低资源青岛方言语音识别方法研究

相紫涵,谷　潇,饶崇郅,渐　令
(中国石油大学(华东) 经济管理学院,山东 青岛 266580)

摘　要:方言识别是语音识别的重要研究方向,常见的语音识别系统是基于标准语言训练的,导致其方言识别效果不佳。鉴于此,该文选择青岛方言作为应用案例开展方言语音识别研究。为解决方言语料匮乏、训练深度网络模型困难导致识别准确率受限等问题,提出应用数据增强方法,搭建基于改进 Conformer 的方言语音识别模型。首先,收集多源语音数据构建方言小型语料库;其次,采用数据增强技术扩充训练数据,以解决语料匮乏问题;最后,为了更好地提取信息,改进 Conformer 模型的降采样结构,引入膨胀卷积和 Mish 激活函数,实现语音到文本的直接映射。实验结果表明,提出的改进降采样模块的端到端模型结合数据增强方法后字错率可达 25.96%,能有效实现低资源条件下的方言识别。

关键词:语音识别;端到端;低资源;数据增强;青岛方言
中图分类号:TP391.42;TN912.34　　**文献标识码**:A　　**文章编号**:1673-629X(2024)04-0146-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0022

Research on Low-resource Qingdao Dialect Speech Recognition Method

XIANG Zi-han, GU Xiao, RAO Chong-zhi, JIAN Ling
(School of Economics and Management, China University of Petroleum (East China), Qingdao 266580, China)

Abstract: Dialect recognition is an important research direction in automatic speech recognition. Common speech recognition systems are based on standard language training, which results in poor performance in dialect recognition. In view of this, we choose Qingdao dialect as an application case for dialect speech recognition research. In order to solve the problems of lack of dialect corpus and difficulty in training deep network model, which lead to limited recognition accuracy, we propose to apply data augmentation method and build a dialect speech recognition model based on improved Conformer. Firstly, multi-source speech data is collected to construct a small-scale dialect corpus. Secondly, data augmentation techniques are applied to expand the training data to address the problem of data scarcity. Finally, in order to better extract information, the down-sampling structure of the Conformer model is improved, and dilated convolution and Mish activation function are introduced to realize the direct mapping from speech to text. Experimental results show that the character error rate of the end-to-end model with improved down-sampling module combined with data augmentation method can reach 25.96%, which can effectively realize dialect recognition under low resource conditions.

Key words: speech recognition; end-to-end; low resource; data augmentation; Qingdao dialect

0　引　言

近年来,随着计算机硬件能力的提升和神经网络理论的发展,深度学习技术已被广泛应用于自动语音识别[1-2],使其识别准确率得以较大提升,智能语音识别技术的落地可用性正不断突破。而当前研究应用大多局限于普通话、英语等主流语言,深究其因是这些语言资源丰富,使得在语音识别上易获取到大量语料用于模型训练。若资源匮乏或采用小型语料,声学建模需要利用不充分的数据资源训练得到尽可能多的声学特征,建立在大量语料训练基础上的深度学习模型便难以达到理想精度[3]。在使用效果、场景优化方面仍面临以下困难:一方面,在不同的场景下,存在着不同的语音识别需求,多场景数据复用性低,导致标注语料不足难以构建基于深度学习的语音识别模型,以应用于特定应用领域[4];另一方面,某些语言所处地区经济和文化水平相对落后,语言学研究匮乏,也为低资源语音识别研究带来了困难。因此,如何针对低资源语言进行识别已成为一项亟需解决的问题,受到研究者们

收稿日期:2023-06-13　　**修回日期**:2023-10-18
基金项目:国家重点研发计划(2021YFA1000100,2021YFA1000102)
作者简介:相紫涵(1998-),女,硕士研究生,研究方向为数据科学与信息管理、语音识别;通讯作者:渐　令(1981-),男,教授,博导,研究方向为智能决策与商务智能。

越来越多的关注。

针对低资源语音识别问题,很多研究者针对数据增强开展研究:使用语音数据增强方法扩充有限训练数据,模拟复杂语音形式和环境变化,从根本上解决训练样本不足的问题。一方面研究在不改变原始标签的情况下,调查语音的声学特征。如 Huang 等人[5]提出四种数据增强方法:添加速度扰动、添加体积扰动、添加房间脉冲响应和添加噪声,用于增加训练数据的数量和多样性,进而对语音特征进行表征学习。乔栋等人[6]在改进语音处理方法的基础上加入高斯白噪声对数据进行处理以缓解过拟合现象。Google 团队提出时域信号扭曲、频率遮盖、时域信号遮盖直接对频谱图等视觉表示进行增强,来生成更多训练数据,使神经网络更稳健,帮助抵抗时域信号的变形,防止频域信号和时域信号的信息丢失[7]。Kharitonov 等人[8]提出的 Wav-Aug 算法也是一类时域的数据扩充算法,通过混合多个不同强度且具有一定时间差的声音序列来模拟混响效果进而产生新的序列。缪裕青等人[9]提出利用三种方法(剪切、缩放、平移)对语谱图图片进行变换以扩充数据集。另一方面的研究利用语音拼接和语音合成获得新的语音输入,继而产生新的转录标签。如 Gokay 等人[10]结合语音合成扩充土耳其语数据并搭建端到端模型,降低了 14.8% 的字错率。于重重等人[11]提出基于 TSM 技术扩充土家语原始语料,改变语速后将不同语速的语音与原语音拼接生成新的长语音并与其他语料共同训练。

方言识别是语音研究领域的一个重要分支,但由于资源的限制,相关研究还相对较少且不够深入。以青岛为例,目前在公安警务、基层访谈、智能回访、政务应用、智慧养老等多个领域中,语音识别系统仅支持普通话,用户需以近似标准的普通话发音才能准确识别。有关青岛方言识别的理论和技术尚不成熟,限制了其在民生等领域的广泛应用。一方面,缺乏大规模的青岛方言语料库,资源匮乏且实际环境复杂难以预测,限制了青岛方言识别技术的开发;另一方面,缺少统一的汉语转换标准,不同区县之间方言存在差异,也为青岛方言识别带来了挑战。因此,推动青岛方言语料库构建和语音识别研究,对于提高人机交互效率、提升智能化生活水平具有重要价值。

为此,该文提出一种低资源青岛方言语音识别方法,主要内容包括:首先,通过人工录制、转录音频等途径收集多源数据建立青岛方言小型语料库,呈现青岛方言在词汇、语音和句法等不同层面的特征;然后,针对低资源情形下深度网络模型训练困难、泛化能力不足这一问题,对比评估常见的数据增强方法,并选择四种方法(语速扰动、音量增强、移动增强、噪声增强)对语料库进行扩容;接着,构建基于改进 Conformer 的青岛方言语音识别模型;最后,在真实数据集上开展对比实验,证明了该模型相较于其他模型具有收敛速度快、识别率高的优点。

1　青岛方言语音数据库

1.1　青岛方言语料库

所构建的青岛方言语料库由 6 139 条青岛方言发音构成(涵盖市南、市北、胶州、胶南等地)。其中 328 条数据来自于中国方言资源保护平台(男性:1 人,女性:1 人),5 811 条数据来自于志愿者录制(男性:2 人,女性:3 人),共计 7 人音频。同时语料库中所有语音数据配有普通话中文释义及对应字典。

1.1.1　文本语料设计

方言语料的朗读文本主要参考中文语音数据集 AISHELL-1 的文本、青岛方言大全、《青岛话音档》、地方故事进行设计,选取极具有代表性的青岛方言的字词句进行录制,剪辑切分语音句并进行人工标注。

1.1.2　语音录制

将方言文本进行分组,选择母语为青岛方言的志愿者分别单独录制,要求录制人吐字清晰、发音流畅、音量适中。语音录制时在无噪音、回音的环境下进行,均为手机录音语句,采用 cool edit pro2 语音编辑软件对收集的语音进行转录,采样频率设为 16 kHz,量化精度设为 16 bit,采用单声道录音,保存格式为. wav。

1.2　语音数据增强

为训练基于深度学习的端到端网络模型,通常需要大量的语音和文本数据,而低资源条件下现有资源不共享且方言资源获取非常有限。为解决这一问题,可以采用多种语音数据增强技术对有限的数据进行不同方式的变换以生成新数据,以此获得更大规模的数据集,进而降低网络发生过拟合现象的风险,缓解数据稀疏问题。需要特别注意的是,在语音识别问题中,对于语音数据增强技术,还需考虑新生成的语音数据的时序性。

因此,该文采用四种数据增强方式,以增加方言数据量和噪声数据,解决语音识别模型在受到噪声干扰时无法将声音信号解码为正确的文本序列的问题,从而提高模型泛化能力和鲁棒性,以适应各类应用场景,如噪音环境等。

2　基于改进 Conformer 的低资源青岛方言语音识别模型

针对低资源条件下的语音识别受限问题,该文提出基于改进 Conformer 的青岛方言语音识别方法,研究框架如图 1 所示。收集多源数据建立青岛方言语料库

以弥补现存语料库的缺失;同时,充分利用数据增强方法增强模型网络学习能力、提高识别精度,从而改善青岛方言语音识别模型;搭建端到端语音识别模型实现语音到文本的直接映射。

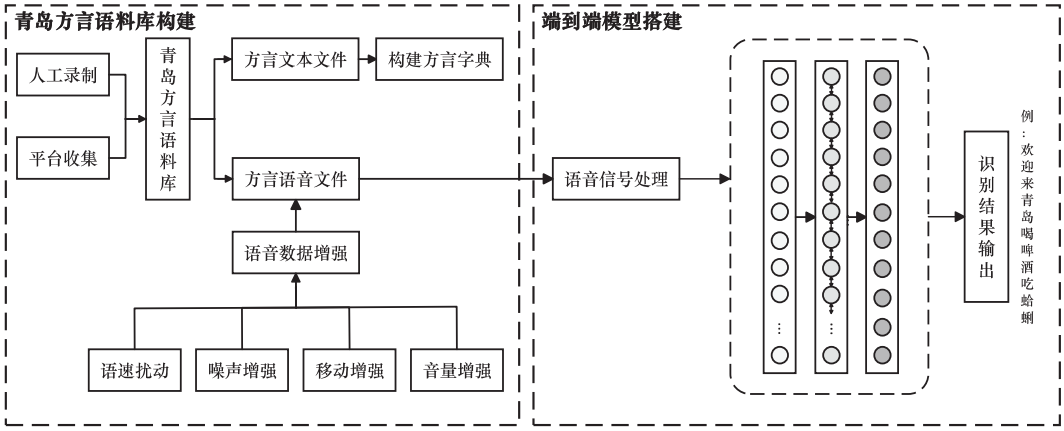

图 1 研究框架

2.1 语音信号处理

语音信号处理模块提取 FBank 特征作为神经网络的输入,具体提取流程如图 2 所示。

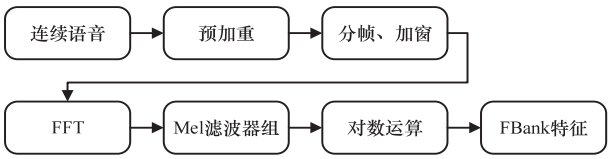

图 2 语音信号处理流程

在进行语音信号处理前应对原始语音信号进行一定程度的修正、预处理以提高语音识别模型的准确性,再进行预加重、分帧、加窗等操作:

(1)预加重:通过一个高通滤波器对输入语音信号进行处理,提高其高频部分的能量,使高频共振峰更加明显,使其更易被处理和分析。

(2)分帧、加窗:将语音信号分为若干固定长度的帧,并对每一帧信号进行窗函数加窗处理,使其在时域上呈现出一定的平滑度增加帧的连续性,以减少频谱泄露。

在预处理完成后,需要根据语音信号波形提取有效声学特征,以提高后续语音识别模型的性能和准确性。实验采用 FBank 特征,相较于梅尔频率倒谱系数(Mel - scale Frequency Cepstral Coefficients, MFCC),FBank 特征具有高相关性和小计算量,更适用于深度神经网络的数据处理。具体流程包括使用傅里叶变换将加窗后的时域信号转换到频域得到频谱,通过频谱取模平方计算出功率谱,然后通过梅尔滤波器组,最后对每个滤波器的输出取对数,以增强较小能量的信息。这样,可以得到具有区分性的语音特征,以进行后续的语音识别模型训练和测试。

2.2 Conformer 网络结构

近几年,研究人员注意到使用自注意力机制的 Transformer 模型在计算机视觉等领域展现出了强劲的性能[12],Dong 等人[13]将 Transformer 模型首次应用于完成语音识别任务。而 Transformer 具有擅长对全局上下文信息进行建模,但难以捕捉细粒度局部信息的特点,卷积神经网络则刚好弥补了这一缺点,更擅长提取局部信息。基于此,谷歌提出使用卷积对 Transformer 模型进行增强的 Conformer 结构,在 Transformer 编码器中增加卷积模块,进而提高了模型对信息的局部处理能力,以对音频序列的局部和全局依赖性进行建模,在语音识别领域中得到了广泛的应用[14]。

Conformer 编码器结构如图 3 所示。

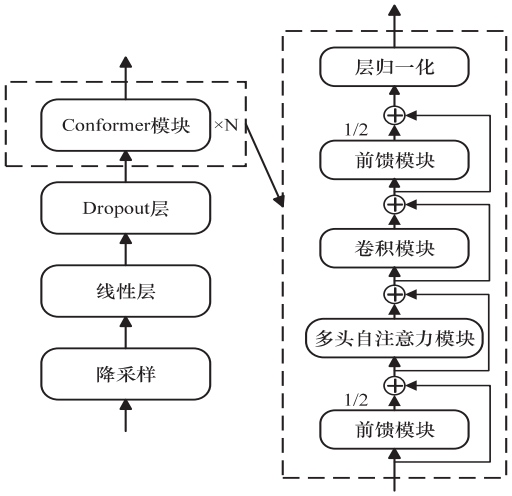

图 3 Conformer 编码器

首先通过二维卷积对数据进行降采样,再通过前馈模块输入 Conformer 模块,利用多头自注意力和卷积模块提取信息。其中,一个标准的 Conformer 模块由 4 个子模块组成(2 个前馈模块(FFN)、1 个自注意力模块、1 个卷积模块)。

与 Transformer 不同的是,Conformer 的多头自注意力模块中,Conformer 继承了 Transformer-XL 中的相对正弦位置编码,让自注意力模块在不同的输入长度上

得到更好的泛化。而卷积模块则是由逐点卷积网络(Pointwise Convolution),门控线性单位激活函数(GLU),1 - D 深度可分离卷积网络(Depthwise Convolution)构成。多头自注意力模块和卷积模块被夹在两个前馈模块中间,采用半步残差链接,并在其他模块上都采用残差单元连接。层归一化结构在 Conformer 模块的最后进行部署,以帮助训练深度模型。通过这种结构,Conformer 将卷积和注意力机制串联起来,进而起到增强的效果。公式表达如下:

$$\tilde{x}_i = x_i + \frac{1}{2}\mathrm{FFN}(x_i) \tag{1}$$

$$x_i' = \tilde{x}_i + \mathrm{MHSA}(\tilde{x}_i) \tag{2}$$

$$x_i'' = x_i' + \mathrm{Conv}(x_i') \tag{3}$$

$$h_i = \mathrm{LayerNorm}(x_i'' + \frac{1}{2}\mathrm{FFN}(x_i'')) \tag{4}$$

其中,FFN 表示前馈网络模块,MHSA 表示多头注意力模块,Conv 表示卷积模块,LayerNorm 表示层归一化。

2.3　改进降采样结构

2.3.1　膨胀卷积

卷积神经网络(Convolutional Neural Networks, CNN)由卷积层、池化层和全连接层组成,通过逐层递进的感受野有效捕捉局部上下文信息。其中卷积层采用权值共享机制减少网络需要训练的参数数量,在降低模型复杂度的同时降低过拟合风险,卷积核用于从输入信号中提取所需的信号特征。并且因其具有良好的平移不变性,能够更好地克服非平稳信号的时变性,因此被广泛应用于语音识别领域[15-16]。然而,CNN 只能通过增大卷积核尺寸的方式增大感受野,以提取深层次的特征信息,弊端在于往往伴随参数急剧增加、训练时间增长等问题。

膨胀卷积(Dilated Convolution)通过在标准卷积核中注入空洞有效增加模型的感受野,同时保持相同的参数量,无需增加计算成本即可捕获更大的区域信息,有效改进了上述问题。经典二维卷积和膨胀卷积的示意图如图 4 所示。

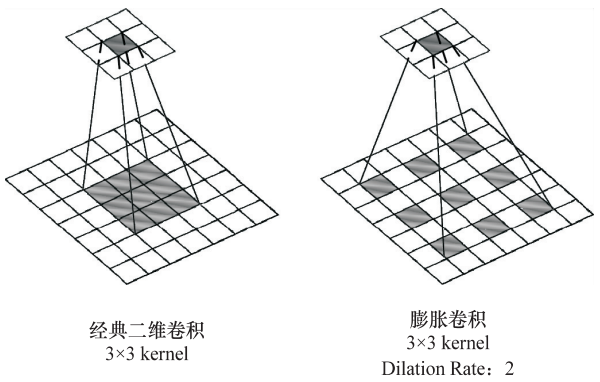

图 4　经典二维卷积和膨胀卷积

相应的,膨胀卷积在增大感受野时,会使提取信息之间存在空隙,导致丢失部分信息。而堆叠多层膨胀卷积以特征映射的方式与前一层的局部感受野连接,每 k 个数据为一组参与卷积运算,且引入膨胀率 d,允许卷积核处理数据时跳过 d 个数据进行处理,不仅降低了空间损失和信息损失,且使得特征提取更为全面,有利于提升模型的整体精度。结构如图 5 所示。

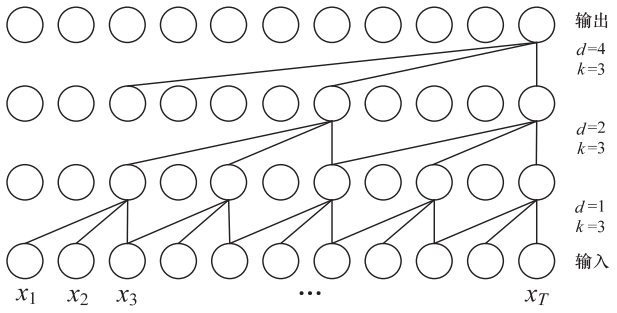

图 5　堆叠膨胀卷积

对此,该文提出使用堆叠膨胀卷积结构的方式对音频序列信号进行更好的编码,第一层膨胀卷积等同于经典二维卷积,以保证底层信息不会丢失,第二层采用空洞率为 2 的膨胀卷积,在保证感受野的情况下利用更多信息,获得长距离的历史信息。

2.3.2　Mish 激活函数

在深度学习领域中,一个好的激活函数可以使梯度更有效地传播,同时不会造成过多额外计算量。为提升模型的识别性能,文中方法将降采样模块中的 ReLU 激活函数替换为 Mish 激活函数[17],公式表达如下:

$$\mathrm{Mish} = x\tanh(\ln(1 + e^x)) \tag{5}$$

随着网络层数的增加,Mish 激活函数在准确性上的表现优于 ReLU 激活函数。这是因为相比 ReLU,Mish 激活函数具有平滑、非单调、上无边界等特点,因而允许更好的信息进入深度神经网络。尽管增大了计算量,但提升效果显著,避免 ReLU 激活函数在负值时保持为零导致训练过程中出现无法学习并收敛缓慢的问题,从而得到更好的准确性和泛化性。两种激活函数的形态如图 6 所示。

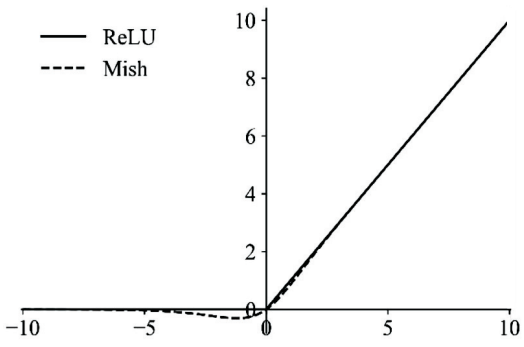

图 6　ReLU 和 Mish 激活函数

2.3.3 具有膨胀卷积的降采样结构

在 Conformer 模型中使用卷积降采样结构对 FBank 特征组成深度图进行缩放,即把原始图中一个窗口内的多个像素变成一个像素,以使语音特征更加集中,提高模型处理能力。基于前述信息,该文对该降采样结构进行改进,结构如图 7 所示。对输入使用膨胀率为 1 和 2、卷积核大小为 3 * 3、步长为 2 的膨胀卷积对特征进行卷积操作。为保证卷积模块的输出序列维度与原模型缩放比例一致,对膨胀卷积模块进行长度为 1 的零填充。每层网络之后使用 Mish 激活函数,以降低模型出现梯度消失等问题的概率。

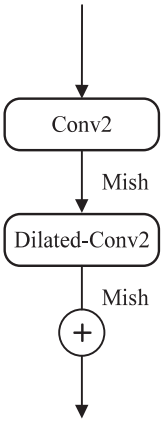

图 7 改进降采样结构

3 实验方法与结果分析

3.1 实验数据

该文自建青岛方言语料库,其中 5 人语音数据作为训练集,扩充训练集语音数据得到 28 836 条语音数据,时长共计 18.72 小时,2 人语音数据构建测试集,共计 1 333 条数据。

3.2 实验环境与实验设置

实验采用 Ubuntu18.04 操作系统,处理器使用 Inter(R) Xeon (R) Gold 6130 CPU @2.10 GHz,GPU 为 NVIDIA GeForce RTX 2080 Ti(11G),实验开发环境为 cuda11.0,python3.8.5。

训练批量大小为 16,轮数为 90,丢弃率为 0.1 并采用学习率衰减策略,初始学习率设置为 $1e^{-3}$,使用 AdamW 作为模型的优化器,并采用 warmup 学习率策略。在编码器前端采用的改进降采样结构使用的卷积核大小为 3 * 3、步幅为 2。编码器使用 12 个 Conformer 块,注意力头数设置为 4。解码器采用 Transformer,解码器块数设置为 3。

该文从原始的 wav 音频文件中提取帧移为 10 ms、窗口大小为 25 ms、维度为 80 的 FBank 特征作为输入,并在训练集上进行全局的倒谱均值方差归一化。

输出采用一个 932 个字符大小的词汇表,该词汇表包括训练集文本的 930 个字符以及额外的 2 个令牌字符,分别为空白字符<blank>、未知字符<unk>。对于测试集的标签,将集外词全部处理为<unk>标签。

训练使用 CTC loss 和 Attention loss 联合优化训练策略[18],避免 CTC 对齐过程中过于随机的同时,加快训练时连速度,使训练过程更加稳定,从而取得更好的识别效果。采用的组合训练损失函数定义如下:

$$L_{\text{CTC-Attention}} = \lambda L_{\text{CTC}} + (1-\lambda) L_{\text{Attention}} \tag{6}$$

其中,L_{CTC} 表示 CTC loss,$L_{\text{Attention}}$ 表示 Attention loss,λ 表示平衡 CTC loss 和 Attention loss 的系数,由于 CTC 一般作为辅助任务,λ 在实验中取 0.3。

3.3 评价指标

采用语音识别领域常用的字错率(Character Error Rate,CER)作为指标来评价模型性能。字错率为输出经过替换、删除、插入三类操作还原为标签所需的字数与总字数之比,越低则模型识别越精准。指标值计算公式如下:

$$\text{CER} = \frac{S + D + I}{N} \tag{7}$$

其中,S 为替换字数,D 为删除字数,I 为插入字数,N 为总字数。

3.4 实验结果与分析

该文利用数据增强技术,基于改进 Conformer 模型构建青岛方言语音识别方法。实验主要围绕以下内容评估文中模型的有效性:(1)验证语音数据增强带来的模型性能提升;(2)通过对比实验验证文中方法在青岛方言数据集上的识别有效性;(3)通过消融实验验证所提方法各模块带来的识别性能提升。

3.4.1 数据增强实验

为解决自建青岛方言语料库语料不足导致的识别性能不佳、训练困难等问题,该文使用以下 4 种方法对原有训练数据进行增强,以提高数据多样性。

(1)语速扰动。

调节所有训练数据的语速,语速增强系数分别为原始语速的 0.9 倍和 1.1 倍,保持自然声音,与实际交流语速相当,以去除说话人语速的影响。

(2)音量增强。

调整所有训练数据的音量,在[0.316,3.16]之间不均匀采样增强训练数据音量,保证增强后音频接近原始音频。

(3)移动增强。

对所有训练数据在时间轴的 ±5% 范围内随机移动,环绕转换的同时保留所有信息,不影响语音段的特征。

(4)噪声增强。

对所有训练样本按一定的间隔分布添加高斯白噪

声,模拟日常生活中的嘈杂环境。

在相同条件下,通过不同的方法对语音进行数据增强,在改进 Conformer 模型上训练进而比较识别精度,共计 6 组对比方案,其中方案 1 使用原始数据集,方案 2 ~5 分别使用在原始数据集上添加语速扰动、音量增强、移动增强、噪声增强对语音数据进行增强后的数据,方案 6 采用原始数据与 4 种方法增强后的数据一同训练。实验结果如图 8 所示。

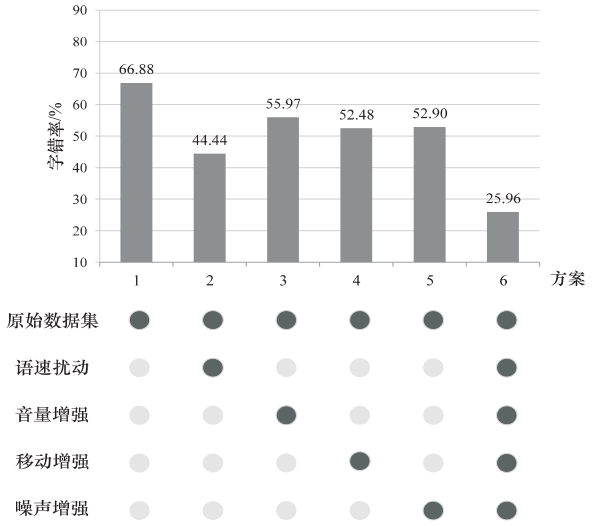

图 8　数据增强对语音识别效果的影响

由图 8 可得如下结论:

(1)方案 1 仅使用原始数据进行训练,由于缺乏数据导致模型训练效果较差。

(2)方案 2 ~4 使用数据增强方法增加数据,语速扰动、音量增强、移动增强的方法在不改变时域、频域对应关系的情况下丰富了样本的多样性,提高了总体准确率。其中,通过语速扰动扩充至三倍数据提高了 22.44 百分点,对于识别效果有最明显的提升;音量增强扩充至两倍数据提升最为有限,提高了 10.91 百分点;移动增强实际上不改变原始语音数据部分,对模型性能提升空间较为有限,仅下降了 14.4 百分点。方案 5 添加高斯噪声的方法在本模型上语音识别性能提升较为明显,下降了 13.98 百分点。高斯白噪声不仅不会破坏语音的连续特征,还有利于提高模型对噪声的适应能力。因此在语音数据增强时,需选择合适的增强方法,只增加训练数据量而忽略数据间的差异并不能取得好的识别效果。

(3)最后将原始数据与数据增强后的数据同时训练,结果表明可大幅提高模型泛化能力。在训练时,性能和训练开销达到了较好的平衡。最终确定训练数据为原始数据与使用 4 种数据增强方式后的组合。

3.4.2　对比实验

为进一步验证文中方法在青岛方言数据集上的有效性,选择 5 种模型(CNN-GRU-CTC、CNN-BiLSTM-CTC、CNN - BiGRU - CTC、DeepSpeech2[19]、Conformer[14])进行对比测试,对比实验结果如表 1 所示。

表 1　各模型对比实验结果

方案	模型	CER/%
1	CNN-GRU-CTC	70.99
2	CNN-BiLSTM-CTC	70.00
3	CNN-BiGRU-CTC	67.06
4	DeepSpeech2	71.16
5	Conformer	29.81
6	文中方法	25.96

由表 1 可得如下结论:

(1)方案 1 ~4 的实验结果表明,在基于循环神经网络的方法中,基于 BiGRU 的模型在本研究低资源方言语料库上有更好的表现。

BiGRU 和 BiLSTM 作为循环神经网络的变体,都能通过各种门保留重要特征,保证在长期信息传播时不丢失信息。但基于 BiGRU 的模型比起 BiLSTM 模型在保持其功能的情况下结构简单、参数量少、复杂度低且易于计算,因此处理少量数据性能更优、不易过拟合。

而使用 BiGRU 比 GRU 效果更优,这是因为基于 BiGRU 的模型可以同时提取过去和未来的特征信息,更好地捕捉序列中的依赖关系,泛化能力更强。

方案 4 采用的 DeepSpeech2 模型使用简单循环神经网络,具有挖掘语音数据中的复杂时序信息、模型简单的优势,训练时长较短。

Conformer 模型通过将 Transformer 模型捕捉长序列信息的能力和 CNN 捕捉局部信息的能力相结合的方式,效果明显优于上述其他模型,字错率为 29.81%,在能力范围里取得了最好的结果。

(2)通过对比实验证明,文中方法在识别性能上明显优于其他模型,选用结合数据增强方法以及改进降采样模块有效提升了 Conformer 模型的性能,在本实验数据集上更具有效果和实用性,有效识别语音信息。

3.4.3　消融实验

为进一步研究各个因素为提出的端到端青岛方言语音识别框架带来的性能提升,在第二节的基础上,通过操作以下 3 组实验条件在青岛方言数据集上进行了消融实验,逐步删除各模块。

(1)改进降采样模块中用 ReLU 替换 Mish 激活函数。

(2)改进降采样模块中去掉堆叠膨胀卷积结构,仅使用 2 个二维卷积进行降采样。

表 2 为消融实验结果。

表 2 消融实验结果

方案	模型	CER/%
1	文中模型	25.96
2	-Mish+ReLU	27.40
3	-堆叠膨胀卷积	29.81

观察表 2 发现,增加堆叠膨胀卷积模块并使用 Mish 激活函数替换 ReLU 激活函数虽然增加了训练时长,但字错率有所降低,验证了文中方法的有效性,实现了低资源条件下的方言语音识别。

4 结束语

针对青岛方言语音识别问题进行研究,提出低资源青岛方言语音识别方法。首先收集多源语音数据建立青岛方言小型语料库,并利用 4 种语音数据增强方法(语速扰动、音量增强、移动增强、噪声增强)扩充数据集;其次,在 Conformer 模型中引入 Mish 激活函数和堆叠膨胀卷积结构,提升模型识别精度,构建基于端到端模型的青岛方言语音识别方法,将语音直接映射为文字。在真实青岛方言语音数据集上的实验结果表明,数据增强可有效降低字错率。本研究可减少地区方言沟通障碍,弥补了青岛方言语音识别研究空缺。下一步将继续扩充青岛语音数据库,探究半监督训练方法,以进一步提升方言语音识别性能。

参考文献:

[1] BAHDANAU D,CHOROWSKI J,SERDYUK D,et al. End-to-end attention-based large vocabulary speech recognition [C]//2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). Shanghai: IEEE, 2016:4945-4949.

[2] ZHOU W,ZHENG Z,SCHLÜTER R,et al. On language model integration for RNN transducer based speech recognition [C]//2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). Singapore: IEEE, 2022:8407-8411.

[3] 马 晗,唐柔冰,张 义,等. 语音识别研究综述[J]. 计算机系统应用,2022,31(1):1-10.

[4] 屈 丹,杨绪魁,闫红刚,等. 低资源少样本连续语音识别最新进展[J]. 郑州大学学报:工学版,2023,44(4):1-9.

[5] HUANG C L. Exploring effective data augmentation with TDNN-LSTM neural network embedding for speaker recognition[C]//2019 IEEE automatic speech recognition and understanding workshop (ASRU). Singapore: IEEE,2019:291-295.

[6] 乔 栋,陈章进,邓 良,等. 基于改进语音处理的卷积神经网络中文语音情感识别方法[J]. 计算机工程,2022,48(2):281-290.

[7] PARK D S,CHAN W,ZHANG Y,et al. SpecAugment:a simple data augmentation method for automatic speech recognition [C]//Proceedings of the international speech communication association. Graz:ISCA,2019:2613-2617.

[8] KHARITONOV E,RIVIÈRE M,SYNNAEVE G,et al. Data augmenting contrastive learning of speech representations in the time domain[C]//2021 IEEE spoken language technology workshop (SLT). Shenzhen:IEEE,2021:215-222.

[9] 缪裕青,邹 巍,刘同来,等. 基于参数迁移和卷积循环神经网络的语音情感识别[J]. 计算机工程与应用,2019,55(10):135-140.

[10] GOKAY R,YALCIN H. Improving low resource Turkish speech recognition with data augmentation and TTS[C]//Proceedings of the 16th international multi-conference on systems,signals & devices (SSD). Istanbul:IEEE,2019:357-360.

[11] 于重重,吴佳佳,陈运兵,等. 基于多头注意力机制的端到端土家语语音识别[J]. 计算机仿真,2022,39(3):258-262.

[12] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Proceedings of the 31st international conference on neural information processing systems (NIPS). Long Beach:MIT,2017:6000-6010.

[13] DONG L,XU S,XU B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition[C]//2018 IEEE international conference on acoustics,speech and signal processing (ICASSP). Calgary: IEEE, 2018: 5884-5888.

[14] GULATI A,QIN J,CHIU C C,et al. Conformer:convolution augmented transformer for speech recognition[C]//Proceedings of the 21st annual conference of the international speech communication association. Shanghai: ISCA, 2020: 5036-5040.

[15] NEWATIA S,AGGARWAL R K. Convolutional neural network for ASR[C]//Proceedings of the 2nd international conference on electronics,communication and aerospace technology. Coimbatore:IEEE,2018:638-642.

[16] PASSRICHA V,AGGARWAL R K. A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR[J]. Journal of Ambient Intelligence and Humanized Computing,2020,11(2):675-691.

[17] MISRA D. Mish:a self regularized non-monotonic activation function[C]//British machine vision conference (BMVC). Manchester:[s. n.],2020:1-14.

[18] MA D,HOU N,XU H,et al. Multitask-based joint learning approach to robust asr for radio communication speech[C]//2021 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). Tokyo: IEEE,2021:497-502.

[19] AMODEI D,ANANTHANARAYANAN S,ANUBHAI R,et al. Deep speech 2:end-to-end speech recognition in english and mandarin[C]//Proceedings of the 33rd international conference on machine learning. New York:ACM,2016:173-182.