DOI:10.16157/j.issn.0258-7998.2006.01.021

基于短时能量的语音/音乐快速分类

陈 功[1],王振力[1],张建兵[2]
(1.解放军理工大学 通信工程学院,江苏 南京 210007;
2.常州工学院,江苏 常州 213002)

摘 要: 针对音频信号的特点,基于语音/音乐信号的短时能量进行快速分段和分类,并给出了方法和步骤。

关键词: 短时能量 语音/音乐 分段 分类

语音和音乐是两类最重要的音频信号。语音和音乐的自动分类在基于内容的音频检索、视频的摘要以及语音识别等众多领域都有重要的应用价值。

目前,国内外语音信号识别多采用基于包括感觉特征和过零率、功率谱、MFCC 系数等多种语音/音乐的分类特征技术来实现。例如 Scheirer 和 Slaney 利用能量、平均过零率、基频和功率谱峰值等多种音频信号特征对语音和音乐信号进行分类。然而这些识别技术所需要的参数特征值较多、过程复杂,不具备实时性,而且有的技术识别可靠性较差。所以上述方法在实际应用中存在一定的局限性。

因此,对利用少量音频数据特征值准确并快速地进行语音和音乐的分类并用来提取音频内容语义和结构的研究已日益引起人们的重视。本文通过对短时能量建模分析,实现了音频信号的分段,并结合短时能量方差值实现语音音乐信号的分类。

1 短时能量建模分析

选取一段无停顿的语音和音乐信号作为待分段的音频信号。在 5~20ms 的时间间隔内可以认为音频信号特征基本保持不变。采用短时能量均方根的概率统计方法来提取语音和音乐信号的特征。图 1 分别为语音和音乐信号短时能量均方根(RMS)的时域波形。其采样频率均为 11 025Hz,矩形窗长度 N 为 10ms,时间长度为 30s。

$$\mathrm{RMS}=\sqrt{\sum_{n=1}^{N}x^2(n)} \qquad (1)$$

式(1)中,$x(n)$为音频信号,矩形窗序列沿音频样点序列逐帧移动,每段帧长度为 N。

30s 的 RMS 的概率分布,即信号分布频数直方图如图 1 所示。由图 1 可知两分布有较明显的差异,可以作为识别语音和音乐信号的特征依据。进一步研究发现,其概率分布服从不同参数时的广义 x^2 分布。概率分布的表达式如下:

$$p(x)=\frac{x^a\exp(-bx)}{b^{a+1}\Gamma(a+1)} \qquad x\geq 0 \qquad (2)$$

参数 $a=(\frac{\mu}{\sigma})^2-1$,$b=\frac{\sigma^2}{\mu}$,其中 μ、σ^2 分别为 RMS 概率分布的均值和方差。

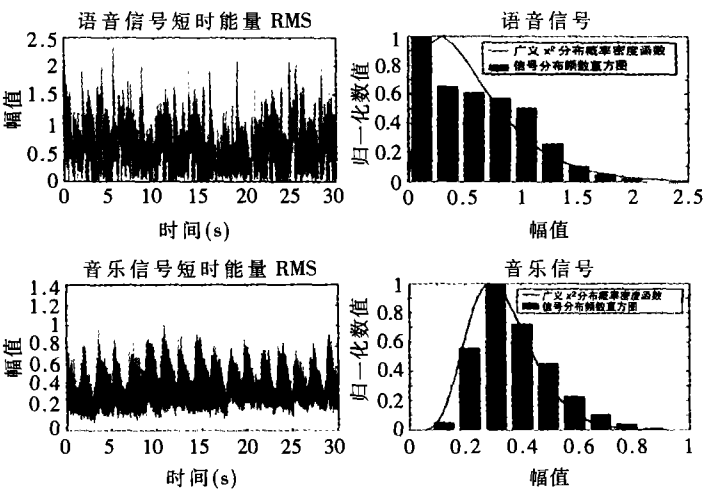

图 1 30s 语音/音乐 RMS 和相应频数直方图

2 语音/音乐信号的分段

在基于 RMS 概率分布实现语音/音乐信号的实时分段中,取总窗长度为 1s、间隔时间为 10ms 的音频信号作为 RMS 建模的时间长度。分别在每 1s 的窗长度时间内计算 RMS 的均值和方差。分段算法的实现分两个时间段:一是将音频信号的分段精度精确到 1s;二是在音频信号变化的有限时间范围内,将分段的时间精度精确到 10ms。

(1)第一阶段的实现过程为:基于 RMS 分析所需检测该秒时间的前一秒和后一秒的信号分类。首先定义相关函数为:

$$\rho(p_1,p_2)=\int\sqrt{p_1(x)p_2(x)}\,\mathrm{d}x \qquad (3)$$

由广义 x^2 分布所表征的 RMS 的概率分布,其相关函数为:

$$\rho(p_1,p_2)=\frac{\Gamma\left(\frac{a_1+a_2}{2}+1\right)2^{\frac{a_1+a_2}{2}+1}\,b_1^{\frac{a_2+1}{2}}\,b_2^{\frac{a_1+1}{2}}}{\sqrt{\Gamma(a_1+1)\Gamma(a_2+1)(b_1+b_2)^{\frac{a_1+a_2}{2}+1}}} \quad (4)$$

定义衡量音频信号变化的距离为：

$$D(i)=1-\rho(p_{i-1},p_{i+1}) \quad (5)$$

在第 i 时间窗内，由于第 $i-1$ 和 $i+1$ 时间窗的音频信号类型发生变化，此时相关函数近似为 0，变化距离则趋于 1；当音频信号无变化时，相关函数趋于 1，变化距离比较小。

考虑到音频信号的时变性，文献[6]给出了滤波及归一化的距离公式：

$$D_n(i)=\frac{D(i)V(i)}{D_M(i)} \quad (6)$$

式中 $V(i)$的作用是滤波。若 $D(i)>D(i-1)$，则 $V(i)=D(i)-D(i-1)$，反之 $V(i)=0$。式中 $D_M(i)$是距离的归一化实现，定义为 $D_M(i)=\max(D(i-1),D(i+1))$。

由于语音信号存在单词和音节的停顿，而音乐信号一般呈现连续性，所以语音信号的方差值大于音乐信号，这也就给上述分段方法赋予了物理意义。

图 2 给出语音段和音乐段的分段结果。由图 2 可知，由于音频信号的时变性，$D(i)$和 $D_M(i)$的分段存在个别时间点的误差，但 $D(i)$和 $D_M(i)$的分段基本与实际分段吻合。而且经滤波和归一后，$D_M(i)$的分段效果比用 $D(i)$分段要明显。

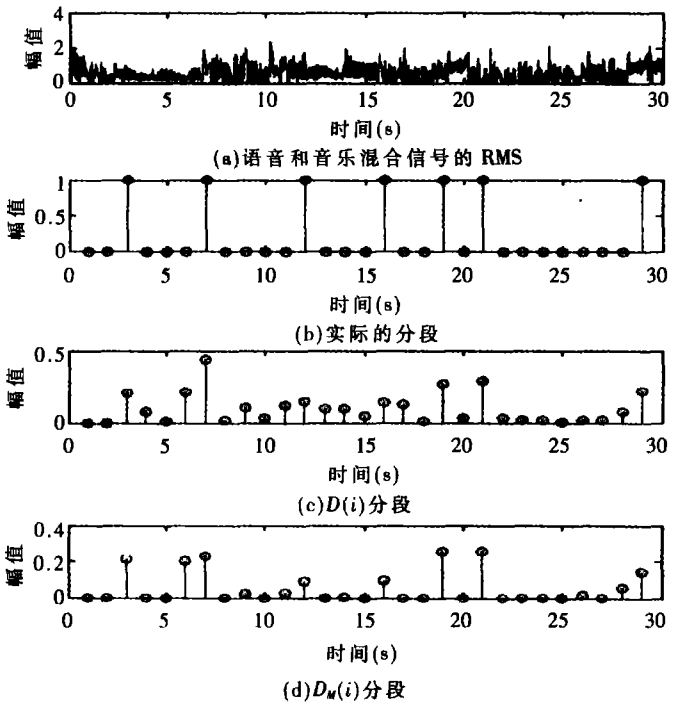

图 2 语音段和音乐段分段图

(2)第二阶段的实现过程为：基于 RMS 分析信号改变的有限时间内连续时间段的分类。该过程的分析时间间隔为 10ms。时间段 5.5~8.5s 分段结果如图 3 所示。

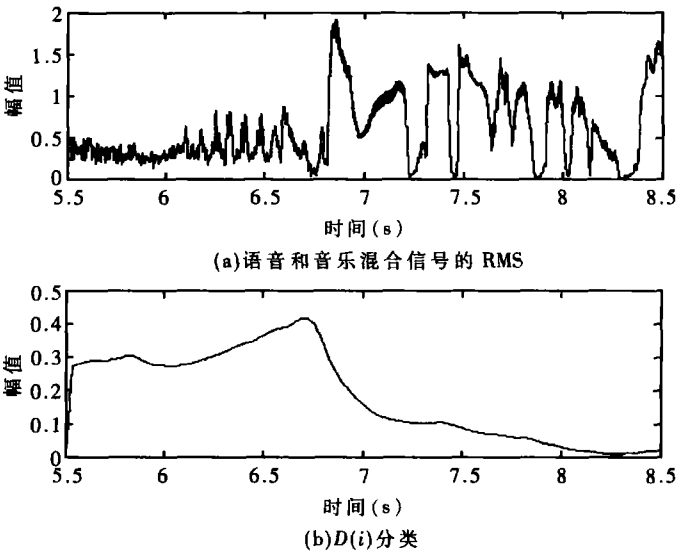

图 3 5.5~8.5s 间隔内语音段和音乐段分类图

3 语音/音乐信号的分类

3.1 标准 RMS 方差值

标准 RMS 方差值可以用来提取具有单词、音节停顿特征的语音信号，并对音频信号进行分类。标准 RMS 方差为：

$$\sigma_A^2=\sigma^2/\mu^2 \quad (7)$$

参数意义同式(2)。

图 4 为一组典型的时间长度为 1s 的语音和音乐信号的标准 RMS 方差值的信号分布频数直方图。注意到概率分布均是非交叠的，因而利用该特征可以准确地区分两类信号。在分类实现时，设定方差阈值为 0.21。在本组数据中，大于阈值的语音信号标准 RMS 方差所占比例为 95.42%，而小于阈值的音乐信号标准 RMS 方差占 98.94%。通过与阈值比较可实现两种信号的识别。

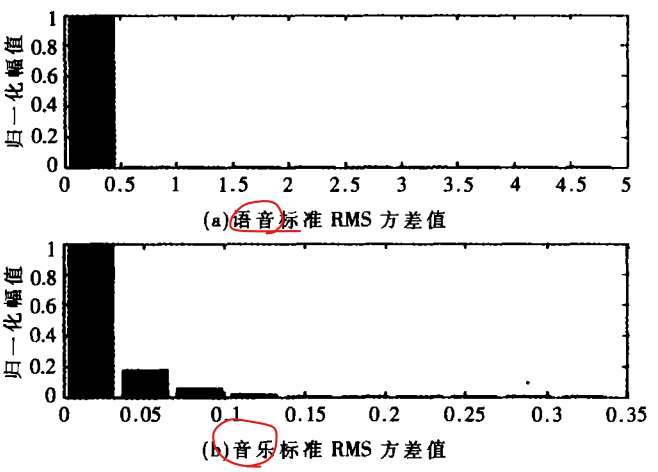

图 4 语音/音乐信号标准 RMS 方差值

3.2 分类结果

上述方法可以用来对含有语音段和音乐段的音频信号分类。考虑到音频信号分类的精度和准确性，选取

1s 的时间长度提取特征值，实现分类过程。

分类步骤为：设定阈值 0.21，若方差值大于阈值，则判断为语音信号，否则判断为音乐信号。

图 5 是一组语音段/音乐段 $D_M(i)$分段及 RMS 分类图。事实上，由于音频信号的时变性，个别时间段语音/音乐信号的短时 RMS 特征具有相似性，并且考虑到特征提取的时间有限(1s)，因此造成图 5(b)中第 6 秒的虚警和图 5(c)中 RMS 分段第 19 秒的错误分类。图 6 为该组语音段和音乐段实际分类图及 RMS 分类图。图 6 所给的该音频信号分类的准确率(定义为相同时间长度下 RMS 正确分类数与实际分类数的比值)为 26/30≈86.7%。

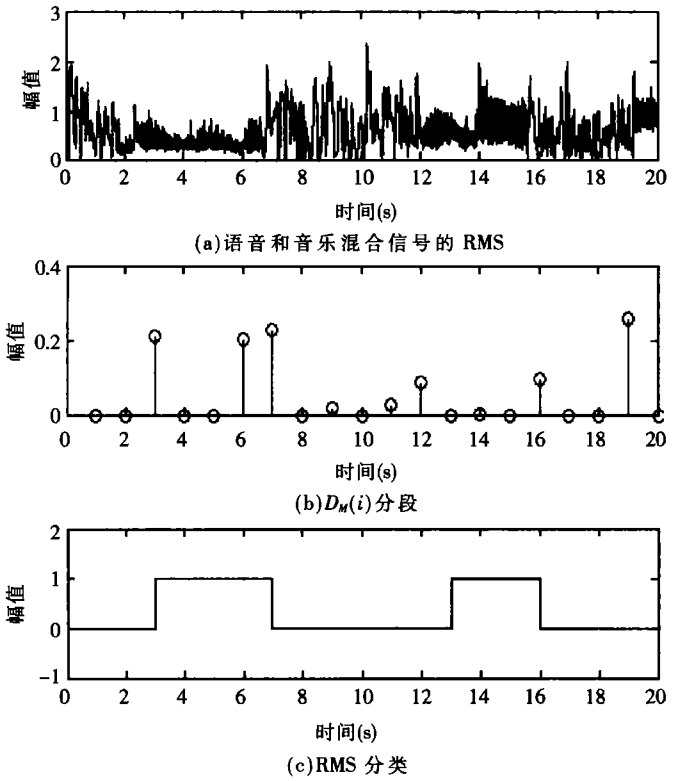

图 5 语音/音乐段分段、分类图

由图 7 所给的另一组音频信号分类图可知，其语音/音乐段分类的准确率为 28/30≈93.3%。从上述各结果可知，分类准确率的指标已达到音频检索、视频的摘要等应用领域的要求。因此，基于短时能量的语音/音乐段的快速分类是可行的。

本文给出了一种基于音频信号短时能量特征的语音和音乐的快速分类方法。该方法对含有独立语音音乐段的音频信号进行了分段和分类。仿真结果表明：利用少量音频数据特征值，可以较准确并快速地进行语音和音乐的分类。所以该分类方法具有一定的可行性和实时性。

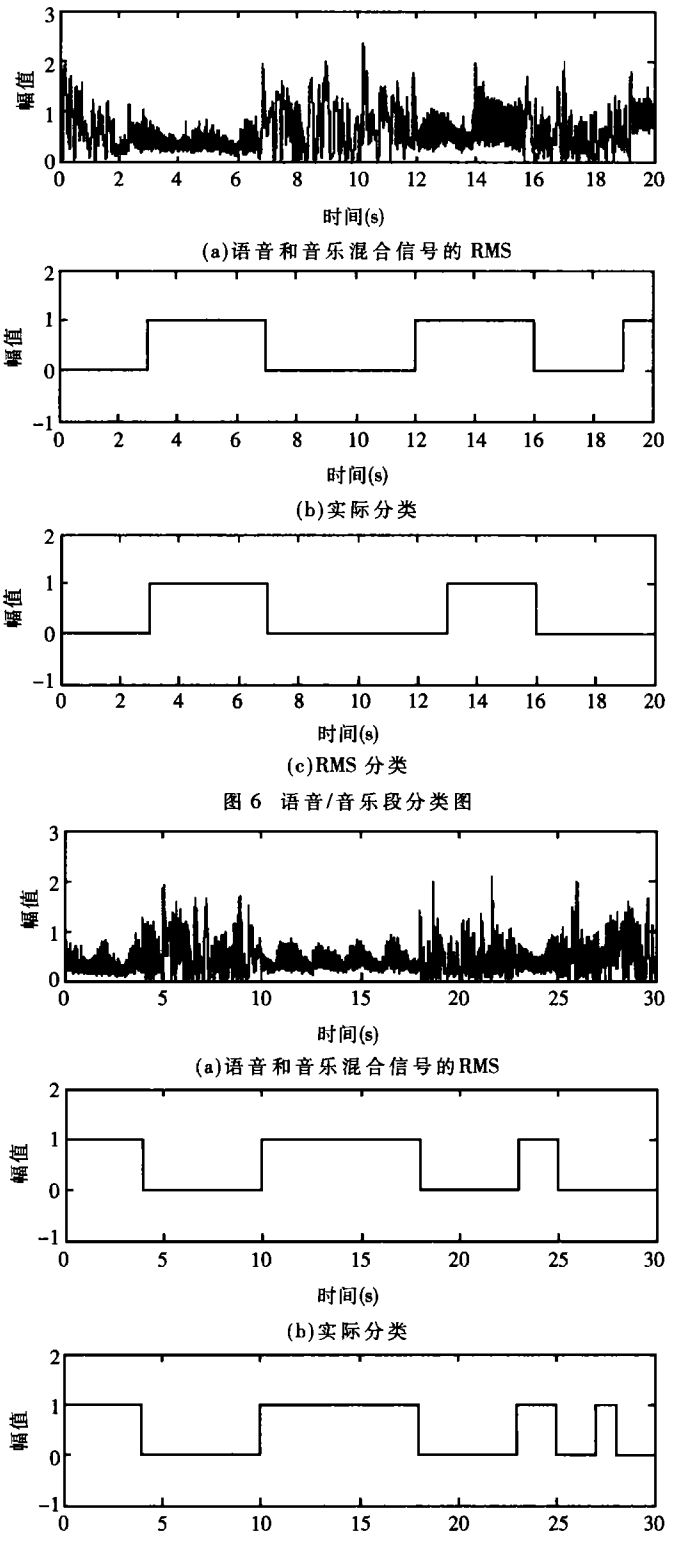

图 6 语音/音乐段分类图

图 7 语音/音乐段分类图

参考文献

1 Saunders J. Real-time discrimination of broadcast speech/music. In: Proc. IEEE ICASSP, 1996

2 Scheier E, Slaney M. Construction and evaluation of a robust multifeature speech/music discriminator. In: Proc. IEEE ICASSP, 1997

(下转第 63 页)

```
        pTimeStart, IN REFERENCE_TIME * pTimeEnd)
{if(pTimeStart==NULL) return S_OK;
   REFERENCE_TIME m_rtStart= * pTimeStart;
   if(pTimeEnd ! =NULL) REFERENCE_TIME m_rtEnd=
                                    ( * pTimeEnd);
}
```

4.4 运行测试

在网络传输速率为150KBps时，即使数据传输量较大，丢包率低于2%，时延低于250ms，音视频的连续性和同步仍较好。当传输速率为80KBps左右时，数据丢包率偶尔高于5%，视频停止发送，随后，当丢包率低于2%时，视频重新发送。而丢包率介于2%~5%时，视频非关键帧被停止传送，传输时延为250ms左右，音频效果较好，视频则随丢包率不同而有所变化。发送端能根据接收终端反映的实时接收情况自动调整发送速率，效果令人满意。

4.5 编写过滤器应注意的问题

(1)锁

DirectShow应用程序至少包含2条线程：主线程和数据传输线程。既然是多线程，肯定需要解决线程同步的问题。过滤器有2种锁：过滤器对象锁和数据流锁。过滤器对象锁用于过滤器级别的如过滤器状态转换、BeginFlush、EndFlush等；数据流锁用于数据处理线程内，如Receive、EndOfStream等。如果没有搞清楚这2种锁，很容易产生程序的死锁。

(2)EndOfStream

当过滤器接收到这个“消息”时，意味着上一级过滤器的数据都已经发送完毕。此后，如果Receive再有数据接收，也不应该去理睬它。如果过滤器对输入针上的数据进行了缓存，在接收到EndOfStream后应确保所有缓存的数据都已经处理过了才能返回。

(3)Media Seeking

一般情况下，只需要在过滤器的输出针上实现NonDelegatingQueryInterface方法，当用户申请得到IID_IMediaPosition接口或IID_IMediaSeeking接口时，将请求往上一级过滤器的输出针上传递。当过滤器图进行MediaSeeking时，一般会调用过滤器上的BeginFlush、EndFlush和NewSegment。如果过滤器对数据进行了缓存，就要重载它们，并做出相应处理。如果过滤器负责给发送出去的样本打时间戳，则在MediaSeeking之后应该重新从零开始打时戳。

(4)使用专门的线程

如果使用了专门的线程进行数据的处理和发送，就需要特别小心，不要让线程进入死循环，并且要让线程处理函数能够实时检查线程命令。应该确保在过滤器结束工作时，线程也能正常地结束。但有时把GraphEdit程序关掉后GraphEdit进程仍在内存中。这是因为数据线程没有安全关闭。

(5)从媒体类型中获取信息

若要在输入针连接的媒体类型中获取视频图像的宽、高等信息，则应该在输入针的CompleteConnect方法中实现，而不要在SetMediaType中。

本文介绍了DirectShow的基本原理，针对多媒体数据的网络实时传输和回放，研究了开发发送和接收过滤器的基本方法，并应用视频关键帧和时间戳技术进行了相应的传输服务质量控制。这些方法可应用于多媒体网络系统中，如视频会议系统等，实践证明这些方法是行之有效的。此外，根据当前接收端RTCP报文的丢包数反馈，在应用程序中还可通过发送和接收缓冲区的动态设置以及对发送速率的动态调整等技术来缓解网络丢包和时延抖动的情况，提高音视频数据传送的服务质量。

参考文献

1 Schulzrinne H, Casner S, Frederick R et al. RTP: A Transport Protocol for Real-Time Applications. IETF RFC 1889, 1996
2 陆其明. DirectShow开发指南. 北京：清华大学出版社，2003
3 朱多智，金天，卢剑. 视频会议系统中基于Mpeg4视频流的带宽控制. 华中科技大学学报(自然科学版)，2003；31(10)
4 刘浩，胡栋. 基于RTP/RTCP协议的IP视频系统设计与实现. 计算机应用研究，2002；19(10)
5 董科军，阎保平：流媒体传输的质量控制技术研究. 微电子学与计算机，2003；20(5)

(收稿日期：2005-07-07)

(上接第55页)

3 Zhang T, Kuo J. Audio content analysis for on-line audio visual data segmentation and classification. IEEE Trans. Speech Audio Process, 2001；9(5)
4 Panagiotakis C, Tziritas G. A Speech/Music Discriminator Based on RMS and Zero-Crossings. IEEE Transactions on Multimedia, 2005；7(2)
5 卢坚，陈毅松. 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，2002；14(3)
6 Young T, Fu K-S. Handbook of Pattern Recognition and Image Processing. Eds, Academic, New York, 1986
7 Wold E, Blum T, Keislar D et al. Content-based classification, search, and retrieval of audio. IEEE Multimedia Mag, 1996；3

(收稿日期：2005-07-18)