吸收峰混叠的太赫兹光谱区间组合特征提取算法

何伟健，程良伦*，邓广水

广东工业大学自动化学院，广东 广州 510006

摘 要 太赫兹光谱是物质识别的前沿方法之一。由于不同物质的分子组成或结构各异，许多物质的太赫兹吸收谱会在特定频率上出现吸收峰，可以作为混合物成分检测的重要特征。有效准确地提取这些吸收峰的参数，是提高识别率的关键。多峰拟合算法将光谱曲线拟合成若干个标准峰函数之和，能够同时提取到吸收峰的频率、峰高、峰宽等信息。但是该算法以寻峰算法结果为基础确定吸收峰的大致位置和数量，寻峰结果不一定是最优的拟合结果，而且很难准确识别定位混叠状态的吸收峰。为了提高混叠光谱中吸收峰的识别定位精度，提出以大幅度平滑后的曲线波谷为分界点，将预处理后的光谱分成若干个子区间。然后将子区间组合起来进行多峰拟合，通过遗传算法得到最优的拟合子区间组合和吸收峰频率近似值，拟合时每个子区间中通过峰数递增最优化方法确定拟合的吸收峰数，最后微调优化得到最优的吸收峰频率、峰高值。为了实现物质的识别，通过密度聚类算法得到同一类纯净物在多次测量中的共同吸收峰，以此作为标准数据，通过提出的基于吸收峰特征的光谱匹配算法实现了纯净物和不同含量混合物的快速识别。对10类纯净物的实际光谱数据进行拟合聚类，得到其吸收峰参数，结果与太赫兹光谱数据库一致。通过识别算法对纯净物测试集进行识别的识别率为100%，证明了特征提取和物质识别算法的有效性。对于含有混叠峰的混合物光谱，二阶导数法对葡萄糖-乳糖混合物光谱中被掩盖吸收峰(1.280 THz)的识别率仅为70%，提取到的频率平均值为1.316 THz；而该算法提高识别率至95%，频率平均值为1.281 THz，该算法提高了对混叠峰的分辨能力，能够精确定位混叠峰。对10类纯净物构成的6类不同程度混叠的二元混合物前二、三识别率分别达到90.8%和98.3%，提取到的特征能够有效应用于混合物的成分检测。该算法能够以纯净物数据为标准数据实现成分各异的混合物成分检测，对于太赫兹光谱混合物成分检测有重要意义。

关键词 太赫兹光谱；混叠峰；多峰拟合；遗传算法；特征提取；区间组合

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)02-0403-07

引 言

大部分物质的振动和转动能级都处于太赫兹波段范围，不同物质在该波段的吸收光谱位置、形状存在差异。同时太赫兹光子的能量很低，一般不会破坏被检测物质。因此太赫兹光谱是进行物质识别的重要手段，在食品安全、医药健康、公共安全等方面有重要应用[1-3]。

目前太赫兹光谱提取特征常用的方法有主成分分析、偏最小二乘法、神经网络、深度置信网络等方法，然后再使用学习器对提取到的特征进行学习[4]。虽然这些方法在进行物质种类较少时的物质识别有效，但提取到的特征并没有明确的物理意义，很难用于对成分含量不同的混合物进行成分检测。由于分子内原子的三维排列、低频运动以及非共价化学键的影响，许多物质的太赫兹吸收谱在特定频率上会出现吸收峰[5]。吸收峰特征由分子内部结构引起，是物质的固有属性，可以作为混合物成分检测时的重要特征，有效准确地提取这些特定吸收峰的参数，是提高识别率的关键。

曲线拟合是进行吸收峰提取的重要方法。多峰拟合算法可以将复杂曲线拟合成多个标准峰函数之和，如高斯函数、洛伦兹函数等，从而提取到光谱曲线中存在的吸收峰参数[6]。纯净物的吸收峰参数容易提取，但混合物的光谱曲线

收稿日期：2019-01-08，**修订日期**：2019-04-05

基金项目：国家重点研发计划项目(2016YFB1200402-019)，广东省应用型科技研发专项资金项目(2015B090923004)，广东省信息物理融合系统重点实验室项目(2016B030301008)，国家自然科学基金广东联合基金项目(U1801263)，NSFC-广东联合基金项目(U1701262)资助

作者简介：何伟健，1993年生，广东工业大学自动化学院硕士研究生 e-mail：2267560508@qq.com

*通讯联系人 e-mail：llcheng@gdut.edu.cn

是多类物质光谱的近似线性叠加，可能会出现吸收峰相互重叠的情况。目前多峰拟合算法以各种寻峰算法结果为基础确定吸收峰所在的大致位置和数量，在混叠光谱中很难准确定位吸收峰[7-8]。

为了提高混叠状态的吸收峰定位精确度，可以分别对寻峰和拟合两个过程进行优化。Cui 等[9]利用遗传算法对高斯拟合进行优化，得到葡萄糖水溶液近红外光谱的吸收峰参数。Mukoyama[10]在 X 光谱中成功利用模拟退火算法优化高斯拟合算法。通过引入遗传或退火算法进行优化，确定吸收峰参数的最优值，能够有效减少多峰拟合的误差。但这些方法需要在优化前事先指定吸收峰的数目，无法实现对各种光谱曲线的自适应处理。另一方面可以优化寻峰算法的准确性，有研究使用基于小波变换的脊线寻峰法和一种新型脊线校正方法，能够更准确地提取出激光光谱中的混叠峰。而毕云峰等先用对称零面积变换寻峰方法确定吸收峰的个数和初步的峰参数，再以 Levenberg-Marquardt 方法对获得的参数进行拟合优化。上述方法先寻找到吸收峰的位置，再进行高斯峰拟合，通过优化寻峰方法提高了对混叠峰的分辨能力。但这些方法分成了寻峰和拟合两个步骤，寻峰过程中得到的吸收峰信息不一定是最优的拟合结果。

为了更准确地识别定位混叠峰，研究中利用了遗传算法，但与其他文献中使用遗传算法直接求解最优拟合参数的方法不同，是将寻峰和拟合两个步骤融合。首先以大幅度平滑后的曲线波谷为分界点分割光谱曲线，通过遗传算法选择最佳的区间组合进行拟合，得到吸收峰频率近似值，拟合时用峰数递增条件计算每个区间内最佳峰数。最后对吸收峰频率进行微调优化，输出吸收峰频率、峰高参数。使用太赫兹光谱仪测量 10 类纯净物及其构成的混合物并处理。结果表明，该算法能够提高在混叠光谱中吸收峰的分辨能力，对吸收峰的定位更加准确，且具有一定对不同物质自适应处理的能力。

1 算法与原理

1.1 多峰拟合算法

光谱仪测量得到的频谱曲线是一条由离散数据点组成的曲线。多峰拟合算法是光谱曲线拟合的常用算法，通常使用高斯峰或洛伦兹峰对光谱曲线进行拟合。本文采用高斯峰对光谱曲线进行拟合，但本文算法对洛伦兹峰和其他峰型也同样适用。通过一系列单个高斯峰去拟合光谱曲线，单个高斯峰方程为

$$\mathrm{Gau}_i(x) = \exp\left(-\frac{(x-b_i)^2}{2c_i^2}\right)A_i \tag{1}$$

式(1)中 A_i，b_i 和 c_i 分别是第 i 个高斯峰的峰高、频率位置、峰宽。由于未知的参数较多，直接拟合所有参数计算复杂。传统的高斯多峰算法需要先用寻峰算法找到高斯峰的大致位置，确定高斯峰的数目，以寻峰结果为起点迭代优化使得拟合误差最小。

由于本算法在进行拟合前，先用寻峰算法确定高斯峰大致位置及数量，寻峰算法的结果不一定是拟合的最优解。而且寻峰算法通常忽视了峰之间的相互关系，导致对混叠峰的辨别能力和定位能力较弱，峰高、峰宽等参数存在较大偏差。

1.2 遗传算法求解最优拟合子区间

将寻峰过程和拟合过程融合在一起，可以有效地提高多峰拟合提取吸收峰的精度。但是在未确定吸收峰数和吸收峰位置的情况下，对整个光谱同时进行处理计算相当复杂。将光谱划分成若干个子区间，分别进行曲线拟合是个比较好的方法。

假设所有高斯峰幅值均为正值，光谱曲线中不含有噪声，当曲线中每两个高斯峰之间均没有重叠的时候，两个高斯峰可以通过曲线的波谷(幅值相同的多个极小点在本文中只算作一个波谷点)进行划分。因此提出以光谱曲线波谷作为分界点，将曲线划分成一个个子区间。然后子区间之间再进行组合，相互依赖关系较大的区间组合一起进行高斯多峰拟合，使用遗传算法得到最佳的子区间组合，即获取拟合误差最小的区间组合。

有相关文献在光谱特征提取中使用遗传算法，但大多都是用来选择其中若干个特征最明显的区间或频率点[11-12]。也有文献采用遗传算法直接最优化高斯峰参数，实现拟合误差的最小化，这类方法需要事先指定吸收峰的个数[9]。由于不同物质的光谱曲线吸收峰个数一般不同，直接指定吸收峰个数无法实现对物质特征的自适应提取。而本方法与这些方法有所不同，通过波谷将光谱区间分成若干子区间，再通过遗传算法找到子区间的最优组合，实现高斯多峰拟合最优区间的选择。详细算法如下

(1) 种群初始化。基因序列 S 初始化即长度为 r 的全 1 二进制序列，其中 r 为光谱曲线的波谷数。r 个波谷加上光谱曲线的两端一共将曲线分成 $r+1$ 个子区间。每个二进制数字分别依次代表一个子区间的状态(第一个或最后一个子区间的状态固定，组合后至少要有一个子区间)，1 代表该子区间独立作为拟合区间，0 代表该子区间与上一子区间合并作为一个整体拟合区间。全 1 序列代表初始状态是直接以波谷划分区间，区间无任何组合。

(2) 个体变异。每段基因序列的每一个基因都有一定概率进行取反运算。不同于传统的遗传算法，由于种群初始化不是随机值，而是相同的值。因此首先进行基因变异运算。

(3) 种群交叉。每段基因序列之间有一定概率交换范围随机的若干片段。为保证最优个体的存在，每次变异之后加入上一代的最优基因序列，最优基因序列初始化为全 1 序列。

(4) 基因译码。将基因序列转换为子区间组合。设 L 为组合后的子区间分界点，若某个基因为 1，则将对应的分界点加入到 L 中，若为 0 则跳过。

(5) 计算适应度。使用高斯多峰拟合算法对每一种区间组合进行测试，每个子区间在拟合时使用 1.3 节所述的最优吸收峰数判断方法确定峰数。设在某一区间组合下拟合得到的均方根误差为 RMSE，则适应度 fit[见式(2)]

$$fit = \frac{1}{\mathrm{RMSE}} \tag{2}$$

判断是否已经达到迭代最大次数，若是则输出当前最佳适应度对应的区间组合和对应的吸收峰频率值，遗传算法结束。

(6) 自然选择。根据每个基因片段适应度与所有基因片段适应度之和的比例确定每个基因片段的复制次数，同时限制复制最大值、最小值。当新一代种群个体数不足时，复制当前最好个体至固定数目。跳回到第(2)步进行下一代的变异。

1.3　最优吸收峰数判断

通过波谷可以将光谱曲线划分为若干个子区间，但每个子区间内的拟合吸收峰数并不确定。一般来说，吸收峰越多，拟合精度越高，但越有可能使吸收峰的信息受损。例如将一个峰高较大的高斯峰拆分成两个较小的高斯峰。一般做法是通过寻峰算法确定高斯峰数，但结果不一定是拟合的最优解。

由于峰高较大的高斯峰结果较为可信，因此研究中采用了包含峰高在内的复合判据判断最优吸收峰数。通过下列峰数递增最优化方法确定每个子区间内的拟合高斯峰数，即通过改变实际拟合的高斯峰数 m 使得式(3)取最大值

$$P_1 = \begin{cases} \sum_i^n A_i - k_2 m + k_3(1-\text{error}) & m \leqslant n \\ \sum_i^n A_i + k_1 \sum_{n+1}^m A_i - k_2 m + k_3(1-\text{error}) & m > n \end{cases} \tag{3}$$

式(3)中 A_i 表示第 i 个高斯峰的峰高，error 是拟合误差，n 和 m 分别是该子区间在大幅度平滑情况下的波峰数和实际拟合的波峰数。k_1，k_2 和 k_3 是给定常数。

在未确定峰位置时进行高斯多峰拟合，拟合高斯峰的中心位置有可能严重偏离拟合区间，因此在式(3)的基础上，还需要根据偏离程度进行衰减，通过改变 m 使得 P_2 取最大值，见式(4)—式(6)

$$d_i = \frac{\left| b_i - f_{\text{start}} - \dfrac{f_{\text{end}} - f_{\text{start}}}{2} \right|}{f_{\text{end}} - f_{\text{start}}} \tag{4}$$

$$C_i = \begin{cases} 1 & 0 \leqslant d_i \leqslant 1 \\ g_1(d_i) & 1 < d_i \leqslant 2 \\ 0 & d_i > 2 \end{cases} \tag{5}$$

$$P_2 = P_1 \prod_i C_i \tag{6}$$

式(4)中 f_{start} 和 f_{end} 分别为该拟合区间的起始频率和结束频率，式(5)中 $g_1(d_i)$ 是在[1，2]内单调递减的函数，研究中采用线性递减函数。$0 \leqslant C_i \leqslant 1$。

1.4　特征提取算法整体流程

特征提取算法整体流程图如图 1 所示。首先采用惩罚最小二乘法平滑光谱曲线，得到大幅度平滑情况下的曲线波谷点和波峰数。然后通过非对称最小二乘法得到基线，用小幅度平滑时得到的曲线减去基线并去掉其他光谱测量中的异常数据，得到预处理后的光谱曲线。

以上述曲线波谷点为分界点将预处理后的曲线划分为若干个区间，光谱幅值取对数后通过遗传算法不断多峰拟合测试得到最优的子区间组合，同时得到各个区间内的吸收峰频率近似值。拟合时每个区间内的拟合峰数通过最优峰数判断方法确定。最后用预处理后的光谱数据对吸收峰频率、峰高等参数进行微调优化，得到最终拟合的吸收峰参数。

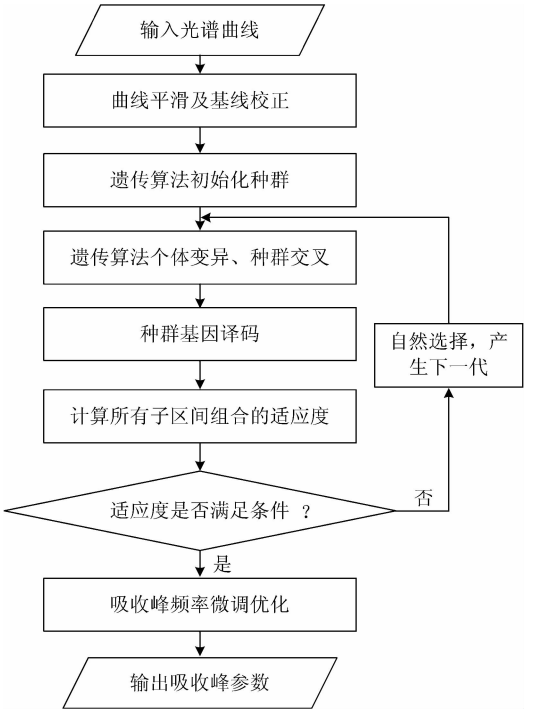

图 1　本文特征提取算法流程图

Fig. 1　The feature extraction algorithm flow chart in this thesis

1.5　吸收峰匹配识别算法

由于每个物质的吸收峰数目不等，很难将从光谱曲线提取到的特征直接输入常用分类器中进行处理。本文提出一种吸收峰匹配识别算法，实现了不同吸收峰数目间光谱曲线的匹配。

设待定物质的吸收峰集合 $P=\{f_1, f_2, \cdots, f_m\}$，从标准物质数据库中取其中一种标准物质进行对比，其吸收峰集合 $P'=\{(f'_1, \varepsilon_1), (f'_2, \varepsilon_2), \cdots, (f'_n, \varepsilon_n)\}$。其中 f_i 是待定物质第 i 个吸收峰的频率，f'_i 是标准物质第 i 个吸收峰的频率，ε_i 是第 i 个标准吸收峰的类间平均频率距离。

依次取出标准物质的第 i 个吸收峰信息，找到待定物质中频率位置最近的吸收峰，即使得式(7)成立

$$\begin{cases} f^* = \underset{f_j \in P}{\operatorname{argmin}} \mid f'_i - f_j \mid \\ \min \mid f'_i - f_j \mid < \varepsilon_0 \end{cases} \tag{7}$$

其中 ε_0 为给定常数，当找不到 f_j 使式(7)成立时，该吸收峰匹配程度为 0，继续计算下一吸收峰。否则该吸收峰的匹配程度计算如式(8)

$$t_i = g_2(\mid f^* - f'_i \mid, \varepsilon_i) \tag{8}$$

其中 $g_2(x, y)$ 是单调递减函数，研究中采用椭圆函数，见式(9)

$$g_2(x,y)=\sqrt{\frac{1-x^2}{9y^2}} \tag{9}$$

循环直至该标准物质的所有吸收峰均判断完毕，待定物质与该标准物质的相似度计算如式(10)

$$\text{similarity}=\frac{\sqrt{1-\left(1-\frac{C_{\text{match}}}{m}\right)}}{n}\sum_{i=1}^{n}t_i \tag{10}$$

式中 C_{match} 是待定物质中已经被匹配的吸收峰数目。依次将该待定物质与数据库中的标准物质比较，待定物质即为相似度最大的标准物质。

2 结果与讨论

2.1 仪器及方法

实验选用日本爱德万公司的 TAS7400 太赫兹光谱仪系统，如图 2 所示。光谱仪频率范围为 0～5 THz，频率分辨率为 7.6 GHz。在 25 ℃、8%相对湿度的相对稳定环境下进行实验。实验中所用的化学试剂均采购于广州东巨实验仪器等专业化学试剂公司，纯度均在 99%以上。固体粉末样品制备过程如下：

(1)取样。用电子分析天平称量一定量粉末。

(2)研磨。用玛瑙研钵研磨粉末，充分研细。

(3)烘干。将固体粉末和其他工具置入烘干机中，在 45 ℃下烘干 1～2 h。

(4)混合。将待混合的固体粉末置于试管中，将试管固定在旋涡混合器上 5 min，充分混合均匀。

(5)压片。使用 10 t 千斤顶压片机和压片模具对固体粉末加压 2 min，压成厚度约 1.5 mm，直径约 13 mm 的固体压片。

(6)测量厚度。用电子游标卡尺测量压片厚度，在不同位置测量三次取平均值。

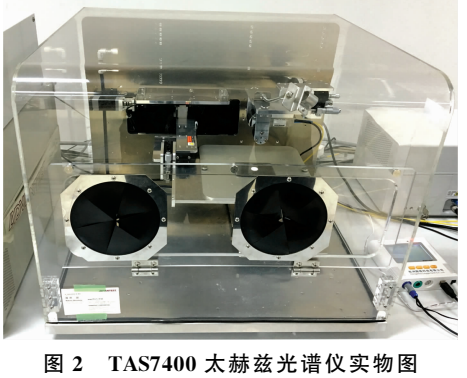

图 2 TAS7400 太赫兹光谱仪实物图

Fig. 2 Photo of the TAS7400 terahertz spectrograph

2.2 纯净物特征提取与识别

取葡萄糖、乳糖、果糖等 10 种常见纯净物按照上述制样方法制作压片，重复测量 30 次。随机选取 20 次测量数据作为训练集，其余 10 次数据作为测试集。采用 1.5 节提出的算法对光谱数据进行处理，由于纯净物的吸收峰一般没有混叠，用较少的标准峰去拟合即可，式(3)中的参数设置为 $k_1=0.4$，$k_2=3$，$k_3=10$。其中乳糖单次测量的吸收系数谱拟合结果如图 3 所示，截取其中信噪比较高的频段[0.3，2]，可以看到在预处理后乳糖的吸收系数谱主要由四个吸收峰组成。拟合后可以同时得到吸收峰频率、峰高、峰宽等吸收峰信息。

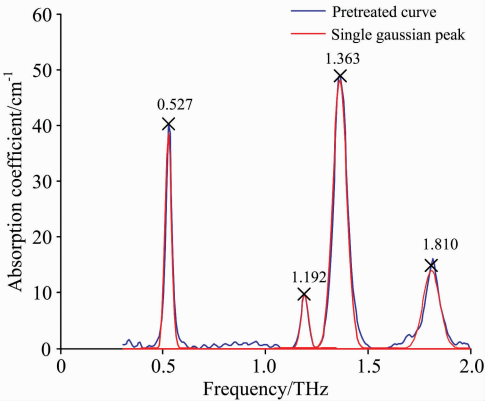

图 3 本算法处理乳糖吸收系数谱的结果

Fig. 3 The result of the algorithm in the thesis for processing the lactose absorption coefficient spectrum

由于在测量中存在各种误差，同一物质每次测量时提取到的吸收峰特征不尽相同，吸收峰的位置、参数可能会出现不同程度的漂移，甚至出现峰高较大的假峰。为了确定某一物质的特征吸收峰数量，并尽可能精确地确定纯净物光谱中的吸收峰，提出使用密度聚类算法确定同一物质在多次测量中的共同吸收峰。由于吸收峰峰高波动较大，仅根据吸收峰频率位置确定吸收峰之间的密度关系，当两个吸收峰频率差值小于 0.01 THz 时，则认为它们密度相连。乳糖 20 次测量的吸收峰聚类效果如图 4 所示，取同一类吸收峰的平均频率作为该物质的特征吸收峰。通过这一方法可以得到纯净物的吸收峰平均频率、平均峰高、平均频率距离等参数，其中吸收峰频率汇总如表 1 所示。由于光谱仪设备所限，频率较高、吸收性较强的吸收峰难以检测。表中所列吸收峰频率与太赫兹光谱数据库(http://thzdb.org/)中的吸收峰频率在误差允许范围内一致[13]。

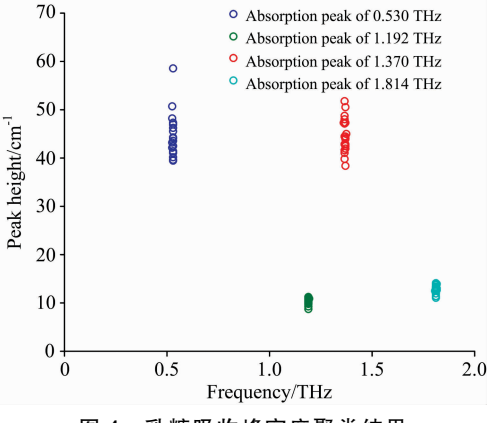

图 4 乳糖吸收峰密度聚类结果

Fig. 4 The result of density cluster absorption peaks of lactose

表 1　纯净物吸收峰汇总

Table 1　Summary of pure substances absorption peak

物质	吸收峰频率/THz
葡萄糖	1.281，1.435，2.080
乳糖	0.530，1.192，1.370，1.814
果糖	1.303
蔗糖	1.442
麦芽糖	1.094，1.587，1.990
山梨酸钾	0.937
苯甲酸	0.629，0.876，1.083
苯甲酸钠	0.833
叔丁基氢醌	0.857，1.047
滑石粉	1.152

以表 1 中所列的数据作为标准吸收峰数据，采用 1.5 节所述的吸收峰匹配识别算法对测试集进行识别，识别结果汇总如表 2 所示。表中的物质含量如果为 100%，说明该样本仅由该物质组成，不含有聚乙烯。麦芽糖、山梨酸钾、滑石粉等物质由于吸收性太强或者不易压片成型，需要加入一定量的聚乙烯，这些样本的剩余含量均为聚乙烯。聚乙烯对太赫兹波的透过性很强，只起到稀释、粘结的作用，基本不影响被测物质的吸收峰频率。表 2 中还列出了 10 次识别过程平均相似度前二的物质，由于葡萄糖、蔗糖具有相似的吸收峰：1.435 和 1.442 THz，使得两者的相似度相当高，但吸收峰匹配算法均能将两者识别开来。总体来说，对 10 种纯净物(不考虑聚乙烯)的识别率为 100%，最大相似度均在 90%以上，证明了所提出吸收峰特征的准确性和吸收峰匹配算法的有效性。

表 2　纯净物识别结果汇总

Table 2　Summary of pure substances recognition results

物质	最大平均相似度	第二平均相似度	识别率
100%乳糖	乳糖(98.8%)	无(0%)	100%
100%葡萄糖	葡萄糖(99.2%)	蔗糖(67.2%)	100%
100%果糖	果糖(98.1%)	葡萄糖(8.87%)	100%
100%蔗糖	蔗糖(92.7%)	葡萄糖(29.7%)	100%
50%麦芽糖	麦芽糖(96.1%)	苯甲酸(15.2%)	100%
50%山梨酸钾	山梨酸钾(94.3%)	乳糖(0.88%)	100%
100%苯甲酸	苯甲酸(94.9%)	麦芽糖(21.6%)	100%
100%苯甲酸钠	苯甲酸钠(98.6%)	叔丁基氢醌(1.33%)	100%
90%叔丁基氢醌	叔丁基氢醌(98.4%)	苯甲酸(21.0%)	100%
75%滑石粉	滑石粉(99.2%)	无(0%)	100%

2.3　混合物特征提取与识别

取上述纯净物若干按照特定的比例制作若干种二元混合物压片，重复测量 20 次。首先对混合物光谱进行预处理，然后分别采用本文算法和二阶导数寻峰算法进行处理。式(3)中的参数设置为 $k_1=0.7$，$k_2=0.5$，$k_3=50$，以尽可能分辨混叠光谱中的吸收峰。图 5 为本文算法处理 40%葡萄糖 40%乳糖压片的结果，剩余含量为聚乙烯；图 6 是将同样预处理后的光谱导入 Origin9.4 中，采用二阶导数寻峰和高斯多峰拟合后剔除明显假峰的拟合结果。两种方法提取的吸收峰结果统计如表 3 所示，表中二阶导数法简称为导数，由于光谱仪频率分辨率为 7.6 GHz，所以平均偏差在 8GHz 左右均为正常范围。光谱中没有混叠的峰很容易被识别出来，对于这种吸收峰两种方法得到的频率非常接近，但二阶导数法存在的假峰更多(图 6 中的 1.794 和 2.127 THz)。而葡萄糖 1.280 THz 的小峰被乳糖 1.370 THz 的吸收峰所掩盖，一般方法难以检测。二阶导数法只能检测出 70%的混叠峰，且定位精度较差；而本算法不仅提高对该混叠峰的识别率至 95%，而且精确到了 1.281 THz，与葡萄糖纯净物的提取结果 1.280 THz 非常接近。

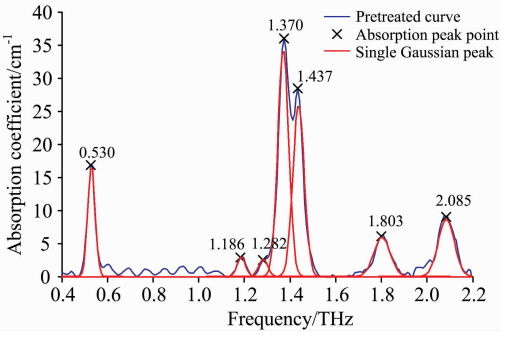

图 5　对葡萄糖-乳糖混合物的多峰拟合结果

Fig. 5　Multi-peak fitting result by the method in this thesis of glucose-lactose mixture

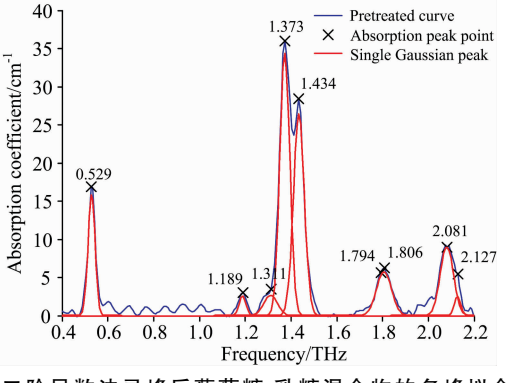

图 6　二阶导数法寻峰后葡萄糖-乳糖混合物的多峰拟合结果

Fig. 6　Multi-peak fitting result of glucose-lactose mixture after peak finding by second derivative method

表 3　不同方法提取的吸收峰频率结果对比

Table 3　Comparison of absorption peaks frequency calculated by different methods

标准值/THz	平均值/THz		平均偏差/GHz		识别率/%	
	本文	导数	本文	导数	本文	导数
0.529	0.529	0.529	1.453	0.811	100	100
1.190	1.188	1.189	2.822	1.471	100	95
1.280	1.281	1.316	2.096	36.16	95	70
1.368	1.370	1.374	3.930	5.595	100	100
1.435	1.446	1.434	10.81	1.374	100	100
1.810	1.813	1.808	6.426	5.704	100	100
2.080	2.080	2.082	3.319	4.910	100	100

表 4 混合物识别结果汇总

Table 4 Summary of mixture recognition results

混合物	最大平均相似度	第二平均相似度	第三平均相似度	前二识别率/%	前三识别率/%
40%葡萄糖 40%乳糖	乳糖 (73.9%)	葡萄糖 (65.2%)	蔗糖 (39.9%)	100	100
50%果糖 50%蔗糖	果糖 (61.0%)	蔗糖 (57.4%)	葡萄糖 (22.1%)	90	95
40%葡萄糖 40%麦芽糖	葡萄糖 (67.1%)	麦芽糖 (58.0%)	蔗糖 (43.0%)	75	100
40%乳糖 40%麦芽糖	乳糖 (61.0%)	麦芽糖 (57.0%)	苯甲酸 (8.85%)	100	100
50%苯甲酸 50%苯甲酸钠	苯甲酸 (73.5%)	苯甲酸钠 (47.4%)	叔丁基氢醌 (22.3%)	90	95
70%苯甲酸 30%山梨酸钾	山梨酸钾 (59.1%)	苯甲酸 (40.3%)	叔丁基氢醌 (8.26%)	90	100

对混合物的其他识别结果汇总如表 4 所示。表中 6 类混合物光谱均存在不同程度的混叠，同时列举了 20 次测量结果平均相似度前三的物质以及前二、前三识别率。由于能够检测到的蔗糖吸收峰只有一个且与葡萄糖非常接近，很难在混合物中确定是否含有蔗糖，导致葡萄糖-麦芽糖的前二识别率偏低。综上所述，对 6 类混合物的平均前二、三识别率分别达到 90.8%和 98.3%，证明了提取到的吸收峰特征能够有效应用于混合物成分检测。

3 结 论

传统多峰拟合算法难以检测和精确定位混叠状态的吸收峰。为了解决上述问题，提出以大幅度平滑后的曲线波谷为分界点将曲线划分成若干个子区间，然后通过遗传算法获得多峰拟合的最优子区间组合和吸收峰频率，在每个子区间内通过峰数递增最优化方法确定拟合的最优峰数。最后以纯净物吸收峰参数为标准数据，通过基于吸收峰特征的光谱匹配识别算法对物质进行识别。实验结果表明，本算法对混叠光谱中混叠吸收峰的分辨能力和定位精度比二阶导数寻峰后拟合的传统算法更优，提取到的吸收峰特征能够准确识别在光谱混叠情况下的混合物成分。该算法对于太赫兹光谱混合物成分检测有重要意义。

但是该方法仅将吸收峰频率特征用于物质识别，很难区分含有频率接近的吸收峰的物质。下一步工作需要研究如何更精确提取混叠状态下的峰高、峰宽等吸收峰其他特征，并设计包含吸收峰频率、峰高、峰宽等特征的物质成分检测算法。

References

[1] Baxter J B, Guglietta G W. Analytical Chemistry, 2011, 83: 4342.
[2] Hangyo M. Japanese Journal of Applied Physics, 2015, 54: 120101.
[3] Liu W, Liu C H, Yu J J, et al. Food Chemistry, 2018, 251: 86.
[4] Huang J L, Liu J S, Wang K J, et al. Spectrochimica Acta Part A-Molecular And Biomolecular Spectroscopy, 2018, 198: 198.
[5] Qiao L B, Wang Y X, Zhao Z R, et al. Optical Engineering, 2014, 53(7): 1.
[6] Zhang J H, Liu Q, Chen Y M, et al. Acta Physico-Chimica Sinica, 2012, 28(5): 1030.
[7] Zhang B, Yu H B, Sun L X, et al. Applied Spectroscopy, 2013, 67(9): 1087.
[8] FENG Fei, WANG Fu-bei, XIE Fei, et al(冯 飞, 王府北, 谢 非, 等). Photonics Journal(光子学报), 2015, 44(6): 113.
[9] Cui X Y, Cai W S, Shao X G. RSC Advances, 2016, 6(107): 105729.
[10] Mukoyama T. X-Ray Spectrometry, 2017, 46(1): 63.
[11] Li Z, Guan A H, Ge H Y, et al. Microchemical Journal, 2017, 132: 185.
[12] Lin P, Chen Y M, He Y. Food and Bioprocess Technology, 2012, 5(1): 235.
[13] Notake T, Endo R, Fukunaga K, et al. IEEE Transactions on Terahertz Science and Technology, 2014, 4(1): 110.

Terahertz Spectral Interval Combination Feature Extraction Algorithm in the Case of Aliasing Absorption Peak

HE Wei-jian, CHENG Liang-lun*, DENG Guang-shui
School of Automation, Guangdong University of Technology, Guangzhou 510006, China

Abstract Terahertz spectrum is an advanced method for material recognition. Due to the different molecular organizations and structures of different substances, the terahertz absorption spectrum of many substances will have many absorption peaks at certain frequency, which can be used as important features of the mixture for component detection. Effective and accurate extraction of the parameters of these absorption peaks is the key to improving the recognition rate. The multi-peak fitting algorithm fits the spectral curve into the sum of several standard peak functions, which can extract the frequency, wave height and wave width of the absorption peaks at the same time. However, based on the results of the peak finding algorithm, fitting algorithm determines the approximate position and number of the absorption peaks before fitting. The peak finding result is not necessarily the optimal fitting result, and it is difficult to accurately identify the aliasing absorption peaks. In order to improve the recognition and positioning accuracy of the absorption peaks in the aliasing spectrum, this thesis proposes to divide the pre-processed spectrum into several sub-intervals by the wave troughs of sharp smoothed curve. Then the sub-intervals are combined for multi-peak fitting, and the optimal fitting sub-interval combination and the approximate value of the absorption peak frequency are obtained by genetic algorithm. The number of absorption peaks is determined by the peak number increment optimization method in each subinterval during fitting. In order to realize the identification of matter, the density clustering algorithm is used to obtain the common absorption peaks of the same kind of pure substance in multiple measurements. Using those peak data as the standard data, the proposed spectral matching algorithm based on the absorption peak characteristics enables rapid identification of pure substances and mixtures of different contents. The actual spectral data of ten kinds of pure substance are fitted and clustered to obtain parameters of absorption peaks, which are basically consistent with the terahertz spectral database. The recognition rate for identifying the test set of pure substances by the recognition algorithm of this thesis is 100%, which proves the effectiveness of the feature extraction algorithm and material recognition algorithm. For the spectrum of mixtures with aliasing peaks, the recognition rate of the second derivative method for the masked absorption peak (1.280 THz) in the glucose-lactose mixture spectrum is only 70%, and the extracted frequency average value is 1.316 THz; The algorithm in this thesis improves the recognition rate to 95% and the average frequency is 1.281 THz, that is to say, this method improves the resolution of the aliasing peak and can accurately locate the aliasing peak. The Top-2 and Top-3 accuracy of the six types of binary mixtures which have different degrees of aliasing and consist of 10 pure materials are 90.8% and 98.3%, respectively. The extracted features can be effectively applied to the component detection of the mixture. The algorithm in this thesis can realize the component detection of mixture by using the data of pure substances as the standard data, which is of great significance to the component detection of mixture in terahertz spectroscopy.

Keywords Terahertz spectroscopy; Overlapped peak; Multi-peak fitting; Genetic algorithm; Feature extraction; Interval selection

(Received Jan. 8, 2019; accepted Apr. 5, 2019)

* Corresponding author