文章编号：2095-4980(2019)03-0482-07

声乐主旋律的自动提取

陆　雄，夏秀渝*，蔡　良，孙文慧

(四川大学 电子信息学院，四川 成都 610065)

摘　要：提出一种基于多候选基频提取和歌声基频判别的声乐主旋律提取算法。该算法可以有效降低旋律定位虚警率，提高整体准确率。利用度量距离(DIS)算法对音乐进行音符切分，并用方差法实现浊音段检测；采用幅度压缩基音估计滤波器(PEFAC)多基频提取技术，通过计算音高显著度提取每个浊音帧的多个候选基频。最后用维特比算法跟踪浊音段主导基频轨迹，并用基频判别模型进行歌声主旋律判别。在MIR-1K数据集上进行的实验表明，在信干比为5 dB和0 dB的情况下，本文算法提取的声乐主旋律整体准确率分别达到了86.22%和77.4%，相比于其他算法至少提高了3.79%和2.01%。

关键词：主旋律；音符切分；维特比算法；基频判别模型

中图分类号：TN911.72　　**文献标志码：**A　　**doi：**10.11805/TKYDA201903.0482

Automatic extraction of vocal music theme

LU Xiong, XIA Xiuyu*, CAI Liang, SUN Wenhui

(College of Electronic and Information Engineering, Sichuan University, Chengdu Sichuan 610065, China)

Abstract: This paper presents a vocal themes extraction algorithm based on multi-candidate fundamental frequency extraction and singing voice fundamental frequency discrimination. The algorithm can effectively reduce the voicing false alarm rate and improve the overall accuracy. First, using the Distance(DIS) metric distance algorithm to achieve note segmentation, and using the variance method to detect voiced segments. Then Pitch Estimation Filter with Amplitude Compression(PEFAC) multi-fundamental frequency extraction technology is utilized to extract multiple candidate fundamental frequencies of each voiced frame by calculating the pitch saliency. Finally, the dominant fundamental frequency trajectory of the voiced segment is tracked by the Viterbi algorithm, and the main melody of the singing voice is determined by the fundamental frequency discrimination model. Experiments conducted on the MIR-1K dataset show that the overall accuracies of the vocal themes extracted by the proposed algorithm reach 86.22% and 77.4%, respectively, at the signal to interference ratio of 5 dB and 0 dB, which are increased by at least 3.79% and 2.01% respectively compared to other algorithms.

Keywords: theme; note segmentation; Viterbi algorithm; fundamental frequency discrimination model

旋律是最重要的音乐要素，由不同音高和时值的单音连续构成。主旋律可分为声乐主旋律和通用主旋律[1]。如果复调音乐中包含歌声，则歌声的音高序列就认为是声乐主旋律；如果不存在歌声，则能量占主导地位的乐器演奏音的音高序列作为器乐主旋律。从物理学角度，音高由基频决定。本文声乐主旋律自动提取是指提取复调音乐中歌声的基频并将其连成旋律线。声乐主旋律自动提取在很多领域都有应用，如音乐结构分析、音乐检索、音乐风格分类、翻唱识别等，也可作为推导其他语义标签的中间步骤。声乐主旋律自动提取面临以下两点挑战[2]：a) 一首音乐由歌声和各种乐器伴奏混合而成，不同声源的频谱相互重叠，很难将某个频率分量归于单个声源。b) 当得到音乐的音高时，仍需要决定哪个音高值属于歌声旋律而不是伴奏。声乐主旋律自动提取是一项非常有意义和挑战性的研究工作。

近二十年来，主旋律自动提取方法的研究取得了显著进步。Masataka Goto[3]是第一个从现实的CD录音中提

收稿日期：2018-04-23；修回日期：2018-07-15

*通信作者：夏秀渝　email:xiaxxy@163.com

取出旋律线和低音线的人，提出了著名主导基频估计法(Predominant-F0 Estimation Method，PreFEst)。该算法不依赖于基频F0的不可靠频率分量，而是在特定的频率范围内通过其谐波获得最显著的F0。Karin Dressler[4]在其主旋律自动提取算法中引入了基于规则的主旋律线识别方法，有效降低了旋律定位虚警率，提高了整体准确率。Yukara Ikemiya等[5]提出一种基于语音分离的主旋律提取方法。实验表明引入了语音分离的主旋律提取算法性能均要高于未引入的算法。Sangeun Kum等[6]用基于数据驱动的方法，用充足的带标签数据为每个基频标签训练了一个深度神经网络模型。由于采用了简单的歌声旋律定位算法，该算法的整体准确率略低于其他算法。但是该算法的原始音高准确率相比于其他算法有所提高，所以只要加入更有效的歌声旋律定位算法，整体准确率就会显著提高。Juan J Bosch等[7]结合了源–滤波器模型和谐波加权两种显著度函数计算方法，根据旋律的连续性提取多条候选旋律线，最后根据候选旋律线的一系列特征进行歌声旋律定位和主旋律提取。通过以上研究现状，发现对音高显著度函数算法的研究贯穿了整个主旋律自动提取算法的研究进程,而对歌声旋律定位算法的研究近几年才引起学者们的关注，尚未达到性能上限。实验表明，引入简单歌声旋律定位算法对提升整体准确率有帮助，进一步提高歌声旋律定位算法的准确度将有利于提高主旋律自动提取算法的整体准确率。

结合前人研究的优势与不足，本文在基于音高显著度函数的旋律提取框架下，针对性地引入音符切分和基频判别模型。基于音符的时间持续性加入基于频谱距离的音符切分，将歌曲切分成一个个频谱相对稳定的段落，有利于旋律跟踪和歌声旋律定位。在歌声旋律定位部分基于歌声和伴奏的音色不同，加入采用神经网络的基频判别模型，基于分段统计主导基频轨迹属于歌声旋律的概率，可以有效降低旋律定位虚警率，提高整体准确率。

1　算法原理

1.1　算法总体框架

声乐主旋律自动提取的总体框架如图 1 所示。整个算法由 4 部分组成，第一部分进行音频预处理、音符切分和浊音段检测，首先对输入的音乐信号归一化、分帧、加窗和时频域变换，然后进行音符切分和浊音(特指有谐波结构的音频)段检测。第二部分进行多候选基频提取，首先利用优化的梳齿滤波器计算浊音段音频的音高显著度函数，然后每帧提取出多个候选基频。第三部分利用维特比算法在每个浊音段进行主导基频轨迹跟踪。第四部分是歌声主旋律判别，利用训练好的基频判别模型统计判断各段主导基频轨迹是否为歌声旋律，是，则保留；不是，则判为伴奏旋律，最后将各段的歌声旋律连接成为声乐主旋律。

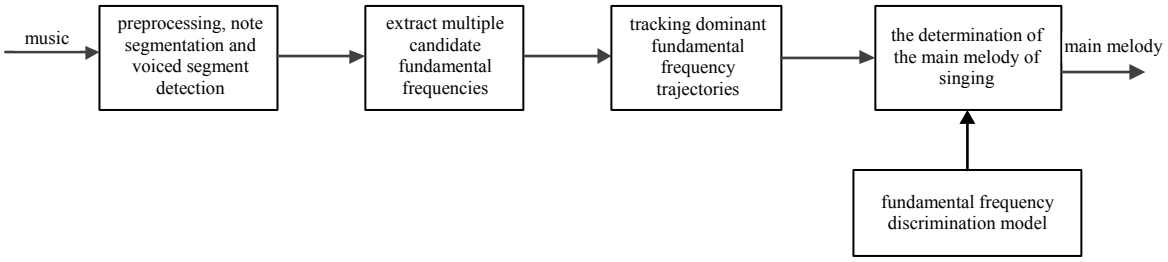

Fig.1 Overall framework of the algorithm
图 1　算法总体框架

1.2　预处理、音符切分及浊音段检测

预处理包括音频的降采样、归一化、分帧、加窗和时频域变换等。通常音乐中高于 4 kHz 的人声谐波分量占比相对很小，所以对原始音乐信号降采样至 8 kHz，这样也可减少后续处理的计算量。音频信号是短时平稳的，还要对音频进行分帧加窗操作。本文采用汉明窗，每帧信号取 320 个样点。本文采用短时傅里叶变换对信号进行时频变换。

音乐由歌声和伴奏组成，它们均由一个个有一定时值的音符组成，每个音符具有相对稳定的频谱特征，反映在语谱图上就是一系列频谱段——其段内差异小，段间差异大，本文采用文献[8]提出的度量距离(DIS)算法来分割音符。DIS 距离是一种综合了数据段段间均值和方差的距离度量方法，用来表征音频段落之间的差距。由于本文前后 2 个数据窗长均取为 5 帧，所以 DIS 度量距离可以简写为式(1)。

$$DIS=\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{tr(\sum_1)+tr(\sum_2)} \tag{1}$$

式中：$\boldsymbol{\mu}_1$和$\boldsymbol{\mu}_2$分别表示前后两段音频特征的均值矢量，$tr\left(\sum_1\right)$和$tr\left(\sum_2\right)$分别表示前后两段音频特征协方差矩阵的迹。当两段音频段之间特征均值差异较大，段内特征方差较小时，DIS 越大，表明两段音频段间距离越大。

本文特征参数采用短时幅度谱。通过按帧滑动数据窗，计算得到关于帧数 t 的 DIS 距离函数 $DIS(t)$：

$$DIS(t)=\frac{\left(\boldsymbol{\mu}_{t,1}-\boldsymbol{\mu}_{t,2}\right)^{\mathrm{T}}\left(\boldsymbol{\mu}_{t,1}-\boldsymbol{\mu}_{t,2}\right)}{tr\left(\sum_{t,1}\right)+tr\left(\sum_{t,2}\right)} \tag{2}$$

式中：$\boldsymbol{\mu}_{t,1}$和$\boldsymbol{\mu}_{t,2}$分别表示 t 帧前后两段音频特征的均值矢量；$tr\left(\sum_{t,1}\right)$和$tr\left(\sum_{t,2}\right)$分别表示 t 帧前后两段音频特征协方差矩阵的迹。

寻找 $DIS(t)$ 中所有极大值点，设置阈值 T_1 为 $DIS(t)$ 的均值，删除小于阈值 T_1 的极大值点。此外快节奏的音乐中四分音符的持续时间大约是 0.5 s，考虑到八分音符、十六分音符的持续时间是四分音符的 1/2 和 1/4，本文设段距不小于 100 ms，否则就去除对应的极大值点，这样剩下的极大值点就是音符切分点。

由于音乐中不仅存在浊音段，还存在非浊音段，所以在切分后用浊音段检测算法标记浊音段和非浊音段。由于浊音段相较于非浊音段频谱方差更大，本文采用频谱方差作为浊音段检测的特征参数。

具体步骤如下：

1) 计算每帧频谱的方差；

2) 设置一个阈值 T_2，在每个分段上统计方差小于阈值 T_2 的帧数 f_n；

3) 如果 f_n 大于分段帧数的一半，则判定该段为非浊音段；否则，判定为浊音段。

1.3 多候选基频提取

幅度压缩基音估计滤波器(PEFAC)是一种鲁棒的多基频提取方法[9-10]，本文利用该算法通过对数域梳齿滤波器与对数域频谱相卷积计算音高显著度函数。

一个基频为 f_0 的浊音信号，它的频域表达式为：

$$Y(f)=\sum_{k=1}^{K}a_k\delta(f-kf_0) \tag{3}$$

式中 a_k 为 k 次谐波的系数。

对数频域可表示为：

$$Y(q)=\sum_{k=1}^{K}a_k\delta(q-\log k-\log f_0) \tag{4}$$

式中 $q=\log f$。在对数频域里，谐波间的间隔与 f_0 无关，因此当它与一个冲激响应如式(5)的滤波器卷积时，

$$h(q)=\sum_{k=1}^{K}\delta(q-\log k) \tag{5}$$

卷积结果 $Y(q)*h(-q)$ 将会在 $q_0=\log f_0$ 的位置产生一个峰，由峰的位置即可确定浊音基频。式(5)是一个由很多 δ 函数组成的理想滤波器，而在实际中由于加窗分析，谐波峰的宽度会展宽，所以实际采用的滤波器如式(6)所示：

$$h(q)=\frac{1}{\gamma-\cos\left(2\pi\exp(q)\right)}-\beta \tag{6}$$

式中：参数 γ 控制着峰的宽度；选取参数 β 使得 $\int h(q)\mathrm{d}q=0$；$\log(0.5)<q<\log(K+0.5)$，否则 $h(q)=0$，K 代表峰的数量。实验表明，该滤波器可以抑制频谱较平滑的噪声。

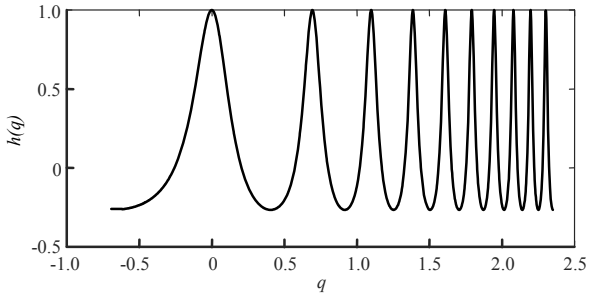

Fig.2 Comb filter
图 2 梳齿滤波器

选取合适的滤波器参数，得到如图 2 所示梳齿滤波器。用该梳齿滤波器与某帧信号对数域频谱相卷积，得到该帧信号的基频显著度函数：

$$S(q)=Y(q)*h(-q) \tag{7}$$

找出该函数中所有峰值对应的频率集合$\{\psi_1\}$作为候选基频。由于人歌唱的音域有限，经统计人声基频一般在 70~1 000 Hz 范围内。将$\{\psi_1\}$中超出此范围的峰值频率全部删去，形成集合$\{\psi_2\}$；然后去除八度错误，利用歌声基频或伴奏主基频对应的显著度值必然大于其半频

和倍频处显著度值的特点，在$\{\psi_2\}$中提取出一个最大显著度值的频率放入集合$\{\psi_3\}$，并将$\{\psi_2\}$中剩下的频率中与此频率成倍频或半频关系的频率删去，重复上述过程直至$\{\psi_2\}=\varnothing$。最后从$\{\psi_3\}$中筛选出 3 个显著度最大的作为最终候选基频集合$\{\psi\}$。

1.4 主导基频轨迹跟踪

提取多候选基频之后，运用维特比算法[11]在段内(每个音符内)进行主导基频轨迹跟踪。本文维特比算法采用音高似然度和音高转移概率，音高似然度定义为：

$$p(f_m \mid \boldsymbol{X}_t)=\frac{s_{t,m}}{\sum_n s_{t,n}} \tag{8}$$

式中：$\boldsymbol{X}_t$为第 t 帧的幅度谱；$s_{t,m}$为 t 帧的第 m 个候选基频 f_m时的音高显著度值；$\sum_n s_{t,n}$为第 t 帧所有候选基频的音高显著度值求和。

文中音高转移概率由带标注的音乐库统计得到。定义相邻帧音高变化率 Δf 为：

$$\Delta f=\frac{f_t-f_{t-1}}{f_{t-1}} \tag{9}$$

式中：f_t为当前帧的音高；f_{t-1}为前一帧的音高。统计库中所有音乐的段内(音符内) Δf 的概率分布并归一化作为音高转移概率，记为 $p(\Delta f)$。

为防止因某帧目标基频丢失出现无法跟踪正确主导基频的情况，将统计得到的音高转移概率做适当修改。将音高转移概率小于 0.001 的概率值均统一设置为 0.001。修正的音高转移概率如图 3 所示。

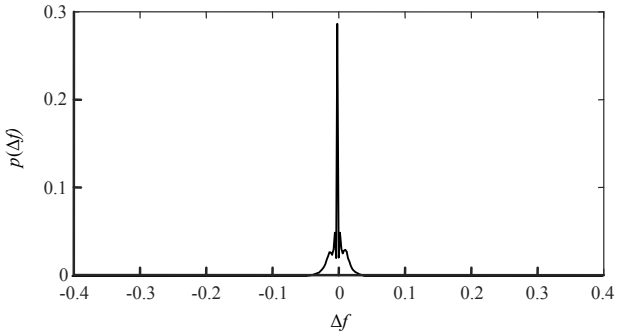

Fig.3 Pitch shift probability
图 3 音高转移概率

每个浊音段采用维特比算法提取该段的最优基音轨迹(主导基频轨迹)，主导基频轨迹应满足式(10)：

$$F=\underset{(f_1,f_2,\cdots,f_T)}{\arg\max}\left\{\alpha\sum_{t=1}^{T}\lg p(f_t \mid \boldsymbol{X}_t)+\beta\sum_{t=2}^{T}\lg p(\Delta f_t)\right\} \tag{10}$$

式中：T 为分段长度；$p(f_t \mid \boldsymbol{X}_t)$为音高似然度；$p(\Delta f_t)$为音高转移概率；$\alpha$ 和 β 为权值。

相比于传统的维特比搜索方法，本文只对每一帧筛选出来的少量候选基频进行维特比跟踪，极大地减少了计算量。

1.5 歌声主旋律判别

提取出每个音符段落(浊音段)主导基频轨迹后，接下来判断该主导基频轨迹是属于歌声还是属于伴奏，删除不属于歌声主旋律的基频轨迹。基于此目的，该部分加入了基频判别模型。

歌声、器乐具有不同的音色，研究表明音色主要决定于声音的频谱，即由声音基音和各次谐音的相对强度决定。梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients，MFCC)[12]考虑到了人类的听觉特性，是一种能够反映声音谱包络特征的参数。基于 BP 神经网络[13]建立基频判别模型。

另外歌声、器乐声都具有谐波结构，所以混合音乐频谱具有近似稀疏性，可以根据主导基频利用梳齿滤波器提取对应声源的谐波谱，由提取信号的 MFCC 送入神经网络判断对应基频是否为歌声。

主旋律判别实现步骤：

1) 由主导基频 F_0 构造出一个频率范围为 0~4 kHz 的梳齿滤波器，如式(11)所示。

$$h(f)=\sum_{k=1}^{K}\delta(f-kF_0)*b(f) \tag{11}$$

式中：K 为 0~4 kHz 范围内谐波的个数；$b(f)$为梳齿滤波器的基本波形(本文采用矩形)。

2) 用梳齿滤波器对信号幅度谱滤波，得到 F_0 对应的谐波谱，并提取其对应的 MFCC 参数。

3) 将 MFCC 送入神经网络识别，判断该主导基频 F_0 是否为歌声基频。

4) 在每个浊音段统计歌声基频的帧数，如果大于该浊音段总帧数的一半，则判定该浊音段的主导基频轨迹为歌声主旋律。

本文神经网络结构为 12–30–12–2，即输入层 12 个神经元，中间隐藏层分别为 30 和 12 个神经元，输出层 2 个神经元，分别代表歌声基频和器乐基频。

为保证模型训练和识别时输入参数的一致性，提取训练数据中歌声和器乐的 MFCC 参数时采用了和识别时类似的操作，即先根据歌声或器乐的基频用梳齿滤波器提取出对应的谐波谱，再计算其 MFCC 参数，并训练神经网络。

2 实验和分析

2.1 实验数据

音乐数据采用 MIR-1K[14]数据集，该数据集包含 1 000 段采样率为 16 kHz，歌声、伴奏可分的音乐，包含时间间隔为 10 ms 的歌声基频标签。本文实验从 MIR-1K 数据集中随机选出 500 段音乐作为训练集,剩余的 500段音乐作为主旋律提取测试集。

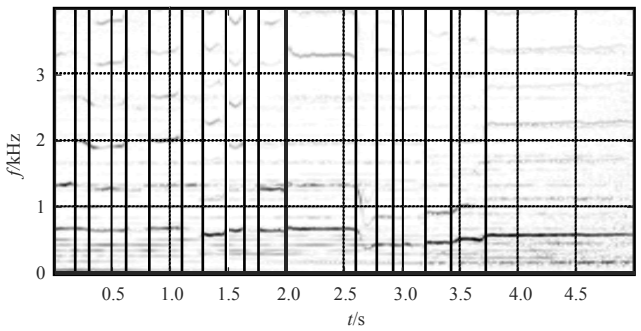

Fig.4 Segmentation effect of a piece of music
图 4 一段音乐的切分效果

2.2 音符切分实验

利用 DIS 度量距离算法对音乐进行音符切分。一段音乐的切分效果如图 4 所示。

以手工分割标注点为基准，分析自动音符分割的效果。采用音符分割准确率 α 作为评价指标。

$$\alpha = \frac{N_{\mathrm{DIS|REL}}}{N_{\mathrm{REL}}} \tag{12}$$

式中：$N_{\mathrm{DIS|REL}}$ 为相邻 2 帧内存在自动分割点的手工分割标注点的数目；N_{REL} 为手动分割标注点的数目。

在测试歌曲总时间为 1 125 s 的测试数据上，统计得到音符分割准确率 α 达到了 92.3%，表明 DIS 度量距离算法能有效地实现音符切分。

2.3 基频判别模型训练与测试实验

为训练基频判别模型，本文从 MIR-1K 数据集中随机选出 500 段音乐作为训练集。为保证模型训练和后续识别时特征输入的一致性，训练数据在提取 MFCC 时采用了后续识别时类似的处理步骤，即先由歌声或伴奏的主基频通过梳齿滤波器提取出对应声源的谐波谱，然后再提取其 MFCC 参数用于训练。总共提取了 274 354 帧训练数据，其中歌声主基频对应的训练数据有 142 483 帧，伴奏主基频对应的训练数据有 131 871 帧。用上述训练数据对 BP 神经网络进行训练得到基频判别模型。

为了验证该模型的有效性，用从测试数据集中提取的 317 278 帧 MFCC 参数及对应标签对该模型进行测试，以识别准确率 β 作为模型的评价指标。

$$\beta = \frac{N_1}{N} \tag{13}$$

式中：N_1 为判别正确的帧数；N 为测试数据总帧数。实验表明该模型识别准确率达到了 80.1%，可用于基于统计的歌声主旋律判别，并有助于提高算法的整体准确率。

2.4 声乐主旋律提取实验

为验证声乐主旋律自动提取算法的准确率，本节用测试集中 500 段音乐作为测试数据，在信干比分别为 0 dB 和 5 dB 的情况下进行实验。此处信干比定义为：

$$SIR = 10\lg\left(E_1/E_2\right) \tag{14}$$

式中：E_1 表示歌声的能量；E_2 表示伴奏的能量。

采用以下 5 种性能指标评价算法性能[15]。

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

2) 旋律定位虚警率(Voicing False Alarm Rate，VFAR)：算法错误地估计为旋律的帧数量/基准数据中被标记为非旋律的帧数量。

3) 原始音高准确率(Raw Pitch Accuracy，RPA)：在基准数据的旋律帧中，算法正确估计音高(检测音高与基

准音高的差别在半个半音之内)的帧的比例。

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

5) 整体准确率(Overall Accuracy，OA)：算法正确估计所有帧的比例，即算法将非旋律帧标记为非旋律，将旋律帧标记为旋律并对于旋律帧检测音高与基准音高的差别在半个半音之内。该度量给出算法的整体性能评分。

为进行实验对比，参考 2016 年 MIREX 中复调音乐主旋律提取比赛的结果[16]，信干比为 5 dB 和 0 dB 的结果分别见表 1 和表 2。

表 1 信干比为 5 dB 时的实验结果(%)

Table1 Experimental results with a SIR of 5 dB(%)

method	VRR	VFAR	RPA	RCA	OA
KON1	94.85	34.66	87.01	87.62	79.26
WFJY1	93.54	27.32	91.71	92.44	82.43
IIY1	96.23	33.25	87.92	88.50	79.35
BG2	82.02	34.19	70.83	77.58	66.72
proposed algorithm	90.62	15.60	86.91	86.99	86.22

表 2 信干比为 0 dB 时的实验结果(%)

Table2 Experimental results with a SIR of 0 dB(%)

method	VRR	VFAR	RPA	RCA	OA
KON1	94.96	52.5	81.83	83.27	70.05
WFJY1	84.73	26.68	86.20	87.60	75.39
IIY1	96.57	55.27	82.00	83.22	67.99
BG2	78.18	40.69	61.18	68.56	58.47
proposed algorithm	79.43	14.30	73.29	73.52	77.40

本文在多候选基频提取阶段，加入了去除八度错误的算法。原始色度准确率减去原始音高准确率可以反映算法八度错误的占比情况，根据表 1、表 2，可以算出在信干比为 5 dB 和 0 dB 时的占比情况分别为 0.08%和 0.23%，相比其他算法略有减少。

由于本文引入了基频判别模型，且运用统计方法来判断每个浊音段的主导基频轨迹是否属于歌声主旋律，所以在极少数情况下会出现将歌声旋律段误判为伴奏旋律段，降低了旋律定位查全率，但是在大多数情况下，不会将伴奏段旋律判断为歌声段旋律，降低了旋律定位虚警率，对提升算法的整体准确率有所帮助。根据表 1、表 2，本文方法的旋律定位查全率低于某些方法，但是其旋律定位虚警率却明显小于其他方法。

整体准确率是最重要的评价指标，本文提出的算法在信干比为 5 dB 和 0 dB 时，其整体准确率分别达到了 86.22%和 77.4%，相比其他算法分别至少提高了 3.79%和 2.01%。

3　结论

本文提出一种基于多候选基频提取和歌声基频判别的声乐主旋律提取算法。相比于传统方法，本文针对性地引入了基于音符持续性特点的音符切分算法和基于歌声乐器音色特点的基频判别模型。其中音符切分算法有利于主导基频轨迹跟踪和歌声旋律定位的准确性，基频判别模型在音符切分的基础上能准确地判断主导基频轨迹是否属于歌声主旋律。实验表明本文主旋律提取算法在信干比为 5 dB 和 0 dB 的情况下，有效地减少了八度错误，旋律定位虚警率明显低于其他算法，整体准确率均高于其他算法，能有效提取声乐主旋律。

参考文献：

[1] 张维维,陈喆,殷福亮,等. 复调音乐主旋律提取方法综述[J]. 电子学报, 2017,45(4):1000-1011. (ZHANG Weiwei, CHEN Zhe,YIN Fuliang,et al. Review on melody extraction from polyphonic music[J]. Acta Electronica Sinica, 2017,45(4): 1000-1011.)

[2] 李伟,冯相宜,吴益明,等. 流行音乐主旋律提取技术综述[J]. 计算机科学, 2017,44(5):1-5. (LI Wei,FENG Xiangyi,WU Yiming,et al. Review on main melody extraction pop music[J]. Computer Science, 2017,44(5):1-5.)

[3] GOTO M. A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings[C]// 2000 IEEE International Conference on Acoustics,Speech,and Signal Processing. Istanbul,Turkey:IEEE, 2000:757-760.

[4] DRESSLER K. Extraction of the melody pitch contour from polyphonic audio[C]// International Symposium/Conference on Music Information Retrieval. London,UK:MIREX, 2005:423-428.

[5] IKEMIYA Y,YOSHII K, ITOYAMA K. Singing voice analysis and editing based on mutually dependent F0 estimation and source separation[C]// 2015 IEEE International Conference on Acoustics,Speech and Signal Processing. Brisbane,QLD, Australia:IEEE, 2015:574-578.

[6] KUM S,OH C,NAM J. Melody extraction on vocal segments using multi-column deep neural networks[C]// International Society for Music Information Retrieval Conference. New York,USA:[s.n.], 2016:819-825.
[7] BOSCH J,GÓMEZ E. Melody extraction in symphonic classical music:a comparative study of mutual agreement between humans and algorithms[C]// 2014 9th Conference on Interdisciplinary Musicology. Berlin,Germany:[s.n.], 2014:212-216.
[8] 孙卫国,夏秀渝,乔立能,等. 面向音频检索的音频分割和标注研究[J]. 微型机与应用, 2017,36(5):38-41. (SUN Weiguo, XIA Xiuyu,QIAO Lineng,et al. Research on audio segmentation and annotation for audio retrieval[J]. Microcomputer & Its Applications, 2017,36(5):38-41.)
[9] GONZALEZ S,BROOKES M. A pitch estimation filter robust to high levels of noise(PEFAC)[C]// 2011 19th European Signal Processing Conference. Barcelona,Spain:IEEE, 2011:451-455.
[10] GONZALEZ S,BROOKES M. PEFAC-a pitch estimation algorithm robust to high levels of noise[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014,22(2):518-530.
[11] 韩纪庆,郑铁然,郑贵滨. 音频信息检索理论与技术[M]. 北京:科学出版社, 2011. (HAN Jiqing,ZHENG Tieran,ZHENG Guibing. Audio information retrieval theory and technology[M]. Beijing:Science Press, 2011.)
[12] 刘加,张卫强. 数字语音处理理论与应用[M]. 北京:电子工业出版社, 2016. (LIU Jia,ZHANG Weiqiang. Theory and applications of digital speech processing[M]. Beijing:Publishing House of Electronics Industry, 2016.)
[13] 黄尚晴,赵志勇,孙立波. BP 神经网络算法改进[J]. 科技创新导报, 2017,14(20):146-147. (HUANG Shangqing,ZHAO Zhiyong,SUN Libo. Improvement of BP neural network algorithm[J]. Science and Technology Innovation Herald, 2017, 14(20):146-147.)
[14] HSU C L,JANG J S R. MIR-1K Dataset[EB/OL]. [2018-01-10]. http://sites.google.com/site/unvoicedsoundseperation/mir-1k. 2009.7.22.
[15] SALAMON J,GOMEZ E,ELLIS D P W,et al. Melody extraction from polyphonic music signals:approaches, applications, and challenges[J]. Signal Processing Magazine IEEE, 2014,31(2):118-134.
[16] BOSCH J J,WANG C C. Audio melody extraction results[EB/OL]. [2018-01-10]. http://nema.lisillinois.edu/ nema_out/mirex 2016/results/Ame/mrx09_p5db/.2016.7.28.

作者简介：

陆　雄(1994-)，男，江苏省泰州市人，在读硕士研究生，主要研究方向为音乐信息检索. email:1215234642@qq.com.

夏秀渝(1970-)，女，四川省广元市人，副教授，主要研究方向为自适应声回波对消、语音增强、计算机听觉场景分析、音乐信息检索.

蔡　良(1992-)，男，湖北省黄冈市人，在读硕士研究生，主要研究方向为语音增强.

孙文慧(1991-)，女，山东省莒南县人，在读硕士研究生，主要研究方向为音乐信息检索.