doi: 10.3969/j.issn.0490-6756.2020.03.016

一种自适应谐波叠加的复调音乐主旋律提取新方法

何甜田，何培宇，陈杰梅
(四川大学电子信息学院，成都 610065)

摘　要：随着音乐数量的迅速增加，对音乐进行数字化的处理已经成为必然趋势. 主旋律反映了音乐的主要思想，提取主旋律在制作计算机音乐，检索分类，哼唱识别等领域具有广泛的应用价值. 本文提出一种自适应谐波叠加的复调音乐主旋律提取算法. 首先，通过声源分离预处理，通过判别基频最小稳定方差改变压缩因子，自适应叠加谐波构建显著函数；然后，对显著函数构建的基频片断采用随机森林模型进行人声检测，组合所有人声帧的最大显著度频率得到音乐的主旋律序列. 实验表明，在 MIR-1K 数据集上得到的结果在高信噪比情况下有显著提升.

关键词：主旋律提取；自适应谐波叠加；随机森林；人声检测
中图分类号：TN912　　**文献标识码**：A　　**文章编号**：0490-6756(2020)03-0519-07

A new method of melody extraction for polyphonic music based on adaptive harmonic superposition

HE Tian-Tian, *HE Pei-Yu*, *CHEN Jie-Mei*
(College of Electronic Information and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: With the rapid increase in thevolume of music, digital processing of music has become an inevitable trend. Melody reflects the main ideas of music, melody extraction has wide application in computer music production, music retrieval and classification, and humming recognition and other fields. In this paper, a music melody extraction algorithm is proposed for polyphonic music based on adaptive harmonic superposition. A saliency function is first constructed by adaptive harmonic superposition through the preprocess of harmonic percussive source separation and the change of compression factor by discriminating the minimum stable variance of fundamental frequency. Then, random forest model is used to voice detection in the fundamental frequency segment constructed by the saliency function. Combining the frequency with maximum significance to get the melody sequence of music. Experiments show that the results obtained from the MIR-1K data set are significantly improved in the case of high signal-to-noise ratio.

Keywords: Melody extraction; Adaptive harmonic superposition; Random forest; Voice detection

1　引　言

随着近年来数字音乐产业的不断发展，人们对获取音乐信息的需求也与日俱增. 旋律是音乐的灵魂和基础，可以表达出音乐的情感意义. 旋律通常是指一个单音的基频序列[1]，但是大多数音乐中，

收稿日期：2019-07-20
基金项目：四川省科技支撑项目(2011SZ0123)；四川省科技支撑项目(2013GZ1043)
作者简介：何甜田(1995—)，女，硕士研究生，研究方向为声学信号处理与应用. E-mail：125944382@qq.com
通讯作者：何培宇. E-mail：hepeiyu@scu.edu.cn

同一时刻的声音通常来自多个不同声源,此类音乐称之为复调音乐. 主旋律提取的目的就是在复调音乐中自动地判别出主导声源的人声或者器乐的旋律. 它可以广泛地应用于哼唱识别、语音合成[2]、内容推荐[3]、制作医疗音乐[4]等.

主旋律提取问题的提出起源于上世纪九十年代,在 2004 年 Goto[5]首次提出了对复调音乐的主旋律提取方法. 他提出计算信号的短时幅度谱,并用加权混合模型对其进行建模,然后计算每帧中具有最大期望概率的基频,构成主旋律. Salamon[6]提出了能量映射叠加与构建音高轮廓线的方法,将线性频域转换到了音乐适用的对数域,使得音高提取更加精准. Ono 等人[7]提出了一种谐波与击打声源分离算法(Harmonic Percussive Source Separation, HPSS),该算法可以分离在时间上平滑连续的和弦分量和在频率上平滑连续的冲击分量. 上述算法虽然各有优点,但大多依赖能量谱的准确性,且无法准确区分人声与伴奏.

因此,本文采用了谐波与击打声源分离作为预处理,将分离后的声源作为输入. 自适应改变压缩因子的值,对谐波进行叠加,在达到最小稳定方差时构建显著度函数进行多基频估计,构建基频片段. 然后通过对训练集提取声学特征,生成随机森林模型[8]. 检测人声存在段映射到基频片段上,选取显著度最大频率作为主旋律. 实验结果表明,在高信噪比情况下整体准确率有显著提升.

2 相关乐理基础介绍

2.1 基频与谐波

一般的声音都是由发音体发出的一系列频率、振幅各不相同的振动复合而成的. 这些振动中有一个频率最低的振动,由它发出的音就是基音,基音的频率称为基频. 谐波存在于基频的整数倍处,也会有较大的能量. 正是不同的谐波分布导致了相同基频的不同发声体的具体音色不同.

2.2 十二平均律

十二平均律是世界上通用的一种音乐定律方法,它将一个八度的音按照频率等比例地分成十二等份,每一等份称为一个半音. 前后两个半音间的频率倍数关系满足:

$$f_{k+1}=2^{\frac{1}{12}}\cdot f_k \tag{1}$$

为了方便计算,本文将所有频率值转换为十二平均律中的音阶. 目前国际标准音是 A4,转换为物理频率是 440 Hz,对应的 midi 音符是 69. 一个半音又定义为 100 音分,文中均以音分为最小单位. 具体转换方式如下.

$$f_{\text{cent}}=6900+1200\log_2\frac{f_{\text{Hz}}}{440} \tag{2}$$

2.3 音乐数据集及评价标准

2.3.1 *音乐数据集* 本文实验中使用的音乐数据集是来自国际音乐信息检索评测比赛中主旋律提取专用的 MIR-1K 数据集. 该数据集专为歌声分离研究而设计,包含有 1 000 首歌曲片段,还包含手动记录的半音音高、有声帧、无声帧、歌词等. 歌曲由非专业的 8 位女性和 11 位男性演唱. 因此本文所针对的均为主导旋律为人声的提取.

2.3.2 *评价标准* 评价标准旨在全面体现主旋律提取算法的性能,主要分为以下 5 个指标.

(1) 人声召回率(Voicing Recall Rate, VRR):提取序列中人声帧占标签序列中人声帧的比例.

(2) 人声虚警率(Voicing False Alarm Rate, VFAR):提取序列中将非人声帧误判为人声帧的比例.

(3) 音高准确率(Raw Pitch Accuracy, RPA):标签序列中人声帧的音高与相应帧的提取序列逐帧比较,音高差值小于 50 音分则为正确.

(4) 音色准确率(Raw Chroma Accuracy, RCA):计算方法与 RPA 大致相同,对比时忽略八度错误.

(5) 总体正确率(Overall Accuracy, OA):提取序列与标签序列所有帧逐帧比较,音高差值小于 50 音分则为正确.

3 主旋律提取

本节主要详细介绍了主旋律提取每一步骤的具体过程,其主要流程如图 1 所示.

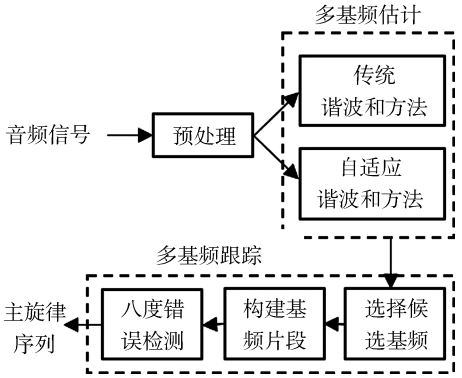

图 1 主旋律提取流程图

Fig. 1 Melody extraction flow chart

3.1　预处理

由于数据集的流行歌曲均为高度混叠的复调音乐，我们很难从中判别出单一的主旋律. 为了降低伴奏带来的影响，文中采用了文献[7]提到的HPSS算法进行预处理. 对于以人声为主旋律的音乐，能有效地筛除掉音乐中平缓的和弦伴奏与节奏感强的击打伴奏，从而增强人声主旋律信号.

3.2　多基频估计

3.2.1　*传统谐波和方法*　谐波理论认为人对声信号的感知是由基频及其一系列谐波共同组成. 在频谱上也可以观察到在信号基频的整数倍处有明显的能量增强，总体呈现梳状结构. 根据这些特点，Hermes 等人[9]提出了谐波和方法.

(1) 将信号降采样后进行短时傅里叶变换(STFT)，得到其频谱 $S(f,t)$；

(2) 计算谐波和：

$$H(f_0,t)=\sum_{n=1}^{N}h^{n-1}S(nf_0,t) \tag{3}$$

其中，h 为压缩因子，取 0.84，作用是使高阶谐波对基频产生的影响更小；N 为最大谐波次数，表示在最大谐波频率范围内出现基频倍数谐波的次数；$H(f_0,t)$为频率在 t 时刻f_0点处的分谐波叠加谱，也称为显著度函数. 在理想范围内取分谐波谱能量最大的频率点作为 t 时刻的基频.

3.2.2　*自适应谐波叠加方法*　经过实验表明，直接由谐波和方法得到的基频序列容易出现半频或倍频错误. 造成该错误主要有两点原因：(1) 是叠加的压缩因子 h 或谐波叠加次数 N 取值不当. 当压缩因子 h 取值过大或者谐波叠加次数 N 过多，低频处能量就会过大，容易出现半频错误；反之则出现倍频错误；(2) 是某一频段处存在较强的伴奏或者噪声. 尽管预处理可以去除部分伴奏和噪声，但残余部分仍会造成基频判别不准确，直接选取能量最大频率点判定基频存在一定误差.

因此，本文提出了一种自适应谐波叠加的方法，根据基频序列的方差特征及其变化趋势自适应改变压缩因子 h. 给定 h 一个初始值为零，根据式(3)计算整体频谱显著度，选取显著度最大频率点作为当前帧基频值，计算所有帧基频序列的方差. 改变 h 的值，若当方差第一次趋于稳定且前后差值小于设定阈值后，则选定此时的 h 为该歌曲谐波叠加的压缩因子. 计算过程如下.

$$f_k=f_0\big|_{H_k=\max(H_k)} \tag{4}$$

$$\mu_k=\sum_{i=1}^{M}f_{ki} \tag{5}$$

$$\sigma_k{}^2=\frac{1}{M}\sum_{i=1}^{M}(f_{ki}-\mu_k)^2 \tag{6}$$

$$h(k+1)=\begin{cases}h(k)+\alpha, & q\sigma_k{}^2>\sigma_{k+1}{}^2\\ h(k), & q\sigma_k{}^2\leqslant\sigma_{k+1}{}^2\end{cases} \tag{7}$$

其中，k 为 h 的迭代次数，M 为该歌曲所有帧数；μ_k 为第 k 次迭代中所有帧的基频序列均值；$\sigma_k{}^2$ 为第 k 次迭代中所有帧的基频序列方差. 步长 α 设置为 0.1，阈值 q 设置为 0.03.

实验表明，大于 5 倍的谐波乘压缩因子后不会对显著函数造成太大影响，且 5 倍以内的谐波和基本包含了所有需要的谐波信息. 因此，将谐波叠加次数 N 的值固定为 5，节约计算成本. 同时基频的选择不再局限于最大能量频率点. 通过分谐波叠加得到整个频谱的显著度函数，在理想基频范围 100～500 Hz 内选出多个候选频率，作为后续处理的输入.

以歌曲 Ani_3_03.wav 为例，图 2 展示了其自适应谐波叠加的显著函数，颜色越亮处说明该频率显著度越高.

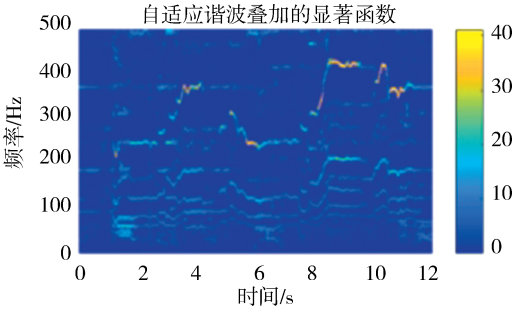

图 2　Ani_3_03.wav 的显著函数

Fig. 2　The saliency function of Ani_3_03.wav

3.3　多音高跟踪

3.3.1　*选择候选基频*　在选取候选基频之前，首先对显著度函数进行处理，只保留峰值点及其邻近两点频率分量，目的是减少非峰值点的干扰，滤除环境噪声. 本文将所有峰值点归为候选基频和补充基频两类. 计算所有峰值点显著度的均值 μ 和标准差 σ，公式如下.

$$\mu=\sum_{i=1}^{P}H_i \tag{8}$$

$$\sigma=\sqrt{(H_i-\mu)^2} \tag{9}$$

其中，P 为该帧峰值点的个数. 将显著度高于 $\mu-\tau\sigma$ 的峰值点归为候选基频，其余峰值点归为补充

基频. 实验表示,τ 取 0.9 时效果最好.

3.3.2 构建基频片段 Justin Salamon 在文献[5]中提到了构建基频片段的方法.

(1) 前后向搜索候选基频,使得一个基频点仅属于一个基频片段,且在时间上连续传递,在频率上平滑变化.

(2) 对所有基频片段的特征进行分类提取,滤除能量、标准差较小的基频片段.

本文在此基础上还计算了能熵比特征[10],并滤除能熵比较小的基频片段. 谱熵反应了声源在频域幅值分布的"无序性",对于噪声和和弦伴奏谱熵较大,能量较小. 计算公式如下.

$$E(n) = \log_{10}\left(1 + \frac{\sum_{n=1}^{N} e(n)^2}{2}\right) \tag{10}$$

$$\mathrm{prob}(n) = \frac{e(n)}{\sum_{n=1}^{N} e(n)} \tag{11}$$

$$H(n) = -\sum_{n=1}^{N} \mathrm{prob}(n)\log_{10}\mathrm{prob}(n) \tag{12}$$

$$E_f = \sqrt{1 + \left|\frac{E(n)}{H(n)}\right|} \tag{13}$$

其中,$E(n)$为基频片段的能量;prob(n)为每个频率分量的归一化谱概率密度函数;$H(n)$为基频片段的谱熵. 以歌曲 Ani_3_03.wav 为例,图 3 展示了通过上述方法构建的基频片段.

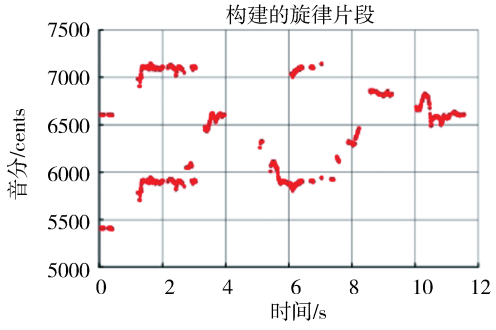

图 3 Ani_3_03.wav 的基频片段

Fig. 3 The fundamental frequency segment of Ani_3_03.wav

3.3.3 八度错误检测 八度错误是指将音高错判为高八度或者低八度的音阶. 对于基频片段,检测及纠正八度错误的步骤如下.

(1) 寻找时间上重合且音高差值在一个八度(1200 音分)左右的两条基频片段;

(2) 按照所有基频片段的能量加权计算每个时间帧的平均音高P_0;

(3) 逐帧计算两条八度错误对音高与P_0的差值,删除差值较大的基频片段.

得到正确基频片段后,按照显著度从大到小排序,每帧取一个基频值,组合得到完整的基频序列. 以歌曲 Ani_3_03.wav 为例,图 4 展示了通过滤除八度错误片段后得到的主旋律序列.

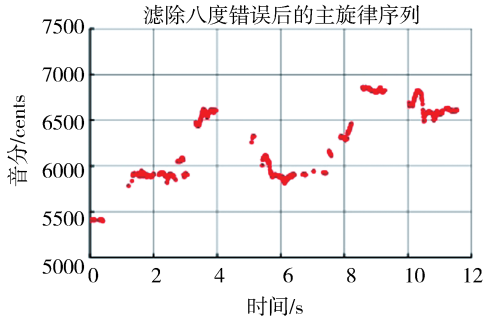

图 4 Ani_3_03.wav 滤除八度错误后的主旋律序列

Fig. 4 The melody sequence of Ani_3_03.wav after eliminating eight-degree errors

3.4 实验结果及分析

本章实验采用 MIR1K 数据集,随机标记 500 首歌曲标记为测试集,另 500 首歌曲将在后续实验标记为训练集使用. 测试歌曲由主旋律与伴奏以 0 dB 的信噪比进行混合,分别计算使用传统谐波和方法与使用自适应谐波叠加方法在测试集上提取主旋律,结果指标参数如表 1 所示.

表 1 传统谐波和方法与自适应谐波叠加方法对比

Tab. 1 Comparison of traditional harmonic sum method and adaptive harmonic superposition method

方法	VRR/%	VFAR/%	RPA/%	RCA/%	OA/%
传统谐波和	81.63	40.79	64.03	64.36	61.52
自适应谐波叠加	78.01	34.36	64.28	64.87	63.63

观察结果可以发现,自适应谐波叠加方法构造的显著函数为后续主旋律提取结果的准确性带来了有效的提升. 但是总体人声召回率较低、虚警率较高导致了整体准确率不佳. 因为数据集中歌曲演唱者是非专业的,并且演唱环境较为嘈杂,人声的能量不够突出,就显著度而言并不占优势,所以在能量筛选的过程中容易出现误判. 因此在后续章节中讨论了人声检测的重要性.

4 人声检测

4.1 声学特征提取

4.1.1 Mel 频率倒谱系数及其 MSDC 系数 人对声音的听觉感知是非线性的,人耳就是一个特殊

的滤波器组，在低频段分布较密，在高频段分布稀疏. 学者根据人耳的特性设计了 Mel 滤波器组来模拟耳蜗模型，并提出了 Mel 频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)[11].

实验中采用的是 24 阶的 Mel 滤波器组，取 2～14 位系数构成 MFCC 特征. 同时取当前帧 MFCC 系数与上一帧 MFCC 系数的差值，加上原始的 13 位 MFCC 系数组成 MSDC 特征，该特征不仅仅局限于当前帧，具有一定的动态性.

4.1.2　*对数频域能量系数*　对数频域能量系数(Log Frequency Power Coefficient, LFPC)[12]取 50～8000 Hz 范围划分成 12 个对数域上等距的子带，代表了子带上能量的分布情况. 计算公式如下.

$$LFPC_t(m) = 10\log_{10}\left[\frac{1}{N_m}\sum_{k\in B_m} X_t^2(k)\right] \quad (14)$$

其中，$X_t^2(k)$是第 t 帧第 k 个频率分量的能量；B_m 是第 m 个子带的频率范围；N_m是该子带内所有频率分量的个数.

4.1.3　*线性预测系数*　在人声中占据大部分能量的都是浊音，而浊音的产生可以等效为单位脉冲序列激励声道管，该过程为线性时不变系统. 一个浊音的采样值可以通过过去若干浊音采样值的线性组合来逼近，在取得最小均方误差时，能够决定唯一的一组线性预测系数(Linear Prediction Coefficient，LPC)[13]. 该特征反应了人声前后时间点的关联性.

4.1.4　*频谱对比度特征*　频谱对比度特征(Spectrum Contrast Features, SCF)[14]将频谱划分为 6 个对数域上等距的子带，计录每个子带内能量峰谷值及其差值. 谱峰主要对应谐波分量，谱谷主要对应非谐波分量和噪声，该特征反应了谐波与非谐波分量的分布情况.

$$P_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x_{k,i}\right) \quad (15)$$

$$V_k = \log\left(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\right) \quad (16)$$

其中，x_k为频谱按照能量降序排列；x'_k为频谱按照能量升序排列；α 为宽度因子，取值为 0.02，表示峰谷值是取附近几点的平均值而定，目的是防止毛刺干扰等.

4.1.5　*频谱形状特征*　频谱形状特征(Spectrum Shape Features, SSF)是通过每帧频谱的形状反应频率分量及能量分布的总体概况. Geoffroy Peeters 在文献[15]中提到可以将频谱形状特征作为判别是否存在人声的依据，并且提出了 8 种特征共同作为一组频谱形状特征向量，包括：谱质心、散度、偏度、峭度、衰减度、滚降频率、谱平坦度、谱突出度.

4.2　随机森林

上世纪八十年代 Breiman 等人发明分类树的算法，实现数据进行分类或回归. 2001 年 Bierman 又把分类树组合成随机森林，即有放回地随机采集多个训练样本，生成多个分类树，每个分类结果都由多个分类树共同投票决定. 比起其他常见的分类方法，如 GMM 分类器、SVM[16]分类器等，随机森林采集部分样本和抽取部分特征寻找最优解，不容易陷入过拟合，对数据适应能力强，且实现简单. 本文通过利用随机森林模型对上述特征进行学习，然后对信号进行分类.

根据 3.4 节实验分类完成的 MIR1K 数据集，将已标记的 500 首训练集中的歌曲及对应的人声标签送入模型进行训练. 每棵决策树通过有放回地选取不同部分特征进行判决，对照人工标注结果训练模型. 训练完成后将另 500 首歌曲送入模型，综合多棵决策树判决结果给出该帧是否属于人声帧的比例，并与人工标注结果进行对比.

4.3　实验结果及分析

4.3.1　*人声特征分类结果*　本实验的目的是验证随机森林模型对不同人声特征组合分类结果的正确性. 实验中树的数量 N 取 100. 表 2 展示了不同特征组合情况下的人声检测分类结果. 最优的特征组合为 MSDC+LFPC，正确率达到了 83.28%.

表 2　不同特征组合下人声检测的结果

Tab. 2　The results of voice detection with different feature combinations

特征组合	VRR/%	VFAR/%	OA/%
MFCC	91.39	50.27	79.92
MSDC	92.01	51.54	79.97
LFPC	90.78	40.60	82.03
LPC	92.52	76.82	72.13
SSF	92.11	65.16	75.22
SCF	88.88	44.43	79.64
MFCC+LPC	92.81	51.04	80.61
MFCC+LFPC	91.74	39.39	83.12
MSDC+LFPC	91.21	37.35	83.28
MFCC+SCF	91.44	44.84	81.44
MFCC+SCF+LFPC	91.66	38.70	83.22

4.3.2 映射主旋律提取结果 将第三节用自适应谐波叠加提取到的主旋律序列,通过 MSDC+LFPC 特征组合得到的随机森林模型,对每一帧信号进行人声检测分类. 根据分类结果将人声帧提取的主旋律保留,非人声帧的主旋律置零. 若人声帧提取的主旋律为零,则提取显著函数中该帧能量最大的频率分量补充到主旋律中.

表 3 是映射主旋律序列的结果. 相比第三章实验结果,总体准确率由 63.63%提升到了 73.25%,召回率明显提升,但虚警率还是较高. 原因是在训练集中人声帧数量远大于非人声帧,生成的随机森林模型分类结果更倾向于人声帧. 为了降低虚警率,我们在之前的映射结果基础上再进行过滤和平滑处理,主要包括过滤能量较小点、删除频率突变点和补充频率缺失点.

表 3 人声检测直接映射主旋律序列结果

Tab. 3 Melody sequences results of voice detection by direct mapping

VRR/%	VFAR/%	RPA/%	RCA/%	OA/%
91.20	37.26	77.07	77.23	73.25

图 5 是歌曲 Ani_3_03.wav 最终提取序列与标签序列对比结果. 为了方便比较,图中将标签序列人为降低了 500 音分呈现. 从图中不难看出,提取的主旋律序列与标签序列基本一致,表明了提出方法的有效性. 表 4 展示了数据集所有音乐主旋律最终提取结果,并与 2018 年 MIREX 主旋律提取的算法 KN3[17]、LS1[18]进行对比. 实验分别采用主旋律与伴奏声以 0、−5、5 dB 三种不同信噪比情况进行混合提取.

在 0 dB 情况下经过平滑过滤后的总体正确率从 73.25%提升到了 76.20%. 通过结合机器学习的方法,将声学特征运用到了人声检测,能有效地减小虚警率. 在高信噪比的条件下,人声与伴奏的特征差异更加明显,有助于分类结果的准确性. 此时的虚警率达到最低 7.09%,总体准确率达到最高 85.04%.

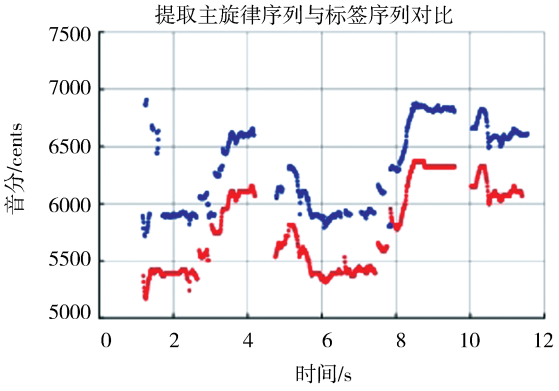

图 5 Ani_3_03.wav 的提取旋律与标签序列对比

Fig. 5 Comparison of melodyextracted from Ani_3_03.wav and label sequences

表 4 本文方法与其它方法结果对比

Tab. 4 Comparison between the results of this method and other method

方法	信噪比/dB	VRR/%	VFAR/%	RPA/%	RCA/%	OA/%
本文方法	0	84.61	18.91	74.78	75.20	76.20
KN3	0	84.76	20.57	74.81	75.72	76.48
LS1	0	89.45	30.60	73.60	74.14	72.19
本文方法	−5	79.49	40.38	54.77	55.28	55.61
KN3	−5	76.80	30.24	61.51	63.22	64.37
LS1	−5	68.30	24.39	51.25	52.24	59.33
本文方法	5	87.18	7.09	82.73	83.09	85.04
KN3	5	88.66	12.77	80.71	81.21	83.09
LS1	5	97.14	34.65	83.63	83.85	77.56

5 结 论

主旋律提取是音乐信号处理的一大重要分支. 本文提出了一种自适应谐波叠加方法构建显著函数,并从中得到候选基频组成了基频片段. 对多种声学特征组合进行分析,结合了随机森林模型进行人声检测. 实验表明在高信噪比情况下主导旋律为人声的音乐集上取得更好效果. 因此,本文的主旋

律提取方法对后续旋律发展、音乐分类、音乐合成等具有一定的借鉴意义.

参考文献：

[1] Pease F, Pease T, Faculty B. Jazz composition: theory and practice [C]//Proceedings of the Consumer Communications and Networking Conference. Las Vegas, USA: IEEE, 2004.

[2] 智鹏鹏，杨鸿武，宋南. 利用说话人自适应实现基于 DNN 的情感语音合成[J]. 重庆邮电大学学报：自然科学版，2018，30：673.

[3] 马彪，李千目. 基于信息差分保护的邻域推荐方法[J]. 江苏大学学报：自然科学版，2019，40：439.

[4] 徐媛媛，何培宇，陈杰梅. 一种基于 IFS 分形算法和分解和弦的耳鸣康复音合成新方法 [J]. 四川大学学报：自然科学版，2017，54：517.

[5] Goto M. A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals[J]. Speech Commun, 2004, 43: 311.

[6] Salamon J, Gomez E. Melody extraction from polyphonic music signals using pitch contour characteristics [J]. IEEE TAudio, Spee, 2012, 20: 1759.

[7] Ono N, Miyamoto K, Roux J L, *et al*. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram [C]//Proceedings of the European Signal Processing Conference. [S. l.]: [s. n.], 2008.

[8] Breiman L. Random forests [J]. Mach Learn, 2001, 45: 5.

[9] Dik J H. Measurement of pitch by subharmonic summation [J]. J Acoust Soc Am, 1988, 83: 257.

[10] 张毅，王可佳，席兵，等. 基于子带能熵比的语音端点检测算法[J]. 计算机科学，2017，44：304.

[11] Rabiner L, Schafer R. 数字语音处理理论与应用[M]. 北京：电子工业出版社，2016.

[12] Nwe T L. Automatic detection of vocal segments in popular songs [C]//Proceedings of the International Society for Music Information Retrieval. Barcelona, Spain: ISMIR, 2004.

[13] 国林，杨武，王巍，等. 数据通信基础[M]. 北京：清华大学出版社，2006.

[14] Jiang D N, Lu L, Zhang H J, *et al*. Music type classification by spectral contrast feature [C]// Proceedings of the IEEE International Conference on Multimedia & Expo. Lausanne, Switzerland: IEEE, 2002.

[15] Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. [C]//Proceedings of the Project Report. Paris, France: IRCAM, 2004.

[16] 郭凯文，潘宏亮，侯阿临. 基于特征选择和聚类的分类算法[J]. 吉林大学学报：理学版，2018，56：395.

[17] Kum S, Nam J. Classification-based singing melody extraction using deep convolutional neural networks [C]//Proceedings of the International Society for Music Information Retrieval. Suzhou, China: ISMIR, 2017.

[18] Su L, Yang Y H. Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music [J]. IEEE/ACM Taudio Spee, 2015, 23: 1600.

引用本文格式：

中　文：何甜田，何培宇，陈杰梅. 一种自适应谐波叠加的复调音乐主旋律提取新方法[J]. 四川大学学报：自然科学版，2020，57：519.

英　文：He T T, He P Y, Chen J M. A new method of melody extraction for polyphonic music based on adaptive harmonic superposition [J]. J Sichuan Univ: Nat Sci Ed, 2020, 57: 519.