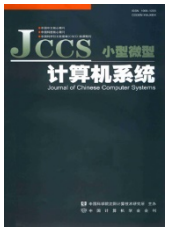


小型微型计算机系统
Journal of Chinese Computer Systems
ISSN 1000-1220,CN 21-1106/TP

《小型微型计算机系统》网络首发论文

题目：强调信息传播和特征分布的说话人验证模型：EIPFD-ResNet
作者：张霞，刘乾，郭倩，梁新彦，钱宇华，畅江
收稿日期：2022-08-17
网络首发日期：2022-11-14
引用格式：张霞，刘乾，郭倩，梁新彦，钱宇华，畅江．强调信息传播和特征分布的说话人验证模型：EIPFD-ResNet[J/OL]．小型微型计算机系统．
https://kns.cnki.net/kcms/detail/21.1106.TP.20221114.0833.002.html

小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems

强调信息传播和特征分布的说话人验证模型：EIPFD-ResNet

张 霞[1,2,3]，刘 乾[1,2,3]，郭 倩[1,2]，梁新彦[1,2]，钱宇华[1,2]，畅 江[1,2]

[1]（山西大学 大数据科学与产业研究院，太原 030006）

[2]（山西省机器视觉与数据挖掘工程研究中心，太原 030006）

[3]（山西大学 计算机与信息技术学院，太原 030006）

E-mail：stitch0507@163.com

摘 要：说话人验证是一种自然、有效的生物特征身份认证方法，其性能很大程度上取决于所提取说话人特征的质量。残差网络（ResNet）具有优越的推理能力，可以提取高质量的说话人特征，因此广泛地应用于说话人验证任务中，然而目前残差网络仍存在音频数据信息利用不充分，提取的特征不利于分类说话人等问题，这些问题大大限制了残差网络的表征能力。本文聚焦于残差网络的模型结构，详细分析了残差块分布比例、激活层、跳跃连接这些结构因素对特征信息提取的影响，以及模型输出特征分布对说话人分类结果的影响，并据此对原始残差块、特征下采样过程以及模型输出头重新设计并构建了一个新的说话人验证模型：EIPFD-ResNet。该模型采用更少激活层的残差块和单独设计的下采样层共同作用来减少音频信号的损失和噪声信息的引入，采用归一化处理后的模型输出头帮助分类损失提供更清晰的分类决策面，并在 3 个公开数据集（VoxCeleb1、VoxCeleb2、Cn-Celeb2）上评估了所提模型的有效性。实验结果证明，本文提出的模型在仅有 7.486M 参数量的情况下，相较于传统 ResNet34 模型，在 3 个数据集上的等错误率(EER)分别降低了 16.4%、33.3%、6.0%，且与强说话人验证模型 ECAPA-TDNN 相比在 VoxCeleb2 和 CN-Celeb2 上 EER 分别降低了 10%和 9.0%。

关键词：说话人验证；声纹识别；说话人嵌入；表征学习；残差网络

中图分类号：TP391 **文献标识码：**A

EIPFD-ResNet: Emphasized Information Propagation and Feature Distribution in ResNet Based Speaker Verification

ZHANG Xia[1,2,3],LIU Qian[1,2,3],GUO Qian[1,2],LIANG Xin-yan[1,2],QIAN Yu-hua[1,2],CHANG Jiang[1,2]

[1]（*Research Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006,China*）

[2]（*Engineering Research Center for Machine Vision and Data Mining of Shanxi Province, Taiyuan 030006,China*）

[3]（*School of Computer and Information Technology, Shanxi University, Taiyuan 030006,China*）

Abstract: Speaker verification is a natural and effective biometric authentication method, and its performance largely depends on the quality of the extracted speaker features. Residual network (ResNet) has superior reasoning ability and can extract high-quality speaker features. Therefore, it is widely used in speaker verification tasks. However, at present, in speaker feature extraction tasks, residual network still has some problems, such as insufficient use of audio data information, and the extracted features are not conducive to classifying speakers. These problems greatly limit the representation ability of residual network. This paper focuses on the model structure of the residual network, analyzes in detail the influence of the structural factors such as residual block distribution proportion, activation layer and shortcut on the feature information extraction, and the influence of the model output feature distribution on the speaker classification results, and redesigns and constructs a new speaker verification model: EIPFD-ResNet. The model uses residual

收稿日期：2022-08-17 收修改稿日期：2022-09-13 基金项目：国家重点研发计划项目(2021ZD0112400,2020AAA0106100)资助;国家自然科学基金重点项目(62136005)资助；山西省重点研发计划项目(201903D421003)资助；山西省高等学校科技创新项目(2019L0034)资助；山西省青年科学基金项目(20210302124556)资助.作者简介：**张 霞，**女，1977 年生，博士，副教授，CCF 会员.研究方向为机器学习和图像处理；**刘 乾，**男，1997 年生，硕士研究生，研究方向为声纹识别和深度学习；**郭 倩，**女，1990 年生，博士，CCF 会员,研究方向为逻辑学习、抽象推理及它们在多图检索上的应用；**梁新彦，**男，1989 年生，博士，CCF 会员,研究方向为多视图机器学习和粒计算；**钱宇华**，1976 年生，博士，教授，博士生导师，CCF 会员.研究方向为人工智能、大数据、机器学习和数据挖掘；**畅 江**，1988 年生，博士，讲师，研究方向为语音信号处理、脑电信号分析、声纹识别和情感识别.

blocks of the less activation layer and a separately designed down sampling layer to reduce the loss of audio signal and the introduction of noise information, and uses the normalized model output to help the classification loss to provide a clearer decision surface of classification. The effectiveness of the proposed model is evaluated on three public datasets (VoxCeleb1,VoxCeleb2 and Cn-Celeb2).The experimental results show that the model proposed in this paper has only 7.486M parameters, compared with the traditional ResNet34 model, the equal error rate (EER) in the three datasets is reduced by 16.4%,33.3%,6.0%,respectively.And compared with the state-of-the-art speaker verification model: ECAPA-TDNN, the EER is reduced by 10% and 9.0% on VoxCeleb2 and CN-Celeb2,respectively.

Key words: speaker verification; speaker recognition; speaker embedding; representation learning; ResNet

1 引　言

说话人验证指的是根据待识别语音的声纹特征识别该段语音是否对应于指定说话人，它是一种自然而有效的生物特征身份认证方法，尤其是文本无关说话人验证[1]（Text Independent Speaker Verification, TI-SV），能够极大的帮助检索目标说话人。目前，其已经被广泛应用在语音认证[2,3]、语音分离[4-6]以及语音合成[7-9]等领域。一般来说，说话人验证任务中最重要的工作是构造一个说话人特征提取器，该提取器应当尽可能地生成具有区分度的固定维说话人嵌入[10,11]。近些年来，随着大量可供训练数据的出现，深度神经网络（Deep Neural Network, DNN）取代传统说话人识别方式[12-15]成为了文本无关说话人验证任务中最广泛使用的说话人表征提取模型。

目前，在端到端的深度学习说话人识别中，基于 DNN 方法的两种主流模型分别是基于时延神经网络[16-18]（Time Delay Neural Network, TDNN）的 x-vector 结构和基于深度卷积神经网络（Convolutional Neural Network, CNN）的 r-vector 结构。x-vector 采用一定空洞率的空洞卷积来提取帧级特征，接着使用池化层将所有帧级特征聚合为一个固定维的向量，最后通过全连接层来提取说话人嵌入。由于深度残差网络[18]对于识别深层信息非常有效，Li C[19,20]等人将其应用在说话人验证任务中，命名为 r-vector。和 x-vector 不同，相比于基于 TDNN 的说话人验证模型，r-vector 接受三维特征作为输入，并采用二维卷积来提取特征，在不同的说话人验证数据集上均取得了良好的效果[21-23]。尽管在 2020VoxSRC[23]挑战赛后，基于 TDNN 的 ECAPA-TDNN[25]模型在说话人验证任务中取得了最优表现，但由于 ResNet 优越的推理速度和不俗的性能在说话人验证任务中仍占据主导地位[26-28]。

由 2021VoxSRC[29]挑战赛结果不难看出，随着不断对 ResNet 层数加深或者通道加宽，基于 ResNet 模型的性能仍可与当前最优说话人验证模型：ECAPA-TDNN 性能持平。例如：2021VoxSRC 竞赛中 Wang J 等人[30]使用 ResNet101 作为特征提取模型。然而为了追求良好的性能，一味的增加网络的深度与宽度，会导致网络优化与学习的难度增加，这对于模型之后部署、应用以及进一步改进带来了巨大的负担。为解决上述问题，本文深入分析了 ResNet 体系架构，通过对网络重新设计，促进信息在网络中的传播，提出了一个新的说话人验证模型 EIPFD-ResNet，在仅使用 7.486M 参数量情况下，取得了目前说话人验证任务中的最优结果。

本文贡献主要包括以下 3 个方面：

1）提出了新的残差块结构与特征图下采样方式。本文提出的残差块允许训练初期的负权值信息通过网络以减少信息损失，重新设计的下采样方式保证了下采样过程中卷积核大小与卷积步长相同从而避免了引入无意义的特征图信息。新的残差块结构与特征图下采样方式显著改善了说话人信息在网络传播过程中的损失情况和噪声引入问题，从而提高了说话人信息在网络中的传播效率，使模型在性能提升的同时加速了收敛。

2）对生成的说话人嵌入特征规范化处理。通过改变说话人嵌入空间中的特征分布，使相同个体的特征更紧凑，不同个体之间的特征更分散，从而提升说话人分类任务的性能。

3）为文本无关说话人验证任务提供强大的基线模型。

2 残差网络结构

基于残差网络的说话人验证模型主要由说话人表征提取模块和分类模块两部分组成。说话人表征提取模块包含帧级特征提取和话语级特征聚合两个部分。帧级特征提取部分包含 4 个阶段，每个阶段包含若干残差块，各阶段中残差块分布比例/数量不同，通常来说，每个基本残差块（ResBlock）包含两个权重层（weight layer）并使用跳跃连接（shortcut）允许信息隔层相加来避免深层网络中的退化问题。每个话语聚合子模块使用特征聚合层将不同长度的帧级说话人特征编码为固定长度的话语级特征[31]，通过模型输出头将固定长度说话人特征送入分类模块，以此训练模型对说话人嵌入的辨别能力，通常将模型输出头后的输出称为说话人嵌入（embedding）。表 1 给出了基于 ResNet 说话人识别模型结构（T 和 F 分别代表特征图的时间维度与频率维度，（3×3），32，1 代表卷积核大小为 3×3，通道数为 32，卷积步长为 1 的卷积层；BN 代表批归一化层；{ResBlock，32，1}×3 代表该阶段由 3 个通道为 32 的步长为 1 的 ResBlock 叠加在一起，FC 代表全连接层）。

鉴于 ResNet 优越的推理速度和不俗的性能，本文以此为基线，展开了不同的改进架构。

3 说话人验证模型 EIPFD-ResNet

为了促进信息在网络中传播，提升模型提取说话人嵌入能力。在本节中分别从基线模型中残差块比例、残差块结构、特征下采样方式以及最后的模型输出头四个方面对原始残差网络进行重新设计，分析由此对说话人验证任务的影响。出于计算量与参数量考虑，最后结合实验给出了基于深度残差网络 Half-ResNet34（通道数为原始 ResNet34 的一半）的更适合于说话人验证任务的模型 EIPFD-ResNet，其整体结构如图 1 所示，其中 IPBlock、下采样层和输出头具体结构分别见图 2（b）、图 3 和图 4（b），表 1 中给出了 EIPFD-ResNet 网络结构。

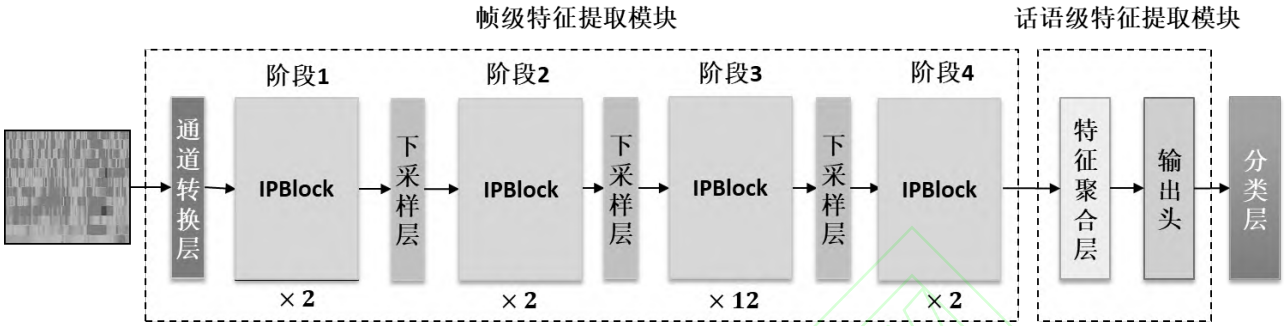


图 1 EIPDF-ResNet 模型整体结构图

Figure1 The architecture of the EIPDF-ResNet

3.1 残差块比率

ResNet 起源于图像领域，其残差块在模型各阶段分布比例主要是根据图像识别任务设计，可能对于说话人验证任务来说不是最优的。受 ConvNext[32]启发，本节以更大第三阶段残差块分布比例修改原始网络中残差块分布，将每个阶段的残差块数量由 Half-ResNet34 中的(3,4,6,3)调整为(2,2,6,2)、(3,3,9,3)、(2,2,12,2)以及(2,2,15,2)。探索残差块分布比例以及由此带来的模型深度与参数量改变对说话人验证任务的影响。

3.2 强调信息传播的残差块（IPBlock）

残差网络使用跳跃连接来解决深层网络产生的退化问题，但在 Ionut C D 等人[33]实验中，随着原始残差块的堆叠，模型深度增加，网络仍表现出优化的困难，这表明原始残差块的设计仍存在不足，过多的残差块仍会影响信息在网络中的传播。本文对原始残差块的结构重新设计，为方便描述，本文将原始残差块命名为 ResBlock，修改后的残差块命名为 IPBlock，图 2(a)给出了原始残差块的例子：在$\mathcal{F}(x^{[l]})$中包含两个卷积层（conv），其卷积核大小均为 3×3、两个批归一化层（BN）和一个激活层（$ReLU$），图中大箭头表示信息传播的最直接路径：主传播路径（在 ResBlock 主传播路径中包含跳跃连接过程），从公式上每个 ResBlock 可以定义为：

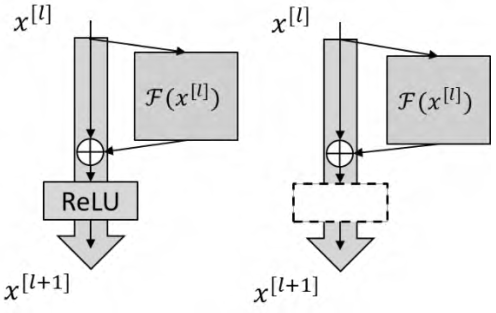
$$\hat{x}^{[l]} = \mathcal{F}\left(x^{[l]}, \{\mathcal{W}^{[l]}\}\right) \tag{1}$$

$$Z^{[l]} = \begin{cases} x^{[l]}, & size\left(\hat{x}^{[l]}\right) = size\left(x^{[l]}\right) \\ \mathcal{W}_p^{[l]} x^{[l]}, & size\left(\hat{x}^{[l]}\right) \neq size\left(x^{[l]}\right) \end{cases} \tag{2}$$

$$x^{[l+1]} = ReLU\left(\hat{x}^{[l]} + Z^{[l]}\right) \tag{3}$$

其中 $x^{[l]}$和$x^{[l+1]}$分别是第l个残差块的输入和输出特征，$ReLU$代表激活函数（激活层），$\mathcal{F}$代表可学习的残差映射函数，$\mathcal{W}^{[l]}$是残差映射中学习到的权重，$\hat{x}^{[l]}$是残差映射的结果，$\mathcal{W}_p^{[l]}$是跳跃连接中一个可学习的权重矩阵，它在$x^{[l]}$与$\hat{x}^{[l]}$尺寸不同时，将二者映射到同等大小，$Z^{[l]}$代表第l个残差块中跳跃连接的输出。对应于图 2（a），公式 1 代表右侧残差映射部分，公式 2 代表跳跃连接过程，公式 3 代表两部分信息在主传播路径中融合并向后传输。

如同在公式 3 和图 2（a）中看到的，负值信号在主传播路径上通过$ReLU$激活层后结果将归于 0，但在初期训练时网络中存在很多负权值，这意味着原始的残差块设计会阻碍特征信息的传递，导致说话人相关信息损失。由此本文分别去掉残差块中残差连接后的激活层以及主干网络中通道转换层中的激活层。通道转换层中的修改在表 1 中体现，去掉激活层的残差块：IPBlock 如图 2（b）所示（虚线框代表去掉了主传播路径中的激活层）。为防止这样设计的网络在特殊情况下（公式 1 结果为 0）主传播路径完全不受约束，给学习带来困难，下文中提到的方式会将信号变得“标准化”，从而稳定学习过程。



(a)ResBlock (b)IPBlock

图 2 原始 ResBlock 与 IPBlock 的区别

Figure2 The difference between the original ResBlock and the modified IPBlock

3.3 独立的下采样过程

使用跳跃连接来使主传播路径中$x^{[l]}$的维度与残差映射$\mathcal{F}$的输出：$\hat{x}^{[l]}$的维度对齐对于模型来说是有害的。在说话人验证任务中，基于残差网络的模型在进行下采样过程中通常在$\mathcal{F}$中采用步长为 2，卷积核大小为 3×3，边缘填充为 1 的

表 1 Half-ResNet34 结构与本文提出的 EIPFD-ResNet 结构对比

Table 1 The Structure Difference Between the Half-ResNet34 and the EIPFD-ResNet

层	Half-ResNet34		EIPFD-ResNet	
	结构	特征图输出尺寸	结构	特征图输出尺寸
输入	-	$1\times F\times T$	-	$1\times F\times T$
通道数转换	$(3\times 3),32,1$ ReLU BN	$32\times F\times T$	$(3\times 3),32,1$ BN	$32\times F\times T$
阶段 1	$\{ResBlock,32,1\}\times 3$	$32\times F\times T$	$\{IPBlock,32,1\}\times 2$	$32\times F\times T$
阶段 2	$\{ResBlock,64,1\}\times 4$	$64\times\frac{F}{2}\times\frac{T}{2}$	$\{IPBlock,64,1\}\times 2$	$64\times\frac{F}{2}\times\frac{T}{2}$
阶段 3	$\{ResBlock,128,1\}\times 6$	$128\times\frac{F}{4}\times\frac{T}{4}$	$\{IPBlock,128,1\}\times 12$	$128\times\frac{F}{4}\times\frac{T}{4}$
阶段 4	$\{ResBlock,256,1\}\times 3$	$256\times\frac{F}{8}\times\frac{T}{8}$	$\{IPBlock,256,1\}\times 2$	$256\times\frac{F}{8}\times\frac{T}{8}$
特征聚合	注意力统计池化[34]	$64F\times 1$	注意力统计池化	$64F\times 1$
输出头	$FC(64F,256)$	256	FDHead	256
损失函数	AAM	说话人数量	AAM	说话人数量

卷积操作，以及在主传播路径中使用步长为 2 卷积核大小为 1×1 的跳跃连接，以此保持下采样过程中$x^{[l]}$的维度与$\hat{x}^{[l]}$的维度匹配，即将$x^{[l]}$与$\hat{x}^{[l]}$的时频维度以及信道维度对齐。不难考虑到在跳跃连接中，由于卷积核大小仅有 1×1 但卷积步长却为 2，代表着在特征维度改变过程中，跳跃连接过程使得$x^{[l]}$失去了 75%的激活，这将会导致大量的信息损失[33]，同时$x^{[l]}$剩余激活部分在选择过程没有经过约束，因此不能保证其激活后的输出是有意义的。最后跳跃连接的结果会添加到对应残差块的输出中，意味着主干信息流中将会引入噪声和信息损耗，对网络中的信息造成负面影响。

为解决上述问题，如图 3 所示，本文将下采样操作从残差块中剥离开，使用单独的下采样层来满足维度变换的需求。在第 1、2 和 3 阶段结束时采用步长为 2，卷积核大小为 2×2 的卷积层来对时频维度和信道维度变换。通过使卷积核大小与步长大小一致来考虑$x^{[l]}$中所有的信息，使元素间的过度更平滑，减少信息损失。批归一化层用来规范信号，减少模型学习困难，保持模型训练过程的稳定性[32]。此外下采样层的作用还在于防止 3.2 中提到的去掉主传播路径上所有激活层后，信息在极端情况下不受任何约束的通过网络。在实验部分展示了单独使用下采样层与 3.2 中方法结合在性能上的好处。

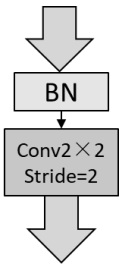


图 3 下采样层结构

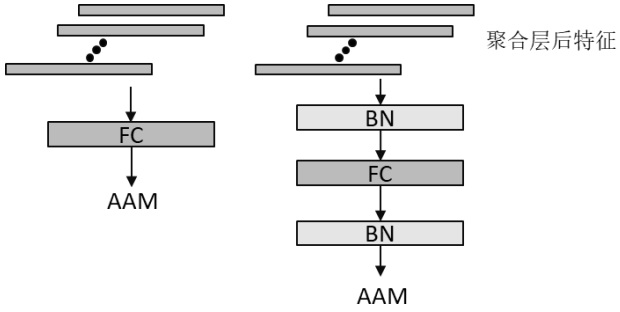
Figure3 The architecture of down sampling layer

3.4 强调特征分布的模型输出头（FDHead）

如图 4（a）所示，许多最先进的说话人验证模型在模型输出头后使用 AAM [35]（Additive Angular Margin Softmax，AAM）来约束说话人嵌入。AAM 如公式 4 所示：

$$L=-\frac{1}{n}\sum_{i=1}^{n}log\frac{e^{s(\cos(\theta_{y^i}+m))}}{e^{s(\cos(\theta_{y^i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{scos\theta_j}} \quad (4)$$

其中 n 代表说话人个数，θ_{y^i} 是当前语句嵌入与其对应说话人类中心夹角，θ_j是当前语句嵌入与其他说话人类中心夹角，s 和 m 是两个超参数，s 代表尺度，该参数目的是将 cos值增大s倍，方便 AAM 提高差异性，m 为子空间角度间隔，间隔越大则表明不同说话人之间的分类间隔越大，越利于分类。

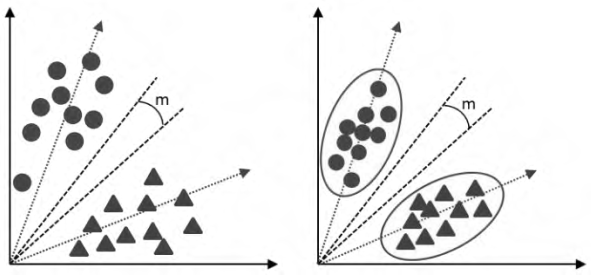


（a）原始输出头　　（b）FDHead 输出头

图 4 原始输出头与 FDHead 的区别

Figure4 The difference between original head and FDHead

由公式 4 可知，AAM 在特征空间内使用余弦角度构造一系列决策边界，把不同说话人的特征分配到角度间隔为 m 的不同子空间中，如图 5（a）所示。最理想的情况是最小类间角度大于最大类内角度，即除 AAM 强制类间存在角度间隔外，希望类内特征分布尽可能紧凑，然而聚合层后的特征在欧式空间内，特征分布较为松散[36]，这可能会给 AAM 优化带来困难。受 Liu W 等人启发[37]，如图 4（b）所示，本文在生成说话人嵌入的全连接层前后分别添加 BN 层来平滑嵌入空间的特征分布[36]，减少特征分布的自由区域。对于 AAM，Softmax 使得特征倾向于仿射状分布时，导致靠近仿射中心的特征缺乏清晰的决策面并且难以区分，但 BN 层可以使特征保持紧凑分布[36]的同时，使得特征空间内话语特征更靠近其对应的说话人类中心（图 5 中虚线箭头），从而得到更清晰的分类决策面，帮助 AAM 更好的约束特征。同时 BN 也能起到正则化效果，预防过拟合，经过 FDHead 后的特征分布示意图如图 5（b）所示。



(a)原始特征分布　　(b)FDHead 后特征分布

图 5 说话人嵌入特征分布示意图

Figure5 The illustration of speaker embedding feature distribution

4 实验设置与细节

4.1 数据集与特征提取

(1)数据集。为了评估本文所提出方法对说话人特征提取的有效性，本文在三个公开数据集 CN-Celeb[38]、VoxCeleb1 和 VoxCeleb2 数据集[39]进行了实验。

VoxCeleb：如表 2 所示，包含 VoxCeleb1 和 VoxCeleb2 两个数据集。这两个数据集均是从 Youtube 网站中提取的大规模文本独立数据集。VoxCeleb1 开发集包含 1221 位名人的 148642 条访问语音，VoxCeleb2 开发集包含 5994 位名人的 1092009 条访问语音。两个数据集之间没有重复。由于算力原因，本文绝大部分实验基于 VoxCeleb1 训练，使用 VoxCeleb-O 评估。为了与目前最优说话人验证模型对比，本文最后以 VoxCeleb2 为训练集，分别在 VoxCeleb-O、VoxCeleb-E 和 VoxCeleb-H 验证集上做了实验验证。VoxCeleb-O 评估集包含 40 名说话人，VoxCeleb-E 评估集包含了整个 VoxCeleb1 开发集与 VoxCeleb-O 中所有说话人，其测试语句更多，结果更具代表性。而 VoxCeleb-H 内则是包含 VoxCeleb1 中相同国籍，相同性别的说话人，对于说话人验证任务来说这个评估集相较于另外两个更困难。

表 2 VoxCeleb：训练集与评估集

Table 2 VoxCeleb: Training Set and Evaluation Set

数据集	说话人	话语条数	验证对
VoxCeleb1-dev	1211	148642	-
VoxCeleb2-dev	5994	1092009	-
VoxCeleb-O	40	4708	37611
VoxCeleb-E	1251	145160	579818
VoxCeleb-H	1190	137924	550894

CN-Celeb：如表 3 所示，同样是 CN-Celeb1 和 CN-Celeb2 两个数据集，本文使用 CN-Celeb2 作为训练集，它是从哔哩哔哩、网易云、喜马拉雅、抖音以及唱吧等平台收集的包含娱乐、访问、唱歌、戏剧、电影、视频博客、现场直播等 11 个不同场景下的声音数据，涉及许多真实的噪音、信道失配和真实的讲话风格，包含 1996 人的超过 500000 条语音，相较于 VoxCeleb 中只包含访问类型的语音，CN-Celeb 更具无约束条件的代表性。本文实验中使用来自 CN-Celeb1 中的 CN-Celeb(E)作为评估集，包含 200 个说话人的 18024 条语音。此外该评估集的验证对中注册语句与测试语句的场景不匹配，以及数据集中包含的大量短语音数据使得该数据集在说话人验证任务中非常具有挑战性。

表 3 CN-Celeb2：训练集与评估集

Table 3 CN-Celeb2:Training Set and Evaluation Set

数据集	说话人	话语条数	验证对
CN-Celeb2	1996	524787	-
CN-Celeb(E)	200	18024	3484292

(2)特征提取。本文所有基于 ResNet 的模型使用 64 维对数梅尔滤波器能量(F-bank)作为输入特征。使用长度为 25ms,窗口长度为 10ms 的汉明窗从输入音频中提取 F-bank。每段音频使用 200 帧的随机块作为网络输入,并且不应用语音活动检测(Voice Activity Detection, VAD)。输入特征是在帧级别上的均值。所有实验在特征提取阶段均结合噪音数据集[40]（气泡音,噪音，混响）做数据增强[41]。最后对提取出的 F-bank 应用频谱增强[42]。

4.2 训练设置

实验是基于 Pytorch 深度学习框架下完成的，本文使用说话人验证任务中常用的模型 Half-ResNet34 作为基线模型，采用注意力统计池化[33]（Attention Statistics Pooling，ASP）对模型提取出的帧级特征聚合，中间通道维度设置为 128。使用 AAM 作为模型的监督损失。

在训练阶段，每次迭代批大小设置为 256，学习率初始值设置为 1e-3 并以每个周期 0.02 的衰减率衰减,选用 Adam 优化器并将其权重衰减设置为 2e-5。AAM 的超参数尺度和间隔分别设置为 30 和 0.2。在评估阶段应用了测试时间增强（Test Time Augment，TTA）方法[43]，通过重叠裁剪从单个话语中提取 10 个说话人嵌入,将 10 个嵌入的平均值作为最终的说话人嵌入。本文使用余弦相似度作为评判标准。EER 和最小检测代价函数[44]（Minimum Detection Cost Function，minDCF）作为性能指标，P_{target}=0.01,C_{miss} = C_a=1，参数量用来衡量模型大小，下文中 EER，minDCF,参数量三个评价指标越小均代表模型越具优越性。

5 结果与讨论

为了彻底评估本文所提出的方法，在本节中首先在 VoxCeleb1 上实验，证明了第三节中提出的 4 种方法的有效性。接着在三个不同数据集上分别实验与其余说话人验证模型对比，进一步证明 EIPFD-ResNet 的优越性。

表 4 内容展示的是在原始 Half-ResNet 模型基础上逐步增加第三节中提出的残差块分布比例、IPBlock、单独下采样层以及 FDHead 模块，并给出了实验结果。

（1）从模型深度与残差块分布比例结果可以看出原始 ResNet 中残差块的分布相对于更改后的残差块分布在说话人验证任务中表现并不出色；从(2,2,6,2)分布与(3,3,9,3)分布结果对比可知,尽管增加模型深度对模型会有一定增益,但这部分增益可能是由于参数量增加带来的；通过观察残差块按照(3,3,9,3)分布与(2,2,12,2)分布结果可以发现，在参数量相对一定时，更大的第三阶段残差块比例对与说话人验证任务来说更有益，这是由于该阶段可以保留特征结构信息前提下使其推理能力达到最强；在(2,2,12,2)分布基础上进一步扩展第三阶段残差块数量后，EER 没有进一步改善，原因如本文 3.2 中描述的使用原始残差块构建网络时，主路径上激活层数量与网络的深度呈正比关系，在网络堆叠过深时会妨碍信息传输，导致网络优化困难。因此本文基于(2,2,12,2)分布完成后续实验。

（2）减少激活层数量后，即使用 IPBlock 替代 ResBlock 并去掉通道数转换层中的激活层，EER 进一步达到了 2.37，相对于修改残差块分布后的模型降低了 4.4%，证明主传播路径中没有激活层对语音特征在模型中的传递更有益。

（3）尽管在使用单独下采样层来缩放模型特征图后 EER 改善不明显，但同时使用单独下采样层和减少激活层数量后，EER 相对降低了 8.5%，大于两者对于模型提升之和，证明了 3.3 中所描述的下采样层可以对主路径信息起到约束作用，从而与 IPBlock 和去掉激活层的通道转换层达到互补效果。此外使用下采样层后，模型可以更好的收敛并加快收敛过程，图 6 展示了使用单独下采样层策略后相对于原始下采样方式收敛情况对比，为了表示更清晰，图中没有展示前 10 轮迭代的收敛过程。图中横轴代表迭代轮数，纵轴代表训练损失，实线是使用原始下采样方式结果，虚线是使用下采样层的结果。

（4）表 4 的最后在上述结果基础上使用本文提出的输出头来生成说话人嵌入，EER 进一步下降了 0.07，这表明 FC 层前后的批归一化层对模型提取特征有效，增强了损失函数对模型的约束。为了更好表现模型提取出的特征分布，本节对模型输出特征作了可视化处理，但由于语音基线数据集中

表 4 本文中提出方法在 VoxCeleb1 数据集上的实验结果

Table 4 The Experimental Results of the Method Proposed in this Paper on VoxCeleb1 Dataset

	残差块分布	EER（%）	minDCF	参数量
方法	Half-ResNet	2.70	0.3220	6.579M
模型深度与残差块分布比例	3 4 6 3	2.70	0.3220	6.579M
	2 2 6 2	2.64	0.3398	**5.295M**
	3 3 9 3	2.57	0.2947	7.354M
	2 2 12 2	**2.48**	**0.3048**	6.985M
	2 2 15 2	2.49	0.3057	7.830M

A.更少激活层	2 2 12 2	**2.37**	0.2946	6.985M
B.单独下采样层	2 2 12 2	**2.45**	0.3186	7.486M
A+B	2 2 12 2	**2.27**	**0.2532**	7.486M
A+B+ FDHead	EIPFD-ResNet	**2.20**	0.2587	7.486M

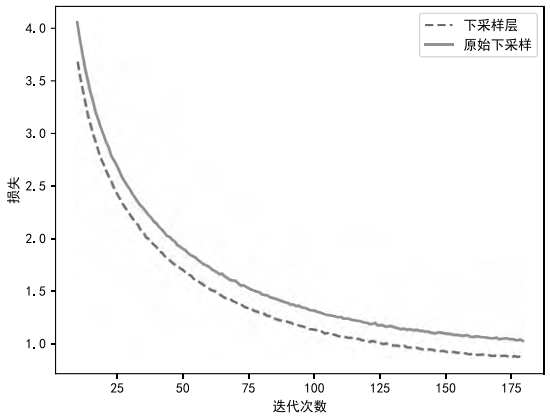


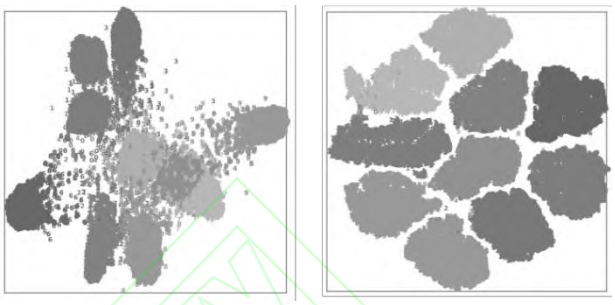
图 6 原始下采样方式与下采样层收敛情况对比

Figure6 The difference of convergence between original down sampling method and down sampling layer

人物数量多，每个人对应的话语很少且分布不均匀，可视化过程中会产生很大的噪音。参考 Hao Luo 等人解决方式[35]，本节使用 MNIST 数据集训练模型来可视化特征空间内的特征分布，原因是相比于语音基线数据集，MNIST 数据集仅有 10 个类别且每个类由平均 600 个样本组成，可使特征分布清晰稳健。图 7 展示了在 MINIST 数据集上特征可视化后的结果，可以看出，经过 FDHead 输出头后的特征分布相对于改进前特征分布更紧凑，决策面相对更清晰。

相较于改进前的模型 Half-ResNet，第三节提出的四个方法改进后的模型：EIPDF-ResNet 在 VoxCeleb1 数据集上 EER 达到了 2.20，总体获得了 17.3%的提升。

为进一步展示 EIPDF-ResNet 的优越性，本节根据数据集大小及数据复杂程度，分别在 VoxCeleb1（规模小）、VoxCeleb2（规模大）、CnCeleb2（复杂场景）三个数据集上进行模型评估。



（a）原始输出头后特征分布（b）FDHead 输出头后特征分布

图 7 原始输出头与修改后的输出头在 MNIST 数据集的特征分布上的可视化比较

Figure7 The visual difference of feature distribution between the original output header and the modified output header on MNIST dataset

（1）表 5 给出了 EIPFD-ResNet 在大数据集下的实验结果，可以发现本文提出的 EIPFD-ResNet 尽管仅使用了 7.486M 的参数量，但仍表现出强大的表征能力，取得了最优结果。相对于该数据集中最广泛使用的说话人验证模型：ResNet34-SE[30]，EIPFD-ResNet 的 EER/minDCF 在 VoxCeleb-O、VoxCeleb-E 和 VoxCeleb-H 评估集上分别相对降低了 19.1%/6.1%、43.5%/46.7%和 38.8%/34.1%，并且显著优于当前最优说话人验证基线模型 ECAPA-TDNN，在三个评估集上，EER/minDCF 分别相对降低了 9.7%/33.3%、8.5%/13.0%和 13.2%/14.3%。

（2）表 6 给出了 EIPFD-ResNet 在小数据集上与常用的基于残差网络结构的话人验证模型对比。可以发现本文所提出的 EIPFD-ResNet 在性能上显著优于其余基于残差网络的说话人验证模型。相较于传统 ResNet34 模型，EER/minDCF 相对改善了 16.4%/18.4%。

表 5 在 VoxCeleb2 数据集上实验结果

Table 5 Experimental Results on VoxCeleb2 Dataset

模型	频率维度	参数量	VoxCeleb-O		VoxCeleb-E		VoxCeleb-H	
			EER(%)	minDCF	EER(%)	minDCF	EER(%)	minDCF
ResNet-SE	64	23.599M	1.26	0.1264	2.09	0.2548	3.53	0.3420
ECAPA-TDNN	80	14.729M	1.13	0.1780	1.29	0.1557	2.49	0.2633
EIPFD-ResNet	64	**7.486M**	**1.02**	**0.1187**	**1.18**	**0.1355**	**2.16**	**0.2254**

表 6 在 VoxCeleb1 数据集上实验结果

Table 6 Experimental Results on VoxCeleb1 Dataset

模型	EER(%)	minDCF	参数量
Half-ResNet34	2.70	0.3220	**6.579M**
ResNet34[23]	2.63	0.3172	23.295M
ResNet18[30]	2.64	0.3019	13.655M
EIPFD-ResNet	**2.20**	**0.2587**	7.486M

（3）表 7 给出了 EIPFD-ResNet 在复杂场景下的性能表现。EIPFD-ResNet 在 CN-Celeb2 上的 EER/minDCF 结果达到了 9.02/0.5233，相比于 ResNet-34 模型与 ECAPA-TDNN 模型，EER /minDCF 分别相对降低了 6.0%/7.8%和 9.0%/8.4%。

表 7 在 CN-Celeb2 数据集上实验结果

Table 7 Experimental Results on CN-Celeb2 Dataset

模型	EER(%)	minDCF
ResNet34[16]	9.60	0.5671
ECAPA-TDNN	9.91	0.5710
EIPFD-ResNet	**9.02**	**0.5233**

6 结　语

残差网络被广泛应用于说话人验证任务中。本文对残差中的信息流及输出特征进行分析，针对其存在的信息传播受限，容易引入噪声信息，提取出的特征难以分类等问题，对残差块分布、残差块结构、特征下采样以及模型输出头进行了更合理的设计，提出了基于残差网络的说话人特征提取模型 EIPFD-ResNet。在保持优越推理速度的同时提高了捕捉说话人本质特征的能力，并在多个数据集上均取得了显著效果，为说话人验证任务提供了一个强有力的基线模型。未来计划从语音信息中特有的时间与频率信息出发对全局信息建模，从而进一步提升模型性能。

References:

[1] Bimbot F, Bonastre J F, Fredouille C, et al. A tutorial on text-independent speaker verification[J].EURASIP Journal on Advances in Signal Processing,2004,4(1):1-22.

[2] Lee K A, Larcher A, Thai H, et al. Joint application of speech and speaker recognition for automation and security in smart home[C]//Proceedings of Annual Conference of the International Speech Communication Association,2011:3317-3318.

[3] Crocco M, Cristani M, Trucco A, et al. Audio surveillance: a systematic review[J].ACM Computing Surveys,2016,48(4):1-46.

[4] Hershey J R, Chen Z, Le ROUX J, et al. Deep clustering: discriminative embeddings for segmentation and separation [C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP),IEEE,2016:31-35.

[5] Chen Z, Luo Y, Mesgarani N. Deep attractor network (DANet) for single-channel speech separation[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing(ICASSP),IEEE,2017:246-250.

[6] Yu D, Kolbæk M, Tan Z H, et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing(ICASSP),2017:241-245.

[7] Arik S, Chen J, Peng K, et al. Neural voice cloning with a few samples[C]//Advances in Neural Information Processing Systems(NIPS),2018:10019-10029.

[8] Jia Y, Zhang Y, Weiss R, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C]//Proceedings of Advances in Neural Information Processing Systems(NIPS),2018:4480-2290.

[9] Xu Zhi-hang, Chen Bo, Zhang Hui, et al. Speech synthesis adaption method based on phoneme-level speaker embedding under small data[J].Chinese Journal of Computers.2022,5(45):1003-1017.

[10] Wang Quan, Muckenhirn H, Wilson K, et al. VoiceFilter: targeted voice separation by speaker-conditioned spectrogram masking[C]//Annual Conference of the International Speech Communication Association(Interspeech),2019:2728-2732.

[11] Žmolíková K, Delcroix M, Kinoshita K, et al. SpeakerBeam: speaker aware neural network for target speaker extraction in speech mixtures[J].IEEE Journal of Selected Topics in Signal Processing,2019,13(4):800-814.

[12] Mclaughlin J, Reynolds D A, Gleason T P.A study of computation speed-UPS of the GMM-UBM speaker recognition system[C]//Proceedings of Sixth European Conference on Speech Communication and Technology,1999.

[13] Matějka P, Glembek O, Castaldo F, et al. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),IEEE,2011:4828-4831.

[14] Huang Jun, Jiang Bing, Li Xian-gang, et al. I-Vector clustering dictionary and attention mechanism framework for speaker adaptation[J].Journal of Chinese Computer Systems,2019,40(2):460-464.

[15] He Liang, Yang Li, Liu Jia. TLS-NAP algorithm for text-independent speaker recognition[J].Pattern Recognition and Artificial Intelligence.2012,6(25):917-921.

[16] Snyder D, Garcia-Romero D, Povey D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Proceedings of Conference of the International Speech Communication Association (Interspeech),2017:999-1003.

[17] Long Hua, Qu Yu-quan, Duan Ying. Short utterance speaker embedding vector algorithm based on kernel canonical correlation analysis[J].Journal of Chinese Computer Systems,2021,42(11):2269-2275.

[18] Hossein Z, Wang Shuai, Anna S et al. But system description

to voxceleb speaker recognition challenge 2019[R].The VoxCeleb Challange Workshop,2019.

[19] Zhou Jian-feng, Jiang T, Li Z, et al. Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function[C]//Annual Conference of the International Speech Communication Association (Interspeech),2019:2883-2887.

[20] Li Chao, Ma X, Jiang B, et al. Deep speaker: an end-to-end neural speaker embedding system[J].arXiv preprint arXiv:1705.02304,2017.

[21] Chung J S, Huh J, Mun S. Delving into VoxCeleb: environment invariant speaker recognition[J].A arXiv preprint arXiv:1910.11238,2019.

[22] Joon S C, Jaesung H, Seongkyu M, et al. In defence of metric learning for speaker recognition[C]//Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2020:2977-2981.

[23] Nikita Torgashov.Id r&d system description to VoxCeleb speaker recognition challenge 2020[J].A arXiv preprint arXiv:2010.12468,2020.

[24] Nagrani A, Chung J S, Huh J, et al. VoxSRC 2020:the second VoxCeleb speaker recognition challenge[J].arXiv preprint arXiv:2012.06867,2020.

[25] 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 the International Speech Communication Association,2020:3830-3834.

[26] Heo H S, Lee B J, Huh J, et al. Clova baseline system for the VoxCeleb speaker recognition challenge 2020[J].arXiv preprint arXiv:2009.14153,2020.

[27] Yao Wei, Chen Shen, Cui Jia-min, et al. Multi-stream Convolutional neural network with frequency selection for robust speaker verification[J].arXiv preprint arXiv:2012.11159,2020.

[28] Jenthe T, Brecht D, Kris D, Integrating frequency translational invariance in TDNNs and frequency positional information in 2D ResNets to enhance speaker verification[C]//Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2021:2302-2306.

[29] Brown A, Huh J, Chung J S, et al. VoxSRC 2021:the 3rd VoxCeleb speaker recognition challenge[J].arXiv preprint arXiv:2201.04583,2022.

[30] Wang Jie, Tong F, Chen Z, et al. XMUSPEECH system for VoxCeleb speaker recognition challenge 2021[J].arXiv preprint arXiv:2109.02549,2021.

[31] Chen Chen, Han Ji-qing, Chen De-yun, et al. Utterance-level feature extraction in text-independent speaker recognition: a review[J].Acta Automatica Sinica,2022,3(48):664-688.

[32] Liu Zhuang, Mao H, Wu C Y, et al. A ConvNet for the 2020s[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2022:11976-11986.

[33] Duta I C, Liu L, Zhu F, et al. Improved residual networks for image and video recognition[C]//Proceedings of International Conference on Pattern Recognition (ICPR),IEEE,2021:9415-9422.

[34] Okabe K, Koshinaka T, Shinoda K. Attentive statistics pooling for deep speaker embedding[C]//Proceedings of Annual Conference of the International Speech Communication Association(Interspeech),2018:2252-2256.

[35] Joon Son Chung, Jaesung H, Seongkyu M, et al. In defence of metric learning for speaker recognition[C]//Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2020:2977-2981.

[36] Luo Hao, Jiang Wei, Gu Youzhi, et al. A strong baseline and batch normalization neck for deep person re-identification[J].IEEE Transactions on Multimedia,2019,22(10):2597-2609.

[37] Liu Wei-yang, Wen Ya-dong, Yu Zhi-ding, et al. Sphereface: deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:212-220.

[38] Li Lan-tian, Liu Rui-qi, Kang Jia-wen, et al. CN-Celeb: multi-genre speaker recognition[J].Speech Communication,2022,137(1):77-91.

[39] Nagrani A, Chung J S, Xie W, et al. Voxceleb: large-scale speaker verification in the wild[J].Computer Speech & Language,2020,60(1):101027.doi:10.1016/j.csl.2019.101027.

[40] Snyder D, Chen G, Povey D. Musan: a music, speech, and noise corpus[J].arXiv preprint arXiv:1510.08484,2015.

[41] Ko T, Peddinti V, Povey D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing(ICASSP),IEEE,2017:5220-5224.

[42] Park D S, Chan W, Zhang Y, et al. Specaugment: a simple data augmentation method for automatic speech

recognition[C]// Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2019,2613-2617.

[43] Chung J S,Nagrani A ,Zisserman A.VoxCeleb2:deep speaker recognition[C]//Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2018,1086-1090.

[44] Sadjadi S O, Kheyrkhah T, Tong A, et al. The 2016 NIST speaker recognition evaluation[C]//Proceedings of Annual Conference of the International Speech Communication Association (Interspeech),2017:1353-1357.

附中文参考文献：

[9] 徐志航,陈　博,张　辉,等.小数据下的因素级别说话人嵌入的语音合成自适应方法[J].计算机学报,2022,5(45):1003-1017.

[14] 黄　俊,蒋　兵,李先刚,等.I-vector聚类字典及注意力机制框架的说话人自适应[J].小型微型计算机系统,2019,40(2):460-464.

[15] 何　亮，杨　毅，刘　加.基于TLS-NAP的文本无关说话人识别算法[J].模式识别与人工智能,2012,25(6):916-921.

[17] 龙　华,瞿于荃,段　荧.一种基于核典型关联分析的短语音说话人嵌入向量算法[J].小型微型计算机系统,2021,42(11):2269-2275.

[31] 陈　晨，韩纪庆，陈德运,等.文本无关说话人识别中句级特征提取方法研究综述[J].自动化学报,2022,48(3):664-688.