◎数据库、数据挖掘、机器学习◎

基于哼唱的音乐检索应用系统

华 斌,尹文慧,张奕林
HUA Bin, YIN Wenhui, ZHANG Yilin

天津财经大学 信息科学与技术系,天津 300222
Department of Information Science and Technology, Tianjin University of Finance and Economics, Tianjin 300222, China

HUA Bin, YIN Wenhui, ZHANG Yilin. Music retrieval system based on query by humming. Computer Engineering and Applications, 2014, 50(22):141-144.

Abstract: Through analyzing the humming melodies pitch extraction and retrieval algorithm, a complete system framework of Query By Humming(QBH) is proposed. It includes the melody feature extraction and approximate melody matching in MIDI music database. The Mel Frequency Cepstral Coefficients(MFCC) is extracted. Through analyzing the theory of DTW algorithm, the cosine similarity of the delta duration sequence is added with the characteristics of the sound for performance improvement of the system. Experiments are conducted in a test set of 340 MIDI songs. The system gets a success rate of top-3 increased by 3.7% and a 16% time reduction.
Key words: query by humming; pitch track; melody match; dynamic time wrapping

摘 要:通过研究哼唱旋律基频提取和检索算法,给出了一个完整的基于哼唱的音乐检索系统框架。系统主要分析了旋律特征提取和近似旋律匹配部分。旋律特征提取部分采用基于差分Mel倒谱法求基频;旋律匹配部分对经典的动态时间弯折算法原理分析后,根据声音特征引入音长差序列的余弦相似度,提高了检索效率和精度。在340首MIDI歌曲的测试集上,前三位识别效率提高3.7%,用时降低16%,系统的性能有明显改善。
关键词:哼唱检索;基频提取;旋律匹配;动态时间弯折
文献标志码:A **中图分类号:**TP391.3 doi:10.3778/j.issn.1002-8331.1212-0357

1 引言

用哼唱检索音乐能够让用户寻找到他仅仅只知道旋律的部分音调的一首歌。用户只是简单地通过电脑的麦克风哼唱出这段音调,然后系统通过查询包括这段音调的歌曲旋律数据库,返回一个查询结果的相关歌曲列表。这种查找方式相比于文本查找(曲名、演唱者等)更自然,也更方便,拥有很大的商业发展潜力,成为近年来很多人研究的热点[1-3]。

Ghias最早研究哼唱检索[4]。他提出的方案是用三个符号U(升高)、R(不变)、D(降低)表示旋律相邻两个音符的音高变化。然后应用近似字符串匹配方法来比较两段旋律的相似度。然而这种描述旋律信息的方法相对简略,在辨识度上也不太理想。Kosugi[5]等采用把音高和节奏信息结合的旋律表示方法,可以适应大型乐曲库的检索。但是该系统要求用户必须伴随一个节拍器哼唱,使用起来不方便。Shih等在QBH系统中使用了隐马尔科夫模型(HMM)比较哼唱旋律和目标歌曲的相似度,实验表明该方法对音高不准比较敏感,但是对节奏上的哼唱误差有较高的容忍度。台湾清华大学在哼唱检索系统中采用DTW和Line Scaling多级匹配算法,但由于DTW,Line Scaling直接使用基频曲线进行匹配,系统的检索速度较慢。

因此如何有效地利用旋律音高和音长特征进行高效检索,是人们进一步研究的方向。针对此问题,本文提出了一种改进的动态时间弯折(Dynamic Time Warping,DTW)算法,引入音长差序列的余弦相似度,将旋律音高和音长同时进行平移,并在此基础上实现了一个哼唱音乐检索系统的原型。另外,有效地提取哼唱旋律的基频也是检索系统的关键步骤,系统模型中采用了差分Mel倒谱法获得帧的基频达到了较好的效果。

作者简介:华斌(1963—),男,博士,教授,主要研究领域为多媒体处理、计算机仿真、决策支持系统、管理信息系统、创新管理;尹文慧(1987—),女,硕士研究生;张奕林(1987—),男,硕士研究生。E-mail:yin12803@163.com
收稿日期:2012-12-29 **修回日期:**2013-05-02 **文章编号:**1002-8331(2014)22-0141-04
CNKI**网络优先出版:**2013-05-24, http://www.cnki.net/kcms/detail/11.2127.TP.20130524.1509.003.html

2　哼唱旋律特征提取

2.1　哼唱片段预处理

音频信号属于“短时平稳过程”每一帧信号视为平稳过程,即统计特性平稳,所以哼唱声音信号处理全过程一般都使用短时处理技术。为了更好地提取哼唱信号的旋律信息,在对哼唱片段进行分析前要进行预处理,包括哼唱旋律的预滤波、预加重、加窗和分帧及语音增强等。下面主要介绍本系统中分帧加窗和语音增强技术。

(1)分帧和加窗

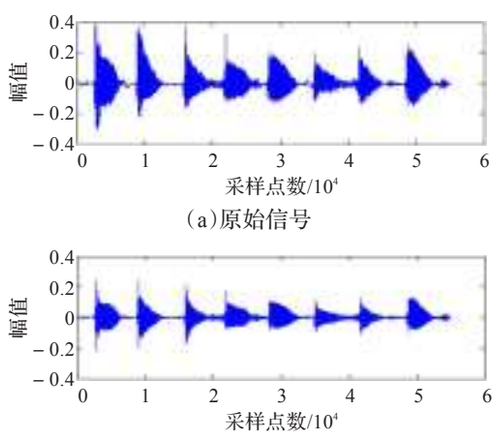
分帧就是把哼唱片段分成一帧一帧的短时信号。一般采用交叠分段的方法,这样可以使帧与帧之间较好保持连续性,因为后一帧都保留了前一帧的部分信息。把前一帧和后一帧的交叠部分称为帧移,帧移与帧长的比值一般取0~0.5。本系统采样频率为11 kHz,考虑频率分辨率和时间分辨率的平衡,采用256位的帧长,128位的帧移。因为傅里叶变换对应的是无限信号,信号经过分帧后变成有限信号,分帧的信号再进行傅里叶变换后,高频部分将有“泄露”,所以要进行加窗处理。窗函数有很多种,如矩形窗、汉宁窗、汉明窗等,各自有其优缺点,本系统采用汉明窗,可以较好的保留波形细节。汉明窗函数为:

$$W(n,\alpha)=(1-\alpha)-\alpha\cos(\frac{2\pi n}{N-1}) \tag{1}$$

其中,$0\leqslant n\leqslant N-1$,一般情况下,$\alpha$取0.46。

(2)语音增强

用户哼唱的声音信号不可避免地含有噪声,严重的会影响哼唱信号的质量。另外信号在传输过程中也会产生各种噪声[6]。噪声分为加性和非加性的。加性噪声一般有宽带噪声、周期噪声和语音干扰等,非加性噪声一般是输入残响和传输过程中的电路噪声,其中非加性噪声可以通过一些技术如同态滤波转为加性噪声。目前语音增强能从携带噪声的语音信号中获得较纯净的语音信号,是解决噪声污染比较有效的方法。语音增强技术一般采用滤波法、自相关抗噪法、减谱法、中心消波法等。本系统采用减谱法。如图1为哼唱原始信号波形图(a)和经过预处理降噪之后的信号波形图(b),可以看出比较好的去噪效果。



(a)原始信号

(b)预处理后信号

图1　原始信号和预处理后信号对比图

2.2　基频提取和切分音符

系统采用了速度较快的差分Mel倒谱法得到基音频率MFCC并转换为帧的半音高,采用两级切分音符的方法将音符分割,然后将帧音高加权处理得到每个音符的音高。旋律特征提取步骤如图2所示。

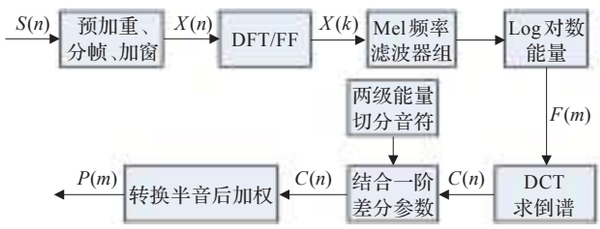


图2　旋律特征提取过程

在图2所示的特征提取过程中,经过离散余弦变换(DCT)后得到每一帧的倒谱参数$c(n)$,提取过程如公式(2)所示[7]:

$$c(n)=\sum_{m=1}^{M}S(m)\cos(\frac{\pi n(m-0.5)}{M}) \tag{2}$$

其中$0\leqslant n<M$,n为所取的MFCC个数;$c(n)$为第n个MFCC系数;M为三角滤波器个数;$S(m)$为声音信号的对数能量;本系统中M取24,n取12。

Mel倒谱充分模拟了人耳的听觉特性,根据声音的性质,将频谱转化为基于Mel频标的非线性频谱,最后变换到倒谱域上。由于没有任何前提假设,因此MFCC系数拥有较好的分辨率。然而传统标准的MFCC系数仅仅反映了语音特性的静态参数,缺少动态特性的描述,因此采用了多个参数特征组合方式来更有效地表征说话人的特征。系统中采用了一阶差分倒谱参数来描述动态特性,差分参数的表达式[8]为:

$$d(n)=\frac{1}{\sqrt{\sum_{i=-k}^{k}i^2}}\sum_{i=-k}^{k}i*c(n+i) \tag{3}$$

其中$c(n)$为已得出的MFCC系数,k是常数,根据经验取2。

这样提取到更精确的基频特征参数。基频f提取出来后,利用半音转换公式Semitone(半音)=69+12×lb(f/440),可以把每帧音高的频率都转换为半音单位来表示。通常一个音符要包含连续的X帧,通过上述方法得到帧的音高后,系统采用了一种加权求特征值的方法来获得一个音符的音高值。原理如下:已知一个音符由x帧组成,帧音高序列为$\{H1,H2,\cdots,Hx\}$。定义每帧的权重值为:$W_i=1-\cos(2\prod\times i/(x+1))$,$1\leqslant i\leqslant x$。然后将具有相同帧音高值权重累加,所得值最大就是这个音符的音高值。考虑到哼唱旋律音符的能量分布是中间高两端低,因此权重函数设计效果为中间加强而两端减弱。

3　旋律匹配

检索算法是研究设计音乐检索系统的核心问题。算法不仅要考虑旋律节奏的变化,要能最大限度地包容用户哼唱发音不准等,还要将检索用时控制在用户可接收范围内,它的性能将直接影响检索系统的精确性和鲁棒性。常用的旋律匹配[9]方法有近似字符串匹配、轮廓(音高曲线)匹配和基于统计模型的检索匹配。近似字符串匹配主要有BF(Brute Force)算法又称蛮力匹配算法[10]、KMP(Knuth-Morris-Pratt)算法[11]等,这是最早被研究和运用的一种方案,但是这类算法的辨识率比较低,对字符串的移位、伸缩都非常敏感。基于统计模型的隐马尔可夫模型相对成熟,算法依据的是频率提取特征向量在统计上的规律,需要经过大量的训练过程。相比之下,动态时间弯折(DTW)算法考虑了旋律节奏的变化,能适应哼唱者相对原始旋律的节奏变化和音高变化,有较高的容错能力,是目前基于内容的音乐检索研究的热点。本文在分析了经典DTW算法原理后对其进行了改进。

3.1　动态时间弯折(DTW)

DTW(Dynamic Time Warping,动态时间弯折)算法,是一种在一定路径的限制下,寻找最优化路径的方法,是动态规划理论在哼唱检索领域最常用的方法[12-14],本文将DTW作为旋律匹配的核心模块。在DTW中,路径方式和代价函数的选取对DTW的性能起着重要的作用,本文选取的路径方式如图3所示。

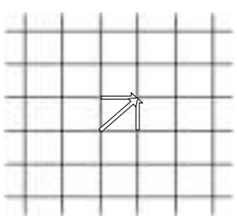


图3　DTW算法—3条路径

则DTW算法可表示为:

$$D(i,j)=\min\begin{cases}D(i-1,j)+d_{i-1,j}\\D(i-1,j-1)+d_{i-1,j-1}\\D(i,j-1)+d_{i,j-1}\end{cases}\tag{4}$$

其中,$D(i,j)$表示哼唱旋律的第i帧与模板旋律的第j帧之间的最小距离,$d(i,j)$为哼唱旋律的第i帧与模板旋律的第j帧之间的距离。

3.2　改进的DTW算法

传统的DTW算法只计算旋律特征的音高系数,忽略了旋律特征的音长部分。如图4所示,两边所表达的旋律特征音高部分相同,而音长却不同,但通过DTW算法计算结果是一样的。

如何将音长特征引入到DTW算法中,有学者做了研究[15],将音长直接与音高相加,但音高和音长的特征从本质来说是不同的,因此本文对引入方式进行了改进。

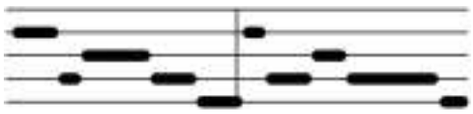


图4　音高相同、音长不同的旋律对比图

首先记录目标旋律的音长差序列$\{R1,R2,\cdots,Rn\}$,哼唱片段旋律的音长差序列$\{Q1,Q2,\cdots,Qn\}$。两者的余弦相似度D'为:

$$D'=\frac{\sum_{i=1}^{n}R_iQ_i}{\sqrt{\sum_{i=1}^{n}R_i^2\sum_{i=1}^{n}Q_i^2}}\tag{5}$$

考虑到余弦相似度的特点,为了修正类似向量(1,2)、(4,5)的余弦相似度过高这种不合理的情况,需要在各维度减去一个均值。这里采用模板旋律和哼唱旋律的音长差均值,即引入:

$$\alpha=\frac{\sum_{i=1}^{n}R_i+\sum_{i=1}^{n}Q_i}{2n}\tag{6}$$

为模板旋律和哼唱旋律音长的特征值,改进的D'为:

$$D'=\frac{\sum_{i=1}^{n}(R_i-\alpha)(Q_i-\alpha)}{\sqrt{\sum_{i=1}^{n}(R_i-\alpha)^2\sum_{i=1}^{n}(Q_i-\alpha)^2}}\tag{7}$$

最后通过对基因音高差的DTW相似度距离和音长相似度加权求和,得到

$$D^I=D+\beta D'\tag{8}$$

$\beta<0$,根据经验值,取-5。

4　哼唱检索应用系统实现

本文实现的哼唱检索系统的整体工作框架如图5所示。

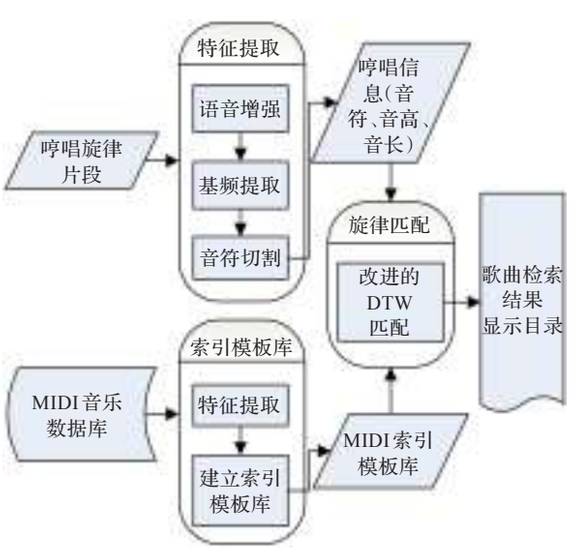


图5　基于哼唱的音乐检索应用系统框架图

系统包括三个主要模块,分别为哼唱旋律特征提

取、MIDI音乐索引模块库建立和近似旋律匹配模块。用户通过麦克风等输入设备进行哼唱采样,将哼唱片段预处理、去噪后进行特征提取,得到哼唱片段的旋律信息(音高和音长);MIDI 格式的文件易于提取音高、音长等旋律特征信息,因此被广泛用于现有的哼唱检索系统中[16],本系统即是在MIDI 格式的歌曲数据库中进行特征提取后建立音乐模板库;当用户哼唱输入完成后,系统将提取的旋律信息与音乐模板库中模板片段进行匹配,然后将按照匹配程度从高到低用列表方式显示给用户。系统原型如图6所示。

图6　基于哼唱的音乐检索应用系统原型

5　实验结果及分析

系统采用的歌曲数据库来自网络搜集的MIDI 格式音乐文件,单音轨模式,共340首歌曲;实验哼唱片段要求用户采用"Da"声哼唱,其中采集参数如下:采样频率为11 025 Hz,单声道采样,量化位数为8位。这种方式的特点是音符之间会留出低能量间隔,系统通过判断信号能量随时间的变化,能够更精确地进行音符划分。

实验中,请到了来自实验室同学和老师共12名实验者,哼唱水平都为普通水平。其中女生6名,男生6名,实验者的年龄分布情况为20~30岁男女各3名,40~50岁男女各3名。分别哼唱10次不同歌曲,共120个哼唱片段,哼唱时间要求为10~15 s。如图7所示为采用本系统改进的DTW算法下的不同性别和不同年龄段的实验者的歌曲命中率前三位、前十位、前二十位的比较。根据实验结果可知,该算法对性别和年龄的影响很小,证明该算法对音频变化和音高变化有比较好的容忍性。为比较新算法的效果,分别针对BF算法、KMP算法、传统的DTW算法、音高音长直接相加改进的DTW算法(以下记作DTW改进一)、基于帧的改进DTW算法(以下记作:DTW改进二)和本文改进的DTW算法进行了测试,测试结果如表1所示。表1列出了120个哼唱片段在六种匹配算法下的目标歌曲排名前三位、前10位和前20位的命中率及运行时间比例结果。

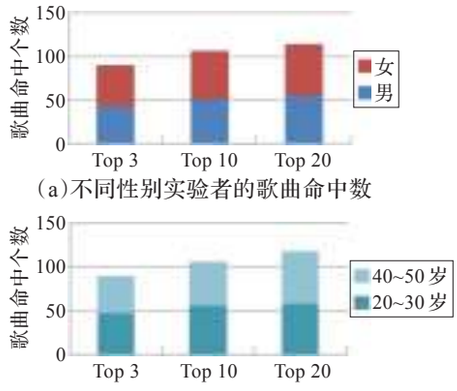


图7　分类统计实验者在改进DTW算法下的歌曲命中数

表1　六种匹配算法的实验结果对比

算法	前三位	前十位	前二十位	用时/DTW用时/(%)
BF	11.6%	21.6%	31.7%	3.79
KMP	18.3%	28.3%	41.6%	2.31
动态时间弯折	72.5%	87.8%	95.4%	1.00
DTW改进一	73.3%	87.8%	92.6%	0.98
DTW改进二	76.9%	89.3%	96.8%	2.41
加权改进的DTW	75.2%	88.4%	95.5%	0.84

根据表1可知,DTW算法相比BF算法和KMP算法的准确率和效率都有显著优势。而将音高与音长直接相加的改进效果不明显,基于帧的改进DTW算法,由于避免了音符切割,在准确度上较本文的改进的算法有提高,但是用时是本文的三倍,这在大数据集上是行不通的。本文经过改进的动态时间弯折算法在速度和精度上都有提高,虽然在命中率上的提高微小,但是经过改进的DTW算法用时有了显著减少,系统性能显著提高。这对以后在大规模数据库上应用检索提供了可行方案。

6　结束语

本文给出了一个基于哼唱的音乐检索应用系统模型。在哼唱旋律特征提取部分,采用了差分Mel倒谱法提取帧的基频后加权处理得到音符的音高;在旋律匹配部分,用经典的动态时间弯折法引入音长信息,根据哼唱片段音高和音长提取的精确度不同,提出音高和音长比加不同权重处理,经过试验表明,该算法可以有效提高系统的精度和运行时间。但是本系统目前仍然是实验系统,在去除噪声处理方面和搜索效率上还有很多待完善的地方,同时会继续研究更自然的哼唱方式不局限用爆破音哼唱,并研究在不同格式数据库中的算法改进以进一步提高系统的精确性和鲁棒性。

参考文献:

[1] 郭敏,刘加.一个基于哼唱的歌曲检索系统[J].语音技术,2009,33(12):63-64.

(下转169页)