基于深层信息散度最大化的说话人确认方法

陈晨[1,2]，彤娅峰[1]，季超群[1]，陈德运[1,2]，何勇军[1]
（1. 哈尔滨理工大学计算机科学与技术学院，黑龙江 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，黑龙江 哈尔滨 150080）

摘 要：针对说话人确认中无法准确捕获特征间非线性关系的问题，提出了一种基于深层信息散度最大化的目标函数表示方法。该方法能通过计算特征所在分布之间相似度，来对特征间的非线性关系进行隐性表示，并在最大化这种统计相关性的优化目标指导下，使深度神经网络向着同类数据更紧凑、异类数据更分散的方向优化，最终达到提升深层特征空间区分性的目标。实验结果表明，相对于其他深度学习方法，所提方法的相对等错误率（EER）最多降低了 15.80%，显著提升了系统性能。

关键词：说话人确认；目标函数；深层信息散度；特征表示学习
中图分类号：TP391.4
文献标识码：A
DOI: 10.11959/j.issn.1000−436x.2021133

Speaker verification method based on deep information divergence maximization

CHEN Chen[1,2], RONG Yafeng[1], JI Chaoqun[1], CHEN Deyun[1,2], HE Yongjun[1]
1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
2. Postdoctoral Research Station of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

Abstract: To solve the problem that the nonlinear relationship between speaker representations cannot be accurately captured in speaker verification, an objective function based on depth information divergence maximization was proposed. It could implicitly represent the nonlinear relationship between speaker representations by calculating the similarity between their distributions. Under the supervision of the optimization goal of maximizing the statistical correlation, the deep neural network was optimized towards the direction that the within-class data was more compact and the between-class data were far away from each other, and finally the discrimination of deep speaker representation space could be effectively improved. Experimental results show that compared with other deep learning methods, the relative EER of the proposed method is reduced by 15.80% at most, which significantly improves the system performance.

Keywords: speaker verification, objective function, deep information divergence, representation learning

1 引言

近年来，以生物识别技术为基础的身份认证方式正在逐渐取代传统的静态身份认证手段。随着科技的发展，以指纹识别、人脸识别及说话人确认为代表的一系列生物识别技术已在多种认证场景中取得了较广泛的应用。其中，说话人确认技术能够根据说话人的声音特性来有效识别其身份。由于每

收稿日期：2021−03−26；**修回日期**：2021−06−15
基金项目：国家自然科学基金资助项目（No.61673142）；黑龙江省自然科学基金资助项目（No.JJ2019JQ0013）；黑龙江省博士后专项基金资助项目（No.LBH-Z20020）；黑龙江省普通高校基本科研业务费专项资金资助项目（No.2020-KYYWF-0341）
Foundation Items: The National Natural Science Foundation of China (No.61673142), The Natural Science Foundation of Heilongjiang Province (No.JJ2019JQ0013), Heilongjiang Postdoctoral Fund (No.LBH-Z20020), The Fundamental Research Founds for the Central Universities of Heilongjiang Province (No.2020-KYYWF-0341)

个人在说话过程中所蕴含的语音特质与发音习惯几乎独一无二，因此说话人确认技术兼具生理特性与行为特性，从而使其相较于其他生物识别技术的仿冒难度更大、安全性更高[1]。与此同时，“无接触”的说话人确认技术能够有效阻断“人传人”的传播链条，为个人健康提供更可靠的保障。

说话人确认能够通过对说话人语音信号的分析处理，来充分结合知识、数据、算法与算力，是迈向第三代人工智能[2]的典型代表。如何从大量语音数据中凝练出准确的说话人身份信息，则是说话人确认任务中最值得关注的研究焦点。为此，需要深入研究能够直接代表说话人身份特性的特征表示问题，研究者也针对该问题提出了大量有效的说话人特征表示学习模型。其中，以身份−矢量（I-vector, identity-vector）[3]方法为基础的一系列特征空间学习方法应运而生，它们均能将具有不同时长的语音信号映射为固定维度的低秩 I-vector 特征表示。在这类方法中，因子分析（FA, factor analysis）[4]、广义变化模型（GVM, generalized variability model）[5]、任务驱动多层结构（TDMF, task-driven multilevel framework）[6]等方法为典型代表。此外，为了去除语音信号中的会话差异性信息（如语音内容间的差异、噪声、信道畸变等），还需要对 I-vector 特征进行会话补偿[7-8]等操作。

除此之外，随着深度神经网络在图像处理、音频处理等方面取得的突破进展，基于深度神经网络的特征表示方法也逐渐出现在说话人确认研究中。例如，D-vector 方法[9]采用深度神经网络（DNN, deep neural network）来提取说话人语音对应的嵌入（embedding）特征，为端到端（E2E, end-to-end）说话人确认方法的发展奠定了基础。X-vector 方法[10-11]则利用时延神经网络（TDNN, time-delay neural network）[12]、统计池化层与全连接层来提取表示说话人身份的 X-vector 特征。由于 X-vector 方法能够取得优良的性能，在此基础上又出现了基于分解 TDNN（F-TDNN, factorized TDNN）[13]、扩展 TDNN（E-TDNN, extended TDNN）[14]、聚合残差扩展 TDNN（ARE-TDNN, aggregated residual extended TDNN）[15]以及稠密连接 TDNN（DC-TDNN, densely connected TDNN）[16]的 X-vector 特征提取方法。此外，视觉几何组−中等（VGG-M, visual geometry group-medium）[17]网络则通过多层的卷积层与池化层的叠加来进行说话人特征表示的学习。

以上方法均通过构建不同的网络结构来学习说话人的特征表示，考虑到目标函数能够对网络描述能力的提升起到重要的指导作用，因此，设计出有的放矢的目标函数能够使所提取的特征更适用于当前任务。在这些目标函数中，一类目标函数以多分类为目标，例如 softmax 损失、交叉熵损失（CE loss，cross entropy loss）；另一类目标函数以度量特征表示之间的相似度为目标，例如对比损失（contrastive loss）[18]与三元组损失（triplet loss）[19-20]等。也有一些目标函数在多分类目标的基础上加入度量学习的限制，例如角 softmax（A-softmax, angular softmax）损失[21-22]、加性边沿 softmax（AM-softmax, additive margin softmax）损失[23]与加性边沿质心（AM-centroid, additive margin centroid）损失[24]等。

由于目标函数是整个任务目标的最直观体现，它能直接影响网络参数的优化方向，因此一个优秀的目标函数将为网络的特征表示能力带来大幅提升。目前，说话人确认研究中所采用的目标函数均基于这一原则取得了卓有成效的成绩。然而，说话人的类别不胜枚举，并无法保证训练数据能够涵盖全部待识别语音的类别，因此采用以多分类为目标的目标函数往往会导致模型的泛化能力不强；反之，以度量学习为目标的目标函数则通过分别控制同类、异类说话人深层特征间的相关性，来驱使网络朝着提升类内相似性与类间差异性的方向优化，从而为网络带来更强的泛化性与普适性。目前，基于度量学习的目标函数大多仅通过简单的欧氏距离或余弦距离来衡量特征间的相关性，并无法准确捕获特征间复杂的非线性关系。而此非线性关系才是特征间相关性的真实写照，其对特征在特征空间的可区分性表示具有十分重要的指导性作用。因此，如何有效度量这种非线性关系是目前亟待解决的关键问题。

针对上述问题，考虑到非线性关系无法通过显性的表达式进行表示，但能够以计算特征所在分布之间相似度的方式进行隐性表示，因此本文将能够计算分布间相似度的信息散度（ID, information divergence）[25-26]引入目标函数的表示过程中，提出基于深层信息散度最大化的说话人确认方法。其将最大化特征之间的统计相关性作为优化目标，并以此来控制神经网络挖掘同类特征之间必然存在的相容性信息、提升异类特征在特征空间的差异性，最终有效提升深层特征空间的区分性。

2　深层信息散度理论

2.1　信息散度表示

在说话人确认任务中，目标函数的定义对区分性网络学习具有至关重要的作用。同时，由于说话人确认系统应具备开集测试的能力，因此定义基于同类、异类说话人间关系的目标函数能够为网络的学习提供普适性更强的下游任务目标。值得注意的是，传统基于距离的相似度度量方式无法有效表示特征间的非线性关系。为此，本文构建了一种基于深层信息散度的目标函数，其能够有效度量同类、异类说话人特征所在分布之间的差异性，从而更加准确地刻画深层特征间的抽象关系。在此目标函数的指导下，神经网络能够向着同类更紧凑、异类更分离的方向进行优化。

定义 s 表示随机采样的样本组，其由 2 个深层特征组成。当样本组中的特征属于同类时，它们的联合分布为 $P(s)=P(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{p}})$；当属于异类时，它们的联合分布为 $Q(s)=Q(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{n}})$，其中 $\boldsymbol{z}_{\mathrm{a}}$ 、$\boldsymbol{z}_{\mathrm{p}}$ 、$\boldsymbol{z}_{\mathrm{n}}$ 分别表示固定（anchor）样本、正例（positive）样本、负例（negative）样本。由于同类、异类样本分布间的差异应尽可能大，因此本文通过最大化 $P(s)$ 与 $Q(s)$ 间的 ID 来达到提升同类、异类差异的目标，此信息散度可以表示为

$$D_{\mathrm{ID}}\left[P(s)\|Q(s)\right]=\int_s P(s)\log\left[\frac{P(s)}{Q(s)}\right]\mathrm{d}s \tag{1}$$

对式(1)进行等价变换，可以得到

$$D_{\mathrm{ID}}\left[P(s)\|Q(s)\right]=\int_s Q(s)\frac{P(s)}{Q(s)}\log\left[\frac{P(s)}{Q(s)}\right]\mathrm{d}s \tag{2}$$

定义 $f(x)=x\log x$，其中 $x=P(s)/Q(s)$，则式(2)可以转换为

$$D_{\mathrm{ID}}\left[P(s)\|Q(s)\right]=\int_s Q(s)f\left[\frac{P(s)}{Q(s)}\right]\mathrm{d}s \tag{3}$$

其中，函数 $f(x)$ 可以由其共轭函数 $f^*(t)$ 进行表示，具体形式为

$$f(x)=\max\left\{xt-f^*(t)\right\} \tag{4}$$

由式(4)可推导出，$f(x)=x\log x$ 的共轭函数为 $f^*(t)=\mathrm{e}^{t-1}$。由于每个 x 都有与其对应的 t，因此 t 是关于 x 的函数，本文将其表示为 $t=d(x)$。将 $f^*(t)$ 与 $d(x)$ 同时代入式(4)，可以得到

$$f(x)=\max\left\{xd(x)-\mathrm{e}^{d(x)-1}\right\} \tag{5}$$

将式(5)代入式(3)，则 $P(s)$ 与 $Q(s)$ 分布之间的信息散度可以进一步表示为

$$\begin{aligned}
&D_{\mathrm{ID}}\left[P(s)\|Q(s)\right]=\\
&\int_s Q(s)\max\left\{\frac{P(s)}{Q(s)}d\left[\frac{P(s)}{Q(s)}\right]-\mathrm{e}^{d\left[\frac{P(s)}{Q(s)}\right]-1}\right\}\mathrm{d}s=\\
&\max\left\{\int_s\left\{P(s)d\left[\frac{P(s)}{Q(s)}\right]-Q(s)\mathrm{e}^{d\left[\frac{P(s)}{Q(s)}\right]-1}\right\}\mathrm{d}s\right\}=\\
&\max\left\{\mathbb{E}_{s\sim P(s)}\left\{d\left[\frac{P(s)}{Q(s)}\right]\right\}-\mathbb{E}_{s\sim Q(s)}\left\{\mathrm{e}^{d\left[\frac{P(s)}{Q(s)}\right]-1}\right\}\right\}
\end{aligned} \tag{6}$$

至此，便得到了基于信息散度表示的目标函数的一般形式。其中，$P(s)/Q(s)$ 为正、负样本组的似然比，是说话人确认中最常见的评价指标之一，当函数 $d(\cdot)$ 作用于其上时，所得到的新形式仍可用于衡量 2 个样本间相关性。

2.2　目标函数表示

本节将在第 2.1 节的基础上，进一步展开讨论函数 $d(\cdot)$ 的具体形式。当 $s\sim P(s)$ 时，s 为正例样本组；当 $s\sim Q(s)$ 时，s 为负例样本组。因此 $\mathbb{E}_{s\sim P(s)}[d(\cdot)]$ 与 $\mathbb{E}_{s\sim Q(s)}\{\exp[d(\cdot)-1]\}$ 分别对应了正、负例样本组的相关性。基于此，式(6)可以进一步表示为

$$\begin{aligned}
&D_{\mathrm{ID}}\left[P(s)\|Q(s)\right]=\\
&\max\left\{\mathbb{E}_{(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{p}})\sim P(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{p}})}\left[d(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{p}})\right]-\mathbb{E}_{(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{n}})\sim Q(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{n}})}\left[\mathrm{e}^{d(\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{n}})-1}\right]\right\}
\end{aligned} \tag{7}$$

为了使 $d(\cdot)$ 继承似然比的作用，其仍然应该具备相似度计算的功能。基于此，本文将其定义为余弦距离打分（CDS, cosine distance score）的形式

$$d(\boldsymbol{z}_1,\boldsymbol{z}_2)=\frac{\langle\boldsymbol{z}_1,\boldsymbol{z}_2\rangle}{\|\boldsymbol{z}_1\|\|\boldsymbol{z}_2\|} \tag{8}$$

其中，$\langle\cdot\rangle$ 表示内积，$\|\cdot\|$ 表示取模。式(7)与式(8)即为本文提出的基于深层信息散度表示的目标函数。在采样过程中，其需要随机地选择固定样本、正例样本与负例样本，组成三元组样本集合。

在网络结构设置方面，考虑到 VGG-M 网络[17]

作为说话人确认领域中的经典网络之一，能够取得良好的性能，且已经得到了很多研究者的实验验证，因此本文采用 VGG-M 网络进行特征表示学习。网络输入采用语谱图特征，对输入特征进行随机的三元组采样，得到样本 $\boldsymbol{x}_{\mathrm{a}}$、$\boldsymbol{x}_{\mathrm{p}}$、$\boldsymbol{x}_{\mathrm{n}}$，它们经 VGG-M 得到的嵌入特征分别表示为 $\boldsymbol{z}_{\mathrm{a}}=g\left(\boldsymbol{x}_{\mathrm{a}}|\Theta\right)$、$\boldsymbol{z}_{\mathrm{p}}=g\left(\boldsymbol{x}_{\mathrm{p}}|\Theta\right)$、$\boldsymbol{z}_{\mathrm{n}}=g\left(\boldsymbol{x}_{\mathrm{n}}|\Theta\right)$，由此可构成正例样本组$\left\{\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{p}}\right\}$、负例样本组$\left\{\boldsymbol{z}_{\mathrm{a}},\boldsymbol{z}_{\mathrm{n}}\right\}$。其中，$g$ 表示 VGG-M 网络，Θ 为 VGG-M 网络的参数。基于深层信息散度最大化与 VGG-M 网络的结构如图 1 所示。

3 实验分析

3.1 数据库与评价指标

本文实验采用 VoxCeleb1 数据库[17]对不同方法的性能进行对比与分析，该数据库的全部音频选自 YouTube 视频网站，是来自复杂场景下的真实语音，包含大量未知噪声。使用该数据库官方说话人确认任务的划分方案：将说话人中不以字母“E”开头的说话人语音作为开发集数据，其中包含 1 211 位说话人、148 642 段语音；以字母“E”开头的说话人语音作为评估集数据，其中包含 50 位说话人、4 874 段语音。测试时采用官方测试计划，总测试数为 37 720 次，目标测试与非目标测试比为 1:1。实验采用等错误率（EER, equal error rate）与最小检测代价函数（Min DCF, minimum detection cost function）来衡量系统的性能，其中 Min DCF 的参数设置为 $C_{\mathrm{miss}}=1, C_{\mathrm{fa}}=1, P_{\mathrm{target}}=0.01$。

为了验证信息散度最大化目标函数的有效性，本文根据如上所述的数据库与实验设置，分别从性能对比与分析、收敛性分析、可视化分析 3 个角度进行实验。

3.2 性能对比与分析

本节将所提方法（简记为 ID-max VGG-M）与其他方法的识别性能进行对比。对比方法除了选择 2 个经典的说话人确认方法，即高斯混合模型–通用背景模型（GMM-UBM, Gaussian mixture model-universal background model）[27]、基于因子分析的 I-vector 方法[3]外，还选择了如下基于深度学习的方法：采用对比（contrastive）损失的孪生（siamese）VGG-M 网络[17]、采用三元组（triplet）损失[19]的 VGG-M 网络与采用 AM-softmax 损失[23]的 VGG-M 网络。为了便于书写，本文将上述方法分别简记为 GMM-UBM、I-vector+PLDA、Siamese VGG-M、Triplet VGG-M 与 AM-softmax VGG-M。

在经典方法的实验中，先对各说话人语音进行语音活动检测处理[28]，以去除语音中的静音部分，然后进行特征提取。前端特征采用梅尔倒谱系数（MFCC, Mel-frequency ceptral coefficient）特征，其维度为 13 维，并计算其一阶、二阶差分，组成 39 维的声学特征。通用背景模型（UBM, universal background model）的高斯混合分量个数为 1 024，总变化空间维度为 400 维，概率线性判别分析模型（PLDA, probabilistic linear discriminant analysis）的子空间维度为 200 维。在识别阶段，GMM-UBM 通过计算测试语音在目标说话人 GMM 上的似然概率密度来获得匹配得分；I-vector+PLDA 方法采用 PLDA 模型作为后端分类器；Siamese VGG-M、Triplet VGG-M 与 AM-softmax VGG-M 方法采用 CDS 方法进行说话人确认匹配。

在深度学习方法的实验中，网络的输入为语谱图特征，为了保证实验对比的公平性与有效性，其参数设置与文献[17]一致，即滑动窗的窗长设置为 25 ms，帧移为 10 ms，快速傅里叶变换的点数为 512 个。基于此，对于一段 3 s 的语音，可以提取 512×300 维的语谱图特征。对于 VGG-M 网络，其结构同样采用文献[17]中的设置，最后一层全连接层的节点数为 1 024 个，由此可得说话人深层特征表示的维度为 1 024 维。训练 VGG-M 网络的优化

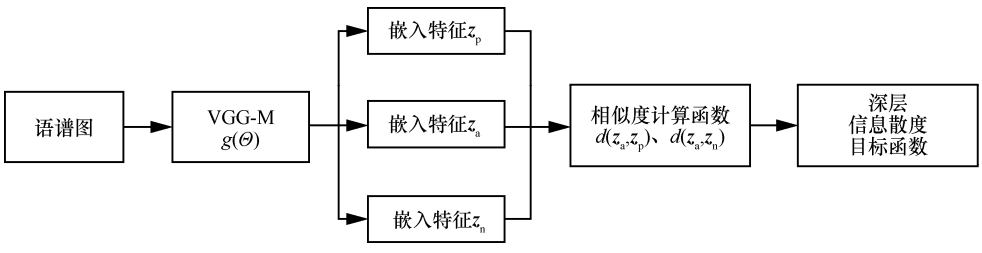


图 1 基于深层信息散度最大化与 VGG-M 网络的结构

器采用随机梯度下降（SGD, stochastic gradient descent）算法，学习率与迭代次数则根据多次的参数调优来确定，最终选择性能最佳时对应的初始学习率、最终学习率与迭代次数，分别为 0.001、0.000 1 与 80。根据上述实验设置，不同方法对应的系统性能情况如表 1 所示。

表 1　　不同方法的性能对比

方法	EER	Min DCF
GMM-UBM	15.00%	0.80
I-vector+PLDA	8.80%	0.73
Siamese VGG-M	7.85%	0.68
Triplet VGG-M	7.71%	0.66
AM-softmax VGG-M	7.35%	0.63
ID-max VGG-M	6.61%	0.61

由表 1 的实验结果可以得出，相比于其他方法，本文提出的 ID-max VGG-M 方法具有更低的 EER。其与 Siamese VGG-M 方法、Triplet VGG-M 方法、AM-softmax VGG-M 方法 3 种方法相比，相对 EER 分别降低了 10.1%、15.8%、14.3%。这也这也验证了本文所提出的 ID-max 目标函数能够指导网络学习更具表示能力的说话人深层特征。

3.3　收敛性分析

本节将对 ID-max VGG-M 方法的收敛性进行验证与分析，通过记录每次 VGG-M 网络训练时在评估集数据上的 EER，来绘制收敛性曲线。根据上述的实验设置，4 种方法的收敛性曲线如图 2 所示。

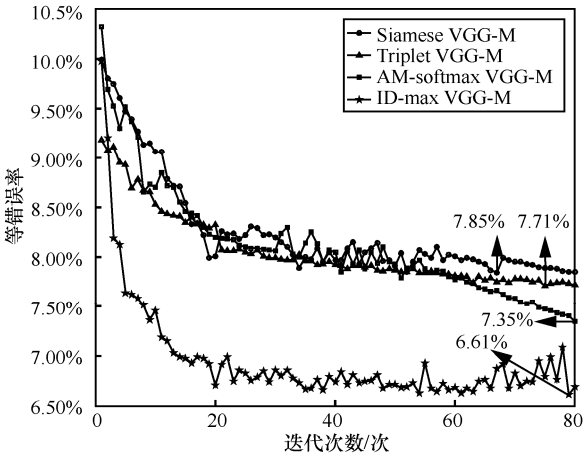


图 2　收敛性曲线对比

从图 2 中可得到以下结论。

1) 从整体上看，随着迭代次数的增加，这 4 种方法对应的等错误率呈下降趋势，系统性能逐渐上升。相比于其他 3 种方法，ID-max VGG-M 方法的等错误率更低。

2) 这 4 种方法均能够在有限的迭代次数内达到收敛状态，其中 ID-max VGG-M 方法在第 79 次迭代时，等错误率达到最低，为 6.61%，这是说话人确认系统最优的性能。

3.4　可视化分析

本节将采用 t-SNE 方法[29]对提取的深层特征表示（embedding）进行 2D 可视化处理，其中 t-SNE 初始降维的维度为 30 维，困惑度为 10。在评估集中随机选择 5 位说话人，并从这 5 位说话人的全部数据中随机选择 80 段语音，各方法均采用以上设置进行数据选择。根据上述设置，不同方法对应的可视化图像如图 3 所示，其中，不同灰度的点代表不同说话人。将所对比方法的说话人特征表示分别记为 I-vector 特征、PLDA 说话人隐变量、Siamese VGG-M embedding 特征、Triplet VGG-M embedding 特征、AM-softmax VGG-M embedding 特征与 ID-max VGG-M embedding 特征。

由图 3 中的实验结果可以得出以下结论。

1) 由图 3(a)与图 3(b)可知，相同类别的说话人特征能够在一定程度上聚集在一起，这是因为经典的 I-vector 特征与 PLDA 隐变量已具有一定的区分能力。但是同类数据仍然较分散，异类数据之间也有相互交叠。

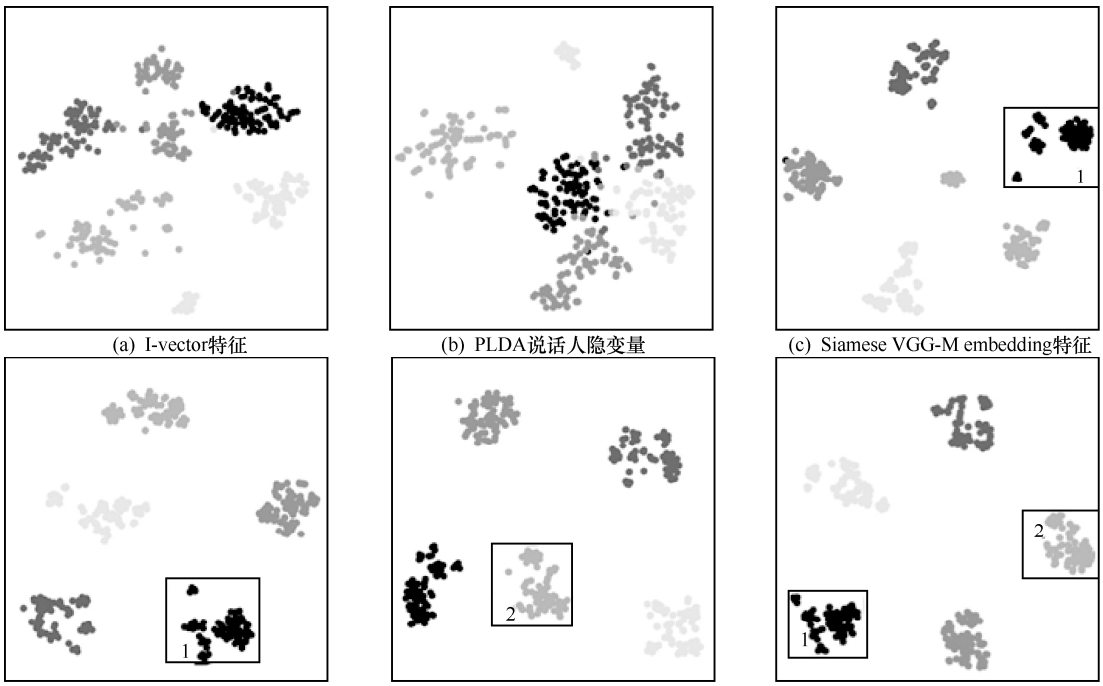
2) 对比图 3(c)、图 3(d)与图 3(f)可知，图 3(f)中的同类特征点更加紧凑。矩形框 1 内的这一现象尤其明显：图 3(c)与图 3(d)中的特征点分散在多个簇内，而图 3(f)中的特征点则相对更加集中。

3) 由图 3(e)与图 3(f)可知，与 ID-max 目标函数相比，当以 AM-softmax 为目标函数提取说话人特征时，同类特征点在空间中仍然较分散，图 3(e)矩形框 2 中的特征点分散得尤其明显。

由此可见，本文提出的基于深层信息散度最大化的目标函数能够使同类的说话人特征表示更加紧凑，异类的特征更加分散。由此得到的说话人特征表示的区分性更强，相应说话人确认系统的性能也能更优。

4　结束语

本文提出一种基于深层信息散度最大化的目标函数表示方法，其将最大化同类、异类说话人特



(a) I-vector特征　(b) PLDA说话人隐变量　(c) Siamese VGG-M embedding特征

(d) Triplet VGG-M embedding特征　(e) AM-softmax VGG-M embedding特征　(f) ID max VGG-M embedding特征

图 3　说话人特征表示的可视化图像对比

征表示所在分布之间的信息散度作为优化目标，挖掘其中存在的非线性关联信息。并以此来控制神经网络挖掘同类样本之间相关性信息，从而有效提升不同说话人数据在特征空间的区分性。实验结果表明，与其他方法相比，所提方法能够有效改善说话人确认系统的性能。

参考文献：

[1] 郑方, 李蓝天, 张慧, 等. 声纹识别技术及其应用现状[J]. 信息安全研究, 2016, 2(1): 44-57.
ZHENG F, LI L T, ZHANG H, et al. Overview of voiceprint recognition technology and applications[J]. Journal of Information Security Research, 2016, 2(1): 44-57.

[2] 张钹, 朱军, 苏航. 迈向第三代人工智能[J]. 中国科学: 信息科学, 2020, 50(9): 1281-1302.
ZHANG B, ZHU J, SU H. Toward the third generation of artificial intelligence[J]. Scientia Sinica (Informationis), 2020, 50(9): 1281-1302.

[3] 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.

[4] VESTMAN V, KINNUNEN T. Supervector compression strategies to speed up I-vector system development[C]//Odyssey 2018 The Speaker and Language Recognition Workshop. [S.n.:s.l.], 2018: 357-364.

[5] MA J B, SETHU V, AMBIKAIRAJAH E, et al. Generalized variability model for speaker verification[J]. IEEE Signal Processing Letters, 2018, 25(12): 1775-1779.

[6] CHEN C, HAN J Q. TDMF: task-driven multilevel framework for end-to-end speaker verification[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 6809-6813.

[7] 高荣春, 韩纪庆, 张磊. 说话人识别中基于最大后验概率的通道补偿方法[J]. 通信学报, 2009, 30(3): 99-103.
GAO R C, HAN J Q, ZHANG L. Channel compensation of speaker identification based on maximum a posteriori[J]. Journal on Communications, 2009, 30(3): 99-103.

[8] 汪海彬, 郭剑毅, 毛存礼, 等. 基于通用背景-联合估计(UB-JE)的说话人识别方法[J]. 自动化学报, 2018, 44(10): 1888-1895.
WANG H B, GUO J Y, MAO C L, et al. Speaker recognition based on universal background-joint estimation(UB-JE)[J]. Acta Automatica Sinica, 2018, 44(10): 1888-1895.

[9] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2014: 4052-4056.

[10] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Interspeech 2017. Piscataway: IEEE Press, 2017: 999-1003.

[11] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 5329-5333.

[12] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[13] VILLALBA J, CHEN N, SNYDER D, et al. State-of-the-art speaker recognition for telephone and video speech[C]//Proceeding of the Twenty Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2019: 1488-1492.

[14] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recogni-

tion for multi-speaker conversations using X-vectors[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 5796-5800.

[15] ZHANG R T, WEI J G, LU W H, et al. ARET: aggregated residual extended time-delay neural networks for speaker verification[C]//Interspeech 2020. Piscataway: IEEE Press, 2020: 946-950.

[16] YU Y Q, LI W J. Densely connected time delay neural network for speaker verification[C]//Interspeech 2020. Piscataway: IEEE Press, 2020: 921-925.

[17] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: a large-scale speaker identification dataset[C]//Interspeech 2017. Piscataway: IEEE Press, 2017: 2616-2620.

[18] BHATTACHARYA G, ALAM M J, GUPTA V, et al. Deeply fused speaker embeddings for text-independent speaker verification[C]//Interspeech 2018. Piscataway: IEEE Press, 2018: 3588- 3592.

[19] ZHANG C L, KOISHIDA 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.

[20] 陈莹, 陈湟康. 基于多模态生成对抗网络和三元组损失的说话人识别[J]. 电子与信息学报, 2020, 42(2): 379-385.
CHEN Y, CHEN H K. Speaker recognition based on multimodal generative adversarial nets with triplet-loss[J]. Journal of Electronics & Information Technology, 2020, 42(2): 379-385.

[21] HUANG Z L, WANG S, YU K. Angular softmax for short-duration text-independent speaker verification[C]//Interspeech 2018. Piscataway: IEEE Press, 2018: 3623-3627.

[22] NOVOSELOV S, SHULIPA A, KREMNEV I, et al. On deep speaker embeddings for text-independent speaker recognition[C]//Odyssey 2018 The Speaker and Language Recognition Workshop. Piscataway: IEEE Press, 2018: 378-385.

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

[24] WEI Y H, DU J Z, LIU H. Angular margin centroid loss for text-independent speaker recognition[C]//Interspeech 2020. Piscataway: IEEE Press, 2020: 3820-3824.

[25] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

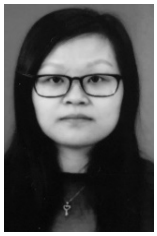
[26] BELGHAZI M. I, BARATIN A, RAJESHWAR S, et al. Mutual information neural estimation[C]//Proceeding of the Thirty-Fifth International Conference on Machine Learning. Piscataway: IEEE Press 2018: 531-540.

[27] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41.

[28] 龙华, 杨明亮, 邵玉斌. 基于特征流融合的带噪语音检测算法[J]. 通信学报, 2020, 41(4): 134-142.
LONG H, YANG M L, SHAO Y B. Noisy voice detection algorithm based on feature stream fusion[J]. Journal on Communications, 2020, 41(4): 134-142.

[29] MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

[作者简介]



陈晨（1990− ），女，黑龙江哈尔滨人，博士，哈尔滨理工大学讲师、硕士生导师，主要研究方向为语音信号处理、音频信息分析、说话人识别等。

彤娅峰（1997− ），女，河南南阳人，哈尔滨理工大学硕士生，主要研究方向为说话人识别、语音信号处理等。

季超群（1995− ），男，黑龙江绥化人，哈尔滨理工大学硕士生，主要研究方向为说话人识别、语音信号处理等。

陈德运（1962− ），男，黑龙江哈尔滨人，博士，哈尔滨理工大学教授、博士生导师，主要研究方向为模式识别、机器学习等。

何勇军（1980− ），男，四川南充人，博士，哈尔滨理工大学教授、博士生导师，主要研究方向为语音信号处理、图像处理等。

《通信学报》第十届编辑委员会

收录声明

本刊对发表的文章，拥有出版电子版、网络版版权，并拥有和其他网站交换信息的权利。本刊支付的稿酬中已经包含上述费用。

Journal on Communications has the copyright to publish electronic edition, online edition of the published articles, and has the right to exchange information with other sites. The expenses have been included in the fee paid by editorial department.

道德声明

本刊发表的论文是作者独立取得的原创性研究成果，无一稿多投；论文内容不涉及国家机密；未曾以任何形式用任何文种在国内外公开发表过；论文内容不侵犯他人著作权和其他权利。若发生一稿多投、侵权、泄密等问题，论文作者将承担全部责任。

The authors of *Journal on Communications* guarantee that their submitted articles are original and contain nothing confidential. The said article is only submitted to *Journal on Communications*. The said article has not been published before and has not been submitted elsewhere for print or electronic publication consideration. The said article is no way whatever a violation or an infringement of any existing copyright or license from the third party. Otherwise, the authors of the said article shall take the blame for the violation or infringement of the related copyright and the leakage of secrets.

通信学报

Journal on Communications

发行代号：国内2-676 / 国外M395

2021年7月25日出版　定价：98.00元

ISSN 1000-436X

CN 11-2102/TN