第41卷 第6A期 计 算 机 科 学 Vol. 41 No. 6A
2014年6月 Computer Science June 2014

Chroma特征的鲁棒性验证

张 秀 李念祖 李 伟
(复旦大学计算机学院 上海 201203)

摘 要 基于内容的多版本音乐识别是近些年来音乐信息检索领域一个比较热门的研究课题。考虑到多版本音乐可能在节奏、速度、音调、音色以及结构等方面的变化,该研究的关键在于选取能反映音乐主要旋律走向的相对稳定的音频特征,在不同的音乐版本之间进行相似度的比较。Chroma特征反映了音频能量在各个音调类间的相对分布,考虑了和声信息、与音色无关、对噪声鲁棒,所以成为多数多版本音乐识别算法使用的特征。通过设计和实验,探究不同的音频干扰形式对Chroma特征的影响,就Chroma特征对音调无关因素的鲁棒性进行验证。

关键词 Chroma特征,鲁棒性,音调不变性,多版本音乐识别
中图法分类号 TP399 **文献标识码** A

Verification for Robustness of Chroma Feature

ZHANG Xiu LI Nian-zu LI Wei
(Department of Computer Science and Technology, Fudan University, Shanghai 201203, China)

Abstract Content-based cover song detection is a hot research program in the field of music information retrieval in recent years. With the consideration that the different versions of music may change in many aspects, such as tempo, speed, pitch, timbre and structure, the key factor in this research is to extract robust audio feature, which can represent songs' main melody progress, and compare their similarities between cover songs. Chroma features express the distribution of audio energy among different pitch class, consider presence of harmonics , timbre independence and robustness to noise, so are used by a lot of cover song detection systems. Based on the theory, we designed and conducted the experiment to explore the effect of different forms of audio interference, and verified Chroma's robustness in the aspect of pitch invariance.

Keywords Chroma feature, Robustness, Pitch invariance, Cover song detection

1 引言

在物质生活水平不断提高的基础上,人类对于高质量精神文化生活的需求愈发重视。而音乐欣赏也正是人类的文化娱乐生活中一个非常重要的方面。尤其是随着信息技术的进步,互联网的海量资源优势为人们能够简单快速获取目标音乐提供了可能性,同时也对于能够快速准确地进行音乐的检索提出了新要求。因此,基于内容的音乐信息检索(CBMIR, Content-based Music Information Retrieval)以及基于内容的音乐认证技术[1]在工业界和学界都得到人们越来越多的关注,相应的新产品、新技术、新算法也层出不穷,并在近些年取得了迅速发展。

基于内容的音乐信息检索,是针对传统的基于文本描述的检索技术而言的。它是指一个利用从音乐本身中提取出来的音频信息在众多候选资源中找到满足用户所需目标的匹配、定位的过程。为了取得好的音乐信息检索的效果,一个关键的技术要点在于提取鲁棒性好且区分性强的音频特征来表征音乐本身的信息。早期对于音乐信息检索的研究借鉴语音信息处理过程,使用能量或基于谱的音色特征[9,14]。基于目前对于该方向的研究,大都是对一些底层的音频频谱信号进行进一步的处理,采用能够一定程度上反映音乐旋律走向和声信息、对噪声鲁棒,同时更有利于检索的中层特征。本文主要研究的Chroma特征就是一种在音乐信息检索领域,尤其是多版本音乐识别课题中应用广泛的中层特征。

多版本音乐就是一个源自原始音乐的新的演唱、表演、或重新录音版本。我们平时所说的"翻唱",就是多版本音乐中最常见的一种表现形式。针对其目标,一般来说就是给定一段音乐作为查询,从数据库中返回一个按相似度高低排序的歌曲列表。

现有绝大多数多版本音乐识别的研究都选用Chroma作为基本的音频旋律特征来进行算法的设计。其主要原因是:该特征被普遍认为考虑到了和声的存在,减小了噪声和非音调声音的干扰,与绝对音调、歌唱人音色、演奏乐器、音量和力度无关,从而能很好地表现多版本音乐中核心音乐要素信息。由此可见Chroma特征对于保证在多版本音乐识别算法的性能方面发挥着巨大作用,这也是验证和尽量提高Chroma特

本文受863国家自然科学基金项目(2011AA01A109),国家自然科学基金(61171128)资助。
张 秀(1989—),硕士生,主要研究方向为音频信息处理、说话人识别;李念祖(1990—),硕士生,主要研究方向为音频信息处理;李 伟(1970—),博士,教授,博士生导师,主要研究方向为信息隐藏与数字水印信息隐藏、音乐信息检索、语音识别。

征的有效性和鲁棒性的重要意义所在。

本文的主要研究目的就是通过实验验证 Chroma 特征对各类音频信号攻击的鲁棒性，即当原始音乐信号中的某些音频要素发生变化时，Chroma 特征抵御无关干扰、保持不变性的能力。在许多相关研究中，Chroma 特征都被认为具有对噪声和非调性声音的鲁棒性，与音色、力度无关等优点[2,13]。然而，在目前的论文中并没有通过数据来直接表明 Chroma 特征在原始音频信号中各音乐信息要素在受到多大程度的改变时，仍然能够保持不变性。本文的贡献就在于，首次通过实验说明不同的音频干扰形式对 Chroma 特征的影响。

本文第 2 节介绍 Chroma 特征相关原理与提取方法；第 3 节包含 Chroma 特征鲁棒性的实验方法，最后是结果和讨论。

2 Chroma 特征

Chroma 特征，也称和声音调类分布（HPCP，Harmonic Pitch Class Profile），是目前音乐信息检索领域，尤其是多版本音乐识别研究中应用最普遍的一种音频旋律特征[8]。本章将对该特征相关原理和提取算法进行介绍和讨论。

2.1 Chroma 特征的乐理基础

根据音乐认知学和心理声学相关理论，人耳对于声音音调的感知是有周期性的。发声频率为二倍关系的两个音对于人的听觉而言具有相似性[12]。基于该事实，现代音乐理论中定义这样满足二倍频关系的两个音具有相同的音名，而它们之间的音程，即音高的差距称为一个八度。十二平均律音阶体系将八度的音程按照频率，等比例地划分成 12 等份，每一等份称为一个半音，分别记为 C、C♯、D、D♯、E、F、F♯、G、G♯、A、A♯、B 共 12 个音名。

Chroma 特征就是依据十二平均律音阶体系而设计的。其基本原理是，将原始音频信号的频谱能量量化成 12 个与八度无关的半阶音调类，使得每个半阶音调类中包含了该类中同音名的所有八度的能量。一个 Chroma 特征通常表示为一个 12 维的向量 $v=[V(1),V(2),V(3),\cdots,V(12)]$，其中每一个向量元素都与{C、C♯、D、D♯、E、F、F♯、G、G♯、A、A♯、B}中的一个音名对应，反映了声音信号的局部能量在这 12 个音名所代表的半阶音调类上的分布。

图 1 中是钢琴单键基本发音的频谱及其对应 Chroma 特征图谱的一个示例。

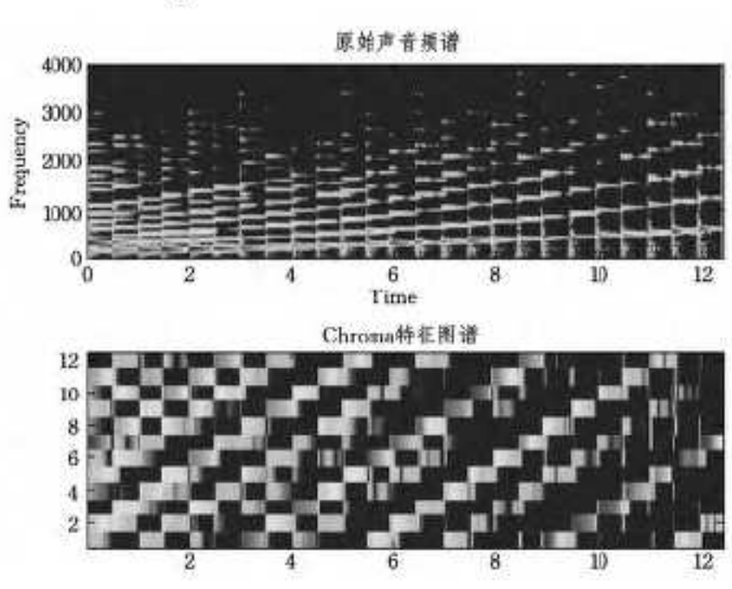


图 1 声音频谱与 Chroma 特征图谱的对比

2.2 Chroma 特征提取的算法实现

Fujishima 在一个和弦识别系统中实现了音调类分布 PCP（Pitch Class Profile）的提取算法[10]，这可以认为是较早的 Chroma 特征提取算法版本。E. Gómez 考虑了和声加权的影响，在该算法的基础上进行改进，提取得到和声音调类分布特征 HPCP[11]，也即更符合现在通常意义标准的 Chroma 特征。

该特征提取算法大致遵循图 2 所示的基本过程。

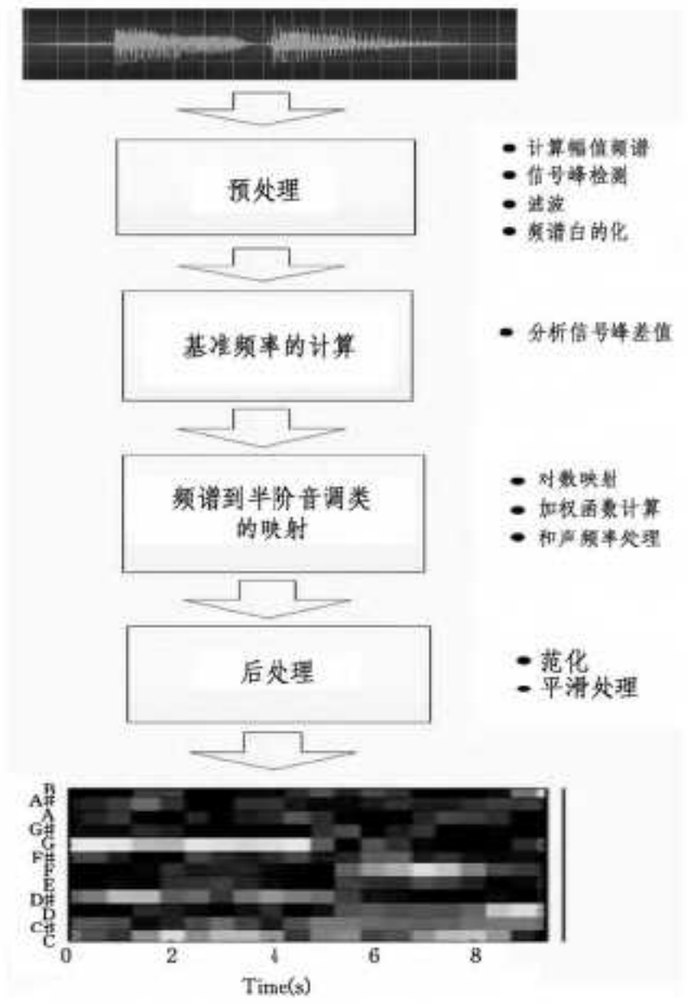


图 2 Chroma 特征提取过程框图

在特征提取的前处理阶段，首先需要把原始的音频输入转换为固定采样率的单通道信号。双声道立体声一般采用信道平均的方法进行转换。对于转换好的信号，通常取连续的、部分重叠（如 75%）的短时帧（如 100ms）逐帧进行计算，通过短时傅里叶变换（STFT，Short-Time Fourier Transform）得到信号的幅值频谱[3,8,11]。也有一些算法是用恒值 Q 变换（CQT，Constant-Q Transform）代替傅里叶变换计算频谱[4]。考虑到人耳对声音的感知能力是有限的，通常还会接着进行滤波处理，即只保留给定频率范围内（如 40～5000Hz）的频谱图。

得到幅值频谱后，就可以进行峰值检测处理，即在频谱中找出确定固定频率范围内最大的 nPeaks 个峰值，并计算它们之间的差值，然后结合十二平均律中规定的 A4 调标准频率为 440 Hz，得到原来音乐信号的基频 f_{ref}。

考虑到音乐中和声的因素，算法还有一个重要的步骤，就是在最后计算特征向量之前，针对各个峰值的频谱进行白化（Whitening）处理。这也是该算法相较于 Fujishima 的版本[10]最大的改进。其具体细节可以参考文献[11]，鉴于篇幅限制，本文在此不过多介绍。

最后通过频谱峰值能量映射各个半截音调类得到 Chroma 特征向量。具体来说，向量中每一维元素的值可以通过下列公式计算得到：

$$\mathrm{Chroma}(n)=\sum_{i=1}^{nPeaks} w(n,f_i)\cdot a_i^2,\ n=1,2,\cdots,12 \qquad (1)$$

其中，a_i 和 f_i 分别是第 i 个信号峰的幅值与频率。$w(n,f_i)$ 是频率 f_i 的信号对于半阶音调类 n 的权重，它的计算方法如下：

先确定每个半音调类的中心基准频率 f_n：

$$f_n = f_{ref} \cdot 2^{\frac{n}{12}}, n=1,2,\cdots,12 \quad (2)$$

那么定义每个信号峰频率 f_i 和半音调类的中心基准频率 f_n 之间的音程距离为：

$$d = 12 \cdot \log_2(\frac{f_i}{f_n}) + 12 \cdot m \quad (3)$$

其中，m 是一个整数值的调整因子，使得 $|d|$ 的值最小。然后可以按照下式计算得到权重值：

$$w(n,f_i) = \begin{cases} -\cos^2(\frac{\pi}{2} \cdot \frac{d}{0.5 \cdot l}), & |d| \leqslant 0.5 \\ 0 & |d| > 0.5 \end{cases} \quad (4)$$

其中，l 是预先设定的加权窗长度，通常 $\frac{4}{3}$ 取半音音程。

除 Gómez 提出的算法外，目前还有许多其他的 Chroma 特征提取算法的实现版本：Ellis 等人在文献[8]中，考虑到音乐的节拍变化因素，提出并实现了一种节拍对齐的 Chroma 特征提取方法，使 Chroma 特征对于音乐的节奏速率变化具有一定的稳定性。文献[13]受到梅尔频率倒谱系数(MFCC，Mel-frequency Cepstral Coefficients)提取过程的启发，改进了 Chroma 特征。算法在对音频信号音调对数化处理后，先进行离散余弦变换(DCT，Discrete Cosine Transform)，然后只保留与音色相关度较小的上层系数进行反余弦变换得到最终频谱，最后映射到各个半阶音调类上得到 CRP(Chroma DCT-reduced Log Pitch)特征向量。论文通过实验证明，CRP 特征的提取更符合人耳的听觉感受，对音色变化具有鲁棒性。此外，许多算法还会对提取得到的 Chroma 特征在时间域上做平滑(Smoothing)[5]后处理，以减小瞬时噪声，提高特征的性能。

3 Chroma 特征鲁棒性实验

本章主要叙述 Chroma 特征鲁棒性验证的实验内容。实验对于所用数据集中每一首歌曲的原始音频信号，采用多种不同的形式进行处理，得到修改后的歌曲版本。对于这些修改的歌曲，计算得到它们的 Chroma 特征矩阵，与原始歌曲的 Chroma 特征矩阵对齐后比较得到变化率，从而验证 Chroma 特征对于这些修改处理的鲁棒性。

3.1 对原始音频信号的修改处理

对原始音频信号的修改处理主要通过 Gold Wave 音频处理软件和由 Andreas Lang[1)] 发布的 StirMark Benchmark for Audio 工具箱完成，修改形式大致可以分为与音调无关的处理和与音调相关的处理两类。

3.1.1 与音调无关的修改处理

与音调无关的修改处理通过 StirMark Benchmark for Audio 工具箱完成，主要包括如下内容：

Addnoise：在原始音频信号中加入白噪声；

Addsinus：在原始音频信号中加入正弦波信号；

AddBrumm：在原始音频信号中加入蜂鸣声；

Amplify：改变原始音频的响度，即对原始音频信号的振幅以固定的倍数进行放大；

Compressor：类似于动态范围压缩器，增加或减小原始信号中较安静信号的响度；

Invert：反转原始音频信号，即对原始音频信号的相位做 180 度的偏移；

Normalizer：把原始音频信号幅度归一化到定值；

ZeroCross：类似于限幅器，对于原始音频信号小于一定阈值的采样部分，设置其幅度值为 0；

RC-HighPass：对原始音频信号作高通滤波处理；

RC-LowPass：对原始音频信号进行低通滤波处理。

为了更直观清楚地说明这些处理方式，我们选择了上述处理形式中的 4 种，在表 1 中通过图例展示了其对原始音频波形的改变。而表 2 则主要罗列了和这些处理有关的一些参数值的设定。

表 1 与音调无关的修改处理波形变化图例

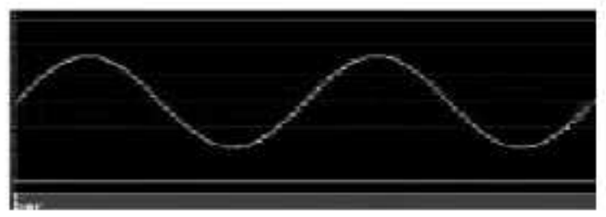
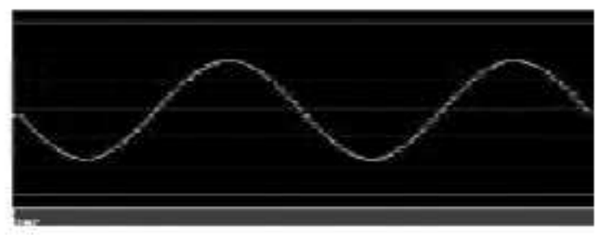
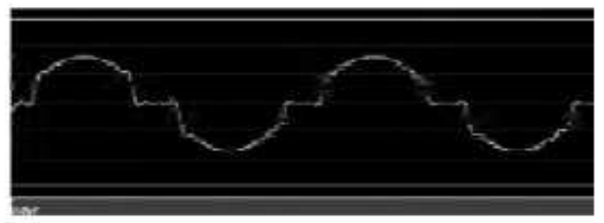
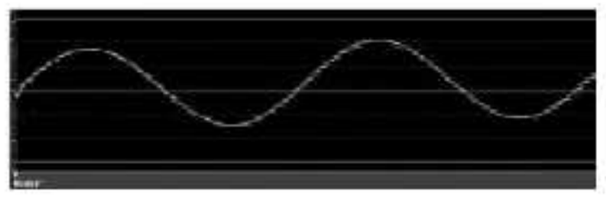
	信号波形
原始信号	
Invert	
Normalize	
ZeroCross	
Addbrumm	

表 2 与音调无关的修改处理参数设定

	参数	设置值
AddBrumm	〈强度〉	〈100～10100〉
	〈频率〉	〈55〉
AddNoise	〈强度〉	〈100～900〉
AddSinus	〈幅值〉	〈1300〉
	〈频率〉	〈900〉
Amplify	〈放大倍数〉	〈50〉
Compressor	〈分贝阈值〉	〈−6.123〉
	〈压缩值〉	〈2.1〉
Invert	无	
Normalizer	〈窗长度〉	〈2048〉
	〈最大值〉	〈28000〉
RC-HighPass	〈高通限制频率〉	〈200〉
RC-LowPass	〈低通限制频率〉	〈9000〉
ZeroCross	〈过零阈值〉	〈1000〉

3.1.2 与音调有关的修改处理

与音调有关的修改处理利用 GoldWave 音频处理软件完

1) http://wwwiti.cs.uni-magdeburg.de/~alang/smba.php#smba-DP

成。对每一首原始音乐的音调分别做降低和升高 1%、2%、3%和 5%的处理,得到与音调有关的 8 个修改版本。

3.2 实验设置与评价标准

实验所用数据集由 160 首歌曲组成,其中包含中文流行音乐和不同类型英文歌曲各 80 首。每首歌曲都有 27 个修改后的版本,一共是 4320 首。每只歌曲均为双声道立体声,采用 WAV 编码,时间长度在 2～6 分钟范围内不等,采样率为 44.1kHz,16 位量化数字。

实验应用 Ellis[1] 等人实现的 Chroma-IF 算法,采用不同的傅里叶变换窗长,从歌曲音频信号中提取 Chroma 特征。

关于 Chroma 特征的鲁棒性的度量,实验定义了两个数值标准,分别是修改后歌曲的 Chroma 特征较原始歌曲的 Chroma 特征相比的平均错误率 MER(Mean Error Rate)和平均相关系数的均值 MAC(Mean Average Correlation)。前者的取值在 0～1 范围内,后者的取值在 -1～1 范围内。若 MER 的值接近 0,MAC 的值接近 1,则表示 Chroma 特征发生的变化小,鲁棒性强。接下来对二者的计算方法进行介绍。

本文在第 2 节中已经提到,Chroma 特征反映了音频信号的局部能量在 12 个半阶音符类中的分布。所以,我们在使用 Chroma 特征时,往往更关注其各个向量元素之间的相对关系,而不是它们绝度值的大小。基于这种考虑,我们在计算平均错误率时,先对 Chroma 特征进行归一范化处理。其具体方法是:对 Chroma 向量中的 12 维向量元素按照值的大小降序排列,然后以排序后各元素在原始 Chroma 特征向量中的位置索引代替原始数值,得到新的特征向量,该向量的每个元素值都是[1,12]范围内的整数。记原始的 12 维 Chroma 特征向量为 $C=[e_1,e_2,,e_{12}]$,那么归一范化处理后的向量可以表示为:

$$SI=[I_1,I_2,\cdots,I_{12}]\,|\,e_{I_1}\geqslant e_{I_2}\geqslant\cdots\geqslant e_{I_{12}} \tag{5}$$

对于一首原始未修改的歌曲,它的 Chroma 特征矩阵记为 $M=[C_1,C_2,\cdots,C_L]$,归一范化处理得到 $norm(M)=[SI_1,SI_2,\cdots,SI_L]$,则对于该歌曲修改后的版本,对应有 $M^{\mathrm{mod}}=[C_1^{\mathrm{mod}},C_2^{\mathrm{mod}},\cdots,C_L^{\mathrm{mod}}]$ 和 $norm(M^{\mathrm{mod}})=[SI_1^{\mathrm{mod}},SI_2^{\mathrm{mod}},\cdots,SI_L^{\mathrm{mod}}]$。

那么可以计算得到修改后歌曲 Chroma 特征相对于原始歌曲 Chroma 特征的错误率 ER(Error Rate):

$$ER=\frac{\sum_{i=1}^{L}Dis(C_i,C_i^{\mathrm{mod}})}{L} \tag{6}$$

其中,$Dis(C_i,C_i^{\mathrm{mod}})$ 的值的计算方法如下:

$$Dis(C_i,C_i^{\mathrm{mod}})=\begin{cases}0, & SI_i=SI_i^{\mathrm{mod}}\\ 1, & SI_i\neq SI_i^{\mathrm{mod}}\end{cases} \tag{7}$$

在此基础上,进一步得到平均错误率:

$$MER=\frac{\sum_{m=1}^{N}ER_m}{N} \tag{8}$$

其中,N 表示数据库中的歌曲总数。

平均相关系数均值计算时,对于每一首原始歌曲和其修改后的版本,已知 Chroma 特征矩阵分别为 $M=[C_1,\cdots,C_L]$ 和 $M^{\mathrm{mod}}=[C_1^{\mathrm{mod}},\cdots,C_L^{\mathrm{mod}}]$,则先计算其平均相关系数 AC(Average Correlation),即它们特征矩阵包含的所有对应 Chroma 向量的相关系数的平均数:

$$AC=\frac{\sum_{i=1}^{L}corr(C_i,C_i^{\mathrm{mod}})}{L} \tag{9}$$

其中,$corr(C_i,C_i^{\mathrm{mod}})$ 表示特征向量 C_i 和 C_i^{mod} 的相关系数。那么平均相关系数均值 MAC 可按照以下公式计算得到:

$$MAC=\frac{\sum_{m=1}^{N}AC_m}{N} \tag{10}$$

其中,N 表示数据库中的歌曲总数。

3.3 实验结果与讨论

对于音调无关的修改处理,实验结果如图 3—图 5 所示。

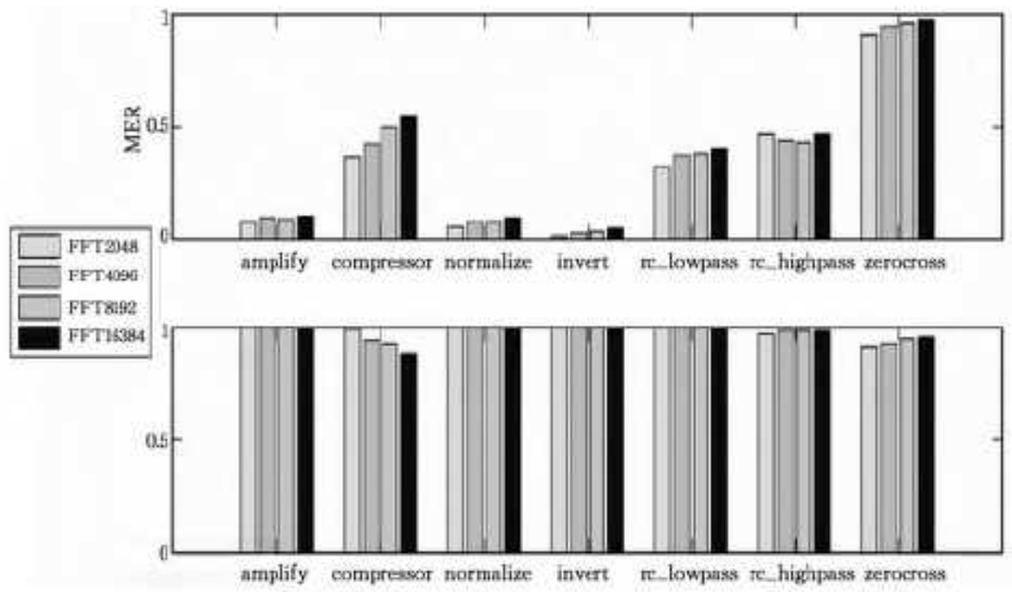


图 3 Chroma 特征鲁棒性实验结果(1)

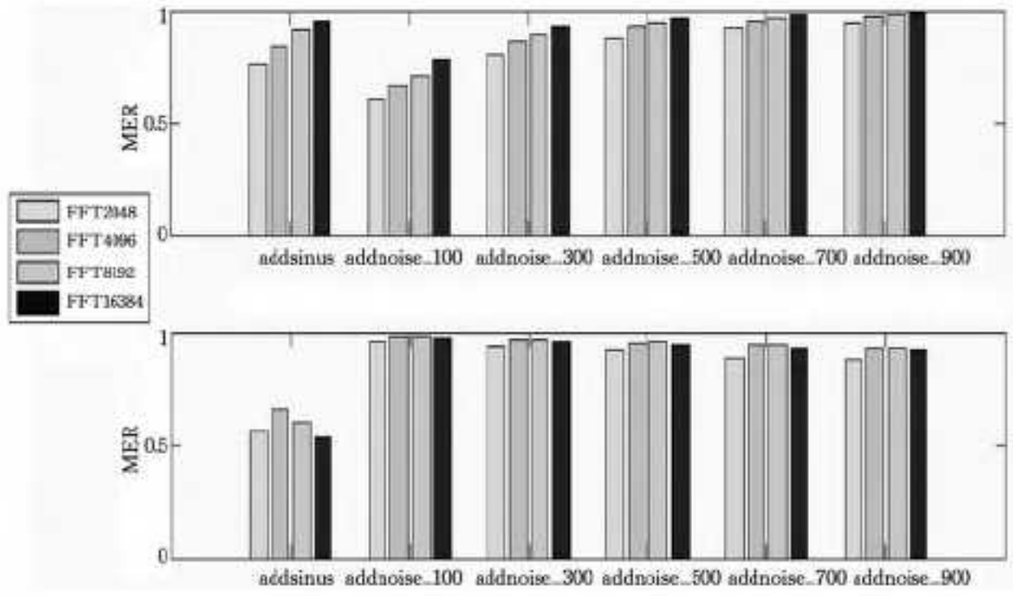


图 4 Chroma 特征鲁棒性实验结果(2)

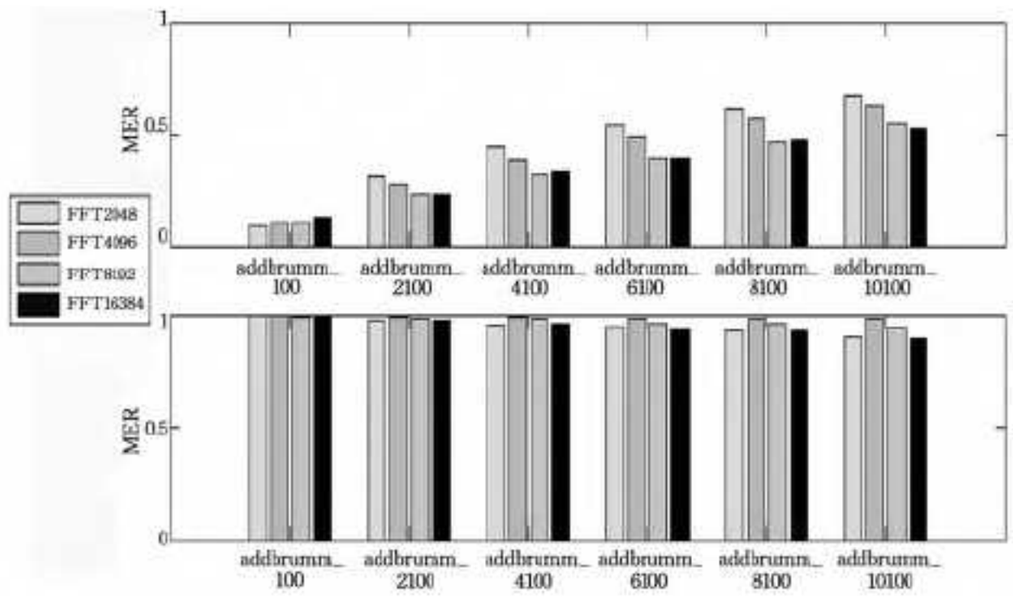


图 5 Chroma 特征鲁棒性实验结果(3)

从图 3 中可以看到,经过 Amplify,Invert 以及 Normali-

[1] http://www.ee.columbia.edu/~dpwe/resources/Matlab/chroma-ansyn/

zer 3 种类型的修改后，MER 和 MAC 分别接近 0 和 1，表明 Chroma 特征对于这些干扰是鲁棒的。我们容易分析得到其中原因：理论上决定半阶音符类分布的因素只有音频信号的频率，而这 3 种处理形式都不会改变原始频率，所以理应不对 Chroma 特征信息产生明显的影响。而在实验结果中，所示 Chroma 特征的微小改变，可能是在处理音频信号产生噪声和计算特征向量时的随机误差造成的。对于 RC-LowPass 和 RC-HighPass 两种滤波处理，原始音频中的一部分频率范围中的信号被去掉，则这部分信号的能量也随之消减，所以反应局部能量在各个半阶音符类上分部的 Chroma 特征理论上必然会受到一定的影响。

图 3 的实验结果也能证明这点。而使用高通滤波的影响略大于低通滤波，这是因为在人耳的听力范围内，人耳对于音乐低频部分的感知比高频部分更敏感。依据图 3 所示，造成 Chroma 特征最明显变化的是 Compressor 和 ZeroCross，这是因为这两种处理都对原始音频中部分信号的强度进行了改变，进而改变了局部能量在各个音阶上的分布。Addnoise，Addsinus 和 Addbrumm 3 种处理都在原始音频中加入了新的干扰信号，考虑到 Chroma 特征的计算基本依据是音频信号的局部能量，不难推断，这些干扰对 Chroma 特征的影响程度主要取决于干扰信号的强度。如图 4、图 5 所示，当外加的干扰信号强度越大时，Chroma 特征的改变越显著。

图 6 展示了与音调有关的修改处理的实验结果。从该实验结果中可以看到，Chroma 特征对于声音频率，即音调的改变是十分敏感的，即使是 1% 的音调偏移，MER 就已经很高，表明歌曲特征矩阵中的绝大部分 Chroma 向量已经发生改变。当达到 5% 的频率移动时，MER 接近于 1，MAC 的值在 0 附近，说明 Chroma 特征就已经近乎完全发生改变。根据十二平均律，$2^{1/12}$ 倍频关系为一个半音音程，$2^{1/12}\approx1.0594$，说明当音乐信号的频率发生 5% 的偏移时，音调就已经发生接近一个半音的改变，所以本实验结果与音乐理论是相合的。

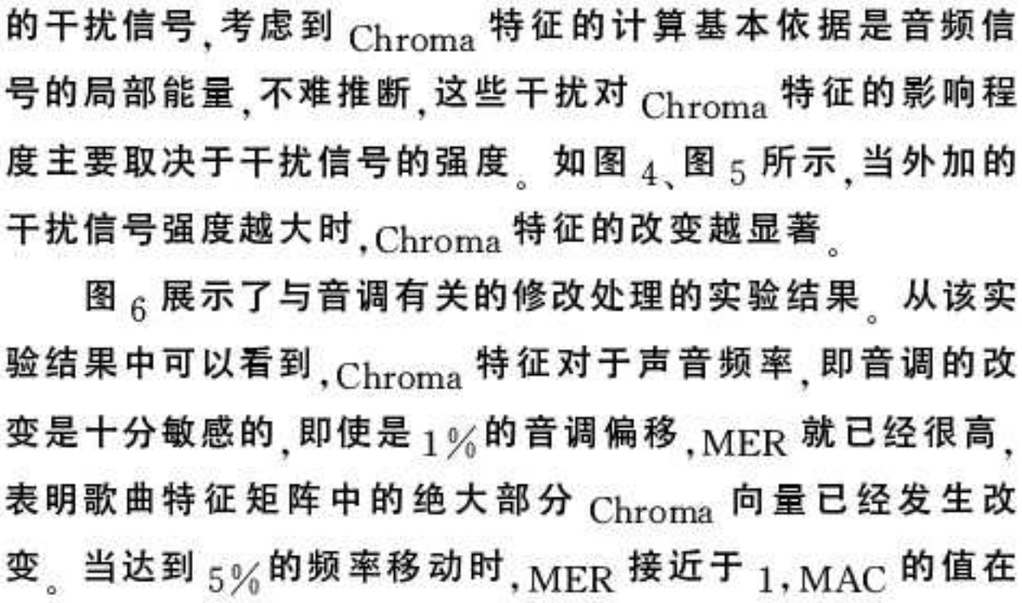


图 6 Chroma 特征鲁棒性实验结果(4)

综合以上结果可以得到关于 Chroma 特征鲁棒性的基本结论：Chroma 特征对于非音调性因素和噪声的干扰能在一定程度上保持不变性，可以认为是相对鲁棒的，而对于音乐的音调变化则是敏感的，特征本身并不具备音调不变性。

结束语 本文针对多版本音乐识别中广泛使用的 Chroma 特征进行分析，首次通过实验，测试在原始音频进行音调偏移和加入与音调无关的信号干扰对 Chroma 的影响。验证了 Chroma 特征的稳定性。

参 考 文 献

[1] 李伟，汪竹蓉，李晓强，等. 数字音频认证研究综述[J]. 计算机科学，2009，36(10)：21-24

[2] Ahonen T E. Combining Chroma Features For Cover Version Identification[C]// Proc. International Symposium on Music Information Retrieval (ISMIR). Utrecht, Netherlands, 2010：165-170

[3] Bartsch M A, Wakefield G H. Audio thumbnailing of popular music using chroma-based representations[J]. IEEE Transactions on Multimedia, 2005, 7(1)：96-104

[4] Brown J C. Calculation of a constant Q spectral transform[J]. The Journal of the Acoustical Society of America, 1991, 89：425

[5] Cho T, Bello J P. A feature smoothing method for chord recognition using recurrence plots[C]// Proceedings of the 12th International Society for Music Information Retrieval Conference. 2011：651-656

[6] Csurka G, Dance C, Fan L, et al. Visual categorization with bags of keypoints[C]// Workshop on statistical learning in computer vision, ECCV. 2004, 1：22

[7] Ellis D P W, Cotton C V. The 2007 labrosa cover song detection system[C]// MIREX 2007. 2007

[8] Ellis D P W, Poliner G E. Identifying 'cover songs' with chroma features and dynamic programming beat tracking[C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2007(ICASSP 2007). IEEE, 2007, 4

[9] Foote J. Arthur: Retrieving orchestral music by long-term structure[C]// International Symposium on Music Information Retrieval. 2000, 1

[10] Fujishima T. Apparatus and method for recognizing musical chords: U. S. Patent 6, 057, 502[P]. 2000-5-2

[11] Gómez E. Tonal description of music audio signals[C]// Unpublished doctoral dissertation, Universitat Pompeu Fabra. Barcelona, Spain, 2006

[12] Gómez E, Herrera P. The song remains the same: Identifying versions of the same piece using tonal descriptors[C]// Proceedings of 7th Intl. Conference on Music Information Retrieval. 2006

[13] Muller M, Ewert S. Towards timbre-invariant audio features for harmony-based music[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(3)：649-662

[14] Yang C. Music database retrieval based on spectral similarity [R]. Stanford, 2001