一种音乐情绪参数化的方法*

吴忻生[1,2] 徐凯春[1] 戚其丰[1†] 高红霞[1,2]

（1 华南理工大学自动化科学与工程学院 广州 510640）
（2 华南理工大学精密电子制造装备教育部工程研究中心 广州 510640）

摘要 针对目前基于情绪的音乐分类研究存在的弊端，为了方便音乐检索，本文提出一种音乐情绪参数化的方法。该方法通过提取反映音乐情绪的特征向量，然后利用 fisher 算法进行维数压缩，再通过大量的音乐样本训练得到节奏、音调和音色 3 个描述音乐情绪的参数，参数的大小反映了情绪的强弱。实验结果表明，音乐情绪参数化的结果符合音乐实际的情绪。

关键词 音乐检索，音乐情绪，参数化，fisher 算法

中图分类号：TN912 **文献标识码**：A **文章编号**：1000-310X(2013)01-0028-06

A parameterized method of musical emotion

WU Xinsheng[1,2] XU Kaichun[1] QI Qifeng[1] GAO Hongxia[1,2]

(1 *College of Automation Science and Engineering, South China University of Technology, Guangzhou* 510640, *China*)
(2 *Engineering Research Center of Precision Electronic Manufacturing Equipments, Ministry of Education, South China University of Technology, Guangzhou* 510640, *China*)

Abstract Considering the disadvantages of music classification researches based on musical emotion, this paper proposes a parameterized method of musical emotion for music retrieval. This method extracts the feature vector of music, then compresses dimensions using the fisher algorithm, finally obtains rhythm, tone and timbre three parameters by training to describe musical emotion. The bigger parameter reflects the stronger musical emotion. Experimental results show that the parameterized result corresponds with the real emotion of music.

Key words Musical retrieve, Musical emotion, Parameterized, Fisher algorithm

1 引言

近年来，随着个人计算机的广泛普及，互联网技术和信息产业的不断发展，音乐市场的持续繁荣，人们有可能拥有越来越多的音乐文件，对大量音乐文件进行管理是一个令人头痛的问题。传统途

2012-05-21 收稿; 2012-09-21 定稿

*中央高校基本科研业务费专项资金(2012ZZ0107)

作者简介：吴忻生 (1961-), 男, 浙江省建德县人, 副教授, 研究方向为智能检测与智能控制、自动化技术和智能系统的研究及其工程应用等。
徐凯春 (1986-), 男, 硕士研究生。
戚其丰 (1976-), 男, 讲师。
高红霞 (1975-), 女, 副教授。

†通讯作者：戚其丰, E-mail：qqfeng@gmail.com

径通过歌曲家、唱片名、歌曲名等信息来对音乐进行分类管理，这种方式简单有效，但是当文件数量很大的时候，就很难管理。另一方面，有时候人们需要根据个人情绪寻找合适的音乐，比如，人们在睡觉前，需要听一些比较低缓的促进睡眠的音乐；当人们参加一些聚会的时候，就需要一些快节奏的、令人兴奋的音乐来活跃气氛。因此，基于内容的音乐分类研究具有很好的现实意义。

20 世纪 90 年代国际上开始了基于内容的音频检索技术研究，到 90 年代末达到高潮。在音乐分类方面取得不少的成果，Wold 等人最先提出利用音乐样本进行频域处理抽取一些特征，然后对音乐进行分类(Muscle Fish 方法)[1]；Foote 根据人耳听觉特性，从音乐中提取 12 阶的 MFCC 特征参数，并使用人工神经网络对音乐进行分类[2]；Li 在 MFCC 基础上，选择了最近特征线(Nearest Feature Line)方法设计分类器[3]；Lin 等把小波变换理论应用到音乐特征提取，并使用 SVM 方法设计分类器，分类正确率得到很大提高[4]；Munoz 等通过实验发现，混合高斯模型与遗传模糊系统的搭配使用，分类效果比混合高斯模型要好很多[5]。

以上的方法通过特征提取和分类算法将音乐分成几个特定的类别，有利于音乐的检索。但很多时候人们想通过音乐情绪中的某些参数大小来挑选合适的歌曲，比如节奏快慢，音调高低等，以上的方法就显得不够灵活，因此，本文提出了一种音乐情绪参数化的方法，首先从音乐中提取合适的时域和频域特征参数，通过 fisher 线性判别算法，将高维的特征向量映射到一维空间，形成音乐的节奏、音调和音色三个参数，参数的大小表示情绪的强弱。这样，人们就可以通过调整这三个参数的大小来管理和检索歌曲。系统的总体框架如图 1 所示。

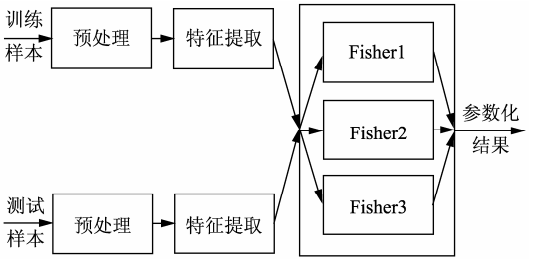

图 1　系统总体框架图

2　音乐特征提取

为了实现每一首歌曲的情绪参数化，首先要做的工作是从歌曲中提取一些能够反映音乐情绪的特诊参数。特征提取的对象为完整的歌曲，平均一首歌是 4~5 分钟。从每一首歌中提取出 124 维的特征向量。本文用于训练和测试的音乐都是从 google music 网站下载，MP3 格式，位速为 192 kbps，采样率为 44.1 kHz，16 bit/sample，2 个声道的立体声。

2.1　音乐预处理

音乐一般是以 MP3 格式存储在个人电脑或者网络服务器，因此，在进行特征提取之前需要对音乐文件进行预处理。针对 MP3 文件，首先对音乐文件进行 MP3 解码，得到音乐采样样本 2 个通道的数据流。由于本文不处理立体声效果，为了方便处理，将 2 个通道的数据求平均值，得到了单声道的数据流。

接着对数据流进行分帧，每一帧长为 512 个采样点（11.6 ms），每一帧都进行加汉明窗处理，然后再做短时傅里叶变换（STFT），得到音乐的频谱信号。将频谱信号送到特征提取模块进行特征提取。预处理的流程如图 2 所示。

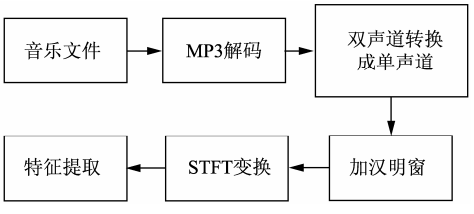

图 2　预处理框架图

2.2　音乐帧级特征提取

本文提取的音乐特征包括时域特征（Time Domain Feature）和频域特征（Spectral Feature）两类。时域特征是过零率（预处理过程不需要加窗和 STFT 变换）；频域特征包括：(1)频谱中心；(2)频谱衰减值；(3)频谱流量；(4) MFCC 倒谱系数；(5)音节系数。

具体的特征介绍如下：

时域过零率（Time Domain Zero Crossings）：对于时域的离散信号，前后两个采样点数据的代数符号不一致，称之为过零。过零的速率称为过零率，具体计算如下：

$$Z_t = \frac{1}{2(N-1)}\sum_{m=1}^{N-1}\left|\mathrm{sgn}\left[x(m+1)\right]-\mathrm{sgn}\left[x(m)\right]\right|, \quad (1)$$

其中，$x(m)$ 为音乐的时域离散信号。人类的发音是由清音和浊音组成，而纯音乐不会有这种现象，因此语音的过零率要高于纯音乐的过零率[6]。在包含语音的音乐中，过零率在一定程度上反映了歌唱者的语速，语速越快，语音越密集，过零率越高。

频谱中心（Spectral Centroid）：频谱中心定义为，音频信号通过短时傅里叶变化（short-time Fourier transform，STFT）得到的频谱的中心[7]。

$$C_t = \frac{\sum_{n=1}^{N} M_t(n)\times n}{\sum_{n=1}^{N} M_t(n)}, \quad (2)$$

其中，$M_t(n)$ 是 t 时刻音频帧的傅里叶变换。频谱中心是频谱形状的一个指标，频谱中心值越大，表明音乐信号具有更多的高频成分。

频谱衰减值（Spectral Rolloff）：频谱衰减值是指到小于频率 R_t 的频谱能量，刚好等于总频谱能量的 85%[7]。

$$\mathrm{Rolloff} = \sum_{n=1}^{R_t} M_t(n) = 0.85\sum_{n=1}^{N} M_t(n), \quad (3)$$

频谱衰减值反映频谱的形状。

频谱流量（Spectral Flux）：频谱流量定义为连续的频谱分布的标准差[7]。

$$F_t = \sum_{n=1}^{N}(N_t[n]\text{-}N_{t\text{-}1}[n])^2, \quad (4)$$

其中，$N_t[n]$ 和 $N_{t\text{-}1}[n]$ 分别是 t 时刻和 t-1 时刻音频帧的傅里叶变换。这一个指标表示局部频谱变化的总量。在一定程度上反映了频谱的形状。

Mel 倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）：前人通过研究发现，人耳对于声音频率的分辨不是完全线性的，在 1000 Hz 以下为线性的，在 1000 Hz 以上接近对数关系[8]。

$$\mathrm{Mel}(f) = 2595\log(1+\frac{f}{700}). \quad (5)$$

经过该公式转换后的频谱称为倒谱域。Mel 倒谱系数是音频信号映射到倒谱域后，提取出来的一组反映音频特征的参数，计算过程是：对音频信号进行 DFT 变换，再将信号映射到倒谱域，接着用 13 个等带宽的三角滤波器滤波，将各个滤波器的输出能量进行累计。最后通过以下公式计算 MFCC 系数：

$$C_n = \sqrt{\frac{2}{K}}\sum_{k=1}^{K}\{\lg[x(k)]\times\cos\frac{n(k-0.5)}{K}\},$$

（n =1,2,…,13） (6)

其中，$x[k]$ 为第 k 个滤波器的输出能量。

音节系数（Chroma）：根据乐理知识，把一个纯八度十二等分，并将每一等分之间的音高距离定为半音，这种方式称为十二平均律。一些最新的研究表明，音乐结构分析问题更多的是取决于音节系数，Wakefield 及 Goto 采用音节系数来描述 12 个同性质的音高类之间的功率谱能量分布[9]。笔者依据该方法，从对数频谱尺度上的功率谱中提取 12 维的音节系数，$v_c(t)$ 表示某个八度上的第 c 个（c=1,2…,12）个音高类：

$$v_c(t) = \sum_{Oct_L}^{Oct_H}\int_{-\infty}^{+\infty} B_{BPF_{c,h}}(f)\psi(f,t)\mathrm{d}f, \quad (7)$$

其中，$\psi(f,t)$ 表示时刻 t 上的对数尺度频率 f，通过短时傅里叶变换（STFT）得到；Oct_L 和 Oct_H 表示八度音的范围，分别为 3 和 8，两者覆盖的频率范围为 130 Hz~8000 Hz；$B_{BPF_{c,h}}(f)$ 是一个汉宁窗带通滤波器，允许对数尺度频率 $F_{c,h}$ 通过。

$$B_{BPF_{c,h}}(f) = \frac{1}{2}(1-\cos\frac{2\pi(f-(F_{c,h}(f)-100))}{200}), \quad (8)$$

$F_{c,h}$ 表示八度位置为 h 的第 c 个音高类上的对数尺度频率，表达式为：

$$F_{c,h} = 1200h + 100(c-1). \quad (9)$$

在得到十二个音节系数以后，为了反映出音节最大值与均值、最小值的差别大小，定义峰值-均值比例和峰值-谷值比例如下：

$$R_{\text{max-avg}} = \frac{v_{\max}}{v_{avg}}, \quad (10)$$

$$R_{\text{max-min}} = \frac{v_{\max}}{v_{\min}}, \tag{11}$$

其中，$v_{\max}$、$v_{\min}$ 和 v_{avg} 为十二个音节系数中的最大值、最小值和平均值，最终得到 14 维的音节系数。

2.3 特征向量的组成

根据 2.2 对每一帧音乐进行特征提取，可以得到 31 维的帧级特征（Frame Feature），具体如表 1 所示。

表 1　帧级特征向量的组成

	帧级特征名称	特征维数
时域	Zero Crossings	1
频域	Spectral Centroid	1
	Spectral Rolloff	1
	Spectral Flux	1
	MFCC	13
	Chroma	14

帧级特征只反映了音乐帧级（11.6 ms）的特征形态，无法得到一小段音乐的特征。因此，本文提出音乐段特征（Clip Feature）的概念，段特征的计算过程如下：对连续的 40 个音乐帧进行帧级特征提取，然后通过计算平均值和标准差来得到段特征。

从段特征中可以得到音乐在 464 ms 时间内的特征，为了提取更长时间的特征，本文又提出了音乐镜头特征（Shot Feature）的概念，镜头特征的计算过程如下：在原来段特征的基础上，对 5 个连续的音乐段进行求均值和标准差，得到镜头特征（2.32 s）。

最后，我们可以从 31 维的帧级特征拓展到 124 维的镜头特征，组成本文的音乐特征向量，如图 3 所示。

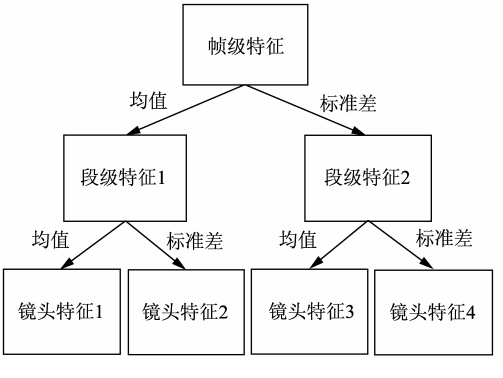

图 3　特征拓展示意图

3　Fisher 线性判别原理

根据上一节的特征提取方法，对每一首歌进行特征提取，可以得到 124 维的特征向量。接着如何用参数大小的方法来表示音乐的特征，比如音乐的节奏、音调、音色等。因此，必须进行维数压缩，笔记采用 fisher 算法解决这个问题。

fisher 算法是一种成熟的线性分类算法，在各个领域有广泛的应用，该算法实际上就是把高维的特征向量空间压缩到一维。fisher 算法是找一个投影方向，使得投影后的点类内离散度最小而类间离散度最大。

设有 w_1 和 w_2 两个类，n 个 m 维的训练样本 $x_k(k=1,2,\dots,n)$，其中 $n_1 \in w_1$，$n_2 \in w_2$，$n_1+n_2=n$。令线性变换矩阵 $\boldsymbol{W} \in R^{n\times m}$，得 $y_k \in \boldsymbol{W}^{\mathrm{T}} x_k$，当 $\boldsymbol{W}$ 的值使得类间离散度与类内离散度的比值最大时，分类效果达到最优。

w_1 和 w_2 在 m 维特征空间里的样本均值向量：

$$\boldsymbol{M_i} = \frac{1}{n_i}\sum_{x_k \in X_i} x_k \text{.（}i\text{=1,2）} \tag{12}$$

通过变换 $\boldsymbol{W}$ 映射到一维特征空间后，平均值分别为：

$$m_i = \frac{1}{n_i}\sum_{y_k \in Y_i} y_k, \text{（}i\text{=1,2）} \tag{13}$$

映射后，定义类内离散度矩阵为：

$$\boldsymbol{S_i^2} = \sum_{y_k \in Y_i} (y_k - m_i)^2. \text{（}i\text{=1,2）} \tag{14}$$

为了使 2 个类的平均值之间的距离越大越好，而类内离散度越小越好，定义 fisher 准则函数：

$$J_F(\boldsymbol{W}) = \frac{|m_1 - m_2|^2}{\boldsymbol{S_1^2} + \boldsymbol{S_2^2}}. \tag{15}$$

用拉格朗日乘子法求解 $J_F(\boldsymbol{W})$ 的极大值点得到最优解：$\boldsymbol{W}^* = \boldsymbol{S}_w^{-1}(M_1 - M_2)$，其中 $\boldsymbol{S}_w$ 为类内离散度矩阵的总和，即 $\boldsymbol{S}_w = \boldsymbol{S_1^2} + \boldsymbol{S_2^2}$。

音乐特征向量通过 $\boldsymbol{W}$ 变换就可以得到一个一维的参数，根据训练样本的不同，可以得到不同音乐情绪的参数大小。笔者选择了“舒缓-强烈”、“低沉-高亢”和“丰富-单纯”三组样本，分别训练出节奏、音调、音色三种音乐情绪参数。最后做归一

化处理，使参数范围为[0,1]。

4 实验与分析

4.1 系统训练

实验开始前，从 google music 网站下载大量音乐训练样本并手动做好分类，如表 2 所示。节奏的划分根据歌曲的节奏快慢和明显程度；音调的划分根据歌手的声调；音色的划分根据演奏的乐器多少丰富程度。

表 2 音乐训练样本分类

划分方式	负样本数	正样本数
节奏	舒缓（190）	强烈（192）
音调	低沉（193）	高亢（194）
音色	丰富（192）	单纯（196）

(1) 取第一组样本“舒缓-强烈”对 fisher1 算法结构进行训练，得到最佳变换矩阵 $\boldsymbol{W}_1$；

(2) 取第二组样本“低沉-高亢”对 fisher2 算法结构进行训练，得到最佳变换矩阵 $\boldsymbol{W}_2$；

(3) 取第三组样本“丰富-单纯”对 fisher3 算法结构进行训练，得到最佳变换矩阵 $\boldsymbol{W}_3$；

三组样本分别经过矩阵 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 和 $\boldsymbol{W}_3$ 变换后的直方图分布情况如图 4~6 所示。

分析以上的直方图可以看出，尽管中间有部分重叠，属于同一训练组的 2 个类别音乐经过 fisher 算法映射后，形成了 2 个很明显的波峰，代表了音乐情绪的强弱。因此，本文提出的方法是有效的。至此，得到三个变换矩阵 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 和 $\boldsymbol{W}_3$，训练过程结束。

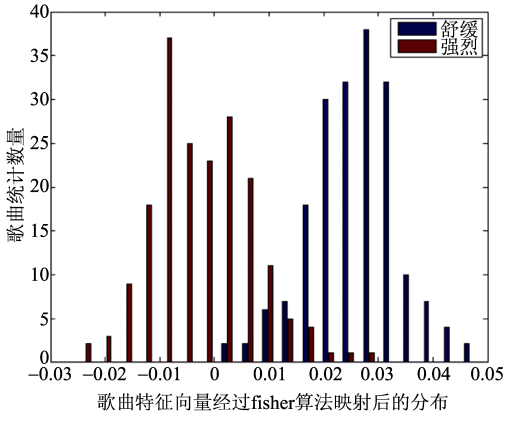

图 4 “舒缓-强烈”分类结果直方图

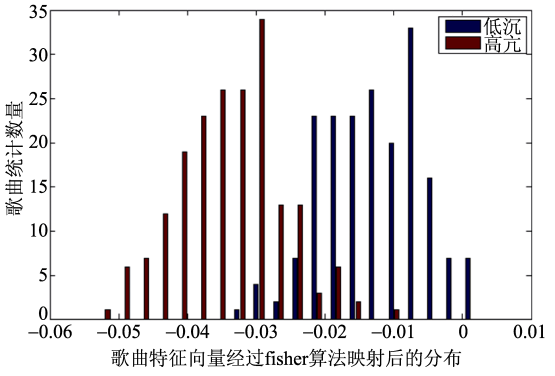

图 5 “低沉-高亢”分类结果直方图

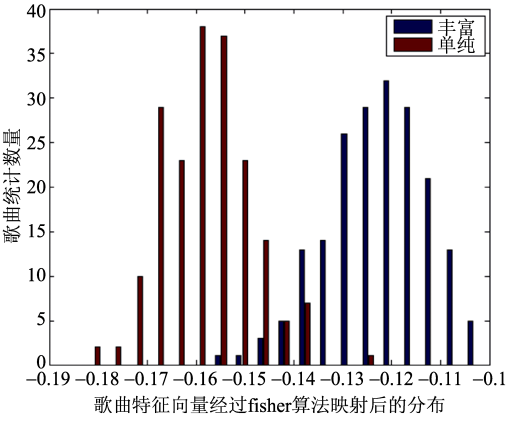

图 6 “丰富-单纯”分类结果直方图

4.2 系统测试

为了对系统进行测试，从 google music 网站下载不同于训练歌曲的各种情绪歌曲，每种情绪都是 100 首歌。然后，将所有测试音乐输入到系统进行参数化，分别统计参数化结果，如表 3 所示。

表 3 歌曲测试样本参数化结果

歌曲情绪	参数化结果统计		正确率（%）
	[0,0.5)	[0.5,1]	
强烈	96	4	96
舒缓	5	95	95
低沉	4	96	96
高亢	94	6	94
丰富	3	97	97
单纯	96	4	96

从以上测试结果可以看出，绝大部分歌曲情绪的识别是正确的，少量的歌曲情绪识别出错了。一方面是因为训练样本还不够充足，导致训练结果有一定的误差；另一方面跟歌曲本身有关，同一首歌

的不同位置情绪不一致，影响了参数化的结果。总的来看，不同情绪歌曲的识别率可以达到 94%以上，满足了实际应用需求，本文提出的音乐情绪参数化方法是有效的。

4.3 特征相对作用实验

为了反映本文提取的特征对分类结果的相对作用，笔者做了特征相对作用实验。根据表 1 的特征向量，按顺序分别去除其中的一个特征，然后按照以上的训练方法对系统重新进行训练，接着使用未参加训练的不同情绪歌曲进行测试（每种情绪 100 歌曲），最后统计得到分类的正确率如表 4 所示。

表 4　分别去除一个特征后的分类正确率（%）

	Zero Crossings	Spectral Centroid	Spectral Rolloff	Spectral Flux	MFCC	Chroma
强烈	94	96	95	95	89	90
舒缓	95	93	96	95	90	90
低沉	95	94	92	93	84	89
高亢	93	92	91	92	85	91
丰富	93	94	97	95	88	91
单纯	95	93	95	94	90	92

从以上表格可以看出，缺少 MFCC 和 Chroma 两个特征对分类结果的影响最大，特别是对于低沉-高亢类别；缺少其他特征也会使分类的正确率有所降低。这说明了本文所提取的特征，都对音乐情绪的正确分类有贡献。

5 结论与展望

本文提出了一种音乐情绪参数化的方法，通过提取音乐有用的时域和频域特征，组成音乐情绪特征向量，然后通过 fisher 算法训练出 3 个变换矩阵，音乐特征经过变换矩阵运算，映射到一维空间，再经过归一化处理，形成节奏、音调和音色 3 个特征参数。实验结果表明，虽然对某些歌曲存在一定的误差，本文提出的方法能够用参数化的方法来描述音乐的情绪。

后续工作包括：(1)对音乐时域和频域特征进行研究，提取更能反映音乐情绪的特征；(2)使用混合高斯模型或者人工神经网络等高级算法来实现音乐特征的维数压缩；(3)音乐情绪的分段识别和参数化。

参 考 文 献

[1] WOLD E, BLUM T, KESLAR D, et al. Content-based classification, search, and retrieval of audio[J]. IEEE Multimedia. Fall, 1996: 27-36.

[2] FOOTE J. Content-based retrieval of music and audio[J]. Multimedia Storage Archiving Syst, 1997, II. Vol. 3229: 138-147.

[3] LI S. Content-based classification and retrieval of audio using the nearest feature line method[J]. IEEE Trans. Speech Audio Processing, Sept. 2000, Vol. 8: 619-625.

[4] LIN C C, CHEN S H, TRUONG T K, et al. Audio classification and categorization based on wavelets and support vector machine[J]. IEEE Trans. Speech Audio Processing, Vol. 13.Sept. 2005: 644-651.

[5] MOLLA K I, HIROSE K. On the Effectiveness of MFCCs and their Statistical Distribution Properties in Speaker Identification[J]. IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurement System, July 2004: 136-141.

[6] 杨行峻, 迟慧生. 语音信号数字处理[M]. 北京: 电子工业出版社, 1995, 24-27.

[7] TZANETAKIS G, COOK P. Musical Genre Classification of Audio Signals[J]. IEEE Trans. Speech Audio Processing, Vol. 10, No. 5, July 2002: 293-302.

[8] LOGAN B. Mel Frequency Cepstral Coefficients for Music Modeling[J]. International Symposium on Music Information Retrieval, Vol. 28, 2000.

[9] GOTO M. A Chorus-section Detecting Method for Musical Audio Signals[J]. In ICASSP, Hong Kong, April 2003: 437-440.