基于支持向量机的音频分类与分割

白 亮 老松杨 陈剑赟 吴玲达
(国防科技大学多媒体研发中心 长沙410073)

摘 要 音频分类与分割是提取音频结构和内容语义的重要手段,是基于内容的音频、视频检索和分析的基础。支持向量机(SVM)是一种有效的统计学习方法。本文提出了一种基于SVM的音频分类算法。将音频分为5类:静音、噪音、音乐、纯语音和带背景音的语音。在分类的基础上,采用3个平滑规则对分类结果进行平滑。分析了SVM分类器的分类性能,同时也评估了本文提出的新的音频特征在SVM分类器上的分类效果。实验结果显示,基于SVM的音频分类算法分类效果良好,平滑处理后的音频分割结果比较准确。

关键词 音频分类与分割,支持向量机

Audio Classification and Segmentation Based on Support Vector Machines

BAI Liang LAO Song-Yang CHEN Jian-Yun WU Ling-Da
(Multimedia Research and Development Center, National Univ. of Defense Technology, Changsha 410073)

Abstract Audio classification and segmentation are an important access to extract audio structure and content, and are a basis for further audio/video retrieval and analysis. Support vector machines (SVM) is a valid statistic learning method. In this paper, the work on audio classification based on SVM is presented. Five audio classes are considered in this paper: silence, noise, music, pure speech and speech over background sound. Three smooth rules are presented and applied in the final segmentation. The performance of SVM on audio classification is evaluated. The effectiveness of some new proposed features is also evaluated. Experiment results show that SVM performs very well for audio classification and segmentation accuracy is good with the proposed three smooth rules.

Keywords Audio classification and segmentation, Support vector machine

1 引言

音频数据是很多多媒体应用中的重要组成部分。原始音频数据本身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。音频自动分类与分割是提取音频中结构化信息和语义内容的重要手段,是音频和视频内容理解、分析与检索的基础[1,2]。

从本质上讲,音频数据的分类是一个模式识别的问题,它包括两个基本方面:特征选择和分类。音频分割是在音频分类的基础上从音频流中提取出不同的音频类别,也就是说在时间轴上对音频流按类别进行划分。分类是分割的前提和基础,对音频流的准确分割是最终的目的。

很多研究者在这个领域做了大量的工作,提出了不同的音频特征和分类方法。主要存在以下两个方面的问题:

第一,这些研究中[3~5]大多采用相对简单的特征,同时处理的分类问题也比较单一,通常只是对语音和音乐进行分类。在简单的分类中分类精度比较满意,但如果分类对象增加,比如加入环境音、非纯语音等或者取较小的窗口,则只使用简单特征进行分类,精度很低。

第二,传统的音频分类算法主要是采用基于规则的分类算法[6~8],即根据一种或者几种音频特征及其阈值判定音频所属的类别。但是,这种方法存在以下缺点:(1)决策规则和分类顺序并不一定是最优的。(2)上层的决策错误会积累到下一层而形成"雪球"效应。(3)分类误差大,需要人的先验知识和试验分析,特别是阈值的确定。所以基于规则得分类算法难于满足不同条件下的不同应用。

一些研究中,提出了很多新的分类算法。文[9]采用NN (nearest neighbor)算法构造分类器。文[10]提出了一种基于神经网络的分类器用于电视节目的分类。文[5]中使用了不同算法构造分类器,包括Gaussian Mixture Model和K-Nearest Neighbor算法。

SVM是在统计学习理论基础上发展起来的一种新的机器学习方法[11],它通过学习在特征空间中寻找最优的分类超平面,可以克服基于规则的分类算法的缺点。因此,本文提出了一种基于SVM的音频分类方法,分类对象是纯语音、带背景音的语音、音乐、静音、噪音。从帧层次上和片段(短时的音频片段,clip)层次上考察了这几类音频的区别性特征并提出了几种新的音频特征。在分类的基础上应用,根据音频流特性提出了3个clip平滑准则,减小了分类误差。在对clip的分类结果进行平滑的基础上,得到最后的音频分割结果。

2 音频特征分析和抽取

在音频分类中,所选取的特征应该能够充分刻画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

白 亮 硕士研究生,主要研究方向为音频分类分段和基于内容的音频信息检索。老松杨 教授,主要研究方向为基于内容的多媒体信息检索。陈剑赟 博士研究生,主要研究方向为视频内容分析、视频摘要。吴玲达 教授,博士生导师,主要研究方向为基于内容的多媒体信息检索和虚拟现实技术。

性。

特征提取前，需要对原始音频数据（采样频率为22.050kHz）做预处理：首先对原始音频信号做预加重(re-emphasized)处理，减少尖锐噪音声影响，提升高频信号，$x(n)$为原始信号，处理后信号$y(n)$：

$$y(n)=x(n)-\text{参数}*x(n-1) \quad (1)$$

参数通常取0.98或0.97。然后将音频分为1000ms(22050个采样)的clip，相邻clip间无重叠部分，再对每个clip加25ms的hamming窗形成帧。相邻帧间有12.5ms的重叠部分。

2.1 基于帧(frame)的音频特征

帧是我们处理的音频信号的最小单位，常用的帧层次上音频特征有以下几种。

2.1.1 MFCC (Mel-frequency cepstral coefficients) MFCC是在Mel标度频率域提取出来的倒谱参数。Mel标度描述了人耳对频率感知的非线性特性。MFCC是语音识别和说话人识别中十分重要的特征，MFCC用在音频分类中也有很好的效果[13~15]，可以提高音频分类的精度。

2.1.2 频域能量(frequency energy) 频域能量定义如下：

$$E=\log\left(\int_0^{\omega_0}|F(\omega)|^2 d\omega\right) \quad (2)$$

$F(\omega)$是该帧的FFT变换系数，ω_0等于采样频率的一半。利用频域能量E来判断静音帧，如果某一帧的频域能量小于阈值，则将该帧标记为静音帧，否则为非静音帧。

2.1.3 子带能量比 将频域划分为4个子带，分别为$[0,\omega_0/8]$,$[\omega_0/8,\omega_0/4]$,$[\omega_0/4,\omega_0/2]$,$[\omega_0/2,\omega_0]$，并计算各子带能量的分布，计算公式如式(3)：

$$D=\frac{1}{E}\int_{Lj}^{Hj}|F(\omega)|^2 d\omega \quad (3)$$

Lj和Hj为子带的上下边界频率。不同类型的音频，其能量在各个子带区间的分布有所不同。音乐的频域能量在上述各个子带区间的分布比较均匀；而语音中，能量主要集中在第1个子带，约有80%以上。

2.1.4 过零率 过零率是描述过零的速度，是信号频率量的一个简单的度量，计算公式如式(4)：

$$ZCR=\frac{1}{2(N-1)}\sum_{n=1}^{N-1}|\text{sgn}[x(m+1)]-\text{sgn}[x(m)]| \quad (4)$$

$x(m)$为离散音频信号。

2.1.5 频率中心(frequency centroid) 频率中心是度量音频亮度(brightness)的指标，一般来说，音乐的频率中心比语音要高。计算方法如式(5)：

$$FC=\frac{\int_0^{\omega_0}\omega|F(\omega)|^2 d\omega}{\int_0^{\omega_0}|F(\omega)|^2 d\omega} \quad (5)$$

2.1.6 带宽(bandwidth) 带宽是衡量音频频域范围的指标。一般来说，语音的带宽范围在0.3kHz～3.4kHz左右，而音乐的带宽范围比较宽，在22.05kHz左右。计算公式如式(6)：

$$BW=\sqrt{\frac{\int_0^{\omega_0}(\omega-FC)^2|F(\omega)|^2 d\omega}{\int_0^{\omega_0}|F(\omega)|^2 d\omega}} \quad (6)$$

2.1.7 基音频率(pitch frequency) 基音频率是衡量音调高低的单位。本文中基音频率采用中心消波(参数为0.70)短时自相关函数的波峰检测算法计算。

2.2 基于片段(clip)的音频特征

clip是分类的单元。根据上面计算的7类帧层次的基本特征，可以计算片clip层次上的特征。计算子带能量比、带宽、亮度等几个通用特征在一个clip中的均值作为该clip的相应特征，同时为了提高分类的准确性，提出下列几个新的clip层次上的特征。

2.2.1 静音比例(silence ratio) 静音比例定义为一个clip中的静音帧占所有帧数的比例。

$$\text{静音比例}=\frac{clip\text{ 中静音帧的数目}}{clip\text{ 中帧的总数}} \quad (7)$$

语音中一般会经常有停顿的地方，所以其静音比例会比音乐高。

2.2.2 High ZCR 比率 根据上面对ZCR特征的分析，在一个音节中语音由清音和浊音交替构成，而音乐不具有这种结构。因此，对于语音信号，其过零率的变化率要高于音乐信号，也就是说一个clip中某个高ZCR值的帧所占的比例高于音乐信号，如图1所示(取160秒长的语音和音乐)。这里取这个ZCR值为一个clip中ZCR平均值的1.5倍。则本文定义特征HZCRR如下：

$$HZCRR=\frac{1}{2N}\sum_{n=0}^{N-1}[\text{sgn}(ZCR(n)-1.5avZCR)+1] \quad (8)$$

N为一个clip中帧数，ZCR(n)是第n帧的过零率。

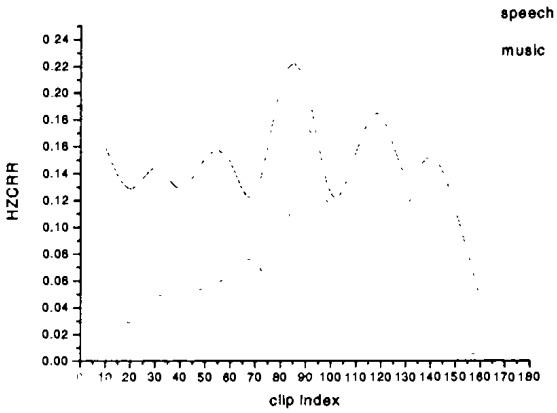

图1 语音/音乐HZCRR图

2.2.3 Low Frequency Energy 比率 与音乐信号相比，语音信号中含有更多的静音帧。因此，语音信号中频域能量低于某个阈值的帧所占的比率要高于音乐信号，如图2所示(取160秒长的语音和音乐)。基于此可以定义特征LFER，这里阈值取一个clip中各帧频域能量平均值的0.5倍。

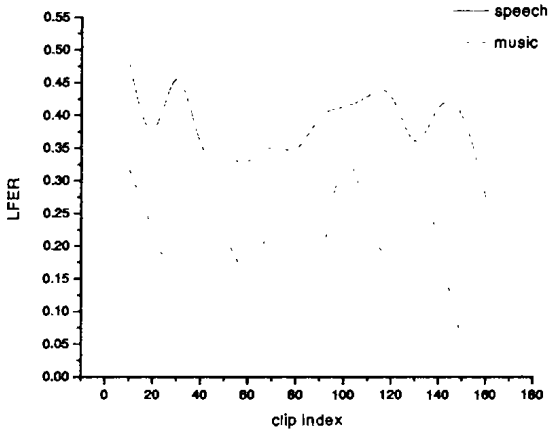

图2 语音/音乐LFER图

$$LFER=\frac{1}{2N}\sum_{n=0}^{N-1}[\text{sgn}(0.5avE-E(n))+1] \quad (9)$$

N 为一个 clip 中帧数，$E(n)$是第 n 帧的频域能量。

2.2.4 *频谱流量* 频谱流量定义为一个 clip 中，相邻两帧之间频谱变化量的均值。计算公式如下：

$$SF=\frac{1}{(N-1)(K-1)}\times\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log A(n,k)-\log A(n-1,k)]^2 \tag{10}$$

其中：

$$A(n,k)=\left|\sum_{m=-\infty}^{\infty}x(m)\omega(nL-m)e^{j\frac{2\pi}{L}km}\right| \tag{11}$$

$x(m)$是输入的离散音频信号，$w(m)$是窗函数，L 是窗长，K 是离散付里叶变换的阶数，N 是 clip 中音频帧数。由于语音信号由清音和浊音交替构成，因此语音信号的 SF 值要大于音乐信号，因而该特征可以有效地区别语音和音乐，如图3所示，0～160秒是一段语音，160～320秒是一段音乐。

文[16]研究表明，清音可以认为是随机噪声激励一个线性时不变系统产生的，而浊音可以认为是一种准周期冲击串激励一个线性时不变系统产生的，二者的基音周期会有显著不同。所以，语音的基音频率曲线应该呈现高低错落，谐成分所占的比例小。而音乐不具有语音的这种结构，常会出现一些基音周期比较平滑的音频段。据此，下列几个特征可以有效地区分语音和音乐。

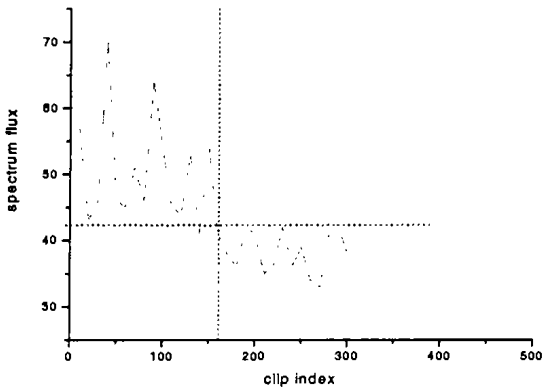

图3 语音/音乐频谱流图

2.2.5 *基音频率标准方差* 一个 clip 中，基音频率的标准方差，衡量基音频率的变化范围大小。

2.2.6 *和谐度* 定义一个 clip 中基音频率不等于0的帧数所占的比例大小为和谐度，比例大和谐度高。

2.2.7 *平滑基音比* 若第 i 帧的基音频率不等于0，并且其与第 $i-1$帧的基音频率差值小于一定的阈值，则第 i 帧为基音平滑帧。一个 clip 中平滑帧的数目与其中基音频率大于0的总帧数之比为平滑基音比。

2.3 特征集的构造

根据上述的特征分析，可以构造音频分类的特征集合。由于不同音频特征的值有很大的差别，因此进行归一化处理：

$x_i'=(x_i-\mu_i)/\sigma_i$，($\mu_i$ 为均值，σ_i 为方差)

由于 MFCC 归一化处理后的实验结果不理想，因此对 MFCC 不做归一化处理，对一个 clip 中的各帧计算12维 MFCC 系数，然后在 clip 内对各维取平均值，作为该 clip 的 MFCC 特征值。这样由13维段层次的基本特征加上12维 MFCC 特征值组成25维的特征向量集，作为 SVM 分类器的输入。

3 分类与分割

对音频数据进行特征提取之后，利用基于规则的算法识别静音 clip 和噪音 clip，对非静噪音 clip，采用基于 SVM 的分类器分为音乐、纯语音和带背景音语音三类。然后对分类结果进行平滑处理，得到最终的音频分割结果。处理流程如图4所示。

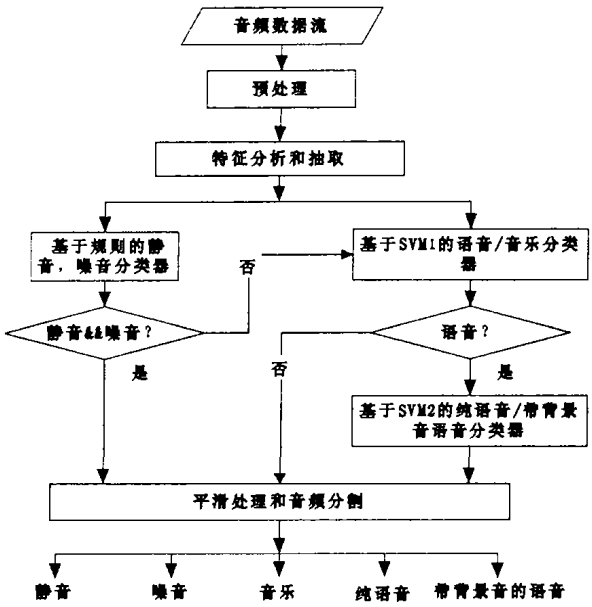

图4 音频分类与分割流程图

3.1 基于 SVM 的分类算法

3.1.1 *SVM 基本原理* 支持向量机的原理是用分类超平面将空间中两类样本点正确分离，并取得最大边缘(正样本与负样本到超平面的最小距离)[12]。该问题可归结为一个二次型方程求解问题[11]，其数学形式为：

$$Minimize\ \Phi(\omega,b)=\frac{1}{2}\|\omega\|^2$$

$$y_i(x_i\cdot\omega+b)-1\geqslant 0\quad i=1,2,\cdots,l \tag{12}$$

范数最小的满足约束的 w 就是最优分类超平面的法向量。目标函数是严格上凹的二次型，约束函数是下凹的，这是一个严格凸规划。按照最优化理论中凸二次规划的解法，可以转化为 Wolfe 对偶问题来求解：

$$Maxmize\ W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j}^{l}\alpha_i\alpha_j y_i y_j x_i\cdot x_j$$

$$s.t.\ \sum_{i=1}^{l}\alpha_i y_i=0$$

$$\alpha_i\geqslant 0\quad i=1,\cdots,l \tag{13}$$

其中 α_i 是样本点 x_i 的 Lagrange 乘子。根据 Kuhn-Tunker 条件，无效约束所对应的 Lagrange 乘子为0，分类规则就仅由恰好在超平面边缘上的少数支持向量决定，而与其它样本无关。

对于线性不可分问题，则将其映射到一个高维空间中，在该空间中线性可分，并且线性判决只需在高维空间中进行内积运算，甚至不需要知道采用的非线性变换的形式，所以避开了高维变化计算问题，使问题大大简化。根据 Hilbert-Schmidt 原理，只要一种运算满足 Mercer 条件，就可以作为内积函数使用。目前常用的内积函数主要有：

(1) 多项式内积函数：

$$K(x,y)=[(x\cdot y)+1]^q \tag{14}$$

(2) 径向基内积函数：

$$K(x,y)=\exp\{-\frac{|x-y|^2}{2\sigma^2}\} \tag{15}$$

(3) Sigmoid 内积函数：

$$K(x,y)=\tanh\{v(x\cdot y)+c\} \tag{16}$$

根据不同的分类问题，可以选用不同的核函数，但目前还没有一个对特定问题选择最佳核函数的有效方法。

3.1.2 基于SVM的分类器的设计 首先选取能量和过零率特征，利用基于规则的分类方法识别出静音和噪音clip。静音的识别规则如下：

① $avZCR<ZCR_{threshold}$

② $avE<E_{threshold}$

$avZCR$ 为该clip所有音频帧 ZCR 的均值，avE 为该clip所有音频帧频域能量均值。同时满足上述两个规则，则认为该clip为静音clip。

本文认为噪音是不包含任何语义内容的音频clip，主要考虑宽带噪声。宽带噪声是比较普遍的一类噪声，其来源很多，如热噪声、气流(如风、呼吸等)噪声及各种随机噪声源等。其在频域上与语音中的辅音频谱相似，宽带噪声的 ZCR 很高，这是因为其高频分量的能量较大，在时域上表现为信号比较杂乱、无规律。其能量谱在较长的时间内(一般大于2s)变化较小。噪声的判别主要依据 ZCR 和能量谱，其判断规则如下：

①对∀ $0\leqslant i<N, ZCR_i>ZCR_{threshold}$

② $\sigma_{Energy}<\sigma_{threshold}$

其中，N 为此段的帧数；ZCR_i 表示该clip中第 I 帧的 ZCR；σ_{Energy} 是此段的能量方差；$\sigma_{threshold}$ 是噪声能量方差的阈值。满足上述两个规则，则该clip是噪音。

然后，对于非静音和非噪音clip采用SVM分类器进行分类，首先构造SVM1区分出音乐和语音clip，然后构造SVM2对语音clip识别，区分出纯语音和带有背景音的语音。本文训练2个径向基函数核支持向量机作为音乐、纯语音和带背景音的语音的自动分类器，径向基函数参数 σ 选择为10。

3.2 平滑与分割

通常，音频流是连续的，音频类型不会频繁或突然交替地改变。基于这个特性，提出下述准则，对不同类别的clip进行平滑处理，提高分类精度。

准则1：假设 $c1$、$c2$、$c3$ 是三个相邻的clip，如果 $c1$ 和 $c3$ 属于同一个音频类别，而 $c2$ 与 $c1$、$c3$ 类别不同，则认为 $c2$ 的类别判断错误，$c2$ 应与 $c1$、$c3$ 类别相同。比如，$c1$、$c3$ 是音乐 $c2$ 是语音，则 $c2$ 分类错误，应该是音乐。

准则2：假设 $c1$、$c2$、$c3$ 是三个相邻的clip，如果它们的类别各不相同，则认为 $c2$ 的分类应该与 $c1$ 相同。比如：$c1$ 是纯语音，$c2$ 是音乐，$c3$ 是带背景音语音，则 $c2$ 应与 $c1$ 分类一致，$c2$ 应为纯语音。

准则3：上述两条准则对静音和噪音不适用，因为静噪音可能突然或频繁出现。比如，如果三个相邻clip分别为语音、静音、语音，则认为这种情况是合理的，不应用上述两个准则。

经过平滑处理，将相邻的相同类别的clip合并在一起，就得到了最终的分割结果。

4 实验结果和分析

试验音频数据来源于CCTV-5体育新闻节目、广告节目的音频内容和CD音乐。采样率是22.050kHz，精度16位，存储格式为wav格式，时间长度为300min，特征提取后共得clip序列总数18000，其中静音1584，噪音438，音乐7440，纯语音5178，带背景音的语音3360。100min数据做训练集，200min数据做测试集。分类的准确度采用分类精度(accuracy)来衡量：

$$\text{分类精度}=\frac{\text{分类正确的样本clip数}}{\text{样本clip总数}}\times\% \quad (17)$$

4.1 分类结果分析

静音和噪音采用基于规则的分类算法。实验结果如表1。

表1 静音、噪音分类结果

	正确分类	误分类	分类精度
静音	1524	141	96.21%
噪音	282	228	64.38%

从试验中可以看到，静音的识别准确率很高；而对于噪音的识别准确率不尽如人意，分析原因是不同音频类别中出现的噪音源不尽相同，所以其噪音的时频特性也不一样，采用单一的阈值判断缺乏普适性，因而试验中噪音判断的准确率低，同时误识率高，所以噪音的识别不宜于采用基于规则的方法。

使用训练集分别训练2个SVM分类器对音乐、纯语音和带背景音的语音分类，用测试集数据进行测试。试验结果如表2。

表2 不同类音频分类结果

类别	分类精度
Music/speech	94.5%
Pure-speech/speech over backgroud sound	91.2%

从试验中可以看到：支持向量机分类器的分类准确率很高，音乐/语音分类精度平均可以达到95.8%，纯语音/带背景音的语音分类精度平均达到94.6%。

4.2 平滑与分割结果分析

对clip分类之后，得到标记了音频类别的clip序列，采用前面提到的准则对分类结果进行平滑，然后合并相同类别的clip，得到分割结果。分割精度用正确的分割边界占实际分割边界的比率衡量，分割边界由clip索引号表示。试验结果如表3。

表3 平滑前后的分割试验结果

音频类别	平滑前	平滑后
静音	95.4%	96.8%
噪音	55.3%	63.7%
音乐	90.2%	93.5%
纯语音	86.6%	89.4%
带背景音语音	83.7%	86.1%

从试验结果可以看出，经过平滑处理，有效地降低了各个音频类的误分率，所以分割精度都有一定的提高，说明本文提出的3个平滑准则有效；静音的分割结果准确率很高，原因是静音的误分类率很低；而噪音的误分率过高，导致了分割精度很低；同时，由于噪音容易被误判为语音，因此纯语音和带背景音语音的分割结果受影响，分割精度相对较低；而音乐受噪音误分类影响小，所以分割精度较高。总体的分割结果，除噪音外，均比较理想，进一步说明了SVM对于音频分类与分割的有效性和适用性。

4.3 特征有效性分析

为了验证本文提出的新特征的有效性，设计如下试验：将比较常用的MFCC、子带能量比均值、带宽均值、频谱中心均值组成基本音频特征集(B set)；基音频率标准方差、和谐度、平滑基音比3类特征刻画了不同类型音频基音周期的变化特

(下转封四)

(上接第 90 页)

性,这里称为基音特征集(P set);本文提出的静音率、High ZCR 比率、Low Frequency Energy 比率、频谱流 4 个新特征组成新特征集(N set)。为了验证本文选取的特征集的有效性,去不同的特征集组合来进行实验分析,试验结果如表 4。

表 4 特征有效性分析试验结果

精度(%)	B set	B set+P set	B set+N set	B set+All
音乐/	86.3	92.7	94.4	96.3
语音		(+6.4)	(+8.1)	(+10.0)
纯语音/	89.1	93.4	91.3	94.2
非纯语音		(+4.3)	(+2.2)	(+5.1)

从上面的实验结果可以看出:基本特征集对于这三类音频的识别具有一定的有效性和准确性;加入基音周期特征集后,对语音/音乐和纯语音/非纯语音的识别精确度都有大幅度的提高,分别提高了 6.4 和 4.3 个百分点;加入新特征集后,对音乐/语音分类精确度影响明显,提高了 8.1 个百分点,原因是音乐和语音结构差别很大,这 3 个特征可以有效刻画音乐和语音的结构特征差别。同时发现,新特征对纯语音和非纯语音的分类效果不明显,精确度只提高了 2.2 个百分点。带背景音语音的频谱是语音与背景音频谱的叠加,因而其与纯语音的显著区别性特征应该是频谱流,我们做了只加入频谱流特征到基本特征集的试验,分类精度提高了 2 个百分点,这一结果验证了我们的想法;全特征集的识别精确度分别达到了 96.3%和 94.2%,说明了本文选取的特征集合理、有效。

结论 音频分类与分割是解决音频结构化问题和提取音频结构化信息和内容语义的关键,是当前基于内容的音频分析领域中一个研究热点,在音频检索、视频摘要和辅助视频分析等方面都有重要的应用价值。本文提出了一种基于 SVM 的音频分类算法,将音频分为静音、噪音、音乐、纯语音和带背景音的语音 5 类。在分类的基础上提出了三个有效的平滑准则,对分类结果进行平滑处理,最终实现对音频流按音频类别的分割。同时,分析了不同类音频的区别性特征,提出了静音率、High ZCR 比率、Low Energy 比率、频谱流 4 个新特征,综合考察了不同特征集在基于 SVM 分类器中的分类性能和精确度。试验结果表明,基于 SVM 的分类算法分类效果良好,分类精度较高;平滑处理进一步提高了分类精度,降低了误分率,分割结果比较准确。

未来的研究方向应放在两个方面,一个是噪音识别,从试验中可以看到,本文的方法对噪音的识别效果不好,如何针对不同音频环境下噪音的特性提取区别性特征来识别噪音是未来工作的一个重点;另一个方面,音频中除了语音和音乐外还有其他一些含有重要语义的音频类别,比如一些典型的环境声音,如雨声、交通工具声、枪炮声等等,未来的研究工作应考虑更多的音频类别作为分类对象。

参考文献

1 Foote J. Content-base retrieval of music and audio. In:C. C. J. Kuo,et al. eds. Multimedia Storage and Archiving Systems II, Proc. of SPIE, volume 3229,1997. 138～147
2 Foote J. An overview of audio information retrieval. ACM-Springer Multimedia Systems, 1998
3 Pfeiffer S, Fischer S, Effelsberg W. Automatic Audio Coutent Analysis. In:Proc. of the fourth ACM intl. conf. on Multimedia, 1996. 21～30
4 Saunders J. Real-time Discrimination of Broadcast Speech/ Music. In:Proc. of ICASSP96, Vol. Ⅱ, Atlanta, May, 1996. 993～996
5 Scheirer E, Slaney M. Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator. In: Proc. of ICASSP 97, vol II,1997. 1331～1334
6 Jiang Hao, Lin Tony, Zhang Hongjiang. Video segmentation with the support of audio segmentation and classification. In:Proc. of ICME'2000-IEEE Intl. Conf. on Multimedia and Expo, New York,2000,3:1507～1510
7 Zhang Tong, Kuo C J. Heuristic Approach for Generic Audio Data Segmentation and Annotation. In:Proc. of the 7 th ACM Intl. Conf. on Multimedia, Orlando,1999. 67～76
8 Srinivasan S, Petkovic D, Ponceleon D. Towards robust features for classifying audio in the cudeVideo system. In:Proc. of the 7 th ACM Intl. Conf. on Multimedia, Orlando,1999. 393～400
9 Wold E, Blum T, Keislar D, Wheaton J. Contemt -based classification, search and retrieval of audio. IEEE Multimedia Magazine,1996,3(3): 27～36
10 Liu Z, Huang J, Wang Y, Chen T. Audio feature extraction and analysis for scene classification. In:IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing
11 Vapnik V. The Nature of Statistical Learning Theory. New York:Springer-Verlag, 1995
12 Cortes C, Vapnik V. Support Vector Networks. Machine Learning, 1995,20:273～297
13 Foote J. Content-base retrieval of music and audio. In:C. C. J. Kuo,et al. eds. Multimedia Storage and Archiving Systems II, Proc. of SPIE, volume 3229, 1997. 138～147
14 Li S H. Content-based classification and retrieval of audio using the nearest featue line method. IEEE Transactions on Speech and Audio Processing, Sep. 2000
15 Moreno P J, Rifkin R. Using the Fisher Kernel Method for Web Audio Classificaion. In: Proc. of ICASSP2000, Vol. IV. June2000. 2417～2440
16 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. 计算机辅助设计与图形学学报,2003,14(3)

计算机科学

(1974 年 1 月创刊)
第 32 卷第 4 期 (月刊)
2005 年 4 月 25 日出版

ISSN 1002－137X
CN50－1075/TP

定价: 25.00 元 国外定价: 5 美元
邮发代号: 78－68
发行范围: 国内外公开

主管单位: 国家科学技术部
主办单位: 国家科技部西南信息中心
编辑出版: 《计算机科学》杂志社
重庆市渝中区胜利路 132 号 邮政编码: 400013
电话: (023) 63500828 E-mail: jsjkx@swic.ac.cn
网址: www.jsjkx.com
社 长: 牟炳林
主 编: 彭 丹
副主编: 朱宗元
主编助理: 徐书令
印刷者: 重庆科情印务有限公司
总发行处: 重庆市邮政局
订购处: 全国各地邮政局
国外总发行: 中国国际图书贸易总公司 (北京399 信箱)
国外代号: 6210—MO