双传感器特征融合的语音识别系统研究*

余伟，陈向东，丁星

（西南交通大学 信息科学与技术学院，成都 610031）

摘要：针对噪声环境下麦克风系统识别率降低的问题，提出一种基于双传感器特征融合的语音识别系统。利用 STM32 单片机同时采集说话人发声时的皮肤振动语音信息和麦克风语音信息，通过 WiFi 发送至上位机，将双路语音特征 MFCC 参数融合并与隐马尔可夫模型结合用于孤立词识别研究。实验结果表明，在安静和噪声环境下，与单一麦克风语音识别系统相比，此系统具有更高的识别率和更强的抗噪能力，鲁棒性更好。

关键词：STM32；麦克风语音；隐马尔可夫模型；孤立词识别；ESP8266

中图分类号：TP391　　**文献标识码**：A

Research on Speech Recognition System Based on Dual-sensor Feature Fusion

Yu Wei，Chen Xiangdong，Ding Xing

（School of Information Science & Technology，Southwest Jiaotong University，Chengdu 610031，China）

Abstract：Aiming at the problem that the recognition rate of microphone system decreases in noisy environment，a speech recognition system based on dual-sensor feature fusion is proposed. The STM32 single-chip microcomputer is used to simultaneously collect the skin vibration voice information and the microphone voice information of the speaker and send them to the host computer through WiFi. The MFCC parameters of the two-way voice feature are fused and combined with the hidden Markov model for isolated word recognition research. The experiment results show that compared with the single microphone speech recognition system，this system has higher recognition rate，stronger anti-noise ability and better robustness in quiet and noisy environments.

Keywords：STM32；microphone speech；hidden Markov model；isolated word recognition；ESP8266

0 引　言

语音识别是人机交流的自然接口来源，也是人机交流最自然的接口之一。然而，在环境中存在各种各样的噪声，语音信号的频谱结构在不同类型和强度的环境噪声作用下会产生不同程度变化。当含噪语音特征输入语音识别系统时，往往造成系统的非预期输出，从而导致系统识别精度下降。因此，提高语音识别系统在嘈杂环境下的识别效果具有重要的研究意义。

目前，研究人员通过语音增强、鲁棒性特征提取和鲁棒性声学建模等方法来提高语音识别系统的鲁棒性[1]，但这些方法都是建立在麦克风语音数据基础之上。随着技术的发展，研究人员提出利用多个语音信息源来提高语音识别系统的鲁棒性。参考文献[2]提出一种结合喉部与标准麦克风利用概率最优滤波器（POF）映射算法从带噪语音中估计干净语音特征用于提升语音系统的鲁棒性，但该系统算法需要一个同时包含干净和嘈杂录音的数据库。参考文献[3]提出一种结合喉部语音和空气传导语音以及视觉唇读特征的多模态英文数字语音识别系统，但该系统忽略了噪声对麦克风语音的干扰，准确率仅达到 94%。

本文提出一种基于双传感器的孤立词识别系统，融合了皮肤语音和降噪后的麦克风语音特征，在 0 dB 以上的带噪环境下识别率高于 80%，安静环境识别率可达 96.8%。

1 双传感器语音识别系统设计

本文提出的双传感器语音识别系统原理是在发声过程中通过柔性压电薄膜传感器和麦克风分别获取下颚皮肤的微弱振动压力语音信号（皮肤语音）和经空气传播的语音信号（气导语音），通过融合两种语音信息源特征并利用算法实现孤立词的识别，以期获得较高的识别率与较强的抗噪能力。本系统主要包括信号调理模块、单片机控制模块和 Matlab 上位机三部分，系统框图如图 1 所示。

2 系统硬件设计

2.1 信号调理电路设计

信号调理电路的功能是将双传感器获取的语音信号幅值变换至可采集的电压预设区间内，设计框图如图 2 所示。

* 基金项目：气体钻井安全监测的前兆预警关键传感器研究（61731016）。

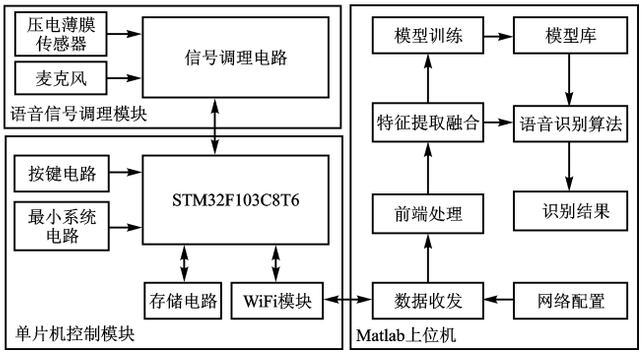


图 1　系统框图

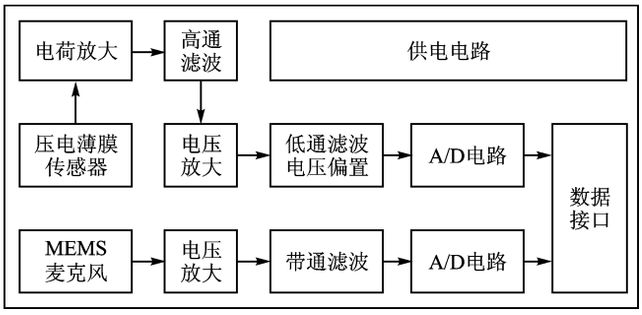


图 2　语音信号调理电路框图

皮肤语音信号调理电路主要包括电荷放大、高通滤波、电压放大、低通滤波电压偏置，模/数转换等模块。其中，电荷放大电路是将压电薄膜受到皮肤振动压力后产生的电荷信号转换成电压信号，由于压电薄膜内阻较大，设计采用高输入阻抗(＞10 TΩ)、低偏置电流(10 fA)的精密CMOS运放LMC6081实现阻抗匹配。

气导语音调理电路主要包括电压放大、带通滤波和模/数转换等。其中气导语音通过歌尔全向MEMS硅麦克风拾取；放大电路采用放大倍数分别为9和10的单级放大电路级联实现；带通滤波器由二阶压控高通和低通滤波器构成，通带为100～4 000 Hz；模/数转换电路采用两颗16位高速ADS8860芯片，参考电压设计为4.096 V。

2.2　单片机控制模块设计

单片机控制模块采用STM32F103C8T6芯片作为主控制器，最小系统电路为主控芯片提供时钟输入、复位、程序下载/调试接口。数据存储电路采用W25Q64芯片通过主控内置SPI2控制器实现语音数据存储功能。通信电路使用ESP8266 WiFi模块实现与上位机的数据传输。为了降低数据连续传输带宽限制，系统采用按键控制语音数据的采集和发送。语音采集流程如图3所示。

图4展示了在安静和嘈杂环境下通过语音采集系统硬件同时录制的两路语音信号波形以及对应的时频图。由图4(a)可知，气导语音受环境噪声影响较大，而皮肤语

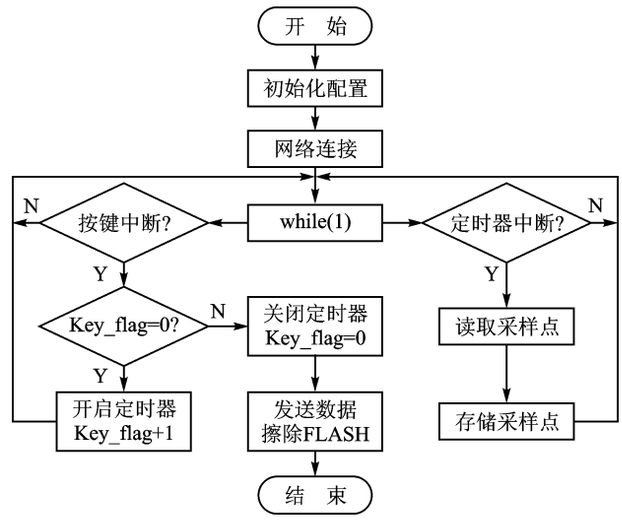
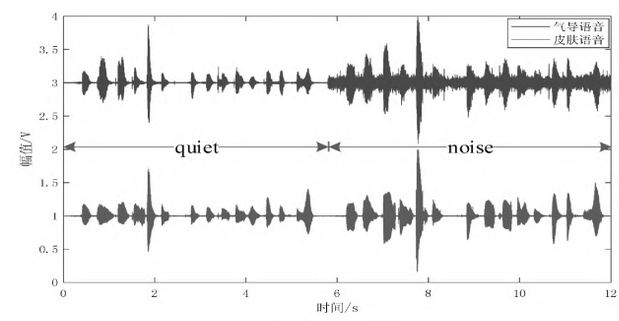
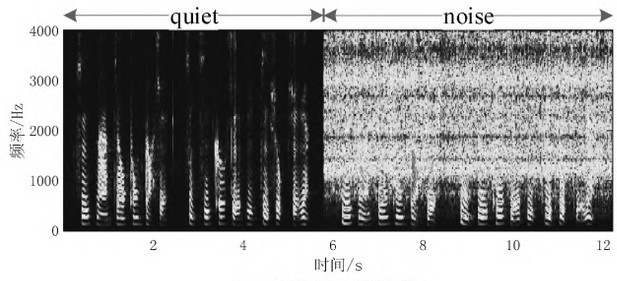


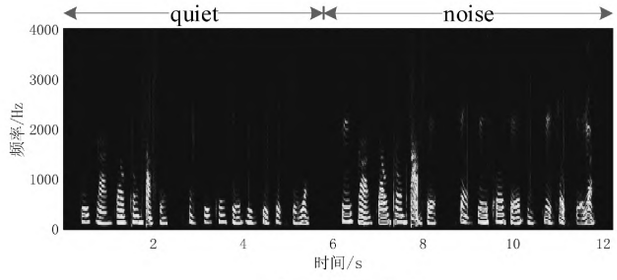
图 3　语音采集流程



(a) 语音波形图



(b) 气导语音语谱图



(c) 皮肤语音语谱图

图 4　安静与嘈杂环境下的语音信号

音不受外界环境噪声的影响。从图4(b)中可以看出，在安静环境下经麦克风采集的语音信号频谱纹路结构清晰，

但其在嘈杂环境下受到了较为严重的干扰。而由图 4(c)可知，无论是安静还是嘈杂环境下，皮肤语音的频谱结构都相对清晰明显，但其频谱能量集中在 2 500 Hz 以下。

3 系统算法设计

3.1 数据前端处理

(1) 语音预处理

上位机接收到语音数据后，首先将语音数字量化值从[0,65535]转换到[−1,1]，消除发音幅度差异。接着对语音信号进行预加重，提升语音高频信息[4]。预加重过程可由式(1)表示。式中 x(n)、y(n)分别表示预加重前后的语音数据。

$$y(n)=x(n)-0.95x(n-1) \tag{1}$$

由于语音采样频率为 8 kHz，本系统以 256 点为一帧，帧重叠点数为 128，加窗函数采用 hamming 窗以减少频谱泄漏[7]。式(2)和式(3)表示加窗过程：

$$y_i(n)=x_i(n)\times w(n) \quad (n=0,1,\cdots,N-1) \tag{2}$$

$$w(n)=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (n=0,1,\cdots,N-1) \tag{3}$$

式中，$x_i(n)$、$y_i(n)$分别表示加窗前后的第 i 帧数据；w(n)表示加窗函数。

(2) 语音降噪

为了减弱气导语音中的噪声干扰，提升语音系统整体识别率，系统采用改进型谱减法实现对气导语音的降噪[5]。改进型谱减法流程如图 5 所示。

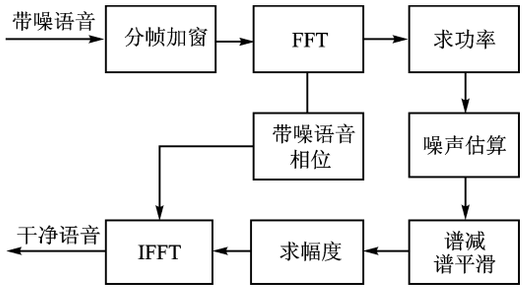


图 5 改进型谱减法降噪流程

(3) 端点检测

语音端点检测是从一段音频数据中正确检测出有效发声部分，可降低系统的计算量[6]。本系统利用皮肤语音不受环境噪声影响的优点，代替传统基于麦克风语音的端点检测。系统采用基于短时能量和短时过零率方法来分割有效发音的起止点。短时能量和短时过零率可由式(4)、式(5)表示：

$$E=\sum_{n=1}^{N}|x(n)| \tag{4}$$

$$Z=\frac{1}{2}\sum_{n=1}^{N}|\mathrm{sgn}[x(n)]-\mathrm{sgn}[x(n-1)]| \tag{5}$$

式中，x(n)为一帧语音信号，sgn[]为符号函数。sgn[]定义如式(6)所示：

$$\mathrm{sgn}[x]=\begin{cases}1 & x>0\\ -1 & x\leqslant 0\end{cases} \tag{6}$$

3.2 特征参数提取

语音信号具有多种特征参数，本系统采用梅尔频率倒谱参数(MFCC)作为语音信号特征参数[7]。系统分别提取双传感器语音 MFCC 参数并将参数合并形成融合特征参数作为后续模型训练与识别的输入特征向量。MFCC 特征提取过程如下：

① 输入一帧语音，对其做 FFT 变换得到频域幅值并平方得到频谱能量 E(k)：

$$E(k)=[|X(k)|]^2 \quad (k=0,1,\cdots,N/2) \tag{7}$$

式中，|X(k)|表示频谱中各频点 k 的幅值。

② 将频谱能量通过 24 个三角带通梅尔滤波器组得到梅尔滤波输出 S(m)：

$$S(m)=\sum_{k=0}^{N-1}E(k)H_m(k) \quad (0\leqslant m<M) \tag{8}$$

式中，M 表示滤波器个数，$H_m(k)$表示三角滤波器传递函数。

③ 将梅尔频谱能量 S(m)取对数并通过 DCT 变换即可得到静态 MFCC 参数 c(i)：

$$c(i)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log[S(m)]\cos\left[\frac{\pi i(2m-1)}{2M}\right] \quad (i=0,1,\cdots,M-1) \tag{9}$$

④ 通过式(10)对静态 MFCC 参数求取差分得到反映语音动态特征 d(i)：

$$d(i)=\frac{1}{\sqrt{\sum_{n=-k}^{k}n^2}}\sum_{n=-k}^{k}nc(i+n) \tag{10}$$

其中，k 取值为 2，将静态 MFCC 参数与动态 MFCC 参数合并后得到 1 帧语音的特征向量。

3.3 模型训练与识别

本系统采用隐马尔可夫模型(Hidden Markov Model，HMM)对孤立词进行建模。HMM 模型参数 λ 由初始状态概率 π、状态转移概率矩阵 A 以及发射概率 B 组成，即 λ=(π,A,B)。

本系统 HMM 模型采用自左向右无跨越结构，使用式(11)所示的混合高斯概率密度函数[9]表示发射概率 B：

$$b_i(o_t)=\sum_{k=1}^{K}\omega_{ik}N(o_t,\mu_{ik},\Sigma_{ik}) \tag{11}$$

式中，o_t 表示观测序列向量，K 为混合高斯数，ω_{ik}、μ_{ik}、Σ_{ik} 分别表示第 i 个状态的第 k 个高斯密度函数的权重、均值和协方差矩阵，混合高斯权重需要满足 $\sum_{k=1}^{K}\omega_{ik}=1$。

系统为每个单词分别建立一个具有5个状态的HMM模型，每个状态含有4个混合高斯。利用Baum-Welch算法调整训练集的单词模型参数得到模型参数，Baum-Welch算法是在给定初始HMM参数$\lambda=(\pi,A,B)$和训练数据条件下，经过多次参数调整得到模型新参数$\lambda'=(\pi',A',B')$，使得$P(O|\lambda')\geqslant P(O|\lambda)$。其中，设定最大训练次数为50，收敛条件阈值为1e−5，训练流程如图6所示。

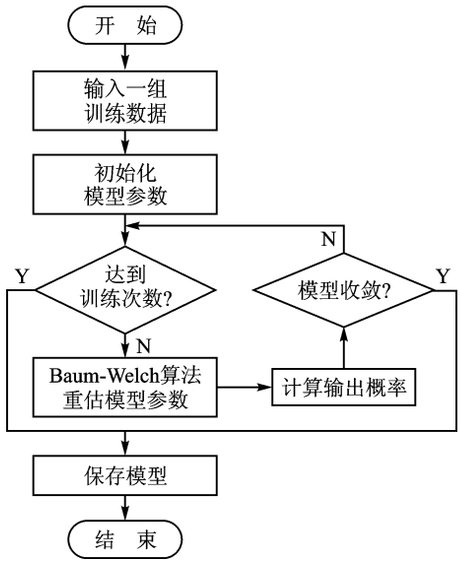


图6 单词模型训练流程

在系统建立好单词模型后，通过Viterbi算法计算融合序列在各个词条模型的输出概率，判断概率最大值的模型标签即可得到输出结果。

3.4 融合特征维度筛选

皮肤语音高频分量受到皮肤组织的衰减，拼接融合不同阶次L的皮肤语音MFCC特征参数可能会对系统识别率造成不同影响。为了找到最佳的皮肤语音特征融合阶次，本文提取了皮肤语音MFCC特征不同阶次(L取值6，8，10，12，14)与气导语音24维MFCC参数进行融合，进行皮肤语音融合特征维度筛选实验，为后续实验中的融合特征提供参考标准。实验数据由10人使用本系统的语音硬件采集终端在实验室安静环境下对数字0～9各录制20次构成，训练集和测试集各1000条。实验结果如图7所示。

从图7中可以看出，当麦克风语音12维MFCC参数与皮肤语音不同维度的MFCC参数阶次进行融合时，系统识别率随着皮肤语音特征参数阶次的增大而增大，当皮肤语音MFCC特征参数阶次取12时，系统平均识别率最高达97.2%，说明L值的增大能够提供更多的鉴别信息，提升了单词的区分度。随着L继续增加，系统识别率开始下降，原因可能是皮肤语音的高频信息不足导致系统对

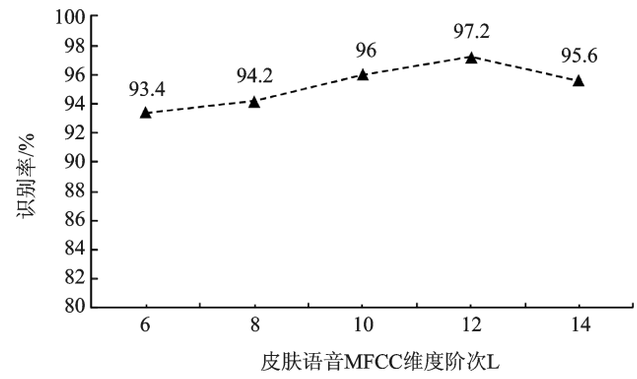


图7 皮肤语音特征阶次对识别率的影响

单词的区分度下降。因此，本系统最终选取皮肤语音12维MFCC＋12维ΔMFCC与气导语音12维MFCC＋12维ΔMFCC组合成48维特征向量作为融合特征。

4 实验结果及讨论

为了更直观地显示数据波形和识别结果，利用Matlab开发了用户显示界面。如图8所示，上位机显示界面主要功能包括网络配置、语音播放、语音数据存储、语音波形显示、识别结果显示以及信息打印等。

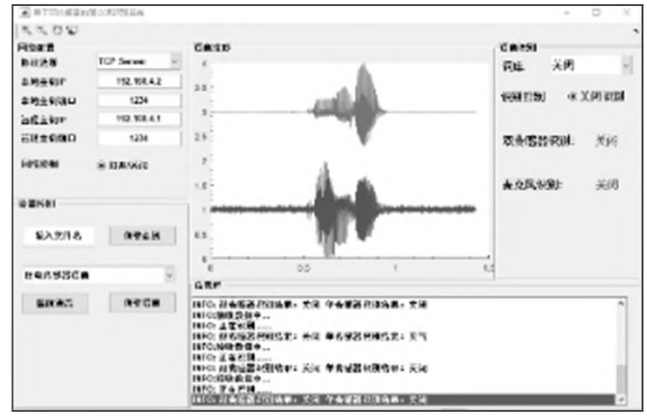


图8 Matlab上位机界面

实验选取20个孤立词作为实验词表，5位实验人员在安静环境下对每类词录音10次(每类词50条)进行模型训练，同时将训练词条中的麦克风和皮肤语音单独进行训练对比。由于HMM是基于说话人无关的模型，因此测试词条由另外5人对词表各录音10次，实验结果如表1所列。

由表1可知，在安静环境下单一麦克风语音识别率为93.5%，皮肤语音识别率为88.7%。相较于皮肤语音，经麦克风采集的气导语音具有更丰富的频率信息，能够更好地表现单词之间的差异性，因此识别率高于皮肤语音识别系统。在融合两种语音特征后的平均识别率达96.8%，高于单一特征的识别系统，说明融合特征的模型在训练过程中进一步提升了单词间的分布差异，从而达到系统整体

识别率的提升。

表 1　安静环境下测试结果

词条	气导语音特征识别率/%	皮肤语音特征识别率/%	融合特征识别率/%
打开	96	86	98
关闭	98	78	96
确认	96	90	98
返回	100	90	100
取消	92	86	98
查看	92	84	98
探测	94	76	96
左移	86	90	94
右移	86	90	100
引导	92	76	96
爬升	92	96	100
下降	96	92	96
速度	84	92	94
高度	92	100	96
航向	100	92	94
前翻页	96	96	98
后翻页	88	94	100
锁定目标	94	90	94
水平飞行	96	90	96
系统状态	100	86	94
平均识别率/%	93.5	88.7	96.8

为了测试融合特征语音识别系统的抗噪性，本文选取 Noise－92 噪声库中的高斯白噪声对测试词条中的麦克风语音加噪，组成－10～30 dB 带噪语音测试集。其测试结果如表 2 所列。

表 2　不同信噪比下的识别结果

信噪比/dB	气导语音特征识别率/%	融合特征识别率/%	信噪比/dB	气导语音特征识别率/%	融合特征识别率/%
－10	6.3	20.5	15	74.4	89.4
－5	13.1	44.8	20	78.0	90.0
0	38.0	81.7	25	84.8	91.0
5	56.7	87.9	30	89.5	91.9
10	69.2	89.0	安静	93.5	96.8

从表 2 中可以看出，气导语音识别系统的识别率随着信噪比的降低而下降：当带噪测试集的信噪比由 30 dB 下降至 5 dB 时，识别率由 89.5%降低至 56.7%；当信噪比下降至－10 dB 时，识别率仅为 6.3%。与单一气导语音识别系统相比，融合特征语音识别系统在信噪比由 30 dB 下降至 5 dB 时，识别率仅由 91.9%下降至 87.9%，且在 0 dB 以上的识别率高于 80%，表明在噪声环境下融合特征中的皮肤语音提供的特征鉴别信息有效抑制了噪声对识别率的影响，提升了识别系统的鲁棒性。

5　结　语

本文结合气导语音和皮肤语音的优势设计了一种基于双传感器特征融合的语音识别系统。系统在安静环境下识别率可达 96.8%，在 0～30 dB 信噪比环境下具有高于 80%的识别率，与单一麦克风语音识别系统相比，具有更高的识别率和鲁棒性，在嘈杂环境下具有一定的应用价值。

参考文献

[1] 黄志东. 鲁棒性语音识别技术研究综述[J]. 信息通信，2019(11)：20－22.

[2] Graciarena M，Franco H，Sonmez K，et al. Combining standard and throat microphones for robust speech recognition[J]. IEEE Signal Processing Letters，2003，10(3)：72－74.

[3] Radha N，Shahina A，Nayeemulla Khan A. Improving Recognition of Speech System Using Multimodal Approach[C]// International Conference on Innovative Computing and Communications. Springer，Singapore，2019：397－410.

[4] 张海宁. 强噪声环境下语音识别方法的研究[D]. 哈尔滨：黑龙江大学，2021.

[5] 吴卫鹏. 基于改进谱减的语音增强算法研究[D]. 南京：南京邮电大学，2019.

[6] 陈锡锻. 一种双门限语音端点检测算法[J]. 浙江工贸职业技术学院学报，2021，21(2)：43－46.

[7] 胡洋霞. 基于 DTW 模型的非特定人孤立词语音识别研究[D]. 天津：河北工业大学，2015.

[8] 曹冠彬. 基于 HMM 的连续语音识别技术研究[D]. 南京：南京理工大学，2018.

[9] Cheng X，Duan Q. Speech emotion recognition using gaussian mixture model[C]//The 2nd international conference on computer application and system modeling，2012：1222－1225.

余伟(硕士研究生)，主要研究方向为信息感知、获取与处理技术；陈向东(教授)，主要研究方向为新型传感器与智能信息获取。通信作者：陈向东，xdchen@home.swjtu.edu.cn。

(责任编辑：薛士然　收稿日期：2022-03-28)