ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1062－1084
DOI:10. 16337/j. 1004-9037. 2024. 05. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于深度学习的说话人确认方法研究现状及展望

李建琛，韩纪庆

(哈尔滨工业大学计算机科学与技术学院,哈尔滨 150001)

摘　要: 随着深度学习的不断发展,说话人确认(Speaker verification)技术已经取得了长足的进步。该技术相较于其他生物特征识别技术,具有可远程操作、成本低和易于人机交互等优势,在公安刑侦、金融服务等领域展现出广泛的应用前景。本文系统综述了基于深度学习的说话人确认技术的发展脉络。首先,介绍了基于深度学习的说话人特征表示模型在模型输入与结构、池化层、有监督损失函数和自监督学习与预训练模型4个方面的发展历程和研究现状;其次,探讨了说话人确认技术在实际应用中面临的跨域不匹配问题,如噪声干扰、信道不匹配和远场语音等,并概述了相应的领域自适应和领域泛化方法;最后,指出了进一步的研究方向。

关键词: 说话人识别;说话人确认;深度学习;领域不匹配;自监督学习

中图分类号: TN912　　**文献标志码:** A

State of the Art and Prospects of Deep Learning-Based Speaker Verification

LI Jianchen, HAN Jiqing

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: With the development of deep learning, speaker verification has made great progress. Compared with other biometric identification technologies, this technology has advantages of remote operation, low cost, easy human-computer interaction, etc., thus it shows a wide range of application prospects in the fields of public security, criminal investigation, and financial services. A systematic overview of the development lineage of deep learning-based speaker verification techniques is provided. Firstly, the development history and research status of deep learning-based speaker representation model are introduced in four aspects: Model input and structure, pooling layer, supervised loss function, and self-supervised learning and pre-training model. Then, the challenges faced by speaker verification are discussed, such as cross-domain mismatch problems like noise interference, channel mismatch and far-field speech, and the corresponding domain adaptation and domain generalization methods are outlined. Finally, the further research directions are presented.

Key words: speaker recognition; speaker verification; deep learning; domain mismatch; self-supervised learning

---

基金项目: 国家自然科学基金 (62376071)。
收稿日期: 2024-08-09;**修订日期:** 2024-09-13

引　言

随着大数据时代的到来以及科学技术的发展,人们对账户安全和身份认证等问题日益重视,同时,在公安刑侦等领域也对身份认证有着迫切需求,因此开展计算机自动身份认证的工作有着重要的意义以及广阔的应用前景。说话人确认(Speaker verification)是根据语音信号中能够表征说话人身份的声纹信息,实现自动身份认证的技术。与人脸识别、指纹识别和虹膜识别等身份认证技术相比,说话人确认继承了语音信号的以下优势:可双向传递、易于实现人机交互、不需要接触、采集成本低廉,可以进行远程身份认证。由于这些优势以及现实中迫切的需求,使得说话人确认技术得以快速发展。

说话人确认技术在几十年的研究历程中,经历了从早期基于底层声学特征的确认到目前基于高层说话人特征表示的确认、从早期基于帧级模型的确认到目前基于段级模型的确认、从早期实验室环境下的确认到目前现实复杂环境下确认的过程。多年来,国内外工业界以及学术界都对说话人确认技术开展了非常多的研究与应用。中国信息产业部在2008年公布了《自动声纹识别技术规范》[1],以促进和规范说话人确认技术在公共安全、信息产业等领域的应用。中国人工智能产业发展联盟也与得意音通声纹联合实验室等发布了中国声纹识别产业发展第一份白皮书[2]。在研究方面,国内外许多知名大学与研究机构都深入开展了说话人确认的研究工作,例如约翰霍普金斯大学的x-vector系统[3]、厦门大学的ASV-subtools工具[4]等。另外,腾讯、百度以及阿里巴巴等知名企业也对说话人确认技术开展了广泛的研究,此外还有科大讯飞、得意音通、快商通,以及思必驰等以语音技术为主导业务的企业。在应用方面,中国建设银行、邮储银行和花旗银行等都推出手机声纹认证服务,作为目前单一的数字密码认证方式的补充;支付宝和微信提供了声纹锁服务,为账号与支付安全保驾护航;微软的语音助手小娜以及苹果公司的Siri均提供了声纹唤醒的服务。

随着说话人确认技术的不断发展,其理论与技术已经取得了长足的进步。目前,说话人确认技术已经发展到了基于深度学习的说话人特征表示(Deep speaker representations)模型。本文在第1部分综述了训练和测试相匹配的实验室环境下的说话人特征表示模型。首先介绍了基础的网络输入与结构,然后介绍了池化层的发展历程,其用于生成段级说话人特征表示;其次,概述了说话人特征表示模型的有监督和自监督训练方法,并介绍了最新的基于大规模语音预训练模型的方法。这些新技术有效减少了对大量有标注训练数据的依赖,为说话人确认技术的发展提供了新思路。在第2部分,对说话人确认技术在实际应用中面临的训练和测试不匹配问题进行了探讨,例如噪声干扰、信道不匹配和远场语音等,并概述了相应的领域自适应和领域泛化等跨域说话人确认方法。最后,对说话人确认未来的研究方向进行了展望。

1　基于深度学习的说话人特征表示模型

说话人确认是通过挖掘声音中存在的说话人线索,判断测试语音是否属于注册说话人的技术。说话人确认的训练过程是在大规模说话人语料库上训练说话人特征表示(Speaker representations)模型,训练完成后可以用来提取语音段的特征表示。在测试过程中,由注册说话人录入语音并建立注册说话人模型,然后将测试语音与注册说话人模型进行匹配打分,其一般框架如图1所示。

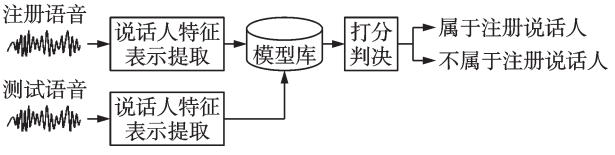

图1　说话人确认的一般框架

Fig.1　General framework for speaker verification

近年来随着深度学习热潮的到来,说话人确认方法已经由传统的i-vector[5]特征表示模型发展到了目前最新基于深度学习的说话人特征表示模型。由于深度神经网络需要固定维度的输入,因此在早期深度说话人特征表示模型d-vector[6]中,是先将当前帧与其前后固定的若

干帧作为网络输入，以得到当前帧的帧级特征表示（Frame-level representations），之后聚合输入语音段中所有的帧级特征表示作为段级说话人特征表示（Segment-level representations）。随后，Snyder等[3]提出x-vector特征表示模型，通过引入统计池化（Pooling）层，能使模型处理任意长度的输入语音段，成为目前广泛使用的基准模型。受d-vector和x-vector工作的启发，后续研究者继续在网络输入与结构、池化层和有监督损失函数上对说话人特征表示模型进行持续改进。此外，基于自监督学习和预训练模型的方法也受到广泛关注。下面分别从这4个方面进行展开。

1.1 网络输入与结构

对于网络的输入，通常是由一维声音信号通过傅里叶变换等步骤得到的二维声学特征，例如梅尔频率倒谱系数（Mel-frequency cepstral cofficients, MFCC）、滤波器组（Filter bank, Fbank）和幅度谱（Magnitude spectrum）。另外还有一些特殊的网络直接输入原始的声音信号，通过网络前几层的学习来模拟传统声学特征提取的过程[7]。

对于网络结构，时延神经网络（Time delay neural network, TDNN）[3]、残差网络（ResNet）[8]，以及转换器网络（Transformer）[9]是目前最常用的网络类型。TDNN是一种受生物学启发的神经网络结构，其本质上是通过在声学特征的时间维度上进行一维卷积以提取局部说话人特征表示。后续研究者继续对TDNN网络进行改进。例如E-TDNN[10]在TDNN基础上将网络层数进行扩展并插入仿射层，获得了比TDNN更大的时域窗长；F-TDNN[11]将TDNN中每层的权重矩阵分解为两个低秩的矩阵以减少参数量；D-TDNN[12]引入瓶颈层和跳层连接，以更少的参数量获得了比E-TDNN和F-TDNN更好的性能；ECAPA-TDNN[13]进一步融合了深度网络研究的近期成果，主要改进可归纳如下：(1)引入挤压-激励（Squeeze-excitation, SE）通道注意机制[14]以自动突出重要的说话人信息；(2)通过Res2Net[15]中的多尺度特征学习机制来捕捉不同时间分辨率下的特征模式；(3)通过多层次特征聚合来充分利用不同网络层的信息；(4)通过残差连接来避免梯度消失和梯度爆炸等问题；(5)通过密集连接更好地保留网络浅层的信息。目前ECAPA-TDNN已成为说话人确认中性能最好的网络之一，其结构如图2所示。

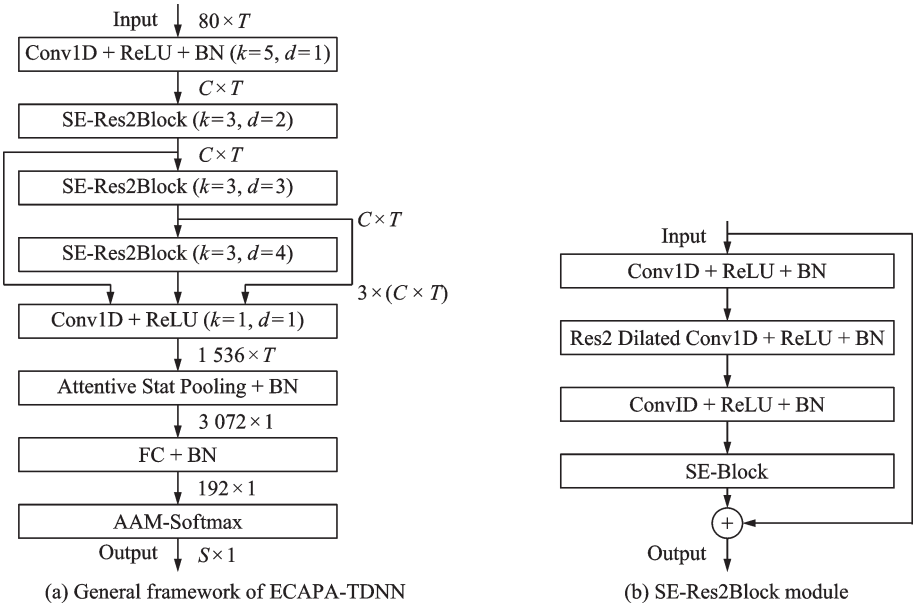

图2 基于ECAPA-TDNN网络的说话人特征表示模型

Fig.2 ECAPA-TDNN based speaker representation model

ResNet同样在说话人确认研究中得到广泛应用,它是在二维时频声学特征矩阵上进行二维卷积的神经网络。与TDNN中的一维卷积相比,在时间和频率维度上进行二维卷积可同时捕捉时间和频率维度上的复杂依赖关系,而TDNN可能无法完全捕获不同频率之间的关系。在ResNet应用的早期,Chung等[16]完全采用了图像处理中的ResNet34和ResNet50网络结构,因此性能一般。此后,Li等[17]引入Inception-ResNet结构,并探讨了其在不同持续时间下的鲁棒性。在文献[18-19]中,虽然使用的ResNet结构在参数设置上有所不同,但都将ResNet中简单的全局平均池化(Global average pooling)替换为更复杂的池化层模块,因此取得了更好的性能提升。随后,ResNet的各种变体,如Res2Net[20-21]、ResNext[20,22]、DF-ResNet[23-24]、ERes2Net[25],都在说话人确认任务中得到了应用。

Transformer最近在自然语言处理(Natural language processing, NLP)[26]、计算视觉(Compute vision, CV)[27]和语音识别(Speech recognition)[28]任务中均表现出了出色的性能。与递归神经网络(Recurrent neural networks, RNN)、TDNN和ResNet等网络相比,Transformer的优势在于其自注意机制强大的全局信息建模和并行计算能力[9]。为了探索Transformer模型对说话人信息的建模能力,Mary等[29]和Safari等[30]尝试将原始的Transformer模型引入说话人确认任务。然而,由于Transformer模型缺乏对局部信息的建模能力,因此其性能并不能令人满意。为了缓解这一问题,Han等[31]和Wang等[32]探索了通过限制注意力头的感受野并与CNN结合的局部注意力机制,以同时建模全局和局部信息。随后,许多工作都在探索如何更好地整合Transformer的全局信息建模能力和CNN网络的局部信息建模能力,其中一些工作试图将CNN插入自注意力模块[33-36],而另外一些工作则使用并行的双分支建模,然后交叉融合两个分支[37-38]。目前,基于Transformer的说话人特征表示模型已经取得了与ECAPA-TDNN和ResNet模型相当的性能。

1.2 池化层

池化层在说话人特征表示模型中起到一个承上启下的作用,其输入为上述网络得到的帧级特征表示,输出为汇集得到的段级特征表示,目标损失函数在段级特征表示上完成对整个模型的优化。可以看出,池化层是一个序列到点的过程,假设池化层的输入和输出分别为$\mathcal{H}=\{\boldsymbol{h}_t\in\mathbf{R}^{d_1}|\,t=1,2,\cdots,T\}$和$\boldsymbol{u}$,其中$\boldsymbol{h}_t$为第$t$个帧级特征表示,$d_1$为特征表示的维度。下面将列出几种常见的池化层结构并描述它们的计算过程。

1.2.1 平均池化层

平均池化层(Average pooling)[16]是最常见的池化层结构,其计算过程为

$$\boldsymbol{u}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t \tag{1}$$

1.2.2 统计池化层

除了计算帧级特征表示的均值$\boldsymbol{m}$,统计池化层(Statistics pooling)[3]还计算了$\mathcal{H}$的标准差$\boldsymbol{d}$,表达式为

$$\boldsymbol{m}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t \tag{2}$$

$$\boldsymbol{d}=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{h}_t\odot\boldsymbol{h}_t-\boldsymbol{m}\odot\boldsymbol{m}} \tag{3}$$

式中$\odot$表示哈达玛积(Hadamard product)。统计池化层的输出为均值向量和标准差向量的拼接,即$\boldsymbol{u}=\left[\boldsymbol{m}^{\mathrm{T}},\boldsymbol{d}^{\mathrm{T}}\right]^{\mathrm{T}}$。

1.2.3 单头注意力平均池化层

平均池化层和统计池化层都假设$\mathcal{H}$中各个帧级特征表示的权重一样,这会导致无法充分利用富含

说话人信息的语音帧。为此,后续研究在池化层中引入了注意力机制来学习各个帧级特征表示的权重或得分。假设第t个帧级特征表示$\boldsymbol{h}_t$的注意力得分为$s^k, k=1,2,\cdots,K$,其中K为注意力头的个数。在单头注意力平均池化层(Single-head attentive average pooling)[39]中,$K=1$。若$K\geqslant 2$,通常称为多头注意力机制。s^k的计算过程如下

$$s^k(\boldsymbol{h}_t)=(\boldsymbol{v}^k)^{\mathrm{T}}\tanh\left(\boldsymbol{W}^k\boldsymbol{h}_t+\boldsymbol{g}^k\right)+b^k \quad k=1,2,\cdots,K \tag{4}$$

式中$\boldsymbol{W}\in\mathbf{R}^{d_2\times d_1}$、$\boldsymbol{g}^k\in\mathbf{R}^{d_2}$、$\boldsymbol{v}^k\in\mathbf{R}^{d_2}$,以及$b^k\in\mathbf{R}$均为可学习的参数。计算完成后,需要对注意力得分进行归一化,有

$$\alpha_t^k=\frac{\exp\left(s_t^k\right)}{\sum_{t'=1}^{T}\exp\left(s_{t'}^k\right)} \quad k=1,2,\cdots,K \tag{5}$$

这保证了注意力得分在[0,1]之间。有了归一化后的注意得分,就可以计算$\mathcal{H}$中帧级特征表示的加权均值为

$$\boldsymbol{m}^k=\sum_{t=1}^{T}\alpha_t^k\boldsymbol{h}_t \quad k=1,2,\cdots,K \tag{6}$$

在单头注意力平均池化层中,由于$K=1$,所以汇集后的段级说话人特征表示为

$$\boldsymbol{u}=\boldsymbol{m}^1 \tag{7}$$

1.2.4 单头注意力统计池化层

在单头注意力统计池化层(Single-head attentive statistics pooling)[40]中,除了计算帧级特征表示的加权均值,还要计算加权标准差,即

$$\boldsymbol{d}^k=\sqrt{\sum_{t=1}^{T}\alpha_t^k\boldsymbol{h}_t\odot\boldsymbol{h}_t-\boldsymbol{m}^k\odot\boldsymbol{m}^k} \quad k=1,2,\cdots,K \tag{8}$$

将加权均值和加权标准差拼接起来即可得到段级说话人特征表示

$$\boldsymbol{u}=\left[(\boldsymbol{m}^1)^{\mathrm{T}},(\boldsymbol{d}^1)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{9}$$

1.2.5 全局多头注意力平均池化层

全局多头注意力平均池化层(Global multi-head attentive average pooling)[41]首次引入多头注意力机制,其中每个帧级特征表示包含K个注意力得分$s^k, k=1,2,\cdots,K, K\geqslant 2$。因此段级说话人特征表示为$K$个加权均值向量的拼接,即

$$\boldsymbol{u}=\left[(\boldsymbol{m}^1)^{\mathrm{T}},(\boldsymbol{m}^2)^{\mathrm{T}},\cdots,(\boldsymbol{m}^K)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{10}$$

值得注意的是,段级特征表示的维度增大了K倍,即$\boldsymbol{u}\in\mathbf{R}^{Kd_1}$,这导致了模型参数量和计算复杂度的增加。

1.2.6 基于子向量的多头注意力平均池化层

基于子向量的多头注意力平均池化层(Sub-vectors based multi-head attentive average pooling)[42]将帧级特征表示$\boldsymbol{h}_t$分成了K个不重叠的子向量$\boldsymbol{h}_t=[(\boldsymbol{h}_t^{(1)})^{\mathrm{T}},(\boldsymbol{h}_t^{(2)})^{\mathrm{T}},\cdots,(\boldsymbol{h}_t^{(K)})^{\mathrm{T}}]^{\mathrm{T}}$。然后,对每个子向量都应用单头注意力机制,即

$$\boldsymbol{u}^k=\sum_{t=1}^{T}\alpha_t^k\boldsymbol{h}_t^k \quad k=1,2,\cdots,K \tag{11}$$

将K个子向量对应的段级特征表示输出进行拼接,即可得到最终的段级说话人特征表示,即

$$\boldsymbol{u}=\left[(\boldsymbol{u}^1)^{\mathrm{T}},(\boldsymbol{u}^2)^{\mathrm{T}},\cdots,(\boldsymbol{u}^K)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{12}$$

由于$\boldsymbol{u}\in\mathbf{R}^{d_1/K}$,因此段级特征表示$\boldsymbol{u}$的维度与帧级特征表示保持了一致。

1.2.7 多分辨率多头注意力平均池化层

在多分辨率多头注意力平均池化层(Multi-resolution multi-head attentive average pooling)[43]中,作者提出用温度参数来控制权重分配的分辨率,将Softmax函数修改为

$$\alpha_t^k=\frac{\exp\left(s_t^k/r_k\right)}{\sum_{t'=1}^{T}\exp\left(s_{t'}^k/r_k\right)}\quad k=1,2,\cdots,K \tag{13}$$

式中r为温度系数。r越大,权重分配的曲线越平缓,即分辨率越低。段级特征表示$\boldsymbol{u}$的计算与全局多头注意力平局池化层类似。

基于注意力机制的池化层通常可以获得比传统的平均池化层和统计池化层更好的性能。表1从是否为多头注意力、是否计算帧级特征表示的标准差、池化层输入和输出的维度是否一致,以及计算注意力得分时是否有温度系数4个角度,对几种常见的注意力池化层进行了区分。在实际中可根据在不同数据集上的表现灵活选择合适的注意力池化层结构。

表1 不同注意力池化层的对比

Table 1 Comparison of various attention-based pooling layers

方法	多头	标准差	维度一致	温度项
单头注意力平均池化层	×	×	√	×
单头注意力统计池化层	×	√	×	×
全局多头注意力平均池化层	√	×	×	×
基于子向量的多头注意力平均池化层	√	×	√	×
多分辨率多头注意力平均池化层	√	×	√	√

1.3 有监督损失函数

损失函数在很大程度上决定了神经网络的性能,所以选择合适的损失函数很重要。由于说话人确认是一个度量学习任务,因此目前常用的损失函数都基于度量学习,可以分为两类:一类是基于成对训练数据的损失(Pair-based loss)[16,44-45],另一类是基于间隔的Softmax损失(Margin-based Softmax loss)[46-47]。下面将分别介绍这两类损失函数。

1.3.1 基于成对训练数据的损失

基于成对训练数据的损失需将训练数据组织成正样本对和负样本对的形式,其优化目标就是使得正样本对相互靠近,而负样本对相互远离。对比损失(Contrastive loss)[16]是其中最简单直观的一种损失函数,其思想为:(1)选取一个正样本对,其对应的损失应该等于它们特征表示之间的欧氏距离,最小化损失就是最小化正样本对之间的距离;(2)选取一个负样本对,那么它们之间的距离应尽可能大,如果大于某个给定的阈值m,则停止训练。根据这一思想,可以得到如下形式的对比损失

$$\mathcal{L}=\sum_{y_{ij}=1}d_{ij}+\sum_{y_{ij}=0}\left[m-d_{ij}\right]_+ \tag{14}$$

式中:d_{ij}表示i和j之间的距离;$y_{ij}=1$表示这两个样本属于同一个说话人,即正样本对,$y_{ij}=0$表示这两个样本属于不同说话人,即负样本对;$[\cdot]_+$为hinge函数。

对比损失使得正样本对之间的距离尽可能的小,负样本对之间的距离尽可能大。而三元组损失(Triplet loss)[44]的思想是使负样本对之间的距离大于正样本对之间的距离。在训练过程中会同时选取一对正样本和负样本,且正负样本对中有一个样本是相同的,称为锚点(Anchor)样本。其损失函数形

式如下

$$\mathcal{L}=\sum_{y_{ap}=1,y_{an}=0}\left[d_{ap}-d_{an}+m\right]_{+} \tag{15}$$

式中：a、n和p分别表示锚点样本、负样本和正样本；d_{ap}表示锚点与正样本之间的距离；d_{an}表示锚点与负样本之间的距离；m为人工设定的间隔。

三元组损失使得负样本对的距离超过正样本对的距离一定间隔时就停止优化，但每次迭代只能关注一个负样本对。为了加快模型收敛的速度，N-对损失(N-pair loss)[45]同时考虑了包含训练集中所有说话人类别的多个负样本对，并计算锚点样本与每个负样本之间的距离。具体来说，假设训练集中有N个说话人，则每个正样本对都对应了$N-1$个负样本对。基于N个样本对和$N(N-1)$个负样本对计算的N-pair损失形式如下

$$\mathcal{L}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}\log\left(1+\exp\left(d_{ii}-d_{ij}\right)\right) \tag{16}$$

式中：d_{ii}为第i个说话人类别中锚点样本和正样本之间的距离；d_{ij}为锚点样本与第j个说话人类别中负样本之间的距离。N-pair损失还用SoftPlus激活函数$f(x)=\log\left(1+\mathrm{e}^{x}\right)$代替了hinge函数，可以有效缓解梯度消失或爆炸的问题。

虽然基于成对训练数据的损失可以取得很好的结果，但需要在每次迭代过程中选择合适的负样本对，距离太小或太大的负样本对都会对模型优化产生不利的影响，这在一定程度上限制了此类损失函数的应用。

1.3.2 基于间隔的Softmax损失

除了基于成对样本来计算损失函数，还可以直接对样本进行分类，将说话人特征表示模型的优化变成多分类问题。多分类问题通常以最小化交叉熵作为优化目标，并以Softmax为输出层的激活函数。该损失函数可简称为Softmax损失，其形式如下

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+b_{y_i}\right)}{\sum_{c=1}^{C}\exp\left(\boldsymbol{w}_{c}^{\mathrm{T}}\boldsymbol{x}_i+b_{c}\right)} \tag{17}$$

式中：$y_i\in\{1,2,\cdots,C\}$为样本$\boldsymbol{x}_i$的说话人标签；C为训练集中的说话人类别数；N为每次迭代时小批量样本的数量；$\boldsymbol{w}$为网络最后一层分类层的权重向量；b为偏置。虽然Softmax损失可以将不同说话人的样本进行有效区分，但无法进一步减小相同说话人样本之间的距离。为此，研究者在Softmax损失中引入不同形式的间隔(Margin)，以增强类间可区分性并减小类内方差。下面介绍两种最常见的带有间隔的Softmax损失：加性间隔Softmax损失(Additive margin Softmax, AM-Softmax)[46]和加性角度间隔Softmax损失(Additive angular margin Softmax, AAMSoftmax)[47]。

加性间隔Softmax损失[46]先将$\boldsymbol{w}$和$\boldsymbol{x}_i$进行归一化，因此式(17)中的内积$\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i$和$\boldsymbol{w}_{c}^{\mathrm{T}}\boldsymbol{x}_i$都变成了余弦相似度的形式。然后，在$\cos\left\langle\boldsymbol{w}_{y_i},\boldsymbol{x}_i\right\rangle$中引入一个间隔$m$，即$\cos\left\langle\boldsymbol{w}_{y_i},\boldsymbol{x}_i\right\rangle-m$。引入$m$后，样本$\boldsymbol{x}_i$与目标类别权重$\boldsymbol{w}_{y_i}$之间的相似度得分就会减小，因此网络会继续优化以使$\boldsymbol{x}_i$与$\boldsymbol{w}_{y_i}$的相似度继续增大。AM-Softmax损失的形式如下

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(s\left(\cos\left\langle\boldsymbol{w}_{y_i},\boldsymbol{x}_i\right\rangle-m\right)\right)}{\exp\left(s\left(\cos\left\langle\boldsymbol{w}_{y_i},\boldsymbol{x}_i\right\rangle-m\right)\right)+\sum_{c=1,c\neq y_i}^{C}\exp\left(s\left(\cos\langle\boldsymbol{w}_{c},\boldsymbol{x}_i\rangle\right)\right)} \tag{18}$$

式中s为一个缩放系数，可以防止训练时梯度过小。

加性角度间隔Softmax损失[47]在角度上引入间隔,即将式(18)中的$\cos\langle \boldsymbol{w}_{y_i}, \boldsymbol{x}_i\rangle - m$替换为$\cos(\theta_{y_i,i}+m)$,有

$$\mathcal{L}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\Big(s\Big(\cos\big(\theta_{y_i,i}+m\big)\Big)\Big)}{\exp\Big(s\Big(\cos\big(\theta_{y_i,i}+m\big)\Big)\Big)+\sum\limits_{c=1,c\neq y_i}^{C}\exp\Big(s\big(\cos\theta_{c,i}\big)\Big)} \tag{19}$$

式中$\theta_{y_i,i}$表示$\boldsymbol{w}_{y_i}$和$\boldsymbol{x}_i$之间的角度。

在实践中,基于间隔的Softmax损失无需复杂的负样本选择策略,而且可以获得很好的结果,因此成为目前说话人确认中研究的主流。

1.4 自监督学习和预训练模型

传统基于深度学习的说话人特征表示模型通常以有监督的方式从头开始训练,但收集大规模带有说话人标签的数据非常耗时,有时还会涉及隐私问题。因此,从数据本身挖掘潜在的标签和内部结构并设计有效的自监督学习(Self-supervised learning)目标受到了越来越多的关注[48-51]。说话人确认中的自监督学习可分为两类:对比式[48-49]和非对比式[50-51]。对比式方法通常假设一句语音只有一个说话人,因此可以从同一句语音中采样正样本对;同时,假设不同的语音包含不同的说话人,从而可以采样得到负样本对。最后,就可以设计不同的对比损失函数来最小化正样本对之间的距离,同时最大化负样本对之间的距离。例如,使用L_1距离来度量样本对之间的距离[52];或者最大化正样本并最小化负样本之间的互信息[53]。为使模型对信道的变化更具鲁棒性,可以在采样时使用不同的数据增强方式,以保证正样本对之间的共享信息不包括信道等无关信息[54]。此外,为有效丰富训练时小批量数据的多样性,研究者还采用动态缓存的方式来存储大量的负样本[55]。

尽管基于对比学习的自监督方法可以有效地学习说话人特征表示,但假设不同语音包含不同的说话人可能会带来假负样本对。为此,研究者提出了无需构建负样本对的Barlow Twins方法[56],从减少说话人特征表示各维度之间冗余的角度进行学习。此外,无需标签的自蒸馏方法(Self-distillation with no labels,DINO)[57-58]逐渐成为目前的主流,其包括学生和教师两个平行网络,具体架构如图3所示。训练时,正样本对分别输入学生和教师网络,然后最小化两个网络输出分布的交叉熵。优化时,学生网络通过梯度反向传播来更新,而教师网络采用滑动平均的方式由学生网络得到。DINO方法在说话人确认任务中的表现优于之前的对比式自监督学习方法[48-49],但距离有监督方法的性能仍有较大的差距。为此,研究者基于DINO方法提出相应的改进策略[59-60]。其中,由DINO方法训练得到的初始模型用于提取每句语音的说话人特征表示,然后使用聚类策略来为每句语音分配一个伪标签。根据伪标签,就可以采用有监督学习的方式在初始模型上继续训练以得到更好的模型。上述生成伪标签并继续训练的方式可以一直进行下去,直到模型的性能不再进一步提升。在这个迭代训练的过程中,伪标签的质量决定了最终模型的性能。

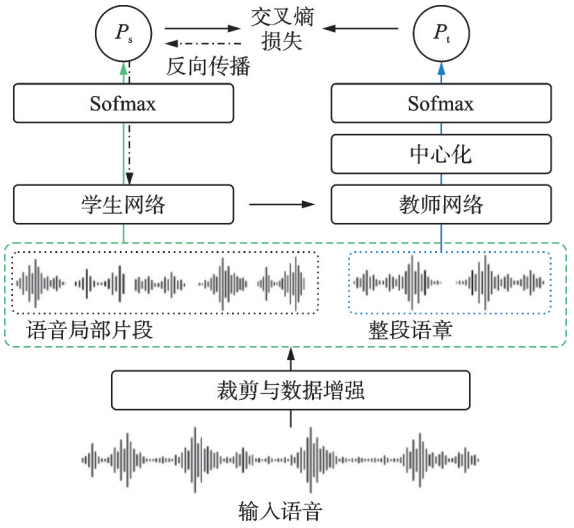

图3 基于DINO的自监督说话人确认方法

Fig.3 DINO-based self-supervised speaker verification

因此,除了从自监督说话人特征表示模型中获取伪标签,研究者还从通用语音预训练模型中获取伪标签[61]。此外,采用额外的视觉信息来提取更好的伪标签[62],以及在训练过程中过滤掉错误的伪标签[63]都取得了进一步的性能提升。

近年来,大规模语音预训练模型[64-66]受到了广泛关注,研究者首先在大规模无标注数据上预训练模型,然后将其应用于不同的语音下游任务。基于语音预训练模型的说话人确认方法如图4所示。早期工作通过在Wav2vec 2.0[64]等语音预训练模型的后端添加简单的池化层和线性层,并在下游说话人数据集上微调模型,证明了Wav2vec 2.0可以应用于说话人确认任务[67-69]。由于语音中的说话人信息主要为短时信息,而基于Transformer架构的语音预训练模型主要在挖掘全局的长时信息,因此研究者在预训练模型后端添加一些卷积层以学习短时信息[70-71]。随着卷积网络的不断加深,目前在语音预训练模型后端使用ECAPA-TDNN网络已成为主流[72]。此外,相关研究表明,说话人信息主要包含在语音预训练模型的浅层,因此将下游ECAPA-TDNN的输入由预训练模型的最后一层替换成所有层的加权和,可以进一步提升说话人确认的性能[73]。除了语音预训练模型常用的Transformer网络,研究者发现语音识别任务中的Conformer预训练模型同样可以应用于说话人确认任务[74]。这表明语音识别和说话人识别任务可以在一定程度上相互促进。此外,Conformer模型中的卷积层也有利于挖掘语音中的短时说话人信息。

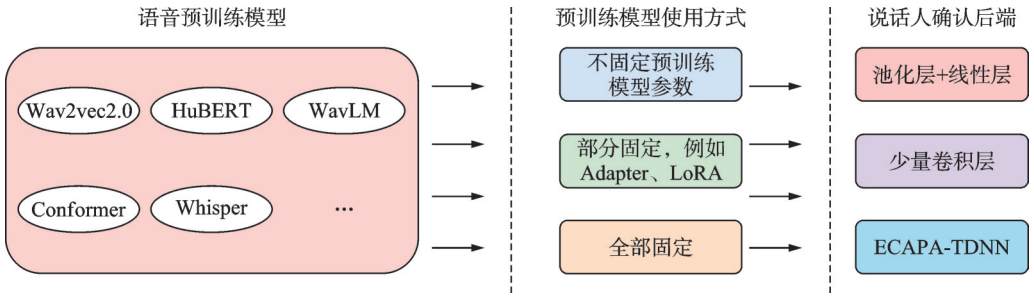

图4 基于语音预训练模型的说话人确认方法

Fig.4 Speech pre-trained model-based speaker verification

2 基于深度学习的跨域说话人确认方法

说话人特征表示模型易受跨域不匹配问题的干扰,所谓领域是指一堆数据和生成这些数据的概率分布[75]。在跨域问题中涉及两个基本的领域:源域(Source domain)和目标域(Target domain),源域指的是拥有大量有标注数据的领域,通常在源域上训练说话人特征表示模型;目标域就是指测试集所在的领域。当源域和目标域测试集出现信道、远场、语种和噪声等不匹配情况时,性能会急剧下降,为此研究者提出各种方法来解决跨域问题。根据目标域是否有训练数据,可将跨域问题分为领域自适应和领域泛化。下面分别介绍说话人确认中这两类问题的研究现状。

2.1 领域自适应

领域自适应(Domain adaptation,DA)是指根据目标域的训练数据将在源域预训练好的说话人特征表示模型自适应到目标域[76],其中目标域训练数据没有标注时的场景是目前研究的难点和热点,即无监督领域自适应(Unsupervised DA,UDA)。与传统的无监督领域自适应问题不同的是,说话人确认中的领域自适应是一个开集问题,即源域和目标域中的说话人类别完全不同。这意味着源域和目标域之间说话人特征表示空间的差异不仅来自领域不匹配,还来自说话人的差异。因此增加了解决说话人确

认中领域自适应问题的难度。目前,主流的无监督领域自适应方法可分为3类:领域对齐(Domain alignment)、伪标签微调(Pseudo-label fine-tuning),以及领域转换(Domain transformation)。这些方法分别从不同的角度以及技术路线出发以提升模型在目标域上的性能,本节将对其进行详细的介绍。

2.1.1 基于领域对齐的领域自适应方法

领域对齐通过最小化源域和目标域说话人特征表示分布之间的差异,以学习一个源域和目标域分布一致的公共说话人特征表示空间。其损失函数可以表示为

$$\mathcal{L}(\mathcal{D}_{\mathrm{s}},\mathcal{D}_{\mathrm{t}})=\mathcal{L}_{\mathrm{div}}(\mathcal{D}_{\mathrm{s}},\mathcal{D}_{\mathrm{t}})+\mathcal{L}_{\mathrm{sup}}(\mathcal{D}_{\mathrm{s}}) \tag{20}$$

式中:$\mathcal{D}_{\mathrm{s}}$和$\mathcal{D}_{\mathrm{t}}$分别表示源域和目标域;$\mathcal{L}_{\mathrm{div}}$表示源域和目标域之间的分布差异度量损失函数;$\mathcal{L}_{\mathrm{sup}}$表示源域上的有监督损失函数,例如AM-Softmax。为度量源域和目标域之间的分布差异,Ganin等[77]提出领域对抗神经网络(Domain-adversarial neural network, DANN)方法,通过一个判别器(Discriminator)网络来度量源域和目标域之间的JS散度,而编码器与判别器则通过梯度反转层(Gradient reversal layer)以对抗的形式来共同训练。DANN建立了一个通用的框架来处理领域自适应问题,并使用梯度下降法进行训练。随后,Wang等[78]将DANN应用到i-vector特征表示空间以学习一个信道不变(Channel invariant)的说话人编码网络。此外,Luu等[79]又将DANN应用到x-vector中,以孪生网络的方式学习更具区分性的信道不变说话人特征表示。除了梯度反转层,也有研究者使用了生成对抗网络(Generative adversarial network, GNN)[80]中的反转标签损失(Inverted-label loss)来学习DANN,可以为编码器网络提供更强的梯度信息[81]。由于在说话人确认后端打分过程中,常用的PLDA模型通常假设说话人特征表示为高斯分布,所以研究者们在DANN中加入了变分自编码器(Variational autoencoder, VAE)分支,以约束学到的说话人特征表示符合高斯分布[82-83]。

除了DANN,通过优化最大均值差异(Maximum mean discrepancy, MMD)[84]损失来最小化源域和目标域特征表示分布的距离也是一种常用的方式,其避免了DANN训练困难的缺点。MMD的核心思想是通过再生核希尔伯特空间(Reproducing kernel Hilbert space, RKHS)中的均值嵌入来度量两个分布的差异。在实践中,MMD的计算通过核函数来简化,以避免直接在高维空间中操作的复杂性。具体来说,若给定RBF核函数$k(x,y)$,则MMD可以表示为

$$\mathcal{L}_{\mathrm{div}}(\mathcal{D}_{\mathrm{s}},\mathcal{D}_{\mathrm{t}})=\mathbb{E}_{x,x'\sim\mathcal{D}_{\mathrm{s}}}\left[k(x,x')\right]+\mathbb{E}_{y,y'\sim\mathcal{D}_{\mathrm{t}}}\left[k\left(y,y'\right)\right]-2\mathbb{E}_{x\sim\mathcal{D}_{\mathrm{s}},y\sim\mathcal{D}_{\mathrm{t}}}\left[k(x,y)\right] \tag{21}$$

式中x和y分别来自源域$\mathcal{D}_{\mathrm{s}}$和目标域$\mathcal{D}_{\mathrm{t}}$。在说话人确认中,Lin等[85]首次将MMD引入自编码器(Auto-encoders)的损失函数中,以学习跨语种不变(Language-invariant)的说话人特征表示。此外,还考虑了源域数据可能是由不同子域构成的多源域领域自适应问题,因此进一步提升了方法的性能[86]。在后续改进中,又提出使用MMD同时约束帧级和段级特征表示的分布,使网络对训练和测试数据之间的领域差异更加鲁棒[87]。

虽然领域对齐方法可以有效学习源域和目标域分布一致的公共说话人特征表示空间,但由于说话人确认中源域和目标域的说话人类别完全不同,因此直接对齐源域和目标域会导致不同说话人特征表示的混合,从而降低其区分说话人的能力。为解决该问题,Li等[88-89]提出在距离空间中对齐源域和目标域,其中距离空间中的每一个点都由成对说话人特征表示计算其余弦距离得到。由于源域和目标域在距离空间中都只有两个类别,即相同说话人和不同说话人,因此其分布差异完全来自领域不匹配,通过在距离空间中优化MMD损失即可完成领域自适应。具体框架如图5所示,其中s^{ws}和s^{bs}分别表示源域中相同说话人(Within-speaker)和不同说话人(Between-speaker)的距离样本,t^{ws}和t^{bs}分别表示目标域中相同说话人和不同说话人的距离样本。除此之外,Hu等[90]直接对齐源域和目标域距离分布的一阶和二阶统计量,同样证明了距离分布自适应方法的有效性。

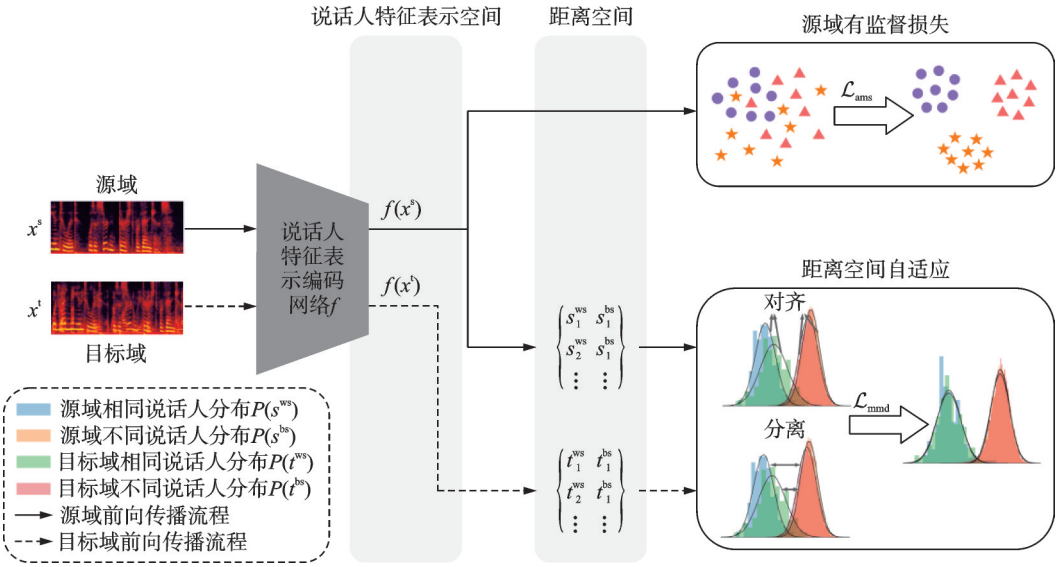

图5 基于距离分布对齐的领域自适应方法

Fig.5 Distance distribution-based domain adaptation method

2.1.2 基于伪标签微调的领域自适应方法

若目标域数据有说话人标签，那么可直接在目标域上微调源域训好的说话人特征表示模型。类似的，对于无监督领域自适应，可通过生成目标域伪标签并微调说话人特征表示模型的方式来完成领域自适应。伪标签微调方法在近几年的说话人确认挑战赛中占据主导地位。例如，在VoxSRC[91]比赛中，大部分参赛队伍都采用了伪标签微调的方法来解决领域自适应问题。为生成目标域伪标签，需使用源域训好的说话人模型提取目标域所有数据的说话人特征表示，并基于特定的聚类方法为每个目标域样本分配唯一的伪标签。在得到伪标签后，目标域微调的损失函数为

$$\mathcal{L}(\mathcal{D}_s,\mathcal{D}_t)=\mathcal{L}_{sup}(\mathcal{D}_t)+\mathcal{L}_{sup}(\mathcal{D}_s) \tag{22}$$

式中$\mathcal{L}_{sup}(\mathcal{D}_s)$和$\mathcal{L}_{sup}(\mathcal{D}_t)$分别表示源域和目标域上的有监督损失函数，例如AM-Softmax。在伪标签微调方法中，伪标签的质量决定着微调后说话人特征表示模型的性能。因此，需使用一个好的聚类方法以产生高质量的目标域伪标签。在众多聚类方法中，由于K-均值(K-means)[92-96]和层次聚类方法(Agglomerative hierarchical clustering，AHC)[59,97]的简单性和有效性，因此其成为最常使用的说话人聚类方法。然而，K-means和AHC只是简单地利用了样本之间的距离关系，没有充分考虑样本的局部结构信息，因此聚类性能有限。为此，Chen等[61]使用Infomap聚类方法来引入样本的局部结构信息；Li等[98]提出了一种渐进子图聚类方法，根据样本的k-近邻信息以产生目标域的伪标签。此外，Li等[89]提出利用图卷积网络(Graph convolutional network，GCN)[99]来挖掘样本的局部结构信息，通过在源域上训练基于GCN的聚类模型，可以在目标域上生成高质量伪标签。

由于伪标签微调方法可充分利用目标域中潜在的样本级标签信息，因此取得了很好的结果。此外，该方法不受领域自适应中开集问题的影响。但是，该方法的性能严重依赖于伪标签的质量，而且忽略了源和目标域分布间的一致性关系。

2.1.3 基于领域转换的领域自适应方法

由于目标域的训练数据通常规模较小且无标注，因此只能从源域迁移训好的说话人特征表示模型

而无法在目标域中从头开始训练。领域转换方法的思想就是基于生成式模型学习源域与目标域数据之间的一个映射,然后在保证转换前后说话人信息一致的基础上将源域中的有标注数据转到目标域,以完成对目标域训练数据的扩充。Shon等[100]基于降噪自编码器(Denoising autoencoder)将源域中的i-vector特征表示转换到目标域,并结合转换后的有标注i-vector以及目标域中的无标注i-vector来估计PLDA中的参数。Nidadavolu等[101]基于循环生成对抗网络(Cycle consistent GANs,cycle-GANs)[102]将麦克风信道的源域声学特征转换到电话信道目标域,实验表明可有效提升x-vector模型在电话信道条件下的性能。此外,由于cycle-GANs不依赖并行的训练数据,因此该方法不需要同时收集说话人在麦克风和电话信道下的数据。在后续改进中,研究者将其扩展到带有噪声和混响的场景中以提升x-vector模型对噪声和混响的鲁棒性[103-104]。此外,Su等[105]引入最新的扩散概率模型(Diffusion probabilistic model,DPM)[106],进一步提升了领域转换的性能。除了将源域数据转换到目标域并重训说话人特征表示模型,也有研究者将目标域数据转换到源域,并利用源域已有的说话人特征表示模型完成目标域上的说话人确认。由于不涉及说话人特征表示模型的微调或训练,因此有效提升了领域自适应的效率[107]。

虽然领域转换方法可以有效解决目标域缺乏训练数据的问题,但其性能严重依赖于转换后数据或特征的质量。此外,还需保证转换前后说话人信息的一致性,但目前的方法都没有明确考虑这一点。

2.1.4 分析对比

上述3种领域自适应方法分别从不同的角度出发来解决领域自适应问题,其各有优缺点。表2详细对比了这3种方法。在实际使用中,可根据具体的应用场景和需求来选择合适的领域自适应方法。也可以结合多种方法来提升模型的性能,例如将领域对齐和伪标签微调方法结合,以充分利用目标域中的有标注数据以及源域和目标域之间的一致性关系[89]。

表2 基于深度学习的说话人领域自适应方法对比

Table 2 Comparison of deep learning-based speaker domain adaptation methods

方法	基本原理	优点	缺点
领域对齐	对齐源和目标域之间的说话人特征表示分布或距离分布	不依赖特定模型结构,通用性强,对有标注目标域数据需求少	对齐说话人特征表示分布不适用于开集领域自适应,对抗训练计算复杂
伪标签微调	生成目标域数据的伪标签并微调源域训好的模型	能充分利用目标域潜在的标签信息,适用于开集领域自适应	易受伪标签质量的影响,忽略了源和目标域间的一致性关系
领域转换	将源域有标注数据转换到目标域以扩充目标域训练集	可有效扩充目标域训练数据,适用于开集领域自适应	依赖于生成式模型,转换可能会丢失说话人信息

2.1.5 其他方法

除了上述方法之外,还有一些方法从其他角度出发来解决领域自适应问题。例如,Wang等[108]提出在说话人特征表示模型中插入基于挤压激励(Squeeze-and-excitation,SE)和批归一化(Batch normalization, BN)的适配器(Adapter)模块,并在自适应过程中固定说话人特征表示模型参数,且只更新适配器模块的参数,以此实现快速且轻量化的领域自适应。Cumani等[109]参考后端得分校正中的AS-norm方法对说话人特征表示进行归一化,可以在不增加大量计算成本的前提下提升在目标域上的性能。

2.2 领域泛化

现有基于领域自适应的跨域说话人确认方法仍需要目标域的训练数据,且只能在单一目标域下得到比较好的性能。然而在实际应用中可能无法事先得知测试数据的所属领域并收集一定的训练数据。

为此，研究者们提出领域泛化(Domain generalization)[110]方法，以得到无需自适应也能在未知目标域中有良好泛化能力的模型。领域泛化方法的基本假设是训练数据中包含多个源域的有标注训练数据，但没有目标域的训练数据。若将语音中所包含的信息分为说话人信息以及与说话人无关的领域信息，为使说话人特征表示模型在未知目标域上有良好的泛化能力，就需要在模型训练过程中防止其编码与说话人无关的领域信息，从而只编码说话人信息。现有的说话人领域泛化方法分别从不同的角度出发来防止说话人模型编码领域信息，总体可分为：语音增强(Speech enhancement)、域不变表示学习(Domain-invariant representation learning)、特征解耦(Feature disentanglement)。

2.2.1 基于语音增强的领域泛化方法

基于语音增强的方法针对的是噪声鲁棒性这一类领域不匹配问题，数据集中不同类型的带噪语音即为不同领域的数据。如图6所示，语音增强模块在使用时位于说话人特征表示模型前端，其作用是将输入的带噪语音转换成干净语音，使说话人特征表示模型在各种噪声环境下都能得到干净的输入。

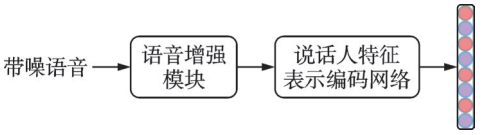

图6 基于语音增强的领域泛化方法

Fig.6 Speech enhancement-based domain generalization

在众多方法中，基于掩蔽(Mask-based)的语音增强技术受到很多关注，其在语音质量和语音可懂度方面表现出很好的性能。该方法使用DNN来估计带噪语音的时频掩码，并使用掩码恢复相应的干净语音。早期工作使用与说话人确认模型相互独立的语音增强模块进行降噪[111-113]。文献[114-115]进一步提出将基于掩蔽的语音增强模块与说话人确认模型联合优化，以得到更适合于说话人确认模型的干净数据。

基于编码-解码架构的语音增强方法是另外一种常用的技术，它能将带噪语音直接映射为对应的干净语音。例如，Plchot等[116]训练了一个基于DNN的自编码器，通过优化最小化均方误差损失，将带噪语音的对数幅度频谱映射到对应的干净频谱，并通过实验证明了该方法在i-vector特征表示模型中的有效性。Novotny等[117-118]探索了与文献[116]类似的增强方法，并将其应用于x-vector模型。除了单独增强幅度谱，OO等[119]还增强相位谱，并使用修改后的群延迟阶谱系数(Modified group delay cepstral coefficients)作为相位谱特征，实验表明同时增强幅度谱和相位谱可以带来更大的性能提升。此外，Gao等[120]引入了基于UNet架构的语音增强方法，并在语音增强模块和说话人特征表示模型之间插入一个DenseNet模块，以防止语音增强模块过度增强带噪语音从而丢失说话人信息。Kim等[121]进一步提出ExU-Net架构来提升基于UNet语音增强方法的性能，并使用说话人度量损失和语音增强损失来联合训练说话人特征表示模型和语音增强模块。Ma等[122]设计了一种梯度加权机制，以减少基于UNet语音增强模块所产生的人工噪声。

此外，也有研究者使用生成对抗网络[80]和扩散模型[106]等生成式模型来得到增强后的干净语音。例如，Michelsanti等[123]使用条件GAN(Conditional GAN)学习从带噪频谱到干净频谱的映射。条件GAN由1个生成器和1个判别器组成，并以对抗方式进行训练。生成器用来增强带噪频谱，判别器则以带噪频谱为条件，学习区分增强后频谱与对应的干净频谱。Yu等[124]同样基于对抗训练来获得噪声鲁棒的特征表示。此外，Dowerah等[125]提出了基于扩散概率模型的两阶段语音增强方法，在第一阶段单独训练语音增强模块，在第二阶段使用自监督学习联合训练ECAPA-TDNN模型和语音增强模块。Kim等[126]除了使用扩散概率模型，还设计了一个辅助增强模块来进一步去除噪声。

2.2.2 基于域不变表示学习的领域泛化方法

文献[76]从理论上证明,如果特征表示对不同领域保持不变,那么这些特征表示就是通用的,可以迁移到其他领域。基于这一理论,人们提出了基于域不变表示学习的领域泛化方法,通过减小说话人特征表示在多个源域之间的差异,使其在未知目标域上有良好的泛化能力。值得注意的是,该方法需要收集相同说话人在各个源域上的数据,否则无法最小化不同源域间的差异。实现域不变表示学习的一种常用方式是通过领域对抗训练来对齐不同源域间的说话人特征表示分布。例如,Zhou等[127]使用基于噪声类型分类的Softmax损失作为判别器的损失函数,来学习对噪声类型鲁棒的说话人特征表示,Meng等[128]使用基于噪声类型和信噪比分类的Softmax损失作为判别器的损失函数以学习鲁棒的说话人特征表示。除了领域对抗训练,Kang等[129]从信息论的角度出发,通过最小化说话人特征表示与领域标签的互信息,来去除说话人特征表示中的领域信息;Huang等[130]在说话人特征表示模型中插入轻量级的领域适配器(Domain adapter),并使用领域适配器来融合同一说话人在不同源域中的说话人特征表示,以提升在未知目标域上的泛化能力;Cai等[131]通过最小化同一说话人在不同噪声环境下说话人特征表示之间的余弦距离,来学习噪声鲁棒的说话人特征表示;Li等[132]通过对齐同一说话人不同源域样本上的梯度,来学习领域不变的说话人特征表示。

元学习(Meta-learning)[133]是一种通过“学习如何学习”来提升模型泛化能力的技术,它在解决领域泛化问题时非常有用。元学习的核心思想是从多个小任务中学习能在新任务上快速适应的模型。在领域泛化中,源域可以看作是不同的小任务,而目标域是一个新的、未见过的任务。通过在多个源域上使用元学习训练说话人特征表示模型,就能得到域不变的说话人特征表示。例如,Kang等[134]提出了一种基于模型无关元学习(Model-agnostic meta-learning, MAML)的领域泛化方法,可以学得在未知目标域上有良好泛化能力的说话人特征表示模型。Zhang等[135]提出了一种元表示学习(Meta representation learning)方法,通过构造多个小任务训练的方式来提高领域泛化能力。此外,Yang等[136]设计了一个基于特定领域网络和领域聚合网络的元学习框架,以学习领域不变的说话人特征表示。

2.2.3 基于特征解耦的领域泛化方法

特征解耦(Feature disentanglement)是另一种常用的领域泛化方法,其核心思想是通过一个说话人编码网络和一个领域编码网络分别学习说话人相关特征表示和领域相关特征表示,并约束两个特征表示相互独立以使说话人特征表示中不含有领域信息。例如,Tong等[137]通过一个注意力模块将原始特征表示解耦为说话人相关特征表示和说话速率相关特征表示,并最小化两个特征表示的余弦相似度;Qin等[138]将说话人特征表示和年龄相关特征表示进行线性解耦,并通过对抗训练去除说话人特征表示中的年龄信息;Nam等[139]通过对抗训练分别学习说话人相关特征表示和语种相关特征表示,并约束两种特征表示的皮尔逊相关(Pearson's correlation)系数来去除说话人特征表示中的语种信息;Mun等[140]通过最小化说话人特征表示和领域特征表示的互信息来约束两种特征表示相互独立,其中领域编码网络通过领域分类损失来学习;Li等[141]用互信息最大化损失来代替领域分类损失,提出了一种基于互信息最大化和最小化的特征解耦方法,该方法无需收集相同说话人在不同源域中的数据,因此更具实用性。

2.2.4 分析对比

以上分别介绍了基于语音增强、域不变表示学习和特征解耦的领域泛化方法,其中基于语音增强的方法主要解决噪声鲁棒问题,而其他两种方法还可以用于解决噪声鲁棒之外的其他领域泛化问题。如表3所示,这3种方法各有优缺点,在实际处理噪声鲁棒问题时,可同时应用语音增强方法和另外两种方法,以分别在前端和说话人特征表示模型中提升噪声鲁棒性。在处理其他领域泛化问题时,可根据在实际数据集中的表现来选择域不变表示学习或特征解耦方法。

表3 基于深度学习的说话人领域泛化方法对比

Table 3 Comparison of deep learning-based speaker domain generalization methods

方法	基本原理	优点	缺点
语音增强	先对带噪语音进行处理,再将增强后的语音用于说话人特征表示模型	可充分利用语音增强任务中的最新成果	只适用于噪声鲁棒问题,不适用于其他类型的领域泛化问题
域不变表示学习	通过最小化不同源域间的差异,以学习对领域变化不敏感的说话人特征表示空间	有理论支撑,适用于各种领域泛化问题	领域对抗训练计算复杂,元学习需计算二阶梯度
特征解耦	将领域相关的信息从说话人特征表示中分离	可解释性强,适用于各种领域泛化问题	模型设计复杂,需额外设计领域编码器和损失函数

3 总结与展望

说话人确认作为生物识别技术的一个重要分支,其研究和应用正在得到快速发展。本文首先综述了说话人特征表示模型在网络输入与结构、池化层,以及有监督损失函数等方面的进展。此外,还介绍了基于自监督学习以及预训练模型的说话人特征表示模型,其利用大规模无标注数据来训练模型,为说话人确认技术提供了新的可能性。上述方法都是在训练与测试集同分布的假设下进行的,但说话人确认系统在实际应用中往往面临着噪声干扰、信道差异和远场识别等跨域不匹配问题。为解决该问题,研究者们提出了多种跨域说话人确认方法,包括领域自适应和领域泛化。这些方法通过将模型迁移到目标域,或者学习跨域不变的说话人特征表示,有效提升了模型在目标域上的性能。尽管目前说话人确认技术已经取得很大进展,但仍需在以下几个方面开展进一步的研究。

(1) 多模态融合:随着多模态技术的发展,研究者们开始探索多模态融合在身份验证中的应用。一般来说,融合说话人语音和人脸图像等多模态信息可显著提高身份验证的性能。未来的研究可以进一步探索如何融合说话人确认和人脸确认等技术,以实现更加准确和全面的身份验证。

(2) 反欺骗技术:近年来,针对说话人确认系统的攻击日益增多。攻击者通过录音重放、语音合成,以及语音转换等手段来欺骗说话人确认系统,从而实现非法访问。为提高说话人确认系统的安全性,研究者从声学特征选取、模型架构等角度出发,深入研究了相应的语音反欺骗技术,这在一定程度上提高了说话人确认系统的安全性[142-143]。未来的研究可以进一步探索更先进的反欺骗技术,并将其与说话人确认系统相结合。

(3) 轻量化模型设计:在实际应用中,说话人确认系统往往需要在嵌入式设备上运行,因此需要设计轻量化的模型。为提高模型的性能和效率,未来的研究可进一步在模型量化、知识蒸馏,以及轻量化网络等方面进行探索,以设计更加高效的说话人确认模型。

参考文献:

[1] 中华人民共和国信息产业部. 自动声纹识别(说话人识别)技术规范: SJ/T 11380—2008[S]. 北京: 信息技术与标准化, 2008.

[2] 郑方, 孙明俊. 中国声纹识别产业发展白皮书[EB/OL]. (2019-04-22). https://www.globalhha.com/doclib/data/upload/doc_con/5e50c8dec7d09.pdf.

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

[4] TONG F, ZHAO M, ZHOU J, et al. ASV-SUBTOOLS: Open source toolkit for automatic speaker verification[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6184-6188.

[5] DEHAK N, KENNY P J, 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.

[6] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 4052-4056.

[7] MUCKENHIRN H, MAGIMAI-DOSS M, MARCELL S. Towards directly modeling raw speech signal for speaker verification using CNNs[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 4884-4888.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, California, USA: MIT Press, 2017.

[10] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recognition for multi-speaker conversations using X-vectors [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 5796-5800.

[11] POVEY D, CHENG G, WANG Y, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 3743-3747.

[12] YU Y Q, LI W J. Densely connected time delay neural network for speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Shanghai, China: ISCA, 2020: 921-925.

[13] DESPLANQUES B, THIENPONDT J, DEMUYNCK K.ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Shanghai, China: ISCA, 2020: 3830-3834.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[15] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[16] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 1086-1090.

[17] LI N, TUO D, SU D, et al. Deep discriminative embeddings for duration robust speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 2262-2266.

[18] CAI W, CHEN J, LI M. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system [EB/OL]. (2018-04-14). https://doi.org/10.48550/arXiv.1804.05160.

[19] ZEINALI H, WANG S, SILNOVA A, et al. But system description to voxceleb speaker recognition challenge 2019 [EB/OL]. (2019-10-16). https://doi.org/10.48550/arXiv.1910.12592.

[20] ZHOU T, ZHAO Y, WU J. ResNeXt and Res2Net structures for speaker verification[C]//Proceedings of 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen, China: IEEE, 2021: 301-307.

[21] ROY M K, KESHWALA U. Res2Net based Text Independent Speaker recognition system[C]//Proceedings of 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence). Noida, India: IEEE, 2022: 612-616.

[22] YAN H, LEI Z, LIU C, et al. GMM-ResNext: Combining generative and discriminative models for speaker verification[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 11706-11710.

[23] LIU B, CHEN Z, WANG S, et al. DF-ResNet: Boosting speaker verification performance with depth-first design[C]//Proceedings of Annual Conference of International Speech Communications Association. Incheon, Korea: ISCA, 2022: 296-300.

[24] LIU T, LEE K A, WANG Q, et al. Golden Gemini is all you need: Finding the sweet spots for speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2324-2337.

[25] CHEN Y, ZHENG S, WANG H, et al. An enhanced Res2Net with local and global feature fusion for speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Dublin, Ireland: ISCA, 2023: 2228-2232.

[26] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24). https://doi.org/10.48550/arXiv.1810.04805.

[27] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of International Conference on Learning Representations. [S.l.]: [s.n.], 2020.

[28] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Shanghai, China: ISCA, 2020: 5036-5040.

[29] MARY N J M S, UMESH S, KATTA S V. S-vectors and TESA: Speaker embeddings and a speaker authenticator based on transformer encoder[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 404-413.

[30] SAFARI P, INDIA M, HERNANDO J. Self-attention encoding and pooling for speaker recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Shanghai, China: ISCA, 2020: 941-945.

[31] HAN B, CHEN Z, QIAN Y. Local information modeling with self-attention for speaker verification[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6727-6731.

[32] WANG R, AO J, ZHOU L, et al. Multi-view self-attention based transformer for speaker recognition[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6732-6736.

[33] ZHANG Y, LV Z, WU H, et al. MFA-conformer: Multi-scale feature aggregation conformer for automatic speaker verification [C]//Proceedings of Annual Conference of International Speech Communications Association. Incheon, Korea: ISCA, 2022: 306-310.

[34] CHOI J H, YANG J Y, JEOUNG Y R, et al. Improved CNN-transformer using broadcasted residual learning for text-independent speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Incheon, Korea: ISCA, 2022: 2223-2227.

[35] WANG H, LIN X, ZHANG J. A lightweight CNN-conformer model for automatic speaker verification[J]. IEEE Signal Processing Letters, 2023, 31: 56-60.

[36] SANG M, ZHAO Y, LIU G, et al. Improving transformer-based networks with locality for automatic speaker verification[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[37] YAO J, LIANG C, PENG Z, et al. Branch-ECAPA-TDNN: A parallel branch architecture to capture local and global features for speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Dublin, Ireland: ISCA, 2023: 1943-1947.

[38] SUN Y, LI C, LI B. Branchformer-based TDNN for automatic speaker verification[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 10981-10985.

[39] BHATTACHARYA G, ALAM M J, GUPTA V, et al. Deeply fused speaker embeddings for text-independent speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 3588-3592.

[40] OKABE K, KOSHINAKA T, SHINODA K. Attentive statistics pooling for deep speaker embedding[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 2252-2256.

[41] ZHU Y, KO T, SNYDER D, et al. Self-attentive speaker embeddings for text-independent speaker verification[C]//Proceedings of Annual Conference of International Speech Communications Association. Hyderabad, India: ISCA, 2018: 3573-3577.

[42] INDIA M, SAFARI P, HERNANDO J. Self multi-head attention for speaker recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Graz, Austria: ISCA, 2019: 4305-4309.

[43] WANG Z, YAO K, LI X, et al. Multi-resolution multi-head attention in deep speaker embedding[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6464-6468.

[44] LI C, MA X, JIANG B, et al. Deep speaker: An end-to-end neural speaker embedding system[EB/OL]. (2017-05-05). https://doi.org//10.48550/arXiv.1705.02304.

[45] SOHN K. Improved deep metric learning with multi-class N-pair loss objective[C]//Proceedings of Advanced Neural Information Processing Systems. Barcelona, Spain: MIT Press, 2016.

[46] WANG F, CHENG J, LIU W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[47] DENG J, GUO J, XUE N, et al. ArcFace: Additive angular margin loss for deep face recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 4685-4694.

[48] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations [C]//Proceedings of International Conference on Machine Learning. Vienna, Austria: PMLR, 2020: 1597-1607.

[49] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 9726-9735.

[50] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 9630-9640.

[51] ZBONTAR J, JING L, MISRA I, et al. Barlow twins: Self-supervised learning via redundancy reduction[EB/OL]. (2021-06-14). https://doi.org/10.48550/arXiv.2103.03230.

[52] JATI A, GEORGIOU P. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(10): 1577-1589.

[53] RAVANELLI M, BENGIO Y. Learning speaker representations with mutual information[C]//Proceedings of Annual Conference of International Speech Communications Association. Graz, Austria: ISCA, 2019: 1153-1157.

[54] ZHANG H, ZOU Y, WANG H. Contrastive self-supervised learning for text-independent speaker verification[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6713-6717.

[55] XIA W, ZHANG C, WENG C, et al. Self-supervised text-independent speaker verification using prototypical momentum contrastive learning[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6723-6727.

[56] MOHAMMADAMINI M, MATROUF D, BONASTRE J F, et al. Barlow twins self-supervised learning for robust speaker recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Incheon, Korea: ISCA, 2022: 4033-4037.

[57] CHEN Z, QIAN Y, HAN B, et al. A comprehensive study on self-supervised distillation for speaker representation learning [C]//Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE, 2023: 599-604.

[58] JUNG J W, KIM Y, HEO H S, et al. Pushing the limits of raw waveform speaker recognition[C]//Proceedings of Annual Conference of International Speech Communications Association. Incheon, Korea: ISCA, 2022: 2228-2232.

[59] THIENPONDT J, DESPLANQUES B, DEMUYNCK K. The IDLAB voxceleb speaker recognition challenge 2020 system description[EB/OL]. (2020-10-23). https://arxiv.org/10.48550/arXiv.2010.12468.

[60] CHO J, VILLALBA J, DEHAK N. The JHU submission to VoxSRC-21: Track 3[EB/OL]. (2021-09-28). https://doi.org/10.48550/arXiv.2109.13425.

[61] CHEN Z, WANG J, HU W, et al. Unsupervised speaker verification using pre-trained model and label correction[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[62] CAI D, WANG W, LI M. Incorporating visual information in audio based self-supervised speaker recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1422-1435.

[63] TAO R, LEE K A, KUMAR DAS R, et al. Self-supervised speaker recognition with loss-gated learning[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6142-6146.

[64] BAEVSKI A, ZHOU Y, MOHAMED A, et al. Wav2vec 2.0: A framework for self-supervised learning of speech representa-

tions[C]//Proceedings of Advanced Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2020: 12449-12460.

[65] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[66] CHEN S, WANG C, CHEN Z, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[67] FAN Z, LI M, ZHOU S, et al. Exploring Wav2vec 2.0 on speaker verification and language identification[EB/OL]. (2020-12-11). https://doi.org/10.48550/arXiv.2012.06185.

[68] VAESSEN N, VAN LEEUWEN D A. Fine-tuning Wav2Vec2 for speaker recognition[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7967-7971.

[69] VAESSEN N, VAN LEEUWEN D A. Training speaker recognition systems with limited data[EB/OL]. (2022-03-28). https://doi.org/10.48550/arXiv.2203.14688.

[70] STAFYLAKIS T, MOŠNER L, KAKOUROS S, et al. Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations[C]//Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE, 2023: 1136-1143.

[71] NOVOSELOV S, LAVRENTYEVA G, AVDEEVA A, et al. On the robustness of Wav2vec 2.0 based speaker recognition systems[C]//Proceedings of Annual Conference of International Speech Communications Association. Dublin, Ireland: ISCA, 2023: 3177-3181.

[72] CHEN Z, CHEN S, WU Y, et al. Large-scale self-supervised speech representation learning for automatic speaker verification [C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6147-6151.

[73] CHEN S, WU Y, WANG C, et al. Why does self-supervised learning for speech recognition benefit speaker recognition?[EB/OL]. (2022-06-27). https://doi.org/10.48550/arXiv.2204.12765.

[74] CAI D, LI M. Leveraging ASR pretrained conformers for speaker verification through transfer learning and knowledge distillation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3532-3545.

[75] 王晋东. 迁移学习简明手册[EB/OL]. (2018-04-11). https://www.labxing.com/files/lab_publications/615-1533737180-LiEa0mQe.pdf.

[76] BEN-DAVID S, BLITZER J, CRAMMER K, et al. Analysis of representations for domain adaptation[C]//Proceedings of the 2006 Conference of Advances in Neural Information Processing Systems. New Orleans, USA: MIT Press, 2007: 137-144.

[77] GANIN Y, LEMPITSKY V. Unsupervised domain adaptation by backpropagation[C]//Proceedings of International Conference in Machine Learning. Lille, France: PRML, 2015: 1180-1189.

[78] WANG Q, RAO W, SUN S, et al. Unsupervised domain adaptation via domain adversarial training for speaker recognition [C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018: 4889-4893.

[79] LUU C, BELL P, RENALS S. Channel adversarial training for speaker verification and diarization[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7094-7098.

[80] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Advanced Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014: 2672-2680.

[81] BHATTACHARYA G, MONTEIRO J, ALAM J, et al. Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6226-6230.

[82] TU Y, MAK M W, CHIEN J T. Variational domain adversarial learning for speaker verification[C]//Proceedings of Annual Conference of Information Speech Communications Association. Graz, Austria: ISCA, 2019: 4315-4319.

[83] TU Y, MAK M W, CHIEN J T. Information maximized variational domain adversarial learning for speaker verification[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain:

IEEE, 2020: 6449-6453.

[84] GRETTON A, BORGWARDT K M, RASCH M, et al. A kernel method for the two-sample-problem[C]//Proceedings of the 2006 Conference of Advances in Neural Information Processing Systems. New Orleans, USA: The MIT Press, 2007: 513-520.

[85] LIN W, MAK M W, LI L, et al. Reducing domain mismatch by maximum mean discrepancy based autoencoders[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2018). Lincoln, USA: ISCA, 2018: 162-167.

[86] LIN W W, MAK M W, CHIEN J T. Multisource I-vectors domain adaptation using maximum mean discrepancy based autoencoders[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2412-2422.

[87] LIN W, MAK M M, LI N, et al. Multi-level deep neural network adaptation for speaker verification using MMD and consistency regularization[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6839-6843.

[88] LI J, HAN J, SONG H. CDMA: Cross-domain distance metric adaptation for speaker verification[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7197-7201.

[89] LI J, HAN J, QIAN F, et al. Distance metric-based open-set domain adaptation for speaker verification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2378-2390.

[90] HU H R, SONG Y, DAI L R, et al. Class-aware distribution alignment based unsupervised domain adaptation for speaker verification[C]//Proceedings of Annual Conference of Information Speech Communications Association. Incheon, Korea: ISCA, 2022: 3689-3693.

[91] HUH J, BROWN A, JUNG J W, et al. VoxSRC 2022: The fourth voxceleb speaker recognition challenge[EB/OL]. (2023-03-06). https://doi.org/10.48550/arXiv.2302.10248.

[92] QIN X, CAI D, LI M. Robust multi-channel far-field speaker verification under different in-domain data availability scenarios [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 71-85.

[93] ZHANG L, LI Y, WANG N, et al. NPU-HC speaker verification system for far-field speaker verification challenge 2022[C]//Proceedings of Conference of the International Speech Communication Association. [S.l.]: ISCA, 2022.

[94] SLAVÍČEK J, SWART A, KLČO M, et al. The phonexia voxceleb speaker recognition challenge 2021 system description [EB/OL]. (2021-09-08). https://doi.org/10.48550/arXiv.2109.02052.

[95] QIN X, LI N, LIN Y, et al. The DKU-tencent system for the voxceleb speaker recognition challenge 2022[EB/OL]. (2022-10-11). https://doi.org/10.48550/arXiv.2210.05092.

[96] CAI D, WANG W, LI M. An iterative framework for self-supervised deep speaker representation learning[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE, 2021: 6728-6732.

[97] MCCREE A, SHUM S, REYNOLDS D, et al. Unsupervised clustering approaches for domain adaptation in speaker recognition systems[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2014). Joensuu, Finland: ISCA, 2014: 265-272.

[98] LI Z, LU J, ZHAO Z, et al. Progressive sub-graph clustering algorithm for semi-supervised domain adaptation speaker verification[EB/OL]. (2023-05-22). https://doi.org/10.48550/arXiv.2305.12703.

[99] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-02). https://doi.org/10.48550/arXiv.1609.02907.

[100] SHON S, MUN S, KIM W, et al. Autoencoder based domain adaptation for speaker recognition under insufficient channel information[C]//Proceedings of Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017: 1014-1018.

[101] NIDADAVOLU P S, VILLALBA J, DEHAK N. Cycle-GANs for domain adaptation of acoustic features for speaker recognition[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6206-6210.

[102] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2242-2251.

[103] NIDADAVOLU P S, KATARIA S, VILLALBA J, et al. Low-resource domain adaptation for speaker recognition using cycle-GANs[C]//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore: IEEE, 2019: 710-717.

[104] NIDADAVOLU P S, KATARIA S, VILLALBA J, et al. Unsupervised feature enhancement for speaker verification[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7599-7603.

[105] SU X, XIE X, ZHANG F, et al. Adversarial diffusion probability model for cross-domain speaker verification integrating contrastive loss[C]//Proceedings of Conference of the International Speech Communication Association. Dublin, Ireland: ISCA, 2023: 5336-5340.

[106] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015: 2256-2265.

[107] LI J, LIU W, LEE T. EDITNet: A lightweight network for unsupervised domain adaptation in speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Incheon, Korea: ISCA, 2022: 3694-3698.

[108] WANG T, LI L, WANG D. Se/Bn adapter: Parametric efficient domain adaptation for speaker recognition[EB/OL]. (2024-06-12). https://doi.org/10.48550/arXiv.2406.07832.

[109] CUMANI S, SARNI S. From adaptive score normalization to adaptive data normalization for speaker verification systems[C]//Proceedings of Conference of the International Speech Communication Association. Dublin, Ireland: ISCA, 2023: 5296-5300.

[110] WANG J, LAN C, LIU C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 8052-8072.

[111] ZHAO X, WANG Y, WANG D. Robust speaker identification in noisy and reverberant conditions[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 3997-4001.

[112] KOLBŒK M, TAN Z H, JENSEN J. Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification[C]//Proceedings of 2016 IEEE Spoken Language Technology Workshop (SLT). San Diego, USA: IEEE, 2016: 305-311.

[113] CHANG J, WANG D. Robust speaker recognition based on DNN/i-vectors and speech separation[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, USA: IEEE, 2017: 5415-5419.

[114] ZHAO F, LI H, ZHANG X. A robust text-independent speaker verification method based on speech separation and deep speaker[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6101-6105.

[115] SHON S, TANG H, GLASS J. VoiceID loss: Speech enhancement for speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 2888-2892.

[116] PLCHOT O, BURGET L, ARONOWITZ H, et al. Audio enhancing with DNN autoencoder for speaker recognition[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 5090-5094.

[117] NOVOTNY O, PLCHOT O, MATEJKA P, et al. On the use of DNN autoencoder for robust speaker recognition[EB/OL]. (2018-11-07). https://doi.org/10.48550/arXiv.1811.02938.

[118] NOVOTNY O, PLCHOT O, GLEMBEK O, et al. Analysis of DNN speech signal enhancement for robust speaker recognition[J]. Computer Speech & Language, 2019, 58: 403-421.

[119] OO Z, KAWAKAMI Y, WANG L, et al. DNN-based amplitude and phase feature enhancement for noise robust speaker identification[C]//Proceedings of Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016: 2204-2208.

[120] GAO Z, MAK M, LIN W. UNet-DenseNet for robust far-field speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Incheon, Korea: ISCA, 2022: 3714-3718.

[121] KIM J H, HEO J, SHIM H J, et al. Extended U-Net for speaker verification in noisy environments[C]//Proceedings of Conference of the International Speech Communication Association. Incheon, Korea: ISCA, 2022: 590-594.

[122] MA Y, LEE K A, HAUTAMÄKI V, et al. Gradient weighting for speaker verification in extremely low signal-to-noise ratio [C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 11311-11315.

[123] MICHELSANTI D, TAN Z H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017: 2008-2012.

[124] YU H, TAN Z H, MA Z, et al. Adversarial network bottleneck features for noise robust speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017: 1492-1496.

[125] DOWERAH S, KULKARNI A, SERIZEL R, et al. Self-supervised learning with diffusion-based multichannel speech enhancement for speaker verification under noisy conditions[C]//Proceedings of Conference of the International Speech Communication Association. Dublin, Ireland: ISCA, 2023: 3849-3853.

[126] KIM J H, HEO J, SHIN H S, et al. Diff-SV: A unified hierarchical framework for noise-robust speaker verification using score-based diffusion probabilistic models[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 10341-10345.

[127] ZHOU J, JIANG T, LI L, et al. Training multi-task adversarial network for extracting noise-robust speaker embedding[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6196-6200.

[128] MENG Z, ZHAO Y, LI J, et al. Adversarial speaker verification[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6216-6220.

[129] KANG W, ALAM M J, FATHAN A. MIM-DG: Mutual information minimization-based domain generalization for speaker verification[C]//Proceedings of Conference of the International Speech Communication Association. Incheon, Korea: ISCA, 2022: 3674-3678.

[130] HUANG W, HAN B, WANG S, et al. Robust cross-domain speaker verification with multi-level domain adapters[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 11781-11785.

[131] CAI D, CAI W, LI M. Within-sample variability-invariant loss for robust speaker recognition under noisy environments[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6469-6473.

[132] LI J, HAN J, SONG H. Gradient regularization for noise-robust speaker verification[C]//Proceedings of Conference of the International Speech Communications Association. Brno, Czech Republic: ISCA, 2021: 1074-1078.

[133] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of International Conference on Machine Learning. Sydney, Australia: PRML, 2017: 1126-1135.

[134] KANG J, LIU R, LI L, et al. Domain-invariant speaker vector projection by model-agnostic meta-learning[C]//Proceedings of Conference of the International Speech Communications Association. Shanghai, China: ISCA, 2020: 3825-3829.

[135] ZHANG J T, SONG Y, LI J, et al. Meta representation learning method for robust speaker verification in unseen domains [C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024: 11301-11305.

[136] YANG S, DAS D, CHO J, et al. Domain agnostic few-shot learning for speaker verification[C]//Proceedings of Conference of the International Speech Communications Association. Incheon, Korea: ISCA, 2022: 595-599.

[137] TONG F, ZHENG S, ZHOU H, et al. Deep representation decomposition for rate-invariant speaker verification[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2022). Beijing, China: ISCA, 2022: 228-232.

[138] QIN X, LI N, CHAO W, et al. Cross-age speaker verification: Learning age-invariant speaker embeddings[C]//Proceedings of Conference of the International Speech Communications Association. Incheon, Korea: ISCA, 2022: 1436-1440.

[139] NAM K, KIM Y, HUH J, et al. Disentangled representation learning for multilingual speaker recognition[C]//Proceedings of

Conference of the International Speech Communications Association. Dublin, Ireland: ISCA, 2023: 5316-5320.

[140] MUN S H, HAN M H, KIM M, et al. Disentangled speaker representation learning via mutual information minimization[C]// Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Chiang Mai, Thailand: IEEE, 2022: 89-96.

[141] LI J, HAN J, DENG S, et al. Mutual information-based embedding decoupling for generalizable speaker verification[C]//Proceedings of Conference of the International Speech Communications Association. Dublin, Ireland: ISCA, 2023: 3147-3151.

[142] 张雄伟, 张星昱, 孙蒙, 等. 说话人验证系统攻击方法的研究现状及展望[J]. 数据采集与处理, 2021, 36(5): 831-849.
ZHANG Xiongwei, ZHANG Xingyu, SUN Meng, et al. Attack methods in speaker verification system: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2021, 36(5): 831-849.

[143] 张雄伟, 李嘉康, 孙蒙, 等. 语音欺骗检测方法的研究现状及展望[J]. 数据采集与处理, 2020, 35(5): 807-823.
ZHANG Xiongwei, LI Jiakang, SUN Meng, et al. Speech anti-spoofing: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 807-823.

作者简介:

李建琛(1996-),男,博士研究生,研究方向:说话人识别和领域不匹配,E-mail:lijianchen@hit.edu.cn。

韩纪庆(1964-),**通信作者**,男,教授,研究方向:智能语音处理和机器学习,E-mail:jqhan@hit.edu.cn。

(编辑:王静)