摘 要

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。音频信号包括语音和非语音(Non-speech)两类信号。一直以来，音频信号的处理主要集中于语音识别、说话者识别等语音处理方面的研究。基于内容的音频信息识别技术的研究还不多。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，是基于内容的音频检索技术能否得以实用的关键所在。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。

本文提出一种基于均值 MFCC 的音频信号识别算法，对 MFCC 系数进运算行了深入的分析。均值 MFCC 系数作为音频特征，采用动态时间规整识别算法，经过大量实验证明，这种方法能有效地对单一音频信号进行识别。

本文的主要工作及研究成果如下：

1、研究了常见的音频数据处理技术，对音频信号进行分帧处理，既考虑音频信号的短时平稳特性，又考虑音频信号本质非平稳特性，同时音频信号具有连续性，不同音频信号根据采样率不同来进行分帧和帧间迭加处理。

2、研究了音频信号主要时、频特征；对部分时域、频域特征进行仿真，分析各种特征的应用情况。对音频信号的短时过零率、短时能量、MEL 倒谱系数等进行主要分析。提出均值 MFCC 系数作为音频特征的方法。

3、研究了音频信号的分层分割方法，基于短时能量和短时过零率对音频信号进行分类的方法，能够对语音、静音、谐音等进行分类。

4、研究了单一音频信号的识别方法。通过对 MFCC 系数进行分析，均值 MFCC 系数作为音频特征，采用动态时间规整识别算法，能够对单一音频进行识别，对已有数据源进行测试，有较高的识别率。

音频信号的处理作为项目的一部分，根据要求实现了对单一音频信号的识别，用 VC6.0 来实现。另外，进行音频信号仿真，对连续音频信号进行分层分割、音频信号的检索方面进行了较深入研究，为以后在这方面的研究奠定良好的基础。

关键词：音频短时帧，过零率，MFCC，动态时间规整

ABSTRACT

The research of audio retrieval based on content, as a newly arisen field, is still being researched and investigated at home and abroad. The audio frequency signal includes two types of signals , speech and non-speech. Always, the processing of the audio frequency signal is mainly concentrated on speech processing, such as speech recognition and talks identify etc. The research of audio retrieval based on content is still not much. How to get the structure information and content meaning , make audio frequency signal having the same semantic classes is the key of the research of audio retrieval based on content. Only there is breakthrough in audio frequency signal recognition based on physics characteristic, more and deeper research of sound index could be done.

In this paper, we present a way of audio recognition based on MFCC and analyze MEL coefficient deeply. Through experiments, it proved that using average MEL coefficients as characters, DTW arithmetic is an efficient arithmetic for recognizing single sound signal.

The major work and achievement of this paper are presented as follows: (1) we review the main methods of audio retrieval at home and abroad. Study the familiar audio frequency data processing technique and general methods of audio retrieval. (2)Study the character of audio signal. Analyze the zero-crossing rate and MFCC. A mean Mel coefficients is proposed, it can be used to recognized different audio signal. (3)Study the segmentation and recognition of audio frequency signal. Audio signal can be divided into segments based on zero-crossing rate. (4)A audio recognition arithmetic based on MFCC is proposed. Through this arithmetic, audio clip can be identified effectively.

The simulation and data analysis in PC with VC6.0 software platform are carried out. And more simulation of audio segmentation and retrieval is done with MATLAB software platform.

Keyword: audio frame, zero-crossing rate , MFCC, dynamic time warping

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，也不包含为获得电子科技大学或其它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。

签名：[signature] 日期：06年6月23日

关于论文使用授权的说明

本学位论文作者完全了解电子科技大学有关保留、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。本人授权电子科技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应遵守此规定）

签名：[signature] 导师签名：[signature]

日期：06年04月23日

第一章 绪论

1.1 课题来源和研究任务

本课题来源于绵阳九洲横向课题，课题要求利用红外、可见光和声音三种传感器的数据进行融合来识别野外环境下的运动车辆。本论文作为该项目的一部分，主要研究音频信号的识别。

本课题涉及的关键技术与研究任务包括：

1、可见光信号分割与特征提取

2、红外信号特征提取与目标跟踪

3、音频信号分割与识别

4、三种信号信息融合及识别

课题研究期间，我参加了以下工作：系统整体方案设计、红外、可见光、音频数据采集，音频信号分割与识别算法研究及 PC 仿真，以及各阶段音频部分报告总结等工作。

1.2 基于内容的音频检索的背景及其意义

大量的网络数据中包含多媒体数据，如视频、图像和音频，人们已经不再满足于通过一般的属性(如名字、年月、价格等)进行检索，对图像和视频，可以采用主色调、纹理等视觉特征来检索；同样，对于音频，需要通过听觉特征进行检索。AOD、VOD、音频解析等系统的实用与推广，都需要高效的音频信息检索技术。音频信息检索技术已经成为信息检索技术的研究重点之一。

自然界的声音极其广泛，如音乐声、风雨声、动物叫声、机器轰鸣声等等，要从数以千万计的音频数据中提取所需的信息，常规的基于文本检索方法是行不通的，这就需要新的技术。只有从广泛的音频数据中提取特征信息，才能对不同音频数据进行分类和检索，这就要用到基于内容检索(Content Based Retrieval ,CBR)的方法。

在基于内容的音频检索技术从 20 世纪 90 年代末兴起以前，一种对语音文本(Speech Document Retrieval)检索的技术已经存在。在这种方法中，先把记录在磁带、(Speech Document Retrieval)检索的技术已经存在。在这种方法中，先把记录在磁带、

录音设备或视频流中的语音信号转录(Transcribe)成文本信号，如将广播新闻转录成文本，然后对文本信息进行关键字提取，进行与关键字相关联的语音文本查询。这种方法一般用于新闻资料、历史资料或教学语言的查询，因为这种方法本质上是对语音信号进行识别，它不涉及去研究如何有效识别出爆炸和掌声等非语音信号，然后基于这些非语音信号进行文本检索[1]。

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

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

音频是声音信号形式。作为一种信息载体，音频可以分为三种类型：

波形声音：对模拟声音数字化得到的数字音频信号，它可以代表语音、音乐、自然界和合成的声响。

语音：具有词字、语法等语素，是一种高度抽象的概念交流媒体。语音经过识别可以转换为文本。文本是语音的一种脚本形式。

音乐：具有节奏、旋律或和声等要素，是人声或和乐器音响等配合所构成的一种声音。音乐可以用乐谱来表示。

不同的类型将具有不同的内在的内容。但从整体看，音频的内容分为三个级别：最低层的物理样本级、中间层的声学特征级和最高层的语义级同，如图(1-1)所示。从低级到高级，内容逐级抽象，内容的表示逐级概括。

在物理样本级，音频内容呈现的是流媒体形式，用户可以通过时间刻度，检索或调用音频的样本数据。例如现在常见的音频录放程序接口。在下一个较高层是声学特征级。声学特征是从音频数据中自动抽取的。一些听觉特征表达用户对音频的感知，可以直接用于检索；一些特征用于语音的识别或检测，支持更高层的内容表示。另外还有音频的时空结构。

最高层是语义级，是音频内容、音频对象的概念级描述。具体来说，在这个级别上，音频的内容是语音识别、检测、辨别的结果；音乐旋律和叙事的说明；以及音频对象和概念的描述。高两层是基于内容的音频检索技术最关心的。在这两个层次上，用户可以提交概念查询，或按照听觉感知来查询[1]。

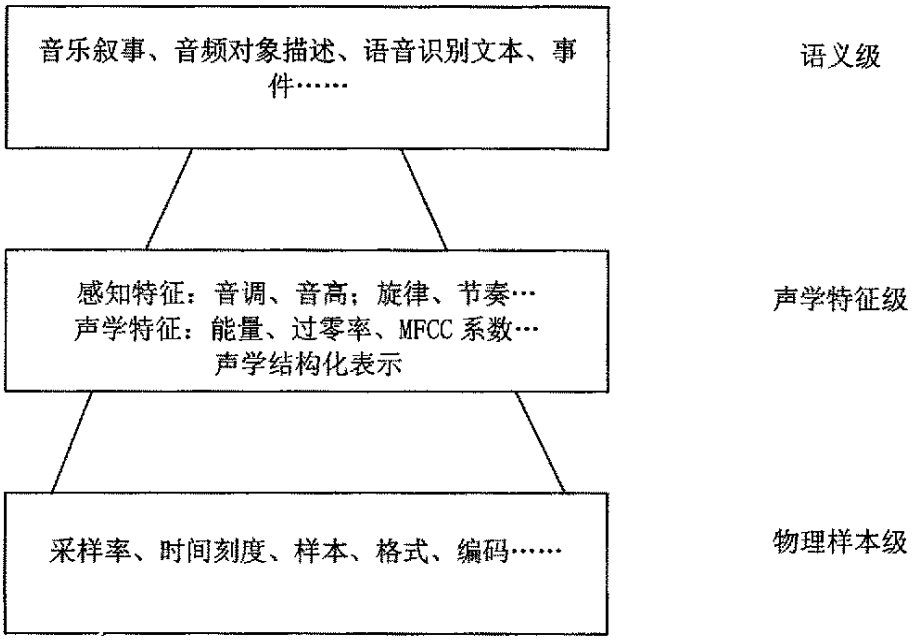


图 1-1 音频内容分层描述模型

音频的听觉特性决定其查询方式不同于常规的信息检索系统.。基于内容的查询是一种相似查询，它实际上是检索出与用户指定的要求非常相似的所有声音。查询中可以指定返回的声音数，或指定相似度的大小。另外可以强调或关闭(忽略)某些特征成分，甚至可以施加逻辑“非”(或模糊的 less 匹配关系)来指定检索条件，检索那些不具有或少有某种特征成分(例如指定没有“尖锐”或少有“尖锐”)的声音。另外还可以对给定的一组声音，按照声学特征进行排序(例如这些声音的嘈杂程度怎样？)。在查询接口上,用户可以采用以下形式提交查询：

示例-用户选择一个声音例子表达其查询的要求，查找出与该声音在某些特征方面相似的所有声音。例如查询与飞机的轰鸣声相似的所有声音。

直喻-通过选择一些声学感知物理特性来描述查询要求，例如亮度、音调和音量等。这种方式类似于可视查询中的描绘查询。

拟声-发出与要查找的声音性质相似的声音来表达查询的要求。例如，用户可以发出嗡嗡声来查找蜜蜂或电气嘈杂声。

主观特征-用个人的描述语言来描述声音。这需要训练系统理解这些描述术语的含义。例如一个用户可能要寻找“欢快”的声音。

浏览-浏览是信息发现的一种重要手段，尤其是对于音频这种时基媒体。除了在分类的基础之上浏览目录之外，重要的基于音频的结构进行浏览。根据对音频

媒体的划分，可以知道语音、音乐和其它声响具有显著不同的特性，因而目前的处理方法可以分为相应的三种：处理包含语音的音频和不包含语音的音频，后者又把音乐单独划分出来。换句话说，第一种是利用自动语音识别技术，后两种是利用更一般性的音频分析，以适合更广泛的音频媒体，如音乐和声音效果，当然也包含数字化语音信号。音频信息检索于是分为：

语音检索-以语音为中心的检索，采用语音识别等处理技术。例如电台节目、电话交谈、会议录音等；

音乐检索-以音乐为中心的检索，利用音乐的音符和旋律等音乐特性来检索。例如检索乐器、声乐作品等；

音频检索-以波形声音为对象的检索,这里的音频可以是汽车发动机声、雨声、鸟叫声，也可以是语音和音乐等，这些音频都统一用声学特征来检索。

1.3 基于内容的音频检索的现状及发展概况

对于人的感官来说，有视觉、听觉、触觉和味觉等方面的感知。在视觉方面，可以感知位置、运动、颜色、纹理、形状、符号等；在听觉方面，可以感知位置、运动、音调、音量、旋律等；还有触觉(机械的、热的、电的、肌肉运动方面的)和嗅觉(气味、味道等)。除了视觉，人们可以从听觉中获得许多的信息，例如我们日常收听的电台节目中欣赏的音乐、聆听的自然声响等。

音频是多媒体中的一种重要媒体。我们能够听见的音频频率范围是 60Hz 到 20kHz，其中语音大约分布在 300Hz 到 4000Hz 之内，而音乐和其它自然声响是全范围分布。声音经过模拟设备记录或再生，成为模拟音频，它们经数字化成为数字音频。数字化时的采样率必须高于信号带宽的二倍，才能正确恢复信号。样本可用 8 位或 16 位比特表示。

以前的许多研究工作涉及到语音信号的处理，例如语音识别.机器容易自动识别孤立的字词，例如用在专用的听写和电话应用方面。连续的语音识别较困难，错误较多，但是目前在这方面已经取得了突破性的进展。还研究了说话人的辨别技术。这些研究成果将为音频信息的检索提供很大的帮助。

常规的信息检索(IR)研究主要是基于文本，例如我们已经非常熟悉诸如 Yahoo 和 AltaVista 这样的搜索引擎。经典的 IR 问题是利用一组关键字组成的查询来定位需要的文本文档。即定位文档中的查询关键字来发现匹配的文档。如果一个文档包含较多的查询项，那么它就被认为比其它包含较少查询项的文档更“相关”。于

是文档可以按照“相关”度来排序，并显示给用户以便进一步搜索。虽然这种一般的 IR 过程是为文本设计的，但是显然也适用于音频或其它多媒体信息的检索。然而，如果我们把数字音频当成一种不透明的位流来管理，虽然可以赋予名字、文件格式、采样率等属性,然而其中没有可以确认的词，或可比较的实体，因此不能向文本那样搜索或检索其内部的内容。对于音乐和非语音声响也是这样。

国外研究机构对音频检索进行了多方面的研究。Muscle Fish[2]是一个商业化的基于音频感知特征的音频检索引擎。Carnegie Mellon 大学的 Informedia 项目[3]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。Cambridge 大学的 VMR(视频邮件检索)小组[4]利用基于网格的词组发现技术检索视频邮件中的消息。Maryland 大学的 VoiceGraph[5]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[6]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。[7]的作者研究了音乐曲调和旋律的检索。另外，MIT[8]、Cornell 大学、南加州大学[9]、澳大利亚 Wollongong 大学、欧洲 EUROMEDIA 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、结构化音频表示和基于说话人的分割和索引等方面的研究。

1.4 论文安排

第一章：本章介绍课题来源，基于内容的音频检索研究背景和发展概况。

第二章：本章讨论了音频信号去噪的常用方法，包括谱减法、中值滤波和预加重滤波，研究了音频信号分帧和加窗处理，进行参数分析。

第三章：本章详细介绍了音频信号的时域特征，包括短时平均能量、短时过零率；音频信号频域特征算法原理和应用。对短时平均能量、短时过零率进行研究，可能根据这两种特征对音频信号进行粗分类。着重研究了音频信号的频域特征，提出了一种均值 MFCC 系数算法，对单一音频信号进行分析时，可以将均值 MFCC 系数作为音频信号的特征。

第四章：本章主要介绍了音频分割和识别算法。研究了音频分层分割，单一音频例子识别两个方面。结合上一章音频特征提取，对音频信号进行识别实验，验证了 Mel 倒谱系数作为单一音频信号特征进行音频识别的可行性。

第五章：介绍音频检索算法，以及音频检索的分类和音频检索的未来和挑战。

第二章　音频信号预处理

2.1 音频处理技术介绍

由于音频本质是信息的载体，在音频检索研究中，一般对信号需要进行三方面的研究，如图(2-1)：（1）研究音频信号如何产生的，这方面研究集中在为音频信号建立产生模型，通过产生模型提取音频特征；（2）音频如何传播，也就是说，音频信号如何通过另外介质传播到人的耳朵里，目前音频检索中在这方面的研究较少；（3）音频信号如何被再形成音频场景。如果要使用计算机取代人，对音频信号大脑皮层中的感知器官处理，而后再形成音频场景。如果要使用计算机取代人，对音频信号进行自动理解，就必须研究人对音频信号的感知机制，使计算机能够像人一样对音频信号自动理解与分析，极大地方便了人们对数据的组织与管理。

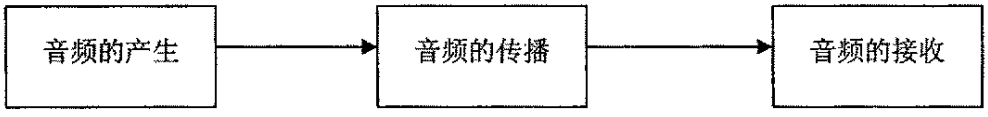


图 2-1 音频是信息的载体

在音频检索中，需要经过特征提取、音频分割、音频识别分类和索引检索这几个关键步骤[10]，如图(2-2)

图 2-2 基于听觉内容的音频检索

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，其中语音大约分布在300Hz-4kHz之内，而音乐和其他自然声响可以分布在60Hz-20kHz任何区域。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。要说明的是，音频信号数字化时的采样必须高于信号带宽的 2 倍，才能正确恢复信号（即奈魁斯特采样频率）[11]。

2.2 音频去噪

音频信号往往伴随着各种不同的噪声，这些噪声可能是外界环境因素造成的，也可能是录音系统本身的某些因素造成的。对音频信号进行去噪，目的就是减小噪声影响，提高信号信噪比。音频去噪的常用方法有谱减法、最小均方误差、中值滤波法等。

谱减法降噪处理，音频增强中的谱减法，假定噪声是加性的，因为大多数噪声短时平稳，以无音频信号声段为参考波动减去音频信号中的加性干扰。线性谱减法如下：

$$|S(W)|=\begin{cases}|X(W)-|N(W)|, |X(W)|-|N(W)|\geq \varepsilon \\ \varepsilon, |X(W)|-|N(W)|<\varepsilon\end{cases} \tag{2-1}$$

$X(W)$、$N(W)$ 为含噪音频和背景噪声的频谱，$S(W)$ 表示经过噪声消除后的音频频谱。为了对前景噪声描述精确，$N(W)$ 取音频前的 N 帧背景噪声的统计平均值。

中值滤波是一种非线性滤波，在 1971 年由 J. W. Tukey 首先提出并应用在一维信号处理技术（时间序列分析）中，中值滤波一般采用一个含有奇数个点的滑动窗口，将窗口中各点值的中值来替代指定点（一般是窗口的中心点）的值。对于奇数个元素，中值是按大小排序后，中间的数值；对于偶数个元素，中值是按排序后中间两个元素值的平均值。三点中值滤波的公式为：

$$\hat{S}(n)=mid\{Y(n-1),Y(n),Y(n+1)\} \tag{2-2}$$

$Y(n)$ 是含噪音频信号，$\hat{S}(n)$ 是音频信号估计。五点中值滤波和七点中值滤波公式和三点中值滤波公式相似。

对于上述滤波算法，通过实验，发现对于特定环境下的连续音频信号，分别对环境音和信号音频进行采集，中值滤波得法在消除一部分环境音的同时，音频信号也被消弱，而采用谱减法能更好的去除环境噪声，并不影响音频信号强度。本文对音频信号的去噪算法，采用了谱减法。图(2-3)为音频信号进行谱减法处理的结果。

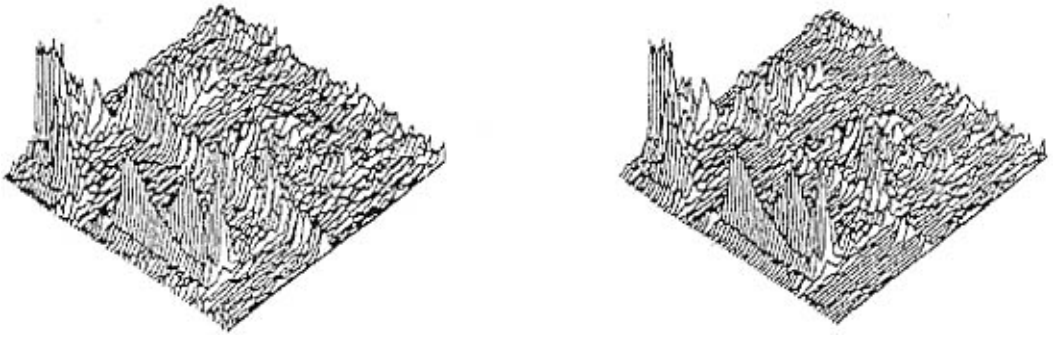
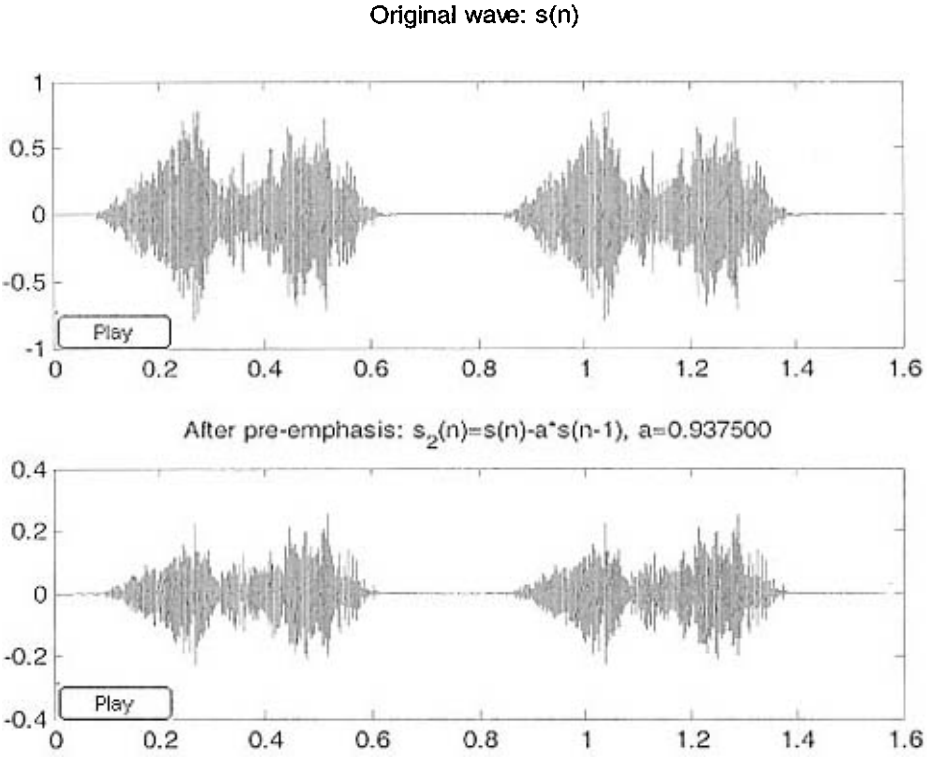


图 2-3　线性谱减前后音频信号

同时结合人耳听觉机理，人耳能够听见的音频频率范围是 60Hz-20kHz，进行音频信号处理时，对音频信号进行预加重处理，经过[1－∂]滤波器进行滤波，滤除 60Hz 以下音频信号。对于不同的采样率，∂取值不同。采样率为 8KHz 音频信号，∂取 0.9375，图(2-4)是对音频信号进行预加重处理结果。



图(2-4) 预加重处理结果

2.3 音频分帧处理

音频信号处理时，音频信号特性在很短时间区间内变化是很缓慢的，所以在这个变化缓慢的时间内所提取的音频特征保持稳定。这样，对音频信号进行处理时，首先就是将离散音频信号分成一定长度单位进行处理，即将离散音频采样点分成一个个音频帧（窗口）。这种方法就是信号‘短时’处理方法，一般一个‘短时’音频帧持续时间长度约为几个到几十毫秒。

一段连续音频信号流 x 采样后的离散音频信号可以表示为 $x=(x(1),...,x(n)..x(K))$，这意味着从此连续音频信号中得到了 K 个采样数据，其中 $x(n)$ 是时刻 $n(1\le n\le k)$ 得到的数据。在“短时”处理时候，假设将这 K 个数据分成 L 组，每一组就是一帧，每一帧包含 $[K/L]$ 个采样点（当然，一般相邻帧间有迭加，其迭加率为 50%-70%左右）。从每一组帧的 $[K/L]$ 个采样点可以提取 $nFeature$ 个特征，最后得到 $L\times nFeature$ 个特征就构成了音频数据 x 的特征，这些特征被用来对音频数据流 x 进行分割、识别与检索。

通过上面的分析可以知道，音频信号“短时”特征处理方法是从采样点集合中提取特征，而不是像视频处理时，从每个“关键”采样点(即视频关键帧)中提取的特征来表示视频数据(在视频处理中，需要从每个视频图像帧中，提取特征进行镜头分割，然后用“关键帧”的特征去表征视频数据)。

本文对音频信号进行分帧处理，主要进行了以下考虑：实验过程中，录制音频信号采样率为 44100Hz，对音频例子进行分帧处理时，考虑音频信号采样率为 44100Hz，取 1024 个采样点做为一个“短时帧”，1024 个采样点约为 25 毫秒，能够满足音频信号短时稳定的特性。帧迭加 512 个采样点，迭加率 50%，同时考虑到音频信号的连续性。音频信号分帧实验结果如图(2-5)所示。

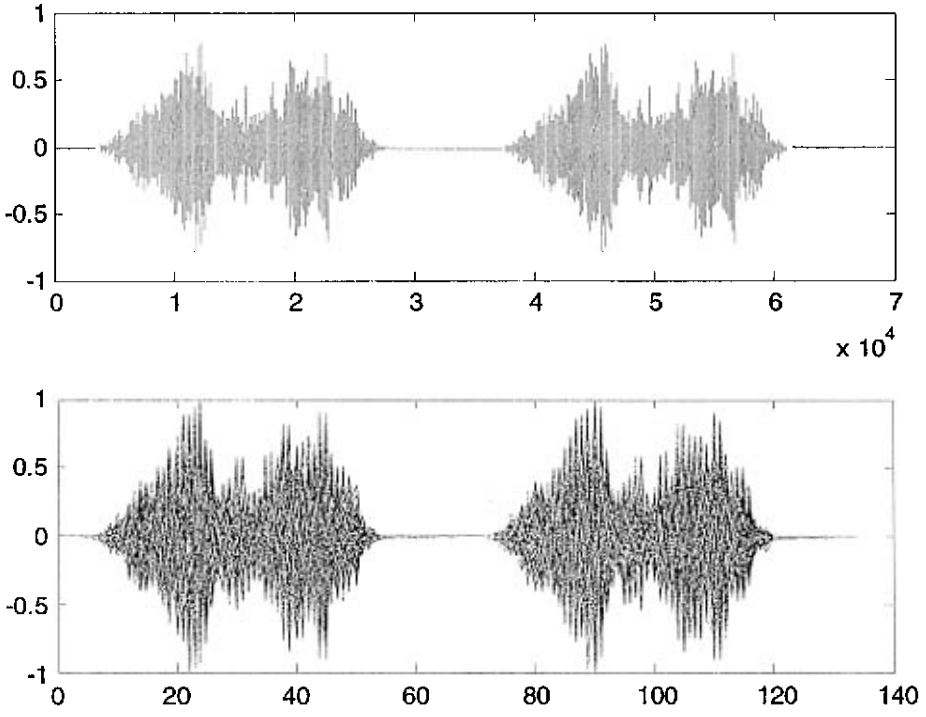


图 2-5 音频例子分帧处理

2.4 音频加窗处理

对于采样得到的$x(n)(1 \le n \le K)$音频信号，考虑到信号在短时间内的连贯性，首先把音频信号的K个采样点分割成前后迭加的帧(每个音频帧内包含几百个采样点)，相邻帧间的迭加率一般为 50%-70%，音频处理中的“短时帧”均是这样得到的[11]。离散时间信号总是有限长的，因此不可避免地要遇到数据截短问题。在信号处理中，对离散信号序列的截短是通过离散信号序列与窗口函数相乘来实现的。设$x(i:i+N)$是一个含N个采样点的短时帧，$w(i)$是长度为N的窗函数，用$w(i)$截短$x(i:i+N)$，得到N点序列$\overline{x(i:i+N)}$，即$\overline{x(i:i+N)} = x(i:i+N)w(n)$，通过这样的途径，先前第个短时帧中的$N$个采样点$x(i:i+N)$被转换成$\overline{x(i:i+N)}$。由于时域上信号做卷积计算，相当于频域上相乘，因此窗口函数计算也可以如下表示：

$$X_N(e^{jw}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\theta})W(e^{j(w-\theta)})d\theta \qquad (2\text{-}3)$$

其中，X 和 W 分别表示频谱。

由此可见，窗口函数 $w(n)$ 不仅影响原信号在时域上的波形，而且也影响其频域的形状。常用的窗口函数有矩形窗、巴特立特(Bartlett)窗、三角窗、海明(Hamming)窗、汉宁(Hanning)窗、切比雪夫(Chebyshev)窗、布莱克曼(Blackman)窗、凯泽(Kaiser)窗等[12]。

矩形窗：

$$w(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & n=else\end{cases} \tag{2-4}$$

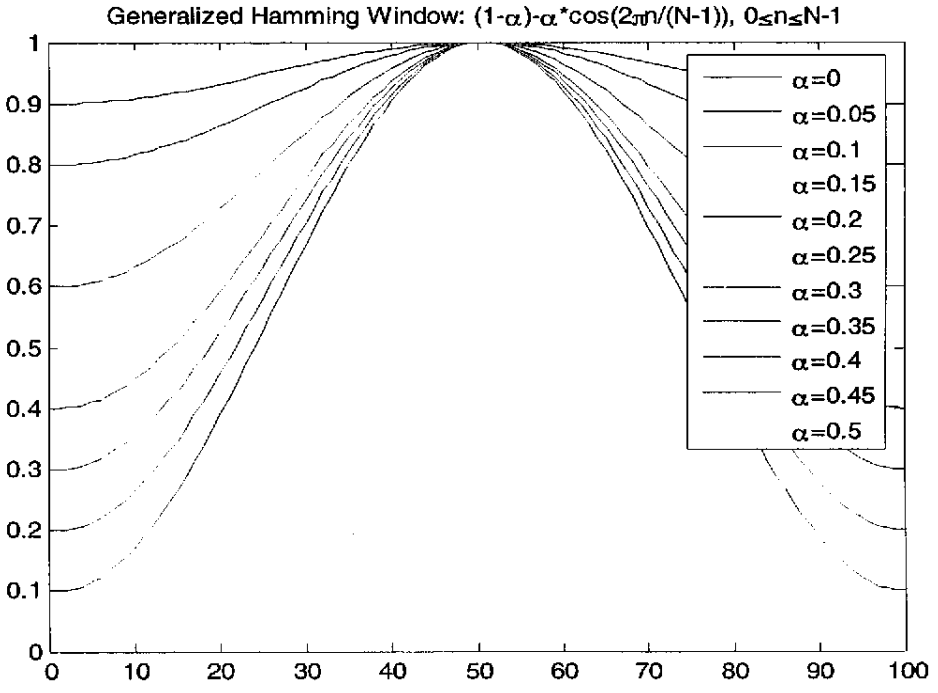
汉明窗：

$$w(n)=\begin{cases}(1-\partial)-\partial\cos[2\pi n/N-1], & 0\le n\le N-1\\ 0, & n=else\end{cases} \tag{2-5}$$

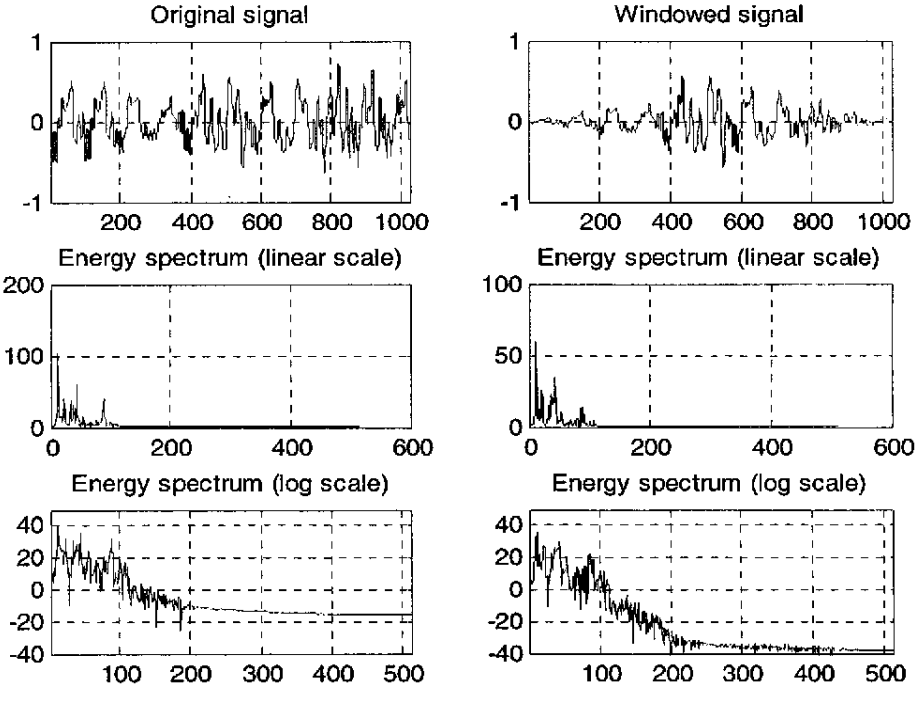
窗函数形状和长度的选择，对于短时分析参数的特性影响很大，为此应选择合适的窗口，使其短时参数更好的反映音频信号的特性变化。矩形窗的谱平滑性较好，但损失了高频成分，使波形细节丢失，并且矩形窗会产生泄漏现象；而汉明窗可以有效的克服泄漏现象，应用范围也最广泛[13]。窗口长度 N 如果很大，则它等效于很窄的低通滤波器，音频信号通过时，反映波形细节的高频部分被阻碍，短时能量随时间变化很小，不能真实的反映语音信号的幅度变化；反之，如果 N 太小，滤波器的通带变宽，短时能量随时间有急剧的变化，不能得到平滑的能量函数，因此，窗口的长度选择应该合适。

根据上面的分析，本文对汉明窗不同的 ∂ 取值进行实验，不同 ∂ 值时，窗口函数分析如图(2-6)。

实验过程中，$\partial=0.46$ 时，对音频信号进行加窗处理，能够很好的反映音频信号的特性变化。图(2-7)为加窗后结果。



图(2-6)汉明窗变化图



图(2-7) 音频信号加窗分析

2.5 本章小结

本章讨论了音频信号去噪的常用方法，包括谱减法、中值滤波和预加重滤波，对于连续音频信号，谱减法能很好的减少环境噪声。结合人耳特点，预加重处理能减少低频噪声。同时，根据项目的要求，对音频信号进行分帧和加窗处理，进行参数分析，找到合适的处理参数。

第三章 音频特征提取

3.1 引言

连续音频信号经过采样，变成离散信号后，按照对$[K/L]$个采样点提取特征方式不同，可以从音频信号中提取三类基本特征：时域特征、频域特征和时频特征。这三类特征空间从不同角度刻划了音频信号的实质，构成了音频信号的描述算子。

另外，按照特征提取单位长短的不同，也可以从音频信号x中提取音频帧特征(Audio Frame)和音频例子(Audio Clip)特征两种不同形式的特征。x的音频帧特征就是从每个$[K/L]$个采样点中分别提取特征，所有$[K/L]$个点中提取的特征就构成了x的特征向量。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和，如爆炸和掌声会持续几秒。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，所以在这种方法中，直接对x提取特征，也就是把x的所有采样点只看成一个“短时帧”，但是这样处理的结果过于粗糙。实际中，对于x的所有采样点$x(n)(1 \le n \le K)$，为了既考虑音频短时平稳特性，又考虑音频信号本质非平稳特性，一般先提取每个含$[K/L]$个采样点的音频帧特征，然后计算音频帧的统计特征(如平均值和方差等)，作为x的音频例子特征。

3.2 音频时域特征提取

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。考察音频信号在这个坐标轴上的能量幅度，对短时平均能量，过零率和线性预测系数等时域特征进行验证。

3.2.1 短时平均能量

短时平均能量指在一个短时音频帧内采样点信号所聚集的平均能量。假定一段连续音频信号流 x 到 K 个采样点，这 K 个采样点被分割成迭加率为 50%的 M 个短时帧。每个短时帧和窗口函数大小假定为 N，对于第 m 个短时帧，其短时平均能量可以使用下面公式计算：

$$E_m = \frac{1}{N}\sum_m [x(n)w(n-m)]^2 \tag{3-1}$$

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。

本文根据上面短时能量公式，对音频例子进行短时平均能量计算，并经过实验，进行阈值设定，如图(3-1)所示，发现可以通过短时平均能量来进行有声/静音检测。

当然，分析发现，在不同的音频信号时，也有可能出现异常。比如，爆炸一般只持续几个短时帧，并且在爆炸音发生前后的短时音频帧所带的能量都很低。如果只用短时能量进行有声/静音检测算法，则会存在把爆炸音频例子判定为静音的问题。

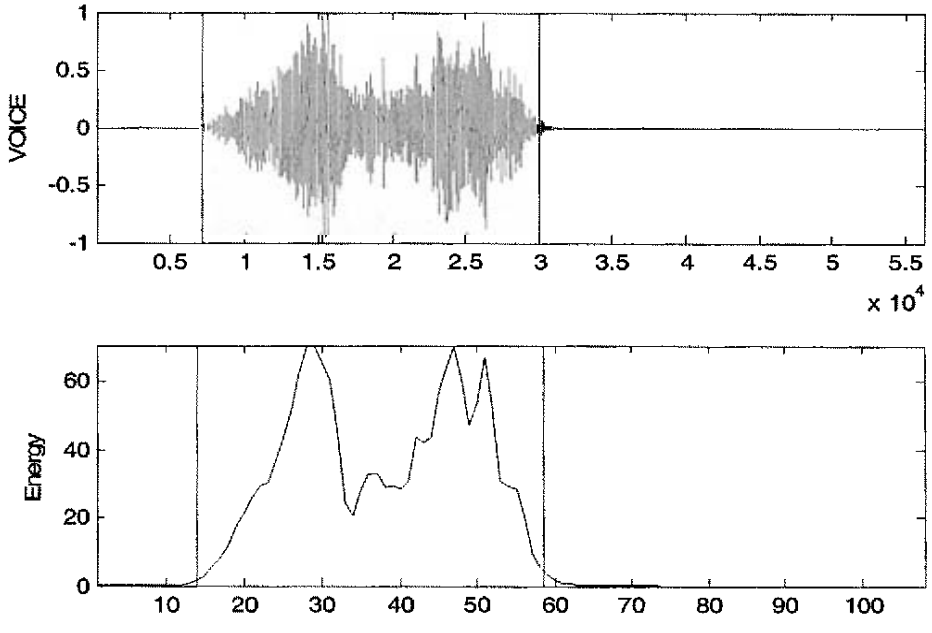


图 3-1 短时平均能量分析

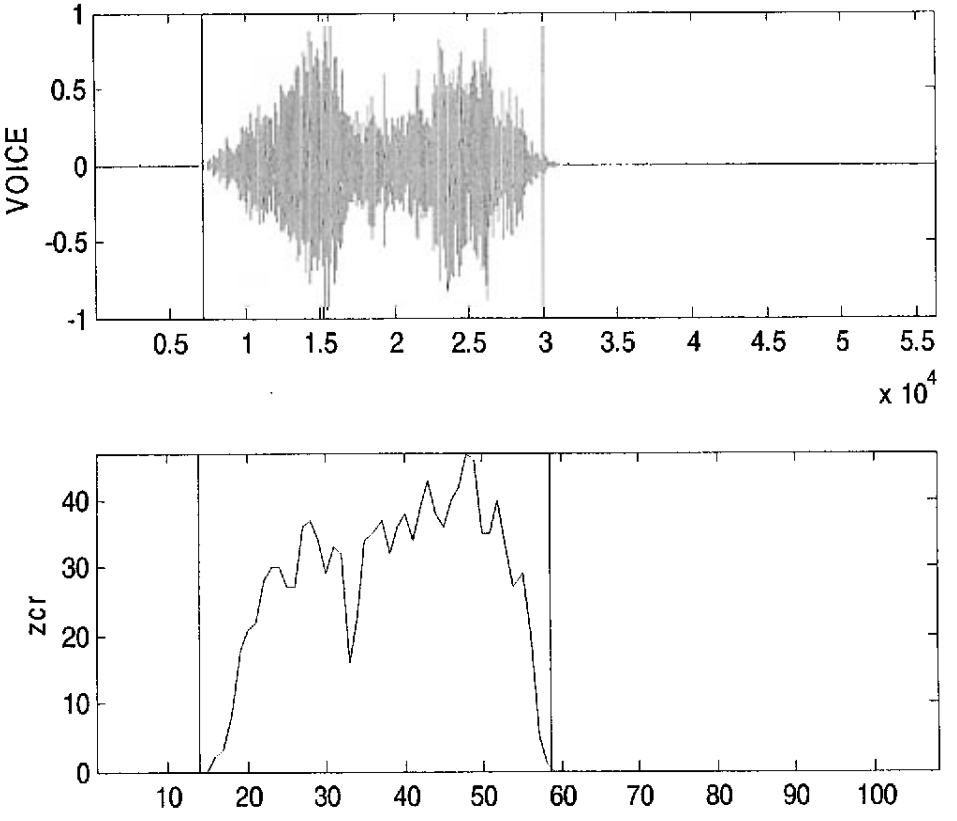
3.2.2 短时过零率

“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短时帧内里的平均频率[14]。对于音频信号流 x 中第 m 帧，其过零率计算如下：

$$Z_m = \frac{1}{2}\sum_m \left|sign[x(n)] - sign[x(n-1)]\right| w(n-m) \tag{3-2}$$

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。当$x(n) \geq 0$时，$sign[x(n)] = 1$；否则$sign[x(n)] = 0$。

如图(3-2)所示，对音频例子进行过零率分析，利用过零率变化，对音频信号进行端点检测。



图(3-2) 过零率分析

通过实验发现，语音的过零率和谐音有着显著的不同，探讨其原因如下：

当人发音时，声音使声门产生振动，发出每秒几十次到百多次的声波脉冲，

然后经过由喉管和口腔组成的将声波帮适当的变形，由口腔送出。在这个过程中，舌的位置和口腔的缩小对声波都有影响，鼻管则起着对声波的旁路作用。当发不同声音时，由于这些器官的位置和形状影响，声波会受到不同的变形，加之声门发出的周期不同，就会发出不同的声音。

比较而言，语音信号比较规范，一般由几个单词构成，每个单词又由元音和辅音交替的章节组成。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，由于爱声道阻碍较小，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，辅音信号的过零率低，而元音信号的过零率就高。

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过零率总会有显著升高，所以利用过零率可以判断语音是否开始和结束。

另外，大多数音乐信号集中在低频部分，其过零率不表现出突然升高或降落的跌宕特性，所以有时候也用过零率来区分语音和音乐两种不同音频信号。

3.2.3 线性预测系数

对于采样后得到的信号序列 $x=(x(1),...x(n),...x(K))$，人们总想用一个模型来模拟它的产生，好比用正弦函数来模拟形状如正弦曲线的声波。

如果用有限个参数的数学模型来线性挖表示音频序列 $x(n)(1\le n\le K)$，这些参数就成为 $x(n)$ 的重要特征，叫做线性预测系数。

记模拟音频信号 $x(n)(1\le n\le K)$ 的数学模型为 $x'(n)$，则：

$$x'(n)=\sum_{k=1}^{p}a_k x(n-K) \tag{3-3}$$

其中 $x(n-K)$ 为音频采样信号，$\{a_k\}$ 为模型参数（又称线性预测系数），p 为模型阶数。从公式可以看出，可以用信号前面的一些采样值（即处暑信号采样值）加权后叠加作为产生音频序列 $x(n)(1\le n\le K)$ 数学模型，也就是用前面的采样信号点去表示后面的采样信号。

借助模型 $x'(n)$，只要知道了前面 p 个采样点，则所有采样点的值可以计算（预测）出来。出于上面的运算属于线性迭加运算，因此是用线性模型去为音频信号序列 $x(n)(1\le n\le K)$ 建立产生模型，其系数 $\{a_k\}$ 就叫做“线性预测系数（Linear Predictive Coefficient, LPC）”。由于 $\{a_k\}$ 反映了音频信号的变化形状，因此可

以代表音频特征。

实际中，不是为音频信号流 x 的全部 K 个采样点建立一个线性产生模型，而是为每个音频帧建立一个线性预测模型。每个短时音频帧有 p 个系数，将这 p 个系数作为这个短时音频帧的特征。在计算模型系数时，采用如下最小均方误差解法，即定义音频短时帧的平均预测误差 E_m 为:

$$E_m = \sum_n (x_m(n) - \sum_{k=1}^{p} a_i x_m(n-i))^2 \qquad (3-4)$$

其中 $x_m(n) = x(n+m)$。在上式中，令 $\partial E_m / \partial a_k = 0, k = 1,2,...p$，就可以得到一组线性方程组，解之即可得出最佳的模型参数。

线性预测模型最大的优点是模型求解是一个线性问题，容易计算。缺点是模型精度不高，只是近似计算，通过模型模拟(预测)的信号其与原始信号存在误差。

事实上，由于音频信号的产生是一个非线性过程，其中存在着混沌随机因素，因而用线性模型来描述语音信号在理论上是不合适的，它不能描述音频信号的非线性本质，因而导致音频处理效果不理想。非线性模型虽有可能更精确模拟音频信号的产生过程，但其求解是一个非线性寻优过程，计算量大，收敛性及稳定性均无法保证。并且，得出一个通用的非线性模型是很困难的。

即使这样，近年来，人们已越来越重视音频非线性分析方法的研究，如神经网络方法。但是由于非线性模型分析方法普遍受到计算复杂和稳定性差等问题的困扰，目前还未得出一种令人接受的音频信号非线性模型分析方法。

上面介绍了从音频短时帧中提取最常用音频时域特征的方法，这些短时帧时域特征还可以衍生出其它特征，比如特征均值和方差等。

如果一段连续音频信号流 x 得到 K 个采样数据，将这 K 个数据分成 M 个迭加短时帧，从每个短时帧中提取平均能量和过零率特征，并且为每一短时帧建立一个 4 阶线性预测模型，则每一短时帧可以提取 6 个时域特征，从 x 中总共提取 $6 \times M$ 个特征向量，就可以使用这 $6 \times M$ 个特征向量对音频数据流 x 进行分割、识别与检索。

3.3 音频频域特征提取

在上面提取音频信号流 $x = (x(1),...x(n),...x(K))$ 时域特征时，将每个时刻的采样点 $x(n)(1 \le n \le K)$ 看成音频在这个时刻信息的全部。

其实，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，人们之所以能够感受到音频信号，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。

音频信号频域特征有多种，常用的有线性预测(LPC)倒谱系数或MEL频率倒谱系数(Mel-frequency cepstrum coefficients-MFCC)。通常MFCC参数比LPC倒谱系数更符合人耳的听觉特性，在有停产噪声和频谱失真情况下，能产生更高的识别精度。

3.3.1 抽样定理

音频信号$x(t)$很多都是连续信号，t的取值是从$-\infty$连续变化到$+\infty$。但是用计算机对这些信号处理时，必须对连续信号采样，即按一定的时间间隔△进行取值，得到$x(n\Delta)(n=...,-2,-1,0,1,2...)$，称△为抽样(或采样)间隔，称$x(n\Delta)$为离散信号或时间序列。问题是：是否能够由离散信号序列$x(n\Delta)$恢复出连续信号$x(t)$。对于最基本的正弦波$x(t)=A\sin(2\pi \times ft+\varphi)$，其中 f>0，按抽样间隔△抽样得到离散信号$x(n\Delta)$，有如下正弦波抽样定理：当$f<\frac{1}{2\Delta}$时，离散信号$x(n\Delta)$可以惟一确定 x(t)；否则不能惟一确定。

对于一般的连贯信号$x(t)$，可以表示为无限多个谐波的叠加，其中频率为 f 的谐波，振幅和初相位可由频谱$X(f)$表示。

由傅立叶积分公式$x(t)=\int_{-\infty}^{+\infty}X(f)e^{i2\pi\times ft}df$知道，对于频率 f，当$X(f)\neq 0$时，就表示连续信号$x(t)$包含了频率为 f 的谐波成分，当$X(f)=0$时，表示 x(t)不包含了频率为 f 的谐波成分。由于离散信号$x(n\Delta)$恢复出连续信号$x(t)$，意味着$x(t)$所包含的所有谐波都可以通过抽样间隔为△的离散谐波惟一恢复出来。也就是说，对于频率 f，只要$X(f)\neq 0$，f 和△都必须满足关系$f<\frac{1}{2\Delta}$，就可以恢复频率为 f 连续谐波信号。如果$X(f)\neq 0$的频率 f 任意大，那么采样时间间隔△就接近 0，表示连续信号$x(t)$不可能由离散信号恢复出来。因此，要由$x(n\Delta)$恢复出连续信号$x(t)$，频谱$X(f)$和抽样间隔△必须满足如下条件：

$X(f)$有截频f_c，即当$|f|\ge f_c$时，$X(f)=0$；并且$\Delta\le\frac{1}{2f_c}$或者$f_c\le\frac{1}{2\Delta}$。也就是说，信号$x(t)$的频率范围是有限的，只包含低于f_c的频率部分。$x(t)$表示为谐

波的叠加，这些谐波的频率 f 在$0 \le f < f_c$之间，谐波的周期 T 在$\frac{1}{f_c} < T < +\infty$之间，$T_c = \frac{1}{f_c}$为极小周期，抽样间隔不能超过极小周期$T_c$的一半。

在离散信号$x(n\Delta)$的频谱中，频率$\frac{1}{2\Delta}$起着重要作用，称为奈魁斯特(Nyquist)频率。

3.3.2 连续信号的滤波和卷积

实际工程中的连续信号$x(t)$含有效信号$s(t)$和干扰信号$n(t)$两部分。对信号处理的一个重要目的是增强有效信号$s(t)$，削弱干扰信号$n(t)$。在许多情况下，干扰信号$n(t)$的频谱$N(f)$与有效信号$s(t)$的频谱$S(f)$是不同的。

用一个频率函数$H(f)$与信号$x(t)$的频谱$X(f)$相乘得到$Y(f) = X(f)H(f)$，这个过程叫做滤波。而频率函数$H(f)$的作用就起到削弱干扰信号$n(t)$频率的作用。

设原始信号$x(t)$的频谱为$X(f)$，用来滤波的频谱$H(f)$所对应的时间函数为$h(t)$，滤波后的频谱$Y(f) = X(f)H(f)$所对应的时间函数为$y(t)$。现在推导滤波后的信号$y(t)$与原始信号$x(t)$和用于滤波的时间函数$h(t)$之间的关系。由傅立叶变换知道：

$$\begin{aligned} y(t) &= \int_{-\infty}^{+\infty} Y(f)e^{j2\pi \times ft}df = \int_{-\infty}^{+\infty} X(f)H(f)e^{j2\pi \times ft}df \\ &= \int_{-\infty}^{+\infty} X(f)[\int_{-\infty}^{+\infty} h(l)e^{-j2\pi \times fl}dl]e^{j2\pi \times ft}df = \int_{-\infty}^{+\infty} h(l)[\int_{-\infty}^{+\infty} X(f)e^{j2\pi \times f(t-l)}df]dl \\ &= \int_{-\infty}^{+\infty} h(l)x(t-l)dl \end{aligned} \tag{3-5}$$

$y(t)$通过由$x(t)$和$h(t)$通过上式形式的积分得到，把上式表示的$y(t)$称为$x(t)$与$h(t)$的卷积，并记为：$y(t) = x(t) \times h(t)$。这样，原始信号函数，滤波函数和滤波后的函数满足如下式子：$\begin{aligned} Y(f) &= X(f)H(f) \\ y(t) &= x(t) \times h(t) \end{aligned}$。

从数学角度看，上面两个式子的意义是：两个频谱相乘，其时间函数就是相应的两个时间函数进行卷积；反之，两个时间函数卷积，其频谱就是相应的两个频谱相乘。

从滤波的角度看，上面两个式子的意义是：滤波可以通过两种方式实现。是在频率域实现，将频谱$H(f)$与$X(f)$相乘得到$Y(f)$，再由$Y(f)$做反傅立叶变换

得到$y(t)$；二是在时间域上实现，将时间函数$x(t)$与$h(t)$通过卷积得到$y(t)$。

3.3.3 能谱特征

如果信号$x(t)$表示电压，假定电阻是 1，则瞬时功率为$x^2(t)$，总的能量就是$\int_{-\infty}^{+\infty}x^2(t)dt$。

按照卷积理论，如果$y(t)$为$x(t)$与$h(t)$的卷积，即$y(t)=x(t)\times h(t)$，相应的频谱关系为$Y(f)=X(f)H(f)$。按照信号与频谱的关系$y(t)=\int_{-\infty}^{+\infty}Y(f)e^{j2\pi\times ft}df$，有：

$$\int_{-\infty}^{+\infty}x(l)h(t-l)dl=\int_{-\infty}^{+\infty}X(f)H(f)e^{j2\pi\times ft}df \tag{3-6}$$

(3-6)式中，取$t=0$，有：

$$\int_{-\infty}^{+\infty}x(l)h(-l)dl=\int_{-\infty}^{+\infty}X(f)H(f)df \tag{3-7}$$

(3-7)式中，取$h(l)=x(-l)$。$h(l)$的频谱为：

$H(f)=\int_{-\infty}^{+\infty}x(-l)e^{-i2\pi\times fl}dl=\int_{-\infty}^{+\infty}x(s)e^{j2\pi\times sf}ds=\overline{X(f)}$。这时，(2.8)变为：

$$\int_{-\infty}^{+\infty}x^2(l)dl=\int_{-\infty}^{+\infty}|X(f)|^2df \tag{3-8}$$

(3-8)式又叫能量等式，也称为帕舍瓦尔(Parseval)等式。这样可以知道，$x(t)$的能量可通过$|X(f)|^2$表示出来，为此，称$|X(f)|^2$为$x(t)$的能谱。

3.3.4 平均功率与功率谱特征

当连续信号$x(t)$的总能量为无限时，就要考虑功率和功率谱。称$\frac{1}{T_2-T_1}\int_{T_1}^{T_2}|x(t)|^2dt$为$x(t)$在区间$[T_1,T_2]$上的平均功率。

现在求$x(t)$在区间$[T_1,T_2]$上的功率谱。设$x(t)$在区间$[T_1,T_2]$外取值为 0，$x(t)$的频谱为$X(f)=\int_{-\infty}^{+\infty}x(t)e^{-i2\pi\times ft}dt=\int_{T_1}^{T_2}x(t)e^{-i2\pi\times ft}dt$。按照能量等式，$x(t)$的能量和能说有关系：

$$\int_{-\infty}^{+\infty} x^2(t)dt = \int_{-\infty}^{+\infty} |X(f)^2| df \tag{3-9}$$

从(3-9)式可以得到：

$$\frac{1}{T_2 - T_1}\int_{T_1}^{T_2} x^2(t)dt = \int_{-\infty}^{+\infty} \frac{1}{T_2 - T_1}\left|\int_{T_1}^{T_2} x(t)e^{-i2\pi \times ft}dt\right|^2 df \tag{3-10}$$

(3-10)式的左端是$x(t)$在区间$[T_1, T_2]$上的平均功率，称$\frac{1}{T_2 - T_1}\left|\int_{T_1}^{T_2} x(t)e^{-i2\pi \times ft}dt\right|^2$为$x(t)$在区间$[T_1, T_2]$上的功率谱。

3.3.5 倒谱特征分析

倒频谱分析是一种非线性信号处理技术，其基本要领是由 Bogert, Healy 和 Tukey 在 1963 年引入的。它是同态系统理论的基础，是专门处理通过卷积组合在一起的信号的。由于倒谱分析与信号处理的 Z 变换有关，先简单介绍 Z 变换。

在模拟信号系统中，用傅立叶变换进行频域分析，拉普拉斯变换作为傅立叶变换的推广，可以对信号进行复频域分析(也可以通过抽样定理，把模拟信号转换为离散信号，用傅立叶变换进行频域分析)。同样，在离散时间信号系统中，信号的傅立叶变换用作频域分析，而 Z 变换作为傅立叶变换的推广，用作复频域分析。

设x_n为离散序列，由以前的讨论可以知道，x_n的频谱为：

$$X(f) = \sum_{n=-\infty}^{+\infty} x_n e^{-i2\pi \times n\Delta f} \tag{3-11}$$

如果已经知道频谱$X(f)$，则可知x_n为$x_n = \Delta \int_{\frac{-1}{2\Delta}}^{\frac{1}{2\Delta}} X(f)e^{i2\pi \times n\Delta f} df$。把$e^{-i2\pi \times n\Delta f}$表示为：$(e^{-i2\pi \times \Delta f})^n$，令$Z = e^{-i2\pi\Delta f}$，则$X(f) = \sum_{n=-\infty}^{+\infty} x_n Z^n$。这种表示比较简洁，即：

$$\hat{X}(Z) = \sum_{N=-\infty}^{+\infty} x_n Z^n \tag{3-12}$$

$\hat{X}(Z)$称为x_n的 Z 变换。其中 Z 是一个复变量，它所在的复平面叫做 Z 平面。

Z 变换与频谱变换的关系是：把频谱$X(f)$中的$e^{-i2\pi \times \Delta f}$换成 Z 就得到了 Z 变换$\hat{X}(Z)$：反之，把 Z 变换$\hat{X}(Z)$变换成$e^{-i2\pi \times \Delta f}$就得到频谱$X(f)$。由于频谱与 Z 变换之间只是一种符号的，其实质并没有改变。

离散序列x_n的频谱为$X(f)$，Z 变换为$\hat{X}(Z)$。(3-11)和(3-12)分别叫做x_n的频谱展开式和 Z 变换展开式，这两个展开式存在惟一性。设离散序列x_n的频谱为$X(f)$，Z 变换为$\hat{X}(Z)$，若$X(f)=\sum_{n=-\infty}^{+\infty}c_n e^{-i2\pi\times n\Delta f}$，$\hat{X}(Z)=\sum_{N=-\infty}^{+\infty}c_n Z^n$，则离散序列$x_n=c_n$。

如果已经知道信号x_n的 Z 变换，去求一信信号序列$\hat{x}_n$的过程叫 Z 反变换(inverse z-transform)。信号x_n的Z反变换数学表达式为：

$$\hat{x}_n=\frac{1}{2\pi\times j}\oint\hat{X}(Z)z^{n-1}dz \tag{3-13}$$

式中$\oint$表示在半径为 r，以原点为中心的封闭圆上沿逆时钟方向环绕一周的积分。r 可以是使$\hat{X}(Z)$收敛的任何值。对于信号x_n，如果它的 Z 变换为$\hat{X}(Z)$，$\hat{X}(Z)$相应的自然对数为$Ln[\hat{X}(z)]$。则x_n的复倒谱(Complex cepstrum)$c_x(n)$是$\hat{X}(Z)$自然对数$Ln[\hat{X}(z)]$的 Z 反变换。若$Ln[\hat{X}(Z)]$收敛，则复倒谱存在，其定义为：

$$c_x(n)=\frac{1}{2\pi\times j}\oint Ln[\hat{X}(z)]z^{n-1}dz \tag{3-14}$$

(3-14)中的$\oint$的积分域为使$Ln[\hat{X}(z)]$收敛的任意圆周。应该指出的是，在这里取x_nZ 变换$\hat{X}(Z)$自然对数来求解复倒谱。实际中可以根据需要取对数的底。

与复倒谱对应，信号x_n的实倒谱(Real Cepstrum，有时也叫导谱)$r_x(n)$定义为其傅立叶变换幅值对数的傅立叶反变换：

$$r_x(n)=\frac{1}{2\pi}\int_{-\pi}^{+\pi}Ln\left|X(e^{jw})\right|e^{jwn}dw \tag{3-15}$$

由于实倒谱与相位无关，所以比复倒谱计算起来更加容易。但是，实倒谱只依赖于傅立叶变换幅值，所以它是不可逆的，也就是说x_n不能用$r_x(n)$去恢复。而复倒谱更为通用。

3.3.6 LPC 倒谱和 Mel 系数

上面主要介绍了信号处理的数学基础，并且介绍了几个音频频域特征。在一般音频识别中，最常用的还是 LPC 和 Mel 倒谱系数。

在进一步介绍之前，先看看频域特征和时域特征的区别和联系：(1)时域特征和频域特征都是从短时音频帧提取的。(2)时域特征是直接在原始信号基础上所提取的特征，而频域是把原始信号先进行傅立叶变换，将原始信号转换到频域，然

后在频域上提取特征。比如，连续音频信号流 x 采样后的离散音频信号 $x=(x(1),...,x(n),...,x(K))$ 分成 M 帧，每帧包含 $[K/M]$ 个采样点(如果帧间有迭加，采样点不是这么多)，那么就是对这 $[K/M]$ 个采样点傅立叶变换，则傅立叶变换得到的系数就表示这个短时音频帧在不同频率上所带能量大小。

图(3-3)给出了对音频短时帧进行时域和频域处理的示意图：可以看到，当把连续音频信号流分成音频短时帧大小后，在时域处理时，对每个短时帧原始信号直接提取特征(这些特征可以是平均能量、过零率和线性预测系数等)；而频域处理时，先对短时帧中 $[K/M]$ 个采样信号点进行傅立叶变换，得到傅立叶系数，也就分别得到在短时帧时间内每个频率带上的能量总和，这样原始信号就从时域转换到了频域。

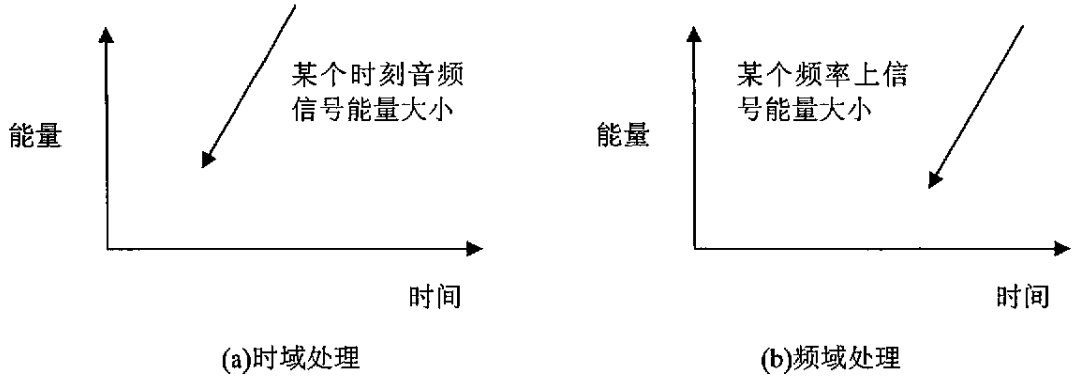


图 3-3　短时音频帧时域和频域特征提取

所以，图(3-3)(a)中的横坐标是时间，它表示 $[K/M]$ 个采样点，纵坐标表示这 $[K/M]$ 个采样点的能量幅度值；图(3-3)(b)中的横坐标是频率，它和采样频率有关，是采样频率的一半，纵坐标表示 $[K/M]$ 个采样点在每个频率上能量的总和。

基于上面的理解，对于某个短时音频帧，其 LPC 倒谱系数提取过程如下：首先用数字滤波器对音频帧所包含的 $[K/M]$ 个采样点进行预加重处理，对预加重处理后的音频帧内信号加窗口函数，然后对它进行自相关分析，把这个结果施以 p 阶线性预测运算，得到长度为 p 的信号序列 x_p，就是音频帧的 LPC 派生倒谱系数；如果对得到的 LPC 派生倒谱系数继续进行 Delta 加权处理，就得到 Delta 倒谱系数。

Mel 倒谱系数(MFCC)是建立在傅立叶和倒谱分析基础上的：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 频率带上具有能量，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 的频率带划分为若干个子带。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 $Mef(f)=2595\lg(1+\frac{f}{500})$；非线性划分中第个频率子带宽度的划分就比较复杂了。无论是线性划分子带，还是非线性划分子带，如果整个频率带被划分就比较复杂了。无论是线性划分子带，还是非线性划分子带，如果整个频率带被 划分为 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(也叫 Mel 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。

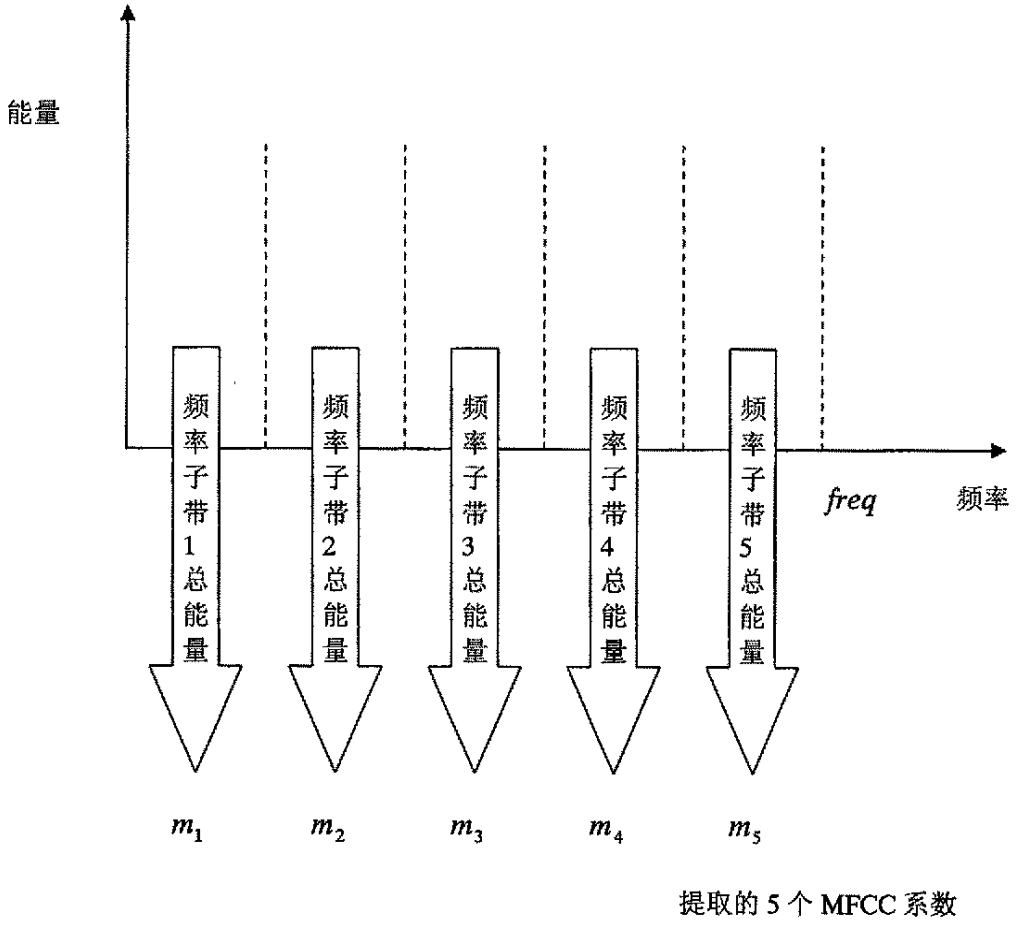


图 3-4 音频短时帧线性 MFCC 系数提取

图(3-4)给出音频短时帧 MFCC 系数提取示意图：在图中，假设这个音频短时帧经过傅立叶变换后，其频谱宽度为 freq(也就是这个音频短时帧最大频率。在前面知道，这个值是音频采样频率的一半)。把这个短时帧频谱均匀线性划分为 5 个频率子带，每个频率子带宽度为 freq/5，每个频率子带上所带能量的总和就构成了这个频率子带的 MFCC 系数，所有 5 个频率子带能量总和就是从这个短时音频帧提取的 MFCC 系数。

在提取 Mel 系数时要注意以下几点：(1)所谓将频率带非线性分为若干频率子带，是指每个子带上的频率宽度是不一样的；(2)生理学研究表明，人耳就是一个滤波器。人耳这个滤波器对某些频率子带的能量敏感，对某些频率子带的能量不敏感。在求 Mel 系数时，如何仿照人耳机制对频率带进行非线性划分，是目前提取音频感知特征研究的热点。

由于 LPC、MFCC 和 Delta 倒谱特征是从每个短时音频帧中提取出来的，它们主要反映的是音频在很短时刻内的静态特征，音频信号的动态特征可以用这些静态特征的差分来描述，如从前后相邻帧提取的 MFCC 特征相减，就是可以反映这个音频 MFCC 特征的动态特性。

把这些动态特征和静态特征一起组成音频的特征向量空间，能够相互互补，很大程度可以提高改善系统的识别性能。

3.3.7 其他频域特征

除了上面介绍的频域特征外，还可以提取其他频域特征，如熵(Entropy)特征和子带组合特征。熵是用来衡量信息复杂度的一个重要指标，其定义如下：

$$\begin{aligned} Etr &= -\sum_{i=1}^{32} P(i)\lg P(i); \\ P(i) &= |M(i)|^2 \Big/ \sum_{i=1}^{32} |M(i)|^2 \end{aligned} \tag{3-16}$$

其中 $M(i)$ 是指将音频帧的频率带划分为 32 个频率子带后，第 i 个频率子带上的能量。

在语音分析中发现，人讲话的音频信号总是集中在某些频率带上，而音乐和自然声音可以分布在所有子带上，所以可以将某些子带上的能量组合起来，判断音频信号是否是语音或音乐，这就构成了子带组合特征。

作为音频信号特征的一种补充，小波系数在音频(音乐)检索中已有应用，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，小波系数都是在短时音频帧中的采样信号经过小波变换后得到的。

音频信号中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 16000Hz 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。

小波变换的等效的频域表示是：

$$W_f(a,b) = < f, \Psi_{a,b} > = |a|^{-1/2} \Big/ 2\pi \int_R F(w)\Psi(aw)e^{jw\pi}dw$$

如果 $\Psi(w)$ 是幅频特性比较集中的带通函数，则小波变换便具有表征待分析信号 $F(w)$ 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分出它们。

如果对未加噪声音频做不同尺度下的小波变换，可以发现在小尺度下，小波变换的波形相减较大，这种较大的差别说明了噪声主要反映了小尺度下的小波变换上。环境背景音乐本身含的噪声频带较宽，其高频成分较多，而从小波变换原理的角度来分析，小尺度变换也就是说小波函数的窗口尺度较小，能很好地反映信号的高频信息，所以经过小波变换后，噪声的主要能量分布在小尺度下的小波变换上。相反，大尺度变换能很好地反映信号的低频分量，在这里主要是指未加噪声下的机电信号分量。由于加噪声的机电信号和不加噪声的机电信号，其主要区别在于高频噪声分量，所以环境背景音乐小波变换后的波开和语音、音乐变换后的波形在小尺度小波变换表现出来明显差别。

小波系数除了具有通过频率子带含有的能量把语音、音乐和一些环境背景音区分开来的特性外，小波系数提取还可以基于压缩域直接完成，从而加快音频处理，提高效率。

3.4 音频例子特征提取

通过对音频时域，频域特征的分析，发现提取音频特征时，都是将很长的音频信号先处理成(迭加)短时音频帧，然后在短时音频帧上提取时域、频域和时频等特征。这是因为按照语音处理理论，音频信号是短时平稳的，而长时间上是剧烈变化的，所以在很短时间的音频帧上提取特征(短时音频帧一般为 4 微秒左右，相邻帧之间的迭加为 2-3 微秒)，能够使提取出来的音频特征保持稳定。

但是，也可以从长时间音频信号 中提取特征，如从音频例子中提取特征(音

频例子的时间长度一般是 1-4 秒)。这种方式考虑的是任何语义都有时间延续性(如枪声会持续 3 秒等)，长时间刻度可以更好反映语义，从持续较长时间的音频例子中得到的特征就叫做音频例子特征。

在实验中，为了使提取的特征既音频信号短时平稳的特性，又满足音频信号有语义持续的性质，本文对音频信号特征提取采用从短时音频帧特征的统计值得到音频信号特征。

音频例子特征提取方式如下：将几秒钟(一般是 1-4 秒)的音频例子分成含迭加的短时音频帧，提取每个短时音频帧特征，形成帧特征向量，然后把短时音频帧特征向量的统计值作为音频例子特征。

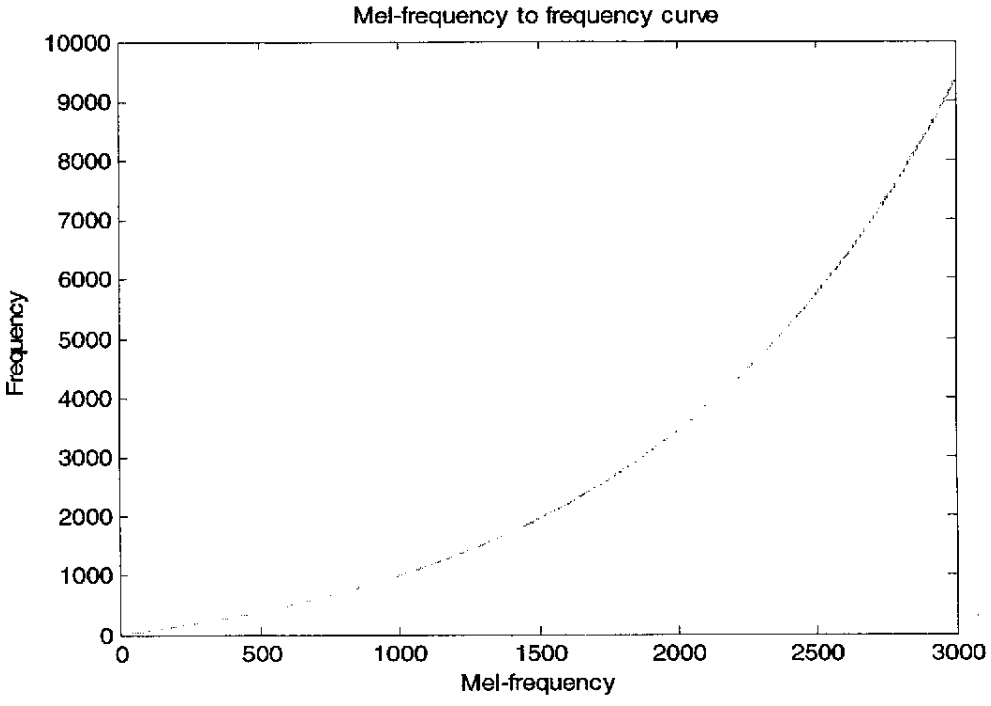
假设某个音频例子被分成 M 个迭加短时音频帧，从每个短时音频帧中提取短时能量，过零率和 Mel 系数特征，那么得到 3*M 特征矩阵。这个矩阵每一行分别表示短时能量、过零率和 Mel 系数三个特征，矩阵的每一列分别表示每一短时帧。对矩阵每一行数据分别取均值，就得到 短时能量标准方差、过零率标准方差和 Mel 系数标准方差。

这样，就可以使用短时能量均值和标准方差、过零率均值和标准方差，以及 Mel 系数均值和标准方差作为音频例子特征。我们主要以均值 Mel 倒谱系数作为音频例子的特征。

3.5 实验结果与分析

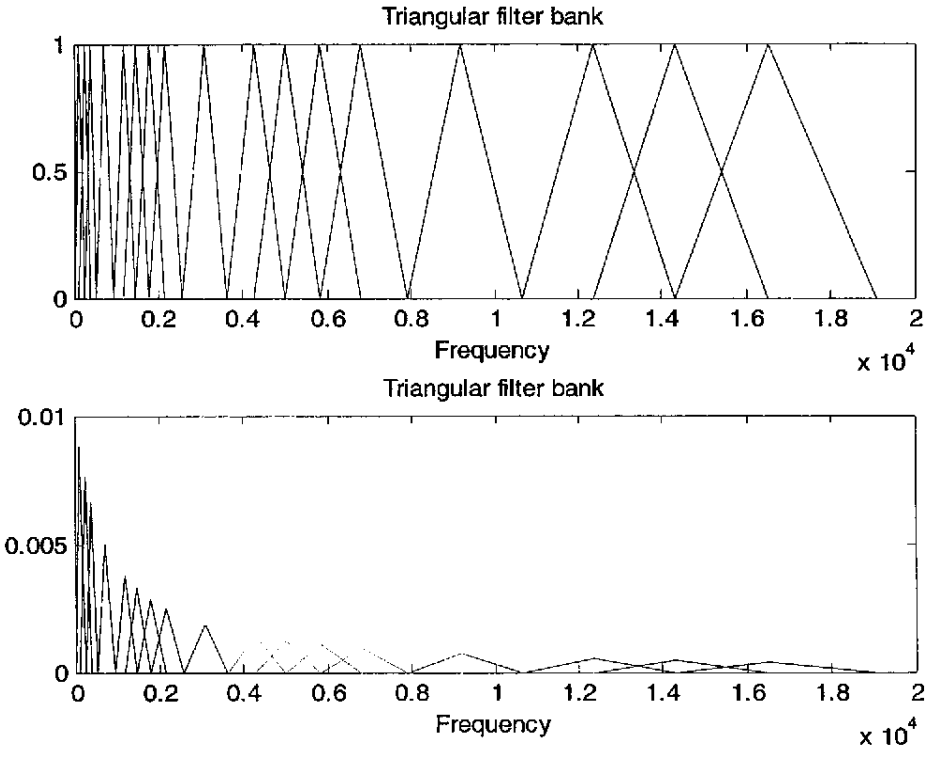
通过对音频信号短时能量、短时过零率、MEL 倒谱系数进行大量实验，分析是否可以用来区分音频信号，作为音频信号的特征进行音频识别，实验结果表明，本文算法能够作为音频信号特征，对音频信号进行识别。下面详细介绍:

Mel 倒谱系数（MFCC）是建立在傅立叶和倒谱分析基础上的，对短时音频帧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。整个频率划分成若干个子带，每个子带的宽度可以取为 $Mef(f)=2595\log_{10}(1+\frac{f}{700})$。MEL 频率与一般频率关系如图(3-5)。

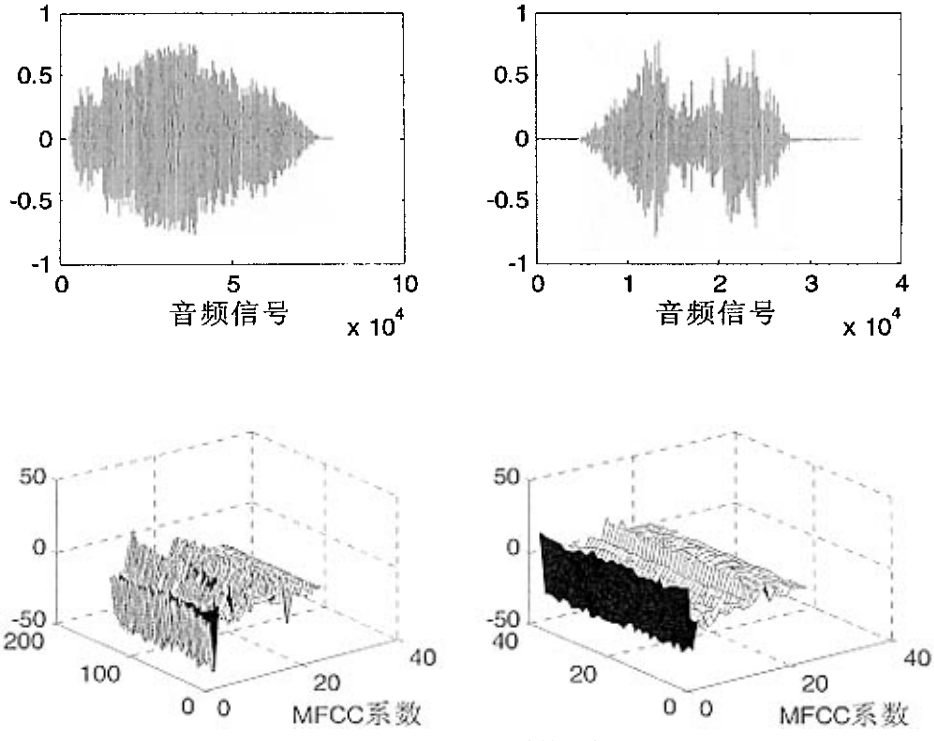


图(3-5) MEL 频率与一般频率关系

如果整个频率带被划分为 K 个子带，分别计算这 K 个子带上的能量，就构成了这个短时音频帧的 K 个 Mel 系数。对提取出来的 Mel 系数计算其对应的倒谱系数就是 Mel 倒谱系数。MFCC 是从 N 个短时音频帧中提取出来的，主要反映的是音频在很短时刻内的静态特征。本实验采用三角带通滤波器，主要是对频谱进行平滑化，并消除谐波作用，凸显音频信号的共振峰。子带化分示例见图（3-6)。实验过程中，对不同的音频信号进行 MFCC 分析，信号采样频率为 44100Hz，整个频带根据 Mel 算法进行子带化分。依次从 8 阶，12 阶，16 阶，24 阶进行实验，发现 24 阶时能够较好的反应信号特征。如图(3-7)所示。

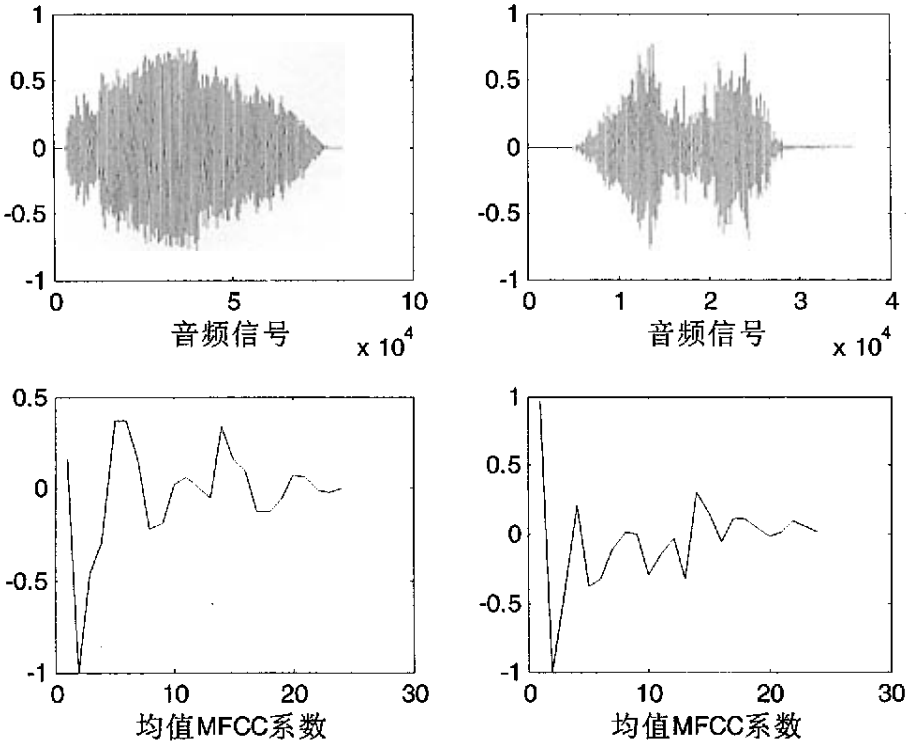


图(3-6) 三角子带化分图示



图(3-7) MFCC 系数分析

由于短时帧的分割数量是巨大的，每个短时帧提取 24 个 Mel 系数，其运算量巨大，如果声音种类较多，音频库数据太大，识别时间难以满足实际的应用要求。因此简单地进行 MFCC 提取不适于音频识别算法。为减少特征数量，采用对所有的短时帧 Mel 系数进行均值处理的方法。每个音频信号得到 24 个均值系数，将均值系数作为识别特征，每种声音有 24 个特征点，极大地减少了运算量，运算速度提高显著，经过实验识别，同样能够达到 100%识别单一的声音。图（3-8）是对两种声音均值系数比较。



图(3-8) 均值 MFCC 系数分析

3.6 本章小结

本章首先介绍了音频信号的特征参数，然后详细介绍了音频信号的时域特征，包括短时平均能量、短时过零率；音频信号频域特征算法原理和应用。对短时平均能量、短时过零率进行研究，可能根据这两种特征对音频信号进行粗分类。着重研究了音频信号的频域特征，提出了一种均值 MFCC 系数算法，对单一音频信号进行分析时，可以将均值 MFCC 系数作为音频信号的特征。

第四章 音频分割与识别

4.1 引言

音频是连续的时间序列信号，犹如不可能对几十分钟或几十个小时视频一起处理一样，也不可能对持续时间很长的音频处理，所以需要对连续的音频流首先进行分割。将连续音频信号流分割成长短不一的音频单元后，需要对每个音频单元进行识别，将它们归属为不同的音频类别，如语音、音乐和环境背景音等。

在视频流中，为了将视频流分成不同的镜头单元，需要寻找纹理、颜色和运动等视觉特征突变的地方，视觉特征突变就是镜头发生了转换(shot detection)，意味着视频内容从一组镜头转换到另外一组镜头，而特征发生突变的地方就叫做镜头边界(shot boundary)。得到镜头边界后，可以将每组镜头的第一个图片帧或最后一个图片帧，作为这组镜头的关键帧，用来表示这组镜头，然后继续对视频进行处理。

音频与视频一样，当从一种类型的音频信号转换到另外一种类型的音频信号时，某些听觉特征会发生变换，前后差别较大(如从语音部分转换到音乐部分，音频的 MFCC 特征会发生明显变化)，所以，也需要在音频特征发生突变的地方对连续的音频流进行切分，把连续的音频流变成不同时间长短的音频例子(Audio Clip)，然后对这些得到的音频例子进行识别，把相似的例子归属到同一类别。

在视频中，是比较相邻两个或几个图片帧之间的特征差异，来判断是否发生了镜头转换(把一秒钟所处理的图像帧总数称为视频采样率)。在音频中，也有“帧”这个处理单元概念，只不过是前面介绍的“短时音频帧”。在音频分割中，每次是比较相信两个或几个短时音频帧特征，寻找特征发生突变的地方，然后在特征突变地方对连续音频进行切分(一段音频要提取多少短时音频帧与音频采样率有关)。

下面主要介绍两种音频信号流的分割算法：(1)音频分层分割。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，这就是音频分层分割算法。(2)音频模板分割算法。前一种音频分割算法是基于一个给定的阈值的，也就是说，通过判断相信音频短时帧或连续几个短时帧特征之间的差是否超过了实

现给定阈值，如果超过了实现给定阈值，就对音频进行分割。如果依靠阈值对音频进行分割，则阈值选取是否适当会极大地影响分割结果。一般而言，一个对任何音频信号流均合适的分割阈值基本不可能得到。为了避免对阈值的极大依赖，基于模板分割算法是为一段音频流建立一个模板，使用这个模板去模拟音频信号流的时序变化，达到音频数据流分割目的[15]。

对分割出来的音频例子进行分类属于模式识别的问题，其任务是通过相似度匹配算法将相似物体归属到一类。在图像和视频相似匹配中，介绍了很多匹配算法，如欧拉距离等。

应该指出的是，音频分割和识别算法其实都是分类过程，分割与识别都是把音频流分类成事先定义好的类别。它们的区别是：音频分割只是粗分过程，如把音频流分割分类成音乐、语音和静音等片段(音频例子)；而音频识别却是对音频进行细分，如把音频分类成“掌声”和“爆炸”等。

从基于内容的音频检索角度讲，音频分割是给每个音频例子标注上了低级语义，而音频识别则是给每个音频例子标注上了中级语义。

4.2 音频分割算法

不同类别音频信号的听觉特征之间存在不同程度差异。比如，静音信号所带的能量就比非静音信号要小的多；语音信号的 Mel 系数与音乐信号的 Mel 系数差别较大。

在连续音频信号流中，当从一类音频信号转换到另外一类音频信号时，这两类音频信号在某些相应音频特征上会发生明显突变，音频分割的思想就是要利用连续音频信号流在发生转变时，听觉特征之间存在差异的现象，把变化出现的地方作为分割点，将音频流切分开，从而将连续音频信号分割成长短不一的音频例子，再进行后续处理。

当音频信号在不同种类音频之间进行转换时，不同特征之间所存在的差异是很不同的。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_j这个音频类别转换到c_k这个音频类别时，特征f_j和f_k之间的差异较大，特征f_i和f_j之间的差别不很明显。所以只需要比较f_j和f_k之间的差异是否较大，不需要比较其他听觉特征之间的差异，如果特征f_j和f_k之间的差异

较大，则对音频流进行切分。

这样，在进行音频分割时，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，这是分层分割音频信号流的基本思想。当然，在实际中，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进来就比较明显了，所以，在分层分割中，也常常使用组合音频特征的分割。

在特征差异比较时，所提取的听觉特征 f_i、f_j和 f_k 往往是前面介绍的音频时域、频域、短时帧和音频例子特征。

在音频分割处理过程中，对不同的音频特征(或组合音频特征)，需要预先分别确定不同特征之间变化的阈值，这样，可以根据确定的阈值，去判断音频数据流之间是否发生了变化。不同特征之间阈值的选取比较困难，特别是一个稳定普遍阈值的获取基本是不可能的。当一个分割阈值在某些应用表现良好，如果使用到另外些应用中去时，往往会产生不理想的结果，使这种基于阈值的音频分割算法不鲁棒。

由于音频是时序数据，为了到达不使用阈值目的，可以训练一个模型去模拟某类音频动态变化，然后根据这个模型的动态变化，达到音频数据流自动分割的目的，这是采用模板对音频信号进行分割的思路。

在基于模板的音频信号流分割方法中，目前比较成功的是应用训练好的隐马尔可夫链，通过 Viterbi 算法求出最佳状态序列，然后实现对不同话者的语音分割。这种方法不需要任何阈值，自动实现音频信号流分割。但是，这种方法也有本身的局限性，就是需要对分割模板进行反复训练。

4.2.1 音频分层分割

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例子。

[10]从每个短时音频帧中提取短时平均能量、过零率协方差、基本频率能量比和过零率周期率四个特征，然后比较前后相邻若干个短时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，分别得到静音(Silence)、对话(Dialog)、和谐背景音乐(Harmonious Music)和环境背景音(Environmental Sound)等音频例子。

本文分别对静音，语音，和谐音乐的短时能量，短时过零率和过零率协方差进行实验，可以发现可以通过短时能量和短时过零率对连续的音频信号进行分层

分割。下面详细说明实验情况。

在前面音频特征提取中介绍过，短时平均能量指在一个短时音频帧音频内信号所聚集的平均能量，可以用短时音频帧的平均能量来把静音与对话、和谐背景音乐和环境背景音音频区别出来，因为相比较，静音的短时平均能量基本为零。短时平均能量的计算在(4-1)给出：

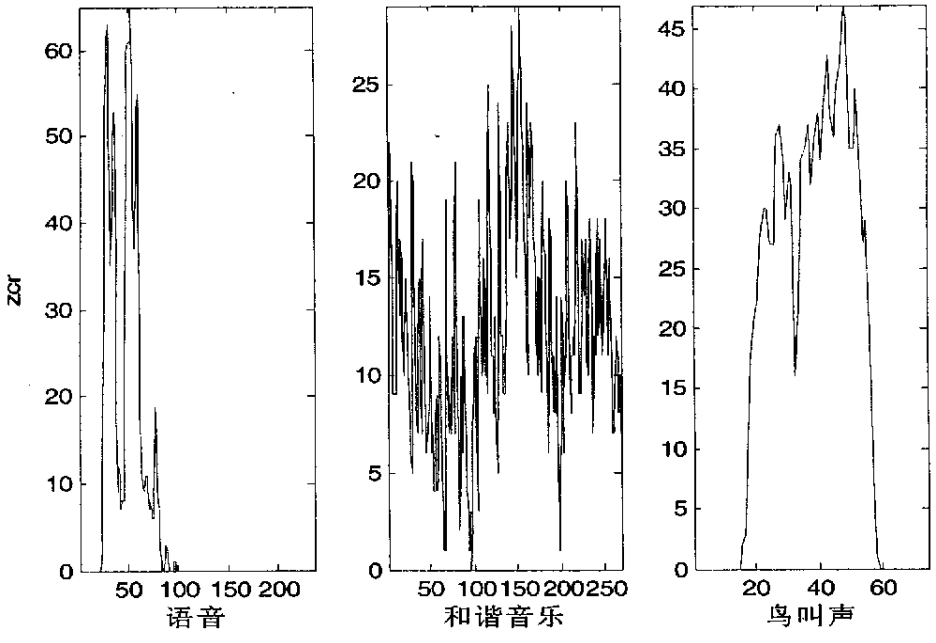
$$E_n = \frac{1}{N}\sum_m [x(m)w(n-m)]^2 \tag{4-1}$$

其中，$x(m)$表示音频信号的离散采样时间，n 表示短时平均能量特征的时间索引，$w(n)$是长度为 N 的 hamming 窗口函数，当$0 \le n \le N-1$时，$w(n)=1$；否则$w(n)=0$。

过零率指音频信号在一个短时间音频帧内能通过零点的次数，其计算由(4-2)给出：

$$Z_m = \frac{1}{2}\sum_m |sign[x(n)] - sign[x(n-1)]|w(n-m) \tag{4-2}$$

其中，当$x(n) \ge 0$时，$sign[x(n)]=1$，否则$sign[x(n)]=0$。$w(n)$是长度为 N 的 hamming 窗口函数，当$0 \le n \le N-1$时，$w(n)=1$；否则$w(n)=0$。



图(4-1) 不同音频的过零率特征比较

过零率协方差计算由(4-3)式给出，其中x_i表示每个时间窗口的过零率，u表示所有时间窗口过零率的数学期望：

$$Cov = \sum_{i=1}^{|Z_m|} E(x_i - u)^2 \tag{4-3}$$

和谐音乐信号的频率始终稳定在一定频率范围，过零率变化缓慢，所以过零率协方差很低[16]。图(4-1)给出了含连续若干个短时帧和谐音乐、语音和背景钟声的过零率特征，横坐标为窗口时间，纵坐标为过零率。从图(4-1)的纵坐标知道，和谐音乐的过零率很低。因此，可以根据Cov值来判断音频信号是否为和谐背景音乐：如果Cov很低，则为和谐背景音乐。

由于对话信号能量主要集中在$200Hz \sim 3.4kHz$之间，其它音频信号的频率分布比较广或集中在高频部分。在分割出会话音频片段时，把频率分为基本频率BF和非基本频率两种。基本频率指$0 \sim 1.5kHz$范围内的频率，其余为非基本频率。定义基本频率能量比率为基本频率所带的能量占整个频率所带能量的比值，其余为非基本频率。定义基本频率能量比率为基本频率所带的能量占整个频率所带能量的比值，其计算由(3-4)给出：

$$BF\_Ratio = \frac{\sum_{j=1}^{T}\sum_{k=1}^{1500} x_j(k)^2}{\sum_{j=1}^{T}\sum_{i=1}^{n} x_j(i)^2} \tag{4-4}$$

式中T表示信号的时间长度，n表示频谱分析时得到的信号频率范围，$x_j(k)$表示在时间j频率范围在基本频率以内的信号所带能量，$x_j(i)$表示在时间j所有频率带的能量。

另外，对话音频流中每个单词开始和结束时的平均过零率显著提高，而单词发音过程中平均过零率基本保持稳定。所以含有多个单词的对话信号过零率变化很大，其过零率协方差也就大。如果一串音频信号流的过零率协方差和基本频率能量同时很高，就可以判断这段音频信号为对话。

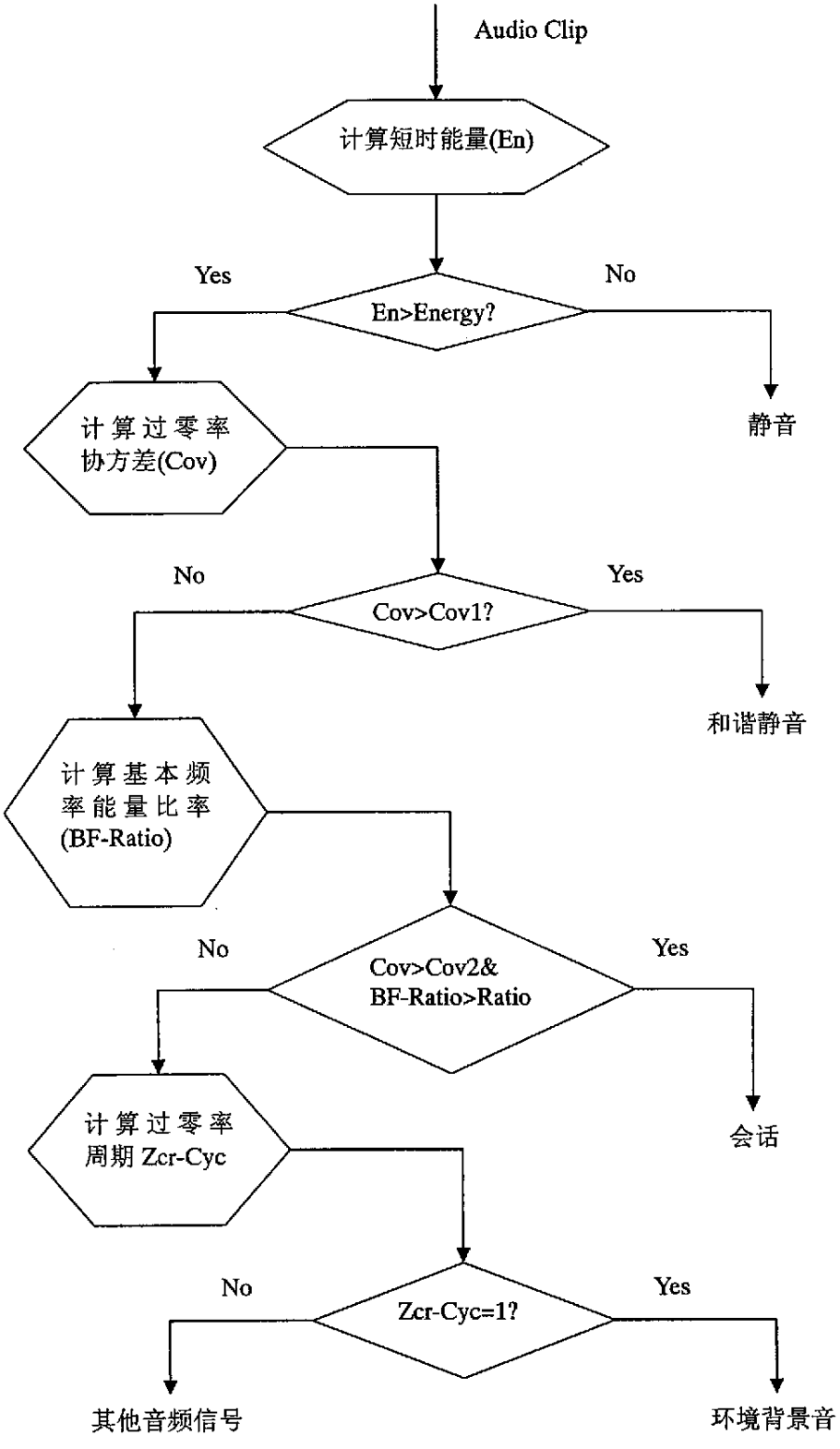


图 4-2 音频数据流分层分割

对于连续的枪声，钟声和瀑布声等环境背景音乐，由于其具有周期性很强的特性。如在 3 秒连续的枪声中，每隔 10 微秒会有次枪击声，然后是 2 微秒的停顿，其过零率表现出周期性规律。过零率周期计算如下：

如果过零率特征存在周期性，$Zcr\_Cyc=1$；否则$Zcr_Cyc=0$。

4.2.2 实验结果与分析

根据上节的研究分析，提取如上音频特征，就可以按照分层原则分割识别出音频信号，具体步骤如图(4-2)所示。其中，Energy、Cov1、Cov2、Ratio 为实验中得到的判断阈值。

为了测试上面分层分割算法效果，在 Matlab 下实现了基于层次的音频模块，这个模块完成音频粗分归类(模块每次比较相邻 10 个音频短时帧之间的特征)，表 4-1 给出了分层次粗分的结果，实验中静音与和谐音乐数据都不存在其他杂音，识别效果理想。但是，对话和环境背景音中存在其他杂质音，所以识别正确率稍微低。

表 4-1 基于分层的音频数据识别结果

粗分类	静音	谐音	对话	环境音	其他
静音	3				
谐音	8				
对话			12	2	
环境音	25	2			3
其他					7

应该指出，上面音频分层分割算法只是完成了对音频的粗分，但是这种算法不能对音频进行进一步细分。比如，在粗分过程中只能够把枪声、钟声、鸟叫声等笼统归结为环境背景音，不能识别出到底是哪一种环境背景音；还有，虽然这种方法可以把语音和音乐分割开来，但是到底是谁在说话是识别不出来的。

4.2.3 基于模板的音频分割

在分层音频分割中，为了将连续音频流分割开，需要使用一个事先得到的阈值去判断相邻若干音频短时帧之间的特征是否发生了突变。

对于不同的音频数据流，判断阈值不相同，所以每个音频样本，需要带我选定不同的分割阈值，对音频分割造成了极大不方便。因此，人们很自然想到是否存在一种自动音频分割模型，使音频分割避免使用阈值。

下面介绍，如果一段语音是由几个不同人的语音构成，那么如何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。

音频信号本质上时序变化数据，对于同类音频信号(如某人的语音)，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，则把它与其他类别音频分割开来，这个用来进行模拟的动态过程就叫做语音模板(template).

对于若干个人，提取其语音特征，然后为他们分别模板，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。

[17]使用隐马尔可夫链模拟不同人的语音，对于一段语音，使用隐马尔可夫链将这段语音信号分割开来，使不同的分割单元只包含同一话者语音，其目的在于使用话者身分对音频例子进行标注。由于语音不是研究的主要方面，这里只简略介绍隐马尔可夫链模型在语音音频例子自动分割中的应用。

如果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，也就是说，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那个话者，就是计算o属于哪个λ_i的概率大，然后把概率最大所归属的λ_i作为识别结果。

比如，对于一个两类分类问题，λ_i和λ_j分别表示训练好的隐马尔可夫链参数模板，对应鸟声和钟声两类音频。现在有一个时间特征序列o，要判断o属于哪类声音，就是计算o分别对应λ_i和λ_j的概率，如果o对应λ_i的概率值最大，则判断o为鸟声，反之为钟声。同时，也可以设定一个阈值，使低于这个阈值的时间特征序列不属于任何一类信号。

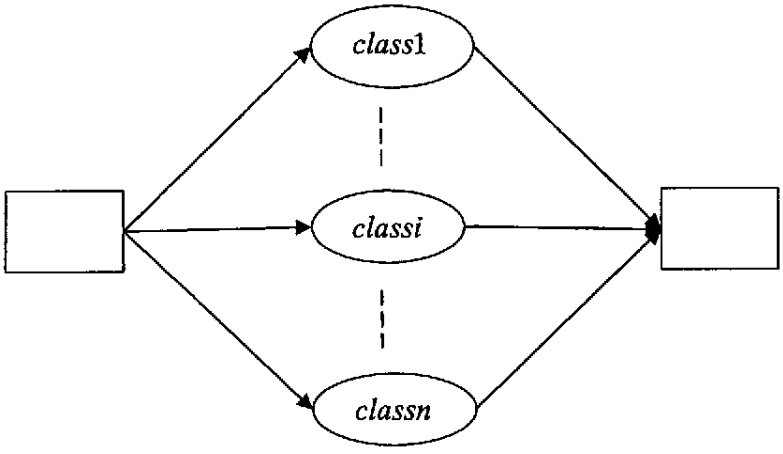
对于隐马尔可夫链λ_i和观测向量o，有三个任务要完成：(1)采用 Bauml-welch 算法训练计算出识别模型的最佳参数；(2)采用前向(Forward)算法计算$P(o|\lambda_i)$的概率值；(3)应用 Viterbi 算法求出观测向量o所对应的最佳状态序列。

在[18]中，把一段音频流中不同话者语音自动分割开来应用了上面第三个任务的输出结果，这样，在分割中，只需要训练一个隐马尔可夫链，而不是n隐马尔可夫链。

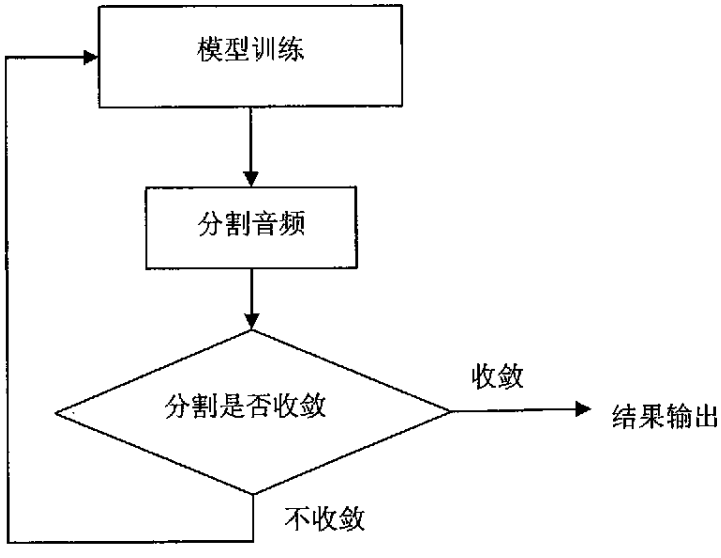
其具体算法如下：采集n个人讲话的样本，训练一个带有n个状态的隐马尔可

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，交语音信号从转换处分割开来。最后，连续语音音频自动被分割成了若干音频单元，每个单元包含且只包含一个话者语音。

图(4-3)显示了把语音信号流切分成不同讲话者音频单元的隐马尔可夫链模板结构，其中图中的每个$class_i$为隐马尔可夫链中的一个状态，其代表一个话者身分，状态和状态之间通过转移概率矩阵进行连接(转移概率矩阵值通过训练得到)。



分割结构图



分割流程图

图(4-3) 基于隐马尔可夫链模板的话者音频分割

在实际切分时，为了防止一次切分结果不精确，所以对同一样本，进行连续切分，比较前后切分结果是否存在很大差异。即，如果前次切分出来的最佳状态序列与目前得到的最佳状态序列基本一致，则表示切分是稳定的，把切分的最佳状态序列作为结果输出。相同状态序列对应的音频单元就是同一话者语音。

基于模板的语音数据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，最后得到隐马尔可夫链模板，指导分割完成。

在训练过程中，从语音数据流中所提取的特征可以是前面介绍的时域、频域、时频域等任何短时帧特征。但是，由于 MFCC 系数在语音分析中获得了成功，最好使用 MFCC 特征进行切分。

4.3 音频例子识别模型

4.3.1 音频例子识别分析

前面介绍了在连续音频数据流的特征发生突变时，将其切分成若干长短不一音频单元的方法。这些长短不一的音频单元好比视频中最小物理单元“镜头”，从现在起，将使用“音频例子(Audio Clip)”来指代这些长短不一的音频单元。

音频例子与视频镜头存在差异，也有相似地方，它们相似之处在于：(1)音频例子和视频镜头均是最小物理单元，它们都表示了一定低级语义，如视频每个镜头表示了单一场景，音频每个音频例子表示了语音和音乐等；(2)在这些物理单元中，其对应多媒体特征基本保持稳定，如一组视频镜头中，其每个图像帧之间的左别很小。在分割得到的音频例子中，其每个短时音频帧之间的听觉特征相差很小。

视频镜头和音频例子之间也存在相同之处[19]：(1)视频镜头和音频例子的表示方式不同。在视频镜头中，使用关键帧来表示一组视频镜头，为了简单，很多时候就用每个镜头中第一个和最后一个视频图像帧来代表整个镜头；在音频例子中，有两种方法去表示音频例子，一是从每个短时音频帧提取的特征向量，用这个 所有短时音频帧向量表示这个音频例子。也就是说，如果这个音频例子里包含 *nFrame* 个短时音频帧，每个短时音频帧提取 *nFeature* 个特征，那么使用 $nFrame \times nFeature$ 这样的向量来表示这个音频例子；另外就是使用聚类算法得到向量空间 $nFrame \times nFeature$ 的聚类质心，使用聚类质心来表示这个音频例子。无论哪种表示方法，在音频例子中，不存在“关键音频帧”的概念。(2)音频例子和视频镜头试题方法的不同。因为在视频镜头中，只有若干个关键视频帧，为了比较两组镜头之间的相似度，只需要比较这些关键帧相似度就可以了，所以经常采用的相似度匹配算法是欧拉距离等几何相似度匹配算法。而在音频例子中，由于每个音频例子所包含的短时音频帧是巨大的，提取的特征相当多，依次比较每个短时音频帧之间的相似性不太合适。所以在音频例子相似度比较多，更多是为相似音频例子建立模板，用一个模板去模拟音频例子数据之间的动态变换。通过比

较模板相似性，达到相似音频例子寻找目的。需要指出的是，视频和音频相似性比较，都不能完全达到主观相似性，所以“视频和音频相关反馈技术”被采用，来弥补主观相似性与几何/时间相似性之间的差距。

本节主要是研究音频例子之间相似度匹配模型的，并且重点研究如何比较音频例子动态相似性。对于两个音频例子$audioClip_i$和$audioClip_j$，假设从每个短时帧中提取$nFeature$个特征。由于每个音频例子持续时间长短不一，如果$audioClip_i$和$audioClip_j$分别包含$nFrame_i$和$nFrame_j$个短时帧，下面主要研究如何判断$nFrame_i \times nFeature$与$nFrame_j \times nFeature$是否相似，就完成了音频例子$audioClip_i$和$audioClip_j$之间相似性比较。

通过音频例子相似性比较，就将每个音频例子归类到了不同音频类别，也就完成了音频例子的识别。

对于分割出来的音频例子，总是想识别它。如，对于环境背景音，想识别它为鸟声还是钟声；而对于音乐，想判断其是哪种乐器了出的音乐；对于对话，想分别出是男的声音，还是女的声音，或者是某个人的语音。从这个角度看，识别就变成了分类，就是将未知数据归属到某一类，使同属于某一类的数据之间存在“相似”之处。

如果说音频分割是对音频数据流的粗分，则音频识别分类是对音频例子进行中级语义标注的过程，能够为多媒体信息查找提供方便。

要对未知数据进行识别，需要先有一个学习的过程。通过学习，构造一个分类器，然后利用这个分类器对未知数据进行识别。因此，基于数据统计规律的机器学习是现代智能识别技术中的一个重要方面，从观测数据(样本)出发寻找规律，利用这些规律对未来数据或无法观测的数据进行识别预测，是包括模式识别和神经网络等学科重要的研究思想。

可以这样描述一类问题的学习分类过程：为了识别 A 类带我，先把自己认为属于 A 类的数据收集进来(这个所收集的数据称为训练样本)，提取每个训练样本的特征，然后使用所提取的特征去训练一个分类器，使这个分类器不公把训练样本中所有属于 A 的数据识别出来，而且对于不属于训练样本的，但是的确属于 A 的未知实际数据也识别出来。

在这个训练过程中，有如下几个步骤需要提醒注意：

一是训练样本的形成。为了训练识别某类物体的分类器，就要选择属于这个类别的数据。在选择数据时，哪些数据是“正样本”，哪些数据是“负样本”必须由自己定义，这个定义的过程叫做“标注”。所谓负样本，指那些不属于某类物体

的数据：如果训练一个识别人脸的分类器，那么人脸图像是这个分类器的正样本，而“树”、“汽车”等图像就是这个分类器的负样本；如果样本选取不好，导致样本不具有普遍代表性，那么训练得到的分类器的识别效果就不会很好。

二是对样本需要提取什么特征。由于特征是用来数据的，在分类器训练与识别过程中，不是使用原始样本数据，而是使用所得到的特征向量，所以良好的特征对识别正确率的提高起到了很重要作用。在这里，假定已经获取了足够好的音频特征，主要关注如何根据这些特征去训练一个良好的分类学习机。

三是选择怎样的学习过程。这主要是如何选择一个良好的学习分类模型，使这个模型不公对训练样本能够取得良好的分类能力，在实际中，对未知数据也能够取得良好识别正确率。

现在考虑，基于训练好的模型，如何实现两类音频数据识别：假设C_1和C_2表示两个通过训练音频数据样本得到的模板，代表两类不同的音频例子。现在要使用C_1和C_2对分割出来的未知音频例子进行识别分类，将所有属于C_1的音频例子归属为一类，属于C_2的音频例子归属为另外一类。令X代表从任一未知音频例子中所提取的音频特征向量(因此X可以用来表征这个未知音频数据)，由贝叶斯理论知道，判断X属于C_1，还是属于C_2的过程，就是判断X对应于哪个模板类别的后验概率最大，即计算$\arg\max_i P(C_i \mid X)$。

对于模板$C_I(1 \le i \le 2)$，其后验概率$P(C_i \mid X)$如下计算：

$$P(C_I \mid X) = \frac{P(C_i)P(X \mid C_i)}{P(X)} \qquad (4\text{-}5)$$

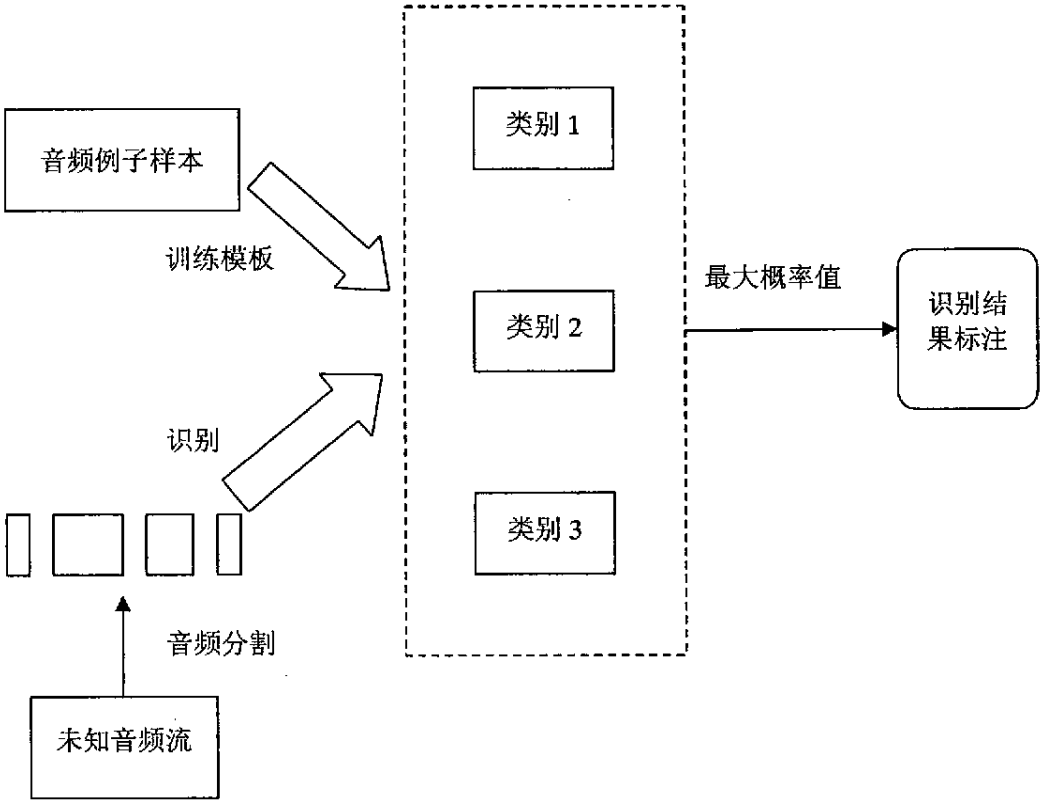
在计算(4-5)中的后验概率时，一般假定不同类别C_i出现的概率$P(C_I)$相等，而对于每个类别而言，$P(X)$总是相等的。所以，后验概率$P(C_i \mid X)$就约等于$P(X \mid C_i)$，这样，对后验概率$P(C_i \mid X)$的计算就直接转换成先验概率$P(X \mid C_i)$的计算。也就是去计算，对于每个训练得到的分类器C_i，X属于它的概率$P(X \mid C_i)$是多少。

于是，对分割出来的音频例子进行识别分类时，先训练一些模板C_i去模拟和代表某类音频(如爆炸、鼓掌声和音乐等)，然后对于未知数据X，计算哪个C_i所对应的$P(X \mid C_i)$最大，最后把未知数据X判定属于$P(X \mid C_i)$值最大时所对应的类别C_i。

训练类别模板C_i和计算$P(X \mid C_i)$的方法可以通过不同机制完成。如隐马尔可夫链模型(Hidden Markov Model，HMM)，支持向量机模型(Support Vector Machine，SVM)等。应该指出，这些分类器的理论基础都来自数据统计理论，而且这些分类

器不仅仅局限于音频信号的识别，它们同样应用于图像、视频和文本识别等领域。

图(4-4)显示了音频例子的识别过程：先是选择训练样本，训练生成表示某类音频的模板；对于未知数据，提取音频特征后，使用前面介绍的方法完成训练音频流的分割，得到一个个长短不一的音频例子；然后使用每个训练得到的模板对分割出来的音频例子分别进行识别，将这些音频例子归属到概率值最大的音频类别中，完成对音频例子识别与标注。



图(4-4) 音频例子识别过程

4.3.2　动态时间规整算法

基于动态时间归整匹配的 DTW 算法从目前来看，可能是一个最为小巧的音频识别的算法。其系统开销小，识别速度快，在对付少量单一音频控制系统中是一个非常有效的算法。主要理论简单说明如下：

假设存储的一个模板包括 M 帧倒谱特征 $R=\{r(m);m=1,2...,M\}$；识别特征序列包括 N 帧倒谱特征 $T=\{t(n);n=1,2,...,N\}$。在 $r(i)$ 和 $t(i)$ 之间定义帧局部失真 $D(i,j)$，$D(i,j)=2|r(i)-t(i)|$，通过动态规划过程，在搜索路径中找到累积失真最

小的路径，即最优的匹配结果。

对称形式 DTW：

$$S(i,j)=\min\begin{cases}S(i-1,j-2)+2D(i,j-1)+D(i,j)\\ S(i-1,j-1)+2D(i,j)\\ S(i-2,j-1)+2D(i-1,j)+D(i,j)\end{cases} \tag{4-6}$$

其中 $S(i,j)$ 是累积失真，$D(i,j)$ 是局部失真。当动态规划过程计算到固定结点 (N,M) 时，可以计算出该模板动态匹配的归一化距离，识别结果即该归一化距离最小的模板：$x=\arg\min\{S(N,M_X)\}$。

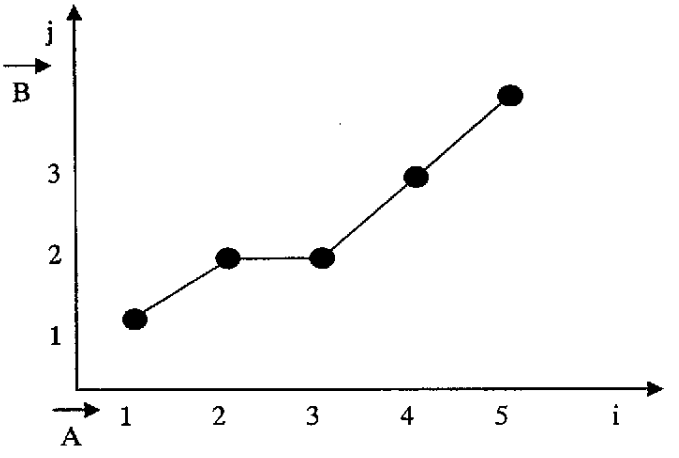
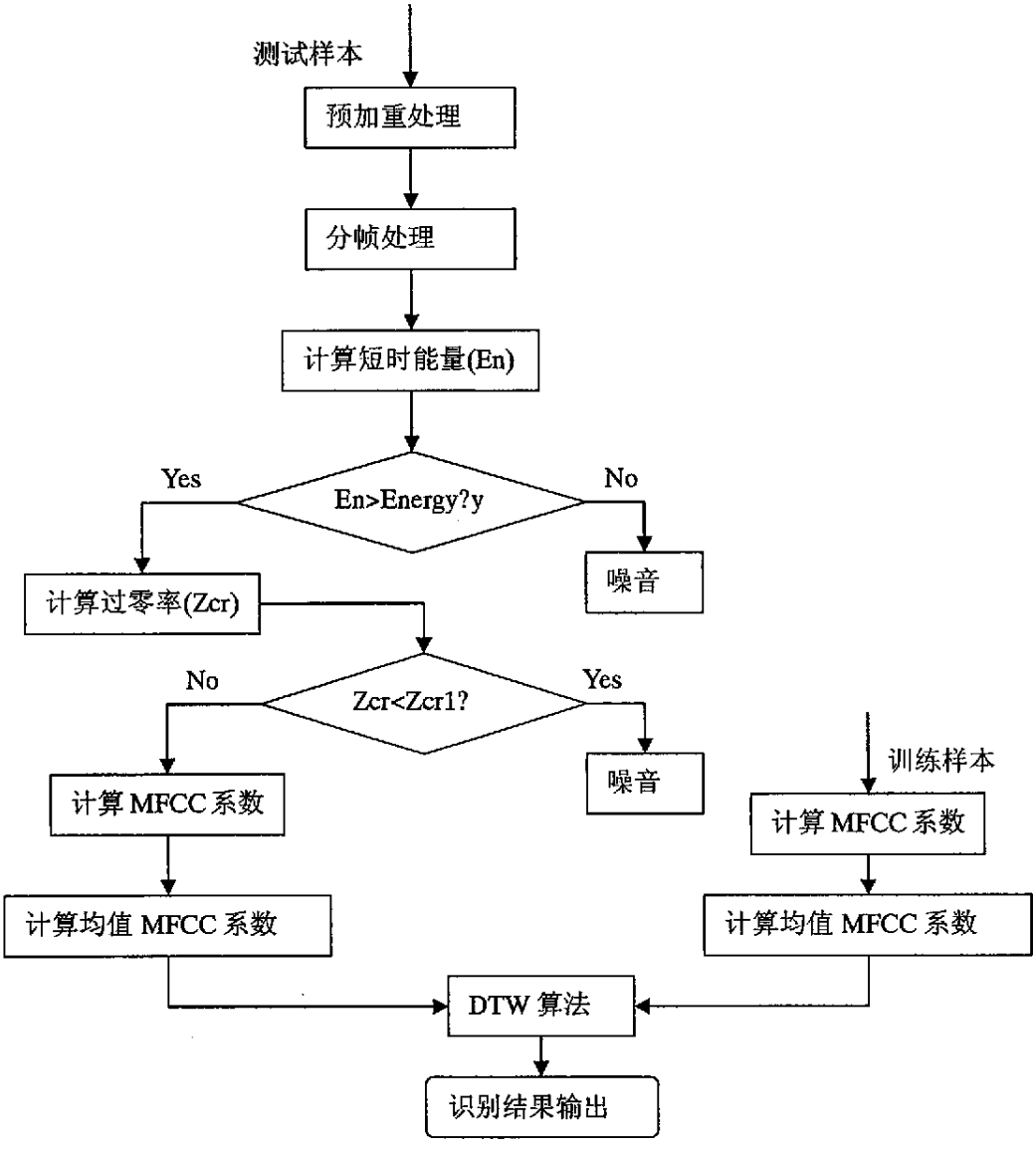


图 4-5 DTW 算法搜索路径

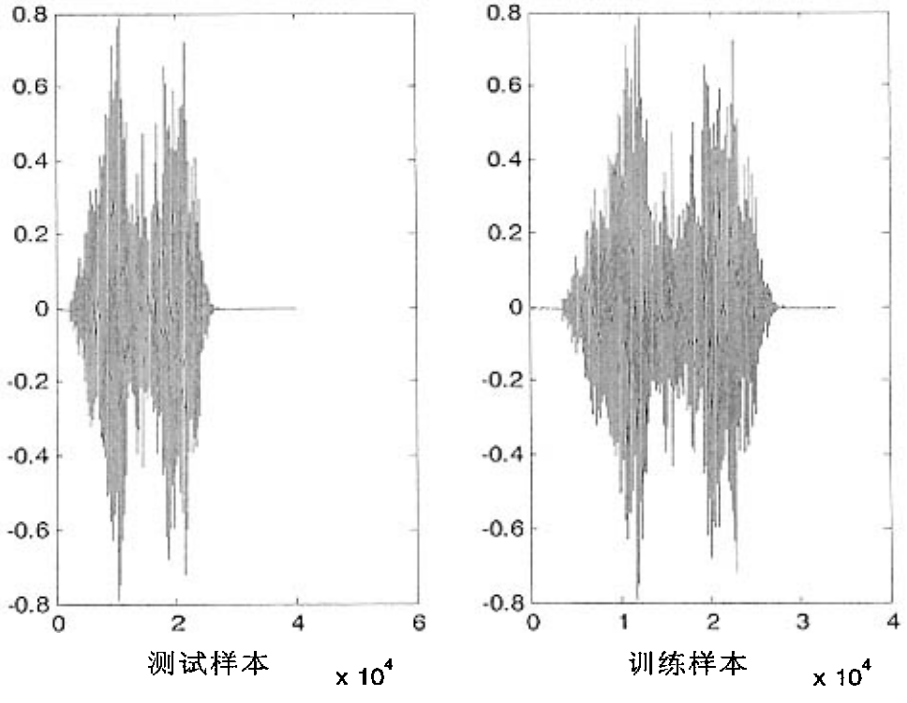
4.3.3 实验结果

根据第三章对音频信号 Mel 倒谱系数的分析，采用 DTW 算法对单一音频例子进行音频识别分析。样本库有 5 种音频信号，对待测样本进行样本匹配，匹配时，采用 DTW 搜索算法，计算测试声音和样本声音特征之间的最小距离，并确定适当的阈值，进行音频信号识别。识别流程图如图(4-6)所示。

图(4-7)为待测信号匹配结果显示。可以看出，通过 Mel 倒谱系数的分析，对单一的音频下例子进行识别，能够准确的识别出待测音频信号。


测试样本
预加重处理
分帧处理
计算短时能量(En)
Yes
En>Energy?y
No
噪音
计算过零率(Zcr)
No
Zcr<Zcr1?
Yes
噪音
计算 MFCC 系数
计算均值 MFCC 系数
训练样本
计算 MFCC 系数
计算均值 MFCC 系数
DTW 算法
识别结果输出

图(4-6) 程序流程图



图(4-7) 单一音频识别结果显示

4.4 本章小结

本章主要介绍了音频分割和识别算法。研究了音频分层分割，单一音频例子识别两个方面。结合上一章音频特征提取，对音频信号进行识别实验，验证了 Mel 倒谱系数作为单一音频信号特征进行音频识别的可行性。

第五章 基于内容的音频检索技术

5.1 引言

前面分别介绍了音频特征提取、连续音频流分割和音频例子识别等内容，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莫过于听觉内容相似的音频检索，如果要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。

这里面主要的一个原因在于，在音频处理时，直接将所提取的音频特征交给计算机处理，让计算机去分析与理解。但是，在人脑对音频信号处理时，耳涡接收到的音频信号(如平均能量、过零率和 Mel 系数等)要被 大脑皮层感知器官先处理，然后基于处理的结果去形成音频信号的理解，而就是形成高级语义。

可以看出，计算机自动对音频信号的处理，与人脑对音频信号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，还是向基于语义音频检索迈出了坚实的一步，因为，任何语义一定有存在形式。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，这也是为什么基于听觉音频检索技术逐渐被 研究的目的。

在基于内容的音频检索技术从 20 世纪 90 年代末兴起以前，一种对语音文本(Speech Document Retrieval)检索的技术已经存在。在这种方法中，先把记录在磁带、录音设备或视频流中的语音信号转录(Transcribe)成文本信号，如将广播新闻转录成文本，然后对文本信息进行关键字提取，进行与关键字相关联的语音文本查询。这种方法一般用于新闻资料、历史资料或教学语言的查询，因为这种方法本质上是对语音信号进行识别，它不涉及去研究如何有效识别出爆炸和掌声等非语音信号，然后莫过于这些非语音信号进行文本检索。由于这种方法前端采取的是语音识别技术，后端对转录生成的文字进行检索，与基于听觉内容的音频检索关系不

大，不做研究。

目前在 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。这种寻找相似旋律和相似风格音乐的方式在网上购买音乐方面用途较大，譬如，人们并不知道某首歌曲的名字和主唱，但是对某些歌曲的旋律和风格非常熟悉，于是人们可以通过嘴巴将他熟悉的旋律“哼”出来。这些旋律通过麦克风数字化输入给计算机，计算机就可以使用搜索引擎去寻找一些歌曲，使反馈给用户的歌曲中包含用户所“哼”的旋律或风格，这种方式称为使用“哼”进行音乐检索(Audio Retrieval by Humming)。

对于广播等音频数据[20]，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，比如，有些人只关心“新闻摘要”，那么他只需要听主持人主题新闻就可以了；有些从对新闻详细分析也感兴起，那么他需要听主持人主题新闻和新闻详细报告两部分；可以说，很少有人喜欢广告的，所以广告可以尽可能从新闻中去除。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，人们对相邻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，这就是相似音频例子的检索。

在对以上三个方面进行介绍开始前，考虑一个问题，在实现相似风格“歌曲”和相似音频例子检索的时候，人们所提交的检索信息会是什么。

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”中获取完事语义。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。在这里，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。

二是提交一个音频例子，提取这个音频例子的特征，按照前面介绍的音频例子识别方法判断这个音频例子属于哪一类，然后把识别出的这类所包含的若干样本按序返回给用户，这是基于例子的音频检索。

第三种是使用“哼(Humming)”作为输入。比如，用户自己哼一段想找寻的音

乐，然后基于用户的“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。这其实也是一种基于例子的音频检索方法，不过其例子是靠“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)。(图 5-1)

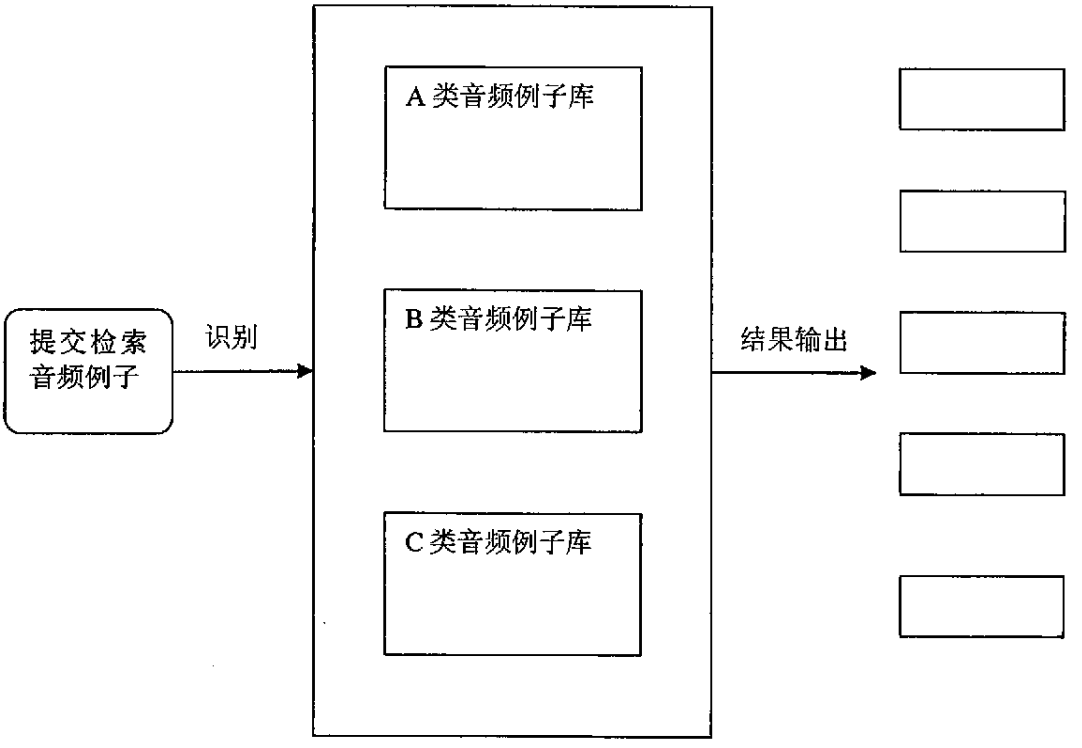


图 5-1 相似音频检索过程

基于例子的音频检索与视频检索一样，用构造的分类模型将用户提交检索的音频例子归属到某类音频，最后按照排序方法返回给用户属于这个音频类的若干相似音频例子。在这种方法中，提取音频例子特征、对音频例子进行判别和构造某类音频模板，都可以用前面介绍的技术完成，关键是如何管理和构造一个庞大的音频例子库。

另外，在相似音频例子检索过程中，应该给用户提供一种机制，让用户可以对反馈结果进行在线评估，然后将评估结果反馈给检索系统，让检索系统根据用户评估，重新进行检索，直到用户满意，这叫做“音频例子相关反馈”。

5.2 相似音频例子检索

这一节中主要介绍如何基于用户提交的查询音频例子，得到相似音频例子。基于例子的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。下面将简单介绍使用隐马乐可夫链对每类音频建立模板，实现相似音频检索。(2)不对每类音频建立模板，而是对每个音频例子建立模板。这里，所谓为每个音频例子建立模板，就是如何寻找一种良好的方式去表征音频例子。音频例子是用听觉特征表示的，在一百万音频例子特征表示时，从每个音频例子的短时音频帧中提取听觉特征，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，使用模糊算法得到每个音频例子的质心，使用质心表示每个音频例子，然后进行相似匹配。

5.2.1 基于分类模型的音频例子检索

图(5-1)给出了基于分类模型的音频例子检索过程：在这个过程中，首先将相似的音频例子组成一个个音频例子数据库，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 $auCorpus$ 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，首先从这个音频例子的短时(迭加)音频帧中提取特征向量 $X=\{x_1,x_2...,x_T\}$，然后采用前向(Forward)算法，计算每个隐马尔可夫链(听觉内容)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。

在基于模型的音频例子检索中，可以使用除隐马尔可夫链以外的其他任何分类模型[21]，如支持向量机也是可以的。

5.2.2 基于模糊音频例子检索

按照学习分类机制的不同，音频例子检索方式可以分为监督检索和非监督检索两类。所谓监督机制指，在生成音频类别模板中，要对每类音频例子进行标注，然后所有属于同一类别音频例子被用来训练表示某类音频的模板。在模板训练生成中，训练样本和训练策略的选择，将大大影响检索效率。如果训练样本不具有代表性或者重复训练导致训练模板过于复杂，将损害检索正确率。，当有新训练样本加入时，需要重新训练模板。

其实，音频例子检索主要要完成三个任务：(1)原始音频数据流如何表征；(2)基于这样的表征如何进行快速相似度匹配；(3)目前相似度匹配都是基于音频特征时间序列在几何意义上的相似性，非主观内容相似性，所以需要找到一种机制，弥补语义相似性与几何拓扑相似性之间的差距。

由于音频例子特征是从音频帧中得到的，如果每帧 11 提取个音频特征，对于采样率为 800Hz 的 3 秒长音频例子，其特征将为 60×11 矢量，导致原始音频数据流表示复杂。但是，如果将提取出来的特征进行聚类，直接使用聚类质心表示音频例子，将简化音频例子的表示形式，并且可以基于这种简化表示，快速试题音频例子之间的相似性。

下面，介绍如何使用模糊聚类方式对每个音频例子形成聚类质心，然后使用聚类质心表示每个音频例子，并且进行相似度匹配与音频例子相关反馈。

5.2.2.1 基于约束模糊聚类的音频检索与相关反馈理论

与上面基于分类模型进行音频例子检索方式不同，从分类学习机制上讲，基于模糊聚类的音频例子检索算法属于非监督检索：音频数据库中的每个音频例子自动提取其音频特征，然后对特征向量进行模糊聚类，用形成的聚类质心去表征数据库中的每个音频例子。对于用户提交要检索的音频例子，在得到其聚类质心后，将提交音频例子的聚类质心与数据库中所有音频例子聚类质心进行快速匹配，取最相似的几个音频例子反馈给用户。如果用户对检索结果不满意，还可以根据用户反馈，修改调整检索结果。

从检索过程可以看出，音频特征提取、模糊聚类质心形成。聚类质心快速匹配和音频例子相减反馈是约束模糊聚类算法的核心，在前面对音频特征提取已经介绍过，假设从每个音频例子的短时音频帧中提取了质心、衰减截止频率、频谱流量和均方根四个压缩域特征，每个短时帧的四个压缩域特征组成的向量就构成

了这个音频例子的特征[22]。

下面将分别对模糊聚类质心形成、聚类质心快速匹配和音频例子相减反馈进行分析。

5.2.2.2 音频例子聚类质心提取

由于几秒钟长度的音频例子就提取出数百个短时帧，而每个短时帧又有四个特征向量来描述。因此在提取出压缩特征以后，如果直接使用这些压缩域特征进行检索，则工作量非常大。

为了缩减数据量而又不失掉每个音频原来的特性，可以对域特征进行聚类分析，提取出固定数目的质心来表征音频，也就是说，用聚类质心来为音频例子建立索引。

通常的聚类方法是将众多归为几个子类，子类之间的界限非常明确，每个属于其中惟一的一个子类，如 k 平均聚类和混合高斯聚类算法[23]。但对于一段音频来说，由于音频本质上是非平稳信号(短时帧内信号平稳)，对它进行明确的聚类不是一件容易的事情：比如很多歌曲，究竟它们属于音乐，还是属于语音就很难确定。

因此，引入了模糊聚类概念实现音频例子压缩域特征表达[24]，即，被聚类对象可以属于第一类别，只是隶属度不同。并且，在聚类过程中，受音频特征变化剧烈影响，某个聚类子集内的可能数目较少，或者某些聚类质心的距离很近，可用时空约束规则来进行调整。

5.3 音频检索的类别

目前以音频为中心实现对多媒体内容的检索可以分为五类。

5.3.1 音频事件提取和查询

在这种方法中，通过对音频信号(如语音、背景音乐和环境噪音等)的检测、分割和识别，达到对视频事件的索引。如文[25]把音频信号分为语音、静音和背景音乐三类，以音频信号的 Mel-频谱系数、短时能量系数和 *delta* 系数作为特征向量，为每一类音频信号建立一条隐马尔可夫链(高斯密度函数作为每个状态的观测概率)去识别这三种音频信号，每个隐马尔可夫链用一条语义信息加以标注(如爆炸、枪响等)，它不仅成功识别分割出了上述三种音频事件，而且用隐马尔可夫链识别了存在多种音频信道源的事件。

音频事件提取和查询还有一个重要作用，就是通过音频信号识别找到多媒体数据流中“重要场景”，而这些重要场景是用户普遍感兴起的。如在足球比赛和棒球比赛等节目，其精彩场景出现时候，往往伴随观众呼喊声和解说员兴奋解说词等，如果能够从足球比赛和棒球比赛中正确识别观众呼喊声解说员兴奋解说词等音频信号，就可以将这些重要场景提取出来，建立索引，提供给感兴趣用户。实现使用音频信号流实现对视频信号流检索。

5.3.2 基于元语言(Metadata)映射结构的多媒体多模态事件提取和查询

在这种方法中，多媒体信息被分成音频、视频和文本三种不同信息，对每个信息独立建立索引结构。在[26]中，声频事件的特征向量用短时能量、过零率、基本频率和频谱共振峰来表示，用基于启发式方法采用隐马尔可夫链按先后顺序分别识别出静音、人语、音乐和环境音等，把识别出来的音频标注为“音频场景”，对不同的音频场景按时间先后关系建立索引图：视频图像首先按照直方图和运动矢量特征被 分割成描述不同语义的镜头，然后在每个镜头中按照着色、纹理和窖等特征提取关键帧。被提取的关键帧同样按照其在整个多媒体位流中出现的时间先后顺序建立索引表；纯文本信息经过字体识别，提取关键字后为每个关键文本字建立起索引表(这样的索引表被称为元语言结构)。这样，一串无结构的音频、视频和文本三态信息就被结构化，以后可以通过对这三个索引表目的查询实现多媒体多模态信息查询。

5.3.3 音频到视频事件的映射

视频和音频特征信号映射的研究起源于实体(Linguistic Entity)，音频被切割成音素(Phonemes)序列，然后再被映射到发出各个音素时所表现出来的嘴唇素(Viseme)。在这种转换方式中，要首先建立音素和视素对应的查找表，以便把每个与相应音素相关的嘴唇视素拼接起来，形成人讲话的动作。

另外一种方式是把音频视为物理现象，把重点放在音频信号参数和视觉参数转换上，而不是视觉图像上[27]。实现这种转化的算法包括矢量量化，神经网络和带高斯混合密度的隐马尔可夫链。

5.3.4 声频音频特征向量融合表征多事件

上面谈到的方法是把视频、音频和文字当作分离的媒体信息，独立加以处理，所以存在不足。[28]把颜色直方图图像特征、12-维频谱向量音频特征和运动矢量作

为三个独立的高斯分布函数集成到一个隐马尔可夫链中(考虑运动矢量的原因在于它能够反应出像机运动，如全景和镜头缩放等)，隐马尔可夫链的状态分为镜头、镜头切换、镜头淡进淡出、全景镜头和镜头缩放 5 种。在多媒体信息流帧和帧之间计算上述三种特征，训练隐马尔可夫链参数。以后一旦获得一列特征向量，就可以由训练好隐马尔可夫链判断产生对应这列特征向量的最大状态列，从而把包含了视频、音频和运动矢量信息的多模态多媒体信息流按照如上不同的状态切割识别出来。

5.3.5 音乐分析(Music Analysis)

从广义上说，任意和谐的声音都能称为音乐。音乐与人的听觉感知紧密相关，它不同于语音，因此不能像分析语音那样，通过分析语音所对应的单词了解语音所要传递的语义上的信息。本质上，音乐更多的是传达了一种感情，一种很难量化的情结，或喜或悲，或紧张或舒缓，音乐的这种特性要求的分析采用不同于语音及其他音频的方法。

首先来分析一下音乐的某些物理特征。与以前分析的时域特征、频域特征或者时频特征不同，音乐的特征是一种感知特征，不单单基于音频的物理波形，而是与人耳的构造以及人的听觉经验相关。一般地，常用的音乐特征有音调(Pitch)、音质(Timbre)、节奏(Rhythm)等，不同的音乐特征应用于不同的领域。

目前音乐分析的研究方向主要有乐谱转录和乐器识别，其应用方向是音乐检索。乐谱转录主要研究如何识别不同音符的音调，从而达到转录的目的。当同一个音符由不同的乐器演奏时，会存在不同的效果，定义这种与乐器有关的特定效果为乐器的音质，乐器识别正是利用了乐器在音质上的区别来达到辨认的目的。

5.4 音频检索未来与挑战

音频处理和检索是与信号处理、人感知心理研究和模式识别等学科相联的研究领域，其面临的挑战较多。有些挑战是在基于内容的图像检索中同样存在的，如基于语义检索，而非特征相似检索。

5.4.1 感知特征提取

从音频数据中可以提取的特征很多，前面把音频特征按照处理空间的不同，分成时域、频域；同时，在提取音频特征时，是考虑音频长时非平稳和短时平稳的本质，还是考虑第一种语义需要持续一定时间，将音频特征分为短时音频帧和

音频例子特征两类。

但是，从前面的分析中知道，目前基于内容的音频检索之所以不能自动实现语义检索的原因在于音频低级特征与高级语义之间存在一条鸿沟。为了填补这个鸿沟，就要研究人脑对音频信号的感知机理。

因此，按照是否对音频信号进行感知变换，将音频特征分成感知和物理两类。物理特征还可以被分为时域和变换域特征。音频感知特征是模拟人的听觉感知器官对信号先进行处理，然后从处理之后的数据中提取特征，把这样得到的特征叫做感知特征。

考察人听觉感知机制对音频数据处理的过程：声波通过空气到达耳蜗，耳蜗通过基底膜里面的毛状细胞把声波转换成时/频表示，然后这种时/频信号序列送给大脑进行处理，产生声音场景。

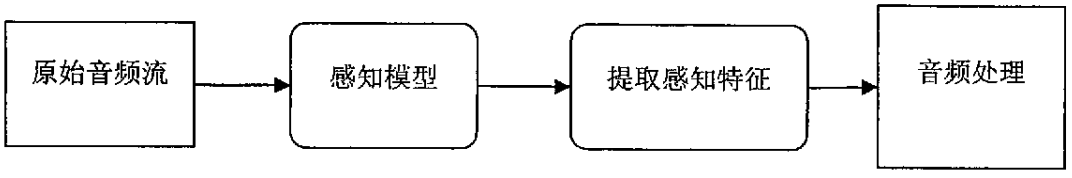


图 5-2 音频感知特征提取

按照人的耳蜗结构，模拟设计耳蜗的功能(听觉感知模型)，让音频信号先通过模拟的耳蜗，然后对经过处理了的信号提取特征，进而让计算机进行理解，可以缩短计算机和人脑对音频场景理解之间的差距。如图(5-2)

如何更好按照人的生理系统构造感知模型，实现感知特征的提取，是面临的一个研究困难。

5.4.2 音频场景分析

一般而言，人耳处理的音频信号来源于几个地方，是混合的音频信号。比如，在一个比较喧闹的环境中录制一段音频信号，在这段音频信号中混合了音乐、人讲话和汽车各种混合声音。然后把这段音频信号播放给某个人听，即使没有看见过这样的场景，并且这三个声音来自不同的信号源，但是人耳能够把混合的声音区分出来，并且在大脑中形成这三个信号所产生的不同的场景，如有关音乐播放的场景，有关人讲话的场景和有关汽车驶过的场景。“听觉场景分析”目的就是研究如何把来自不同声源混合在一起的音频切分，然后分别形成这些声音所表示的场景。

这样，听觉场景分析是为了探讨大脑如何把混合的音频切分成不同的音频事

件。前面讲到，音频原始信号应该首先被人的耳蜗处理赔笑不同停产上不同频率带的信号。显然，对于不同的声源信号，如何把它们转换生成的频率带上的信号组合到一起，让大脑处理是很关键的。

加拿大 McGill 大学的 Bregman 教授在其 1990 年出版的“听觉场景分析：对声音进行感知层次组织”的书中介绍[29]：听觉场景分析用来研究听觉系统如何对外界掣进行组织与加工的。其任务有两个：一是找出那些能够使声谱成人组合到一想或使它们分离成独立的听觉流或表象的声学特征；二是基于这些特征，把不同声源的音频信号分组。场景分析包含两个阶段一是把不同感觉元素分配到相应组中；另一阶段是对知觉组织进行验证和修复。这两个阶段分别对应于自下而上和自上而下两个加工过程。

在第一个阶段中，先把听觉信号分割成许多独立的单元，这些单元与声谱中特定时域和频域相对应。然后，对这些单元进行分组或分离。分组是指听觉系统把某些具有相似特征或时间接近的音知觉分为一个流，使之从复杂的环境声中突出出来。分离则是从复杂环境声中辨别出声音的不同来源或区分不同声音。分离和分组是一对统一的概念，如果出现了分组，也就意味着流与流之间产生了分离。分组是把在不同时间内顺序出现的谱成份纳入一个知觉流，以便计算环境中声音的序列特性。而后者则把同时出现的成分分开，将它们放入不同的流中。

在人的感知信号形成过程中，人一般会把环境中特定的声音信号，如言语、音乐以及其他熟悉的声音等存入记忆中，形成认知单元，这些认知单元就是认知图式。当混合的音频信号流被按照特征分离并且分组后，听觉系统获得的信息模式与图式相同时，图式将被激活，并且通过图式对模式的其余部分进行推测。图式还可以被与其相关联的其他图式激活。图式加工是一个自上而下的加工过程。

听觉场景分析把人对来自混合声源的信号所代表的不同事件分离开来，目前有些学者开始天空如何对混合的声音场景进行分割检测：如 Eric D.Scheirer 利用相关图市制技术对混合信号中的音频场景进行分割与动态检测[30]。

5.4.3 更好分类器研究

所谓分类器指将数据分成不同范畴或类别的一种装置或过程。对于音频研究而言，所谓分类就是将音频例子按照其特征分成不同音频类别，如“爆炸”、“掌声”和“语音”类。

分类在音频检索中起了很重要作用，如隐马尔可夫链、支持向量机、最近特征线和混合模型等分类机学习训练过程。隐马尔可夫链和支持向量机属于监督学

习分类模型。在训练一个用这些模型表示的音频类别前，自己要选择一批属于这个类别的音频例子，然后进行训练[31]。

应该讲，一个良好的分类器的得到，需要考虑具体的应用，不同的分类器在不同的应用领域可以获得成功应用。

5.5 本章小结

本章主要介绍音频检索算法， 包括相似音频例子检索，基于模糊的音频例子检索。探讨了音频检索的分类和音频检索的未来和挑战。音频检索分类主要包括了音频事件提取和查询，音频到视频事件的映射及音乐分析等。

第六章 总结

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。音频信号的处理作为研究项目的一个部分，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：

- 音频信号的基本处理方法，对音频信号进行分帧和加窗处理。
- 音频信号主要的时域、频域特征提取；并对音频信号的时域、频域特征的应用情况进行分析。提出了一种均值 MFCC 系数特征提取方法。
- 研究了音频信号的分割和识别算法。音频信号的分割部分主要研究了音频信号的分层分割方法，对音频信号进行粗分类；音频信号的识别部分主要研究了动态时间规整匹配算法，对单一的音频信号进行识别。
- 研究分析当前音频检索的分类情况和面临的挑战。

本文的研究作为研究课题的一部分，是信号源中三种信号的一种，主要技术要求是对单一目标的识别起到辅助作用。同时基于内容的音频信号的研究是一个新兴的领域，还处于研究的初级阶段，对音频信号的研究主要进行仿真处理，采用 MATLAB 来实现。音频信号数据格式为.wav 文件，主是用 GOLDWAV 软件，进行采样率和格式转化。

本论文进行的音频仿真部分，完成了对单一音频信号的识别，对连续音频信号进行分层分割的功能；对单一音频信号的识别部分进行程序移植，采用 VC6.0 来实现，达到了整个项目对音频部分的技术要求。另外，对音频信号的检索方面进行了较深入研究，为以后在这方面的研究奠定良好的基础。

参考文献

[1] 李国辉 李恒峰 基于内容的音频检索：概念和方法[J] 小型微型计算机系统 2000 年 11 月 p1173-1177

[2] Erling Wood el, Content based classification, search, and retrieval of audio. IEEE Multimedia, 1996

[3] Witbrock M. J. and Hauptmann A. G., Speech recognition and information retrieval. Proceedings of the 1997 DARPA Speech Recognition Workshop, February 2-5, 1997

[4] S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, Acoustic Indexing for Multimedia Retrieval and Browsing. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, 199-202, Munich, Germany, April 1997

[5] Laura Slaughter, A graphical interface for speech-based retrieval. The third ACM Digital library conference, Phi, June, 1998

[6] Smith M. and Kanade T., Video skimming for quick browsing based on audio and image Characterization. Submitted to the Special Issue of the IEEE Transactions on Pattern Analysis and Machine Intelligence on Digital Libraries : Representation and Retrieval, 1995. Also as Technical Report CMU-CS-95-186

[7] McNab R. J., Smith L. A., Witten I. H., Henderson C. L. and Cunningham S. . J., Towards the digital music library : tune retrieval from acoustic input. Proc Digital Libraries' 96, pp 11-18 1996

[8] Kenney Ng. Towards robust methods for spoken document retrieval . Proceedings of Int. Conf. on Spoken Language Processing, 1998

[9] Tong Zhang and C.-C. JayKuo. Content-based classification and retrieval of audio . Proceedings of SPIE' s Conference on Advanced Signal Processing Algorithms, Architectures, and ImplementationsVIII, SanDiego, July, 1998

[10] 庄越挺 潘云鹤 吴飞编著 《网上多媒体信息分析与检索》[M] 清华大学出版社 2002 年 9 月

[11] 邢伟利，周明全 基于内容的音频检索技术[J]，西北大学学报，2004 年 4 月，第 2 卷第 4 期

[12] 韩萍 仓储物害虫声音的模式识别[J], 计算机工程 ,2003 年，第 29 卷第 22 期 151-153

页

[13] 韩纪庆 语音信号处理 清华出版社 2004

[14] 卢 坚，陈毅松，孙正兴等. 基于隐马尔可夫模型的音频自动分类[J]. 软件学报，2002，13(8)：1 593- 1 597.

[15] 吴飞 庄越挺 郑科 刘骏伟 基于压缩域特征话者识别的电视节目分类检索[J] 《人工智能与模式识别》 2002年3月 第15卷 第1期

[16] Kedem, B. Spectral Analysis and Discrimination by Zero-Crossings. Proceedings of the IEEE, Nov 1986, Vol 74, No11, pp1477-1493

[17] Wilcox, L D, F R Chen, D Kimber and V Balasubramanian. Segmentation of Speech Using Speaker Identification. Adelaide, Australia: Proc Int Conf Acoustics, Speech and Signal Processing, April 1994

[18] A K Jain and F Farroknia. Unsuervised texture segmentation using Gabor filters. Pattern Recognition, 1991,24(12), pp 1167-1186

[19] 王 辰 张宪海 老松杨 胡晓峰，基于声、像特征的视频暴力场面的探测[J]，小型微型计算机系统 第22卷第4期

[20] 贾磊 穆向禺 徐波 广播语音的音频分割[J] 中文信息学报，第16卷第1期 37-42页

[21] 张杰，余志刚，黄志同 语音识别中广义模型及其算法收敛性分析[J]计算机工程与应用 2000 Vol.36 No.2 P.60-62

[22] 张春林，杨玉红，胡瑞敏 音频内容分割与聚类的研究[J] 计算机工程 ，第 28 卷第 7 期.173-174页

[23] George Tzanetakis, Perry Cook. Multi feature audio segmentation for browsing and annotation. New Palz, NY: in Proc 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA99, 1999

[24] Gerald Salton, Edward Fox, Harry Wu. Extended Boolean Information Retrieval. Communications of the ACM, December 1983, Vol 26, No 11, page 1022

[25] Zhang T and Kuo C C. Audio-Guided Audiovisual Data Segmentation, Indexing and Retrieval. San Jose, California: Part of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Database VII, January 1999, SPIE Vol 3656, pp 316-327

[26] Zhang, T and C-C Jay Kuo. Integrated approach to multimodal media content analysis. San Jose: IS&T/SPIE' s Symposium on Electronic Imaging: Science & Technology-Conference on Storage and Retrieval for Image and Video Database VIII, Jan 2000

[27] Ram R Rao, Russell M & Mersereau Tsuhan Chen. Using HMM' s for Audio-to-Visual Conversion. Princeton, New Jersey, USA Electronic Proceedings: IEEE Signal Processing Society 1977 Workshop on Multimedia Signal Processing June, 1977

[28] J S Boreczky and L D Wilcox. A hidden Markov model framework for video segmentation using audio and image features. Seattle, WA: in Proc Int Conf Acoustics, Speech and Signal Processing(ICASSP-98), May 12-15 1998, Vol 6, pp 3741-3744

[29] Bregman A(1990). Audiotory Scene Analysis. Cambridge MA:MIT press

[30] Eric DSchirer. Sound Scene Segmentation by Dynamic Detection of Correlogram Comodulation. Mit Media Laboratory Perceptual Computing Section Technical Report No 491, April 1999

[31] 吴飞 庄越挺 张引 潘云鹤 基于隐马尔可夫链的音频语义检索[J] 《人工智能与模式识别》 2001年3月 第14卷 第1期 p104-108

致 谢

在此，谨向为本论文倾注了大量心血、提供了大量帮助的老师和同学表示深深的感谢。

首先，要感谢我的导师解梅教授。她在学习和研究上给我们创造了一个和谐自由的环境，给予我们充分的积极性和发挥的空间，使我们终生受益。在研究生学习期间，解老师对我思想上、学习上的循循善诱和谆谆教诲。本论文从选题到方案的确定，从工作的逐步实施到最终的审定都无不倾注了她大量的心血和辛劳。解老师学术精深、治学严谨、工作勤奋，她的言传身教将一直指导我以后的工作和学习。

在攻读硕士学位期间，我得到了教研室其他同学的无私帮助，特别是与我同一个项目组的贾海涛、胡柳和黄宇同学，我们在研究和学习的过程中互相学习，互相帮助，最终共同完成了科研项目，在此也向他们表示我深深的谢意。

最后，我要特别感谢我的父母和妻子，他们一直在我身后默默的关心我、支持我和鼓励我，他们为我付出得太多太多。

个人简历

许刚

1977 年 1 月出生于山东省邹城市

1998 年 7 月获电子科技大学电子材料与元器件专业学士学位

1998 年 7 月至 2002 年 3 月工作于空军第一实验训练基地

2002 年 3 月至 2003 年 9 月工作于兰空通信修配厂

2003 年 9 月至今在电子科技大学电子工程学院攻读硕士学位

发表的学术论文：

[1] 许刚. MPEG-4 对象分割与编码算法. 中国科教博览，2004 年第 11 期

[2] 许刚，解梅. 基于内容的音频检索研究. 科学学报，2006 年第 1 期

[3] 许刚，解梅. 音频检索特征选取探究. 自然科学研究(中文版)，Volume 3 ，number 1， February 2006