申请上海交通大学硕士学位论文

基于音频的视频内容检索

——面向流媒体内容监控的音频检索关键技术研究

学　　校：上海交通大学

院　　系：电子信息与电气工程学院

班　　级：B0603495

学　　号：1060349170

姓　　名：时金

专　　业：信号与信息系统

导　　师：周军

上海交通大学电子信息与电气工程学院

2009 年 1 月

A Dissertation Submitted to Shanghai Jiao Tong University for the

Degree of Philosophy Master

Video Content Retrieval Based on Audio

Author: Jim Z. Shi

Specialty: Signal and Information System

Advisor: Asso. Prof. Zhou Jun

School of Electronics and Electric Engineering

Shanghai Jiao Tong University

Shanghai, P.R.China

January, 2009

上海交通大学
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本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和汇编本学位论文。

保密□，在_____年解密后适用本授权书。

本学位论文属于

不保密☑。

（请在以上方框内打“√”）

学位论文作者签名：时彦丰　　指导教师签名：周军

日期：2009 年 1 月 15 日　　日期：2009 年 1 月 15 日

摘 要

近年来，随着多媒体和网络技术的迅猛发展，数字视频信息出现了飞速膨胀。新的视频应用，例如视频邮件、视频点播、数字电视等，已经为越来越多的人所熟悉和接受。如何有效检索这些视频内容成为了研究的热点问题。与此同时，许多个人、团体以互联网为平台，通过音视频下载、流媒体广播等方式向社会发布各类信息。社会安全保障行动中，对信息的监控必不可少。如何对海量的实时流媒体视频进行有效的监控，及时发现、过滤出这些信息中包含的敏感、有害内容，对维护互联网空间和谐、乃至社会安定起着至关重要的作用。因此，流媒体视频内容实时性监测在国民经济和社会发展中有重大需求，是当今信息社会亟待解决的问题之一。

然而，视频检索和流媒体内容监测技术面临着巨大挑战。目前许多视频检索网站（如 Google Video 等）以及流媒体点播网站（如 YouTube.com 等）基本采用的是基于媒体标题、标签关键字文字的检索和监控，即通过人们对视频的文字描述，把视频当作文档来处理。如何从视频内在的图像序列、音频信息、字幕中提取语义从而加以检索和监控具有很强的挑战和吸引力。同时，网络视频信息量非常大，并且还在急剧的膨胀中。如何能提高检索和监控的实时性和有效性，也成为影响基于内容的视频检索和监控技术应用于实际的至关重要的因素。本项研究工作就是在这样的背景下，探索了如何从音频检索角度出发，辅助基于内容的视频检索和流媒体监控，以提高其有效性和实时性。

论文首先从音频模态进行视频内容检索的方法进行了探讨和分析，研究总结了基于音频的视频检索中音频特征的有效选取和门限值的确定，并通过实验，给出了一组有效的音乐特征以及门限值选取的方法；其次，在向量模板分类算法的基础上，提出了一种分层的向量模板分类算法（HCMBVT），通过分级分类，减少分类过程中的计算冗余，提高了分类效率；在此基础上，结合传统基于欧式空间距离的匹配算法，提出一种改进的前向加权序列匹配算法（FWDM），实验结果表明这两种算法可以有效地提高视频检索的实时性；通过实验探讨了基于音频的视频内容检索系统的优缺点及可行性。最后给出了实验结果及分析。

关键词：视频检索，音频检索，音频分类，特征提取，HCMBVT，FWDM

Abstract

In recent years, with the rapid development of multimedia and network technology, digital video information appeared in the rapid expansion. New video applications such as video mail, video on demand, digital TV and so on, have been familiar with and accepted by more and more people. How to effectively search these video contents has become the hot issue of the study. At the same time, many individuals, groups released all types of information to the community through audio and video downloads, streaming media broadcast etc, using the Internet as a platform. Monitoring and control of information is essential in social security operations. How to monitor the massive real-time streaming video effectively, timely detect and filter out the information which contains the sensitive and harmful content, plays a vital role in the maintenance of the Internet space harmony and even social stability. Therefore, the national economy and social development have a significant demand for streaming video content real-time monitoring which is to be settled urgently in the information society nowadays.

However, the video retrieval and streaming media content monitoring technology facing an enormous challenge. At present, many video search sites (such as Google Video, etc.) as well as on-demand streaming media Web site (such as YouTube.com, etc.) is based on the basic use of the search and monitoring of media title, tags keyword text. which deals with video as text document throngh the text description of video content. How to extract semantic from intrinsic image sequences, audio information, subtitles from video and thus to implement semantic retrieval and monitoring is a strong challenge. At the same time, network video is very large amount of information, and is in rapid expansion. Finding ways of improving the effectiveness of information retrieval and real-time monitoring is also a crucial factor of how content-based video retrieval and monitoring technology are used in practice. This study work is in such a context, from the audio point of view, to explore how to retrieve a supplement to content-based video retrieval and streaming media monitoring, in order to enhance its effectiveness and real-time.

In this thesis, we'll discuss and analysis the content based video retrieval from audio modality, study the audio feature selection and the gate value in video retrieval system based on audio, and supply a set of audio features and method to determine the threshold of gate value; after that, we introduce a hierarchical classification method based on vector template (HCMBVT), which improves the

classification efficiency through level separation, reduction in the process of calculating the classification of redundancy. On this basis and the combination of the traditional Euclidean space-based distance-matching algorithm, we introduce an improved match method named Forwarding-sequential Weighted-feature distance measuring (FWDM); the experimental results show that these two algorithms can effectively improve the real-time video retrieval. Finally, experimental results and analysis are shown.

Keywords: video retrieval, audio retrieval, audio classification, feature extraction, HCMBVT, FWDM

目 录

基于音频的视频内容检索.....0

摘 要 I

ABSTRACT.....III

第一章 绪论 1

1.1 课题研究背景与意义 1

1.2 国内外研究现状 3

1.3 论文的主要研究成果和论文结构 5

第二章 基于音频的视频检索系统概述 8

2.1 视频的结构化信息 8

2.2 典型的多模态视频检索框架.....9

2.3 基于音频的视频流媒体内容检索框架.....11

2.3.1 基于文本和关键词的音频检索.....12

2.3.2 基于音频特征的音频检索.....13

2.4 本章小结 14

第三章 视频检索中的音频特征分析与提取.....16

3.1 音频特征的描述.....16

3.1.1 时域音频特征 17

3.1.2 频域音频特征 21

3.1.3 系数域音频特征 24

3.2 音频特征分析和提取相关技术[31] 30

3.2.1 音频短时处理技术 31

3.2.2 同态处理技术 32

3.2.3 预处理技术 33

3.3 音频特征的选取及实验.....35

3.4 本章小结 37

第四章 视频检索中的音频分类与匹配 39

4.1 音频分类算法.....39

4.2 音频匹配算法.....40

4.3 改进的分类和匹配算法.....41

4.3.1 分层的向量模板分类算法.....*41*
4.3.2 前向加权序列的匹配算法.....*45*
4.3.3 算法融合及系统设计.....*46*
4.4 实验结果及分析.....50
4.5 本章小结.....54
第五章 总结与展望.....55
5.1 本文总结.....55
5.2 未来工作展望.....55
参考文献.....57
附 录.....61
附录 1：图片目录.....61
附录 2：表格目录.....62
附录 3：缩略语.....63
致谢.....65
攻读学位期间发表的学术论文目录.....66

第一章 绪论

1.1 课题研究背景与意义

近年来，随着多媒体和网络技术的迅猛发展，数字视频信息出现了飞速膨胀。新的视频应用，例如视频邮件、视频点播、数字电视等，已经为越来越多的人所接受和熟悉。如何有效检索这些视频内容成为了研究的热点问题[9,35]。与此同时，许多个人、团体以互联网为平台，通过音视频上传、下载、流媒体广播等方式向社会发布各类信息。社会安全保障行动中，对信息的监控必不可少。如何对海量的实时流媒体视频进行有效的监控，及时发现、过滤出这些信息中包含的敏感、有害内容，对维护互联网空间和谐、乃至社会安定起着至关重要的作用。因此，流媒体视频内容实时性监测在国民经济和社会发展中有重大需求，是当今信息社会亟待解决的问题之一[46,47]。然而，视频检索和流媒体内容监测技术面临着巨大挑战。目前许多视频检索网站以及流媒体点播网站基本采用的是基于媒体标题、标签关键字文字的检索和监控[38]。通过人们对视频的文字描述，把视频当作文档来处理，其实质依然是利用传统的基于文本的检索方案。如何从视频内在的图像序列、音频信息、字幕中提取语义从而加以检索和监控具有很强的挑战和吸引力。

Kin-Wai Sze 等研究者[10,12,17]在视频检索领域做了大量工作，并取得了令人满意的结果，他们的研究主要在视频图像单一模态领域。然而我们注意到从视频图像中提取和分析视频的语义存在以下难度：

- 视频是一个完整和连续的信息流，是集图像、声音和文字为一体的综合性媒体信息流，而这个信息流本身并不包含明确的结构信息。
- 视频本身对应的数据量非常巨大，如果没有一个有效的组织方式，很难对这些海量数据进行管理和检索。

过去的几年里，一些研究[11,14,15]表明多模态综合信息中进行检索，可以取得比过去单一模态更好的结果。由此，多模态（Multi-mode）[9,19,38,44]信息融合的视频检索方法便成为视频检索领域重要的组成部分。该方法针对视频图像、音频流中包含的多模态信息分别进行查询，并通过有效融合得到优于任何单一检索模块的查询结果。语音作为多媒体信息的一种模态，是传播信息的一种重要载体，具有清晰的语义。由于语音的存在，音频流所包含的语义信息往往比图像流丰富，提取语义信息也更加直观方便。

人和人交流时，对信息的理解是视听双态，这被称为“McGurk 现象”[29]。“McGurk 现象”指出，听觉信号适合于表述描述性(Descriptive)语义，视觉信号适合于表达指令性(Manipulative)语义，视觉和听觉综合才能表达一个完整丰富的语义信息，对两者的割裂将使完整语义信息丢失。因此，在对综合了视觉和听觉信号的视频信息查询时，视觉和听觉信号的综合交互(Audio-Visual Interactiom)尤为重要，“McGurk 现象”为基于音频的视频内容检索的综合分析研究提供了理论基础。在其他的研究中，Hoi 等人从较高的层面讨论了利用音频、语音和信号处理技术进行多模态视频检索的系统框架[9,19]；Shepherd 等人[7]研究了利用音频特征在多媒体数据库中进行基于内容的音乐检索；Chung-Hsien Wu 等人[8]讨论了从音频语音信息中进行关键词抽取和语义确认方面的方法；Ki-Man Kim 等人[30]则着重探讨了利用音频特征进行快速音频检索的方法；利用音频处理技术为基于内容的视频检索提供辅助解决方案，正吸引了越来越多研究者的目光。这是本论文研究的理论背景。

同时，本论文的工作也有其具体的现实意义。随着我国网络基础设施的的不断完善，越来越多的人通过网络链接起来。个人的意见、信息可以通过网络以文字、音视频等多媒体在更短的时间内传播给更广泛的群体。越来越多的个人、团体以互联网为平台，通过音视频下载、流媒体广播等方式向社会发布各类信息。这给网络使用者带来巨大便利的同时，也给网络监管部门提出了一个极具挑战的课题，即如何能够及时对网络上传播的多媒体信息流进行内容检索，以维护网络社区的和谐环境最终达到社会的安全保障，在这个过程中对信息的监控必不可少[46,47]。从这点上来看，研究基于音频的视频内容检索尤其现实的应用前景。

另一方面，随着研究的不断深入，基于内容检索的速度成为了制约其在生活中被广泛应用的瓶颈。目前大多成熟的商业性视频网站都还没有采用基于内容的音视频检索技术的一个主要原因就是基于内容的多媒体信息检索需要耗费过于巨大的系统资源。众所周知，计算机、网络通信技术的日趋成熟，各种实用型音视频技术的不断完善，语音信箱、视频聊天、音视频会议和数字视频点播系统等产品得到了越来越广泛的应用，这意味着网络音视频信息量非常庞大，并且依然在急剧的膨胀中。这无不给检索任务提出了难题，即如何能提高检索和监控的实时性和有效性，成为影响基于内容的视频检索和监控技术应用于实际的至关重要的因素。也正因为此，本论文在前人研究成果的基础上，重点讨论了如何减少检索过程的计算量和提高检索速度，这在技术应用上也有其现实意义。

本论文就是在这样的背景下，探索了如何从音频角度出发，辅助基于内容的视频检索和流媒体监控，重点研究了如何在检索过程中减少计算冗余、提高速度，以提高其有效性和实时性。

1.2 国内外研究现状

关于基于内容的视频分析与检索，前人已经取得了很多研究成果。目前，国外已研发出多个基于内容的视频检索系统。比较著名的有：

- 主要有 IBM 公司研究中心开发的基于内容的检索系统(IBM's Query By Image Content, QBIC)[28]。它可以对图像、视频、文本和语音进行检索。主要由两部分组成即数据库生成部分和查询部分。在数据库生成时成每一个图像对象和视频对象的内容特征包括：颜色，纹理，形状，以及摄像机和对象的移动等，都被提取出并存进数据库中。当查询时，数据库查询部分把用户利用图形化方法提供的对象特征与数据库中存储的对象内容特征进行比较匹配，寻找出具有相似特征的图像和视频。
- 美国哥伦比亚大学图像和高级电视实验室开发的 VisualSEEK 系统是基于内容的图像、视频检索系统[27]。利用 VisualSEEK 工具可以在网络上搜索和检索图像与视频，通过用户接口工具表示出要查询的图像的主要可视特征，将其送到检索服务器，服务器查找和检索出最佳匹配图像，并返回给用户。VisualSEEK 自动进行特征抽取，而不是依赖于人工输入的文本和关键字。它提供可移植、易用并具有可视查询能力的用户接口，使用户容易直接提交简捷的基于内容的查询。有经验的用户还可以构造更复杂的查询。
- 意大利 Palermo 大学开发的 JACOB[26]是一个基于内容的视频查询系统，系统可进行视频自动分段并从中抽取关键帧，并可按彩色及纹理特征以关键帧描述基于内容的检索。
- 新加坡国立大学开发的一个基于内容的检索机 COKE 系统。其显著技术特色包括:多种特征提取方法、多种基于内容检索方法、使用自组织神经网络对复杂特征度量、建立基于内容索引的新方法以及对多媒体信息进行模糊检索的新技术。

还有其他一些系统，如美国哥伦比亚大学研究实现的 VideoQ 系统[23]；Virage, Inc.公司的 Virage Search Engine[25]，由 UIUC (University- of Illinois at Urbana Champaign)开发的 MARS 等。国内的主要研究单位如清华大学、上海交通大学、微软亚洲研究院、国防科技大学多媒体研究中心等单位，也开展了对基于内容的视频检索技术的研究，获得了一定的成果。例如，微软亚洲研究院的张宏江博士所带领的小组研制出的 Ifind 信息检索系统[45]等。

以上这些系统都是用于理论研究性质的系统，至今尚未有正式的商业网站采用基于内容的视频检索技术。原因在于该技术目前检索的速度和准确率还不尽人意，所以离实际应用还存在一定距离。美国国家标准与技术协会（NIST）为提高视频检索的性能，每年举办一次关于视频检索的竞赛活动 TRECVID。举行的视频检索竞赛为世界各地研究者提供了交流探讨的平台。TRECVID 的前身是 Text Retrieval Conference(TREC)，致力于信息检索的研究。TRECVID 为参赛者提供了大量测试视频以及统一的评分流程，每年都有几十个团队参赛，国外知名的机构包括 IBM、哥伦比亚大学、卡内基梅隆大学、微软亚洲研究院等，国内有清华大学、上海交通大学等。此外，MPEG-7 即“多媒体内容描述接口"(Multimedia Content Description Interface)，作为 MPEG 组织提出的新标准，其目标是制定一组标准的描述符及其描述模式(定义描述子的结构和相互关系),内容描述与媒体内容结合，使用户能够快速准确地进行检索,这也注定了其在未来通用的视频检索中将扮演主要角色、发挥重要的桥梁作用。

利用音频模态进行视频内容检索的系统框架基本类似[9,19,36,38],即主要分五个步骤：1)从流媒体中解复用出音频信息；2)从音频流中提取出音频特征；3）对音频特征进行分类；4）利用音频特征进行匹配；5）对匹配结果进行融合后得到最终检索结果。其中与检索结果关系最密切的是中间三步即特征提取、音频分类和匹配。研究者在这些方面做了大量研究。

音频特征的选取是音频检索最基本的过程。它是指在众多声学特征当中选取出最适合音频检索的那些特征，作为音频分类和匹配的依据。他对音频检索的影响主要有两个方面，即检索的速度和正确率。Ki-Man Kim[30]最初只选取 ZCR 一个特征，获得了理想的速度，然而准确率较低；后来又选取了 FC 等 4 个特征使得准确率获得了略微的提升。中科院语音研究所的李恒峰等人[39]选取了多达 98 维的特征向量，获得了理想的分类和检索结果，然而检索时间较长。可以说音频分类和匹配过程中计算时间和准确率这对矛盾量的根本原因就体现在音频特征的选取上。

音频的分类是指利用选取的音频特征对音频帧进行内容分类（语音、音乐等）；音频匹配则是指把相同类型的音频片段的特征值进行比较以找到相同或相似的音频帧。对音频帧先进行分类是为了避免不必要的匹配过程以提高检索过程的整体效率。Foote J.等人在研究中采用了针对音频特征值的分类方法[1]，这种方法速度快、操作简单，但准确率不高只适用于简单类型音频分类；Weld E，Chung-Hsien Wu[14,18,35,37]等人在研究中都采用了基于音频特征的统计特征的分类方法，这种方法的准确率最高，但是检索前需要大量样本进行学习且计算量较大；澳大利亚人工智能研究院的 Elias Pampalk [32]等人开发的基于 SOM

(Self-Organizing Maps) 的音乐聚类系统以及中国国防科技大学多媒体实验室的李恒峰、李国辉[39]开发的基于内容的音频分类与检索系统 ARS 采用了基于音频特征向量模板的分类算法获得了较好的实验效果，但这种方法在分类之前需要计算特征向量模板中所有特征值，存在一定计算冗余； 另外 Hoi 等人在[9,19]中详细讨论了基于音频和其他模态的视频检索系统的基本框架。

1.3 论文的主要研究成果和论文结构

基于内容的视频检索理论近年来在国外得到了较快的发展，目前该方向还有一些问题有待解决。本论文的主要研究对象是该理论的一个分支，即从音频检索的角度出发，对视频进行辅助检索，主要研究对流媒体视频进行实时的、有效的监控，及时发现、过滤出其中包含的敏感信息。从而能够在流媒体这样一个对实时性要求较高的特定应用场景下，满足对敏感信息检出、过滤的应用要求。本论文所采取的方法是从音频着手，认真选取有效的音频特征，通过提高音频分类和匹配速度，在保证准确性的前提下，提高检索、监控的效率以提高处理的实时性。

本质上讲，音频检索是一个模式识别的问题。它包括几个方面，语义的描述、特征的提取和特称的匹配。语义检测当中最直接的语义描述就是基于关键词语音样本检索（key word query）。工作的基础在于音频特征的选取，重点是音频的分类和匹配。这些问题与传统的语音检索工作基本一致，对于流媒体中的实时语音检测具有重要参考价值。可以说特征的有效选择和提取是前提基础，关键词样本和流媒体中有效信息的匹配是最终目的。同时我们注意到流媒体中语音检测也有其特殊性：语音信号处理的实时性具有较高要求，且大量的流媒体信息具有单边性，即对当前时刻以后所接受到的内容具有未知性。传统的音频检索方法[2,3,4,5]在实际应用中主要存在以下两个方面的问题：1）这些研究中所采用的媒体样本具有简单的分类特征。传统采用的视频文件具有很鲜明的类型特征，例如或者完全是语音新闻节目，或者完全是音乐的片段，这种条件下检索的准确度和速度还比较满意，但如果对象是各种类型片段混合的码流时结果不是很理想。2）在进行特征匹配的时候仅针对一种或者几种音频特征，采用某种匹配算法进行分类、匹配运算。如果选取较少特征则精度低，如果选取较多特征进行联合运算则运算量大，速度慢，不利于实时性分析。本论文目的就是通过学习国内外相关研究成果[1,30,32,35,39]，选取合理的音频特征，在此基础上研究出复杂度相对较低，计算量较小、检索较快并且实用的音频检索特征分类和匹配算法，取得的主要研究成果包括：

1.针对本论文研究重点简化了现有的多模态视频检索模型，给出了论文使用的基于音频的流媒体内容检索系统框架。Hoi 等人在[9,40]中提出的框架是基于

图像、音频、文本等多模态的，相对来说比较复杂。而本论文主要讨论的是基于音频模态对视频流媒体进行检索监控，为此，在原有模型的基础上给出了一个简化模型，使得讨论问题更加集中。

2. 选取了在音频分类匹配过程中有良好表现的音频特征，使音频检索过程更加有效。Ki-Man Kim 在[30]中选取了 4 个特征进行检索，计算量小、速度快，但是准确率不高；李恒峰等人[39]中采用了 98 维的音频特征向量获得了较高的准确率，但同时牺牲了检索时间。本文通过研究各特征值对音频分类的影响程度，选取了 25 维的音频特征向量该方法有在保证了分类准确率的条件下降低了特征向量维数，减少了检索过程的复杂度。

3. 在对音频进行分类时，改进了基于音频特征向量模板的分类方法，使其减少了计算冗余，提高了算法的实时性。传统的向量模板分类方法需要一次性计算所有特征值，再与向量模板进行比较从而对音频进行分类。经过改进后的算法引入了分层的概念，对音频进行逐次分类，在确认与样本音频不同类的同时避免了后序特征值的计算从而减少了检索过程的计算量，提高了效率。

4. 在音频分类时的分层思想基础上，在匹配过程中采用了前向有序匹配以提高匹配过程的速度，从而 进一步提高检索算法实时性。传统匹配方法采用计算特征向量之间欧式空间距离的方法进行匹配。本文在此基础上，融合了分类时的分层思想，在层中对特征值进行排序，然后在分类进行的同时进行特征前向匹配，通过门限判定减少不必要的后序计算，从而提高了算法速度。

论文首先研究了基于音频的视频检索的背景与意义；然后介绍本文使用的基于音频的视频流媒体检索框架；接着探讨了音频检索模块中的三个步骤：音频特征选取、音频分类和音频匹配并给出了实验结果。概括起来本文的章节安排如下：

第一章，绪论。主要阐述基于音频的视频内容检索的课题背景、研究现状和论文内容、结构。

第二章，基于音频的视频检索系统概述。介绍当前视频检索系统的架构以及本论文所使用的基于音频的流媒体视频检索系统框架，并对主要难点和做了简要介绍。

第三章，视频处理和检索中的音频特征。首先介绍不同种类的音频特征，然后阐述了针对音频帧和音频段的特征选取方法。

第四章，视频处理和检索中的音频分类与匹配。介绍如何利用所选取的特征值对视频进行快速分类匹配。针对流媒体的特征，提出了一种分层的向量模板分类算法（Hierarchical Classify Method Based on Vector Templates, HCMBVT）和改进的前向序列特征加权距离（Forwarding-sequential weight-feature distance

measuring, FWDM）匹配算法，通过实验给出了两种算法在性能方面的提升结果，并在最后给出了两种算法的融合以及系统平台的实现。

第五章，结论与展望。对本文进行了总结，并对今后的基于内容的视频检索研究进行了展望。

第二章 基于音频的视频检索系统概述

2.1 视频的结构化信息

一部视频持续的时间或长或短，长到几个小时的电影，短到几秒的单一镜头。如今那些视频网站上有大量的短视频，包含的内容单一，用到的检索方法并非基于视频内容的，而是基于视频外部的诸如标题、标签等的文本信息，这些信息已对视频从高于内容的层次作了人工的概括，因此以整个视频为单位还是可行的。但是一旦视频的检索范围变大，包括了长篇的新闻、电影等这些内容丰富的视频时，再以整个视频为检索单位明显是不合适的[38]。反之，以每一帧作为检索单位也是不合适的，因为即使再短的视频也会包含上百上千帧，视频一长这种检索方式更是不切实际。因此找到合适的检索单位，构建从底层特征到高层语义之间的桥梁使得检索更有效率是视频检索中的第一步。

通常，一段视频可以划分为几个场景，每个场景包含一个或多个镜头。每个镜头又由一系列连续记录的图像帧组成。因此，原始视频可以按照由粗到细的顺序划分为四个层次结构：视频(video)、场景(scene)、镜头(shot)、和图像帧(frame)，如图 1 所示：

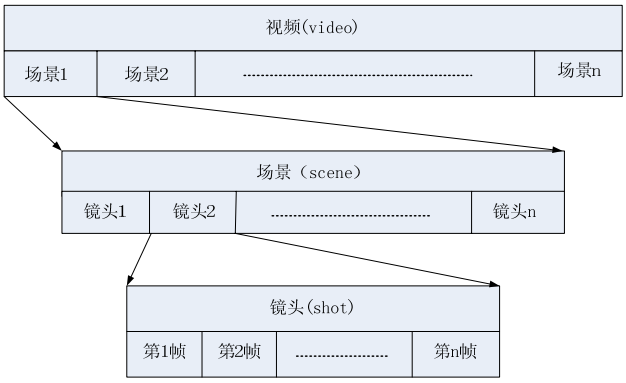


图 1 视频的结构化信息[36]

Figure-1 : Structure Information of Video [36]

帧(frame)是视频数据的最小单元，是一幅静止的画面。镜头(shot)是视频数据的基本单位，它是摄像头的一次连续的动作，只能拍摄相邻地点连续发生的事情。场景(scene)由内容相近的镜头组成，从不同的角度描述同一个事件。而视频(video)序列则由许多场景组成，叙述一个完整的故事结构。其中镜头是常用的一个层次，因为同一个镜头的场景是单一的或连续变化的，因此描述的内容往往比

较单一。一部较长的内容复杂的视频都是由多个内容单一的镜头组合而成，因此将视频分割成镜头，并以镜头为检索单位是合适的。然而一个镜头仍然会包含成百上千帧，甚至更多，如何利用视觉信息依然是个问题。Kin-Wai Sze 等人[10,12]从每个镜头中提取一幅或几幅关键帧来代表该镜头，然后可对关键帧进行图像分析以提取有用的视觉信息即图像的低层特征。而在多模态的视频检索系统中，经常加上将语音识别的结果，按时间点把每个词对应到各自的关键帧，于是视频检索被转化为结合了语音等信息的图像检索，其中还可以包括时间序列上的信息，比如运动信息和光照变化等。现代的基于内容的视频检索技术在检索过程中往往采用这种多模态的方式[8,9]。

2.2 典型的多模态视频检索框架

一般的视频文件中同时包含有视频信息和音频信息，在音频和视频数据中又各自包含了不同的语义和特征。如在视频信息中包含物体形状、颜色纹理、文字、运动方向以及拍摄相机的运动等等；而在音频信息中则包含有说话人信息、文字信息、语音类别以及语音物理特征等等。多模态的 CBVR 是联合以上这些信息特征（模态）进行检索的一种检索方法[40]，是近几年该领域的研究热点。视频信息流本质上是由文本、图像序列和音频等多态媒质交互融合形成的。由于视频是文本、图像序列和音频等多模态信息的综合体，每一模态都表示了丰富的语义信息，所以只有多模态的综合才能表示视频所蕴涵的完整语义信息。

Hoi 等人[9,19]研究了多模态的视频检索系统，并给出了一个典型的基于内容的多模态视频检索系统模型如图 2 所示：

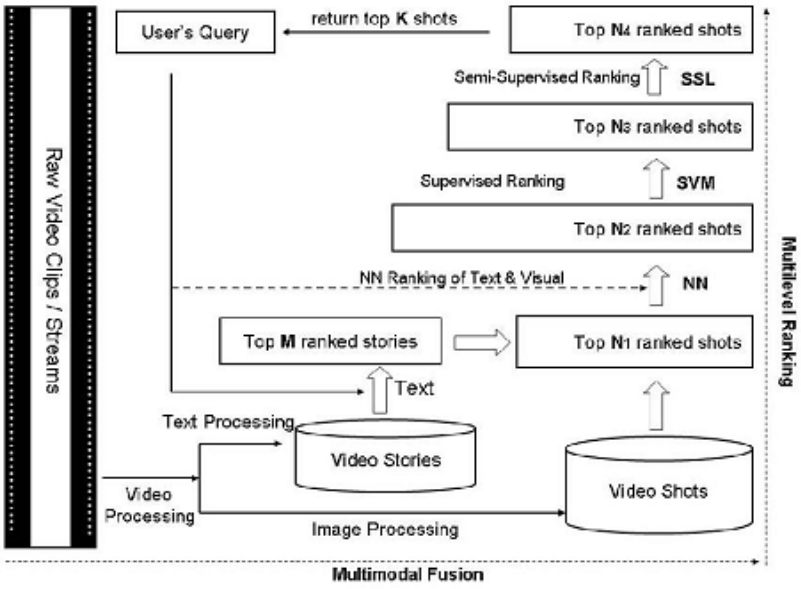


图 2 多模态的基于内容的视频检索的一般模型[9]

Figure-2 General Model of Multi-Mode Video Retrieval System based on Content [9]

根据本论文所研究的特殊应用，结合张静、俞辉等人的研究成果[40]，在此基础之上，我们可以概括出一个经过简化后的融合音频和视频特征信息的视频检索系统，如图 3 所示：

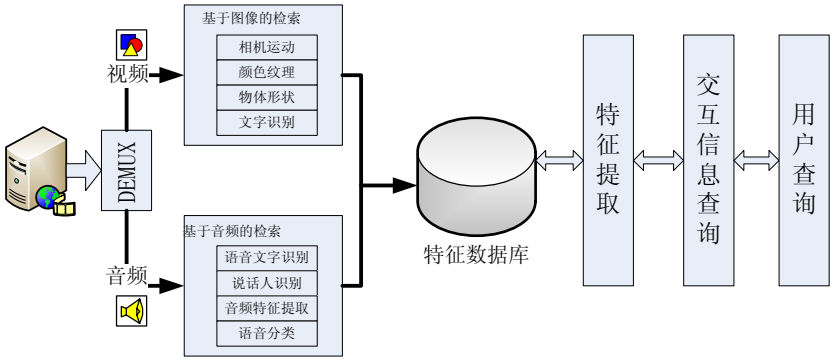


图 3 多模态的基于内容的视频检索的一般模型[40]

Figure-3 General Model of Multi-Mode Video Retrieval System based on Content

媒体数据首先通过解复用模块分成音频数据和视频数据，接下来分别通过各自的特征提取模块提取相应特征、语义，经索引后存入媒体特征数据库。当用户通过查询模块进行检索时，首先进过结构化查询解释模块提取查询条件，而后进行特征提取，最后把待查询的特征值与数据库中的相匹配，通过交互模块呈现返回给用户。由此可以看出该框架中有两个关键步骤：一方面对海量的媒体信息需要进行索引，把索引结果存入数据库；另一方面需要对用户输入的特征在索引文件中进行匹配。如何提高这两个步骤的执行速度成为本论文研究的重点问题。解

决问题的侧重点在于利用媒体文件中的音频信息，采用基于内容的音频检索技术，来获得较快的视频索引和匹配速度。

2.3 基于音频的视频流媒体内容检索框架

在 2.2 中，我们已经介绍了一种典型的多模态视频检索系统模型。在该模型中，利用了视频和音频的多模态检索方法，达到较理想的结果。但是该方法中采用的模态较多，计算量较大，检索、匹配时间均比较长，不适合在高实时性要求的环境下使用。

基于内容的音频检索技术是基于内容的视频检索的重要辅助方法。视频信息结构复杂，信息量具大，对其进行有效的组织管理并实现方便的查询检索面临很大的困难。开始时，研究者[10,12,16,17,34]的主要工作集中对视频流的处理，如镜头边界检测、关键帧提取和场景的合并等工作中，并且取得了许多卓越的成果。但是这些成果在实际运用中遇到了难题。从空间复杂度上来看，视频流是连续的静止图像组成的，以至于表达这些二维的图像所需要的数据量是同时语音数据量的成十甚至上百倍，所以视频索引的数据量要远远高于音频检索的数据量。其次，由于图像是二维数据，在进行图像处理、模式识别、机器视觉等等任务时涉及大量复杂的矩阵浮点运算 ，这无疑大大提高了视频检索的时间复杂度。基于视频域的检索需要从底层图像数据中提取出高层语义，但是如何从图像中抽取出所包含的物体并进行识别、语义提取一直是该领域的难题。与此相比，语音数据在这些方面有着天然的优势：语音是一维数据流，从数据量角度来说远远少于视频流；语音处理多数情况下只涉及线性运算，时间复杂度也要低于视频；最后，由于语言自身就包含有丰富的语义信息，从语音中提取出人类可以理解的语义比从图像中提取高层语义要简单的多。同时，音频检索技术在这样的背景下，研究者[2,3,4,36]把目光投向了音频，而基于内容的音频检索技术成为了视频检索的一种重要辅助方法。考虑到本论文研究的重点，在图 3 所示系统的基础上对其进行了简化，形成基于音频的视频内容检索系统模型，如图 4 所示：

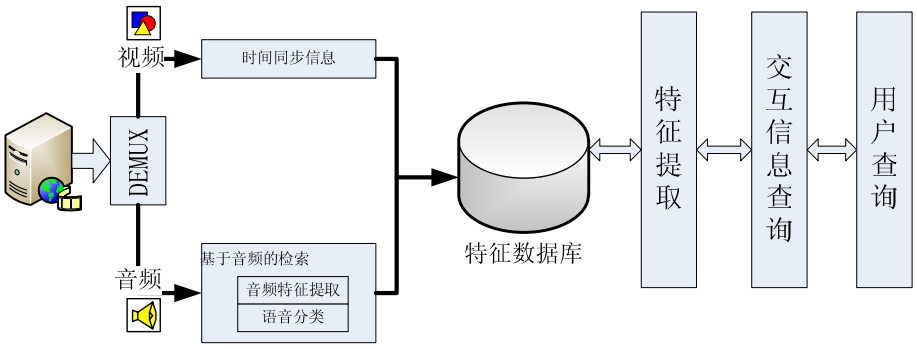


图 4 基于音频的视频内容检索系统模型

Figure-4 General model of Content based video retrieval system

由于本论文主要讨论基于音频的视频内容检索，所以，在图 3 的基础上，简化了视频特征提取模块，仅保留视频同步时间轴信息。而在基于音频的检索模块中，仅保留了音频特征提取和语音分类两种处理方法。这样，大大减少了系统的时间、空间复杂度，提高了系统的实时性响应性能。经实验验证，该系统也同样能获得较好的匹配结果。基于内容的音频检索技术在视频检索领域的应用主要有两种方式：基于文本和关键词的音频检索和基于语音特征的音频检索。

2.3.1 基于文本和关键词的音频检索

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[14]。这种检索方法的重要特征是需要利用语音识别技术把音频数据转换成文本形式：在索引阶段，引擎把媒体文件中的音频流中所包含的语音信息经过语音识别转化为文字信息，索引在数据库中；检索阶段引擎通过用户提供的关键字在数据库中进行文本查找。用户提供的文本关键字可以从三个渠道获得：

- 用户直接提供文本关键字，这是最直接也是最常用的方法；
- 通过分析用户检索需求中的文字描述而获得关键字；
- 用户提供查询样例，引擎从样例中利用语音识别技术提取文本关键字；

由此可以看出，基于文本和关键词的音频检索其本质依然是文字检索，问题的核心是从语音到文字的映射，这涉及到语音识别技术。这种方法的优点是比较直观，可以利用文本索引、检索这一目前已经非常成熟的技术，在各种平台上都有现成的开发包以供使用。缺点是索引、检索过程中都需要引入语音识别的过程，这增加了检索任务的空间、时间复杂度。由于语音识别与语言密切相关，如果需要检索不同语言的媒体文件，则需要提供不同语言中所有文字的语音特征库，这增加了识别的空间复杂度；而随着特征库的增大，在特征库中进行检索所需要的

时间也必然随之增加，这无疑又增加了时间复杂度。并且，这种方法仅针对语音文本检索有效，而不适用于非语音类音频。一种解决的方法是针对非语音文本类的音频文件，采用人工输入音频属性和描述的方法来得到关键字。目前网络上的MP3 搜索引擎大多数采用的就是这种方法。但是基于人工输入的属性和描述来进行音频检索也存在显见的缺点：

- 当数据量越来越多时，人工注释的工作量加大；
- 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；
- 不支持非语音类音频；
- 不能支持实时音频数据流的检索。

基于这些原因，基于文本和关键词的音频检索技术在实际应用中受到了很大的限制。从而产生了另外一种音频检索技术：

2.3.2 基于音频特征的音频检索

为解决上述问题，基于语音特征的音频检索技术应运而生[30]。基于语音特征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，音频可以分为以下几种类型：

- 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。
- 乐音：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。
- 静音：静默声音。
- 噪音：嘈杂无序的声音。
- 环境音：对模拟声音数字化得到的数字音频信号，它包括除语音、乐音之外的自然界的其他声音和合成的声音。
- 混合音：以上几种种声音中的一种或几种混合而成的音频。

不同的类型将具有不同的内在的内容。基于语音特征的音频检索，充分利用音频的听觉特性，直接针对音频特征进行处理，不要求目标音频中包含语音，从而拓宽了该检索方法可以应用的领域。不同于基于文本的音频检索方法，基于语音特征的音频检索可以用于：

- 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[8]；
- 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[7]；
- 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[30,35]；

正是由于这些原因，本论文图 4 中的基于音频检索模块中，也采用基于音频特征的内容检索技术。同时，我们注意到，要想在实时流媒体的应用环境下应用基于语音的视频内容检索技术，还应该对该系统做进一步改进。在写进特征数据库的同时，进行特征匹配，以实时的监测流媒体出现的匹配关键字。如图 5 所示：

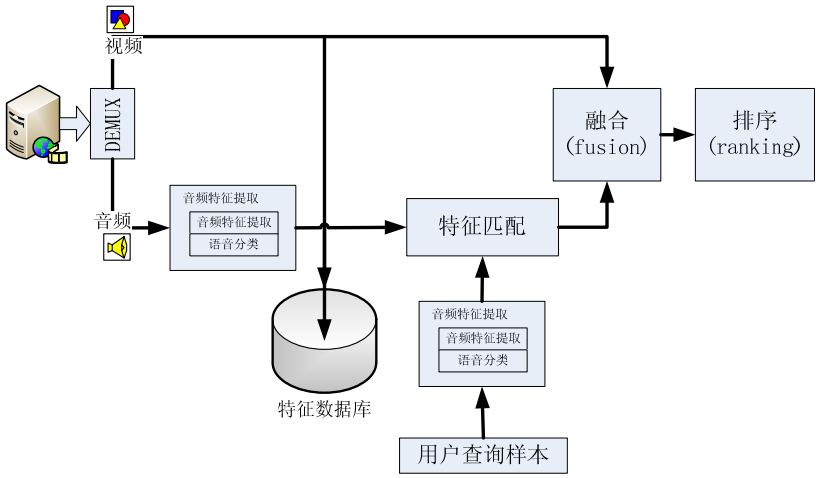


图 5 基于音频的视频流媒体内容检索的一般模型

Figure-5 General mode of video stream content retrieval system based on audio

由于所讨论的重点不同，本论文所采用的系统对视频流并不进行处理。音频流经过特征提取模块后，与视频同步时间轴信息一起写入特征数据库，以待以后查询使用。同时，针对时间窗内的音频类型、特征，与从用户输入的样本中提取出的类型信息和特征值进行匹配，针对匹配得到的结果与视频相融合（fusion），最后经过排序（ranking）后，把结果返回给用户。

2.4 本章小结

在本章中，我们首先介绍了视频检索系统的一般框架，分析了其特点与不足。在此基础之上介绍了最新的多模态视频检索系统模型。

接下来，针对本论文的研究方向我们着重讨论了基于内容的视频检索技术中的音频处理技术与难点。其中，介绍了音频检索的分类与各类型的特点与不足，分析了在本论文应用场景下所应采用的技术。

最后，结合以上所讨论的结果，给出了本论文采用的基于音频的视频流媒体内容检索系统框架。

第三章 视频检索中的音频特征分析与提取

音频特征分析与抽取是音频处理的基础，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和一般性。本章简要介绍了音频特征抽取的相关技术，重点分析不同特征域中音频的区别性特征及其抽取方法，最后对本文提出的特征的有效性进行实验分析。

3.1 音频特征的描述

要研究音频信息在视频中的应用，首先要用一些音频特征来刻画视频中的音频信息。由于音频信号具有短时性，即在一段短时间间隔里音频信号才可以保持相对稳定一致的特征。因此，在传统的音频处理和语音处理研究领域，更多的是考虑音频信号的短时特征[1~4,15]。用来提取短时特征的这些短的时间间隔称为帧（Frame），一般一帧的长度在 10~40ms 之间。如果过短，分析后将得到粒度过细的信息而不能反映各类音频的区别特性；如果过长，则容易导致音频特征平均化以后不能反映特征的时序变化特性。为了区别于视频中的帧的概念，在本文中用“音频帧”（Audio Frame）来表示提取短时音频特征的一小段声音。在音频帧上提取的短时音频特征，称之为音频帧特征（Audio Frame Features）。对于大多数的视频应用而言，通常要在一段相对较长的时间间隔内考察音频随时间变化的规律和统计特征。这些相对较长的时间间隔称为音频段（Audio Clips），一般一个音频段的长度为 1~2 秒。在某些特殊的应用中，也会用音频分割的结果作为音频段的边界。在音频段上提取的音频特征，称之为音频段特征（Audio Clip Features）。

如何有效地选取音频特征，是音频处理领域的一个重点研究方向，许多研究者在这方面做了大量工作[1~4,24,30,35,45]。音频特征值可以分为时域、频域和系数域特征，文章[45]作者为了分析音频特征值与音频分类的关系总结出了一个比较完整的音频特征值集合。所有这些特征值从不同角度描述了音频的特定特征，每种特征值的计算量不同并且每种特征值对在不同应用场合所起到的所用也有较大差距，文献[24,35,36]都对这方面进行了较深入的研究。在这些研究中，浦剑涛等人[45]选取了 98 维的音频特征向量，实验结果获得了较高的命中率和分类正确率，另一方面 Ki-Man [30]等人为了探讨快速音频检索的可行性仅选取了 4 个音频特征，也获得了比较理想的实验结果。

分析这些研究者的研究成果我们发现，在特征值的选取上有一个矛盾的地方，即向[45]那样想获得较理想的匹配命中率和分类准确率，我们不得不选取更多的音频特征作为实验判断的依据，但是这样会造成计算量的显著增加，不利于应用在实践当中；另一方面，如果如[30]所述那样选取较少的特征值，固然可以较好的满足实时性要求，但是匹配命中率和分类准确率在实践中不是非常理想。论文所针对的是流媒体实时内容监控这一特定的应用场景，要求我们在实时性和准确性之间寻求一个适用于实际应用的平衡点。正是基于这个目的，本文根据前人的研究成果[24,35,36]，首先选取了 38 与音频分类检索关系比较密切的特征值，在此基础上通过理论分析以及实验，最后简化为 25 维的特征向量作为本课题的实验基础。

论文主要分析、考察了四个方面的影响：首先是特征值主要用于区分何种音频；接着得到该特征值属于哪个取值区间的时候可以判定为其所属音频类型；然后给出了在分类模板中的取值；最后给出在匹配过程中，样本特征与音频特征之间的置信范围，即落在置信范围之内的音频特征匹配获得成功。为验证不同特征值与不同音频类型（纯语音、音乐、噪音）之间的相互联系，我们制作了一段特殊音频 test.wav。该段音频 0~15s 是纯语音，为一语言类节目的片段；15~30s 是纯音乐，为一交响乐片段；30~45s 为一段背景噪音。通过计算对比这个音频不同区间的音频特征值，可以容易的发现音频特征与音频类型之间的相互联系：

3.1.1 时域音频特征

- 均方根（Root Mean Square，RMS）：这是一个帧特征。该特征描述了语音帧的平均响度：

$$RMS[n]=\sqrt{\frac{\sum_{i=0}^{L_f-1}(frame[n,i])^2}{L_f}} \tag{3-1}$$

其中，$frame[n,i]$表示第 n 帧中的第 i 个采样值，L_f表示帧长度，下同。

RMS 的特点是计算量小，物理意义非常直观。RMS 是一个帧特征，表示了音频帧的平均响度，可以非常有效判断该音频帧是否是静音。通过实验可以得到 RMS 对音频分类的影响如表 1 所示。在其中分类命中区间是指归一化特征值落在该区间当中时认为音频帧为区分类别；均值指在基于模板分类方法中所采用的模板值；而置信范围是指在匹配时与样本特征值的最大误差范围。

表 1 RMS 对分类的影响

Figure-1 effect to audio classification of RMS

特征名称	区分类别	分类命中区间	均值	置信范围
RMS	静音	<0.1	0.05	±47%

在 RMS 的基础上，可以计算出一个音频段特征，静音比（Silence Ratio, SR）：该特征定义为 clip 中静音帧的数目与 clip 帧长之间的比例:

$$SR[n]=\frac{\sum_{i=0}^{L_c-1}Gate[i]}{L_c},$$
$$Gate[n]=\begin{cases}1, & when\ RMS[n]<\Gamma;\\ 0, & otherwise;\end{cases} \tag{3-2}$$

其中 Γ 为预设门限值。SR 的计算依赖于帧 RMS，故可以作为 RMS 的后序（见 4.3.2）音频特征来判断音频段是否为静音。

表 2 SR 对分类的影响

Figure-2 effect to audio classification of SR

特征名称	区分类别	分类命中区间	均值	置信范围
SR	静音	>0.9	0.95	±10%

- 过零率（Zero Crossing Ratio， ZCR）：语音、音乐以及背景音的过零率有很大不同，因此这是一个用途非常广泛的特征。定义为：

$$ZCR[n]=\frac{1}{2(L-1)}\cdot\sum_{i=1}^{L-1}\left|\mathrm{sgn}\left(x'[i+1]\right)-\mathrm{sgn}\left(x'[i]\right)\right|, \tag{3-3}$$
其中， $L=L_f\ or\ L_c$

其中 L_c 指音频段长度，下同。ZCR 是一个常用的音频帧特征，该特征计算量很小，却可以有效判断音频帧是否为噪音[24,35,36]，故在实践中应用广泛。

表 3 ZCR 对分类的影响

Figure-3 effect to audio classification of ZCR

特征名称	区分类别	分类命中区间	均值	置信范围
ZCR	噪音	>0.9	0.95	±10%

在 ZCR 的基础上，存在一种音频段特征，高过零率帧的比例 HZCRR（High Zero Crossing Rate Ratio）：音频段中过零率高于 1.5 倍平均过零率的音频帧的比例。

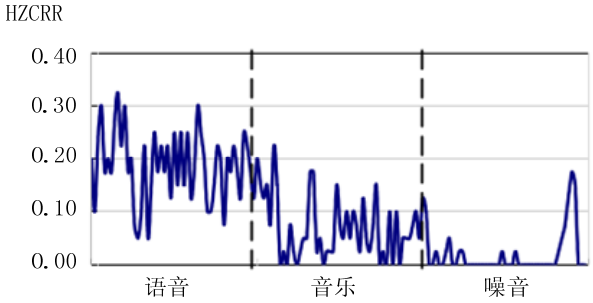


图 6 语音、音乐、噪音的 HZCRR 对比

Figure-6 Comparison of voice, music, noise, HZCRR

图 6 给出了语音、音乐和噪音的高过零率帧的比例 HZCRR 的变化曲线。可以发现语音信号的 HZCRR 均值 0.19 为最高，音乐 HZCRR 均值为 0.08 次之，噪音信号的 HZCRR 均值为 0.02 则基本接近于 0。需要指出的是，噪音信号的 HZCRR 为 0 时，意味着这时的噪声是白噪声。如果出现其他类型的噪声时，HZCRR 也可能会变得很大。HZCRR 的计算依赖于 ZCR，故 HZCRR 是 ZCR 的一个后序特征。

表 4 HZCRR 对分类的影响

Figure-4 effect to audio classification of ZCR

特征名称	区分类别	分类命中区间	均值	置信范围
HZCRR	噪音	<0.5	0.25	±25%

- 短时能量（Short Time Energy, STE）：

$$STE[n]=\sum_{i=0}^{L_f} frame[n,i]^2 \tag{3-4}$$

STE 的特点是计算量较小，物理含义明确，表示音频帧的能量，也即响度。可以有效区分是否为静音帧。对比(3-1)和(3-4)容易发现，RMS 的计算依赖于 STE，故可以在计算 STE 后再计算 RMS 以减少重复计算，也就是说 RMS 是 STE 的后序音频特征。

表 5 STE 对分类的影响

Figure-5 effect to audio classification of STE

特征名称	区分类别	分类命中区间	均值	置信范围

STE	静音	<0.1	0.05	±43%

在短时能量 STE 的基础上，我们采用了一种音频段特征，低短时能量率（Low Short Time Energy Ratio，LSTER）：音频段中，短时能量低于 0.5 倍平均短时能量的音频帧的比例。短时能量可以用来衡量音频信号的强度。我们知道噪声、音乐以及语音中的浊音都具有比较高的能量。但是对于语音来说，除了浊音之外，还包含清音和一些由于说话停顿而引起的短时间的静音，而这些清音和静音的能量比较低。因此我们可以计算在一段音频信号内短时能量低于某个值的音频帧的比例。

图 7 给出了语音、音乐和噪音的低短时能量率 LSTER 的变化曲线。与上面的分析相一致， LSTER 均值语音信号为 0.31 最高，音乐次之为 0.13，噪音最小仅为 0.02。

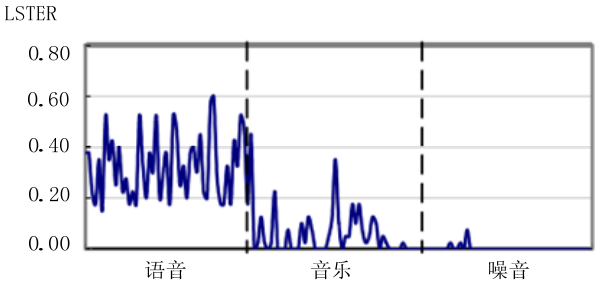


图 7 语音、音乐、噪音的 LSTER 对比

Figure-7 Comparison of voice, music, noise, LSTER

从图中可以发现，LSTER 可以有效判断音频段是否为语音。

表 6 LSTER 对分类的影响

Figure-6 effect to audio classification of LSTER

特征名称	区分类别	分类命中区间	均值	置信范围
LSTER	语音	>0.22	0.40	±35%

- 自相关系数（Auto Correlation Coefficients, ACC）：

$$ACC[n,l]=\sum_{i=0}^{L_f-l-1} frame[n,i]\,frame[n,i+l] \tag{3-5}$$

自相关系数反映了一段音频信号的规律性。自相关系数越大，音频信号越有规律。当一个音频帧的最大自相关系数小于一定的阈值时，我们认为这个音频帧是一个噪声音频帧。虽然 ACC 计算量很小，但通常不直接使用该特征值。在 ACC 此基础上，可以导出另一个音频段的特征，噪音率（Noise Frame

Ratio, NFR）:音频段中所含有噪音帧的比率。噪声率一定程度上反映了音频信号的时间特性。噪声信号没有很明显的规律，因此，一段噪声音频中的噪声音频帧的比例（即噪声率）要远远高于语音和音乐。从图 8 所示的噪声率曲线也可以看出这一规律。NFR 是判断音频段是否是噪音的重要特征。

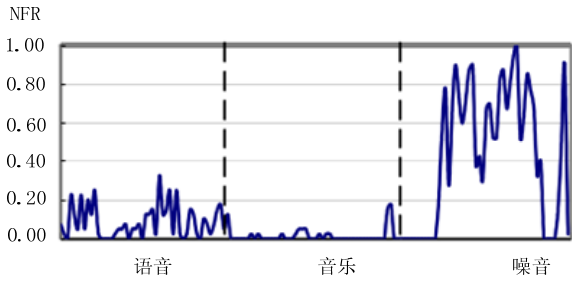


图 8 语音、音乐、噪音的 NFR 对比

Figure-8 Comparison of voice, music, noise, NFR

比较(3-3)和(3-5)可以看出，ZCR、HZCRR 的计算量要小于 NFR，故可以把 NFR 作为前两者的后序特征，在 ZCR、HZCRR 进行判断后计算 NFR 以提高分类、匹配准确率同时减少前期计算量。

表 7 NFR 对分类的影响

Figure-7 effect to audio classification of NFR

特征名称	区分类别	分类命中区间	均值	置信范围
NFR	噪音	>0.5	0.65	±25%

3.1.2 频域音频特征

相比较时域特征而言，频域音频特征的计算量明显增大，故大部分频域特征均作为时域特征的后序特征，以减少不必要计算量，这点在 4.3.1 中也有体现。DFT 系数是计算频域特征的基本量。DFT 系数：DFT(n,k)表示第 n 帧的第 k 个离散傅立叶变换系数。DFT 系数是计算其他频域特征的基础，通常用快速傅立叶变换（FFT）来计算：

$$DFT[n,k]=\sum_{m=0}^{M-1} frame[n,m]e^{-j\frac{2\pi m}{M}k},\quad k=1,...,M \tag{3-6}$$

其中，k=1,…,M，M 是 DFT 系数的阶。

- 频谱质心（Spectral Centroid, SC）：又称为音频亮度（audio brightness），指一个音频帧的频谱能量分布的平均点：

$$F_c[n]=\frac{\sum_{k=1}^{K}k\left|DFT[n,k]\right|^2}{\sum_{k=1}^{K}\left|DFT[n,k]\right|^2} \tag{3-7}$$

频谱质心 SC 反映了音频信号在频谱能量分布上的特性。在语音中，不同的音素由于频率特性的不同，其频谱能量的分布也不同，因此在语音段中频谱质心的变化会比较频繁。而音乐和噪声一般具有比较连续的频谱能量特性，频谱质心在一段时间内会比较稳定。因此我们可以用一个音频段内所有音频帧的频谱质心的方差来分辨语音和非语音。

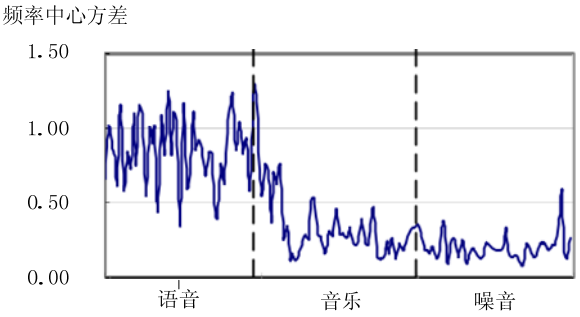


图 9 语音、音乐、噪音的 SC 方差对比

Figure-9 Comparison of voice, music, noise, SC

从图 9 中的频率中心曲线可以看出，语音片段的频谱质心方差要远远大于音乐和噪音。

表 8 SC 方差对分类的影响

Figure-8 effect to audio classification of SC variance

特征名称	区分类别	分类命中区间	均值	置信范围
SC 方差	语音	>0.5	0.75	±25%

- 带宽（bandwidth）：频域各频率值与频率中心差值的能量加权平均值，衡量音频频域范围指标。

$$BW[n]=\sqrt{\frac{\sum_{k=1}^{K}(k-F_c[n])^2\left|DFT[n,k]\right|^2}{\sum_{k=1}^{K}\left|DFT[n,k]\right|^2}} \tag{3-8}$$

一般来说，语音的带宽范围在 0.3kHz~3.4kHz；音乐的带宽较宽，为 22.05kHz 左右，可以用这个特性来区分语音和乐音。

表 9 BW 对分类的影响

Figure-9 effect to audio classification of BW

特征名称	区分类别	分类命中区间	均值	置信范围
BW	语音	<0.16	0.1	±27%

- 子带能量比（Sub-Band Energy Ratio, SBER）：把频域分成四个子带：$\left[0,\frac{\omega_0}{8}\right]$，$\left[\frac{\omega_0}{8},\frac{\omega_0}{4}\right]$，$\left[\frac{\omega_0}{4},\frac{\omega_0}{2}\right]$，$\left[\frac{\omega_0}{2},\ \omega_0\right]$，分别计算各子区间的能量比。能量比衡量了能量在各子带的分布，不同类型语音其分布有所不同：

$$SBER[n,i]=\frac{1}{E[n]}\sum_{k=L_i}^{H_i}\left|DFT[n,k]\right|^2$$
$$E[n]=\sum_{k=1}^{K}\left|DFT[n,k]\right|^2 \tag{3-9}$$

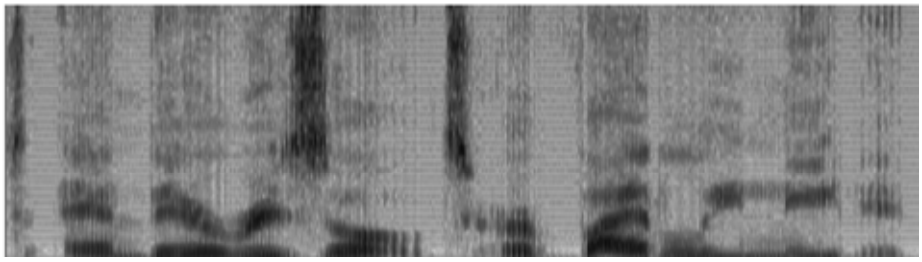


图 10 语音的频谱图

Figure-10 Spectrum of voice

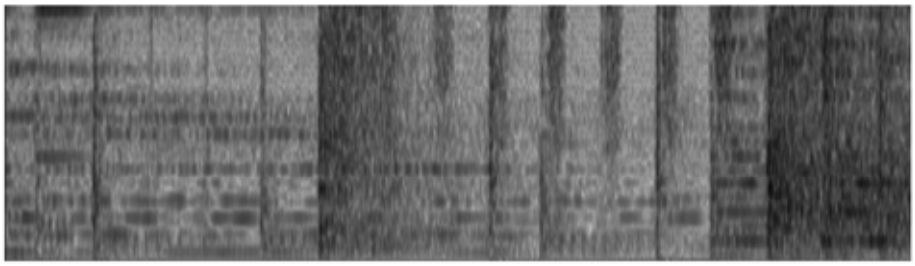


图 11 音乐的频谱图

Figure-11 Spectrum of music

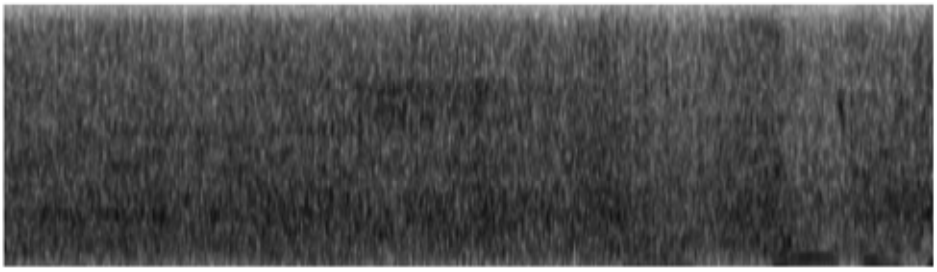


图 12 噪音的频谱图

Figure-12 Spectrum of noise

图 10~图 12 是三段长度为 2 秒的语音、音乐和噪声的频谱图（spectrogram）。频谱图直观地反映了音频信号不同频率上的能量分布。其中横轴代表时间，纵轴代表频率。图中颜色的深浅代表某一时刻在某个频率上的能量的大小。颜色越深，能量越大。 3-9 定义的子带谱能量从某种程度上可以看作是对频谱图的量化。它先把整个频带分成若干个子带，然后计算每个子带能量的大小。从图 10~图 12 可以看出，噪声信号在整个频带内的能量分布比较均匀，而语音信号（尤其是浊音）和音乐信号在某几个频段内具有比较高的能量。所不同的是，音乐信号的高能量频带分布更具有规律性。

3.1.3 系数域音频特征

从人类发声的角度来看，不同说话人之间的差异由先天因素引起，同时受到后天因素的影响。人的发声器官的基本构造是相同的，但是由于人的多样性，不同的人发声器官的生理尺寸会有所差别。比如成年男性说话人的声带比较宽厚，语音的基音频率就比较低，大致分布在 60~200Hz；而女性说话人和小孩的声带较窄，语音的基音频率就比较高，大致分布在 200~450Hz。除了先天的发声器官的差异之外，后天的环境因素也会影响说话人。由于社会环境、人生经历、性格等一些外在因素因人而异，不同的人在说话时会有不同的习惯和风格，即使在发同样的音时，发声器官的动作也会有所差异。同时，人在说话时的场景和心理状态也会影响实际的发音。从人类语音感知的角度来看，人往往不是从声音本身的差别来分辨不同说话人的。诸如说话的习惯和风格等一些间接的高层次的说话人特点对人区分不同说话人起到了很关键的作用[31]。

尽管物理上的一些音频特征并非人类分辨说话人的主要依据。但是通过近些年的研究[6]，还是找到了一些具有一定说话人表征能力的物理特征。这些特征可以在大多数的应用场合起到分辨说话人信息的作用。目前公认比较有效的用于刻画说话人信息的物理特征主要有三个：线性预测系数（Linear Prediction Coefficients, LPCs）、Mel 频率倒谱系数（Mel-Frequency Cepstrum Coefficients）

和基音频率（Pitch）。线性预测系数是一种语音信号的短时衡量指标，它把语音信号看成全极点滤波器的输出，而 LPCs 为滤波器系数。

相比通过对人的发声机理的研究而得到的声学特征 LPCs，Mel 频率倒谱系数是受人的听觉系统研究成果推动而导出的声学特征。由于充分模拟了人的听觉特性，而且没有任何前提假设，MFCCs 具有识别性能和抗噪能力。经常用到的 MFCC 有 12 维、13 维(加入 F_0 能量)、39 维(13 维的 MFCCs 加上 13 维的一阶导和 13 维的 二阶导)。

3.1.3.1 线性预测系数

线性预测[31]作为一种工具，几乎普遍应用于语音信号处理的各个方面。这种方法是最有效的和最流行的语音分析技术之一。在各种语音分析技术中，它是第一个真正得到实际应用的技术。语音处理又有许多突破，但这种技术目前仍然是唯一的最重要的分析技术基础。线性预测的基本思想是：对音频信号的各个取样值，可以用它过去的若干个取样值的线性组合来表示；各加权系数的确定原则是使预测误差的均方值最小。预测误差定义为真实取样值与预测值之差。如果利用过去 p 个取样值来进行预测，称为 p 阶线性预测。

参数模型法是现代谱估计的主要内容，经常采用的模型由三种：

1) 自回归(auto-regressive, AR)模型是一个全极点的模型；

2) 移动平均(move-average, MA)模型是一个全零点的模型；

3) 自回归-移动平均(ARMA)模型是一个既有零点，又有极点的模型。

从数字信号处理的知识可知，AR 模型易反映频谱中的峰值，MA 模型易反映频谱中的谷值，而 ARMA 模型可以反映以上两者。在实际上最常用的模型是全极点模型。

根据参数模型功率谱估计的思想，可以将语音信号 x(n)看成是由一个输入序列 u(n)激励一个全极点的系统模型 H(z)而产生的输出：

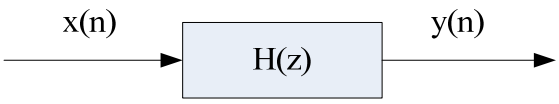


图 13 语音信号的线性模型

Figure-13 Linear model of voice signal

对于全极点函数模型，传递矩阵为：

$$H(z)=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}} \tag{3-10}$$

其中，G 为常数，为系统增益；a_i 为实系数；p 为系统模型的阶数。用系数$\{a_i\}$来定义一个 p 阶的线性预测器：

$$F(z)=\sum_{i=1}^{p}a_i z^{-i} \tag{3-11}$$

可将这个 p 阶线性预测器从时域的角度理解为，用信号的前 p 个样本来预测当前的样本得到如下预测值：

$$\hat{x}(n)=\sum_{i=1}^{p}a_i x(n-i) \tag{3-12}$$

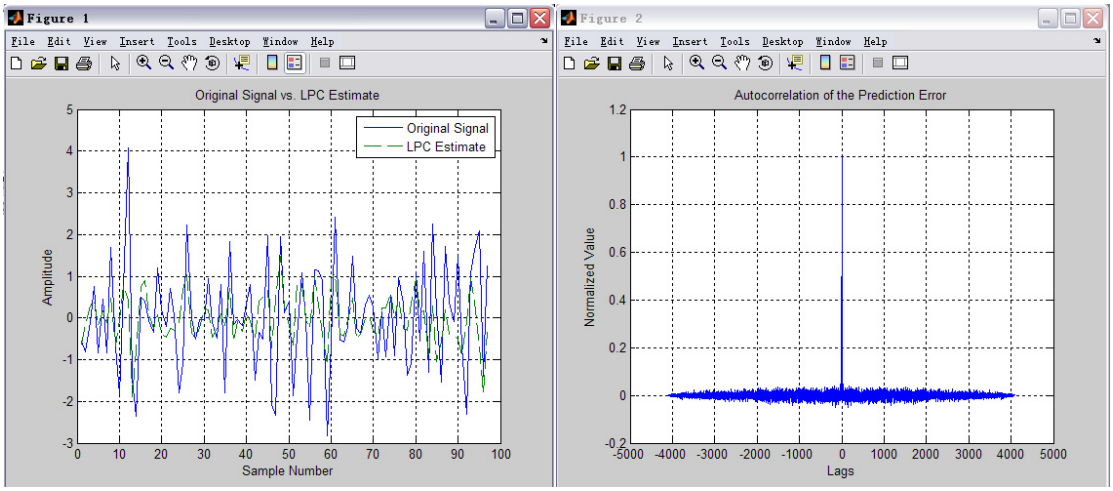
因为线性预测器$F(z)$使用 AR 模型的系数$\{a_i\}$来构造的，而 AR 模型是在最小均方意义上对数据的拟合，所以线性预测器$F(z)$必然是一个最佳预测器，即此时预测器的预测误差短时能量最小。

语音信号的线性预测分析就是根据这一性质，从语音信号 x(n)出发，依据最小均方误差准则，估计出一组线性预测器的系数$\{a_i\}$，这就是所要求的信号 AR 模型的系数，$\{a_i\}$被称为线性预测系数或 LPC 系数。预测器的预测误差为：

$$e(n)=x(n)-\hat{x}(n)=x(n)-\sum_{i=1}^{p}a_i x(n-i) \tag{3-13}$$

为了在最小均方误差意义上计算最佳预测系数，定义短时预测均方误差为：

$$\begin{aligned}E[n]=\sum_{i=1}^{L_f}e[n,i]^2&=\sum_{i=1}^{L_f}\left(frame[n,i]-frame'[n,i]\right)^2\\&=\sum_{i=1}^{L_f}\left(frame[n,i]-\sum_{j=1}^{p}a_j frame[n,i-j]\right)^2\end{aligned} \tag{3-14}$$



（左）LPC 预测信号与原信号比较　　（右）预测误差

图 14 LPC 预测

Figure-14 LPC

由于语音信号的时变特性，线性预测分析应该在短时的语音段上进行，即按帧进行。因此，上式的求和通常也是在一帧语音的范围内进行。使上式中的达到最小，$\{a_i\}$ 必须满足：

$$\frac{\partial E[n]}{\partial a_k} = 0 \tag{3-15}$$

即：

$$\begin{aligned}\frac{\partial E[n]}{\partial a_k} &= -2\Big(\sum_{i=1}^{L_f} frame[n,i]\,frame[n,i-k] \\ &\quad -\sum_{j=1}^{p} a_j \sum_{i=1}^{L_f} frame[n,i-k]\,frame[n,i-j]\Big) \\ &= 0\end{aligned} \tag{3-16}$$

根据(3-16)得到：

$$\sum_{i=1}^{L_f} frame[n,i]\,frame[n,i-k] = \sum_{j=1}^{p} a_j \sum_{i=1}^{L_f} frame[n,i-k]\,frame[n,i-j] \tag{3-17}$$

这样可以得到以 $\{a_i\}$ 为变量的线性方程组。若定义：

$$\varphi(n,k,i) = \sum_{j=1}^{L_f} frame[n,j-k]\,frame[n,j-i] \tag{3-18}$$

则(3-17)可以写成：

$$\varphi(n,0,k)=\sum_{j=1}^{p}a_j\varphi(n,j,k) \tag{3-19}$$

(3-19)为 p 个包含 p 个未知数的线性方程组，求解方程组就可得到线性预测系数 $\{\hat{a}_i\}$。

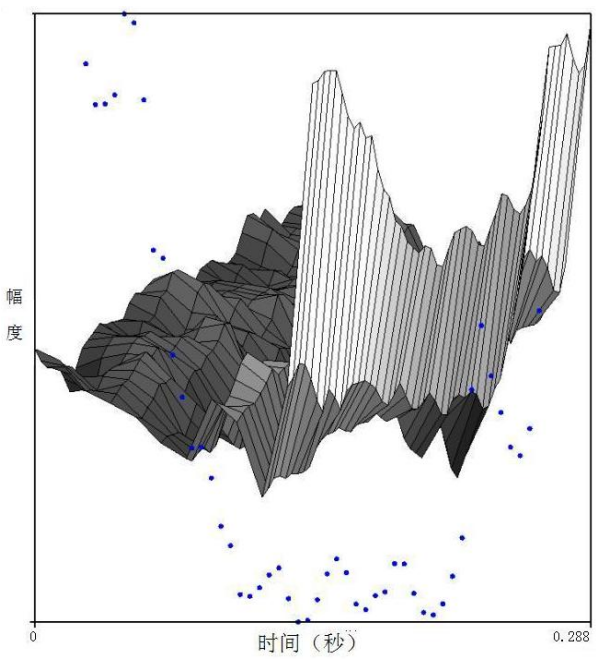


图 15 “帐篷”语音短时 LPCC 系数图

Figure-15 short-time LPCC of “zhangpeng”

图 15 显示了一段语音的短时 LPCC 系数图。横轴为时间轴，纵轴为系数幅度。Z 轴为系数维度坐标轴。

3.1.3.2 Mel 频率倒谱系数

Mel 频率倒谱系数(MFCC)[4,33]是在频谱上采用滤波器组的方法计算出来的，将语音频率划分成一系列三角形的滤波器序列,这组滤波器在频率的 Mel 坐标上是等带宽的。这是因为人类在对 1000Hz 以下的声音频率范围的感知遵循近似线性关系；对 1000Hz 以上的声音频率范围的感知不遵循线性关系，而是遵循在对数频率坐标上的近似线性关系。

与普通实际频率倒谱分析不同，MFCC 的分析着眼于人耳的听觉特性，因为,人耳所能听到的声音的高低与声音的频率并不成线性正比关系,而是 MEL 频率尺度则更符合人耳的听觉特性。所谓 MEL 频率尺度,它的值大体上对应于实际频率的对数分布关系。Mel 频率与实际频率的关系可用下式近似表示：

$$Mel(f)=2595\cdot\log_{10}\left(1+\frac{f}{700}\right) \tag{3-20}$$

或者是

$$Mel(f)=1127\cdot\ln\left(1+\frac{f}{700}\right) \tag{3-21}$$

式中 f 为频率，单位是 Hz。其提取的流程框图如下：

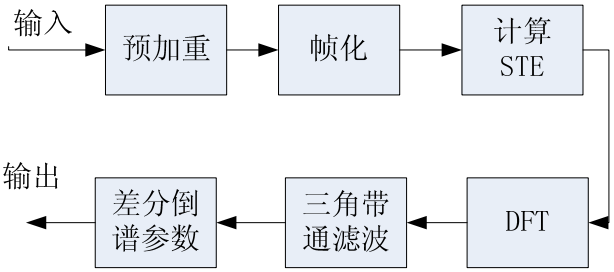


图 16 MFCC 计算的流程框图

Figure-16 Flowchart of MFCC calculation

帧化以后的信号首先需要计算短时能量（STE）。短时能量代表着音量的高低，亦即声音振幅的大小，可以根据此能量的值来过滤掉语音信号中的一些细微噪声。短时能量的定义见(3-4)。由于语音信号的时域上的变化快速而不稳定，所以通过 DFT 将它转化到频域上来观察，此时它的频域会随着时间的变化作缓慢的变化。所以通常将加窗后的帧经过 FFT 求出每帧的频谱参数。

将每帧的叛逆谱参数通过一组 M 组三角带通滤波器(一般 M 为 24~40 组)所组成的梅尔刻度滤波器，在频域对能量谱进行带通滤波。Mel 频率滤波器组为在语音的频谱范围内设置的若干个带通滤波器 $H_m(f)$，其中心频率为 $f_c[m]$，m=1,2,…M，每个滤波器具有三角形滤波特性，m 值小时相邻 $f_c[m]$ 之间的间隔也小，随着 m 的增加相邻 $f_c[m]$ 的间隔逐渐变大，每个带通滤波器的传递函数为：

$$H_m(f)=\begin{cases}0 & ,\ f\le f_c[m-1]\\ \dfrac{2\left(f-f_c[m-1]\right)}{\left(f_c[m+1]-f_c[m-1]\right)\left(f_c[m]-f_c[m-1]\right)} & ,\ f_c[m-1]\le f\le f_c[m]\\ \dfrac{2\left(f_c[m+1]-f\right)}{\left(f_c[m+1]-f_c[m-1]\right)\left(f_c[m+1]-f_c[m]\right)} & ,\ f_c[m]\le f\le f_c[m+1]\\ 0 & ,\ f\ge f_c[m+1]\end{cases} \tag{3-22}$$

将每个频带三角滤波的输出取对数，求出每一个输出的对数能量再将此 M 组参数进行余弦变换，求出 M 阶的梅尔倒谱系数。余弦转换公式为：

$$C_m = \sum_{i=1} STE[i] \cos\left(m\left(i - \frac{1}{2} \right) \frac{\pi}{N} \right), m = 1, 2, ..., M \qquad (3\text{-}23)$$

语音信号的 Mel 频率倒谱具有如下性质：

- 倒谱的低时部分对应语音信号的声道分量，且按 1/n 的趋势随 n 的增加而衰减，故用维数不多的倒谱向量足以表征语音的声道分量。
- 倒谱的高时部分对应于语音信号的音源激励分量。

这表明最前若干维以及最后若干维的 MFCC 系数对语音区分性能影响较大，MFCC 系数个数通常取最低的 12~16 阶。与线性预测倒谱分析相比，MFCC 的优点是不依赖全极点语音信号产生模型的假定,因而在噪声环境下表现出更强的鲁棒性，在非特定人说话人识别方面有利于减小因说话人不同的差异可能带来的影响[4,33]。

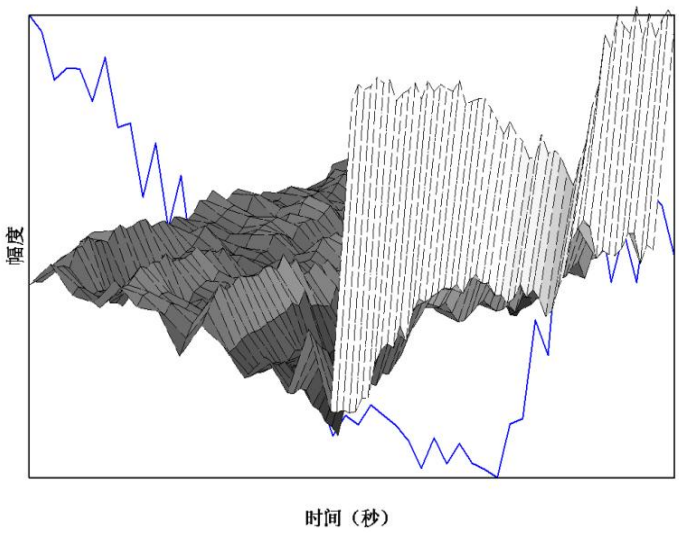


图 17 “帐篷”语音短时 MFCC 系数图

Figure-17 short-time MFCC of “zhangpeng”

图 17 显示了一段语音的短时 MFCC 系数图。横轴为时间轴，纵轴为系数幅度，Z 轴为系数维度坐标轴。

3.2 音频特征分析和提取相关技术[31]

语音信号是一种典型的非平稳信号，对非平稳信号采用短时平稳方法来分析，经研究在 10ms～30ms 内，人的发音模型是相对稳定的，所以在这样一个短的时

间间隔内，可用平稳信号的分析方法来处理声音信号。实际应用中，可以利用数字信号处理技术和信号系统理论来抽取音频的物理特征。对音频特征的抽取要用到多种方法，其中短时处理技术（包括时域和频域的短时处理技术）和同态处理技术是最基本、最典型的技术。

3.2.1 音频短时处理技术

音频是一个非平稳随机过程，其特性是随时间变化的，但这种变化是很缓慢的。鉴于此，可以将音频信号分成一些连续的短小片段，以保证在这些短小片段内音频特性保持平稳，从而对这个平稳的音频段进行处理，这就是短时处理技术。这些短小片段一般长为 20～30ms，称为音频帧。注意这里说的音频帧与视频流中帧的概念是不同的。相邻帧可以有部分重叠，以保持帧间音频特性的平稳过渡。每一帧可以看成是从一个具有固定特性的持续音频中截取出来的，这个持续音频通常认为是由该音频帧周期性重复得到的。因此，对每个音频帧进行处理就等效于对持续音频的一个周期进行处理，或者说等效于对固定特性的持续音频进行处理。

短时处理技术根据在研究域上的不同分为短时时域处理技术和短时频域处理技术。语音信号本身就是时域信号，因而时域分析法是最早使用、理解最直观、应用范围最广的一种语音分析方法。时域处理主要是计算音频的短时能量、短时平均幅度、短时平均过零率和短时自相关函数。这些计算都是以音频信号的时域抽样为基础的。因为语音是一个非平稳过程，因此适用于周期、平稳随机信号的标准傅立叶变换不能用来直接表示语音信号，而应该用短时傅立叶变换对语音信号的频谱进行分析，相应的频谱称为“短时谱”。

傅里叶变换的短时谱：

$$F_n\left(e^{j\omega}\right)=\sum_{m=0}^{L_f} frame[n,m]e^{-j\omega m} \tag{3-24}$$

相应的功率谱为：

$$S_n\left(e^{j\omega}\right)=\left|F_n\left(e^{j\omega}\right)\right|^2 \tag{3-25}$$

在语音信号数字处理中，功率谱具有重要的意义，在一些语音应用系统中，往往都是利用语音信号的功率谱。

时域分析与频域分析都具有一定的局限性，语音信号是一种复杂的信号，单纯的时域分析其能力是有限的，频域分析则会丢掉语音信号的时变特性。为此，人们一直都在探索时域和频域相结合的分析方法，如语音分析中常用的语谱图，

以及近年来新型的小波时频分析方法，它们都是将时域和频域特性相结合来分析语音信号的。

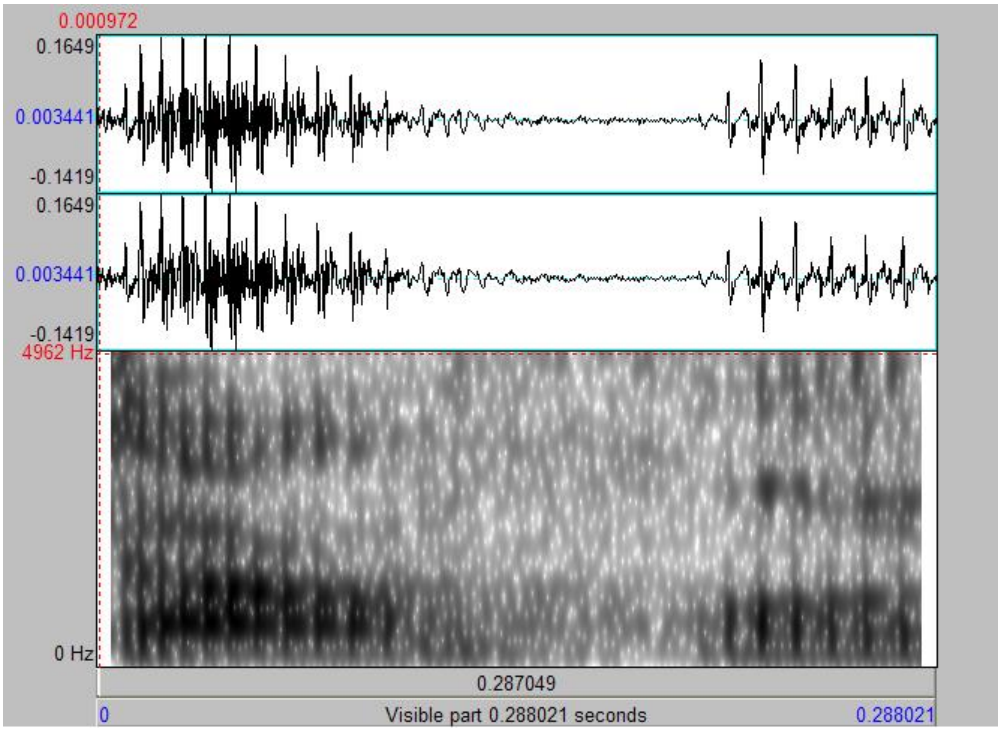


图 18 “帐篷”语音的波形图和语谱图

Figure-18 Wavelet and Spectrum of “Zhangpeng”

从图中可以看出，语谱图上因其不同的灰度，形成不同的纹路，称之为“声纹”。声纹因人而异，可以作为说话人识别的一种依据。

3.2.2 同态处理技术

根据声音信号的产生模型，音乐和语音都可以看作是一个线性非时变因果稳定系统 H[Z]受到信号 E[Z]激励后产生的输出。对乐器而言，不同音调的音，有的是激励信号 E[Z]的变化产生的，如弦乐器，有的是系统函数 H[Z]的变化产生的，如吹拉乐器。有的是 E[Z]和 H[Z]同时变化产生的，如钢琴。对于语音来说，声音的变化是由系统函数 H[Z]和激励信号 E[Z]的共同作用产生的。为我们有必要采用一定的方法将这两者有效的分开，这个方法就是同态滤波。滤波的过程是将卷积处理化为乘积，然后作对数处理，使之化为可分离的相加成分，结果就形成了倒谱。倒谱描述了说话人的声道分量，故是非常有效的说话人个性特征参数。

在时域上，这些音频信号 v[n]可以看成是系统的单位冲激响应 h(n)和激励信号 e(n)的卷积。由卷积信号求得参与卷积的各个信号是数字信号处理领域中普遍遇到的一项共同的任务。解决此任务的算法称为解卷积算法，同态信号处理是解卷积算法的一种。

3.2.3 预处理技术

- 预加重

因为发声过程中声带和嘴唇的效应，使得高频共振峰的振幅低于低频共振峰的振幅，进行预加重的目的就是为了消除声带和嘴唇的效应，来补偿语音信号的高频部分。方法是将经采样的数字语音信号 x[n]通过一个高通滤波器(High Pass Filter)：

$$H(z)=1-\alpha z^{-1},\ 0.9\le\alpha\le1.0 \tag{3-26}$$

其中α为与采样率相关的经验值，一般取 0.95。经过预加重的信号为：

$$x'[n]=x[n]-\alpha x[n-1] \tag{3-27}$$

- 帧化和加窗

音频帧(frame)是音频处理中的基本单位。取 10~20ms 为一帧。为了避免相邻两帧的变化过大，所以帧与帧之间需要重叠一部，一般为二分之一或三分之一，也就是每次位移一帧的二分之一或三分之一后再取下一帧，这样可以避免帧与帧之间的特征变化太大。

$$\begin{aligned} &frame[n]=\{i\in\theta\mid x'[i]w[i]\},\\ &\theta=\{n(L_f-\delta),n(L_f-\delta)+1,...,n(L_f-\delta)+L_f-1\} \end{aligned} \tag{3-28}$$

其中，$w[n]$为窗函数，L_f为帧长度，δ为帧间重叠部分的长度。将每一帧代入窗函数，其目的是消除各帧两端可能造成的信号不连续性，常用的窗函数有方窗、汉明窗和汉宁窗等，根据窗函数的频域特性常采用汉明窗。$w[n]$定义如下：

$$w[n]=\begin{cases}0.45-0.46\cos\left(\dfrac{2\pi n}{L_f-1}\right), & 0\le n\le L_f\\ 0, & otherwise\end{cases} \tag{3-29}$$

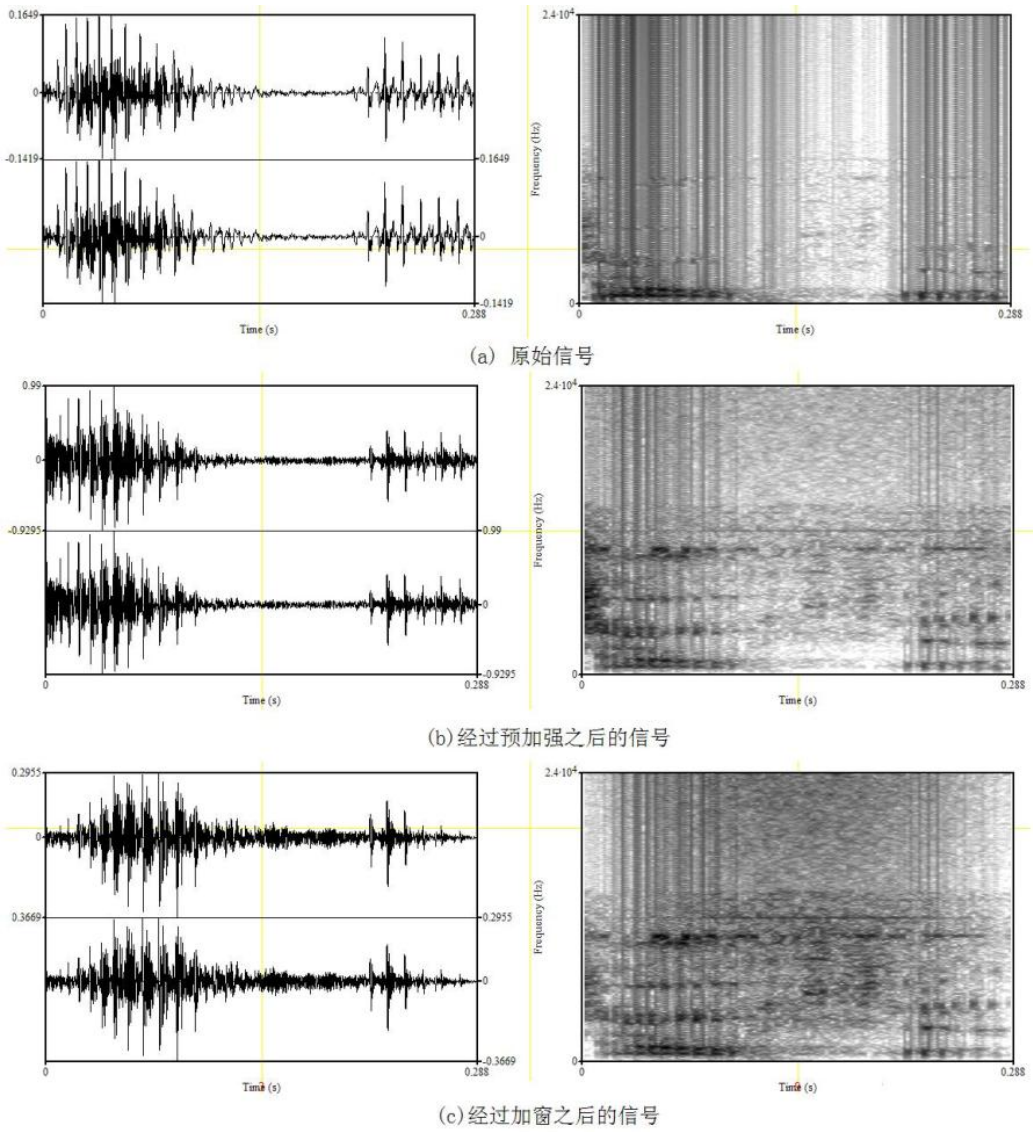


图 19 “帐篷”语音信号预处理前后波形、频谱的对比

Figure- Comparison of wavelet and spectrum of “zhangpeng”

- 片段化

片段（clip）是本实验中语音处理的抽象语义单位。是进行语义分析和匹配的基本单位，一般取 1~2s 为一个音频段。音频段的定义如下：

$$clip[n]=\{i\in\theta\mid frame[i]\},\theta=\{nL_c,nL_c+1,...,(n+1)L_c-1\} \qquad (3\text{-}30)$$

- 音频特征归一化

对音频帧和段计算得到其音频特征以后，需要对得到结果进行归一化处理，以消除不同音频特征值之间的差别。方法如下：

$$\phi_i' = \frac{\phi_i - \mu_i}{\sigma_i} \tag{3-31}$$

ϕ_i为所选取的第 i 个特征值，ϕ_i'为归一化后的特征值。实验表明系数域特征值归一化后效果不理想，所以对系数域特征值不进行归一化处理。

3.3 音频特征的选取及实验

在音频处理过程中，通常用到许多特征量，本论文首先对这些特征量组成的特征向量进行降维处理以减少在实验过程中的计算量。结合 Shieh, J 和冯哲等人的研究成果[4,35,36,45]，通过对媒体库中已知类型的媒体样本文件进行统计分析，我们选取的特征全集是：过零率(Zero-Crossing Rate, ZCR)、高过零率帧比率(High Zero-Crossing Rate Ratio, HZCRR)、短时能量(Short-Time Energy, STE)、短时能量均方值(Root-Mean Square, RMS)、低能量帧比率(Low Short-Time Energy Ratio, LSTER)、静音帧比率(Silent Frame Ratio, SFR)、频谱差分幅度(Spectrum Flux, SF)、频谱质心(Spectral Centroid, SC)、频谱宽度(Band Width, BW)、频谱截止频率(Spectral Rolloff Frequency, SRF)、噪音帧比率(Noise Frame Ratio, NFR)、子带能量比(Sub-Band Engergy Ratio, SBER)、线性预测倒谱系数(LPCC)、线谱对(Line Spectrum Pair, LSP)、梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)。16 种音频特征，其中 SBER 采用 4 个子带，LPCC 为 10 阶，MFCC 为 13 阶，

表 10 音频特征全集

Table-10 audio feature set

维数	特征	维数	特征
1	ZCR	8	SC 方差
2	HZCRR	9	BW
3	STE	10	SRF
4	RMS	11	NFR
5	LSTER	12~15	SBER
6	SFR	16~25	LPCC
7	SF	26~38	MFCC

在 3.1 中，在 Shieh, J 和冯哲等人的研究[4,35,36,45]基础上，通过分析各特征值定义所代表的声学含义以及实验的统计分析结果，我们得到了这些特征与音频分类之间的联系。其中类型均值是指当音频帧属于区分类别时其特征值的统计均值，也是基于向量模板分类方法中所使用的模板值；命中区间选择的是均值附近包含 90%样本点的区域范围；而置信范围则是样本点在均值附近分布的标准方差，结果如表 11 所示，其中各特征值已经用 3.2.3 所述预处理技术归一化：

表 11 音频特征值对分类的影响

Figure-11 effect to audio classification of multiple audio features

特征名称	区分类别	分类命中区间	类型均值
RMS	静音	<0.1	0.05
SR	静音	>0.9	0.95
ZCR	噪音	>0.9	0.95
HZCRR	噪音	<0.42	0.24
STE	静音	<0.1	0.05
LSTER	语音	>0.22	0.42
NFR	噪音	>0.5	0.68
SC 方差	语音	>0.5	0.73
BW	语音	<0.16	0.1

比较这些特征值，按照其分类功能对其进行分类得到：

表 12 实验选取的特征向量及分类结果

Table-12 result of hierarchical feature set

区分	维数	特征
静音	1~3	STE,RMS,SR
噪音	4~6	ZCR,HZCCR,NFR
语音	7~9	LSTER,SC 方差,BW
音乐	10~25	SBER(4),MFCC(12)

由 3.1 章的特征比较图可以看出，RMS，SR，STE 三个特征可以较明显区别音频帧是否是静音帧，同时由于他们都是时域音频特征，计算量比较小取得结果较好，所以放在第一层，作为特征向量的第 1~3 维分量；接下来，通过分析和实验可以得知，NFR，ZCR，HZCCR 在区分音频帧是否为噪音时可以起到突出作用（如图 6、图 8 实验结果所示），同时也可以看出，NFR，ZCR，HZCCR 三者也只涉及到时域样本的计算，这点在分层向量算法中有着重要的作用（4.3 节讲介绍）；通过实验（图 7 实验结果所示）以及已有的关于语音、音乐频率范围[36]的知识，我们可以得知 LSTER，SC，BW 在区分语音和音乐方面起到了重要作用，其中 LSTER 只涉及时域计算、而 SC、BW 则涉及到了 DFT 等频域计算，这一组特征的计算量明显增大；最后，选取 SBER，MFCC 二组特征作为区分音频帧是否为音乐的关键特征。所选取的特征与检索效率之间的关系将在 4.3 节做进一步探讨。

为验证所选取特征的有效性。实验采用上海文广实时电视节目作为媒体数据库。为便于分析对比，把实时码流用 VLC 转存到文件，其中视频格式为 MPEG-1 Video，1024Kbit/s；音频格式为 MPEG Audio，192Kbit/s；封装格式为 MPEG 1，每个片段时长约 10min。用户输入“帐篷”的语音样本保存为 wav 格式，利用表 12 所示特征值对媒体数据库片段进行分类，并与[30,45]所采用的特征值进行比较，实验均采用基于特征向量模板的分类方法，得到的结果如表 13 所示：

表 13 特征有效性实验结果

Table-13 result of feature efficiency

	纯语音	语音（音乐背	语音（噪音背	纯音乐	噪音
文[30]特征向量（4 维）	90.5%	67.7%	77.31	89.1%	88.73
本文特征向量（28 维）	94.1%	71.32%	83.62%	95.3%	90.2%
文[45]所用特征向量（90 维）	96.74%	75.44%	88.76%	96.67%	92.73%

3.4 本章小结

音频特征分析与抽取是音频处理的基础，是基于语音的视频检索中的关键技术。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和一般性。本章首先简要介绍了音频短时处理（时域、频域）和同态处理这两种特征分析和提取的重要技术，引入了本系统用到的一些关键概念。在此基础上，从时域、频域、系数域三种不同角度介绍了本系统中所采用的音频特征，重点介绍了 LPCC 和 MFCC 这两种应用十分广泛的音频特征，并阐述

了这些特征所表达的声学特点以及他们在语音分类和匹配过程中所扮演的重要角色。为下一章视频检索中音频分类与匹配内容做好了理论铺垫。

第四章 视频检索中的音频分类与匹配

音频特征的分析和提取是检索的理论基础，而分类和匹配则是音频检索实现的两个关键技术。一个典型的音频检索系统具有如下流程：

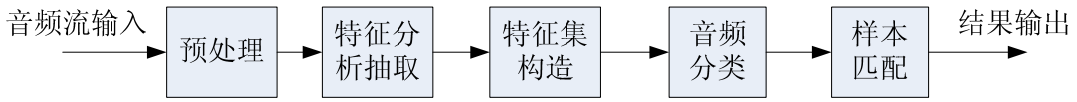


图 20 音频检索系统典型框架

Figure-20 General framework of audio retrieval system

从图 20 可以清晰看出，音频分类和样本匹配都是基于音频特征集的基础上的。特征的分析、抽取相关技术在第 3 章中已经详细阐述，在本章中我们将在后续章节对音频的分类和匹配技术加以介绍，并在本章的最后提出一种适用于本论文系统的改进音频分类、匹配方法。

4.1 音频分类算法

音频分类技术从设计思想上可以划分为三类：

- 针对音频特征值的分类方法[1]；

针对音频特征值的分类方法是最简单也是比较常用的方法。该方法的基本思路是：选取可以识别某种音频类别的合适的特征，然后设定该特征的一个阈值，根据事先约定的规则，用实际计算的特征值与阈值比较，来识别音频类别。这种方法操作简单，但也由于其简单，所以只适用于识别特征简单的音频类型，比如静音。这种方法存在以下缺点：决策规则和分类顺序并不一定是最优的；上层的决策错误会积累到下一层而形成“雪球”效应；分类误差大，需要人的先验知识和实验分析，特别是阈值的确定。所以基于规则的分类方法分类精度较低，只适用于区别性明显的简单音频类别的分类工作，难于满足复杂的、多特征的音频分类应用。但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基于规则的分类器应用广泛，J. Foote[1]采用的一种有监督贪心算法构造分类决策树就是其中的代表。

- 针对音频特征向量的分类方法；

该分类器利用向量匹配的思想，为每一个音频类型建立一个标准向量模板，然后计算实际音频帧的特征向量，用特征向量匹配标准向量模板，通过计算它们在向量空间中的距离，来判别音频类型。在澳大利亚人工智能研究院的 Elias

Pampa[32]等人开发的基于 SOM(Self-Organizing Maps)的音乐聚类系统中就采用了模板匹配的类型判断方法，通过计算模板向量和特征向量的欧拉距离来进行匹配；国防科大多媒体实验室的李恒峰、李国辉[39]开发的基于内容的音频分类与检索系统 ARS 也采用基于模板的音频检索算法。Ki-Man Kim[30]在一种快速语音分类算法中采用了基于特征向量的思想。针对音频特征统计特征的分类方法；

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，它为自动和自主学习分类的实现提供了一种行之有效的途径，是目前和未来该领域研究的主要方向。早期的基于统计学习算法的音频分类研究主要集中在神经网络算法的应用上，其代表是 Feiten.B，Frank.R 等人的研究工作[20]，他们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。近几年随着人工智能，机器学习领域的快速发展，为开展具有自主学习能力和自动音频分类研究工作提供了很好的基础，越来越多的研究者将隐马尔可夫模型[35]、K 阶最近邻算法[14,37]和高斯混合模型[15,18]等统计学习算法应用到了音频分类研究中。支持向量机(SVM)也是一种基于统计学习的算法，是 Vapni, V.N.等人[21,22]提出的以结构风险最小化原理为基础的一种分类方法，这种方法在九十年代末开始被研究人员关注，由于其泛化能力较好，在小样本的情况下就可以达到较好的精度，它在音频分类上已经有所应用。

4.2 音频匹配算法

无论采取 4.1 中所讨论的哪种音频分类技术，最终采用的音频匹配方法都是相同的。在现代音频匹配技术中，通常采用向量距离门限判定的算法[30,32]。即首先把在分类中所采用的特征值，组成一个高维的特征向量，分别计算样本和媒体序列的向量值，然后在向量空间中计算这两个向量代表的点之间的距离，根据门限判定是否向量匹配。

设所选取的特征组成的向量为$\vec{\phi}$：

$$\vec{\phi}=\{\phi_1,\phi_2,...,\phi_n\} \tag{4-1}$$

n 为选取的特征数目。计算向量之间距离比较常用的有两种方法，即欧式距离(Euclidean distance)和马氏距离(Mahalanobis distance)。因马氏距离需要计算个坐标之间的统计相关值，计算较复杂，所以在音频匹配技术中则通常采用欧式距离：

$$D=\sqrt{\sum_{i=1}^{n}\left(\vec{\phi}_s[i]-\vec{\phi}_m[i]\right)^2} \tag{4-2}$$

$\vec{\phi}_s[i]$表示样本特征向量中第 i 个特征值，$\vec{\phi}_m[i]$表示媒体特征向量中第 i 个特征值。计算得到距离后，与门限值进行比较得到匹配结果：

$$J=\begin{cases}1, \ when\ D<\Gamma\\ 0, otherwise\end{cases} \tag{4-3}$$

4.3 改进的分类和匹配算法

本论文重点研究流媒体的实时内容检测，为此，我们对 4.1 和 4.2 章所介绍的技术进了改进，以满足流媒体的高实时性要求。本论文研究的重点包含三个方面：特征的有效性选择、分层的向量模板分类技术以及前向加权序列的匹配技术。有效的选取音频特征是论文任务完成的关键。特征选取过多会造成计算量成几何级数递增，从而影响任务完成的实时性；特征选取过少则无法确切表征语音样本的语义特征，使得匹配准确率下降，无法获得满意结果。本文通过实验的方法得到了一个有效性比较高的特征集，在实时性和准确性之间获得了较好的平衡。

音频特征的选择是任务完成的基础，音频的分类和匹配则是任务完成的关键。下面，就本论文针对任务的特殊性所提出的两种改进方案做比较详尽的阐述。

4.3.1 分层的向量模板分类算法

在音频检索中，进行样本匹配之前通常要对目标媒体进行音频分类，如果目标媒体与样本属于统一类型，则进行匹配；若不是同一类型，则跳过匹配过程。这样可以大大缩短音频检测过程的计算量，如图 21 所示：

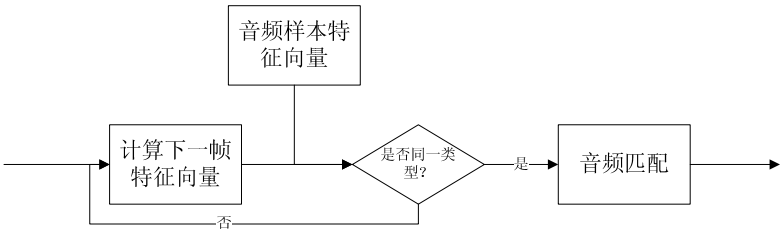


图 21 音频检索流程图

Figure-21 flowchart of audio retrieval

本论文结合基于规则的分类方法和基于模板的分类思想，针对传统的分类过程进行优化，提出一种分层的向量模板分类技术（Hierarchical Classify Method

Based on Vector Templates, HCMBVT）。该方法的理论出发点是：所选取的特征值在音频分类任务中，起到的作用有所不同。也就是说不同特征值在区分不同类别音频时起到的作用不同。例如本文 3.2.1 章介绍的 HZCRR 在区分噪音和非噪音的时候非常有用，而在区别语音和音乐的时候则效果不是很明显；而在 3.2.2 中介绍的 SC 方差在区分语音和非语音时差别明显，而在其他场合则效果不大。为此，本文在分类过程中采用了一种新的方法。首先通过实验分析所选取特征集合中的每一个特征，根据其特点分为四层，如图 22：

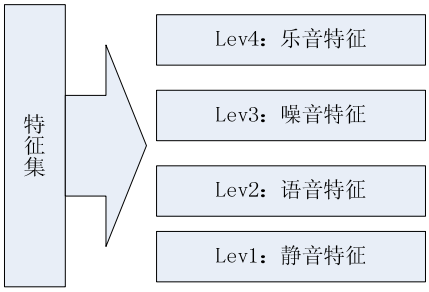


图 22 特征集合的分层

Figure-22: layers of audio features

把每一类中的特征值组成一个特征向量，分别记为 $\vec{\phi}_{lev1}$， $\vec{\phi}_{lev2}$， $\vec{\phi}_{lev3}$ 和 $\vec{\phi}_{lev4}$。

在 检测，先在音频库中计算四层特征向量的分类模板；检测时根据图 23 所示流程进行分类：

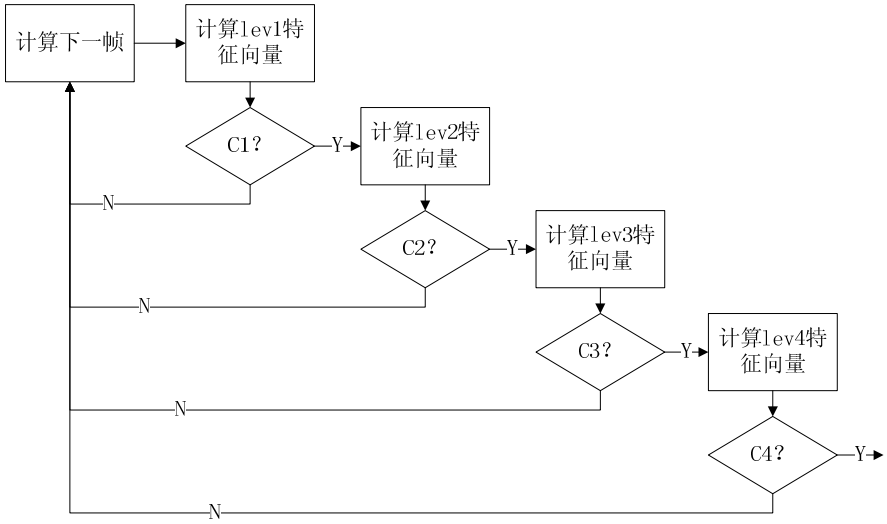


图 23 分层向量模板分类流程图

Figure-23 flowchart of hierarchical classification based on templates

其中，每一个音频帧对应一个分类向量 $\vec{C}=\{C1,C2,C3,C4\}$，C1 表示是否静音，C2 表示是否语音，C3 表示是否乐音，C4 表示是否噪音。C1~C4 之间有一定相互限定，如是静音就不可能再是语音或乐音了。向量 $\vec{C}$ 可能取值如下表：

表 14 分类向量取值表

Table-14 classification vector table

含义	C1	C2	C3	C4
乐音	0	0	0	1
噪音	0	0	1	1
语音	0	1	0	0
静音	1	0	0	0

根据表 14，可以对图 23 表示的流程图进行进一步优化，以减少不必要的计算量，如图 24 所示：

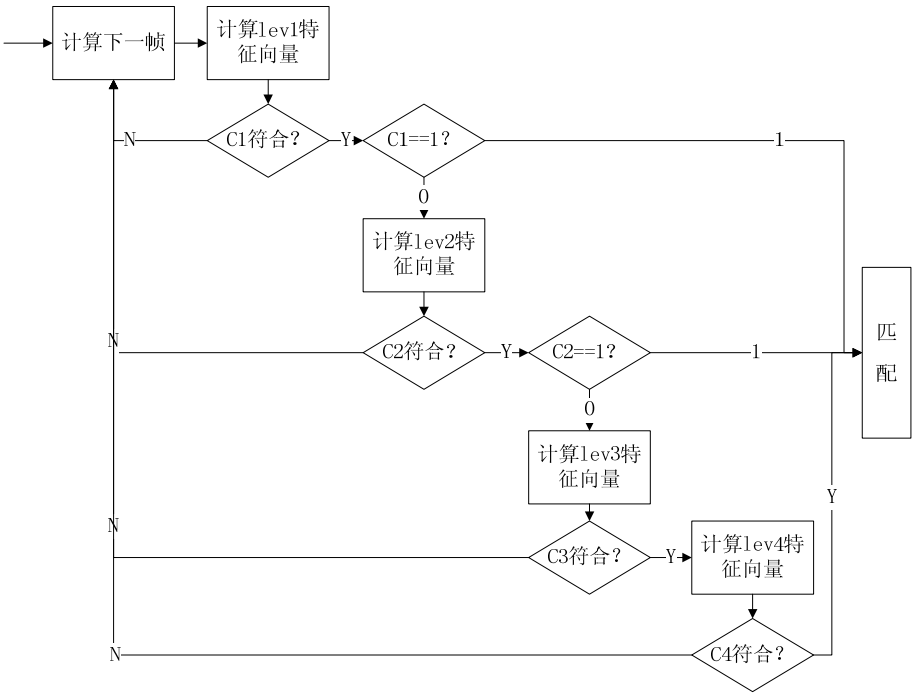


图 24 改进的分层分类流程图

Figure-24 flowchart of improved HCMBVT

设每层计算的计算量为 $q_i, i=1,2,3,4$，不同类型音频片段的统计分布为均匀分布，计算一般分类算法、分层的向量模板分类和改进的分层分类算法的计算量分别如下：

表 15 分类算法计算量比较（1）

Table-15 comparison of calculation amount of classification method (1)

算法名称	平均计算量	结果

一般分类	$q_1+q_2+q_3+q_4$	$q_1+q_2+q_3+q_4$
HCMBVTv1	$q_1+\frac{1}{2}\left(q_2+\frac{1}{2}\left(q_3+\frac{1}{2}q_4\right)\right)$	$q_1+\frac{1}{2}q_2+\frac{1}{4}q_3+\frac{1}{8}q_4$
HCMBVTv2	$q_1+\frac{1}{2}\left(\frac{1}{2}\left(q_2+\frac{1}{2}\left(\frac{1}{2}\left(q_3+\frac{1}{2}q_4\right)\right)\right)\right)$	$q_1+\frac{1}{4}q_2+\frac{1}{16}q_3+\frac{1}{32}q_4$

由表 15 可以清晰看到与一般分类算法相比，HCMBVT 算法在q_1没有减少，而在q_2、q_2、q_3处均获得较大幅度的减少。这给我们分层分类器设计做出了有力指导。即要把层计算量较小的放在较低层，而把层计算量大的放在较高层中。现假设每层计算量均为 1，对算法优化程度进行估算，

表 16 分类算法计算量比较（2）

Table-16 comparison of calculation amount of classification method (2)

算法名称	平均计算量	理论减少比例
一般分类	4	-
HCMBVTv1	1.875	53.125%
HCMBVTv2	1.34375	66.4%

注意以上减少比例仅是限定在分类阶段的特征值计算量，不是整个媒体检索系统的计算量减少比例。该算法对整个系统性能的提升结果见第 5 章实验部分，在下一章，我们也会注意到采用 HCMBVTv2 提高实时性相应的同时所付出的代价。

表 12 已经给出了我们实验中所用到的所有音频特征，结合分层的向量模板分类算法所提到的特征分层理论，现把表 12 所示特征作如下分层：

表 17 实验选取的特征向量及分层结果

Table-17 result of hierarchical feature set

level	区分	维数	特征
1	静音	1~3	STE,RMS,SR
2	语音	4~6	LSTER,SC,BW

3	噪音	7~9	ZCR,HZCCR,NFR
4	音乐	10~25	SBER(4),LPC(10),MFCC(13)

特别注意到由层 1 至层 4 的计算量大体呈递增趋势的（分析见 3.3 节），结合表 15 所示分析结果可以看出，如此分层可以有效减少检索所需的计算量。

4.3.2 前向加权序列的匹配算法

音频匹配的过程，实质上是音频特征向量的匹配过程。在一般多媒体检索系统中所采用的音频匹配算法通常是(4-2)式所表示的欧式空间距离匹配算法[30, 35]。该算法概念清晰，易于理解。但是其不足之处在于该算法要求一次性的对特征向量所有坐标进行计算，这造成了在信息检索过程中的计算冗余。而针对流媒体内容检测所要求的高实时性，如何减少匹配算法中的计算量就自然成为了本论文的研究重点之一。本文在检测中提出一种改进的匹配算法：前向序列特征加权距离算法（Forwarding-sequential Weighted-feature distance measuring, FWDM）如图 25 所示：

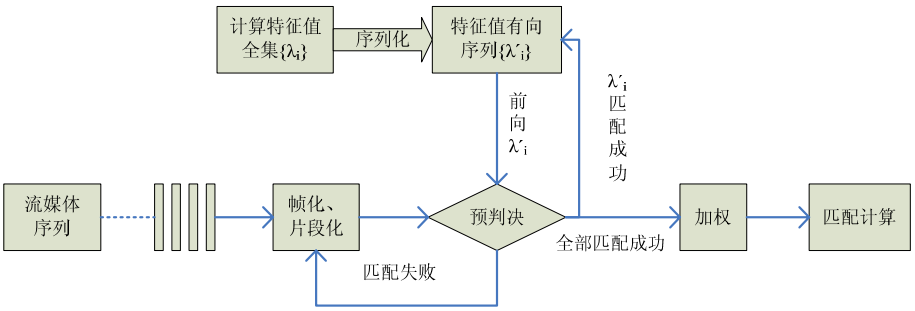


图 25 FWDM 算法流程图

Figure-25 flowchart of FWDM

针对选取的特征向量，FWDM 算法不是针对音频段一次性计算所有特征值之后与样本向量之间距离得到匹配结果，而是在每计算一个特征向量之后便通过加权然后与匹配门限值进行比较得到预判决结果，如果成功则进行下次匹配，如果不成功则进行下一音频片段的计算过程。同时，特征向量的排序与最后检索的准确率和速度也有关系，根据 4.3.1 中讨论可以得出，计算量较小、可以有效区分样本与媒体音频段之间不同的特征值如 RMS，ZCR，SR，BW 等应该优先计算，而计算量较大对类型不敏感、对语音内容语义敏感的特征值如 SBER，MFCC 等应该放在后面计算。这样，对第三章所选取的特征值进行排序，组成一个有序的特征向量依次进行判决，可以有效的提高匹配的速度。预判决函数为：

$$J_i = \begin{cases} 0, \ when\ \alpha_i \dfrac{\left|\phi_s[i]-\phi_m[i]\right|}{\phi_s[i]} > \varepsilon \\ 1, \ otherwise \end{cases} \tag{4-4}$$

ε 为系统门限值，α_i 的取值大小根据 3.1 章中的置信区间大小来设定，为消除各特征值置信区间不同的影响，取置信区间的倒数。如果 $\phi_s[i]$ 匹配成功则顺次尝试匹配 $\phi_s[i+1]$ 直至全部特征值匹配。否则，跳过当前帧，对下一音频帧进行匹配。以此达到提高速度的目的。判决结束后，可以按如下公式计算其加权后向量之间的距离，最终获得最后匹配结果：

$$\begin{aligned} D[n] &= \sqrt{\sum_{i=1}^{N}\left[\alpha_i\left(\bar{\mu}[n,i]-\bar{\psi}[n,i]\right)\right]^2}, \\ Score[n] &= \begin{cases} Rank(D[n]), & when\ D[n] < \Gamma, \\ 0, & otherwise, \end{cases} \end{aligned} \tag{4-5}$$

其中，Γ 为判决门限，$\bar{\mu}$ 和 $\bar{\psi}$ 分别为流媒体帧和样本帧的归一化特征值。通过这样的两步、分级的匹配过程，达到快速、准确的实现流媒体中基于语音的实时语义检测。同时，为提高用户使用体验，在真或假的二元判决基础上，在满足判决门限条件下，可以根据计算所得距离进行评分（ranking），最后把分数高于阈值的片段按照分数由高到低配列出来以供用户浏览。

4.3.3 算法融合及系统设计

4.3.1 和 4.3.2 小节分别介绍了本论文改进的分类和匹配算法，即分层的向量模板分类算法和前向加权序列匹配算法。不难发现，这两种算法都有一种相似的设计思想即惰性计算。所谓惰性计算就是在检索的过程中，并不是一次性的计算出所需要的全部特征值再进行针对特征值全集的分类和匹配。这种算法造成许多不必要的计算冗余和资源浪费。针对此，我们提出了按需计算、推迟计算等方法改进了分类和匹配算法，减少了检索过程中的计算量。同时我们也注意到，由于改进后的 HCMBVT 和 FWDM 都是由一种设计思想设计出来的，所以在实现过程上自然的可以融为一体，从而可以进一步减少系统的计算量。可以通过两次融合来达到这个目的。

未融合前的检索系统如图 26 所示：

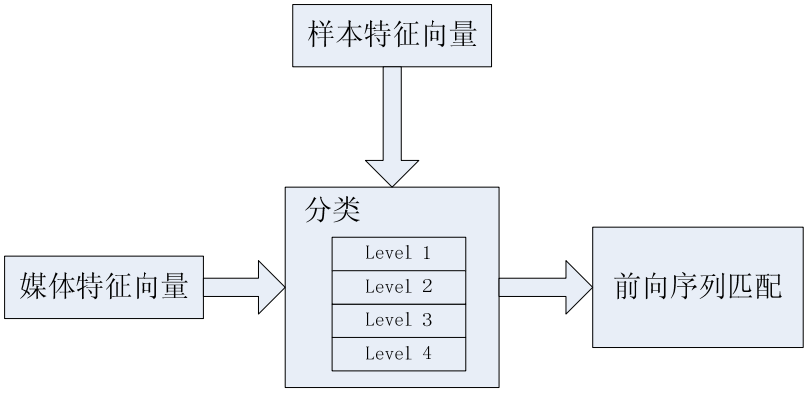


图 26 未融合前系统框图

Figure-26 unmerged system framework

本系统首先对媒体特征向量和样本特征向量进行分层分类，如果属于同一类型再对两特征向量进行前向序列匹配，最终获得匹配结果。我们利用分类和匹配过程都对特征向量集合进行分组这一共同点，进行第一步融合。具体方法是把匹配过程融入到每层的分类过程当中。即不是在分类过程之后进行匹配计算，而是在每层分类成功配对后进行该层内的匹配，如图 27 所示：

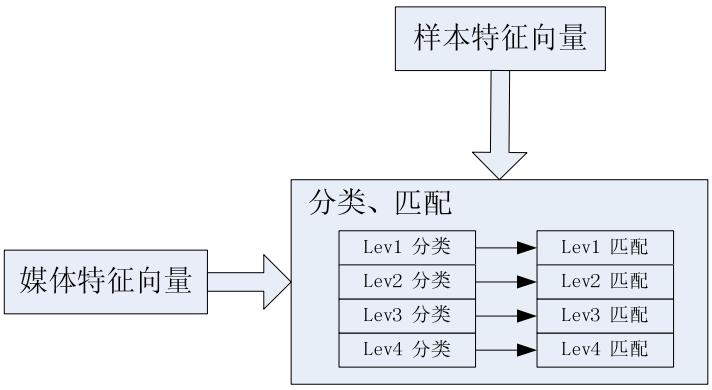


图 27 融合（1）后的系统框图

Figure-27 merged (1) system framework

经过第一次融合后的系统引入了改进算法的分层的概念，这有效的减少了不同类型音频检索所需要的计算量。但是融合后系统并没有包含有 FWDM 算法中前向加权的思想，所以在图 27 的基础之上进行再融合，对每层的特征向量进行排序，并进行前向加权匹配。经过融合以后的最终系统如图 28 所示：

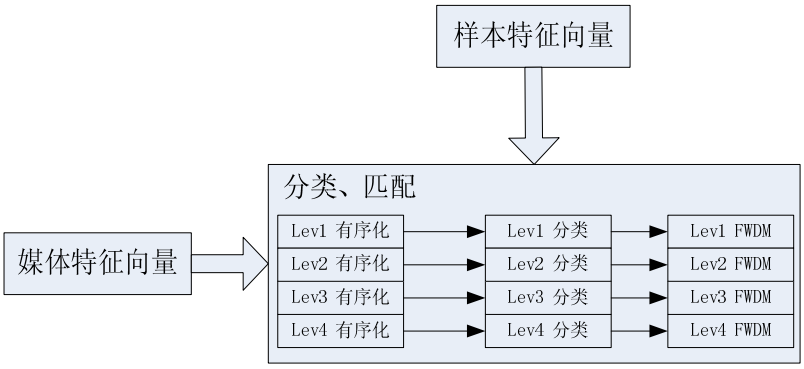


图 28 融合（2）后的系统框图

Figure-28 merged (2) system framework

本实验系统的设计使用场景是针对流媒体的实时性内容检测。系统处理对象是实时发送的多媒体流，如本上海文广的 IPTV 数据流、上海交通大学 COMIC 网站上的 VOD 点播数据流等。要达到的效果是实时检测音频数据流，统计音频内容特征写入数据库，同时针对用户输入的语音样本提供数据库检索或者实时语义监控的功能。

本课题的目的有以下两点：

首先，研究基于音频的视频内容检索系统，验证多模态媒体检索的可行性，提供多模态媒体内容检索一种实现的可能性方案；

其次，探索如何在对准确率影响不是很大的情况下，提高多媒体内容检索的时间效率、实时性，以增强多模态信息检索系统的实用性；

4.3.3.1 系统基本功能

目标系统需要达到以下基本功能：

- 视频流媒体处理：能够接收实时流媒体信息，从中分别解复用出音频、视频流信息；
- 音频特征提取：对音频流数据进行帧化、段化，提取各种音频特征；
- 音频特征分类：对音频进行分类；
- 音频特征匹配：与用户输入的语音样本进行匹配；
- 基于音频的流媒体视频检索系统的设计实现：同时把音频内容信息写入数据库，向用户提供基于音频的视频内容关键词样本检索功能；

4.3.3.2 系统平台

主机：浪潮服务器

CPU ： Intel(R) Xeon(TM) CPU 2.80GHz × 4

内存：2.0G

开发语言：C++语言

操作系统：Fedora 6 Linux Kernel 2.6.16

开发平台：Eclipse

编译器： GCC

软件开发包： FFMPEG，libsvm

4.3.3.3 系统架构

本实验系统的框图如图所示：

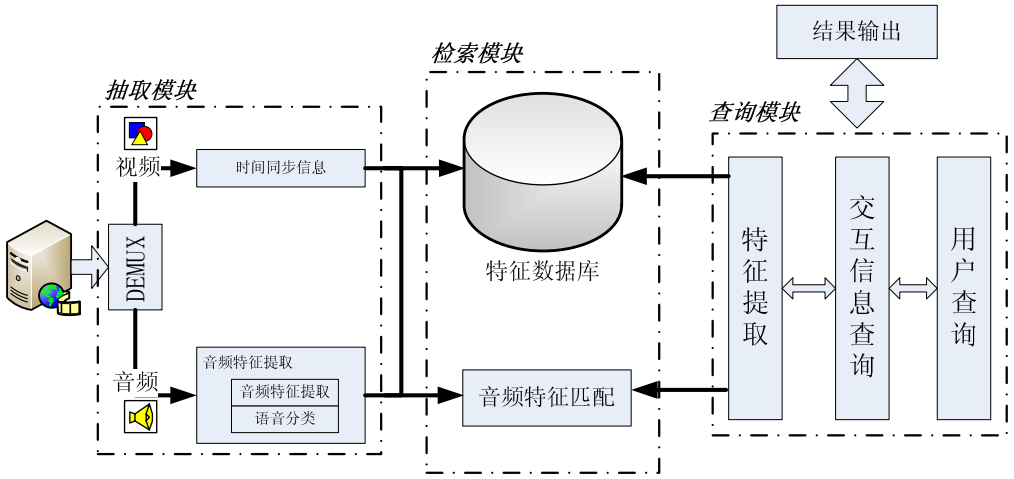


图 29 实验系统框图

Figure-29 framework of the experiment system

本实验系统主要划分为三个模块：

- 抽取模块：该模块负责读入流媒体信息，分别解复用出音频流和视频流；对音频流进行音频特征提取和分类，从视频流中抽取时间同步信息；把时间同步信息和音频特征输出到检索模块；
- 查询模块：读入用户输入关键词样本，抽取音频特征，输入到检索模块进行检索操作；交互信息模块则负责处理用户定制的查询以及查询的结果的返回；
- 检索模块：接受抽取模块以及查询模块输入的音频特征信息，分两部分进行操作；一部分把媒体信息写入数据库以提供非实时的信息检索功能；另一部分实时的进行音频匹配，以提供实时语义检测功能；

系统主要的功能模块为音频特征提取模块、音频分类模块和音频特征匹配模块。这三个模块也是本课题理论研究重点的具体实现，其算法的理论基础在 4.3 节已经详述。

由于本课题主要研究的重点是对音频流的处理，所以我们着重描述一下本实验系统音频流处理流程。本流程在经典处理方法中融合了 4.3.3 中提到的改进的分层向量模板分类算法和前向序列匹配算法。详细的流程图如图 30 所示：

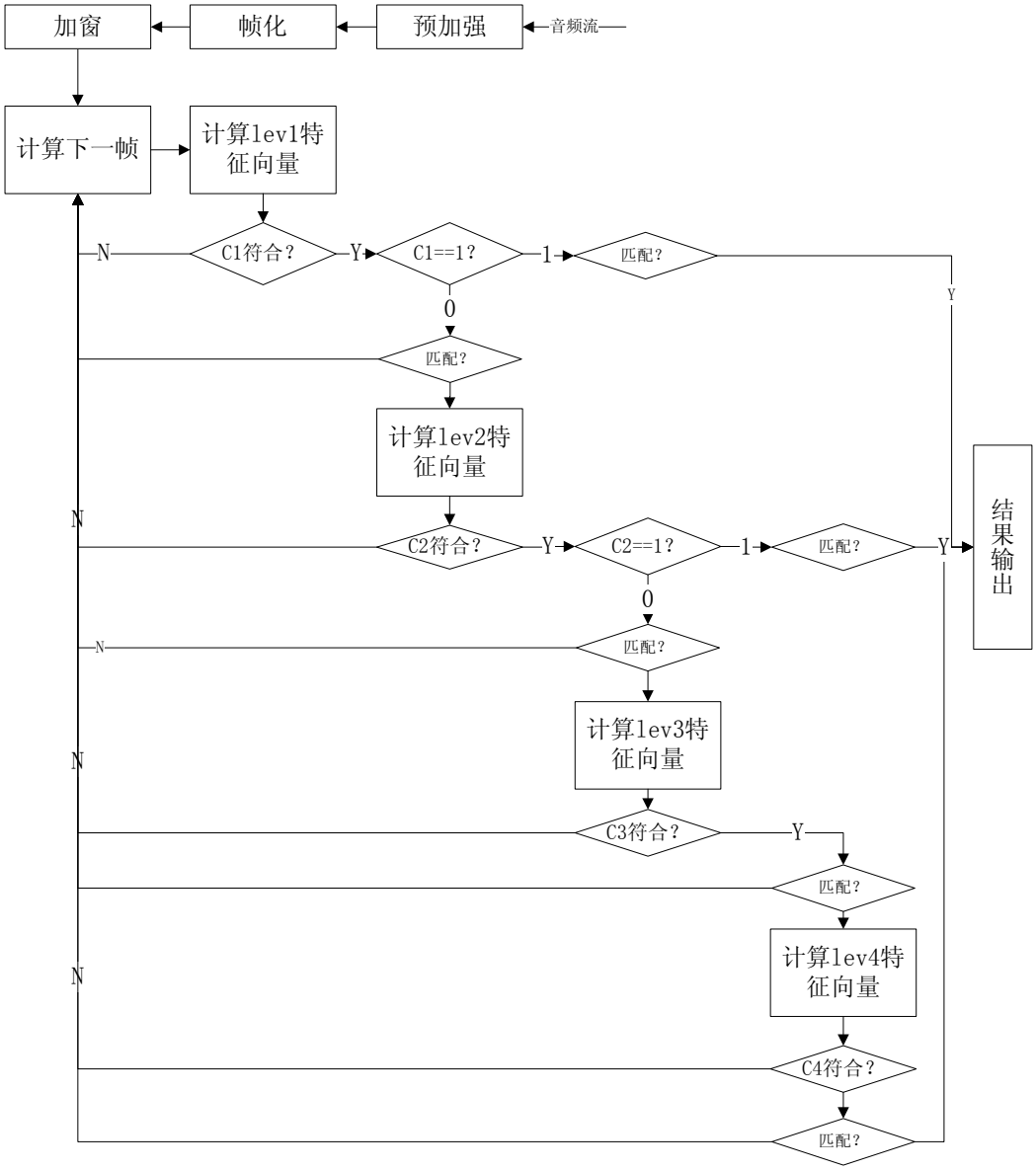


图 30 实验系统音频流处理流程图

Figure-30 flowchart of audio processing in experiment system

4.4 实验结果及分析

实验采用上海文广实时电视节目作为媒体数据库。为便于分析对比，把实时码流用 VLC 转存到文件，其中视频格式为 MPEG-1 Video，1024Kbit/s；音频格式为 MPEG Audio，192Kbit/s；封装格式为 MPEG 1，每个片段时长约 10m。用户输入“毛主席”的语音样本保存为 wav 格式，对数据库中媒体文件进行检索。实验首先比较了 FWDM 算法在音频匹配阶段对性能提升的影响，结果如表 18 所示：

表 18 匹配算法效率比较

Table-18 efficiency comparison of match methods

	纯语音	语音（音乐背景）	语音（噪音背景）	总时
一般算法[35]	555s	476s	493s	1524s
FWDM	492s	425s	450s	1367s
提高				10.3%

结果分析：从结果中可以看出，在采用了 FWDM 算法后，匹配速度有了较明显的提高，由于没有背景噪音的影响，对纯语音的片段的检索速度提高最多，命中率效果也最好。相比之下，带有背景噪音的实验提升效果较差。同时发现，配有音乐的歌唱中搜索命中率不高。

为验证 HCMBVT 算法的有效性，采用不同方法对数据库中媒体文件进行检索。实验 1 采用[30,32]中提到的模板向量分类算法没有采取分层机制；实验 2 采用本论文提出的 HCMBVTv1、HCMBVTv2 算法；实验 3 采用[35]SVM 分类算法对媒体库片段进行分类，结果如下：

表 19 分类算法可靠性比较

Table-19 availability of classification methods

正确率	纯语音	语音（音乐背景）	语音（噪音背景）	纯音乐	背景音
一般算法[30]	90.4%	78.0%	87.1%	87%	79.8%
HCMBVT	89.1%	74.44%	83.76%	85.3%	78.7%
SVM[35]	91.9%	82.23%	89.4%	88.4%	82.4%

分析表 19 的实验数据，可以看出 SVM 算法的可靠性相对而言是最好的，向量模板次之而 HCMBVT 表现比较差。这是因为 HCMBVT 采取的分层分类机制每次分类所选取的特征值均少于其他二种方法，以至于在准确性上相比较而言稍逊。由于本论文针对流媒体检索这一特定的应用场景，可以看出 HCMBVT 在实

时性上面表现最好。在以后的研究当中可以考虑结合其他算法进一步提高准确性。对整个媒体库进行一次检索操作的时间如图所示：

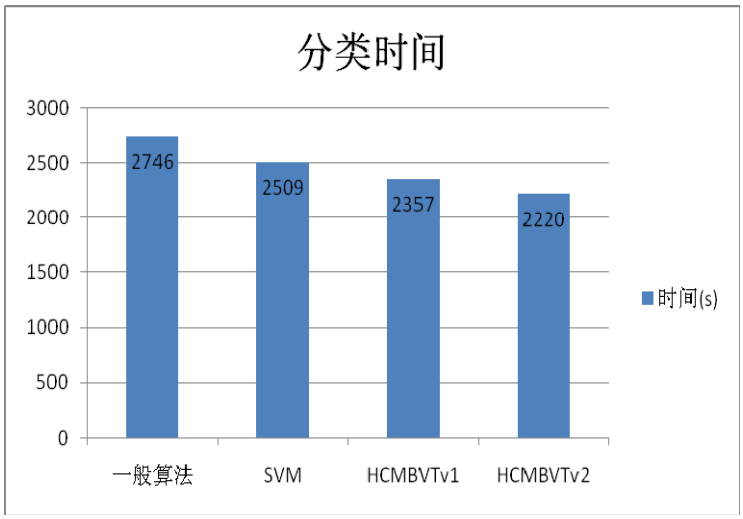


图 31 分类算法效率对比图

Figure-31 efficiency comparison of classification

结果分析：本实验比较了一般基于向量模板的分类算法、SVM 分类算法、HCMBVTv1 以及 HCMBVTv2 这四种算法之间的时间效率和可靠性。通过实验可以看出，对媒体数据库进行一次检索所需要的时间 HCMTVTv2 的时间最少，SVM 方法次之，最慢的是一般基于向量模板的分类算法，HCMBVTv2 的时间效率比一般向量模板方法提高了 19.1%，比 SVM 算法提高了 11.5%。分析其原因是一般算法和 SVM 算法在进行分类操作之前，均需要对音频帧计算出其对应的所有音频特征分量，而 HCMBVT 算法采取了分层的惰性计算，在试验中所需要时间有较明显的减少。同时，根据 4.3.1 小节所分析，HCMBVTv2 比 HCMBVTv1 具有更大的惰性，所以前者效率也比后者有轻微的提升。我们注意到，对于带有音乐和噪音背景的语音分类，几种算法的表现均较差。分析其原因是这类音频带有语音和音乐或者噪音两种类型的特性，使得其片段在音频特征上更倾向于背景音或噪音，这给分类器正确分类带来了很大困难。

以下三组实验分别比较了本论文改进的算法和其他检索算法。实验 1 采用向量模板分类+向量模板匹配方法；实验 2 采用了 SVM 分类+向量模板匹配方法；实验 3 采用了 HCMBVT+FWDM 方法；每组实验中分别进行 2 次检索，第一次在一段新闻当中检索语音“帐篷”单词，第二次在一段足球比赛录像当中检索裁判的哨音样本；得到实验数据如下：

表 20 检索方法可靠性比较

Figure-20 availability of retrieval methods

	语音“帐篷”		裁判哨音	
	准确率	时间(s)	准确率	时间(s)
向量模板分类 向量模板匹配	90.10%	28.83	78.80%	21.63
SVM 分类 向量模板匹配	91.29%	29.81	83.63%	22.66
HCMBVT FWDM	89.40%	24.75	75.25%	18.32

可以看出，三个方案中实验 3 采用了本论文改进的分类和匹配算法，所用时间最短，实时性提升明显，但准确率相比略低这是由于采用了分层思想的分类和匹配算法可以有效减少检索过程的计算量，然而由于没层中采用的特征值较少，影响了最后的准确率；准确率表现最好的是 SVM 分类匹配的算法，但是这种方法要求之前提供 SVM 学习所需样本并且额外的运算时间。

同时，我们也考察了各特征门限值的选取对实验结果的影响。选取 phonix-info.mpg 作为检索对象，pingji.wav 作为查询样本，选取不同门限值进行检索得到的命中率、检索时间如下图所示：

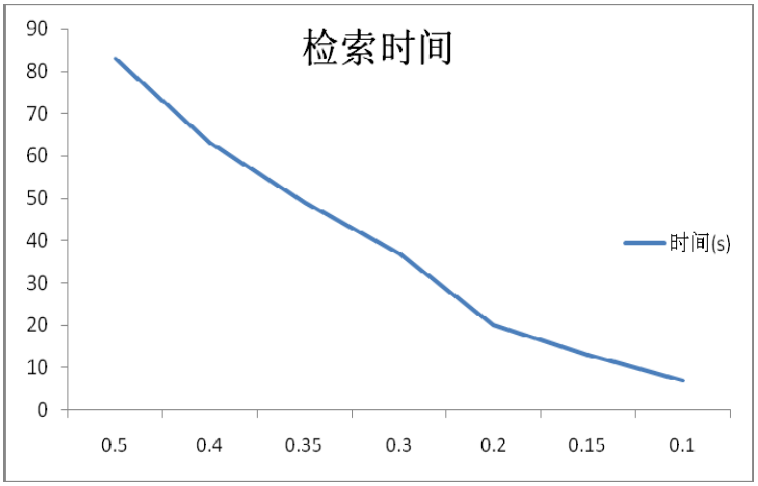


图 32 匹配门限值与检索时间关系图

Figure-32 Relationship between gate value and retrieval time

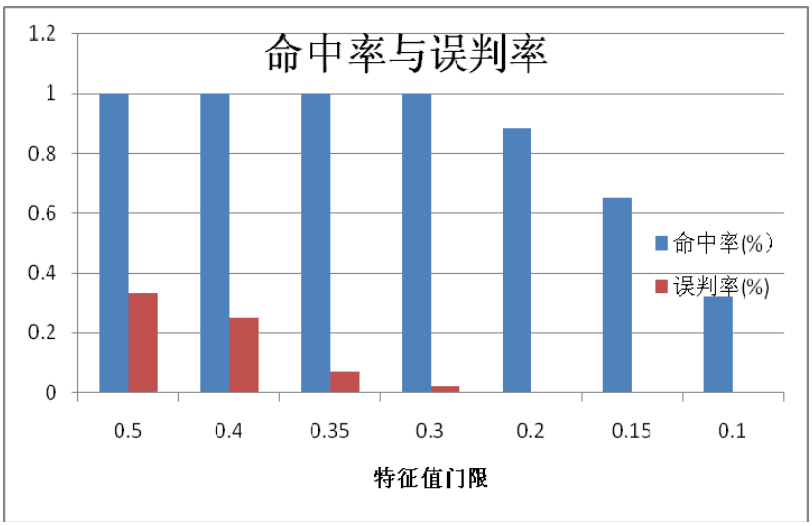


图 33 特征值门限和命中率、误判率关系图

Figure-33 relationship between gate value and accuracy

结果分析：如何选取有效的音频特征，是决定实验准确率和效率的关键问题。本实验从音频特征门限角度给出了检索时间和可靠性的关系。从实验结果上可以明显看出，当门限值逐渐减小的时候，意味着评判条件更加苛刻，大部分计算在前面几层就被终止了，所以检索的计算量、时间明显减少。同时我们也注意到，随着门限值的递减，算法的命中率也会逐渐降低。这是因为随着门限条件越来越苛刻，一些原本符合特征匹配条件的音频帧也被判为不合格，造成漏判；相对的，随着条件的愈加严格，误判的概率逐渐减小，直至没有误判的情况发生。实验表明，在选择好特征值的基础上，需要根据不同的需求来选取判定门限。当需要效率优先的应用时，可选取较严格的门限；而可靠性成为问题的关键时可以选取较宽松的 门限。从实验可以看出，当门限值在 0.3 附近时误判率最低，而命中率依然保持 100%，同时对比图 32 此时检测时间也比较快，故最佳门限值应在 0.3 附近选取。实验选取 0.29 作为本文所使用门限。

4.5 本章小结

音频特征的分析和提取是检索的理论基础，而分类和匹配则是音频检索实现的两个关键技术。本章首先介绍了典型的音频分类和匹配算法，在此基础之上，针对本文所讨论的特殊应用环境，分别提出了改进的分类算法 HCMBVT 和改进的匹配算法 FWDM，并从理论上比较了改进算法和传统算法之间的优缺点。最后，融合这两种改进的检索技术提出了本论文检索系统的系统框图。

第五章 总结与展望

5.1 本文总结

视频检索技术应用前景十分广泛，无论是在娱乐、数据监控、管理等方面都有着重要的意义。然而现有的视频信息检索技术大部分集中在对视频图像方面的研究，即利用图像信息提取高层语义进行视频信息检索。本文在此基础之上，结合国内外的研究成果，着重研究了从音频角度来进行媒体内容的多模态检索，并重点研究了如何提高基于音频的视频内容检索的效率问题。

本文首先回顾和总结了国内外在多模态媒体信息检索领域的研究状况，特别的围绕基于音频模态的视频媒体检索进行了重点介绍；第二章中，我们介绍了视频检索系统的一般框架，分析了其特点与不足，在此基础之上介绍了最新的多模态视频检索系统模型。针对本论文的研究方向我们着重讨论了基于内容的视频检索技术中的音频处理技术与难点。其中，介绍了音频检索的分类与各类型的特点与不足，分析了在本论文应用场景下所应采用的技术，并结合以上所讨论的结果，给出了本论文采用的基于音频的视频内容检索系统框架。

接下来的第三章、第四章则详细介绍了音频特征提取、音频分类和匹配这三个系统中的关键模块的技术、作用；其中本文的贡献主要在于：对从音频模态进行视频内容检索的方法进行了探讨、实现和分析；研究了基于音频的视频检索中音频特征的有效选取和门限值的确定；提出了一种分层的向量模板分类算法（HCMBVT）；提出了一种改进的前向加权序列的匹配算法（FWDM）；通过实验探讨了基于音频的视频内容检索系统的优缺点及可行性。

5.2 未来工作展望

目前基于音频的视频媒体检索的主体框架已经趋于定型，但是在一些具体的技术上还存在很大的提升空间。本论文并没有实现一个很完备的视频内容检索系统，我觉得在该课题还有以下一些问题需要将来的研究者继续探讨：

- 由于不同说话人的某些语音特征很大差别，如何解决用户输入样本与媒体文件说话人特征不同问题；
- 如何解决输入样本与媒体文件中语音速率不同的问题；
- 如何解决音频语义和视频语义相适应的问题；

- 如何把基于音频特征的技术与语音识别等其他相关技术结合起来进行更深入的研究；
- 如何把视频场景分析与音频语义分析结合起来，从而进行更有效的媒体语义分析和信息抽取；

参考文献

[1] Foote J. An overview of audio information retrieval, ACM Springer Multimedia Systems, 1998

[2] Liu Z, Huang J, Wang Y, Chen T., “Audio feature extraction and analysis for scene classification.”, Multimedia Signal Processing, 1997., IEEE First Workshop on 23-25 June 1997 Page(s):343 - 348

[3] S. Kiranyaz and M. Gabbouj, “Generic content-based audio indexing and retrieval framework”, IEEE Image Sigal Process, Vol. 153, No. 3, June 2006

[4] Shieh, J.-R.J.; Audio content based feature extraction on subband domain, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on March 31 2008-April 4 2008 Page(s):217 - 220

[5] Zabih R., Miller J., Mai K., A Feature-Based Algorithm for Detecting and Classifying Scene Breaks, ACM Multimedia 95, Nov 1995, pp. 189-200.

[6] Roy D.，Malamud C. Speaker identification based text to audio alignment for an audio retrieval system. IEEE International Conference on Acoustics Speech and Signal Processing(ICASSP’97)，1997.4，2:1099-1102.

[7] Shen, J.; Shepherd, J.; Ngu, A.H.H.; Towards Effective Content-Based Music Retrieval With Multiple Acoustic Feature Combination Multimedia, IEEE Transactions on Volume 8, Issue 6, Dec. 2006 Page(s):1179 - 1189

[8] Chung-Hsien Wu; Chien-Lin Huang; Chin-Shun Hsu; Kuei-Ming Lee; Speech retrieval using spoken keyword extraction and semantic verification , TENCON 2007 - 2007 IEEE Region 10 Conference Oct. 30 2007-Nov. 2 2007 Page(s):1 - 4

[9] Hoi, S.C.H.; Lyu, M.R.; A Multimodal and Multilevel Ranking Framework for Content-Based Video Retrieval. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on Volume 4, 15-20 April 2007 Page(s):IV-1225 - IV-1228

[10] Kin-Wai Sze; Kin-Man Lam; Guoping Qiu; An optimal key frame representation for video shot retrieval. Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on 20-22 Oct. 2004 Page(s):270 - 273

[11] Rong Yan, Jun Yang, and Alexander G. Hauptmann, "Learning query-class dependent weights in automatic retrieval," in Proc. ACM International Conference on Multimedia, New York, NY, USA, 2004, pp.548-555

[12] Eung Kwan Kang; Sung Joo Kim; Joon Soo Choi; Video retrieval based on scene change detection in compressed streams. Consumer Electronics, IEEE Transactions on Volume 45, Issue 3, Aug. 1999 Page(s):932 - 936

[13] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders; Early versus late fusion in semantic video analysis; in Proc. ACM International Conference on Multimedia, Singapore, 2005, pp.399-402

[14] Chung-Hsien Wu; Chien-Lin Huang; Chin-Shun Hsu; Kuei-Ming Lee; Speech retrieval using spoken keyword extraction and semantic verification， TENCON 2007 - 2007 IEEE Region 10 Conference Oct. 30 2007-Nov. 2 2007 Page(s):1 - 4

[15] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang; Video search re-ranking via information bottleneck principle; in Proc. ACM International Conference on Multimedia, Santa Barbara, CA, USA, 2006, pp.35-44

[16] Huang-Chia Shih; Chung-Lin Huang; Content-Based Multi-Functional Video Retrieval System, Consumer Electronics, 2005. ICCE. 2005 Digest of Technical Papers. International Conference on 8-12 Jan. 2005 Page(s):383 – 384

[17] Eung Kwan Kang; Sung Joo Kim; Jong Soo Choi; Video retrieval based on scene change detection in compressed streams, Consumer Electronics, 1999. ICCE. International Conference on 22-24 June 1999 Page(s):224 - 225

[18]Weld E, Blum T, Kreisler D, et al. Content-based Classification, Search and Retrieval of Audio. IEEE Multimedia，1996(3):27-36

[19] Hoi, S.C.H.; Lyu, M.R.; A Multimodal and Multilevel Ranking Scheme for Large-Scale Video Retrieval; Multimedia, IEEE Transactions on Volume 10, Issue 4, June 2008 Page(s):607 - 619

[20] Feiten, B., Frank, R., Ungvary, T. Organization of sounds with neural nets. In: Proceedings of the 1991 International Computer Music Conference, International Computer Music Association. San Francisco, 1991. 441444

[21] CHRISTOPHER J.C.BURGES, Bell Laboratories, Lucent Technologies, A Tutorial on Support Vector Machine for Pattern Recognition.

[22] Vapnik, V.N.; An overview of statistical learning theory, Neural Networks, IEEE Transactions on Volume 10, Issue 5, Sept. 1999 Page(s):988 - 999

[23]VideoQ, persia.ee.columbia.edu:8080

[24] Mingchun Liu; Chunru Wan; A study on content-based classification and retrieval of audio database, Database Engineering & Applications, 2001 International Symposium on. 16-18 July 2001 Page(s):339 - 345

[25] Virage Creates World's First Video Search Engine for the World Wide Web, http://www.autonomy.com/content/News/Releases/1998/0919a.en.html

[26] http://danadler.com/jacob/.

[27] VisualSEEk - A joint spaital-feature image search engine

http://www.ee.columbia.edu/ln/dvmm/researchProjects/MultimediaIndexing/VisualSEEk/VisualSEEk.htm

[28] IBM's Query By Image Content, http://wwwqbic.almaden.ibm.com/

http://infolab.stanford.edu/~wangz/project/imsearch/review/JOUR/datta.pdf.

[29] McGurk, Harry; and MacDonald, John (1976); "Hearing lips and seeing voices," Nature, Vol 264(5588), pp. 746–748

[30] Ki-Man Kim; Se-Young Kim; Jae-Kuk Jeon; Kyu-Sik Park; Quick audio retrieval using multiple feature vectors. Consumer Electronics, IEEE Transactions on Volume 52, Issue 1, Feb. 2006 Page(s):200 - 205

[31]Thomas F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Publishing House of Electronics Industry, 2004

[32] Elias Pampalk, Andreas Rauber, Dieter Merkl; Content-based organization and visualization of music archives; December 2002 MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia Pages: 570 - 579

[33]郭春霞、裘雪红 “基于 MFCC 的说话人识别系统研究”， 西安电子科技大学 2006 年 1 月

[34] 陆伟艳、夏定元、刘毅，“基于内容的视频检索的关键帧提取”，微计算机信息，Vol.23, No.11-3, 298~300

[35] 白亮、老松杨，“基于支持向量机的音频分类与分割”，计算机科学，Vol.32，No.4，87~91， 2006 第 12 期

[36] 冯哲、吴立德，基于内容的视频检索中的音频处理，复旦大学，2004年11月

[37]邢伟利.基于内容的音频检索技术研究与实现 西北大学学报，2004年6月

[38]卢铮 杨小康，基于内容的多模态视频检索 ，上海交通大学学报，2008年1月

[39] 李恒峰; 李国辉; 基于内容的音频检索与分类 计算机工程与应用, COMPUTER ENGINEERING AND APPLICATIONS, 编辑部邮箱 2000年 07期

[40]张静 俞辉，一种多模态信息融合的视频检索模型，计算机应用 Vol.28 No.1 Jan.2008 Page.199-213

[41] 吴春辉，黄胤科，钟宝荣.基于内容的音频检索技术.《现代计算机（专业版）》2006年 第4期

[42]富亮，薛向阳，郭跃飞. 基于内容的音乐检索研究 复旦大学，2005年2月

[43]宋文静 基于隐性马尔可夫模型的音频检索 南京理工大学，2004 年 3月

[44] 吕紫东， 基于内容的视频检索，现代计算机 第275期，51-54

[45] 浦剑涛、王辉、姜红尘, 徐波 语音和非语音的音频特征分析. 计算机学报, 声学技术， 2006（24）91-104

[46] 刘林海 陈永炜; 视频检索技术在有线电视监测系统中的应用; 广播与电视技术 2008年第6期 33-38

[47] 曹蓁 李建华 李翔; 网络媒体内容监控技术与方法研究; 信息网络安全 2005年11期 50-52

附　录

附录 1：图片目录

图 1 视频的结构化信息[36] 8
图 2 多模态的基于内容的视频检索的一般模型[9] 10
图 3 多模态的基于内容的视频检索的一般模型[40] 10
图 4 基于音频的视频内容检索系统模型 12
图 5 基于音频的视频流媒体内容检索的一般模型 14
图 6 语音、音乐、噪音的 HZCRR 对比 19
图 7 语音、音乐、噪音的 LSTER 对比 20
图 8 语音、音乐、噪音的 NFR 对比 21
图 9 语音、音乐、噪音的 SC 方差对比 22
图 10 语音的频谱图 23
图 11 音乐的频谱图 23
图 12 噪音的频谱图 24
图 13 语音信号的线性模型 25
图 14 LPC 预测 27
图 15 “帐篷”语音短时 LPCC 系数图 28
图 16 MFCC 计算的流程框图 29
图 17 “帐篷”语音短时 MFCC 系数图 30
图 18 “帐篷”语音的波形图和语谱图 32
图 19 “帐篷”语音信号预处理前后波形、频谱的对比 34
图 20 音频检索系统典型框架 39
图 21 音频检索流程图 41
图 22 特征集合的分层 42
图 23 分层向量模板分类流程图 42
图 24 改进的分层分类流程图 43

图 25 FWDM 算法流程图 45
图 26 未融合前系统框图 47
图 27 融合（1）后的系统框图 47
图 28 融合（2）后的系统框图 48
图 29 实验系统框图 49
图 30 实验系统音频流处理流程图 50
图 31 分类算法效率对比图 52
图 32 匹配门限值与检索时间关系图 53
图 33 特征值门限和命中率、误判率关系图 54

附录 2：表格目录

表 1 RMS 对分类的影响 18
表 2 SR 对分类的影响 18
表 3 ZCR 对分类的影响 18
表 4 HZCRR 对分类的影响 19
表 5 STE 对分类的影响 19
表 6 LSTER 对分类的影响 20
表 7 NFR 对分类的影响 21
表 8 SC 方差对分类的影响 22
表 9 BW 对分类的影响 23
表 10 音频特征全集 35
表 11 音频特征值对分类的影响 36
表 12 实验选取的特征向量及分类结果 36
表 13 特征有效性实验结果 37
表 14 分类向量取值表 43
表 15 分类算法计算量比较（1） 43

表 16 分类算法计算量比较（2） 44
表 17 实验选取的特征向量及分层结果 44
表 18 匹配算法效率比较 51
表 19 检索算法可靠性比较 51
表 20 检索方法可靠性比较 53

附录 3：缩略语

VR： Video Retrieval 视频检索

CBVR：Content Based Video Retrival 基于内容的视频检索

ZCR: Zero-Crossing Rate 过零率

HZCRR: High Zero-Crossing Rate Ratio 高过零率帧比率

STE: Short-Time Energy 短时能量

RMS: Root-Mean Square 短时能量均方值

LSTER: Low Short-Time Energy Ratio 低能量帧比率

SFR: Silent Frame Ratio 静音帧比率

SF: Spectrum Flux 频谱差分幅度

SC: Spectral Centroid 频谱质心

BW: Band Width 频谱宽度

SRF: Spectral Rolloff Frequency 频谱截止频率

NFR: Noise Frame Ratio 噪音帧比率

SBER: Sub-Band Engergy Ratio 子带能量比

LPCC：Linear Prediction Cepstral Coefficient 线性预测倒谱系数

LSP: Line Spectrum Pair 线谱对

MFCC: Mel-Frequency Cepstral Coefficient 梅尔倒谱系数

HCMBVT: Hierarchical Classify Method Based on Vector Templates 分层的向量模板分类技术

FWDM: Forwarding-sequential Weighted-feature distance measuring 前向序列特征加权距离算法

致谢

在攻读硕士的两年半期间，感谢上海交通大学图像通信研究所为我提供的良好的科研环境，为我的科研工作创造了必备的条件，在这个团结一心、积极进取的大集体里，令我感受到崇高的荣誉感。

感谢杨小康老师、周军老师、胡剑凌老师、方向忠老师给予我的指导，他们广博的学术造诣、严谨的治学态度令我深深敬佩，给我留下了深刻的印象。感谢支诤老师给予的无私关怀与帮助。

感谢刘晓东、王国忠、纪志胜、郑剑锋、彭媛、何子由等同学给予我的帮助，正是他们的友爱互助的精神，促使我在学习的道路上取得重大进步，克服了诸多困难，最终完成了硕士论文。

感谢我的家人，他们给予了我最重要的无私支持。

最后感谢所有给予我帮助的人们，我将在以后的学习工作中做出最大的贡献来感谢他们！

攻读学位期间发表的学术论文目录

[1]时金、周军， 流媒体中的实时语义检测，信息技术，2009 年第 3 期