基于互信息自适应估计的说话人确认方法

陈 晨[1,2*]，季超群[1]，李文文[1]，陈德运[1,2]，王莉莉[1,2]，杨海陆[1,2]

(1. 哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080；2. 哈尔滨理工大学计算机科学与技术博士后流动站 哈尔滨 150080)

【摘要】为了更准确地度量特征间的关系，提出了一种基于互信息自适应估计的目标函数表示方法。将具有自适应特性的度量方法引入到目标函数中，该目标函数以最大化类内相似度、最小化类间相似度为目标，并能根据深层特征的真实分布情况对相似度进行动态的调整，从而使深度神经网络朝着区分性更强的方向进行优化。此自适应度量方式还被用于特征筛选，其能够根据特征的特点进行有针对性的参数更新，使得选取的特征具有典型性，提升目标函数对于深度神经网络优化方向的指导能力。实验结果表明，相比于其他深度神经网络方法，该方法的相对等错误率最多降低了 28%，显著提升了说话人确认系统的性能。

关 键 词 互信息估计； 目标函数； 自适应学习； 特征表示学习； 说话人确认

中图分类号 TP391.4 **文献标志码** A doi:10.12178/1001-0548.2022174

Mutual Information Adaptive Estimation for Speaker Verification

CHEN Chen[1,2*], JI Chaoqun[1], LI Wenwen[1], CHEN Deyun[1,2], WANG Lili[1,2], and YANG Hailu[1,2]

(1. School of Computer Science and Technology, Harbin University of Science and Technology Harbin 150080;

2. Postdoctoral Research Station of Computer Science and Technology, Harbin University of Science and Technology Harbin 150080)

Abstract In order to measure the relationship between features more accurately, an objective function representation method based on mutual information adaptive estimation is proposed for speaker verification systems. This objective function introduces an adaptive metric learning method, and the optimization objective is maximizing the intra-class similarity and minimizing the inter-class similarity. Meanwhile, the objective function can dynamically adjust the similarity according to the real distribution of deep features. Based on dynamically adjusting, the deep neural networks can be optimized towards the direction of stronger discrimination. In addition, the adaptive metric method is used for feature sampling and update the parameters according to the characteristics of the features. Thus, the feature can be more typical and beneficial to improve the supervised ability of the optimization direction of the deep neural networks. Experimental results show that, compared with other deep neural networks, the relative equal error rate of the proposed method is reduced by up to 28%, and the performance of the speaker verification system is significantly improved.

Key words mutual information estimation; objective function; representation learning; self-adaption; speaker verification

生物特征识别是一项根据人类自身的生物特性进行身份鉴别的技术。近年来随着人工智能、大数据、云计算等技术的飞速发展，生物特征识别技术正越来越广泛地应用于监控、监视、网络安全和执法等方面[1]。在众多生物特征识别技术中，说话人确认[2]技术因兼顾生物特征的生理特性与行为特性，具有更高的安全性，备受研究者的广泛关注。

随着深度学习的快速发展，深度神经网络在很多领域都取得了较好的效果。视觉几何组−中等 (visual geometry group-middle, VGG-M) 网络[3]最初应用于图像处理领域，由于其在图像处理领域的优异表现被各界关注，并应用于说话人确认任务的特征提取阶段[4]。深层残差网络 (deep residual networks, ResNet)[5]则可将浅层数据直接传递到深层网络，

收稿日期：2022－06－08；修回日期：2022－09－21

基金项目：国家自然科学基金 (62101163)；黑龙江省自然科学基金 (LH2021F029)；中国博士后科学基金 (2021M701020)；黑龙江省博士后专项 (LBH-Z20020)；黑龙江省普通高校基本科研业务费 (2020-KYYWF-0341)

作者简介：陈晨 (1990－)，女，博士，主要从事语音信号处理、音频信息分析和说话人识别等方面的研究.

*通信作者：陈晨，E-mail：chenc@hrbust.edu.cn

有利于梯度优化并加快网络的训练效率。

在目标函数方面，最初以分类为目标的目标函数最为常见[6]。这类目标函数主要围绕 softmax 损失从两个角度开展研究，一是通过增加不同类别决策边界间的距离来提升其区分能力，包括其变形角-softmax(angular softmax, A-softmax) 损失[7]、加性间隔 softmax(additive margin softmax, AM-softmax) 损失[8]、动态加性间隔 softmax(dynamic-additive margin softmax)[9]、加性角间隔 softmax(additive angular margin softmax, AAM-softmax) 损失[10] 等；二是通过正则化的形式来增加 softmax 损失的区分性，这类方法通常以加权的形式建立起正则化器与 softmax 损失的联系，使用的正则化器一般也是可独立使用的损失函数，如中心 (center) 损失[11]、环 (ring) 损失[12] 等。度量学习侧重于考虑特征间的类间与类内关系，能够帮助以分类为目标的目标函数更全面地计算特征间的相关度与区分度，是开放集度量学习问题。因此，以度量学习为目标的目标函数更适合确认任务。常见的以度量学习为目标的目标函数包括二元交叉熵损失[13]、对比 (contrastive) 损失[14]、三元组 (triplet) 损失[15]、四元组损失[16]、基于互信息 (mutual information, MI) 的目标函数[17] 等。且随着采样技术的研究与发展，仅以度量学习为优化目标的方法也能够具有理想的性能，与分类结合度量学习的方法具有相仿的效果[18]。

以度量学习为目标的目标函数能够深度挖掘同类特征和异类特征相关性，使网络朝着类内相似和类间差异的方向进行更新。度量学习在计算距离时，通常采用传统的相似度计算方式，如欧氏距离打分、余弦距离打分等。由于其不具备参数，使得在相似度计算方面存在灵活性弱、适应性差等问题。当把这些传统的相似度计算方式应用于目标函数中时，并不能对特征间复杂的非线性关系进行有效表示。针对这一问题，可以有针对性地开发度量学习方法中的自适应能力，从而使目标函数能够根据特征的特点进行动态调整，并在此目标的指引下提升网络对特征表示的区分能力。考虑到自适应性的度量方式能够根据类内和类间的特征分布进行有针对性的参数更新，使得在该度量方式下选取的特征更具有典型性，更有利于目标函数对于网络的特征表示。基于此，本文利用互信息来衡量同类特征之间的相似性信息和异类特征之间的差异性信息，并将一种能够进行自适应学习的度量方法——神经概率线性判别分析 (neural PLDA, NPLDA)[19] 引入到目标函数的表示中。经过 NPLDA 对 embedding 特征的真实情况进行动态调整后，基于互信息的目标函数能够更好地指引网络朝着类内相似化、类间差异化的方向更新。本文将此方法命名为互信息自适应估计 (mutual information adaptive estimation, MIAD)，其将最大化互信息作为神经网络的优化目标。

1 互信息自适应估计

1.1 目标函数表示

本文方法的过程示意图如图 1 所示。本文利用互信息来衡量同类、异类说话人特征所在分布之间的差异性。并利用 NPLDA 模型对特征间的相似性进行自适应表示，从而保证在每轮更新中，根据 embedding 特征的分布特性，有针对性地进行特征间的相似性表示。考虑到需要对同类与异类进行表示，本文所提出的目标函数需以度量学习为目标，并通过三元组数据进行表示，此方法的过程示意图如图 1 所示。定义由神经网络提取的 embedding 特征 $\boldsymbol{x}_a$、$\boldsymbol{x}_p$、$\boldsymbol{x}_n$分别为基准 (anchor) 样本、正例 (positive) 样本、负例 (negative) 样本，基准样本与正例样本所属的说话人类别相同，与负例样本所属的类别不同。根据上述符号定义，本文所提出的目标函数可以表示为：

$$f_{\mathrm{MIAD}}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{e}^{S_i(\boldsymbol{x}_a,\boldsymbol{x}_n)}-\frac{1}{N}\sum_{i=1}^{N}S_i(\boldsymbol{x}_a,\boldsymbol{x}_p) \tag{1}$$

式中，N表示三元组的个数；$S_i(\boldsymbol{x}_a,\boldsymbol{x}_n)$表示第$i$个三元组中$\boldsymbol{x}_a$与$\boldsymbol{x}_n$的相似度；$S_i(\boldsymbol{x}_a,\boldsymbol{x}_p)$表示第$i$个三元组中$\boldsymbol{x}_a$与$\boldsymbol{x}_p$的相似度。通过最小化$f_{\mathrm{MIAD}}$，可以使基准$\boldsymbol{x}_a$与正例$\boldsymbol{x}_p$的相似度达到最大、与负例$\boldsymbol{x}_n$的相似度达到最小，从而达到最大化类间相似度、最小化类内相似度的目标。

对于式 (1) 中的相似度$S_i(\cdot)$，简单的相似度度量方法 (如欧式距离、余弦距离等) 无法保证能准确地衡量 embedding 特征间的关系，因此需要根据特征的真实情况来对相似度进行动态调整。基于此，本文将具有验证识别代价能力的 NPLDA 引入，并将其用作相似度度量方法。其能够根据同类漏报率、异类误报率进行参数的自适应调整。NPLDA 的相似度计算方式与传统 PLDA 的对数似然比打分类似，均能够表示为：

$$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}_i+\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}_j+2\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x}_j \tag{2}$$

式中，$\boldsymbol{x}_i$、$\boldsymbol{x}_j$为进行相似度计算的 embedding 特征；$\boldsymbol{P}$、$\boldsymbol{Q}$为 NPLDA 模型的参数，它们的初始值是随机生成的 0～1 之间呈均匀分布的矩阵，能随着 embedding 特征的改变而进行动态调整。

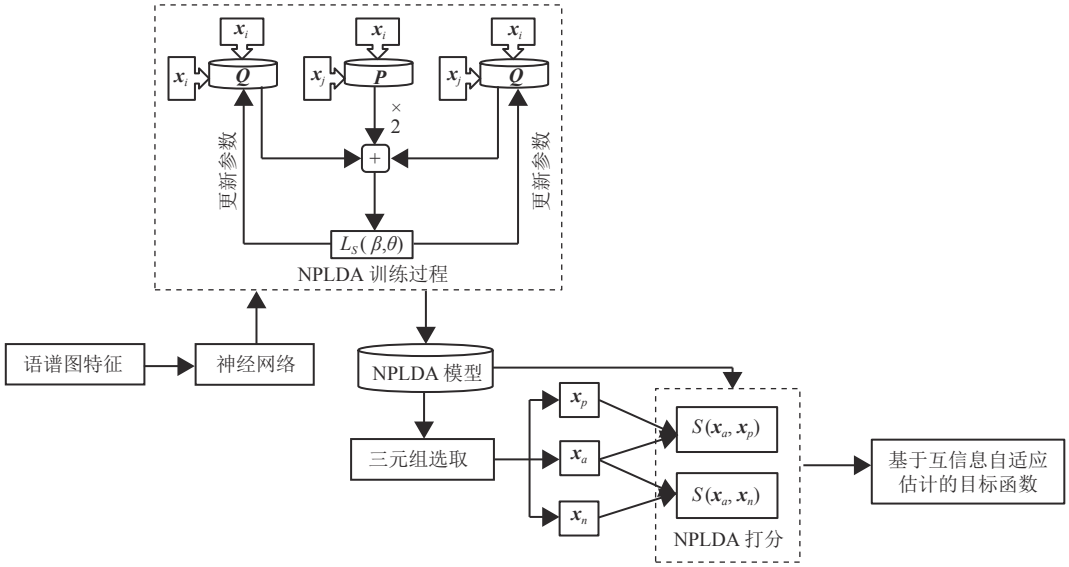


图 1　本文所提出方法的过程示意图

在 NPLDA 的训练过程中，需要对同类漏报率、异类误报率进行评价。漏报率与误报率越大，模型损失越大，因此可将最小化它们的加权和当作模型的优化目标。同时，由于漏报与误报针对的识别任务是确认任务 (即目标与非目标的二分类问题)，因此需要对 NPLDA 的训练数据进行划分，以组成以“对”为单位的样本组。针对这一问题，本文采用随机抽样生成标签的方式进行样本组的划分。基于上述描述，NPLDA 的目标函数可以表示为：

$$L_s(\beta,\theta) = P_{\text{Miss}}(\theta) + \beta P_{\text{FA}}(\theta) \tag{3}$$

式中，θ为相似度阈值，当两个 embedding 的相似度大于θ时，则判定二者为同类，反之为异类；$P_{\text{Miss}}(\theta)$、$P_{\text{FA}}(\theta)$分别对应阈值为θ时的漏报率与误报率；β为权重系数：

$$\beta = \frac{C_{\text{FA}}(1-P_{\text{prior}})}{C_{\text{Miss}}P_{\text{prior}}} \tag{4}$$

式中，C_{Miss}、C_{FA}分别为漏报与误报的代价成本，本文将其设置为 1；P_{prior}为样本组中两个 embedding 特征为同类的先验概率。

需要注意的是，简单通过训练样本中漏报与误报样本所占比率来计算$P_{\text{Miss}}(\theta)$与$P_{\text{FA}}(\theta)$，并无法保证目标函数$L_s(\beta,\theta)$连续且可导。因此，需要对$P_{\text{Miss}}(\theta)$与$P_{\text{FA}}(\theta)$进行近似表示：

$$P_{\text{Miss}}(\theta) = \frac{\sum_{i=1}^{N} t_i[1-\text{sigmoid}(\rho(S_i-\theta))]}{\sum_{i=1}^{N} t_i} \tag{5}$$

$$P_{\text{FA}}(\theta) = \frac{\sum_{i=1}^{N} (1-t_i)\text{sigmoid}(\rho(S_i-\theta))}{\sum_{i=1}^{N} (1-t_i)} \tag{6}$$

式中，S_i为第i个样本组的相似度；t_i为样本组的标签，当样本组中两个 embedding 特征为同类时，$t_i=1$，反之$t_i=0$；ρ为翘曲系数，当ρ值足够大时，$L_s(\beta,\theta)$的近似值能够逼近原始值，本文将ρ设置为 15。

1.2　三元组选取

在本文所提出的目标函数中，需要采用 NPLDA 以计算 embedding 特征的相似度，而在计算目标函数前，还需通过 embedding 特征间的相似度以选取三元组。为了统一目标函数与三元组选取时的相似度度量方法，本文在进行三元组选取时，同样采用 NPLDA 计算 embedding 特征间的相似度，以确保不同环节中相似度的一致性。

在三元组选取时，对于每个类别的 embedding 特征$\boldsymbol{x}_a$，首先均需计算其类内相似度$S(\boldsymbol{x}_a,\boldsymbol{x}_p)$与类间相似度$S(\boldsymbol{x}_a,\boldsymbol{x}_n)$。然后，再从全部备选特征中，

选择符合要求的三元组。具体而言，若当前三元组中类内相似度大于类间相似度，则该三元组中的样本为易区分样本，在筛选时应尽量减少对这类三元组的选择。为了加快网络的收敛速度，应选取类内相似度小于类间相似度的三元组，如此便可更直观地向网络传递误差信息，加快网络的收敛速度。同样地，类内相似度与类间相似度相差不大的三元组对于网络参数的更新也具有正向的促进作用，为了能够区分这一情况下的三元组，引入间隔 (Margin) 变量α，根据经验α值一般设置在 0.1～1 之间。引入间隔后的三元组选取规则如下：

$$S(\boldsymbol{x}_a,\boldsymbol{x}_n)-S(\boldsymbol{x}_a,\boldsymbol{x}_p)+\alpha<0 \tag{7}$$

待选择的三元组若不满足式 (7)，则说明当前网络不能将该三元组进行正确分类，选择该三元组进入网络中学习，使网络在后续的训练中能够对其进行正确的分类。在三元组选取时，需要有针对性地选择训练数据、构建数据组，此过程需要一定的调参经验，对于方法的复现存在少许挑战。

1.3 特征匹配

在说话人确认的测试阶段，需从网络中提取 embedding 特征用于后续的特征匹配。定义网络提取的目标说话人 embedding 特征为$\boldsymbol{x}_{\text{target}}=(y_1,y_2,\cdots,y_D)^{\mathrm{T}}$，测试说话人 embedding 特征为$\boldsymbol{x}_{\text{test}}=(b_1,b_2,\cdots,b_D)^{\mathrm{T}}$。本文采用余弦距离打分 (CDS)进行相似度计算，CDS 可表示为：

$$d_{\text{CDS}}(\boldsymbol{x}_{\text{target}},\boldsymbol{x}_{\text{test}})=\frac{\sum_{d=1}^{D}y_d b_d}{\sqrt{\sum_{d=1}^{D}y_d^2}\sqrt{\sum_{d=1}^{D}b_d^2}} \tag{8}$$

2 实验结果及分析

2.1 实验数据库和评价标准

为了验证本文方法在真实应用场景中的有效性，实验采用语音质量参差不齐的大规模说话人识别数据库 VoxCeleb1[4]。数据库中的音频均提取自 YouTube 视频网站，这些音频取自多种复杂环境，包含各类噪音。数据库的开发集包含 1 211 位说话人 (690 男，561 女) 提供的 148 642 段语音音频。评估集则包含开发集类别以外的 40 位说话人，共计 4 874 条语音。测试时采用官方测试计划列表，总测试数为 37 720 次，非目标测试与目标测试比为 1:1。评价标准采用等错误率 (equal error rate, EER) 与最小检测代价函数 (minimum detection cost function, minDCF)，其中 minDCF 的参数采用官方设置。EER 与 minDCF 的数值越低，说明性能越好。实验将从性能、收敛性及特征可视化 3 方面，对所提出方法的性能进行定量与定性的多方位对比分析。

2.2 实验性能对比与分析

本节将对比本文所提方法 (MIAD) 与其他各类方法的性能，对比的方法包括说话人确认中传统的统计模型与深度神经网络模型。其中，统计模型类方法包括高斯混合模型–通用背景模型 (Gaussian mixture model-universal background model, GMM-UBM)[20]、身份–矢量 (identity-vector, I-vector) 结合概率线性判别分析 (probabilistic linear discriminate analysis, PLDA)，简写为 I-vector+PLDA[21]。GMM-UBM 的前端声学特征分别采用梅尔倒谱系数 (mel-frequency cepstral coefficient, MFCC) 特征[2,22]、修改幂归一化倒谱系数 (modified power-normalized cepstral coefficients, MPNCC) 特征[23]、基于仿射变换与特征转换 (affine transform and feature switching, ATFS) 的特征[23]。深度神经网络模型则包括以 VGG-M、ResNet34[5] 为网络结构，并分别以对比损失、三元组损失、AM-softmax 损失为目标函数的 6 种说话人识别系统。上述 6 种方法均采用 CDS 来进行说话人匹配，分别简写为 VGG-M+Contrastive、VGG-M+Triplet、VGG-M+AM-softmax、ResNet34+Contrastive、ResNet34+Triplet、ResNet34+AM-softmax。对于上述 6 种使用 VGG-M 网络、ResNet34 网络的方法，还分别提取了 embedding 特征，并利用 NPLDA 作为后端分类器，分别简写为 VGG-M+Contrastive+NPLDA、VGG-M+Triplet+NPLDA、VGG-M+AM-softmax+NPLDA、ResNet34+Contrastive+NPLDA、ResNet34+Triplet+NPLDA、ResNet34+AM-softmax +NPLDA。此外，对比方法还包括：基于 CNN 的方法 (AutoSpeech)[24]、基于 VGG 的网络[25]、SincNet 网络[26]、基于 VGG-M+MI[17] 的方法。

上述方法的参数设置如下：在统计模型方面，MFCC 特征、MPNCC 特征、ATFS 特征的维度分别为 13 维、9 维、9 维，且上述 3 种特征均采用一阶、二阶差分。GMM-UBM 的高斯分量个数为 1 024 个，i-vector 维度为 400 维，PLDA 模型的子空间维度为 200 维。在深度神经网络模型方面，首先对输入的语音信号预加重、分帧、加窗等预处理

操作。预加重系数设置为 0.97，加窗的窗长为 25 ms，帧移为 10 ms，FFT 的点数设置为 512 个。经过以上操作后可以获得一个 512×300 维的语谱图特征。VGG-M 网络、ResNet34 网络最后一层全连接层的维度为 1 024 维，其对应的 embedding 特征亦为 1 024 维。在三元组选取时，间隔α设置为 0.3。VGG-M、ResNet34 的优化算法采用随机梯度下降 (stochastic gradient descent, SGD) 算法，初始学习率为10^{-3}，最终学习率为10^{-4}。在 MIAD 目标函数中的 NPLDA 模型则使用适应性矩估计 (adaptive moment estimation, Adam) 算法作为优化器。基于以上参数设置，不同方法的实验性能如表 1 所示。

表 1　不同方法的性能对比

模型	方法	EER/%	minDCF
统计模型	MFCC+GMM-UBM[20]	15.00	0.80
	MPNCC+GMM-UBM[23]	8.05	0.86
	ATFS+GMM-UBM[23]	7.23	0.76
	MFCC+i-vector+PLDA[21]	8.80	0.73
	AutoSpeech(N=8,C=128)[24]	8.95	-
深度神经网络	VGG[25]	7.00	0.68
	SincNet[26]	7.20	-
	VGG-M+Contrastive+NPLDA	8.37	0.82
	VGG-M+Triplet+NPLDA	7.40	0.72
	VGG-M+AM-softmax+NPLDA	7.83	0.71
	VGG-M+Contrastive	7.62	0.66
	VGG-M+Triplet	7.59	0.66
	VGG-M+AM-softmax	7.52	0.65
	VGG-M+MI[17]	**6.61**	**0.61**
	VGG-M+MIAD	**6.60**	0.62
	ResNet34+Contrastive+NPLDA	8.56	0.84
	ResNet34+Triplet+NPLDA	8.13	0.79
	ResNet34+AM-softmax+NPLDA	7.57	0.70
	ResNet34+Contrastive	7.98	0.69
	ResNet34+Triplet	7.87	0.68
	ResNet34+AM-softmax	7.34	0.62
	ResNet34+MIAD	**6.44**	**0.60**

从表中可以看出以下几点。

1) VGG-M+MIAD 方法、ResNet34+MIAD 方法的性能明显优于使用相同网络的其他方法，EER 明显降低。在相同网络结构的条件下，MIAD 能够取得优于其他目标函数的性能。

2) 相比于 VGG-M+MI，本文所提方法的 EER 虽然只有小幅度降低，但相比于其他目标函数的性能提升明显，EER 最多降低了 2.35%。且所提方法的亮点在于能够有针对性地开发度量学习的自适应能力，能使目标函数根据特征的特点进行动态调整，还能消除三元组选取阶段和目标函数相似度度量方法不一致的隐患。

3) ResNet34+MIAD 相比于其他深度神经网络方法，相对等错误率最多降低了 28%。本文所提的 MIAD 目标函数能够有效地衡量同类、异类说话人特征所在分布之间的差异性，引入自适应方法能够更有针对性地对 embedding 特征进行表示，有效提升了识别系统的性能。

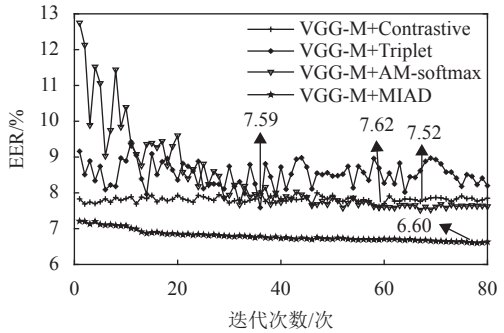
2.3　收敛性对比与分析

本节将对比具有相同网络结构的不同目标函数方法的收敛性。网络结构分别为 VGG-M、ResNet34，目标函数则包含 AM-softmax 损失、三元组损失、对比损失、MIAD 损失。收敛性曲线采用 EER 和 minDCF 作为性能评价指标，上述所有方法均使用相同的预训练模型。4 种方法的收敛性曲线图如图 2 所示，从图中可以看出以下几点。

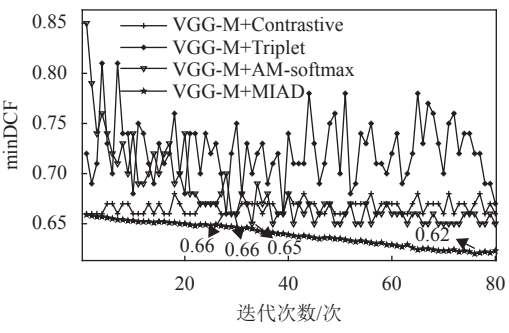
1) 随着迭代次数的增加，全部方法的等错误率和 minDCF 均有下降趋势。本文的 MIAD 方法在使用两种网络结构的情况下，等错误率和 minDCF 更低。

2) 本文方法 VGG-M+MIAD 在经过 78 轮迭代后等错误率达到最低，数值为 6.60%，ResNet34+MIAD 在经过 67 轮迭代后等错误率达到最低，数值为 6.44%，相比于其他使用相同网络结构的方法性能更好。可以证明本文方法能够提升说话人识别系统的性能。

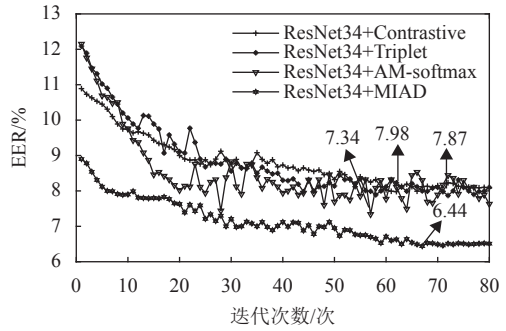
3) 本文方法在使用相同网络的情况下，均拥有更低的 minDCF，VGG-M+MIAD 数值为 0.62，ResNet34+MIAD 数值为 0.60。进一步证明了本文方法具有更好的性能。



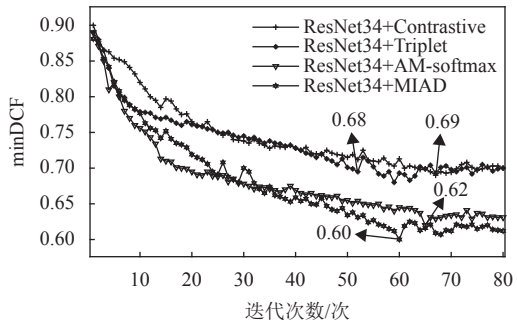
a. 使用 VGG-M 网络 EER 收敛性曲线



b. 使用 VGG-M 网络 minDCF 收敛性曲线



c. 使用 ResNet34 网络 EER 收敛曲线



d. 使用 ResNet34 网络 minDCF 收敛性曲线

图 2 收敛性曲线对比图

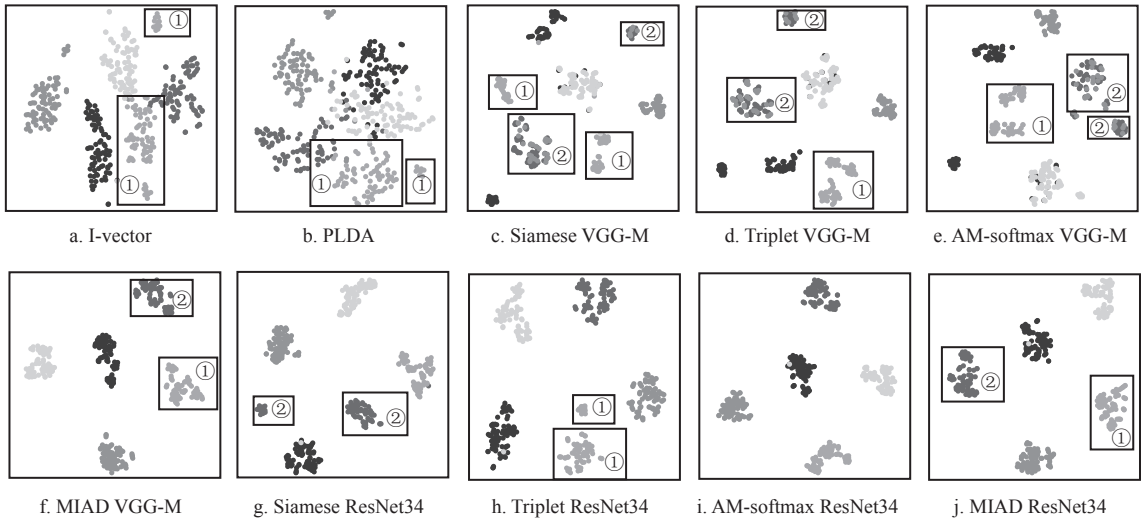
2.4 可视化分析

为了更直观地衡量本文方法的有效性，使用 t-SNE[27] 方法对不同方法进行可视化表示。对比方法包括 i-vector 特征、PLDA 说话人隐变量、VGG-M+Contrastive 的 embedding 特征、VGG-M+Triplet 的 embedding 特征、VGG-M+AM-softmax 的 embedding 特征、VGG-M+MIAD 的 embedding 特征、ResNet34+Contrastive 的 embedding 特征、ResNet34+Triplet 的 embedding 特征、ResNet34+AM-softmax 的 embedding 特征、ResNet34+MIAD 的 embedding 特征。从评估集中随机选择 5 位说话人进行可视化表示，每位说话人包含 80 段语音，不同类别的说话人对应不同灰度的点。t-SNE 方法的各项参数设置为：维度 30 维，困惑度 10。

基于上述实验设置，不同方法的可视化对比图如图 3 所示。从图中可以看出以下两点。

1) 相比于图 3a～3e、3g～3j，图 3f、3j 中的可视化特征聚集得更紧凑。由此可见，本文方法能够更好地捕获同类特征的相似性。

2) 在各子图的矩形框①中，图 3a-3e、3h 中的同类特征均被聚到 2 簇中，但图 3f、3j 却能很好地聚到同一簇中。同样地，在子图的矩形框②中，图 3c-3e、3g 中的同类特征均被聚到 2 簇中，但图 3f、3j 却能很好地聚到同一簇中。由此可见，对于那些类内差异性大的特征，本文方法仍然能够很好地对其同类相似性进行表示。



a. I-vector　b. PLDA　c. Siamese VGG-M　d. Triplet VGG-M　e. AM-softmax VGG-M

f. MIAD VGG-M　g. Siamese ResNet34　h. Triplet ResNet34　i. AM-softmax ResNet34　j. MIAD ResNet34

图 3 不同特征的可视化对比图

3 结 束 语

本文提出了一种基于互信息自适应估计的目标函数，该目标函数能够根据特征的实际情况进行动态调整，使得互信息估计能够挖掘到更有价值的同类、异类特征信息。该方法还将具有自适应能力的度量方法 NPLDA 应用于特征选取阶段，NPLDA 能够根据特征的真实情况有针对性地更新参数，使选取的特征更典型，从而有效提升在此目标函数监督下网络的表示能力。从性能、收敛性、特征可视化 3 个方面的对比分析可以证明，本文方法在说话人确认任务上具有良好表现。在后续的研究工作中，考虑到 NPLDA 中的漏报与误报对应的是目标/

非目标的确认任务，因此可以将其目标函数改进为基于互信息的损失，从而为整个网络的优化带来正向提升。

参 考 文 献

[1] ZHANG D D , ZUO W. Computational intelligence-based biometric technologies[J]. IEEE Computational Intelligence Magazine, 2007, 2(2): 26-36.

[2] 李明, 张勇, 李军权, 等. 改进 PSO-SVM 在说话人识别中的应用[J]. 电子科技大学学报, 2007, 36(6): 1345-1349.
LI M, ZHANG Y, LI J Q, et al. Application of improved PSO-SVM approach in speaker recognition[J]. Journal of University of Electronic Science and Technology of China, 2007, 36(6): 1345-1349.

[3] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//Proceedings of the British Machine Vision Conference 2014. Nottingham: BMVC, 2014: 1-12.

[4] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[C]//Proceeding of the Annual Conference of the International Speech Communication Association. Stockholm: Interspeech, 2017: 2610-2620.

[5] CHUNG J S, NAGRANI A, ZISSERMAN A. Voxceleb2: Deep speaker recognition[C]//Proceeding of the Annual Conference of the International Speech Communication Association. Hyderabad: Interspeech, 2018: 1086-1090.

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

[7] HUANG Z L, WANG S, YU K. Angular softmax for short-duration text-independent speaker verification[C]// Proceeding of the Annual Conference of the International Speech Communication Association. Hyderabad: Interspeech, 2018: 3623-3627.

[8] YU Y Q, FAN L, LI W J. Ensemble additive margin softmax for speaker verification[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 6046-6050.

[9] ZHOU D, WANG L, LEE K A, et al. Dynamic margin softmax loss for speaker verification[C]// Proceeding of the Annual Conference of the International Speech Communication Association. Shanghai: Interspeech, 2020: 3800-3804.

[10] ZHONG Q, DAI R, ZHANG H, et al. Text-independent speaker recognition based on adaptive course learning loss and deep residual network[J]. EURASIP Journal on Advances in Signal Processing, 2021(1): 1-16.

[11] LI N, TUO D, SU D, et al. Deep discriminative embeddings for duration robust speaker verification[C] //Proceeding of the Annual Conference of the International Speech Communication Association. Hyderabad: Interspeech, 2018: 2262-2266.

[12] LIU Y, HE L, LIU J. Large margin softmax loss for speaker verification[C]//Proceeding of the Annual Conference of the International Speech Communication Association. Graz: Interspeech, 2019: 2873-2877.

[13] ZHANG Y, YU M, LI N, et al. Seq2seq attentional Siamese neural networks for text-dependent speaker verification[C]//ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 6131-6135.

[14] BHATTACHARYA G, ALAM M J, GUPTA V, et al. Deeply fused speaker embeddings for text-independent speaker verification[C]//Proceeding of the Annual Conference of the International Speech Communication Association. Hyderabad: Interspeech, 2018: 3588-3592.

[15] CHUNLEI Z, KAZUHITO K, HANSEN J H L. Text-Independent speaker verification based on triplet convolutional neural network embeddings[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1633-1644.

[16] BAI Z, ZHANG X L, CHENG J. Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification[C]//ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 6819-6823.

[17] 陈晨, 彤娅峰, 季超群, 等. 基于深层信息散度最大化的说话人确认方法[J]. 通信学报, 2021, 42(7): 231-237.
CHEN C, RONG Y F, JI C Q, et al. Speaker verification method based on deep information divergence maximization[J]. Journal on Communications, 2021, 42(7): 231-237.

[18] KYE S M, JUNG Y, LEE H B, et al. Meta-learning for short utterance speaker recognition with imbalance length pairs[C]//Proceeding of the Annual Conference of the International Speech Communication Association. Shanghai: Interspeech, 2020: 2982-2986.

[19] RAMOJI S, KRISHNAN P, GANAPATHY S. NPLDA: A deep neural PLDA model for speaker verification[C]//Odyssey: The Speaker and Language Recognition Workshop. Tokyo: Odyssey, 2020: 202-209.

[20] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

[21] DEHAK N, KENNY P J, DEHAK R, et al. Front-End factor analysis for speaker verification[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(4): 788-798.

[22] 杨成福, 章毅. 相关向量机及在说话人识别应用中的研究[J]. 电子科技大学学报, 2010, 39(2): 311-315.
YANG C F, ZHANG Y. Study to speaker recognition using RVM[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(2): 311-315.

[23] ATHULYA M S, SATHIDEVI P S. Speaker verification from codec-distorted speech through combination of affine transform and feature switching[J]. Circuits, Systems, and Signal Processing, 2021, 40(12): 6016-6034.

[24] DING S, CHEN T, GONG X, et al. AutoSpeech: Neural architecture search for speaker recognition[C]// Proceeding of the Annual Conference of the International Speech Communication Association. Shanghai: Interspeech, 2020: 916-920.

[25] SHON S, TANG H, GLASS J. Frame-Level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model[C]//2018 IEEE Spoken Language Technology Workshop (SLT). Athens: IEEE, 2018: 1007-1013.

[26] RAVANELLI M, BENGIO Y. Speaker recognition from raw waveform with Sincnet[C]//2018 IEEE Spoken Language Technology Workshop (SLT). Athens: IEEE, 2018: 1021-1028.

[27] LAURENS V D M, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

编　辑　叶　芳