文本无关说话人识别的一种多尺度特征提取方法

陈志高[①②] 李 鹏[③] 肖润秋[①②] 黎 塔[①] 王文超[*①]

[①](中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)

[②](中国科学院大学 北京 100049)

[③](国家计算机网络应急技术处理协调中心 北京 100029)

摘 要：近些年来，多种基于卷积神经网络(CNNs)的模型结构表现出越来越强的多尺度特征表达能力，在说话人识别的各项任务中取得了持续的性能提升。然而，目前大多数方法只能利用更深更宽的网络结构来提升性能。该文引入一种更高效的多尺度说话人特征提取框架Res2Net，并对它的模块结构进行了改进。它以一种更细粒化的工作方式，获得多种感受野的组合，从而获得多种不同尺度组合的特征表达。实验表明，该方法在参数量几乎不变的情况下，等错误率(EER)相较ResNet有20%的下降，并且在VoxCeleb, SITW等多种不同录制环境和识别任务中都有稳定的性能提升，证明了该方法的高效性和鲁棒性。改进后的全连接模块结构能更充分利用训练信息，在数据充足和任务复杂时性能提升明显。具体代码可以在https://github.com/czg0326/Res2Net-Speaker-Recognition获得。

关键词：说话人识别；多尺度特征；鲁棒性；高效性

中图分类号：TN912.34 **文献标识码**：A **文章编号**：1009-5896(2021)11-3266-06

DOI: 10.11999/JEIT200917

A Multiscale Feature Extraction Method for Text-independent Speaker Recognition

CHEN Zhigao[①②] LI Peng[③] XIAO Runqiu[①②] LI Ta[①] WANG Wenchao[①]

[①](*Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100190, *China*)

[②](*University of Chinese Academy of Sciences, Beijing* 100049, *China*)

[③](*National Computer Network Emergency Response Technical Team/ Coordination Center of China, Beijing* 100029, *China*)

Abstract: Recently in speaker recognition tasks, consistent performance gains have been continually achieved by various Convolutional Neural Networks (CNNs), which have shown increasingly stronger multiscale representation abilities. However, most existing methods enhance their strength with more layers and deeper structures. In this paper, a unique multiscale backbone architecture, Res2Net, is introduced for speaker recognition tasks, and its blocks are modified for assessment. This architecture works at a more granular level than most layer-wise networks. It improves the system by combining many equivalent receptive fields, resulting in a combination of different feature scales. The experiments results demonstrate that this architecture steadily achieves a 20% improvement on the Equal Error Rate (EER) over the baseline without additional computational burden. Its effectiveness and robustness are also verified in different environments and tasks, such as VoxCeleb and Speakers In The Wild (SITW). The modified full-connection block can make sure a more sufficient use of information and improves the performance obviously in more complex tasks. The code is available at https://github.com/czg0326/Res2Net-Speaker-Recognition.

Key words: Speaker recognition; Multiscale features; Robustness; Efficiency

收稿日期：2020-10-26；改回日期：2021-03-13；网络出版：2021-03-25
*通信作者： 王文超 wangwenchao@hccl.ioa.ac.cn
基金项目：国家自然科学基金(11590772, 11590774, 11590770)
Foundation Items: The National Natural Science Foundation of China (11590772, 11590774, 11590770)

1 引言

过去10年里，说话人识别技术得益于深度学习方法的帮助而发展迅速。深度神经网络(Deep Neural Networks, DNN)通过强大的抽象表达能力，在

多种模式识别任务中都表现出十分可观的性能。同时，在说话人任务中，基于深度神经网络的技术方法也在一定程度上逐渐超过了传统的因子分析框架[1]。

受深度神经网络在自动语音识别(Automatic Speech Recognition, ASR)领域的成功启发，Variani等人[2]提出了利用深度神经网络提取说话人特征的方法，通过建立一个帧级别的神经网络说话人分类器，区分每一帧输入语音的说话人身份。网络训练收敛之后，最后一层隐含层的输出用作每一帧输入语音的说话人特征表达，称为d-vector。此后，为了获得输入语音更多的上下文信息，时间延迟神经网络(Time-Delay Neural Networks, TDNN)也被引入说话人识别任务，它把网络的输入从单一的帧级别提升到一个更大的时间尺度。

类似于d-vector的思路，Snyder等人[3]提出了x-vector说话人识别系统，期望在文本无关的短时语音任务中，取得更好的表现。其中最重要的改进是在网络结构中引入了一个统计层，计算前面网络层输出的统计信息，从而把特征表达从帧级别提升到句级别，积累更多的说话人信息。最终，x-vector系统在短时说话人识别任务中取得显著的性能提升，并成为后来多数说话人识别研究的基线系统。

近几年，卷积神经网络在计算机视觉和语音识别任务中取得了长足的进展，在说话人识别任务中的作用也更加显著。王文超等人[4]通过时间延迟神经网络和卷积神经网络的融合，同时关注语音的局部和全局信息，目的同样是在学习多尺度的特征表达。

早些时候的卷积神经网络框架，例如AlexNet和视觉几何组网络(the Visual Geometry Group Network, VGGNet)，在图像分类任务中效果显著。Nagrani等人[5]、Huang等人[6]、Yadav等人[7]把VGGNet引入说话人识别领域，并同样证明有效。在说话人识别任务中，VGGNet结构通过增加网络层数来增强表征说话人信息的能力，然而与其他领域一样，过多的网络层数可能会导致梯度消失的问题。

为了应对梯度消失，He等人[8]提出了一种深度残差神经网络(Residual neural Network, ResNet)。它由多个残差模块堆积而成，每个模块的输入和输出层之间建立一个捷径链接。通过恒等映射的引入，有效避免了网络的退化。此外，与Inception-Nets类似，ResNet也利用不同的卷积核大小来提升多尺度特征表达能力。综合上述优势，ResNet(例如ResNet-34, ResNet-50, ResNet-101)已经成为目前最通用的卷积神经网络结构之一。然而，它提升表达能力的主要方法仍依靠增多网络层的数目。

Gao等人[9]此前在计算机视觉任务中提出了一种多尺度特征提取结构，Res2Net，取得了可观的性能提升。然而其对信息的发掘可能不够充分，网络结构的删减也缺乏原理性解释。本文引入文本无关说话人识别任务中，验证其在语音任务的有效性和鲁棒性，并对该结构做进一步改进，最大限度利用训练数据，应对更复杂的识别任务。

第2节将介绍基于卷积神经网络的框架结构，尤其是本文的基线系统，ResNet框架。而新引入的Res2Net结构以及改进将在第3节详细描述。第4节是实验的具体设置和结果分析。最后的结论将在第5节阐述。

2　相关工作

2.1　卷积神经网络

卷积神经网络设计初衷是为从多尺度表达特征，以此为基础的网络结构已经在大量模式识别任务中表现出最优的性能[10]。AlexNet通过滤波器的堆叠获得可观的性能表现，但是它的网络层数有限，因此感受野相对较小。此后VGGNet通过更深层的网络结构来增大感受野，而且由于卷积核更小，VGGNet在性能超过AlexNet的同时，参数量也更少。然而，随着网络层数的增加，梯度消失成为难以避免的问题。

2.2　深度残差网络

为了应对梯度消失的问题，ResNet引入了残差块结构。与传统的单线结构不同的是，每一个残差块的输入和输出之间建立了捷径连接，直接把输入x传到输出作为初始结果，输出结果为$H(x)=F(x)+x$，当$F(x)=0$时，$H(x)=x$，也就是恒等映射。这使得网络加深的时候，准确率不会下降，而且学习目标由原本的完整输出$H(x)$变为$H(x)$和x的残差$F(x)$, $F(x)=0$也更容易学习。图1是经典的深度残差网络的残差块结构。

Cai等人[11]、Heo等人[12]和Chung等人[13]将ResNet结构引入了说话人识别任务，然而目前这些系

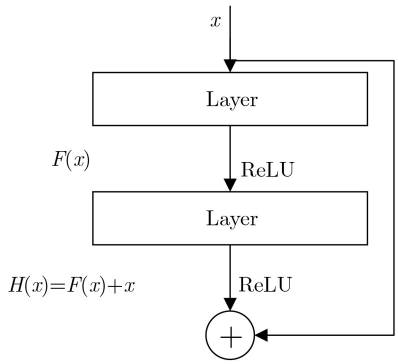

图 1 深度残差网络的残差块

统提升识别性能的主要方法还是在于增加网络层数。

3 模型结构

本文引入说话人识别任务的结构Res2Net，核心思想是寻求更多尺度的特征表达。相比于传统的增加网络深度(depth)和宽度(width)[14]，它用一种更细粒化的工作方式，更高效地获得更多的感受野。

图2是文献[9]的Res2Net模块示意图，本文中称为简化的Res2Net模块。如图2所示，Res2Net调整的是ResNet-50里的3×3模块。一共C个通道的3×3滤波器被切分成s个滤波器组，每组有w个通道($C=s\times w$)，特征图也被切分成s组，并由此引入了一个新的维度：s(scale)。与ResNet和ResNeXt网络结构相比，scale是继depth,width和基数(cardinality)[15]之后的第4个维度。在切分操作之后，每一组$x_i(i\in\{1,2,\cdots,s\})$都有$w$个通道的输入特征图，输出$y_i$可以表示为

$$y_i=\begin{cases}C_i(x_i), & i=1\\ C_i(x_i+y_{i-1}), & 1<i<s\\ x_i, & i=s\end{cases}\tag{1}$$

其中，小的3×3卷积操作记作C_i。在图2所示的简化Res2Net模块中，当一个C_i接收到前一个C_i的信息，相应的感受野会变大。而ResNet主体网络有若干个卷积层，上述切分操作层层作用，最终Res2Net网络的输出将获得多种感受野大小的组合，从而有效地从多尺度表达语音特征。

然而，简化的Res2Net模块删除了最后一个C_i，并且每一组输出只传递给相邻的下一组。文献[9]没有给出原理性的解释，可能是避免过拟合现象。本文改进的全连接的Res2Net模块结构，如图3所示。

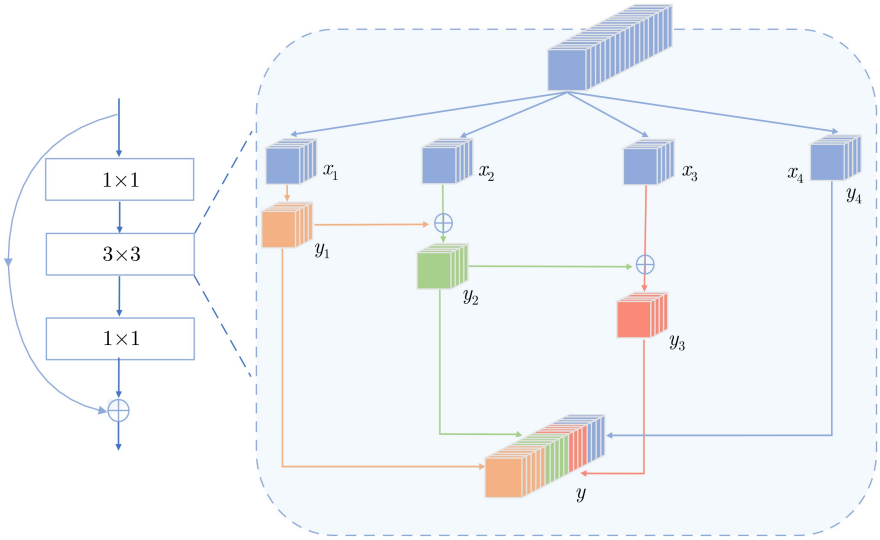

图 2 简化的Res2Net模块示意图

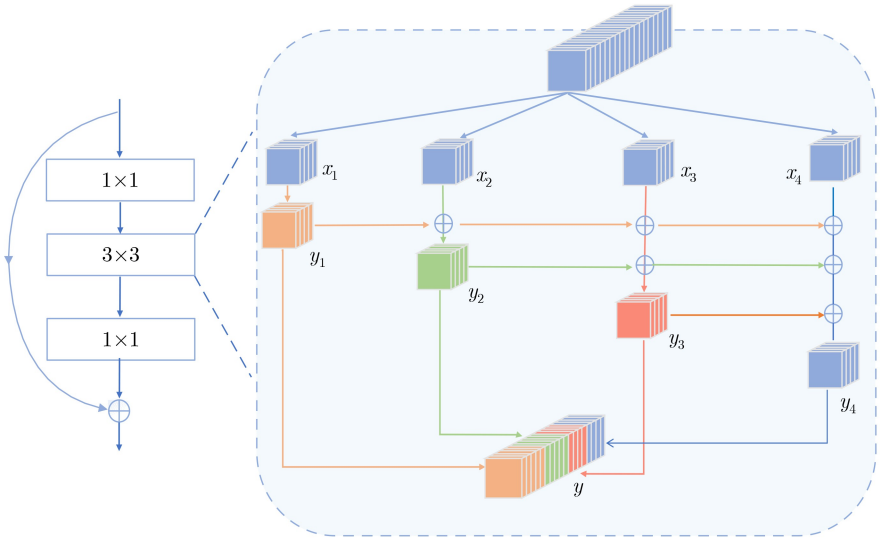

图 3 全连接的Res2Net模块示意图

公式相应地调整为

$$y_i = \begin{cases} C_i(x_i), & i = 1 \\ C_i(x_i + y_{i-1} + \cdots + y_1), & 1 < i \le s \end{cases} \tag{2}$$

为了更充分地发掘训练数据中的信息，本文改进的全连接的Res2Net模块结构与简化的Res2Net模块不同的是，在训练过程中，每一组滤波器会接收前面所有输出，以求获得更多的感受野尺寸组合。如图3所示，特征图在分组后，会先融合前面的滤波器输出，再由自身的滤波器提取特征，同时传送给后面的组以及最后的输出。最终，所有分组的特征图拼接回一起，传送给后面的1×1滤波器，融合之后进入下一个模块。

这样每一份输入特征，都有多种路径到达最后的输出特征。每一条可能的路径，都会有不同的信息融合。因为相比于图像分类任务中一些固定的物品种类，说话人识别面对的是相对更难的开集(open set)问题，期望训练数据有更强的覆盖和泛化性，以及关注到各种环境信息、噪声干扰等。因此，本文改进的全连接的Res2Net结构能更充分地利用语音信息，预期比简化的Res2Net模块有着更强大的多尺度特征表达能力，有望在更复杂的识别任务中发挥作用，第4节的实验将会对两种结构的模块进行性能比较。

4　实验

本节首先介绍实验用到的数据集，然后详细描述实验系统的设置，最后是结果的展示与分析。

4.1　实验数据集

本文实验的训练集用的是近些年说话人识别通用的VoxCeleb1数据集[5]和VoxCeleb2数据集[13]。其中VoxCeleb1训练集包含1211名说话人的148642条训练语句，VoxCeleb2则有5994名说话人的超过百万条训练语句。本文实验主体的测试集是VoxCeleb1的评估集，包含40名说话人的4874条语音，构成37720个测试对。

为了验证文章方法在不同录制环境等复杂任务的鲁棒性，实验还测试了在SITW(Speakers In The Wild,)数据集[16]上的性能。SITW是由斯坦福国际研究院(Stanford Research Institute international, SRI international)开发的现实环境录制的语音数据集，包含299位名人的语音。其中，本文用到的它的评估集包含180位名人的2883条语音，且有4种不同的测试场景，core-core, core-multi, assist-core以及assist-multi。Core代表注册或测试的语音中只包含单一说话人是语音，而assist和multi则分别代表注册和测试的语音中包含多于一个人的说话语音，干扰因素更多。

4.2　系统设置

为了验证网络结构对特征表达能力的提升，本文实验采用朴素的训练方法。基线系统是x-vector和ResNet-50网络，前端没有语音增强，后端没有复杂的处理和损失函数技巧。所有实验系统在PyTorch平台搭建，均采用批标准化和线性整流函数(ReLU)加速收敛。语音特征提取的是64维滤波器组特征(Filter Banks, FBank)，最终输出的说话人特征向量为512维。最终采用简单的余弦距离打分，计算等错误率(Equal Error Rate, EER)和最小检测代价函数(minimum Detection Cost Function, minDCF)来评估识别性能。EER是指检测的虚警率(False Alarm Probability, FAP)和漏检率(Miss Probability, MP)相等时的指标，检测代价函数也是在特定目标先验概率的情况下虚警和漏检的代价加权和。

4.3　实验结果

4.3.1　性能比较

主体实验的第1部分，测试不同的网络系统在VoxCeleb1上的识别性能，本文引入的Res2Net方法由于更丰富的感受野组合，预期会比基线系统有更强的特征表达能力。

如表1所示，简化的Res2Net-50系统(Res2Net-50-sim)表现出了最好的识别性能，并且比基线ResNet-50系统有相对12%的提升，符合此前的预期，而全连接的模块结构(Res2Net-50-full)比简化的模块结构性能略差。

实验接下来采用规模更大的VoxCeleb2训练集，再次测试上述系统的识别性能。

如表2所示，各网络系统在接收更丰富的训练

表 1　VoxCeleb1测试集各系统性能表现(训练集：VoxCeleb1)

系统	等错误率EER(%)	最小检测代价函数minDCF		
		P=0.1	P=0.01	P=0.001
x-vector	4.189	0.212	0.391	0.512
ResNet-50	3.955	0.212	0.404	0.483
Res2Net-50-sim	**3.484**	**0.194**	**0.370**	0.481
Res2Net-50-full	3.633	0.201	0.373	**0.477**

表 2　VoxCeleb1测试集各系统性能表现(训练集：VoxCeleb2)

系统	等错误率EER(%)	最小检测代价函数minDCF		
		P=0.1	P=0.01	P=0.001
x-vector	2.985	0.179	0.336	0.465
ResNet-50	2.243	0.158	0.299	0.391
Res2Net-50-sim	1.729	0.143	0.271	0.405
Res2Net-50-full	**1.403**	**0.136**	**0.259**	**0.364**

数据之后，识别性能都有大幅提升，Res2Net-50系统相比于ResNet-50基线，有超过20%的性能提高。而且此时全连接的模块结构，表现出了较明显的优势，各项指标相比于简化的模块结构都有10%左右的领先。

参照其他研究者在VoxCeleb1测试集上一些可以复现的性能，本文引入的Res2Net-50同样有可观的表现。Okabe等人通过引入注意力机制的池化方法(attentive pooling)达到3.85%的等错率，训练集改用规模更大的VoxCeleb2后，Zeinali等人[17]使用多达160层的ResNet以及加性角度间隔损失函数，得到了1.31%的等错率。而本文引入的Res2Net，仅用50层和余弦距离打分即可达到1.4%的等错率。具体性能对比见表3。

为了验证Res2Net的鲁棒性，本实验又在开放环境数据集SITW上测试各系统的性能。如表4所示，Res2Net-50系统保持了比ResNet-50基线相对约10% 的性能提升。X-vector系统也表现不俗，体现出在跨数据集情况下的鲁棒性。与表2结果类似，当有更充足的训练数据供学习，以及更复杂的识别任务时，全连接的模块结构更能发挥优势。尤其是VoxCeleb2训练时，4种测试场景相对简化的模块结构都有9%～15%的性能提升。

4.3.2 效率探究

首先值得一提的是，实验中Res2Net-50相较于ResNet-50有稳定的超过10%的性能提升，但模型大小和参数量仅仅是ResNet-50的1.06倍，而采用全连接模块结构的Res2Net-50-full也只是简化的Res2Net-50-sim的1.16倍，同等训练数据时几乎不增加计算负载。

识别性能的提升，得益于更多的感受野为Res2Net系统带来了更强大的多尺度特征提取能力。而在不增加网络深度的前提下，直接关系到感受野的就是width和scale两个维度。Scale表示特征图被切分的组数，width表示每一组的通道数，增大scale和width都可以起到增大感受野的作用。接下来的实验将选取小训练集性能较好的简化Res2Net作为框架，探究scale和width对识别性能的影响。

在表5和表6中，7w4s表示width为7, scale为4的Res2Net-50网络，16w表示width增大为16, 8s则表示scale设为8。实验结果可以看出，二者都可以降低识别的等错误率，只是width增大1倍多，性能提升十分有限，而scale设为8则使得性能有相对10%的提升。因此，本文引入Res2Net中新增的scale维度，相比传统ResNet的width维度，能更高效地增强网络的特征提取能力。

表 3 系统VoxCeleb测试集性能

	训练集	等错率(%)
Nagrani等人[5]	VoxCeleb1	7.80
Okabe等人[18]	VoxCeleb1	**3.85**
Heo等人[12]	VoxCeleb1	5.50
Chung等人[13]	VoxCeleb2	3.95
Heo等人[12]	VoxCeleb2	2.66
Zeinali等人[17]	VoxCeleb2	**1.31**
本文系统	VoxCeleb1	**3.266**
本文系统	VoxCeleb2	1.403

表 5 Res2Net-50调整width和scale在VoxCeleb性能表现

参数设置	等错误率EER(%)	最小检测代价函数minDCF		
		P=0.1	P=0.01	P=0.001
7w4s	3.484	0.194	0.370	0.481
16w4s	3.446	**0.186**	0.357	0.491
7w8s	**3.266**	0.188	**0.347**	**0.475**

表 6 Res2Net-50调整width和scale在SITW性能表现

系统	SITW测试集EER(%)			
	Core-core	Core-multi	Assist-core	Assist-multi
7w4s	6.483	8.520	8.306	9.740
16w4s	6.370	8.382	8.601	9.411
7w8s	**5.549**	**7.726**	**7.699**	**9.122**

表 4 SITW 4种测试条件下各系统性能表现

系统	训练集	SITW测试集EER(%)			
		Core-core	Core-multi	Assist-core	Assist-multi
x-vector	VoxCeleb1	6.698	8.661	8.476	9.920
ResNet-50		7.217	9.358	9.282	10.972
Res2Net-50-sim		**6.483**	**8.520**	8.306	9.740
Res2Net-50-full		6.603	8.575	**8.297**	**9.516**
Res2Net-50-sim	VoxCeleb2	3.258	4.765	4.613	5.706
Res2Net-50-full		**2.952**	**4.201**	**3.931**	**4.833**

5　结论

本文引入了一种简单、高效的Res2Net结构到说话人识别领域。相比于传统ResNet结构，它通过更加细粒化的工作方式，在参数量只增加0.06倍的情况下，获得了更多的感受野组合和更多尺度的特征表达，从而取得了超过20%的性能提升。在此基础上改进的全连接Res2Net结构确保信息利用更充分，在学习数据充足或复杂的识别任务中，又有相对10%左右的提升，并且在没有任何语音增强和后端处理的情况下，VoxCeleb1测试集上达到了1.4%的等错率。鲁棒性上，本方法在不同录制环境(SITW)的识别任务中都有稳定的超过10%的性能提升。此外，新的scale维度也被证明相比于传统的网络深度和宽度，在多尺度表达特征的能力上更加高效，值得引入其他神经网络框架进行尝试。

参 考 文 献

[1] 郭武, 戴礼荣, 王仁华. 采用因子分析和支持向量机的说话人确认系统[J]. 电子与信息学报, 2009, 31(2): 302–305. doi: 10.3724/SP.J.1146.2007.01289.
GUO Wu, DAI Lirong, and WANG Renhua. Speaker verification based on factor analysis and SVM[J]. *Journal of Electronics & Information Technology*, 2009, 31(2): 302–305. doi: 10.3724/SP.J.1146.2007.01289.

[2] VARIANI E, LEI Xin, MCDERMOTT E, *et al*. Deep neural networks for small footprint text-dependent speaker verification[C]. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014: 4052–4056.

[3] SNYDER D, GARCIA-ROMERO D, POVEY D, *et al*. Deep neural network embeddings for text-independent speaker verification[C]. The Interspeech 2017, Stockholm, Sweden, 2017: 999–1003.

[4] 王文超, 黎塔. 基于多时间尺度的深层说话人特征提取研究[J]. 网络新媒体技术, 2019, 8(5): 21–26.
WANG Wenchao and LI Ta. Research on deep speaker embeddings extraction based on multiple temporal scales[J]. *Journal of Network New Media*, 2019, 8(5): 21–26.

[5] NAGRANI A, CHUNG J S, and ZISSERMAN A. Voxceleb: A large-scale speaker identification dataset[EB/OL]. https://arxiv.org/abs/1706.08612, 2017.

[6] HUANG Zili, WANG Shuai, and YU Kai. Angular softmax for short-duration text-independent speaker verification[C]. The Interspeech 2018, Hyderabad, India, 2018: 3623–3627.

[7] YADAV S and RAI A. Learning discriminative features for speaker identification and verification[C]. The Interspeech 2018, Hyderabad, India, 2018: 2237–2241.

[8] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[9] GAO Shanghua, CHENG Mingming, ZHAO Kai, *et al*. Res2net: A new multi-scale backbone architecture[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 43(2): 652–662.

[10] 柳长源, 王琪, 毕晓君. 基于多通道多尺度卷积神经网络的单幅图像去雨方法[J]. 电子与信息学报, 2020, 42(9): 2285–2292. doi: 10.11999/JEIT190755.
LIU Changyuan, WANG Qi, and BI Xiaojun. Research on rain removal method for single image based on multi-channel and multi-scale CNN[J]. *Journal of Electronics & Information Technology*, 2020, 42(9): 2285–2292. doi: 10.11999/JEIT190755.

[11] CAI Weicheng, CHEN Jinkun, and LI Ming. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[EB/OL]. https://arxiv.org/abs/1804.05160, 2018.

[12] HEO H S, JUNG J W, YANG I H, *et al*. End-to-end losses based on speaker basis vectors and all-speaker hard negative mining for speaker verification[EB/OL]. https://arxiv.org/abs/1902.02455, 2019.

[13] CHUNG J S, NAGRANI A, and ZISSERMAN A. Voxceleb2: Deep speaker recognition[EB/OL]. https://arxiv.org/abs/1806.05622, 2018.

[14] ZAGORUYKO S and KOMODAKIS N. Wide residual networks[EB/OL]. https://arxiv.org/abs/1605.07146, 2016.

[15] XIE Saining, GIRSHICK R, DOLLÁR P, *et al*. Aggregated residual transformations for deep neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1492–1500.

[16] MCLAREN M, FERRER L, CASTAN D, *et al*. The speakers in the wild (SITW) speaker recognition database[C]. The Interspeech 2016, San Francisco, USA, 2016: 818–822.

[17] ZEINALI H, WANG Shuai, SILNOVA A, *et al*. BUT system description to VoxCeleb speaker recognition challenge 2019[EB/OL]. https://arxiv.org/abs/1910.12592, 2019.

[18] OKABE K, KOSHINAKA T, and SHINODA K. Attentive statistics pooling for deep speaker embedding[EB/OL]. https://arxiv.org/abs/1803.10963, 2018.

陈志高：男，1994年生，博士生，研究方向为说话人识别、语音信号处理、语种识别等.

李　鹏：男，1983年生，高级工程师，研究方向为网络与信息安全等.

肖润秋：男，1995年生，博士生，研究方向为鲁棒说话人识别、语音信号处理等.

黎　塔：男，1983年生，研究员，研究方向为语音信号处理、大词汇自然口语语音识别、关键词识别等.

王文超：男，1991年生，助理研究员，研究方向为语音信号处理、说话人识别、语种识别等.

责任编辑：余　蓉