文章编号：0427-7104(2018)03-0271-43

理解数字音乐——音乐信息检索技术综述

李 伟[1,2]，李子晋[3]，高永伟[1]

（1.复旦大学 计算机科学技术学院，上海 201203；2.复旦大学 上海市智能信息处理重点实验室，上海 200433；
3.中国音乐学院 音乐科技系，北京 100101）

摘 要：近 20 年来，音频压缩技术的成熟及互联网的普及使得音乐迅速从磁带和激光唱盘（CD）转变为互联网上以 MP3 为代表的数字音乐.海量数字音乐带来分类组织、查询检索、内容理解与分析等一系列问题，促使产生了一个新兴的交叉学科，即基于内容的音乐信息检索（Content-based Music Information Retrieval，MIR）.本文阐述了 MIR 与音乐科技、声音与音乐计算、计算机听觉、语音信息处理、音乐声学等各个相关领域概念的区别与联系，将 MIR 技术的数十个研究领域按照与音乐要素的密切程度划分为核心层与应用层.分类总结了各领域的概念、原理、应用、基本技术框架及典型文献，同时介绍了研究中常用的音乐领域知识并明确了中英文术语.最后总结 MIR 领域存在的各方面问题，并展望其未来发展趋势.

关键词：音乐科技；声音与音乐计算；计算机听觉；音乐信息检索；音乐声学

中图分类号：TP391 文献标志码：A

DOI:10.15943/j.cnki.fdxb-jns.2018.03.001

1 音乐科技概述

音乐与科技的融合具有悠久的历史.早在 20 世纪 50 年代，一些不同国家的作曲家、工程师和科学家已经开始探索利用新的数字技术来处理音乐，并逐渐形成了音乐科技/计算机音乐（Music Technology/Computer Music）这一交叉学科.20 世纪 70 年代之后，欧美各国相继建立了多个大型计算机音乐研究机构，如 1975 年建立的美国斯坦福大学 CCRMA（Center for Computer Research in Music and Acoustics）、1977 年建立的法国巴黎 IRCAM（Institute for Research and Coordination Acoustic/Music）、1994 年成立的西班牙巴塞罗那 UPF（Universitat Pompeu Fabra）的 MTG（Music Technology Group）以及 2001 年成立的英国伦敦女王大学 C4DM（Center for Digital Music）等.在欧美之后，音乐科技在世界各地都逐渐发展起来，欧洲由于其浓厚的人文和艺术气息成为该领域的世界中心.该学科在中国大陆发展较晚，大约 20 世纪 90 年代中期开始有零散的研究，由于各方面的限制，至今仍处于起步阶段[1].

音乐科技分为两个子领域：一是基于科技的音乐创作；二是数字音频与音乐技术的科学技术研究.本文内容限于后一领域.音乐科技具有众多应用，例如数字乐器、音乐制作与编辑、音乐信息检索、数字音乐图书馆、交互式多媒体、音频接口、辅助医学治疗等.这些应用背后的科学研究通常称为声音与音乐计算（Sound and Music Computing，SMC），在 20 世纪 90 年代中期被定义为国际计算机学会（Association for Computing Machinery，ACM）的标准术语[2].SMC 是一个多学科交叉的研究领域.在科技方面涉及到声学（Acoustics）、音频信号处理（Audio Signal Processing）、机器学习（Machine Learning）、人机交互（Human-Machine Interaction）等学科；在音乐方面涉及作曲（Composition）、音乐制作（Music Creation）、声音设计（Sound Design）等学科.国际上已有多个侧重点不同的国际会议和期刊，如 1972 年创刊的 JNMR（Journal of New Music Research）、1974 年建立的 ICMC（International Conference on Computer Music）、1977 年创刊的 CMJ（Computer Music Journal）、2000 年建立的 ISMIR（International Society for Music Information Retrieval Conference）等.

收稿日期：2017-12-18
基金项目：国家自然科学基金（61671156）
作者简介：李 伟（1970—），男，教授，博士生导师，E-mail：weili-fudan@fudan.edu.cn.

SMC 是一个庞大的研究领域,可细化为以下 4 个学科分支.(1) 声音与音乐信号处理:用于声音和音乐的信号分析、变换及合成,例如频谱分析(Spectral Analysis)、调幅(Magnitude Modulation)、调频(Frequency Modulation)、低通/高通/带通/带阻滤波(Low-pass/High-pass/Band-pass/Band-stop Filtering)、转码(Transcoding)、无损/有损压缩(Lossless/Lossy Compression)、重采样(Resampling)、回声(Echo)、混音(Remixing)、去噪(Denoising)、变调 PS(Pitch Shifting)、保持音高不变的时间伸缩(Time-Scale Modification/Time Stretching, TSM)、线性时间缩放(Time Scaling)等.该分支相对比较成熟,已有多款商业软件如 Gold Wave、Adobe Audition/Cool Edit、Cubase、Sonar/Cakewalk、EarMaster 等.(2) 声音与音乐的理解分析:使用计算方法对数字化声音与音乐的内容进行理解和分析,例如音乐识谱、旋律提取、节奏分析、和弦识别、音频检索、流派分类、情感分析、歌手识别、歌唱评价、歌声分离等.该分支在 20 世纪 90 年代末随着互联网上数字音频和音乐的急剧增加而发展起来,研究难度大,多项研究内容至今仍在持续进行中.与计算机视觉(Computer Vision, CV)对应,该分支也可称为计算机听觉(Computer Audition, CA)或机器听觉(Machine Listening, ML)[3].注意计算机听觉是用来理解分析而不是处理音频和音乐[4],且不包括语音.语音信息处理的历史要更早数十年,发展相对成熟,已独立成为一门学科,包含语音识别、说话人识别、语种识别、语音分离、计算语言学等多个研究领域.CA 若剔除一般声音而局限于音乐,则可称为音乐信息检索(Music Information Retrieval, MIR),这也是本文主要的介绍内容.(3) 音乐与计算机的接口设计:包括音响及多声道声音系统的开发与设计、声音装置等.该分支偏向音频工程应用.(4) 计算机辅助音乐创作:包括算法作曲、计算机音乐制作、音效及声音设计等.该分支偏向艺术创作.

与音乐有关但是与 SMC 不同的另一个历史更悠久的学科是音乐声学(Music Acoustics).音乐声学是研究在音乐这种声音振动中存在的物理问题的科学,是音乐学与物理学的交叉学科.音乐声学主要研究乐音与噪声的区别、音高音强和音色的物理本质、基于电磁振荡的电声学、听觉器官的声波感受机制、乐器声学、人类发声机制、音律学、与音乐有关的室内声学等.从学科的角度看,一部分音乐声学知识也是 SMC 的基础,但 SMC 研究更依赖于音频信号处理和机器学习这两门学科.同时,研究内容面向音频与音乐的信号处理、内容分析和理解,与更偏重于解决振动相关物理问题的音乐声学也有较大区别.为更清楚地理解各学科之间的区别与联系,我们将音乐科技及听觉研究各领域关系分别示于图 1 和图 2.

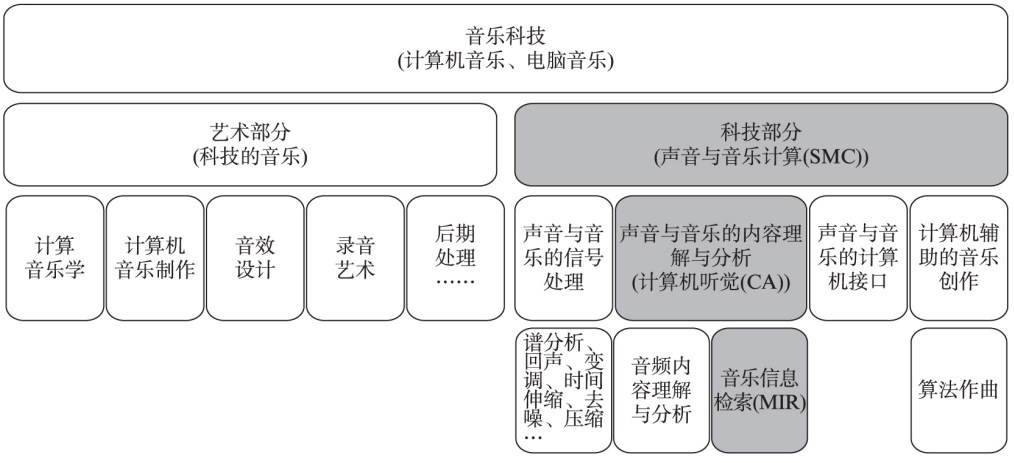

图 1 音乐科技各领域关系图

Fig.1 A relation graph of different music technology fields

2 基于内容的音乐信息检索

从 20 世纪 90 年代中期开始,互联网在世界范围内迅速普及.同时,以 MP3(MPEG-1 Layer 3)为代表的音频压缩技术开始大规模应用.此外,半导体技术和工艺的迅猛发展使得硬盘等存储设备的容量越来越大.这几大因素使得传统的黑胶唱片、磁带、CD 光盘等音乐介质几乎消失,取而代之的是在电脑硬盘上存储,在互联网上传输、下载和聆听的数字音乐.海量的数字音乐直接促使了音乐信息检索(MIR)技术的产

生，其内涵早已从最初的狭义音乐检索扩展到使用计算手段对数字音乐进行内容分析理解的大型科研领域，包含数十项研究课题. 2000 年国际音乐信息检索学术会议(ISMIR)的建立可以视为这一领域的正式创建.

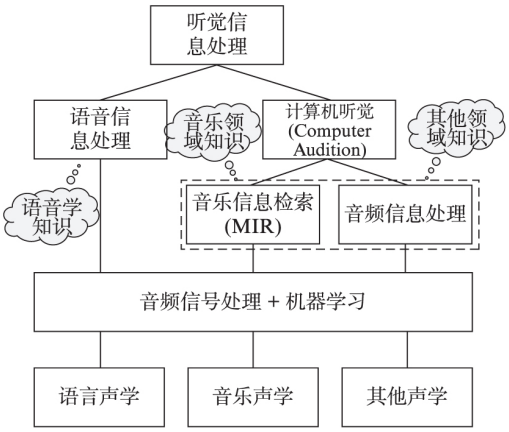

图 2　听觉研究各领域关系图

Fig.2　A relation graph of different auditory research fields

基于内容的音乐信息检索(及相关音乐科技)有很多应用.在娱乐相关领域，典型应用包括听歌识曲、哼唱/歌唱检索、翻唱检索、曲风分类、音乐情感计算、音乐推荐、彩铃制作、卡拉 OK 应用、伴奏生成、自动配乐、音乐内容标注、歌手识别、模仿秀评价、歌唱评价、歌声合成及转换、智能作曲、数字乐器、音频/音乐编辑制作等.在音乐教育及科研领域，典型应用包括计算音乐学、视唱练耳及乐理辅助教学、声乐及各种乐器辅助教学、数字音频/音乐图书馆等.在日常生活、心理及医疗、知识产权等其他领域，还包括乐器音质评价及辅助购买、音乐理疗及辅助医疗、音乐版权保护及盗版追踪等应用.此外，在电影及很多视频中，音频及音乐都可以用来辅助视觉内容进行分析.以上应用均可以在电脑、智能手机、音乐机器人等各种平台上进行实现.

早期的 MIR 技术以符号音乐(Symbolic Music)如 MIDI(Musical Instrument Digital Interface)为研究对象.由于其具有准确的音高、时间等信息，很快就发展得比较成熟.后续研究很快转为以音频信号为研究对象，研究难度急剧上升.随着该领域研究的不断深入，如今 MIR 技术已经不仅仅指早期狭义的音乐搜索，而从更广泛的角度上包含了音乐信息处理的所有子领域.我们根据自己的理解，将 MIR 领域的几十个研究课题归纳为核心层和应用层共 9 个部分(图 3).核心层包含与各大音乐要素(如音高与旋律、音乐节奏、音乐和声等)及歌声信息处理相关的子领域，应用层则包含在核心层基础上更偏向应用的子领域(如音乐搜索、音乐情感计算、音乐推荐等).下面依次对其概念、原理、基本技术框架以及典型算法进行介绍.

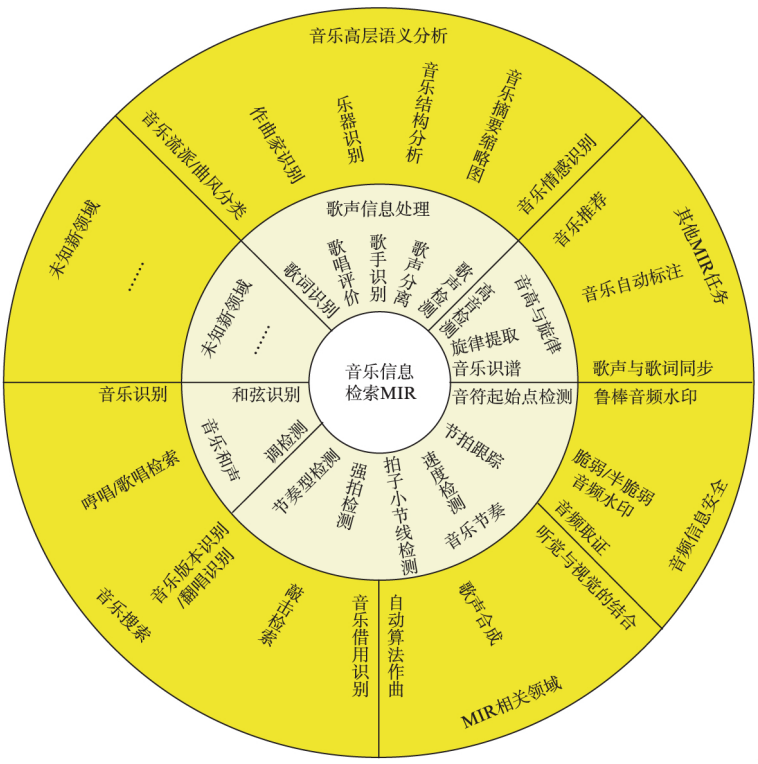

图 3　音乐信息检索(MIR)的研究领域

Fig.3　Illustration of MIR research topics

MIR 研究领域的科研涉及一些基本的乐理知识,在图 4 中汇总显示,在下述文字中可参照理解.

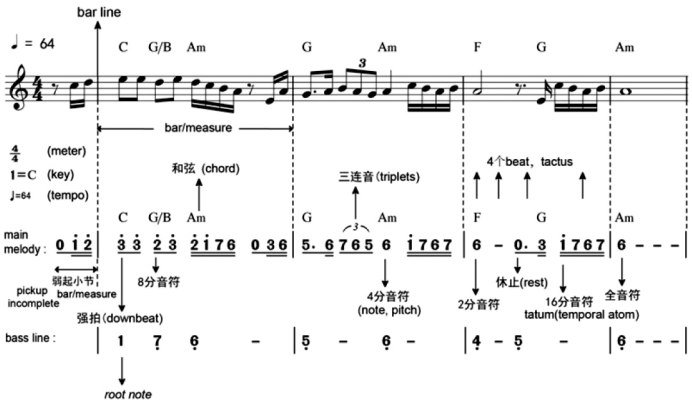

图 4 MIR 的研究领域常用的基本乐理常识

Fig.4 Musical knowledge commonly used in MIR

2.1 音高与旋律

音乐中每个音符都具有一定的音高属性.若干个音符经过艺术构思按照节奏及和声结构(Harmonic Structure)形成多个序列,其中反映音乐主旨的序列称为主旋律,是最重要的音乐要素,其余序列分别为位于高、中、低音声部的伴奏.该子领域主要包括音高检测、旋律提取和音乐识谱等任务.

2.1.1 音高检测

音高(Pitch)由周期性声音波形的最低频率即基频决定,是声音的重要特性.音高检测(Pitch Detection)也称为基频估计(Fundamental Frequency/f_0 Estimation),是语音及音频、音乐信息处理中的关键技术之一.音高检测技术最早面向语音信号,在时域包括经典的自相关算法[5]、YIN 算法[6]、最大似然算法[7]、SIFT(Simplified Inverse Filter Tracking)滤波器算法[8]以及超分辨率算法[9]等;在频域包括基于正弦波模型[10]、倒谱变换[11]、小波变换[12]等的各种方法.一个好的算法应该对声音偏低偏高者都适用,而且对噪音鲁棒.

在 MIR 技术中,音高检测被扩展到多声部/多音音乐(Polyphonic Music)中的歌声信号.由于各种乐器伴奏的存在,检测歌声的音高更加具有挑战性.直观上首先进行歌声与伴奏分离有助于更准确的检测歌声音高[13],估计每个音频帧(Frame)上的歌声音高范围也可以减少乐器或歌声泛音(Partial)引起的错误尤其是八度错误(Octave Errors)[14-15],融合几个音高跟踪器的结果也有希望得到更高的准确率[16].此外,相邻音符并非孤立存在,而是按照旋律与和声有机地连接,可用隐马尔科夫模型(Hidden Markov Model, HMM)等时序建模工具进行纠错[17].除了歌声,Goto 等使用期望最大化(Expectation Maximization, EM)方法并结合时域连续性来估计旋律和低音线的基频[18].

2.1.2 旋律提取

旋律提取(Melody Extraction)从多声部/多音音乐信号中提取单声部/单音(Monophonic)主旋律,是 MIR 领域的核心问题之一.在音乐检索、抄袭检测、歌唱评价、作曲家风格分析等多个子领域中具有重要应用.从音乐信号中提取主旋律的方法主要分为 3 类:即音高重要性法(Pitch-salience based Melody Extraction)[19-20]、歌声分离法(Singing Separation based Melody Extraction)[21-22]及数据驱动的音符分类法(Data-driven Note Classification)[23].第一类方法依赖于每个音频帧上的旋律音高提取,这本身就是一个极困难的问题.此外还涉及旋律包络线的选择和聚集等后处理问题.第三类方法单纯依赖于统计分类器,难以处理各种各样的复杂多声部/多音音乐信号.相比之下,我们认为第二类方法具有更好的前景.这里并不需要完全彻底的音源分离,而只需要像文献[21]那样根据波动性和短时性特点进行旋律成分增强或像文献[22]那样通过概率隐藏成分分析(Probabilistic Latent Component Analysis, PLCA)学习非歌声部分的统计模型进行伴奏成分消减,之后即可采用自相关等音高检测方法提取主旋律线(Predominant Melody Lines).以上各种方法还面临一些共同的困难,如八度错误,如何提取纯器乐的主旋律等[24].

2.1.3　音乐识谱

音乐可分为单声部/单音和多声部/多音.单声部/单音音乐在某一时刻只有一个乐器或歌唱的声音，使用 2.1.1 节中的音高检测技术即可进行比较准确的单声部/单音音乐识谱(Monophonic Music Transcription).目前急需解决的是多声部/多音音乐识谱(Polyphonic Music Transcription)，即从一段音乐信号中识别每个时刻同时发声的各个音符，形成乐谱并记录下来，俗称扒带子.由于音乐信号包含多种按和声结构存在的乐器和歌声，频谱重叠现象普遍，音乐识谱(Music Transcription)极具挑战性，是 MIR 领域的核心问题之一.同时，音乐识谱具有很多应用，如音乐信息检索、音乐教育、乐器及多说话人音源分离[25]、颤音和滑音(Glissando)标注等[26].

多声部/多音音乐识谱系统首先将音乐信号分割为时间单元序列，然后对每个时间单元进行多音高/多基频估计(Multiple Pitch/Fundamental Frequency Estimation).再根据 MIDI 音符表将各基频转换为对应音符的音名，最后利用音乐领域知识或规则对音符、时值等结果进行后处理校正，结合速度和调高估计输出正确的乐谱.

多音高/多基频估计是进行音乐识谱的核心功能，经常使用对音乐信号的短时幅度谱[27]或常数 Q 变换(Constant-Q Transform)[28]进行矩阵分解的方法，如独立成分分析(Independent Component Analysis，ICA)[29]、非负矩阵分解(Non-negative Matrix Factorization，NMF)[30]、概率隐藏成分分析(PLCA)[31]等.与此思路不同，文献[32]基于迭代方法，首先估计最重要音源的基频，从混合物中将其减去，然后再重复处理残余信号.文献[33]使用重要性函数(Salience Function)来选择音高候选者，并使用一个结合候选音高的频谱和时间特性的打分函数来选择每个时间帧的最佳音高组合.由于多声部/多音音乐信号中当前音频帧的谱内容在很大程度上依赖于以前的帧，最后还需使用谱平滑性(Spectral Smoothness)[34]、HMM、条件随机场(Conditional Random Fields，CRFs)等进行纠错.

音乐识谱的研究虽然早在 30 年前就已开始[35]，但目前仍是 MIR 的研究领域中一个难以解决的问题，只能在简单情况下获得一定的结果.随着并发音符数量的增加，检测难度急剧上升，而且性能严重低于人类专家.主要原因在于当前识谱方法使用通用的模型，无法适应各个场景下的复杂音乐信号.一个可能的改进方法是使用乐谱、乐器类型等辅助信息进行半自动识谱[36]，或者进行多个算法的决策融合[37].

2.2　音乐节奏

音乐节奏是一个广义词，包含与时间有关的所有因素.把音符有规律地组织到一起，按照一定的长短和强弱有序进行，从而产生律动的感觉.MIR 领域与节奏相关的子领域包括：音符起始点检测、速度检测、节拍跟踪、拍子/小节线检测及强拍估计、节奏型检测.由于与英文对应的中文翻译混乱，本文采用文献[38]中的术语.

2.2.1　音符起始点检测

音符起始点(Note Onset)是音乐中某一音符开始的时间[39]，如图 5 所示.对于钢琴、吉他、贝斯等具有脉冲信号特征的乐器的音符，其起始(Attack)阶段能量突然上升，称为硬音符起始点(Hard Note Onset).而对于小提琴、大提琴、萨克斯、小号等弦乐或吹奏类乐器演奏的音符，则通常没有明显的能量上升，称为软音符起始点(Soft Note Onset).音符起始点检测(Note Onset Detection)通常是进行各种音乐节奏分析的预处理步骤[40]，在音乐混音(Music Remixing)、音频修复(Audio Restoration)、歌词识别、TSM、音频编码及合成(Audio Coding and Synthesis)中也都有应用.

在单声部/单音音乐信号中检测音符起始点并不难，尤其是对弹拨或击奏类乐器，简单地定位信号幅度包络线的峰值即可得到很高的准确率.但是在多声部/多音音乐信号中，检测整体信号失去效果，通常需要进行基于短时傅里叶变换(Short-time Fourier Transform，STFT)、小波变换(Wavelet Transform，WT)、听觉滤波器组的子带(Subband)分解.如文献[41]在高频

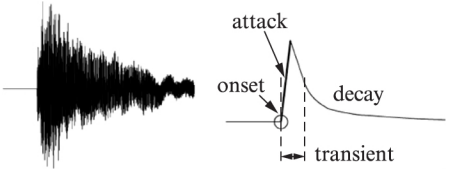

图 5　理想情况下一个音符的时间域信息描述：起始点(Onset)、上升(Attack)、过渡(Transient)和下降(Decay)

Fig.5　The description of Onset, Attack, Transient and Decay in an ideal note

子带基于能量的峰值挑选(Peak-picking)来检测强的瞬态事件,在低频子带使用一个基于频率的距离度量来提高软音符起始点的检测准确性.文献[42]没有检测能量峰值,而是通过观察相位在各个音频帧的分布也可以进行准确检测.除了常规子带分解,还有其他的分解形式.如文献[43]对音乐信号的频谱进行 NMF 分解,在得到的线性时域基(Linear Temporal Bases)上构造音符起始点检测函数.文献[44]将音乐信号基于 MT(Matching Pursuit)方法进行稀疏分解(Sparse Decomposition),通过稀疏系数的模式判断信号是稳定还是非稳定,之后自适应地通过峰值挑选得到 Onset 矢量.

除了基于信号处理的方法,后来又发展了多种基于机器学习的检测方法.机器学习主要用于分类,但具体应用方式并不相同.例如,文献[45]使用人工神经网络(Artificial Neural Network, ANN)对候选峰值进行分类,确定哪些峰值对应于音符起始点,哪些由噪声或打击乐器引起.希望避免峰值挑选方法中的门限问题.文献[46]则使用神经网络将信号每帧的频谱图(Spectrogram)分类为 Onsets 和 Non-onsets,对前者使用简单的峰值挑选算法.

2.2.2 速度检测

速度检测/感应(Tempo Detection/Induction)获取音乐进行的快慢信息,是 MIR 节奏类的基本任务之一.通常用每分钟多少拍(Beats Per Minute, bpm)来表示.速度检测是音乐情感分析(如欢快、悲伤等)中的一个重要因素.另一个有趣的应用是给帕金森病人播放与其走路速度一致的音乐,从而辅助其恢复行动机能[47].

进行音乐速度检测通常首先进行信号分解.核心思想是在节奏复杂的音乐中,某些成分会比整体混合物具有更规律的节奏,从而使速度检测更容易.如文献[48]将混合信号分解为和声部分和噪声部分两个子空间(Subspace),文献[49]将混合信号分解到多个子带.打击乐器控制速度进行,文献[50]使用非负矩阵分解(NMF)将混合信号分解为不同成分(Component),希望把不同的鼓声甚至频谱可能重叠的底鼓和贝斯声分解到不同的成分.与此思想类似,文献[51]使用概率隐藏成分分析(PLCA)将混合音乐信号分解到不同的成分.

针对各个子空间或子带的不同信号特性,采用不同的软、硬音符起始点函数,使用自相关、动态规划等方法分别计算周期性,再对候选速度值进行选择[52].速度检测方法基本上都基于音频信号处理方法,文献[53]提出了一种基于机器学习的方法,该方法使用听觉谱特征和谱距离,在一个已训练好的双向长短时记忆单元-递归神经网络(Bidirectional Long Short Term Memory-Recurrent Neural Network, BLSTM-RNN)上预测节拍,通过自相关进行速度计算,训练集包含不同音乐流派而且足够大.以上方法对于节奏稳定、打击乐或弹拨击奏类乐器较强的西方音乐,速度检测准确性已经很高,而对于打击乐器不存在或偏弱的音乐准确性则较差.

在处理弦乐等抒情音乐(Expressive Music),或速度发生渐快(Accelerando)、渐慢(Rallentando)时,速度检测仍然具有很大的研究难度,需为每个短时窗口估计主局部周期(Predominant Local Periodicity, PLP)进行局部化处理[54].文献[55]使用概率模型来处理抒情音乐中的时间偏差,用连续的隐藏变量对应于速度(Tempo),形式化为最大后验概率(Maximum A Posteriori, MAP)状态估计问题,用蒙特卡洛方法(Monte Carlo)求解.文献[56]基于谱能量通量(Spectral Energy Flux)建立一个 Onset 函数,采用自相关函数估计每个时间帧的主局部周期,然后使用维特比(Viterbi)算法来检测最可能的速度值序列.

以上方法都是分析原始格式音频,还有少量算法可以对 AAC(Advanced Audio Coding)等压缩格式音乐在完全解压、半解压、完全压缩等不同条件下进行速度估计[57].无论原始域还是压缩域速度检测算法,目前对于抒情音乐、速度变化、非西方音乐、速度的八度错误(减半或加倍/Halve or Double)等问题仍然没有很好的解决办法.

2.2.3 节拍跟踪

节拍(Beat)是指某种具有固定时长(Duration)的音符,通常以四分音符或八分音符为一拍.节拍跟踪/感应(Beat Tracking/Induction)是计算机对人们在听音乐时会无意识地踮脚或拍手的现象的模拟,经常用于对音乐信号按节拍进行分割[58],是理解音乐节奏的基础和很多 MIR 应用及多媒体系统如视频编辑、音乐可视化、舞台灯光控制等的重要步骤.早期的算法只能处理 MIDI 形式的符号音乐或者少数几种

乐器的声音信号，而且不能实时工作.20 世纪 90 年代中期以后，开始出现能处理包含各种乐器和歌声的流行音乐声音信号的算法，基本思想是通过检测控制节奏的鼓声来进行节拍跟踪[59].节拍跟踪可在线或离线进行，前者只能使用过去的音频数据，后者则可以使用完整的音频，难度有所降低[60].

节拍跟踪通常与速度检测同时进行[61-63]，首先在速度图(Tempogram)中挑选稳定的局部区域[59].下一步就是检测候选的节拍点，方法各不相同.文献[61]将节拍经过带通滤波等预处理后，对每个子带计算其幅度包络线和导数，与一组事先定义好的梳状滤波器(Comb Filter)进行卷积，对所有子带上的能量求和后得到一系列峰值.更多的方法依赖于音符起始点、打击乐器及其他时间域局域化事件的检测[62].如果音乐偏重抒情，没有打击乐器或不明显，可采用和弦改变点(无需识别和弦名字)作为候选点[63-64].

以候选节拍点为基础，即可进行节拍识别.文献[61]用最高的峰值对应于速度并进一步提取节拍.文献[65]基于感知设立门限并得到节拍输出.文献[66]使用简单有效的动态规划(Dynamic Programming，DP)方法来找到最好的节拍时间.文献[67]采用机器学习中的条件随机场(CRF)这种复杂的时域模型，将节拍位置估计模拟为时序标注问题.在一个短时窗口中通过 CRF 指定的候选者来捕捉局部速度变化并定位节拍.

对大多数流行音乐来讲，速度及节拍基本维持稳定，很多算法都可以得到不错的结果，但具体的定量性能比较依赖于具体评测方法的选择[68].对于少数复杂的流行音乐(如速度渐慢或渐快、每小节拍子发生变化等)和绝大多数古典音乐、交响乐、歌剧、东方民乐等，节拍跟踪仍然是一个研究难题.

2.2.4　拍子检测、小节线检测及强拍估计

音乐中有很多强弱不同的音符，在由小节线划分的相同时间间隔内，按照一定的次序重复出现，形成有规律的强弱变化即拍子(Meter/Time Signature).换句话说，拍子是音乐中表示固定单位时值和强弱规律的组织形式.在乐谱开头用节拍号(如 4/4、3/4 等)标记.拍子是组成小节(Bar/Measure)的基本单位，小节则是划分乐句、乐段、整首乐曲的基本单位.在乐谱中用小节线划分，小节内第一拍是强拍(Downbeat).拍子和小节提供了高层(High-level)的节奏信息，拍子检测/估计/推理(Meter Detection/Estimation/Inference)、小节线检测(Bar line/Measure Detection)及强拍检测/估计(Downbeat Detection/Estimation)在音乐识谱、和弦分析等很多 MIR 任务中都有重要应用.

一个典型的检测拍子的方法是首先计算节拍相似性矩阵(Beat Similarity Matrix)，利用它来识别不同部分的重复相似节拍结构[69].利用类似的思路，即节拍相似性矩阵，可进行小节线检测.使用之前小节线的位置和估计的小节长度来预测下一个小节线的位置.该方法不依赖于打击乐器，而且可以在一定程度上容忍速度的变化[70].对于没有鼓声的音乐信号，通过检测和弦变化的时间位置，利用基于四分音符的启发式音乐知识进行小节线检测[64].音乐并不一定都是匀速进行，经常会出现渐快、渐慢等抒情表现形式，甚至出现 4/4 或 3/4 拍子穿插进行的复杂小节结构(如额尔古纳乐队演唱的莫尼山)，这给小节推理(Meter Inference)算法带来巨大困难.文献[71]提出一个基于稀疏 NMF(Sparse NMF)的非监督方法来检测小节结构的改变，并进行基于小节的分割.

强拍估计可以确定小节的起始位置，并通过周期性分析进一步获得小节内强拍和弱拍的位置，从而得到拍子结构[72].对拍子和小节线检测都非常有益.文献[73]将强拍检测和传统的节拍相似性矩阵结合，进行小节级别(Bar-level)的自动节奏分析.早期强拍序列预测方法采用 Onset、Beat 等经典节奏特征及回归模型[74]，近年随着深度学习(Deep Learning)技术的成熟，出现了数个数据驱动的强拍检测算法.文献[75]使用深度神经网络(Deep Neural Networks，DNN)在音色、和声、节奏型等传统音乐特征上进行自动特征学习(Feature Learning)，得到更能反映节奏本质的高层抽象表示，使用 Viterbi 算法进行时域解码后得到强拍序列.类似地，文献[76]从和声、节奏、主旋律(Main Melody)和贝斯(Bass)4 个音乐特征出发进行表示高层语义的深度特征(Deep Feature)学习，使用条件随机场(CRF)模型进行时域解码得到强拍序列.文献[77]使用两个递归神经网络(Recurrent Neural Networks，RNN)作为前端，一个在各子带对节奏建模，一个对和声建模.输出被结合送进作为节奏语言模型(Rhythmical Language Model)的动态贝叶斯网络(Dynamic Bayesian Network，DBN)，从节拍对齐的音频特征流中提取强拍序列.

2.2.5 节奏型检测

音乐节奏的主体由经常反复出现的具有一定特征的节奏型(Rhythmic Pattern)组成.节奏型也可以叫做节拍直方图(Beat Histogram),在音乐表现中具有重要意义,使人易于感受便于记忆,有助于音乐结构的统一和音乐形象的确立.节奏型经常可以清楚地表明音乐的流派类型,如布鲁斯、华尔兹等.

该子领域的研究不多,但早在 1990 年就提出了经典的基于模板匹配的节奏型检测方法[78].另一项工作也使用基于模板匹配的思路,对现场音乐信号进行节奏型的实时检测,注意检测时需要比节奏型更长的音频流.该系统能区分某个节奏型的准确和不准确的演奏,能区分以不同乐器演奏的同样的节奏型,以及以不同速度演奏的节奏型[79].鼓是控制音乐节奏的重要乐器,文献[80]通过分析音频信号中鼓声的节奏信息进行节奏型检测.打击乐器的节奏信息通常可由音乐信号不同子带的时域包络线进行自相关来获得,具有速度依赖性.文献[81]对自相关包络线的时间延迟(Time-lag)轴取对数,抛弃速度相关的部分,得到速度不变的节奏特征.除了以上信号处理类的方法,基于机器学习的方法也被应用于节奏型检测.文献[82]使用神经网络模型自动提取单声部/单音或多声部/多音符号音乐的节奏型.文献[83]基于隐马尔科夫模型从一个大的标注节拍和小节信息的舞曲数据集中直接学习节奏型,并同时提取节拍、速度、强拍、节奏型、小节线.

2.3 音乐和声

音乐通常是多声部/多音,包括复调音乐(Polyphony)和主调音乐(Homophony)两种主要形式.复调音乐拥有漫长的历史,从公元 9 世纪到 18 世纪前半叶流行于欧洲.18 世纪后半叶开始到现在,主调音乐逐渐取代了复调音乐的主要地位,成为最主要的音乐思维形式.复调音乐含有两条或以上的独立旋律,通过技术处理和谐地结合在一起.主调音乐以某一个声部作为主旋律,其他声部以和声或节奏等手法进行陪衬和伴奏.特点是音乐形象明显,感情表达明确,欣赏者比较容易融入.其中,和声(Harmony)是主调音乐最重要的要素之一.和声是指两个或两个以上不同的音符按照一定的规则同时发声而构成的声音组合[84].与和声相关的 MIR 子领域有和弦识别及调高检测.

2.3.1 和弦识别

和弦(Chord)是音乐和声的基本素材,由 3 个或以上不同的音按照三度重叠或其他音程(Pitch Interval)结合构成,这是和声的纵向结构.在流行音乐和爵士乐中,一串和弦标签经常是歌曲的唯一标记,称为所谓的主旋律谱(Lead Sheets).此外,和弦的连接(Chord Progressions)表示和声的横向运动.和声具有明显的浓、淡、厚、薄的色彩作用,还能构成乐句、乐段,包含了大量的音乐属性信息.和弦识别在音乐版本识别、音乐结构分析等多个领域具有重要作用,它另一个有趣的应用是为用户的哼唱旋律自动配置和弦伴奏[85].

典型的和弦识别(Chord Detection)算法包括音频特征提取和识别模型两部分.音高类轮廓(Pitch Class Profile, PCP)是描述音乐色彩的半音类(Chroma)特征的一个经典实现,是一个 12 维的矢量.由于其在 12 个半音类(C、$^{\#}$C/$^{\flat}$D、D、$^{\#}$D/$^{\flat}$E、E、F、$^{\#}$F/$^{\flat}$G、G、$^{\#}$G/$^{\flat}$A、A、$^{\#}$A/$^{\flat}$B、B)上与八度无关的谱能量聚集特性,成为描述和弦及和弦进行的首要特征(如图 6 所示,所有灰色的音符 C 或 c,不管其在哪个组或八度,其频谱能量均相加得到 C 半音类的 Chroma 值.其余依此类推).对传统 PCP 特征进行各种改进后,又提出 HPCP(Harmonic PCP)、EPCP(Enhanced PCP)、MPCP(Mel PCP)等增强型特征.这些特征在一定程度上克服了传统 PCP 特征在低频段由于各半音频率相距太近而引起的特征混肴的缺陷,而且增强了抗噪能力[86].文献[87]没有使用所有的频率来计算 Chroma,它首先检测重要频率(Salient Frequencies),转换为音符后按照心理物理学(Psychophysics)规则为泛音加权.近年来,随着深度学习的流行,文献[88]采用其特征学习能力自动获取更抽象的高层和声特征.

常规的和弦检测算法以固定长度的 Frame 进行音频特征计算,不符合音乐常识.更多的算法基于 Beat 级别的分割进行和弦检测[89].这符合和弦基本都是在小节开始或各个节拍处发生改变的音乐常识,也通常具有更好的实验结果[90].例如,文献[91]使用一个和声改变检测函数在各 Beat 位置分割时间轴,对每个片段的平均 Chroma 再进行和弦识别.流行音乐包含鼓、钹等各种打击乐器,对只依赖于和声成分的和弦检测带来干扰.文献[92]使用 HPSS(Harmonic/Percussive Sound Separation)技术进行预处理,压制

鼓声，强调基于和声成分的 Chroma 特征和 Delta-Chroma 特征.

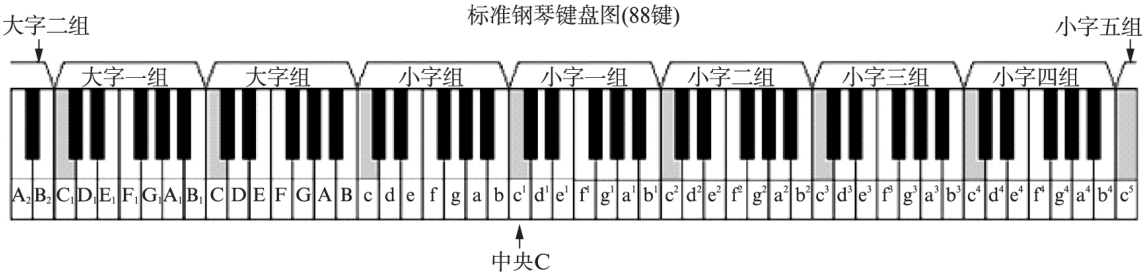

图 6　PCP(Chroma)特征计算原理图

Fig.6　Principle of PCP(Chroma) feature calculation

在识别阶段，早期的方法采用基于余弦距离等的模式匹配(Pattern Matching)[89,93].随着研究的深入，隐马尔科夫模型(HMM)、条件随机场(CRF)、支持向量机(Support Vector Machine，SVM)、递归神经网络(RNN)等机器学习分类方法陆续被引入来建立和弦识别模型[94].和弦序列与根音(Bass Notes)、小节线位置(Measure Positions)及调高(Key)等音乐要素互相关联，文献[95]使用一个 6 层的动态贝叶斯网络(DBN)从音频波形中对它们同时估计，其中 4 个隐藏层联合模拟和弦、调高、根音、小节线位置，两个观察层模拟对应于低音(Bass)和高音(Treble)音色内容的低层(Low-level)特征.

音乐和弦并非独立存在，而是按照特定的音乐规则在时间轴上连接和行进[96].使用音乐上下文(Context)信息进行时间序列后处理通常可以提高性能.常用的方法有考虑和弦相对稳定性的平滑算法[89]、时间序列解码 Viterbi 算法[97]、递归神经网络[94]等.目前的和弦检测算法通常只能识别 12 个大小调的 24 个常用三和弦，对于更复杂的和弦如七和弦只有少数算法[98-101]才能实现，而且随着分类数量的增加，性能有所下降.

2.3.2　调高检测

调性(Tonality)是西方音乐的一个重要方面.调性包括调高(Key)和大小调(Major/Minor)，是调性分析(Tonality Analysis)的一个重要任务.在乐理知识中，如图 7 所示，C、D、E、F、G、A、B 是音名(Pitch Names)，对应于钢琴上真实的键.Do、Re、Mi、Fa、So La、Si 是唱名(Syllable Names)，随着音乐的调高而变化.例如 C 大调中的 Do 就是 C，对应的音阶(Scale)是 C、D、E、F、G、A、B；D 大调的 Do 就是 D，对应的音阶是 D、E、$^{\sharp}$F、G、A、B、$^{\sharp}$C.检测一首曲子的调高及调高的变化对于音乐识谱、和弦检测、音乐情感计算、自动伴奏、音乐结构分析、音乐检索等领域都有重要作用[102-105].

一个典型的调高检测(Key Detection)模型由两部分组成，即特征提取与调高分类(Classification)[106]或聚类(Clustering)[107].目前用来描述对调高感知的音频特征基本都是 PCP(Chroma)，还有基于音乐理论和感知设计的特征[108].尽管缺乏音乐理论支持，心理学实验表明这是一种有效的方法.比如著名的 Krumhansl-Schmukler 模型就是通过提取 PCP(Chroma)特征来描述人们对调高的认知机制[109].调高检测通常使用的调高分类器包括人工神经网络(ANN)[110-111]、隐马尔科夫模型(HMM)[112]、支持向量机(SVM)[108]、集成学习(Ensemble Learning-AdaBoost)[113]，有的算法还使用平滑方法减少调高的波动[111].大多数流行歌曲只有一个固定的调高，少数流行歌曲会在副歌这样具有情感提升的部分变调.古典音乐则经常发生更多的变调.对于这些复杂情况，需要进行调高的局部化检测.

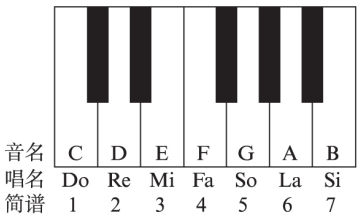

图 7　音名、唱名对照图

Fig.7　The correspondence plot of pitch names and syllable names

2.4　歌声信息处理

音乐包括声乐和器乐两种形式.在声乐中，歌声通常承载音乐的主旋律，与节奏及和声一起构成音乐的 3 大要素.在 MIR 中与歌声信息处理(Singing Information Processing)相关的子领域包括歌声检测、歌声分离、歌手识别、歌唱评价、歌词识别等.其中歌声检测是其他子领域的前处理步骤.

2.4.1 歌声检测

歌声检测(Vocal/Singing Voice Detection)的任务是判定整首歌曲中哪些部分是歌声,哪些部分是纯乐器伴奏.歌声检测算法一般包含以下几个步骤:将歌曲分成几十毫秒长的音频帧(Frame),从各帧中提取能够有效区分歌声和伴奏的音频特征,利用基于规则的门限方法或者基于机器学习的统计分类方法对特征进行歌声/非歌声(Vocal/Nonvocal)分类,考虑到歌声的连续性,最后还需对帧级别的分类结果进行平滑纠错处理.除了 Frame,还可以节拍(Beat)为单位进行歌声检测[114].

歌声检测算法使用的音频特征包括谱特征(Spectral Features)、和声、频率颤音(Vibrato)、幅度颤音(Tremolo)、音色(Timbre)、歌声共振峰(Singing Formant)、主音高(Predominant Pitch)、梅尔频率倒谱系数(Mel-frequency Cepstral Coefficients, MFCC)、线性预测倒谱系数(Linear Prediction Cepstral Coefficients, LPCC)、线性预测系数(Linear Prediction Coefficients, LPC)等[115-122],使用的分类器也从高斯混合模型(Gaussian Mixture Model, GMM)[123]、HMM[115]、SVM[114]扩展到 RNN[124]等深度学习手段.

基于监督学习(Supervised Learning)的歌声检测方法需要大量歌声/非歌声片段的手工标注,费时费力,价格昂贵.为解决这个问题,文献[125]将主动学习(Active Learning)技术集成进传统的基于 SVM 的监督学习方法,极大地减少了标注数据量.文献[126]提出另一个有趣的方法,利用 MIDI 文件中音符 Onset/Offset(起始点/截止点)提供的准确的 Vocal/Nonvocal 边界,将 MIDI 合成的音频使用动态时间规整(Dynamic Time Warping, DTW)与真实音频进行对齐,从而用极少的代价获得大量的真实音频的歌声/非歌声标注训练数据.文献[127]使用 Vibrato、Attack-decay、MFCC 等特征对基于 HMM 的 Vocal/Nonvocal 分类器进行协同训练(Co-training).首先,使用 Vibrato 训练 HMM,从自动标记的测试歌曲片段中挑选最可靠的部分加入到标注歌曲训练集;然后,对 Attack-decay 重复同样的过程;最后,用 MFCC 得到最后的 Vocal/Nonvocal 片段.协同训练充分利用丰富的未标注歌曲,大大减少了手工标注的工作量和计算代价.

进行歌声检测还有少数非监督学习(Unsupervised Learning)算法.文献[128]使用 K 奇异值分解法(K-Singular Value Decomposition, K-SVD)进行短时特征稀疏表示的字典学习(Dictionary Learning),估计每个码字(Code Word)出现的概率,并用其计算每帧加权函数的值,还使用一个二值门限将音乐分割为 Vocal 和 Nonvocal 片段.高阶泛音的时域波动是表征歌声的明显特征[119].文献[129]用正弦波模型(Sinusoid Model)将泛音随时间变化的频率模拟为一个缓慢变化的频率加上一个正弦调制.对泛音相似性矩阵(Partial Similarity Matrix)进行聚类,将和声相关的泛音聚类到各个音源,其中一个音源很可能是歌声,有利于对歌声检测和分离.

上述 Frame-level 的歌声检测经常会出现碎片化的结果.考虑到音乐中的歌唱区域前后关联并非独立存在,还需进行时域平滑后处理(Post-processing)去掉短时突变点.一类方法是通过一阶、二阶差分或方差等表示特征级(Feature-level)时域上下文关系[118],另一类方法是使用中值滤波、自回归滑动平均(Auto Regressive Moving Average, ARMA)[123]、考虑过去时域信息的模型如 LSTM-RNN[118]、或同时考虑过去和未来时域信息的模型如 BLSTM-RNN[124].

2.4.2 歌声分离

歌声分离(Vocal/Singing Voice Separation)是指将歌声与背景音乐伴奏进行分离的技术.在旋律提取、歌手识别、哼唱/歌唱检索、卡拉 OK 伴奏、歌词识别、歌唱语种识别等应用领域有十分重要的作用.由于歌声与由多种音调类乐器组成的伴奏都是和声的(不能被视为噪声),并按照和声结构耦合在一起(即基频或泛音重叠),而且还有多种打击类乐器的干扰,因此歌声分离具有相当大的难度和挑战性.简单的去噪(Denoise)方法并不适用[130].一个通用的方法可以将之视为盲源分离(Blind Source Separation, BSS)问题,如文献[131]结合独立成分分析(ICA)和小波门限方法(Wavelet Thresholding)来分离歌声.但是此类方法没有利用任何音乐信号本身的信息,通常效果较差.专门的歌声分离算法根据输入音乐信号音轨(Vocal Track)的数量可分为立体声歌声分离(Stereo Vocal Separation)和单声道歌声分离(Monaural Vocal Separation).注意不论音轨数量,输入都是多声部/多音音乐.

从立体声中分离歌声的传统方法假设歌声位于中央信道,利用声源的空间差异(Spatial Diversity)来

定位和分离歌声.空间方法的结果可以接受,但有很多由于中央信道估计不准带来的失真和虚假部分(Distortions and Artifacts).文献[132]将基频 f_0 信息集成进来.首先使用 MuLeTs 立体声分离算法预分离,得到它的 f_0 序列,然后使用 HMM 对音高包络线进行平滑后再分离歌声及非歌声区域.文献[133]首先利用双耳信息即信道间强度差(Inter-channel Level Difference, ILD)和信道间相位差(Inter-channel Phase Difference, IPD)进行位于中央的歌声粗略分离,之后使用 GMM 对混合信号频域的低层分布进行聚类.文献[134]将歌声部分模拟为源/滤波器模型(Source/Filter Model),伴奏模拟为 NMF 的成分混合,立体声信号被假设为歌声与伴奏的瞬时混合,然后使用最大似然法(Maximum Likelihood)联合估计两个信道的所有参数.

近年来更多的研究集中于从单声道真实录音中提取歌声,与立体声歌声分离相比更加困难.单声道音乐信号分离主要包括以下几种技术框架:

(1) 基于音高推理(Pitch-based Inference)获得歌声泛音结构的分离技术[135].具体地说,就是从混合音乐信号中首先估计歌声基频 f_0 包络线,然后通过 f_0 及其泛音成分来提取歌声.相反,如果歌声首先被从混合音乐信号中准确地提取出来,f_0 的估计也会更容易.这实际是一个鸡生蛋、蛋生鸡的问题.为克服该限制,文献[136-137]提出一种迭代的方法,即首先用鲁棒主成分分析(Robust Principal Component Analysis, RPCA)算法初步分离歌声,从分离的歌声信号中初步估计歌声主旋律的 f_0 包络线,并在基频重要性频谱图(f_0 Salience Spectrogram)中寻找最佳时域路径,之后将 RPCA 的时频掩蔽(Time-frequency Mask)和基于 f_0 和声结构的时频掩蔽相结合,与上一步骤迭代进行更准确的歌声分离.

(2) 基于矩阵分解技术,如非负矩阵分解(NMF)、鲁棒主成分分析(RPCA)等的分离技术.基于 NMF 的算法分解音乐信号频谱,选择分别属于歌声和伴奏的成分并合成为时域信号[138-139].RPCA 算法将混合音乐信号分解为一个低秩成分(Low-rank Component)和一个稀疏成分(Sparse Component).因为伴奏本身是重复的,所以被认为是低秩子空间.而歌声根据其特点则被认为是中等稀疏的[140].基于此思想,更适合音频的 RPCA 的非负变种即 RNMF(Robust Low-rank Non-negative Matrix Factorization)也被引入进行歌声分离[141].采用矩阵分解技术有不同的粒度,粗粒度方式如 RPCA 利用歌声的特定性质直接将音乐信号分解为歌声和伴奏,细粒度方式如 NMF 将音乐信号分解为一套细化(Fine-grained)的成分,并重新组合来产生目标声源估计.文献[142]将两种方式以级联的方式结合,将音乐信号分解为一套中粒度成分(Mid-level Components).该分解足够细来模拟歌声的不同性质,又足够粗以保持该成分的语义,使得能直接组装出歌声和伴奏.

(3) 基于计算听觉场景分析(Computational Auditory Scene Analysis, CASA)的分离技术.CASA 是一种新兴的声音分离计算方法[143],它的基本原理来自于 Bregman 提出的听觉场景分析(Auditory Scene Analysis, ASA)[144],主要思路就是利用感知线索(Cue)把混合音频信号分别组织进对应于不同声源的感知数据流.听觉场景分析(ASA)认为人类听觉器官的感知过程一般可以分成两个主要过程: 分割阶段(Segmentation)和聚集阶段(Grouping).在分割阶段,声学输入信号经各种时频变换(Time-frequency Transforms)被分解为时频单元(T-F Units),各单元被组织进片段(Segments),每个片段可近似认为来自单一声源.在聚集阶段,来自同一音源的 Segments 根据一套规则被合并到一起.分割片段和聚集使用的 Cues 包括时频近似度(T-F Proximity)、和声、音高、Onset/Offset、幅度/频率调制、空间信息等.受 ASA 启发,CASA 利用从人类听觉系统中获得的知识来进行声音分离,希望获得接近人类水平的分离性能.与其他声音分离方法相比,CASA 利用声音的内在性质进行分离,对声源进行了最小的假设,展现出极大的潜力.

文献[145-146]是基于 CASA 的单声道歌声分离的创始性算法.首先进行 Vocal/Nonvocal 部分检测,之后进行主音高检测得到歌声区域的歌声包络线,最后基于分割和聚集进行歌声分离.在分割阶段,基于时域连续性(Temporal Continuity)和交叉信道相关(Cross-channel Correlation)将时域连续的 T-F 单元合并成 Segments.在聚集阶段,直接应用检测到的音高包络线.简言之,如果一个 T-F 单元的局部周期性与该帧检测的音高相匹配,那么该单元被标记为歌声主导(Singing Dominant).如果一个 Frame 中大多数的 T-F 单元被标记为歌声主导,那么这个 Frame 就被标记为歌声主导.如果一个 Segment 有超过一半的

Frames 是歌声主导,那么这个 Segment 就是歌声主导.标记为歌声的 Segments 被聚集为代表歌声的音频流.文献[147]使用 CASA 框架设计一个串联(Tandem)算法,迭代地估计歌声音高(Singing Pitch)并分离歌声.首先估计粗略的音高,然后考虑和声与时域连续性线索(Harmonicity and Temporal Continuity Cues)用它来分离目标歌声,分离的歌声再用来估计音高,如此迭代进行.为提高性能,提出一个趋势估计(Trend Estimation)算法来检测每个 Frame 的歌声音高范围,去除了大量错误的伴奏或歌声泛音产生的音高候选者.掩蔽函数(Masking Functions)是 CASA 类分离算法的核心,最常用的一个叫做理想二值化掩蔽(Ideal Binary Mask, IBM).标记为歌声的时频块具有紧密的时频域上下文关系,文献[148-149]使用深度递归神经网络(RNN)对掩蔽块的时频域连接进行优化,把分离过程作为最后一层的非线性过程,并进行网络的联合优化.

2.4.3 歌手识别

歌手识别(Singer/Artist Identification)是指判断一首歌曲是由集合中的哪个歌手演唱的.在歌手的分类管理、音乐索引和检索、版权管理、音乐推荐等领域都有重要应用.受乐器伴奏、录音质量、伴唱等的影响,歌手识别是一个十分困难的问题.此外,一个歌手每个专辑中的歌曲通常有很多类似之处.比如风格(Style)、配器(Instrumentation)、后处理方式(Post-production)等,使得训练和测试数据分布在同一唱片中时会产生偏高的识别结果,称为唱片效应(Album Effect)[150].除了少量算法采用半解压状态的修正余弦变换(Modified Discrete Cosine Transform, MDCT)系数作为特征直接在 MP3 压缩域上进行歌手分类识别[151-152],绝大多数算法都是以原始格式音频作为输入.

歌手识别借鉴了说话人/声纹识别(Speaker/Voiceprint Recognition)的整体技术框架[153].人类听觉系统(Human Auditory System, HAS)到底通过什么样的感知特征来识别特定的歌声,目前仍不得而知.由于歌声和语音之间在时频结构上的巨大差别,除了用于声纹识别的典型的音频特征如感知线性预测(Perceptual Linear Prediction, PLP)[154]、MFCC[155]等,还需引入更多表示歌声特性的特征,如音色[156]、反应歌手个性化风格的频率颤音倒谱系数(Vibrato Cepstral Coefficients)[157]等.音色是人耳区分不同音乐声音的基础[158].文献[159]从声音的正弦泛音的瞬时幅度和频率估计谱包络线(Spectral Envelope),作为特征输入分类器识别歌手.使用的分类器有 GMM[160]、卷积神经网络(Convolutional Neural Network, CNN)[161]、混合模型[162]等.

由于伴奏的干扰,在进行歌手识别前通常需要进行歌声增强(Singing Enhancement)或伴奏消减(Accompany Reduction)等预处理[163-164].文献[163]采用基于 NMF 分解的歌声分离技术作为预处理.文献[164]首先提取主旋律(Predominant Melody)的和声结构,使用正弦波模型将这些成分重新合成为主旋律,并估计可靠的旋律帧从而实现伴奏消减.虽然增强或分离多声部/多音音乐中的歌声部分作为前处理被相信是一个提高歌手识别任务性能的有效方式,但是因为不可避免地存在失真,而且会传递到后续的特征提取和分类阶段,所以只能带来有限的提高.文献[165]是一种基于 CASA 的歌声增强预处理措施.在每个 Frame 上的二元时频掩蔽(Binary T-F Mask)包括可靠的歌声时频单元,和另外一些不可靠或丢失的歌声时频单元,频谱不完整.为减轻失真,使用两个缺失特征(Missing Feature)方法即重构(Reconstruction)和边缘化(Marginalization)处理不完整的歌声频谱(Vocal Spectrum),再进行歌手识别.与上述主要基于音频分离的方法思路不同,文献[166]提出了一个有趣的方法.从一个很大的卡拉 OK 数据集中手工混合左右声道的清唱歌声和伴奏,并研究清唱和带伴奏歌声之间在倒谱上的变换模型.当一个未知的带伴奏的歌声出现时,即可把带伴奏歌声的倒谱转换到接近清唱的水平,从而有助于后续的歌手分类识别.

当前已有的歌手识别算法在整体框架上与说话人/声纹识别相同,需要事先搜集大量的歌声清唱数据并建立歌声模型.但是与语音数据主要来自普通人群并相对容易获取不同,歌唱主要源自各种级别的艺术家,而且几乎都是带有乐器伴奏.大量搜集每个歌手的无伴奏清唱数据几乎是不可能的.许多歌手识别算法使用带伴奏的歌声进行训练,但效果并不如人意.文献[167]研究了用语音数据代替歌唱数据来刻画歌唱者声音的可能性,结论是很难用说话完全代替歌唱的特性,原因在于大多数人的说话和歌唱具有很大的差异.两个可能的解决思路是:(1) 将歌手的语音数据转换为歌唱数据增加训练数据量,然后使用歌

声驱动的模型进行识别；(2) 将歌手的歌唱数据转换为对应的语音，然后使用大量语音数据训练的模型进行识别.

2.4.4　歌唱评价

歌唱评价(Singing Evaluation)是对演唱的歌声片段做出各方面的正面或负面描述，一直以来都是音乐学界所关注的课题之一.在歌唱表演、歌唱比赛、卡拉 OK 娱乐、声乐教育等场合都具有重要应用.目前已有的自动歌唱评价系统基本集中于卡拉 OK 场景.

早期的算法(如 20 世纪 90 年代中前期)受计算机软硬件的限制，只采用音量(Volume/Loudness)作为评价标准，经常跟人类评价的结果大相径庭.随着 MIR 技术及计算资源的发展，后续算法有了很大发展.常规思路是计算两段歌声中的各种音频特征如音量、旋律线的音高、音准(Intonation)、音程、音符时长、节奏、颤音、音频特征包络线、统计特征等之间的相似度，并给出一个用户表现的评价，如好/坏(Good/Poor)两个质量分类[168]、高/中/低(High/Medium/Low)3 个质量分类[169]或总体评分[170-171].常用分类器有 SVM 等，测量相似性可使用动态时间规整(DTW)技术，由不同特征得到的相似性分数可使用权重结合到一起[172].

因为绝大多数歌曲都包含伴奏，在很多情况下无法找到歌星的清唱录音和用户歌声进行比较.文献[173]直接以歌手的 CD/MP3 带伴奏歌曲作为参考基准，以音高、强弱、节奏为特征与用户的歌声进行比较，除了给出用户的总体歌唱评价，还指出哪里唱的好和不好.该技术比较接近于人类评价方式.类似地，文献[174]的评价包括是否跑调或走音(Off Key/Be in Tune)，这需要以很高的频率分辨率进行基频估计，通常小于几个音分(Cent).以上算法虽有一定效果，但很少考虑各个评价要素之间的关系.以系统评价和人类评价之间的相关系数定量衡量，经常与人类评价结果相去甚远[175].

另一个难题是如何进行歌唱的高级评价.即在跟准节奏和音高的前提下，如何像人类专家那样对音色具有何种特点、是否有辨识度，音域(Pitch Range)是否合适，吐字是否清晰，演唱是否感情饱满等进行高级评价.这方面的研究工作很少.文献[176]采用歌声相关的特征如频率颤音、和声噪音比(Harmonic-to-noise Ratio)针对中文流行歌曲的 6 种歌唱音色进行分类，这 6 种音色是浑厚(Deep)、沙哑(Gravelly)、有力(Powerful)、甜美(Sweet)、空灵(Ethereal)、高亢(High pitched).文献[177]使用基频、共振峰、和声及残差谱(Residual Spectrum)作为特征，识别用户的发音区域并分析其声音质量.推荐一个更合适的音域，避免唱破音出现嗓子疼痛等声带健康(Vocal Health)问题.对于歌曲来讲，歌词的正确发音也是非常重要的一个方面.好的歌手如邓丽君经常被评价为字正腔圆.文献[178]使用谱包络线和音高作为特征，混合高斯模型(GMM)和线性回归(Linear Regression)作为分类器，自动对歌唱元音(Singing Vowel)的质量进行分类评价.除了以上基于声学的歌唱评价方法，文献[179]通过测量嘴和下颌(Mandible)移动的肌电图(Electromyography，EMG)进行歌唱评价.

2.4.5　歌词识别

歌词识别(Lyrics Recognition/Transcription)与语音识别(Speech Recognition)问题的总体目标类似，都是把语言转换为文本；总体技术框架也类似，都包括声学模型和语言模型.但是由于歌唱和说话在声学和语言特性上的巨大差别，具体技术实现上需针对歌唱的特点进行更有针对性的设计.从声学模型来看，歌唱是一种特殊的语音.日常语音基本可视为匀速进行，音高变化范围很小.歌唱则需要根据音乐的旋律和节奏，以及颤音、转音等艺术技巧控制声带的发声方式、时间和气息的稳定性.一个普遍的现象是同一个人歌唱和说话的音色具有很大不同.从语言模型来看，歌词具有一定的艺术性，还需要押韵，也与日常交流的语言具有很大区别.此外，与单纯的语音不同，歌声几乎都是与各种乐器伴奏混合在一起，涉及信号分离的难题.而且搜集清唱歌声与歌词对应的训练数据十分困难.这些特点使得传统的语音识别模型无法直接使用，带来巨大挑战.

虽然很多歌曲的歌词已经被人工上传到各个音乐网站，但是仍然有很多歌曲缺少歌词.此外，如果歌词识别具有一定的准确率，则可以将基于声学特征(Acoustic Features)的歌唱检索转换为更成熟的基于文本的检索[180]，或者帮助基于音频特征的歌唱检索.另外，歌词识别在歌曲分类、歌词与音频或口型对齐上也有一定应用.

目前为止,仅有极少数歌词识别的算法被提出,而且识别准确率很低.类似于经典语音识别框架,文献[181]使用音乐特征及 HMM 模型进行歌词识别,用合成的歌声解决缺乏训练数据的问题,在无乐器伴奏的歌声中(A Cappella)中识别音节(Syllable),并且为了响应音素(Phoneme)长度的变化,构建一个依赖于长度(Duration Dependent)的 HMM.文献[182]利用类似的语音识别模型,用有限状态自动机(Finite State Automaton, FSA)描述识别语法,将待识别的歌词约束为数据库中存储的歌词.文献[183]对音乐中的音素进行识别(Phoneme Recognition)以帮助歌词识别.在有视觉信息的情况下,还可以发展基于视觉的歌词识别方法.文献[184]利用光学字符识别(Optical Character Recognition, OCR)引擎自动识别 YouTube 网站上下载的视频中的歌词.文献[185]通过口型信息帮助歌词识别,但效果有限.

2.5 音乐搜索

音乐搜索(Music Retrieval)是指在给定某种形式的查询(Query)时,在数据库中检索与之匹配的结果集并按相关性从高到低返回的过程.按查询输入的形式,可进一步分为 5 个子领域:音乐识别、哼唱及歌唱检索、音乐版本识别或翻唱识别、节拍检索、音乐借用.

2.5.1 音乐识别

音乐识别(Music Identification/Recognition)在产品上又称为听歌识曲.通常用手机或麦克风录制 10 s 左右的音乐作为查询(Query)片段,计算其音频指纹后与后台音频指纹库中的记录进行匹配,并将最相似记录的歌曲名字、词曲作者、歌唱者甚至歌词等相关元数据返回.音频指纹是指可以代表一段音频重要声学特征的基于内容的紧致数字签名,音频指纹技术(Audio Fingerprinting)是音乐识别的核心,当扩展到一般音频时也可称为基于例子的音频检索(Query by Example, QBE).音频指纹算法首先提取各种时频域音频特征,对其建模后得到指纹,之后在指纹库中进行基于相似性的快速匹配和查找[186].

常用的音频特征有 Chroma[187]、节奏直方图(Rhythm Histogram)[187]、节拍[188]、经 KD 树量化的旋律线字符集[189]、音高与时长[190]、树量化的 MFCC 峰值和谱峰值(Spectral Peaks)字符集[191]、MPEG-7 描述子[192]、频谱图局部峰值[193]、从音乐的二维频谱图上学习到的图像特征[194]、MP3 压缩域的听觉 Zernike 矩(Auditory Zernike Moment)[195]等.各种特征经常融合在一起使用.

录制的输入片段有可能经受保持音高不变的时间伸缩(TSM)和变调(PS)这两种失真.与一般的噪声类失真影响音频质量不同,这两种失真主要会引起音频指纹在时频域上的移动,产生同步失真(Desynchronization),从而使查询片段指纹与数据库音频指纹集匹配失败.文献[196]从二维频谱图上提取计算机视觉中的 SIFT(Scale Invariant Feature Transform)描述子作为音频指纹,利用其局部对齐(Local Alignment)能力抵抗 TSM 和 PS 失真.

为使提取的特征更鲁棒,还需进行去噪、回声消除(Echo Cancellation)[188]等预处理.常用的指纹匹配方法包括经典的 TF-IDF(Term Frequency-Inverse Document Frequency)打分匹配、局部敏感哈希(Local Sensitive Hashing, LSH)、倒排表(Inverted Table)等[189,191].文献[197]使用一个有趣的方法加速查询匹配,即首先自动识别 Query 和数据库歌曲的流派,在匹配阶段只计算 Query 与数据库中具有同样流派的歌曲的相似性.

2.5.2 哼唱/歌唱检索

哼唱/歌唱检索(Query by Humming/Singing, QBH/QBS)通常用麦克风录制长短不一的哼唱或歌唱声音作为查询片段,计算音频特征后在数据库中进行相似性匹配,并按匹配度高低返回结果列表,最理想的目标是正确的歌曲排名第一返回.典型的应用场景是卡拉 OK 智能点歌.与上述音乐识别技术相比,哼唱/歌唱检索不仅同样面临由于在空气中录音而引起的信号质量下降,还面临跑调、节拍跟不上等新的困难.哼唱/歌唱检索的结果与用户的哼唱/歌唱查询片段质量高度相关,存在很大的不确定性.

除了常规时频域音频特征,能够在一定程度上反应音乐主旋律走向的中高层音频特征更适合于哼唱/歌唱检索.早期的哼唱检索系统[198]根据音高序列的相对高低(Relative Pitch Changes)用 3 个字符即方向信息来表示旋律包络线,并区分不同的旋律.这 3 个字符是'U'、'D'、'S',表示当前音符的音高分别高于、低于、等于前一个音符的音高.3 个字符表示的主要问题比较粗糙,于是文献[199]把旋律包络线扩展为用 24 个半音(Semitone)字符来表示,即当前音的一个正负八度.类似地,文献[200]也采用音程作为

特征矢量，并增加了分辨率.与以上均不相同，文献[201]没有使用明确的音符信息，而是使用音符出现的概率来表示旋律信息.除了音高类的特征，文献[202]利用音长和音长变化对旋律进行编码，以获得更加精确的旋律表示.文献[203]采用符合 MPEG-7 的旋律序列，即一系列音符长度和长度比例作为 Query.与上述不同，文献[204]从数据驱动出发，用音素级别(Phoneme-level)的 HMM 模拟哼唱/歌唱波形的音符片段，用 GMM 模拟能量、音高等特征，更加鲁棒地表示用户哼唱/歌唱的旋律.

由于用户音乐水平不一，哼唱及歌唱的查询片段经常出现音符走音、整体跑调、速度及节拍跟不上伴奏等现象.即使用户输入没有问题，音符及节拍的分割和音高识别算法也不会 100%准确，可能产生插入或删除等错误[205].给定一个不完美的 Query，如何准确地从大规模数据库里检索是一个巨大挑战.为克服音符走音现象，文献[206]对查询片段的音高序列进行平滑，去除音高检测或用户歌唱/哼唱产生的异常点(Outlier).绝大多数算法采用相对音高序列，而不是绝对音高.只要保持相对音高序列的正确，那么对于个别音符走音现象以上算法都是具有一定容错性的[207].为克服哼唱和原唱之间的速度差异，文献[208]首先使用原始 Query 来检索候选歌曲，如果结果不可靠，Query 片段被线性缩放两倍重新检索.如果还不可靠，则缩放更多倍数.日本的卡拉 OK 歌曲选择系统 Sound Compass[209]也采用类似的时间伸缩方法.为克服整体跑调现象，绝大多数算法采用调高平移(Key Transposition)办法进行纠错[208-209].用户哼唱的各种错误经常在开头和结尾处出现.基于此假设，文献[210]认为只有 Query 的中间部分是属于某个音乐的子序列.对两个流行的局部对齐方法即线性伸缩(Linear Scaling，LS)和 DTW 扩展后进行匹配识别.为抵抗 Query 可能存在的各种错误，文献[205]采用音频指纹系统描述重要的旋律信息，更好地比较 Query 和数据库歌曲.

另一个影响匹配性能的因素是如何切割 Query 和数据库完整歌曲的时间单元.文献[211]表明基于音符的切割比基于帧的分割处理更快.文献[200]表明所有音乐信息基于节拍分割会比基于音符分割对于输入错误更加鲁棒.文献[212]进一步提出一种基于乐句(Music Phrase)分割和匹配的新方法，使得匹配准确率大大提升.基于节拍和乐句的分割不仅性能更好，而且也更符合音乐语义.类似地，文献[213]也采用乐句尺度的分段线性伸缩(Phrase-level Piecewise Linear Scaling)，基于 DTW 或递归对齐(Recursive Alignment)进行旋律匹配，并将每个乐句的旋律片段约束在一个有限范围内调整.

在旋律相似性匹配方面，最直接的技术是字符串近似匹配/对齐技术[203].随着研究的深入开始引入动态时间规整(DTW)、后辍树索引、隐马尔科夫模型(HMM)、K 近邻(K-nearest Neighbor)、N-gram、基于距离的相似性(Distance-based Similarity)、基于量化的相似性(Quantization-based Similarity)、基于模糊量化的相似性(Fuzzy Quantization-based Similarity)等各种方法[209,214-215].大多数哼唱/歌唱检索方法在时间域测量音符序列之间的距离，文献[216]在快速傅里叶变换(Fast Fourier Transform，FFT)频域计算音符序列之间的欧氏距离(Euclidean Distance)，匹配速度有所加快.基于动态规划(Dynamic Programming，DP)的局部对齐方法穷尽两个音乐片段之间所有可能的匹配，返回最佳局部对齐，缺点是计算代价太高，在指数增长的数据库规模下，准确的局部对齐已不可能.因此，对于大规模数据库需要更高效的检索办法.文献[217]实现基于音符音高的 LSH 索引算法，筛选候选片段，使用线性伸缩来定位候选者的准确边界.文献[218]使用多谱聚类(Multiple Spectral Hashing，MSH)得到特征矢量，之后用改进的 DTW 进行相似性匹配.文献[219]利用显卡 GPU(Graphic Processing Unit)硬件强大的计算能力，实现局部对齐的快速算法，速度提高超过 100 倍.基于不同匹配策略可得到不同的结果，文献[220]在打分级别(Score-level)融合多个结果.

2.5.3　音乐版本识别或翻唱识别

音乐版本识别(Cover Song Identification，CSI)或翻唱识别判断两首音乐是否具有同样的本源.众所周知，很多音乐经过重新编曲、演唱和演奏后会形成很多版本.这些不同的版本会保持主旋律基本相同，但是音乐结构、调高、节奏(速度)、配器(音色)、歌唱者性别、语言等都可能会发生巨大变化[221].人类大脑具有高度的抽象思维、逻辑推理能力，识别多版本音乐轻而易举.但是音频数据的改变，却使机器识别相当困难.

绝大多数音乐版本识别算法使用一个通用框架，包括特征提取和模式匹配两步.在特征提取阶段，对

应于各个音乐要素的高层特征很难准确计算.因此,直接采用高层特征的算法难以得到希望的效果.低层特征无法反映音乐语义,仅有少数 CSI 算法采用低层特征,如音色形状序列(Timbral Shape Sequences),即一首歌曲的经过量化的平滑频谱的相对改变[222].为弥补高、低层音频表示之间的鸿沟,在 CSI 中经常使用既能在一定程度上反映高层特征又能比较准确计算的中层特征(Mid-level Feature).文献[223]提出一个集成旋律和节奏特性的中层特征.首先进行节拍跟踪,产生独立于速度变化的节拍对齐(Beat-synchronous)表示.之后在连续音频节拍上进行多音高检测,在检测的音高上提取旋律线用于后续检索.文献[224]也类似地采用主旋律作为特征.多版本音乐保持主旋律基本不变,会使在此基础上配置的和声在很多情况下也保持基本不变.所以,和声类的特征也可以用于 CSI.文献[225]采用半音类 Chroma/PCP 及其变种.文献[226]采用随时间变化的半音类图(Chromagram).文献[227]在 Chromagram 中去掉相位信息,并应用指数分布的频带得到一个对于乐器、速度、时移不敏感的特征矩阵.文献[228]受心理物理学启发,根据人类听觉对相对音高比绝对音高敏感的事实,采用描述相对音高的动态 Chroma 特征(Chroma-based Dynamic Feature).文献[229]在 Chroma 基础上用深度学习中的自编码器(Auto Encoder)学习一个更能刻画音乐本质的中间表示.文献[230]使用在 MIDI 训练的 HMM 识别的和弦序列.与以上基于原始格式音频输入的 CSI 算法不同,文献[231]在压缩域直接从 AAC 文件中提取一个低复杂度的有效特征,即在半解码(Semi-decoding)的情况下,直接将 MDCT 系数映射到 12 维 Chroma 特征上.

以上主旋律与和声类的中层特征对配器(音色)、歌唱者性别、语言等变化都具有较强的内在鲁棒性[232].对于节奏或速度变化,可采用基于节拍分割对齐的同步方法[226].对于调高变化,一般采用调高平移的措施[226].更严重的挑战来自多版本音乐中可能存在的音乐结构变化.一个典型的例子是苏芮演唱的“一样的月光”.该歌曲有两个版本,相差一段长 1 分多钟的副歌,前奏也完全改变.更多的例子体现在 CD 版歌曲和演唱会版本的歌曲,通常演唱会版本会在间奏处增加一大段乐手独奏的即兴表演.在音乐结构发生变化时进行 CSI 必须采用额外措施.首先按照前奏、主歌(一般 2 个)、副歌(通常大于 2 个)、桥段、间奏、结尾等部分分割,之后再调用上边的办法在各个局部进行音乐版本识别.文献[233]基于音乐结构分析提出一个新的 CSI 算法,只匹配重要部分(如主歌、副歌)而忽略次要部分,使用加权平均来集成各部分的相似性.

相似性度量一般包括互相关、归一化的 Frobenius 范数、欧氏距离、点积、动态时间规整(DTW)、Smith-Waterman 对齐算法、隐马尔科夫模型(HMM)的最可能隐含状态序列、近似最近邻检索(Approximate Nearest Neighbor Search)等[226-227,230,232,234],在大规模匹配时还可采用局部敏感哈希(LSH)进行索引[223].除了以上常规方法,文献[235]基于信息论(Information Theory)方法研究 CSI 中的音频时间序列之间的相似性问题.在离散情况下计算归一化压缩距离(Normalized Compression Distance, NCD),在连续情况下计算基于信息的相似性度量(Information-based Measures of Similarity).文献[222]基于矢量量化(Vector Quantization)或聚类的思想设计一种 CSI 相似性匹配方法.将滑动窗口的各音频特征映射到以时间排序(Time-ordered)的高维空间中的点云(Point Cloud),同一歌曲不同版本对应的点云可近似认为由旋转(Rotation)、平移(Translation)或伸缩(Scaling)得到.通常融合不同的 CSI 检测算法能得到更好的结果,如文献[236]结合了 DTW 和 Qmax 的结果.

文献[234]研究了诸多要素对 CSI 的影响.实验表明:Chroma 特征比 MFCC 等音色类特征更适合此任务,而且增加分辨率可提高识别准确性.余弦距离比欧氏距离更适合计算 Chroma 序列相似性.最佳平移索引(Optimal Transposition Index, OTI)对调高循环移位,以获得两首歌曲之间的最大相似性,比首先估计调高再平移的方法更准确.采用节拍跟踪、调高估计、旋律提取、摘要提取等中间步骤进行 CSI,由于它们本身并不完全可靠,反而可能会使性能降低.考虑到在各个版本音乐间可能出现较大的歌曲结构改变,进行局部相似性计算或局部对齐是唯一可行的 CSI 检测方法.

2.5.4 敲击检索

敲击检索(Query By Tapping, QBT)是根据输入的节拍信息,从数据库中返回按节拍相似度高低排序的音乐列表.在整个检索过程中没有利用音高信息.随着个人手持设备的普及,QBT 提供了一个通过摇晃或敲击设备的新颖有趣的音乐检索方式.现在该领域的研究成果还较少.

一个典型的方法是提取 Query 中音符时长矢量[237]作为特征，归一化处理后采用动态规划方法在输入时间矢量和数据库时间矢量特征间进行比对并排序返回.类似地，文献[238]采用与文献[237]同样的特征，只是把音符时长的名字改为音符起始点间距(Inter-Onset Interval, IOI)，建立 IOI 比例矩阵(IOI Ratio Matrix)，通过动态规划与数据库歌曲匹配.文献[239]提出的系统叫做 BeatBank，用户在 MIDI 键盘或电子鼓上敲击一首歌的节奏，输入被转化为符合 MPEG-7 的节拍描述子(Beat Description Scheme)作为特征.文献[240]以音频中计算出的峰度(Kurtosis)的变化(Variations)作为节奏特征，并采用局部对齐算法匹配计算相似性分数.

2.5.5　音乐借用

音乐借用(Music Borrowing)有着长期的历史.特别是在当今数字化时代，一个艺术家可以轻易地截取别人作品的某些部分，并将其集成到自己的歌曲中来.借用和被借用歌曲之间共享一个旋律相似的片段.通常音乐界认为适当的借用在艺术创作中是允许的，但是超过一定长度(如 8 小节)就涉嫌抄袭侵权行为.例如，台湾女子组合 S.H.E 的《波斯猫》借用了柯特尔比的《波斯市场》作为副歌部分的旋律，流行歌手卓亚君的《洛丽塔》的前奏来自钢琴曲《致爱丽丝》，李健的《贝加尔湖畔》则在桥段借用了一段俄罗斯民歌的旋律，这些属于正常引用的例子.而王力宏的《大城小爱》和周华健的《让我欢喜让我忧》全曲旋律非常相似，花儿乐队两张专辑中的 13 首歌曲也与其他歌曲旋律高度相似，这些都被指控涉嫌抄袭.从用户的角度来说，发现音乐引用会给用户带来不少乐趣，也可以进行作曲技巧分析.从法律的角度来看，检测歌曲之间的相似片段将有助于音乐作品的版权维护和盗版检测，具有重要的现实意义.

音乐借用最显著的特点是两个音乐作品中的相似片段往往都比较短，而且起止位置随机.对计算机处理而言这是一个难题.音乐借用与翻唱检索和音乐识别之间有区别也有联系，如图 8 所示.它们的相似点都是检测不同歌曲之间的旋律相似的部分，而区别在于翻唱检索基本以整首歌曲为单位，有较长的相似部分，如图 8(a)所示；音乐识别则是固定一个短的片段，在另一首歌曲中检索与之相似的短片段，片段之间只是音频质量不同，如图 8(b)所示；音乐借用也是检测相似短片段，但难点在于不知道相似片段在歌曲中的起始位置及长度，而且，类似于 CSI，片段之间的音色、强弱、配器、速度、调高、语言等都可能不同，如图 8(c)所示.据我们所知该问题在国内外 MIR 领域尚未被研究过.

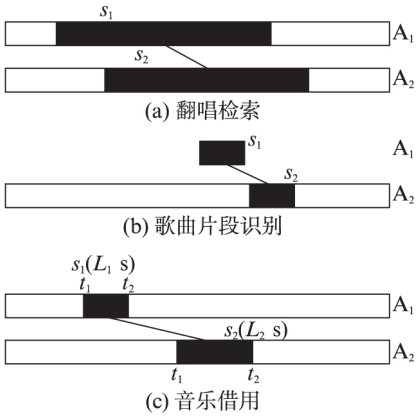

图 8　音乐借用识别与翻唱检索、歌曲片段识别的区别与联系

Fig.8　Difference and relation between music borrowing, cover song identification and music identification

2.6　音乐高层语义分析

如果把音高、旋律、节奏、和弦、调式、歌声等核心层的 MIR 领域理解为低层音乐语义(Low-level Music Semantics)，那么音乐流派/曲风、情感、结构、摘要/缩略图、作曲家及乐器识别等应用层的 MIR 领域则可以理解为高层音乐语义(High-level Music Semantics).低层音乐语义的研究有助于高层音乐语义的分析理解和自动标注，进而基于此进行更有效的音乐搜索和推荐.

2.6.1　音乐流派/曲风分类

音乐流派/曲风(Music Genre/Style)指的是音乐的不同风格.西方音乐通常划分为流行(Pop)、摇滚(Rock)、爵士(Jazz)、乡村(Country)、古典(Classical)、蓝调(Blues)、嘻哈(Hip-hop)和迪斯科(Disco)等类别.如果考虑世界各地的民族音乐，那么划分的类别将更多更复杂.这些分类方法主观性强而且有争议，目前还没有一种通用的绝对标准.音乐流派分类在音乐组织管理、浏览、检索、情感计算和推荐中都有重要应用.

音乐流派分类是一个典型的模式识别问题，通常包括特征提取和统计分类两步.常用特征包括谱特征、倒谱特征、MFCC、频谱图上计算出的纹理特征、音高直方图(Pitch Histograms)等[241-242].特征的时域

特性对分类也很重要[243].常用的分类器有 GMM、SVM[244]、深度神经网络(DNN)[245]、卷积神经网络(CNN)[246]等.为解决流派类别定义的模糊性及有监督学习训练的困难,另一个思路是根据音乐内容分析比如节奏进行聚类,性能与有监督学习的方法相差不多[247].

2.6.2 作曲家识别

作曲家识别(Music Composer Recognition)是指通过听一段音乐并分析音频数据,识别出相应的作曲家信息,基本应用于音乐理论分析等专业场景.该领域的典型方法仍然是音频特征加上统计分类器.用常规低层音频特征刻画作曲家风格和技巧存在较大缺陷,提取高层音乐特征会更有利于挖掘作曲的内在风格和技巧[248].近些年,随着深度学习方法的流行,其从大数据(Big Data)中自动学习特征的能力也被用来深入研究作曲家的风格和技巧[249].文献中使用的分类器还包括决策树(Decision Trees)、基于规则的分类(Rule-based Classification)、SVM 等[250].

2.6.3 智能乐器识别

每个国家、民族都有自己独特的乐器,种类繁多.如西方的风笛(Bagpipes)、单簧管(Clarinet)、长笛(Flute)、羽管键琴(Harpsichord)、管风琴(Organ)、钢琴(Piano)、长号(Trombone)、小号(Trumpet)、小提琴(Violin)、吉他(Guitar)等管弦乐器(Orchestral Instruments),还有鼓(Drum)、铙钹(Cymbal)等各种打击乐器(Percussive Instruments).到现代,还出现了各种模拟和扩展相应声学乐器音色的电声乐器(Electronic Instruments).中国也有很多的民族乐器,如古筝(Guzheng)、古琴(Guqin)、扬琴(Yangqin)、琵琶(Pipa)、二胡(Erhu)、马头琴(Horse-head String Instrument)等.准确识别乐器种类是音乐制作、乐器制造及评估等领域人士必备的专业技能,对音乐搜索、音乐流派识别、音乐识谱等任务都十分有益.

随着音频信息处理技术(Audio Information Processing)及人工智能技术(Artificial Intelligence, AI)的发展,出现了一些智能乐器识别(Intelligent Instrument Recognition)方法.主要的框架仍然是特征+分类器的通用模式识别框架.已用的特征包括 LPC、MFCC、基于常数 Q 变换的倒谱系数、基于频谱图时频分析的音色特征、基于稀疏特征学习(Sparse Feature Learning)得到的特征;已用的分类器包括 GMM、SVM、贝叶斯决策(Bayesian Decision)等[251-255].大多数方法识别单一乐器的声音输入.但是,在现实中音乐基本上都是多种乐器的混合.识别多声部/多音音乐中的各种乐器是一个重要而且更具挑战性的任务.文献[256]基于卷积神经网络(CNN)进行真实多声部/多音音乐中的主乐器(Predominant Instrument)的识别,网络以具有单一标签的固定长度的主乐器音乐片段训练,从可变长度的音频信号中估计主乐器,并在测试音频中采用滑动窗口,对信息输出进行融合.

2.6.4 音乐结构分析

音乐通常由按照层次结构组织的多个重复片段组成.音乐结构分析(Music Structure Analysis)的目的就是把音乐信号分割为一系列时间区域,并把这些区域聚集到具有音乐意义的类别[257].这些类别一般包括前奏(Intro)、主歌(Verse)、副歌(Chorus/Refrain)、间奏(Interlude)、桥段(Bridge)和结尾(Outro).注意主歌、副歌和间奏通常有多个段落,前奏、桥段、结尾则只有一个,而且副歌比主歌具有更高的相似度[258].音乐结构分析既可用于加深对音乐本身的理解,也可以辅助多个其他研究内容如音乐版本识别、乐句划分、音乐摘要、内容自适应的音频水印[259]等.

音乐内部具有高度的重复性,基于音频特征构造自相似矩阵(Self-similarity Matrix)成为结构分析的主要方法.可使用的音频特征包括音色、Chroma/PCP 等[260-261].对自相似矩阵进行基于阈值的 0/1 二值化(Binarization)即得到递归图(Recurrence Plot)[262],这种量化处理可以使音乐中的重复模式(Repetition Patterns)对速度、乐器、调高的改变具有更强的鲁棒性.基于相似性思路文献[263]提出另一个更复杂的方法.首先在节拍帧上检测和弦,用动态规划匹配和弦后得到和声相似的区域,从而划分音乐结构.与以上基于重复模式相似性的方法不同,另一种结构分析的典型思路是基于对音频特征(如音色)的子空间聚类(Subspace Clustering),并假设每个子空间对应于一个音乐段落[264-265].

2.6.5 音乐摘要/缩略图

音频摘要/缩略图(Music Summary/Thumbnail)是指找到音乐中可听的最具代表性的音频片段.但是何谓最具代表性并没有很好定义,可以有多个选择.音乐摘要/缩略图有多个应用,比如制作彩铃、浏览、

检索、购买数字音乐等[266].

获取音乐摘要/缩略图与音乐结构分析密切相关.一类方法只进行初步的结构分析,即基于音频特征计算各片段之间的相似性,再寻找最合适的片段集作为摘要/缩略图.使用的音频特征包括和声特征序列[267]及其直方图[268]、调性分析[269]等.另一类方法首先进行完整的音乐结构分析,之后从不同结构部分中提取摘要/缩略图.文献[270]使用副歌和它之前或之后的乐句合并组成摘要/缩略图,以保证摘要/缩略图开始和结束点位于有意义的乐句边界.文献[271]从流行音乐主歌和副歌中选择两个最具代表性的部分作为双摘要/缩略图.

2.6.6　音乐情感识别

音乐很容易和人产生情感共鸣,在不同的时间和环境下可能会需要带有不同感情色彩(如雄壮、欢快、轻松、悲伤、恐怖等)的音乐.音乐情感识别(Music Emotion Recognition, MER)在音乐选择与推荐、影视配乐、音乐理疗等场景都有重要应用,是近年来 MIR 领域的研究热点.

音乐情感识别最初被模拟为单标签[272]/多标签分类(Single/Multi-label Classification)问题[273].MER 需要建立符合人类认知特点的分类模型,经典的如 Hevner 情感模型和 Thayer 情感模型[274].为克服分类字典的模糊性,另一个思路是将音乐情感识别模拟为一个回归预测(Regression Prediction)问题[275].由 Arousal 和 Valence(AV)值构成二维 AV 情感空间,每个音乐信号成为情感平面上的一个点.心理学实验表明,Arousal 和 Valence 两个变量可以表达所有情绪的变化.Arousal 在心理学上可翻译为"活跃度",表示某种情绪含有能量的大小或活跃的程度,文献中表达类似含义的还有 Activation, Energy, Tension 等单词.Valence 在心理学上可翻译为"诱发力",表示感到舒适或愉悦的程度,文献中表达类似含义的还有 Pleasant, Good Mood, Positive Mood 等词汇[276].

目前,绝大多数 MER 算法都是基于音乐信号的低层声学特征如短时能量、谱特征等.虽然计算方便,但是与音乐情感没有直接关系,效果往往并不理想.有些文章利用歌词作为辅助信息对音乐情感进行分析[277].未来需要在音乐领域知识的指导下,研究音乐高层特征(如旋律走向、速度、强弱、调性、配器等)与音乐情感之间的关系,而且需要与音乐心理学(Music Psychology)进行更紧密的结合,引入更先进的情感模型.

2.7　其他 MIR 领域

MIR 还存在一些相关子领域,例如音乐推荐、音乐自动标注、歌声与歌词同步等.

2.7.1　音乐推荐

音乐推荐(Music Recommendation)通过分析用户历史行为,挖掘用户潜在兴趣,发现适合其喜好的音乐并进行推送.音乐推荐已在国内外多个音乐网站实现为产品,中文网站通常将此功能起名为"猜你喜欢".据少量调查,目前的音乐推荐产品用户体验不佳.客户需求高度个性化,如何根据不同的时间、地点、年龄、性别、民族、学历、爱好、经历、心情等因素进行精准个性化推荐仍是未解决的研究难题.

主流的推荐技术主要有 3 种:(1) 协同过滤推荐(Collaborative Filtering Recommendation),认为用户会倾向于欣赏同自己有相似偏好的用户群所聆听的音乐[278].换句话说,如果用户 A 和 B 有相似的音乐喜好,那么 B 喜欢但是还没有被 A 考虑的歌曲就将被推荐给 A.协同过滤推荐最主要的问题是不能给评分信息很少的新用户或新歌曲进行推荐,即冷启动(Cold-start)现象[279].(2) 基于内容的推荐(Content-based Recommendation),根据音乐间的元数据或声学特征的相似性推荐音乐[280].如果用户 A 喜欢歌曲 S,那么具有与 S 相似音乐特征的歌曲都将被推荐给 A.基于内容的推荐方法在一定程度上缓解了冷启动问题,更适用于新系统.(3) 混合型推荐(Hybrid Recommendation),除了传统的用户评价信息,还使用多模态数据如几何位置、用户场景、微博等社交媒体信息以及各种音乐标签如流派、情感、乐器和质量等[281].

除了从聆听历史中挖掘个人兴趣爱好以进行音乐推荐,现实生活中还需要其他种类的音乐推荐.例如,在缓解个人精神压力,为家庭录像选择最佳配乐、公共场合选择背景音乐时,需要基于情感计算进行推荐[282];在日常生活的不同场合(如工作、睡觉、运动)下通常需要不同种类的音乐.

2.7.2　音乐自动标注

近年来,互联网上出现了数以百万甚至千万计的数字音乐和音频,这也激发用户产生各种各样复杂

的音乐发现(Music Discovery)的需求.例如,在一个怀旧的夜晚检索“八十年代温柔男女对唱”,在结婚纪念日找“纪念结婚的乐曲”,或者“萨克斯伴奏的悠扬浪漫的女生独唱”[283].这种查询本身是复杂甚至模糊的,与之前具有确定查询形式的音乐识别、哼唱/歌唱检索、翻唱检索等具有本质区别.

给音乐和音频赋予描述性的关键字(Descriptive Keywords)或标签(Tags)是建立符合这样需求的检索引擎的一个可行办法.音乐标签属于社会标签(Social Tags)的一种,可由用户人工标注或通过学习音频内容与标签之间的关系进行自动标注.音乐标签除了检索特定要求的音乐,还有很多应用,如建立语义相似的歌唱播放列表(Playlist)、音效(Sound Effect)库管理、音乐推荐等[284].

用户人工标注通常采用有趣的游戏方式,如文献[285]中的“Herd It”.用机器自动标注通常使用机器学习方式.文献[286]采用梅尔尺度频谱图(Mel-spectrogram)作为进行自动标注的一个有效的时频表示,使用完全卷积网络(Fully Convolutional Neural Networks, FCNNs)进行基于内容的音乐自动标注.采用更多的层数和更多的训练数据会得到更好的结果.

鉴于待标注的标签内容本身无法确定[287](包括音乐情感、歌手、乐手、流派、乐器、语言、音色、声部、场景、风格、年代、唱片公司、歌词主题、流行度、民族、乐队、词曲作者等无法穷尽的描述),对于海量数据也很难采用人类专家之外的客观评价,目前该类方法还处于不是很有效的状态.

音乐标注(Music Annotation/Tagging/Labelling)的主要挑战是如何减少人类劳动并建立可靠的分类标签.一个经典的方法是利用主动学习进行自动标注,即选择少数最有信息量的样本进行人工标注并加入到训练集.对于二类分类问题,倾向于在每次迭代中选择单个的未标注样本.主动学习的问题是在每次标注样本后都要重新训练,很容易使用户失去耐心.文献[288]提出一个新的多类主动学习(Multi-class Active Learning)算法,在每个循环选择多个样本进行标注.实现中需要注意减少冗余,避免选择异常点,以使每个样本为模型的改进提供独特的信息.

2.7.3 歌声与歌词同步

在一个制作优良的电影电视、卡拉 OK 或音乐电视(Music TV, MTV)节目中,歌手演唱的声音、歌手的口型(Mouth Shapes)、屏幕显示的歌词这三者之间必须保持同步,否则将严重影响观众的欣赏质量.三者同步涉及视频、音频、文本之间的跨媒体(Cross-media)研究,目前在文献中尚未发现完整的工作,仅有少量研究集中于歌声与歌词之间的同步(Singing/Lyrics Synchronization).

文献[289-290]设计了一个叫做 LyricAlly 的系统,将歌唱的声音信号与对应的文本歌词自动进行时间配对对齐.音频处理部分结合低层音频特征和高层音乐知识来确定层次性的节奏结构和歌声部分.文本处理部分使用歌词来近似得到歌唱部分的长度.文献[291]改进了 LyricAlly 系统,把行级别(Line-level)的对齐改进到音节级别(Syllabic-level)的对齐,该方法同样使用动态规划,但是使用音乐知识来约束动态规划的路径搜索.

在语音信息处理技术中,使用 Viterbi 方法可以有效地对齐单声部/单音语音和相应的文本.但是,该方法却不能直接应用于 CD 录音进行音乐信号和相应歌词之间的自动对齐,因为歌声几乎都是和乐器伴奏混杂在一起.为解决该问题,文献[292]首先检测歌声部分并进行分离,之后对分离的歌声采用语音识别中的音素模型(Phoneme Model)识别发声单元,再与文本对齐.文献[293-294]采用了一系列措施对以上基于 Phoneme 的模型进行改进:包括将歌声元音(Singing Vowels)和歌词的音素网络(Phoneme Network)对齐;检测音频中摩擦辅音(Fricative Consonant)不存在的地方,阻止歌词中的摩擦音素(Fricative Phonemes)对齐到这些区域;忽略乐句之间(Inter-phrase)不属于歌词的元音发音;引入新的特征矢量进行歌声检测.文献[295]提出一个基于信号处理而不是模型的有趣方法.该方法首选使用文本到语音(Text-To-Speech,TTS)转换系统将歌词合成为语音,这样将音乐与文本歌词的自动对齐问题转化为两个音频信号之间的对齐问题,在词(Word)级别上进行.

2.8 与 MIR 相关的其他音乐科技领域

传统的 MIR 并不包括算法作曲、歌声合成、音视频融合、音频信息安全等内容.但我们考虑到 MIR 本身也是处于不断进化的过程,将音频音乐技术领域里其他十分重要的算法作曲、歌声合成、音视频融合,甚至音频信息安全等内容纳入到扩展的 MIR 范畴将会是未来的发展趋势.

2.8.1　自动/算法/AI 作曲

世界上的音乐，无论东方还是西方，均可以进行一定程度的形式化表示[296]．这为引入计算机技术参与创作提供了理论基础．自动作曲(Automated Composition)也称算法作曲(Algorithmic Composition)或人工智能作曲(AI Composition)，就是在音乐创作时部分或全部使用计算机技术，减轻人(或作曲家)的介入程度，用编程的方式来生成音乐．研究算法作曲一方面可以让我们了解和模拟作曲家在音乐创作中的思维方式，另一方面创作的音乐作品同样可以供人欣赏．

AI 和艺术领域差距巨大，尤其在中国被文理分割得更为厉害．两个领域的研究者说着不同甚至非常不同的语言，使用不同的方法，目标也各不相同，在合作和思想交换上产生巨大困难[297]．自动作曲研究中存在的主要问题有：音乐的知识表达问题，创造性和人机交互性问题，音乐创作风格问题，以及系统生成作品的质量评估问题．

自从 20 世纪 50 年代，AI 领域的不同技术已经被用来进行算法作曲．这些技术包括语法表示(Grammatical Representations)、概率方法(Probability Method)、人工神经网络、基于符号规则的系统(Symbolic Rule-based Systems)、约束规划(Constraint Programming)和进化算法(Evolutionary Algorithms)、马尔科夫链(Markov Chains)、随机过程(Random Process)、基于音乐规则的知识库系统(Music-rule based Knowledge System)等[298]．算法作曲系统将受益于多种方法融合的混合型系统(Hybrid System)，而且应在音乐创作的各个层面提供灵活的人机交互，以提高系统的实用性和有效性．

下边举一些有趣的例子．文献[299]提供一个交互式终端用户接口环境，可以实时对声音进行参数控制，从而使用进化计算(Evolutionary Computation)进行算法作曲，用遗传算法(Genetic Algorithms)来产生和评价 MIDI 演奏的一系列和弦．文献[300]从一段文本或诗歌出发，给每个句子分配一个表示高兴或悲伤的情绪(Mood)，使用基于马尔科夫链的算法作曲技术来产生具有感情的旋律线，然后采用某些歌声合成软件如 Vocaloid 进行输出．马尔科夫链的当前状态只与前一个状态有关，而旋律预测有较长的历史时间依赖性，因此该方法具有先天不足．如何设计一个既容易训练又能产生长期时域相关性的算法作曲模型成为一个大的挑战．文献[301]利用最新的深度学习技术，即深度递归神经网络(RNN)的加门递归单元(Gated Recurrent Unit，GRU)网络模型，在一个大的旋律数据集上训练，并自动产生新的符合训练旋律风格的旋律．GRU 尤其善于学习具有任意时间延迟的复杂时序关系的时间域序列，该模型能并行处理旋律和节奏，同时模拟它们之间的关系．该模型能产生有趣的完整的旋律，或预测一个符合当前旋律片段特性的可能的后续片段．

2.8.2　歌声合成

歌声本质上也是语音，所以歌声合成技术(Singing Voice Synthesis，SVS)的研究基本沿着语音合成(Speech Synthesis)的框架进行．语音合成的主要形式为文本到语音的转换．歌声则更加复杂，需要将文本形式的歌词按照乐谱有感情，有技巧的歌唱出来．因此，歌声合成的主要形式为歌词＋乐谱到歌声(Lyrics＋Score to Singing，LSTS)的转换．歌声和语音在发音机制、应用场景上有重大区别，歌声合成不仅需要语音合成的清晰性(Clarity)、自然性(Naturalness)、连续性等要求，而且要具备艺术性．

歌声合成涉及音乐声学、信号处理、语言学(Linguistics)、人工智能、音乐感知和认知(Music Perception and Cognition)、音乐信息检索、表演(Performance)等学科．在虚拟歌手、玩具、练唱软件、歌唱的模拟组合、音色转换、作词谱曲、唱片制作、个人娱乐、音乐机器人等领域都有很多的应用．

跟语音合成类似，歌声合成早期以共振峰参数合成法为主．共振峰(Formant)是声道(Vocal Tract)的传输特性即频率响应(Frequency Response)上的极点(Pole)，歌唱共振峰通常表现为在 3 kHz 左右的频谱包络线上的显著峰值．共振峰频率的分布决定语音/歌声的音色．以具有明确物理意义的共振峰频率及其带宽为参数，可以构成共振峰滤波器组，比较准确地模拟声道的传输特性．精心调整参数，能合成出自然度较高的语音/歌声．共振峰模型的缺点是，虽然能描述语音/歌声中最重要的元音，但不能表征其他影响自然度的细微成分．而且，共振峰模型的控制参数往往达到几十个，准确提取相当困难．因此共振峰参数合成法整体合成的音质达到实用要求还有距离[302]．由上可知，高质量的估计声道过滤器(Vocal Tract Filter，VTF)即谱包络线的共振态/反共振态(Resonances/Anti-resonances)对歌声合成非常有益．已有算

法经常使用基于单帧分析(Single-Frame Analysis, SFA)的离散傅里叶变换(Discrete Fourier Transform, DFT)来计算谱包络线.文献[303]将多帧分析(Multiple-Frame Analysis, MFA)应用于音乐信号的 VTF 构型的估计.一个具有表现力(Expressive)的歌手,在歌唱过程中利用各种技巧来修改其歌声频谱包络线.为得到更好的表现力和自然度,文献[304]研究共振峰偏移(Formant Excursion)问题,用元音的语义依赖约束共振峰的偏移范围.

与上述基于对发声过程建模的方法不同,采样合成/波形拼接合成(Sampling Synthesis/Concatenated-based Singing Voice Synthesis)技术从歌声语料库(Singing Corpus)中按照歌词挑选合适的录音采样,根据乐谱及下文要求对歌声的音高、时长进行调整,并进行颤音、演唱风格、情感等艺术处理后加以拼接.该方法使得合成歌声的清晰度和自然度大大提高,但需要大量的时间和精力来准备歌声语料库,而且占用空间很大.这类方法也称为基音同步叠加(Pitch Synchronous Overlap Add, PSOLA),以下列举几个例子.

基于西班牙巴塞罗那 UPF 大学 MTG 与日本雅马哈(Yamaha)公司联合研制的 Vocaloid 歌声合成引擎,第三方公司出品了风靡世界的虚拟歌唱软件——初音未来.该系统[305]事先将真人声优的歌声录制成包含各种元音、辅音片段的歌声语料库,包括目标语言音素所有可能的组合,数量大概是 2 000 个样本/每音高.用户编辑输入歌词和旋律音高后,合成引擎按照歌词从歌声语料库中挑选合适的采样片段,根据乐谱采用频谱伸缩方法(Spectrum Scaling)将样本音高转换到旋律音高,并在各拼接样本之间进行音色平滑.样本时间调整自动进行,以使一个歌词音节的元音 Onset 严格与音符 Onset 位置对齐.文献[306]采用类似思路,在细节上稍有不同,而是采用重采样的方法进行样本到旋律的音高转换,基于基音周期的检测算法扩展音长.

与早期主要基于信号处理的方法不同,后期的歌声合成算法大量使用机器学习技术.基于上下文相关 HMM(Context-dependent HMM)的歌声合成技术一度成为主流,用其联合模拟歌声的频谱、颤音、时长等[307].近年来,随着深度学习的流行,更适合刻画复杂映射关系的 DNN 技术被引入到歌声合成中.文献[308]用 DNN 逐帧模拟乐谱上下文特征(Contextual Features)和其对应声学特征之间的关系,得到比 HMM 更好的合成效果.文献[309]采用了另一种方法,没有用 DNN 直接模拟歌声频谱等,而是以歌词、音高、时长等为输入端,用 HMM 合成歌声和自然歌声的声学特征的区别为输出端,在它们之间用 DNN 模拟复杂的映射关系.歌唱是一种艺术,除了保持最基本的音高和节奏准确,还有很多艺术技巧如 Vibrato、滑音等.这些技巧表现为歌声基频包络线(f_0 Contour)的波动,充分反映了歌手的歌唱风格.文献[310]使用深度学习中的 LSTM-RNN 模型来模拟复杂的音乐时间序列,自动产生 f_0 序列.以乐谱中给定的音乐上下文和真实的歌声配对组成训练数据集,训练两个 RNN,根据音乐上下文分别学习 f_0 的音高和 Vibrato 部分,并捕捉人类歌手的表现力(Expressiveness)和自然性.对于一个乐谱来说,可能有很多风格迥异的歌唱版本,目前的歌声合成算法只能集中于模拟特定的歌唱风格.文献[311]首先从带标注的实际录音中提取 f_0 参数和音素长度,结合丰富的上下文信息构建一个参数化模板数据库.然后,根据目标上下文选择合适的参数化模板,进行具有某种歌唱风格的歌声合成.

除了以上歌词+乐谱到歌声的转换,还有一类语音+乐谱到歌声的转换,即语音到歌声转换(Speech-To-Singing Conversion, STSC).这是在两个音频信号之间的转换,避免了以前声道特性估计不准或需要预先录制大规模歌唱语料库的困难,开辟了一条新的思路.文献[312]提出了一个简单的语音到歌声转换算法.首先分割语音信号,得到一系列语音基本单元.之后确定每个基本单元和对应音符之间的同步映射,并根据音符的音高对该单元的基频进行调整.最后根据对应音符的时长调整当前语音基本单元的长度.文献[313]采用类似思路,但是对用户输入进行了一定约束.输入的语音信号是用户依据歌曲的某段旋律(如一个乐句),按照节拍诵读或哼唱歌词产生的.因此每段语音信号可以更准确地在时间上对应于该旋律片段.后续处理包括按旋律线调整语音信号的音高,按音符时长进行语音单元时间伸缩,平滑处理音高包络线,加入颤音、滑音、回声等各种艺术处理.文献[314]设计了一个新的歌声合成系统"SingBySpeaking",输入信息为读歌词的语音信号和乐谱,并假设已经对齐.为构造听觉自然的歌声,该系统具有 3 个控制模块:基频控制模块,按照乐谱将语音信号的 f_0 序列调整为歌声的 f_0 包络线,同时调整颤音等影响歌声自然度的 f_0 波动;谱序列控制模块,修改歌声共振峰并调制共振峰的幅度,将语音的频谱形状(Spectral

Shapes)转换为歌声的频谱形状；时长控制模块，根据音符长度将语音音素的长度伸缩到歌声的音素长度.频谱特征可以直接反映音色特性，f_0 包络线、音符时长以及强弱(Dynamics)等组成韵律特征(Prosodic Features)反映时域特性.为得到高自然度的歌声合成，文献[315]使用适于模拟高维特征的 DNN 对这些特征进行从语音到歌声的联合模拟转换.

2.8.3　听觉与视觉的结合

人类接收信息的方式主要来源于视觉和听觉，现代电影和电视节目、多媒体作品几乎都是声音、音乐、语音和图像、视频的统一.绝大多数视频里都存在声音信息，很多的音乐节目如音乐电视里也存在视频信息.音视频密不可分，互相补充，进行基于信息融合的跨媒体研究对很多应用场景都是十分必要的.下边列举一些音视频结合研究的例子.

音乐可视化(Music Visualization)是指为音乐生成一个能反映其内容(如旋律、节奏、强弱、情感等)的图像或动画的技术，从而使听众得到更加生动有趣的艺术感受.早期的音乐播放器基于速度或强度变化进行简单的音乐可视化，在速度快或有打击乐器的地方，条形图或火焰等图形形状会跳得更快或更高.Herman 等提出的音乐可视化理论[316]假定音高和颜色之间具有一定的关系，基于此理论使用光栅图形学(Raster Graphics)来产生音符、和弦及和弦连接的图形显示.音符或和弦的时域相邻性被映射为颜色的空间临近性，经常显示为按中心分布的方块或圆圈.电影是人类历史上最重要的娱乐方式之一，是一种典型的具有艺术性的音视频相结合的媒体.相比于早期的无声电影，在现代电影中声音和音乐对于情节的铺垫、观众情绪的感染、整体艺术水平的升华起到了无可替代的作用.文献[317]基于视频速度(Video Tempo)和音乐情感(Music Mood)进行电影情感事件检测，并对声音轨迹的进程进行了可视化研究.文献[318]结合 MIR 中的节拍检测和计算机视觉技术，融合音视频输入信息对机器人音乐家和它的人类对应者的动作进行同步.文献[319]分别计算图片和音乐表达的情感，对情感表达相近的图片和音乐进行匹配，从而自动生成基于情感的家庭音乐相册(如婚礼的图片搭配浪漫的背景音乐).文献[320]将音视频信息结合进行运动视频的语义事件检测(Semantic Event Detection)，以方便访问和浏览.该文定义了一系列与运动员(Players)、裁判员(Referees)、评论员(Commentators)和观众(Audience)高度相关的音频关键字(Audio Keywords).这些音频关键字视为中层特征，可以从低层音频特征中用 SVM 学习出来.与视频镜头相结合，可有效地用 HMM 进行运动视频的语义事件检测.此外，还有电影配乐、MTV 中口型与歌声和歌词同步等有趣的应用.

2.8.4　音频信息安全

音频信息安全(Audio Information Security)主要包括音频版权保护(Audio Copyright Protection)和音频认证(Audio Authentication)两个子领域.核心技术手段为数字音频水印(Digital Audio Watermarking)和数字音频指纹.音频水印是一种在不影响原始音频质量的条件下向其中嵌入具有特定意义且易于提取的信息的技术.音频指纹技术的细节可参照 2.5.1 节，本部分主要介绍数字音频水印技术.

2.8.4.1　音频版权保护

数字音频作品(通常指音乐)的版权保护主要采用鲁棒数字音频水印(Robust Audio Watermarking)技术.除了版权保护，鲁棒音频水印还可用于广播监控(Broadcast Monitoring)、盗版追踪(Piracy Tracing)、拷贝控制(Copy Control)、内容标注(Content Labeling)等.它要求嵌入的水印能够经受各种时频域的音频信号失真[321].鲁棒数字音频水印技术按照作用域可分为时间域和频率域算法两类.时域算法鲁棒性一般较差.频域算法充分利用人类听觉特性，主流思路是在听觉重要的中低频带上嵌入水印，从而获得对常规信号失真的鲁棒性.

早期的鲁棒音频水印算法主要集中于获得嵌入水印的不可听性(Inaudibility)或称感知透明性(Perceptual Transparency)，和对常规音频信号处理失真(如压缩、噪声、滤波、回声等)的鲁棒性.如文献[322]把音频切成小片段，直接修改音频样本进行水印嵌入.水印按照音频内容被感知塑形(Perceptually-shaped)，利用时域和频域感知掩蔽(Temporal and Frequency Perceptual Masking)来保证不可听性和鲁棒性.文献[323]将通信系统中借鉴来的直接序列扩频(Direct Sequence Spread Spectrum，DSSS)思想成功应用于数字水印技术中.将一个数字水印序列与高速伪随机码相乘后叠加到原始音频信号上，并利用人

类听觉系统(Human Auditory System, HAS)的掩蔽效应(Masking Effect)进一步整形水印信号以保证其不可听到.为在感知质量(Perceptual Quality)、鲁棒性(Robustness)、水印负载(Watermark Payload)等相互冲突的因素中间达到平衡,文献[324]根据音频信号的内容进行自适应的水印嵌入,将一些低层音频特征用 PCA(Principal Component Analysis)提取主成分后,使用数学模型来评价在感知透明性约束下的水印嵌入度.水印自适应地嵌入到小波域的第三层细节系数(Detailed Coefficients).

常规的音频信号失真主要通过降低音频质量来消除水印,这个问题很快被解决.后来的挑战主要集中于抵抗时频域的同步失真(Synchronization Distortions).这种失真通过对时频分量的剪切、插入等操作,使水印检测器(Watermark Detector)找不到水印的嵌入位置,从而使检测失败.抵抗同步失真主要有穷举搜索(Exhaustive Search)、同步码(Synchronization Code)、恒定水印(Invariant Watermark)和隐含同步(Implicit Synchronization)等 4 种方法[325].后两种方法因为明确地利用了音频内容分析,与之前将水印嵌入到时间域样本或频率域变换系数的算法不同,被称为第二代数字水印(Second-generation Digital Watermarking)技术[326].文献[327]基于恒定水印的思想,提出一种第二代数字音频水印算法,即通过调整音频信号每个帧的小波域系数平均值的符号来嵌入水印数据,从而使水印检测器对同步结构的变化不敏感.文献[328]基于隐含同步的思想,提出基于音乐内容分析的局部化数字音频水印算法.通过 3 种不同方法提取出代表音乐边缘(Music Edges)的局部区域,利用其感知重要性和局部性获得对信号失真和同步失真的免疫力(Immunity),然后通过交换系数法在其中嵌入水印.类似地,文献[329]基于音频内容分析和傅里叶变换,在时频域能量峰值点周围的 ROI(Region of Interest)区域进行水印嵌入,以抵抗音频编辑和恶意随机剪切(Malicious Random Cropping)引起的同步失真.

除了音频水印,2.5.1 节所述的音频指纹技术也可以用于版权保护.因其不需要往信号里加入额外信息,也称为被动水印(Passive Watermarking)技术.此外还有一些别名,如音频鲁棒感知哈希(Audio Robust Perceptual Hashing)、基于内容的数字签名(Content-based Digital Signatures)、基于内容的音频识别(Content-based Audio Identification)等.

2.8.4.2 音频认证

音频伪造(Audio Forgery)在当今的数字音频时代已经变得极其容易.对于重要录音信息(比如电话交谈的金融信息、领导人讲话、军事指令、时间地点等)进行恶意篡改(Malicious Tampering),或插入虚假信息,删除关键片段,制造虚假质量(Fake Quality)的音频等都会给政治、经济、军事、法律、商业等各个领域带来极大的影响.

脆弱及半脆弱数字音频水印(Fragile/Semi-fragile Audio Watermarking)主要用于数字音频作品的真实性(Authenticity)和完整性(Integrity)保护.脆弱水印在宿主数据发生任何变化时都会无法检测到,类似于密码学里的哈希值,典型的例子如 LSB(Least Significant Bit)方法[330].半脆弱水印则融合了鲁棒水印与脆弱水印的特性,在能够抵抗有损压缩、噪声、滤波、重采样等可允许操作(Acceptable/Admissible Operations)的同时,对剪切、插入、替换等恶意操作(Malicious Operations)敏感[331].水印需要逐段嵌入,以便于在发生恶意操作时进行定位.基于水印的音频认证需要嵌入水印信息,也称为主动音频认证(Active Audio Authentication).

但是,在现实应用场景中,给所有的音频内容都预先嵌入水印是不可能的.因此,利用被动音频认证(Passive Audio Authentication),也称为音频取证(Audio Forensics),具有更大的应用前景.音频取证的基本方式包含听觉测试(Listening Test)、频谱图分析(Spectrogram Analysis)和频谱分析(Spectrum Analysis)等.高级方式利用音调(Tones)、相位、ENF(Electric Network Frequency)、LPCs、MFCCs、MDCT 等各种音频特征及机器学习方法进行判断[332].文献[333]基于音频特征及朴素贝叶斯(Naïve Bayes)分类器确定数字音频使用的麦克风和录音环境.文献[334]在 MP3 格式的音频中发现,从低比特率(Low Bit-rate)转码得到的虚假高比特率(High Bit-rate)的 MP3 比正常 MP3 有更少的小数值 MDCT 系数.因此,小数值 MDCT 系数的个数可作为一个有效的特征来区分虚假质量 MP3 和正常质量 MP3.

由于音频信号可能很长,有时需要判断其中的某一段是否被恶意篡改过,称为片段认证(Fragment Authentication)问题.文献[335]第一次提出一个解决方法,在音频频谱图上计算来自计算机视觉的 SIFT

描述子，利用其强大的局部对齐能力将待认证片段对齐到原始音频的相应位置，通过音频指纹比对检测可允许操作和恶意操作，精确进行篡改定位及分类.该方法需要保留原始音频，具有较大的局限性.

3　总结与展望

基于内容的音乐信息检索以数字音乐为研究对象，覆盖几乎一切与数字音乐内容分析理解相关的研究课题，是多媒体、信号处理、人工智能、音乐学相结合的重要学科分支.相似的技术框架扩展到一般音频后，统称为计算机听觉，也可称为音频与音乐计算.从学科角度讲，计算机听觉与语音信息处理最为相关，而且都以物理声学为基础.

本文介绍了音乐信息检索技术的发展历史、学科架构，将几十个研究课题按照与各音乐要素的紧密程度分类归入核心层与应用层.对每个研究课题，概述其研究目的和应用场景，总结主要的技术框架及典型算法.

与自然语言处理、计算机视觉、语音信息处理等相关领域相比，计算机听觉在国内外发展都比较缓慢.几个可能的原因包括：(1) 数字音乐涉及版权问题无法公开，各种音频数据都源自特定场合和物体，难以搜集和标注.近 20 年来，计算机听觉跟其他学科一样，绝大多数方法都是基于机器学习框架.数据的获取及公开困难严重影响了算法的研究及比较.(2) 音乐和音频信号几乎都是多种声音混合在一起，很少有单独存在的情况.音乐中的各种乐器和歌声在音高上形成和声，在时间上形成节奏，耦合成多层次的复杂音频流，难以甚至无法分离处理.(3) 计算机听觉几乎都是交叉学科，进行音乐信息检索研究需要了解最基本的音乐理论知识，进行音频信息处理则需要了解相关各领域的专业知识和经验.(4) 此外，作为新兴学科，还存在社会发展水平、科研环境、科技评价、人员储备等各种非技术类原因阻碍计算机听觉技术的发展.

MIR 在娱乐、音乐教学、心理疏导、医学辅助治疗、公共及家庭环境监控、目标检测识别、智能交通、设备故障检测等方面具有很多应用，而且理论上所有视频应用都需要和音频结合，是一门非常实用具有广阔前景的技术.

致谢：感谢 4 位匿名审稿人和北京大学陈晓鸥教授提出的宝贵修改意见！

参考文献：

[1] CSMT会议组织委员会.2016 CSMT 会议论文集序言[J].复旦学报(自然科学版)，2017，**56**(2)：135.

[2] CAMURRI A，DE POLI G，ROCCHESSO D. A taxonomy for sound and music computing[J]. *Computer Music Journal*，1995，**19**(2)：4-5.

[3] DUBNOV S. Computer audition：An introduction and research survey[C]//ACM International Conference on Multimedia(ACM MM). California，USA：ACM，2006：9-9.

[4] GERHARD D. Computer music analysis[R]. Surrey，UK：Simon Fraser University，1997.

[5] RABINER L. On the use of autocorrelation analysis for pitch detection[J]. *IEEE Transactions on Acoustics，Speech，and Signal Processing*，1977，**25**(1)：24-33.

[6] CHEVEIGNÉ A D，KAWAHARA H. YIN，a fundamental frequency estimator for speech and music[J]. *Journal of the Acoustical Society of America*，2002，**111**(4)：1917-1930.

[7] RODET X，DOCVAL B. Maximum-likelihood harmonic matching for fundamental frequency estimation[J]. *Journal of the Acoustical Society of America*，1992，**92**(4)：2428-2429.

[8] MARKEL J. The SIFT algorithm for fundamental frequency estimation[J]. *IEEE Transactions on Audio and Electroacoustics*，1972，**20**(5)：367-377.

[9] MEDAN Y，YAIR E，CHAZAN D. Super resolution pitch determination of speech signals[J]. *IEEE Transactions on Signal Processing*，1991，**39**(1)：40-48.

[10] MCAULAY R J，QUATIERI T F. Pitch estimation and voicing detection based on a sinusoidal speech model[C]//IEEE International Conference on Acoustics，Speech and Signal Processing(IEEE ICASSP). New Maxico，USA：IEEE，1990：249-252.

[11] RABINER L, CHENG M, ROSENBERG A, et al. A comparative performance study of several pitch detection algorithms[J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976, **24**(5): 399-418.

[12] KADAMBE S, BOUDREAUX-BARTELS G F. Application of the wavelet transform for pitch detection of speech signals[J]. *IEEE Transactions on Information Theory*, 1992, **38**(2): 917-924.

[13] FAN Z C, JANG J S R, LU C L. Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking[C]//IEEE Second International Conference on Multimedia Big Data(IEEE BigMM). Taipei, China Taiwan: BigMM, 2016: 178-185.

[14] HSU C L, WANG D L, JANG J S R. A trend estimation algorithm for singing pitch detection in musical recordings[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic: ICASSP, 2011: 393-396.

[15] DZIUBINSKI M, KOSTEK B. Octave error immune and instantaneous pitch detection algorithm[J]. *Journal of New Music Research (JNMR)*, 2005, **34**(3): 273-292.

[16] 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]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Kyoto, Japan: ICASSP, 2012: 457-460.

[17] HSU C L, CHEN L Y, JANG J S R, et al. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement[C]//International Society for Music Information Retrieval Conference(ISMIR). Kobe, Japan: ISMIR, 2009: 201-206.

[18] GOTO M. f_0 estimation of melody and bass lines in real-world musical audio signals[J]. *Information Processing Society of Japan SIG Notes*, 1999(68): 91-98.

[19] PAIVA R P, MENDES T, CARDOSO A. On the detection of melody notes in polyphonic audio[C]// International Society for Music Information Retrieval Conference(ISMIR). London, UK: ISMIR, 2005: 175-182.

[20] **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.

[21] TACHIBANA H, ONO T, ONO N, et al. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Dallas, USA: ICASSP, 2010: 425-428.

[22] HAN J, CHEN C W. Improving melody extraction using probabilistic latent component analysis[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Prague, Czech Republic: ICASSP, 2011: 33-36.

[23] POLINER G E, ELLIS D P W. A classification approach to melody transcription[C]//International Society for Music Information Retrieval Conference(ISMIR). London, UK: ISMIR, 2005: 161-166.

[24] SALAMON J, GOMEZ E, ELLIS D P W, et al. Melody extraction from polyphonic music signals: Approaches, applications, and challenges[J]. *IEEE Signal Processing Magazine*, 2014, **31**(2): 118-134.

[25] KARJALAINEN M, TOLONEN T. Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Phoenix, USA: ICASSP, 1999, **2**: 929-932.

[26] MIRYALA S S, BALI K, BHAGWAN R, et al. Automatically identifying vocal expressions for music transcription[C]//International Society for Music Information Retrieval Conference(ISMIR). Curitiba, Brazil: ISMIR, 2013: 239-244.

[27] VINCENT E, BERTIN N, BADEAU R. Adaptive harmonic spectral decomposition for multiple pitch estimation [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, **18**(3): 528-537.

[28] ARGENTI F, NESI P, PANTALEO G. Automatic transcription of polyphonic music based on the Constant-Q bispectral analysis[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(6): 1610-1630.

[29] ABDALLAH S A, PLUMBLER M D. An independent component analysis approach to automatic music transcription[C]//Audio Engineering Society(AES) Convention. Amsterdam, The Netherlands: AES, 2003: 1-7.

[30] SMARAGDIS P, BROWN J C. Non-negative matrix factorization for polyphonic music transcription[C]// IEEE Workshop on Applications of Signal Processing to Audio and Acoustics(WASPAA). New York, USA: WASPAA, 2003: 177-180.

[31] BENETOS E, DIXON S. A shift-invariant latent variable model for automatic music transcription[J]. *Computer Music Journal*, 2012, **36**(4): 81-94.

[32] KLAPURI A P. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness[J]. *IEEE Transactions on Speech and Audio Processing*, 2003, **11**(6): 804-816.

[33] BENETOS E, DIXON S. Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2011, **5**(6): 1111-1123.

[34] PERTUSA A, INESTA J M. Multiple fundamental frequency estimation using Gaussian smoothness[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 105-108.

[35] KATAYOSE H, INOKUCHI S. Intelligent music transcription system[J]. *Journal of Japanese Society for Artificial Intelligence*, 1990, **5**(1): 59-66.

[36] KIRCHHOFF H, DIXON S, KLAPURI A. Shift-variant non-negative matrix deconvolution for music transcription[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Kyoto, Japan: ICASSP, 2012: 125-128.

[37] BENETOS E, DIXON S, GIANNOULIS D, et al. Automatic music transcription: Challenges and future directions[J]. *Journal of Intelligent Information Systems*, 2013, **41**(3): 407-434.

[38] **陈桂华.严格区分节奏、节奏型、节拍、拍子的意义**[J].**乐府新声**, 1990, **8**(2): 50-53.

[39] BELLO J P, DAUDET L, ABDALLAH S, et al. A tutorial on onset detection in music signals[J]. *IEEE Transactions on Speech and Audio Processing*, 2005, **13**(5): 1035-1047.

[40] **桂文明.音符起始点检测算法研究**[D].**南京：南京理工大学**, 2013.

[41] DUXBURY C, SANDLER M, DAVIES M. A hybrid approach to musical note onset detection[C]// International Conference on Digital Audio Effects(DAFx). Hamburg, Germany: DAFx, 2002: 33-38.

[42] BELLO J P, ANDLER M. Phase-based note onset detection for music signals[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Hong Kong, China: ICASSP, 2003: 441-444.

[43] WANG W W, LUO Y H, CHAMBERS J A, et al. Note onset detection via nonnegative factorization of magnitude spectrum[J]. *EURASIP Journal on Advances in Signal Processing*, 2008, **2008**: 1-15.

[44] SHAO X, GUI W, XU C. Note onset detection based on sparse decomposition[J]. *Multimedia Tools and Applications*, 2016, **75**(5): 2613-2631.

[45] MAROLT M, KAVCIC A, PRIVOSNIK M. Neural networks for note onset detection in piano music [EB/OL].[2017-07-09]. https://www.researchgate.net/publication/2473938_Neural_Networks_for_Note_Onset_Detection_in_Piano_Music.

[46] LACOSTE A, ECK D. A supervised classification algorithm for note onset detection[J]. *EURASIP Journal on Advances in Signal Processing*, 2007, **2007**: 1-13.

[47] YU Y, ZHOU Y, WANG Y. A tempo-sensitive music search engine with multimodal inputs[C]// International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies(MIRUM). Scottsdale, USA: ACM, 2011: 13-18.

[48] ALONSO M, RICHARD G, DAVID B. Accurate tempo estimation based on harmonic+noise decomposition[J]. *EURASIP Journal on Advances in Signal Processing*, 2007, **2007**: 1-14.

[49] GAINZA M, COYLE E. Tempo detection using a hybrid multiband approach[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(1): 57-68.

[50] GÄRTNER D. Tempo detection of urban music using tatum grid non-negative matrix factorization[C]// International Society for Music Information Retrieval Conference(ISMIR). Curitiba, Brazil: ISMIR, 2013: 311-316.

[51] CHORDIA P, RAE A. Using source separation to improve tempo detection[C]//International Society for Music Information Retrieval Conference(ISMIR). Kobe, Japan: ISMIR, 2009: 183-188.

[52] GAINZA M. On the use of a dynamic hybrid tempo detection model for beat tracking[C]//IEEE International Conference on Multimedia and Expo(ICME). Suntec, Singapore: ICME, 2010: 552-557.

[53] EYBEN F, BÖCK S, SCHULLER B, et al. Universal onset detection with bidirectional long-short term memory neural networks[C]//International Society for Music Information Retrieval Conference(ISMIR). Utrecht, The Netherlands: ISMIR, 2010: 589-594.

[54] GROSCHE P, MULLER M. Computing predominant local periodicity information in music recordings [C]//IEEE Workshop on Applications of Signal Processing to Audio and Acoustics(WASPAA). New York, USA: WASPAA, 2009: 33-36.

[55] CEMGIL A T, KAPPEN B. Monte Carlo methods for tempo tracking and rhythm quantization[J]. *Journal of Artificial Intelligence Research*, 2003, **18**(1): 45-81.

[56] PEETERS G. Time variable tempo detection and beat marking[C]//International Computer Music Conference(ICMC), Barcelona, Spain: ICMC, 2005: 1-4.

[57] HOLLOSI D, BISWAS A. Complexity scalable perceptual tempo estimation from HE-AAC encoded music[C]//Audio Engineering Society(AES) Convention. London, UK: AES, 2010: 8109.

[58] STARK A M, DAVIES M E P, PLUMBLEY M D. Real-time beat-synchronous analysis of musical audio[C]//International Conference on Digital Audio Effects(DAFx). Como, Italy: DAFx, 2009: 299-304.

[59] GOTO M, MURAOKA Y. A beat tracking system for acoustic signals of music[C]//ACM International Conference on Multimedia(ACM MM). San Francisco, USA: ACM, 1994: 365-372.

[60] LAROCHE J. Offline and online tempo detection and beat tracking[J]. *Journal of the Acoustical Society of America*, 2002, **111**(5): 2417-2417.

[61] SCHEIRER E D. Tempo and beat analysis of acoustic musical signals[J]. *Journal of the Acoustical Society of America*, 1998, **103**(1): 588-601.

[62] LAROCHE J. Estimating tempo, swing and beat locations in audio recordings[C]//IEEE Workshop on Applications of Signal Processing to Audio and Acoustics(WASPAA). New York, USA: WASPAA, 2001: 135-138.

[63] GOTO M. An audio-based real-time beat tracking system for music with or without drum-sounds[J]. *Journal of New Music Research*, 2001, **30**(2): 159-171.

[64] GOTO M, MURAOKA Y. Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions[J]. *Speech Communication*, 1999, **27**(3): 311-335.

[65] ZAPATA J R, HOLZAPFEL A, DAVIES M E P, et al. Assigning a confidence threshold on automatic beat annotation in large database[C]//International Society for Music Information Retrieval Conference (ISMIR). Porto, Portugal: ISMIR, 2012: 157-162.

[66] ELLIS D P W. Beat tracking by dynamic programming[J]. *Journal of New Music Research*, 2007, **36**(1): 51-60.

[67] FILLON T, JODER C, DURAND S, et al. A conditional random field system for beat tracking[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brisbane, Australia: ICASSP, 2015: 424-428.

[68] DAVIES M E P, DEGARA N, PLUMBLEY M D. Evaluation methods for musical audio beat tracking algorithms[R]. London, UK: Queen Mary University of London, 2009.

[69] GAINZA M. Automatic musical meter detection[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Taipei, China Taiwan: ICASSP, 2009: 329-332.

[70] GAINZA M, BARRY D, COYLE E. Automatic bar line segmentation [C]//Audio Engineering Society (AES) Convention. New York, USA: AES, 2007: 1-7.

[71] QUINTON E, O'HANLON K, DIXON S, et al. Tracking metrical structure changes with sparse-NMF [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: ICASSP, 2017: 41-45.

[72] HAAS W B, VOLK A. Meter detection in symbolic music using inner metric analysis [C]//International Society for Music Information Retrieval Conference(ISMIR). New York, USA: ISMIR, 2016: 441-447.

[73] DURAND S, DAVID B, RICHARD G. Enhancing downbeat detection when facing different music styles [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: ICASSP, 2014: 3156-3160.

[74] HOCKMAN J A, DAVIES M E P, FUJINAGA I. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass [C]//International Society for Music Information Retrieval Conference (ISMIR). Porto, Portugal: ISMIR, 2012: 169-174.

[75] DURAND S, BELLO J P, DAVID B. Downbeat tracking with multiple features and deep neural networks [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brisbane, Australia: ICASSP, 2015: 409-413.

[76] DURAND S, ESSID S. Downbeat detection with conditional random fields and deep learned features [C]// International Society for Music Information Retrieval Conference(ISMIR). New York, USA: ISMIR, 2016: 386-392.

[77] KREBS F, BÖCK S, DORFER M, et al. Downbeat tracking using beat-syncronous features and recurrent neural networks [C]//International Society for Music Information Retrieval Conference(ISMIR). New York, USA: ISMIR, 2016: 129-135.

[78] BLOSTEIN D, HAKEN L. Template matching for rhythmic analysis of music keyboard input [C]// International Conference on Pattern Recognition(ICPR). Atlantic, USA: ICPR, 1990, 1: 767-770.

[79] MAGUIRE R. Real-time rhythmic pattern detection from audio via template matching[EB/OL]. [2017-11-10]. http://ryanmaguiremusic.com/media_files/pdf/RealTimeRhythmicPatternDetection.pdf.

[80] GRUHNE M, DITTMAR C, GAERTNER D, et al. An evaluation of pre-processing algorithms for rhythmic pattern analysis [C]//Audio Engineering Society (AES) Convention. San Francisco, USA: AES, 2008: 7542.

[81] GRUHNE M, DITTMAR C. Improving rhythmic pattern features based on logarithmic preprocessing [C]// Audio Engineering Society(AES) Convention. Munich, Germany: AES, 2009: 7817.

[82] COCA A E, ZHAO L. Musical rhythmic pattern extraction using relevance of communities in networks [J]. *Information Sciences*, 2016, **329**: 819-848.

[83] KREBS F, BÖCK S, WIDMER G. Rhythmic pattern modeling for beat and downbeat tracking in musical audio [C]. International Society for Music Information Retrieval Conference(ISMIR). Curitiba, Brazil: ISMIR, 2013: 227-232.

[84] **孙云鹰.主调音乐与复调音乐**[J].**音乐学习与研究**, 1985, **2**: 46-47.

[85] LEE H R, JANG J S R. I-Ring: A system for humming transcription and chord generation [C]//IEEE International Conference on Multimedia and Expo (ICME). Taipei, China Taiwan: ICME, 2004: 1031-1034.

[86] JIANG N, GROSCHE P, KONZ V, et al. Analyzing Chroma feature types for automated chord recognition [C]//International Conference: Semantic Audio. Ilmenau, Germany: Audio Engineering Society, 2011: 1-10.

[87] VAREWYCK M, PAUWELS J, MARTENS J P. A novel Chroma representation of polyphonic music based on multiple pitch tracking techniques [C]//ACM International Conference on Multimedia (ACM MM). Vancouver, Canada: ACM, 2008: 667-670.

[88] ZHOU X, LERCH A. Chord detection using deep learning [C]//International Society for Music Information Retrieval Conference(ISMIR). Malaga, Spain: ISMIR, 2015: 52-58.

[89] ZENZ V, RAUBER A. Automatic chord detection incorporating beat and key detection [C]//IEEE International Conference on Signal Processing and Communications (ICSPC). Dubai, United Arab Emirates: ICSPC, 2007: 1175-1178.

[90] MADDAGE N C, KANKANHALLI M S, LI H. Effectiveness of signal segmentation for music content representation [C]//International Multimedia Modeling Conference (MMM). Kyoto, Japan: MMM, 2008: 477-486.

[91] HARTE C, SANDLER M. Automatic chord recognition using quantised Chroma and harmonic change segmentation [C]//Music Information Retrieval Evaluation eXchange (MIREX). Kobe, Japan: MIREX, 2009: 1-3.

[92] UEDA Y, UCHIYAMA Y, NISHIMOTO T, et al. HMM-based approach for automatic chord detection using refined acoustic features [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, USA: ICASSP, 2010: 5518-5521.

[93] MAZHAR F. Automatic guitar chord detection [D]. Tampere, Finland: Tampere University of Technology, 2012.

[94] SIGTIA S, BOULANGER-LEWANDOWSKI N, DIXON S. Audio chord recognition with a hybrid recurrent neural network [C]//International Society for Music Information Retrieval Conference (ISMIR). Malaga, Spain: ISMIR, 2015: 127-133.

[95] MAUCH M, DIXON S. Simultaneous estimation of chords and musical context from audio [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, **18**(6): 1280-1289.

[96] ROLLAND J B. Chord detection using Chromagram optimized by extracting additional features [C]// Music Information Retrieval Evaluation eXchange (MIREX). Taipei, China Taiwan: MIREX, 2014: 27-31.

[97] KHADKEVICH M, OMOLOGO M. Improved automatic chord recognition [C]//Music Information Retrieval Evaluation eXchange (MIREX). Kobe, Japan: MIREX, 2009: 1-3.

[98] MAUCH M, DIXON S. Approximate note transcription for the improved identification of difficult chords [C]// International Society for Music Information Retrieval Conference (ISMIR), 2010: 135-140.

[99] DENG J, KWOK Y K. A hybrid Gaussian-HMM-Deep-Learning approach for automatic chord estimation with very large vocabulary [C]//International Society for Music Information Retrieval Conference (ISMIR). New York, USA: ISMIR, 2016: 812-818.

[100] JIANG J, LI W, WU Y. Chord recognition using random forest model [C]//Music Information Retrieval Evaluation eXchange (MIREX). Suzhou, China: MIREX, 2017: 1-4.

[101] WU Y, LI W. Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model [C]//Music Information Retrieval Evaluation eXchange (MIREX). Suzhou, China: MIREX, 2017: 9-12.

[102] CHAI W, VERCOE B. Detection of key change in classical piano music [C]//International Society for Music Information Retrieval Conference (ISMIR). London, UK: ISMIR, 2005: 468-473.

[103] ZHU Y, KANKANHALLI M. Key-based melody segmentation for popular songs [C]//International Conference on Pattern Recognition (ICPR). Stockholm, Sweden: ICPR, 2004: 862-865.

[104] BENETOS E, JANSSON A, WEYDE T. Improving automatic music transcription through key detection [C]//International Conferenc: Semantic Audio. London, UK: AES, 2014: 1-7.

[105] IKEDA T, SUZUKI S. Automatic accompaniment device having a function for controlling accompaniment tone on the basis of musical key detection [P]. United States Patent 5412156, 1995.

[106] CAMPBELL S E. Automatic key detection of musical excerpts from audio [D]. Montreal, Quebec, Canada: McGill University, 2010.

[107] ZHU Y, KANKANHALLI M S. Precise pitch profile feature extraction from musical audio for key detection [J]. *IEEE Transactions on Multimedia*, 2006, **8**(3): 575-584.

[108] SCHULLER B, GOLLAN B. Music theoretic and perception-based features for audio key determination [J]. *Journal of New Music Research*, 2012, **41**(2): 175-193.

[109] KRUMHANSL C L. Cognitive foundations of musical pitch [M]. Oxford, UK: Oxford University Press, 1990.

[110] SUN J, LI H, LEI L. Key detection through pitch class distribution model and ANN [C]//IEEE International Conference on Digital Signal Processing(DSP). Santorini-Hellas, Greece: DSP, 2009: 1-6.

[111] SUN J, LI H, MA L. A music key detection method based on pitch class distribution theory[J]. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 2011,**15**(3): 165-175.

[112] SAITO S, TAKEDA H, NISHIMOTO T, et al. Key detection of music audio signals via HMM using Chroma vector through specmurt analysis [J]. *Information Processing Society of Japan SIG Notes*, 2005,**2005**: 85-90.

[113] CHUAN C H. A temporal multi-view approach for audio key finding using adaboost [C]//IEEE International Conference on Multimedia and Expo(ICME). San Jose, USA: ICME, 2013: 1-4.

[114] WU F, SUN S, ZHANG J, et al. Singing voice detection of popular music using beat tracking and SVM classification[C]//IEEE/ACIS International Conference on Computer and Information Science(ICIS). Las Vegas, USA: ICIS, 2015: 525-528.

[115] YOU S D, WU Y C. Comparative study of singing voice detection methods [M]//Book Chapter of Lecture Notes in Electrical Engineering(LNEE). Berlin-Heidelberg: Springer-Verlag, 2015: 1291-1298.

[116] RAMONA M, RICHARD G, DAVID B. Vocal detection in music with support vector machines[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 1885-1888.

[117] REGNIER L, PEETERS G. Singing voice detection in music tracks using direct voice vibrato detection [C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Taipei, China Taiwan: ICASSP, 2009: 1685-1688.

[118] LEHNER B, WIDMER G, BOCK S. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks[C]//IEEE European Signal Processing Conference(EUSIPCO). Nice, France: EUSIPCO, 2015: 21-25.

[119] SANTOSH N, RAMAKRISHNAN S, RAO V, et al. Improving singing voice detection in presence of pitched accompaniment[C]//India National Conference on Communications(NCC). Guwahati, India: NCC, 2009: 276-280.

[120] NEW T L, LI H. Singing voice detection using perceptually-motivated features[C]//ACM International Conference on Multimedia(ACM MM). Augsburg, Germany: ACM, 2007: 309-312.

[121] RAO V, RAMAKRISHNAN S, RAO P. Singing voice detection in polyphonic music using predominant pitch[C]//Conference of the International Speech Communication Association(InterSpeech). Brighton, UK: InterSpeech, 2009: 1131-1134.

[122] WIKIPEDIA. Fundamental frequenc[EB/OL]. [2017-06-18]. https://en.wikipedia.org/wiki/Fundamental_frequency.

[123] LUKASHEVICH H, GRUHNE M, DITTMAR C. Effective singing voice detection in popular music using ARMA filtering [C]//International Conference on Digital Audio Effects(DAFx). Bordeaux, France: DAFx, 2007: 1-4.

[124] LEGLAIVE S, HENNEQUIN R, BADEAU R. Singing voice detection with deep recurrent neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brisbane, Australia: ICASSP,2015: 121-125.

[125] LI W, FENG X, XUE M. Reducing manual labeling in singing voice detection: An active learning approach[C]//IEEE International Conference on Multimedia and Expo(ICME). Seattle, USA: ICME, 2016: 1-5.

[126] LEE K, CREMER M. Automatic labeling of training data for singing voice detection in musical audio[C]// International Conference on Signal Processing, Pattern Recognition and Applications(SPPRA). Innsbruck, Austria: SPPRA, 2009: 1-9.

[127] KHINE S Z K, NEW T L, LI H. Singing voice detection in pop songs using co-training algorithm[C]//

IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 1629-1632.

[128] PIKRAKIS A, KOPSINIS Y, KROHER N, et al. Unsupervised singing voice detection using dictionary learning [C] // IEEE European Signal Processing Conference (EUSIPCO). Budapest, Hungary: EUSIPCO, 2016: 1212-1216.

[129] REGNIER L, PEETERS G. Partial clustering using a time-varying frequency model for singing voice detection [C] //IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, USA: ICASSP, 2010: 441-444.

[130] SHAMILI P, SUNNY K, THASLEEMA T M. A Survey on singing voice separation techniques [J]. *International Journal of Advanced Research in Computer Engineering and Technology*, 2015, **4**(2): 358-361.

[131] SHAMILI P, SUNNY A K, THASLEEMA T M. Singing voice separation using hybrid ICA and wavelet thresholding [J]. *International Journal of Advanced Computer Technology*, 2015, **4**(3): 74-76.

[132] PABLO C M, DAMIÁN M M, MAXIMO C, et al. Singing voice separation from stereo recordings using spatial clues and robust f_0 estimation [C] //International Conference: Semantic Audio. Ilmenau, Germany: AES, 2011: 5.

[133] KIM M, BEACK S, CHOI K. Gaussian mixture model for singing voice separation from stereophonic music [C] //International Conference: Audio for Wirelessly Networked Personal Devices. Pohang, Republic of Korea: Audio Engineering Society, 2011: 1-6.

[134] DURRIEU J L, OZEROV A, FÉVOTTE C, et al. Main instrument separation from stereophonic audio signals using a source/filter model [C] //IEEE European Signal Processing Conference (EUSIPCO). Glasgow, Scotland: EUSIPCO, 2009: 15-19.

[135] HSU C L, JANG J S R. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, **18**(2): 310-319.

[136] IKEMIYA Y, YOSHII K, ITOYAMA K. Singing voice analysis and editing based on mutually dependent f_0 estimation and source separation [C] //IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brisbane, Australia: ICASSP, 2015: 574-578.

[137] IKEMIYA Y, ITOYAMA K, YOSHII K. Singing voice separation and vocal f_0 estimation based on mutual combination of robust principal component analysis and subharmonic summation [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2016, **24**(11): 2084-2095.

[138] ZHU B, LI W, LI R. Multi-stage non-negative matrix factorization for monaural singing voice separation [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2013, **21**(10): 2096-2107.

[139] CHANRUNGUTAI A, RATANAMAHATANA C A. Singing voice separation for mono-channel music using non-negative matrix factorization [C] //IEEE International Conference on Advanced Technologies for Communications(ATC). Hanoi, Vietnam: ATC, 2008: 243-246.

[140] HUANG P S, CHEN S D, SMARAGDIS P, et al. Singing voice separation from monaural recordings using robust principal component analysis [C] //IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Kyoto, Japan: ICASSP, 2012: 57-60.

[141] SPRECHMANN P, BRONSTEIN A, SAPIRO G. Real-time online singing voice separation from monaural recordings using robust low-rank modeling [C] //International Society for Music Information Retrieval Conference(ISMIR). Porto, Portugal: ISMIR, 2012: 67-72.

[142] DRIEDGER J, MÜLLER M. Extracting singing voice from music recordings by cascading audio decomposition techniques [C] // IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brisbane, Australia: ICASSP, 2015: 126-130.

[143] WANG D L, BROWN G J. Computational auditory scene analysis: Principles, algorithms, and applications [M]. New York, USA: Wiley-IEEE Press, 2006.

[144] BREGMAN A S. Auditory scene analysis [M]. Cambridge, USA: MIT Press, 1990.

[145] LI Y, WANG D L. Singing voice separation from monaural recordings [C]//International Society for Music Information Retrieval Conference(ISMIR). Victoria, Canada: ISMIR, 2006: 176-179.

[146] LI Y, WANG D L. Separation of singing voice from music accompaniment for monaural recordings [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(4): 1475-1487.

[147] HSU C L, WANG D L, JANG J S R, et al. A tandem algorithm for singing pitch extraction and voice separation from music accompaniment [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, **20**(5): 1482-1491.

[148] HUANG P S, KIM M, HASEGAWA-JOHNSON M, et al. Joint optimization of masks and deep recurrent neural networks for monaural source separation [J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015, **23**(12): 2136-2147.

[149] HUANG P S, KIM M, HASEGAWA-JOHNSON M, et al. Singing-voice separation from monaural recordings using deep recurrent neural networks [C]//International Society for Music Information Retrieval Conference(ISMIR). Taipei, China Taiwan: ISMIR, 2014: 477-482.

[150] KIM Y E, WILLIAMSON D S, PILLI S. Towards quantifying the "album effect" in artist identification [C]//International Society for Music Information Retrieval Conference (ISMIR). Victoria, Canada: ISMIR, 2006: 393-394.

[151] TSAI T H, HUANG Y S, LIU P Y, et al. Content-based singer classification on compressed domain audio data [J]. *Multimedia Tools and Applications*, 2015, **74**(4): 1489-1509.

[152] LIU C C, HUANG C S. A singer identification technique for content-based classification of MP3 music objects [C]//ACM International Conference on Information and Knowledge Management (CIKM). McLean, USA: ACM, 2002: 438-445.

[153] ZHANG T. Automatic singer identification [C]//IEEE Conference on Multimedia and Expo (ICME). Baltimore, USA: ICME, 2003: 33-36.

[154] DHARINI D, REVATHY A. Singer identification using clustering algorithm [C]//IEEE International Conference on Communications and Signal Processing (ICCSP). Melmaruvathur, India: ICCSP, 2014: 1927-1931.

[155] MESAROS A, ASTOLA J. The Mel-frequency cepstral coefficients in the context of singer identification [C]//International Society for Music Information Retrieval Conference(ISMIR). London, UK: ISMIR, 2005: 610-613.

[156] KHINE S Z K, NEW T L, LI H. Exploring perceptual based timbre feature for singer identification [C]// International Symposium on Computer Music Modeling and Retrieval (CMMR). Copenhagen, Denmark: Springer-Verlag, 2007: 159-171.

[157] NEW T L, LI H. Exploring vibrato-motivated acoustic features for singer identification [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, **15**(2): 519-530.

[158] NEW T L, LI H. On fusion of timbre-motivated features for singing voice detection and singer identification [C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 2225-2228.

[159] BARTSCH M A, WAKEFIELD G H. Singing voice identification using spectral envelope estimation [J]. *IEEE Transactions on Speech and Audio Processing*, 2004, **12**(2): 100-109.

[160] MADDAGE N C, XU C, WANG Y. Singer identification based on vocal and instrumental models [C]// IEEE International Conference on Pattern Recognition(ICPR). Cambridge, UK: ICPR, 2004: 375-378.

[161] **何灼彬.基于卷积深度置信网络的歌手识别** [D].**广州：华南理工大学**,2015.

[162] HU Y, LIU G. Separation of singing voice using nonnegative matrix partial co-factorization for singer identification [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015, **24**(3): 643-653.

[163] SHEN J, CUI B, SHEPHERD J, et al. Towards efficient automated singer identification in large music

databases[C]//International ACM Conference on Research and Development in Information Retrieval (SIGIR). Seattle, USA: SIGIR, 2006: 59-66.

[164] FUJIHARA H, GOTO M, KITAHARA T, et al. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2010, **18**(3): 638-648.

[165] HU Y, LIU G. Automatic singer identification using missing feature methods[C]//IEEE International Conference on Multimedia and Expo(ICME). San Jose, USA: ICME, 2013: 9-14.

[166] TSAI W H, LIN H P. Background music removal based on cepstrum transformation for popular singer identification[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(5): 1196-1205.

[167] TSAI W H, LEE H C. Singer identification based on spoken data in voice characterization[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, **20**(8): 2291-2300.

[168] NAKANO T, GOTO M, HIRAGA Y. Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features[C]//International Conference on Spoken Language Processing(ICSLP). Pittsburgh, USA: ICSLP, 2006: 1706-1709.

[169] MAKA T. Attributes of audio feature contours for automatic singing evaluation[C]//IEEE International Conference on Telecommunications and Signal Processing (ICTSP). Rome, Italy: ICTSP, 2013: 517-520.

[170] TSAI W H, LEE H C. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, **20**(4): 1233-1243.

[171] KATAOKA Y, ITOH K, IKEDA M, et al. Analysis and evaluation of singing voice to produce a support system for singing[J]. *Information Processing Society of Japan SIG Notes*, 1998, **1998**: 23-30.

[172] LIN C H, LEE Y S, CHEN M Y. Automatic singing evaluating system based on acoustic features and rhythm[C]//IEEE International Conference on Orange Technologies(ICOT). Xi'an, China: ICOT, 2014: 165-168.

[173] TSAI W H, MA C H, HSU Y P. Automatic singing performance evaluation using accompanied vocals as reference bases[J]. *Journal of Information Science and Engineering*, 2015, **31**(3): 821-838.

[174] TAKEUCHI H, HOGURO M, UMEZAKI T. A pitch extraction method with high frequency resolution for singing evaluation[J]. *IEEJ Transactions on Electronics Information and Systems*, 2009, **129**(10): 1889-1901.

[175] TAKEUCHI H, HOGURO M, UMEZAKI T. A karaoke system singing evaluation method that more closely matches human evaluation[J]. *IEEJ Transactions on Electronics Information and Systems*, 2010, **130**(6): 1042-1053.

[176] SHA C Y, YANG Y H, LIN Y C, et al. Singing voice timbre classification of Chinese popular music [C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Vancouver, Canada: ICASSP, 2013: 734-738.

[177] KENTARO H, KATUNOBU I. Discrimination method of voice register and voice quality in high pitch for popular singing[J]. *Information Processing Society of Japan SIG Notes*, 2012, **2012**: 1-6.

[178] JHA M V, RAO P. Assessing vowel quality for singing evaluation[C]//India National Conference on Communications(NCC). Kharagpur, India: NCC, 2012.

[179] ASANUMA K, OKAZAKI S, ITOH K, et al. Measurement technique in local area of body surface for singing evaluation[R]. Tokyo, Japan: IEICE Technical Report, 2010, **110**(211): 29-34.

[180] MESAROS A. Singing voice identification and lyrics transcription for music information retrieval[C]// International Conference on Speech Technology & Human-computer Dialogue(SpeD). Cluj-Napoca, Romania: SpeD, 2013.

[181] AWATA S, SAKO S, KITAMURA T. Vowel duration dependent hidden Markov model for automatic

lyrics recognition[J]. *Journal of the Acoustical Society of America*, 2016,**140**(4): 3427-3427.

[182] HOSOYA T, SUZUKI M, ITO A, et al. Lyrics recognition from a singing voice based on finite state automaton for music information retrieval[C]//International Society for Music Information Retrieval Conference(ISMIR). London, UK: ISMIR, 2005: 532-535.

[183] SZEPANNEK G, GRUHNE M, BISCHL B, et al. Perceptually based phoneme recognition in popular music, classification data analysis and knowledge organization[M]. Berlin-Heidelberg: Springer-Verlag, 2010: 751-758.

[184] HADIZADEH H, FATOURECHI M, BAJIC I V. An automatic lyrics recognition system for digital videos[EB/OL]. [2017-05-29]. http://www.sfu.ca/~ibajic/pubs/hfb_mmsp2012.pdf.

[185] MASSARO D W, JESSE A. Read my lips: Speech distortions in musical lyrics can be overcome (slightly) by facial information[J]. *Speech Communication*, 2009,**51**(7): 604-621.

[186] CANO P, BATLE E, KALKER T, et al. A review of algorithms for audio fingerprinting[C]//IEEE Workshop on Multimedia Signal Processing(MSP). St. Thomas, USA: MSP, 2002: 169-173.

[187] MIOTTO R, MONTECCHIO N. Integration of Chroma and rhythm histogram features in a music identification system[C]//Workshop on Exploring Musical Information Spaces(WEMIS). Corfu, Greece: WEMIS, 2009: 41-45.

[188] MURATA K, NAKADAI K, YOSHII K, et al. A robot singer with music recognition based on real-time beat tracking[C]//International Society for Music Information Retrieval Conference(ISMIR). Philadelphia, USA: ISMIR, 2008: 199-204.

[189] XU T, JIA R, LI H, et al. Music identification with KD-tree and melody-line[C]//IEEE International Conference on Multimedia Technology(ICMT). Hanzhou, China: ICMT, 2011: 576-580.

[190] HÉBERT S, PERETZ I. Recognition of music in long-term memory: Are melodic and temporal patterns equal partners?[J]. *Memory & Cognition*, 1997,**25**(4): 518-533.

[191] XU T, YU A W, LIU X, et al. Music identification via vocabulary tree with MFCC peaks[C]//ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies(MIRUM). Scottsdale, USA: MIRUM, 2011: 21-26.

[192] CRYSANDT H. Music identification with MPEG-7[C]// Storage and Retrieval Methods and Applications for Multimedia. San Jose, California: Society of Photo-Optical Instrumentation Engineers(SPIE), 2003: 117-125.

[193] WANG A. The Shazam music recognition service[J]. *Communications of the ACM-Music Information Retrieval*, 2006,**49**(8): 44-48.

[194] KE Y, HOIEM D, SUKTHANKAR R. Computer vision for music identification[C]//IEEE Internatinoal Conference on Computer Vision and Pattern Recognition(CVPR). San Diego, USA: CVPR, 2005: 597-604.

[195] LI W, XIAO C, LIU Y. Low-order auditory Zernike moment: A novel approach for robust music identification in the compressed domain[J]. *EURASIP Journal on Advances in Signal Processing*, 2013,**2013**: 132.

[196] ZHU B, LI W, WANG Z. A novel audio fingerprinting method robust to time scale modification and pitch shifting[C]//ACM International Conference on Multimedia(ACM MM). Firenze, Italy: ACM, 2010: 987-990.

[197] BORJIAN N, KABIR E, SEYEDIN S, et al. A query-by-example music retrieval system using feature and decision fusion[J]. *Multimedia Tools and Applications*, 2018,**77**(5): 6165-6189.

[198] GHIAS A, LOGAN J, CHAMBERLIN D, et al. Query by humming: Musical information retrieval in an audio database[C]//ACM International Conference on Multimedia(ACM MM). San Francisco, USA: ACM, 1995: 231-236.

[199] PARKER C. Towards intelligent string matching in query-by-humming systems[C]//IEEE International

Conference on Multimedia and Expo(ICME). Baltimore, USA: ICME, 2003: 25-28.

[200] KOSUGI N, NISHIHARA Y, SAKATA T, et al. A practical query-by-humming system for a large music database[C]//ACM International Conference on Multimedia(ACM MM). Marina del Rey, USA: ACM, 2000: 333-342.

[201] SONG J, BAE S Y, YOON K. Mid-level music melody representation of polyphonic audio for query-by-humming system[C]//International Society for Music Information Retrieval Conference(ISMIR). Porto, Portugal: ISIMIR, 2002: 133-139.

[202] YU H M, TSAI W H, WANG H M. A query-by-singing system for retrieving karaoke music[J]. *IEEE Transactions on Multimedia*, 2008, **10**(8): 1626-1637.

[203] SAILER C. Using string alignment in a query-by-humming system for real world applications[J]. *Journal of the Acoustical Society of America*, 2005, **118**(3): 2032-2032.

[204] SHIH H H, NARAYANAN S S, KUO C C J. Multidimensional humming transcription using hidden Markov models for query by humming systems[C]//IEEE International Conference on Multimedia and Expo(ICME). Baltimore, USA: ICME, 2003: 541-544.

[205] UNAL E, CHEW E, GEORGIOU P G. Challenging uncertainty in query by humming systems: A fingerprinting approach[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2008, **16**(2): 359-371.

[206] DUC T N T, NHAT M L, HOANG H N D, et al. Continuous pitch contour as an improvement feature for music information retrieval by humming/singing[C]//Pacific Rim International Conference on Artificial Intelligence(PRICAI). Hanoi, Vietnam: Springer, 2008: 1086-1091.

[207] ANANTHAKRISHNAN G, RAMAKRISHNAN A G. Relative pitch tracking for singing voice as an application in query by humming systems[C]//IASTED International Conference on Signal Processing, Pattern Recognition and Applications(SPPRA). Innsbruck, Austria: ACTA Press, 2007: 275-280.

[208] WANG Q, GUO Z, LI B, et al. Tempo variation based multilayer filters for query by humming[C]// International Conference on Pattern Recognition(ICPR). Tsukuba, Japan: ICPR, 2012: 3034-3037.

[209] KOSUGI N, SAKURAI Y, MORIMOTO M. SoundCompass: A practical query-by-humming system[C]// ACM International Conference on Management of Data(SIGMOD). Paris, France: SIGMOD, 2004: 881-886.

[210] WANG Q, GUO Z, LIU G, et al. Local alignment for query by humming[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Vancouver, Canada: ICASSP, 2013: 3711-3715.

[211] YANG J, LIU J, ZHANG W Q. A fast query by humming system based on notes[C]//Conference of the International Speech Communication Association (InterSpeech). Makuhari, Chiba, Japan: InterSpeech, 2010: 2898-2901.

[212] LI M, ZHAO Z, SHI P. Query by humming based on music phrase segmentation and matching[C]// IEEE International Conference on Fuzzy Systems and Knowledge Discovery(FSKD). Zhangjiajie, China: FSKD, 2015: 1966-1970.

[213] CAO W, JIANG D, HOU J, et al. A phrase-level piecewise linear scaling algorithm for melody match in query-by-humming systems[C]//IEEE International Conference on Multimedia and Expo(ICME). New York, USA: ICME, 2009: 942-945.

[214] HOU J, JIANG D, CAO W, et al. Effectiveness of n-gram fast match for query-by-humming systems [C]//IEEE International Conference on Multimedia and Expo(ICME). New York, USA: ICME, 2009: 1310-1313.

[215] ITO A, HEO S P, SUZUKI M, et al. Comparison of features for DP-matching based query-by-humming system [C]//International Society for Music Information Retrieval Conference (ISMIR), Barcelona, Spain: ISMIR, 2004: 1-6.

[216] TSAI W H, TU Y M, MA C H. An FFT-based fast melody comparison method for query-by-singing/

humming systems[J]. *Pattern Recognition Letters*, 2012,**33**(16): 2285-2291.

[217] GUO Z, WANG Q, LIU G. A query by humming system based on locality sensitive hashing indexes[J]. *Signal Processing*, 2013,**93**(8): 2229-2243.

[218] PARK C H. Query by humming based on multiple spectral hashing and scaled open-end dynamic time warping[J]. *Signal Processing*, 2015,**108**: 220-225.

[219] FERRARO P, HANNA P, IMBERT L, et al. Accelerating query-by-humming on GPU [C]// International Society for Music Information Retrieval Conference(ISMIR). Kobe, Japan: ISMIR, 2009: 279-284.

[220] NAM G P, LUONG T T T, NAM H H, et al. Intelligent query by humming system based on score level fusion of multiple classifiers[J]. *EURASIP Journal on Advances in Signal Processing*, 2011, **2011**: 21.

[221] SERRÀ J, GÓMEZ E, HERRERA P. Audio cover song identification and similarity: Background, approaches, evaluation, and beyond, book chapter of advances in music information retrieval[M]. Berlin-Heidelberg: Springer-Verlag, 2010: 307-332.

[222] TRALIE C J, BENDICH P. Cover song identification with timbral shape sequences[C]//International Society for Music Information Retrieval Conference(ISMIR). Malaga, Spain: ISMIR, 2015: 38-44.

[223] MAROLT M. Melody-based retrieval in audio collections[J]. *Journal of the Acoustical Society of America*, 2007,**122**(5): 2962.

[224] SAILER C, DRESSLER K. Finding cover songs by melodic similarity[C]//Music Information Retrieval Evaluation eXchange(MIREX). Victoria, Canada: MIREX, 2006: 1-3.

[225] KIM Y E, PERELSTEIN D. Audio cover song detection using chroma features and hidden Markov model[C]//Music Information Retrieval Evaluation eXchange(MIREX). Vienna, Austria: MIREX, 2007: 27-28.

[226] WANG C L, ZHONG Q, WANG S Y, et al. Cover song identification by sequence alignment algorithms [C]//International Conference on Graphic and Image Processing(ICGIP). Cairo, Egypt: Society of Photo-Optical Instrumentation Engineers(SPIE), 2011: 1-7.

[227] JENSEN J H, CHRISTENSEN M G, ELLIS D P W, et al. A tempo-insensitive distance measure for cover song identification based on Chroma features[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 2209-2212.

[228] KIM S, NARAYANAN S. Dynamic Chroma feature vectors with applications to cover song identification [C]//IEEE Workshop on Multimedia Signal Processing(MSP). Cairns, Australia: MSP, 2008: 984-987.

[229] FANG J T, DAY C T, CHANG P C. Deep feature learning for cover song identification[J]. *Multimedia Tools and Applications*, 2017,**76**(22): 23225-23238.

[230] LEE K. Identifying cover songs from audio using harmonic representation[C]//Music Information Retrieval Evaluation eXchange(MIREX). Victoria, Canada: MIREX, 2006: 12-14.

[231] CHANG T M, CHEN E T, HSIEH C B, et al. Cover song identification with direct chroma feature extraction from AAC files[C]//IEEE Global Conference on Consumer Electronics(GCCE). Tokyo, Japan: GCCE, 2013: 55-56.

[232] TAVENARD R, JÉGOU H, LAGRANGE M. Efficient cover song identification using approximate nearest neighbors[EB/OL]. [2017-05-12]. https://hal.archives-ouvertes.fr/file/index/docid/672897/filename/hal2012.pdf.

[233] CAI K, YANG D, CHEN X. Cross-similarity measurement of music sections: A framework for large-scale cover song identification[C]//International Conference on Intelligent Information Hiding and Multimedia Signal Processing(IIH-MSP). Kaohsiung, China Taiwan: Springer, 2016: 151-158.

[234] 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.

[235] FOSTER P, DIXON S, KLAPURI A. Identifying cover songs using information-theoretic measures of

similarity[J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015, **23**(6): 993-1005.

[236] DEGANI A, DALAI M, LEONARDI R, et al. A heuristic for distance fusion in cover song identification[C]//IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). Paris, France: WIAMIS, 2013: 1-4.

[237] JANG J S R, LEE H R, YEH C H. Query by tapping: A new paradigm for content-based music retrieval from acoustic input [C] //Pacific-Rim Conference on Multimedia (PCM). Bejing, China: Springer, 2001: 590-597.

[238] CHEN C T, JANG J S R, LU C H. Improved query-by-tapping via tempo alignment[C]//International Society for Music Information Retrieval Conference (ISMIR). Taipei, China Taiwan: ISMIR, 2014: 289-294.

[239] EISENBERG G, BATKE J M, SIKORA T. BeatBank—An MPEG-7 compliant query by tapping system [C]//Audio Engineering Society(AES) Convention. Berlin, Germany: AES, 2004: 1-7.

[240] HANNA P, ROBINE M. Query by tapping system based on alignment algorithm [C] //IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Taipei, China Taiwan: ICASSP, 2009: 1881-1884.

[241] TZANETAKIS G, ERMOLINSKYI A, COOK P. Pitch histograms in audio and symbolic music information retrieval[J]. *Journal of New Music Research*, 2003, **32**(2): 143-152.

[242] LEE C H, SHIH J L, YU K M, et al. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features[J]. *IEEE Transactions on Multimedia*, 2009, **11**(4): 670-682.

[243] MCKINNEY M F, BREEBAART J. Features for audio and music classification [C] //International Society for Music Information Retrieval Conference(ISMIR). Baltimore, USA: ISMIR, 2003: 151-158.

[244] MENG A, SHAWE-TAYLOR J. An investigation of feature models for music genre classification using the support vector classifier [C] //International Society for Music Information Retrieval Conference (ISMIR). London, UK: ISMIR, 2005: 604-609.

[245] RAJANNA A R, ARYAFAR K, SHOKOUFANDEH A, et al. Deep neural networks: A case study for music genre classification [C] //IEEE International Conference on Machine Learning and Applications (ICMLA). Miami, USA: ICMLA, 2015: 655-660.

[246] ZHANG W, LEI W, XU X, et al. Improved music genre classification with convolutional neural networks[C] //Conference of the International Speech Communication Association (InterSpeech). San Francisco, USA: InterSpeech, 2016: 3304-3308.

[247] SHAO X, XU C, KANKANHALLI M S. Unsupervised classification of music genre using hidden Markov model[C] //IEEE International Conference on Multimedia and Expo (ICME). Taipei, China Taiwan: ICME, 2004: 2023-2026.

[248] MEARNS L, TIDHAR D, DIXON S. Characterization of composer style using high-level musical features[C]//International Workshop on Machine Learning and Music(MML). Firenze, Italy: MML, 2010: 37-40.

[249] **胡振,傅昆,张长水.基于深度学习的作曲家分类问题**[J].**计算机研究与发展**,2014,**51**(9): 1945-1954.

[250] HERREMANS D, MARTENS D, SÖRENSEN K. Composer classification models for music-theory building, book chapter of computational music analysis [M]. New York, USA: Springer, 2016: 369-392.

[251] MARTIN K D, KIM Y E. Musical instrument identification: A pattern-recognition approach [J]. *Journal of the Acoustical Society of America*, 1998, **104**(3): 1768-1768.

[252] MARQUES J, MORENO P J. A study of musical instrument classification using Gaussian mixture models and support vector machines[R]. Cambridge, UK: Cambridge Research Laboratory, 1999.

[253] BROWN J C. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features[J]. *Journal of the Acoustical Society of America*, 1999, **105**(3): 1933-1941.

[254] PATIL K, ELHILALI M. Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases[J]. *EURASIP Journal on Audio Speech and Music Processing*, 2015, **2015**: 27.

[255] HAN Y, LEE S, NAM J, et al. Sparse feature learning for instrument identification: Effects of sampling and pooling methods[J]. *Journal of the Acoustical Society of America*, 2016, **139**(5): 2290-2298.

[256] HAN Y, KIM J, LEE K, et al. Deep convolutional neural networks for predominant instrument recognition in polyphonic music[J]. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2017, **25**(1): 208-221.

[257] PAULUS J, MÜLLER M, KLAPURI A. State of the art report: Audio-based music structure analysis[C]//International Society for Music Information Retrieval Conference(ISMIR). Utrecht, Netherlands: ISMIR, 2010: 625-636.

[258] SHIU Y, JEONG H, KUO C C J. Similar segment detection for music structure analysis via Viterbi algorithm[C]//IEEE International Conference on Multimedia and Expo(ICME). Toronto, Canada: ICME, 2006: 789-792.

[259] XU C, MADDAGE N C, SHAO X, et al. Content-adaptive digital music watermarking based on music structure analysis[J]. *ACM Transactions on Multimedia Computing Communications and Applications*, 2007, **2007**: 1-16.

[260] JUN S, RHO S, HWANG E. Music structure analysis using self-similarity matrix and two-stage categorization[J]. *Multimedia Tools and Applications*, 2015, **74**(1): 287-302.

[261] SHIU Y, JEONG H, KUO C C J. Musical structure analysis using similarity matrix and dynamic programming[C]//Multimedia Systems and Applications VIII. Boston, United States: Society of Photo-Optical Instrumentation Engineers(SPIE), 2005, **6015**: 398-409.

[262] BELLO J P. Measuring structural similarity in music[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **19**(7): 2013-2025.

[263] MADDAGE N C, XU C, KANKANHALLI M S, et al. Content-based music structure analysis with applications to music semantics understanding[C]//ACM International Conference on Multimedia(ACM MM). New York, USA: ACM, 2004: 112-119.

[264] **李相莲,李明,刘若伦,等.基于音色单元分布的音乐结构分析**[J].**声学学报**,2010,**35**(2): 276-281.

[265] PANAGAKIS Y, KOTROPOULOS C. Music structure analysis by subspace modeling[C]//IEEE European Signal Processing Conference(EUSIPCO). Bucharest, Romania: EUSIPCO, 2012: 1459-1463.

[266] ZHANG T, FONG C K, XIAO L, et al. Automatic and instant ring tone generation based on music structure analysis[C]//ACM International Conference on Multimedia(ACM MM). Beijing, China: ACM, 2009: 593-596.

[267] NIETO O, HUMPHREY E J, BELLO J P. Compressing music recordings into audio summaries[C]// International Society for Music Information Retrieval Conference(ISMIR). Porto, Portugal: ISMIR, 2012: 313-318.

[268] TIAN A, LI W, XIAO L, et al. Histogram matching for music repetition detection[C]//IEEE International Conference on Multimedia and Expo(ICME). New York, USA: ICME, 2009: 662-665.

[269] CHAI W. Semantic segmentation and summarization of music: Methods based on tonality and recurrent structure[J]. *IEEE Signal Processing Magazine*, 2006, **23**(2): 124-132.

[270] SHAO X, MADDAGE M C, XU C. Automatic music summarization based on music structure analysis[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Philadelphia, USA: ICASSP, 2005: 1169-1172.

[271] JIANG N, MÜLLER M. Estimating double thumbnails for music recordings[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brisbane, Australia: ICASSP, 2015: 146-150.

[272] LU L, LIU D, ZHANG H J. Automatic mood detection and tracking of music audio signals[J]. *IEEE*

Transactions on Audio, Speech and Language Processing, 2006,**14**(1): 5-18.

[273] WIECZORKOWSKA A, SYNAK P, RAS Z W. Multi-label classification of emotions in music[C]// International Conference on Intelligent Information Processing and Web Mining(IIPWM). Ustron, Poland: Springer, 2006: 307-315.

[274] KIM Y E, SCHMIDT E M, MIGNECO R, et al. Music emotion recognition: A state of the art review [C]//International Society for Music Information Retrieval Conference(ISMIR). Utrecht, Netherlands: ISMIR, 2010: 255-266.

[275] YANG Y H, LIN Y C, SU Y F, et al. A regression approach to music emotion recognition[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2008,**16**(2): 448-457.

[276] **孙晓煜.自动化音乐情感分类问题的研究**[D].**杭州: 浙江大学**,2010.

[277] WANG X, CHEN X, YANG D, et al. Music emotion classification of Chinese songs based on lyrics using TF * IDF and rhyme [C] //International Society for Music Information Retrieval Conference (ISMIR). Miami, Florida: ISMIR, 2011: 765-770.

[278] SHARDANAND U, MAES P. Social information filtering: algorithms for automating word of mouth[C]// SIGCHI conference on Human factors in computing systems. Denver, Colorado: ACM, 1995: 210-217.

[279] SCHEIN A I, POPESCUL A, UNGAR L H, et al. Methods and metrics for cold-start recommendation [C]//International ACM Conference on Research and Development in Information Retrieval(SIGIR). Tampere, Finland: SIGIR, 2002: 253-260.

[280] CHEN Z S, JANG J S R, LEE C H. A kernel framework for content-based artist recommendation system in music[J]. *IEEE Transactions on Multimedia*, 2011,**13**(6): 1371-1380.

[281] **刘珊珊.音频特征与社会标签相结合的音乐推荐系统**[D].**武汉: 华中科技大学**,2011.

[282] KUO F F, CHIANG M F, SHAN M K, et al. Emotion-based music recommendation by association discovery from film music [C] //ACM International Conference on Multimedia(ACM MM). Hilton, Singapore: ACM, 2005: 507-510.

[283] SORDO M, COVIELLO E. Tutorial on music auto-tagging [C] //International Society for Music Information Retrieval Conference(ISMIR), Curitiba, Brazil: ISMIR, 2013.

[284] BERTIN-MAHIEUX T, ECK D, MANDEL M. Automatic tagging of audio: the state-of-the-art[J]. *Machine Audition: Principles, Algorithms and Systems*, 2010,**2010**: 334-352.

[285] BARRINGTON L, O'MALLEY D, TURNBULL D et al. User-centered design of a social game to tag music [J]. ACM SIGKDD Workshop on Human Computation(HCOMP). Paris, France: HCOMP, 2009: 7-10.

[286] CHOI K, FAZEKAS G, SANDLER M. Automatic tagging using deep convolutional neural networks[C]// International Society for Music Information Retrieval Conference(ISMIR). New York, USA: ISMIR, 2016: 805-811.

[287] WANG J C, YANG Y H, WANG H M, et al. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval[C]//ACM International Conference on Multimedia(ACM MM). Nara, Japan: ACM, 2012: 89-98.

[288] CHEN G, WANG T, GONG L, et al. Multi-class support vector machine active learning for music annotation[J]. *International Journal of Innovative Computing Information and Control*, 2010,**6**(3): 921-930.

[289] WANG Y, KAN M Y, NEW T L, et al. LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics[C]//ACM international conference on Multimedia(ACM MM). New York, USA: ACM, 2004: 212-219.

[290] KAN M Y, WANG Y, ISKANDAR D, et al. LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2008,**16**(2): 338-349.

[291] ISKANDAR D, WANG Y, KAN M Y, et al. Syllabic level automatic synchronization of music signals and text lyrics[C]//ACM International Conference on Multimedia(ACM MM). Santa Barbara, USA: ACM, 2006: 659-662.

[292] FUJIHARA H, GOTO M, OGATA J. Automatic synchronization between lyrics and music CD

recordings based on viterbi alignment of segregated vocal signals[C]//IEEE International Symposium on Multimedia(ISM). San Diego, USA: ISM, 2006: 257-264.

[293] FUJIHARA H, GOTO M, OGATA J. LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics[J]. *IEEE Journal of Selected Topics in Signal Processing*, 2011, **5**(6): 1252-1261.

[294] FUJIHARA H, GOTO M. Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection[C]// IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Las Vegas, USA: ICASSP, 2008: 69-72.

[295] MADDAGE N C, SIM K C, LI H. Word level automatic alignment of music and lyrics using vocal synthesis[J]. *ACM Transactions on Multimedia Computing Communications and Applications*, 2010, **6**(3): 1-16.

[296] **申涛.算法作曲中对节奏控制的若干方式[D].武汉：武汉音乐学院**,2009.

[297] FERNÁNDEZ J D, VICO F. AI methods in algorithmic composition: A comprehensive survey[J]. *Journal of Artificial Intelligence Research(JAIR)*, 2013, **48**(1): 513-582.

[298] **冯寅,周昌乐.算法作曲的研究进展[J].软件学报**,2006, **17**(2): 209-215.

[299] MORONI A, MANZOLLI J, ZUBEN F V, et al. Evolutionary computation applied to algorithmic composition[C]//Congress on Evolutionary Computation(CEC). Washington, USA: CEC, 1999: 807-811.

[300] SIEVERS J, WAGENIUS R. Algorithmic composition from text: How well can a computer generated song express emotion?[EB/OL].[2017-05-29]. http://www.diva-portal.org/smash/get/diva2:723667/FULLTEXT01.

[301] COLOMBO F, MUSCINELLI S P, SEEHOLZER A, et al. Algorithmic composition of melodies with deep recurrent neural networks[C]//Conference on Computer Simulation of Musical Creativity(CSMC). Huddersfield, England: Huddersfield University, 2016: 1-12.

[302] TERNSTRÖM S, SUNDBERG J. Formant-based synthesis of singing[C]//Conference of the International Speech Communication Association(InterSpeech). Antwerp, Belgium: InterSpeech, 2007: 4013-4014.

[303] DEGOTTEX G, ARDAILLON L, ROEBEL A. Multi-frame amplitude envelope estimation for modification of singing voice[J]. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2016, **24**(7): 1242-1254.

[304] CHAN P Y, DONG M, LIM Y Q, et al. Formant excursion in singing synthesis[C]//IEEE International Conference on Digital Signal Processing(DSP). Singapore: DSP, 2015: 168-172.

[305] KENMOCHI H, OHSHITA H. Vocaloid-commercial singing synthesizer based on sample concatenation[C]//Conference of the International Speech Communication Association(InterSpeech). Antwerp, Belgium: InterSpeech, 2007: 4009-4010.

[306] **李娟.基于语料库的歌声合成技术[D].北京：北京师范大学信息科学与技术学院**,2011.

[307] SHIROTA K, NAKAMURA K, HASHIMOTO K, et al. Integration of speaker and pitch adaptive training for HMM-based singing voice synthesis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Florence, Italy: ICASSP, 2014: 2578-2582.

[308] NISHIMURA M, HASHIMOTO K, OURA K, et al. Singing voice synthesis based on deep neural networks[C]//Conference of the International Speech Communication Association(InterSpeech). San Francisco, USA: InterSpeech, 2016: 2478-2482.

[309] HONGO K, NOSE T, ITO A. Spectral and pitch modeling with hybrid approach to singing voice synthesis using hidden semi-Markov model and deep neural network[J]. *Journal of the Acoustical Society of America*, 2016, **140**(40): 2962-2962.

[310] ÖZER S. f_0 modeling for singing voice synthesizers with LSTM recurrent neural networks[D]. Barcelona, Spain: Universitat Pompeu Fabra, 2015.

[311] ARDAILLON L, CHABOT-CANET C, ROEBEL A. Expressive control of singing voice synthesis

using musical contexts and a parametric f_0 model [C]//Conference of the International Speech Communication Association(InterSpeech). San Francisco, USA: InterSpeech, 2016: 1250-1254.

[312] SUN J, LING Z, JIANG Y, et al. Method and device for converting speaking voice into singing[P]. WIPO Patent Application, CN2012/08799,2014. .

[313] **张智星,徐志浩,李宏儒,等.歌声合成系统、方法以及装置**[P].CN102024453A,2012.

[314] SAITOU T, GOTO M, UNOKI M, et al. SingBySpeaking: Singing voice conversion system from speaking voice by controlling acoustic features affecting singing voice perception [J]. *Information Processing Society of Japan SIG Notes*, 2008,**2008**: 25-32.

[315] NGUYEN H Q, LEE S W, TIAN X. High quality voice conversion using prosodic and high-resolution spectral features[J]. *Multimedia Tools and Applications*, 2016,**75**(9): 5265-5285.

[316] MITROO J B, HERMAN N, BADLER N I. Movies from music: Visualizing musical compositions[C]//ACM Annual Conference on Computer Graphics and Interactive Techniques(SIGGRAPH). Chicago, Illinois: SIGGRAPH, 1979,**13**(2): 218-225.

[317] CHEN Y H, KUO J H, CHU W T, et al. Movie emotional event detection based on music mood and video tempo[C]//International Conference on Consumer Electronics(ICCE). Las Vegas, USA: ICCE, 2006: 151-152.

[318] BERMAN D R. AVISARME: Audio-visual synchronization algorithm for a robotic musician ensemble[D]// Washington: University of Maryland, Department of Mechanical Engineering, 2012.

[319] **邵曦,刘君芳,季茜成.基于情感的家庭音乐相册自动生成研究**[J].**复旦学报(自然科学版)**,2017,**56**(2): 149-158.

[320] XU M, XU C, DUAN L, et al. Audio keywords generation for sports video analysis [J]. *ACM Transactions on Multimedia Computing Communications and Applications*, 2008,**4**(2): 11.

[321] **李伟,袁一群,李晓强,等.数字音频水印技术综述**[J].**通信学报**,2005,**26**(2): 100-111.

[322] SWANSON M D, ZHU B, TEWFIK A H, et al. Robust audio watermarking using perceptual masking [J]. *Signal Processing*,1998,**66**(3): 337-355.

[323] Cox I J, KILIAN J, LEIGHTON F T, et al. Secure spread spectrum watermarking for multimedia[J]. *IEEE Transactions on Image Processing*, 1997,**6**(12): 1673-1687.

[324] KAUR A, DUTTA M K, SONI K M, et al. Localized & self-adaptive audio watermarking algorithm in the wavelet domain[J]. *Journal of Information Security and Applications*, 2017,33: 1-15.

[325] COX I J, MILLER M L, BLOOM J A. **数字水印**[M].**王颖,黄志蓓译.北京: 电子工业出版社**,2003.

[326] KUTTER M, BHATTACHARJEE S K, EBRAHIMI T. Towards second generation watermarking schemes[C]//IEEE International Conference on Image Processing(ICIP). Kobe, Japan: ICIP, 1999: 320-323.

[327] LI W, XUE X. An audio watermarking technique that is robust against random cropping[J]. *Computer Music Journal*, 2003,**27**(4): 58-68.

[328] LI W, XUE X, LU P. Localized audio watermarking technique robust against time-scale modification [J]. *IEEE Transactions on Multimedia*, 2006,**8**(1): 60-69.

[329] WU C P, SU P C, KUO C C J. Robust audio watermarking for copyright protection[C]//Advanced Signal Processing Algorithms, Architectures and Implementations. Denver, Colorado: Society of Photo-Optical Instrumentation Engineers(SPIE), 1999,**3807**: 387-397.

[330] BENDER W, GRUHL D, MORIMOTO N, et al. Techniques for data hiding [J]. *IBM Systems Journal*, 1996,**35**(3.4): 313-336.

[331] LI W, ZHANG X, WANG Z. Music content authentication based on beat segmentation and fuzzy classification[J]. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013,**2013**: 11.

[332] GUPTA S, CHO S, KUO C C J. Current developments and future trends in audio authentication[J]. *IEEE Multimedia*, 2012,**19**(1): 50-59.

[333] KRAETZER C, OERMANN A, DITTMANN J, et al. Digital audio forensics: A first practical evaluation on microphone and environment classification [C]//ACM Workshop on Multimedia and

Security(MM&Sec). Dallas, USA: ACM, 2007: 63-74.

[334] YANG R, SHI Y Q, HUANG J. Defeating fake-quality MP3[C]//ACM Workshop on Multimedia and Security(MM&Sec). Princeton, USA: ACM, 2009: 117-124.

[335] XUE X, LI W, YIN Y. Towards content-based audio fragment authentication[C]//ACM international conference on Multimedia(ACM MM). Scottsdale, USA: ACM, 2011: 1249-1252.

Understanding Digital Music—A Review of Music Information Retrieval Technology

LI Wei[1,2], LI Zijin[3], GAO Yongwei[1]

(1. *School of Computer Science and Technology, Fudan University, Shanghai* 201203, *China*;
2. *Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai* 200433, *China*;
3. *Department of Music Technology, China Conservatory of Music, Beijing* 100101, *China*)

Abstract: In recent twenty years, with the mature of audio compression technique and the proliferation of the Internet, the format of music has quickly changed from tapes and compact discs to digital music on the Internet. Vast amount of digital music has led to a series of problems on classification, organization, retrieval, content understanding and analysis etc, which has fostered a new interdisciplinary research field, i.e., content-based music information retrieval(MIR). This paper clarifies the differences and relations between MIR and music technology, sound and music computing, computer audition, speech information processing, and music acoustics. We divide dozens of MIR research fields into a core(inner) circle and an application(outer) circle, the core circle is more related to fundamental music elements such as pitch, melody, rhythm and harmonic etc. With regards to these research fields, we concisely introduce the basic concept, applications, principles, technical frameworks and typical algorithms. Also, related music domain knowledge is illustrated, and the Chinese translation of technical terms are standardized. At last, we point out the possible problems that exist in computer audition, and prospect the future development.

Keywords: music technology; sound and music computing; computer audition; music information retrieval; music acoustics