第 37 卷 第 10 期
2021 年 10 月

信 号 处 理
Journal of Signal Processing

Vol. 37 No. 10
Oct. 2021

文章编号：1003-0530(2021)10-1843-09

采用音质特征和 VLAD 编码的新冠肺炎检测算法

张昊然 韩易辰 谭咏梅 李 雅
(北京邮电大学，人工智能学院，北京 100876)

摘 要： 2020 年，世界卫生组织宣布 COVID-19 疫情为大流行病。为了实现 COVID-19 快速地、可靠地检测，本研究通过语音信号分析技术来寻找感染 COVID-19 的语音信号特征，利用咳嗽声片段和语音片段对是否感染 COVID-19 做出自动判断。在 INTERSPEECH 2021 ComParE 竞赛提供的相关数据集和 baseline 的基础上，本文首先利用语音端点检测技术对数据集进行增广，其次在特征集中加入语音质量特征，使相关 baseline 结果得到了提升，证明了语音质量特征在对 COVID-19 自动语音检测任务上的有效性。同时，引入局部聚合描述子向量对低级别特征进行编码，当字典大小较小时，有效地提升了系统的分类性能。最后，对多种算法得到的分类结果进行融合，进一步提升分类效果，最终在两个子任务中的验证集上 UAR 分别取得了 73.9% 和 77.2%。

关键词： COVID-19 自动检测；语音切分；语音质量特征；局部聚合描述子向量；情感识别
中图分类号： TP181 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2021.10.007

引用格式： 张昊然，韩易辰，谭咏梅，等. 采用音质特征和 VLAD 编码的新冠肺炎检测算法[J]. 信号处理，2021，37(10)：1843-1851. DOI：10.16798/j.issn.1003-0530.2021.10.007.

Reference format: ZHANG Haoran，HAN Yichen，TAN Yongmei，et al. COVID-19 detection algorithm using voice quality features and VLAD coding[J]. Journal of Signal Processing，2021，37(10)：1843-1851. DOI：10.16798/j.issn.1003-0530.2021.10.007.

COVID-19 Detection Algorithm Using Voice Quality Features and VLAD Coding

ZHANG Haoran HAN Yichen TAN Yongmei LI Ya
(Beijing University of Posts and Telecommunications，Artificial Intelligence Academy，Beijing 100876，China)

Abstract: In 2020，the World Health Organization declared the COVID-19 outbreak a pandemic. In order to promote the rapid and reliable detection of COVID-19，this research introduced voice signal processing technology to find the voice signal characteristics of COVID-19 infection，and automatically judges whether it is infected with COVID-19 using cough fragments and speech fragments. On the basis of the relevant data set and baseline provided by INTERSPEECH 2021 ComParE，firstly，the audio segmentation technology was used to augment the data set. And secondly，voice quality features were added to the feature set，which improved baseline results and proved that the voice quality features are effective on the task of automatic speech detection for COVID-19. At the same time，Vector of Locally Aggregated Descriptors is introduced to encode low-level features. When the dictionary size is small，the classification performance of the system is effectively improved. Finally，the classification results obtained by multiple algorithms are fused to further improve the final classification effect. The UAR for CCS and CSS sub-challenges are 73.9% and 77.2%，respectively.

Key words: COVID-19 automatic detection; voice segmentation; voice quality; vector of locally aggregated descriptors; emotion recognition

收稿日期： 2021-08-13；**修回日期：** 2021-09-12
基金项目： 国家自然科学基金(61773379)；北京邮电大学新进教师人才项目(2021RC37)支持

1 引言

2020年3月11日，世界卫生组织宣布COVID-19疫情为大流行病。随着疫情的快速国际化发展，它在许多方面多多少少改变了我们的生活。医生和科学家努力寻找COVID-19的线索/指标[1]，希望能够找到快速的检测方法，控制疫情的蔓延。除了通过CT和X射线扫描识别的肺部损伤[2-3]，Tandan等人发现最常见的症状是发烧、咳嗽、肺炎、喉咙痛[4]，而Alzubaidi等人[5]使用不同的特征选择和聚类方法，发现发烧、咳嗽、疲劳、喉咙痛和呼吸短促是几个相对重要的新冠肺炎感染特征。他们的发现证实了咳嗽和喉咙痛在COVID-19中的普遍性。这使得我们想到使用语音信号处理技术来寻找这种特征，并提供一种可靠和快速的COVID-19检测方法。因为从语言生成理论来看，喉部和肺部的任何变化都可以从声音中得到体现。

在基于语音的COVID-19识别方面人们已经做了很多工作。首先是学者们收集了由COVID-19患者的语音语料。AI4COVID-19[6]于2020年开始实施，这是一个由人工智能驱动的COVID-19感染筛查解决方案，可通过智能手机应用程序进行部署。在这个项目通过录制3秒钟的咳嗽声，通过云计算返回COVID-19的测试结果，整个过程在两分钟内完成，提高了COVID-19的检测效率。剑桥COVID-19声音数据库[7]也通过APP收集用户的简短朗读语音和几声咳嗽声。Wei等人[8]提出了一个用于COVID-19感染风险评估的实时机器人，它集成了语音识别、温度测量、关键词检测、咳嗽检测等功能。卷积神经网络和SVM被用于咳嗽分类。Laguarta等人[9]将肌肉退化、声带变化、情绪/心情变化、肺部和呼吸道的变化作为特征，对咳嗽进行自动分类，该模型对使用官方测试诊断的受试者的COVID-19敏感性达到98.5%，特异性达到94.2%。

基于剑桥COVID-19声音数据库，学者们组织了公开挑战赛——INTERSPEECH 2021计算语言学挑战赛[10]，以加速这一研究领域的进展。该比赛有两个子挑战，即COVID-19 Cough Sub-Challenge (CCS)和COVID-19 Speech Sub-Challenge (CSS)，其中CCS子挑战提供了397名志愿者的语音，每名志愿者录制了1～3声咳嗽声，CSS子挑战提供了366名志愿者的语音，每名志愿者录制相同的文字内容1～3遍，即“I hope my data can help to manage the virus pandemic”。这些语料被用来对是否感染COVID-19进行二元分类。同时挑战赛官方还提供了基线(baseline)算法，包括openSMILE[11]，openXBOW[12]，DeepSpectrum[13]，auDeep[14]和End2You[15]。

在基线的基础上，我们的工作对这两个子挑战的贡献有三个方面。

首先，我们使用语音端点检测方法将长语料分割成若干短语料做语料增广，这是因为挑战中只有大约600个训练样本。语料库规模限制是COVID-19诊断以及其他基于语音的医疗问题中常见的问题，如何在小语料库的基础上提升模型性能，是这类问题首先要解决的问题。

第二，尽管基于深度学习的端到端的识别方法在大多数任务上能取得显著的性能提升，但对于小数据集，具有区分性的特征提取、特征编码等传统方法仍有一定的应用价值。除了基线特征外，本文还将语音质量(voice quality，VQ)特征引入基于语音的COVID-19识别中。语音质量用来描述人类声音的听觉色彩[16]，简称音质。音质参数反映发音时声门波形状的变化，反映了声音质量的变化，如声道肌肉紧张程度等。主要包括共振峰、谐波噪声比、松紧度、粗糙度、清晰度、明亮度、喉化度和呼吸声等等。VQ与发音器官变化紧密相关，可以成为检测COVID-19的有效特征，这也是这项工作的出发点。

第三，对基线提取的低水平特征进行局部聚合描述符(Vector of Locally Aggregated Descriptors，VLAD)编码[17]。VLAD特征编码在图像分类、动作识别等领域被证明可以对局部特征进行更精确的表示，因为从统计角度看，一般认为残差能够比频率包含更多的信息。同时还能对特征数据起到降维的效果。所以，我们对数据集中的低级别特征引入VLAD编码，对特征进行更深层次的表示，以提高系统的分类准确率。在情感识别领域，Balaji的人的工作[18]表明在人脸中层级特征加VLAD，能够提升模型性能，在EmotiW 2017 group-level情感识别竞赛中取得了76.5%的效果。

本文的组织结构如下。第2节介绍本文提出的

方法和系统框架,包括数据增广方法、语音质量特征分析和 VLAD 编码。在第 3 节中,提出一系列的实验对比并且对实验结果进行分析。第 4 节中对本文发现进行了总结。

2　系统框架

图 1 展示了本文所用的算法流程框架。首先输入原始的音频文件对音频进行数据增广。之后从音频文件中提取特征,除了 baseline 所提供的 openSMILE、openXBOW、DeepSpectrum、auDeep、End2You 外,还添加了本文所使用的 VQ 特征和 VLAD 编码。得到编码结果后,将特征编码输入到线性 SVM 分类器中得到每一种特征编码的分类结果。最后对这些结果进行基于投票机制的后端融合(Late fusion),得到最终的分类结果,即该段音频的录制者是否感染新冠肺炎。数据增广、baseline 算法、VQ 特征和 VLAD 编码部分将在本章节中具体讨论。

2.1　数据增广

CCS 任务训练集有 289 组数据,测试集有 231 组数据; CSS 任务训练集有 315 组数据,测试集有 295 组数据。新冠检测阳性样本在 CCS 任务样本中占总数的 23.0%,在 CSS 任务样本中占总数的 35.1%。样本数量少且不同类别间样本数分布不均匀,所以首先我们将长音频文件切分为若干个段音频文件。对于音频文件的切分,一般采用两种方法: 等时间间隔切分和根据是否为静音片段对音频进行切分。对于等时间间隔,原始咳嗽声音频长度大约为 7 ~ 10 s,每段音频包含 1 ~ 3 声的咳嗽声,而每声咳嗽声又包含 2 ~ 3 声轻重不同的“咳”声,所以 CCS 任务我们选取每 500 ms 对音频文件进行切割,这样切割不会使同一声咳嗽切割后过于细碎,也不会因为切割间隔过长使数据增广效果不明显。对于 CSS 任务由于音频更长每 600 ms 对音频文件进行切割。对于根据是否为静音片段进行切割,我们计算每一帧的能量,通过 GMM 模型确定是否为静音片段,对长音频进行切割。只有包含咳嗽或讲话的文件被保留,而静音段的音频被丢弃。

2.2　baseline 算法

baseline 利用 openSMILE、openXBOW、DeepSpectrum、auDeep 和 End2You 工具对音频文件进行特征提取,分别得到 ComParE_2016 特征集、音频词袋特征(Bag-of-Audio-Words, BoAW)、DeepSpectrum、auDeep 和 End2You 特征集。其中 openSMILE、openXBOW、DeepSpectrum、auDeep 所得到的特征向量需要输入到线性 SVM 中进行分类。

(1) openSMILE: 与 INTERSPEECH 官方提供的基线特征集相同,使用 ComParE_2016 特征集。该特征集包含由低级描述符(Low-Level descriptor, LLD)通过统计学计算得出的 6373 维特征数据。

(2) openXBOW: 该工具对音频文件提取音频词袋特征(Bag-of-Audio-Words, BoAW)。BoAW 在语音情感识别和声学时间检测领域已经被成功应用。通过从 openSMILE 提取的 LLDs 随机取样学习得到码本。再在这些码本上对 LLDs 进行频率统计,得到最终对音频文件表示的 BoAW 编码。

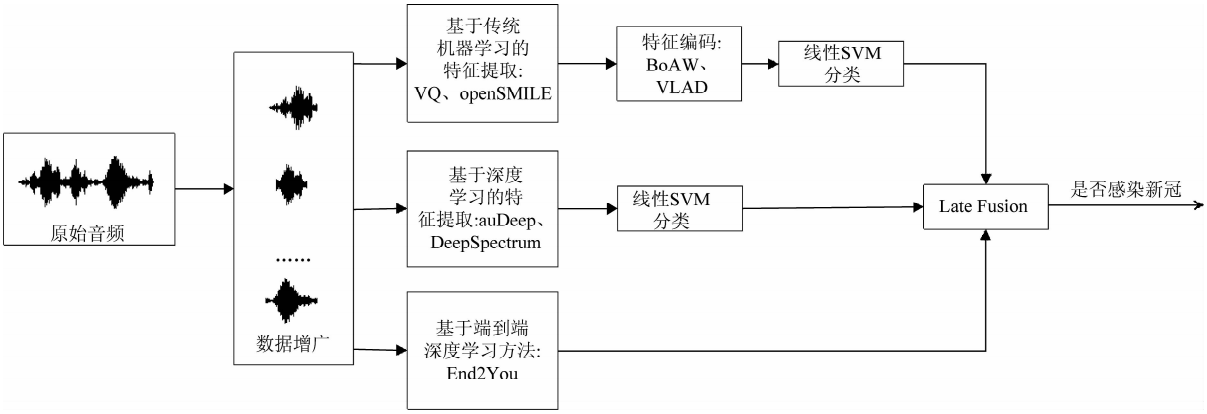


图 1　系统算法框图

Fig. 1　System algorithm block diagram

(3) DeepSpectrum: 利用该工具对输入的音频文件使用汉明加窗转化为梅尔频谱图,汉明窗宽度为32 ms、重叠为16 ms。将频谱图输入到预训练好的CNN模型DenseNet121中。提取网络中的平均池化层的激励结果,得到一个2048维的特征向量对音频文件进行描述。

(4) auDeep: 对原始数据的波形提取梅尔频谱图。用四个不同的功率水平阈值分别对频谱图进行剪切,通过这样的方法可以消除一定的背景噪声。之后,通过无监督学习的方式,在这些频谱图上学习一个无重复的、循环的序列到序列的音频编码器。并且将学习到的频谱图的描述符提取出来作为相关音频文件的特征向量。最后将在不同阈值切割的频谱图上提取的特征向量拼接起来,得到一个音频文件最终的特征向量。

(5) End2You: 利用该工具包进行端到端的学习。与上述工具不同的是,End2You利用Emo-18[19]深度神经网络进行特征的提取与分类。该网络首先使用一个卷积网络从原始的时域表示提取特征,之后用带有门控循环单元(Gated Recurrent Unit,GRU)的循环网络进行最终的分类。而上述的其他工具提取得到的特征向量还需输入到另外的分类器中进一步分类,才能得到分类结果。

2.3 音质特征

上述的baseline中所提供的特征集大都通过统计学得出,忽略了对语音片段音质的分析,音质特征能够对说话者发声器官的一些物理变化进行描述。而感染新冠肺炎的病人往往在喉部发生一些变化,所以我们引入音质特征对语音片段进行描述。常用的音质特征可以分为时域和频域的特征。从时间角度分析来看,基频抖动(Jitter)是描述测量到的基频值的变化程度,是由相邻一段时间内的基频值来推测出当前的基频值这个预测出来的结果和实际基频之间的差。当声带肌肉紧张,声道表面变硬等,都会使得基频抖动较大。振幅抖动(Shimmer)与基频抖动类似,反映的是振幅周期间的变化。开商(open quotient, OQ)描述了喉头开放阶段的相对持续时间。准开商(quasi-open quotient, QOQ)作为与标准开商有关的更稳健的测量方法,它通常用于VQ分析。另一组参数是从声门脉冲的峰值振幅或其导数中提取的。可以分别定义为连续周期之间的基本频率和振幅的变化。第一和第二谐波振幅差(H1-H2)[20]是一种频谱测量,用于描述谐波结构,一般反映声门收缩,数值越低表示收缩越大。噪声一般都是对应H1-H2的低值。振幅商(amplitude quotient, AQ)是声门信号振幅与声门信号导数的最小值之间的比率,而归一化振幅商(normalized amplitude quotient, NAQ)是基于基本频率AQ的归一化形式,它比AQ更稳健。也有研究表明,NAQ是区分呼吸音和紧张音的重要声门特征[21]。从频谱分析角度来看,低于第一共振峰的能量对声音质量也是重要的,因为这些谐波中蕴含的能量也较大。谐波丰富因子(harmonic richness factor, HRF)[16]是由高于第一谐波的谐波振幅之和除以基频振幅来衡量的,它量化了喉音源的幅度谱中的谐波数量[22]。Maxima Dispersion Quotient(MDQ)[23]被提出来用于区分呼吸音和紧张音。以前的研究表明,音质的变化可以揭示说话者的情绪状态[24]、说话者的社会关系[25]和个人特征[26]等。

本文所使用的音质特征利用VQ工具包[1]提取。提取出的VQ特征包含了上述的NAQ、QOQ、H1-H2,HRF,MDQ特征,并计算了它们的最大、最小、均值、方差、范围等5个统计参数,共26维。

2.4 VLAD编码

局部聚合描述子向量——VLAD是一种特征编码方法,能够将低级别的特征进行更高层次的表示,在图像分类、动作识别等领域被广泛运用。所以,我们对提取的LLDs特征进行VLAD编码,对特征进行更加精确的描述。

VLAD编码过程如下:

(1) 用openSMILE工具从n个音频文件中提取出LLDs,将LLDs作为音频文件的局部特征$b_i \in R^d$;

(2) 从每个音频文件的LLDs中随机选取m个局部特征组合成特征矩阵$\boldsymbol{X}=\{x_1, x_2, \cdots, x_{nm}\} \in R^{d\times nm}$;

(3) 将$\boldsymbol{X}$作为K-Means聚类算法的输入进行字典学习,得到k个聚类中心$\boldsymbol{W}=\{w_1, w_2, \cdots, w_k\} \in R^{d\times k}$,做为字典中心;

(4) 对每个音频文件的每一个局部特征b_i寻找与之距离最近的字典中心w_j,将向量残差进行累加,得:

[1] https://github.com/jckane/Voice_Analysis_Toolkit

$$c_j = \sum_{i;NN(b_i)=w_j}^{k} (b_i - w_j) \tag{1}$$

函数 $NN(b_i)$ 用于求与 b_i 距离最近的字典中心，c_j 为一个字典中心对应的残差向量。将所有的残差向量拼接在一起，就可得到编码 $\boldsymbol{C}=\{c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, \cdots, c_k^{\mathrm{T}}\} \in R^{dk}$。

(5) 对得到的编码 $\boldsymbol{C}$ 分别进行功率归一化:

$$c_{i,j} = \mathrm{sign}(c_{i,j}) * \sqrt{|c_{i,j}|} \tag{2}$$

$$\mathrm{sign}(c_{i,j}) = \begin{cases} 1, & \text{if } c_{i,j} > 1 \\ 0, & \text{if } c_{i,j} = 0 \\ -1, & \text{if } c_{i,j} < 0 \end{cases} \tag{3}$$

和 L2 归一化:

$$c_j = \frac{c_j}{\| c_j \|_2} \tag{4}$$

得到最终的 VLAD 编码 $\boldsymbol{C}$。

3　实验结果

3.1　分类器设置

本文进行的分类实验分类器采用的是线性支持向量机(Support Vector Machine，SVM) 分类器。线性 SVM 是机器学习领域常用的分类器之一。由于仅需判断语音的录制者是否感染新冠肺炎，所以本文实验使用线性 SVM 进行的是二分类任务，设置归一化参数 C 为 0.1，最大迭代次数为 10000，选择损失函数为 squared_hinge:

$$\mathrm{squared\_hinge}(y_{\mathrm{true}}, y_{\mathrm{pred}}) = \frac{1}{d}\sum_{i=1}^{d} (\max(1 - y_{\mathrm{true}}^{(i)} * y_{\mathrm{pred}}^{(i)}, 0))^2 \tag{5}$$

其中 y_{true} 为样本真实标签，值为 0 或 1。y_{pred} 为预测结果。d 为样本总数。

3.2　不同数据增广方式

为了对比两种切分方法的效果，我们利用 openSMILE 中的 function: ComParE_16 进行特征提取，线性 SVM 分类器对特征向量进行分类，得到未加权平均召回率(Unweighted Average Recall，UAR) 对分类结果进行评估。

表 1 展示了两个数据集上不同切分方法以及未切分的原始音频在特征集上的分类实验结果。由表 1 对比可知，等时间间隔切分得到的音频对分类准确率的提高不如根据是否为静音片段切分后的识别率提升效果好，甚至在 CCS 任务中低于原始音频的分类效果。这是因为在所用数据集中，有效语音片段在整个语音段中占比较小，等时间间隔切分后会存在大量的静音片段。在这些静音音频段上进行特征提取，得到的特征会使数据集中存在过多的干扰点，降低了所提取特征的区分度，导致分类结果较差。所以我们仅保留根据是否为静音片段切割的音频数据，进行接下来的实验。经过数据增广后，train 和 dev 集的样本数列举如表 2 所示。

表 1　不同切分方法和原始音频在数据集上的 UAR

Tab. 1　UAR of different segmentation methods and original audio on the data sets

		UAR/%
CCS	未切分	61.4
	等时间间隔切分(500 ms)	59.9
	是否为静音片段切分	63.1
CSS	未切分	57.9
	等时间间隔切分(600 ms)	62.0
	是否为静音片段切分	65.4

表 2　train、dev 集中每个类别的样本数

Tab. 2　The number of samples in each category in the train and dev sets

		Train 切分后/原始未切分	Dev 切分后/原始未切分	Sum 切分后/原始未切分
CCS	Negative	503/215	410/183	913/398
	Positive	181/71	106/48	287/119
CSS	Negative	388/243	238/153	626/396
	Positive	119/72	208/142	327/214

3.3　VLAD 编码实验

本文选择对用 openSMILE 工具 function: ComParE_16 从 CCS 和 CSS 数据集提取出的音频文件 LLDs 进行 VLAD 编码。再将编码结果输入到线性 SVM 中进行分类。字典大小为 125、250、500、1000、2000 时的分类结果如表 3 所示: 当字典大小较小时，VLAD 编码的分类表现在未切分的音频上优于表 1 中直接使用 openSMILE 特征集的结果。但是由于每一个局部 LLDs 是一个 130 维的特征向量，当字典码本数较大时，进行 VLAD 编码后每一个音频文件的特征向量维数会达到数十万维，降低了特征之间的区分度，导致最终分类效果下降。所以，当字典大小为 500、1000、2000 时，利用 VLAD 编码结果进行分类

的准确率较字典大小为125、250时大幅下降。

表3 VLAD编码分类结果. *N*: 字典码本数

Tab. 3 Results of VLAD coding. *N*: The number of codebooks

N	未切分 CCS UAR/%	未切分 CSS UAR/%
125	61.5	**69.6**
250	**64.3**	66.8
500	62.6	58.4
1000	58.5	60.2
2000	62.0	59.6

字典中心是通过特征集合聚类而来的聚类中心，对最终的VLAD编码结果有直接的影响。所以字典中心的选取对于VLAD编码最终的分类效果会产生不同的影响。为了探寻字典大小取值对分类效果的影响并且进一步提高系统性能，我们从10到250对字典大小进行遍历，步长取10，得到的结果如图2所示。可以看出，当字典大小较小时，VLAD编码的分类结果在两个任务上都取得了较好的效果。随着字典大小的增加，VLAD编码取得的分类效果整体上呈现逐渐下降的趋势，这也符合VLAD编码适用于字典码本数较小的情况的特性。同样地，从图中也可以观察出最终的分类效果与字典中心的个数也并无直接的线性映射关系，所以无法通过提前预估字典大小的取值来取得最佳的分类效果。

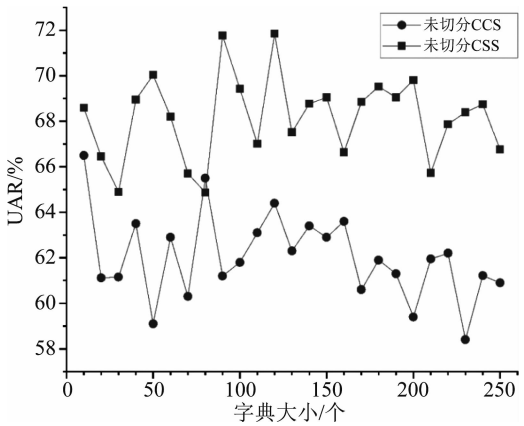


图2 不同字典大小VLAD编码的分类结果

Fig. 2 Classification results of VLAD with different dictionary sizes

对于CCS任务，当字典大小为10时，VLAD编码取得了最佳的分类结果，分类UAR达到了66.5%。同时字典大小为10时，每一个音频文件所对应的特征向量维度仅为1300，远小于openSMILE等基线特征集，运行占用更小的内存，计算速度更快。对于CSS任务，当字典大小为120时，分类效果最佳，UAR为71.8%，远高于表1中使用openSMILE特征集的未切分CSS任务的结果。综上，在两个数据集上的分类实验结果证实了VLAD编码在该任务上的可行性。

3.4 不同特征集和Late fusion分类结果

我们采用线性SVM对提取的数据特征进行分类实验，对比实验结果。利用集成学习的思想，采用投票机制，对各个特征集的分类结果进行Late fusion。使用在dev集上的UAR作为最终分类结果的评价指标。

如表4所列举的实验结果可知，CCS任务中除了VQ和DeepSpectrum外，音频切割后带来的数据量的提升有效地提升了分类的准确率，其中openXBOW取得了较大的提高，字典取值为125时较未切分音频的结果UAR提高了11.7%。同时也说明了去除了静音段的音频片段后提取的音频特征能够对音频进行更好的描述，从而取得更高的分类准确率。我们引入的VQ特征虽然在单独引用时并不能提高分类准确率，但是在与openSMILE进行了前端的融合后，在切分前和切分后都使分类的准确率得到了提高，对比单独使用openSMILE特征，分别提高了0.3%和0.6%。说明引入的VQ特征有效地补充了openSMILE对语音质量的特征描述的缺失，使得特征描述更加完整。

同样在CSS任务中openSMILE由于数据量的增加结果有显著的提升，openXBOW经过切分数据量增加后最好的结果有了明显的提升，成为所有方法中最好的方法，DeepSpectrum、auDeep数据量增加后有一定提升，但效果不明显，End2You结果有所下降但依然接近最好的结果。但是在Late fusion后结果大幅提高。VQ切分后结果比较差，但是整合openSMILE的特征后相比openSMILE有了一定的提升，说明VQ中的特征对特征描述起到了一定程度的补充作用。

最后经过Late fusion后，两个任务集上系统分类准确率都取得了最大值，CCS任务达到73.9%，CSS任务达到了77.2%，相较baseline结果都取得了较大提升。

表 4　音频分类结果. *N*：openXBOW 字典码本数，并且将输入划分为两个大小相同的码本（ComParE-LLDs/ComParE-LLDs-deltas）. *X*：auDeep 中对频率图进行剪切的不同分贝值. VLAD 编码分类结果仅列举最佳结果. Late fusion 采用的是投票机制. Baseline 为 INTERSPEECH ComParE 所提供的 baseline

Tab. 4　Audio classification results. *N*: The number of openXBOW dictionary codebooks, and divide the input into two codebooks of the same size (ComParE-LLDs/ ComParE-LLDs-deltas). *X*: Different decibel values for clipping the frequency map in auDeep. Only the best classification results of VLAD coding are listed. Late fusion uses a voting mechanism. Baseline is the baseline provided by INTERSPEECH ComParE

	未切分 CCS UAR/%	切分后的 CCS UAR/%	未切分 CSS UAR/%	切分后的 CSS UAR/%	Baseline	
					CCS UAR/%	CSS UAR%
	Dev	Dev	Dev	Dev	Dev	Dev
	VQ+SVM					
	59.0	56.1	59.0	61.2	–	–
	openSMILE: ComParE_2016 + SVM					
	61.4	63.1	57.9	65.4	61.4	57.9
	VQ+openSMILE: ComParE_2016+SVM					
	61.7	63.7	57.7	68.3	–	–
N	openXBOW+SVM					
125	60.7	72.4	64.2	70.6	60.7	66.0
250	60.7	69.2	65.4	71.0	60.7	60.6
500	66.4	68.3	66.3	70.9	66.4	64.2
1000	66.2	67.2	67.3	70.9	66.2	62.6
2000	64.7	68.1	67.5	70.8	64.7	66.3
	DeepSpectrum+SVM					
	63.6	61.9	57.4	59.3	63.6	56.0
X/dB	auDeep+SVM					
−30	61.1	62.0	54.0	63.0	60.7	65.8
−45	59.9	63.2	62.4	63.6	64.1	66.3
−60	60.9	66.2	59.3	62.4	67.6	59.4
−75	57.4	65.1	61.6	60.5	64.0	58.4
Fused	59.3	64.6	60.1	62.1	65.4	62.2
	End2You					
	58.5	62.3	76.3	69.2	61.8	70.5
	VLAD					
	66.5	65.4	71.8	68.3	–	–
	Late fusion					
	72.9	**73.9**	**76.7**	**77.2**	–	–

4　结论

为了应对 COVID-19 的快速检测需求，本文提出了基于语音内容分析的新冠肺炎自动识别方法，并在 INTERSPEECH 2021 ComParE 竞赛提供的数据集上进行了验证。针对新冠肺炎数据量小的问题，本文通过语音端点检测方法对数据进行切分增广，提升了小数据集上的分类效果。此外，本文引入语音质量特征，用于对发声器官例如喉、声带等器官的变化进行建模，补充 openSMILE 所提的 LLDs 对音质特征提取的不足。受到 ComParE 竞赛基线系统中 BoAW 编码的启发，本文还引入 VLAD 编码

对低水平特征进行更深层次的描述，使得在字典规模较小时，就能获得更好得分类效果，证实了VLAD编码在语音特征编码中的有效性。引入VLAD编码能够在获得更快系统响应速度的同时占用相对较小的内存空间，便于检测系统在各种功能移动终端上部署，并能快速计算得到新冠检测结果。但是字典的具体大小并不能提前确定，需要进行多次尝试确定最佳取值。在未来的研究中，可以从深度学习的角度引入新的特征提取网络、分类模型对基于语音的新冠肺炎识别检测任务进行提升。

参考文献

[1] ANTONELLI M, CAPDEVILA J, CHAUDHARI A, et al. Optimal symptom combinations to aid COVID-19 case identification: analysis from a community-based, prospective, observational cohort [J]. Journal of Infection, 2021, 82(3): 384-390.

[2] CLEVERLEY J, PIPER J, JONES M M. The role of chest radiography in confirming COVID-19 pneumonia [J]. BMJ (Clinical Research ed.), 2020, 370: m2426-m2426.

[3] BORAKATI A, PERERA A, JOHNSON J, et al. Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study [J]. BMJ Open, 2020, 10(11): e042946-e042946.

[4] TANDAN M, ACHARYA Y, POKHAREL S, et al. Discovering symptom patterns of COVID-19 patients using association rule mining [J]. Computers in Biology and Medicine, 2021, 131: 104249.

[5] ALZUBAIDI M A, OTOOM M, OTOUM N, et al. A novel computational method for assigning weights of importance to symptoms of COVID-19 patients [J]. Artificial Intelligence in Medicine, 2021, 112: 102018-102018.

[6] 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.

[7] 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 (ICASSP). IEEE, 2021: 8328-8332.

[8] WEI Wenqi, WANG Jianzong, MA Jiteng, et al. A real-time robot-based auxiliary system for risk evaluation of COVID-19 infection [J]. arXiv Preprint arXiv: 2008.07695, 2020.

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

[10] SCHULLER B W, BATLINER A, BERGLER C, et al. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates [J]. arXiv Preprint arXiv: 2102.13468, 2021.

[11] FLORIAN E, MARTIN W, BJÖRN S. Opensmile: the munich versatile and fast open-source audio feature extractor [C]. Proceedings of the 18th ACM International Conference on Multimedia, 2010: 1459-1462.

[12] MAXIMILIAN S, BJÖRN S. openXBOW-Introducing the Passau open-source crossmodal bag-of-words toolkit [J]. The Journal of Machine Learning Research, October 2017, 18(96): 1-5.

[13] AMIRIPARIAN S, GERCZUK M, OTTL S, et al. Snore sound classification using image-based deep spectrum features [J]. Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, (Stockholm, Sweden), ISCA, August 2017: 3512-3516.

[14] AMIRIPARIAN S, FREITAG M, CUMMINS N, et al. Sequence to sequence autoencoders for unsupervised representation learning from audio [M]. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop, 2017: 17-21.

[15] TZIRAKIS P, ZAFEIRIOU S, SCHULLER B W. End2You—the Imperial toolkit for multimodal profiling by end-to-end learning [J]. arXiv Preprint arXiv: 1802.01115, 2018.

[16] MAIDMENT J A. The phonetic description of voice quality [J]. Journal of the International Phonetic Association, 1981, 11(2): 78-84.

[17] 朱建清，林露馨，沈飞，等. 采用SIFT和VLAD特征编码的布匹检索算法 [J]. 信号处理，2019，35(10)：1725-1731.
ZHU Jianqing, LIN Luxin, SHEN Fei, et al. Fabric retrieval algorithm using SIFT and VLAD feature coding [J]. Journal of Signal Processing, 2019, 35(10): 1725-1731. (in Chinese)

[18] BALAJI B, ORUGANTI V R M. Multi-level feature fu-

sion for group-level emotion recognition [C]. Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017: 583-586.

[19] TZIRAKIS P, ZHANG Jiehao, SCHULLER B W. End-to-end speech emotion recognition using deep neural networks [C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5089-5093.

[20] MATTHEW G, PETER L. Phonation types: a cross-linguistic overview [J]. Journal of Phonetics, 2001, 29(4): 383-406.

[21] ALKU P, BÄCKSTRÖM T, VILKMAN E. Normalized amplitude quotient for parametrization of the glottal flow [J]. The Journal of the Acoustical Society of America, 2002, 112(2): 701-710.

[22] CHILDERS D G, LEE C K. Vocal quality factors: analysis, synthesis, and perception [J]. The Journal of the Acoustical Society of America, 1991, 90(5): 2394-2410.

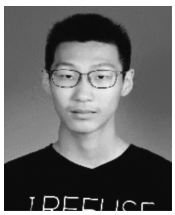
[23] KANE J. Tools for analysing the voice: developments in glottal source and quality analysis [D]. Dublin, Ireland, Trinity College, 2012.

[24] CHRISTER G, AILBHE N C. The role of voice quality in communicating emotion, mood and attitude [J]. Speech Communication, 2003, 40(1): 189-212.

[25] LI Ya, NICK C, TAO Jianhua. Voice quality: not only about "you" but also about "your interlocutor" [C]. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015): 4739-4743.

[26] CAMPBELL N. Listening between the lines: a study of paralinguistic information carried by tone-of-voice [C]. International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004.

作者简介



张昊然　男，1999 年生，江苏南通人。北京邮电大学研究生，主要研究方向为语音语言处理、机器学习。

E-mail: zhanghaoran@bupt.edu.cn

韩易辰　男，1997 年生，内蒙古自治区包头市人。北京邮电大学研究生，主要研究方向为语音合成、情感计算。

E-mail: adelacvgaoiro@bupt.edu.cn

谭咏梅　女，1975 年生，云南丽江人。北京邮电大学副教授，博士，主要研究方向为自然语言处理、机器学习。

E-mail: ymtan@bupt.edu.cn

李　雅（通讯作者）　女，1984 年生，陕西西安人。北京邮电大学副教授，博士，主要研究方向为语音交互、多模态情感计算。

E-mail: yli01@bupt.edu.cn