基于 Conformer 的咳嗽声检测

乔凯，李哲宏，姚袁，徐金诚，王明欢
广州城建职业学院，广东 广州 511660
DOI: 10.61369/TACS.2025080019

摘　　要： 通过声音信号检测咳嗽对于医疗和健康监护的应用至关重要，包括疾病诊断和远程患者监测。虽然很多深度学习算法在实验数据上能够取得90% 以上的准确率，但是一旦应用于实际环境中，咳嗽检测的准确率就会大大降低。很多种类的声音（例如：短促的说话声、笑声、关门声等）都被识别为咳嗽。卷积神经网络（CNN）联合循环神经网络（RNN）虽能能够提升检测性能，但在捕捉咳嗽各个阶段的依赖关系和时序动态特征方面仍面临挑战。为此，本文提出用于咳嗽检测的 Cough-Conformer（卷积增强型 Transformer）架构，通过卷积层实现局部特征提取，并结合自注意力机制进行全局上下文建模。我们在咳嗽数据集 COSWARA、语音数据集、噪声数据集、和笑声数据集的基础上提取声音数据作为实验数据集；在该数据集上训练 Cough-Conformer，准确率达到了97.64%，F1得分为 0.98，然后在计算机房录制的音频数据集上验证，其中准确率达到了87.01%，F1得分为 0.87。实验结果表明，Cough-Conformer 在咳嗽检测任务中相比于传统 CNN 和 RNN 模型，在准确率和 F1得分均有显著提升，尤其在复杂背景噪声环境下表现出更强的健壮性。通过引入多头自注意力机制，模型能够有效捕捉咳嗽声音的时序动态特征与上下文依赖关系。进一步分析显示，卷积层与 Transformer 模块的协同作用提升了对咳嗽不同阶段特征的辨识能力，为远程患者监测中的咳嗽检测提供了更优秀的解决方案。

关 键 词： 咳嗽声检测；Cough-Conformer 模型；时序数据

Robust Cough Detection Using Conformer-Based Models

Qiao Kai, Li Zhehong, Yao Yuan, Xu Jincheng, Wang Minghuan
Guangzhou Urban Construction Vocational College, Guangzhou, Guangdong 511660

Abstract: Cough detection from audio signals is critical for healthcare applications, including disease diagnosis and remote patient monitoring. Although many deep learning algorithms can achieve an accuracy of over 90% on experimental data, once applied in practical environments, the accuracy of cough detection will greatly decrease. Many types of sounds (such as short speech, laughter, closing doors, etc.) are mistaken for coughing. Convolutional Neural Networks (CNN) combined with Recurrent Neural Networks (RNN) can improve detection performance, but they still face challenges in capturing the long-range dependency relationships and temporal dynamic features of a cough across its various stages. This paper introduces the Cough-Conformer(Convolution-augmented Transformer for Cough Detection) architecture for cough detection, combining convolutional layers for local feature extraction with self-attention mechanisms for global context modeling. We extracted sound data as the experimental dataset based on the cough dataset COSWARA, speech dataset, noise dataset, and laughter dataset. Training Cough-Conformer on this dataset achieved an accuracy rate of 97.64% and F1-score is 0.98. Then evaluated on the cough datasets recorded during our computer room, our Cough-Conformer achieves excellent results, with 87.01% accuracy and F1-score is 0.87. The experimental results show that Cough-Conformer has significantly improved accuracy and F1-score compared to traditional CNN and RNN models in cough detection tasks, especially exhibiting stronger robustness in complex background noise environments. Because a multi-head self-attention mechanism, the model can effectively capture the temporal dynamic features and contextual dependencies of cough sounds. Further analysis shows that the synergistic effect of convolutional layers and Transformer modules enhances the ability to identify features of different cough stages, providing a better solution for cough detection in remote patient monitoring.

Keywords: cough detection; cough conformer; time-series data

一、绪论

很多呼吸道疾病的诊断都需要通过咳嗽，例如：哮喘、肺结核和新冠肺炎等等。咳嗽时间与频率是医生洞察病情发展及评估治疗方案效果重要依据，而用于咳嗽声检查的设备可以将医生从手写咳嗽记录的繁琐工作中解救出来。近几年伴随着新冠肺炎疫

情的结束，关于咳嗽声研究的火热已经淡去，然而我相信咳嗽仍然值得研究。首先，很多用于咳嗽声检测的深度学习模型虽然在实验数据上取得了90%以上的准确率，但是一旦应用到实际环境，其性能将会大打折扣，很多短促的说话声、笑声、敲击声会被卷积神经网络误识别为咳嗽。其次，作为时序数据的咳嗽声包含了大量的信息。这要从咳嗽的3个阶段说起，咳嗽大体分为3个阶段[1]：第一个阶段努力吸气的吸气阶段，第二个阶段是对着闭合的声门用力呼气的压缩阶段，然后是通过主动声门打开和快速呼气流的排出阶段。虽然咳嗽声的三个阶段加起来不到1秒钟，但是每个阶段的时序信息变化量非常大，同时伴有大量的能量信息和频率信息，如图1所示。一次咳嗽的时序信息量不亚于一段5个字左右的语音。咳嗽声虽然时间短暂，但是包含的时序信息依然非常丰富。

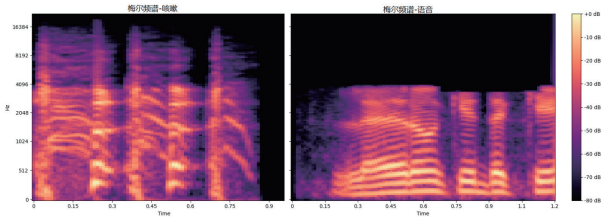

图1 咳嗽与说话声梅尔频谱的对比图

在人工智能领域，特别是在语音识别和环境声音分类任务中，咳嗽常被归类为一种噪声，比如在ESC-50数据集中[2]，咳嗽被标注为噪声。咳嗽虽然不被视为语音[3]，但它作为人类声音的一种形式，具备了语音的部分特征和咳嗽独有的特征。从图1可以看出咳嗽声与语音有明显的不同，首先，语音源于发音器官的协调运动，并表现出更强的规律性，因此，语音具有清晰谐波结构；其次，咳嗽具有特定的生理功能：清除气道中的阻塞物。咳嗽的产生过程从人体进行短暂深吸气后开始，然后声门紧闭，呼吸肌强烈收缩，从而排出肺部高压空气，这是一个复杂而精细的过程，需要呼吸系统多个部分协同工作[4]。这种反应通常更为剧烈，因此产生的声音往往比说话声更刺耳、更响亮。在频谱上，咳嗽的谐波结构并不明显。在高频区域（4096Hz以上的区域），咳嗽声仍存在较强的能量分布，而语音在该区域的能量分布十分微弱。我们看到的语音梅尔频谱图，在4096Hz以上的区域之所以没有能量分布，那是因为采取了去噪声处理（主要指高频噪声）[5,6]，而咳嗽本身就包含噪声，如果采取去噪声，会失去咳嗽声独特的特征。

卷积神经网络的特长不是获取时序信息，循环神经网络获取时序信息的长度非常有限，CRNN虽然在应对时序信息时以轻量级的参数量取得了较好的结果，但在应对长序列信息时能力有限。而transformer类的模型就是为了获取长文本信息的，已经成功地应用到声音事件检测领域被称为“Audio Spectrogram Transformer”[2]。因此，可以推断transformer类的模型能够获取咳嗽声短时间内丰富的时序信息。我们采用梅尔频谱作为咳嗽声的特征提取，首次将Conformer模型应用到咳嗽检测，并且通过多次实验，我们发现轻量化的Cough-Conformer效果仍然优秀，其参数量仅仅是Conformer的三分之一。

二、相关研究

早在2016年，论文[7]就提出了有关深度学习应用于咳嗽声检测的研究，文中探讨了针对短时傅立叶变换STFT（Short-Time Fourier Transform）两种解决方案：一种是采用卷积神经网络CNN（Convolutional Neural Network），准确率达到了87.6%；另一种是采用循环神经网络RNN（Recurrent Neural Network），准确率达到了79.7%。2021年，论文[8]提出了针对梅尔频谱（Mel-spectrogram）的卷积神经网络CNN，F1得分达到了98.18%。同年的论文[9]在梅尔频率倒谱系数MFCC（Mel Frequency Cepstral Coefficients）的特征数据上提出了一种卷积神经网络CNN，F1得分达到了98.17%。而2022年，同济大学的论文[10]介绍的模型C-BiLSTM是一种卷积神经网络CNN与循环神经网络RNN相结合的方法，在梅尔频率倒谱系数MFCC的特征数据上，准确率最高达到了99.64%。2024年，论文[11]提出了应用说话声检查方法（Voice Activity Detection）来对声音分段，然后再用分类模型(Multi-Layer Perceptron)识别咳嗽声，咳嗽事件检测错误率（Detection Error-rate Scores）得分为0.31。论文[12]提出的CRNN咳嗽声识别的准确率接近了98%

然而，上述研究都是在实验数据上获取的实验结果，如果遇到噪声干扰，干扰类声音，势必将严重影响模型的性能。由于以上模型所采用的循环神经网络在时间序列数据的处理上都无法与transformer类模型相比。而同时具备卷积和transformer模块的混合模型：Conformer模型在语音识别领域表现出色，但其在咳嗽检测方面的应用仍属空白。所以我们有依据可以推断Conformer不仅能够胜任咳嗽检测，而且依靠transformer模块的优势，可以轻易辨别干扰类声音。

三、实验介绍

本实验的基线模型选取论文[9]的CNN和论文[12]的CRNN。采用声音事件检测方法来检测所有发生的声音事件。对声音事件做一秒的分段（类似于论文[9]描述的方法），接着交给模型来判断是否为咳嗽。为了增强模型的抗干扰能力。选取了多种干扰声音与咳嗽声一同训练。所有模型训练30轮，每轮训练后在验证集上评估准确率、F1分数，30轮结束后挑出综合评估最好的模型保存，然后在测试集上计算准确率、F1分数。对比Cough-Conformer与CNN、CRNN的实验结果，验证Cough-Conformer是否真的优于其它模型。并在自己录制的机房音频数据集上，针对三个模型分别展开咳嗽检测，看看这真实场景中模型的泛化能力怎么样。

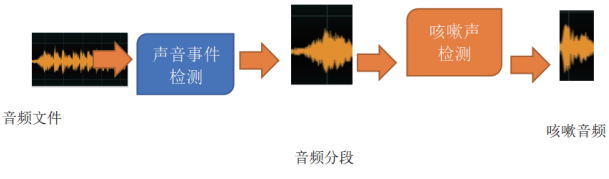

图2 咳嗽声检测流程示意图

（一）数据集

实验数据集源自咳嗽数据集Coswara，噪声数据集UrbanSound8K和ESC-50[2]（包含咳嗽声）、普通话语音数据集“Free ST Chinese Mandarin Corpus”、还有由笑声、叹气、咳嗽、清嗓子、打喷嚏和嗅闻声组成的VocalSound数据集（包含咳嗽声）。我们在此5个数据集的基础上进一步做了数据清洗，对所有声音事件统一以1秒为时间单位，应用声音处理包librosa的onsets函数做音频分段，整理出咳嗽声11962条记录、笑声8479条、语音207913条、汽车喇叭952条、狗吠声音2473条。然后采取房间冲激响应（Room Impulse Response），添加噪声、时间域增强（SpecAugment算法的时域近似，采用SpeechBrain工具包的TimeDomainSpecAugment函数）3种方式对数据样本量少的声音种类开展了声音数据增强，最终整理出咳嗽声35886条记录、笑声9079条、语音9000条、汽车喇叭8952条、狗吠声音9073条，组成了实验数据集，咳嗽与非咳嗽声的样本量如图所示。咳嗽声（35886条）与非咳嗽声音（36104条）数据样本量比例接近于1:1。这样在训练模型时，不会出现由于数据样本量不均衡而导致的模型偏向于样本量大的那一类数据。实验数据集经分层抽样，然后随机打乱顺序，按照8:1:1的比例划分为训练集（57590条），验证集（7197条）和测试集（7203条），实验数据集只有两个标签，一个是咳嗽，另一个是非咳嗽。

为了进一步，验证模型的泛化能力，我们在机房录制了含有咳嗽的音频文件。以此音频文件作为真实环境的数据集对模型展开测试。该数据集总共有18声咳嗽，159条非咳嗽事件。

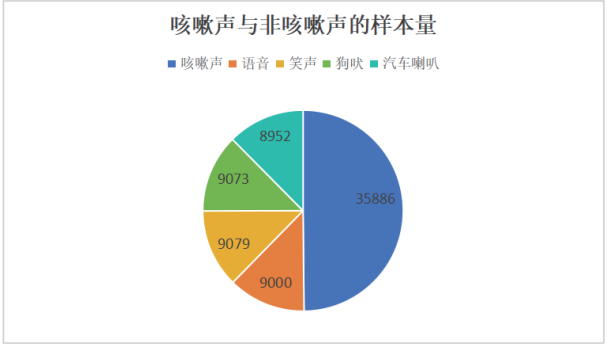

图3 实验数据集的咳嗽声与非咳嗽声的样本量

（二）特征提取

所有的数据除了语音外，都采用高于16kHz的采样率，因此预处理将所有非16kHz的采样率音频数据重采样至16kH。在特征表示上，摒弃了在语音识别中广用的梅尔频率倒谱系数（MFCC），因为咳嗽声与语音有本质的区别，咳嗽在高频和低频区域仍然有很强的能量信息分布。采用声音事件检测（Sound Event Detection）领域的梅尔频谱图作为本研究的特征输入，以保留更全面的声音的频率域信息。该频谱图经128个梅尔滤波器组生成，采用窗口长为25毫秒，帧移为10毫秒。并且在提取特征前不采取去噪声措施。

（三）Cough-Conformer

在Conformer提出之前，语音识别领域主要有两种强大的模型架构，一个是Transformer类的模型，它们基于自注意力机制，擅长捕捉序列中长距离的全局依赖关系；另一个是卷积神经网络（CNN）基于卷积神经网络，擅长提取局部、位置相关的特征。然而，它们各有不足，Transformer对局部细节的建模相对较弱，而CNN捕捉长距离依赖关系的能力不如自注意力机制。Conformer（Convolution-augmented Transformer）模型，其名字本身就是Convolution和Transformer的组合，其最关键的部分是Conformer模块。该模块内集成了多头自注意力（MHSA），有效捕捉音频序列中的全局上下文信息；卷积模块（Conv）能够高效地提取局部特征；前馈网络（FFN）提供非线性变换。

通过反复实验在Conformer模型的基础上做了改进得到了Cough-Conformer，仅仅堆叠了3个Conformer模块；嵌入层（Embedding）的维度减半为256；多头自注意力机制（MHSA）采用32个注意力头；设置Dropout为0.2，在训练时对全连接层随机“丢弃”20%的神经元，让它们临时失效，以防止过拟合。总体参数量仅有28.3M比Conformer（L）的118M小得多。

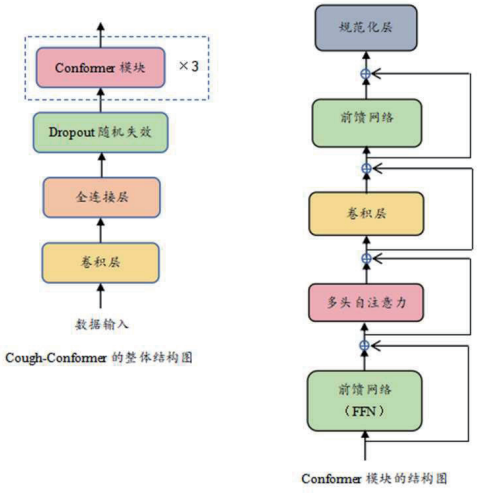

图4 Cough-Conformer模型结构示意图，左图为Cough-Conformer的整体结构图，右图为Conformer模块的结构图。

（四）损失函数与优化器

本文采用BCEWithLogitsLoss（The Binary Cross-entropy Loss with Logits）损失函数和AdamW优化器。

Pytorch中二分类常用的损失函数BCEWithLogitsLoss（The Binary Cross-entropy Loss with Logits）将Sigmoid激活函数和二元交叉熵损失结合在了一起。该函数直接接收原始logits作为输入，而非经过Sigmoid处理后的概率值。这种实现方式比先应用Sigmoid函数再计算二元交叉熵损失具有更好的数值稳定性，尤其在logits值极大或极小时优势更为明显。针对单个样本的BCEWithLogitsLoss计算如公式（1）所示。x是计算得到的logits，y是真实的目标标签（0或1）。

$$BCEWithLogits(x,y)=max(x,0)-x\cdot y+log\left(1+e^{-|x|}\right) \quad (1)$$

AdamW优化器纠正了Adam中一个权重衰减受自适应梯度更新影响的问题，这种正则化实现方式泛化能力差。而AdamW通过回归到最原始的权重衰减概念，并使其与自适应

梯度更新解耦，从而极大地提升了算法的泛化能力和实用性。这也是为什么 AdamW 如今已成为训练深度学习模型的首选优化器。

（五）评价指标

本研究的评价指标采用准确率和 F1-scroe，在模型训练完成后，选取最优模型在测试集上验证。

论文[7,10,12]都采用了准确率（Accuracy）作为咳嗽声检测的评价指标之一。准确率是用于衡量模型预测结果与实际结果之间的匹配程度，它衡量了模型预测结果为正例的可靠性，是基础的、最常用且直观的评价指标之一。准确率的定义为：对于给定的测试数据集，分类器正确分类的样本数与总样本数之比。具体来说，假设我们有 N 个咳嗽声，其中正确预测咳嗽人的咳嗽声样本数量为 A，那么咳嗽人识别的准确率 Acc 就可以表示为：

$$Acc = \frac{A}{N} \tag{2}$$

早在2017年，F1得分 (F1 score) 就应用于声音事件检测的实验中，2021年声音事件检测的综述论文再次采用了 F1得分。论文[8,9]也选择 F1得分咳嗽声检测的评价指标。F1得分常用于深度学习的分类任务，F1得分涵盖了精确率 (Precision) 和召回率 (Recall)。我们先从 TP、TN、FP 和 FN 的定义开始了解 F1得分。

- TP（True Positive）：正确被检测为咳嗽的样本。
- TN（True Negative）：正确被检测为非咳嗽的样本。
- FP（False Positive）：错误地被检测为咳嗽的非咳嗽样本。
- FN（False Negative）：被检测为非咳嗽的咳嗽样本。

精确率 (Precision) 和召回率 (Recall) 以及 F1得分可以由如下公式求得。

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

（六）模型训练

基线模型和 Cough-Conformer 都采用相同的超参数。即 AdamW 优化器的学习率设置为1e-3，权重衰减为1e-4。训练过程中使用余弦退火进行学习率调整，初始预热阶段为5个 epoch。批量大小设为16，总训练轮数为30。输入音频数据经计算得到梅尔频谱，然后送入模型的网络训练，损失函数 BCEWithLogitsLoss 计算模型预测值与真实值之间差异程度，AdamW 优化器依据损失值（Loss）引导模型向正确的方向调整。每轮训练后在验证集上评估准确率、F1分数，30轮训练结束后，选取最佳模型在测试集上进行验证。

四、实验结果分析

通过实验我们得到了准确率和 F1得分的数据表，模型在训练集上和测验证集上训练结束时绘制的损失曲线。以及训练结束后，在测试集上计算得到的 roc 曲线。

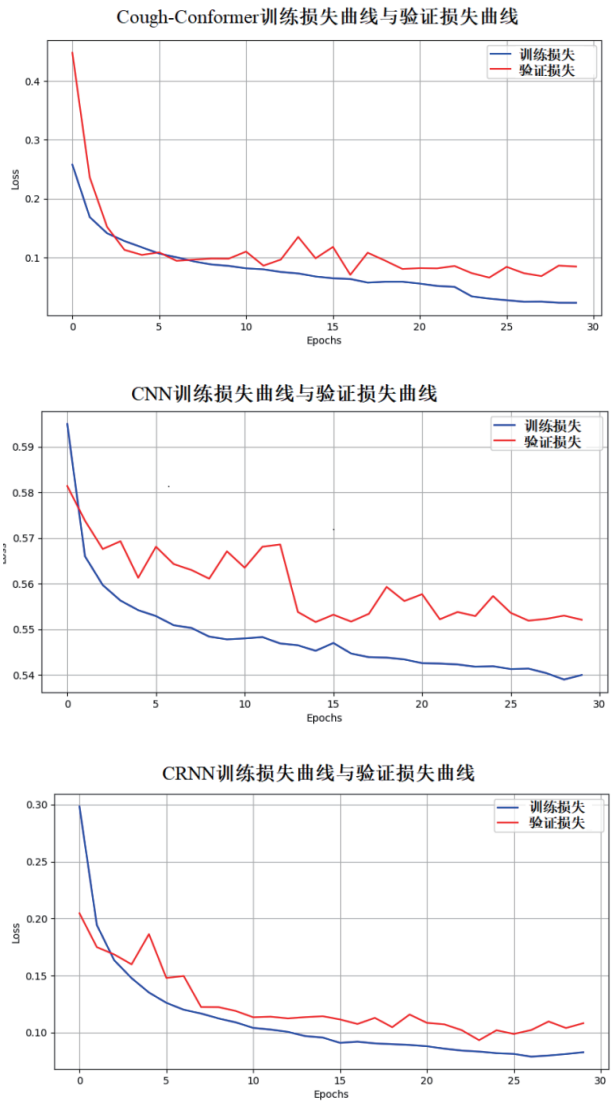

图5 Cough-Conformer 与基线模型的损失曲线

从图5可以看出，Cough-Conformer 仅仅在30轮训练就可以将训练损失曲线和验证损失曲线同时降到损失值0.1以下。CNN 由于训练次数不足导致损失曲线一直徘徊在损失值0.55左右。CRNN 表现较好，训练损失曲线已经达到了0.1以下，但是验证损失曲线主体都在0.1以上。由此可见，Cough-Conformer 不仅可以在较少的训练次数下就将损失曲线降下来，而且验证损失曲线和训练损失曲线的距离很近，表明其具有较好的泛化能力。

表1 咳嗽检测模型对比实验

模型	数据集	准确率	F1得分	精确率	召回率
Cough-Conformer	测试集	97.64%	0.98	97.19%	98.11%
CNN	测试集	90.62%	0.91	91.46%	89.53%
CRNN	测试集	95.88%	0.96	95.29%	96.49%
Cough-Conformer	机房实录	87.01%	0.87	86.49%	87.01%
CNN	机房实录	70.06%	0.71	72.04%	70.06%
CRNN	机房实录	81.92%	0.83	83.16%	81.92%

通过表一可以看出，在测试集上 Cough-Conformer 无论是准确率还是 F1 得分都取得了最好的成绩。本文一开始就提到的虽然深度学习可以在实验数据上取得90% 以上的准确率，但是一旦换成真实环境的音频数据，性能将会大大降低。因此，Cough-Conformer 的性能在真实的机房录制的音频数据上，性能也下降了，但是准确率为87.01% 已经接近90%。这样的成绩已经非常优秀了。

五、结论

本文首次提出了基于 Conformer 网络结构的咳嗽声检测模型 Cough-Conformer，模型的参数量不到 Conformer 的一半。Cough-Conformer 综合利用卷积神经网络的局部性特征提取优势与基于注意力的全局上下文信息。不仅能够胜任咳嗽检测，而且依靠 transformer 类模型的优势，可以轻易辨别干扰类声音。不同于 CNN 和 CRNN 模型需要训练上百次才能找到最优解，Cough-Conformer 仅训练30次就可以在验证集上将损失（Loss）降到0.1以下，并且泛化能力优异，在我们自行在机房录制的音频数据上，咳嗽声检测的准确率达到87.01% 的，可以应用于真实场景。

参考文献

[1]Zhang P C , Wang Y H , Liu X ,et al.Conformational study of 8-C-glucosyl-prunetin by dynamic NMR spectroscopy[J].Chinese Chemical Letters, 2002, 13(7):645-648. DOI:10.1021/cm020249a.

[2] Li-Xin Y .Conformation Analysis and Comparison of Epristeride(17β-N-t-Butylcarboxamide-androst-3,5-diene-3-carboxylic Acid) and Its Analogs[J].高等学校化学研究：英文版 , 2005, 21(5):3.DOI:CNKI:SUN:GHYJ.0.2005-05-005.

[3]ZHANG, Wang P C , Liu Y H ,et al.Conformational Study on 8-C-glucosyl-prunetin by Dynamic NMR Spectroscopy[J].Acta Chimica Sinica, 2003.

[4] 俞涵 . 中英文混合的民航空管语音识别研究 [D]. 厦门大学 ,2022.

[5] Sun Y , Zhang F , Zhang L ,et al.Synthesis of calix[4]arene derivatives via a Pd-catalyzed Sonogashira reaction and their recognition properties towards phenols[J]. 中国化学快报 (英文版), 2014.

[6] Zhu Y B .STEREOCHEMICAL CONTROL IN PROPYLENE POLYMERIZATION CATALYZED BY UNBRIDGED METALLOCENE CATALYSTS[J]. 高分子科学 (英文版), 2001.

[7]Sun Y , Pan W , Fu J ,et al.Conformation preference and related intramolecular noncovalent interaction of selected short chain chlorinated paraffins[J]. 中国科学：化学 (英文版), 2016.

[8] Zge B , Grkan K , Cemal P ,et al.Vibrational investigation of 1-cyclopentylpiperazine:A combined experimental and theoretical study[J]. 中国科学：物理学 力学 天文学 (英文版), 2014.

[9]STUDIES ON THE CONFORMATIONS OF SUBEROGORGIN AND ITS METHYL ESTER BY MNDO METHOD[J]. 科学通报：英文版 , 1990(23):4.DOI:CNKI:SUN:JXTW.0.1990-23-016.

[10]Renqing L , Zuogang C , Guoping S .Ab Initio Calculation of Room Temperature Ionic Liquid 1-Ethyl-3-Methyl-Imidazolium and AlCl3[J].China Petroleum Processing & Petrochemical Technology, 2007, 16(3):51-56.DOI:10.1007/s10553-007-0078-7.

[11] Hongwei K E , Li R , Amp X X .Density functional theory study of 1:1 glycine - water complexes in the gas phase and in solution[J]. 中国科学：化学 (英文版), 2010.

[12]Subhasish,Bandyopadhyay,Asit,et al.Intra-species sequence variability in 28s rRNA gene of Oesophagostomum venulosum isolated from goats of West Bengal,India.[J]. 亚太热带医药杂志：英文版 , 2010(7):515-515.

