

自动音乐标注系统综述 *

郭 奕¹, 徐红兵¹, 唐继勇¹, 何 俐²

(1. 电子科技大学 自动化工程学院, 成都 611731; 2. 四川省计算机研究院, 成都 610041)

摘 要: 通过对大量文献资料的查询、阅读和分析,对自动音乐标注系统进行了详细的总结,内容包括自动音乐标注技术的概念和研究内容、系统结构和发展等,并详细总结了其中的关键技术:音符切分和多基频估计,最后对目前的自动音乐标注系统的优缺点进行了总结,并指出了未来该技术的发展方向。在资料整理过程中还发现这是国内外第一篇专门针对自动音乐标注系统进行的综述性文章,具有重要的参考价值。

关键词: 自动音乐标注; 音乐信息检索; 音符切分; 多基频估计

中图分类号: TP23

文献标志码: A

文章编号: 1001-3695(2011)04-1201-05

doi:10.3969/j.issn.1001-3695.2011.04.001

Review of automatic music transcription system

GUO Yi¹, XU Hong-bing¹, TANG Ji-yong¹, HE Li²

(1. College of Automation, University of Electronic Science & Technology of China, Chengdu 611731, China; 2. Sichuan Institute of Computer Sciences, Chengdu 610041, China)

Abstract: Through querying, reading and analyzing a large number of literature, this paper gave a detailed summary on automatic music transcription system, including the concepts and contents, system architecture and development of the automatic music transcription system. And then gave a detailed summary of the key technology: onset detection and multipitch estimation. Finally, summarized the advantages and disadvantages of the current automatic music transcription system and pointed out the future direction of development of the technology. And also found that this was the first review article of automatic music transcription system and it had an important reference value.

Key words: automatic music transcription; music information retrieval; onset detection; multipitch estimation

音乐是人们经常接触的媒体,它有很多种存储形式,比如 MIDI、MP3 和各种压缩音乐制品、实时的音乐广播等。音乐数据的表示形式可以分为结构化的符号表示形式(如 MIDI 音乐等)、音频形式(wav、mp3 等)和乐谱形式三类^[1]。自动音乐标注,也称为自动音乐记谱或乐谱自动识别,从广义上讲,是指将音乐演奏的声学信号所对应的乐谱信息自动翻译出来,即将音乐数据的表示形式从其他形式转换成乐谱形式^[2]。它是音乐分析领域中的一项十分重要同时也是十分困难的工作,在音乐检索、音乐可视化、音乐辅助教学等领域具有广阔的应用前景。例如人们学习歌唱或者乐曲演奏时,可以通过自动标注系统的输出结果对其学习效果进行评价,从而加快学习的效果和效率。在基于内容的音乐检索技术中,自动标注技术是实现音频音乐语义级检索的基础,具有重要地位,是一项非常有意义的研究工作^[3,4]。

因为 MIDI 格式本身包含着大量的乐谱信息,所以基于 MIDI 格式的音乐分析工作可以做得比较深入,技术也相对成熟。而目前国际上提到的自动音乐标注技术,通常是指采用计算的方法从音乐的声音信号中提取音符数量、音高等信息,生成类似 MIDI 这样的符号化音乐表示形式或者乐谱^[5],即指将音频数据从音频形式转换为符号表示形式或乐谱形式。如果能找到一种有效的、通用的方法,将音频数据转换为 MIDI 这样的符号化音乐表示形式,那么很多对 MIDI 的处理技术也就

可以得到应用。

1 研究内容

通常音乐记谱需要受过专门训练的音乐工作者人工完成,需要耗费大量的时间和人力。自动音乐标注技术极大地方便了自动化音乐分析和注释,可被广泛用于基于内容的音乐检索、低位率的压缩编码和自动音乐伴奏系统等,是音乐信号处理中的关键技术。从研究内容上看,音乐自动标注技术包括以下几个方面。

1.1 音高估计与音符识别

音符(note)是乐音的最基本要素,是记录乐音及其时值长短的符号。从物理学角度看,单乐音音符主要由基频、振幅及倍频三个要素构成。乐器发出的乐音通过人耳听觉系统反映到听觉神经中枢,引起听者的主观感觉,这种感觉形成心理学上的单乐音音符的三要素:音调(又称音高)、响度、音色,这三个特性分别与三个物理量密切相关。音调(即音高,pitch)是由这个音符的基频决定的,频率高则音调高,频率低则音调低;响度是由音符振幅决定的;而音色则由单乐音音符频谱(即各谐波成分比例)构成决定^[6]。

音高估计或者音符识别是自动音乐标注工作的基础。由于音高作为一个重要的音频特征,在语音识别以及一般音频分

收稿日期: 2010-10-20; 修回日期: 2010-11-22 基金项目: 国家自然科学基金重点资助项目(60972107)

作者简介: 郭奕(1982-),男,博士,主要研究方向为计算机多媒体技术(lpngy@126.com);徐红兵(1967-),男,教授,博导,主要研究方向为智能控制系统;唐继勇(1965-),男,副教授,主要研究方向为广播电视技术、多媒体技术;何俐(1982-),男,主要研究方向为计算机应用。

类、分析等工作中都有广泛应用。有关音高估计的文献比较多,也存在一些经典的方法,如自相关法、倒谱法等。目前对于单乐器在实验室环境下所演奏的单个音符的音高估计是比较准确的,但是对实际的音乐作品或复杂条件下作音高估计则相当困难。

1.2 节拍或节奏识别

将音符按照一定的规律组织起来叫做节奏,而节拍(time)则是指相同时值的强拍与弱拍有规律的循环出现。节拍用拍号表示,如 2/4 表示以四分音符为一拍,每小节两拍^[6]。在研究成果方面,Goto 等人^[7]提出了一种针对不含有打击乐器(如鼓)的音乐节拍跟踪算法,该方法不仅能够对间隔不超过四分之一音符的节拍信息进行检测,还能够在较大的时间尺度内对音乐节拍的结构进行分析,从而找到“和旋”等高级音乐结构的改变点。Smith^[8]则利用一种基于线性相位 Gabor 变换的小波分析技术对音乐的节奏进行分析,他将可能的节奏信号分解为 0.1 Hz ~ 100 Hz 的若干分量,再通过相位一致性约束对其进行分析。此方法对单声道的打击类乐器所表达的音乐节奏比较有效。

1.3 旋律或和声提取

旋律是指经过艺术构思而形成的若干乐音的有组织、有节奏的和谐运动。它建立在一定的调式和节拍的基础上,是按一定的音高、时值和音量构成的、具有逻辑因素的单声部进行。音乐作品中的旋律与和声是一种非常重要的高级信息,但是从基于一般音频信号的音乐记录中将其提取出来是十分困难的。Su 等人^[9]提出了一种基于小波变换和自组织神经网络的多音色和声自动识别方法,该方法可以直接对和声音频片段进行分类而无须知道具体的乐器音色和音符序列。其他大多数方法都是针对 MIDI 文件进行处理的,如何从一般的 WAVE 文件中寻找或表示主旋律仍是一个难题。

1.4 复调音乐乐谱识别

复调音乐(polyphony music)指由几个声部构成的多声部音乐,即同一时刻存在多个旋律同时发音,与单声音乐相对。在实际的音乐作品中存在着大量的复调成分,这也是为什么自动音乐标注技术始终无法真正实用化的重要原因之一。对复调音乐进行处理的时候,由于需要识别音乐中同时发声的多个音符,该技术又叫做多基频估计技术。多基频估计是自动音乐标注的核心技术与主要难点,许多文献通常不加区分地将多基频估计称为自动音乐标注。有关多基频估计的内容将在下文作详细介绍。

2 自动音乐标注系统概述

早期的音乐标注系统有着严重的限制,如音乐所包含的声音个数、声音的频率范围、演奏音乐的乐器和音乐的风格等。早在 20 世纪 70 年代,斯坦福大学开发了第一个复音音乐标注系统^[10],此系统仅被用于有两个声音的二重奏。随后包括伊利诺斯大学、大阪大学、东京大学和麻省理工大学等研究组都开发了自己的复音音乐标注系统^[11~14]。

进入 21 世纪,自动音乐标注的研究日趋活跃,众多新技术被引入。Marolt^[15]的记谱系统用听觉模型作为信号前端,然后采用耦合自适应共振网络来跟踪谐波,最后采用人工神经网络

作音符识别。Abdallah 等人^[16]利用音乐信号的稀疏性,使用独立成分分析(independent component analysis,ICA)和非负矩阵分解(non-negative matrix factorization,NMF)把钢琴信号的频谱分解成独立自适应谱基,该系统在钢琴信号上取得了不错的结果。Vincent 等人^[17]采用类似的想法对音乐信号进行非线性的独立子空间(independent subspace analysis,ISA)分解,并使用因子化的隐马尔可夫模型(hidden Markov model,HMM)来生成音符。Bello 等人^[18]针对钢琴音乐记谱采用了基于先验知识的方法,假设音乐是由独立的音符线性叠加得到,这些独立音符事先存储于数据库中,从而把记谱任务转换为估计叠加系数的问题。Klapuri^[19]的记谱系统分为节奏检测和音高估计两个部分,其音高估计部分采用修正的听觉模型,取得了很好的效果。Ryynanen 等人^[20]采用了类似的音高估计方法,但在音符识别过程中引入了音符模型,值得一提的是,该系统允许音乐信号中有打击乐器存在。此外,产生式模型^[21]、泊松点过程分析^[22]也被引入了自动记谱研究中。

大部分自动音乐标注系统是非常复杂的,它们常常结合不同的知识和原理,很难用单一的分类学对这些系统进行总结和分类。许多自动音乐标注系统实际上是在着重解决以下两个问题:如何将自动标注问题分解成更容易实施的子问题;如何将音频信号逐步抽象成为音符标志。许多系统多少都利用了人耳的听觉模型,特别是人耳的音高感觉模型(pitch perception model)。主要存在两种有关音高感觉的理论:频谱理论(spectral theory)和时序理论(temporal theory)^[23~25]。两种理论都认为耳蜗相当于一个频谱分析仪。时序理论认为,耳蜗滤波以后,音高感觉是通过跟踪耳蜗滤波输出的时序包络的周期性来实现的;而频谱理论认为,在进行耳蜗频谱分析后,人耳感觉音高是通过频谱进行模式识别来实现的^[26~28]。

根据对大量文献资料的分析发现,目前系统的处理流程主要有两类:

a)直接对复调音乐进行处理,其流程如图 1 所示。在此类系统中,往往先采用起始点检测算法在输入音频数据中检测音符发声的起始点,并以音符发声起始点为边界将输入音频分割成片段,使每个片段中只在开始处有音符发声的起始点。然后采用多基频估计技术检测每个片段中同时存在的多个基频。重复音符检测模块的功能是判别相邻两个片段中的同名音符是来自于时值较长的一个音符,还是该音符的两次演奏。在随后的模块中,估算每个音符的时值、响度等,最后生成类似乐谱形式的标注结果^[29]。在有的系统中又把起始点检测和重复音符检测以及时值估计结合在一起叫做音符切分。

b)本着将复杂的问题分解成若干简单问题的思路,首先将复音音乐分离成若干个单音音乐,然后再针对单音音乐进行音值检测、音高提取、响度分析等标注操作,最后将结果整合得到最终乐谱。系统流程如图 2 所示。由于对单音音乐的标注技术相对成熟,有许多经典方法来处理,如 HPS(harmonic product spectrum)、Cepstrum、CBHPS(cepstrum biased HPS)、最大似然(maximum likelihood)和自相关(autocorrelation)算法等^[30~35]。采用这种思路构建的自动音乐标注系统的核心问题是如何有效地将复音音乐分离为若干单音音乐。文献[36~38]介绍了几种语音分离的方法,文献[39,40]又总结了专门针对音乐文件的分离方法。

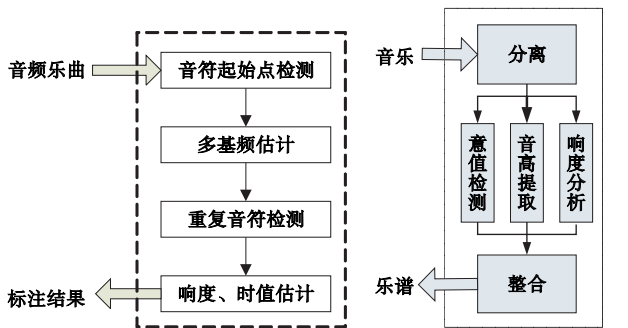


图1 音乐标注系统的流程一 图2 音乐标注系统的流程二

3 关键技术

3.1 音符切分

音符切分的主要工作是检测一个音符的起始点和持续时间或终止点等信息,以此来将一段音乐分割成单个音符。而其中的关键和核心就是音符的起始点检测(onset detection)技术。

通常的音符起始点检测方法是寻找音乐信号中的瞬时(transient)区域。其中瞬时区域有很多定义,比如能量突变点、信号短时频谱的变化等,由此出现了多种不同原理的检测算法,主要有基于时域的和基于频域的起始点检测方法两类。

基于时域特征的音值检测算法是最简单和常用的音值检测方法。当观察一个简单音乐信号时域上的变化时,可以非常容易地发现在起始点产生的位置上往往伴随着信号幅值的激增。基于这种观察结果的起始点检测算法均使用了跟踪信号幅值包络(amplitude envelope)的函数^[41]。通过校正和平滑音乐信号很容易得到这样一个幅值跟踪函数:

$$E_0 = \frac{1}{N} \sum_{m=-N/2}^{N/2-1} |x(n+m)|w(m) \tag{1}$$

其中: $w(m)$ 是一个长度为 N 、以 $m=0$ 为中心的窗函数。如果将局部能量作为跟踪对象,则有

$$E_0 = \frac{1}{N} \sum_{m=-N/2}^{N/2-1} |x(n+m)|^2 w(m) \tag{2}$$

找到式(2)中的突变点即可。

还有很多起始点检测算法是基于信号频谱结构的,于是定义了频谱音值检测函数。设信号 $x(n)$ 的傅里叶变换为

$$X_k(n) = \frac{1}{N} \sum_{m=-N/2}^{N/2-1} x(nh+m)w(m)e^{-2j\pi mk/N} \tag{3}$$

其中: $w(m)$ 是一个 N 点的窗函数, h 是相邻窗之间的距离。在频域中,能量的增加与瞬时的出现相关。因为信号的能量通常主要集中在低频部分,而瞬时所引起的能量变化却往往发生在高频的地方,所以可以在频谱能量相加之前对高频部分使用高的权值。Masri 提出了一种高频率内容函数^[42],此函数在出现瞬时的情况下产生尖峰值,由此检测出起始点。

另外,文献[43]比较了现有各类音符起始点检测算法,发现对于有调击弦打击类乐器(如钢琴)、有调弓弦非打击类乐器(如提琴)、无调打击类乐器(如鼓)及复杂混合乐音,加权高频成分的算法 HFC(high frequency content)和自适应统计模型 ICA(independent component analysis)算法具有更为优良的总体检测性能。文献[44]结合听觉滤波器组,改进 Mel 频率倒谱系数(mel frequency cepstrum coefficients, MFCC),提出了全相位差分 MFCC,由此得到的音符起点检测函数具有适合音乐类型广、执行效率高等优点。

3.2 多基频估计

多基频估计(multipitch estimation)技术用于识别音乐中同时发声的多个音符,是自动音乐标注系统的核心技术与主要难点。大部分音乐都是多声部的,同时有多个音符发声,存在多个基频,而音符频谱叠加后的情况相当复杂,大大增加了多基频估计的难度。

为了使算法取得更好的性能,在多基频估计中需要考虑如下因素:

a)不谐和系数(inharmonicity factor)。不谐和性(inharmonicity)是一种常见的物理现象。如果乐器的弦是理想的柔顺弦,它的 k 次谐波频率是基频的 k 倍。可事实上,包括钢琴在内的许多弦乐器的琴弦具有一定的硬度。在弦的振动过程中,回复力的一部分来源于弦的张力,另一部分来源于弦自身的硬度。而后者导致频谱中随着谐波次数的升高,谐波的频率间隔将逐渐拉大,高于理想数值而不再是基音频率的整数倍,即存在不谐和性。除琴弦材质(硬度)的因素外,钢琴等乐器还受锈蚀程度、蠕变(塑性变形)程度、弦列设计、琴码质量以及铁骨尺寸精度等诸多因素的影响,其谐波成分具有明显的不谐和性。考虑不谐和性后,谐波频率与基音频率间的关系可近似地表示为

$$f_k = k \times f_1 \sqrt{1 + (k^2 - 1)B} \tag{4}$$

其中: f_1 是基频; k 是谐波次数; B 是不谐和系数,其数值与乐器、音符有关。低音区音符的谐波数量多,且中频信号比低频信号更易被人耳感知,人耳对谐波的听觉敏感度更高些,所以,不谐和系数对低音区的音符影响较大。

b)频率重叠。理想情况下,基音频率呈有理数关系的两个音符,它们的很多谐波会重叠,即有相同的频率。在实际情况下,尽管乐器发声存在频率误差,很难发生精确的谐波重叠,但算法的频率分辨能力往往很难达到可以精确辨别这些谐波频率差别的程度,从而导致不同音符的部分频率成分仍在分析结果中出现重叠的效果。发生频率重叠时,如果把先检测出的音符基频及其谐波从信号频谱中去除,则剩余未检出的音符会由于频谱被严重破坏而难以正确分析。

c)基频丢失。它是指在音符的频谱中不存在基频成分或者基频成分的幅度远小于其他谐波的幅度。基频丢失大多发生在低音音符中。这主要是由于大多数乐器对低频振动的共振性能不好,致使低频声音信号不能像高频声音信号那样通过共振来有效地增强振动幅度,从而导致包括基频在内的低次谐波振动幅度很小,发生基频丢失。但人耳对音高的感知存在“虚音高”现象,即对于同时发声的、频率差距近似等差分布的多个声音,音高的感知受频率差值的影响。因此,当发生基频丢失时,从人耳音高感知的效果上看,仍然需要将丢失的基频检测出来。而一些算法在进行多基频估计时都将频谱中的频率峰值作为候选基频,默认为基频存在并且幅度较高。因此,当音符发生基频丢失时,这些算法都将难以正确检测。

d)谐波丢失。与基频丢失相反,大多数乐器的高音区音符的频谱中基频是最主要的成分,相比之下谐波的幅度很低甚至没有谐波出现。

e)频率误差。一方面,许多钢琴家和指挥家使用的音符频率与国际标准不同。例如,为了使演奏的音乐更加明亮、铿锵,故意调高音符的频率。另一方面,在使用过程中,由于受温

度、湿度、使用强度等因素的影响,乐器的发声频率也会出现偏移,产生误差。如果这两方面因素产生的误差具有相同的误差偏移方向,累计误差会更大。再加上不谐和系数的影响,频谱中的谐波会具有很大的频率偏移。

近年来许多研究者在多基频估计方面做了大量的研究,也取得了不少成果。

文献[45]采用了一种时域分析方法,给出了信号的一个产生式模型:

$$y(t) = \sum_{n=1}^N \sum_{m=1}^{M_n} [a_{n,m} \cos(m\omega_n t) + b_{n,m} \sin(m\omega_n t)] + e(t) \quad (5)$$

其中: N 是同时出现的音符个数; M_n 是声音的谐波个数; ω_n 是第 n 个声音的基频; $a_{n,m}$ 和 $b_{n,m}$ 一起决定了每个谐波的幅度和相位; $e(t)$ 是残余噪声项。接下来需要从观测 $y(t)$ 出发来估计式(5)中的所有参数,利用音乐信号的先验知识和贝叶斯方法来估计。文中定义了一个在模型参数给定时关于 $y(t)$ 的似然函数,然后精心选择了先验分布的参数。对于一个输入信号来说,它首先被分成小的片段,然后每个小片段的信号模型参数在时域中得到估计。该方法从理论上非常漂亮,但对于实际音乐信号,其参数空间非常大,后验分布也存在大量的局部极值。文中使用可逆跳转马尔科夫链蒙特卡罗方法(Markov chain Monte Carlo, MCMC)来对后验概率采样,这导致算法非常慢。尽管如此,据文献记载该算法的性能对于音符个数不超过三个的情况还是相当鲁棒的。

此外,相当多的多基频估计算法都是基于频域的迭代算法。在此类算法中,首先要估计声音的数目;然后主导基频被预测,再从混合声音的频谱中减去对应于主导基频的声音;最后对剩余声音再预测主导基频。如此进行反复迭代。

Goto^[46]用加权混合音调模型(tone model)来建模音乐信号的短时频谱。每个音调模型包含固定数目的谐波成分,每个谐波成分则用中心位于基频整数倍位置的高斯分布来建模。Goto设计了一个计算有效的期望最大化(EM)算法来迭代地更新音调模型和它们的权重。他还用一个多代理结构来跟踪连续帧的基频权重。其算法成功地跟踪了真实音乐唱片中的旋律和基音线条。尽管这个系统整体来说比较复杂,其核心EM算法却是容易实现的。该算法估计了所有基频的权重,但他在实验中通常只估计出那个首要基频。

Kameoka等人^[47]的算法继续了Goto的音调模型的思路,但他们不是对音调模型中的每个谐波单独建模,而是对一个音调模型的一组谐波统一建模。他们假定每个音符都有其独特而稳定的谐波结构,即各次谐波之间的相对幅度关系。虽然音符的幅度会发生变化,但这个谐波结构是比较稳定的。他们用一个带约束的混合高斯模型来建模谐波结构。这里的约束就是谐波之间的频率关系。可以看到经过这样的约束,系统待估计的参数就减少了很多。对于音符的发音过程,他们也用混合高斯模型建模。这样在频域建立了信号的产生式模型,最后使用期望最大化算法来求解参数。

Poliner等人^[48]的算法采取了与其他方法完全不同的思路,他们把音高估计问题看做一个多标号的分类问题。他们把每个信号帧看做一个样本,其中的基频看做标号,特征选取的是该帧信号的频谱。然后对于每个基频,收集含有这个基频的所有样本,训练一个一对多(one versus all)的支持向量机(support vector machine, SVM)分类器。乐器音符有几十个(钢琴键

盘有88个按键),因此需要训练出几十个分类器。对于测试样本,每一个分类器都会对其作二值分类。最后音高估计结果就是这些分类器分类结果的综合。他们在钢琴信号上作了测试,取得了不错的结果。该算法思路很新颖,但它面临两个问题:a)多标号分类问题的参数空间实际上是很大的,采用一对多分类器处理虽然极大地缩小了参数空间,但却割裂了基频分类器之间的联系,实际上多音音乐的音符之间是有很密切的联系;b)即便这样,还是需要训练几十个分类器,这个计算量也是很大的。

文献[49]提出了一种通过构建基于音乐场景自适应的谐波字典来完成无监督的音符事件检测的方法。在此方法中,乐器的先验知识并非是从此类乐器演奏的其他音乐文件中获得的,而是直接从待分析的音乐文件中去获取。因为即使是同一个乐器,也会因为演奏环境、演奏家的演奏风格、麦克风的位置等因素造成其谐波结构的改变,如果不加区分地统一处理,势必会影响识别结果。文中通过分析每个音符对人耳的感知特性、频谱平滑性、稳定性以及相同音符在不同音乐帧的距离等特征,获取每个音符的谱包络,以此来达到无监督和自适应的目的。但是目前的算法还不能很好地处理多个音符的情况。

文献[50]提出了一种通过分析复音音乐的声音信号的基于帧的多基频估计系统。文中从谐和乐器的物理特性出发,充分考虑其谐和性、频谱平滑性和同步性,重点解决了噪声估计、让谐波匹配适应不和谐的谐音、重叠部分的处理、预防次谐波和过谐波等关键问题,在实验中取得了不错的效果。

文献[51]提出了一种非迭代的多基频估计算法。算法首先作一个初步的预测,期望能发现所有的候选基频,然后按照频谱不规律性和谐音结构的知识消除错误预测。与迭代方法相比,文中介绍的方法有着明显的速度优势。另外,此系统还采用了一种快速时频分析法——回声器时频分析。由于此系统从时频分析、音符切分和多基频估计等各个阶段都仔细考虑了计算效率问题,以至于该系统相比别的音乐记谱系统在保持良好性能的情况下有着明显的速度优势。但此系统完全采用信号处理的方法,没有结合任何音乐学知识,如果能结合一些音乐学知识,其性能应该还有所提高。

4 结束语

目前的自动标注系统能够较好地完成实验室条件下的单音音乐的标注工作,但要真正面向应用,还有大量工作需要完成。目前的研究成果大多只针对某种情况下的自动音乐识谱取得了不错的成就,如针对钢琴或吉他等,还没有一种比较通用的方法。虽然目前国际上的音频处理研究成果能够从复杂的音乐中滤除背景音乐,保留人声,但都是在基于一定信噪比的前提下所得到的结果,而从任意音乐中较准确地分离人声和背景音乐、分离背景音乐中的多种音调成分仍然是一个难题,也说明这是一个非常有前景的方向。同时在音乐中有很多非常个性化的表现手法,如颤音(vibrato)、滑音等,在目前的自动识谱系统中仍然无法实现,故仍然是一个有前景的方向。再者自动识谱的研究过程中所产生的算法和思想大多可以直接用于音乐检索、音乐情景分析等热点领域,说明自动识谱是一个非常具有前景的研究领域。

参考文献:

- [1] 刘林. 自动音乐识谱系统中的音符检测与流派分类[D]. 成都: 电子科技大学, 2008.
- [2] 张一彬, 周杰, 边肇祺, 等. 基于内容的音频与音乐分析综述[J]. 计算机学报, 2007, 30(5): 712-728.
- [3] CASEY M A, VELTKAMP R, GOTO M, *et al.* Content-based music information retrieval: current directions and future challenges[J]. *Proceedings of the IEEE*, 2008, 96(4): 668-696.
- [4] BROSSIER P, SANDLER M, PLUMBLEY M. Real time object based coding[C]//Proc of the 114th Convention. 2003.
- [5] PLUMBLEY M D, ABDALLAH S A, BELLO J P, *et al.* Automatic music transcription and audio source separation[J]. *Cybernetics and Systems*, 2002, 33(6): 603-627.
- [6] 刘伟. 音乐音符识别的方法[D]. 长春: 吉林大学, 2008.
- [7] GOTO M, MURAOKA Y. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions[J]. *Speech Communication*, 1999, 27(3/4): 311-335.
- [8] SMITH L M. Listening to musical rhythms with progressive wavelets [C]//Proc of IEEE TENCON Digital Signal Processing Applications. 1996: 508-513.
- [9] SU B, JENG S K. Multi-timbre chord classification using wavelet transform and self-organized map neural networks[C]//Proc of IEEE International Conference on Alousitics, Speech, and Signal Processing. 2001: 3377-3380.
- [10] MOORER J A. On the transcription of musical sound by computer [J]. *Computer Music Journal*, 1977, 1(4): 32-38.
- [11] MAHER C F. Evaluation of a method for separating digitized duet signals[J]. *Journal of the Audio Engineering Society*, 1990, 38(12): 956-979.
- [12] KATAYOSE H, INOKUCHI S. The Kansei music system[J]. *Computer Music Journal*, 1989, 13(4): 44-53.
- [13] KASHINO K, TANAKA H. A sound source separation system with the ability of automatic tone modeling [C]//Proc of International Computer Music Conference. 1993: 248-255.
- [14] MARTIN K D. Automatic transcription of simple polyphonic music: robust front end processing [R]. [S. l.]: MIT Media Laboratory, 1996.
- [15] MAROLT M. Transcription of polyphonic piano music with neural networks[C]//Proc of the 10th Mediterranean Electro Technical Conference. 2000: 512-515.
- [16] ABDALLAH S, PLUMBLEY M. Polyphonic transcription by non-negative sparse coding of power spectra[C]//Proc of the 5th International Conference on Music Information Retrieval. 2004: 318-325.
- [17] VINCENT E, RODET X. Music transcription with ISA and HMM [C]//Proc of Lecture Notes in Computer Science. 2004: 1197-1204.
- [18] BELLO J P, DAUDET L, SANDLER M B. Automatic piano transcription using frequency and time-domain information [J]. *IEEE Trans on Audio, Speech, and Language Processing*, 2006, 14(6): 2242-2251.
- [19] KLAURI A. Signal processing methods for the automatic transcription of music [D]. Tampere: Tampere University of Technology, 2004.
- [20] RYYNANEN M, KLAURI A. Polyphonic music transcription using note event modeling[C]//Proc of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2005: 319-322.
- [21] CEMGIL A T, KAPPEN H J, BARBER D. A generative model for music transcription[J]. *IEEE Trans on Audio, Speech and Language Processing*, 2006, 12(2): 679-694.
- [22] PEELING P H, LI C, GODSILL S J. Poisson point process modeling for polyphonic music transcription [J]. *Journal of the Acoustical Society of America Express Letters*, 2007, 121(4): EL168-EL175.
- [23] GOLDSTEIN J. An optimum processor theory for the central formation of the pitch of complex tones[J]. *Journal of the Acoustical Society of America*, 1973, 54(6): 1496-1516.
- [24] De ALAIN C. Pitch perception model[C]//Proc of ICASSP IEEE International Conference on Acoust; Speech Signal Process. 1986: 897-900.
- [25] De ALAIN C, HIDEKI K. Multiple period estimation and pitch perception model [J]. *Speech Communication*, 1999, 27(3): 175-185.
- [26] MAROLT M A. A connectionist approach to automatic transcription of polyphonic piano music [J]. *IEEE Trans on Multimedia*, 2004, 6(3): 439-449.
- [27] ZHOU R, ZOIA G. Polyphonic music analysis by signal processing and support vector machines[C]//Proc of the 8th International Conference on Digital Audio Effects. 2005: 1-9.
- [28] PERTUSA A, INESTA J. Polyphonic music transcription through dynamic networks and spectral pattern identification [C]//Proc of International Conference on Artificial Neural Network in Pattern Recognition Acoustics. 2003: 19-25.
- [29] 郑贵滨. 基于内容的音频信息检索技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2006.
- [30] CHANG W C, SU A W Y. A novel recurrent network based pitch detection technique for quasi-periodic/pitch varying signals [C]//Proc of IEEE International Joint Conference on Neural Networks. 2002: 816-821.
- [31] KASHIO K, MURASE H. Music recognition using note transition context [C]//Proc of IEEE ICASSP'98. 1998: 3593-3596.
- [32] BLAEK T R, DONOHUE K D. Pitch determination of music signal using the generalized spectrum [C]//Proc of IEEE Southeast Conference. 2000: 104-109.
- [33] ZHU Yong-wei, KANKANHALLI M S, GAO Sheng. Music key detection for musical audio [C]//Proc of the 11th International Multimedia Modeling Conference. 2005: 30-37.
- [34] ZENG Yu-min, WU Zhen-yang. Speech pitch detection algorithm based on modified cepstrum model[J]. *Journal of Nanjing University of Science and Technology*, 2007, 31(4): 503-508.
- [35] BOTROS N M, ADAMJEE R S. Speech-pitch detection using maximum likelihood algorithm [J]. *Control and Intelligent Systems*, 2002, 30(2): 52-58.
- [36] 王晓伟, 林锁. 基于独立分量分析的混合声音信号分离[J]. 兵工自动化, 2007, 26(6): 48-50.
- [37] 李辉, 戴蓓, 陆伟. 基于前置滤波和小波变换的带噪语音基音周期检测方法[J]. 数据采集与处理, 2005, 20(1): 100-104.
- [38] 钟静, 傅彦. 基于快速ICA的混合语音信号分离[J]. 计算机应用, 2006, 26(5): 1120-1124.
- [39] LIU Ruo-lun, LI Su-ping. A review on music source separation [C]//Proc of IEEE Youth Conference on Information Computing Telecommun. 2009: 343-346.
- [40] EMMANUEL V. Musical source separation using time-frequency source priors[J]. *IEEE Trans on Audio Speech and Language Processing*, 2006, 14(1): 91-98.
- [41] LI Yi-peng, WANG De-ling. Separation of singing voice from music accompaniment for monaural recordings[J]. *IEEE Trans on Audio, Speech, and Language Processing*, 2007, 15(4): 1475-1487.
- [42] LU Lie, JIANG Hao, ZHANG Hong-jiang. A robust audio classification and segmentation method [C]//Proc of the 9th ACM International on Multimedia Conference. 2001: 203-211. (下转第1210页)

- [14] KO T, RAYTHEON C, ARLINGTON V A. A survey on behavior analysis in video surveillance for homeland security applications[C]//Proc of the 37th Applied Imagery Pattern Recognition Workshop. 2008.
- [15] HU Wei-ming, TAN Tie-niu, WANG Ling, *et al.* A survey on visual surveillance of object motion and behaviors[J]. *IEEE Trans on Systems, Man, and Cybernetics—Part C*, 2004, 34(3): 334-352.
- [16] 侯志强, 韩崇昭. 视觉跟踪技术综述[J]. *自动化学报*, 2006, 32(4): 603-617.
- [17] BETSER A, VELA P, TANNENBAUM A. Automatic tracking of flying vehicles using geodesic snakes and Kalman filtering[C]//Proc of the 43rd IEEE Conference on Decision and Control. Piscataway: IEEE Press, 2004: 1649-1654.
- [18] MICHAEL K, ANDREW W, DEMETRI T. Snakes: active contour models[C]//Proc of the 1st International Conference on Computer Vision. Boston: Kluwer Academic Publishers, 1988: 259-268.
- [19] PARAGOPS N, DERICHE R. Geodesic active contours and level sets for the detection and tracking of moving objects[J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2000, 22(3): 266-280.
- [20] HU Wei-ming, XIAO Xue-juan, XIE Dan, *et al.* Traffic accident prediction using 3-D model-based vehicle tracking[J]. *IEEE Trans on Vehicular Technology*, 2004, 53(3): 677-693.
- [21] ROBERT K. Video-based traffic monitoring at day and night[C]//Proc of the 12th International IEEE Conference on Intelligent Transportation Systems. 2009: 285-290.
- [22] BRAGATTO T A C, RUAS G I S, BENSO V A P, *et al.* A new approach to multiple vehicle tracking in intersections using Harris corners and adaptive background subtraction[C]//Proc of IEEE Intelligent Vehicles Symposium. 2008: 548-553.
- [23] 王亮, 胡卫明, 谭铁牛. 人运动的视觉分析综述[J]. *计算机学报*, 2002, 25(3): 225-233.
- [24] CHENG Hui, YANG Chang-jiang, HAN Feng, *et al.* HO2: a new feature for multi-agent event detection and recognition[C]//Proc of Computer Vision and Pattern Recognition Workshop. 2008.
- [25] PICIARELLI C, MICHELONI C, FORESTI G L. Anomalous trajectory patterns detection[C]//Proc of International Conferences on Pattern Recognition. 2009.
- [26] MYERS C, RABINIER L, ROSENBERG A. Performance tradeoffs in dynamic time warping algorithm for isolated word recognition[J]. *IEEE Trans on Acoustics, Speech, and Signal Processing*, 1980, 28(6): 623-635.
- [27] KIM C, SEO K D. Robust DTW-based recognition algorithm for hand-held consumer devices[C]//Proc of International Conference on Consumer Electronics. 2005: 433-434.
- [28] BOBICK A F, WILSON A D. A state-based approach to the representation and recognition of gesture[J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 1997, 19(12): 1325-1337.
- [29] 施毅. 基于视频的交通事件检测方法研究[D]. 南京: 东南大学, 2007.
- [30] LI Xiao-kun, PORIKLI F M. A hidden Markov model framework for traffic event detection using video features[C]//Proc of ICIP. 2004: 2901-2904.
- [31] BASHIR F I, KHOKHAR A A, SCHONFELD D. Object trajectory-based activity classification and recognition using hidden Markov models[J]. *IEEE Trans on Image Processing*, 2007, 16(7): 1912-1919.
- [32] NAIR V, CLARK J J. Automated visual surveillance using hidden Markov models[C]//Proc of the 15th Vision Interface Conference. Calgary: [s. n.], 2002: 88-92.
- [33] KAMIJO S, HARADA M. Event recognitions from traffic images based on spatio-temporal Markov random field model[C]//Proc of the 8th World Congress on Intelligent Transport Systems. 2001: 1-12.
- [34] KAMIJO S, HARADA M. Traffic monitoring and accident detection at intersections[J]. *IEEE Trans on Intelligent Transportation Systems*, 2000, 1(2): 108-118.
- [35] SAUNIER N, SAYED T. Automated road safety analysis using video data[C]//Proc of the 86th Annual Meeting of the Transportation Research Board. 2007.
- [36] SAUNIER N, SAYED T. A feature-based tracking algorithm for vehicles in intersections[C]//Proc of the 3rd Canadian Conference on Computer and Robot Vision. [S. l.]: IEEE Press, 2006.
- [37] SAUNIER N, SAYED T. Clustering vehicle trajectories with hidden Markov models application to automated traffic safety analysis[C]//Proc of International Joint Conference on Neural Networks. 2006: 4123-4138.
- [38] KOHONEN T. The self-organizing maps[M]. Berlin: Springer-Verlag, 1995.
- [39] OWENS J, HUNTER A. Application of the self-organizing map to trajectory classification[C]//Proc of the 3rd IEEE International Workshop on Visual Surveillance. [S. l.]: IEEE Press, 2000: 77-83.

(上接第1205页)

- [43] BELLO J P, DAUDET L, ABDALLAH S, *et al.* A tutorial on onset detection in music signals[J]. *IEEE Trans on Speech and Audio Processing*, 2005, 13(5): 1035-1047.
- [44] 关欣, 李镔, 田洪伟. 基于差分全相位 MFCC 的音符起点自动检测[J]. *计算机工程*, 2010, 36(11): 25-29.
- [45] DAVY M, GODSILL S J, IDIER J. Bayesian analysis of western tonal music[J]. *Journal of Acoustical Society of America*, 2006, 119(4): 2498-2517.
- [46] GOTO M. A predominant-F0 estimation method for real-world musical audio signals: MAP estimation for incorporating prior knowledge about F0s and tone models[C]//Proc of Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis. 2001.
- [47] KAMEOKA H, NISHIMOTO T, SAGAYAMA S. A multipitch analyzer based on harmonic temporal structured clustering[J]. *IEEE Trans on Audio, Speech and Language Processing*, 2007, 15(3): 982-994.
- [48] POLINER G E, ELLIS D P W. A discriminative model for polyphonic piano transcription[J]. *EURASIP Journal on Applied Signal Processing*, 2007, 2007(1): 154.
- [49] CARABIAS-ORTI J J, VERA-CANDEAS P, C  NDAS-QUESADA F J, *et al.* Music scene-adaptive harmonic dictionary for unsupervised note-event detection[J]. *IEEE Trans on Audio, Speech and Language Processing*, 2010, 18(3): 473-486.
- [50] YE H C, ROEBEL A, RODET X. Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals[J]. *IEEE Trans on Audio, Speech and Language Processing*, 2010, 18(6): 1116-1126.
- [51] 周若华, 颜永红. 一个快速自动音乐记谱方法[J]. *声学学报*, 2010, 35(2): 282-286.