文章编号：2096-1618(2020)04-0373-05

基于稀疏自编码神经网络的声乐主旋律提取

孙文慧[1,2]， 夏秀渝[1]， 陆 雄[1]
(1.四川大学电子信息学院，四川 成都 610065；2.武警警官学院 四川 成都 610213)

摘要：旋律是音乐中最重要的要素，音乐主旋律提取是音乐检索的核心技术之一。复调音乐中歌声的音高序列构成了声乐主旋律。提出一种声乐主旋律自动提取改进算法，根据声乐信号的频谱特点，改进音高显著度函数的计算方法，降低计算复杂度，减少声乐主旋律提取时间。改用性能更优的稀疏自编码神经网络替代原算法的浅层BP神经网络作为基频判别模型，提高主旋律模型的识别准确率，降低旋律定位虚警率，从而提高声乐主旋律提取整体的准确率。在MIR-1K数据集上进行的实验表明，改进算法提取的声乐主旋律整体准确率比原算法至少提高了1.51%，提取主旋律的平均提取时间要比原算法减少大约0.12 s。

关 键 词：主旋律提取；音高显著度；稀疏自编码；基频判别
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2020.04.002

0 引言

近年来，计算机和网络技术进入一个新信息时代，多媒体信息数据成倍增长，传统靠人工方式对音乐进行文本标注来检索音乐已经不能够满足需求。鉴于此，更贴近生活基于内容的音频检索技术成为研究热点。基于内容的音频检索技术主要是利用音乐的基本特征：节奏、旋律、频谱、幅度、音高等表述音乐，以便实现音频的匹配。文中研究声乐主旋律提取，为后续音频检索打下基础，具有一定的研究意义。

旋律(melody)是音乐的基本特征，当前学术界普遍认同的主旋律的定义由Poliner等[1]提出：主旋律是听者根据一段复调音乐(polyphonic music)感知的，并被听者识别为音乐本质的单音高序列。旋律就好比是音乐的灵魂和基础，在音乐表现中具有重要意义[2]。旋律包含了大量音乐中极具价值的重要信息，音乐主旋律提取的效果直接影响到音频检索的效率，因此研究声乐主旋律提取具有极大的应用价值。

针对课题组前期提出的声乐主旋律提取算法[3]进行改进，一是根据声乐频谱的特点，改进音高显著度计算方法，以降低算法复杂度。二是声乐主旋律判别模块改用稀疏自编码神经网络(SAENN)[4]和softmax[5]分类器代替BP神经网络作为基频判别模型，达到提高基频识别准确率，降低旋律定位虚警率的目的。

1 算法原理

1.1 算法总体框架

算法的总体框架，如图1所示。

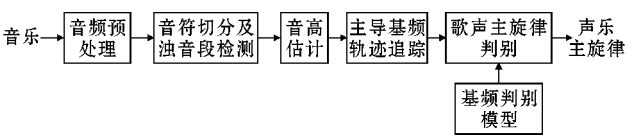

图1 声乐主旋律提取算法总体框架

根据总体框架，声乐主旋律的提取步骤如下：第一步，对输入音乐进行预处理，主要工作是对信号降采样、归一化、分帧、加窗和时频域变换；第二步，通过频谱分析进行音符切分和浊音段检测；第三步，音高估计，在浊音段每一帧上计算音高显著度，根据基频特征和音高显著度函数筛选3个候选基频；第四步，主导基频轨迹追踪，在每个浊音段内利用维特比算法提取一条主旋律线；第五步，歌声主旋律判别，由于提取的主旋律并非都为歌声主导，所以采用一个歌声判别模型判断主导基频是否属于歌声，连接所有属于歌声的音高序列形成声乐主旋律。

1.2 信号预处理

歌声是一种时变非平稳的语音信号，为更好地分析其携带的信息，通常采取一系列措施先对其进行预处理，包括对输入的音频信号降采样、归一化、分帧、加窗和STFT等。

考虑到语音信号的频率范围集中在300～3400 Hz，为降低后续信号处理的计算量，所用的音频信号统一降采样到8 kHz。另外对输入信号做归一化处理，保证进一步处理的一致性。

语音信号在短时间内近乎平稳，基于此，可以对语音信号进行分帧加窗处理，文中采用汉明窗，其公式为

$$w(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)] & 0\le n\le N\\ 0 & \text{else}\end{cases} \quad (1)$$

收稿日期：2019-11-22

式中,N 为窗长,帧长40 ms,帧移20 ms。根据需要,分帧后可以采用 STFT(短时傅立叶变换)对信号进行时频变换。

1.3 音符切分及浊音段检测

音乐是由一个个音符组成的,每个音符具有相对稳定的频谱特性,对音乐进行音符切分,基于音符估计音高可以获得更可靠准确的音高特征。采用一种基于度量距离的音频切分算法(*DIS* 算法)[6]分割音符。

音符分割采用的特征参数为短时幅度谱,按帧滑动数据窗,计算第 t 帧音频前后两个数据窗音频特征的 *DIS* 度量距离,如下:

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

式中,$\boldsymbol{\mu}_{t,1}$和$\boldsymbol{\mu}_{t,2}$代表相邻两段音频特征的均值矢量,$tr(\sum_{t,1})$和 $tr(\sum_{t,2})$代表相邻两段音频特征矩阵的迹。

得到 *DIS* 度量距离曲线后,通过寻找 $DIS(t)$ 极大值点并按照一定的筛选规则筛选音符切分点。筛选规则为:利用标准差阈值与极大值点进行比较,若极大值点大于标准差阈值则为音符切分点,否则舍去。

另外,音乐信号中存在非浊音段,非浊音段是不存在基频的。所以,基频估计前应排除这些段。文中采用频谱方差法判断有声段中的浊音段和非浊音段[7]。

1.4 多基音估计

复调音乐中往往同时存在多个基频,系统采用音高显著度函数进行多候选基频提取。具体采用的是基于幅度压缩基音估计滤波器(pitch estimation filter with amplitude compression,PEFAC)[8-9]的基频提取方法。

若一段浊音信号的基频为 f_0,此信号在频域可表示为

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

其中 a_k 为 k 次谐波的系数。对 f 取对数可得:

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

其中 $q=\log f$。将它与一个梳状滤波器卷积,卷积结果 $Y(q)\cdot h(-q)$,即基频显著度函数将会在 $q_0=\log f_0$ 的位置产生一个峰,峰值在频域上的位置也就是所需信号浊音段的基频。

文中实际采用的梳状滤波器为

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

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

其中参数 γ 控制梳状滤波器谱峰的宽度,K 代表谱峰的个数。

定义每帧信号基频显著度函数为

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

针对多基频信号,找出 $S(q)$ 所有峰值位置上对应的频率集合 $\{\psi_1\}$ 作为候选基频。另外,歌声的基频范围通常为 70 ~ 1000 Hz,在每个频点均按式(7)计算 $S(q)$ 的运算量较大。为降低计算复杂度,根据歌声频谱的特点,只筛选信号频谱幅值大的一些频点计算显著度值,可以大大降低计算量。

改进的音高显著度计算实现如下:设置一个阈值 T,针对 70 ~ 1000 Hz的频点,判断信号对数幅度谱的幅值是否大于 T。若大于 T,则按(8)式计算显著度值;否则,将该频点对应的显著度值直接置 0。即:

$$S(q)=\begin{cases}\sum_{q=q_0}Y(q)\cdot h(q-q_0) & Y(q)\geq T\\ 0 & Y(q)<T\end{cases} \tag{8}$$

式中 q 为对数域频点,$Y(q)$ 为对数域频谱,$h(q)$ 为对数域梳齿滤波器函数;$S(q)$ 为音高显著度函数。

由于只计算 70 ~ 1000 Hz能量较大的有限个频点的显著度值,因此计算复杂度显著降低。最后,从所有频点中找出最大的 3 个互相不成为倍频或者半频的频率作为最终的候选基频。其中,f_n 表示第 n 个候选基频的频率位置,s_n 表示第 n 个候选基频的显著度值。

1.5 主导基频轨迹跟踪

采用文献[3]类似的方法进行主导基频轨迹跟踪。根据音高显著度函数的峰值提取出多个候选基频,然后运用维特比算法[10]在浊音段内(每个音符内)进行主导基频轨迹跟踪[3]。

每个浊音段采用基于音高似然度和音高转移概率的 Viterbi 算法提取该段的最优基音序列,即主导基频轨迹,使其满足公式:

$$F=\underset{(f_1,f_2,\cdots,f_T)}{\arg\max}\{\alpha\sum_{t=1}^{T_1}\lg p(f_t\mid\boldsymbol{X}_t)+\beta\sum_{t=2}^{T_1}\lg p(\Delta f_t)\} \tag{9}$$

式中,T_1 表示浊音段的分段长度,$p(f_t\mid\boldsymbol{X}_t)$ 和 $p(\Delta f_t)$ 分别表示音高似然度和转移概率,α 和 β 分别为各自对应的权值。

1.6 歌声主旋律判别

声乐主旋律指的是人唱的歌声旋律,而上述主旋律提取方法提取的每个音符段的主导基频轨迹可能属于歌声,也有可能属于伴奏乐器。因此,需要引入基频判别模型判断每段音频的主导基频轨迹是否属于歌声旋律。

原算法采用了浅层 BP 神经网络[11]作为基频判别模型,目前深度神经网络在分类任务中展现出更好的

性能,在语音信号处理领域也得到了广泛应用。因此,将使用稀疏自编码神经网络(SAENN)和 softmax 分类器来建立训练基频的判别模型,以达到更好的声乐基频识别效果。

稀疏自编码(SAE)神经网络在自编码(AE)神经网络的基础上加了稀疏限定约束,采用一种无监督性质的特征学习算法[12]。该算法把稀疏性限制用在中间隐藏层的神经元,训练好的稀疏降噪自编码神经网络相当于一个非线性滤波器,利用少量的隐层激活单元表征原有信号,自动地筛选信号显著性原子,这对声乐基频识别率的提高具有重要意义。

具体网络结构如图 2 所示。

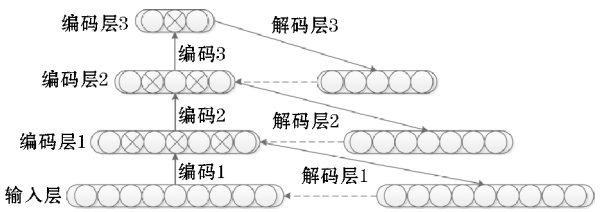

图 2　SAE 结构图

基频判别模型前端为稀疏自编码神经网络,其结构为 12-200-50-16,即输入为能够反映声音谱包络特征的 12 维 MFCC 特征,中间使用的隐藏层神经元的分别为 200、50 和 16,最后一个隐层神经元的后端接 softmax 分类器含有两个输出端,分别代表歌声基频和器乐基频。

此外,超参数设置如下:权值惩罚参数 λ 设为 0.0001;期望稀疏参数 ρ 为0.05;稀疏惩罚项权值 β 为 3;预训练与 softmax 分类器的最大迭代次数为 500;SAENN(稀疏自编码神经网络) 整体网络所能够承受的微调的最大迭代次数为 2000。

训练时采用 L-BFGS 算法进行权值优化计算,该算法只保存并利用最近 m 次迭代的曲率信息构造海森矩阵的近似矩阵,每次迭代的开销较小,每一次的迭代都能保证矩阵的正定性,算法具有良好的鲁棒性。

基于以上基频判别模型,判断每个浊音段的主导基频是否为歌声主旋律。具体步骤如下:

(1)首先构造一个以矩形波 $b(f)$ 为基本波形的梳齿滤波器,滤波器频率范围 0 ~4000 Hz,公式为

$$h(f)=\sum_{k=1}^{K}\delta(f-kF_0)\cdot b(f) \tag{10}$$

其中 K 代表频率范围内波形的个数。

(2)求基频 F_0 对应的谐波谱:用上一步构造的梳齿滤波器对信号的幅度谱滤波,提取谐波谱的梅尔倒谱参数 MFCC[13](声音产生机制和人耳听觉感知特性相结合的产物,是音频信号分析的重要特征参数之一[14])。

(3)将 MFCC 参数送入上述训练好的基频判别模型进行识别,得到 F_0 是否为歌声基频的判断结果。

(4)通过各浊音段中被判断为歌声基频统计得到的帧数,若其结果小于一段中总帧数的 1/2,则 F_0 的轨迹不是歌声主旋律,反之为歌声主旋律。

2　实验及结果分析

2.1　实验数据

实验所采用的音乐数据集均取自 Hsu 提供的 MIR-1K[15]数据集,它由 110 首中国卡拉 OK 歌曲组成,演唱部分为业余歌曲爱好者录制的 1000 首采样率为16 kHz的音乐片段,背景音乐与歌声分别录入左右声道,包含时间间隔为10 ms的歌声基频标签。

随机选出 500 段音乐用于训练主导基频判别模型的神经网络,其他 500 个音乐片段作为主旋律提取测试集。

2.2　基频判别模型训练与测试

为了与原算法的基频判别模型[3]进行对比,采用相同的实验条件来训练和测试基于 SAE 的基频判别模型。所使用的数据情况如表 1 所示。

表 1　模型训练数据和测试数据

	歌声主导基频帧数	伴奏主导基频帧数	总帧数
训练数据	142483	131871	274354
测试数据	162379	154899	317278

利用以上训练数据,采用 L-BFGS 算法对基于稀疏自编码神经网络的基频判别模型进行训练。完成模型训练后,再利用测试集数据进行性能测试。首先提取测试数据每帧信号的 MFCC 参数及其对应标签,然后将每帧信号的 MFCC 参数送入基频判别模型,得到是否为人声基频的判决结果并与标签结果对比,最后统计识别准确率 β:

$$\beta=\frac{N_1}{N} \tag{11}$$

式中 β 为识别准确率,N_1 为测试数据集中被判断正确的帧数,N 为总测试数据的帧数。

实验表明该模型识别准确率达85.1%,性能指标相比原算法提高约 5%,优于原算法的基频判别模型。后续实验表明,应用该模型可以降低旋律定位虚警率和提高算法的整体准确率。

2.3　音乐主旋律提取实验

声乐主旋律提取有两个基本任务:一是判断旋律

是否真正存在,二是准确估计主旋律的音高(要求音高的估计值和参考值之间的差别不超过半个半音范围)。围绕基本目标任务,同样采用文献[16]使用的5种性能指标评价算法性能[3]:旋律定位查全率(voicing recall rate,VR),旋律定位虚警率(voicing false alarm rate,VFAR),原始音高准确率(raw pitch accuracy,RPA),原始色度准确率(raw chroma accuracy,RCA),整体准确率(overall accuracy,OA)。在前5种常用的评价指标基础上,增加了第6种评价指标"平均提取时间",用来对比改进算法和原算法的计算量。平均提取时间(average extraction time,AET)定义为程序的运行时间与主旋律基频数量之比。

所用的测试数据为随机选取的500段音乐,以下实验分别在信干比为0 dB和信干比为5 dB的情况下进行,具体结果如表2和表3所示。

表2 SIR为5dB时的实验结果

性能指标 / 算法	VRR/%	VFAR/%	RPA/%	RCA/%	OA/%	AET/s
原算法	90.62	15.6	86.91	86.99	86.22	0.58
改进算法	92.31	13.7	88.05	88.14	87.73	0.46

表3 SIR为0dB时的实验结果

性能指标 / 算法	VRR/%	VFAR/%	RPA/%	RCA/%	OA/%	AET/s
原算法	79.43	14.3	73.29	73.52	77.4	0.58
改进算法	81.25	12.2	75.63	75.95	79.1	0.46

由表2和表3可以看出,改进算法的所有性能指标均要优于原算法,旋律定位查全率(VRR)、原始音高准确率(RPA)、原始色度准确率(RCA)、整体准确率(OA)4个指标在不同信干比条件下均有所提升,而旋律定位虚警率(VFAR)下降了2%左右,说明改进算法能更准确地识别出伴奏和歌声的旋律,这是由于改进算法中的基频判别模型的识别准确率要高于原算法。此外,改进算法的AET要比原算法减少了大约0.12 s,验证了改进算法在提取每段音乐旋律时,音高显著度函数的计算过程和复杂度要低于原算法。

3 结论

提出一种基于稀疏自编码神经网络的主旋律提取改进算法。在原算法的基础上,改进了音高显著度函数的计算方法和声乐主旋律判别模块。其中,根据声乐信号频谱的特点改进了音高显著度函数的计算方法,降低算法计算复杂度,减少声乐主旋律提取时间;声乐主旋律判别模块改用基于稀疏自编码深度神经网络的分类器进行基频类别判断,提高模型的识别准确率,降低旋律定位虚警率,从而提高了算法整体准确率。改进算法提取主旋律的平均提取时间比原算法减少了大约0.12 s,提取的声乐主旋律整体准确率至少提高了1.51%。

参考文献:

[1] Poliner GE, Ellis DP, Ehmann AF, et al. Melody tran-scription from music audio: approaches and evaluation [J]. IEEE Transactions on Audi, Speech, and Language Processing, 2007, 15(4): 1247-1256.

[2] 李重光. 基本乐理通用教材[M]. 北京:高等教育出版社,2004.

[3] 陆雄,夏秀渝,蔡良,等. 声乐主旋律的自动提取[J]. 太赫兹科学与电子信息学报,2019(3):482-488.

[4] 仲志丹,樊浩杰,李鹏辉. 基于稀疏自编码神经网络的抽油机井故障诊断[J]. 西安科技大学学报,2018,38(4):669-675.

[5] 王勇,赵俭辉,章登义,等. 基于稀疏自编码深度神经网络的林火图像分类[J]. 计算机工程与应用,2014,50(24):173-177.

[6] 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(5):38-41.

[7] 宋知用. MATLAB在语音信号分析与合成中的应用[M]. 北京:北京航空航天大学出版社,2013.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 朱啸天,张艳珠,王凡迪. 一种基于稀疏自编码网络的数据降维方法研究[J]. 沈阳理工大学学报,2016,35(5):39-43.

[13] 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.

[14] 韩纪庆,张磊,郑铁然. 语音信号处理[M]. 北京:清华大学出版社,2004.

[15] Hsu C L,Jang J S R. MIR-1K Dataset[EB/OL]. http://sites. google. com/site/unvoicedsoundseperation/mir-1k. 2009. 7. 22,2018-01-10.

[16] 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.

Vocal Main Melody Extraction based on Neural Network of Sparse Autoencoder

SUN Wenhui[1,2], XIA Xiuyu[1], LU Xiong[1]

(1. College of Electronic and Information Engineering,Sichuan University,Chengdu 610065,China;2. Officers College of PAP,Chengdu 610213,China)

Abstract:Melody is the most important element in music, and the main melody extraction is one of the core technologies of music retrieval. The pitch sequence of singing voice in polyphonic music constitutes the main theme of vocal music. This paper presents an improved algorithm for automatic extraction of vocal main melody. Firstly, according to the spectrum characteristics of the vocal signal, the calculation method of the pitch saliency function is improved, and the complexity of calculation and the time of vocal main melody extraction are reduced. Secondly, instead of the shallow BP neural network, the sparse self-encoding neural network with better performance is used as the fundamental frequency discrimination model, which improves the recognition accuracy of the main melody model and reduces the false alarm rate of the melody, thus it improves the overall accuracy of the vocal main melody extraction rate. Experiments conducted on the MIR-1K dataset show that the overall accuracy of the vocal themes extracted by the algorithm is at least 1.51% higher than that by the original algorithm, and the average extraction time of the improved algorithm is about 0.12 s less than that of the original algorithm.

Keywords:main melody extraction; pitch saliency; sparse autoencoder; fundamental frequency discrimination