DOI:10.16136/j.joel.2022.06.0632

# 基于CRNN混合神经网络的多语种识别

王 瑶,龙 华*,邵玉斌,杜庆治,王延凯

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

**摘要**:在语种识别过程中,为提取语音信号中的空间特征以及时序特征,从而达到提高多语种识别准确率的目的,提出了一种利用卷积循环神经网络(convolutional recurrent neural network,CRNN)混合神经网络的多语种识别模型。该模型首先提取语音信号的声学特征;然后将特征输入到卷积神经网络(convolutional neural network,CNN)提取低维度的空间特征;再通过空间金字塔池化层(spatial pyramid pooling layer,SPP layer)对空间特征进行规整,得到固定长度的一维特征;最后将其输入到循环神经网络(recurrenrt neural network,CNN)来判别语种信息。为验证模型的鲁棒性,实验分别在3个数据集上进行,结果表明:相比于传统的CNN和RNN,CRNN混合神经网络对不同数据集的语种识别准确率均有提高,其中在8语种数据集中时长为5 s的语音上最为明显,分别提高了5.3%和6.1%。

**关键词**:语种识别;卷积循环神经网络混合神经网络;卷积神经网络;循环神经网络

**中图分类号**:TP391 **文献标识码**:A **文章编号**:1005-0086(2022)06-0620-09

# Multilingual recognition based on CRNN hybrid neural network

WANG Yao, LONG Hua*, SHAO Yubin, DU Qingzhi, WANG Yankai

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

**Abstract**: In the process of language recognition, a multilingual recognition model using convolutional recurrent neural network (CRNN) hybrid neural network is proposed to extract spatial features and temporal features of speech signals and improve the accuracy of multilingual recognition. The model firstly extracts the acoustic features of speech signals. Then the features are input into the convolutional neural network (CNN) to extract the low-dimensional spatial features. Next the spatial features are structured through the spatial pyramid pooling layer (SPP layer), and the fixed length one-dimensional features are obtained. Finally, it is input into the recurrent neural network (RNN) to identify the language information. To test and verify the robustness of the model, experiments are conducted on three data sets, the results show that compared with the conventional convolution neural network and recurrent neural network, CRNN hybrid neural network language recognition accuracy of different data sets are increased, the data sets in eight languages with time about 5 s have the voice of the most obvious, which are increased by 5.3% and 6.1% respectively.

**Key words**: language recognition; convolutional recurrent neural network (CRNN) hybrid neural network; convolutional neural network (CNN); recurrent neural network (RNN)

## 1 引 言

语种识别(language identification,LID)是指计算机根据不同语种之间的差异来判别语音样本中所用语言的种类。传统的语种识别模型有高斯混合-通用背景模型(Gaussian mixture model-universal background model,GMM-UBM)[1]、高斯超向量-支持向量机(Gaussian super vector-support vector machines,GSV-SVM)[2]和基于全差异(total variability,TV)空间的i-vector系统[3]等。近些年来,长短时记忆网络(long short-term memory network,LSTM)[4]和卷积神经网络(convolutional

* E-mail:1670931890@qq.com

收稿日期:2021-09-06 修订日期:2021-09-28

基金项目:国家自然科学基金(61761025)资助项目

neural network,CNN)[5]等深度神经网络也在语种识别模型中得到了广泛的应用。文献[6]使用LSTM联合训练多个语种识别任务,将这些任务的隐藏信息进行共享来提高语种识别的准确率,从而提出基于多任务学习的语种识别系统。文献[7]结合注意力机制和长短时记忆递归神经网络(long short term memory-recurrent neural network,LSTM-RNN)[8]搭建了端到端的语种识别模型,得到了不错的语种识别准确率。文献[9]首先融合语音信号的梅尔频率倒谱系数(mel-frequency cepstral coefficients,MFCC)[10]和位移差分倒谱系数(shifted delta cepstral,SDC)[11],然后将融合后的特征转换为图像,最后将图像输入到 CNN 网络中进行语种识别,同样也取得了较好的语种识别准确率。相比于传统模型,基于神经网络的系统更加适应语种识别任务,其不仅能够提取语音信号的特征,还可以提供更高的语种识别准确率[12]。

由于语音是一种在时间序列上的信号,在语音序列上存在很强的时序联系,而在神经网络中,循环神经网络(recurrent neural network,RNN)[13]可以对时间序列进行建模。所以基于 RNN 的语音信号处理可以增强语音信号上下文之间的关联性,同时保留一些重要信息,但是 RNN 网络还存两个缺点:1) 当输入序列较长时,由于时间的迭代乘法,训练速度可能非常缓慢[14];2) 训练过程中可能会出现梯度消失和梯度爆炸的问题[14]。相比于 RNN 网络,CNN 网络的局部感受野和权值共享特性在减少计算复杂度的同时,还可以将高维的特征向量变换为低维度的特征向量,并且,ABDEL-HAMID 等[15]认为将 CNN 网络用于语音识别有 3 个重要的优势:1) 局部感受野可增强对非白噪声的鲁棒性;2) 权值共享可以进一步增强模型的鲁棒性;3) 池化操作可以抵抗频带带来的扰动。

本文中结合 CNN 对空间特征的提取能力和 RNN 对时序特征的提取能力,提出了基于卷积循环神经网络(convolutional recurrent neural network,CRNN)混合神经网络的多语种识别模型。该模型由CNN网络、空间金字塔池化层(spatial pyramid pooling layer,SPP layer)[16]和 RNN 网络组成。CRNN 混合神经网络最早被应用于文本识别,在文献[17]中首先通过 CNN 中的 VGG 网络对图片进行特征提取,再采用 RNN 对特征序列进行预测,最后通过连接时序性分类方法(connectionist temporal classification,CTC)得到最终结果,而随着网络层数的增加,VGG 网络可能会出现梯度消失和梯度爆炸等问题,从而影响整个系统的性能。为了减少梯度消失和梯度爆炸带来的影响,本文中 CNN 网络采用深度残差网络 Resnet18,长短时记忆网络选用双向长短时记忆网络(bidirectional long short-term memory network,Bi-LSTM)。为了使得模型可以对不同时长语音进行语种识别,在 CNN 网络后端引入了 SPP 层,将 CNN 网络输出的低维度空间特征转换为固定长度的一维特征,同时为了更好地提取语音信号的时序特征,将 SPP 层输出的一维特征输入到 RNN 网络中,最终得到语音信号的语种。

## 2　特征提取

在语种识别中,通常提取语音信号的 MFCC、对数 Mel 尺度滤波器组能量(log mel-scale filter bank energies,Fbank)[18]以及语谱图[19]作为语种特征。因此本文选用这 3 种特征作为 CRNN 混合神经网络的输入。相比于 MFCC 的提取过程,Fbank 减少了一步离散余弦变换(discrete cosine transform,DCT),因此具有更多的原始信息。在特征提取的过程中,各个特征的帧长和帧移均为 512 和 256,其中 Fbank 提取 40 维特征,MFCC 提取 20 维特征,语谱图横轴代表时域信息,纵轴代表频域信息。

## 3　CRNN 语种识别模型

CRNN 混合神经网络前端为 CNN 网络中的深度残差网络 Resnet18,并且为了满足混合神经网络处理可变时长语音的要求,在 Resnet18 网络后端引入 SPP 层,然后将 Resnet18 网络输出的固定维度的空间特征输入到 Bi-LSTM 网络中,得到语音信号的时序特征。最后通过 softmax 分类函数来判别语音信号的语种。整体结构图如图 1 所示。

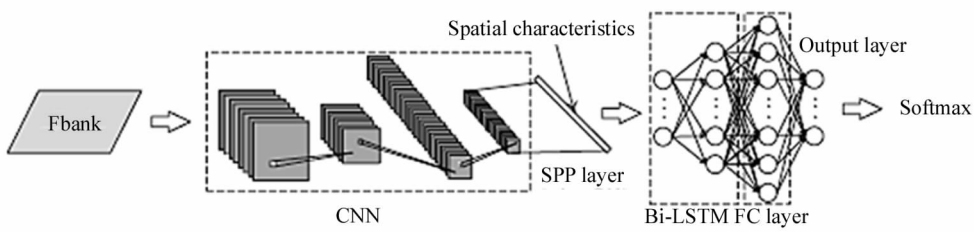


**图 1　CRNN 语种识别模型**

**Fig. 1　CRNN language recognition model**

### 3.1 CNN 网络

在文献[17]中，CRNN 网络中的 CNN 网络采用的是 VGG 网络。在 VGG 网络内部使用多个的卷积核代替其他大尺度的卷积核，其优点在于，保证相同感知野的条件下，不仅提升了网络的深度，在一定程度上也提升了神经网络的效果。然而 VGG 网络拥有 3 个全连接层，将使用大量的参数，因此它会消耗大量的计算资源。同时随着网络层数的增加，系统可能出现梯度消失和梯度爆炸的问题，也会影响神经网络的性能。

相比于 VGG 网络，本文中使用的深度残差网络 Resnet18 引入了残差单元[20]，很好地解决了深度神经网络中出现的梯度消失和梯度爆炸问题，残差单元可以表示为：

$$y = F(x, \boldsymbol{W}_i) + x, \tag{1}$$

$$F(x) = \boldsymbol{W}_2\sigma(\boldsymbol{W}_1 x), \tag{2}$$

式中，$x$ 和 $y$ 分别表示所在网络层的输入和输出结果，$F(x, \boldsymbol{W}_i)$表示要学习的残差映射，$F(x)$代表残差函数。$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 代表图 2 中第 1 个网络层和第 2 个网络层的权重向量，$\sigma$ 代表 relu 激活函数。最后残差单元的输出为 $\sigma(F(x)+x)$。

当残差函数 $F(x)=0$ 时，此时堆积层做了恒等映射，网络的性能不会随着网络层数的增加而下降，事实上残差函数不会为 0，因此堆积层在输入特征基础上还可以学习到新的特征，从而拥有更好的性能。如图 2 所示为残差单元结构示意图。

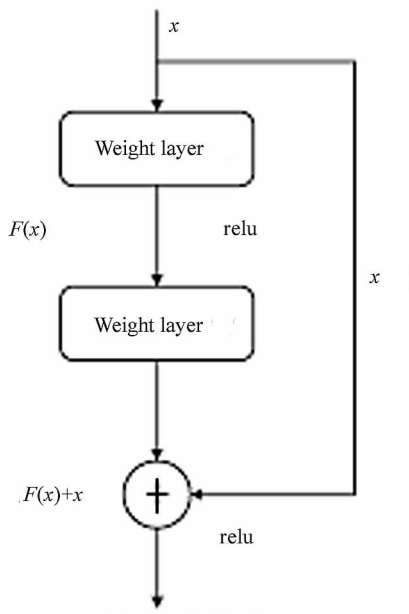


**图 2 残差单元**

**Fig. 2 Residual units**

Resnet18 网络被广泛应用到了计算机视觉和图像识别中。其简化模型如图 3 所示。

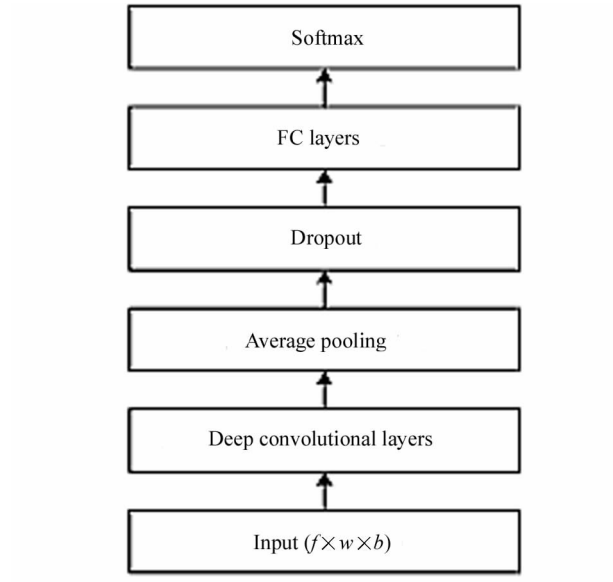


**图 3 Resnet18 网络模型**

**Fig. 3 Resnet18 network model**

图 3 中($f\times w\times b$)代表输入特征的维度，语音信号经过特征提取，得到帧数为 $f$，维数为 $w$ 的声学特征，$b$ 代表输入特征的数量。

由于 Resnet18 网络后端的全连接层需要固定长度的输入，而在语种识别中，语音信号的时长不是固定的，因此在 Resnet18 网络后端引入空间金字塔池化层，将不同时长的语音池化成相同维度。引入 SPP 层后的网络结构图如图 4 所示。

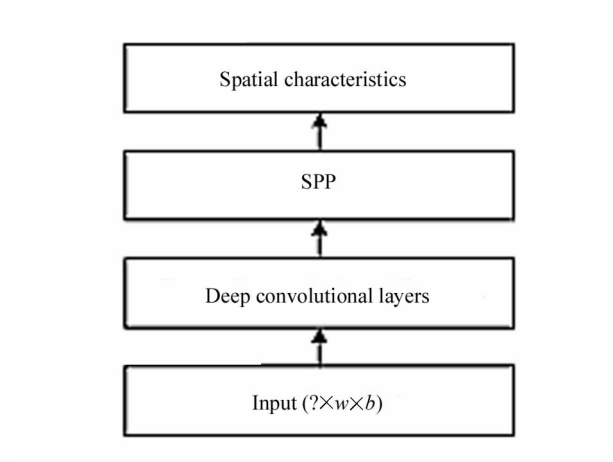


**图 4 引入 SPP 层的 Resnet18 网络**

**Fig. 4 Resnet18 network introducing the pooling layer of the spatial pyramid**

引入的 SPP 层替换了 Resnet18 网络中原有的平均池化层，并去掉全连接层，SPP 层的输出直接作

为 Bi-LSTM 网络的输入。由于输入语音的时长可变,在提取特征的过程中,分帧得到的帧数是不确定的,因此神经网络的输入特征维度为(? ×$w$×$b$),?代表不确定的帧数。

### 3.2　SPP 层

在传统的卷积层输出中,滑动窗口的大小是固定不变的,因此滑动窗口的数量由输入特征的维度决定,当输入特征的维度不等时,由于所需要的滑动窗口数量不等,输出的特征维度也不等。而在 SPP 层中,滑动窗口的大小与输入特征的大小成比例,滑动窗口的数量固定不变,即使输入特征的维度不等,也可以保证输出的特征维度相等。其结构如图5所示。

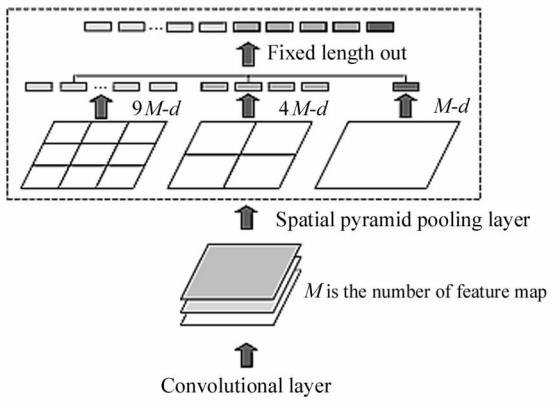


**图 5　SPP 层结构图**

**Fig. 5　Structure diagram of SPP layer**

从图中可以看出,卷积层的输出可以根据网络需求被划分为 1×1,2×2 和 3×3 等多个小方块,然后对这些小方块进行平均池化和最大池化,最后将池化得到的数据拼接起来,形成一个固定长度的序列。设卷积层输出特征图的维度为($c$,$h_{in}$,$w_{in}$),池化数量为($n$,$n$),其中 $c$,$h_{\text{in}}$,$w_{\text{in}}$分别代表通道数、高度和宽度。经过空间金字塔池化的矩阵大小($h_{\text{new}}$,$w_{\text{new}}$)可以表示为:

$$\begin{cases} S_h = \left\lceil \dfrac{h_{\text{in}}}{n} \right\rceil \\ K_h = \left\lceil \dfrac{h_{\text{in}}}{n} \right\rceil \\ p_h = \left\lfloor \dfrac{k_h \times n - h_{\text{in}} + 1}{2} \right\rfloor \\ h_{\text{new}} = \left\lfloor \dfrac{(2 \times p_h + h_{\text{in}} - k_h)}{s_h} + 1 \right\rfloor \end{cases}, \tag{3}$$

$$\begin{cases} S_w = \left\lceil \dfrac{w_{\text{in}}}{n} \right\rceil \\ K_w = \left\lceil \dfrac{w_{\text{in}}}{n} \right\rceil \\ p_w = \left\lfloor \dfrac{k_w \times n - h_{\text{in}} + 1}{2} \right\rfloor \\ w_{\text{new}} = \left\lfloor \dfrac{(2 \times p_w + h_{\text{in}} - k_w)}{s_w} + 1 \right\rfloor \end{cases}, \tag{4}$$

式中,$k_h$,$k_w$ 分别代表滑动窗口的高和宽,$s_h$,$s_w$ 分别代表高和宽方向上的步长,$p_h$,$p_w$ 分别代表高和宽方向上填充数量。

SPP 层最早在文献[16]中被应用于图像识别,为保留图像信息的完整性,在池化过程中采用平均池化,本文考虑到语音产生过程中句与句之间会出现停顿以及不同语种单个字发音的不同,会造成声学特征中能量谱变化较大的情况,因此在 SPP 过程中采用平均池化和最大池化相结合的方式,来更好地提取能量谱信息。

### 3.3　Bi-LSTM 网络

在 CRNN 语种识别模型的后端,为了更好地提取语音信号中的时序特征,通过 Bi-LSTM 网络来对 CNN 网络输出的空间特征进行建模,如图 1 所示。Bi-LSTM 网络具体包括,输入层、LSTM 层和一个全连接层,首先将带有标签的空间特征依次输入到 Pytorch 封装的 LSTM 层和 1 个全连接层,将输出结果映射到 softmax 函数中,计算语音信号被判别为每个语种的概率:

$$p_J = \frac{c_I}{\sum_{g=1}^{G} c_g}, \tag{5}$$

式中,$p_J$ 表示判别为第 $J$ 类语种的概率,$G$ 代表语种个数,$c_I$、$c_g$ 分别代表第 $I$ 个节点和第 $g$ 个节点的输出值。

## 4　实验分析

### 4.1　实验设置

本文在 3 个数据集上进行了实验。数据集 1 从 LibriVox 网站下载,共 5 种语言,分别为英语、法语、德语、西班牙语和意大利语。采样率为 16 kHz,精度为 16 bit,声道为单声道。每种语言包含 3 种时长,分别为 3 s、5 s 和 10 s,各时长语音的条数相等,均为 2 500 条,其中训练集 2000 条,测试集 500 条;数据集

2从国际广播电台中录制，共8种语言，分别是普通话、缅甸语、越南语、柬埔寨语、老挝语、韩语、藏语、维吾尔语。采样率为16 kHz，精度为16 bit，声道为单声道。每种语言包含3种时长，分别为3 s、5 s和10 s，各时长语音的条数相等，均为1 250条，其中训练集1 000条，测试集250条；数据集3来自科大讯飞AI开发者大赛提供的10种方言数据，每种方言包含40位说话者6 h的语音数据，语音采样率为16 kHz，精度为16 bit，声道为单声道。训练集每个6 000条，包含30位说话人，其中男性和女性各15人，每个说话人200条；测试集中每个语种500条，包含10个说话人。训练集和测试集按语音的时长分为≤3 s的短时语音和>3 s的长时语音。将各个数据集对应的语种打上标签，如表1所示。

本文中语种识别的测试标准采用识别准确率(Recognition Accuracy，$A_R$)来评价。

$$A_R=\frac{\sum_{g=1}^{G}a_g}{\sum_{g=1}^{G}b_g},\tag{6}$$

式中，$G$代表语种个数，$a_g$是第$g$个语种识别正确的语音个数，$b_g$代表第$g$个语种总的语音数，$A_g$代表准确率。

在实验过程中，主要分为4个部分：1) 为了探究不同语种特征以及网络参数对语种识别准确率的影响，首先以数据集1和数据集2中时长为5 s的语音进行实验，分别提取MFCC、Fbank以及语谱图特征作为神经网络的输入，来确定当语种识别准确率最高时，CNN网络输出空间特征的维度、Bi-LSTM网络隐藏层的层数和每层网络神经元的个数；2) 探究在空间金字塔池化中采用不同的池化方式对语种识别准确率的影响；3) 对比最优网络参数下，CNN网络、Bi-LSTM网络、VGG+Bi-LSTM混合神经网络以及Resnet18+Bi-LSTM混合神经网络的语种识别准确率；4) 通过在数据集3上进行实验，对比Resnet18+Bi-LSTM混合神经网络语种识别模型与文献[6]中多任务语种识别模型的性能。

**表1　实验数据集设置**

**Tab. 1　Setting of experimental data set**

| Data set 1 | | Data set 2 | | Data set 3 | |
|---|---|---|---|---|---|
| Languages | Label | Languages | Label | Languages | Label |
| English | 0 | Chinese | 0 | Minnan | 0 |
| French | 1 | Burmese | 1 | Kejia | 1 |
| German | 2 | Vietnamese | 2 | Shanghai | 2 |
| Spanish | 3 | Cambodian | 3 | Hefei | 3 |
| Italian | 4 | Lao | 4 | Shaanxi | 4 |
| | | Korean | 5 | Ningxia | 5 |
| | | Tibetan | 6 | Changsha | 6 |
| | | Uyghur | 7 | Hebei | 7 |
| | | | | Nanchang | 8 |
| | | | | Sichuan | 9 |

## 4.2　实验结果

CNN网络输出空间特征的维度由SPP层决定，实验中，分别将卷积层的输出按图5中的(1×1)、(1×1,2×2)、(1×1,2×2,3×3)和(1×1)、(2×2,3×3,4×4)进行划分，得到的空间特征维度可以表示1×$M$、(1×1+2×2)×$M$、(1×1+2×2+3×3)×$M$和(1×1+2×2+3×3+4×4)×$M$，其中$M$代表输出特征图的数量。在数据集1和数据集2的实验结果如图6所示。对比不同划分方式的效果，发现并不是CNN网络输出的维度越高，语种识别准确率越高。在数据集1和数据集2中，3种不同的特征都是在输出维度为(1×1+2×2+3×3)×$M$时，语种识别准确率最高，其中Fbank相比于MFCC和语谱图特征又具有更高的语种识别准确率，在数据集1和数据集2上分别达到了87.6%和80.4%，其次是输出维度为(1×1+2×2)×$M$时，Fbank在数据集1和数据集2的语种识别准确率分别在87.1%和80.2%，但考虑到输出维度越高，网络训练所需的时间和消耗的计算资源都会随之增大，因此为了提升语种识别模型的综合性能，CNN网络输出空间特征的维度设定为(1×1+2×2)×$M$。

在确定了空间特征的维度之后，再继续探究Bi-LSTM网络隐藏层的层数对语种识别准确率的影响。依次增加隐藏层的层数由1—6层，实验结果如图7所示。在一定程度上，增加隐藏层的层数可以

提高语种识别准确率,但层数与语种识别准确率的关系不成正比,随着网络层数的增加,神经网络会出现过拟合。当具有 4 层隐藏层时,Fbank 在数据集 1 和数据集 2 上的语种识别准确率最高,分别达到了 88.2%和 81.2%。

进一步探究隐藏层中神经元的个数对语种识别准确率的影响,依次增加神经元的个数从256—4 096,实验结果如图 8 所示。

从图 8 中可以看出,在一定范围内,随着神经元数目的增加,语种识别准确率也随之增加,但并不是神经元的个数越多,语种识别准确率就越高。当神经元个数为 1 024 时,Fbank 在数据集 1 和数据集 2 上的语种识别准确率都达到了最高,分别为 90.8%和 85.6%;但神经元个数超过 1 024 之后,语种识别准确率开始逐渐下降,主要原因是神经元数量过多,神经网络出现了过拟合。因此,并不能一味地通过增加隐藏层中神经元的个数来提高语种识别准确率,当神经元的数目过大,会导致网络训练变得非常缓慢,同时也会出现内存不足的问题。

综合上述 3 组实验,初步确认了 CRNN 的网络参数,同时选用 Fbank 作为网络的特征输入以达到更高的语种识别准确率。为了探究在空间金字塔池化中采用不同的池化方式对语种识别准确率的影响,本文将 SPP 层的池化方式按如下分配:1)(1×1+2×2)×$M$ 中的(1×1)×$M$ 部分采用平均池化,(2×2)×$M$ 部分采用最大池化;2) 全部采用最大池化;3) 全部采用平均池化,实验在数据集 3 上进行,结果如图 9 所示。

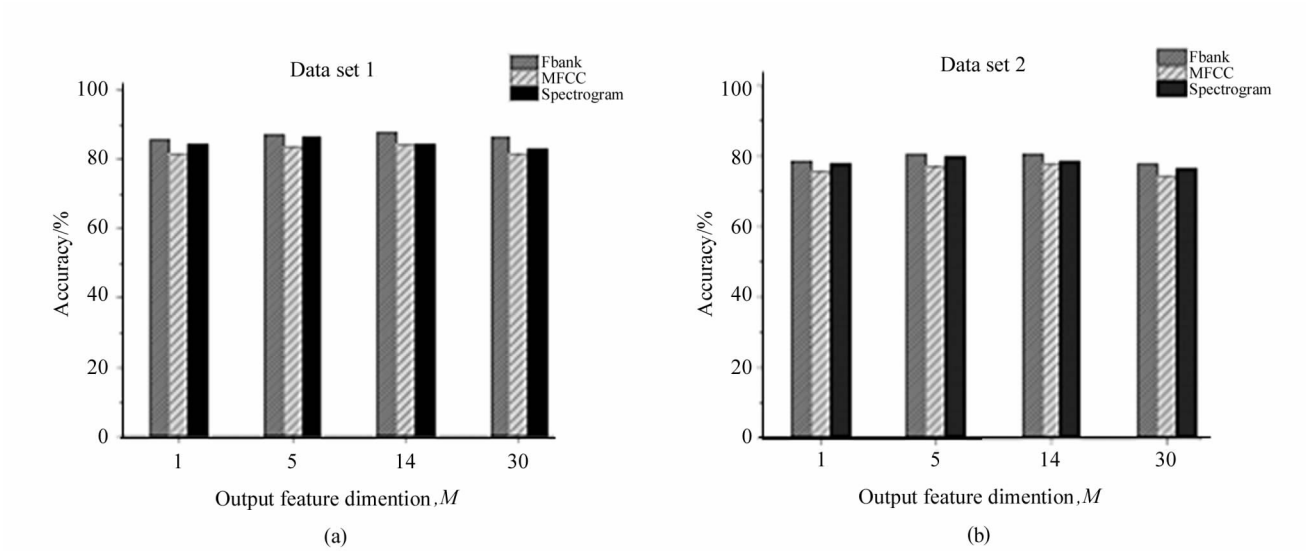


**图 6　空间特征维度对语种识别准确率的影响**

**Fig. 6　Influence of spatial feature dimension on language recognition accuracy**

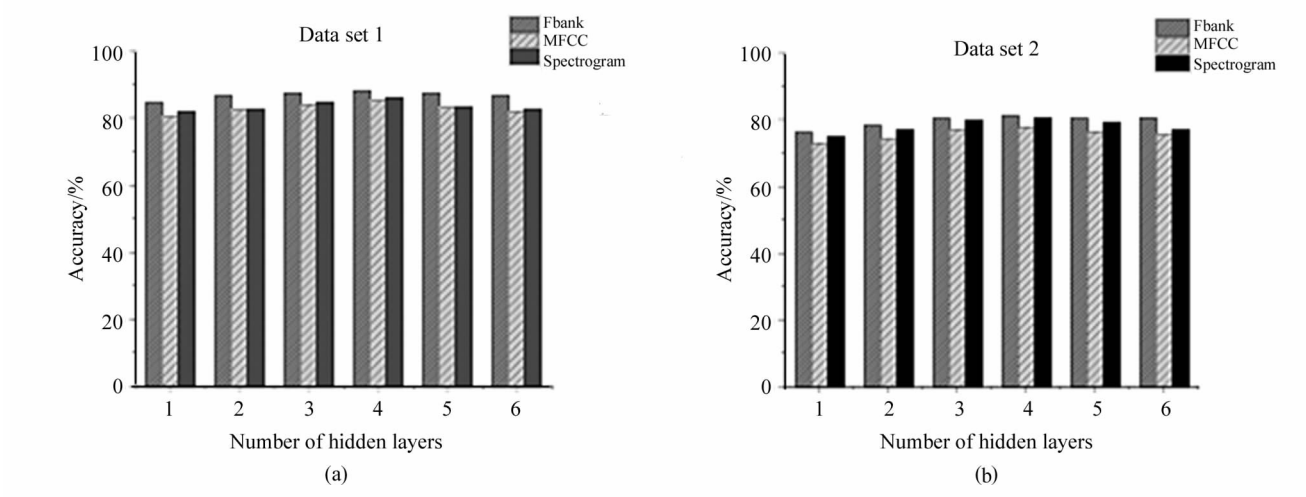


**图 7　隐藏层的层数对语种识别准确率的影响**

**Fig. 7　Influence of the number of hidden layers on the accuracy of language recognition**

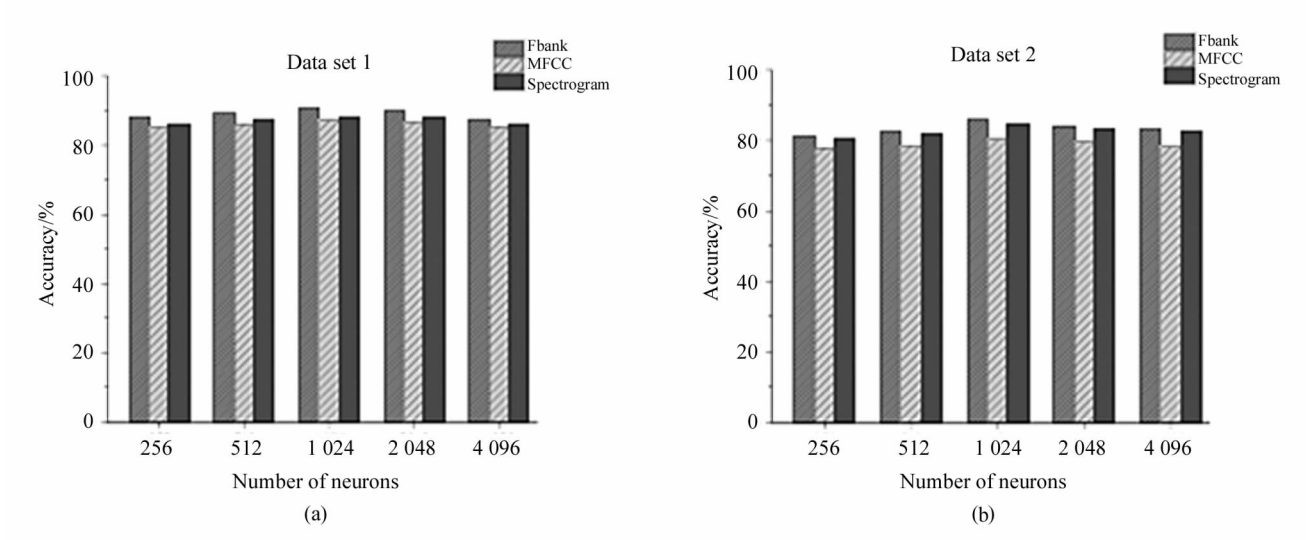


**图 8　隐藏层神经元的个数对语种识别准确率的影响**

**Fig. 8　The influence of the number of hidden layer neurons on the accuracy of language recognition**

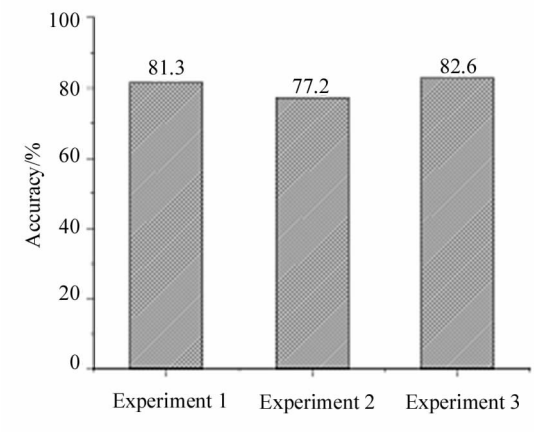


**图 9　SPP 层不同池化方式对语种识别准确率的影响**

**Fig. 9　Influence of different pooling modes on accuracy of language recognition in SPP layer**

从实验结果可以得出，采用平均池化与最大池化相结合的方式并没有在原有平均池化的基础上提升语种识别准确率。因此本文在后续的实验中，空间金字塔池化层依然与文献[16]保持一致，全部采用平均池化的方式。

本文为了进一步探究 Resnet18＋Bi-LSTM 混合神经网络相比于单独的 Resnet18 网络、Bi-LSTM 网络、文献[6]中的多任务语种识别模型以及文献[17]中的 VGG＋Bi-LSTM 网络对语种识别准确率的影响，实验分别在数据集 1、数据集 2 和数据集 3 上进行了实验，实验结果如表 2 所示。

从表 2 中可以看出，随着语音时长的增加，各个模型在不同数据集上的语种识别准确率也随之增加。相比于其他网络模型，Resnet18＋Bi-LSTM 混合神经网络在数据集 1、数据集 2 和数据集 3 上的语种识别率最高，其中在数据集2中时长为5s的语音

**表 2　不同网络模型的语种识别准确率**

**Tab. 2　Language recognition accuracy of different network models**

| Model | Data set 1 | | | Data set 2 | | | Data set 3 |
|---|---|---|---|---|---|---|---|
| | Test utterance length/s | | | | | | |
| | 3 s | 5 s | 10 s | 3 s | 5 s | 10 s | |
| Resnet18 | 79.5% | 88.1% | 93.2% | 76.4% | 80.3% | 86.3% | 78.1% |
| Bi-LSTM | 79.1% | 87.4% | 91.9% | 74.9% | 79.5% | 85.2% | 75.0% |
| Resnet18＋Bi-LSTM | **81.7%** | **90.8%** | **94.2%** | **79.1%** | **85.6%** | **89.8%** | **82.6%** |
| VGG＋Bi-LSTM | 78.5% | 87.9% | 82.3% | 77.3% | 81.0% | 87.1% | 77.5% |
| Multi-task | —— | —— | —— | —— | —— | —— | 80.2% |

上提升最大，相比于 CNN 网络、Bi-LSTM 网络和 VGG+Bi-LSTM 网络分别提升了 5.3%、6.1%和 4.6%；相比于文献[6]提出的多任务语种识别模型，在相同的数据集 3 上也提升了 2.4%。

## 5　结　论

本文提出的基于 CRNN 混合神经网络的语种识别模型相比于单独的 CNN 网络和 Bi-LSTM 网络具有更高的语种识别准确率，通过提取语音信号的 Fbank 作为语种特征，分别在 3 个数据集上进行了实验，其中在数据集 2 中时长为 5 s 的语音上语种识别准确率提升最为明显，分别提升了 5.3%和 6.1%，主要原因是 CRNN 混合神经网络可以更好地提取语音信号中的空间特征和时序特征。但是该网络在数据集较大的情况下训练会比较缓慢，因此，在后续的研究过程中还需要进一步改进。

## 参考文献：


[1] REYNOLDS D A，QUATIERI T F，DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Process，2000，10(1-3)：19-41.

[2] CAMPBELL W M，STURIM D E，REYNOLDS D A. Support vector machines using GMM supervectors for speakers verification[J]. IEEE Signal Processing Letters，2006，13(5)：308-311.

[3] DEHAK N，TORRES-CARRASQUILLO P A，REYNOLDS D A，et al. Language recognition via i-vectors and dimensionality reduction[C]//Proceedings of the 12th Annual Conference of the Interational Speech Communication Association，INTERSPEECH 2011，August 28-31，2011，Florence，Italy. Baixas：International Speech Communication Association，2011：857-860.

[4] CHEN H K，CHEN Y. Speaker recognition based on multimodal short-term memory network with depth gate[J]. Progress in Laser and Optoelectronics，2019，56(3)：031007.
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.

[5] LI C，MA X，JIANG B，et al. Deep Speaker：an end-to-end neural speaker embedding system[EB/OL]. (2017-05-05)[2021-09-06]. https://arxiv.org/abs/1705.02304.

[6] QIN C G，WANG H，REN J. Dialect language recognition based on multi-task learning[J]. Computer Research and Development，2019，56(12)：2632-2640.
秦晨光，王海，任杰. 基于多任务学习的方言语种识别[J]. 计算机研究与发展. 2019，56(12)：2632-2640.

[7] GENG W，WANG W F，ZHAO Y Y，et al. End-to-end language identification using attention-based recurrent neural networks[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association，September 8-12，2016，San Francisco，CA，USA. Baixas：International Speech and Communication Association，2016：2944-2948.

[8] GONZALEZ-DOMINGUEZ J，LOPEZ-MORENO I，Sak H，et al. Automatic language identification using long short-term memory recurrent neural networks[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association，September 14-18，2014，Singapore. Baixas：International Speech and Communication Association，2014：2155-2159.

[9] RUBEN Z，ALICIA L D，JAVIER G D，et al. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks[J]. Plos One，2016，11(1)：e0146917.

[10] LEU F Y，LIN G. An MFCC-based speaker identification system[C]//2017 IEEE 31st International Conference on Advanced Information Networking and Application (AINA)，March 27-29，2017，Taipei，Taiwan，China. New York：IEEE，2017：1055-1062.

[11] MIAO X X，ZHANG J，SUO H B，et al. Time extension method applied to short-term speech language recognition[J]. Journal of Tsinghua University (Natural Science Edition)，2018，58(3)：254-259.
苗晓晓，张健，索宏彬，等. 应用于短时语音语种识别的时长扩展方法[J]. 清华大学学报(自然科学版)，2018，58(3)：254-259.

[12] BARTZ C，HEROLD T，YANG H，et al. Language identification using deep convolutional recurrent neural networks[EB/OL]. (2017-08-16)[2021-09-06]. https://arxiv.org/abs/1708.04811.

[13] LEE J，TASHEV I. High-level feature representation using recurrent neural network for speech emotion recognition[C]//Proceedings of the 16th Annual Conference of International Speech Communication Association，INTERSPEECH 2015，September 6-10，Dresden，Germany. Baixas：International Speech Communication Association，2015：1572-1577.

[14] ZHANG Y, PEZESHKI M, BRAKEL P, et al. Towards end-to-end speech recognition with deep convolutional neural networks[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association, INTERSPEECH, 2016, September 8-12, San Francisco, CA, USA. Baixas: International Speech Communication Association, 2016: 410-414.

[15] ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545.

[16] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-16.

[17] SHI B, XIANG B, CONG Y. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(11): 2298-2304.

[18] YU H, TAN Z H, ZHANG Y M, et al. DNN filter bank cepstral coefficients for spoofing detection[J]. IEEE Access, 2017, 99(5): 4779-4787.

[19] WANG Y K, LONG H, SHAO Y B, et al. Based on combination of endpoint detection and dynamic range control language recognition [J/OL]. Laser and optoelectronics progress: 1-14 [2021-08-22]. http://kns.cnki.net/kcms/detail/31.1690.tn.20210816.1530.052.html.
王延凯, 龙华, 邵玉斌, 等. 基于联合端点检测和动态范围控制的语种识别[J/OL]. 激光与光电子学进展: 1-14 [2021-08-22]. http://kns.cnki.net/kcms/detail/31.1690.tn.20210816.1530.052.html.

[20] ALAEDDINE H, JIHENE M. Deep residual network in network[J]. Computational Intelligence and Neuroscience, 2021, 21(3): 1182-1191.

**作者简介：**

**龙 华** (1963—)，女，博士，教授，硕士研究生导师，云南省通信协会常任理事，主要研究为网络通信和信号处理.