激光与光电子学进展

基于联合端点检测和动态范围控制的语种识别

王延凯，龙华*，邵玉斌，杜庆治，王瑶
昆明理工大学信息工程与自动化学院，云南 昆明 650500

摘要 在语种识别系统中，静音段干扰、话音分贝范围不一致均会导致语种识别性能下降。此外，利用语谱图进行语种识别的算法由于无法有效展现其低频部分的信息，也会导致语种识别性能无法提升。为此，提出了一种基于联合端点检测和动态范围控制的语种识别方法。首先提取语音梅尔倒谱系数的第一维系数，随后使用中值滤波对特征参数进行平滑处理并进行端点检测以去除语音中静音段干扰，其次使用动态范围控制来调整不同语音的分贝值范围，最后将log刻度语谱图输入到卷积神经网络中进行分类。实验结果表明，在ResNeSt网络中，在VoxForge公共语料库下，所提算法相比传统的基于语谱图的语种识别算法性能提升了7.16个百分点。此外，在相同实验设置下，log刻度语谱图的识别性能也优于其他主流特征，充分验证了所提算法和特征的有效性与优越性。

关键词 傅里叶光学与信号处理；语种识别；端点检测；动态范围控制；语谱图；卷积神经网络
中图分类号 TN912.34 **文献标志码** A **DOI**：10.3788/LOP202259.1307001

Language Identification Using Joint Voice Activity Detection and Dynamic Range Control

Wang Yankai, Long Hua*, Shao Yubin, Du Qingzhi, Wang Yao
Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, *Yunnan, China*

Abstract In the language identification system, the interference of silent segments and the inconsistency of voice decibel range leads to a decline in language identification. Additionally, algorithms using spectrograms for language identification cannot effectively show the information of its low-frequency part, which results in performance failure. To mitigate this, we proposed a language identification method based on joint voice activity detection and dynamic range control. First, we extracted the first dimension coefficient of the Mel-scale frequency cepstral coefficients. Second, we applied median filtering to smooth the feature parameters and perform voice activity detection to remove the silent segment of the voice. Next, we used the dynamic range control to adjust the decibel range of different voices. Finally, we put the log scale spectrogram into the convolutional neural network for classification. The experimental results show that the proposed algorithm improved performance by 7.16 percentage points as compared with the traditional language identification algorithm using spectrogram in the VoxForge public corpus under the ResNeSt network. Additionally, under the same experimental settings, the recognition performance of the log scale spectrogram showed superiority over other mainstream features, which fully validates the effectiveness and superiority of the proposed algorithm and features.

Key words Fourier optics and signal processing; language identification; voice activity detection; dynamic range control; spectrogram; convolutional neural network

1 引 言

语种识别的任务就是正确识别语音序列所属的语言类别，是近年来语音研究的一个热点方向。它一般作为其他语音处理的前端，在电台监测、语音识别、信息检索、机器自动翻译等方面有着广泛的应用。

针对语种识别的研究，国内外的研究者首先对语音特征展开了研究，十几年来先后提出使用梅尔频率倒谱系数(MFCC)[1]，偏移差分倒谱(SDC)[2]，伽马频

收稿日期：2021-07-12；**修回日期**：2021-08-05；**录用日期**：2021-08-13
基金项目：国家自然科学基金(61761025)
通信作者：*2748373869@qq.com

率倒谱系数(GFCC)[3],矢量因子(I-Vector)[4]等底层声学特征进行语种识别。

随着神经网络的发展,底层声学特征又分别作为卷积神经网络(CNN)[5]、深度神经网络(DNN)[6]的输入,通过不断地迭代和网络学习能力,取得了比经典的通用背景模型(GMM-UBM)[7]更高的识别准确率。蒋兵等[8]借助DNN出色的特征提取能力,提取了深瓶颈特征,有效提升了短时语音和易混淆方言的识别性能;Bhanja等[9]将色度特征和MFCC特征融合,通过提高特征维度牺牲算力来提升识别性能;Bhowmick等[10]提出使用GMM-UBM和N元语言模型(N-Gram)提取语音的超向量,并用奇异值分解(SVD)对数据进行降维,在保证准确率的同时也提高了语种识别速度;Garain等[11]将底层声学特征转化为图像并在CNN上进行识别;但这些方法的特征表现形式都比较单一,且方法的鲁棒性能不佳。Montavon等[5]将同时含有时域和频域信息的语谱图作为CNN的输入,取得了比声学特征更高的识别准确率;文献[12-13]将滤波器组参数(Fbank)特征转化为图像特征,并将频率刻度转为log形式,提升了语种识别性能。目前的语种识别方法通过新的语种特征或提升底层声学特征维度来提升语种识别准确率,但这些方法并未考虑到在实际生活中语音信号往往受到静音段、语音分贝范围不一致等问题的干扰,从而造成语音质量和可懂度降低[14-15],此外一些语种识别算法的输入特征能否有效展现语音信息也会影响语种识别性能。

为了改善语种识别过程中静音段的干扰,本文首先通过对第一维的MFCC($MFCC_0$)特征进行中值滤波,将有声段和无声段的特征进行平滑处理,以便更好地进行端点检测(VAD)并去除语音中的静音段;其次,针对不同语音分贝值差异问题,提出使用动态范围控制(DRC)单元来调整语音分贝值在同一动态范围;然后针对传统linear刻度语谱图无法有效展示其低频部分信息的问题,提出将语谱图的频率刻度设置为log形式,在有效展示低频部分信息的同时也不缺失高频部分的信息;最后将log刻度语谱图输入到分类网络中进行分类。实验结果表明,本文所提方法在VoxForge公共语料库中取得了97.94%的识别准确率,优于相同实验设置下MFCC-SDC特征、MFCC特征、GFCC特征、Fbank特征和linear刻度语谱图的识别准确率。

2 语谱图有效信息量分析

作为一种包含了语音的时间、频率和能量信息的特征量,语谱图广泛应用于众多语种识别系统中[5,16]。但这些方法对于语谱图的使用并未考虑到静音段的干扰、不同语音的分贝范围以及是否有效展现低频部分的语谱图信息等情况对于识别结果的影响。本文以此为出发点,分别在以下3个方面展开研究:1)利用VAD单元对语音进行端点检测并去除静音段的干扰;2)使用DRC单元控制语音分贝值在同一动态范围;3)提出使用log刻度语谱图,以充分展现低频部分的语谱图信息。

2.1 语音端点检测VAD单元

语音序列中包含静音点和有声点。假设在长度为S的语音序列中静音点的个数为L_1,有声点的个数为L_2,且满足$L_1+L_2=S$。当序列中某采样点$x(i)$,$(1\leqslant i\leqslant S)$为静音时,随机变量$\xi$取值为0,表示静音点无法带来有用信息,当$x(i)$为有声点时,随机变量$\xi$取值为1,表示有声点能够带来有用信息,此时随机变量ξ的期望为$E_1=E(\xi)=L_2/S$,表示此段语音可以带来的平均信息;若此时去掉原序列中L_3个静音点,为使不同的语音具有相同的序列长度,则需要对应补充L_3个有声点,此时总语音点数保持S不变,有声点的个数变为L_2+L_3,静音点的个数变为L_1-L_3,期望对应变为$E_2=E(\xi)=(L_2+L_3)/S$。去除静音前后满足$E_1<E_2$,可见在去除静音点后语音可以带来的平均信息增多(如表1所示),所以在众多的语音应用中,经常采用端点检测技术去除语音中的静音段。

表1 VAD前后概率分布变化

Table 1 Probability distribution change before and after VAD

Probability distribution before VAD	ξ	1	0
	P	$\frac{L_2}{S}$	$\frac{L_1}{S}$
Probability distribution after VAD	ξ	1	0
	P	$\frac{L_2+L_3}{S}$	$\frac{L_1-L_3}{S}$

$MFCC_0$对语音端点具有较好的跟踪能力,但是在有声段字间的波动较大,很容易在双门限法的端点检测中出现误识别[17]。结合中值滤波平滑序列的作用[18],本文提出使用中值滤波对$MFCC_0$参数进行平滑处理,从而使静音段和有声段之间的区别更加明显。其平滑过程如下:

对输入长度为M的$MFCC_0$特征序列$C[n]$,选定一个长度L(L为奇数)的移动滑块。特征序列$C[n]$中值滤波的输出$M_L\{C[n]\}$是L个信号值$\{C[n],C[n-1],\cdots,C[n-L+1]\}$的中值,具体表示为

$$\hat{C}[n]=M_L\{C[n]\}=\mathrm{med}_{m=0}^{L-1}C[n-m], \quad (1)$$

式中,$\hat{C}[n]$表示中值滤波后的$MFCC_0$特征参数,其中$0<n\leqslant M$。实验中L设置为11。

从图1可以看到,中值滤波后的$MFCC_0$特征对于静音段和有声段的区分更加明显。依此,将滤波后的$MFCC_0$特征通过单参数的双门限法进行端点检测,并根据检测结果去除语音中静音段的干扰。

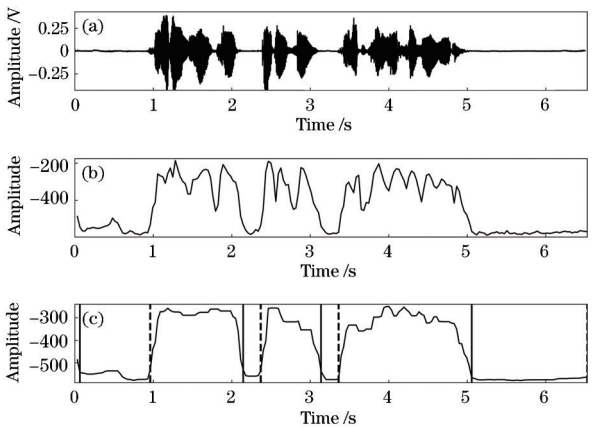

图 1　$MFCC_0$特征端点检测。(a) 语音波形;(b) $MFCC_0$特征;(c) 中值滤波后$MFCC_0$特征端点检测结果

Fig. 1　$MFCC_0$ feature voice activity detection. (a) Voice waveform; (b) $MFCC_0$ features; (c) $MFCC_0$ feature voice activity detection result after median filtering

2.2　动态范围控制DRC单元

2017年叶中付等[19]在进行语种识别任务中发现语种类内多样性可以导致测试样本与训练样本不匹配,进而影响到了后端的语种识别性能。事实上,在大多数的语种识别任务中,大量的人参与了语料库的制作,同一语种往往包含不同的说话人$\{S_1, S_2, \cdots, S_Q\}$,且相同语种不同语音$\{T_1, T_2, \cdots, T_Z\}$的分贝范围(−55～−4 dB)具有较大差异,增大了后续语种识别的难度。因此本文提出对每个语音序列乘以动态参数k,来使不同的语音都在同一分贝范围内波动。式(2)～(3)展示了语音序列的分贝值计算方法:

$$x=\sum_{i=0}^{(N/L)-1}20\lg\frac{P_{\mathrm{rms}}}{P_{\mathrm{ref}}},\tag{2}$$

$$P_{\mathrm{rms}}=\sqrt{\frac{\sum_{j=iL}^{(i+1)L}x_j^{\,2}}{L}},\tag{3}$$

式中:N表示语音序列的长度;L为计算语音均方根的序列长度(本文取$L=N$);P_{ref}为语音的最大振幅值(语音位深度为16,则$P_{\mathrm{ref}}=2^{16}$);P_{rms}为当前声音序列的均方根(RMS);x_j表示语音第j个采样点。

实验中表明,人耳可以接受的语音分贝范围为−25～−15 dB,当分贝值低于−25 dB时人耳很难听清话音内容,而高于−15 dB则又会出现爆破音,两种情况都会干扰人耳获取语音信息。因此针对此人耳的听觉特性,提出动态范围控制单元DRC,实现对高分贝值的语音进行抑制,对低分贝值的语音进行提升的功能。因为DRC单元和机器学习算法中的sigmoid函数趋势类似,所以对sigmoid函数按照上述分析内容进行参数调整:

$$D_{\mathrm{out}}=\frac{10}{1+0.3\mathrm{e}^{(-8-0.5D_{\mathrm{in}})}}-24,\tag{4}$$

式中:D_{in}代表输入语音的分贝值,该值根据式(2)～(3)确定;D_{out}代表输出语音分贝值。

图2中DRC单元通过对原始语音信号同乘动态常数k,实现对高分贝值语音的抑制,低分贝值语音的提升,从而保证不同语音的分贝值在同一范围内波动。

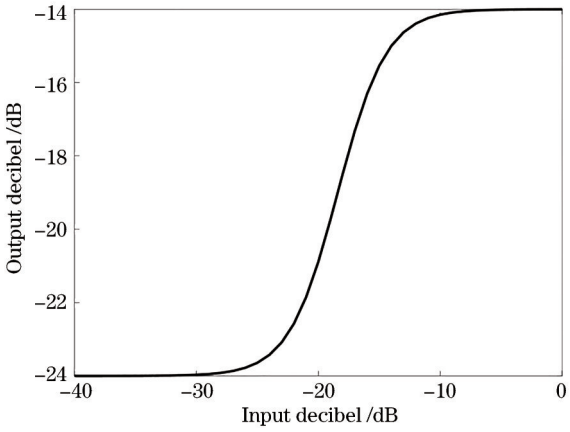

图 2　DRC输入输出处理单元

Fig. 2　DRC input/output processing unit

通过联合式(2)～(4),计算$D_{\mathrm{out}}-D_{\mathrm{in}}$,便可得到动态常数$k$和输入分贝值$D_{\mathrm{in}}$、输出分贝值$D_{\mathrm{out}}$之间的关系:

$$D_{\mathrm{out}}-D_{\mathrm{in}}=\sum_{i=0}^{(N/L)-1}20\lg k,\tag{5}$$

$$k=\sqrt{10^{\frac{D_{\mathrm{out}}-D_{\mathrm{in}}}{10}}},\tag{6}$$

式(6)表示简化后的动态常数k,其中$D_{\mathrm{out}}-D_{\mathrm{in}}$表示调整前后语音的分贝值增益变化。

$$y(n)=\sqrt{10^{\frac{D_{\mathrm{out}}-D_{\mathrm{in}}}{10}}}\cdot x(n),\tag{7}$$

式中,$x(n)$、$y(n)$$(1\leqslant n\leqslant N)$分别表示经过DRC单元前、后语音的第$n$个采样点的采样值。

图3为DRC处理前后语音的变化对比。

2.3　语谱图频率尺度分析

语谱图中包含了语音的时间、频率和能量信息,因为人耳的敏感区间集中在300～3400 Hz,所以linear刻度的语谱图频率刻度并不能直观地展现出300～3400 Hz的语音细节信息。本文提出将语谱图的频率刻度改为log形式,以突显语谱图中低频部分的细节信息。

图4(a)展示的是在linear刻度下某条语音的语谱图,图4(b)展示的是在log刻度下同一条语音的语谱图,相比于图4(a),图4(b)在突显低频部分语谱图的同时,也保留了高频部分的语谱图信息。

3　语种识别及参数设置

在公共语料库环境下,不同说话人的语音文件在性别、年龄、信道、语音分贝、静默时长、背景环境等方

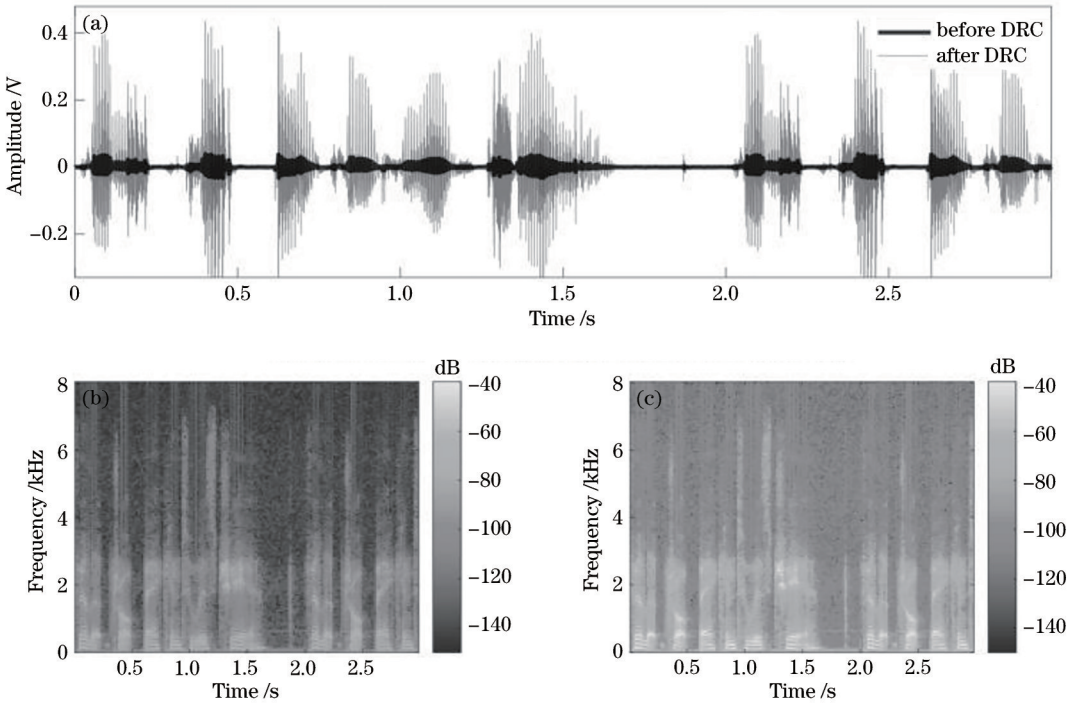

图 3　DRC 处理前后语音变化。(a) DRC 处理前后语音波形变化;(b) DRC 处理前语谱图;(c) DRC 处理后语谱图

Fig. 3　Voice changes before and after DRC processing. (a) Voice waveform changes before and after DRC processing; (b) spectropram before DRC processing; (c) spectropram after DRC processing

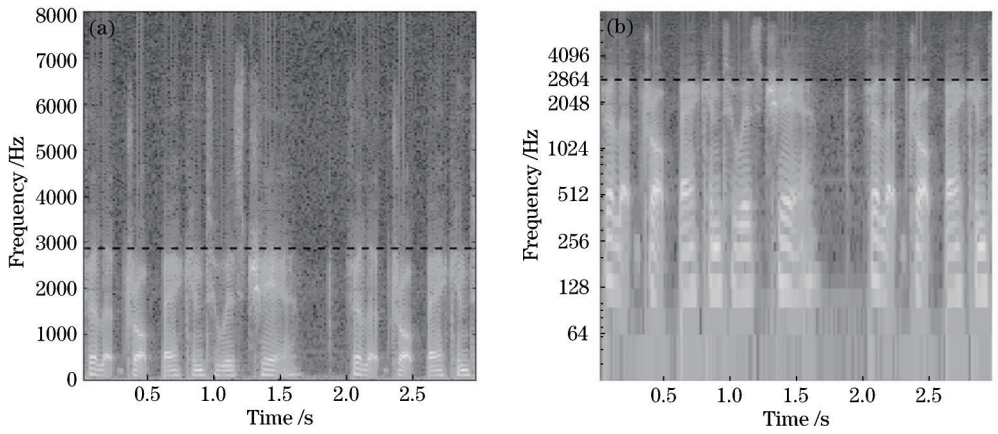

图 4　不同频率尺度比较。(a) linear 刻度语谱图;(b) log 刻度语谱图

Fig. 4　Comparison of different frequency scales. (a) Linear scale spectrogram; (b) log scale spectrogram

面存在较大差异。本文在实验过程中发现语音中静音段以及语音分贝浮动对于识别结果有较大的干扰。为消除此类干扰,分别提出了使用 VAD 单元对中值滤波后的 $MFCC_0$ 特征进行单参数的双门限端点检测并根据检测结果去除语音中的静音段;使用 DRC 单元来控制语音的分贝范围在人耳正常范围内,最后将 log 刻度语谱图送入分类网络得到识别结果。整体算法如图 5 所示。

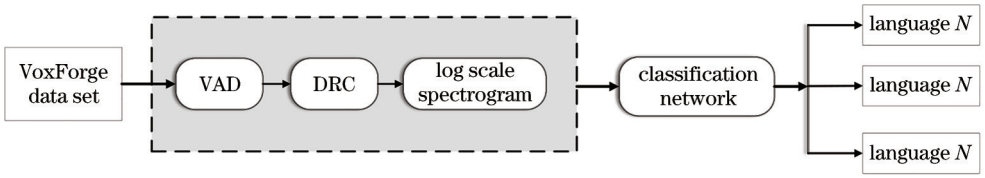

图 5　语种识别流程图

Fig. 5　Flow chart of language recognition

3.1　数据准备

实验中所有语音文件来自VoxForge公共语料库，并选取(French，German，Spanish，Italian，Russian)作为本文所用数据，每条语音均采用16 k采样，16 bit量化，脉冲编码调制(PCM)，语音长度为3 s。每语种选取1500条语音，其中1200条用来训练，300条用来测试，训练集和测试集没有任何交集，表2展示的是数据的分配以及每种语言参与的人数。

表2　训练集、测试集数据分配

Table 2　Data allocation of training set and testing set

Language type	Training set		Testing set		Total wav number	Duration /s
	Wav number	People number	Wav number	People number		
French	1200	150	300	149	1500	3
German	1200	150	300	150	1500	3
Spanish	1200	151	300	151	1500	3
English	1200	169	300	154	1500	3
Italian	1200	151	300	150	1500	3
Russian	1200	150	300	148	1500	3
Total	7200	921	1800	902	9000	

3.2　参数设置

在基于语谱图进行分类的识别系统中，其帧长、帧移的设置对分类的结果也会产生影响，文献[20]对语谱图的不同帧长、帧移设置了对比实验，并在帧长设置为256、帧移设置为128时取得了最好的分类效果。故本文语谱图的帧长帧移也按照此参数设置。

语谱图中linear刻度的频率设置为{0, 1000, 2000, …, 8000}，log刻度的频率设置为{0, 64, 128, …, 8192}。

本文的分类模型采用两种不同分类网络，一种是经典的残差神经网络(Resnet)[21]，被广泛应用于图像分类；另一种是拆分注意力网络(ResNeSt)[22]。两种分类网络均采用101层结构，参数设置(learning_rate为0.0001，epoch为100，batchsize为50)一致，且两种网络的输入特征图大小均为200 pixel×200 pixel。

3.3　测量指标

本文测试指标采用了机器学习领域常用的分类准确率评价指标F_{1_score}和准确率$A_{accuracy}$。如图6所示，与二分类的混淆矩阵不同，多分类任务将分类结果分为5个类别：

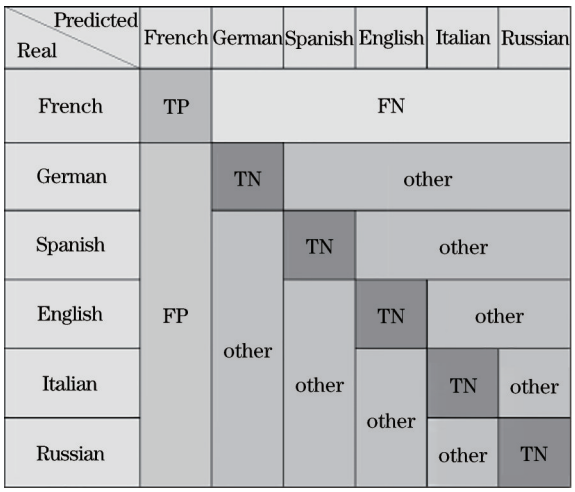

图6　多分类任务评价参数

Fig. 6　Multi-classification task evaluation parameters

图中，TP(true positive)为预测是正确的正样本，S_{TP}；TN(true negative)为预测是正确的负样本，S_{TN}；FP(false positive)为预测是错误的正样本，S_{FP}；FN(false negative)为预测是错误的负样本，S_{FN}；other为与当前预测无关的错误样本，S_{other}。

以下是测量指标F_{1_score}和$A_{accuracy}$的求取步骤。

步骤1：求取每语种的精确率P_k和召回率R_k

$$P_k=\frac{S_{TP}}{S_{TP}+S_{FP}},\tag{8}$$

$$R_k=\frac{S_{TP}}{S_{TP}+S_{FN}},\tag{9}$$

式中，P_k和R_k分别代表的是第k个语种的识别精确率和召回率，$k=1,2,\cdots,K$，K表示的是语种的数目，精确率表示的是识别出的语种有多少比例是正确的，召回率表示的是正确识别的语种占真实值的比例。

步骤2：求取第k个语种的为精确率和召回率的调和均值

$$F_{1\_k}=\frac{2P_kR_k}{P_k+R_k}。\tag{10}$$

步骤3：对所有语种的调和均值F_{1_k}求均值，并得到最终的F_{1_score}指标

$$F_{1\_score}=\frac{1}{K}F_{1\_k}。\tag{11}$$

步骤4：计算准确率$A_{accuracy}$

准确率表示分类正确的样本个数占所有样本个数的比例

$$A_{accuracy}=\frac{S_{TP}+S_{TN}}{S_{TP}+S_{TN}+S_{FP}+S_{FN}+S_{other}}。\tag{12}$$

本文实验结果采用了$A_{accuracy}$和F_{1_score}作为识别结果的评价指标。

4　实验结果及分析

在网络参数设置一致的情况下，本文实验分为3部分：1）在未采用VAD、DRC的识别系统中，分别检验了语谱图频率坐标刻度（linear、log）对于识别结果的影响，具体实验结果见图7；2）分别验证了VAD单元、DRC单元在两种分类网络下对识别结果的影响，具体实验结果见图8、图9；3）在同样采用了VAD单元、DRC单元、相同的帧长帧移以及ResNeSt的情况下，分别将log刻度语谱图与MFCC、MFCC-SDC、GFCC、Fbank特征、linear刻度语谱图特征进行对比，实验结果见表3。

4.1　实验1

为验证语谱图频率坐标对于实验结果的影响，本文在未采用VAD单元、DRC单元的情况下，分别在两种分类网络下检验了linear刻度语谱图和log刻度语谱图的识别结果，实验结果如图7所示。

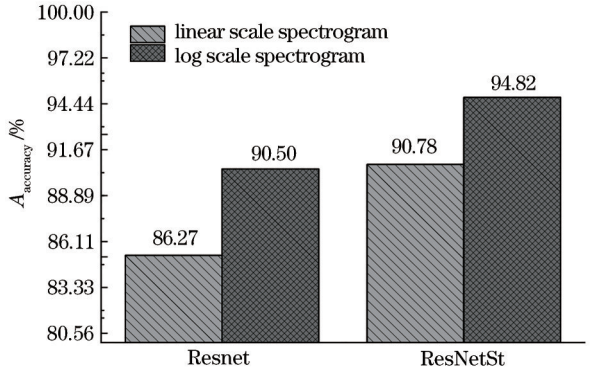

图7　不同频率坐标尺度结果

Fig. 7　Results of different frequency coordinate scales

相比linear刻度的识别结果，log刻度语谱图因为更加符合人耳的听觉机制，突出显示了语谱图中中低频的信息，使得后续的分类网络可以充分挖掘隐藏在语谱图中的信息，所以识别准确率分别在Resnet分类器下提升4.23个百分点，在ResNeSt分类器下提升4.04个百分点。充分验证了log刻度语谱图的有效性。

4.2　实验2

为测试本文提出的VAD单元和DRC单元对于识别结果的影响，分别在Resnet和ResNeSt两种分类网络下对以下4种情况进行了实验：1）log刻度语谱图；2）VAD+log刻度语谱图；3）DRC+log刻度语谱图；4）VAD+DRC+log刻度语谱图。

由图8数据可知，在经典的Resnet下，相比只采用log刻度语谱图的识别结果，应用VAD单元后准确率提高3.44个百分点，应用DRC单元后准确率提高3.28个百分点，同时应用VAD单元和DRC单元后准确率提高3.83个百分点。

由图9数据可知，在ResNeSt下，应用VAD单元后识别准确率提高2.79个百分点，应用DRC单元后准确率提高2.68个百分点，同时应用VAD单元和DRC单元后，准确率提高3.12个百分点。

分析图8、图9的识别结果可知，本文提出的VAD单元和DRC单元在两种分类网络下的识别准确率均有不同程度的提升。结合图7中的linear刻度语谱图在两种分类网络中的语种识别结果可知，图8、图9中，linear刻度语谱图在经过VAD+DRC和log刻度语谱图的操作处理后，其语种识别准确率在Resnet和ResNeSt两种分类网络下分别提升了8.06个百分点和7.16个百分点；此外对比两种分类网络的识别结果，因为ResNeSt对特征图的不同通道赋予不同权重，相比Resnet具有更多的参数，所以其识别性能相比Resnet普遍有约4个百分点的提升，但两种不同网络下的识别结果，都充分验证了本文所提的VAD单元和DRC单元的有效性。

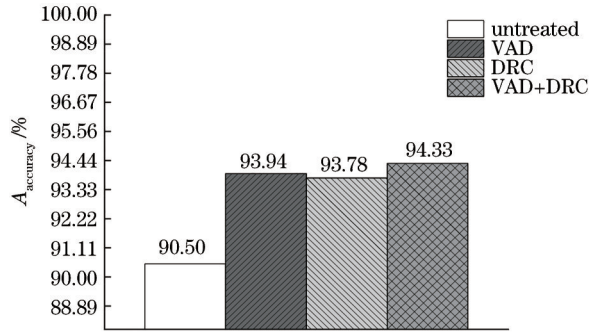

图8　Resnet分类结果

Fig. 8　Resnet classification results

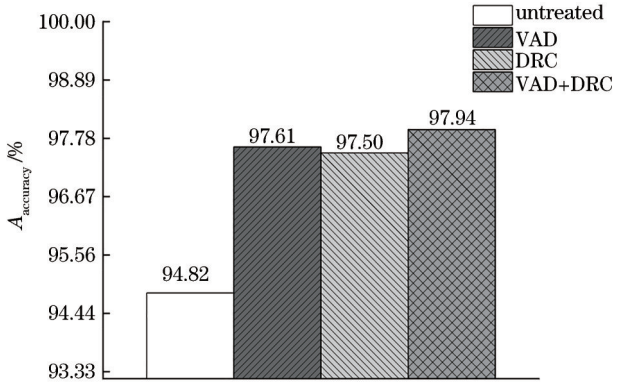

图9　ResNeSt分类结果

Fig. 9　ResNeSt classification results

4.3　实验3

为测试不同语种特征的语种识别结果，分别将近几年常用的MFCC[1]、MFCC-SDC[2]、GFCC[3]、Fbank[5]、linear刻度语谱图[13]以及本文提出的log刻度语谱图绘制成红绿蓝（RGB）彩色图以测试不同特征的识别准确率（如表3所示）。其中语料库使用2.1节的VoxForge数据，且均经过提出的VAD单元和DRC单元处理；分类网络采用2.2节ResNeSt；语音信号的帧长、帧移设置为2.2节所提到的256和128。

表 3　不同特征语种识别结果比较

Table 3　Comparison of language identification results of several different features

Feature	(Frame_number, Data_dimension)	$A_{accuracy}$ /%
MFCC-SDC	(374, 56)	65.72
MFCC	(374, 39)	80.88
GFCC	(374, 32)	85.44
Log scale Fbank feature	(374, 64)	93.05
Linear scale spectrogram	(374, 128)	93.66
Log scale spectrogram (proposed)	(374, 128)	97.94

分析发现 MFCC-SDC 特征、MFCC 特征、GFCC 特征、Fbank 特征、linear 刻度语谱图特征的识别准确率总体呈递增趋势，其中 56 维的 MFCC-SDC 特征识别效果较差，因为其特征是由 7 维 MFCC 特征和 49 维 SDC 特征组合而成，两种特征值大小存在较大差异，所以在将数值映射为图像时两种特征会有较大颜色和亮度的差别，从而使得后续的分类网络难以准确区别出语种信息；此外，GFCC 特征的识别准确率高于 MFCC 特征 4.56 个百分点，log 刻度 Fbank 特征的识别准确率高于 GFCC 特征 7.61 个百分点，linear 刻度语谱图结果略高于 log 刻度的 Fbank 特征 0.61 个百分点。通过分析各特征的提取过程可以发现，语谱图和 Fbank 特征相差梅尔滤波；MFCC 和 GFCC 特征则是通过相应的梅尔滤波、伽玛通滤波后去相关化得到，其中伽玛通滤波器相比梅尔滤波器更加符合人耳的听觉机制，所以 GFCC 特征的识别结果高于 MFCC 特征；SDC 特征则是对 MFCC 前 7 维特征作近一步处理。根据数据处理定理（信息论中平均交互信息量的不增性），语谱图相比 Fbank 特征和 MFCC、GFCC、SDC 特征含有更多的信息量。此外 3.1 节实验 1 结果也表明，log 刻度的语谱图在不丢失信息量的同时突出了低频部分的语谱图信息，从而使得识别准确率高于 linear 刻度的识别准确率。所以本文提出的 log 刻度语谱图取得了最好的识别结果 97.94%，分别高于 MFCC-SDC 特征 32.22 个百分点、MFCC 特征 17.06 个百分点、GFCC 特征 12.50 个百分点、Fbank 特征 4.89 个百分点、linear 刻度语谱图 4.28 个百分点，充分验证了提出的 log 刻度语谱图特征的有效性。

4.4　识别结果分析

本文在 VoxForge 公开数据集上进行训练，并对 6 个语种（French, German, English, Spanish, Italian, Russian）每语种抽取 300 条测试语音，共计 1800 条，在 VAD＋DRC＋log 刻度语谱图＋ResNeSt 情况下取得了最好的识别结果为 97.94%。图 10 中表示的是当前情况下的语种识别结果，每一列表示的是识别为该语种的情况。

分析实验结果可以得知，French 和 Italian 的召回率最高为 0.9900，Spanish 的召回率最低为 0.9667，召回率的数值反映了该语种被算法正确识别的比例；

Real \ Predicted	French	German	Spanish	English	Italian	Russian	Recall	Precision	F_{1_k}
French	297	0	1	1	1	0	0.9900	0.9706	0.9802
German	3	293	1	2	1	0	0.9767	0.9670	0.9718
Spanish	3	3	290	1	3	0	0.9667	0.9898	0.9781
English	1	5	0	294	0	0	0.9800	0.9800	0.9800
Italian	0	1	0	0	297	2	0.9900	0.9770	0.9834
Russian	2	1	1	2	2	292	0.9733	0.9932	0.9832
F_{1_score}									0.9794
$A_{accuracy}$									0.9794

图 10　语种识别结果混淆矩阵

Fig. 10　Language recognition result confusion matrix

Russian 的精确率最高为 0.9932，German 的精确率最低为 0.9670，精确率的数值反映了算法识别出的该语种有多少比例是正确的。总体的 F_{1_score} 和 $A_{accuracy}$ 指标达到了 97.94%。

同时还可以在语系角度对图 10 的语种识别混淆矩阵做出分析。VoxForge 所选的 6 个语种均为印欧语系，其中英语和德语均属于日耳曼语族，这也解释了在混淆矩阵中 English 有 5 条识别为 German；意大利语、西班牙语、法语均属于罗曼语族，这为 Spanish 分别有 3 条识别为 Italian 和 French 提供了合理解释；俄语独立属于斯拉夫语族，在图 10 的识别结果中，Russian 的错误识别比较均衡，与日耳曼语族相比更倾向于罗曼语族；这也为 Italian 有 2 条错误识别为 Russian 提供了一些解释。

5　结　　论

分别提出：1）使用 VAD 单元去除语音中静音段干扰，以提升语谱图中所携带的信息量；2）使用 DRC 单元动态调整不同语音的分贝值，以减小相同语种不同语音文件的类内差距；3）使用 log 刻度语谱图突出显示低频部分的信息，以便分类网络捕捉到不同语种之间的类间差距。实验表明，提出的 VAD 单元、DRC 单元以及 log 刻度语谱图，对语种的识别准确率均有所提升，并在 VoxForge 的 6 语种语料库中取得了 97.94% 的识别准确率，优于相同实验设置下的其他语种特征识别方法，充分证明了该方法的有效性。未来工作中，需要继续加强对于复杂环境

下语种识别的研究，例如语音时长不同的说话环境、含背景噪声的说话环境等，以进一步提升语种识别性能的鲁棒性。

参 考 文 献

[1] Gunawan T S, Husain R, Kartiwi M. Development of language identification system using MFCC and vector quantization[C]//2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application, November 28-30, 2017, Putrajaya, Malaysia. New York: IEEE Press, 2017: 17618163.

[2] Torres-Carrasquillo P A, Reynolds D A, Deller J R. Language identification using Gaussian mixture model tokenization[C]//2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, May 13-17, 2002, Orlando, FL, USA. New York: IEEE Press, 2002: 757-760.

[3] 张卫强, 刘加. 基于听感知特征的语种识别[J]. 清华大学学报(自然科学版), 2009, 49(1): 78-81.
Zhang W Q, Liu J. Language identification based on auditory features[J]. Journal of Tsinghua University (Science and Technology), 2009, 49(1): 78-81.

[4] Martínez D, Burget L, Ferrer L, et al. iVector-based prosodic system for language identification[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing, March 25-30, 2012, Kyoto, Japan. New York: IEEE Press, 2012: 4861-4864.

[5] Montavon G. Deep learning for spoken language identification[C]//NIPS Workshop on deep learning for speech recognition and related applications, 2009: 1-4.

[6] Lopez-Moreno I, Gonzalez-Dominguez J, Plchot O, et al. Automatic language identification using deep neural networks[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, 2014, Florence, Italy. New York: IEEE Press, 2014: 5337-5341.

[7] Xu Y H, Yang J, Chen J. Methods to improve Gaussian mixture model for language identification[C]//2010 International Conference on Measuring Technology and Mechatronics Automation, March 13-14, 2010, Changsha, China. New York: IEEE Press, 2010: 656-659.

[8] Jiang B, Song Y, Wei S, et al. Deep bottleneck features for spoken language identification[J]. PLoS One, 2014, 9(7): e100795.

[9] Bhanja C C, Bisharad D, Laskar R H. Deep residual networks for pre-classification based Indian language identification[J]. Journal of Intelligent & Fuzzy Systems, 2019, 36(3): 2207-2218.

[10] Bhowmick A, Biswas A, AnveshKumar N, et al. Identification/segmentation of Indian regional languages with singular value decomposition based feature embedding [J]. Applied Acoustics, 2021, 176: 107864-107873.

[11] Garain A, Singh P K, Sarkar R. FuzzyGCP: a deep learning architecture for automatic spoken language identification from speech signals[J]. Expert Systems With Applications, 2021, 168: 114416-114429.

[12] Revay S, Teschke M. Multiclass language identification using deep learning on spectral images of audio signals [EB/OL]. (2019-05-10)[2021-03-05]. https://arxiv.org/abs/1905.04348.

[13] Shukla S, Mittal G. Spoken language identification using ConvNets[M]//Chatzigiannakis I, de Ruyter B, Mavrommati I. Ambient intelligence. Lecture notes in computer science. Cham: Springer, 2019, 11912: 252-265.

[14] 陈湟康, 陈莹. 基于具有深度门的多模态长短期记忆网络的说话人识别[J]. 激光与光电子学进展, 2019, 56(3): 031007.
Chen H K, Chen Y. Speaker identification based on multimodal long short-term memory with depth-gate[J]. Laser & Optoelectronics Progress, 2019, 56(3): 031007.

[15] 任凯龙, 汪毅, 陈晓冬, 等. 用于腹腔镜扶持器控制的特定人语音识别算法[J]. 激光与光电子学进展, 2020, 57(18): 181702.
Ren K L, Wang Y, Chen X D, et al. Speaker-dependent speech recognition algorithm for laparoscopic supporter control[J]. Laser & Optoelectronics Progress, 2020, 57(18): 181702.

[16] Draghici A, Abeßer J, Lukashevich H. A study on spoken language identification using deep neural networks [C]//Proceedings of the 15th International Conference on Audio Mostly, September 14-17, 2020, Graz, Austria. New York: ACM Press, 2020: 253-256.

[17] 吴新忠, 夏令祥, 张旭, 等. 基于谱熵梅尔积的语音端点检测方法[J]. 北京邮电大学学报, 2019, 42(2): 83-89.
Wu X Z, Xia L X, Zhang X, et al. Voice activity detection method based on MFPH[J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(2): 83-89.

[18] 方斌, 陈家益. 去除脉冲噪声的小波阈值去噪算法[J]. 激光与光电子学进展, 2021, 58(22): 2210016.
Fang B, Chen J Y. Wavelet threshold denoising algorithm for impulse noise removal[J]. Laser & Optoelectronics Progress, 2021, 58(22): 2210016.

[19] 叶中付, 戚婷, 李赛峰, 等. 基于LDOF准则的自适应高斯后端语种识别方法[J]. 通信学报, 2017, 38(4): 17-24.
Ye Z F, Qi T, Li S F, et al. Adaptive Gaussian back-end based on LDOF criterion for language recognition[J]. Journal on Communications, 2017, 38(4): 17-24.

[20] Cai Y N, Xu W L. The best input feature when using convolutional neural network for cough recognition[J]. Journal of Physics: Conference Series, 2021, 1865(4): 042111.

[21] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[22] Zhang H, Wu C R, Zhang Z Y, et al. ResNeSt: split-attention networks[EB/OL]. (2020-05-19)[2021-03-04].