基于内容的音乐节拍跟踪

陈 哲, 许洁萍
(中国人民大学信息学院计算机系, 北京 100872)

摘 要: 节奏是音乐的三大要素之一, 对其准确的分析和提取具有重要的研究意义. 节奏特征主要分为音乐节拍和速度. 本文首先提出了一种利用自相关相位-熵序列分析音乐节拍结构及音乐速度的方法. 利用该方法对 50 首流行歌曲及 50 首纯乐器音乐速度的分析结果可达到 97%; 在速度分析结果基础上, 文中还给出了节拍点求解过程的近似贝叶斯模型, 使得节拍点序列在整体上与音乐信号的长时速度保持一致; 文中在最后给出了利用动态规划思想进行音乐节拍跟踪的新方法, 完成了音乐节拍跟踪实验, 并通过与其它实验的结果比较, 验证了算法的有效性.

关键词: 节奏特征; 节拍跟踪; 速度提取; 自相关-熵序列; 动态规划
中图分类号: TP37 文献标识码: A 文章编号: 0372-2112 (2009) 4A-156-05

Content Based Music Beat Tracking

CHEN Zhe, XU Jie-ping
(*Computer Department of School of Information, Renmin University, Beijing* 100872, *China*)

Abstract: Rhythm is one of the three main factors of music. It's of great importance to analyze and extract rhythm features from music signal which mainly includes beat and tempo. In this paper, firstly we introduce a method which uses autocorrelation phase-entropy, to analyze meter and tempo of music. With this method we get a 97 percent of accuracy degree in music tempo induction on a database of 100 songs(including 50 popular and 50 Chinese folk); Then based on the result of tempo induction, we propose a roughly Bayesian model to keep the beats sequence consistent with the global tempo of music in the music beats locating; Finally we introduce a new beat tracking algorithm using dynamic programming to finish the beat tracking. By comparing its result with others', this algorithm proves to be very effective.

Key words: rhythm features; beat tracking; tempo induction; autocorrelation phase-entropy; dynamic programming

1 引言

《韦氏大词典》将节奏定义为"音乐的一个方面它包括了与乐音向前进行有关的所有因素(如重音、节拍和速度)". "进行"在这里是个关键词, 节奏即是音乐的进行. 我们用脚拍打出的就是节奏, 它是"曲调的快慢缓急". 节奏是所有音乐都不可缺少的组成部分, 它就像音乐的脉搏, 音乐如果没有了节奏就失去了表达音乐思维的能力. 同时节奏作为人理解音乐的基础, 也是许多音乐分析、研究、应用的基础.

节奏最明显的表现就是节拍, 因此节拍跟踪即根据人的听感将一段音乐的拍子显现出来, 在计算机音乐研究中引起了很多学者的关注. Goto 和 Muraoka 较早的提出了一系列的以声学信号为输入的节拍跟踪模型, 首先基于鼓点提出了一个节拍跟踪模型[1], 后来又开发了一个基于和声变化监测的系统[2], 但他们只在一首歌的 40 个音乐片段上做了实验, 而且没有给出每个节拍点的精确实验结果. Scheirer 提出使用音符切分算法, 像人类听觉系统一样先分频带处理然后再组合处理来进行节拍跟踪, 并在 7 首音乐上通过实验[3]证明了对音乐信号分频带处理可以得到较好的节拍跟踪结果, 但没有进行大数据量的分析验证. 2006 年, Ellis 使用 40 维梅尔频率滤波器组求取了音符起始点的包络曲线, 在包络曲线上采用动态规划的方法进行了节拍点的提取[4], 根据该想法完成的系统, 参加了音乐信息检索国际会议 MIREX-06 的实验比赛, 取得了 77% 准确率的最好实验结果.

纵观已有的研究方法, 研究者提出的研究模型都只考虑了信号的频谱能量, 而没有考虑信号频谱的相位. 但实际上, 音乐的节奏信息就是音乐长时间内的重复, 而音乐的自相关信号中包含有大量的周期信号信息, 所以很自然的考虑可以从自相关信号中提取节奏信息. 文献[5]的实验也证明这一点: 一个单纯的由打击乐器组成, 节拍明显的音乐信号, 从它的自相关信号中确实能

收稿日期: 2008-11-22; 修回日期: 2008-12-25
基金项目: 国家自然科学基金(No. 60873098); 中国人民大学"985 工程"自由探索项目(No. 21356234)

很容易提取出节奏信息. 因此, 本文提出了结合音乐信号频谱的相位信息来分析音乐片段, 进行节拍跟踪的新方法.

文中首先根据节拍估计结果得到音乐的节拍结构: 属于两拍系列还是三拍系列, 然后利用节拍结构完成了速度估计, 提取出音乐的两个候选速度, 最后利用速度结果实现了节拍跟踪. 本文第 2 部分论述了音乐中自相关相位- 熵序列的提取方法及节拍结构的提取和速度的估计方法; 第3 部分则详细介绍了在速度估计结果基础上, 利用动态规划思想进行节拍跟踪的方法, 文章最后给出了实验结果及总结分析.

2　节拍与速度估计

节拍就是乐曲中表示固定单位时值和强弱规律的组织形式, 亦称拍子. 它有两个特性: 周期性和连续性.

节拍周期性表现为节拍结构(meter), 是乐曲中周期性出现的节奏序列. 节拍连续性表现为音乐的平均速度, 其单位为 BPM (Beat Per Minute: 每分钟多少拍). 因此, 节拍跟踪与音乐的节拍结构和音乐的平均速度有着密不可分的联系. 节拍提取与速度估计是节拍跟踪的前提和基础. 图 1 给出了节拍跟踪实验的整体流程图. 实验分为: 预处理、自相关相位- 熵序列的提取、节拍、速度提取及节拍跟踪几大部分.

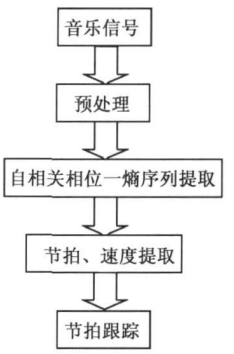

图1　节拍跟踪实验流程图

2.1　预处理

为了方便获取歌曲的标准节奏, 我们选取的实验材料是已知曲谱的 50 首流行音乐和 50 个纯乐器的民族歌曲的片段, 片段长度不等, 最短为 5s, 最长为 30s. 文件格式均为双通道立体声, 44. 1KHz 采样率的 mp3 文件.

实验中, 我们首先将双通道的 mp3 文件转化成了单通道的 wav 格式文件. 由于音乐的节奏信息可以理解为音乐在一个较长时间内显示的特征, 因此, 为了简化运算, 我们减小了信号在时间上的分辨率, 对信号进行降采样, 将采样率有原有的 44. 1KHz 降为 1000Hz.

2.2　自相关相位-熵序列

现实中大部分的音乐不仅包含了打击乐器, 还包含各种弦乐器, 使得很大一部分音乐的节拍并不那么明显, 所以单纯的由自相关信号中提取节奏信息还不能很好的得到节奏信息. 针对自相关信号不能反应相位特征的缺点, 实验中我们采用了自相关相位矩阵及自相关相位-熵序列[6]来进行节奏信息的分析.

2.2.1　自相关相位矩阵

它既包含了周期信号信息, 又包含了信号的相位信息. 具体计算方法如式(1)所示.

$$A(\varphi, l)=\left\{\sum_{i=0}^{[N/l]-1} * X(l * i+\varphi) * X(l * (i+1)+\varphi)\right\}_{\varphi=0}^{l-1} \tag{1}$$

其中: l 代表偏移点, φ 代表同一个偏移点的不同相位).

2.2.2　自相关相位-熵序列

自相关相位-熵序列是自相关相位矩阵中每一个偏移点在不同相位上的信息熵[7]. 具体计算方法如下:

首先对于自相关相位矩阵中每一个偏移点将其所有不同相位相加(式(2)), 即将上一步所得到的自相关相位矩阵按列求和, 得到 Sum 序列:

$$Sum_1=\sum_{i=0}^{N} A_{li} \tag{2}$$

其中: $l=1, 2 \cdots L$.

然后利用式(2) 求出自相关每一个偏移点关于相位的熵, 得到熵序列 $\boldsymbol{H}$.

$$H_l=-\sum_{i=0}^{N}(A_{li}/Sum_l) * \log_2(A_{li}/Sum_l) \tag{3}$$

其中: $l=1, 2 \cdots L$.

最后将得到的熵序列与降采样信号的自相关序列相乘(式(4)), 得到自相关相位-熵序列$\boldsymbol{AH}$. 其中$\boldsymbol{Af}$ 为降采样信号的自相关信号.

$$\boldsymbol{AH}=\boldsymbol{Af} * \boldsymbol{H} \tag{4}$$

因此, 自相关相位-熵序列既能反应音乐信号的节奏信息, 同时也增强了其抗干扰性及鲁棒性.

2.3　节拍结构分析及速度提取

根据以上讨论可知, 节拍跟踪与节拍结构以及速度有着很强的联系, 因此本节首先分析节拍结构, 然后利用节拍结构提取音乐速度.

2.3.1　节拍结构分析

根据自相关序列的原理, 若每 l 采样帧为一个节拍, 那么每 $2l$ 个采样帧为一个小节, $4l$ 个采样帧为两个小节, 均作周期性的重复, 所以如果节拍长度为 l 采样帧, 音乐为两拍系列, 那么在 $2l$, $4l$, $6l$ …等帧处的自相关序列的值都应该较大. 同理, 若节拍长度为采样帧, 音乐为三拍系列, 那么在 $3l$, $6l$, $9l$ …等帧处的自相关序列的也出现峰值. 根据这一特性, 我们提出了两个预测音乐节拍的公式:

$$Duple_l=AH(l)+AH(2l)+AH(4l)+\cdots+AH(2kl) \tag{5}$$

其中: $l=201, 202 \cdots 4000$, k 为正整数, 且 $2kl \leqslant 4000$.

$$Triple_l=AH(l)+AH(3l)+AH(6l)+\cdots+AH(3kl) \tag{6}$$

其中: $l = 201, 202 \cdots 4000$, k 为正整数, 且 $3kl \leqslant 4000$.

根据式(5), 可以求出对于节拍长度为 l 个采样帧, 每一小节有两拍, 即属于两拍系列的得分. 并选取所有点中最高的得分作为两拍系列的最终得分 S_{duple}.

根据式(6), 求出对于节拍长度为 l 个采样帧, 每一小节有三拍, 即属于三拍系列的得分 *Triple*. 然后选取所有点中最高的得分作为三拍系列的最终得分 S_{triple}.

选择 S_{duple}和 S_{triple}中较大的一个作为最后得分, 并且根据这个得分高的作为最后的节拍.

实验中, 我们只取自相关相位–熵序列在 200~ 4000 之间的点, 这是因为音乐的速度大都分布在 20BPM~ 300BPM 之间, 即节拍的长度分布在 200~ 3000ms 之间, 转换成采样帧的则为 210~ 3150 采样帧之间.

2.3.2 速度估计

在 2.2.2 得到的自相关相位–熵序列上做峰值检测, 即得到每一拍所占的采样帧数目. 具体实验步骤如下:

加高斯窗: 在自相关相位–熵序列的对数域上面加高斯窗, 由于大部分的音乐速度均在 120BPM 左右, 所以我们将高斯窗的中心选为 120BPM 所对应点, 即 60 × 1000/120= 500 帧所对应点, 然后根据多次实验对比选择 δ= 100 帧.

计算第一候选速度: 在加了高斯窗的自相关相位–熵序列上求出最大峰值点所在的采样帧数 m, 即为每一拍所占采样帧数, 由于采样帧间隔为 1ms, 所以每一拍的时间为 m 毫秒. 所以一分钟的拍数为 60* 1000/m= 60000/m, 即速度为 60000/m BPM.

求第二候选速度: 应用 2.3.1 中预测的节拍结构求第二侯选速度. 即如果是两拍系列的话, 即在加了高斯窗的自相关相位–熵序列上求出 2 × m 和 m/2 附近的最大峰值, 选取其中的最大峰值所在的点 n 为第二侯选速度决定的拍子长度, 同时求出第二侯选速度 60000/n BPM. 同样, 如果上一步求得的节拍结构是三拍系列的话, 则求出 3× m 和 m/3 附近的最大峰值, 选取其中的最大峰值所在的点 n 第二侯选速度决定的拍子长度, 同时求出第候选速度 60000/n BPM.

3 节拍跟踪

节拍是音乐长时间内表现出的连续性的行为, 这里长时间我们定义为 5s 以上. 相邻两个节拍之间有着紧密的联系, 通过相邻节拍的间距可以控制音乐的速度. 但是为了音乐的听觉需要, 节拍间距并不是完全相等, 而是有着细微差别, 起伏不平的. 所以总体说来, 节拍在长时间上表现出的速度基本恒定, 但是每两个节拍之间的间隔, 即瞬时速度的倒数却都是不一样的. 一般的节拍跟踪方法只考虑了瞬时的波动性, 并没有考虑节拍间距的长时连续性. 本文首先考虑瞬时波动性, 计算 onset 曲线, 然后考虑长时连续性, 计算了累积得分曲线, 最后得出最佳节拍序列.

3.1 计算 onset 曲线

具体步骤如下:

(1) 将 2.1 中降采样后的信号分帧求取每帧的能量, 其中帧长为 32ms(32 个采样点), 帧移为 4ms

(2) 对于每一帧, 求取它与后面一帧的能量差的绝对值, 最后一个帧不计算. 这样就求得了一个粗略的 onset 曲线. 其每个点之间的距离为帧移 4ms, 相当于采样率降为 250Hz.

然而, Onset 曲线仅仅体现了节拍间距的瞬时波动性. 为了保证节拍间距的长时连续性, 本文利用了第 2 部分中提出的速度并与动态规划[4] 方法进行了结合.

3.2 近似贝叶斯模型

为了保证节拍间距的长时连续性, 我们利用长时平均速度, 建立了一个近似贝叶斯模型. 具体过程如下:

(1) 选取上文求出的两个候选速度中较快的速度 Tempomax, 按照公式 60 × 1000/Tempomax (毫秒) 求出总体片段的平均节拍间距 pd.

(2) 将 onset 曲线卷积上一个高斯窗, 其中高斯窗的中心为 0, δ 为pd/32, 窗长为$(-pd, pd)$, 间隔步长为也 pd/32, 所以窗内共有 65 个点.

(3) 选取卷积的结果中间的部分, 即除去卷积带来的信号加长的那部分: 前 32 个点和后 33 个点, 作为节拍跟踪时各个采样帧的瞬时得分曲线 Local Score, 下文简称为 L.

(4) 对于每个采样帧 t, 对 $L(t-2*pd, t-pd/2)$ 在对数域上加上一个高斯窗, 并求出其中峰值以及峰值所在的点 t_1. 高斯窗的中心为点 $t-pd$, 相当于一个近似的贝叶斯模型[8]. 因为按照 2.3.2 得到的速度, 当 t 点为节拍点时, $t-pd$ 点是最有可能的前一个节拍点, 因此, 高斯窗相当于模型的先验概率, 窗中心在 pd, 窗长为$(pd/2, 2*pd)$, δ 作为参数输入. 经过实验比较我们选取 δ 等于 6.

3.3 使用动态规划进行节拍跟踪

若已求得了一个节拍点, 则节拍跟踪, 就是要在已知此节拍点的基础上, 求得下一个节拍点. 由上一步可以看出, 将平均节拍间距作为先验概率, 结合瞬时得分, 我们可以求得相对于节拍点 t, 在$(t-2*pd, t-pd/2)$ 这些点为节拍点的得分, 即条件概率. 选择其中最大的一个点(即上步中的求峰值 *peak* 及其所在点) 作为这个点的前驱点. 这样对于每个点, 都可以求出相对于这个点为节拍点的前驱节拍点. 然而对于每个点, 都有可能是后面多个点的前驱节拍点. 这样, 如果从头往后搜寻节拍点就会遇到很多问题. 因此, 实验中我们采用的是

一个从后向前的求解过程.

很明显, 这是一个多阶段决策的问题, 每个确定的节拍点为一个阶段, 然后求取其前驱到达下个阶段, 所以我们很自然的想到了动态规划的方法. 具体算法如下:

首先令所有点的累积得分(Cumulative Score, 后文简称 CS)等于瞬时得分(式(7)), 即:

$$CS(t) = L(t) \tag{7}$$

然后对于采样点 t, 不是在 $L(t-2*pd, t-pd/2)$ 的瞬时得分上加对数高斯窗求峰值, 而是在累积得分 $CS(t-2*pd, t-pd/2)$ 上加对数高斯窗求取峰值 $peak$ 和峰值所在点 t_1. 为了平衡瞬时得分在累积得分中的影响, 将 $peak$ 乘以一个平衡因子加到当前点的瞬时得分 $L(t)$ 上(式(8)). 根据实验结果, 平衡因子取 0.7.

$$CS(t) = L(t) + 0.7 \times peak \tag{8}$$

同时记录下峰值所在的点 t_1, 作为当前点的前驱 $Pre(t)$ 即:

$$Pre(t) = t_1 \tag{9}$$

若 $CS(t_1)$ 很小, 小于瞬时得分曲线中最大值的 1%, 则将当前点的前驱设为 -1, 表示当前点前面没有前驱了.

经过这样一个动态规划的过程, 我们就得出了每个点的累积得分 $CS(t)$.

最后, 根据已求得的累积积分曲线, 我们可以求出累积得分曲线中所有的峰值. 选取其中的中值作为阀值, 然后选取累积积分曲线全部峰值之中大于阀值并且时间最靠后的一个峰值点作为该片段的最后一个节拍点. 然后根据之前求得这个点的前驱找到前一个节拍点, 就这样按照前驱一直找下去, 当前驱值为 1 时, 停止搜索节拍点, 完成节拍跟踪.

4　实验

4.1　实验数据

本实验采用的数据为 50 个不同节拍, 不同速度的流行音乐片段和 50 个纯乐器的民族歌曲片段, 片段长度不一, 最短为 5s, 最长为 30s. 实验数据的具体分析见表 1, 民族歌曲只有二拍结构的, 流行音乐则考虑了二拍和三拍的情况.

表 1　实验数据分析

	60BPM 以下	60~ 80 BPM	80~ 100 BPM	100~ 120 BPM	120BPM 以上	总体数目
两拍结构流行音乐	6	21	11	4	3	45
三拍结构流行音乐	0	2	3	0	0	5
两拍结构纯乐器民族歌曲	5	36	7	2	0	5

4.2　速度实验结果

由于节拍跟踪要基于速度, 因此有必要分析速度. 表 2 给出速度的估计结果. 我们认为速度和曲谱所标速度 ±5% 即为正确.

表 2　速度估计结果

	流行音乐	纯乐器民族歌曲	所有歌曲
第一候选速度正确	40	40	80
第二候选速度正确	7	10	17
总体正确数	47	50	97
总正确率	94%	100%	97%

实验结果表明: 100 个片段中有 80 个第一候选速度正确即正确率为 80%, 17 个第二候选速度正确, 则总体速度估计正确率为 97%. 另外, 我们发现纯乐器民族歌曲的总体正确率达到了 100%, 高于流行歌曲的 94%. 这是由于纯乐器歌曲除去了人声的影响, 突出了打击乐器的部分, 使得音乐的节奏感更加明显. 这一结果同人的听感结果也是一致的, 同时, 我们还将速度实验结果与文献[4]的速度估计结果进行了对比, 我们所有的速度结果都比其速度结果更接近于曲谱的速度.

4.3　节拍跟踪结果

在速度提取结果基础上, 我们进行了节拍跟踪标注. 图 2 为歌曲"爱情"片段提取出的节拍跟踪结果图. 图中横轴为时间, 纵轴为音乐信号包络, 竖线标注了节拍跟踪得到的节拍点. 由于人的标注结果的获得存在一定的困难, 因此, 我们与文献[4]的实验结果进行了对比. 图中上部分为本实验结果, 下部分为根据文献[4]的方法获得的实验结果. 文献[4]中, 作者对自己的实验结果与人听感获得的节拍标注结果进行了大量的比较研究, 并在 06 年的 ISMIREX 竞赛上获得了非常理想的结果. 本实验中, 歌曲的标注结果与文献[4]的结果非常相似, 并且根据之前速度提取结果的对比, 我们有理由相信我们的结果更加准确.

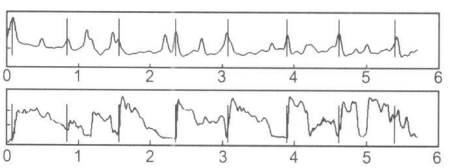

图2　歌曲片段节拍跟踪结果图

5　实验总结及展望

本实验采用了自相关相位-熵矩阵和动态规划方法分析音乐片段的节奏, 包括节拍和速度, 取得了较好的结果, 表明自相关相位矩阵中含有大量的节奏信息, 同时也说明了动态规划的方法适合用来提取音乐中的长时信息.

由于人对节拍跟踪标注结果的获取还存在一定的难度,标注结果的个性化影响也很大,因此下一步的工作:一是增加实验数据的数量和质量(增加不同风格的音乐),增加此方法的可信性.二是进一步分析音乐片段的节拍信息,做到更加准确的判断音乐的节拍,而不仅仅区别出两拍系列和三拍系列,更加准确的估计音乐速度.

参考文献:

[1] Masataka Goto, Yoichi Muraoka. A real-time beat tracking system for audio signals[A]. In Proceedings of the International Computer Music Conference[C]. San Francisco: International Computer Music Association, 1995. 171- 174.

[2] Masataka Goto, Yoichi Muraoka. A real-time beat tracking for drumless audio signals: chord change detection for musical decisions[J]. Speech Communication, 1999, 27: 311- 335.

[3] Eric D Scheirer. Tempo and beat analysis of acoustic musical signals[J]. The Journal of The Acoustical Society America, 1998, 103(1): 588- 601.

[4] Daniel Ellis. Beat tracking by dynamic programming[J]. New Music Research, Special Issue on Beat and Tempo Extraction, 2007, 36(1): 51- 60.

[5] Douglas Eck. Finding meter in music using an autocorrelation phase matrix and shannon entropy[A]. ISMIR[C]. London: University of London, 2005. 504- 509.

[6] Douglas Eck. Finding long-timescale musical structure with an autocorrelation phase matrix[J]. Music Perception, 2006, 24(2): 167- 176.

[7] C E Shannon, (1948). A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27: 379- 423, 623- 656.

[8] Roger B Dannenberg. An intelligent multi-track audio editor[A]. In Proceedings of The 2007 International Computer Music Conference[C]. San Francisco: International Computer Music Association, 2007, 2: 89- 94.

作者简介:

陈 哲 男,1984 年生,硕士研究生.研究方向:多媒体内容分析与检索. E-mail: czmailbox@gmail.com

许洁萍 女,1966 年生,博士,副研究员,硕士生导师.近期主要研究方向:多媒体信号处理、多媒体内容分析与检索、汉语语音标准研究. E-mail: xjieping@ruc.edu.cn