻为 t 时，状态为 j 时观测值为 O_k 的输出概率的大小，其中 $O=O_1O_2\cdots O_T$ 是完整的观测序列。

起始状态概率 $\pi=\{\pi_i\}$，它规定对于一个观测序列 $O=O_1O_2\cdots O_T$，起始时刻 O_1 位于某个状态的概率。由于起始时刻 O_1 可以位于任何一个状态，所以需要知道每个状态为起始状态的概率，用 $\pi_i=P[q_t=S_i]$（$1\le i\le N$）定义每个状态为起始状态的概率。在音频分类与识别中；初始状态可以为 N 个状态中的任何一个。

通常用五元组 $\lambda=(N,M,A,B,\pi)$ 来表示一个隐马尔可夫模型 λ，或者简写为 $\lambda=(A,B,\pi)$。更形象地说，HMM 可分为两部分，一个是 Morkov 链，由 π,A 描述，产生的输出为状态序列，另一个是一个随机过程，由 B 描述，产生的输出为观察值序列，如图 4.1 所示，T 为观察值的时间长度。

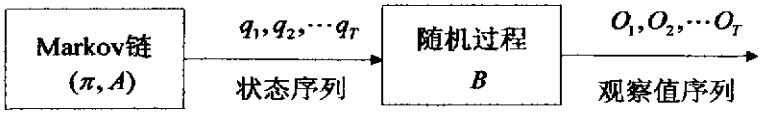


图 4.1　HMM 组成示意图

HMM 需要研究的三个基本问题是：(1)已知 HMM λ 的各参数，求某一观察序列 O 在该模型下的极大似然，即 $p(O|\lambda), O=O_1,\cdots,O_T$，$T$ 为观察序列长度；(2)在给定的 HMM 模型 λ 的条件下，求观察序列 O 最有可能历经的状态序列 S；(3)已知样本集合的条件下，如何根据样本集合训练模型并获得模型参数。问题(1)可以由前向（Forward）或者后向（Backward）算法解决，问题(2)是典型的状态空间搜索问题，经典的算法有基于动态规划的 Viterbi 算法，Beam Search 和 A*算法等，问题(3)是传统统计学习过程，其学习算法有 Baum-Welch 算法和梯度下降算法等。Baum-Welch 算法能够在理论上证明经过有限次迭代算法就能收敛，但它和梯度下降算法同样都会陷入局部极值点，而不能得到全局最优的结果。

4.1.1 概率估计问题

对于一个观察序列 $O=O_1O_2\cdots O_T$ 和模型 λ，如何快速计算出 $P(O|\lambda)$？一般采用的方法有前向算法和后向算法，下面分别介绍前向和后向算法。

在前向算法中，定义函数 $\alpha_t(i)=P(O_1O_2\cdots O_t, q_t=i|\lambda)$，表示在模型 λ 中，当时刻 t 的状态 q 为 i 时，观察序列 $O_1O_2\cdots O_t$ 的条件概率。可以用如下的方法递归求解 $\alpha_t(i)$：

递归初始化：$\alpha_1(i)=\pi_i b_i(O_1)$，$(1\le i\le N)$；

递归：$\alpha_{t+1}(j)=[\sum_{i=1}^{N}\alpha_t(i)\alpha_{ij}]b_j(O_{t+1})$，$(1\le t\le T-1, 1\le j\le N)$；

递归中止：$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i)$；

第一步递归初始化计算的是初始状态为 i，并且状态 i 对应的第一个观察事件 O_1 的概率是 $\pi_i b_i(O_1)$。

第二步递归计算是前向算法的核心。$\alpha_t(i)\alpha_{ij}$ 表示时刻 t 前面出现的观察事件序列是 $O_1O_2\cdots O_T$，并且由时刻 t 时的状态 i 转移到时刻 $t+1$ 时的状态 j 的过程中的概率值。由于从时刻 t 有 N 种可能从状态 $S_i(1\le i\le N)$ 到达状态 j，把这些情况累加起来，再乘以 $t+1$ 时候状态 j 对于观察事件 O_{t+1} 的概率 $b_j(O_{t+1})$，就是 $\alpha_{t+1}(j)$ 的值。

最后一步是递归终止结束，并且计算时刻 T 能到达所有状态的前向概率值：$\alpha_T(i)=P(O_1O_2\cdots O_T,q_T=i|\lambda)$，$(1\le i\le N)$。该算法的时间复杂度为 $O(N^2T)$。

后向算法与前向算法类似，首先定义后向函数 $\beta_t(i)=P(O_{t+1}O_{t+2}\cdots O_T,q_t=i|\lambda)$，表示在模型 λ 中，当时刻 t 的状态 q 为 i 时，观察序列 $O_{t+1}O_{t+2}\cdots O_T$ 的条件概率。可以用如下的方法递归求解 $\beta_t(i)$：

递归初始化：$\beta_T(i)=1$，$(1\le i\le N)$；

递归：$\beta_t(i)=\sum_{i=1}^{N}\alpha_{ij}b_j(O_{t+1})\beta_{t+1}(j)$，$t=T-1,T-2,\cdots,1$；$(1\le j\le N)$；

递归中止：$P(O|\lambda)=\sum_{i=1}^{N}\beta_t(i)$；

在初始化时，对于任意的 i $(1\le i\le N)$，给 $\beta_T(i)$ 赋值为 1。在递归步骤中，考虑时刻 t，当前状态为 i，$\alpha_{ij}b_j(O_{t+1})$ 计算的是从时刻 $t+1$ 到时刻 t，状态由 j 到 i 的转移概率乘以 $t+1$ 时观测 O_{t+1} 属于状态 j 的概率。把所有可能的状态转移序列累加起来，就是 $\beta_t(i)$ 的值。后向算法的计算复杂度接近于 $O(N^2T)$。

4.1.2 Viterbi 算法

为了寻找观测事件序列 $O=O_1O_2\cdots O_T$ 所对应的最佳状态序列 $q=q_1q_2\cdots q_T$，定义函数：$\delta_t(i)=\max_{q_1,q_2\cdots q_{t-1}}P[q_1q_2\cdots q_{t-1},q_t=i,O_1O_2\cdots O_t|\lambda]$，也就是找到一个状态序列，这个状态序列在 t 时状态为 i，并且状态 i 与前面 $t-1$ 个状态构成的状态序列的概

率值最大。在具体实现时，利用一个数据结构$\psi_t(j)$来记录每时刻最佳状态，这样根据动态规划方法，递推公式为：

初始化：$\delta_1(i)=\pi_i b_i(O_1)$，$\psi_1(i)=0$，$(1\le i\le N)$；

递推：$\delta_t(j)=\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$，

$\psi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]$，$(2\le t\le T,1\le j\le N)$；

终止：$P^*=\max_{1\le i\le N}[\delta_T(i)]$，$q_T^*=\arg\max_{1\le i\le N}[\delta_T(i)]$；

最佳状态序列：$q_t^*=\psi_{t+1}(q_{t+1}^*)$，$t=T-1,T-2,\cdots,1$.

4.1.3 参数训练

参数训练的意思是如何找到一个方法使隐马尔可夫模型$\lambda=(N,M,A,B,\pi)$中的参数得到优化，能够模拟实际生活中的随机过程。

对于音频例子的分类和识别来说，所谓的参数优化是指如果要训练一个隐马尔可夫模型λ去识别所有的音频，那么需要先收集一部分音频例子作为训练样本，训练一个隐马尔可夫模型λ，则训练成的隐马尔可夫模型对训练样本取得比较高的识别率，这样训练得到的隐马尔可夫模型中的参数才是“优化的”。

需要指出的是，所谓的“最优化”方法是不存在的，不可能找到“最优化”的参数，只能找到“次优化”或“优化”的参数；在 HMM 中，认定了如下的假设：如果训练成功的识别分类器，对训练样本的识别性能良好，则对实际未知数据的识别性能也良好，也就是使用了训练风险去代替实际风险；HMM 对时序信号的识别具有很强的优势。

实际中，模型参数中的N和M是由用户根据需要指定的，需要训练的是其他三个参数：状态之间的转移概率矩阵A，状态与观察事件之间的概率矩阵B和初始状态概率矩阵π。

定义一个变量$\xi_t(i,j)$，表示t时状态为i以及$t+1$时状态为j的概率：

$$\xi_t(i,j)=P(q_t=i,q_{t+1}=j\mid O,\lambda). \tag{4.3}$$

并且定义$\gamma_t(i)$表示给定模型和观测事件序列的情况下，在时刻t时状态为i的

概率，所以有 $\gamma_t(i)=\sum_{j=1}^{N}\xi_t(i,j)$。如果对 $\gamma_t(i)$ 中的 t 求和，那么 $\sum_{t=1}^{T-1}\gamma_t(i)$ 表示在时刻 T 内发生了多少次状态 i 到其他状态的转移。如果对 $\xi_t(i,j)$ 中的 t 求和，那么 $\sum_{t=1}^{T-1}\xi_t(i,j)$ 表示在时刻 T 内发生了多少次状态 i 到状态 j 的转移。这样模型中的三个变量可以进行如下训练：

$\tilde{\pi}_i=\gamma_1(i)$，表示时刻 1 经过状态 i 次数；

$\tilde{a}_{ij}=\dfrac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$，表示在时刻 T 内，状态 i 转移到状态 j 的总次数，除以在时刻 T 内状态 i 被经过的总次数；

$\tilde{b}_j(k)=\dfrac{\sum_{t=1,O_t=v_k}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$，表示在时刻 T 内，经过状态 j 并且状态 j 对应的观测事件为 v_k 的总数除以时刻 T 内经过状态 j 的总数。

在指定状态数目 N 和每个状态的观测事件数目 M 后，按照上面的方法对其他参数训练，得到 λ，然后按照 EM 期望最大值算法估计 $P(O|\lambda)$，为了判断得到的参数是否训练好，需要继续训练，得到新的 $\tilde{\lambda}$，然后再计算 $P(O|\tilde{\lambda})$，如果 $P(O|\lambda)>P(O|\tilde{\lambda})$，也就是说继续训练得到的参数 $\tilde{\lambda}$ 比前次得到的参数 λ 要“优化”些，训练过程一定会最终收敛。实际中需要设定一个阈值 ξ，如果 $P(O|\lambda)-P(O|\tilde{\lambda})|<\xi$，则意味着训练下去意义不大，训练终止，将 $\tilde{\lambda}$ 作为训练得到的这类音频的隐马尔可夫模板。

在训练某类音频模板时，如果训练样本为 nTraining，从每个训练样本中提取特征向量为 $O_i(1\le i\le nTraining)$。每个训练样本所包含的短时帧的数目可能不一样多，但是从每个音频帧中提取的特征数目一定要一样。训练开始前要分别设定隐马尔可夫链的状态总数和每个状态所含有的高斯概率分布总数，训练得到代表

这类音频的隐马尔可夫模型中概率矩阵 A,B,π 等参数，用这个隐马尔可夫模板来代表这类音频。在识别时，对于未知数据，也应该从每个短时帧中提取与训练时同样数目的音频特征，计算每个隐马尔可夫语义模板 λ_i 对观察序列（即特征向量）的似然概率 $P_i(X \mid \lambda_i)$。令 $j = \arg\max\{P_i, 1 \le i \le k\}$ 则这个音频被识别为隐马尔可夫链 j 所代表的语义。

HMM 是一个进行时序数据识别模拟的分类器。HMM 实现音频识别的本质是从每类音频的观测数据出发，寻找它们的内在规律，利用这些规律来构造模拟识别它们的模型。而寻找规律的过程是利用样本数据，不断学习训练，使 HMM 链中的参数对样本产生的概率值最大，得到“次优化”训练模板，作为每类音频例子的分类器。

4.1.4 HMM 的局限性

HMM 是基于这样的假设的：如果训练好的模型对样本数据取得了良好的识别率，则就假定训练好的模板对实际未知的数据也能取得良好的识别率，但是有时这样的假设不成立。

在训练中，为了让 HMM 对训练样本取得较好的识别率，会对这个模型使用 Baum-Welch 最大预期算法进行反复训练，这样使模型变得很复杂。这样将导致这个复杂模型对训练样本的分类能力很好，对实际未知数据的识别率反而下降，称这种情况为过学习（overfit）问题。

在训练 HMM 时，HMM 的结构是需要自己设定的，如 HMM 状态的数目、每个状态所对应的高斯分布数目，以及状态之间的结构。但是并不知道怎样一个 HMM 结构才能达到最优结果，只能凭经验。

在训练 HMM 过程中，选择多少样本最合适呢？并非是样本越多越好，而是“好样本”越多越好。但是没有一个合适的方法去判断哪些是“好样本”，哪些是“坏样本”，只能凭自己的主观感受决定。

使用 HMM 进行音频分类与识别时需要克服上面的局限性，下面就介绍另外一种机器学习理论——支持向量机（Support Vector Machine）。

4.2 支持向量机（SVM）的基本理论

支持向量机是 Vapnik 等[15]人提出的以结构风险最小化原理为基础的一种新

的分类方法。与 HMM 相比，SVM 具有以下的优势：

(1) SVM 在训练过程中考虑了分类模型自动构造，不需要用户指定，不像在 HMM 训练中，识别模型部分参数和模型拓扑结构需要自己根据实际情况指定。

(2) SVM 的训练可以在小样本前提下完成，训练样本不需要很多，而 HMM 训练中，样本多自然会提高识别精度。SVM 是专门针对有限样本情况的，其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值。

(3) 由于采用了结构风险最小化原理[26]，对 SVM 进行训练的目的是要使其识别的实际风险小，而在 HMM 训练中，只是保证训练得到模型的经验风险小，然后由模型的经验风险小去假定其实际风险也小（这种假设有时候不成立，如产生过学习现象）。

(4) 算法最终将转化为一个二次型寻优问题，从理论上说，得到的将是全局最优点，解决了 HMM 的局部极值问题。

(5) 算法将实际问题通过非线性变换转换到高维的特征空间，通过在高维空间中构造线性判别函数来实现原空间中的非线性判别函数，保证机器有较好的推广能力，同时它巧妙解决了维数问题，其算法复杂度与维数无关。

先标识训练样本如下：$\{x_i, y_i\}, i = 1, \cdots, l, y_i \in \{-1, 1\}, x_i \in R^N$，支持向量机的原理是用分类超平面将空间中两类样本点正确分离，并取得最大边缘。所有在这个超平面上的点 x 满足 $< w, x > + b = 0$，w 是超平面的法向量，那么寻找最优平面的问题为最小化：$\phi(w) = \frac{1}{2} \| w \|^2$，使其满足

$$y_i(< w_i, x_i > + b) \geq 1 \tag{4.4}$$

这里 $\phi(w)$ 是 w 的凸函数，于是上面的问题转换为约束条件下最优化求解问题，可以用拉格朗日方法求解。引入拉格朗日乘子 α_i，则有拉格朗日函数方程：

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j < x_i, x_j > \tag{4.5}$$

约束条件是：$\sum_{i=1}^{l} \alpha_i y_i = 0$，$\alpha_i \geq 0$。

有的时候，训练样本直接使用上面的方法找不到一个超平面实现样本点的分离，根据 Cover 定理[27]：一个复杂的模式识别分类问题，在高维空间比低维空间更容易线性可分，可以通过一个核函数把低维空间中的样本数据映射到高维空间中去。如果存在核函数 K 使 $K(x_i, x_j) = \Phi(x_i) \bullet \Phi(x_j)$，那么在训练中只需要考虑核

函数 K，不必明确知道映射 Φ 是什么。

这样，可以按照前面描述的方法构造最佳分类面，只不过支持向量不是直接来自于输入样本，而是映射后的高维特征空间。则在高维特征空间中所构造的判别函数为：

$$f(x)=sign(\sum_{i=1}^{N_s}\alpha_i y_i \Phi(s_i)\bullet\Phi(x)+b)=sign(\sum_{j=1}^{N_s}\alpha_i y_i K(s_i,x)+b) \tag{4.6}$$

其中，N_s 是高维特征空间的维数，s_i 是支持向量。

如果要寻找最佳的分类面，核函数和核参数的选择是非常重要的，使用最多的几个核函数包括(1)多项式核函数：$K(x,y)=(1+xy)^p$，参数 p 表示多项式的度；(2)高斯核函数：$K(x,y)=e^{-|x-y|^2/2\sigma^2}$，参数 σ 表示高斯函数的方差；(3)Sigmoida 核函数：$K(x,y)=\tanh(x\bullet y-\delta)$。许多分类问题证明高斯核函数的分类效果比多项式核函数和 Sigmoidal 核函数的分类效果要好[28,29,30,31]。Guo 和 Li[32]等人提出了 ERBF 函数作为核函数，并且证明了该核函数能够改善分类效果。ERBF 核函数的数学描述为：$K(x,y)=e^{-|x-y|/2\sigma^2}$。

标准支持向量机并不直接生成先验概率，而是输出的距离。可以应用先验概率对音频进行分类：对要识别的某类音频例子，采集样本，提取特征，训练识别这类音频例子的支持向量机 SVM_c $(1\le c\le N)$，用 SVM_c 代表这类音频的模板，然后用核函数通过下面的公式求得未知音频 x 相对于每一类音频模板 SVM_c 的概率

$$P(SVM_c\mid x)=\frac{1}{1+\exp(fx+B)} \tag{4.7}$$

其中 B 是核函数参数，f 是 x 相对于每一类音频模板 SVM_c 的输出；最后把未知的音频例子 x 归类到最大概率值所对应的支持向量机模板。

但支持向量机也存在一些局限性：

从音频例子特征提取中，已经知道一般进行支持向量机训练时，使用的音频例子统计特征。可是，音频本质上是非平稳的，所以要在极小的时间区域提取音频特征才能反映音频信号变换特性。故此，支持向量机虽是一个良好的识别模型但是在时序信号模拟上存在不足。HMM 在处理时序信号方面则具有优势，因此可以把 HMM 与 SVM 结合起来，取长补短。

每个支持向量机只能完成两类音频的识别问题，但是通常的音频都是对多个

类别进行分类和识别的，这就要构造多个支持向量机来完成多类分类识别问题。

如果一个分类问题 N 类可分，且这 N 类中的任何两类间一定可分，基于这个策略，可以按照下面两种方式使用支持向量机进行多类模式识别与分类：(1)一对多（one-to-all），该方法将多类分类问题抽象为两类分类，即使用算法找出其中一类与除此类之外的所有类的边界，对其中的每一类都如此。设类别数为 K，则需要训练 K 个支持向量机，使每个支持向量机能够区分类别 $Audio_i$ 与不属于 $Audio_i$ 的所有音频；(2)一对一（one-to-one）策略，该方法是对所有类别中的每类样本分别训练，为了识别 $Audio_i$ 等 K 类音频，需要训练 $K(K-1)/2$ 个支持向量机，使总有一个支持向量机能够把 $Audio_i$ 与 $Audio_j$ 区分开来。上面两种方法各有优劣，在很多情况下一对一的分类精度要优于一对多的方法，但是前者学习和分类过程中的计算量要明显高于后者，并且随着类别数的增加，运算量迅速上升。有相关文献对多类分类方法在不同实验数据分类中的性能作了比较以及在两类策略的基础上提出的各种改进，如增广两类分类法 AB（Augment Binary）[33]，将多类分类数据通过扩充维数投影到两类空间中进行分类[34]，能够得到较高的分类精度。

另外，如果已知任意两两可分，则通过一定的组合法则，由两两可分最终实现 N 类可分。基于此思想，可以将 SVM 与二叉树的基本思想结合起来构成多类分类器，实现音频的多类分类。常用的方法有：

(1) 自上而下二叉树的构造。初始状态包含所有的音频类型，然后用递归的方法逐步分离出每一类音频，直到所有的类别都被分离出来。例如，对于一个 4 类分类问题，可以按照自上而下的思想构造二叉树，若各类基本无先验知识，无法指导树叶节点的划分，则采用每次决策分出一类的二叉树结构，如类型 1 所示；若各类有部分先验知识，则可以采用完全二叉树结构，如类型 2 所示。

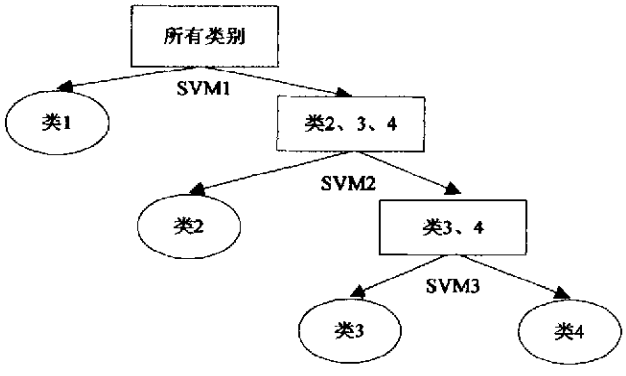


图 4.2 自上而下二叉树 1

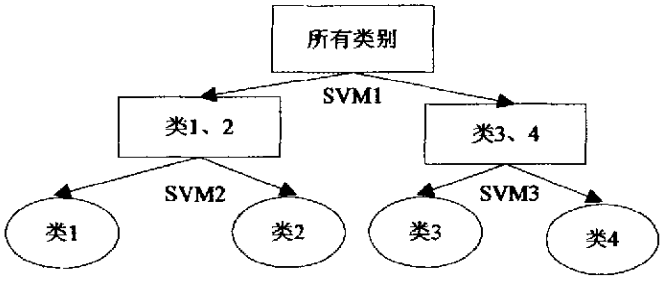


图 4.3 自上而下二叉树 2

(2) 自下而上二叉树的构造。每两类进行递归的比较。如果与测试类型之间的距离比较小，则进行进一步的比较，直到测试类型被分到指定的类中。仍就以四类分类为例，自下而上二叉树的构造过程如下图所示：

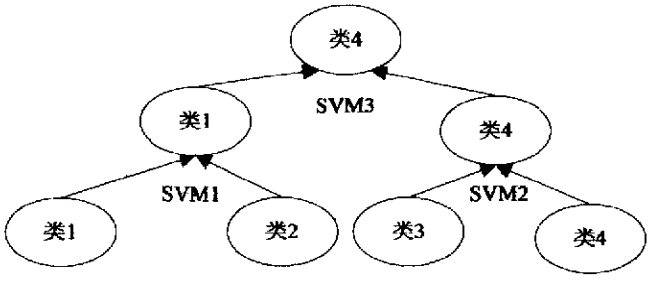


图 4.4 自下而上二叉树

表 4.1 则是对上面四种类型多类分类器的训练复杂度和总的测试时间复杂度的比较。

表 4.1 不同多类分类器的时间复杂度和训练复杂度比较表

复杂度 / 分类器类型	训练复杂度	测试时间复杂度
一对一	$\frac{c(c-1)}{2}T(n,n)$	$\frac{c(c-1)}{2}$
一对多	$cT(n,(c-1)n)$	$c-1$
自上而下二叉树	$\sum_{i=1}^{\log_2 C} 2^{i-1}T(\frac{cn}{2^i},\frac{cn}{2^i})$	$\log_2 c$
自下而上二叉树	$\frac{c(c-1)}{2}T(n,n)$	$c-1$

其中 c 表示类别总数，每一类包含 n 个数据，$T(n,n)$ 表示训练复杂度。

4.3 HMM-SVM 相结合的音频分类算法的研究

前面介绍了隐马尔可夫模型与支持向量机模型的基本理论，两种方法在音频分类与识别的应用中，各有优劣，如 HMM 适用于处理连续的非平稳的随机信号，但在学习过程中需要人为设定一些参数，并且会出现过学习的现象，从而导致实际风险比较大，另外 HMM 的区分能力比较差。SVM 适合于处理分类，具有较好的区分能力和泛化能力，SVM 是一个二类分类器，但音频的分类问题又属于多类分类问题，需要设计基于 SVM 的多类分类算法；再者，SVM 使用的多是音频的统计特征，这很难反映音频的非平稳特性；此外，SVM 也不适于处理大样本数据，在大样本数据集和特征维度比较高时，SVM 的计算复杂度很高。因此，人们很希望能够找到一种 HMM-SVM 的混合算法，克服 HMM 和 SVM 的缺陷，但又不失它们各自的优势，从而改善和提高分类性能。

HMM 适于处理连续信号，SVM 适于处理分类问题，同时，HMM 更多的表达了类别内部的相似性，而 SVM 则很大程度上反映了类别之间的差异，因而根据两者不同的侧重点，使其组合能够获得良好的效果[17,18,30]。文献[17]的基本思想是首先用基于 HMM 的分类器进行分类，接下来使用 SVM 解决那些用 HMM 不能分类的不确定的部分。其间用到了动态时间卷积核函数（DTWK）并把欧几里德距离应用到高斯核函数中：$K(X,Y)=\exp\{-\lambda D(X,Y)\}$。这种方法提高了手势语言的识别精度。文献[18]则综合应用 SVM 的判别力与 HMM 的短时性能，用 SVM 的概率输出代替 HMM 中的高斯混合输出。用小波变换方法提取观测向量，减少了数据的维数并且提高了鲁棒性。实验结果表明，SVM 与 HMM 的混合算法同样能够提高识别精度。文献[35]将支持向量机的输出通过 Sigmoid 函数和高斯模型转化为概率，并作为隐马尔可夫模型中各个隐状态的输出概率，将每个 HMM 的训练样本分成 N 个聚类，计算聚类中心，把得到的聚类中心用于 SVM 的训练，实验结果表明，HMM-SVM 的这种结合能够降低接受错误的概率。

本文用如下的方法将 HMM-SVM 结合起来构造音频分类器。首先提取短时音频帧特征作为 HMM 的观察序列，为每一类音频训练 HMM，并计算每个样本在对应的 HMM 输出概率，将样本的输出概率和音频片段特征一起作为 SVM 的输入，训练成每一类的 SVM。分类器性能的好坏采用下面的评价标准来衡量：各类 $Audio_i$ 的分类精度 C_i 和平均分类精度 C，分别定义如下：

$$C_i=\frac{\text{分类正确的}C_i\text{样本数}}{\text{预测为}C_i\text{的样本总数}}\times 100\%. \tag{4.8}$$

$$C=\frac{\text{分类正确的样本数}}{\text{预测样本的总数}}\times 100\%. \tag{4.9}$$

算法实现的流程图如下：

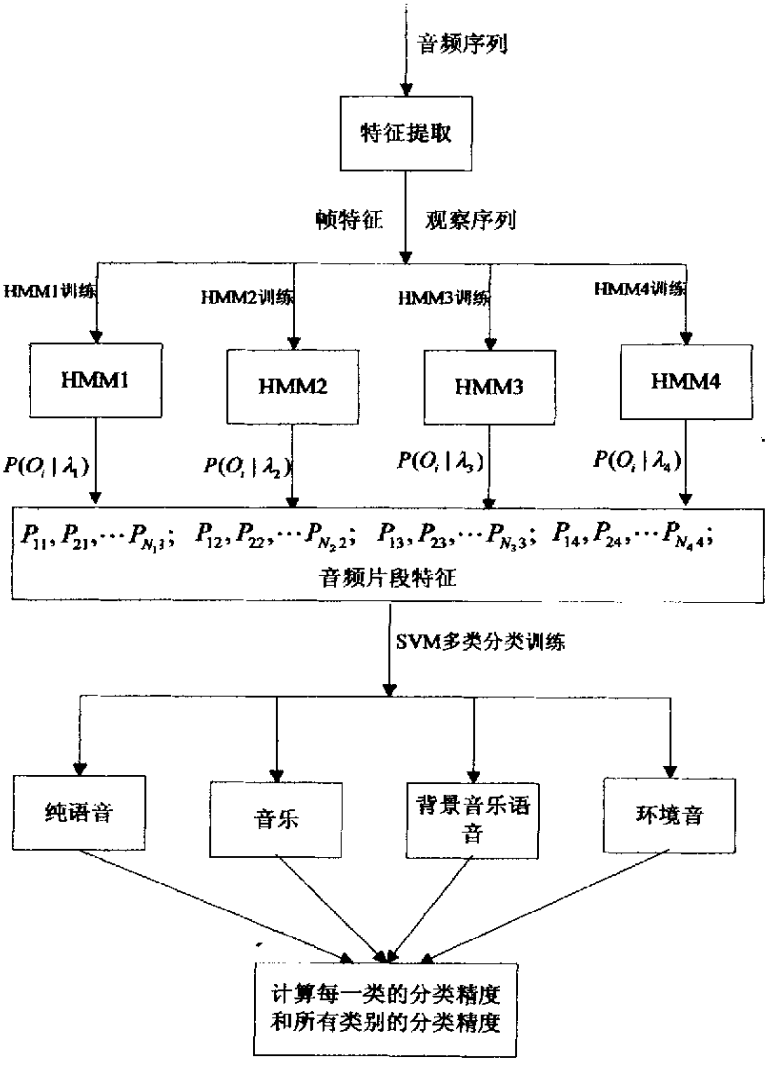


图 4.5 HMM 与 SVM 相结合的算法流程图

4.3.1 音频特征的提取与选择

在进行分类器的训练时，特征的选择是关键，因为分类算法再好，如果选择的特征不能够有效的区分不同种类的音频，那么就不能得到好的分类器，并且有可能会出现大的偏差，导致分类结果的误差较大。

音频特征选择有两种方法，一种是直接法，可以从每个特征值对不同音频类

型的概率统计曲线的相似性，看是否有明显的区分性，比如带宽特征是由质心特征经过运算得到的，从图 3.1 中可以看到带宽特征的区别更明显，所以在选择特征时就没有必要选择质心带宽两种特征，只需要选择带宽特征即可。另一种是间接法，即可以先使用某些特征构造 Baseline 系统，然后补充新特征，看分类正确率是否提高，比如可以先选择 MFCC 特征构造 Baselie 系统，这是由于 MFCC 特征能够反映听觉特性，很多文献的实验结果都说明该特征在音频分类和识别中能够取得较好的结果，然后再加入其他的特征，比如过零率、短时帧能量等特征，看是否能够更好的区分不同的音频种类。

在特征抽取的基础上构造音频分类的特征的集合，由于不同音频特征的值有很大的差别，所以要对特征集进行归一化处理：

$$x_i' = (x_i - \mu_i)/\sigma_i . \tag{4.10}$$

式中，μ_i为均值，σ_i为方差。由于 MFCC 归一化处理后维数值差别过小，试验结果不理想，所以对 MFCC 不作归一化处理。对每个音频片段的各帧计算 12 维 MFCC 系数，然后对各维取平均值，作为该片段的 MFCC 特征值。

4.3.2 HMM 训练与概率特征提取

在对每一类音频进行 HMM 训练时，充分考虑到信号的短时平稳性，利用 HMM 适于处理连续信号，表现内部特性的性质，需要选择音频帧特征。同时又考虑到计算复杂度和特征的区分效果，在进行 HMM 训练时，选择的特征是基于小波变换的带宽、基频、过零率以及反映听觉相似性的 MFCC 特征，这些特征都是基于短时音频帧提取的。

为每一类音频训练各自的 HMM。选择每一类音频的 50%作为训练样本，其他的作为测试样本。对每一个样本进行分帧，每一帧包含 256 个采样点，相邻帧之间有 80 个采样点的叠加。计算样本每一帧的带宽、基频、过零率以及 12 阶的 MFCC，组成 15 维特征向量作为 HMM 样本数据。在训练之前，首先用 K-Means 算法对提取的带宽、基频、过零率特征进行归一化处理，指定训练样本的数量，状态参数，以及观察序列的长度，HMM 的参数 A, B 和 π 的初始值用随机生成的方法生成。接下来基于高斯混合模型采用基于最大期望值的估计方法逐步优化隐马尔可夫模型的各个参数，包括状态转移矩阵 A，状态与观察事件间的概率矩阵 B 以及初始状态概率矩阵 π。训练好 HMM 以后，根据 $P' = \log(P(O|\lambda))$，计算每个

样本在 HMM 下的概率值。按照上述方法分别训练表示纯语音的 HMM_s，表示音乐的 HMM_m，表示带背景音乐的语音 $HMM_{s\&m}$ 以及表示环境音的 HMM_e。表 4.2 即表示 music 的部分训练样本在训练好的下的 HMM_m 概率值。

表 4.2　music 训练样本在 HMM_m 下的概率值

训练样本 i	1	2	3	4	5
概率 P'	-1994.39	-2055.56	-1909.97	-1884.25	-1950.36
训练样本 i	6	7	8	9	10
概率 P'	-1969.52	-1974.14	-1930.15	-1967.91	-2011.07
训练样本 i	11	12	13	14	15
概率 P'	-1946.6	-1952.84	-2116.46	-2045.61	-2232.94

同样的方法，我们可以求出每类训练样本 i 在训练好的 HMM_i 下的概率值，即可以得到每类音频样本在各自 HMM 下的概率曲线，如图 4.6 所示：

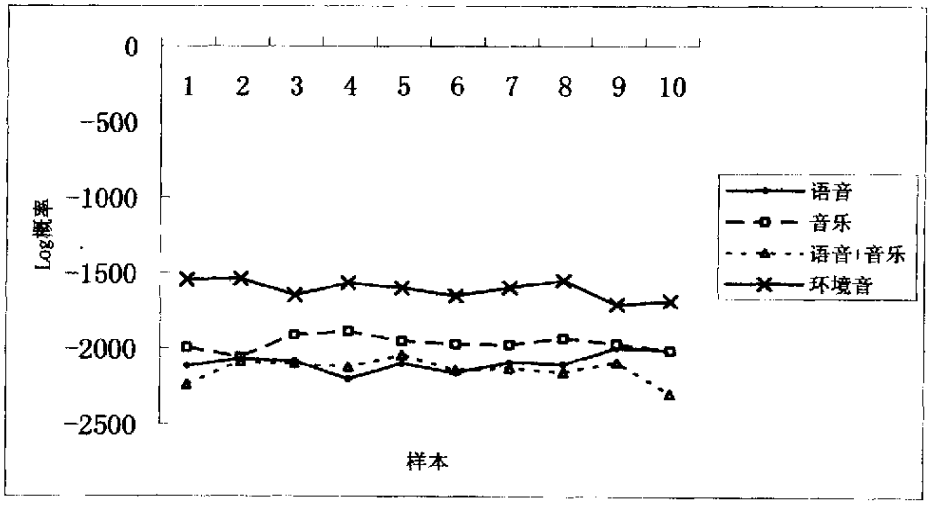


图 4.6 每类样本在各自 HMM 下的概率曲线

从图中可以看出，每个样本在各自 HMM 下的概率区别很大，特别是环境音，音乐与语音（包括纯语音和带背景音乐的语音）。但是对于纯语音和带背景音乐的语音，通过二者的概率分布不能把二者区分开来。但总的来说，概率特征对音频类别的区分还是很明显。因此可以把概率作为音频分类的一个新的特征，应用到

下面的支持向量机的分类中。

4.3.3 支持向量机（SVM）训练

(1) SVM 分类器训练

在进行 SVM 训练时，选用的训练样本和前面 HMM 的训练样本相同，所不同的是选用的提取特征不同。HMM 主要选择基于短时音频帧的特征像 MFCC、基于小波变换的带宽、基频、短时过零率等，而 SVM 则选取基于音频例子的统计特征值。选用的特征包括基于小波变换的带宽均值、方差、小波子带能量比、静音比、基音频率均值、零过零率比、小波子带过零率周期以及前面得到的 HMM 概率值等特征，把这些特征作为 SVM 的输入向量，训练支持向量机。对于一个四类分类的问题，按照前面的分析知宜采用自上而下的多类分类方法，因为该方法训练复杂度和测试时间复杂度总的来说比较小，训练复杂度是 $T(2n,2n)+2T(n,n)$，测试时间复杂度为 2。因此我们选用自上而下的方法构造二叉树。根据 HMM 得到的概率曲线图，可以采用 SVM 二叉树类型 1 进行构造多类分类器。首先经过支持向量机 SVM1 把区别其他种类比较明显的环境音区分开来，接着通过支持向量机 SVM2 把音乐区分出来，最后通过支持向量机 SVM3 把语音和带背景音乐的语音区分开来，从而完成对语音、音乐、带背景音乐的语音以及环境音四种不同类型音频的分类，如图 4.7 所示。选用的核函数是高斯核函数，参数 σ 选择为 0.00002。

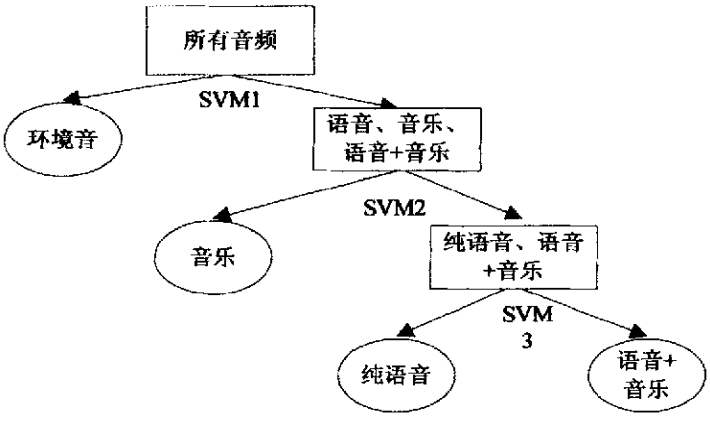


图 4.7　SVM 二叉树的构造

(2) 分类精度

判断一个分类器性能的好坏的指标是分类准确率。即事先知道训练样本所属

的类别，然后设计分类器，再用该分类器对测试样本进行识别，比较测试样本的实际所属类别与分类器输出的类别，进而统计正确识别率。正确识别率是反映分类器性能的主要指标。

把剩下的音频作为测试样本，每一类音频有 20 个测试样本，在训练好的支持向量机模型下进行类别的判别，最后可以得到如表 4.3 所示的分类结果，其中 1 表示纯语音，2 表示音乐，3 表示带背景音乐的语音，4 表示环境音。A 表示测试样本所属类别，$\overline{A}$ 表示 SVM 分类器输出类别。

表 4.3 测试样本 SVM 分类输出类别与实际类别比较

A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\overline{A}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	3	1
A	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
$\overline{A}$	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	3	2	2	2	2
A	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
$\overline{A}$	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	4	1
A	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
$\overline{A}$	4	4	4	4	4	4	4	2	4	4	4	4	1	4	4	4	4	4	1	4

按照(1)中的方法训练支持向量机，对测试样本按训练好的支持向量机进行识别，可以得到各类音频的分类精度如表 4.4 所示：

表 4.4　HMM 与 SVM 相结合得到的四类音频的分类精度表

	speech	music	speech music	environment
speech	0.90	0.00	0.10	0.00
music	0.05	0.90	0.05	0.00
Speech music	0.15	0.00	0.80	0.05
environment	0.10	0.05	0.00	0.85

根据平均分类精度的定义 $C=\frac{\text{分类正确的样本数}}{\text{预测样本的总数}}\times 100\%$，可以得到该方法的 SVM 分类精度为 0.8625。从表中可以看出，该方法对纯语音和音乐的分类精度较高，可达到 90%，对带背景音乐的语音和环境音的分类效果不是很理想。带背景音乐的语音误判为纯语音的概率比较高，达到了 15%，还有 5%的可能被误判为环境音。而环境音误判为语音的概率也比较高，环境音还有可能被误判为音乐。

如果单用支持向量机的方法对音频样本进行训练，采用相同的训练样本，训练用到的输入仍然是基于小波变换的音频统计特征，包括带宽均值、方差、小波子带能量比、静音比、基音频率均值、零过零率比等。可以得到如表 4.5 所示的每一类音频的分类精度，所有音频的分类精度是 0.7。

表 4.5 单用 SVM 得到的四类音频的分类精度表

	speech	music	speech music	environment
speech	0.75	0.05	0.20	0.00
music	0.00	0.80	0.10	0.10
Speech music	0.05	0.35	0.40	0.20
environment	0.00	0.05	0.05	0.90

表 4.4 和表 4.5 相比，纯语音、音乐、带背景音乐的语音三种音频类型的分类精度都有了不同程度的提高，尤其是带背景音乐的语音的分类精度有了大幅度的提高，由 40%上升到 80%，提高了整一倍。这就说明了文中所用的 HMM 与 SVM 相结合的分类算法比单独使用 SVM 的分类效果好。

基于小波变换的特征在提取时间方面也具有很大的优势。在提取短时音频帧特征时，特征提取的时间复杂度主要花费在每一帧特征的计算上，所以帧数的多少决定了提取时间的快慢。在时域和傅立叶变换频域的特征提取是针对整个音频进行的，对整个音频片段进行分帧处理，分成大约 10ms 的短时帧，这样对于 1s～2s 的音频例子，大约要分成几百帧，基于短时帧提取特征时，要对这几百帧的每一帧都进行特征提取运算，如果把对每一帧的特征提取看作一次处理的话，则要经过几百次处理才能完成对所有帧的特征提取。这样计算量就比较大，花费的时间也比较多。而基于小波变换的音频短时帧特征的提取则是首先经过一次多层小波分解，小波变换相当于一个滤波器的作用，滤去音频信号的高频部分，而保留音频信号的低频(近似)部分，然后对近似部分再进行分帧处理，这样就大大减少了短时帧的数目，降低了运算复杂度，节省了特征提取的时间。图 4.8 为每个音频例子基于傅立叶变换和小波变换特征提取时间的比较。同时从表 4.4 中可以看出，基于小波变换特征能够实现纯语音、音乐、带背景音乐的语音和环境音的分类，这说明应用小波变换得到的特征在音频分类中是有效的。

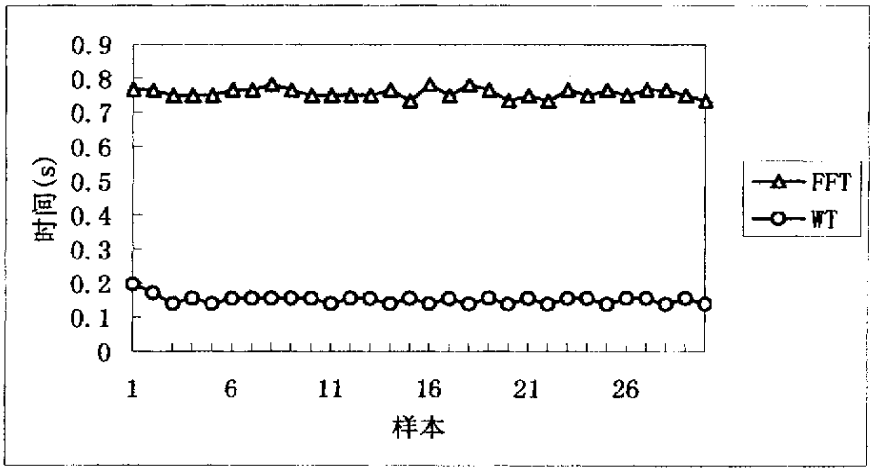


图 4.8 基于小波变换和傅立叶变换特征提取时间的比较

4.4 小结

本章介绍了隐马尔可夫模型（HMM）和支持向量机（SVM）两种常用的音频分类方法的基本原理和方法，同时指出两种方法在音频分类中的优势与局限性，提出一种两种分类方法相结合的分类方法。其中分类训练所用到的特征是基于小波变换提取的特征，充分利用了 HMM 的时间序列性与 SVM 的泛化性，得到的分类效果比在相同特征条件下单独使用 SVM 的好。

第五章 结论及未来的工作

5.1 结论

音频分类技术是音频检索的基础，在视频检索系统和其他多媒体应用系统中也有广泛的应用。目前对音频处理的研究主要集中在语音处理的研究，对广义的音频研究还比较少，对音频的分类研究则是近几年才开始的。基于内容的音频分类与检索本质上是一个模式识别问题，这个问题由两个方面的基本问题组成：选择什么样的特征最能够代表音频信号的内容和如何基于所选特征对音频信号进行分类。一个有效的特征要能够表示具有不同任务、不同环境的鲁棒的音频信号的大部分显著特征，要能够描述任意的音频类型。

小波分析属于时频分析的一种，它在时域和频域同时具有良好的局部化性质，具有多分辨分析的特点，在时频两域都具有表征信号局部特征的能力，广泛的应用于信号处理、图像处理、音乐、雷达、机械诊断及数字电视等领域。由于音频信号是短时平稳信号，所以基于小波变换域提取的特征能够更好的反映音频信号的局部特性。因此，本文主要分析了基于小波变换域提取的音频特征，包括质心、带宽、过零率、静音比以及子带能量等特征。

本文对常用的音频特征基于两种不同的规则进行了归类，首先是基于不同变换域提取的特征描述，主要介绍了基于小波变换域的音频特征的提取以及在小波变换域中各特征的数学描述，并简单的介绍了基于时域和傅立叶变换频域的音频特征的提取；接着是基于不同的时间长度对音频特征进行分析。时间长度分为毫秒级的基于短时音频帧的特征以及秒级的基于音频片段的特征。基于两种规则的不同的音频特征有相当部分的重合，在后面综合运用基于不同规则的特征，如在 HMM 训练中，主要用到基于小波变换域的短时音频帧特征质心、带宽等以及基于傅立叶变换域的 MFCC（短时音频帧）特征等。

实验结果表明，同基于傅立叶变换的特征的提取相比，基于小波变换域提取的特征能够更好的表征不同种类的音频，并且特征提取阶段的计算复杂度较低，花费的时间较少。

音频分类的方法有很多，本文主要介绍了 HMM、SVM 的原理，并把 HMM 和 SVM 结合起来应用于音频分类领域，把音频分类为纯语音、音乐、带背景音乐的语音和环境音。既充分利用 SVM 强的泛化能力和分类能力，同时保留了 HMM 较强的时间序列建模能力。首先用基于小波变换的短时音频帧特征质心、带宽、过零率、基音频率和基于傅立叶变换的短时音频帧特征 MFCC 一起作为 HMM 训练的观察序列，分别训练四种音频类型的 HMM，充分利用了 HMM 具有较强时间建模能力的特点，然后根据每类各自的 HMM 计算训练样本在该 HMM 下的概率值，作为下面 SVM 训练的基础。在进行 SVM 训练时，用到的特征是基于音频片段的特征，包括带宽、基音频率均值，小波子带能量比，高过零率比，静音比等特征，并结合 HMM 训练得到的概率值一起作为 SVM 的输入，训练 SVM 分类器。实验结果表明这种方法得到的分类器对纯语音和音乐的分类精度较高，达到 90%，对环境音的分类精度为 85%，对带背景音乐的语音的分类精度比较低，只有 80%。但是比单独使用 SVM 方法用相同的特征进行训练得到的分类精度要提高 16 个百分点。

5.2 未来的工作

音频分类是音频检索的基础，本文主要介绍了基于小波变换的音频特征的提取，并利用提取得到的特征，采用 HMM 与 SVM 相结合的方法把音频分为纯语音、音乐、带背景音乐的语音和环境音四种类型，分类精度为 86.5%。分类精度同只用 HMM 和 SVM 相比有了大幅度的提高，但同时理论分析和实验结果也显示出了一些不足和问题，有待进一步改善。今后需要进一步研究的问题有：

(1) 新特征的提取与分析

本文主要介绍的特征是在音频和语音中常用的特征，所不同的只是在小波域中，用基于小波变换得到的小波系数重新表示，如质心、带宽、过零率、小波子带能量比等；或是用小波变换的方法重新提取一遍，如基音频率。基本上没有涉及到新的有效的特征提取，因此需要进一步的研究不同种类的音频信号，并充分考虑基于小波变换的音频信号的细节信息，提取一些反映这些细节的有效的特征，应用到下面的分类和检索的研究中去。

(2) 压缩域上的音频特征的提取与音频分类的研究

本文分类研究所用到的音频数据都统一是 wav 格式，也就是音频的解

压缩格式。目前我们所接触到的大部分音频数据都是以 mp3、avi、rm、wma 等压缩格式存储的。如果只是对解压缩的音频进行处理的话就需要对这些压缩格式的音频进行解压缩，解压过程需要大量的时间和占用大量的系统资源，并且在存储时也需要很大的存储空间，这对实际的应用是十分不利的。如何在压缩域上处理音频流，如何分析和提取压缩域上的音频特征，如何将小波变换应用到音频的压缩领域，这些都是未来的研究方向。

(3) 更有效的分类器的设计

本文把 HMM 与 SVM 结合起来训练分类器，虽然取得了比较好的分类效果，同时看出分类效果仍然不是很理想，尤其是对环境音和带背景音乐的语音的效果比较差。如何设计更有效的分类器，其中要考虑多方面的因素，主要有两方面，一是更加有效的音频特征的抽取，二是多种分类算法的融合，充分利用每一种分类算法的优势，克服其不足，从而使新的分类器能够提高分类的精度。

(4) 完整的检索系统的实现

音频分类的目的就是为了实现音频的检索。但是由于时间关系和本人知识能力的限制，目前没有做这一方面的研究工作，这就需要在以后的工作中更进一步，做出一个完善的音频检索系统，能够在实际的检索应用中发挥作用。

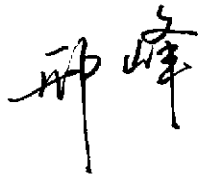
由于作者知识和能力的局限，很难对音频分类与检索的相关技术进行详细深入的讨论与研究，文中的错误与漏洞也难以避免，在此敬请原谅。

致 谢

衷心地感谢我的导师郑继明老师，本文的研究工作是在他的悉心指导下完成的。两年多来，导师为我提供了宽松与自由的研究环境，使我能够专心致志的进行本文的研究。导师严谨的治学态度，认真地做事风格，强烈地责任心，以及平易近人的工作作风，都使我终身受益。在论文完成之际，谨向我的导师致以崇高的敬意和诚挚的谢意。

感谢本文参考和引用的文献的作者们，他们开创性的劳动为本文的研究工作打下了坚实的基础。

向所有关心和支持我的亲戚、朋友和同学致以诚挚的谢意。



2007. 5. 30

攻硕期间从事的科研工作及取得的研究成果

1. 从事的主要科研工作

2005 年 7 月至今参加了重庆市教委项目“基于 MPEG7 标准和数据挖掘的视频检索与分类”，负责音频部分的分析与处理。

2. 发表的论文

邢峰,郑继明,吴渝,李婧. 基于小波变换的音频特征提取与分类[J]. 计算机科学.2006.33(11A):232-234

参考文献

[1] Feiten B., Frank R., Ungvary T. Organization of sounds with neural nets[C]. International Computer Music Conference, International Computer Music Association. San Francisco. 1991. 441-444

[2] Feiten B., Gunzel S. Automatic indexing of a sound database using self-organizing neural nets[J]. Computer Music Journal. 1994.18(3):53-65

[3] Hao Jiang, Tony Lin, Hongjiang Zhang. Video segmentation with the support of audio segmentation and classification[C]. Proceedings of ICME'2000-IEEE International Conference on Multimedia and Expo. New York, 2000.1507-1510

[4] Tong Zhang, C-C Jay Kuo. Heuristic approach for generic audio data segmentation and annotation[C]. Proceedings of the 7 th ACM International Conference on Multimedia. Orlando.1999.67-76

[5] Li,S.Z. Content-based classification and retrieval of audio using the nearest feature line method[J]. IEEE Transactions on Speech and Audio Processing. 2000.8(5):619-625

[6] 卢坚,陈毅松,孙正兴,张福严. 基于隐马尔可夫模型的音频自动分类[J]. 软件学报. 2002.13(8):1593-1597

[7] Chien-Chang Lin, Shi-Huang Chen, Trieu-Kien Truong and Yukon Chang. Audio Classification and Categorization Based on Wavelet and Support Vector Machine[J].IEEE Transactions on Speech and Audio Processing. 2005.13(5):644-651

[8] Wold,E., Blum, T.Keislar, D.,et al. Content-Based classification, search, and retrieval of audio[J]. IEEE Multimedia Magazine. 1996. 3(3):27-36

[9] L.Lu, H.Zhang, and S.Li. Content-based audio classification and segmentation by using support vector machines[J]. ACM Multimedia Systems Journal 8. March 2003. 8(6):482-492

[10] J.Foote. Content-base retrieval of music and audio. Multimedia Storage and Archiving Systems[J].Proc.of SPIE. 1997.32(29):138-147

[11] 李加顺.音频分段技术研究[D].国防科技大学.2001.20-28

[12] 李恒峰,李国辉. 基于内容的音频检索与分类[J].计算机工程与应用.

2007.(7):54-56

[13] Z Liu, J Huang, Y Wang, T Chen. Audio feature extraction and analysis for scene classification[C]. IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing. New Jersey,USA.1997.23-25

[14] 卢坚. 基于内容的音频检索技术[D]. 北京.国家图书馆.2001.31-35

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

[16] 吴飞,庄越挺,潘云鹤. 基于增量学习支持向量机的音频例子识别与检索[J].计算机研究与发展. 2003.1. 40(7): 950-955

[17] Jianjun Ye, Hongxun Yao, Feng Jiang. Based on HMM and SVM Multilayer Architecture Classifier for Chinese Sign Language Recognition with Large Vocabulary[D]. Third International Conference on Image and Graphics.377-381

[18] Liu Jiang-hua, CHEN Jia-pin, CHENG Jun-shi. Hybrid SVM/HMM Method for Face Recognition[J]. Journal of Donghua University. 21(1):34-38

[19] 叶福军.HMM 和 SVM 相结合的音频自动分类[J]. 高性能计算技术. 2005.2. 172:44-46

[20] A Grossmann, J Morlet. Decomposition of Hardy Function into Square Integrable Wavelets of Constant Shap[J]. SIAM J Math.Anal. 1984.15:723-726

[21] 大卫 马尔著. 刘磊,汪云九,姚国正译. 视觉计算理论[M]. 北京:科学出版社.1988.42-57

[21] Phung Quoc Dinh, Chitra Dorai, Svetha Venkatesh. Video Genre Categorization Using Audio Wavelet Coefficients[C]. The Fifth Asian Conference on Computer Vision. 2002.1-6

[22] Guohui Li, Ashfaq A.Khokhar. Content-based Indexing and Retrieval of Audio Data using Wavelet[C]. The IEEE international conference on multimedia and expro. New York. Aug.2000.885-888

[23] S.-H.Chen and J.-F.Wang. Noise-robust pitch detection method using wavelet transform with aliasing compensation[J]. Proc. Inst. Elect. Eng. Vision, Image Signal Process. Dec.2002. 149(6):327-334

[24] C.-T.Hsieh, E.Lai, Y.-C. Wang. Robust speech features based on wavelet transform with application to speaker identification[J]. Proc.Inst. Elect.Eng. Vision, Image Signal Process. Apr.2002.149(2):108-114

[25] Lu L., Zhang H.J., Jiang H. Content analysis for audio classification and segmentation[J]. IEEE Transaction on Speech and Audio Processing. 2002.10(7):504-516

[26] 庄越挺,潘云鹤,吴飞.网上多媒体分析与检索[M].北京:清华大学出版社.2002

[27] T M Cover. Geometrical and statistical properties of systems and linear inequalities with applications in pattern recognition[J]. IEEE Trans on Electronic Computers. 1965.19:326-334

[28] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines[J]. IEEE Trans. Geosci. Remote Sens. 2004. 42(8):1778–1790

[29] P. Clarkson and P. J. Moreno. On the use of support vector machines for phonetic classification[J]. IEEE Int. Conf. Acoustics, Speech, Signal Process. 1999.2:585–588

[30] A. Ganapathiraju, J. E. Hamaker and J. Picone. Applications of support vector machines to speech recognition[J]. IEEE Trans. Signal Process. 2004. 52(8):2348–2355

[31] F. Schwenke. Hierarchical support vector machines for multi-class pattern recognition[C]. IEEE Fourth Int. Conf. Knowledge-Based Intelligent Eng. Syst. Allied Technologies, 2000. 561–565

[32] G. Guo and S. Z. Li. Content-based audio classification and retrieval by support vector machines[J]. IEEE Trans. Neural Networks. 2003.14(1): 209–215

[33] Hsu C W, Lin C J.A. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks. 2002. 13(2):415-425

[34] 郭世杰. 基于支持向量机的多类分类问题的研究[D]. 上海师范大学.2005.16-21

[35] 忻栋,杨莹春,吴朝晖. 基于 SVM-HMM 混合模型的说话人确认[J]. 计算机辅助设计与图形学学报. 2002. 14(11):1080-1082

[36] B.Q.Huang，C.J.Du，Y.B.Zhang and M-T.Kechadi. A Hybrid HMM-SVM Method for Online Handwriting Symbol Recognition[C]. The 6th International Conference on Intelligent Systems Design and Applications. Shandong, China. 2006.887-891