DOI:10.13878/j.cnki.jnuist.20221108003

郭新[1]　邓爱文[2]　罗程方[2]　邓飞其[2]

开放场景下短时语音说话人识别系统的优化设计

摘要

为适应开放场景下说话人识别短时语音的应用需要,本文对说话人识别模型进行优化,提升了模型的准确率和鲁棒性.为了实现对重要频率特征的筛选,提出基于重加权的特征增强层及网络,起到增强特征表达的作用.将人脸识别领域的误分类样本损失函数首次引入到说话人识别领域,提高对困难样本的挖掘能力.提出基于误分类样本挖掘的分类损失与基于小样本学习框架的余弦角度原型损失的组合损失函数,解决了分类损失函数与说话人识别实际评测需求不匹配和度量函数对采样策略依赖性强的问题.实验结果显示,与基准模型相比,性能指标等误率(EER)降低12.45%,最小检测代价函数(minDCF)降低14.09%,取得现有说话人识别领域的优异效果.

关键词

说话人识别;重加权;特征增强层;分类损失函数;度量损失函数

中图分类号 TN912.3;TP18

文献标志码 A

收稿日期 2022-11-08

资助项目 广东省普通高校特色创新类项目(2022KTSCX258,2021KTSCX224);广州市基础研究计划(202002030476);广东交通职业技术学院项目(GDCP-ZX-2021-004-N1)

作者简介

郭新,女,博士,讲师,研究方向为说话人识别.guoxin@gdcp.edu.cn

邓飞其(通信作者),男,博士,教授,研究方向为动力系统的稳定性分析、随机与非线性系统的分析等.aufqdeng@scut.edu.cn

1 广东交通职业技术学院 机电工程学院,广州,510650

2 华南理工大学 自动化科学与工程学院,广州,510641

0 引言

说话人识别通过分析语音中的声纹特征来确认说话人身份,实现这一任务的关键在于如何从语谱图中提取具有足够区分性的说话人特征.说话人识别具有广泛的应用场景,如智能家居唤醒、用户账号登录和电话诈骗破案等.随着电子设备性能的提升,基于深度学习的说话人识别系统性能取得显著进步,但是在开放场景下短时输入语音的识别性能还有待提高,其核心在于如何改进帧级特征提取网络、特征聚合层和损失函数3个关键技术.

帧级特征提取网络方面,主流的网络有TDNN及其变体[1-2]和ResNet结构[3].TDNN系列网络能够提取时序依赖性强的特征,ResNet结构则利用二维卷积神经网络进行特征提取,且一般特征提取网络对输入语谱的不同频率没有区别对待,但并非每个频率范围的特征信息对说话人识别系统模型的确立同等重要,实际上低频率声纹特征具有更高的贡献度[4-5].Zhou等[6]在原始ResNet结构中加入SE(Squeeze-and-Excitation)模块,有效地增强了特征通道维度上的信息交互;Yadav等[7]则是利用基于卷积的频域和时域注意力机制来进一步增强中间特征频域和时域的信息交互性.受此启发,本文提出一种基于重加权的特征增强层及网络.

特征聚合层方面,如基于注意力机制的特征聚合层Self-Attentive Pooling(SAP)[8]和Attentive Statistics Pooling(ASP)[9].而Luo等[10]则将视频理解任务中提出的NeXtVLAD应用到说话人识别中,显著提高了模型的特征聚合效果.

损失函数设计方面,现阶段主要使用基于分类的损失函数,如AMSoftmax[11]和AAMSoftmax[12],通过在余弦角度约束的分类边界上加入间隔(margin)裕度来约束同类别特征的角度变换范围,从而提高类内特征的紧凑性,但是它们都忽视了困难样本信息对于辨别性特征学习的重要性,且本质上是根据分类任务设计的损失函数,训练时目标函数与说话人识别任务本质需求存在一定的不匹配性.

本文改进说话人识别中的两大关键技术,设计出更有效的帧级特征提取网络和使得模型训练更充分的损失函数,进一步提升基于深度学习的说话人识别模型在开放场景下短时语音的识别性能.

1 说话人识别模型基本框架

目前主流的说话人识别算法是在基于embedding向量的深度学

习框架下进行训练和测试的,整体框架如图 1 所示.

训练阶段,原始语音信号经过声学特征提取模块得到声学特征.首先,将提取的声学特征输入到帧级特征提取网络中提取帧级特征序列;然后,利用特征聚合层从帧级特征序列中提取说话人的 embedding 向量形成语句级特征向量;最后,利用说话人标签计算损失函数来优化说话人 embedding 向量,使得类内距离尽可能小,类间距离尽可能大.测试阶段,将训练好模型输出的说话人 embedding 向量输入到后端打分模型,与注册数据库中的特征向量进行相似度打分,根据得分来判断两段语音是否属于同一个说话人.

2 基于重加权的特征增强层的帧级特征提取网络

2.1 基于重加权的特征增强层

从近几年国际大型公开的说话人识别挑战赛[13-14]中可以看到,基于梅尔频谱分析的特征如 Fbank,仍然是最热门和最有竞争力的输入声学特征[15].在大多数的工作中,Fbank 频域维度特征通常会被当成一个整体同等对待而没有考虑不同频率范围特征信息的重要性.然而,不同频率范围的特征信息对于说话人识别模型的性能影响是不同的[4-5].

基于此,本文提出基于重加权的特征增强层(Reweighted-based Feature Enhancement Layer,RFEL),对输入声学特征中不同频率特征赋予不同的重要性权重,RFEL 的结构如图 2 所示.

从图 2 可知,RFEL 结构是对输入特征的重要频率进行增强,它为输入特征频域维度上的每一维频率特征计算一个权重参数,并利用该权重参数与输入特征对应频率上的值相乘,得到频率重加权后的增强特征.图 2 中,输入特征 $\boldsymbol{X}$ 大小为 $T \times F \times C$,这里的 T 表示输入特征的时域维度,F 表示频域维度,C 表示通道维度.$\boldsymbol{V}_{\text{freq}} = [v_1, v_2, \cdots, v_F]$ 表示频率权重向量,参数初值可任意设置,其后可以通过网络学习进行更新.向量 $\boldsymbol{S}_{\text{freq}} = [s_1, s_2, \cdots, s_F]$ 是 $\boldsymbol{V}_{\text{freq}}$ 经过 Sigmoid() 函数得到的向量,为的是将其权值限制在 [0,1] 范围内,维度均为 $1 \times F \times 1$.$\boldsymbol{S}_{\text{freq}}$ 和 $\boldsymbol{V}_{\text{freq}}$ 中每个参数是可由网络学习并更新的.

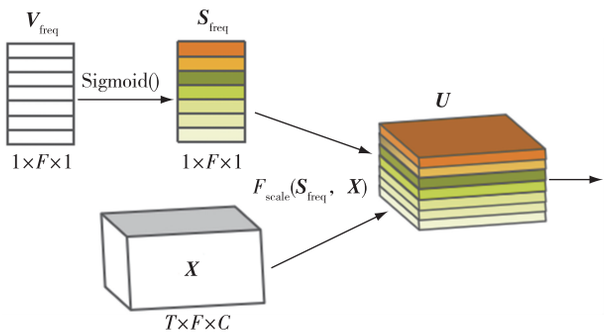

图 2 基于频率重加权的特征增强层(RFEL)结构

Fig. 2 Structure of Reweighted-based Feature Enhancement Layer (RFEL)

输入特征 $\boldsymbol{X}$ 与频率权重向量 $\boldsymbol{S}_{\text{freq}}$ 对应频率特征权重相乘,得到频率重加权后的输出特征为 $\boldsymbol{U}$,维度为 $T \times F \times C$,计算表达式如下:

$$\boldsymbol{U} = F_{\text{scale}}(\boldsymbol{S}_{\text{freq}}, \boldsymbol{X}), \tag{1}$$

其中,F_{scale} 操作表示 $\boldsymbol{S}_{\text{freq}}$ 中的每个权重值与输入特征 $\boldsymbol{X}$ 对应频率维度上的特征进行相乘.

RFEL 设计目的是为了增强输入频谱特征的重要频域维度信息,让模型能够学会分析不同频率范围内特征的重要性.基于此,本文还将 RFEL 用到网

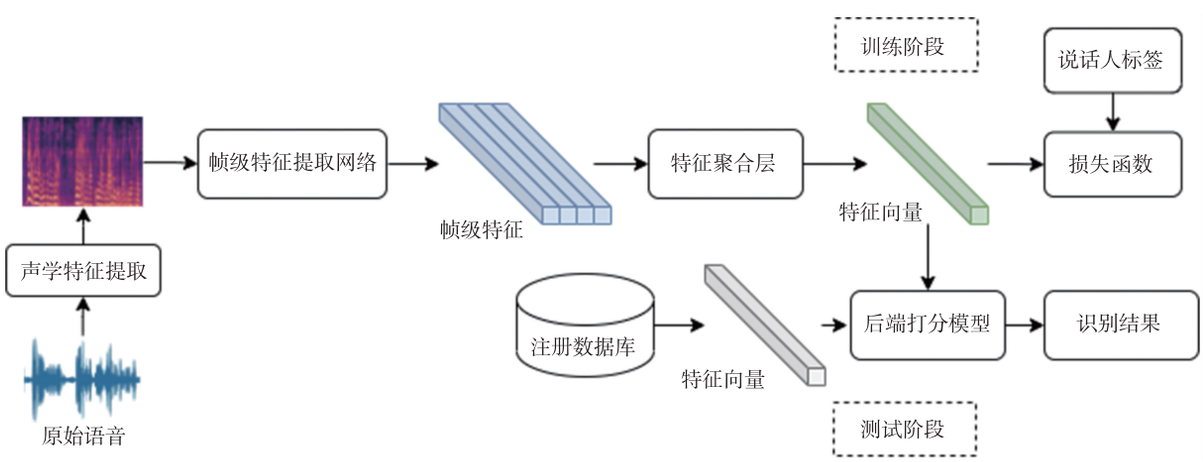

图 1 基于深度学习的说话人识别模型框架

Fig. 1 Framework of deep learning-based speaker recognition model

络中进行网络中间频域特征增强.

2.2 基于重加权的特征增强网络

本文使用的是 Fast ResNet-34 模型结构[16].为了提高轻量化模型的特征提取能力,在 Fast ResNet-34 框架中加入 SE[17] 模块,构成 Fast-SE-ResNet-34 框架,可以通过注意力机制对网络中间输出特征的通道维度进行增强.

本文提出的基于重加权的特征增强网络(Reweighted-based Feature Enhancement Network,RFEN)结构如图 3 所示.输入的频谱特征首先经过 RFEL 进行频域维度特征增强,随后经过第一层卷积神经网络降采样为原来大小的一半,再输入到 Fast-SE-ResNet-34 框架下的 4 个特征提取阶段(stage)中.从图 3 中可以看到,RFEL 可以放在每个 stage 中最后一个残差模块的输出之后,用来增强每个 stage 输出的中间特征,将最后一个 stage 输出的特征输入到特征聚合层中,则可提取到区分性强的说话人特征向量(embedding).

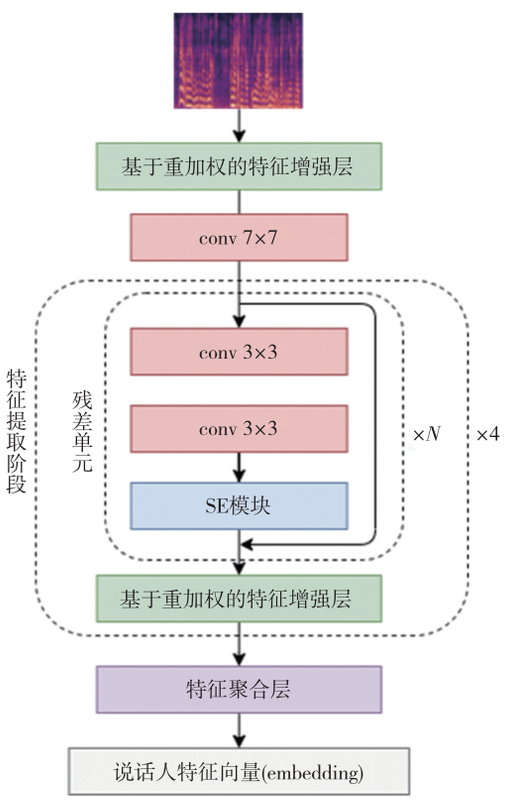

图 3 基于重加权的特征增强网络结构

Fig. 3 Structure of Reweighted-based Feature Enhancement Network (RFEN)

随着网络加深,输出每个点的特征都与周围点特征存在大量信息交互,致使频率特征高度相关,并且部分频率特征在降采样过程中被转换到通道维度,此时再使用 RFEL 对频率特征进行细粒度分析会比较困难.因此,在设计基于多层 RFEL 的帧级特征提取网络时考虑每个 stage 后 RFEL 是可选的,本文在实验部分验证了模型在不同 stage 后加入 RFEL 后的效果,从而找到最优的网络模型结构.

3 组合损失函数

损失函数是为了训练能够让模型提取出区分性强的说话人特征的一组参数.文献[18-19]提出先用 Softmax 损失函数预训练再使用度量学习损失函数微调模型的策略;文献[20-22]提出基于 Triplet loss 的困难样本对挖掘策略;文献[23]将基于小样本学习的原型网络损失引入说话人识别任务中;文献[16]使用查询集与类中心之间的距离度量计算方式,将原始的欧氏距离替换成基于余弦相似度的距离度量,取得了优异的效果.

3.1 基于误分类样本挖掘的分类损失函数

误分类样本就是原本属于类 A 的样本,被误分类到类 B 中.误分类样本大多是难样本,是比较考验模型识别能力的样本.误分类样本对于提高模型特征区分能力有着至关重要的作用.

基于误分类样本挖掘的分类损失函数 MVSoftmax 表达式如下:

$$L_{\mathrm{MV}} = -\log\left(\mathrm{e}^{sf(m,\theta_{w_y,x})}\cdot\left(\mathrm{e}^{sf(m,\theta_{w_y,x})}+\sum_{k\neq y}^{K}h(t,\theta_{w_k,x},I_k)\,\mathrm{e}^{s\cos(\theta_{w_k,x})}\right)^{-1}\right), \tag{2}$$

式中:e 为自然常数,s 为超参数(scale);$\boldsymbol{x}$ 表示当前样本经过最后一层分类层的特征表示,即为当前说话人 embedding;y 表示当前训练样本真实类别标签;$\boldsymbol{w}_k$ 表示网络最后一层分类层的参数向量,$k\in\{1,2,3,4,\cdots,K\}$,K 表示每批训练集中说话人总数;$\theta_{w_k,x}$ 表示参数向量 $\boldsymbol{w}_k$ 和当前输入特征 $\boldsymbol{x}$ 之间的角度值;$\cos(\theta_{w_k,x})$ 表示参数向量 $\boldsymbol{w}_k$ 与当前输入特征 $\boldsymbol{x}$ 之间的余弦相似度;$f(m,\theta_{w_y,x})$ 表示增加了间隔(margin)参数后的余弦相似度函数.

需要注意的是式(2)中的 I_k,它是二值指示函数,表达式如下:

$$I_k=\begin{cases}0, & f(m,\theta_{w_y,x})>\cos(\theta_{w_k,x}),\\ 1, & f(m,\theta_{w_y,x})<\cos(\theta_{w_k,x}).\end{cases} \tag{3}$$

也就是说,当样本 $\boldsymbol{x}$ 与真实说话人类别对应参数向量 $\boldsymbol{w}_y$ 之间的余弦相似度大于和非标签说话人 k

对应参数向量 $\boldsymbol{w}_k$ 之间的余弦相似度时,说明样本 $\boldsymbol{x}$ 没有被误分类到说话人 k 中,此时函数 $I_k = 0$,否则$I_k = 1$.

式(2)中,$h(t,\theta_{w_k,x},I_k)$ 是一个用来对误分类类别进行加权的函数,可以表示为

$$h(t,\theta_{w_k,x},I_k) = \mathrm{e}^{st(\cos(\theta_{w_k,x})+1)I_k}, \tag{4}$$

其中,$t \geqslant 0$ 是预先设置的超参数,当 $t = 0$ 时,基于误分类样本损失函数就变为普通的基于间隔的损失函数.当 $I_k = 0$,即没有误分类时,$h(t,\theta_{w_k,x},I_k) = 1$;当 $I_k = 1$,有被误分类时,$h(t,\theta_{w_k,x},I_k) > 1$,这表示当前样本被误分类为类别 k 时,会进一步加大样本与类别 k 分类边界之间的间隔,使得说话人特征度量空间中当前样本特征与说话人类别 k 特征更加可分.

3.2 基于原型损失的度量损失函数

在小样本学习框架下,训练集和测试集都会被分为两个不重合的子集,即支持集和查询集,用 $\boldsymbol{x}^{\mathrm{s}}$ 和 $\boldsymbol{x}^{\mathrm{q}}$ 分别表示经过特征提取后的支持集和查询集样本中的说话人特征向量,原型损失函数便是小样本学习框架下的损失函数.基于余弦相似度的原型损失函数见文献[10,24],它可表示为

$$L_{\mathrm{AP}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{e}^{w\cos(x_k^{\mathrm{q}},c_k)}}{\sum_{i=1}^{K}\mathrm{e}^{w\cos(x_k^{\mathrm{q}},c_i)}} \tag{5}$$

式中,分子表示查询样本 $\boldsymbol{x}^{\mathrm{q}}$ 与其所属真实说话人类中心向量之间的余弦相似度的自然指数,分母则表示查询样本$\boldsymbol{x}^{\mathrm{q}}$ 与当前训练批次中包含的所有说话人类中心向量之间的余弦相似度的自然指数之和.除此之外,K 表示在每个训练批次中随机抽取的说话人数量,$\boldsymbol{c}_k$ 为每个说话人类中心向量,也称为原型,表示说话人在特征空间中的特征信息,x_k^{q}表示第 k 个说话人的查询样本.

3.3 组合损失函数

分类损失函数能够优化的是样本与分类层参数向量之间的关系,属于实例—代理之间的关系,能够在训练阶段为模型提供较稳定的收敛曲线,而度量损失函数优化的是样本与支持集中真实样本的类中心向量之间的度量关系,属于实例—实例之间的关系.而说话人识别模型实质就是实例—实例之间的关系,度量函数更加符合说话人识别任务的实际应用场景.

文献[15,24]使用分类损失和度量损失的组合函数,本文将其进行优化升级,将难样本挖掘的分类损失函数融入其中,得到的组合损失函数如下:

$$L_{\mathrm{SR}} = L_{\mathrm{MV}} + \alpha L_{\mathrm{AP}}, \tag{6}$$

其中:α 为可学习的权重系数,可训练;L_{MV} 是基于难样本挖掘的分类损失函数;L_{AP} 是基于余弦相似度的原型度量损失函数;L_{SR} 代表说话人识别的组合损失函数.

4 实验结果及分析

4.1 实验设置

本文实验均在 Ubuntu18.04.3 LTS、64 位系统下进行,所采用的深度学习框架是 PyTorch,输入声学模型为 80 维的 Fbank,实验中所有模型均使用 Adam 优化器进行训练,权重衰减率设置为 5e-5,初始学习率为 0.005,batch size 大小设为 256.数据集为 VoxCeleb1,共包含 1 251 个说话人的 153 516 条音频,语音时长总计 352 h,其中训练集包含 1 211 个说话人,测试集包含 40 个说话人.基准模型[10]为基于 Fast-SE-ResNet34、NeXtVLAD 特征聚合层和 AMSoftmax loss+Augular Prototypical loss 组合损失函数的模型.

4.2 评价指标

说话人识别任务中最常用的性能评估指标为等误率(Equal Error Rate,EER)和最小检测代价函数(Minimum Detection Cost Function,minDCF).

等误率 EER 定义为错误拒绝率 FRR 与错误接受率 FAR 相等时的错误率,EER 越小说明系统的性能越好.表达式为

$$\mathrm{EER} = \mathrm{FRR} = \mathrm{FAR}, \tag{7}$$

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}, \tag{8}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}. \tag{9}$$

式中,FN 、TP、FP 和 TN 所代表的含义如表 1 混淆矩阵所示.

表 1 混淆矩阵

Table 1 Confusion matrix

真实类别	预测类别	
	正例	反例
正例	真阳性(TP)	假阴性(FN)
反例	假阳性(FP)	真阴性(TN)

minDCF 指标计算时考虑了实际使用过程中两种错误事件发生的代价,以及真实说话人和冒充者的先验概率,选择阈值使得 DCF 最小时的值为 minDCF.计算表达式为

$$\text{minDCF} = C_{\text{fa}} \times \text{FAR} \times (1 - P_{\text{target}}) + C_{\text{fr}} \times \text{FRR} \times P_{\text{target}}, \quad (10)$$

其中,C_{fa} 和 C_{fr} 分别表示错误接受样本和错误拒绝样本的权重,P_{target} 和 $1 - P_{\text{target}}$ 分别表示真实说话人和冒名顶替者出现的先验概率,minDCF 越小表示系统的风险系数越小,模型性能越好.

4.3 RFEL 有效性实验

本节实验主要验证所提出 RFEL 的有效性,实验结果如表 2 所示.

表 2 基于输入声学特征增强的频率重加权层实验结果

Table 2 Experimental results of RFEL

帧级特征提取网络	EER/%	minDCF
Fast-SE-ResNet34(基准模型)	2.73	0.298
Fast-SE-ResNet34+RFEL	**2.62**	**0.262**

另外 RFEL 还可以放在帧级特征提取网络中增强网络中间特征的频域特征,不同 stage 下 RFEL 组合消融实验结果如表 3 所示.

表 3 不同 stage 下 RFEL 组合的消融实验结果

Table 3 Results of ablation experiments with RFEL combinations at different stages

RFEL 位置	组合 1	组合 2	组合 3	组合 4	组合 5	组合 6
Input	✓	✓	✓	✓	✓	✓
stage1		✓		✓	✓	✓
stage2			✓	✓	✓	✓
stage3					✓	✓
stage4						✓
新增参数量	80	120	100	140	150	160
EER/%	2.62	2.69	2.64	**2.49**	2.52	2.72
minDCF	0.262	0.310	0.302	**0.244**	0.277	0.271

从表 2 中可看出,多层 RFEL 结构的性能都超过基准模型.表 3 中,组合 4 的 EER 为 2.49%,minDCF 为 0.244,与基准模型(EER 为 2.73%,minDCF 为 0.298)相比,EER 相对降低了 8.8%,minDCF 相对降低了 18.12%,获得最好的性能指标,所以 RFEL 的数量和位置选择极其重要.

4.4 组合损失函数有效性实验

为验证组合损失函数的有效性,组合函数参数 α 与基准模型[10]中的参数一致,均设定为 1.

由表 4 可知,使用 MVSoftmax 和 AP 组合损失函数后模型的 EER 比基准模型平均降低了 12.45%,minDCF 平均降低了 14.09%.实验结果表明,使用基于误分类样本分类损失和小样本学习框架损失的组合损失函数训练的模型性能最好.

表 4 损失函数对比实验结果

Table 4 Experimental results of loss function comparison

损失函数	EER/%	minDCF
AMSoftmax	3.11	0.276
MVSoftmax	3.01	0.268
AMSoftmax+AP	2.49	0.244
MVSoftmax+AP	2.39	**0.256**

4.5 短时语音鲁棒性实验

不同音频时长下的测试实验结果如表 5 所示.

从表 5 中可以看出,在所有音频时长测试条件下,本文提出的模型均比基准模型的识别性能要好,尤其是对于短时音频(1~2 s)场景,ΔEER 的值大于使用整段音频的测试场景,说明本文提出的改进方法能够进一步提升模型在短时语音下的性能.

表 5 不同音频时长下的 EER 测试实验结果

Table 5 Experimental results of different audio duration tests (EER) %

模型	音频时长						
	Full	1 s	1.5 s	2 s	2.5 s	3 s	4 s
基准模型	2.73	7.90	5.67	4.53	3.95	3.53	3.10
本文提出的模型	**2.39**	7.21	5.07	4.03	3.53	3.23	2.83
ΔEER	0.34	0.69	0.60	0.50	0.42	0.30	0.27

5 结束语

本文针对开放场景下短时语音的说话人识别系统进行优化,对说话人识别架构中的特征提取网络和损失函数进行优化改进.通过一种基于重加权的特征增强层来增强特征表达,并将其嵌入到网络中来改善说话人特征提取网络中间特征的区分性表示.此外,还将人脸识别中的基于误分类样本损失函数首次引入到说话人识别领域,和原型度量损失函数融合进行模型的训练,解决了分类损失函数与说话人识别本质需求不匹配和度量函数对采样策略依赖性强的问题,大大提升模型的识别精度.实验结果表明,改进后的说话人识别模型性能更加优异.

参考文献

References

[1] Snyder D,Garcia-Romero D,Sell G,et al.Speaker recognition for multi-speaker conversations using x-vectors[C]//2019 IEEE International Conference on Acoustics,Speech and Signal Processing.May 12-17,2019,Brighton,UK.IEEE,2019:5796-5800

[2] Villalba J,Chen N X,Snyder D,et al.State-of-the-art speaker recognition for telephone and video speech:the JHU-MIT submission for NIST SRE18[C]//Interspeech 2019.September 15-19,2019,Graz,Austria.ISCA,2019:1488-1492

[3] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition.June 27-30,2016,Las Vegas,NV,USA.IEEE,2016:770-778

[4] Muckenhirn H,Magimai-Doss M,Marcell S.Towards directly modeling raw speech signal for speaker verification using CNNs[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing.April 15-20,2018,Calgary,AB,Canada.IEEE,2018:4884-4888

[5] Deng A W,Wang S,Kang W X,et al.On the importance of different frequency bins for speaker verification[C]//2022 IEEE International Conference on Acoustics,Speech and Signal Processing.October 27-28,2022,Singapore,Singapore.IEEE,2022:7537-7541

[6] Zhou J F,Jiang T,Li Z,et al.Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function[C]//Interspeech 2019.September 15-19,2019,Graz,Austria.ISCA,2019:2883-2887

[7] Yadav S,Rai A.Frequency and temporal convolutional attention for text-independent speaker recognition[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing.May 4-8,2020,Barcelona,Spain.IEEE,2020:6794-6798

[8] Chowdhury F A R R,Wang Q,Moreno I L,et al.Attention-based models for text-dependent speaker verification[C]//2018 IEEE International Conference on Acoustics,Speech and Signal Processing.April 15-20,2018,Calgary,AB,Canada.IEEE,2018:5359-5363

[9] Okabe K,Koshinaka T,Shinoda K.Attentive statistics pooling for deep speaker embedding[C]//Interspeech 2018.September 2-6,2018,Hyderabad,India.ISCA,2018.Doi:10.21437/interspeech.2018-993

[10] Luo C F,Guo X,Deng A W,et al.Learning discriminative speaker embedding by improving aggregation strategy and loss function for speaker verification[C]//2021 IEEE International Joint Conference on Biometrics.August 4-7,2021,Shenzhen,China.IEEE,2021:1-8

[11] 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

[12] Deng J K,Guo J,Xue N N,et al.ArcFace:additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).June 15-20,2020,Long Beach,CA,USA.IEEE,2019:4685-4694

[13] Sadjadi S O,Kheyrkhah T,Tong A,et al.The 2016 NIST speaker recognition evaluation[C]//Interspeech 2017.August 20-24,2017,Stockholm,Sweden.ISCA,2017:1353-1357

[14] Nagrani A,Chung J S,Huh J,et al.VoxSRC 2020:the second voxceleb speaker recognition challenge[J].arXiv e-print,2020,arXiv:2012.06867

[15] 郭新,罗程方,邓爱文.基于深度学习的开放场景下声纹识别系统的设计与实现[J].南京信息工程大学学报(自然科学版),2021,13(5):526-532
GUO Xin,LUO Chengfang,DENG Aiwen.A deep learning-based speaker recognition system for open set scenarios[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2021,13(5):526-532

[16] Chung J S,Huh J,Mun S,et al.In defence of metric learning for speaker recognition[C]//Interspeech 2020.October 25-29,2020,Shanghai,China.ISCA,2020.Doi:10.21437/interspeech.2020-1064

[17] Hu J,Shen L,Albanie S,et al.Squeeze-and-excitation networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(8):2011-2023

[18] Chung J S,Nagrani A,Zisserman A.VoxCeleb2:deep speaker recognition[C]//Interspeech 2018.September 2-6,2018,Hyderabad,India.ISCA:ISCA,2018.Doi:10.21437/interspeech.2018-1929

[19] Li C,Ma X K,Jiang B,et al.Deep speaker:an end-to-end neural speaker embedding system[J].arXiv e-print,2017,arXiv:1705.02304

[20] Bredin H.TristouNet:triplet loss for speaker turn embedding[C]//2017 IEEE International Conference on Acoustics,Speech and Signal Processing.March 5-9,2017,New Orleans,LA,USA.IEEE,2017:5430-5434

[21] Zhang C L,Koishida K.End-to-end text-independent speaker verification with triplet loss on short utterances[C]//Interspeech 2017.August 20-24,2017,Stockholm,Sweden.ISCA,2017:1487-1491

[22] Jati A,Peri R,Pal M,et al.Multi-task discriminative

training of hybrid DNN-TVM model for speaker verification with noisy and far-field speech[C]//Interspeech 2019. September 15 - 19, 2019, Graz, Austria. ISCA,2019:2463-2467
[23] Wang J X, Wang K C, Law M T, et al. Centroid-based deep metric learning for speaker recognition[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. May 12-17,2019, Brighton, UK. IEEE, 2019:3652-3656
[24] Guo X, Luo C F, Deng A W, et al. DeltaVLAD: an efficient optimization algorithm to discriminate speaker embedding for text-independent speaker verification[J]. AIMS Mathematics,2022,7(4):6381-6395

Optimal design of short-time speech speaker recognition system in open scenarios

GUO Xin[1] DENG Aiwen[2] LUO Chengfang[2] DENG Feiqi[2]
1 School of Electrical and Mechanical Engineering, Guangdong Communication Polytechnic, Guangzhou 510650, China
2 School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China

Abstract To meet the application needs of speaker recognition for short-duration speech in open scenarios, we herein optimize the speaker recognition model in aspects of accuracy and robustness. First, to realize the selection of important frequency features from the input acoustic data, a Reweighted-based Feature Enhancement Layer (RFEL) and a Reweighted-based Feature Enhancement Network (RFEN) are proposed to enhance the feature representation. Second, the loss function of Misclassified Vector guided Softmax loss (MVSoftmax) in face recognition is introduced into the speaker recognition to improve the mining ability towards hard samples. Third, a combined loss function of MVSoftmax and few-shot learning based Angular Prototypical loss (AP) is proposed, which solves the mismatch between the classification loss function and the actual evaluation requirements of speaker recognition, and relieve the strong dependence of the metric function on the sampling strategy. Finally, the experimental results show that the performance metric EER of the proposed model is reduced by 12.45% and the minDCF is decreased by 14.09% compared to the baseline model, achieving excellent performance in speaker recognition.

Key words speaker recognition; reweighted; feature enhancement layer; classification loss function; metric loss function