文章编号:1007-5321(2021)06-0134-07 **DOI**:10.13190/j.jbupt.2021-053

面向真实噪声环境的语种识别

邵玉斌, 刘 晶, 龙 华, 李一民
(昆明理工大学 信息工程与自动化学院, 昆明 650500)

摘要: 语种识别受真实噪声环境的影响较大,识别效果不佳. 为了解决真实噪声环境下语种识别的问题,提出一种基于对数灰度语谱图的图像处理语种识别方法. 根据噪声能量和语音能量在语谱图上的分布规律对真实噪声中的语音信号进行带通滤波;再结合人耳听觉特性提取对数灰度语谱图;然后提取图像主成分特征作为语种特征,采用残差神经网络模型进行训练测试. 实验结果表明,在掠夺者战斗机驾驶舱的环境下,所提方法的平均识别正确率相对于线性灰度语谱图方法提升了 27.5%,在其他噪声环境下的平均识别正确率也有提升.

关 键 词: 语种识别; 真实噪声环境; 对数灰度语谱图; 残差神经网络; 图像处理
中图分类号: TN912.3 **文献标志码**: A

Language Identification in Real Noisy Environments

SHAO Yu-bin, LIU Jing, LONG Hua, LI Yi-min
(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Language identification is heavily influenced by the real noise environment, resulting in poor identification results. To solve this problem, an image processing method for language identification is proposed based on the logarithmic gray-scale speech spectrogram. The logarithmic gray-scale speech spectrogram is obtained by combining the human auditory properties and the voice filtered in real noise environments according to the different distribution of noise and speech on the speech spectrogram. Then, the principal component of the spectrogram is extracted as language features and, a residual neural network model is used for training and testing. Experimental results show that the average identification rate of the proposed method is improved by 27.5% in the noisy cockpit of a Blackburn Buccaneer compared to the linear grey-scale speech spectrogram method. In other noisy environments, the average identification rate is also improved.

Key words: language identification; real noise environment; logarithmic gray-scale speech spectrogram; residual neural network; image processing

语种识别作为机器翻译、多语言信息服务等自然语言处理技术的前端系统,是研究的热点问题,而真实噪声环境下的语种识别是研究的重点和难点[1]. 依靠人工辨别语种已无法满足实际需求,而且在嘈杂环境下人耳可能无法准确地辨别语种类型,语种特征也容易被干扰或者被掩蔽,无法清晰地表达语种信息. 因此,研究真实噪声环境下的语种识别愈发重要. 语种识别的关键是特征提取和语种

收稿日期: 2021-04-02
基金项目: 国家自然科学基金项目(61761025)
作者简介: 邵玉斌(1970—), 男, 教授, 硕士生导师.
通信作者: 刘 晶(1996—), 男, 硕士生, E-mail:liujing@stu.kust.edu.cn.

模型的构建. 目前传统的语种特征主要是基于声学层特征和音素层特征. 主流的声学层特征包括梅尔频率倒谱系数、伽马通频率倒谱系数[2]等. 以上特征在嘈杂环境下受影响较大,导致识别效果不佳. 基于音素层的语种识别方法是将语音分为一段段的音素序列,再根据不同语种之间的音素搭配进行语种识别[3]. 基于音素层的特征受噪声影响较小,但音素切分提取困难,导致识别性能下降.

随着深度学习被引入语种识别领域, Montavon 等[4]提出将语种识别转为图像识别,提取线性灰度语谱图(LGSS, linear gray scale spectrogram)作为语种特征,该特征在噪声环境下会被掩盖掉大量语种特征,导致低信噪比下难以被正确识别. Jiang 等[5]提出了利用深度神经网络的特征抽取能力提取深度瓶颈特征(DBF, deep bottleneck feature)的方法,抑制了说话人的信息,但是提取过程复杂,实用性较低. Lopez-Moreno 等[6]提出了将特征提取、特征变换和分类器融于一个神经网络的端到端语种模型的方法,加快了训练识别速度. Geng 等[7]提出了将注意力机制模型引入语种识别模型进行特征训练识别的方法,降低了无效特征的干扰. Jin 等[8]提出了从网络的中间层提取语种区分性基本单元特征的方法,但是池化层只是对 LID-senone 做了简单的加权平均,造成大量的信息损失,导致识别误差较大. Cai 等[9]提出了基于可学习字典编码层的端对端系统的方法,从底层直接学习语种类别的信息,摒弃传统的声学模型. Deshwal 等[10]提出了一种基于混合特征提取技术和前馈反向传播神经网络分类器的语言识别方法. Li 等[11]提出了基于多特征和多任务模型的深度联合学习策略的识别方法. Bhanja 等[12]提出了基于自动声调和非声调预分类的语种识别方法.

针对上述方法在真实噪声环境下语种识别存在的不足,提出了基于滤波对数语谱图的图像处理语种识别方法,将语音信号进行带通滤波,再提取对数灰度语谱图,最后提取图像主成分特征作为语种特征. 在 8 种噪声环境下,采用残差神经网络语种模型对提出的方法进行了测试. 测试结果表明,提出的方法初步解决了真实噪声环境下识别正确率低的问题,达到了提高语种识别正确率的目的.

1　构建模型

真实环境下的语音含有多种类型的噪声,因此,采用 Nonspeech 公开噪声库的 8 种噪声构建不同噪声源、不同信噪比的测试语料库,模拟真实环境下采集的语音数据. 噪声源分别为白噪声(WN, white noise)、驱逐舰作战室背景噪声(DORBN, destroyer operations room background noise)、军用车辆噪声(MVN, military vehicle noise)、高频信道噪声(HFCN, high frequency channel noise)、粉红噪声(PN, pink noise)、车内噪声(VN, volvo noise)、F16 座舱噪声(F16CN, F16 cockpit noise)和掠夺者战斗机驾驶舱噪声(BFCN, buccaneer fighter cockpit noise). 将残差神经网络的语种识别模型作为语种模型.

1.1　噪声音频产生模型

在真实噪声环境下定义带噪语音信号为 $x(n)=s(n)+w(n)$,其中:$s(n)$为纯净语音信号,$w(n)$为噪声源信号. 平均信噪比定义为

$$L_{\mathrm{SNR}}=10\lg\frac{\sum_{n=1}^{H}s^2(n)}{\sum_{n=1}^{H}\omega^2(n)} \tag{1}$$

其中:$\sum_{n=1}^{H}s^2(n)$为语音信号能量之和,H 为语音的总采样点数;$\sum_{n=1}^{H}\omega^2(n)$为噪声信号能量之和. 每种噪声源环境下的语音信号波形被噪声淹没的面积不同,使信号处理和语种识别比较困难.

1.2　残差神经网络语种模型

残差神经网络语种模型[13]与普通的非线性卷积网络相比,具有一个向前的线性过程,即在原始输入的基础上增加该层的残差结果,共同作为下一层的输入. 残差网络的线性传递不会破坏到网络的非线性关系,并且模型能够达到更深层,同时学习的映射关系也可以更加复杂. 残差神经网络模型主要解决了深度网络退化和梯度下降的问题,误差比普通的卷积网络小,在图像识别领域的误差只有 3.57%.

2　语种特征提取及处理

特征提取是语种识别中非常重要的环节,一个好的特征需要具备鲁棒性高及语种区分性大的特点. 针对真实噪声环境下语种识别的问题,提出基于图像处理的滤波对数语谱图特征,该特征的提取流程如图 1 所示.

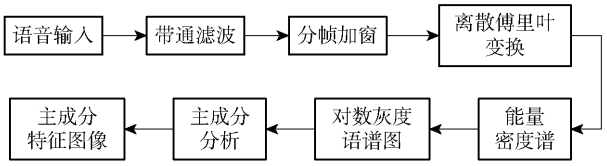


图1 主成分语种特征提取流程

2.1 带通滤波

根据语音学的研究,语音的信息能量主要集中在中低频段,高频部分的信息能量较少,而且人耳对高频信息的分辨率不高[14]. 图2所示为时长为35 min、采样率为8 kHz的一段语音和噪声的功率密度与频率的关系. 由图2可知,语音信息集中在中低频部分,有5种噪声的信息能量在高频部分高于语音信息的能量. 在低频部分有7种噪声的信息能量高于语音信息的能量. 因此,采用带通滤波可以在损失少量语音信息的前提下,滤除大量的高频和少量低频部分的干扰信息. 实验结果表明,通带的取值范围为100~1 500 Hz时的语种识别效果最佳. 图3所示为F16CN源环境下滤波前后的语谱. 由图3可知,高频部分的信息较少,噪声信息较大,滤除高频部分后减少了部分噪声的干扰.

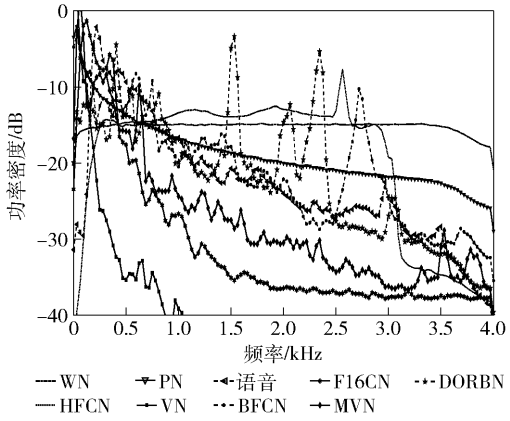
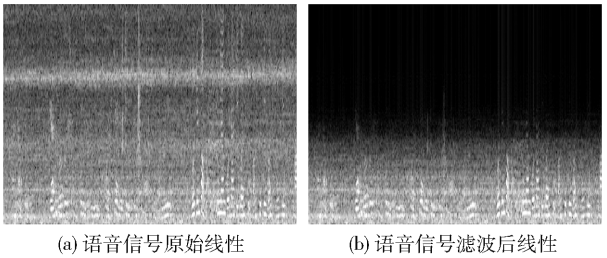


图2 语音和噪声的功率密度与频率的关系



(a) 语音信号原始线性灰度语谱图　(b) 语音信号滤波后线性灰度语谱图

图3 语音信号滤波前后线性灰度语谱图

2.2 对数灰度语谱图生成

语谱图综合了频域和时域特性,包含了大量的语音学信息和语音频谱随时间变化的信息[15]. 语谱图在噪声环境下不会随着噪声的增大而发生内在的变化,只会被噪声掩蔽掉部分信息,从而方便采用图像处理技术进行噪声抑制. 耳蜗的构造决定了听觉频率的空间分布是接近对数的,因此采用对数灰度语谱图(TGSS, logarithmic gray scale spectrogram)特征可以更好地模拟人耳的听觉特性,增强语谱图的辨识度. 滤波对数灰度语谱图(FTGSS, filtered + TGSS)特征的提取包括4个步骤.

1) 对带通滤波后的语音信号 $x(n)$ 分帧加窗,实验的帧长为256,帧移为128,使用的是汉明窗,分帧加窗后的第 i 帧信号为 $x_i(n)$.

2) 对 $x_i(n)$ 进行离散傅里叶变换,有

$$S_i(k) = \sum_{n=1}^{N} x_i(n) e^{-j2\pi kn/N} \tag{2}$$

其中 N 为傅里叶变换的点数.

3) $S_i(k)$ 的能量密度谱为

$$P_i(k) = |S_i(k)|^2 \tag{3}$$

对 $P_i(k)$ 取对数,有

$$P_i^{(dB)}(k) = 10\lg P_i(k) \tag{4}$$

4) 以时间 n 为横轴,对数化频率 $k_1 = \mathrm{lb}(k)$ 为纵轴,$P_i^{(dB)}(k)$ 为灰度级的二维图像绘制的滤波对数灰度语谱图如图4所示.

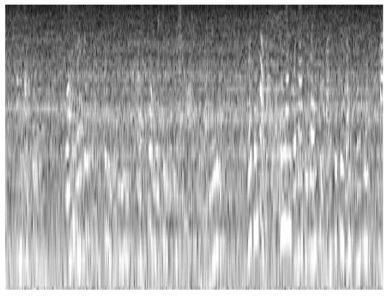


图4 滤波对数灰度语谱图

2.3 主成分语种特征提取

图像是高维的数据,对图像做降维处理可以降低数据的复杂度,从而发现数据的内在规律和本质特征,以便于识别分类[16]. 原始图像包括冗余信息和噪音信息,识别分类时会造成干扰. 通过主成分特征提取可以抑制部分噪声和冗余信息的干扰,从而提高识别的正确率. 采用主成分分析(PCA, principal component analysis)方法[17]对图像进行降维,

将高维的数据映射到低维的空间中表示，使用较少的数据维度保留住较多原数据点的特性. 滤波对数灰度语谱图为 3 通道的，对其进行主成分语种特征提取的步骤如下：

1）将输入的 3 通道 $r \times c \times 3$ 图像矩阵经过变换降维成 $p \times 3$ 的矩阵，其中 $p = r \times c$，r 为图像矩阵的行数，c 为图像矩阵的列数；

2）求出每个通道的均值，再将数据进行中心化处理，经过中心化处理的图像矩阵为

$$\boldsymbol{X}_{(k)} = \boldsymbol{G}_{(k)} - \frac{1}{r \times c} \sum_{q=1}^{r \times c} \boldsymbol{G}_{(k)}(q),\ 0 < k \leqslant 3, 0 < q \leqslant r \times c \tag{5}$$

其中：$\boldsymbol{G}$ 为 $p \times 3$ 的原始图像变换矩阵，k 为矩阵的列数，q 为矩阵的行数；

3）计算样本的协方差矩阵，有

$$\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} \tag{6}$$

4）计算协方差矩阵的特征值并按照降序排列，即特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$，对应的特征向量为 $\boldsymbol{\alpha}_1$，$\boldsymbol{\alpha}_2$，$\boldsymbol{\alpha}_3$；

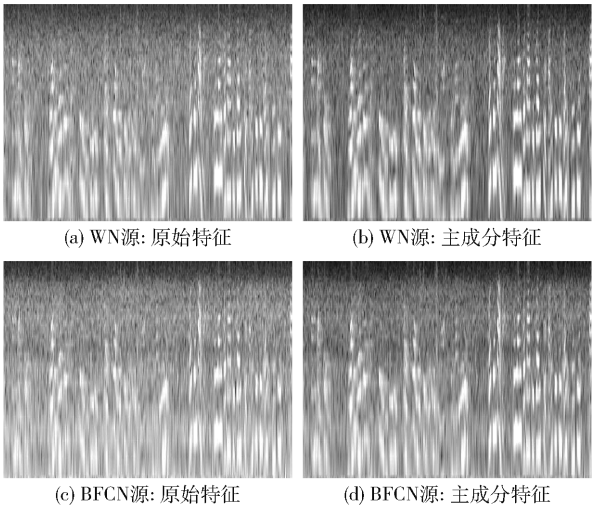
5）$p \times 3$ 的主成分特征矩阵为

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{W} \tag{7}$$

其中 $\boldsymbol{W} = [\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \boldsymbol{\alpha}_3]$ 为特征向量矩阵；

6）根据每个通道的主成分矩阵重构主成分滤波对数灰度语谱图特征（FTGSS + PCA，filtered logarithmic gray scale spectrogram + PCA），从而保留相关度高的语种信息，减少相关度低的噪声信息.

图 5 所示为贡献率最高的第 1 维主成分特征. 从原始图像和主成分图像的对比发现，噪声有所减少，语谱线更清晰.



(a) WN源：原始特征　(b) WN源：主成分特征

(c) BFCN源：原始特征　(d) BFCN源：主成分特征

图 5　原始特征和主成分特征

3　实验结果分析

3.1　实验设计

利用 PyCharm 进行了仿真实验，使用的软件为 pytorch1.4 版，Windows10 操作系统，8 GB 内存，处理器为 Intel-i7-4710MQ. 采用的语料集是中国国际广播电台的广播音频文件，每段音频的持续时间 $t = 10$ s，采样率 $f = 8\ 000$ Hz，单通道，共包括汉语、藏语、维吾尔族语、英语、哈萨克斯坦语等 5 种语种. 测试数据集和训练数据集下的语音数据如表 1 和表 2 所示.

表 1　不同语种、语音（白噪声）下的语音训练数据段

语种	信噪比/dB					
	无噪	25	20	15	10	5
汉语	100	100	100	100	100	100
藏语	100	100	100	100	100	100
维吾尔族语	100	100	100	100	100	100
英语	100	100	100	100	100	100
哈萨克斯坦语	100	100	100	100	100	100

注：训练集共包括 3 000 段语音.

表 2　不同语种、语音（8 种噪声）下的语音测试数据段

语种	信噪比/dB				
	−10	−5	0	5	10
汉语	171	171	171	171	171
藏语	171	171	171	171	171
维吾尔族语	171	171	171	171	171
英语	171	171	171	171	171
哈萨克斯坦语	171	171	171	171	171

注：每种噪声源都需要构建 5 种不同信噪比的测试集，每个测试集包括 855 段语音，共 40 个不同信噪比下的不同噪声源.

采用美国国家标准与技术研究院评测标准中多语种辨识的识别正确率为评价指标，公式为

$$S = \frac{H + Z + W + Y + K}{G} \tag{8}$$

其中：H, Z, W, Y, K 分别为每个语种识别正确的个数，G 为总的测试集个数，S 为识别正确率.

3.2　实验测试与分析

为了测试并分析不同噪声环境下不同方法的语种识别正确率. 基于梅尔尺度滤波器能量（Fbank，log Mel-scale filter bank energies）特征[18]和 DBF[5]，LGSS[4]，TGSS，FTGSS，FTGSS + PCA 设计了 6 组

实验.

3.2.1 模型参数的选取

针对所构建的模型,通过调整模型的网络层数来提升模型的识别正确率. 当网络层数从12层增加到24层时,识别准确率也随之改变;当网络层数为18时,识别的准确率达到最高.

对学习率进行优化,将学习率从0.000 01增加到0.000 3,找出模型的最佳学习率,学习率对识别正确率的影响如图6所示. 由图6可见,随着学习率的改变,识别正确率也随之变化. 学习率过低时,会出现过拟合问题,且引起高偏差现象,从而导致模型不稳定. 实验过程中发现,学习率为0.000 1时,识别正确率最高.

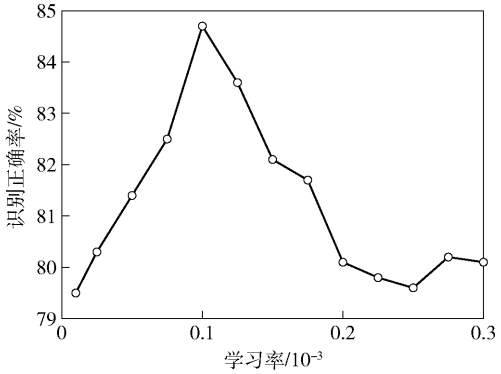


图6 调整学习率对识别正确率的影响

最后进行模型训练迭代周期调试. 训练迭代周期和损失函数的变化如图7所示. 由图7可知,从第28次迭代开始曲线变得平缓,表示已经开始收敛,28~33这个迭代周期比较合理,继续训练会造成模型过拟合现象. 因此,选取迭代周期为30次.

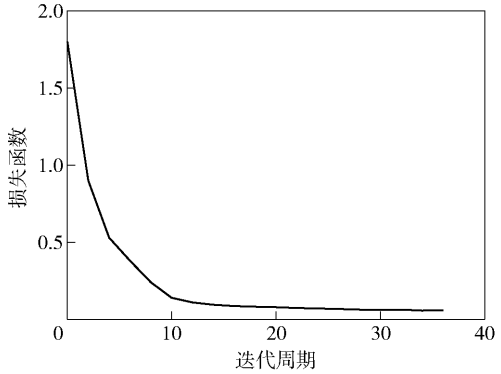


图7 训练迭代周期和损失函数的变化

3.2.2 不同噪声下语种识别的性能

为了测试8种不同噪声源环境下所提方法的鲁棒性和识别性能并分析优劣原因,设计了6组实验.

实验1:提取64维Fbank特征作为语种特征.

实验2:提取LGSS特征作为语种特征.

实验3:提取DBF特征作为语种特征.

实验4:提取TGSS特征作为语种特征,验证该特征相对于LGSS特征的有效性.

实验5:根据不同噪声源和语音信息能量在语谱图上的分布位置,引入带通滤波器,提取FTGSS特征作为语种特征,验证滤波有效抑制部分噪声干扰有效性.

实验6:采用PCA技术对FTGSS特征进行主成分分析,提取FTGSS+PCA特征作为语种特征.

以上6个实验的实验结果如表3所示.

表3 不同噪声源和信噪比下的语种识别正确率 %

噪声源	识别方法	信噪比/dB				
		-10	-5	0	5	10
WN	Fbank	20.3	26.7	33.5	62.8	72.6
	LGSS	20.2	24.2	55.9	69.6	75.1
	DBF	21.3	22.4	67.1	80.4	84.9
	TGSS	23.6	41.3	64.9	80.2	82.9
	FTGSS	27.3	43.7	70.2	81.3	85.7
	FTGSS+PCA	39.4	73.1	84.7	86.6	88.3
DORBN	FTGSS+PCA	28.2	67.9	82.1	82.7	84.9
MVN	FTGSS+PCA	34.1	48.3	64.3	75.2	79.6
HFCN	FTGSS+PCA	36.3	80.9	83.1	83.7	84.3
PN	FTGSS+PCA	38.5	46.3	72.1	81.2	84.1
VN	FTGSS+PCA	52.5	59.9	66.8	74.9	83.7
F16CN	FTGSS+PCA	33.1	40.8	58.4	74.5	81.1
BFCN	FTGSS+PCA	50.1	76.1	78.9	81.6	86.2

由实验1~4可知,对于白噪声环境,在5种信噪比等级下,采用的TGSS特征优于Fbank特征和LGSS特征. 由于TGSS特征更符合人耳的听觉特性,抗干扰的能力更强. 采用TGSS特征比采用DBF特征稍有逊色,DBF特征可以较好地消除说话人信息等与语种无关信息的干扰,但是,当信噪比过低时,DBF特征无法很好地消除噪声的干扰,导致识别效果不佳.

对比实验3~5可知,相对于DBF特征和TGSS特征,采用FTGSS特征在5种信噪比等级下都优于对比方法. 相对于DBF特征,在5种信噪比等级下,识别准确率分别提高了6.0%,21.3%,3.1%,0.9%和0.8%. 由于采用带通滤波可以有效滤除大量高频噪声和低频部分的噪声,相对提高了信噪比.

从实验 5 ~ 6 可知，在白噪声环境下，相对于 FTGSS 而言，FTGSS + PCA 在低信噪比下的识别效果提升明显. 在信噪比为 - 10 dB，- 5 dB 和 0 dB 时，识别正确率分别提升了 12.1%，29.4% 和 14.5%. 由于对 FTGSS 特征进行主成分特征提取可以消除部分贡献率低的噪声干扰，保留了相关度高的语种信息，可间接提高特征的抗干扰能力.

从其他 7 种噪声环境下的识别性能可知，在训练集背景噪声为白噪声，测试集为其他背景噪声的情况下，采用所提 FTGSS + PCA 特征方法依然具有很好的识别效果. 实验结果表明，针对真实噪声环境下的语种识别方法具有较好的鲁棒性和识别性能.

图 8 所示为 8 种噪声源环境下，采用 FTGSS + PCA 特征和 LGSS 特征的平均识别正确率.

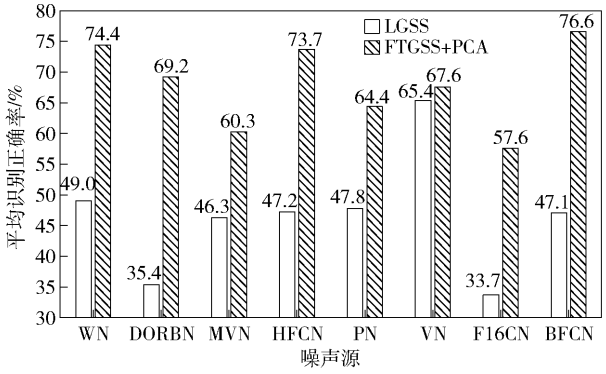


图 8　不同真实噪声源下的平均识别正确率

从图 8 可知，在 8 种噪声源环境下，采用 FTGSS + PCA 特征方法相对于采用 LGSS 特征方法的识别性能均有提升. 在噪声源为 WN，DORBN，HFCN 和 BFCN 下的语种识别正确率分别提升了 25.4%，33.8%，26.5% 和 29.5%. 除 VN 源环境下，其他噪声源环境下语种识别正确率的提升效果也较为明显. VN 源环境下语种识别正确率的提升不明显是因为车内噪声能量在语谱图上的分布是由低频部分决定的. 综上所述，采用 FTGSS + PCA 特征方法很大程度上提高了真实噪声环境下语种识别的正确率，而且还具有较高的鲁棒性.

4　结束语

针对真实噪声环境下语种识别正确率较低和鲁棒性差的问题，提出了一种模拟人耳听觉特性的语谱图和结合图像处理的噪声抑制的语种识别方法. 理论和实验分析结果证明，在 8 种噪声源环境下的语种识别正确率有所提升，且鲁棒性也有所增强，所提出的方法适用于真实噪声环境下的识别. 后续可以探究更适合对语谱图进行噪声抑制的图像处理算法，对不同的噪声源采用不同的处理方法.

参考文献：

[1] Li Haizhou, Ma Bin, Lee K A. Spoken language recognition: from fundamentals to practice[J]. Proceedings of the IEEE, 2013, 101(5): 1136-1159.

[2] 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 Weiqiang, Liu Jia. Language recognition based on auditory perception characteristics [J]. Journal of Tsinghua University (Natural Science Edition), 2009, 49(1): 78-81.

[3] Zissman M A. Comparison of four approaches to automatic language identification of telephone speech [J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(1): 31.

[4] Montavon G. Deep learning for spoken language identification[C]//2009 NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. Vancouver: NIPS Foundation, 2009: 1-4.

[5] Jiang Bing, Song Yan, Wei Si, et al. Performance evaluation of deep bottleneck features for spoken language identification[C]//The 9th International Symposium on Chinese Spoken Language Processing. Singapore: IEEE, 2014: 143-147.

[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 (ICASSP). Florence: IEEE Press, 2014: 5337-5341.

[7] Geng Wang, Wang Wenfu, Zhao Yuanyuan, et al. End-to-end language identification using attention-based recurrent neural networks[C]//Interspeech 2016. San Francisco: ISCA, 2016: 2944-2948.

[8] Jin Ma, Song Yan, Mcloughlin I, et al. LID-senones and their statistics for language identification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 171-183.

[9] Cai Weicheng, Cai Zexin, Liu Wenbo, et al. Insights into end-to-end learning scheme for language identification [J]. IEEE Signal Processing Society Sigport, 2018, 28(2): 202-210.

[10] Deshwal D, Sangwan P, Kumar D. Feature extraction

methods in language identification: a survey[J]. Wireless Personal Communications, 2019, 107(4): 2071-2103.

[11] Li Lin, Li Zheng, Liu Yan, et al. Deep joint learning for language recognition[J]. Neural Networks: the Official Journal of the International Neural Network Society, 2021, 141(9): 72-86.

[12] Bhanja C C, Laskar M A, Laskar R H. Modelling multi-level prosody and spectral features using deep neural network for an automatic tonal and non-tonal pre-classification-based Indian language identification system [J]. Language Resources and Evaluation, 2021, 55(3): 689-730.

[13] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Press, 2016: 770-778.

[14] 刘威. 单通道语音水印与语音增强算法研究[D]. 南京: 东南大学, 2017.

[15] Franzoni V, Biondi G, Milani A. Emotional sounds of crowds: spectrogram-based analysis using deep learning [J]. Multimedia Tools and Applications, 2020, 79(47/48): 36063-36075.

[16] 蓝雯飞, 汪敦志, 张盛兰. 一种新的降维算法 PCA_LLE 在图像识别中的应用[J]. 中南民族大学学报(自然科学版), 2020, 39(1): 85-90.
Lan Wenfei, Wang Dunzhi, Zhang Shenglan. Application of a new dimensionality reduction algorithm PCA_LLE in image recognition[J]. Journal of South-Central University for Nationalities (Natural Science Edition), 2020, 39(1): 85-90.

[17] Qaraei M, Abbaasi S, Ghiasi-Shirazi K. Randomized non-linear PCA networks-science direct[J]. Information Sciences, 2021, 545: 241-253.

[18] Zhu Dong, Huang Ming, Yang Jingjing, et al. Identification of spoken language from webcast using deep convolutional recurrent neural networks[C]//2019 International Conference on Information Technology. Sanya: Electrical and Electronic Engineering (ITEEE 2019), 2019: 1147-1152.