

流行音乐主旋律提取技术综述

李 伟^{1,2} 冯相宜¹ 吴益明¹ 张旭龙¹

(复旦大学计算机科学技术学院 上海 201203)¹

(复旦大学上海市智能信息处理重点实验室 上海 200433)²

摘 要 旋律是最重要的音乐要素之一,多应用于音乐内容分析、音乐创作、音乐教育、抄袭检测等方面。主旋律提取旨在从一段音乐中自动估计对应于主旋律单音音符序列的音高或基频。流行音乐一般属于复杂的多音音乐,因此主旋律提取面临着许多挑战。综述了主旋律提取的研究背景,将主旋律提取的典型方法分别从基于音高显著度、基于声源分离以及其他方法3个类别进行了阐述,最后介绍了主旋律提取的评价指标以及研究进展。

关键词 主旋律提取,多音音乐,音乐信息检索

中图法分类号 TP391.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.001

Review on Main Melody Extraction from Pop Music

LI Wei^{1,2} FENG Xiang-yi¹ WU Yi-ming¹ ZHANG Xu-long¹

(School of Computer Science, Fudan University, Shanghai 201203, China)¹

(Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China)²

Abstract Melody is one of the most important elements of music, with many direct and indirect applications in music content analysis, music creation, music education and protection of music intellectual properties. Main melody extraction aims to produce a sequence of frequency values corresponding to the pitches of the dominant melody from a musical recording. Due to the complexity and specificity of pop music, main melody extraction from pop music turns out to be highly challenging. This paper reviewed research works of main melody extraction, classified and summarized the methods used for main melody extraction. Finally, evaluation methodologies for main melody extraction were presented.

Keywords Main melody extraction, Polyphonic music, Music information retrieval

1 研究背景及意义

音乐是一种历史悠久的艺术形式。近二十年来,音频信息处理技术的发展和互联网的兴起,使得音乐形式逐渐从传统的磁带、黑胶唱片过渡到数字音乐。据中国互联网网络发展状况统计报告,音乐下载、在线聆听、彩铃等网络音乐服务多年来一直位居所有互联网应用的前三名。普通消费者可能会在智能手机或音乐播放器中存储成百上千首歌曲,而云端的音乐服务提供商则经常需要存储几百万首甚至上千万首歌曲。

如此海量的音乐再也无法依靠人工进行管理,需要新的方法对其进行描述、分类、搜索和交互。因此,基于内容的音乐信息检索(Music Information Retrieval, MIR)成为了热门的研究领域,并在最近二十年得到了迅速发展。

音乐包括多个形式要素(即表现手段),如旋律、节奏、和声、音色、力度、速度、调式、曲式等。其中最基本的要素是旋律和节奏。旋律与人类听觉感知相关,但是在音乐学研究中并没有清晰的定义^[1]。在MIR领域,通常认为旋律属于混合

音频信号中的单个声源,这个声源一般是人声或其中最主要的乐器。从信号处理的角度出发,音乐可以被分为单音音乐(monophonic music)和多音音乐(polyphonic music)两种。前者指在任意时刻只有一个音符发声;后者指在同一时刻有两个及两个以上的音符同时发声,音符可以来自不同乐器(例如歌声、吉他或贝斯),或同时演奏多个音符的单个乐器(例如钢琴)。旋律提取技术处理的对象是多音音乐信号,自动估计对应于旋律单音音符序列的主音高(predominant pitch)或主基频(predominant fundamental frequency, f_0)^[2](见图1)。

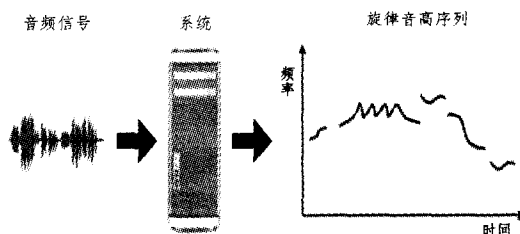


图1 旋律提取:从音频信号中提取旋律音高序列

到稿日期:2016-04-04 返修日期:2016-07-24 本文受国家自然科学基金 NSFC(61671156)资助。

李 伟(1970—),男,博士,教授,博士生导师,主要研究方向为音频信息处理、多媒体信息安全, E-mail: weilifudan@fudan.edu.cn; 冯相宜(1993—),男,硕士生,主要研究方向为音频信息处理; 吴益明(1993—),男,硕士生,主要研究方向为音频信息处理; 张旭龙(1989—),男,博士生,主要研究方向为音频信息处理。

尽管音高是一个感知概念, f_0 是一个物理量^[3], 但是术语 predominant pitch 和 f_0 在音乐计算文献中经常被等价使用。

事实上, 近年来用计算机进行旋律提取(或主基频估计)已经成为数字音乐计算领域的一个重要的研究课题^[4-6]。虽然人可以从混合音乐信号中轻松地辨别旋律, 甚至乐感良好并接受专业训练的人可以将旋律识别为乐谱, 但是由机器来执行这个任务却是复杂且具有极大挑战的。主要原因在于以下两个方面: 1) 一个多音音乐信号由录音中的所有乐器产生的声音波形叠加组成, 很多时候这些乐器同时演奏。把来自不同声源、按照和声结构高度耦合叠加在一起的频谱分开到对应的各个音符是极其困难的, 而后期混响、回声等处理会进一步增加声源的重叠, 模糊音乐信号的起止时间(note onset/offset), 使频谱分离更加困难。2) 即使已经得到音符的基频序列, 还需要判断哪些音高属于旋律, 哪些属于伴奏, 当旋律为歌声但是存在背景和声时, 检测更加困难。

音乐旋律提取技术具有许多直接和间接的应用。直接应用包括音乐识谱(music transcription)^[7]、民族音乐学(ethnomusicology)、音准分析(intonation analysis)^[8]、旋律主题和模式分析(melodic motif and pattern analysis)^[9]、电声作曲(electroacoustic composition)等计算音乐学和音乐教育领域。间接应用主要体现在其作为哼唱检索(query by humming)^[10]、翻唱检索(cover song identification)^[11]、流派分类(genre classification)^[12]、歌声消除(vocal cut)或伴奏自动产生^[13]、歌唱家识别^[14]等其他研究的重要前处理步骤。Adobe 公司开发的旋律提取技术已经被集成到其音乐处理工具 Adobe Audition 中, 但由于商业保密而无法得知技术细节。除了以上应用, 据我们所知还有一个应用即音乐借用(music borrowing)在 MIR 领域尚未被研究过。在音乐借用中, 借用和被借用歌曲之间共享一个持续时间较短的相似旋律片段。通常, 音乐界认为适当的借用和艺术创作中是允许的, 但是超过 8 小节就涉嫌抄袭侵权行为。检测歌曲之间的相似片段对于音乐作品的版权维护和盗版检测具有重要的现实意义。

2 自动旋律提取的国内外研究现状

旋律提取与单音音高估计^[15-16]和多音音高估计(multi-pitch estimation)不同^[17]。其与单音音高估计的区别在于旋律提取处理的是多音音乐信号; 与多音音高估计的区别在于旋律提取只识别旋律线的音高, 同时与其他声源的音高进行区分。

虽然旋律提取包括检测歌声旋律和主乐器演奏的旋律, 但是绝大多数算法主要关注歌声旋律的提取。原因有两点: 1) 大量的流行歌曲具有歌声旋律, 这使得歌声旋律的提取在商业上具有较强的吸引力; 2) 歌声具有不同于大多数乐器的独特性质^[18], 算法可以利用这些性质更准确地识别旋律。

2.1 基于音高显著度(pitch salience)的旋律提取方法

大多数已有的旋律提取方法属于此类。它具有一个通用的结构, 首先在时间上估计音高显著度, 然后应用跟踪规则提取旋律线。具体包括以下几个步骤(见图 2)。

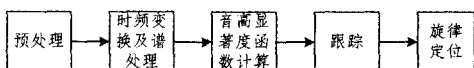


图 2 基于音高显著度的旋律提取方法的一般步骤

(1) 预处理。首先进行预处理来增强旋律可能存在的频率范围。常用的方法包括带通滤波器、感知相关的等响滤波器(equal loudness filter)^[19]、音调类声音与打击乐器声音的源分离技术^[20-21]。

(2) 时频变换及谱处理。将信号分割成时间帧进行时频变换及谱处理。常用的时频变换包括短时傅里叶变换(STFT)^[19, 22-26]、多分辨率变换(multirate filterbank)^[27]、常 Q 变换(constant-Q transform)^[28]以及多分辨率傅里叶变换(multi-resolution FFT, MRFFT)^[20, 21, 29]。与 STFT 相比, 后者时频变换避免了时频分辨率不均匀的缺点, 更接近于人类听觉系统, 在低频有更高的频率分辨率来分解相近的音符, 在高频有更高的时间分辨率来跟踪快速变化的泛音。但是实验表明同一旋律提取算法采用 MRFFT 与采用 STFT 相比并没有显著的提高^[19]。在时频变换后, 大多数方法检测谱峰值并对其进行进一步处理, 例如根据幅度或正弦性(sinusoidality)滤除不代表泛音的峰值^[20-24, 27], 或者应用谱幅度归一化(spectral magnitude normalization)减少音色的影响^[25, 28]。

(3) 音高显著度函数(pitch salience function)的计算。这是此类算法的核心。在通过上述步骤得到的频谱图上计算该函数, 将其峰值作为旋律潜在的 f_0 是发现旋律线音高的重要依据。显著度函数可以由不同的方法定义得到, 大多数方法使用某种形式的泛音求和, 例如将某个音高的显著度计算为它的泛音频率幅度的加权^[19-22, 28, 30], 从训练数据中学习得到权重。文献^[45]在音频信号的频谱中检测泛音显著度, 通过分析泛音显著度指标计算得到多音音频的主基频, 并将其转换为 MIDI 形式的音符序列表示旋律。该方法是目前国内最早尝试提取音乐旋律的工作, 但基本属于重复基于音高显著性的早期旋律提取方法, 几乎没有引入创新。Goto^[27]和 Marolt^[23]使用期望最大化(EM)对观察到的频谱拟合一个音调模型, 该模型的 f_0 对应于某个音高, 其最大后验概率(MAP)估计被定义为该音高的显著度。其他显著度定义方法还包括自相关^[31]、谱峰值的逐对分析^[32]等。

音高显著度函数方法有一个主要缺点, 即检测到的旋律音高 f_0 候选值有时是真实 f_0 的整数倍或分数倍, 成为虚假音高。算法经常会选择输出位于旋律正确音高上下一个八度的 f_0 , 这将导致常见的八度错误(octave errors)。一些算法试图直接减少在音高显著度函数中的虚假音高数量。Dressler 检查可能属于同一个泛音系列的一对谱峰值, 如果有很多高幅度的频率落在其间, 则认为是虚假音高并衰减它们的求和结果^[33]。Cancela 提出了另一种方法, 即对于某个候选音高 f_0 , 若在频率 $2k \cdot f_0$, $3k \cdot f_0/2$ 和 $3k \cdot f_0$ 的谱成分平均幅度高于在 $k \cdot f_0$ 的频率谱成分的平均值, 则认为 f_0 是虚假旋律音高并衰减其泛音求和。这种情况下的虚假旋律 f_0 是真实 f_0 音高值的 $1/2$, $2/3$ 或 $1/3$ ^[28]。Klapuri 计算显著度函数的每个频谱峰值及其泛音序列的谱包络线。真实旋律音符的谱包络线通常比较平滑, 而虚假旋律音高的峰值与其相比则更不规则, 此类显著度函数将被衰减, 使得基于谱平滑度的八度错误减少^[17]。Bosch 等^[48]将平滑瞬时模型与基于音高显著度的旋律提取算法相结合, 提高了原始音高的准确率, 减少了八度错误。以上这些减少八度错误的方法的效果

都不明显,在音高跟踪阶段调整大的音高跳跃通常会取得更为显著的效果。除了八度错误,若歌曲中存在背景和声,也会经常引起旋律音高估计错误,目前对此尚无较好的解决办法。

(4)跟踪。确定音高显著度函数的峰值后,剩下的任务就是判断哪些音高峰值属于旋律。这是每个算法最关键的步骤,也是灵活性最大的步骤,实际上每个算法都使用了不同的方法。大多数方法直接在音高显著度峰值上,或先根据时间、音高和显著度连续等约束得到候选峰值,再由候选峰值聚集而成的音高包络线(也叫片段或轨迹)上跟踪旋律。常用的跟踪技术包括聚类^[23]、基于启发式的跟踪代理^[27,33]、隐马尔可夫模型^[21-22]、动态规划^[20,24]、随机森林^[34]等。

(5)旋律定位。旋律提取的最后一个重要步骤是旋律定位,即确定某个时间点旋律是否存在。大多数方法对每个音帧使用固定的或动态的基于音高显著度函数的门限^[23-25,28,31-33]。对于歌曲,可以使用基于人声与乐器音色的分类来确定旋律是否存在^[20]。这部分经常会出现错误:漏判(false negative),即当旋律实际存在时错误地判断为不存在;虚警(false positive),即将实际上没有旋律的部分错误地判断为存在旋律。

2.2 基于声源分离的旋律提取方法

近年来出现了许多音频源分离技术,其中一些能直接将旋律的主要部分从混合物中分离出来^[35-39]。这些算法通常用分离音频与原始音频之间的信噪比来评价性能。只有下述几种算法延续了单音音符识别并进行了旋律的音符序列估计。

Durrieu等^[35]将混合音频信号的功率谱模拟为主唱和伴奏这两个分布的和。主唱的分布表示为一个源/过滤(source/filter)模型,伴奏的分布表示为具有不同谱形状的任意数量的源的和。源/过滤模型采用两种不同的表示:一种是平滑瞬时混合模型(Smooth Instantaneous Mixture Model, SIMM),另一种是平滑高斯尺度混合模型(Smooth Gaussian Scaled Mixture Model, SGSMM)。前者将主乐器或歌声表示为所有可能的音符的瞬时混合物;后者则在任意时刻只允许一个源/滤波对(source/filter couple)是活动的,虽然计算量更大,但是也更具有实际意义。这两种表示都使用期望最大化来进行模型参数估计,然后使用维特比(Viterbi)算法发现一个平滑的 f_0 轨迹,得到旋律音高序列。最后,在每个帧中通过与能量阈值进行比较来定位歌声旋律。

Tachibana等^[36]通过改变用于STFT分析的窗口长度,分离在时间上持续平滑的声源(音调类乐器分量)和在频率上持续平滑的声源(打击类乐器分量)。经过这两个阶段后,残余信号里的歌声旋律成分被大大加强。在残余信号的频谱图上,候选频率 f 属于歌声旋律的概率正比于它在整数倍泛音上的能量的加权和。通过发现频率序列的最大后验概率MAP的路径,可以直接得到旋律频率序列。最后通过歌声检测进行旋律定位。

除了以上几种直接利用音源分离提取旋律的方法,还有一些作者试图将声源分离和基于音高显著度的方法结合起来。Hsu^[20]和Yeh^[21]都使用Tachibana提出的声源分离方法^[36]进行预处理,减少伴奏分量并增强旋律分量,然后从残余信号的频谱图中计算音高显著度函数,按照2.1节的框架估计旋律。

除了以上两种算法以外,还有几种典型的声源分离算法也可以用于构建旋律提取系统。在构建过程中,这些算法或者与一个单音音高跟踪算法结合,直接从分离的声音信号中估计旋律 f_0 序列,或者作为一个预处理步骤,后接基于音高显著度的旋律检测框架或下述的基于分类的旋律检测框架。Huang等^[37]基于音乐伴奏经常具有重复结构而歌声包含更多变化的事实来进行歌声分离。作者将伴奏的频谱图模拟为一个低秩矩阵,将声音的频谱图模拟为一个稀疏矩阵;使用鲁棒主成分分析(Robust Principal Component Analysis, RPCA)把混合信号的频谱图分解为歌声矩阵和伴奏矩阵。Rafii和Pardo^[38]利用一种不同的重复性方法分离歌声与伴奏。首先计算混合物频谱图的自相关,得到伴奏的重复周期。通过计算连续重复的频谱图的中值得到一个新的频谱图,其中包含的重复信号即被视为伴奏。这个频谱图被用来构造一个时频掩蔽,即可从伴奏中分离出歌声。该方法被Liutkus等进一步扩展到完整歌曲上^[39]。因为重复部分可能在主歌和副歌之间发生变化,通过使用Rafii和Pardo^[38]的方法在局部窗口中搜索歌曲中的局部重复性,并计算自相似矩阵,从而更好地识别歌曲中的重复片段。在未来的研究中,可能会出现更多的基于声源分离的旋律提取算法。

2.3 其他的旋律提取方法

大多数旋律提取方法是基于音高显著度计算或者声源分离技术的,还有一些方法从不同的角度进行旋律提取。Poliner和Ellis^[40]提出了一种基于数据驱动的方法,该方法不需要利用音乐声学知识(例如,计算基于泛音求和的音高显著度函数),而使用训练好的统计分类器直接从频谱图中估计旋律音符。通过STFT计算频谱表示,去除2kHz以上的频率成分,其余频谱的幅度在短时帧内被归一化以减少不同乐器音色的影响。SVM分类器在基于帧的256维特征矢量上训练,每个帧使用60个横跨5个八度的MIDI音符进行标注。最后通过200~1800Hz之间的能量平方的全局阈值进行旋律定位。该算法并不完善,没有采用后处理措施来减少八度错误和虚警率。

Sutton等^[41]提出了另一种完全不同的方法。作者没有试图设计一个算法来处理多音音频信号,而是应用两个不同的单音音高估计器来计算并返回各自的音高序列,然后使用HMM将结果结合起来。该方法依赖于这样一个假设:虽然单音音高估计并不能很好地处理多音音频信号(通常会导致大量的估计错误),但是通过结合不同估计器的输出有可能会得到一个更可靠的结果。

文献[46]首先利用谐波能量互相关的假设对音乐进行多基频提取,然后从多基频数据中筛选出人声演唱的部分作为该段音乐的主旋律。文献[47]首先对音频信号进行歌声检测,然后对有歌声段利用迭代谱减算法进行多基频估计,最后利用每条基频轨迹的特征来确定主旋律。这两种方法的思路类似,均需要进行多基频提取,而这个任务比旋律提取更加困难,因此逻辑上存在疑问。

3 旋律提取评测标准

旋律提取通常采用以下5个全局度量进行性能评测^[2,6]。

(1)旋律定位查全率(voicing recall rate):算法估计正确的旋律帧数量/基准数据中全部被标记为旋律的帧数量。

(2)旋律定位虚警率(voicing false alarm rate):算法错误地估计为旋律的帧数量/基准数据中被标记为非旋律的帧数量。

(3)原始音高准确率(raw pitch accuracy):在基准数据的旋律帧中,算法正确估计音高(检测音高与基准音高的差别在半个半音之内)的帧的比例。

(4)原始 Chroma 准确率(raw Chroma accuracy):在原始音高准确率的基础上,将估计的音高和基准音高都映射到 1 个八度内,即忽略八度错误。

(5)整体准确率(overall accuracy):算法正确估计的所有帧的比例,即算法将非旋律帧标记为非旋律,将旋律帧标记为旋律并且对于旋律帧检测音高与基准音高的差别在半个半音之内。该度量给出系统的整体性能评分。

在完整数据集上的得分由所有片段上的分数平均得到。原始音高的准确率和整体准确率是最重要的两个评价指标。在 2009 年之后原始音高的准确率基本不变,保持在 70%~80%之间;因为在原始音高的准确率的基础上还要考虑旋律定位的因素,所以整体准确率有所下降,保持在 65%~75%之间。在计算代价方面,基于声源分离技术的算法由于内含大量矩阵运算,通常比基于音高显著度的方法更加复杂。

结束语 本文对流行音乐主旋律提取技术进行了综述。首先描述了主旋律提取在数字音乐计算领域的重要意义和研究价值。然后将现有的研究工作划分为基于音高显著度、基于声源分离以及其他方法 3 类,阐述了主旋律提取的方法。最后描述了旋律提取评测的 5 个标准。

近年来,研究者们提出了众多主旋律提取算法,基于音高显著度与基于声源分离的旋律提取算法依然是主流。但是,在各个数据集上的测试结果表明,最新算法的性能并没有得到显著的提高。对主旋律提取的研究仍具有巨大的提升空间和探索的可能性。

参 考 文 献

- [1] KREIMAN J, SITTIS D. Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception [M]. Wiley-Blackwell, 2011: 1-24.
- [2] POLINER G E, ELLIS D P W, EHMAN A F, et al. Melody Transcription From Music Audio: Approaches and Evaluation [J]. IEEE Transactions on Audio Speech & Language Processing, 2007, 15(4): 1247-1256.
- [3] MOORE B C. An Introduction to the Psychology of Hearing [J]. General Information, 1997, 27(1): 3-10.
- [4] KLAPURI A P. Automatic Music Transcription as We Know it Today [J]. Journal of New Music Research, 2004, 33(3): 269-282.
- [5] KLAPURI A, DAVY M, et al. Signal processing methods for music transcription [M]. Springer Science & Business Media, 2007.
- [6] SALAMON J, GOMEZ E, ELLIS D P W, et al. Melody extraction from polyphonic music signals: Approaches, applications, and challenges [J]. Signal Processing Magazine, IEEE, 2014, 31(2): 118-134.
- [7] GÓMEZ E, CAÑADAS-QUESADA F J, SALAMON J, et al. Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing [C]// ISMIR. 2012: 601-606.
- [8] KODURI G K, SERRAÈRRÀ JULIÀ J, SERRA X. Characterization of intonation in carnatic music by parametrizing pitch histograms [C]// Proceedings of the 13th International Society for Music Information Retrieval Conference. Porto, Portugal. 2012: 95-98.
- [9] PIKRAKIS A, GÓMEZ F, OORAMAS S, et al. Tracking Melodic Patterns in Flamenco Singing by Analyzing Polyphonic Music Recordings [C]// ISMIR. 2012: 421-426.
- [10] SALAMON J, SERRA J, GÓMEZ E. Tonal representations for music retrieval: from version identification to query-by-humming [J]. International Journal of Multimedia Information Retrieval, 2013, 2(1): 45-58.
- [11] FOUCARD R, DURRIEU J L, LAGRANGE M, et al. Multimodal similarity between musical streams for cover version detection [C]// 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE, 2010: 5514-5517.
- [12] SALAMON J, ROCHA B, GÓMEZ E. Musical genre classification using melody features extracted from polyphonic music signals [C]// 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012: 81-84.
- [13] DURRIEU J L, RICHARD G, DAVID B. An iterative approach to monaural musical mixture de-soloing [C]// IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009). IEEE, 2009: 105-108.
- [14] MESAROS A, VIRTANEN T, KLAPURI A. Singer Identification in Polyphonic Music Using Vocal Separation and Pattern Recognition Methods [C]// ISMIR. 2007: 375-378.
- [15] Hess W. Pitch determination of speech signals: algorithms and devices [M]. Springer Science & Business Media, 2012.
- [16] GÓMEZ E, KLAPURI A, MEUDIC B. Melody description and extraction in the context of music content processing [J]. Journal of New Music Research, 2003, 32(1): 23-40.
- [17] KLAPURI A. Signal processing methods for the automatic transcription of music [M]. Finland: Tampere University of Technology, 2004.
- [18] SUNDBERG J, ROSSING T D. The Science of the Singing Voice [J]. Journal of the Acoustical Society of America, 1990, 87(1): 462-463.
- [19] SALAMON J, GÓMEZ E. Melody extraction from polyphonic music signals using pitch contour characteristics [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(6): 1759-1770.
- [20] HSU C L, JANG J S R. Singing Pitch Extraction by Voice Vibrato/Tremolo Estimation and Instrument Partial Deletion [C]// ISMIR. 2010: 525-530.
- [21] YEH T C, WU M J, JANG J S R, et al. A hybrid approach to singing pitch extraction based on trend estimation and hidden

- Markov models[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012:457-460.
- [22] RYNNÄNEN M P, KLAURI A P. Automatic transcription of melody, bass line, and chords in polyphonic music[J]. *Computer Music Journal*, 2008, 32(3): 72-86.
- [23] MAROLT M. Gaussian Mixture Models For Extraction Of Melodic Lines From Audio Recordings[C]//ISMIR. 2004.
- [24] RAO V, RAO P. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(8): 2145-2154.
- [25] ARORA V, BEHERA L. On-line melody extraction from polyphonic audio using harmonic cluster tracking[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(3): 520-530.
- [26] SRIDEVI S H, GULHANE S R. Melody Extraction from Polyphonic Music Signal using STFT and Fanchirp Transform[C]//International Journal of Engineering Research and Technology. 2015.
- [27] GOTO M. A real-time music-scene-description system; Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals[J]. *Speech Communication*, 2004, 43(4): 311-329.
- [28] CANCELA P. Tracking melody in polyphonic audio[C]//Proc. of Music Information Retrieval Evaluation eXchange, 2008.
- [29] DRESSLER K. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT[C]//Proc. of 9th Int. Conf. on Digital Audio Effects (DAFx-06). 2006:247-252.
- [30] JO S, JOO S, YOO C D. Melody pitch estimation based on range estimation and candidate extraction using harmonic structure model[C]//Interspeech. 2010:2902-2905.
- [31] PAIVA R P, MENDES T, CARDOSO A. Melody detection in polyphonic musical signals; Exploiting perceptual rules, note salience, and melodic smoothness[J]. *Computer Music Journal*, 2006, 30(4): 80-98.
- [32] DRESSLER K. Pitch estimation by the pair-wise evaluation of spectral peaks[C]//Audio Engineering Society Conference; 42nd International Conference; Semantic Audio. Audio Engineering Society, 2011.
- [33] DRESSLER K. An Auditory Streaming Approach for Melody Extraction from Polyphonic Music[C]//ISMIR. 2011:19-24.
- [34] BITTNER R M, SALAMON J, ESSID S, et al. Melody Extraction By Contour Classification[C]//ISMIR. 2015.
- [35] DURRIEU J L, RICHARD G, DAVID B, et al. Source/filter model for unsupervised main melody extraction from polyphonic audio signals[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(3): 564-575.
- [36] TACHIBANA H, ONO T, ONON, et al. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source[C]//2010 IEEE International Conference on Acoustics speech and signal processing (ICASSP). IEEE, 2010: 425-428.
- [37] HUANG P S, CHEN S D, SMARAGDIS P, et al. Singing-voice separation from monaural recordings using robust principal component analysis[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012:57-60.
- [38] RAFII Z, PARDON B. Repeating pattern extraction technique (REPET): A simple method for music/voice separation[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(1): 73-84.
- [39] LIUTKUS A, RAFII Z, BADEAU R, et al. Adaptive filtering for music/voice separation exploiting the repeating musical structure[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012: 53-56.
- [40] POLINER G E, ELLIS D P W. A classification approach to melody transcription[C]//ISMIR 2005. 2005:161-166.
- [41] SUTTON C, VINCENT E, PLUMBLEY M, et al. Transcription of vocal melodies using voice characteristics and algorithm fusion[C]//Proc. of Music Information Retrieval Evaluation eXchange (MIREX). 2006.
- [42] 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: 1429-1432.
- [43] SERRA J, GÓMEZ E, HERRERA P, et al. Chroma binary similarity and local alignment applied to cover song identification[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, 16(6): 1138-1151.
- [44] ZHU B, LI W, WANG Z, et al. A novel audio fingerprinting method robust to time scale modification and pitch shifting[C]//Proceedings of the International Conference on Multimedia. ACM, 2010:987-990.
- [45] ZHANG J J. Harmonic overtone detection based main melody extraction[D]. Shanghai: Shanghai Jiao Tong University, 2007. (in Chinese)
张俊杰. 基于和谐泛音检测的主旋律提取技术[D]. 上海: 上海交通大学, 2007.
- [46] SONG Y Y. An approach for music melody extraction based on underdetermined single-source speech separation[D]. Beijing: Beijing University of Posts and Telecommunications, 2012. (in Chinese)
宋岳阳. 基于单源欠定语音分离的音乐主旋律提取方法研究[D]. 北京: 北京邮电大学, 2012.
- [47] ZHANG M M. Melody extraction from singing voice of polyphonic music[D]. Jinan: Shandong University, 2015. (in Chinese)
张萌萌. 复合音乐中歌声旋律的提取[D]. 济南: 山东大学, 2015.
- [48] BOSCH J J, BITTNER R M, SALAMON J, et al. A Comparison of Melody Extraction Methods Based on Source-Filter Modelling[C]//Proceedings of International Society for Music Information Retrieval Conference (ISMIR). 2016:571-577.