清 华 大 学

综 合 论 文 训 练

题目：基于音频事件检测的语音分割算法研究

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：吕萌

指导教师：徐明星 副教授

2009 年 6 月 24 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

随着计算机技术和网络技术的不断发展，包括图像、视频、音频在内的多媒体数据已成为了互联网上的主要信息媒体，其中音频信息占有很重要的地位。本文研究的是，基于音频事件检测，将音频流分割成音频事件片段。研究工作包括以下几个内容：

1) 本文选定了五种已知的音频事件类别：Speech，Music，Speech with Music，Advertisement和Song。对于其他不属于这五类的音频事件，统一归为未知类别。本文基于音频分析的层次化结构，根据模式识别的思想，提出了一个基于音频事件检测的语音分割的分层算法。

2) 本文重点讨论了音频特征的选择和提取方法，并考察了不同音频特征对算法性能的影响。本文选择了多种音频特征，包括时域特征（短时能量和短时过零率），频域特征（MFCC、LPCC和PLP），以及韵律特征（基音频率和共振峰）。对应于本文所设计的分层算法，使用了基于音频Frame和音频Clip的音频特征提取和分析方法。

3) 分类器的设计也是本文重点内容之一。本文选择了三种基于统计学习算法得到的分类器：多维高斯分布、高斯混合模型和支持向量机，并考察了不同分类器对算法性能的影响。同时本文还考虑了基于阈值对未知类别音频事件的判断。

关键词：音频事件检测；音频分类；特征提取；高斯混合模型；支持向量机

ABSTRACT

With the development of computer technology, network technology, multimedia data, which includes image, video and audio, becomes primary information media on the Internet. Audio data is an important part of multimedia applications. This work focuses on segmenting audio stream to acoustic event segments, based on Acoustic Event Detection. It can be concluded as follows:

1) This work selects five known classes of acoustic events: speech, music, speech with music, advertisement and song. Any other acoustic event which does not belong to these known classes will be classified to an unknown class of acoustic events. Based on a hiberarchy framework for audio structure analysis and according to the idea of pattern recognition, a layered audio segmentation algorithm based on Acoustic Event Detection is proposed in this work.

2) This work focuses on the methods of selection and extraction of the audio features, and studies the different algorithm performance of using different audio features. This work chooses a variety of audio features, including time-domain features (short-time energy and short-time zero-crossing rate), frequency-domain features (MFCC, LPCC and PLP), and prosodic features (pitch and formant). This work uses both Frame-based and Clip-based audio feature extraction and analysis methods, corresponding to the layered algorithm.

3) The design of classifier is also one of the key elements of this work. Three types of classifier based on statistical learning algorithm are chosen in this work: multi-dimensional Gaussian distribution, Gaussian mixture models and support vector machines. This work studies the different algorithm performance of using different classifiers. At the same time, this work also considered the judgment of unknown class of acoustic audio events using threshold.

Keywords：Acoustic Event Detection; Audio Classification; Feature Extraction; Gaussian Mixture Model; Support Vector Machine

目 录

第1章 引言 1

1.1 研究背景 1

1.2 研究现状 2

1.2.1 音频特征提取 3

1.2.2 音频分类方法 3

1.2.3 音频分割方法 4

1.2.4 音频事件检测 4

1.3 研究内容 5

1.4 论文结构 5

第2章 研究思路与总体安排 6

2.1 研究基础 6

2.1.1 音频分析的层次化结构 6

2.1.2 典型的模式识别过程 7

2.2 总体研究思路 8

2.2.1 技术流程 8

2.2.2 实验思路 12

第3章 音频特征的分类与提取 14

3.1 时域特征 14

3.1.1 短时能量 14

3.1.2 短时过零率 15

3.2 频域特征 16

3.2.1 Mel频率倒谱系数（MFCC） 16

3.2.2 线性预测倒谱系数（LPCC） 17

3.2.3 感知线性预测系数（PLP） 18

3.2.4 HTK工具包介绍 20

3.3 韵律特征 20

3.3.1 基音频率（Pitch） 20

3.3.2 共振峰（Formant） 21

3.3.3 Praat语音学软件介绍 21

3.4 多层音频特征提取 22

第4章 分类器的理论基础与相关算法 24

4.1 多维高斯分布 24

4.1.1 基本概念 24

4.1.2 基于多维高斯分布的分类器 24

4.2 高斯混合模型（GMM） 25

4.2.1 基本概念 25

4.2.2 参数估计 26

4.2.3 识别过程 27

4.3 支持向量机（SVM） 28

4.3.1 理论基础及训练方法 28

4.3.2 LIBSVM软件包介绍 29

4.4 多层分类器 30

4.4.1 多层分类器结构 30

4.4.2 对未知类别的处理 32

第5章 实验设计与结果分析 33

5.1 性能评价指标 33

5.2 实验结果 33

5.3 结论 36

5.3.1 不同音频特征对性能的影响 36

5.3.2 不同分类器对性能的影响 38

5.3.3 判断未知类别的阈值选择 39

第6章 总结与展望 41

6.1 本文内容总结 41

6.2 今后的工作与展望 42

插图索引 43

表格索引 44

参考文献 45

致 谢 47

声 明 48

附录A 书面翻译 49

# 第1章 引言

## 1.1 研究背景

随着现代信息技术的逐渐发展，数据存储能力的不断增强，同时互联网的迅速蔓延，使得人们在日常生活中频繁接触各种数字媒体。以图像、视频、音频为代表的多媒体数据成为了互联网上的主要信息媒体，其中音频信息占有非常重要的地位。因此人们在日常生活中所面临的问题不再是缺少多媒体信息，而是如何在数以万计的多媒体信息中准确并迅速地找到自己所需要的信息，尽管文字搜索引擎已经发展成熟，但却无法应对大量的多媒体信息检索。因此多媒体信息检索技术的发展是现代信息技术发展过程中必经的一条道路，音频信息检索是多媒体信息检索技术中必不可少的一部分。

音频检索作为音频信息检索的重要部分之一，具有广泛的实用价值，可以应用于互联网音频信息搜索、音像馆和图书馆的资料管理以及满足公安、安全部门的业务等诸多需要[1]，是目前研究的一大热点。但由于原始音频数据除了含有采样率、量化精度、编码方法等有限的注册信息外，本身仅仅是一种非语义符号表示和非结构化的二进制流，缺乏内容语义的描述和结构化的组织，极大限制了音频检索的发展[2]。对于音频这类具有信息关联度大、数据结构复杂、数据量大、处理要求高等特点的非结构化数据，如何提取其中的结构化信息和内容语义，使无序的音频数据变得有序，是音频信息深度处理、分析和基于内容检索的关键[3]，是后续音频检索应用的基础。

人们往往关心对自己有用的音频信息，比如从新闻广播中查询特定的音频段（广告、新闻播报、音乐等）。另外，在某些特定的情况下需要锁定特定的声音片段，比如欢呼声、枪声、脚步声等，这些信息对于监控非常有用[4]。音频事件检测的重要性由此而显现出来。基于音频事件检测，将音频流按照音频事件分割成段，有很广泛的应用价值。

第一、音频信息检索基本上都是针对单体声音进行研究和处理，但实际的情况是一段录音常包含多种类别的声音。将音频流分割成独立类别的事件片段，对于音频检索系统，便于在某个片段中对单体声音进行查询和辨认，也可以用于音频检索系统的索引建立，提高检索系统的性能；对于很多识别系统，比如说话人识别系统、音乐识别与检索系统、关键词识别系统等，能够将目标锁定在某个分割后的片段中来进行，提高系统的识别性能。

第二、视频信号的自动切分和分类是多媒体技术的一个重要内容，但由于受技术发展的限制，仅靠现有的图像和视频处理技术，计算机无法“理解”视频的内容，视频镜头切分准确度不高，不能对故事单元进行基于内容的分类。通过对音频流进行适当的分割，切分成不同类别事件片段，可以辅助视频的分割，提高分割的准确率[3]。

综上所述，音频检索在多媒体信息技术中占有重要的地位，但由于音频本身的特点和现有技术的限制，如何提取音频中的结构化信息和内容语义，进一步分析音频数据尤为重要。根据人们对音频信息的需求，音频事件检测的重要性逐渐显现，基于音频事件检测的语音分割对于其他多媒体信息技术都有很重要的意义。

## 1.2 研究现状

音频是一种典型的媒体，是信息的一种常用载体，可以分为三种主要类型：1) 语音，具有字词、语法等语素，经过识别可以转换为文本；2) 音乐，具有节奏、旋律和声音等要素，由人声和乐器音响等配合所构成，可以用乐谱来表示；3) 波形声音，对模拟声音数字化而得到的数字音频信号，可以代表语音、音乐、自然界和合成的各种声音[5]。音频信息检索是多媒体信息检索的重要部分之一。

由于不同类别的音频有不同的特性，相对应的，音频信息检索主要分为三类：1) 话音检索，以语音为中心的检索，采用语音识别等相关处理技术；2) 音乐检索，以音乐为中心的检索，主要根据音符、旋律等音乐特性来进行判断、检索；3) 音频检索，以波形声音为对象的检索，这里的音频可以是各种声音，包括语音和音乐等，这些音频统一根据声学特征来进行检索[5]。

早先的信息检索研究主要是基于文本，并且现在发展已比较成熟，主要是基于对关键字的查询。但对于音频信息，它缺少类似于关键词这样可以用来进行匹配的实体。因此首先可以直接想到的是基于人工标注的属性和描述来进行音频检索，将语音转换为文字，再对文字进行检索。然而这种方法的缺点后来逐渐体现出来，它工作量大，人对音频的感知无法用文字表达清楚，且无法满足实时音频数据流的检索要求[2]。基于内容的音频检索因此出现，从音频数据中提取并分析音频特征信息，进行音频检索。首先对音频数据进行特征提取和分析，建立音频和特征数据库，然后进行音频分割、识别和音频检索。因此音频检索的关键技术包括了音频特征提取、音频分类方法和音频分割技术。音频事件检测则是音频检索中的一个分支。

### 1.2.1 音频特征提取

音频特征提取与分析是音频分类的基础，所选择的特征应该能够充分体现音频在时域和频域上的重要分类特性。

Muscle Fish是一个商业化的基于音频感知特征的音频检索引擎，由Erling Wold等人于1996年提出。Muscle Fish分析了音频数据的听觉特征，包括基音频率、振幅、声音响度、带宽和倒频谱，可以对语音、音乐和其他音频数据分类，并对语音和音乐做深入的分析[6]。

Carnegie Mellon大学的Tsuhan Chen和Polytechnic大学的Zhu Liu等人于1997年对音频检索研究中常用的音频特征进行了详细的分析，包括短时能量，短时过零率，音调，带宽，短时频谱，频谱质心，Mel频率倒谱系数等[7]。

Jonathan Foote于1999年提出了一种基于量化树的方法，它提取音频数据的倒频谱特征MFCC，利用音频数据的频谱表示并构造一个量化树，最后的特征是一个直方图[8]。

随着技术的发展，如何提取特征或者提取何种新特征能够更能准确地表示音频类别特性，将是音频特征提取和分析的研究重点。

### 1.2.2 音频分类方法

音频分类方法目前主要有以下三类：

基于规则的音频分类方法。选取合适的音频特征，设定合适的阈值，根据事先约定的规则，将实际计算得到的特征值和阈值相比，来识别音频的类别。在大部分传统音频分类工作中，这种方法应用广泛[6, 8]。这种方法操作简单，因此也只适用于识别特征简单的音频类型。这种方法的缺点比较明显，由于基于规则的判断层层递进，上层的决策错误往往会积累到下一层，误差随层次的增多而增大，并且阈值的确定需要很多先验知识以及人工实验分析。这种方法分类精度较低，难于满足复杂的、多特征的音频分类应用。

最小距离音频分类。利用模板匹配的思想，所有音频特征组成一个特征向量，为每一个音频类别建立一个模板特征向量，计算实际音频帧的特征向量，通过计算特征向量和模板特征向量在向量空间中的距离来识别音频的类别。Andreas Rauber等人于2002年提出的基于Self-Organizing Maps的音乐聚类系统中采用的就是这种方法[9]。这种方法在说话人识别的研究中也被广泛使用。

基于统计学习算法的音频分类。事先给定一批带有类别标记的训练样本，通过有指导的学习训练来生成分类器，然后对测试样本集中的待分类样本进行测试、分类。其中比较典型的分类方法包括基于支持向量机的分类方法、基于神经网络的分类方法、基于决策树的分类方法和基于隐马尔可夫模型的分类方法等。Erling Wold等人采用最近邻算法构造分类器[6]；Zhu Liu等人采用基于神经网络的分类方法，用于电视节目的分类[7]；Eric Scheirer等人使用了不同算法构造分类器，包括高斯混合模型和K-近邻算法[10]；微软研究院的Lie Lu等人提出了一种鲁棒的分类方法，采用的是基于K-近邻模型的音频分类方法[11]，并提出了一种基于支持向量机的分类方法，用于对静音、音乐、环境音、纯语音和含音乐和噪音的非纯语音的分类[12]；南京大学的卢坚博士等人提出了一种基于隐马尔可夫模型的音频分类方法，用于语音、音乐以及它们的混合声音的分类[13]。基于统计学习算法的音频分类方法是目前音频分类方法研究的重点。

### 1.2.3 音频分割方法

在音频检索的研究中，音频分割方法和音频特征提取及音频分类方法这两类技术相比，研究相对较少。

传统的音频分割方法通常采用简单的滑动窗口方法[6, 8]，用一固定长度的滑窗对音频流进行简单分割，在滑窗内部按“投票规则”将音频流平滑为一个类别，然后将具有相同类别的段合并得到最终的分割结果。

微软研究院的Lie Lu等人根据音频流的特性，提出了一些平滑规则，根据这些规则对音频流进行平滑，然后合并具有相同类别的段。这个方法本质上是对分类结果的修正[11, 12]。

### 1.2.4 音频事件检测

音频事件检测是近年才发展起来的，有非常广泛的应用价值。在早先的研究中，侧重点各有不同。Lie Lu等人对音频分类和分割的研究[14]和Julien Pinquier等人对影视原声中语音和音乐的分类的研究[15]侧重于将音频分割成少数几类；在Daniel P. W. Ellis的研究[16]中侧重于分离几个音频源；在Rui Cui等人的研究[17]中侧重于检测一些突出的音频事件。

Andrey Temko等人给出了在小房间的环境下CHIL方案系统对音频事件检测与分类的评估结果。他们尝试在一个真实的研讨会环境下识别12类音频事件（脚步声、键盘敲击声等）的时间边界和标记[18]。

如何能够对于更多种类别的音频事件、对更普遍的声音进行检测与分类，这将是音频事件检测技术的发展重点。

## 1.3 研究内容

本文的研究目标是，对于给定的时间较长的连续语音，设计并实现算法，能够检测出不同类别音频事件的时间边界，将连续语音分割成若干个小的音频片段，每个片段内包含且仅包含一种类别的音频事件。

音频事件的类别非常多且非常细，我们无法对每种类别给出一个特定的识别、检测算法。如何能够用一个一般性的、对绝大部分音频事件类别适用的算法，对音频事件进行检测并分割成段，是本文需要重点研究的部分。

本文的研究内容主要包括以下两个方面：

1) 音频特征的提取和分析。这是音频分类的基础，而音频分类是音频事件检测的重要部分之一。所选择的特征应该能够充分体现音频在时域和频域上的重要分类特性。本文将使用基于音频Frame和音频Clip的音频特征提取和分析方法。并且本文将考察使用不同特征集的分类性能。

2) 分类器的设计。本文将使用基于统计学习算法的音频分类方法，将考察通过三种不同的训练模型和算法得到的分类器的分类性能，这三种方法分别为基于多维高斯分布、基于高斯混合模型和基于支持向量机的分类方法。

综上所述，通过提取不同的音频特征集、使用不同的音频分类方法，对比它们的性能，最终确定一个相对最优的具有普遍性的算法，使得基于音频事件检测的语音分割的准确率最高。

## 1.4 论文结构

本文共六章。第二章将提出一个分层算法的总研究思路和技术流程，并依此设计实验。第三章将介绍各类音频特征的提取和分析，以及所使用到的相关工具。第四章将介绍三种分类器的理论基础与训练算法，以及所使用到的相关工具。第五章将介绍完整的实验设计思想和方法，并给出所有实验结果以及对结果的分析。最后一章将会总结和评价论文的全部工作，并对未来的研究工作进行展望。

# 第2章 研究思路与总体安排

## 2.1 研究基础

### 2.1.1 音频分析的层次化结构

对音频进行分析，是以短小的音频片段作为分析的结构单元的。结构单元的时间粒度如果较小，则在技术上处理更方便，但由于时间过短很难从中提取有价值的语义信息。结构单元的时间粒度如果过大，则语义信息可以较完整地提取到，但分析起来技术上难以支持。因此对于音频分析，普遍采用一种层次化的结构，如图2.1所示，从低层到高层为不同时间粒度的结构单元，对应不同的分析方法。

音频

音频高层语义单元

Segment 2

Segment 1

Segment n

Clip 2

Clip 1

Clip n

Frame 2

Frame 1

Frame n

…

…

…

特征提取

音频分类

音频分割

…

技术基础

图2.1 音频分析的层次化结构

不同时间粒度的音频结构单元的定义如下：

Frame：音频帧。音频是一个非平稳随机过程，其特性随时间变化，但变化缓慢，因此音频信号是短时平稳的，从短时间的音频帧中提取的特征比较稳定。音频帧是音频处理的最小单元，一般长20~30ms。

Clip：音频小片段。任何语义都有时间延续性，虽然音频信号是短时平稳的，但是从短时音频帧中提取的特征无法反映音频所蕴含的语义信息，所以还需要较大时间粒度的结构单元。Clip由若干帧组成，长度固定。

Segment：音频片段。它由若干个相同类别的Clip组成，时间粒度更大，时间长度不定，是音频分割的结果，具有一定的语义。

音频高层语义单元：由许多Segment的不同组合形成，具有完整、丰富的语义信息。此层结构单元不在本文研究范围之内。

从最小粒度的结构单元开始进行分析。基于Frame和Clip进行音频特征的提取和分析，基于Clip进行音频分类，对分类结果分割成Segment。

### 2.1.2 典型的模式识别过程

音频分类本质上是一个模式识别的过程，满足模式识别的一般性过程，如图2.2所示，其中特征的选择和分类器为关键。

特征提取

预处理

传感器

分类器

后处理

输入

决策

图2.2 模式识别的一般性过程

训练分类器

评价分类器

选择模型

选择特征

采集数据

开始

结束

图2.3 分类器的设计流程

分类器作为模式识别的关键，通常不断重复几个步骤，如图2.3所示。首先需要采集用于训练和测试的数据；然后选择合适的特征、合适的训练模型或算法，训练分类器，确定分类器的参数；最后根据测试评价的结果，通过多次修改参数和模型的选择，得到最优的分类器。

## 2.2 总体研究思路

基于音频分析的层次化结构，根据模式识别的思想，本文设计了一个基于音频事件检测的语音分割的分层算法，并依此设计实验思路。

### 2.2.1 技术流程

本文所处理的音频数据采集方式为：收集网络直播的电台广播、电视节目等的音频，使用Cool Edit内部录制计算机声卡的声音，44.1kHz采样，16位单声道，统一为wav格式。

本文所涉及到的已知音频类别有五种：纯语音（Speech），纯音乐（Music），带背景音乐的语音（Speech with Music），广告（Advertisement），带有人声的音乐/歌曲（Song）。对于其他不属于这五类已知类别的音频段，统一划分为未知类别（Unknown）。

本文设计的基于音频事件检测的语音分割的分层算法，其技术流程如图2.4所示。各部分介绍如下：

1) 预处理

为方便研究和实验，本文中处理的音频对象数据格式全部统一。因此在预处理的时候，主要任务是对音频流加窗成Frame，再加窗成Clip。每一个Frame包含1024个采样点，长约23ms，相邻两个Frame有256个采样点的重叠。基于韵律特征进行分析时每一个Clip包含75帧，相邻两个Clip有30帧的重叠；基于频域特征进行分析时每一个Clip包含50帧，相邻两个Clip有20帧的重叠。对输入音频流进行预处理之后，就可以在此基础上进行后续处理了。

2) 特征提取

音频特征的提取和分析是音频分类的基础，选取合适的特征对于音频分析将会非常有利。在这一部分将会基于Clip提取韵律特征，基于Frame提取频域特征，构造特征集，为后面的分层分类做准备。关于音频特征的选择和提取，将在第三章做详细介绍。

Frame特征提取

Clip特征提取

特征提取

加窗成Clip

加窗成Frame

预处理

静音段端点附近应用分类器

VAD

切分静音部分

基于频域特征的第二层分类

Clip类别标注

对非静音部分分类

基于韵律特征的第一层分类

基于一定规则修正分类结果

合并成Segment

分割

特征提取

预处理

分类器设计

训练分类器

选择模型

评价分类器

测试数据

训练数据

结果

图2.4 基于音频事件检测的语音分割的分层算法技术流程图

3) 分类器的设计

预处理和特征提取两部分，对于训练数据和测试数据是同样处理的。本文将会选择三个不同的模型来训练三个不同的分类器，通过比较各自的测试结果，最终选择性能最好的一个。这三个不同的模型分别是：多维高斯分布，高斯混合模型（Gaussian Mixture Model, GMM），以及支持向量机（Support Vector Machine, SVM）。关于分类器，将在第四章做详细介绍。

4) 切分静音部分

如图2.5所示为音频事件类别的层次划分。切分静音部分是分层算法中的第一层处理，将静音部分切分出来，便于在后续分类中有针对性地对非静音部分进行处理。首先使用基本的VAD算法，将非静音部分切分出来。由于切分必定存在误差，因此在静音片段的端点附近使用分类器，进一步判断该部分属于Silence或者属于Voice。同时，将非静音片段间较短的静音片段归并至非静音片段。因为音频事件分割对于静音和非静音的切分的精度要求并不是很高，只需要将非静音部分找出来、将大部分静音部分切分出去即可，所以这些处理都是可行的。

Speech with Music

Music

Advertisement

Song

Audio

Silence

Voice

With Music

Speech

Without Music

图2.5 音频事件类别的层次划分

VAD，即Voice Activity Detection，是基于短时平均能量（短时平均幅度）和短时过零率这两种时域特征、进行有声和无声的检测算法。其基本原理就是，预先设定两个短时能量的阈值EH（MH）、EL（ML）和一个短时过零率的阈值ZS，根据阈值比较进行前后搜索，确定非静音部分，如图2.6所示。

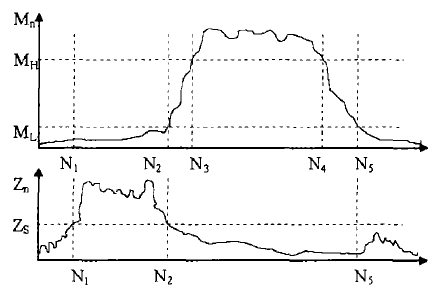


图2.6 基本VAD算法示意图

首先进行第一次搜索比较。当信号的幅度大于MH时，该点所在帧可以确定为非静音帧，记为N3，但N3还不能准确标记非静音部分的起点。接着进行第二次搜索比较，从点N3开始向前寻找第一个幅度等于阈值ML的信号点，记为N2，该点比较准确地确定了非静音部分的起点。但某些清音幅度和静音相差不大，都不高，前两次搜索比较并不能很好地加以区分，因此要根据静音和清音的短时过零率的不同来进行第三次搜索比较。从点N2开始向前搜索过零率刚好低于阈值ZS三倍（经验值）的点，为了减小误差，向前搜索的最大时间长度定为约一帧时长。找到的点就定为非静音部分的起点，同理可以确定非静音部分的终点[19, 20]。

VAD算法的重点在于确定和调整阈值EH（MH）、EL（ML）和ZS。本文对此不做研究，根据文献[20]中所总结的内容，取MH和ML分别为0.168和0.068，或者取EH和EL分别为0.1682和0.0682，取ZS为30。

经过这一层处理，将静音部分切分出来，后续分类中就可以有针对性地对非静音部分进行处理了。

5) 对非静音部分分类

分层算法中的第二层处理，即对非静音部分分类，而在这一部分中又将分为两层进行分类。第一层将非静音部分分为含音乐部分（With Music）和不含音乐部分（Without Music）这两大类，使用韵律特征。第二层将第一层分类得到的含音乐部分进一步分为Music、Speech with Music、Advertisement和Song四小类，使用频域特征，若四类均不属于则划分到Unknown类。不含音乐部分则划分为Speech类。由于韵律特征并非对所有音频都适用，即并非对所有音频数据都能提取韵律特征，因此会有一些非静音部分没有经过第一层分类，需要在第二层的时候直接对其进行处理，同样，若五类均不属于则划分到Unknown类。关于特征的选择和如此分层处理的思想将在第三章介绍，关于Unknown类别的划分将在第四章介绍。

6) 分割

音频的语义内容具有时间连续性，基于这一特性可以设计分割准则，对非静音部分的Clip序列进行平滑，对前面的分类结果进行修正，然后将具有相同类别的Clip合并成音频事件片段。本文采用的分割准则有如下四个：

准则1：设C1、C2和C3为三个相邻的Clip，如果它们的类别各不相同，则认为C2的类别应与C1相同[3]。

准则2：设C1、C2和C3为三个相邻的Clip，如果C1和C3属于同一类别，而C2与C1、C3的类别不同，则认为C2的类别判断有误，C2应与C1、C3的类别相同[3]。

准则3：设C1、C2和C3为三个相邻的Clip，如果C1为一个非静音部分的第一个Clip，C2和C3属于同一类别，而C1与C2、C3的类别不同，则认为C1的类别判断有误，C1应与C2、C3的类别相同。

准则4：设C1、C2和C3为三个相邻的Clip，如果C3为一个非静音部分的最后一个Clip，C1和C2属于同一类别，而C3与C1、C2的类别不同，则认为C3的类别判断有误，C3应与C1、C2的类别相同。

经过分割后得到的结果即若干个分属于不同类别的音频事件片段。

以上详细介绍了本文设计的基于音频事件检测的语音分割的分层算法技术流程。不同类别的音频事件具有不同的特性，选取不同的音频特征，选择不同的模型，将各自的优点结合起来，设计形成一种针对音频事件检测的具有一般性的分层算法。关于音频特征和分类器将在后两章中详细介绍。

### 2.2.2 实验思路

本文的研究重点在音频特征的提取分析和分类器的设计这两个方面上。通过考察不同种类的音频特征对音频事件检测、分割性能的影响，选择最合适的特征组合。同时还考察从不同的训练模型和算法得到的分类器对音频事件检测、分割性能的影响，选择最合适的分类器。

对于音频特征，本文中进行了几种选择和组合，包括MFCC、LPCC、PLP，以及韵律特征中的基频（Pitch）和共振峰（Formant）。对这些音频特征的详细介绍和分析将在第三章中进行。

对于分类器的设计，本文选择了几种不同的训练模型和算法进行考察和研究，包括多维高斯分布，高斯混合模型（GMM），以及支持向量机（SVM），其中重点放在后两种训练模型和算法上。对分类器的详细介绍将在第四章中进行。

根据多种特征、多种训练模型和算法的选择，本文设计了如表2.1和表2.2所示的分组实验方式。

表2.1 单层分类算法的实验设置

|  |  |  |  |
| --- | --- | --- | --- |
|  | 多维高斯分布 | GMM | SVM |
| MFCC | √ | √ | √ |
| LPCC | √ | √ | √ |
| PLP | √ | √ | √ |

表2.2 两层分类算法的实验设置

|  |  |  |  |
| --- | --- | --- | --- |
|  | GMM+GMM | SVM+GMM① | SVM+SVM |
| Pitch+Formant+MFCC② | √ | √ | √ |
| Pitch+Formant+LPCC | √ | √ | √ |
| Pitch+Formant+PLP | √ | √ | √ |

① 第一层分类使用SVM训练算法，第二层分类使用GMM训练模型。

② 第一层分类使用Pitch和Formant特征，第二层分类使用MFCC特征。

# 第3章 音频特征的分类与提取

音频特征的提取和分析是音频分类的基础，而音频分类是音频事件检测的重要部分之一。所选择的特征应该能够充分体现音频在时域和频域上的重要分类特性。选取合适的特征对于音频分析将会非常有利。

音频是一个非平稳随机过程，其特性随时间变化，但变化缓慢，因此音频信号通常被认为是短时平稳的，短时的音频段被认为具有固定的音频特性。这些短时的音频段即音频帧（Frame），是音频处理的最小单元，一般长20~30ms，相邻帧会有部分重叠。根据音频信号的短时平稳特性，每一帧可以认为是一小段具有固定特性的持续音频。对音频帧进行特征提取和分析，即对连续音频流加窗做周期性的处理。

本章将从时域特征、频域特征和韵律特征这三个方面来介绍音频特征的提取和分析，以及所使用到的一些现有工具，并根据各自的特点，针对本文算法来设计实验中使用的音频特征组合。

## 3.1 时域特征

音频的时域特征种类很多，但根据实际研究需求会有不同的选择和组合。本文的研究中只选取了其中的两种，分别为短时能量（Short-Time Energy）和短时过零率（Short-Time Zero-Crossing Rate），用于VAD算法，对静音和非静音部分进行切分。

### 3.1.1 短时能量

短时能量（Short-Time Energy，STE）的计算公式如式3-1所示，其中为一个音频帧的离散信号，包含个采样点。

 （3‑1）

静音的能量和非静音相比要小很多，因此，能量是区分静音和非静音的有效特征。本文使用的VAD算法即采用了短时能量这一特征，检测大部分的非静音音频信号。对于VAD算法中的阈值EH和EL，本文未作研究，根据文献[20]中所总结的内容，取EH为0.1682，取EL为0.062。

仅从能量上来区分静音和非静音，很容易将非静音中清音语音划分到静音部分，因为清音语音的能量也比较低。因此仅提取短时能量这一特征是不够的，还需要提取其他的时域特征。

### 3.1.2 短时过零率

短时过零率（Short-Time Zero-Crossing Rate，STZCR）表示一帧语音中音频信号波形穿过横轴（零电平）的次数。输入离散音频信号，如果相邻采样点的值具有不同的代数符号，则称之为发生了过零。计算公式如式3-2所示，其中为一个音频帧的离散信号，包含个采样点，为符号函数，如式3-3所示。

 （3‑2）

 （3‑3）

语音信号是由浊音和清音交替组成，同时有发音的音节和不发音的音节。浊音语音的能量约集中在低频部分，清音语音的能量则集中在较高频率部分。由于频率的高低与过零率的高低有对应关系，对于清音语音，其过零率高，对于浊音语音，其过零率低，所以语音信号的过零率的变化很大，如图3.1所示。

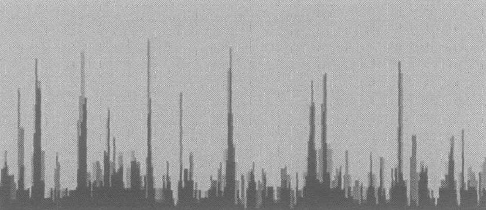


图3.1 一段语音的过零率图

静音的过零率和清音语音相比要小很多，因此使用短时过零率这一特征可以减小将清音语音误分到静音的这种情况。VAD算法结合了短时能量和短时过零率这两种时域特征，以实现对静音和非静音部分的基本切分。对于VAD算法中的阈值ZS，本文未作研究，根据文献[20]中所总结的内容，取ZS为30。

## 3.2 频域特征

本文的研究中使用了Mel频率倒谱系数（Mel-Frequency Cepstrum Coefficient，MFCC）、线性预测倒谱系数（Linear Prediction Cepstrum Coefficient，LPCC）和感知线性预测系数（Perceptual Linear Prediction，PLP）这三类特征。LPCC是基于人的发声模型的特征参数，MFCC和PLP是基于人的听觉模型的特征参数。LPCC和PLP的计算方法有部分相似。因此，这三种特征有较强的相关性。在选择音频特征组合的时候应选择不具有相关性的特征进行组合和分析。所以对于这三种特征，本文将对它们进行横向比较，考察和分析分别选用这三种特征对于算法性能的影响。

### 3.2.1 Mel频率倒谱系数（MFCC）

MFCC是基于人的听觉模型得到的声学特征。根据人耳的听觉机理，当两个音调的频率差小于一个带宽值时，人的主观感觉会发生变化，会把两个音调听成一个，称之为屏蔽效应。这个带宽值称为临界带宽，Mel刻度是对这一临界带宽的度量方法之一。根据人耳听觉的临界带这一特性，依照Mel刻度设置一组带通滤波器，将音频信号的能量谱通过该滤波器组，对输出取对数，并做离散余弦变换（DCT），最终得到MFCC参数。其提取流程如图3.2所示。

Mel滤波器组

取对数

FFT

加窗成帧

DCT

音频信号

MFCC

图3.2 MFCC的提取流程

提取MFCC特征的具体流程如下：

1) 对输入音频信号进行加窗分帧处理，对每帧进行预加重处理后按式3-4做快速傅立叶变换（FFT），获得频谱信息，进一步按式3-5计算得到能量谱。

 （3‑4）

 （3‑5）

2) 将线性频率按式3-6转换到Mel频率，然后基于Mel频率刻度，设置一组共M个具有三角形或正弦形滤波特性的滤波器，按照式3-7计算通过滤波器的输出，并取其自然对数得到，。

 （3‑6）

 （3‑7）

3) 对按式3-8做离散余弦变换（DCT），得到，。M为MFCC参数的阶数。

 （3‑8）

计算MFCC的一阶差分和二阶差分，和MFCC一起作为完整的特征向量。在本文的研究中，提取了12维MFCC和1维能量值，加上其一阶差分和二阶差分，形成共39维的特征向量。

MFCC参数充分考虑了人耳的听觉特性，在有信道噪声和频谱失真的情况下，识别精度仍然较高，抗噪能力也比较强。

### 3.2.2 线性预测倒谱系数（LPCC）

LPCC是基于人的发声模型得到的声学特征。全极点线性预测模型可以对人的声管模型进行很好的描述，n时刻的信号可以用前若干时刻的信号的线性组合来估计。通过使实际语音的采样值和线性预测值之间的均方差最小，可得到线性预测系数（LPC）。LPC的计算方法有自相关法（德宾Durbin法）、协方差法、格型法等。

LPCC并非由原始音频信号直接得到，而是基于LPC系数计算得到的。由原始音频信号得到P阶LPC系数，，通过如式3-9至3-11所示的递推公式计算得到M阶LPCC系数，。

 （3‑9）

 （3‑10）

 （3‑11）

计算LPCC的一阶差分和二阶差分，和LPCC一起作为完整的特征向量。在本文的研究中，提取了12维LPCC和1维能量值，加上其一阶差分和二阶差分，形成共39维的特征向量。

LPCC参数比较彻底地去掉了声音产生过程中的激励信息，主要反映声道特性，计算量小、速度快，但抗噪声的能力较弱。

### 3.2.3 感知线性预测系数（PLP）

PLP是基于人的听觉模型得到的声学特征。它在临界带分析处理、等响曲线预加重、以及信号强度-听觉响度变换这三个层次上模仿了人耳的听觉感知机理。其提取流程如图3.3所示。

一般LPC处理方法

预处理

FFT

临界带分析

等响度预加重

强度-响度转换

IDFT

音频信号

PLP

图3.3 PLP的提取流程

1) 频谱分析。对输入音频信号做加窗、分帧等预处理后，对每一帧做FFT变换，得到频谱信息，然后计算得到能量谱。

2) 临界带分析。临界频带的划分反映了人耳听觉的屏蔽效应，利用公式3-12将频率轴转换到Bark频率，共N个频带。将转换后的能量谱与式3-13中给出的模拟临界带曲线按式3-14做卷积，得到临界带功率谱。

 （3‑12）

 （3‑13）

 （3‑14）

3) 等响度预加重。人耳对不同频率的声音所感受到的响度并不相同，如式3-15所示，约40dB的等响曲线是对人耳听觉在不同频率的敏感性的一种近似表示。为了体现人耳这一特性，需要对信号按式3-16进行等响度预加重。

 （3‑15）

 （3‑16）

4) 强度-响度转换。为了近似模拟声音的强度和人耳感受到的响度之间的非线性关系，按式3-17进行强度-响度转换。

 （3‑17）

5) 最后对进行IDFT变换，得到相应的自相关函数，进一步利用LPC的处理方法，使用Durbin算法计算全极点模型的系数，并计算倒谱系数，最后得到PLP参数。

计算PLP的一阶差分和二阶差分，和PLP一起作为完整的特征向量。在本文的研究中，提取了12维PLP和1维能量值，加上其一阶差分和二阶差分，形成共39维的特征向量。

PLP参数的提取从人耳的屏蔽效应原理出发，考虑了临界带宽的特性，并且考虑了耳蜗分频特性，计算量小，识别精度较高，且具有较高的稳健性。

### 3.2.4 HTK工具包介绍

HTK（Hidden Markov Model Toolkit）是剑桥大学开发的一个专门用于处理隐马尔可夫模型（HMM）的软件工具包，主要用于语音识别，但同时也被用于许多其他模式识别应用。HTK经过剑桥大学、Entropic公司及Microsoft公司的不断增强和改进，在语音识别领域处于世界领先水平。

HTK的大部分功能被集成到一系列函数库模块，这些模块可以使用相同接口方式和外界进行交互。HTK工具包中包括了许多的函数模块，对于语音处理有完整的处理流程。但本文研究内容并非语音识别，仅有针对性地利用HTK工具包中的HCopy模块提取音频的频域特征。在配置文件中设置不同参数，以帧为单位提取得到12维MFCC、LPCC或PLP特征，1维能量值，加上一阶差分和二阶差分，共39维特征向量。

## 3.3 韵律特征

韵律特征（Prosodic Features）是针对语音提出的。每一个音节在它不被解释为另外一个音节的条件下，它的发音特征以及由此导致的声学表现和人的听觉感知结果仍然在一定范围里是可变化的。这个语音中可变化的并且被赋予言语交流意义的重要部分即韵律特征[27]。由于韵律特征是针对语音而提出，对于非语音（比如音乐）提取得到的特征和语音相比差别很大，或者可能无法提取出有意义的特征来。因此韵律特征在本研究中也作为区分不同类别音频事件的特征之一，主要用于区分含音乐和不含音乐的音频。本文中使用到两种韵律特征，基音频率和共振峰，将两者结合，作为韵律特征组合进行第一层音频分类。

### 3.3.1 基音频率（Pitch）

在日常生活中，人们听到的大部分声音不是纯音，即由正弦波得到的声音，而是复合音，这是由多个不同频率和振幅的正弦波叠加而成。在一个自然的复合音里，有一个振幅最大、频率最低的分量，也就是第一谐波分量，或称为基音，它的振动频率称为基音频率，简称基频。对于人工合成的声音或通过滤波处理后的声音，它的基频并不等于第一谐波的频率。

人耳对复合音的音高的感知，主要决定于基频，基频越高，人耳所感受到的音高也越高，反之则越低。一般来讲，女声的基频比男声高，童声的基频比成人声音高。对于不同的声音，其音高不同，所提取到的基频会不同。

本文中提取基频时，窗长设为10ms，截止频率范围设为75~600Hz。

### 3.3.2 共振峰（Formant）

在特定的频率区域聚集大量声音能量的语音表现，称为共振峰。在语音的产生中，任何声道结构都有共振，因此就会出现共振峰。语音信号的共振峰频率反映了发生时声道的特点。

人耳听觉的频率范围为20Hz~20kHz，而人发声的频率范围是300Hz到3.4kHz。因此非语音的音频，它的频率范围可以达到人耳听觉的频率范围。不同的声音有不同的频率范围，其共振峰的分布也会不同。

本文中提取共振峰时，对音频的前四个共振峰频率进行估计，窗长设为10ms，截止频率上限设为8kHz。

### 3.3.3 Praat语音学软件介绍

Praat语音学软件全称为Praat: doing phonetics by computer，通常简称Praat，作者是荷兰阿姆斯特丹大学的Paul Boersma教授和David Weenink助教授。Praat是一款[跨平台](http://zh.wikipedia.org/wiki/%E8%B7%A8%E5%B9%B3%E5%8F%B0)的多功能[语音学](http://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E5%AD%A6)专业[软件](http://zh.wikipedia.org/wiki/%E8%BD%AF%E4%BB%B6)，主要用于对[数字化](http://zh.wikipedia.org/w/index.php?title=%E6%95%B0%E5%AD%97%E5%8C%96&action=edit&redlink=1)的[语音](http://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3)[信号](http://zh.wikipedia.org/wiki/%E4%BF%A1%E5%8F%B7)进行[分析](http://zh.wikipedia.org/w/index.php?title=%E8%AF%AD%E9%9F%B3%E5%88%86%E6%9E%90&action=edit&redlink=1)、标注、[处理](http://zh.wikipedia.org/wiki/%E6%95%B0%E5%AD%97%E4%BF%A1%E5%8F%B7%E5%A4%84%E7%90%86)及[合成](http://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E5%90%88%E6%88%90)等实验，同时生成各种[语图](http://zh.wikipedia.org/w/index.php?title=%E8%AF%AD%E5%9B%BE&action=edit&redlink=1)和文字报表。目前Praat已成为全世界实验语音学、心理语音学、自然语言处理等相关领域的研究人员普遍选用的专业软件。

Praat包含图形界面和命令行两种运行模式，适用于Windows操作系统的两个可执行文件分别为praat.exe和praatcon.exe，两者相互独立。本文仅使用其中的praatcon.exe，通过命令行方式从控制台调用，提取基频和共振峰两种韵律特征。

使用Praat软件提取基频和共振峰时，对每一个输入音频文件根据设定好的窗长，提取所有的基频和共振峰。然后对于每个包含75帧的Clip（时长约1.3s）计算统计值，包括平均值（Mean）、中位值（Median）、标准差（Standard Deviation）、最小值（Minimum）、最大值（Maximum）和变化范围（Range），相邻两个Clip有30帧的重叠（时长约0.5s）。输出文件中每个Clip对应一个统计值的特征向量，其中基频的特征向量为6维，共振峰的特征向量为24维（四个共振峰，各6维）。

## 3.4 多层音频特征提取

基于音频事件检测的语音分割，其特征选择和组合应能够区分各个不同类别的音频事件。在本文涉及的五类音频事件中，除了纯语音，都含有音乐。根据前面介绍的三大类特征的不同特点，设计了如图3.4所示的多层音频特征提取模式。

Speech with Music

Music

Ad

Song

Audio

Silence

Voice

With Music

Speech

Without Music

STE + STZCR

Pitch + Formant

MFCC / LPCC / PLP

Level 2 Classification

Level 1 Classification

VAD

图3.4 多层音频特征提取模式

对此多层音频特征提取模式有如下说明：

1) VAD算法。

对于输入音频信号，以Frame为最小分析单元，每个Frame包含1024个采样点。提取每个Frame的短时能量和短时过零率，使用VAD算法将音频分为静音和非静音。之后在静音段的端点附近使用分类器，进一步提高静音和非静音的切分准确度。

2) 第一层分类。

对于非静音部分音频，以Clip为最小分析单元，每个Clip包含75帧。使用Praat软件提取基频和共振峰，对每个Clip计算统计值，包括平均值、中位值、标准差、最小值、最大值和变化范围。因此得到6维的基频向量和24维共振峰向量（每一个共振峰得到6维），共30维特征向量。使用分类器将非静音部分音频分成With Music和Without Music两大类。

3) 第二层分类。

对于Without Music类，直接判断为Speech。对于With Music类，以Frame为最小分析单元，使用HTK工具包中的HCopy提取MFCC、LPCC或PLP特征，包括一阶差分和二阶差分，共39维特征向量。使用分类器将含音乐部分音频分成Music、Speech with Music、Advertisement和Song四类。

由于并非所有音频都可以提取到基频这一特征，在第一层分类结束后仍会有部分非静音部分并未分到With Music类或Without Music类。因此在第二层分类时，还需对这些第一层分类遗留下来的非静音部分进行分类，使用分类器直接将其分成五类，即本文所涉及到的全部五种音频事件类别。

为和多层分类结果进行对比，还将进行单层分类的实验。单层分类实际上是减少了第一层分类，即只使用频域特征，在将静音部分切分之后，对所有非静音部分使用分类器，将其分成五类。

以上即本文设计的多层分类算法中的特征提取部分。在第五章将会给出各组实验的结果，包括单层分类和多层分类的算法性能对比，以及使用不同频域特征时的算法性能对比。

# 第4章 分类器的理论基础与相关算法

分类器是音频分类的重点，而音频分类是音频事件检测的重要部分之一。分类器性能的好坏直接影响到音频事件检测性能的好坏。选择合适的分类器，将会有效提高分类的准确度，这对基于音频事件检测的语音分割结果有直接的影响。

本文选择的分类器均为基于统计学习算法得到的。本章将介绍三种分类器的基本算法或模型，包括多维高斯分布、高斯混合模型和支持向量机，并会介绍所使用到的现有工具。

## 4.1 多维高斯分布

### 4.1.1 基本概念

多维高斯分布实际上是多个随机变量的联合分布。设为一个维随机向量，其中每个分量服从正态分布，。记均值向量为，协方差矩阵为。则的维高斯密度函数如式4-1所示：

 （4‑1）

### 4.1.2 基于多维高斯分布的分类器

针对本文研究内容，将训练数据中五类音频事件的所有特征向量分别放入相应的集合中，作为五个模型的训练数据集，特征向量维数为。每个集合的向量个数为，。对第个训练数据集，按式4-2和式4-3计算其均值向量和协方差矩阵：

 （4‑2）

 （4‑3）

训练的结果是，每个模型对应两个参数，为均值向量和协方差矩阵。使用分类器识别时，将输入的特征向量带入到式4-1中，得到维高斯密度，然后按式4-4计算得到对数似然函数值最大的一类，即识别结果。

 （4‑4）

## 4.2 高斯混合模型（GMM）

### 4.2.1 基本概念

高斯混合模型（Gaussian Mixture Model，GMM）是由个多维高斯分布的加权叠加而成。它的一个基本理论前提是，只要高斯混合的数目足够多，一个任意的分布就可以在任意的精度下用这些高斯混合的加权叠加来逼近。

GMM的表示如式4-5所示，其中为GMM分量的个数，为维随机向量，为GMM分量，即一个维高斯密度函数，为对应分量的加权系数。满足式4-1，满足式4-6。

 （4‑5）

 （4‑6）

GMM的各个分量可由均值向量和协方差矩阵来描述，因此GMM可以由如式4-7所示参数集来表示。

 （4‑7）

### 4.2.2 参数估计

针对本文的研究，对每一个音频事件类别都用一个GMM来描述，设定混合数M=5。在给定了每类的训练数据集之后，就需要对模型参数进行估计。最常用的一种方法是最大似然（Maximum Likelihood，ML）估计，即在给定训练向量集的情况下，寻找合适的模型参数，使GMM的似然函数值最大。假设可用的训练向量集为，则GMM的似然函数如式4-8所示：

 （4‑8）

由于式4-8是模型参数的非线性函数，很难直接求出上式的最大值。因此常用EM（Expectation Maximization）算法迭代估计参数。通过LBG算法或K-均值算法可以得到参数的一个初始值。EM算法是从该初始值开始，每次迭代都用当前参数计算期望值并最大化期望值，按照式4-10至4-12给出的重估公式估计出一个新的参数，使得新模型参数下的似然函数值比原来增大。当似然函数值的增幅小于预先设定好的一个阈值，或者已经达到规定好的迭代次数时，参数估计结束。式4-9给出了分量的后验概率。

 （4‑9）

 （4‑10）

 （4‑11）

 （4‑12）

本文采用K-均值算法初始GMM的参数。首先选取前个特征向量作为初始聚类中心，按照最小距离准则对所有样本特征向量分类。计算各个新的聚类中心，若聚类中心变化不小于预先设定好的一个阈值，则根据新的聚类中心重新对所有样本分类，否则初始化结束，各聚类的均值和方差作为初始均值和方差，权值则是各聚类中所包含的特征向量个数占总特征向量个数的百分比。

### 4.2.3 识别过程

由于本文涉及到五类已知类别音频事件，因此五种类别分别对应GMM参数。给定待识别的音频特征向量集为，则为第类的后验概率如式4-13所示，为第类条件下的概率密度，为第类的先验概率，为对所有类的概率密度。

 （4‑13）

识别结果可以由最大后验概率准则给出，如式4-14所示，为识别结果。

 （4‑14）

一般情况下每类的先验概率是未知的，可设均为1/5，另外对于每一类，式4-13中的都相等，因此式4-14等价于式4-15。

 （4‑15）

为简化计算一般采用如式4-16所示的对数似然函数。

 （4‑16）

因此最终的识别准则如式4-17所示：

 （4‑17）

使用GMM的时候，混合数的选取比较重要。过小，则识别性能较差；过大，则需要有足够多的数据量，并且需要保证有较长的训练时间，否则模型无法精确建立，识别率反而会急剧降低，而且运算量明显增大。本文选取不同混合数，经过多次测试，最后选择了整体性能较好的一个值，。

## 4.3 支持向量机（SVM）

### 4.3.1 理论基础及训练方法

支持向量机（Support Vector Machine，SVM）是近年来在统计学习理论的基础上发展起来的一种模式分类方法，是一种解决二分类的分类器。它基于结构风险最小化的原则训练分类器，使得对于未见的样本也有很好的分类性能。

对于线性可分的情况，构造一个最优超平面，使得样本空间中两类样本的间隔最大，从而保证分类器在整个样本空间中的错误率上限最小，如图4.1(a)所示。对于线性不可分的情况，利用非线性映射，将低维空间变换到高维空间，使在低维的样本空间中线性不可分的样本在映射到高维空间后变成线性可分，如图4.1(b)所示。非线性映射通过核函数（Kernel Function）来实现。



图4.1 SVM的基本原理

目前常用的核函数主要有：

1) 多项式核函数，如式4-18所示。

 （4‑18）

2) 径向基核函数，如式4-19所示。

 （4‑19）

3) Sigmoid核函数，如式4-20所示。

 （4‑20）

训练SVM的核心思想是在线性约束条件下，求解一个二次规划问题，从而得到最优超平面。该超平面的决策方程如式4-21所示，其中为支持向量，为其类别标签，为支持向量个数，和为最优超平面的参数，为核函数。

 （4‑21）

### 4.3.2 LIBSVM软件包介绍

SVM算法所涉及的数学知识比较难，并且算法实现难度很大。国际上已经有很多科学研究者将他们的成果放在网络上免费提供给用于研究目的，方便大多数研究者。目前有关SVM计算的相关软件有很多，比如LIBSVM、mySVM、SVMLight等。本文使用的是LIBSVM软件包。

LIBSVM是台湾大学林智仁副教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的软件包，其中包含了已编译好的可在Windows操作系统上运行的可执行文件，以及源代码，方便修改或在其他操作系统上使用。此软件包提供了很多的默认参数，并且提供了交互检验的功能。

本文使用的是LIBSVM 2.88版本，其中包含的已编译好的适用于Windows操作系统的可执行文件包括：进行SVM训练的svm-train.exe，根据已获得的SVM模型对数据集进行预测的svm-predict.exe，对训练数据与测试数据进行简单缩放操作的svm-scale.exe，以及一个可视化操作工具svm-toy.exe。

本文针对韵律特征训练了一个SVM模型SVM\_1，针对三种频域特征分别训练了两个SVM模型SVM\_2和SVM\_3。一个典型的使用流程如图4.2所示。

svm-scale

svm-train

svm-predict

trainset.txt

svm-scale

trainset\_scale.txt

svm.model

testset.txt

testset\_scale.txt

scale.range

result.txt

图4.2 典型的LIBSVM使用流程

在训练一个SVM模型之前，首先要按照LIBSVM要求的格式准备训练数据集，并使用svm-scale.exe对训练数据进行简单的缩放操作。之后使用svm-train.exe基于缩放过后的训练数据训练SVM模型，得到模型文件，其中加入了参数-b 1允许使用概率估计值。在使用训练好的SVM分类器进行分类之前，同样要按照LIBSVM要求的格式准备测试数据集，并使用svm-scale.exe对测试数据进行简单的缩放操作，和缩放训练数据使用同种规则。之后使用svm-predict.exe基于训练好的模型对缩放过后的测试数据进行预测识别，输出结果中包含每一个测试样例的预测结果，以及对每类的概率估计值。此概率估计值用于对未知类别进行判断。

## 4.4 多层分类器

### 4.4.1 多层分类器结构

尽管SVM是一种解决二分类的分类器，但使用LIBSVM可以解决多分类问题。本文设计了多层分类的算法，因此需要多个分类器组合使用，可以对每层使用同种分类器，也可以不同种类的分类器交错组合使用。如图4.3所示为本文设计的多层分类器结构。

Speech with Music

Music

Ad

Song

Audio

Silence

Voice

With Music

Speech

Without Music

Level 2 Classification

Level 1 Classification

VAD

Voice

SVM\_1 / GMM

SVM\_2 / GMM

SVM\_3 / GMM

GMM

图4.3 多层分类器

对上述多层分类器结构说明如下：

1) 训练GMM和待分类类别无关，只需给定一类的训练样本集，就可在此基础上训练得到这一类的模型，在识别时使用相应的GMM即可。使用LIBSVM训练SVM分类器则和待分类类别有关，因此针对本文研究，训练三个SVM分类器。

2) 在基本VAD算法结束后，对静音段端点附近使用分类器，可以进一步提高静音和非静音的切分准确度。本文在这里使用Silence和Voice两个GMM。

3) 在第一层分类部分，使用的是韵律特征，可以使用GMM也可以使用SVM分类器。若使用GMM则直接使用相应类别的模型即可，但若使用SVM分类器必须使用特定的分类器，即SVM\_1。

4) 在第二层分类部分，使用的是频域特征，同样可以使用GMM也可以使用SVM分类器。在这一层分类，需要将With Music类分为相应的四小类，还需要将第一层分类结束后遗留下来的非静音部分分为相应的五小类。若使用GMM则直接使用相应类别的模型即可，但若使用SVM分类器则必须使用特定的分类器，分别为SVM\_2和SVM\_3。

为和多层分类结果进行对比，还将进行单层分类的实验，即减少了第一层分类，即只使用频域特征，在将静音部分切分之后，对所有非静音部分使用分类器，将其分成五类，可以使用GMM或SVM分类器。

以上即本文设计的多层分类算法中的分类器部分。在第五章将会给出各组实验的结果，包括单层分类和多层分类的算法性能对比，以及使用不同种分类器时的算法性能对比。

### 4.4.2 对未知类别的处理

音频事件类别多种多样，音频事件检测不可能对所有类别都有先验知识，总会有未知类别的音频出现。本章所介绍的三种分类器，其分类结果均为某个已知类别，但实际情况是有可能音频根本不属于五个已知类别的任一类。如果仅是简单将音频分类到五个已知类别，对于一些音频的分类结果不真实，准确率会降低。

对于这三种分类器，每个特征向量在每个类别中都会有一个值来表示它和这个类别的匹配程度，称之为分数（Score），分类器输出的分类结果为分数最高即匹配程度最高的那一类。对多维高斯分布和GMM而言，这个值为对数似然函数值；对SVM而言，这个值为概率估计值。假如某个输入特征向量并不属于任何已知类别，尽管根据分数大小的比较找到了匹配程度最高的那一类，但很有可能这个分数却并不高，该向量在各类中的分数差距不大。如果和所有类别的匹配程度都相差无几，分数都并不很高，则相当于它并不属于任何类别。

因此，本文分类器输出直接分类结果之前，先根据阈值对结果进行修正。设一个特征向量在各类别中的分数序列为，其中最大值为，以它作为上限，以作为下限，在这个分数区间内如果出现多于3个（包括3个）分数，则说明各类的分数差别不大，判断这个输入样例为未知（Unknown）类别。阈值设定的不同，分类结果会不同，式4-22和式4-23定义了错误接收率（False Accept Rate）和错误拒绝率（False Reject Rate）。

 （4‑22）

 （4‑23）

通过调整阈值，找到使和相等或最接近的阈值。这样基本将不属于已知类别的特征向量统一归到未知类别。在第五章将会比较不同阈值对于算法性能的影响。

# 第5章 实验设计与结果分析

## 5.1 性能评价指标

基于音频事件检测的分段结果中包含两种音频事件片段：已知事件类别的音频段和未知类别的音频段。

对于已知事件类别的音频段，基于时间计算正确分类的准确率，如式5-1所示。每一个类别会得到一个准确率，同时对于所有类别还有一个总准确率。因此对于每一组实验会得到六个准确率值。

（5-1）



对于未知类别的音频事件，实际上是对分类器的输出结果，根据某个规则将不属于任何已知类别的音频事件归为未知类别。在第四章中已经介绍过，通过调整阈值，找到一个最佳阈值，使得式4-22给出的错误接收率和式4-23给出的错误拒绝率相等或最接近，然后使用这个阈值对分类器的直接输出结果进行修正。因此在每一组实验中会动态调整阈值，得到错误接收率和错误拒绝率的曲线图，并很直观地得到期待阈值。

## 5.2 实验结果

本文选择了多种音频特征和多种分类器，用以研究和实现基于音频事件检测的语音分割分层算法，因此本文设计了如表5.1所示的分组实验方式。所有音频数据统一为wav格式，44.1kHz采样，16位单声道。表5.2列出了数据集的设置。

表5.3至表5.7分别给出了五类已知类别的准确率，依次为Speech、Music、Speech with Music、Advertisement和Song。表5.8给出了总准确率。在每一个表中，左上部分给出了在单层分类情况下使用不同音频特征和不同分类器得到的结果；右下部分给出了在两层分类情况下使用不同音频特征和不同分类器组合得到的结果。

表5.1 分层分类算法的实验设置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 单层分类 |  | 多维高斯分布 | GMM | SVM |
| MFCC | √ | √ | √ |
| LPCC | √ | √ | √ |
| PLP | √ | √ | √ |
| 两层分类 |  | GMM+GMM | SVM+GMM | SVM+SVM |
| Pitch+Formant+MFCC | √ | √ | √ |
| Pitch+Formant+LPCC | √ | √ | √ |
| Pitch+Formant+PLP | √ | √ | √ |

表5.2 实验数据设置

|  |  |  |  |
| --- | --- | --- | --- |
|  | 训练数据集 (s) | 测试数据集 (s) | 总计 (s) |
| Speech | 607.0 | 955.3 | 1562.3 |
| Music | 956.9 | 400.4 | 1357.3 |
| Speech with Music | 626.9 | 1056.7 | 1683.6 |
| Advertisement | 603.0 | 417.4 | 1020.4 |
| Song | 612.3 | 1241.6 | 1853.9 |
| 总计 | 3406.1 | 4071.4 | 7477.5 |

表5.3 已知类别Speech的分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 82.7% | 75.8% | 89.9% | / | / | / |
| LPCC | 85.8% | 90.6% | 87.2% | / | / | / |
| PLP | 64.2% | 63.0% | 66.3% | / | / | / |
| P+F+MFCC | / | / | / | 78.5% | 85.8% | 85.8% |
| P+F+LPCC | / | / | / | 79.8% | 86.7% | 86.4% |
| P+F+PLP | / | / | / | 77.6% | 85.4% | 84.8% |

表5.4 已知类别Music的分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 25.7% | 53.0% | 43.8% | / | / | / |
| LPCC | 24.7% | 82.9% | 59.1% | / | / | / |
| PLP | 19.6% | 27.1% | 33.3% | / | / | / |
| P+F+MFCC | / | / | / | 55.5% | 55.4% | 45.3% |
| P+F+LPCC | / | / | / | 83.4% | 84.2% | 57.8% |
| P+F+PLP | / | / | / | 28.0% | 35.6% | 34.8% |

表5.5 已知类别Speech with Music的分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 76.5% | 76.8% | 64.8% | / | / | / |
| LPCC | 75.0% | 75.4% | 58.8% | / | / | / |
| PLP | 77.4% | 78.7% | 65.7% | / | / | / |
| P+F+MFCC | / | / | / | 77.1% | 77.5% | 58.6% |
| P+F+LPCC | / | / | / | 71.1% | 74.3% | 55.4% |
| P+F+PLP | / | / | / | 78.9% | 78.7% | 48.9% |

表5.6 已知类别Advertisement的分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 51.1% | 24.9% | 46.7% | / | / | / |
| LPCC | 38.4% | 17.7% | 33.6% | / | / | / |
| PLP | 49.1% | 35.7% | 43.3% | / | / | / |
| P+F+MFCC | / | / | / | 21.8% | 23.2% | 45.4% |
| P+F+LPCC | / | / | / | 15.6% | 16.4% | 31.4% |
| P+F+PLP | / | / | / | 20.7% | 28.3% | 42.0% |

表5.7 已知类别Song的分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 66.3% | 72.3% | 76.3% | / | / | / |
| LPCC | 85.5% | 75.1% | 83.7% | / | / | / |
| PLP | 53.7% | 72.9% | 74.7% | / | / | / |
| P+F+MFCC | / | / | / | 72.4% | 72.6% | 76.8% |
| P+F+LPCC | / | / | / | 75.3% | 75.1% | 83.6% |
| P+F+PLP | / | / | / | 72.3% | 75.0% | 74.6% |

表5.8 总分类结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 多维高  斯分布 | GMM | SVM | GMM+  GMM | SVM+  GMM | SVM+  SVM |
| MFCC | 60.4% | 57.8% | 66.7% | / | / | / |
| LPCC | 54.4% | 45.8% | 61.1% | / | / | / |
| PLP | 52.5% | 58.2% | 58.6% | / | / | / |
| P+F+MFCC | / | / | / | 53.0% | 55.7% | 63.3% |
| P+F+LPCC | / | / | / | 39.5% | 42.5% | 58.1% |
| P+F+PLP | / | / | / | 50.1% | 59.0% | 55.3% |

## 5.3 结论

表5.3至表5.8给出了所有的实验结果。由于准确率的定义如式5-1所示，分子为正确分类的音频事件段的时间总和，分母为分类结果中已知类别音频时间段的时间总和，实验结果中如果某一个数值突然变得很大，除了实验本身的原因以外，还有可能是由于计算这个准确率时分母并不大，即该类音频事件被检测出的并不多，此数值会有一定的误导性。因此，在对实验结果进行分析时，应该从整体趋势来进行分析，而不是针对单个准确率之间的绝对差距。

### 5.3.1 不同音频特征对性能的影响

不同音频特征对性能的影响趋势是相似的，和分类器无关，因此不妨所有分类器都选择SVM。图5.1展示了单层分类的情况下不同音频特征对性能的影响。图5.2展示了基于同一种音频特征进行单层分类和两层分类时对性能影响的比较。

图5.1 单层分类时不同音频特征对性能的影响

图5.2 单层分类和两层分类时不同音频特征组合对性能的影响

从图5.1所显示的内容来看，并不能明显看出MFCC、LPCC和PLP这三类特征对于算法性能的影响。实际上这三类特征彼此有相关性，且各自有各自的优点，它们都有各自的不太适用或比较适用的音频事件类别。因此如果可以根据它们各自的特点选择最合适于待检测类别的特征，算法性能会有所提高。

从图5.2所显示的内容来看，当使用韵律特征进行第一层分类、使用MFCC进行第二层分类时，算法性能在整体趋势上比只用MFCC进行单层分类的性能要有所下降。可能的原因是，在第二层分类时只是针对相应的大类别分类到相应的小类别，假如在进行第一层分类时已经存在误分类的现象，累积到第二层时准确率会更低。不过从图上来看，有两类的分类准确率稍有提高，分别为Music和Song，从这个角度来看，基频和共振峰这些韵律特征对于区分含音乐和不含音乐是有一定作用的。

### 5.3.2 不同分类器对性能的影响

选择音频特征时，不论使用MFCC、LPCC或PLP中任何一种，不同分类器对性能的影响趋势是类似的，不妨选择MFCC特征。图5.3展示了单层分类时不同分类器对性能的影响。图5.4展示了两层分类时不同分类器组合对性能的影响。

图5.3 单层分类时不同分类器对性能的影响

图5.4 两层分类时不同分类器组合对性能的影响

从图5.3和图5.4所显示的内容上来看，在整体趋势上使用SVM分类器时算法性能稍好一点。

从图5.3可以看到，对于Music类别，GMM明显比多维高斯分布性能好，因为音乐是种复杂的音频，用多个多维高斯分布的加权叠加来逼近这种复杂的分布比仅用一个多维高斯分布效果要好，多维高斯分布实际上相当于混合数M=1的高斯混合模型。

从图5.4可以看到，第二层分类使用GMM时，第一层分类使用GMM或者SVM算法性能差距不太大。可能的原因在于，第一层只是基于韵律特征将非静音分成两大类，使用何种分类器效果不会差很多。同时，在本文的研究中，第一层分类时规定如果一个样例对于两大类的分数之间差距小于1/3的话则不在这层将它分类，将它作为第一层分类遗留下来的非静音部分在第二层分类。所以有了这样一个判断，在第一层使用两种分类器分类的结果更加不会相差很多。

### 5.3.3 判断未知类别的阈值选择

以上分组实验的设计中包括不同音频特征和分类器的选择和组合，在这里选择音频特征MFCC和SVM分类器，基于两层分类进行研究。通过调整阈值，找到一个最佳阈值，使得错误接收率（FAR）和错误拒绝率（FRR）相等或最接近。如图5.5所示为不同阈值下的FAR曲线和FRR曲线。

图5.5 不同阈值下的错误接收率和错误拒绝率

从图中可以看出，错误接收率（FAR）随阈值增大呈下降趋势，而错误拒绝率（FRR）随阈值增大呈上升趋势。这两个错误率，任何一个过大对算法性能都有很大的影响，因此所寻找的阈值就是使得FAR和FRR相等或相差最小的阈值，如图中黑色虚线所示位置，此时阈值，，。

# 第6章 总结与展望

## 6.1 本文内容总结

音频检索在多媒体信息技术中占有重要的地位，具有广泛的实用价值。根据人们对音频信息的需求，音频事件检测在近几年发展起来，其重要性也逐渐显现。基于音频事件检测的语音分割对于其他多媒体信息技术都有很重要的意义。对于音频检索和识别系统，它可以将目标范围缩小到某个分割后的片段中，能很大程度上提高系统性能。对于视频信号的自动切分和分类，它则是非常有效的辅助工具。本文的主要工作包括以下几个方面：

1) 本文选定了几个已知的音频事件类别，分别为Speech，Music，Speech with Music，Advertisement和Song这五种，对于其他不属于这五类的音频事件，统一归为未知类别。本文根据音频分析的层次化结构，对不同时间粒度的结构单元使用不同的分析方法。根据模式识别理论，本质上音频分类是一个模式识别的过程。本文基于音频分析的层次化结构，根据模式识别的思想，设计了一个基于音频事件检测的语音分割的分层算法，对应到音频事件类别的不同层次划分。

2) 音频特征的提取和分析是音频分类的基础，而音频分类是音频事件检测的重要部分之一。本文重点讨论了音频特征的选择和提取方法，并考察了不同音频特征对算法性能的影响。本文选择的音频特征有：时域特征中的短时能量和短时过零率，频域特征中的MFCC、LPCC和PLP，以及韵律特征中的基音频率和共振峰。对应于本文所设计的分层算法，使用了基于音频Frame和音频Clip的音频特征提取和分析方法。

3) 音频分类是音频事件检测的重要部分之一，而分类器是音频分类的重点。分类器性能的好坏直接影响到音频事件检测性能的好坏。选择合适的分类器，将会有效提高分类的准确度，这对基于音频事件检测的语音分割结果有直接的影响。分类器的设计也是本文重点内容之一。本文选择了三种基于统计学习算法得到的分类器：多维高斯分布、高斯混合模型和支持向量机，并考察了不同分类器对算法性能的影响。同时本文还考虑了基于阈值对未知类别的判断。

## 6.2 今后的工作与展望

基于音频分析的层次化结构，根据模式识别的思想，本文选择不同种类的音频特征和分类器，设计了一个基于音频事件检测的语音分割的分层算法，对应到音频事件类别的不同层次划分，对于某些类别性能较好。然而随着音频媒体的数据量越来越庞大，人们对于音频信息的要求也越来越高、越来越多。音频检索技术仍需要进一步发展，且刻不容缓。我认为以下几点是未来基于音频事件检测的语音分割算法的研究重点：

1) 音频事件类别种类繁多且内容富于变化，这从客观上加大了特征分析和抽取的难度。然而本文中所涉及到的音频类别仅包含五种，对于特征的选择也很有限。从本文给的实验结果来看，虽然不同音频特征对算法性能的影响看上去并不太明显，却也能看到一些关于音频特征对性能影响的整体趋势。但是一旦音频事件类别的种类增多，对于特征的选择必须进一步考虑和选择。如何选取更有效的或者新的能够显著区分音频事件类别的音频特征，这将是未来工作的一个重点。

2) 本文中所涉及到的音频类别仅包含五种，并且其中几类为比较宏观的分类，比如Speech，Music等。这种分类宏观且不够全面，实际的音频流中包含比这五类复杂得多的音频类别。在本研究中，对于不属于这五类的音频事件统一归为了未知类别。实际上，更细致的音频事件类别的划分，比如风声、鼓掌声、脚步声等，对于音频处理与人们的需求更有帮助。因此，对音频事件分类的细化和对更多特定类别的音频事件的检测是未来工作的另一个重点。

插图索引

图2.1 音频分析的层次化结构 6

图2.2 模式识别的一般性过程 7

图2.3 分类器的设计流程 7

图2.4 基于音频事件检测的语音分割的分层算法技术流程图 9

图2.5 音频事件类别的层次划分 10

图2.6 基本VAD算法示意图 11

图3.1 一段语音的过零率图 15

图3.2 MFCC的提取流程 16

图3.3 PLP的提取流程 18

图3.4 多层音频特征提取模式 22

图4.1 SVM的基本原理 28

图4.2 典型的LIBSVM使用流程 30

图4.3 多层分类器 31

图5.1 单层分类时不同音频特征对性能的影响 37

图5.2 单层分类和两层分类时不同音频特征组合对性能的影响 37

图5.3 单层分类时不同分类器对性能的影响 38

图5.4 两层分类时不同分类器组合对性能的影响 39

图5.5 不同阈值下的错误接收率和错误拒绝率 40

表格索引

表2.1 单层分类算法的实验设置 13

表2.2 两层分类算法的实验设置 13

表5.1 分层分类算法的实验设置 34

表5.2 实验数据设置 34

表5.3 已知类别Speech的分类结果 34

表5.4 已知类别Music的分类结果 35

表5.5 已知类别Speech with Music的分类结果 35

表5.6 已知类别Advertisement的分类结果 35

表5.7 已知类别Song的分类结果 36

表5.8 总分类结果 36

参考文献

1. 王清亮, 常青, 薛向阳. 音频信息检索综述[J]. 计算机科学, 2004, 31(61): 59-63.
2. 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, 21(11): 1173-1176.
3. 白亮. 音频分类与分割技术研究[D]. 国防科学技术大学, 2004.
4. Xiaodan Zhuang, Xi Zhou, Thomas S. Huang and Mark Hasegawa-Johnson. Feature Analysis and Selection for Acoustic Event Detection [C]. ICASSP, 2008: 17-20.
5. 季春. 音频信息检索技术的发展及应用[J]. 现代情报, 2007, (1): 157-160.
6. Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based Classification, Search, and Retrieval of Audio [J]. IEEE Multimedia, Fall, 1996: 27-36.
7. Zhu Liu, Jincheng Huang, Yao Wang, Tsuhan Chen. Audio Feature Extraction and Analysis for Scene Classification. IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing.
8. Jonathan Foote. An Overview of Audio Information Retrieval [J]. Multimedia Systems, 1999, (7): 2-10.
9. Andreas Rauber, Elias Pampalk, Dieter Merkl. Using Psycho-Acoustic Models and Self-Organizing Maps to Create a Hierarchical Structuring of Music by Sound Similarity. IRCAM, 2002.
10. Eric Scheirer, Malcolm Slaney. Construction and Evaluation of a Robust Multifeature Music/Speech Discriminatior. 1997.
11. Lie Lu, Hao Jiang and HongJiang Zhang. A Robust Audio Classification and Segmentation Method. 2001.
12. Lie Lu, Stan Z. Li and Hong-Jiang Zhang. Content-based Audio Segmentation Using Support Vector Machines. 2001.
13. 卢坚, 陈毅松, 孙正兴, 张福炎. 语音/音乐自动分类中的特征分析[J]. 计算机辅助设计与图形学学报, 2002, 14(3): 233-237.
14. Lie Lu, Hong-Jiang Zhang, and Hao Jiang. Content analysis for audio classification and segmentation. IEEE Trans. Speech and Audio Processing, 2002, 10: 504-516.
15. Julien Pinquier, Jean-Luc Rouas and Régine André-Obrecht. Robust speech / music classification in audio document. ICSLP02, 2002, pp. III: 2005-2008.
16. Daniel P. W. Ellis. Prediction-driven computational auditory scene analysis. Ph.D. thesis, MIT, 1996.
17. Rui Cui, Lie Lu, Hong-Jiang Zhang, and Lian-Hong Cai. Highlight sound effects detection in audio stream. ICME03, 2003, pp. III: 37-40.
18. Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. IV Jornadas en Tecnologia del Habla, Zaragoza, Spain, November, 2006.
19. 蔡莲红, 黄德智, 蔡锐. 现代语音技术基础与应用[M]. 北京: 清华大学出版社, 2003. 24-28.
20. 林帆, 徐明星. 一种改进的基于时域参数的语音切分算法[J]. 计算机科学, 2006, 33(4): 164-167.
21. 王金明, 张雄伟. 话者识别系统中语音特征参数的研究与仿真[J]. 系统仿真学报, 2003, 15(9): 1276-1278.
22. 于明, 袁玉倩, 董浩, 王哲. 一种基于MFCC和LPCC的文本相关说话人识别方法[J]. 计算机应用, 2006, 26(4): 883-885.
23. 许飞. 语音信号特征提取技术简介.
24. 李燕萍, 唐振民, 钱博, 张燕. 基于PLAR特征补偿的鲁棒性说话人识别仿真研究[J]. 系统仿真学报, 2009, 21(2): 409-412.
25. 扈浩. 汉语语音的情感识别研究[D]. 清华大学, 2007.
26. 石现峰, 张学智, 张峰. 基于HTK的语音识别系统设计[J]. 计算机技术与发展, 2006, 16(10): 37-41.
27. 仲晓波, 杨玉芳. 国外关于韵律特征和重音的一些研究[J]. 心理学报, 1999, 31(4): 468-475.
28. 熊子瑜. Praat语音软件使用手册.
29. 王书诏. 基于高斯混合模型的说话人识别系统的研究[D]. 大连理工大学, 2006.
30. John Makhoul, Francis Kubala, Timothy Leek, Daben Liu, Long Nguyen, Richard Schwartz and Amit Srivastava. Speech and Language Technologies for Audio Indexing and Retrieval [J]. Proceeding of the IEEE, 2000, 88(8): 1338-1351.

致 谢

衷心感谢徐明星副教授一学期以来对我毕业设计的精心指导。徐老师严谨细心，对我的研究工作给予了很多的指导和帮助。

感谢徐路师兄和王刚师兄，他们对我的研究工作给予了耐心的帮助与指点，给我的研究工作带来了灵感。感谢和我同组做毕业设计的张锆和张斯驰，他们与我进行了很多次有益的讨论，对我的研究工作有很大启发。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日 期：\_\_\_\_\_\_\_\_\_\_\_\_\_

附录A 书面翻译

听觉事件检测的特征分析与选取

**摘要**

语音感知特征，如Mel频率倒谱系数（MFCC），已经被广泛应用到听觉事件检测当中。然而，语音和听觉事件之间频谱结构不同，使得语音特征集的性能有所降低。我们建议根据近似贝叶斯精度量化每个特征成分的辨别能力，为听觉事件检测生成一个有辨别能力的特征集。和MFCC相比，使用前面提到的方法生成的特征集，在听觉事件检测中可相对提高30%的精确度。

**关键词**：听觉事件检测，特征选取，贝叶斯精度，隐马尔可夫模型

**1. 概述**

听觉事件检测（Acoustic Event Detection, AED）是音频场景分析的一个分支[1, 2, 3, 4, 5, 6]，有非常广泛的应用。特别地，非语音，即（非语音）听觉事件的信息，揭示了人类和社会的活动。比如，会议刚开始时椅子挪动或门的声音[4]，体育赛事中观众的欢呼声[7]，街上的一声枪响[8]，以及疗养院内匆忙的脚步声。这些信息对于监控、多媒体信息检索和智能会议室等应用都非常有用。有些事件是比较一致且突出的，比如欢呼声，而另一些则是细微的，比如在一个铺有地毯的会议室中的脚步声、笔记本电脑的键盘敲击声和包装纸的声音。

先前发表的文章侧重于将音频分割成少数几类[2, 3]、分离几个音频源[1, 9]、和检测一些突出的听觉事件[5]。CHIL和美国国家标准与技术研究院（NIST）在2006年[4]和2007年的CLEAR上举行了对听觉事件检测的评估，尝试在一个真实的研讨会环境下识别12个听觉事件的时间边界和标记。这些听觉事件中很多都是细微的（低信噪比，如脚步声、包装纸声、键盘敲击声），或/且和语音重叠，使得任务特别有挑战性。虽然已经探究了不同的系统结构和特征集[4]，但即使是顶端的听觉事件检测系统，它的性能也相当低[10]。

一个合适的特征集在听觉事件检测中扮演了非常重要的角色。在不同的分析任务中已经提出了各种音频感知特征[1, 11, 5]。在最近的CLEAR对听觉事件检测的评估中，最受欢迎的特征是复杂的语音感知特征集[4, 6]，比如Mel频率倒谱系数（MFCC）和转换频率滤波器参数，后者已被证明能够很好地表现语音频谱结构。然而，这些参数并不一定适合听觉事件检测，理由如下：1) 对于研究听觉事件的频谱结构，有限的工作已经完成，根据语音频谱结构设计的语音特征可能对听觉事件检测来说远达不到最优。2) 信噪比（Signal-to-Noise Ratio, SNR）对听觉事件检测而言很低，尤其是当重叠的语音被看做噪声的时候。所以，对听觉事件检测而言，分析听觉事件的频谱结构和设计合适的特征集是非常重要的。

我们建议根据近似贝叶斯精度量化每个特征成分的辨别能力，为听觉事件检测生成一个有辨别能力的特征集。首先通过主成分分析（Principal Component Analysis, PCA）对在特征库中的所有特征去除相关性。然后对每一个听觉事件，我们对已经被去除相关性的所有特征成分采用非参数分布估计。通过这些估计的分布，我们可以将事件的分类近似到贝叶斯精度，可以量化这些已经去除相关性的特征成分的辨别能力，并指导我们的特征选取。我们证明了前面提出的这个特征分析和选取的框架，可以从传统的为语音识别（原始任务）设计的语音特征库中，很方便地生成为听觉事件检测（新任务）使用的特征集。基于HMM的使用前面得到的特征集的听觉事件检测系统，和使用MFCC和相同数量的参数的基准系统相比，效果要好得多。

**2. 听觉事件检测中的语音感知特征**

过去的几十年间，有很多关于语音感知特征的研究已经完成[12, 13]。这些特征主要是基于语音产生和感知的性质而设计的。谱图的包络（共振结构）被认为拥有了最多的语音信息，而不是谱图的精细结构（谐波结构）。转换频率滤波器参数和MFCC[12]都使用三角形带通滤波器绕过谱图的精细结构。此外，这些语音特征集选定非均匀临界频段，模拟非均匀的从人类听觉感知中观察到的频率辨析率，在低频部分提供高辨析率。

语音感知特征已经被广泛使用在音频分析中[4, 6]。然而，听觉事件和语音的频谱结构是不同的，如图1所示。

图1 听觉事件“键盘声”、“脚步声”和人声的谱图（图见原文）

所以，根据语音的频谱结构所设计的语音特征集，对听觉事件检测而言远达不到最优，对在听觉事件检测中准确使用语音特征集的有效性提出了质疑。比如，他们可能忽略了某些频率部分，其中包含了语音辨别能力较低的信息，但可能对于听觉事件而言却包含了很多有辨别能力的信息。另一方面，他们的目的是特别强调声学和语音音素是不同的。

**3. 辨别能力的度量**

我们建议根据近似贝叶斯精度量化每个特征成分的辨别能力。很直观地，特征成分的辨别能力越高，会有越高的贝叶斯精度。这将帮助我们理解在听觉事件检测任务中使用的语音特征集中的显著的特征成分，并为听觉事件检测设计合适的特征集。

在多类别的情况下，贝叶斯精度定义如下：





 (1)

其中，是第i类的先验概率，是第i类中观测值x的似然度。要注意，在特征空间内定义了一个特殊的区域，在这个区域中第i类有最高的似然度：

(2)



所以我们可以在数据集上近似贝叶斯精度，如下：

 (3)

其中，表示第t个实例的真实性标记，是单位脉冲δ函数。

在对每一个特征成分的分布都没有先验知识的情况下，要计算一个特征成分的贝叶斯错误率，我们采用非参数密度估计。parzen窗密度估计是一种非参数密度估计的技术[14]。给定一个内核函数，给定训练集的分布可以被一个内核的线性组合所近似，这些内核以观测数据点为中心。在这个研究中，我们使用parzen窗和高斯内核函数，来对每一个事件估计每一个特征成分的分布。

图2给出了特征库中52个不同特征成分的贝叶斯精度，这个特征库由26维的转换频率滤波器参数和26维的MFCC组成。

图2 不同特征成分的贝叶斯精度（图见原文）

**4. 特征集的生成**

我们建议根据特征成分在训练数据集上的贝叶斯精度，来从特征库中选择特征成分。为了减少不同特征成分之间的相关性，我们首先对特征库使用主成分分析。然后提出两种基于贝叶斯精度的方法，来量化已去除相关性的特征成分的辨别能力。

第一种方法采用公式3中定义的贝叶斯精度的近似作为目标函数。第二种方法采用所有数据点的真实性标记的似然度的负和作为目标函数，如公式4所定义。我们称这两种方法分别为Hard\_Bayesian和Soft\_Bayesian。

 (4)

已去除相关性的特征成分中，在目标函数中得到值较高的特征成分组成了生成的特征集。

我们总结了为听觉事件检测生成一个特征集的过程，如图3所示。

图3 特征分析与选择框架（图见原文）

**5. 实验**

5.1 基于隐马尔可夫模型的听觉事件检测系统

为了听觉事件的检测与分类，我们实现了一个基于隐马尔可夫模型（HMM）的系统，其中每一个听觉事件被一个隐马尔可夫模型所模拟，这个模型拥有三个发射状态和从左向右的状态转移。状态的观测分布是逐步训练的有五个混合状态的高斯混合模型。更多的关于我们的CLEAR对基于隐马尔可夫模型的听觉事件检测系统的评估的细节上的描述，参见文献[10]。

5.2 数据集与度量标准

我们的听觉事件检测实验使用了2007年的CLEAR对听觉事件检测的评估的官方数据[15]：大约3小时用于系统开发，2小时用于系统评估。所有的数据都是研讨会风格，包括语音、听觉事件、以及部分两者的重叠。许多事件是细微的，且与背景噪音或语音相比有低信噪比。性能使用AED-ACC进行测试，其中AED-ACC被定义为系统输出听觉事件的标记和参照听觉事件的标记的F函数值（精度和召回的调和平均值）。AED-ACC的目标是评价对所有听觉事件实例的检测和分类，面向一些应用，比如智能房间的实时服务和基于音频的监控。

5.3 实验设置

这些实验比较了两种单向的基于隐马尔可夫模型的听觉事件检测系统的性能，它们分别使用了生成的听觉事件检测特征和MFCC基线集。这些听觉事件检测特征集是使用第4节中的方法，从一个常规的语音感知特征（即MFCC和转换频率滤波器参数）库中生成的。所有特征集有相同的成分数（78），并且所有系统有相同的参数个数。

MFCC基线集在语音识别和其他音频应用中被广泛使用。我们使用在0Hz~11000Hz带宽上计算得到的26维MFCC以及它们的一阶回归（delta）系数和二阶回归（加速度）系数（称之为MFCC26DAZ）。

前两个生成的特征集（DERIVE26DAZ\_hard和DERIVE26DAZ\_soft），均由26个成分组成，这些成分是通过使用在第4节中提到的Hard\_Bayesian方法或Soft\_Bayesian方法、从一个包含26维转换频率滤波器参数和26维MFCC的特征库中生成。这些生成的特征成分的一阶回归系数和二阶回归系数也包含在其中。后两个生成的特征集（DERIVE78\_hard和DERIVE78\_soft），均由78个特征成分组成，这些成分是从一个包含26维转换频率滤波器参数、它们的一阶回归系数和二阶回归系数以及所有的78个MFCC26DAZ成分的特征库中生成，使用Hard\_Bayesian方法或Soft\_Bayesian方法。

5.4 实验结果

当我们训练系统时，我们将三小时的数据保留其中的1/3作为Dev集，用以调整一些系统参数。如图4所示，在Dev集和Test集上，使用任何一个生成的特征集的系统，它的性能都比基线系统要好，有30%左右的相对的AED-ACC性能提升。

图4 使用基线系统集MFCC26DAZ和四个生成的特征集的AED-ACC分数（图见原文）

我们同样比较了当特征集被使用在我们的单向的基于隐马尔可夫模型的听觉事件检测的CLEAR评估系统中时的性能[10]，其中使用了所有的数据用于系统的训练。如图5所示，所有的生成的听觉事件检测特征集都比基线好。特别的，DERIVE78\_soft获得了超过30%的相对的AED-ACC性能提升。

图5 使用基线系统集MFCC26DAZ和四个生成的特征集的AED-ACC分数（图见原文）

以上结果说明了，对于听觉事件检测这个任务，广泛使用的完整的MFCC集远远达不到最优，而通过前面提到的方法生成的特征集能够在没有参数增加的听觉事件检测中获得更好的性能。

**6. 结论与讨论**

在这篇文章中，我们建议根据近似贝叶斯精度量化每个特征成分的辨别能力，为听觉事件检测生成一个有辨别能力的特征集。我们证明了我们的方法在CLEAR听觉事件检测评估任务上的有效性。其中被提议的特征分析和选择框架可以方便地为一个新任务（即听觉事件检测）从一个常规的特征库中生成特征集，这个特征库为一个更常规的任务（即语音识别）而被驱动。

音频索引和检索中的语音和语言技术

（注：本文由我和张锆、张斯驰合翻译完成，我翻译的部分为原文1-5页）

**特约论文**

随着本质上无限制的数据存储能力的出现，以及互联网使用的增长，我们可以合理地去想象一个仅敲击几下键盘或者通过声音命令就能够随意获得任何已存储的信息的世界。因为这些数据中大部分会是从各种来源得到的语音的形式，所以发展浏览这些音频数据和为它们建立索引的必要的技术变得很重要。这篇文章将记述一些必须的语音和语言技术，并将介绍一个将这些技术集成为一个系统的叫做“Rough’n’Ready”的成果，它为音频数据建立索引，生成一个结构上的摘要，并提供浏览这些存储数据的浏览工具。在这篇文章中突出的技术包括说话人无关的连续语音识别，说话人分段和识别，名字定位，话题分类，故事分段和信息检索。这个系统会自动地将连续音频输入流按照说话人分段，将同一说话人的音频片段聚类，识别系统已知的说话人，并且转录所说的单词。它还会将输入流根据话题内容分段成故事，然后定位人名、地点和组织。这些结构上的特征将存储在一个数据库中，用来构造高度可选择的搜索查询，以从大量音频存档中对特殊内容进行检索。

***关键词***——音频浏览，音频索引，信息提取，信息检索，命名实体提取，名字定位，说话人辩护检测，说话人聚类，说话人识别，语音识别，故事分段，话题分类。

**1. 概述**

国家科学基金会信息与智能系统部主任M. Lesk在一篇关于世界上有多少信息的文章中总结道：“在短短几年内，我们将能够存储一切——再也不会有信息被丢掉——而典型的信息块将不再为人类所看见。”[1] 大部分这些信息将会是语音的形式，它们从各种来源得到：电视，广播，电话，会议，报告等等。然而，由于在大量音频存档中定位信息很困难，语音并没有作为一种档案来源来被评价。但是，在语音和语言技术在经过十年或更长时间的稳定进步后，现在已经有可能开始建立基于内容的自动索引和检索工具，这将会使语音记录作为一种档案来源和文本一样有价值。

这篇文章记述了一些在开发强有力的音频索引系统中需要使用到的语音和语言处理技术。一个合并了这些技术的原型系统已经被建立，用于广播新闻索引和检索。这个被称为Rough’n’Ready的系统提供了粗略的语音转录，为浏览做好准备。这些合并在这个系统中并在这篇文章中记述到的技术，包括了说话人无关的连续语音识别，说话人分段，说话人聚类，说话人识别，名字定位，话题分类，故事分段和信息（或故事）检索。这些不同技术的综合使得Rough’n’Ready能够生成一个口语的高水准的结构上的摘要，便于浏览数据。

在这篇文章中报告的系统和方法和几种其他不同的多媒体索引系统在今天的发展下有关联。卡内基梅隆大学（CMU）的信息媒体系统[2]-[4] 和MITRE企业的广播新闻导航系统[5], [6]，两者都能对于广播新闻记录自动转录音频信号和调整音频信号的时间，以在转录结果中定位特定的名字，并且运用信息检索技术找到音频。然而，这两个系统的焦点都在视频流的特征上。这些系统证明了从视频中得到的信号对于定位新闻故事之间的边界非常有效。它们还广泛使用了现在美国大部分电视新闻节目附加的封闭字幕文本。

CMU正在开发另一个为视频中的会议做索引和浏览的多媒体系统[7]。这个领域中并没有封闭字幕可使用，所以对自动转录有更强的依赖。但是视频还被利用来检测说话人的变化，并且解释姿态，比如凝视的方向和头/手的运动。

相比而言，Rough’n’Ready系统完全将焦点放在音频信号中所包含的语言上的内容，并且因此由语音信号中获得所有的信息。这是一个特意的选择，旨在让所有开发成果通向有效的提取和摘要，以及音频信息的显示。这给了Rough’n’Ready一个特殊的能力，如果语音时是唯一的知识来源的话。我们的系统的另一个显著的特征是，所有被使用到的语音和语言技术共享了一个公有的统计建模模式，它促进了各种知识来源的综合。

第二节介绍了Rough’n’Ready系统，展示了它的一些索引和浏览的能力。剩余的几节将焦点放在了在这个系统中被使用到的单独的语音和语言技术。第三节介绍了在各种技术中广泛使用的基本统计建模模式。第四节记述了被使用到的语音识别技术，第五节详述了三类说话人识别技术：说话人分段，说话人聚类和说话人识别。之后几节介绍的技术均将语音识别组件生成的文本作为它们的输入。第六至九节依次介绍了以下技术：名字定位，话题分类，故事分段和信息检索。

**2. 用Rough’n’Ready建立索引和浏览**

*A. Rough’n’Ready系统*

Rough’n’Ready系统的体系结构[8] 如图1所示。整个系统由三个子系统组成：索引，服务器和浏览器。索引子系统如图所示为一个技术的级联，需要一个单一音频波形作为输入，并产生输出一个紧凑的结构上的摘要，编码为一个XML文件输入到服务器。输入波形的持续时间为几分钟至几小时不等。整个索引过程在双733-MHz奔腾III处理器上在流模式下实时运行。系统接受连续输入并逐步产生内容索引，关于输入的输出延迟小于30秒。

图1 Rough’n’Ready音频索引和检索系统的分布式体系结构（图见原文）

服务器有两个功能：一个是收集和管理存档，另一个是和浏览器进行交互。服务器接收索引的输出，并将它们逐步加入到它现有的音频存档中。对于每一个索引处理的音频段，音频波形被标准MP3压缩处理并存储在服务器上，以便应对将来从客户端（浏览器）得到的重放请求。XML文件包含了从索引中自动提取的特征，它被上传到一个关系数据库中。最后，所有的音频段中的故事都被建立索引，以便进行快速信息检索。

浏览器是Rough’n’Ready系统中唯一一个与用户有交互的部分。它的主要任务是向服务器发送用户请求并有目的地显示结果。各种各样的浏览，搜索和检索工具，可供浏览音频存档和查找感兴趣的信息。浏览器被设计为一个Action X控件的集合，使得它既能够作为一个独立的应用程序运行，又能够嵌入到其他应用程序中，比如互联网浏览器。

*B. 索引和浏览*

如果我们取一个新闻广播，将音频输入到一个说话人无关的连续语音识别系统中，输出将会是未分化的字词序列。如图2所示为一个电视新闻节目（ABCs *World News Tonight*, Jan. 31, 1998）中的一集的这种输出的开始。即使这个输出结果中并不包含任何识别错误，却也很难浏览它并且通过扫视就能知道这个广播所讲的内容。

图2 由BBN Byblos语音识别系统产生的一个*World News Tonight*音频广播的转录结果（图见原文）

图3 由Rough’n’Ready产生的自动的结构上的摘要的元素（图见原文）

说话人分段和识别结果显示在左边；人、地点和组织的名字在中间部分用彩色显示；

和故事相关的话题显示在右边。所有结果均为从新闻广播中自动提取。

现在，比较图2和图3，图3为Rough’n’Ready浏览器的屏幕截图，显示了一些该系统应用到相同广播时的音频索引组件的结果。原本是未分化的字词序列，现在已经分成像段落一样的片段，片段之间的边界和说话人之间的边界相对应，显示在最左边一列中。这些边界是系统自动提取的。说话人片段被按照性别识别，并且在整整半个小时一集内聚类成组，相同的说话人在相同的标签下。一个说话人Elizabeth Vargas被按照名字识别，使用了说话人特定的听觉模型。系统通过说话人分段，聚类和识别的组件推演获得这些音频剧集的特征。

图3中中间一列中的彩色字词显示了人名，地名和组织名——所有重要的实义词——它们是通过系统的名字定位组件自动找到的。即使转录的结果包含语音识别错误，这里显示的增加的版本容易阅读，并且最小的努力就可以得到故事的要点。

图3中最右边一列显示的是一个话题标签集，这些话题标签是通过系统的话题分类组件自动选择的，以描述新闻广播的第一个故事的主要主题。这些话题标签是从系统已知的超过5500个可能的话题中取出来的。这些话题标签建立了一个非常高水准的基本口语的内容摘要。

图3所示的话题标签实际上被系统应用于一个字词的滑动窗口，然后由此产生的一系列话题标签被系统的故事分段组件用来将整个新闻广播划分成一系列故事。对这一集故事分段的结果如图4所示，是另一个音频浏览器的屏幕截图。

图4 一个音频存档的高水平组织，将一集*Headline News*显示成一系列主题故事，均从新闻广播中自动提取（图见原文）

将连续口语字词流打碎成一系列有边界的、有标签的故事，这是一种新的强有力的能力，使Rough’n’Ready能够有效地将一个大的音频记录存档转换成文件式单元的集合。在图4所示的浏览器的视图中，一个由150小时广播新闻组成的音频存档被按照不同的内容制作商组织成剧集的集合。特定的一集（CNN *Headline News*, Jan. 6, 1998）被扩展以显示系统针对这一特定剧集检测出的故事序列。每一个故事由一个简短的话题标签列表表示，这些话题标签是由系统选择来描述故事的主题的。这种表示的实际结果是，一个人可以通过一个小的具有高度描述性的标签集，很快地获得一个新闻广播的内容要点。

图4所示扩展剧集的第一个故事和Sonny Bono遭遇的重大滑雪事故有关。这个故事的三个重要主题——化学，事故和Sonny Bono——都被系统自动识别。同样重要的是，系统拒绝了这个故事其他的全部5500个话题标签，仅留下了这里显示的简明的四个话题标签的列表，来描述这个故事。值得注意的是，系统从未在进入训练集前将这些话题放在一起观察，因为Bono只死了一次。尽管如此，它能够从一个非常大的可能的集合中选择这个非常详实且简约的话题列表，同时将连续字词流分割成故事序列。

整个广播新闻音频存档以和图4所示扩展剧集相同的方式被自动归纳。这意味着存档可以当做文本文件的集合来对待，可以简单地操作和搜索，就像我们联系因特网搜索和检索操作一样简单。转录结果的每一个字以及系统提取的所有结构特征，在剧集中都和一个时间偏移相联系，这使得原始的音频或视频片段可以在需要的时候从存档中检索到。待检索的实际片段可以很方便地由用户限定范围，作为一个故事，作为一个或多个说话人片段，或者作为在转录中连续字词的任意跨度。这使用户能精确控制待检索的片段。

现在我们回到这篇文章的主要议题，即对Rough’n’Ready系统所使用到的各种语音和语言技术的一个描述，之前对这些技术的一般建模模式做了一个简短的阐述。对那些更新近的贡献的描述会比那些已经发展了很多年的技术更加详细。

**3. 统计建模模式**

这篇文章中描述的技术都遵循相同的统计建模模式，如图5所示。系统有两部分：训练和识别。给定一些感兴趣的数据的统计模型，系统的识别部分首先将输入数据分析为特征序列，或特征向量，然后执行一个搜索且输出序列，此序列是在给定特征序列的情况下最大化输出序列的概率之后得到的输出序列。换句话说，即选择使*P ( output | input, model )*最大的输出，*P ( output | input, model )*为输出的概率，事先给定了输入和统计模型。训练程序根据经过分析的训练数据库以及相应的基本事实（即这些数据的理想的公认的序列），来估计统计模型的参数。统计模型本身由技术开发者所指明。

图5 本文介绍的语音和语言技术所使用到的统计建模模式（图见原文）

这个方法的一些特性如下：

1) 严格概率形式，允许从不同知识来源通过联合它们的概率来进行信息的综合。

2) 基于有注释的训练数据库（注释是提供基本事实的过程）的对模型参数进行估计的自动训练算法。此外，注释是可提供的，仅需要专业知识，且是可以由学生或实习生完成的。

3) 语言无关的训练和识别，仅需要从一个新语言得到的有注释的训练数据。

4) 先进的性能。

5) 面临退化的输入时是鲁棒的。

下面我们将看到这个模式是如何在不同技术中工作的。

**4. 语音识别**

广播新闻的自动转录是一个有挑战性的语音识别问题，因为会有经常性的且不可预知的变化，包括说话人，说话方式，话题，信道和背景条件的变化。Rough’n’Ready中的转录是由BBN Byblos大词汇量的说话人无关的语音识别系统[9] 创造的。经过了几年DARPA广播新闻评估的参与，Byblos系统已经发展成为一个鲁棒的先进的语音识别系统，它能够转录实际生活中的广播新闻音频数据[10]。

Byblos系统遵循了如图5所示的统计模式。在分析部分，系统每隔10ms计算一次偏mel倒谱系数并得到一个15个系数的特征向量，作为一个时间函数。为了有效地处理广播新闻中的连续语音流，数据被分成易处理的片段，有可能是依赖于说话人或者信道特性（对播音员的语音为宽带，或者对电话语音为窄带）。在下一节中描述的基于说话人的分段，其后是更进一步的基于检测到的停顿的分段[11]。

整个统计模型有两部分：听觉模型和语言模型。听觉模型，对每一个声音或音素描述了特征向量的时变发展，使用连续密度的隐马尔可夫模型（HMMs）[12] 来为各种语音前后关系中的每一个音素建立模型。音素模型的前后关系可以扩展到多达两个之前和之后的音素。加权混合高斯密度——所谓的高斯混合模型——用于为每一个HMM状态建立倒谱特征向量的概率密度模型。如果需要，可以将模型设定成依赖性别的和特定信道的，并且可以将模型配置成能够捕获字词间和跨字词的前后关系。能够专门处理自发语音的声学信息，这是普遍的广播新闻算法，该算法被开发以适应自发语音——包括很短时间——的典型发音，以及特殊的停顿和非语音事件的声学模型，这里非语音事件包括音乐，静音/噪音，笑声，呼吸声和咂嘴声等[13]。

系统中所用的语言模型是N维语言模型[14]，每个词的概率是前一个词（二维语言模型）或前两个词（三维模型）的功能。高阶模型通常会得到较高的识别精度，但速度较慢而且需要大的存储量。

为了得到最高得分的词语序列，Byblos系统使用一种多遍识别搜索方案[15], [16]，它通常从一个大致但快速的首字母开始向后传递——首次匹配传递——伴随在那些作用在更小的搜索空间使用改进性更高精度的模型的传递方式使得搜索空间变窄，从而减少整体的运行消耗。对于Rough’n’Ready，系统在首次匹配传递之后使用两种传递：第一种是反向传递（从一个发音的最后到开始）会创建出一个N个最佳单词序列假设的最高得分的列表（N是通常是在100到300之间），而最后一个传递会存储一个如下所述的N个最好的序列。最终的词语最高得分序列被作为识别的输出。

首次匹配传递，从每一个发音的开始到结束，是一种同步搜索用于伴随着一种机器人语音的共享混合（PTM）声学模型和一种两位近似词语语言模型的单个语音树状算法。这个输出是一个带有单词结束次数的词语图表，被用于引导下一阶段。在一个PTM声学模型中，所有与文本相关的模型的所有隐马尔可夫模型状态的音素被结合在一起，共享一个包含256个成分的高斯混合密度；只有密度权值在状态间变化。带有回溯算法的产生N个最佳的过程运用一个更精确的字内状态聚类的共享混合（SCTM）声学模型和一个三维词语的语言模型。一个音素的所有模型的隐马尔可夫模型中的对应状态被聚类到一个类的数值中共享一个具有64个高斯成分的混合密度。一个典型的SCTM系统通常使用大学3000个这样的类。最终的过程会使用一个跨词语的SCTM声学模型和一个三维词语的语言模型重新对N个最佳进行打分，并选择最可能的假设作为识别的输出。

对于每个说话人，无监督适应的Byblos系统可以被用于改进识别的精度。这个过程需要对说话人改变边界的检测。下一节将描述被用于Rough’n’Ready系统中去计算这些边界的说话人切割。Byblos中的适应性表现是依据由剑桥大学开发的最大可能性线性回归方法[18]。

在实际程序中，例如Rough’n’Ready，录音转录进行得尽可能的快是非常重要的。除了上文描述的搜索方法之外，更快的速度有可能在实际中减少运算时间。近些年的主要加速算法包括快速高斯计算法（FGC），语法传播和N个最佳的树状重判方法[19]。

自从高斯数量结合每个隐马尔可夫模型状态是非常大的（通常大致250000）,高斯计算成为了主要瓶颈。Byblos的FGC算法的实现是由IBM开发的基于决定性的FGC的变体。概念上，整个声学空间可以将一个决定树分割成小的区域，这样每个区域和任意高斯码本都只有一个可以覆盖整个区域的短的高斯列表。在识别中，决定树被用于确定一个只有一些高斯量被用于计算可能性的小的声学区域与每个输入特征矢量相符合。FGC算法可使首次匹配算法加速三分之一，使N个最佳加速二点五分之一，并且不损失精确度。

束搜索算法可以被适应在窄的束中运行的非常快。但是，aggressive窄的束经常会在词语边界过早的修剪出正确的方法，由于可能性得分的突然改变导致语言模型的得分被用于这些边界。为了优化着个方法，卫门已经开发出一种算法可以将语言模型的可能性传播到一个词语整个部分去消除这些较大的得分突变[19]。当解码器在一个词语边界过渡时，就是说，从w1到w2，除了用二维概率P(w2|w1)，我们运用概率比P(w2|w1)/P(w2)。之后我们通过P(w2)(1/k)乘以w2内的音与音之间的过渡从而消除分母P(w2)，k是w2中的音数。我们称这个过程为“语法传播”，并且我们发现它允许我们在返回的过程中使用一个窄的束，从而减少二分之一的时间在计算上同时不损失精度。

最终，N个最佳的重判过程也通过使用一个所有N个假设为了被重新打分被排列成树状的同时减少多余计算的树状重判算法，减少二分之一的时间[19]。

当我们以三倍实际时间（3\*RT）在一个450MHz奔腾二处理器运行Byblos时，在DARPA广播新闻的测试数据上，使用一个60000词汇量的单词表时的词语错误率是21.4%。在十倍实际时间时错误率减少到17.5%，而系统运行在230倍实际时间时减少到14.8%。

**5. 说话人识别**

使用准确声音信号的一个主要优势是识别说话人序列的潜能。说话人识别的问题有三个连续的成分：说话人分割、说话人聚类和说话人识别。说话人分割是根据说话人切割音频流；说话人聚类是将来自同一个说话人的声音片段聚集在一起；说话人识别根据系统中已知声音识别说话人的特征。我们接下来介绍每一个方面。

*A. 说话人分割*

说话人分割的目标是在声音信号中确定说话人之间的所有边界。这在广播新闻中是一个困难的问题由于背景音乐、噪声、变化的频道状态。说话人边界的准确识别为语音识别器提供了一个来自于单一说话人的输入片段，使得说话人的标准化和适应性技术每次被高效的用于一个说话人。因此，说话人改变的边界通过识别器将连续的词语流切分成通常在话题中是同类的段落似的单元。

我们已经开发出一种新的两段式说话人改变检测的方法[21]。第一步是检测语音和非语音的边界的同时（从图1标记，在系统中的这点，语音识别还没有开始），第二步在语音片段中进行实际的说话人分割。由于广播新闻中的80%的说话人边界发生在非语音的间隔时间，所以准确的定位非语音片段是非常重要的。

为了检测语音和非语音边界，我们使用一种粗糙的非常快速的与性别无关的音素识别处理输入。我们拆分音素成为三个宽的类（元音、摩擦音、阻塞音），同时我们包含五种不同的典型非语音现象的模型（音乐、静音/噪声、笑声、喘气、抿嘴）。每个语音类是通过一个五种状态的隐马尔可夫模型和一个64维高斯密度混合。模型的参数从声学数据的20h中准确的估计出。识别器在输入的每一个帧进行语音和非语音检测的结果达到90%。

第二步是在第一个步确定的每一个语音边界上假设说话人改变的边界上进行实际的说话人分割。在语音级别上的时间处理允许算法运行的非常快的同时，在每一帧上保持与假设一个边界同样的准确度。说话人改变的判断通过比率检测的一种形式，即无效假设是相邻片段通过同样的基本分布建立的。通过特征矢量xi和yi给定的两个片段x={xi,i=1,...,N}和y={yj,j=1,...,M}，各自的，我们假设x和y是由高斯运算得到的。由于两个片段的均值对于背景影响非常敏感，我们只能使用协方差来建立比率，它来自于[22]



式中的z是x和y的综合，[[]]是估计每个过程的协方差矩阵的最大可能性。它通常导致当我们估计的高斯参数的数据越多，[[]]越高[22]。为了减少偏移，一个标准化参数被提出，所以比率测试变为



式中通过经验确定且大于1.这个标准化可能性比率与Bayesian信息评判标准中所用的相似[23]。然而，在我们的这个问题中，我们可以使用额外的信息，即一个说话人的改变更有可能的发生在非语音的间隔时间，从而改善我们的判断的方式。最终的测试，因此，具有以下形式。



1）在非语音区域：如果，那么片段x和y被认为是来自同一个说话人，此外，t是一个阈值调节使错误接受和错误拒绝的错误最小。



2）在语音区域：测试改为+a，是一个正数阈值进行与1)中同样的调节方式。被用于向非语音区域移动语音/非语音边界从而减少分裂词语的可能性。



我们实现了一个连续的过程，即在一个语音中每次增加一个说话人片段同时利用上述算法假设说话人的改变在每段语音的边界。这个过程是近似因果的，即使只有2秒的前瞻量，足够得到充分的数据进行检测。这个方法用于DARPA广播新闻检测的结果是在100毫秒的正确边界（一个音素的大致持续时间）中找到72%的说话人改变，错误率为20%。大多数遗漏的边界是短暂的问候或感叹词例如早上好或谢谢，同时大多数错误接受是在非语音区域和不重要的部分。

*B. 说话人聚类*

说话人聚类的目标是将一个单一广播或情节中的来自同一个说话人的所有片段识别出来并标记一个单独的标签；它是无监督说话人辨认的一种形式。在广播新闻中的问题是困难的是由于信号极端的可变性以及说话人的正确号码可以非常的大（序列1-100）。我们已经发现一种使用自底向上（聚合法）的聚类方法的可接受的解决方法，通过由于假设类的数量的功能导致的类的总数被限制。

每个说话人片段的特征矢量是由一个单独的高斯密度构造的。式(1)中的可能性比率测试被重复的用于对那些被认为非常相似的单元的所有片段的集体聚类匹配聚集成为一个类并且一个完整的聚类树被建立。在程序的每一个步骤和每一次聚类，一个新的高斯模型被用于那个类[25]。现在说话人聚类的问题减少为找基于最有利的判断标准的方式修剪聚类树。我们选择最小化的判断标准是两个部分的和

式中的k是为了具体修剪聚类树的类的数量，Nj是类j中特征矢量的数量。式(3)中的第一部分是确定类内的方差矩阵的公式[24]，第二部分是用于调整或减少的部分用于补偿，因为方差矩阵的决定因素是关于k单调递减的。最后的聚类是修剪聚类树从而最小化式（3）。的值是根据最优化运行而通过经验确定的；它通常在值域0<<1内。



这个算法在非常广的范围的新闻广播上提供了高效性。它在忽略片段中说话人真正数量上表现的很好，因此产生了类的高纯度。类的纯度被定义为被正确聚类的帧的百分比，其值为95.8%。

*C. 说话人辨认*

在说话人聚类的阶段中，每个说话人的类的创建是通过性别进行辨认的。每个性别的高斯混合模型是通过大量按性别区分的训练数据进行评价得到的。接下来，一个说话人片段的性别通过计算男性和女性模型的可能性比率对数确定的。这种方法在性别检测中有2.3%的错误。

除了性别，如果提供来自说话人大约一分钟的语音，系统可以辨认一个特定的目标。再者，一个高斯混合模型通过训练数据确定并且利用[26]中具体的方法来确定语音中的来自目标说话人的片段。任何数量的目标模型可以被建立而且同时在系统中用于辨认说话人。为了标记判断，目标模型的集合会与来自数以百计的说话人的语音的与文本无关的一大批模型进行对比。每一个目标说话人模型都适应于与文本无关的模型。为了改善不同说话人改变频道的影响，倒频谱平均值消去法使得每一个说话人片段的特征矢量的均值在建模前被减去。

在DARPA广播新闻资料中，20%的说话人片段是来自20个已知的说话人。因此，说话人辨认的问题就是数据包括已知和未知说话人的开集辨认问题，并且系统不得不决定识别已知的说话人，拒绝未知的说话人。使用上述的方法，我们的系统产生以下三个类型的错误：一个已知说话人片段错误辨认为另一个说话人的辨认错误率为0.1%；一个已知说话人片段被归类为未知的错误拒绝率为3.0%；一个未知说话人片段被归类为来自一个已知说话人的错误接受率为0.8%。

**6. 名字辨认**

名字辨认在Rough’n’Ready中的目标是从语音中抽取重要的成分并收集在一个数据库中。目前，系统可以定位人物、地点和组织的名字。这个领域大部分先前的工作只考虑书面语文本资源并且集中设计规则驱动的算法来定位名字。从口语自动抄写中脱离出来要比从书面语中更加困难，因为缺少大小写、标点和句子边界，还有识别错误的存在。在使用规则驱动的系统中存在相当数量的负面影响。为了克服这些问题，我们已经开发出名叫IdentiFinder的一个基于隐马尔可夫模型的名字抽取系统。这个技术只需要我们提供通过打字输入的训练文本和名字在整体中标记的位置。这个系统另外的优势在于它对其他语言很简单的提供了接口，只需要来自一个新语言的带有注解的训练数据集。

名字辨认的问题在图6中已经被解释了。人的名字（Michael Rose, Radovan Karadzic）使用黑体字；地点（Bosnia, Pale, Sarajevo）使用下划线；组织使用斜体字。我们需要找到所有这三个名字的集合并将其他归类为通用语（GL）。

图7表现了IdentiFinder中使用的隐马尔可夫模型对每个类型的名字内容进行建模。这个模型包含了一部针对三个名字实体加另一步对文章中其他所有文字，并通过过渡状态从一个阶段转变到另一个阶段。连接每个状态的是词汇表中的所有词语的二维统计模型，一个不同与为每个阶段进行评价的二维模型。通过认为这个是一个可以产生文本中所有单词的生成性的模型，大多数时间我们在GL阶段中说出通用语词汇。接着在我们像生成一个名字时，我们会转移到名字阶段中的一种；我们会停留在这个阶段生成那个名字。然后我们要不转移到另一个名字阶段，要不更有可能的回到通用语阶段。说出每个词语或转移到另一个阶段的决定是根据前一个单词和先前的状态。这样，模型使用上下文来帮助检测和归类名字。例如，在通用语中单词“Mr.”有可能接着转移到人物的阶段。在人名字产生后，有可能转移到GL阶段并随后产生“said”或者“departed”。这些与上下文有关的影响都被考虑在我们的模型中。

图7中的模型参数是通过带有三个名字集合标记注解的训练数据自动确定的。接着，通过给定的文本，模型被用于对每个单词属于三个名字类之一或不属于的概率进行判断。然后，我们使用Viterbi算法[28]找到通过文章计数的最大可能性的状态序列。结果就是名字类的序列的答案。

图6 一个定义了三个名字类型的句子：人物（Michael Rose, Radovan Karadzic），地点（Bosnia, Pale, Sarajevo），和组织（U.N.）。（图见原文）

图7 隐马尔可夫模型被用于寻找名字的IndentiFinder。每个状态都包含了词汇表中所有词语的二维统计语言模型。（图见原文）

自从我们的系统通过来自新闻广播的一百万个词语的带注解的数据训练后，与文本无关测试集中的许多词语对于名字辨认系统来说成为不可知的，即使它们对于语音识别器来说是已知的。（那些对于语音识别器未知的词语会被错误的识别为一个存在的词语，因此降低效果。）处理这些未知词语是十分重要，因为这些词语中的一部分存在于所期望的名字中，并且我们希望系统能够辨认出它们，即使它们在训练部分从未见过。训练时，我们将训练数据分为两半。在每一半，我们将另一半中未出现的句子替换为“未知”。接着我们对包括未知词汇的词汇估计所有的概率。所有已知词汇的概率由所有数据进行评价。在测试阶段，我们在名字辨认系统中将所有未知的语句替换为“未知”并可以找到状态的最匹配序列。我们发现通过使用特定用途的上下文，许多未知的名字在名字辨认系统中被正确的标记出。

我们进行信息抽取的方法的一个优势在于我们可以学习不同文章类型的统计规律的简单性。例如，我们希望系统工作在缺少关键信息的文章（例如：文章显示为全部大写或全部小写）。从我们的注解文章中移去关键信息并且重新评价模型是非常简单的。如果我们希望使用IdentiFinder作为语音识别器的输出，我们希望文章既没有大小写，也没有标点。此外，没有缩写并且数值是被拼写出来的（例如：TWENTY FOUR要比24更好）。再者，为了学习来自于语音识别器的文本输出模型，我们可以轻易地模仿这个效果在我们的注解文章中。当然，通过给定的来自一个新语言的带注解的数据，在那个语言中训练同样的系统区识别名字是非常简单的。

我们已经进行了若干个实验检测IdentiFinder在寻找名字中的表现。另外，我们还检测了但缺少大小写和标点信息或面对来自自动语音识别错误时的性能损失。在测量系统的精确度时，同时考虑名字的类型和文章中对应词汇的跨度。我们通过名字错误的数量除以参考答案中正确名字的总数来测量空隙错误率，类型和跨度被分别计数为不同的空隙[29]。

在来自于DARPA新闻广播资料的测试中，名字的种类数是7（而不是Rough’n’Ready中所用的3），IdentiFinder对于含有大小写和标点的文章得到了11.4%空隙错误率。当所有的大小写和标点被去掉后，空隙错误率仅仅增加的16.5%。

在最近以语音输入的名字辨认的DARPA评价中，同样是具有7个类别的名字，对于Byblos语音识别器的输出的空隙错误率为26.7%，语音识别词语错误率为14.7%[30]。当所有的识别错误被纠正，不增加任何大小写和标点信息时，空隙错误率下降到14.1%。总体而言，我们发现名字辨认的空隙错误率对于词语错误率大致上是一一对应的线性增加。

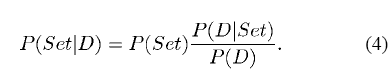
**7. 话题归类**

许多成就已经被实现在话题归类方面，这些不同话题的模型被独立的评估，即使每一个文章中包含着复杂的话题。一个显著的例外来自Yang和Chute的工作，作为他们的模型中的一部分，他们考虑通常上同时复杂的话题是与每一篇文章相联系的。我们对于话题归类的方法在原则上是与Yang和Chute相同的，除了我们使用Bayesian框架[32]取代了基于距离的方法。我们成为OnTopic的话题聚类成分是参数通过给定具有上千个话题标记的训练样本文章得到的概率隐马尔可夫模型。模型允许每一个文章中的词语对文章中标记的每一个话题打不同的分数。来自OnTopic的输出是一个对于给定文章所有可能话题和对应得分排序列表。

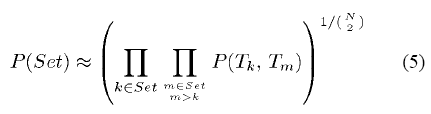
图8 OnTopic中使用的隐马尔可夫模型对一个故事中的话题的集合进行建模。这个模型可以对每一个故事中数以千计的话题进行判断。（图见原文）

*A. 模型*

我们选择对应于给定文章D的话题集合Set，这样事后概率P(Set|D)是最大的。

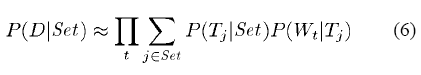


为了排列话题集合的目的，P(D)可以被忽略。事前概率P(Set)真正的结合了在集合中含有所有标记的文档的概率,这个概率可以使用话题同时发生概率近似为P(Tk,Tm)



式中的N是Set中话题的数量，指数使用来处理不同大小的相似话题集合。P(Tk,Tm)通过P(Tk|Tm)和P(Tm)的最大可能性的乘积确定。前者是通过那些一个Tm为话题也以Tk为话题的文档的分数确定的，后者是通过以Tm为话题的文档的分数确定的。

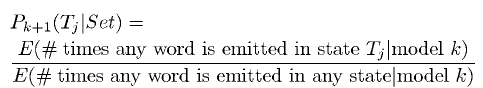
剩下需要计算的是P(D|Set),词语在标记了所有话题的Set文档中的环境概率。我们使用隐马尔可夫模型对这个概率建立模型，其中包含集合中每一个话题和Fig8提出的额外的状态GL.模型在文档中一个一个的产生词语。首先选择一个话题范围根据P(Tj|Set)产生下一个词语，接着根据P(Wt|Tj)选择一个词语，然后在选择另一个话题范围，以此类推。因此，P(D|Set)的公式是



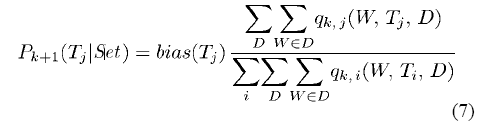
式中t在文章的词语集合中变化。上述等式中的成分是通过下述的训练数据确定的。

*B. 确定隐马尔可夫模型参数*

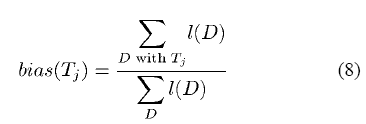
我们使用一个最大期望算法[33]的一个有偏移的形式去评估图8中隐马尔可夫模型的跃变概率P(Tj|Set)和发射概率P(Wt|Tj)。跃变概率定义为



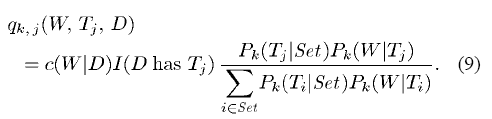
可以通过下式确定



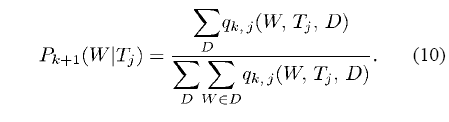
其中



是一个偏移单元，l(D)是文章D中词语的数量，并且



qk，j(W,Tj,D)是在D中用Tj计数的W的数量分数，通过在生成的模型中给定的现有参数的集合；c(W|D)是文章中词语W出现的次数；I(x)是一个指示功能，当它被预测为真和0时，返回1。偏移单元需要将观察数据向GL状态偏移；否则，EM算法会在GL状态下得到一个零跃变概率[31]。偏移的作用是将话题Tj的跃变和发射概率设置为全部词语中这个话题词语的分数粗略的等于bias(Tj)。发射概率通过下式确定



*C. 归类*

为了对给定的文档进行归类，我们需要找到最大话题的集合（4）。但是所有可能集合的总数是



当可能话题的数量M是数以千计的话，这是一个非常大的数。由于计算如此大的数的概率会阻碍计算速度，我们使用一个两过程方法。在第一遍，我们在可能为最佳集合中选择一个小的话题集合。在第二遍时，我们使用（4）计算所有这些候选集合的得分。我们在第一遍假设话题是独立的整体通过分别计算每个话题来选择候选话题，使用将(4)轻微修改的公式

式中当x<0时是0，并且用于过滤出组成文章中话题的负面信息的词语影响。参数被确定用于平衡生成的模型并对训练模型优化。为了补偿文档中对词语的独立假设，参数会使变平（当小于1）或变尖（当大于1）跃变概率的区域。



*D. 实验*

我们 OnTopic 分类器应用了二步过程对一个被主要来源媒体注解和发步的广播新闻报导在上面描述。为每个故事，注解者给予了一些主题标签他们想在故事中表现的主题。通过数字1到13标注每个主题，每一故事平均4.5个主题。集成被分成一年，即42502个故事之内，用来训练，和一月即989个故事，用来测试。训练组包含4627个独特的主题标签。

我们系统的表现和人工注释写下当做主题标签是不同的，因为我们的系统下达被命令的目录所有的主题，每个算出得分，但是注释对每一个故事有一个小的，不排序的目录。图9（图见原文）表示测量表现的不同合理方法。横坐标表示由系统提供的前N个主题。对于每个N的值，我们比较根据系统产生的和那些注释产生的主题。从两个方面来比较。那至少有一正确的曲线表示那部分的故事为至少N个主题之一，给每个故事的标签被包含在注解之中。清楚地，那一个分数随着N的增加而增加。我们见到，举例来说，得分的顶端被认为了正确了76%的时间。在比较的第二个方法中，我们比较因系统而产生的所有的顶端主题与相同的注解的主题和计数的组的不同，然后我们测量正确率和召回率。正确率是系统计算正确的结果（也就是，与人工注释吻合），召回率是人工标注的通过系统计算出来的。正常情况下，二者此消彼长。

我们有指示标准我们采用测量我们的系统不是十分必要。主题注解是不一件容易的任务当主题的数字是大的时候；人们倾向于标签为文件，因为记得这么多主题很困难。在非正式考察中，边界处的主题通常没有人工标注，我们时常发现计算机给的主题相当合理。

应用OnTopic到任何可能的广播新闻（举例来说，对于每个说话者片段）,为了索引的目的，使用主题分类会是更有用的一个方法。这是下一个部分的主题。

**8. 故事分段**

故事分段把每个连续的语音单词分成有标注的许多的音节。在 Rough’n’Ready，每个数据窗口包含200字的，连续的窗口之间有四个字的交叠。对于每个数据窗户，和为已知的5500个主题，我们计算每个窗口可能的主题。5500目录对于每个数据窗口的主题得分自动地只选取最高得分（也就是，最有关的）。得分最高的100个主题被一个高斯过程刻画，而且我们选择当做我们的被修剪的目录那些主题得分是平均分数的两倍。这一个程序的结果在图10（图见原文），表示被描述主题选取程序的结果的一个转变。

问题现在将位于片段的边界。我们定义一个主题窗口为50连续的主题，而且我们计算主题持久当做每个主题标示的发生的数目。我们然后测量在一给定的主题窗口中最大值－持久得分当做最大的持久发现为任何的主题。图11（图见原文）表示最大值－持久得分当做横跨一段插曲的一个主题窗口的功能。最大值50数值典型地在区域期间被到达相同的故事里面。垂直在图11表示不同的片段的真实的交界。藉由设定一个90%的最大值的临限，当做显示图11的水平线, 我们能搜寻那在临限下面的区域的片段交界。

通过注意关键字在文本的位置，片段交界可以更精确。关键字是那些在一个主题窗口中占得分重要因素的部分。我们只观察那在一个故事的大约6%-8%的字组然后提供任何的支持主题标示。我们也观察到那支援字组时常最容易进入二个分组每当他们跨越一个片段交界时。一个组代表前面主题，而另一个代表后面的。我们利用这自动产生使故事交界在稳定的主题区域之间。我们也强迫交界偏爱一个附近的片段而避免音节被分断。较进一步的明细在[34]中提供。

故事分段程序被测试集成了有105段插曲与总共966个故事。如有一个50字组的容许度，故事分段程序正确地发现77%的交界而且有90%的错误接受度,也就是，对所有交界，大约有二个边界被发现了在了两个分段中。较长的故事，应该分在哪个主题，很容易被我们的程序细分，这也是可容错率高的原因。注意，为了索引的目的，如此高的错误的认同率并没有大的影响。图4表示了在一段广播新闻上的故事分段的结果。

**9. 信息识别**

Rough’n’Ready浏览器有能力识别以说话人、主题，和│或名字为基础的重要故事人、地方和组织。另一个浏览器的能力是将找到与给定故事类似的重要故事。要执行这一个任务，我们使用一本小说信息检复（IR）系统，叫做Golden Retriever[35]。信息索引和辨识发生在Rough’n’Ready的服务器。每当newepisode 被处理在索引制作人手边，一个新的撷取索引被产生在那之上编入索引的故事的整个档案库。浏览器给使用者一个有力的例子查询能力，那里一整个的新闻报导交付给Golden Retriever，搜寻引擎当做一个在一个大的声音档案库中查找所有的相似故事。这提供一个有效的方法给一个使用者发现更多的相关的例子。这一个能力在主题分类和音节分割上有利用价值。

Golden Retriever是一个小说检索系统，基于HMM的IR计算一个文件属于有关的系统的可能性，给定一个查询，列出所有在集合中可能的结果。IR反映了我们的主题分类的工作：我们通过例子集合来构建集模型和判断程序。集合有一系列文件，一组自然语言查询（十个一组）, 和一些中肯陈述的裁判是否每个文件与输出查询相关。在文件的集成的一些重要的抽取样品上对于每个查询人工的注解者作中肯审判。我们建立一个统计的模型排名训练文件，有效地藉由他们的标示给定训练查询。

**IR的Bayesian模型**

如有一个查询，可能作为某一个分类的可能性[36]。在其他者里面字组，我们想要使用文件排序的功能：后验概率，可能性文件D是有关的，给定查询Q。我们再次使用Bayes'规定分解后验概率



是有关的一个文件的之前可能性是至任何的查询。现在，我们承担，这更重要者是均匀，虽然，大体而言，我们能制造那居先一个功能文件扮演重要角色。P(Q)是先前的可能性被当前查询引起的。因为这量不改变文件排名，我们能安全地不理睬它。左边是查询的条件概率引起的，在文件是有关的假说之下。我们做模型这个剩馀的量与一不连续的依赖文件的HMM。这将会是我们想到文件HMM的一个模型当做产生查询。HMM的叁数应该被估计在如此的一个方法例如让它变成更有可能的那一文件将会产生一个查询到哪一个它是有关的超过一查询到哪一个它不是有关。



必要HMM的一个简单形成仅仅有二个状态，如图12所示（图见原文）。被标示“D”的状态表现藉由直接地画字组产生查询字组的选择文件。被标示“GL” 的状态表现选择一般的语言的字组，也就是，没有关心文件。大多数的查询包含在场的字组在有关的文件中，但是所有的查询包含许多不是字组有关的文件。

**训练IR HMM**

HMM 的叁数是过渡可能性而且在每个状态为每一个字组的发射可能性。大体而言，我们从例子theEMalgorithm估计这些叁数。在实践中，然而，我们发现没有充足的训练例子为发射可能性找好估计。如此我们为D设定发射可能性和GL状态成为在文件中的字组的unigram分配和整个的集成，分别地。更进一步，我们设定过渡可能性对所有的文件相同，和我们估计a使用EM运算法则。我们发现a=0.7是EM的聚合值。

**演示**

我们已经在本文上测试这简单的二状态的HMM检索算法（TREC-7）集成，有528155个文件[35]。我们轻轻的预加工集成在命令进入字组之内向上分离文件而且形态学地允许相似的字组相配。每个文件的字符串进入字组。然后我们进行合并藉由波特的运算法则[37]。下一步，我们丢弃任何事在一连串的400个“停止”字组中发现的东西。最后，数字的和非字组项目被减少以选出记号（“数目”，“元”等等）。

测试用一个57.6平均包含了 50个查询每一查询的字组。每个查询被预加工了在相同的样子描述在为文件上面。然后，为每个查询，我们为每一个文件计算（12）。结果每个查询的高得分文件检索率在78％以上。

简单的模型在上面描述已经被延长，通过用不同的查询增加较多的状态产生期限的机制（举例来说，同义字、双字母组，主题，无人监督的中肯反馈）, 和藉着文件的包含居先，造成较高的正确率[38]。

**10. 未来方向**

这一篇文章把重心集中在演讲了和语言技术那是对索引和声音的浏览的需要数据。艺术级技术已经被整合在一系统，Rough’n’Ready，自动地获取输入的数据流，并在数据库中检索出合适的结果。系统明确地处理口语了因为说话总是一个信息的富有来源，以在电话中交谈为例子，它是唯一信息。然而除了语音之外，还有更多的信息模式。在一个媒体中比如新闻，很多有价值的信息能被直接地从电视或在荧屏上的本文吸取。在如此的一个媒介，从所有的可得模态得到一个索引很重要。我们没有怀疑一个专业的语音检索系统，比如Rough’n’Ready ，可能有效地在一个广泛的多媒体信息中当做一个索引系统。

在这篇文章中被描述的技术默认在优越的环境条件下，根据工作环境的不同，这种准确度也会改变。即使在广播中的听觉环境新闻可能是非常复杂的――从工作室性质的电话上的面谈或在栏位中――很多的演讲通常是一个喇叭筒的广播。有其他的应用程序，如此会议，听觉的环境是更加挑战性的。除非会议高度地使用特殊化喇叭筒，被捕获的演讲讯号是有可能的高度地杂响。所有的这些效果将会有对检索结果产生干扰。应用于这方面的研究是十分必要的。

我们相信，在这篇文章中被描述的技术有达成了商业的公用程式可能是可能的。在最不久的将来，至少为应用程序类似广播新闻。无疑地，它将会拿一些实际的在技术前的商业的尝试，而且得到广泛地应用。当改良准确度的时候那元件技术将会使如此的系统变成更可使用，它将会是其他的工程学，介面和人类－因数将会决定这些系统的公用程式进入的议题短期。如何将会对来源的系统介面数据？数据将会哪里被储存和在什么花费？使用者将会与系统互动吗？对决意系统必须被调节到特定的应用程序和在什么花费？每个新的应用程序将会需要不同的能力吗那将会必须被包含在系统之中？举例来说，如果应用程序需要，商业广告被识别明确地，然后如此的一个能力可能必须被发展分开地而且包含。简而言之，当我们相信那的时候演讲处理技术已经达成点在商业已经变得可能的地方，实际者商业的程序，一如往常以任何的新技术，蕴藏着许多障碍和有的议题被解决在技术调查它的完整定位之前。

**鸣谢**

在Rough’n’Ready使用的每一场核心内容和语言技术已经被DARPA支援和另外地美国政府的代理商。技术有非常被正式竞争的技术评估获益由 DARPA 赞助而且被 NIST 实行许多年。在发展用的许多数据和技术的评估分配在宾夕凡尼亚州的大学的数据协会。各种不同评估的结果，连同文件一起从参与了评估的不同的位置，能在年度 DARPA 工作室的进行被发现由摩根 Kaufmann 公开。作家想要谢谢书评人为他们的优良又有帮助的评论。

书面翻译对应的原文索引

[4] Xiaodan Zhuang, Xi Zhou, Thomas S. Huang and Mark Hasegawa-Johnson. Feature Analysis and Selection for Acoustic Event Detection [C]. ICASSP, 2008: 17-20.

[30] John Makhoul, Francis Kubala, Timothy Leek, Daben Liu, Long Nguyen, Richard Schwartz and Amit Srivastava. Speech and Language Technologies for Audio Indexing and Retrieval [J]. Proceeding of the IEEE, 2000, 88(8): 1338-1351.