doi:10.3969/j.issn.1001-4616.2022.04.015

基于深度学习的长语音口音识别研究

朱丹浩[1],王　震[2],黄肖宇[3],马　壮[4],徐　杰[4]

(1.江苏警官学院刑事科学技术系,江苏 南京 210031)
(2.江苏警官学院干训部,江苏 南京 210031)
(3.江苏警官学院计算机信息与网络安全系,江苏 南京 210031)
(4.江苏省苏州市张家港市公安局,江苏 苏州 215600)

[**摘要**] 普通话口音识别是物证鉴定的重要技术之一. 目前普通话口音识别技术主要基于传统机器学习方法建立,也未针对长语音做专门设计,识别精度不高. 针对以上问题,本文提出了基于深度学习的长语音口音识别方法. 该方法首先将长语音切分为句子级别的多个短语音,然后使用经过预训练的 X-vectors 模型提取特征,再基于不同方法对句子特征进行融合,最后采用 Amsoftmax 最大化口音类别间隔并进行分类. 在真实的物证口音识别数据集上的实验结果显示,本文方法的识别精确率为 94.1%,比非深度学习的基准方法和基于 X-vectors 的基准方法分别提升了 21.6% 和 2.1%,验证了本文方法的有效性和针对长语音的口音识别能力.

[**关键词**] 深度学习,口音识别,长语音,普通话
[**中图分类号**]TP18;TN912.34　[**文献标志码**]A　[**文章编号**]1001-4616(2022)04-0110-09

Research on Long Speech Accent Recognition Based on Deep Learning

Zhu Danhao[1],Wang Zhen[2],Huang Xiaoyu[3],Ma Zhuang[4],Xu Jie[4]

(1.Department of Criminal Science and Technology,Jiangsu Police Institute,Nanjing 210031,China)
(2.Department of Cadre Training,Jiangsu Police Institute,Nanjing 210031,China)
(3.Department of Computer Information and Network Security,Jiangsu Police Institute,Nanjing 210031,China)
(4.Jiangsu Province Zhangjiagang Public Security Bureau,Suzhou 215600,China)

Abstract:Mandarin accent recognition is one of the important technical tools for identifying judicial evidence. At present, Mandarin accent recognition technology is mainly based on traditional machine learning methods, and is not specially designed for long speech,so the recognition accuracy is not high. To address the above problems,this paper proposes a long speech accent recognition method based on deep learning. The method firstly cuts the long speech into multiple short speech at sentence level,then extracts features using pre-trained X-vectors model,then fuses the sentence features based on different methods, and finally uses Amsoftmax to maximize the accent category interval and perform classification. Experimental results on a real public security accent recognition dataset show that the recognition accuracy of this paper is 94.1%,which is 21.6% and 2.1% better than the non-deep learning benchmark method and the X-vectors-based benchmark method,respectively,verifying the effectiveness of this paper and the accent recognition ability for long speech.

Key words:deep learning,accent recognition,long speech,mandarin

物证鉴定时常要对待鉴定语音进行口音分析,推测说话人的地区和籍贯. 由于我国口音多且复杂,需要专业的物证鉴定专家才能进行口音识别,这影响了口音识别技术在基层的普及程度,相关工作往往错失线索,走了弯路[1].

目前,已有少数学者探索对普通话语音进行自动口音识别[2-4],但面对物证鉴定的具体应用需求,识别精度仍显不足,主要存在两个问题. 问题一:方法和长语音的适配性不高. 待鉴定语音一般是 2 min 左右的长语音,专家进行口音识别时,须从多个句子中找到能够表征口音特征的关键字词或因素. 而现有方法面向的主要是数秒的句子级别的短语音,对长语音的特征提取能力不足. 问题二:模型的学习能力不足,

收稿日期:2022-07-27.
基金项目:国家自然科学基金项目(71974094)、江苏省社科基金项目(19TQD002)、江苏省教育厅自科项目(21KJB520004)、江苏高校优势学科工程资助项目(PAPD).
通讯作者:朱丹浩,博士,讲师,研究方向:深度学习、自然语言处理. E-mail:zhudanhao@jspi.cn

现有研究主要利用传统的高斯混合模型进行学习,而近年来在说话人识别、语音识别领域的进展[5-6]已经证明,深度学习方法具有更强大的声学特征学习能力.

针对以上两个问题,本文构建了基于深度学习的长语音口音识别方法. 对于问题一,提出了针对长语音的学习框架:将长语音切分为多个短语音句子,分别进行特征提取,再基于不同的 Kernel 对提取后的特征进行全局的特征混合. 对于问题二,使用预训练过的 X-vectors 模型[7]作为短语音的特征提取器,并使用 Amsoftmax 函数[8-9]增加模型的口音判别能力,可以有效对不定长度的短语音提取复杂的非线性口音特征. X-vectors[7]是近年来提出的基于深度学习的语音特征提取器,在说话人识别任务上取得了最高水平,而预训练模型可以提升在不同任务上的泛化能力.

本文的贡献主要有两点:第一,提出了基于深度学习的长语音口音识别方法,该方法可针对长语音学习复杂的非线性口音特征;第二,本文方法在物证鉴定口音识别数据集上的识别精度为 94.1%,超过了其它基准方法.

1 相关文献综述

在普通话的口音识别上,Hou 等[2]基于 MFCC 和共振峰频率特征,使用支持向量机(support vector machine,SVM)进行多层特征混合. 庞程等[3]使用同样的特征,利用高斯混合模型进行学习,识别上海、广东、四川等 7 个地区的普通话口音,使用的语音长度在 3-8 s 之间. 杨伟和杨俊杰[4]从方言语音学的视角,挑选音系例字进行口音识别. 总体上,普通话口音识别使用的方法比较简单,难以充分挖掘数据中复杂的特征和关系. 英文口音识别的研究相对更为深入. “口音英语识别挑战赛 2020”(accented english speech recognition challenge 2020,AESRC2020)[10]对美国、中国、日本等 8 个国家的英语口音进行了识别. 主办方提供的官方基准模型为不同层数的 Transformer,研究结果显示,经过预训练的 12 层 Transformer 取得了最好的成绩. Zhang 等[11]使用辅助的自动语音识别任务来提取与语言相关的语音特征,还提出了一种混合结构,该结构融合了固定声学模型和可训练声学模型,使得与语言相关的声学特征更加健壮. Wang 等[12]采用卷积递归神经网络作为前端编码器,并使用递归神经网络集成局部特征,他们在训练过程中添加了语音识别辅助任务缓解过拟合,并引入了人脸识别的损失函数增强特征的鉴别能力. Peng 等[13]将口音识别任务和语音识别任务进行联合训练,他们将不同口音的英语看作不同的语言,基于 DNN-HMM 框架训练了多语言识别系统.

可见,提高口音识别准确率的关键在于构建强大的语音特征编码器,该编码器一般通过在其它语音学习任务上预训练或联合训练而得. 说话人识别任务和口音识别任务较为相似,可作为口音识别的预训练特征编码器. 早先,大多数说话人识别任务基于 i-vectors(Dehak 等,2011)[14]进行,该类方法使用全局背景模型(universal background model,UBM)和投影矩阵对数据进行无监督的拟合,获取到的低维特征表示即为 i-vectors. 近几年,基于 X-vectors[7,15]的方法成为 i-vectors 的替代方案,X-vectors 基于多层的时间延迟神经网络(time delay neural networks,TDNN)[16]、统计池化层和前馈神经网络提取特征. 相比之下,X-vectors 是基于神经网络的架构,在大数据下特征抽取能力更强,在和各类数据增强算法、过拟合算法结合时,可取得更好的性能[5,17-18].

尽管英文的口音识别技术、语音特征提取技术取得了较大进展,但对应到物证鉴定工作所需的普通话长语音识别,仍存在一些问题. 例如,X-vectors 算法直接应用到长语音口音识别效果如何? 能否针对长语音进行改进? 对于普通话的长语音口音识别,如何建立强大的特征提取和分类模型?

2 方法

2.1 框架

首先将长语音切分为多个短语音句子,再基于 X-vectors 方法分别进行特征提取,最后使用不同的 Kernel 对全局特征进行混合,并预测口音.

2.2 长音频切分

当下的大多数音频特征提取器面向的都是 3-5 s 左右的单句短音频,而实际口音识别过程中使用的音频是 2 min 左右的长音频. 因此,本研究首先将每个长音频 $x \in X$ 切分成多个短音频 $x_1, x_2 \cdots x_m$,再进行后续处理.

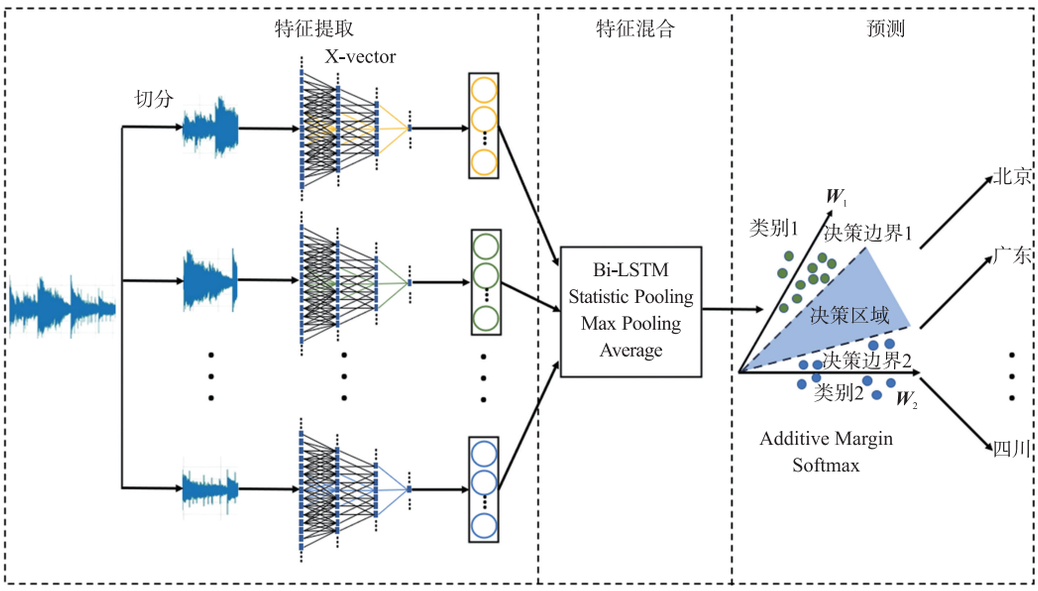

图 1　方法的总体框架图

Fig. 1　General framework diagram of the method

具体而言,本研究使用 pydub 包[19]进行切分,静音阈值设置为−70 dBFS,静音超过 700 ms 则进行切分. 为避免切分出的短音频时长过短或过长,将切分后时长小于 3 s 的音频与后面一段切分音频进行合并. 如果单个切分长度大于 10 s,则直接将其切分为多个长度小于 10 s 的短音频.

2.3　短音频的特征提取

口音识别任务和说话人识别任务同属音频分类问题,具有一定相似性. 近年来,X-vectors 算法[7,15]及其变种在说话人识别任务上大获成功,因此,本研究使用 X-vectors 算法进行短音频特征提取,X-vectors 的网络结构见表 1.

表 1　X-vectors 的网络结构

Table 1　Network structure of X-vectors

层	输入上下文	总感受野	输入/输出向量维度
TDNN1	$[t-2,t+2]$	5	5 * 24/512
TDNN2	$\{t-2,t,t+2\}$	9	512 * 3/512
TDNN3	$\{t-3,t,t+3\}$	15	512 * 3/512
TDNN4	$\{t\}$	15	512/512
TDNN5	$\{t\}$	15	512/1 500
Statistic pooling	$[0,T]$	T	1 500 * T/1 500 * 2
FNN1	$\{0\}$	T	3 000/512
FNN2	$\{0\}$	T	512/512

初始输入层为从原始的短音频中提取的 24 维的 Fbank 特征,通过宽度为 25 ms、间隔为 1 ms 的滑动窗口得到. TDNN1 到 TDNN5 分别为 5 层 TDNN 网络,在音频帧上进行计算,时刻 t 的帧向量由上一层以 t 为中心的帧上下文计算而得,因此,随着层数的增加,感受野逐渐增加,到 TDNN5 时为前后 15 帧. 由于短音频的长度不一,为使提取到的特征维度大小一致,在 TDNN5 后加入了统计池化层(statistic pooling),对所有帧得到的向量计算平均值和方差. 最后,FNN1 和 FNN2 为前馈神经网络层. 短音频 x_i 在经过特征提取后,最终得到的特征向量为 $z_i \in R^{512}$.

要训练出强大的特征提取器,仅依靠口音分类任务稀疏的标签是不充分的[10,12],因此,口音识别任务常依赖于在语音识别或说话人识别任务中得到的预训练特征提取器. 本研究使用了在 Voxceleb 数据集上预训练的 X-vectors 模型作为特征提取器[20]. 也就是说,在进行口音识别模型训练时,表 1 中的参数不是随机初始化,而是直接以文献[20]中的模型参数进行初始化.

2.4　长音频的特征融合

经过切分和短音频特征提取后,从长音频 x 中提取到多个特征向量 $z_1, z_2 \cdots, z_m$,须将其重新融合为单

个的定长特征向量 z,才能进行下一步的口音分类. 要注意的是,不同的长语音会切分出不同数量的短语音,所以 m 的大小并不固定. 本研究尝试了如下 4 种不同的融合不定长特征向量的方法,并将在实验部分验证其效果.

(1)平均值(Average)

z 为 $z_1,z_2\cdots,z_m$ 的每一个维度的平均值.

(2)最大值池化(max pooling)

z 为 $z_1,z_2\cdots,z_m$ 的每一个维度的最大值.

(3)统计池化(statistic pooling)

对 $z_1,z_2\cdots,z_m$ 的每一个维度求平均值和方差,再进行拼接得到 z. 因此,统计池化实质上是比平均值增加了长度一致的方差向量. 文献[18]发现,方差向量可以有效提升说话人识别的精度.

(4)双向长短期记忆网络(Bi-LSTM)

为充分挖掘不同短音频之间的复杂非线性关系,使用循环神经网络进行建模,为解决长序列学习的梯度消失和梯度爆炸现象,使用长短期记忆单元(long-short term memory unit)[21]. 具体实现可见文[22]. 计算公式见式(1).

$$\begin{aligned} z_f &= \mathrm{LSTM}(z_1,z_2,\cdots,z_m),\\ z_b &= \mathrm{LSTM}(z_m,z_{m-1},\cdots,z_1),\\ z &= z_f \,||\, z_b. \end{aligned} \tag{1}$$

式中,z_f、z_b 分别为前向和后向的 LSTM 的最后一个输出向量,最终融合向量 z 为二者的拼接,||表示向量首尾拼接.

2.5 口音分类

口音可看作某个地区人说话的共性特点,在进行分类时应当满足组内特征距离尽可能内聚,组间距离尽可能大. 本研究使用 Amsoftmax[9] 计算特征向量 z 对应于口音 y_j 的概率,可加大不同口音样本之间的间隔,提升分类的区分度.

首先,对特征 z 进行两层前馈神经网络的特征变换:

$$\boldsymbol{h} = \boldsymbol{W}_2\mathrm{LeakRelu}(\boldsymbol{W}_1 z+\boldsymbol{b}_1)+\boldsymbol{b}_2, \tag{2}$$

式中,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 为权重矩阵,$\boldsymbol{b}_1$,$\boldsymbol{b}_2$ 为偏置向量,LeakRelu 为常见的非线性函数.

其次,基于 $\boldsymbol{h}$ 对所属口音进行打分:

$$\mathrm{AMS}(y_j) = \frac{e^{s*(\cos\theta_{y_j}-m)}}{e^{s*(\cos\theta_{y_j}-m)} + \sum_{i,j!=j}^{N} e^{s*\cos\theta_i}}, \tag{3}$$

式中,m 是间隔,要求不同类别之间至少有 m 的间隔,s 用于将后面的打分增加倍数,因为 cos 函数的值域在[0,1]间,不扩大会使不同地区的口音概率区分不明显. $\cos\theta_{y_j}$ 则是 z 和 y_j 口音的余弦夹角:

$$\cos\theta_{y_j} = \frac{\boldsymbol{W}_{y_j}\cdot z}{|\boldsymbol{W}_{y_j}| * |z|} \tag{4}$$

$\boldsymbol{W}_{y_j}$ 是 y_j 口音对应的参数向量.

在得到口音概率得分后,选择概率最大的类为 z 对应的长音频 x 的口音:

$$y(x) = \mathrm{argmax}_{y_j}\mathrm{AMS}(y_j) \tag{5}$$

训练时,使用交叉熵损失函数:

$$\mathrm{Loss}(y_t,x) = -\sum_{x\in\text{训练集}} \log\ \mathrm{AMS}(y_t), \tag{6}$$

y_t 表示训练数据中正确的口音分类.

3 实验

3.1 实验设置

3.1.1 数据集和评测标准

本研究实验用的语音为江苏警官学院物证鉴定中心搜集的普通话长语音,形式和物证鉴定工作中的

常用数据类似. 文件格式为 wav,采样率为 16 kHz. 数据的具体统计见表 2. 切分后短音频的总时长小于原先长音频的总时长,是因为在切分时去除了句子间的静音部分. 包含了 9 个典型区域的口音,数据集总样本数为 3 550,不同口音的数据量并不平衡,如天津只有 100 个样本,而湖北等有 600 个样本,这会对训练结果产生一定影响,3.5 节进行了相应分析.

表 2　数据集统计表

Table 2　Statistical table of the data set

长音频	数量	3 550	不同口音的长音频数量	兰州	450
	平均时长	137 s		福建	400
	最大时长	498 s		广东	300
	最小时长	20 s		湖北	600
	总时长	133 h		北京	600
切分后的短音频	数量	13 451		天津	100
	平均时长	5.6 s		扬州	250
	最大时长	10 s		合肥	600
	最小时长	3 s		银川	250
	总时长	104.5 h			

评测标准为口音识别的精确率,即:

$$准确率=\frac{识别正确的样本数}{总测试样本数}. \tag{7}$$

3.1.2　基准方法

非深度学习方法:基于多特征融合的 GMM 汉语普通话口音识别方法[3],该方法使用高斯混合函数对共振峰和 Fbank 特征进行学习.

X-vectors[7]:直接用 X-vectors 方法对长语音进行特征提取和分类,不考虑长语音的特性,这也是当下语音分类任务的主流基准方法.

Transformer 方法:目前英文口音分类研究使用的主要方法之一[10,23]. 根据文献[10],较小的模型 Transformer-3L 在无预训练数据时准确率较高,较大的模型 Transformer-12L 在有预训练数据时准确率较高,故使用 Transformer-3L 和 Transformer-12L 作为基准方法.

在实验时,本研究还区分了深度学习方法中使用/不使用预训练模型的结果. 在使用预训练模型时,X-vectors 方法和本文方法统一使用了文献[20]在说话人识别任务上的预训练模型,因此,二者结果的比较是公平的. 而 Transformer 方法使用的是语音识别的预训练模型,因此可比性稍弱.

3.1.3　训练过程

采用 5 折交叉验证方法,将数据随机切分为 5 份,每次训练时依次选择 4 份作为训练集,1 份作为测试集,模型的最终表现为 5 次实验的平均结果.

为增加模型的泛化能力,使用数据增强技术辅助训练. 包括速度变换,速度随机调整为(0.9,1.0,1.1)之间. 加入环境噪声,使用 MUSAN 噪音数据集[24],随机在 0-15 db 进行加性噪音混合. 本研究的算法基于 SpeechBrain 进行开发[25]. 使用了 1 块 RTX3090 GPU 加速训练,训练一轮大约需要 4 min,训练一个模型总计需要 2 h 左右.

训练时使用的其它重要参数见表 3.

表 3　本文方法的主要参数配置

Table 3　The main parameters configuration of the method in this paper

参数名	参数值	参数名	参数值和说明
Bi-LSTM 层数	1	批大小	8
Bi-LSTM 维度	512	Amsoftmax-m	0.2
前馈神经网络维度	512	Amsoftmax-s	32
训练轮数	32	初始学习率	0.001
优化算法	Adam	最终学习率	0.0001

3.2　分类结果

本研究将方法分为 3 栏:非深度学习方法、深度学习方法(无预训练)和深度学习方法(有预训练),实验结果见表 4. 在所有方法中,使用预训练数据的本文方法(Bi-LSTM)取得了最高的精度 0.941,对比之前的普通话口音识别方法[3]大幅度提升了 0.216,证实了本文方法的有效性. 同时,在无预训练时,本文方法(Max Pooling)的精度比 X-vectors 提升了 0.069;在有预训练时,本文方法(Bi-LSTM)的精度比 X-vectors 提升了 0.021,这表明相比直接应用 X-vectors 方法,本文的框架可以有效提升在长语音上的口音分类能力.

表 4　分类结果

Table 4　Classification results

	方法	准确率		方法	准确率
非深度学习方法	高斯混合函数[3]	0.725	深度学习方法(有预训练)	X-vectors[7]	0.92
深度学习方法(无预训练)	X-vectors[7]	0.754		Transformer-3L[10]	0.909
	Transformer-3L[10]	0.772		Transformer-12L[10]	0.923
	Transformer-12L[10]	0.749		本文方法(Bi-LSTM)	0.941
	本文方法(Bi-LSTM)	0.761		本文方法(Average)	0.925
	本文方法(Average)	0.817		本文方法(Max pooling)	0.93
	本文方法(Max Pooling)	0.823		本文方法(Statistic Pooling)	0.921
	本文方法(Statistic Pooling)	0.808			

从表 4 中,可以得到以下 3 点结论:

首先,非深度学习方法的准确率明显低于深度学习方法. 方法[3]的准确率为 0.725,比深度学习方法(无预训练)中精度最低的 X-vectors 方法还要低 0.03. 这一现象也符合近几年说话人识别领域的发展趋势,即非深度学习算法逐渐被排除出各类竞赛、论文的基准方案之外. 因此,在普通话口音识别领域应用深度学习算法是必然趋势.

其次,预训练模型可以大幅度提高深度学习方法的口音识别精度. 对于所有的深度学习算法,在使用预训练参数初始化后均取得了 10%-15%的精度提升. 例如 X-vectors 的精度从 0.754 提升到了 0.92,本文方法(max pooling)的精度从 0.823 提升到 0.93.

最后,在无预训练时,小模型的效果反而要超过大模型. 在无预训练时,Transformer-3L 的识别精度要比 Transformer-12L 高 0.023,而有训练时,大模型 Transformer-12L 的优势凸显,比 Transformer-3L 要高 0.014. 同样的,本文方法(Bi-LSTM)的参数量要高于本文方法(average)、本文方法(max pooling)等,在无预训练时精度不如小模型,但在有预训练时,本文方法(Bi-LSTM)的精度有了明显提高,反超了其它小模型. 这一点也和文献[10]的结论一致.

3.3　不同特征融合方法的比较

图 2 将训练数据量调整为原先的 10%、30%、50%、70%和 90%,比较了不同特征融合方法的预测精度. 在不使用预训练模型时,图 2(a)中 Average 方法在大多数数据点上取得了最好的精度,而 Bi-LSTM 方法效果明显低于其它所有方法,但随着数据量的增加,Bi-LSTM 方法的精度提升得较快. 在使用预训练模型时,图 2(b)中的 Bi-LSTM 方法在训练数据量较小的情况下精度低于其它 3 种方法,随着训练数据量的增大,逐渐超过了其它方法. Bi-LSTM 在两图中表现得不一致,很可能是因为监督信息不足以支持参数的优化. Average、Max pooling 和 Statistic pooling 本身并不增加新参数,而 Bi-LSTM 层引入了大量的新参数,在使用预训练模型时,声音特征提取模型的参数由预训练模型的参数初始化,一定程度上缓解了对新参数优化的压力,所以 Bi-LSTM 在训练数据超过 50%时精度最高. 而不使用预训练模型时,要同时优化的参数过多,Bi-LSTM 的效果不佳.

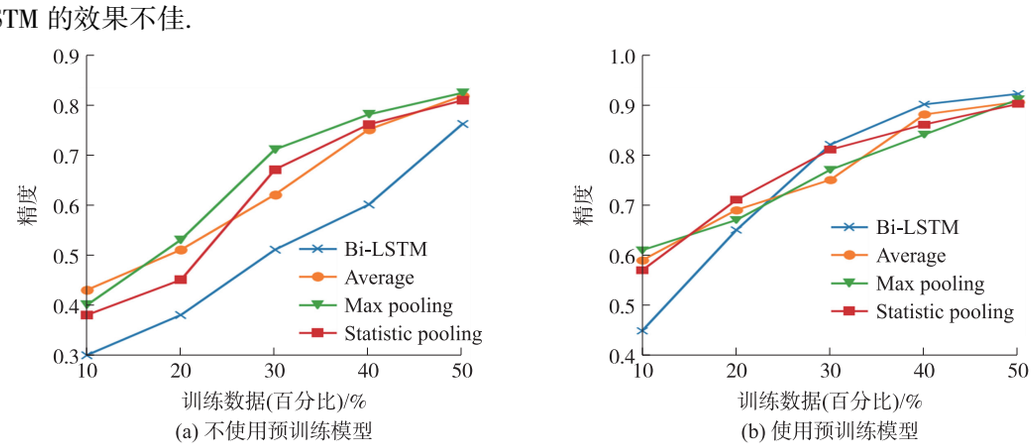

图 2　特征融合方法在不同训练数据量时的精度

Fig. 2　Accuracy of feature fusion methods at different training data amounts

综上,Bi-LSTM 提供了更好的特征融合能力,但需要大量的优化信息(预训练信息+口音分类监督数据);Average、Max pooling 和 Statistic pooling 在不同情况下表现比较稳定,可通过在测试集上的效果进行选择.

3.4　参数敏感性分析

本研究对预训练模型下的 Bi-LSTM 融合方法进行了参数敏感性分析,该方法也是在所有结果中精度最高的方法. 涉及到的参数主要在口音预测层,包括 Bi-LSTM 的层数和维度数,Amsoftmax 中的间隔 m 和缩放系数 s. 结果见图 3.

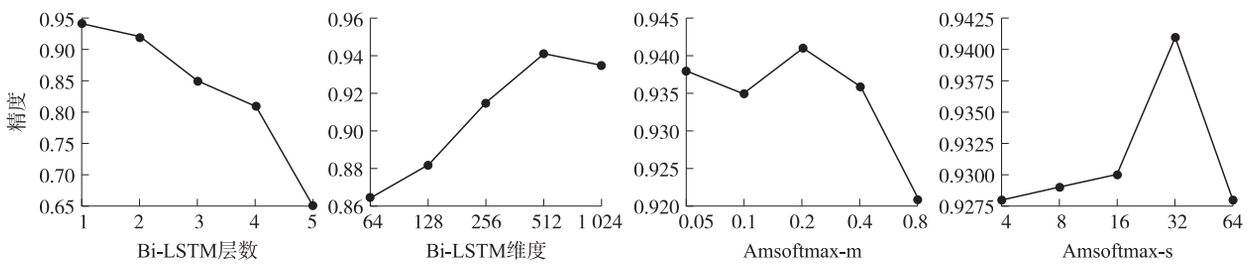

图 3　参数敏感性分析

Fig. 3　Parameter sensitivity analysis

Bi-LSTM 的超参数对结果影响较大,特别是层数上升时,精度急剧下降,在层数到 5 时精度已经只有 0.65 左右. 而维度数上升到 512 时取得最佳效果,继续增加维度精度会下降. 相比之下,Amsoftmax 的两个参数敏感性较低,对精度的影响大约在 2%左右,m 值和 s 值分别取 0.2 和 32 时精度达到峰值. 综上,对于 Bi-LSTM 的参数要小心调整,对最终精度影响很大.

3.5　模型的分类错误分析

图 4 展示了一次 5 折验证结果的混淆矩阵的热力图. 颜色越深表示数值越大,对角线上的色块表示正确分类的口音,其它位置的非 0 色块表示错误分类的口音. 总体来看,兰州、广东、福建和银川的识别准确率很高,只有极少数分类错误. 合肥和扬州的口音识别准确率稍低,合肥的口音容易被识别成湖北或北京口音. 而扬州的口音也有 4 个被识别成了天津. 天津口音的识别错误率相对较高,20 个口音中有 4 个被识别为了北京. 经过请教口音鉴定专家,在专业人士看来,北京和天津的口音是具有较高辨识性的,但由于现在人口迁移速度快,少量的混淆是可能存在的. 从模型学习的角度来看,天津的测试样本只有 20 个,对应的训练样本也远小于其它大类,这种训练数据的不充分、不平衡可能也是造成学习准确率不高的可能原因.

图 5 展示了不同时长的音频的分类平均错误率. 在实际的口音识别中,专家常通过找关键例字来判断口音,例如北京口音中的儿化音,福建口音中“h,f”不分的情况. 当音频过短时,具有区分性的关键例字出现较少,影响识别准确率. 图 5 中小于 1 min 的音频也因此识别错误率较高. 但音频长度超过 3 min 后,

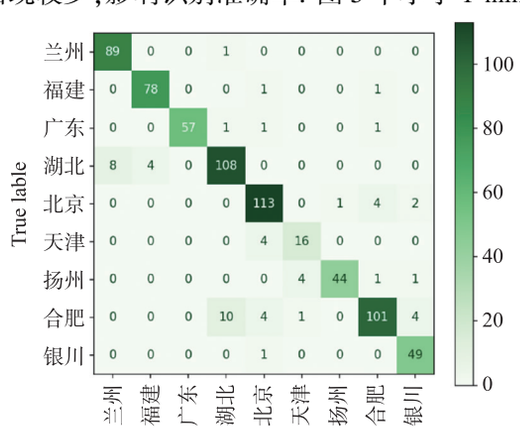

图 4　口音预测结果的混淆矩阵图

Fig. 4　Confusion matrix of accent prediction results

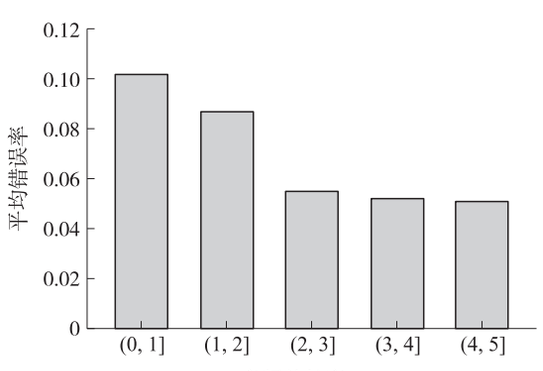

图 5　不同长度的音频的错误率

Fig. 5　The classification error rate on different video lengths

识别精度的下降已经趋缓,不见明显差异,这可能因为关键性例字已经能基本覆盖到. 因此,如果以口音识别为目标,3 min 左右的音频长度可能是性能和成本的较佳平衡点.

4 结论

本文提出了一种面向长语音的普通话口音识别方法,适用于物证鉴定场景的数据特点. 本文方法充分利用了深度学习技术在短语音特征学习中的优势,并建立了长语音学习框架. 在长语音普通话口音数据集上的实验表明,本文方法大幅度提升了普通话口音识别的准确率,优于非深度学习的方法,对比直接应用短语音深度学习方法也有明显优势.

未来我们将从两个方面推进研究. 首先,针对普通话建立预训练模型,目前使用的是在英文语料上的预训练模型,尽管精度提升效果不错,但仍和普通话有一定距离. 其次,建立其它说话人特性识别方法,提升在物证鉴定场景下的可用性. 例如,面向长语音的年龄识别、病理识别、职业识别等,可帮助缩小查找范围,提升办案效率.

[参考文献]

[1] 欧阳国亮,李志芳. 方言识别在侦查应用中面临的问题及对策[J]. 山西警察学院学报,2017,25(1):51-54.

[2] HOU J,LIU Y,ZHENG T F,et al. Multi-layered features with SVM for Chinese accent identification[C]//2010 International Conference on Audio,Language and Image Processing. Shanghai,2010:25-30.

[3] 庞程,王秀玲,张结,等. 基于多特征融合的 GMM 汉语普通话口音识别[J]. 华中科技大学学报(自然科学版),2015(S1):5.

[4] 杨伟,杨俊杰. 基于语言学音系例字的口音自动识别探究[J]. 中国司法鉴定,2021(2):5.

[5] YANG S W,CHI P H,CHUANG Y S,et al. Superb:speech processing universal performance benchmark[DB/OL]. arXiv preprint arXiv:2105.01051. [2021-03-03]. https://doi.org/10.48550.arXiv.2015.01051

[6] BAI Z,ZHANG X L. Speaker recognition based on deep learning:an overview[J]. Neural networks,2021,140:65-99.

[7] SNYDER D,GARCIA-ROMERO D,SELL G,et al. X-vectors:robust dnn embeddings for speaker recognition[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Calgary,Canada:IEEE,2018:5329-5333.

[8] MAHDI H,DENGXIN D. Unified hypersphere embedding for speaker recognition[J]. arXiv preprint arXiv:1807.08312,[2018-07-22]. https://doi.org/10.48550.arXiv.1087.08312

[9] WANG F,CHENG J,LIU W Y,et al. Additive margin softmax for face verification[J]. IEEE signal processing letters,2018,25(7):926-930.

[10] SHI X,YU F,LU Y,et al. The accented english speech recognition challenge 2020:Open datasets,tracks,baselines,results and methods[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Toronto,Canada:IEEE,2021:6918-6922.

[11] ZHANG Z,WANG Y,YANG J. Accent recognition with hybrid phonetic features[J]. Sensors,2021,21(18):6258.

[12] WANG W,ZHANG C,WU X. Deep discriminative feature learning for accent recognition[DB/OL]. arXiv preprint arXiv:2011.12461. [2020-11-25]. https://doi.org/pdf/2011.12461.pdf

[13] PENG Y,ZHANG J,ZHANG H,et al. Multilingual approach to joint speech and accent recognition with DNN-HMM Framework[C]//2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC),Tokyo,Japan:IEEE,2021:1043-1048.

[14] DEHAK N,KENNY P,DEHAK R,et al. Front-end factor analysis for speaker verification[J]. IEEE transactions on audio,speech,and language processing,2011,19(4):788-798.

[15] SNYDER D,GARCIA R D,POVEY D,et al. Deep neural network embeddings for text-independent speaker verification[C]//Interspeech,Stockholm,Sweden,2017:999-1003.

[16] PEDDINTI V,POVEY D,KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Sixteenth Annual Conference of the International Speech Communication Association,Dresden,Germany:2015.

[17] CHUNG J S,NAGRANI A,ZISSERMAN A. Voxceleb2:deep speaker recognition[DB/OL]. arXiv preprint arXiv:1806.05622. [2018-06-14]. https://doi.org/10.21437/Interspeech.2018-1929

[18] OKABE K,KOSHINAKA T,SHINODA K. Attentive statistics pooling for deep speaker embedding[DB/OL]. arXiv preprint arXiv:1803.10963. [2018-03-29]. https://doi.org/10.21437/Interspeech.2018-993

[19] jiaaro.com. Pydub[EB/OL]. https://github.com/jiaaro/pydub. (2021-03-10)[2022-07-04].

[20] Speechbrain. Speaker Verification with xvector embeddings on Voxceleb[EB/OL]. https://huggingface.co/speechbrain/spkrec-xvect-voxceleb,(2021-05-03). [2021-07-04].

[21] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural computation,1997,9(8):1735-80.

[22] ZAREMBA W,SUTSKEVER I,VINYALS O. Recurrent neural network regularization[DB/OL]. arXiv preprint arXiv:1409.2329. [2014-09-08]. https://arXiv.org/pdf/1409.2329.pdf

[23] GAO Q,WU H,SUN Y,et al. An end-to-end speech accent recognition method based on hybrid CTC/attention transformer ASR [C]//ICASSP 2021-2021 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),Toronto,Canada:IEEE,2021:7253-7257.

[24] SNYDER D,HEN G,POVEY D. MUSAN:a music,speech,and noise corpu[DB/OL]. arXiv:1510.08484v1. [2015-10-28]. https://doi.org/10.48550/arXiv.1510.08484

[25] RAVANELLI M,PARCOLLET T,PLANTINGA P,et al. SpeechBrain:a general-purpose speech toolkit[DB/OL]. arXiv preprint arXiv:2106.04624. [2021-06-08]. https://doi.org/10.48550/arXiv.2016.04624

[责任编辑:陆炳新]