DOI:10.16652/j.issn.1004-373x.2024.01.016
引用格式:谢金洪,魏霞.基于ResCNN-BiGRU的四川方言语音识别[J].现代电子技术,2024,47(1):89-93.

基于ResCNN-BiGRU的四川方言语音识别

谢金洪,魏　霞
(新疆大学 电气工程学院,新疆 乌鲁木齐　830017)

摘　要: 由于基于深度卷积神经网络的语音识别模型中缺乏对特定方言音素特征的提取能力,造成方言发音底层特征部分信息丢失,进而导致方言识别准确率不高、鲁棒性差等问题。针对上述问题,提出一种结合残差网络(RestNet)和双向门控循环网络(BiGRU)的模型,该模型以GFCC特征图为输入,同时在残差网络中设计多尺度卷积模块,通过不同大小的卷积核提取特征,然后使用双向门控循环网络捕捉序列数据中的长期依赖关系,最后采用连接时序分类算法进行标签软对齐,实现四川方言语音识别模型。在四川方言语料库上的实验结果表明,提出的模型识别性能优于现有基准模型。

关键词: 四川方言; 音素特征; 双向门控循环网络; 多尺度卷积; 连接时序分类; 标签软对齐
中图分类号: TN912.3-34　　**文献标识码:** A　　**文章编号:** 1004-373X(2024)01-0089-05

Sichuan dialect speech recognition based on ResCNN-BiGRU

XIE Jinhong, WEI Xia
(School of Electrical Engineering, Xinjiang University, Urumqi 830017, China)

Abstract: Due to the lack of extraction ability of phonemic features of specific dialects in the speech recognition model based on deep convolutional neural network (DCNN), part of the information of the underlying features of dialect pronunciation is lost, which in turn leads to problems such as low dialect recognition accuracy and poor robustness. Therefore, a model combining residual network (RestNet) and bidirectional gated recurrent (BiGRU) network is proposed. The GFCC (Gammatone frequency cepstrum coefficient) feature map is taken as the input. A multi-scale convolutional layer is designed in the residual network. The features are extracted by convolution kernels of different sizes. And then, the long-term dependence in the sequence data is captured by the BiGRU network. Finally, the connected time series classification algorithm is used for label soft alignment to realize a Sichuan dialect speech recognition model. Experimental results on the Sichuan dialect corpus show that the recognition performance of the proposed model is better than that of the existing benchmark models.

Keywords: Sichuan dialect; phonemic feature; BiGRU network; multi-scale convolution; connection time series classification; label soft alignment

0　引　言

随着语音识别技术的不断发展,越来越多的人开始关注方言语音识别技术的应用。汉语方言种类繁多,通常分为闽南语、粤语等七大方言[1],它们在发音、语调、语法等方面都与标准普通话有很大的差异。目前,由于大规模的方言语料库制作成本非常高,导致低资源方言语音识别技术研究进展缓慢,正面临着逐渐消失的困境。因此,研究方言语音识别对方言保护、语音信号处理以及自然语言处理等领域具有重要现实意义。

传统的方言语音识别方法从带口音的普通话识别[2-4]方法发展而来,这些方法主要采用概率统计或距离度量的方法进行建模,再利用基因周期、梅尔倒谱系数[5](Mel Frequency Cepstrum Coefficient, MFCC)、线性预测编码[6](Linear Predictive Coding, LPC)、线性预测倒谱系数[7](Linear Predictive Cepstral Coefficient, LPCC)等技术进行语音识别。其中,MFCC根据人耳听觉特性将语音信号建模为非线性时变系统的输出,可以获得较为精确的语音参数估计,但同时也造成部分高频信息丢失。文献[8-14]针对传统MFCC的缺点进行了改进。例如,文献[8]使用Gammatone滤波器组替代传统的Mel滤波器组,在TIMIT数据集上的测试准确率达到了90%;文献[11]使用多正弦窗函数进行频谱估计,减少了频谱泄漏,可以提取到较低方差的语音特征。虽然这些改进特征在一定程度上提高了识别率,但计算成本较大,降低了模型训练速度。因此,一些学者尝试将语音特征当作

收稿日期:2023-06-30　　修回日期:2023-07-20

一张图像进行处理，采用深度学习方法自动抽取隐含的语义特征，显著提升了方言语音识别性能。例如：文献[15]中使用神经网络训练的方法，以CNN为特征提取器，提取到优于MFCC的语音特征，将该特征作为ResNet-BLSTM模型的输入，在Aishell-1语音数据集上有较好的识别效果；文献[16]以Fbank为语音特征，设计了基于深度前馈序列记忆网络与链接时序分类相结合的海南方言语音识别模型，提升了语音识别性能和训练速度；文献[17]将注意力机制引入到CNN网络中对声学模型进行建模，设计了大同方言语音的翻译模型。这些深度语音识别模型在一定程度上提升了模型性能，但却忽略了方言本身特定音素的重要性，导致复杂噪声环境下的方言语音识别性能较差。

针对以上问题，本文提出了一种基于残差卷积网络（ResCNN）与双向门控网络（BiGRU）相结合的混合模型结构（ResCNN-BiGRU）。该模型在残差网络中设计多尺度卷积模块（Multi-scale Convolutional Fusion Network, MCFN），直接对语音信号的特征图进行不同尺度的特征提取，以获得更加鲁棒和准确的语音特征，这样大大增强了CNN的表达能力；其次，在卷积网络后接BiGRU，通过该网络学习序列数据中的长时依赖关系，从而提高模型识别性能。

1 语音特征提取

语音特征作为语音识别系统的重要组成部分，其精确性和稳定性对系统性能的优劣有较大影响。因此，提取高质量的语音特征是语音识别的关键一步。人耳生理学研究表明，内耳中的毛细胞纤毛能够非常敏锐地感知声音信号的细节特征，这些特征包括声音的频率、强度和持续时间等[18]。在语音识别中，通常采用一组相互交叠的带通滤波器组模拟人耳这一机理，本文选用Gammatone滤波器组实现人耳听觉特性。

1.1 Gammatone滤波器

Gammatone滤波器的时域脉冲响应为：

$$g_i(t) = At^{n-1}\mathrm{e}^{-2\pi b_i t}\cos(2\pi f_i + \psi_i)U(t),\quad 1 \leqslant i \leqslant N \tag{1}$$

式中：N表示滤波器个数，本文取N=24；A为滤波器增益；f_i为滤波器的中心频率；$U(t)$为阶跃函数；b_i为滤波器衰减因子，与滤波器的带宽有关，由等效矩形带宽表示为：

$$b_i = 1.019\mathrm{ERB}(f_i) \tag{2}$$

$$\mathrm{ERB}(f_i) = 24.7 \times \left(4.37 \times \frac{f_i}{1\,000} + 1\right) \tag{3}$$

滤波器组由多个滤波器组成，这些滤波器的中心频率呈梳状分布，低频段带宽窄滤波器数量多，高频段滤波器数量少且带宽大，可以更好地抑制噪声干扰，其频率响应如图1所示。

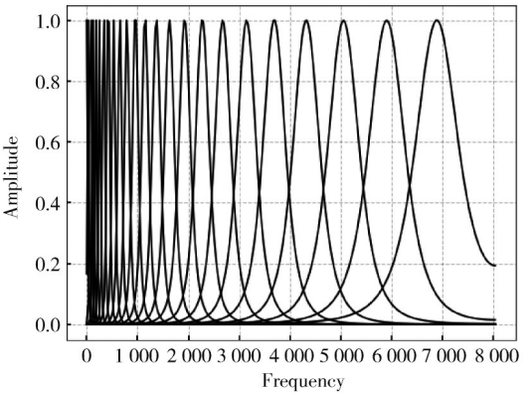

图1 Gammatone滤波器组频率响应曲线

1.2 听觉特征提取

基于Gammatone滤波器组的倒谱系数（Gammatone Frequency Cepstrum Coefficient, GFCC）提取流程如图2所示。

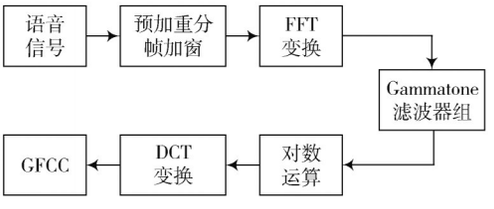

图2 GFCC提取框图

GFCC具体提取流程如下：

1）预加重操作能够增强语音信号的高频分辨率，弥补传输过程中的高频衰减，一般使用一阶FIR高通滤波器来实现，其传递函数为：

$$H(z) = 1 - \alpha z^{-1} \tag{4}$$

式中α为预加重系数，取0.97。

2）将预加重之后的语音信号划分成若干帧，每帧的时长通常为10~30 ms，本文取25 ms。为了增加相邻两帧之间的连续性，对每帧语音信号进行汉明窗加权处理，这样只需要对窗口内的数据进行观察，便于语音特性分析。汉明窗的数学表达式为：

$$W(n) = \begin{cases} 0.54 - 0.46\cos\dfrac{2\pi n}{N-1}, & 0 \leqslant n \leqslant N-1 \\ 0, \quad \text{otherwise} \end{cases} \tag{5}$$

3）加窗后的数据通过FFT变换转换为频域上的能量分布，不同能量分布代表不同的语音特性[19]。然后使用Gammatone滤波器组进行滤波处理，消除谐波影响，凸显共振峰，让频谱更加平滑。再对频谱进行对数运算和离散余弦变换，去除乘性噪声与特征分量之间的相关性，得到GFCC特征。

2　声学模型

2.1　DCNN模型

本文的DCNN主干网络如图3所示。每层卷积网络之后使用批归一化层来降低模型对数据分布的依赖性，再使用池化层对输出特征进行下采样。

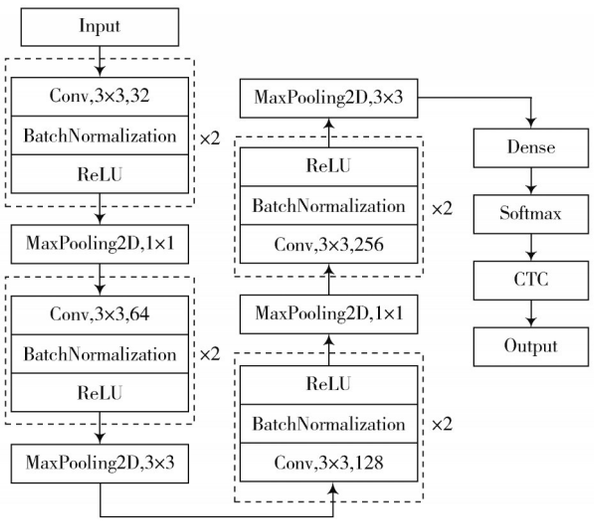

图3　DCNN模型

图3中包括8层卷积，每层卷积核的大小都是3×3，卷积核的数量分别为32、64、128、256。池化核尺寸分别为1×1、3×3，其目的是保留特征图的纹理特征，降低参数量，加快模型训练速度。

2.2　声学模型

2.2.1　残差网络

在DCNN网络中引入残差连接，使得网络的参数优化变得更加容易，比起普通堆叠网络而言，随着网络层数的增加，识别精度也更高。残差网络结构如图4所示。

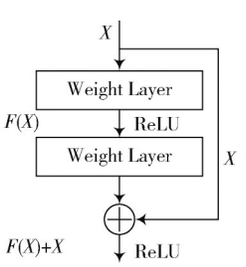

图4　残差网络结构

残差结构可简单的写成如下形式：

$$y = F(X,\boldsymbol{W}_i) + X \tag{6}$$

式中：X表示残差块的输入；y表示输出；$\boldsymbol{W}_i$为权重矩阵；$F(\cdot)$为残差函数。

如果X的维度与残差函数的输出维度不同，需要给X执行一个线性映射来匹配维度，则：

$$y = F(X,\boldsymbol{W}_i) + W_s X \tag{7}$$

2.2.2　MCFN模块设计

为了充分提取方言特定音素的底层特征，本文基于ResNet网络设计多尺度卷积模块作为模型的输入层，通过不同尺寸的卷积操作提取深层次的抽象语音特征。MCFN模块的设计结构如图5所示。

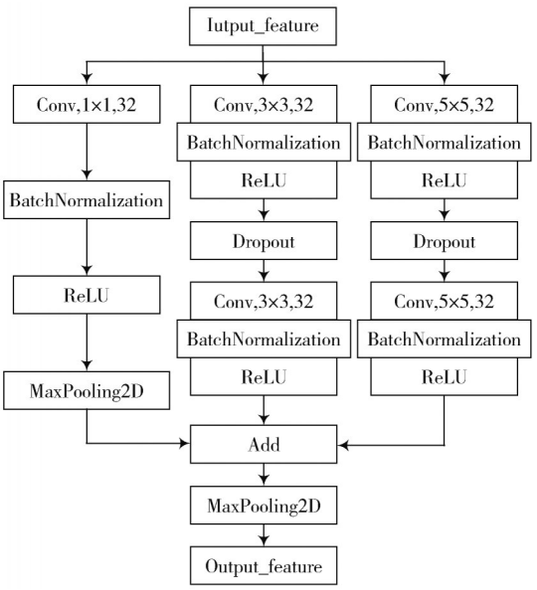

图5　MCFN模块

2.2.3　BiGRU网络

虽然卷积网络具有较强的局部特征提取能力，但是在处理语音信号时缺乏对语音序列长时依赖关系的建模能力。因此，本文结合GRU网络来提取长时特性。基本的GRU单元结构如图6所示。

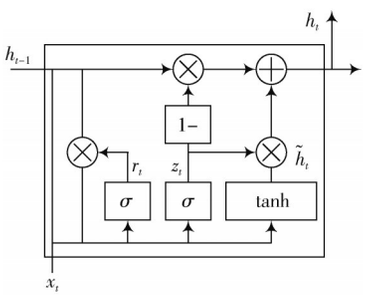

图6　GRU单元结构

GRU单元中有重置门和更新门两个门结构，其中重置门的作用是遗忘前一时刻隐层单元h_{t-1}的信息，而更新门则控制前一时刻隐层状态和当前输入信息的平衡。具体推导公式如下：

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{8}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{9}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \cdot h_{t-1}, x_t] + b_h) \tag{10}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \tag{11}$$

BiGRU网络由两层方向相反的GRU单元组成，该

网络分别在时间维的前向和后向依次处理输入序列，并将每个时间步GRU的输出拼接成为最终的输出层，这样可以让网络有效地学习序列中的上下文信息。

2.2.4 声学模型结构设计

本文设计的声学模型结构如图7所示，将多尺度卷积模块作为模型的输入层，增强模型的特征提取能力，使模型能够充分提取方言特定音素底层特征，再通过3层BiGRU网络提取时序信息，每层BiGRU单元大小设置为256。

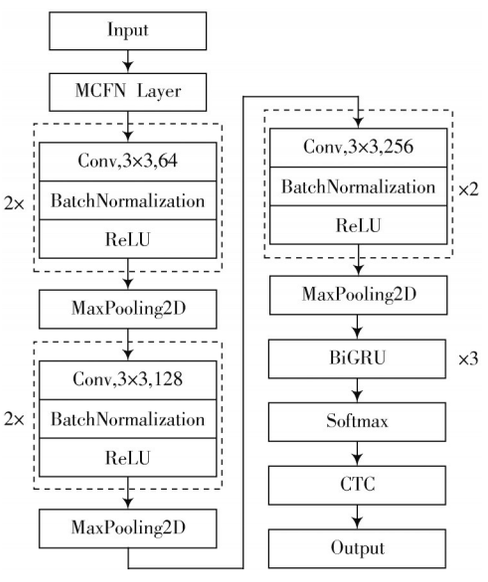

图7 ResCNN-BiGRU模型

3 实验步骤

3.1 实验数据

实验所使用的数据集由四川方言语音数据集和标准普通话数据集Thchs30-tiny构成，前者主要收集自四川当地本土影视作品和日常生活中的语音音频，再通过语速扰动、音量扰动和添加噪声等方法增广后，共收集得到6 208条四川方言数据，总时长约为9.35 h；而Thchs30-tiny由Thchs30中的Test和Dev共3 388条语音数据组成，所有数据均采用WAV格式、单声道，采样频率为16 kHz。

3.2 实验平台及评价指标

使用TensorFlow构建并测试深度学习网络模型，在PC机上运行，其运行内存为10 GB，GPU设备为1台RTX3080。

实验模型的评价指标为字错误率(Word Error Rate, WER)，计算公式如下：

$$\mathrm{WER}=\frac{S+D+I}{U}\times 100\% \tag{12}$$

式中：S表示替换的字数；D表示删除的字数；I表示插入的字数；U表示字符总数。

3.3 模型训练及优化

在模型训练时，使用Adam优化器，初始学习率设置为0.000 8，采用学习率衰减机制，衰减值设置为0.000 5，batch-size设置为16。为了防止过拟合，每层网络添加Dropout，初始值设为0.25，损失函数使用CTC-Loss，其计算公式为：

$$L(T)=-\sum_{(X,Y)\in T}\ln P\left(Y\mid X\right) \tag{13}$$

式中：T为训练集；$P\left(Y\mid X\right)$表示给定X输出Y序列的概率。

为了验证提出模型的有效性，选取并复现了如下基准模型：

模型1：文献[15]设计了基于ResNet-BLSTM的端到端语音识别模型，该模型采用残差连接增加网络结构深度，并利用BLSTM构建声学模型，同时使用CTC计算损失，在Aishell-1数据集上鲁棒性较好。

模型2：文献[20]设计了基于Maxout的语音识别模型，该模型使用Maxout作为激活函数，使用多个线性函数的组合逼近目标函数，提升了模型的泛化能力。

4 实验结果与分析

首先，在无背景噪声的条件下对比分析不同模型的性能。选取13维的GFCC语音特征作为模型的输入，对比模型分别为复现的基准模型1和模型2，以及采用消融思想搭建的模型。不同模型的表现效果如表1所示。

表1 不同模型WER性能表现 %

模型	方言数据集	Thchs30-tiny
基准模型1	4.53	17.57
基准模型2	5.95	20.58
DCNN	12.77	30.85
DCNN+MCFN	9.58	29.10
ResCNN-BiGRU	0.52	10.39
ResCNN-BiLSTM	1.12	12.10

相较于基准模型，本文提出的模型在方言数据集和普通话数据集上的字错误率均最低，这是因为本文基于残差思想设计的MCFN模块能够充分提取输入语音特征的深层抽象信息。

进一步，在方言数据集上测试文中模型的抗噪性能。选取Noisex92噪声数据库中的Babble、Pink和White三种噪声，按照信噪比0 dB、5 dB、10 dB、15 dB给纯净方言语音添加噪声，形成本次实验使用的噪声数据集。不同噪声下的字错误率对比如表2所示。

对比三种噪声下的字错误率，在白噪声环境下的识别效果较好，这是因为白噪声可以当作常数处理，而其他

两种噪声则更加复杂难以预测。总体来说，本文模型对复杂噪声环境有较好的抗噪性能，在0 dB时识别效果显著。

表2　不同噪声下字错误率对比　　%

噪声类型	0 dB	5 dB	10 dB	15 dB
Babble	6.11	2.69	1.65	1.13
Pink	2.77	1.78	1.24	1.08
White	1.97	1.32	1.23	1.02

本文模型在不同信噪比下识别三种噪声的字错误率如图8所示。

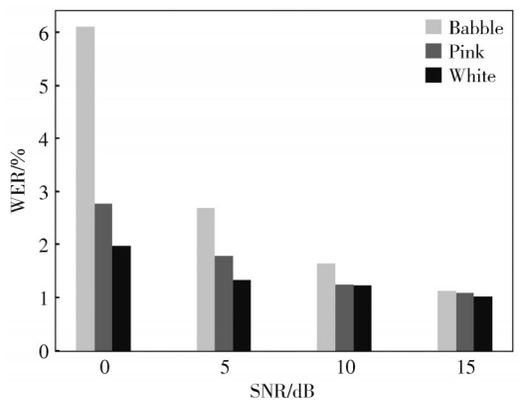

图8　ResCNN-BiGRU模型性能表现效果

5　结　论

本文提出了结合多尺度卷积和双向门控循环网络的方言语音识别模型，该模型充分发挥了MCFN模块提取底层特征的能力，以及BiGRU网络提取序列长时依赖关系的优势。在四川方言语料库上的实验结果表明，本文提出的混合语音识别模型比单一网络结构识别性能更好。下一步工作将尝试提取更有效的方言底层发音特征，并利用这些特征提高模型对复杂噪声环境的适应性。

参　考　文　献

[1] 梁雯.中国语言文化的传承与创新：评《汉语方言概要(第二版)》[J].高教探索，2019(5)：144.

[2] 杨威，胡燕.混合CTC/attention架构端到端带口音普通话识别[J].计算机应用研究，2021，38(3)：755-759.

[3] WANG W, XU W Y, SUI X, et al. Investigations of low resource multi-accent mandarin speech recognition [C]// IEEE International Conference on Information and Automation. New York: IEEE, 2015: 62-66.

[4] 刘林泉，郑方，吴文虎.基于小数据量的方言普通话语音识别声学建模[J].清华大学学报(自然科学版)，2008(4)：604-607.

[5] WINURSITO A, HIDAYAT R, BEJO A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition [C]// International Conference on Information and Communications Technology (ICOIACT). New York: IEEE, 2018: 379-383.

[6] 唐铭，何岩萍，尹恒，等.基于声道特性的腭裂语音高鼻音等级自动识别[J].计算机工程与应用，2018，54(21)：141-147.

[7] 解滔，郑晓东，张奕.基于线性预测倒谱系数的地震相分析[J].地球物理学报，2016，59(11)：4266-4277.

[8] 胡峰松，曹孝玉.基于Gammatone滤波器组的听觉特征提取[J].计算机工程，2012，38(21)：168-170.

[9] ISIDOROS R, PETROS M. Improved frequency modulation features for multichannel distant speech recognition [J]. IEEE journal of selected topics in signal processing, 2019, 13(4): 841-849.

[10] ANURAG C, ARUN R. Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals [J]. IEEE transactions on information forensics and security, 2020, 15: 1616-1629.

[11] KINNUNEN T, SAEIDI R, SEDLAK F, et al. Low-variance multitaper MFCC features: A case study in robust speaker verification [J]. IEEE transactions on audio, speech, and language processing, 2012, 20(7): 1990-2001.

[12] RISANURI H, AGUS B, SUJOKO S, et al. Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system [C]// International Conference on Information Technology and Electrical Engineering (ICITEE). New York: IEEE, 2018: 280-284.

[13] 张怡然，白静，王力.基于多窗频谱估计和平滑幅度谱包络的Mel频率倒谱系数(MFCC)改进算法[J].科学技术与工程，2014，14(19)：253-256.

[14] GONG L, XIE S, ZHANG Y, et al. A robust feature extraction method for sound signals based on Gabor and MFCC [C]// 2022 6th International Conference on Communication and Information Systems (ICCIS). New York: IEEE, 2022: 49-55.

[15] 胡章芳，徐轩，付亚芹，等.基于ResNet-BLSTM的端到端语音识别[J].计算机工程与应用，2020，56(18)：124-130.

[16] 余旭文.基于深度学习的海南方言语音识别[D].海口：海南大学，2020.

[17] 刘晓峰，宋文爱，余本国，等.基于注意力机制的大同方言语音翻译模型研究[J].中北大学学报(自然科学版)，2020，41(3)：238-243.

[18] 王勇，孟华，陈正武，等.基于Gammatone倒谱系数的直升机声信号识别[J].湖南大学学报(自然科学版)，2021，48(6)：74-79.

[19] 田娇，徐勇胜.呼吸音分析在儿童支气管哮喘中的研究进展[J].中国医学科学院学报，2021，43(5)：833-839.

[20] ZHANG Y, PEZESHKI M, BRAKEL P, et al. Towards end-to-end speech recognition with deep convolutional neural networks [EB/OL]. [2017-01-10]. https://arxiv.org/abs/1701.02720.

作者简介：谢金洪(1998—)，男，四川广安人，硕士，主要研究方向为语音识别。

魏　霞(1977—)，女，新疆昌吉人，副教授，主要研究方向为智能数据采集、网络安全、智能运维。