计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200022

一种基于音高显著性增强的主旋律提取方法

金文清　韩　芳
东华大学信息科学与技术学院　上海 201620
(jwq_wem@163.com)

摘　要　在音乐信息检索领域,主旋律的提取是一项非常困难的工作。复调音乐中的不同声源相互影响,导致主旋律音高序列不连续,使旋律原始音高准确率降低。针对这一问题,设计了增强音高显著性表示和自动旋律跟踪的CNN-CRF模型。为了更好地提取谐波信息,提出利用结构化的数据来加强SF-NMF计算的初始显著性表示,并在动态规划框架下结合旋律特征和音高的平滑约束条件在音高空间寻找最优的演变路径。实验表明,所提方法得到了较好的旋律提取结果,且在两个测试数据集上的原始音高准确率均高于其他参考方法,通过对比不同输入验证了结构化数据能加强显著性表示并弥补SF-NMF对音高的误判。

关键词:主旋律提取;音乐信号处理;音高显著性增强;CNN-CRF;音乐信息检索
中图法分类号　TP391

Main Melody Extraction Method Based on Saliency Enhancement

JIN Wen-qing and HAN Fang
School of Information Science and Technology, Donghua University, Shanghai 201620, China

Abstract　In the field of music information retrieval, the extraction of the main melody is a very difficult task. In the polyphonic music, different sound sources interact with each other, leading to discontinuity of the main melody's pitch sequence, which reduces the accuracy of the original pitch of the melody. In response to this problem, a CNN-CRF model with enhanced pitch saliency representation and automatic melody tracking is designed. In order to better extract the harmonic information, it is proposed to enhance the initial saliency representation of the SF-NMF calculation by structured data, and to combine the melody characteristics and the smooth constraint conditions of the pitch under the dynamic programming framework to find the optimal evolution path. Experiments show that the proposed method has better melody extraction results, and the original pitch accuracy on both test data sets is higher than that of other reference methods. Comparing different inputs validates that structured data can enhance the significance representation and make up for the misjudgement of pitch by SF-NMF.

Keywords　Melody extraction, Music signal processing, Pitch saliency enhancement, CNN-CRF, Music information retrieval

1　引言

互联网的兴起和计算机技术的发展,使得数字音乐的创作和传播更加便捷。海量的音乐数据再也无法依靠人工进行管理,需要新的方法对音乐进行检索、分类、推荐和分析,因此音乐信息检索(Music Information Retrieval, MIR)成为了热门的研究领域,并迅速得到了发展。主旋律提取是音乐信息检索领域的一个热点和难点任务。旋律作为音乐的本质,在理解音乐语义学和区分不同音乐作品中起着重要的作用。主旋律提取的主要任务是从复调音乐[1]中自动获取歌声或乐器音高序列,它被广泛应用于音乐检索、体裁分类、翻唱歌曲识别等领域[2]。

主旋律提取方法通常利用音乐信号的频谱特性及音高连续性的特点来进行音高估计。其困难之处在于,复调音乐中的主导歌声或乐器通常都有一个多音伴奏,把来自不同声源按照和声结构高度耦合叠加在一起的频谱分开到对应的音符上是极其困难的。经过二十来年的发展,主旋律提取技术的研究有了重大的进步。频谱谐波加权和(Harmonic Sum Spectrum, HSS)[3]是一种简单的音高显著性度量函数,它以音乐信号的音高能量显著性和各次泛音的谐波性为关键条件,联合音高轮廓的连续性约束来跟踪主旋律的音高序列。Durrieu[4-5]提出了SF-NMF模型来分离歌声旋律和伴奏的频谱,并用Viterbi算法跟踪主旋律音高。该算法具有较高的鲁棒性,但在分析中低频音乐信号时,主旋律提取的准确率较低,容易产生八度错误。Bosch等[6]结合了HSS和源-滤波器模型得到新的显著性表示方法,并依据音高连续性获得多条候选音高轮廓线,通过分析候选音高轮廓线的特点来进行主旋律音高序列的定位。Gong等[7]将隐马尔可夫模型和"谐波乐器/打击乐器声音分离"模型进行结合,在单声道波形文件中实现旋律提取。近年来,深度学习在主旋律提取领域得到发展,Kum等[8]基于数据驱动训练了一个多列深度神经网络进行声乐旋律提取,并用隐马尔可夫模型进行旋律跟踪。之后,卷积神经网络(Convolutional Neural Network, CNN)[9-10]、递归神经网络(Recurrent Neural Network,

基金项目:国家自然科学基金(11572084,11972115)
This work was supported by the National Natural Science Foundation of China (11572084,11972115).
通信作者:韩芳(yadiahan@dhu.edu.cn)

RNN)[11]被用在主旋律提取上。Basaran 等[12]以源-滤波器分离得到的结果作为卷积循环神经网络(Convolutional Recurrent Nerual Network，CRNN)的输入，证明了不需要大量训练数据就能在简单的网络中得到较好的效果。

通过以上研究，发现一个能正确表征音高的显著性函数在主旋律提取中有重要作用，而旋律跟踪算法则能进一步提高结果的正确率。然而，主旋提取的难点在于音乐信号具有典型的谐波性，主旋律和伴奏音符按照和声结构高度耦合，各次谐波位于基频的整数倍处，甚至会出现一些基频丢失的情况，比如强低音伴奏和一些独特的演唱技巧[13]，这都使得主旋律识别变得尤其困难。

结合前人研究的优势与不足，在 SF-NMF 的基础上，本文提出一种 CNN-CRF 模型，这是一种增强音高显著性特征和自动旋律跟踪的方法。由 SF-NMF 方法得到的音高显著性特征，会使旋律轮廓出现跳变，极易产生八度错误。为了改善这种显著性表示方法，本文利用了音乐信号中丰富的谐波分量，并用一种更结构化的数据对其进行表示。首先，在 CNN 增强音高显著性表示的过程中，采用谐波 Q 变换(Harmonic Constant-Q Transform，HCQT)计算更结构化的能量谱密度，将其和基于 SF-NMF 得到的初始音高显著性特征一起输入网络，得到增强的显著性特征图。在最终的旋律跟踪和定位上，利用条件随机场(Conditional Random Fields，CRF)自动学习旋律线中音高之间的平滑约束规则，在增强的显著性特征图中自动跟踪和定位旋律音高，选择输出最佳旋律线。初始音高显著性在最后也参与旋律帧的检测，以得到最终结果。

2 CNN-CRF 的原理

本文设计的主旋律提取算法的框架如图 1 所示。

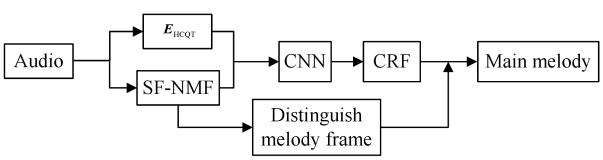


图 1 算法的总体框架

Fig. 1 General framework of algorithm

该算法的输入为一段音乐信号，对其进行预处理后，用 HCQT 计算音乐信号的能量谱密度 $\boldsymbol{E}_{\mathrm{HCQT}}$，并将其输入 CNN 中，以在结构化的数据中学习到谐波关系，并在简单特征中学习旋律的深层表示。在网络训练前，利用平滑瞬时模型估计出初始的音高显著性特征，联合能量谱密度加入网络中，通过神经网络学习局部特征，以进一步增强显著性。CRF 层则结合旋律特征和音高的平滑约束条件，在音高空间寻找最优的演变路径，生成随时间变化的音高轮廓图，并将其中最大可能轨迹线作为最终的主旋律线。最后，将训练前估计出的初始音高显著性特征用于旋律帧的检测，以修正最后的旋律线。

2.1 初始显著性特征

Durrieu 等[5]提出了一种基于音源分离的方法，它是一种平滑的瞬时混合模型，将混合音频信号的功率谱模拟为主唱和伴奏这两个分布的和，并将主乐器或歌声表示为所有可能的音符的瞬时混合物。本文将在该模型的基础上分离得到旋律的初始音高显著性特征，并将其和 $\boldsymbol{E}_{\mathrm{HCQT}}$ 一起输入网络中。

该模型假设主旋律与伴奏的混合是瞬时得到的，因此源-滤波器和伴奏可采用 SF-NMF 模型[12]建模，获得如下表示：

$$\begin{aligned}\boldsymbol{S}\approx\hat{\boldsymbol{S}}&=\boldsymbol{S}^{F_0}\odot\boldsymbol{S}^{\Phi}+\boldsymbol{S}^{B}\\&=\boldsymbol{W}^{F_0}\boldsymbol{H}^{F_0}\odot\boldsymbol{W}^{\Phi}\boldsymbol{H}^{\Phi}+\boldsymbol{W}^{B}\boldsymbol{H}^{B}\\&=\boldsymbol{W}^{F_0}\boldsymbol{H}^{F_0}\odot\boldsymbol{W}^{\Gamma}\boldsymbol{H}^{\Gamma}\boldsymbol{H}^{\Phi}+\boldsymbol{W}^{B}\boldsymbol{H}^{B}\end{aligned}\tag{1}$$

其中，$\boldsymbol{S}$ 表示音乐信号的能量谱密度，$\boldsymbol{S}=\boldsymbol{S}^{F_0}\odot\boldsymbol{S}^{\Phi}$ 表示歌声的能量谱密度估计值，$\boldsymbol{S}^{F_0}$ 和 $\boldsymbol{S}^{\Phi}$ 分别对应于源和滤波器，$\odot$ 表示矩阵的哈达玛积(Hadamard product)，$\boldsymbol{S}^{B}$ 表示伴奏的能量谱密度估计值。利用 NMF 将它们分解为基和幅值矩阵 $\boldsymbol{S}^{F_0}=\boldsymbol{W}^{F_0}\boldsymbol{H}^{F_0}$，$\boldsymbol{S}^{\Phi}=\boldsymbol{W}^{\Phi}\boldsymbol{H}^{\Phi}$，$\boldsymbol{S}^{B}=\boldsymbol{W}^{B}\boldsymbol{H}^{B}$，其中 F_0，Φ 和 B 分别表示源、滤波器和伴奏。

模型假设的滤波器是一种平滑滤波器，但无法直接构造平滑滤波器结构的基 $\boldsymbol{W}^{\Phi}$，因为它取决于主旋律分布。因此，为了构建这样的滤波器，需要将滤波器的基进一步分解为一组带通滤波器的线性组合：$\boldsymbol{W}^{\Phi}=\boldsymbol{W}^{\Gamma}\boldsymbol{H}^{\Gamma}$。$\boldsymbol{W}^{\Gamma}$ 是提前构造的一组带通滤波器，它们的带宽覆盖在不同的频率段，确保了这组滤波器能覆盖所分析的所有频率。模型假设了 $\boldsymbol{S}^{F_0}$ 具有谐波结构，因此需要提前构造具有谐波结构的基 $\boldsymbol{W}^{F_0}$。$\boldsymbol{W}^{F_0}$ 的每一列都代表一个基频的 f_0 谐波结构。设计从最小频率 55 Hz 开始，以对数为间隔，连续 f_0 值之间的比例将是 $2^{1/120}$，即每个八度都有 120 个频率带个数。这样的结构使幅值矩阵 $\boldsymbol{H}^{F_0}$ 中的相应行能够表示特定基频 f_0 的幅值，类似于显著性特征。参数估计则是利用基于最大似然的启发式乘法[4]来更新，每次迭代按照以下顺序更新参数：$\boldsymbol{H}^{F_0}$，$\boldsymbol{H}^{\Gamma}$，$\boldsymbol{H}^{\Phi}$，$\boldsymbol{W}^{B}$ 和 $\boldsymbol{H}^{F_0}$。将得到的初始显著性特征 $\boldsymbol{H}^{F_0}$ 作为输入网络的一个特征。

2.2 结构化的谐波能量谱密度

HCQT 有谐波、时间和频率 3 个维度的结构信息，这样的结构化信息将更容易表现出音乐信号的谐波关系和时间约束。为了更好地捕获音乐信号中的谐波关系，本文采用 HCQT 来计算音乐信号的能量谱密度 $\boldsymbol{E}_{\mathrm{HCQT}}$。HCQT 可以很方便地将谐波表示在一个维度上，不仅符合音符的分布规律，而且能得到更结构化的数据特征。以此作为网络的输入表示之一，使网络更容易学习到深度特征。

标准的常 Q 变换(Constant-Q Transform，CQT)是由一组中心频率按对数间隔的滤波器组成，其中第 k 个频率带的中心频率为：

$$f_k=f_{\min}\cdot 2^{\frac{k}{B}}\tag{2}$$

其中，$k=0,1,2,\cdots,K-1$；$f_{\min}$ 和 $f_{\max}=f_{K-1}$ 分别为频率区间的最小和最大频率值；B 为每个八度频率带的个数。这符合十二平均律中定义的音高分布规律[14]，且相较于短时傅里叶变换(Short-Time Fourier Transform，STFT)，CQT 在低频时能得到更高的频率分辨率，且所计算的数据量更小。通过 CQT 计算音乐信号能量谱密度的第 k 个分量公式如式(3)所示：

$$\begin{aligned}E_{\mathrm{CQT}}(k)&=|X_{\mathrm{CQT}}|^2(k)\\&=\left|\frac{1}{N[k]}\sum_{n=0}^{N[k]-1}w(n,k)x(n)\mathrm{e}^{-\mathrm{j}2\pi Qn/N[k]}\right|^2\end{aligned}\tag{3}$$

其中，Q 为常数因子，大小为 $1/(2^{1/B}-1)$；$x(n)$ 表示一段离散信号；$w(n,k)$ 是 CQT 所用的窗函数，它的长度 $N[k]$ 与 k 有关：

$$N[k]=Q\frac{f_s}{f_k}\tag{4}$$

在 CQT 中，f_k 的 h 次谐波频率 $h\cdot f_k$ 只能在 $h=2^n$(n 为正整数)直接测量，很难捕捉到奇数次的谐波信息。对此，

HCQT 构造了 3 个维度信息来捕获谐波信息：$\boldsymbol{H}[H,T,F]$。除了时间维度 T 和频率维度 F 外，还有包含谐波信息的维度 H，当 $H=1$ 时，$\boldsymbol{H}[1,T,F]$则是一个标准的 CQT 结果。对于 $H=h(h>0)$次谐波，$\boldsymbol{H}[h,T,F]$是最小频率为 $h\cdot f_{\min}$ 的 CQT 结果，并且所有谐波都具有相同频率分辨率和八度的频率带个数。因此，在 $\boldsymbol{H}[h,T,F]$中，第 k 个频率带的中心频率[9]为：

$$f_k=h\cdot f_{\min}\cdot 2^{\frac{k}{B}} \tag{5}$$

其中，$k=0,1,2,\cdots,K-1$。

通过 HCQT 计算音乐信号能量谱密度第 h 维度的第 k 个分量的公式如式(6)所示：

$$E_{\mathrm{HCQT}}(k,h)=|X_{\mathrm{HCQT}}|^2(k,h)=\left|\frac{1}{N[k,h]}\sum_{n=0}^{N[k,h]-1}w(n,k,h)x(n)\mathrm{e}^{-\mathrm{j}2\pi Qn/N[k,h]}\right|^2 \tag{6}$$

对于双音轨文件，其能量计算公式如下所示：

$$E=0.5(E_{\mathrm{L}}^2+E_{\mathrm{R}}^2) \tag{7}$$

其中，E_{L} 和 E_{R} 分别是音乐文件的左、右通道的能量。

通过 HCQT 计算获取谐波的能量信息，得到更结构化的数据，再利用二维卷积神经网络进行建模，可以有效地利用时间、频率和谐波的局部特征及它们之间的关系。

2.3 CNN-CRF 模型的构建

本文设计的 CNN-CRF 模型如图 2 所示。

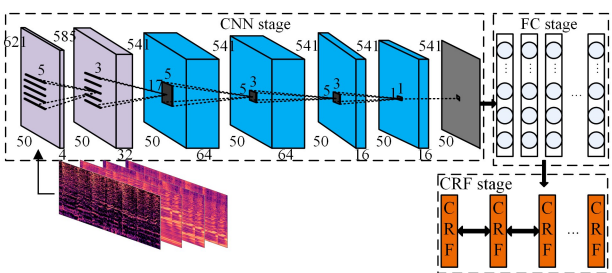


图 2 CNN-CRF 模型

Fig. 2 CNN-CRF model

模型主要分为两个部分：音高显著性增强和动态规划过程。将能量谱密度特征和初始的音高显著性信息输入卷积神经网络中，提取更深层次的旋律特征，以增强音高显著性表示。这里也可以将 CNN 看作编码层，它将特定的音乐信息提取出来并编码输入 CRF 层。CRF 层则在音高空间中学习旋律特征和音高的平滑约束条件，以得到最佳主旋律线。

CNN 过程是一个高维特征提取过程，共使用了 6 个不同维度的卷积层，在卷积过程中保持时间维度不变。在前两层网络中引入空洞卷积，以扩大网络的局部感受野，快速获得全局信息。在第一层卷积层中使用 32 个大小为(5 5)、膨胀系数为(1,10)的空洞卷积块，以提取不同音符之间的特征。在第二层卷积层中使用 64 个大小为(3 5)、膨胀系数为(1,12)的空洞卷积块，以提取在不同八度中的音符特征。最后一层卷积层则使用 1 个大小为(1 1)的卷积核来把旋律特征映射到二维特征图上，方便之后的旋律提取和跟踪。为了保持数据的完整性，在整个卷积网络中没有添加池化层，每个卷积核步长都为 1，并且都设置一个偏置值，采用 ReLU 激活函数。

在卷积层之后使用一个大小为 64 的全连接神经网络，其作用是将 CNN 层增强得到的显著性特征图重新整合并压缩至与旋律标签维数相等的空间。

在全连接层之后使用条件随机场，它是一种无向图模型[15]。在输入给定的一组随机变量序列的条件下，它能得到输出一组随机变量的条件概率分布模型，常被用于动态规划和序列标注。

因此，在给定一组时间序列上的旋律特征时，CRF 可以在全局范围内学习到音高的局部约束规则，包括旋律特征和音高的平滑约束条件之间的关系，然后解码出最优的旋律音高序列值，以保证输出旋律线结果的有效性。

CRF 层的输入是经过 FC 层压缩的显著性特征图，即旋律特征序列。CRF 将计算 t 时刻的每个音高得分和不同时刻的音高转移概率，并根据音高得分输出旋律序列。

给定旋律特征序列：

$X=(x_1,x_2,x_3,\cdots,x_T)$

旋律音高序列标签：

$Y=(y_1,y_2,y_3,\cdots,y_T)$

其预测的序列得分如式(8)所示：

$$P(X,Y)=\sum_{t=1}^{T}M_{y_t,y_{t+1}}+\sum_{t=1}^{T}N_{t,y_t} \tag{8}$$

其中，$\boldsymbol{M}$ 是转移矩阵，$M_{y_t,y_{t+1}}$ 表示从 t 时刻音高 y_t 转移到 $t+1$ 时刻音高 y_{t+1} 的概率，N_{t,y_t} 表示 t 时刻音高为 y_t 的概率。$P(X,Y)$表示输入旋律特征序列 X 被标记为音高序列 Y 的概率分数，求出最大 $P(X,Y)$的值，即得到最大可能的旋律音高序列值。

2.4 旋律帧检测

因为在 CRF 中已经对非旋律帧进行了估计，即音高被标记为 0 的帧，所以本文只是通过简单算法进行旋律帧的辅助检测，实现旋律帧的修正功能。在初始音高显著性特征图中，对每帧的显著值进行求和统计，若该帧的统计值小于所有帧的统计平均值的 α 倍，则认为该帧为非旋律帧，否则为旋律帧。

3 实验结果

3.1 主旋律提取评价指标及数据集

主旋律提取目标有两个：1)输出正确的旋律音高。当输出落在标签的半个半音范围之内时，则认为模型对旋律音高估计正确；若不在这个音程范围内，则认为模型对旋律音高估计错误。2)对旋律帧的判别。当该帧有旋律时，则输出旋律音高的估计值；当该帧不存在旋律时，则输出 0 作为音高值。

旋律提取通常采用以下 5 个全局度量[16]进行性能评测：查全率(voicing recall rate)、虚警率(voicing false alarm rate)、原始音高准确率(raw pitch accuracy)、音色准确率(raw chroma accuracy)和整体准确率(overall accuracy)。

主旋律提取实验采用 MIR-1K 和 MIREX05 两个数据集。

(1)MIR-1K 数据集：由台湾大学咨询工程学系多媒体咨询检索实验室录制，总共包含 1000 段 R&B，pop，jazz，opera 等分格的音乐信号，采样频率为 44.1 kHz，每段录制时长为 8～12 s，由 19 个经过专业训练的演唱者按照流行音乐的唱法录制。整个数据集的旋律音高标签间隔 10 ms，本文随机选取其中的 800 段音乐信号作为训练数据，其余 200 段音乐信号作为测试数据。

(2) MIREX05 数据集：由 Columbia University 的 Labrosa 实验室团队标注，包括多种风格共 13 个音乐片段，其采样频率为 44.1 kHz，每个片段录制时间约 24～39 s。整个数据集的旋律音高标签间隔 10 ms，本文随机选取其中 1 段音

乐作为测试数据，其余部分全部作为训练集数据。

3.2 结果与分析

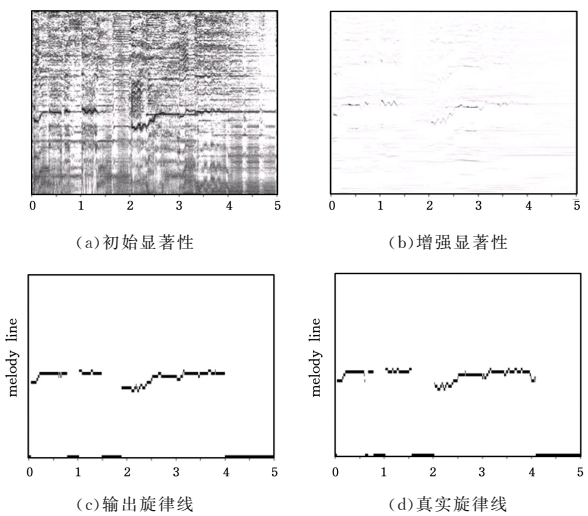
本文计算 HCQT 利用了 3 个谐波分量。为了避免过多的偶数次谐波冗余信息，只计算奇数次谐波的 CQT，其中 $h\in\{1,3,5\}$；并且帧移设置为 10 ms，其大小与数据集中标签的时间间隔相等。当 $h=1$ 时，将最小频率 $f_{\min}$、最大频率 $f_{\max}$ 分别设置为 55 Hz，1976 Hz，即在横跨约 5 个八度的音符范围内分析旋律频率。当 $B=12$ 时，每个频率刚好对应各自音符的基频，本文将每个八度的频率带个数 B 设置为 120，即连续 f_k 之间的比值为 $2^{1/120}$，每次计算 0.5 s 的音乐信号，将得到大小为(3×50×621)的能量谱密度信息。

在 SF-NMF 模型中，对 $\boldsymbol{W}^{F_0}$ 的构建从最小频率 55 Hz 开始，以对数为间隔，即连续的 f_0 值之间的比例将是 $2^{1/120}$。在 $\boldsymbol{W}^{\Gamma}$ 中设置了一组 30 个带通滤波器组成的滤波器组，其中带通滤波器用汉宁窗来模拟。设置迭代次数为 40，将得到大小为(50×621)的幅值矩阵 $\boldsymbol{H}^{F_0}$。

本文是在 55～1976 Hz 范围内分析包括 63 个音符频率的信息，并额外设置一个标签表示无旋律帧，每个时间维度的网络输出大小为 64 维。每次计算 50 ms 的音乐信号的初始特征，将大小为(4×50×621)的特征图输入网络中，其中初始的音高显著性特征占 1 个通道的信息，能量谱密度特征占 3 个通道的信息，最后网络将输出大小为(50×64)的旋律图，其中将包含一条完整的旋律线。

旋律帧检测阶段的 α 值设置为 10%。

图 3 所示为 MIREX05 测试集前 5 s 的显著性特征图及旋律线。图 3(a)是 SF-NMF 得到的初始音高显著性特征图；图 3(b)是加入结构化数据后 CNN 层的输出，即经过 CNN 增强后的音高显著性特征图。可视化之后可以清楚地观察到旋律线的轮廓，虽然其中包含了其他候选旋律线，但也证明了 CNN 增强音高显著性的有效性。图 3(c)为模型最终输出的旋律线；图 3(d)为标签标记的旋律线，除了旋律帧和非旋律帧交界处存在较大差异，模型估计的旋律线十分接近实际旋律线。



(a)初始显著性 (b)增强显著性

(c)输出旋律线 (d)真实旋律线

图 3 显著性特征图及旋律线

Fig. 3 Salient feature map and melody line

表 1 和表 2 分别列出了在数据集 MIR-1K 和 MIREX05 中使用不同旋律提取方法的结果。

表 1 不同方法在 MIR-1K 中的结果

Table 1 Results of different methods in MIR-1K

(单位：%)

Method	*VR*	*VFA*	*RPA*	*RCA*	*OA*
Salamon[1]	85.1	**30.8**	72.9	75.7	**69.6**
Kum[8]	93.4	65.8	72.6	77.0	61.3
Bosch[6]	74.3	50.0	47.3	52.1	48.0
CNN-CRF	**95.5**	53.6	**73.6**	**78.1**	66.1

表 2 不同方法在 MIREX05 中的结果

Table 2 Results of different methods in MIREX05

(单位：%)

Method	*VR*	*VFA*	*RPA*	*RCA*	*OA*
Salamon[1]	77.6	23.9	69.8	**76.9**	67.6
Kum[8]	89.4	58.5	73.3	75.2	61.6
Bosch[6]	85.2	30.9	72.6	76.8	71.2
CNN-CRF	**91.9**	**16.8**	**76.3**	76.6	**78.6**

从表 1 和表 2 中可以看出本文所提出的 CNN-CRF 方法在大部分指标上都优于其他方法。MIR-1K 是一个特殊的数据集，其旋律能量和伴奏能量之比为 1，所以采用信号能量作为特征时，有时会出现将伴奏段误判为旋律段的情况，导致虚警率可能高于其他方法；但大部分情况都能正确识别，且在整体准确率上有较好的结果，特别是在 MIREX05 中明显优于其他方法。

原始音高准确率在主旋律提取中至关重要。由于引入 CNN 强化了初始音高显著性表示，并利用 CRF 进行旋律跟踪，本文方法在旋律查全率和原始音高准确率上较其他方法有了明显提高，其在 MIR-1K 上的旋律查全率和原始音高准确率较其他方法分别提升了 2.1%和 0.7%，在 MIREX05 中的旋律查全率和原始音高准确率较其他方法分别提升了 2.5%和 3.0%。

为了验证谐波能量谱密度对强化显著性表示的有效性，本文尝试在不同输入情况下比较模型所表现的性能，并用音色准确率减去原始音高准确率来表示八度错误。表 3 列出了在两个数据集中输入不同特征时的实验结果。

表 3 不同输入特征的实验结果

Table 3 Experimental results of different input features

(单位：%)

数据集	Method	*RPA*	*RCA*	*OA*	*Octave error*
MIR-1K	$\boldsymbol{E}_{\mathrm{HCQT}}$	62.2	66.8	58.1	4.6
	$\boldsymbol{H}^{F_0}$	62.1	71.3	51.2	9.2
	$\boldsymbol{E}_{\mathrm{HCQT}}+\boldsymbol{H}^{F_0}$	**73.6**	**78.1**	**66.1**	**4.5**
MIREX05	$\boldsymbol{E}_{\mathrm{HCQT}}$	69.6	70.5	60.1	0.9
	$\boldsymbol{H}^{F_0}$	64.2	65.3	71.8	1.1
	$\boldsymbol{E}_{\mathrm{HCQT}}+\boldsymbol{H}^{F_0}$	**76.3**	**76.6**	**78.6**	**0.3**

对比发现，在同时输入 $\boldsymbol{E}_{\mathrm{HCQT}}$ 和 $\boldsymbol{H}^{F_0}$ 的情况下，原始音高准确率、音色准确率和整体准确率有明显提高，而且大大减少了八度错误的产生。这是因为在 CNN 的帮助下，可以提取结构化信息的高维特征来加强显著性表示，弥补 SF-NMF 模型对音高的误判，且谐波能量谱密度中包含丰富的谐波信息，它可以减少在 SF-NMF 模型中产生的八度错误。

结束语 本文提出了一种音高显著性增强和自动旋律跟踪的方法。在基于 SF-NMF 模型的基础上分离得到初始显著性表示，通过计算结构化的数据帮助 CNN 学习特定的音乐信息来增强初始显著性表示。在旋律跟踪上，利用 CRF 自

动学习音高之间的约束规则，提高旋律提取的原始音高准确率。实验表明，模型能有效提取旋律信息并增强其显著性表示，加入结构化数据可以减少 SF-NMF 模型所产生的八度错误，且相较于其他方法，本文提出的方法有更高的旋律查全率和原始音高准确率。但本文只在主旋律为歌声的音乐信号中进行实验，接下来还应改进模型，将其应用于主旋律来自乐器的音乐信号提取。

参 考 文 献

[1] SALAMON J, GOMEZ E. Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(6): 1759-1770.

[2] ZHANG W W, CHEN Z, YIN F L, et al. Review on Melody Extraction from Polyphonic Music[J]. Acta Electronica Sinica, 2017, 45(4): 1000-1011.

[3] KLAPURI A P. Multiple fundamental frequency estimation by summing harmonic amplitudes[C]// 7th International Society for Music Information Retrieval Conference (ISMIR). Victoria: Music Information Retrieval Society, 2006: 216-221.

[4] DURRIEU J L, DAVID B, GAËL R. A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation[J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(6): 1180-1191.

[5] DURRIEU J L, RICHARD G, DAVID B, et al. Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(3): 564-575.

[6] BOSCH J J, BITTNER R M, SALAMON J, et al. A Comparison of Melody Extraction Methods Based on Source-Filter Modelling [C]// 17th International Society for Music Information Retrieval Conference (ISMIR 2016). 2016: 571-577.

[7] GONG J C, LIU G. A Melody Pitch Extraction Algorithm for Waveform File Based On Hidden Markov Mode[J]. Software, 2013, 34(12): 152-155, 177.

[8] 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. 2016: 819-825.

[9] BITTNER R M, MCFEE B, SALAMON J, et al. Deep salience representations for f0 estimation in polyphonicmusic[C]// Proceedings of the International Society for Music Information Retrieval (ISMIR). 2017.

[10] SU L. Vocal melody extraction using patch-based CNN [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 371-375.

[11] PARK H, YOO C D. Melody extraction and detection through LSTM-RNN with harmonic sum loss[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017: 2766-2770.

[12] BASARAN D, ESSID S, PEETERS G. Main melody extraction with source-filter NMF and CRNN[C]// Proceedings of International Society for Music Information Retrieval (ISMIR). 2018: 82-89.

[13] ZHANG W W, CHEN Z, YIN F L. Melody Extraction from Polyphonic Music Combining Modified Euclidean Algorithm and Dynamic Programming[J]. Journal of Signal Processing, 2018, 34(8): 1008-1015.

[14] FANG X Y. Research on Melody Extraction in Polyphonic Music[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.

[15] MCCALLUM A, LI W. Early results fornamed entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learningat HLT-NAACL 2003. Association for Computational Linguistics, 2003: 188-191.

[16] LI W, FENG X Y, WU Y M, ZHANG X L. Review on Main Melody Extraction from Pop Music[J]. Computer Science, 2017, 44(5): 1-5.

JIN Wen-qing, born in 1996, master. His main research interests include deep learning and music information retrieval.

HAN Fang, born in 1981, Ph.D, professor. Her main research interests include intelligent systems and neurodynamics.