基于动静态特征双输入神经网络的咳嗽声诊断COVID-19算法

张永梅,孙 捷
(北方工业大学信息学院,北京 100144)

摘 要: 新型冠状病毒肺炎(COVID-19)已经在世界范围内造成了严重影响,在防控疫情方面学者们进行了大量研究. 利用咳嗽声判断病变部位来诊断新冠肺炎具有非接触、成本低、易获取等优点,但是此类研究在国内较为匮乏. 梅尔倒谱系数(Mel Frequency Cepstral Coefficients,MFCC)特征仅能够表示声音的静态特征,而一阶差分MFCC特征还能反应声音的动态特征. 为了更好地防治新冠肺炎,本文提出了基于动静态特征双输入神经网络的咳嗽声诊断新冠肺炎算法,通过咳嗽声诊断新冠肺炎. 在Coswara数据集基础上,对咳嗽声的音频进行裁剪,提取MFCC和一阶差分MFCC特征训练了一个动静态特征双输入神经网络模型. 本文模型采用统计池化层,可以输入不同长度的MFCC特征. 实验结果表明,与现有模型相比较,本文算法明显提升了识别准确率、召回率、特异性和F1值.

关键词: 深度学习;咳嗽声;新冠肺炎;梅尔倒谱系数;音频技术;卷积神经网络
基金项目: 国家重点研发计划(No.2020YFC0811004)
中图分类号: TN912;TP183 **文献标识码:** A **文章编号:** 0372-2112(2023)01-0202-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211630

A Dynamic-Static Dual Input Deep Neural Network Algorithm for Diagnosing COVID-19 by Cough

ZHANG Yong-mei, SUN Jie
(*School of Information Science and Technology, North China University of Technology, Beijing* 100144, *China*)

Abstract: The COVID-19 (corona virus disease 2019) has caused serious impacts worldwide. Many scholars have done a lot of research on the prevention and control of the epidemic. The diagnosis of COVID-19 by cough is non-contact, low-cost, and easy-access, however, such research is still relatively scarce in China. Mel frequency cepstral coefficients (MFCC) feature can only represent the static sound feature, while the first-order differential MFCC feature can also reflect the dynamic feature of sound. In order to better prevent and treat COVID-19, the paper proposes a dynamic-static dual input deep neural network algorithm for diagnosing COVID-19 by cough. Based on Coswara dataset, cough audio is clipped, MFCC and first-order differential MFCC features are extracted, and a dynamic and static feature dual-input neural network model is trained. The model adopts a statistic pooling layer so that different length of MFCC features can be input. The experiment results show the proposed algorithm can significantly improve the recognition accuracy, recall rate, specificity, and F1-score compared with the existing models.

Key words: deep learning; cough; COVID-19; Mel frequency cepstral coefficients; audio technology; CNN
Foundation Item(s): National Key R&D Program of China (No.2020YFC0811004)

1 引言

新型冠状病毒肺炎(COVID-19)简称“新冠肺炎”,平均潜伏期为5.2天,感染后会引起发烧、咳嗽以及其他类似流感的症状,影响人体多种组织和器官功能. 许多受感染患者会发展为肺炎,并迅速转为严重的急性呼吸衰竭. 研究表明,超过60%的患者一旦病情发展到严重阶段,就会很快死亡,因此,在疾病早期进行严密监控和有效干预变得尤为重要.

截至2022年4月7日,根据世界卫生组织的官网报告,全球已有492 189 439例确诊病例,占全球总人口数

收稿日期:2021-12-08;修回日期:2022-05-17;责任编辑:朱梅玉

的6.23%,其中死亡人数6 159 474人. 随着疫情的发展,一些国家又出现了新冠病毒的变异. 变异新冠病毒的传播为全球抗疫带来新的挑战.

世界各国在抗击疫情时付出了沉重的财力物力代价. 新冠肺炎疫情发生以来,如何在早期及时发现和识别公共卫生事件的发生和流行成为公共卫生领域的重中之重.

新冠肺炎潜伏期长、传染性强,人群普遍易感,尽管大部分病例为轻症,但值得注意的是,仍有少数患者无肺炎症状,甚至无症状. 在新冠肺炎疫情下对疑似病例的"早发现、早隔离"对遏制感染至关重要. 此外,确诊过程中病人的核酸检测一定需要是阳性,换言之,咽拭子、痰液,或者可能取得的支气管灌洗液等检查到了病毒核酸(阳性),才能作为确诊病例.

虽然核酸检测为新冠肺炎诊断的金标准,但是核酸检测对样本采集的时期要求较高,若检测时鼻、咽拭子中所含的病毒载量过小,则可能导致核酸结果呈现"假阴性". 因此,电子计算机断层扫描(Computed Tomography,CT)检查在肺部感染的检出和评估中的价值举足轻重. 然而,胸部CT检查也存在局限性. 由于新冠肺炎缺乏特异性影像学表现,仅依据CT表现很难将新冠肺炎与其他类型病毒引起的肺炎加以鉴别. 例如,新冠肺炎与腺病毒肺炎影像学表现特别相似,胸部CT检查无法鉴别. 每一套胸部CT片子近300个切面,正常情况下,一个有经验的医生也需要15~30分钟才能查看完. 在疫情防控的关键时期,如何提供一种非接触性、准确率高、成本低的新冠肺炎初步诊断方法显得尤为重要.

无发烧、乏力、头晕等明显特征的无症状感染者,已经成为新冠病毒传播和复发的最大威胁. 咳嗽声作为人的生物特征,是新冠肺炎患者的一个普遍症状,新冠肺炎患者的咳嗽声与非感染者不同,人耳很难分辨其中的差别. 咳嗽信号在COVID-19检测呈阳性的患者中有所改变,新型冠状病毒很容易通过咳嗽或打喷嚏的方式释放出来,并传播给他人.

在文献[1]中,麻省理工学院研究人员提出了一种新的人工智能模型,可以通过倾听非感染者和新冠肺炎患者之间咳嗽的细微差别,发现新冠肺炎确诊病例和无症状病例. 当患有新冠肺炎时,产生咳嗽声音的方式会发生变化,即使没有症状[1]. 声音作为自然界的一种物理特征,具备非接触、检测成本低、侵入性小、可以提供快速结果等特点. 例如张小恒等人[2]提出一种将非监督学习用于研究帕金森病的语音诊断方法,也为本文的研究提供了一定的理论基础.

根据医学方面最新的研究进展和新闻报道,COVID-19病毒将可能与人类长期共存,所以研究开发基于咳嗽声的新冠肺炎检测算法具有巨大的潜力和应用前景. 国外已经相继建立了多个采集和研究咳嗽声以及研发咳嗽声诊断新冠肺炎算法的项目,然而国内却缺少这方面的研究.

根据世界卫生组织公布的新冠肺炎症状,新冠肺炎病人的咳嗽声和正常人有细微差别,具体在肌肉退化、声带强度、情绪(怀疑或沮丧等)、呼吸和肺功能几个方面表现出特征差异[3]. 印度班加罗尔科研机构(Indian Institute of Science(IISc) Bangalore)建立了Coswara项目收集咳嗽声,并意图实现新冠肺炎检测. 卡内基梅隆大学建立了收集咳嗽声的项目,并与世界各地的研究者一起研究咳嗽声诊断新冠肺炎的算法.

剑桥大学建立了咳嗽声收集项目,并且获得了欧洲研究委员会项目资助. 该项目中,Brown等人[4]采用咳嗽声的梅尔倒谱系数(Mel Frequency Cepstral Coefficients,MFCC)作为特征训练VGGNet(Visual Geometry Group Network),并结合支持向量机(Support Vector Machine,SVM),取得了80%以上的新冠肺炎检测准确率;Han等人[5]基于SVM训练了一个神经网络Voice Only Model,新冠肺炎识别准确率为77%左右.

Andreu-Perez等人[6]研究了在卷积神经网络(Convolutional Neural Network,CNN)基础上改造的DeepCough2D和DeepCough3D网络,分别利用100×33×1的MFCC特征和100×33×3的MFCC特征(利用3个采集设备的3个不同的100×33×1的MFCC特征进行连接得到),仅凭咳嗽声就能诊断是否感染新冠肺炎,准确率为96%.

Imran等人[7]开发的分类器综合了肺部X光影像和咳嗽声诊断新冠肺炎,咳嗽声分类器(Classical Machine Learning-based MuLti-class classifier,CML-ML)将$M\times N$的MFCC特征先利用主成分分析(Principal Component Analysis,PCA)降维,得到$M\times P$的张量,再求振幅绝对值,得到$M\times 1$的张量,与求均值后的$M\times 1$的MFCC连接,组成$2M\times 1$的特征向量,采用SVM进行分类,该分类器的分类准确率为90%左右.

Bagad[8]提取了咳嗽声音频信号的257×201频谱特征,利用64维的梅尔滤波器组,处理成64×201的对数梅尔频谱特征,输入到残差神经网络ResNet18(Residual Network 18)完成分类. 国外还出现了数个通过检测咳嗽声诊断新冠肺炎的手机或电脑软件.

在国内,目前还没有基于咳嗽声诊断新冠肺炎的相关报道. 赵建等人[9]利用DNN-HMM(Deep Neural Network Hidden Markov Model)语音识别声学模型识别猪咳嗽声,及早发现生猪养殖过程中的呼吸道疾病. 黎煊等人[10]采用深度置信网络(Deep Belief Network,DBN),通过猪咳嗽声检测猪呼吸道疾病,取得了90%

以上的准确率.

目前的咳嗽声诊断新冠肺炎方法一般都直接利用咳嗽声的MFCC特征进行训练.MFCC虽然能较好地体现咳嗽声频率能量的静态特征,但是对于咳嗽声这种多个连续、每一次的轻重缓急都不同的声音类型来说,需要更准确的咳嗽声特征表达.李伟红等人[11]研究的经验小波滤波器组对低信噪比的声学信号取得了较好的特征提取效果.本文利用MFCC和一阶差分MFCC(即ΔMFCC)特征,训练了一种新的动静态特征双输入神经网络模型,结合统计池化层对任意尺寸输入的二维特征向量进行分类,利用咳嗽声分析测试者是否有可能是新冠肺炎患者.图1给出了本文基于动静态特征双输入神经网络的咳嗽声诊断新冠肺炎算法的流程.

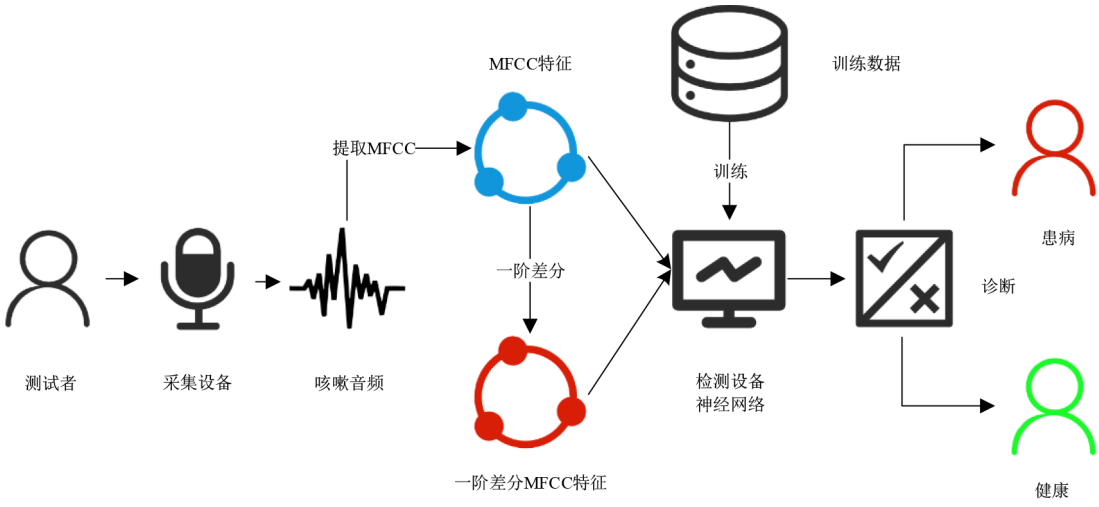


图1 本文算法流程

2 数据集介绍

目前用于诊断新冠肺炎的咳嗽声数据集很少.例如剑桥大学建立了COVID-19之声数据集,截止到2020年11月已经收集到16 000多位贡献者提供的30 000多份录音,这些录音的提供者年龄大部分在20~49岁之间,国籍以意大利居多,英国和巴西次之,其中,男性占比62%.除了统计新冠肺炎患病状态之外,还统计了是否吸烟等影响因素,但是该数据集是非公开数据集.较为易得的数据集比如Github上的一个开源数据集是俄罗斯建立的一个MP3文件数据集[12].该数据集中共有1 322个音频文件,分为有症状感染者、无症状感染者和非感染者三类.681份新冠肺炎患者数据中有294份无症状感染者数据,387份有症状感染者;而非新冠肺炎患者有641份.该数据集使用电话录音,或者电报中心的麦克风录音,录音质量没有保证.由于其他数据集比较难以获取,本文使用俄罗斯的MP3数据集的数据来组建验证集验证本文模型的泛化能力.

DiCOVA(Diagnosing COVID-19 using Acoustics)挑战赛的数据集来自Coswara数据集[13].Coswara由印度班加罗尔科研机构建立,希望通过人类各种声音特征诊断新冠肺炎[14].截至2021年7月14日,Coswara采集公布了1 947份数据,每份数据收集了录制者的深呼吸、浅呼吸、重咳嗽、轻咳嗽、快速数数、常速数数以及发音分别为A,E和O的音频数据.与录音一起,该数据集提供了与每个录音相关的COVID-19的阳性/阴性、个人的性别,以及国籍数据.Coswara是公开数据集中录制比较标准并且被广泛认可的数据集,所以本文主要使用Coswara数据集.

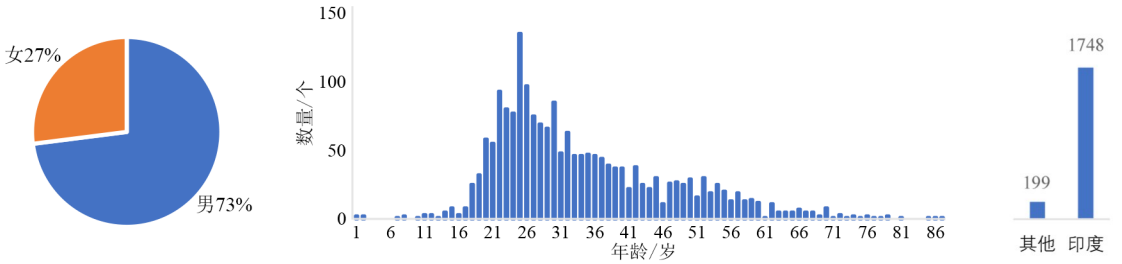
表1是本文使用Coswara数据集的情况,Coswara数据集中的数据包含非感染者1 340份,轻度新冠肺炎患者207份,中度新冠肺炎患者51份,重度新冠肺炎患者37份,未感染呼吸道疾病的157份,有呼吸道疾病但是尚未确诊的90份,患病但完全康复的65份.其中,未感染呼吸道疾病、有呼吸道疾病但是尚未确诊和患病但完全康复的157+90+65=312份数据对于本文来说是模糊的概念,所以不使用这312份数据.随机抽取1 000份非感染者数据,130份轻度患者数据,40份中度患者数据,30份重度患者数据,共1 200份数据,用于训练;剩余340份非患者数据,77份轻度患者数据,11份中度患者数据,7份重度患者数据,共435份数据,用于验证.

图2给出了Coswara数据集中各项数据的分布情况,由于男性和女性的声道在大小和形状上有所不同,从频谱的角度来看,男性和女性产生的咳嗽不一

表1　Coswara数据集使用情况(截至2021年7月14日)

原标签和统计	原数据量	训练数据量	验证数据量	分类
Healthy	1 340	1 000	340	1
positive_mild	207	130	77	0
positive_moderate	51	40	11	0
positive_asymp	37	30	7	0
no_resp_illness_exposed	157	0	0	未使用
resp_illness_not_indentified	90	0	0	未使用
recovered_full	65	0	0	未使用
总计	1 947	1 200	435	未使用312

定相等. 从图2(a)中可以看到,Coswara数据集中男性录音有1 423份,占73%,女性录音有524份,占27%. 从图2(b)可以看到,音频文件的来源年龄大部分分布在18岁到60岁之间,其中,24岁的数量最多. 图2(c)表明录音者大部分来自印度本地,有1 748份,其余的199份来自其他地区. 另外,根据统计,所有轻咳嗽声的音频时间的平均值是5.43 s,由于Coswara数据集中的所有音频的采样率都是48 kHz,所以轻咳嗽声的音频平均长度约为260 000帧,这些数据需要进行预处理.



(a) 性别分布　　(b) 年龄分布　　(c) 地区分布

图2　Coswara数据集的数据分布

3　实验方法

咳嗽声音是一种爆破音,相对于正常说话声音,咳嗽声音具有持续时间短、能量集中、频率高的特点. 根据呼吸系统生理及生物物理原理,声音产生于物体的振动,咳嗽声产生于人体的多个器官——肺部气体排出带动喉咙和声带振动,加上共鸣腔器官的共同作用. 新冠肺炎患者的病变部位往往在肺部和喉咙,所以通过咳嗽声可以提取与肺部、喉咙病变的相关特征,这种特征人耳很难分辨,但是通过统计频率或者能量分布能够甄别.

3.1　平衡数据

在Coswara数据集中,本文已经挑选了比较典型的数据分成训练数据集和验证数据集. 在训练数据集中,正类1 000份(非感染者的1 000份数据),负类200份(包括轻度患者130份数据,中度患者40份数据,重度患者30份数据). 验证数据集中,正类340,负类95. 对于训练数据集来说,这样的数据比例不平衡. 如果使用本文中的1 000份正类和200份负类组成的训练集训练模型,模型只需要将所有数据识别为正类就可以达到83.3%的训练准确率,也会将验证集中的所有样例识别为健康.

本文将训练数据中的1 000个正类分为5份,分别与负类组成一个400份数据的训练数据集,共5个训练集. Coswara数据集中剩余的340份正类和95份负类数据以及俄罗斯的MP3数据集中1 322份数据共同组成本文的验证集,用于验证本文实验方法. 每组数据训练时,会加载之前训练的权重参数作为预训练的参数,当训练中损失下降的幅度低于0.001时,停止训练,此时再使用验证集验证,选取最好的训练结果,图3给出了本文的交叉累计训练过程.

3.2　裁剪静音

数据集中的录音包含一个录音者的多次咳嗽声,每次咳嗽之间有静音部分. 对于音频信号来说,单次咳嗽声持续时间的长短也是判断新冠肺炎患者和非新冠肺炎患者的一个重要特征. 在录制咳嗽声时,几次咳嗽声之间的间隔往往随机,尤其是对于特意发出的咳嗽声来说,几声咳嗽之间的间隔更是随机,而应用场景下往往需要被检测人特意发出咳嗽声. 所以如果能消除掉咳嗽声之间的静音,不仅能减少数据量,还能使特征提取和神经网络更加专注于咳嗽声本身特征,节约计算资源.

在裁剪静音时,静音定义为音频信号中振幅低于最大值5%的帧. 本文每次分析窗口大小为512帧,如果这512帧全部被认为是静音,则删除这些帧,否则不删除,窗口的滑动大小是50帧,即每次往后移动50帧进行分析. 在MFCC特征提取过程中,本文使用的分析窗大小也是512,即每次分析512帧,所以裁剪时,本文也选择512的分析窗大小. 在图4中,图4(a)咳嗽声读入的是一个233 472帧的一维数组,纵轴表示振幅,振幅越大,表示声音越大. 图4(b)是去掉了图4(a)中静

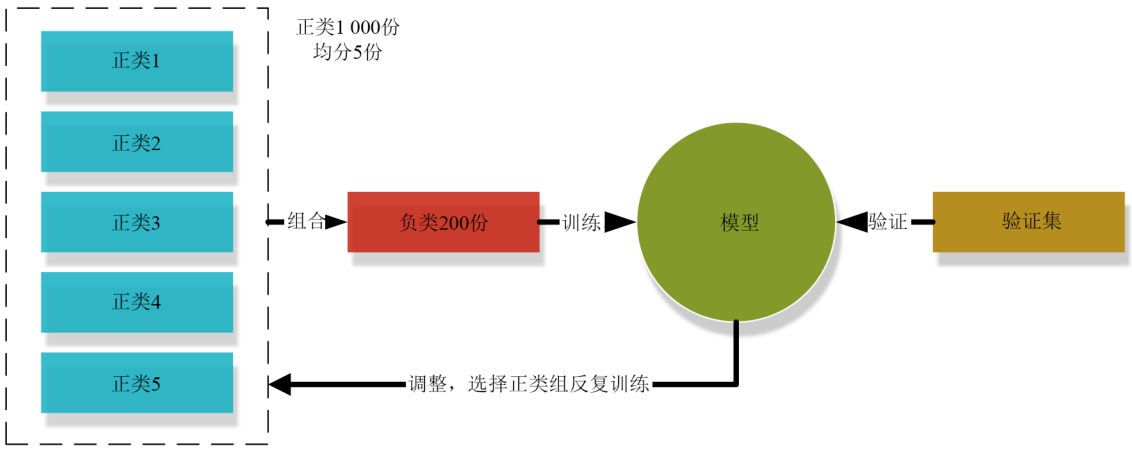
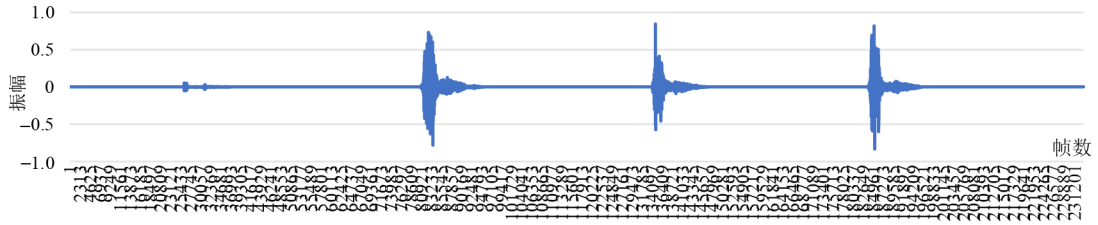
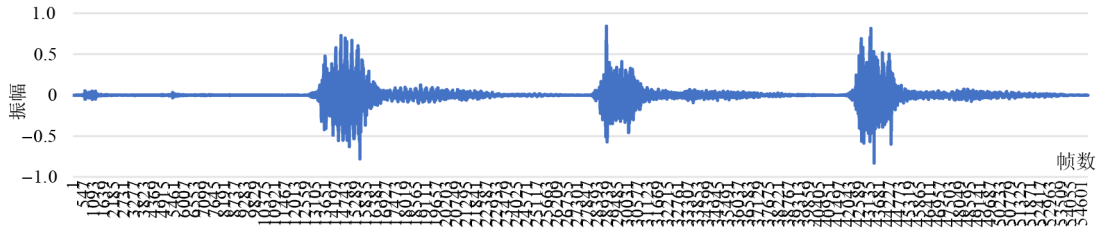


图3 交叉累计训练



(*a*) 咳嗽声信号波形图(共233 472帧)



(*b*) 裁剪后咳嗽声信号波形图(共55 050帧)

图4 去除静音的音频信号前后波形对比

音的裁剪结果,音频信号长度减少到了55 050帧.

3.3 MFCC特征提取

从声学和听觉的角度来看,语音特征提取方法包括线性预测编码(Linear Predictive Coding,LPC)、线性预测倒谱系数(Linear Predictive Cepstral Coefficient,LPCC),以及MFCC等特征.MFCC特征具有较强的抗噪声能力和优异的性能,是目前语音识别中最有效和最广泛应用的特征.

MFCC特征描述了人类听觉系统对于频率为1 kHz以下声音的感知能力表现为线性函数关系,对于频率为1 kHz以上的声音则呈对数关系.根据人耳的听觉激励,通过Mel滤波器组对原始频率进行非线性映射,利用Mel系数代替原始频率,将频谱变换到人耳听觉感知的Mel域[15].MFCC系数反映了声音信号在不同频段的能量分布.从文献[15]对病变处嗓音的建模研究中可以了解到,人类嗓音一般分布在250 Hz左右,所以利用MFCC特征能够较好地体现咳嗽声音的能量分布特征[16].Mel频率与声音信号频率的关系表示为式(1),其中f表示频率.

$$\mathrm{Mel}(f)=2595\lg\left(1+\frac{f}{700}\right) \tag{1}$$

计算MFCC的步骤包括预加重、加窗分帧、快速傅里叶变换(Fast Fourier Transform,FFT)、平方、Mel滤波、能量叠加、取对数,以及离散余弦变换(Discrete Cosine Transform,DCT).利用这些步骤,得到MFCC特征矩阵.本文提取MFCC特征矩阵时,得到的矩阵规模为(40,H),40是设置的倒谱数量,librosa库提取MFCC的默认倒谱数量是20,本文在大量实验的基础上选择倒谱数量为40.列数量H是由音频的帧长度取模512得到,这是由于MFCC计算分析的汉明窗滑动窗口大小是512.

MFCC为语音信号的静态特征,不符合语音动态变化的特征,人耳对于语音的动态特性更加敏感.为了更

好地反映连续咳嗽声的动态特征,本文再求MFCC的一阶差分. 一阶差分是离散函数中连续相邻的两项之差,能够表示当前声音帧和前一帧的关系,体现两帧之间的联系. 此处的帧表示MFCC特征帧,并非上文所述的音频帧. ΔMFCC的计算方法如式(2)所示. 其中Δc_n为特征参数的一阶差分,I=2,c_{n+1}表示n+1帧MFCC的值.

$$\Delta c_n=\frac{\sum_{i=-1}^{I} i+c_{n+1}}{\sum_{i=-1}^{I} i^2} \tag{2}$$

由于不同的咳嗽声录制者使用的采集设备不尽相同,以及采集时离麦克风远近距离不同等各种不可控因素的影响,采集的音频信息可能会有此类硬性误差. 为了减少这种影响,将数据进行Z-Score标准化. Z-Score标准化将不同量级的数据统一化为同一量级,保证数据之间的可比性. 其中Z-Score标准化通过式(3)实现. 式(3)中u表示均值,s表示标准差.

$$z_i=\frac{x_i-u}{s} \tag{3}$$

图5给出了MFCC系数与时间对照的图. 在图5中,横轴表示时间,纵轴表示MFCC矩阵的第一个维度数,也就是40. 图5分别给出了一组典型的新冠肺炎患者和非感染者MFCC和标准化MFCC特征矩阵的可视化图对比,坐标点的值表示此时语音信号的能量,并用颜色深浅来表示能量的大小. 越接近蓝色表示低频能量越多,越接近红色表示高频能量越多. 从图5可以看到,新冠肺炎患者比非感染者的咳嗽声包含更多的高频能量.

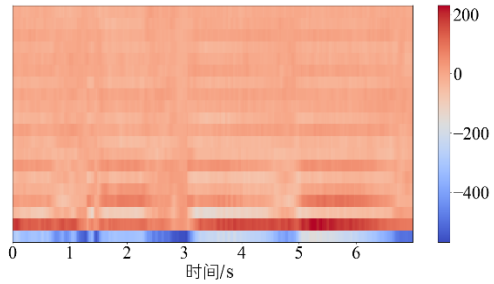
4　动静态特征双输入神经网络

通过对音频数据进行裁剪,减少音频中的静音和低频部分,使特征提取能更专注于咳嗽声而非随机的咳嗽声间隔. 通过提取MFCC特征反应咳嗽声的频率和能量特征,计算ΔMFCC来表示MFCC特征的前后帧之间的关系.

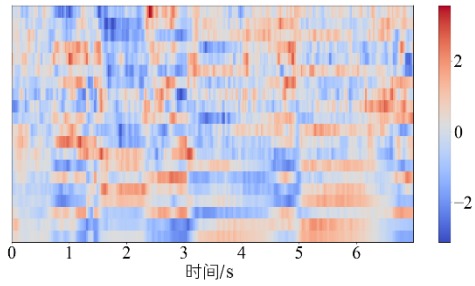
2018年Okabe等人[17]提出了统计池化层(statistics pooling layer),能够将不等长的音频数据输入到模型中,提取i-vector作为说话人识别的身份特征. 本文将统计池化层加入到模型中取得了较好效果. 统计池化层将每个特征张量分别求均值和标准差. 在式(4)和式(5)中,T表示一个特征张量的长度,$\boldsymbol{h}_t$表示特征矩阵,⊙表示两个矩阵相乘,μ表示均值,σ表示标准差. 将μ和σ连接,组成一个二维的新的特征矩阵,这样就能把不定长的输入特征张量变成定长.

$$\mu=\frac{1}{T}\sum_{t}^{T}\boldsymbol{h}_t \tag{4}$$

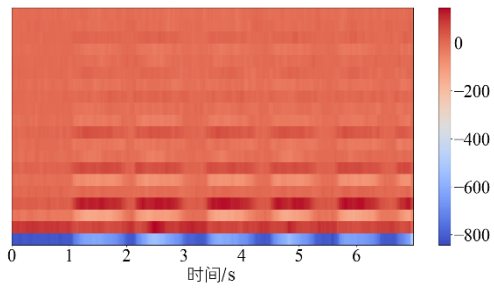
$$\sigma=\sqrt{\left(\frac{1}{T}\sum_{t}^{T}\boldsymbol{h}_t\odot\boldsymbol{h}_t-\mu\odot\mu\right)} \tag{5}$$



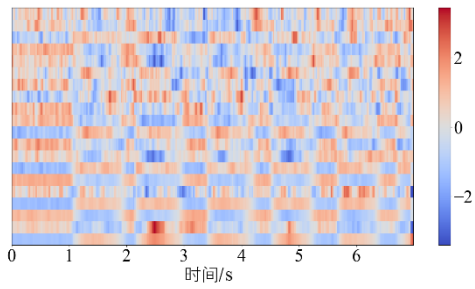
(a) 非感染者咳嗽声MFCC特征



(b) 非感染者咳嗽声标准化MFCC特征



(c) 新冠肺炎患者咳嗽声MFCC特征



(d) 新冠肺炎患者咳嗽声标准化MFCC特征

图5　新冠肺炎患者和非感染者MFCC和标准化MFCC特征对比

将MFCC特征和ΔMFCC特征分别输入到模型中,分别对咳嗽声的静态特征和动态特征进行处理. 本文在CNN基础上搭建了一个可以有不定长输入的动静态特征双输入神经网络模型. 图6给出了本文网络模型的结构.

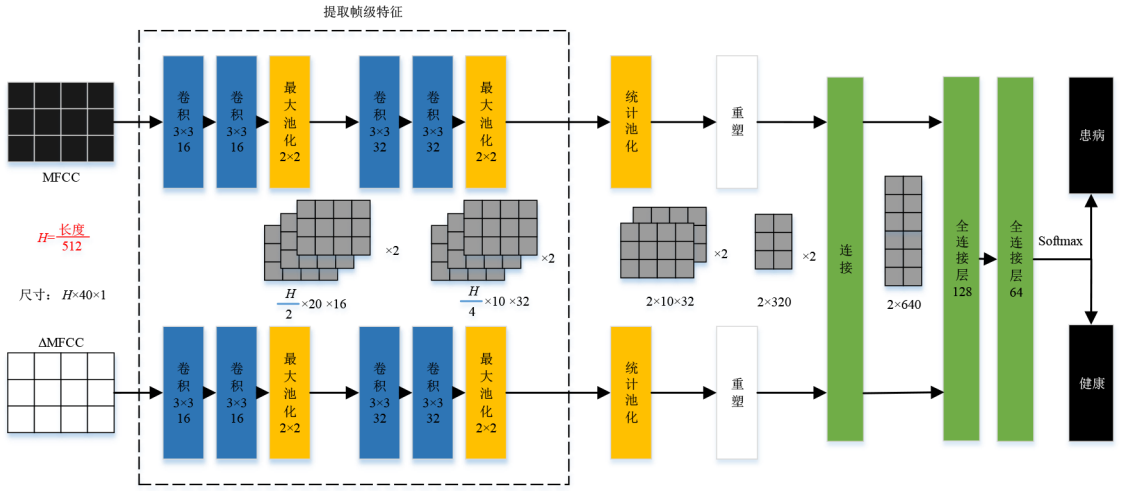


图6 动静态特征双输入神经网络

本文模型采用了一个双输入结构，目前的关于使用咳嗽声诊断新冠肺炎的相关文献中，这样的网络结构比较罕见．在处理图像数据的神经网络中有类似结构，用于不同通道输入不同尺寸的数据，然后综合这些数据的特征进行分类和预测．本文的两个通道的输入特征的尺寸相同，使用单输入的网络结构虽然不会影响参数量，但是单通道网络需要将两个特征作为两个维度输入，在神经网络中表现为4维，统计池化层更难在特定的维度上求均值和方差，可能会造成两种特征互相混淆影响．而输入进两个通道能够单独对两种特征提取特征，使之后的统计池化层容易对MFCC和ΔMFCC的数据总体求均值和方差，保持了两种特征在求均值和方差时的独立性．

在总的参数设置上，本文使用的batch大小是32，学习率为0.000 1，损失函数是二元交叉熵(Binary Cross Entropy)．表2给出了模型详细的参数配置．

表2 动静态双输入神经网络的具体输入输出尺寸和卷积核参数

层	卷积核	尺寸	层	卷积核	尺寸
Input1		H×40×1	Input2		H×40×1
Conv1_1	3×3 16	H×20×16	Conv2_1	3×3 16	H×20×16
Conv1_2	3×3 16	H×20×16	Conv2_2	3×3 16	H×20×16
Max pooling1_1	2×2	H×20×16	Max pooling2_1	2×2	H×20×16
Dropout1_1	0.3	H×20×16	Dropout2_1	0.3	H×20×16
Conv1_3	3×3 32	H×10×32	Conv2_3	3×3 32	H×10×32
Conv1_4	3×3 32	H×10×32	Conv2_4	3×3 32	H×10×32
Max pooling1_2	2×2	H×10×32	Max pooling2_2	2×2	H×10×32
Dropout1_2	0.3	H×10×32	Dropout2_2	0.3	H×10×32
Statistic Pooling1		2×10×32	Statistic Pooling2		2×10×32
Reshape1		2×320	Reshape2		2×320
Concatenate					2×640
Flatten					1×1 280
Dense1					128
Dense2					64
Softmax					2

第一轮训练设置100次迭代次数，之后的训练在loss下降值低于0.001时停止训练．模型的前半部分，输入H×40×1的MFCC特征矩阵和ΔMFCC特征矩阵，H为帧长度取模512，40是倒谱数量，1表示将二维的特征矩阵变成三维，以方便进行二维卷积操作．前两个卷积层的卷积核数量为16，卷积核大小为3×3，输出H×40×16的特征矩阵，经过2×2的最大池化层，特征矩阵的长和宽变成原来的一半．之后的两个卷积层卷积核数量为32，再经过一个2×2的最大池化层之后特征矩阵的长宽又变为原来的一半，但是深度变成了32，这样可以对MFCC特征和ΔMFCC特征进一步提取．两个通道分别输入MFCC和ΔMFCC，先利用卷积层和池化层提取

MFCC的帧级特征.

提取帧级特征之后,利用统计池化层对第1个维度计算特征向量的均值和标准差,每个通道得到一个2×10×32的特征张量,两个维度分别是均值和标准差. 然后将其后两个维度展开成320长度,两个通道分别得到一个2×320的特征向量,并把两个特征向量简单地连接成一个2×640的特征向量. 使用全连接层进一步把特征向量转换成一维,即1 280长度的一维特征向量,并转换为128长度的稠密向量和64长度的一维特征向量. 最后使用Softmax分类,实现二分类任务.

5　实验结果及分析

本文采用准确率(Accuracy)、召回率(Recall/Sensitivity)、特异性(Specificity)和F1值(F1-score)来评价模型对咳嗽声的分类结果. 根据混淆矩阵,计算公式分别为

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{6}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{7}$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{8}$$

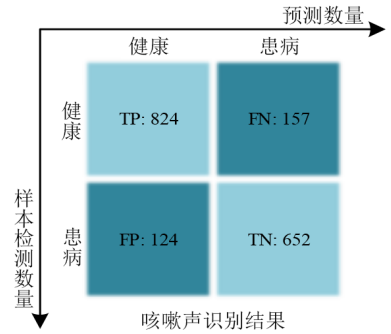
$$\text{F1}=\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{9}$$

本文验证集中新冠肺炎患者数据共776份,是Coswara数据集的95份和俄罗斯MP3数据集的681份之和;非感染者981份,是Coswara数据集的340份和俄罗斯MP3数据集的641份之和. 由于俄罗斯的MP3数据集在数据格式、采样频率上与训练数据集有所不同,所以将其调整为与训练数据集相同的格式之后进行验证实验.

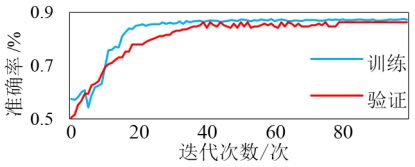
在图7(*a*)混淆矩阵中,TP(True Positive)与FN(False Negative)之和表示验证集中非感染者是981伤,同样,FP(False Positive)与TN(True Negative)之和表示验证集中新冠肺炎患者数据是776. 纵向表示模型的预测结果,TP和FP分别表示预测为非感染者时预测正确和错误的数量,FN和TN分别表示预测为新冠肺炎患者时预测错误和正确的数量. 对于训练集来说,最好的一次训练过程如图7(*b*)和图7(*c*)所示. 从图7(*b*)可以看到,训练准确率最终达到86%左右,在图7(*c*)中,训练的损失最终降到0.02左右. 从图7(*a*)可以看到,在776份新冠肺炎患者数据中,有652份识别正确,124份识别错误,准确率为84.02%.

新冠肺炎的症状变化较多,最近出现很多无症状和轻症感染. 本论文方法适合于有症状、无症状、轻度、中度、重度新冠肺炎患者的诊断.

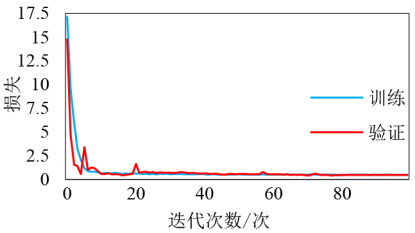
本文的验证集数据来自Coswara数据集和俄罗斯



(*a*) 混淆矩阵



(*b*) 准确率



(*c*) 损失

图7　实验结果

的MP3数据集. Coswara数据集的95份新冠肺炎患者数据包括77份轻度新冠肺炎患者、11份中度患者和7份重度患者数据. 俄罗斯MP3数据集的681份新冠肺炎患者数据包括294份无症状感染者和387份有症状感染者数据. 表3给出了本文方法对于各种新冠肺炎患者的实验结果.

表3　各种新冠肺炎患者的实验结果

患者类型	总数	识别正确数	识别错误数	准确率
有症状	387	335	52	86.56%
无症状	294	233	61	79.25%
轻度	77	67	10	87.01%
中度	11	10	1	90.91%
重度	7	7	0	100%
共计	776	652	124	84.02%

从表3可以看到,本文方法在俄罗斯的MP3数据集上的整体识别准确率低于Coswara数据集,可能是由于MP3数据集录音质量没有保证,本文对MP3数据集进行了格式和频率转化,以及Coswara数据集的新冠肺炎患者验证数据量较少. 本文方法对轻度患者的识别准

确率低于中度和重度患者,对无症状患者的识别准确率低于有症状患者,主要的可能原因是轻度患者和无症状患者咳嗽声的表现特征不明显,但本文方法对无症状患者的识别准确率也能达到79.25%.

本文方法现在处于实验室验证阶段,还需要进一步研究和论证,提高对所有年龄和种族的人的咳嗽声诊断新冠肺炎的准确率.

5.1 系统测试

在本文提出的一系列实验方法中,为了验证这些方法的有效性,进行了一系列消融实验进行实验方法的纵向对比.

首先是交叉训练的实验结果如表4所示. 在表4中,第一次训练使用的是第一个随机数据集,在100个迭代次数下的训练结果. 之后的每次训练都会加载之前的训练参数,并且当损失降低值低于0.001时停止训练. 从表4可以看到,5次的训练效果逐渐提升,交叉训练对于模型提升准确率和泛化具有良好效果.

表4 交叉训练实验结果比较

模型	训练准确率	训练损失	验证准确率
训练1	70.43%	0.446	67.65%
训练2	77.46%	0.102	72.71%
训练3	83.78%	0.055	79.42%
训练4	85.90%	0.024	82.79%
训练5	86.92%	0.021	84.00%

其次是不使用统计池化层和使用统计池化层的比较,统计池化层的作用是调整特征尺寸. 不使用统计池化层时,模型的输出尺寸是通过截断较长特征,填充(用0填充)较短特征来实现一致. 在输入时,使尺寸一致就可以不使用统计池化层. 但是在使用0填充和截断特征的影响下,不使用统计池化层的模型效果不如使用统计池化层的模型. 在使用相同的训练方法之后,最好的验证准确率达到76.32%.

最后是使用单特征输入模型和本文模型的对比情况. 本文模型与单独使用MFCC特征和单独使用ΔMFCC的模型进行了对比. 对比模型有4个:单通道MFCC特征模型、单通道ΔMFCC特征模型、双通道MFCC特征模型,以及双通道ΔMFCC特征模型. 对比结果如表5所示. 从评价结果来看,采用一种特征时,无论是单通道还是多通道的模型,模型的效果都不如本文提出的动静态特征双通道输入模型,所以本文的动静态双输入方法效果最好.

5.2 对比实验

在与其他模型的横向比较中,为了体现比较效果,也由于可获取数据集只有本文数据集,统一将其他各模型也使用本文所用的训练数据集、验证数据集和交叉累计训练方法进行比较和测试. 本文实现了在

表5 单特征模型实验结果比较

模型	准确率	召回率	特异性	F1值
单通道MFCC特征模型	73.98%	75.43%	72.21%	76.40%
单通道ΔMFCC特征模型	71.59%	73.70%	68.94%	74.34%
双通道MFCC特征模型	75.01%	76.75%	72.80%	77.42%
双通道ΔMFCC特征模型	75.41%	77.37%	72.29%	77.84%
本文模型	84.00%	83.99%	84.02%	85.43%

DiCOVA挑战赛中,Kamble等人提出的LightGBM方法[18]、Deshpande等人改进的Bi-LSTM模型[19]和Chang等人提出的在ResNet基础上改进的模型[20],并与本文方法进行比较.

作为DiCOVA挑战赛baseline的这些方法,使用的训练集和验证集都来自Coswara数据集. 在仅有Coswara验证集时,复现的系统在其论文的训练集和验证集上都达到了论文报告的结果. 本文的训练集是Coswara数据集,验证集由Coswara数据集和俄罗斯的MP3数据集共同组成,采用本文的训练集和验证集得到的四种方法实验结果比较如表6所示.

从表6可以看到,在同样的验证数据下,本文模型效果最好. 相比其他方法,本文进行了充分的数据预处理,通过双通道融合动静态MFCC特征,采用统计池化层充分利用提取的特征,提升了本文模型的准确率.

表6 实验结果比较

模型	准确率	召回率	特异性	F1值
本文模型	84.00%	83.99%	84.02%	85.43%
LightGBM	70.40%	72.50%	67.78%	73.14%
Bi-LSTM	70.51%	74.92%	64.94%	73.94%
ResNet	77.23%	79.71%	74.09%	79.63%

6 结论

本文通过裁剪咳嗽声音频信号,并结合咳嗽声的MFCC特征和ΔMFCC特征作为训练特征,训练了一个动静态双输入神经网络模型. 本文模型在识别新冠肺炎患者的咳嗽声的任务中达到了84.02%的准确率,从对比实验可以看到,采用统计池化层实现不定长特征向量的输入,充分利用提取的特征,通过双通道融合MFCC和ΔMFCC特征,提升了模型的准确率.

对于一种准备应用于室外的检测方法来说,不同人录制时,离麦克风远近不同会对音频信号的振幅,也就是声音的大小造成影响. 另外,目前的研究对噪声的处理比较欠缺. 这两个因素很少被考虑到,也还未建立有噪声数据和室外采集的咳嗽声数据. 之后的研究可以通过加入模拟的噪声提升鲁棒性,或者将音频信号的能量值降低来模拟麦克风的远近影响,直接采集建立这样的数据集也会为此类研究的发展做出一定贡献.

目前无症状感染者,以及不断变异的病毒导致的新冠肺炎症状的数据集不够全面. 进一步建立无症状感染者以及变异病毒导致的新冠肺炎公开数据集,有利于提升本文模型的新冠肺炎识别准确率.

参考文献

[1] LAGUARTA J, HUETO F, SUBIRANA B. COVID-19 artificial intelligence diagnosis using only cough recordings [J]. IEEE Open Journal of Engineering in Medicine and Biology, 2020, 1: 275-281.

[2] 张小恒, 张馨月, 李勇明, 等. 面向帕金森病语音诊断的非监督两步式卷积稀疏迁移学习算法[J]. 电子学报, 2022, 50(1): 177-184.
ZHANG X H, ZHANG X Y, LI Y M, et al. An unsupervised two-step convolution sparse transfer learning algorithm for Parkinson's disease speech diagnosis[J]. Acta Electronica Sinica, 2022, 50(1): 177-184. (in Chinese)

[3] 世界卫生组织. 2019冠状病毒病(COVID-19)专题问答[EB/OL]. (2020-11-10)[2021-11-10]. https://www.who.int/zh/news-room/questions-and-answers/item/coronavirus-disease-covid-19.

[4] BROWN C, CHAUHAN J, GRAMMENOS A, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Virtual Conference: ACM, 2020: 3474-3484.

[5] HAN J, BROWN C, CHAUHAN J, et al. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 8328-8332.

[6] ANDREU-PEREZ J, PÉREZ-ESPINOSA H, TIMONET E, et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels[J]. IEEE Transactions on Services Computing, 2022, 15(3): 1220-1232.

[7] IMRAN A, POSOKHOVA I, QURESHI H N, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app[J]. Informatics in Medicine Unlocked, 2020, 20: 100378.

[8] BAGAD P, DALMIA A, DOSHI J, et al. Cough against COVID: Evidence of COVID-19 signature in cough sounds [EB/OL]. (2020-09-23) [2021-12-07]. https://arxiv.org/abs/2009.08790.

[9] 赵建, 黎煊, 刘望宏, 等. 基于DNN-HMM声学模型的连续猪咳嗽声识别[J]. 农业工程技术, 2020, 40(30): 93.
ZHAO J, LI X, LIU W H, et al. DNN-HMM based acoustic model for continuous pig cough sound recognition[J]. Agricultural Engineering Technology, 2020, 40(30): 93. (in Chinese)

[10] 黎煊, 赵建, 高云, 等. 基于深度信念网络的猪咳嗽声识别[J]. 农业机械学报, 2018, 49(3): 179-186.
LI X, ZHAO J, GAO Y, et al. Recognition of pig cough sound based on deep belief nets[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(3): 179-186. (in Chinese)

[11] 李伟红, 王伟冰, 龚卫国. 低信噪比下公共场所异常声音声学特征提取[J]. 声学学报, 2019, 44(5): 934-944.
LI W H, WANG W B, GONG W G. Acoustic features extraction of abnormal sounds in public places with low signal-to-noise ratio[J]. Acta Acustica, 2019, 44(5): 934-944. (in Chinese)

[12] ALEXGEERTSEN, VLADISLAV C. GitHub-covid19-cough/dataset: Dataset of recordings of induced cough [DB/OL]. (2020-12-11) [2021-12-07]. https://github.com/covid19-cough/dataset.

[13] MUGULI A, PINTO L, SHARMA N, et al. DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics[EB/OL]. (2021-06-17)[2021-12-07]. https://arxiv.org/abs/2103.09148.

[14] SHARMA N, KRISHNAN P, KUMAR R, et al. Coswara—A database of breathing, cough, and voice sounds for COVID-19 diagnosis[EB/OL]. (2020-08-11) [2021-12-07]. https://arxiv.org/abs/2005.10548.

[15] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357-366.

[16] 顾玲玲, 张晓俊, 黄程韦, 等. 息肉与麻痹喉声源分类中非线性动力学发声系统模型研究[J]. 声学学报, 2015, 40(6): 878-885.
GU L L, ZHANG X J, HUANG C W, et al. Study on the model of nonlinear dynamics phonation system for the classification of polyps and paralysis phonation[J]. Acta Acustica, 2015, 40(6): 878-885. (in Chinese)

[17] OKABE K, KOSHINAKA T, SHINODA K. Attentive statistics pooling for deep speaker embedding [EB/OL]. (2019-02-25)[2021-12-07]. https://arxiv.org/abs/1803.10963.

[18] KAMBLE M R, GONZALEZ-LOPEZ J A, GRAU T, ET

al. PANACEA cough sound-based diagnosis of COVID-19 for the DiCOVA 2021 Challenge[EB/OL]. (2021-06-07)[2021-12-07]. https://arxiv.org/abs/2106.04423.

[19] DESHPANDE G, SCHULLER B W. The DiCOVA 2021 challenge—An encoder-decoder approach for COVID-19 recognition from coughing audio[C]//Interspeech 2021. Brno: ISCA, 2021: 931-935.

[20] CHANG J, CUI S, FENG M. DiCOVA-Net: Diagnosing covid-19 using acoustics based on deep residual network for the DiCOVA challenge 2021[EB/OL]. (2021-07-11)[2021-12-07]. https://arxiv.org/abs/2107.06126.

作者简介

张永梅 女,1967年1月出生于山西省太原市. 现为北方工业大学信息学院教授、博士生导师. 获省部级科技进步一等奖1项、二等奖3项,获省级教学成果二等奖2项. 授权发明专利19项,申请软件著作权63项. 在国内外发表学术论文212篇,出版专著和教材5部.
E-mail: zhangym@ncut.edu.cn

孙 捷 男,1997年6月出生于山东省临沂市. 现为北方工业大学信息学院硕士研究生. 研究方向为图像处理和人工智能.
E-mail: sunjie0627@163.com