工程科学学报, 第 45 卷, 第 11 期: 1962−1969, 2023 年 11 月
Chinese Journal of Engineering, Vol. 45, No. 11: 1962−1969, November 2023
https://doi.org/10.13374/j.issn2095-9389.2022.09.19.001; http://cje.ustb.edu.cn

基于 SE-DR-Res2Block 的声纹识别方法

李　平[1,2)], 高清源[1,2)], 夏　宇[1,2)], 张小勇[1,2)], 曹　毅[1,2)]✉

1) 江南大学机械工程学院, 无锡 214122　2) 江苏省食品先进制造装备技术重点实验室, 无锡 214122
✉通信作者, E-mail: caoyi@jiangnan.edu.cn

摘　要　针对声纹识别领域中基于传统 Res2Net 模型特征表达能力不足、泛化能力不强的问题, 提出了一种结合稠密连接与残差连接的特征提取模块 SE-DR-Res2Block(Sequeeze and excitation with dense and residual connected Res2Block). 首先, 介绍了应用传统 Res2Block 的 ECAPA-TDNN(Emphasized channel attention, propagation and aggregation in time delay neural network) 网络结构和稠密连接及其工作原理; 然后, 为实现更高效的特征提取, 采用稠密连接进一步实现特征的充分挖掘, 基于 SE-Block(Squeeze and excitation block) 将残差连接和稠密连接相结合, 提出了一种更高效的特征提取模块 SE-DR-Res2Net. 该模块以一种更细粒化的方式获得不同生长速率和多种感受野的组合, 从而获取多尺度的特征表达组合并最大限度上实现特征重用, 以实现对不同层特征的信息进行有效提取, 获取更多尺度的特征信息; 最后, 为验证该模块的有效性, 基于不同网络模型采用 SE-Res2Block(Sequeeze and excitation Res2Block)、FULL-SE-Res2Block(Fully connected sequeeze and excitation Res2Block)、SE-DR-Res2Block、FULL-SE-DR-Res2Block(Fully connected sequeeze and excitation with dense and residual connected Res2Block), 分别在 Voxceleb1 和 SITW(Speakers in the wild) 数据集开展了声纹识别的研究. 实验结果表明, 采用 SE-DR-Res2Block 的 ECAPA-TDNN 网络模型, 最佳等错误率分别达到 2.24% 和 3.65%, 其验证了该模块的特征表达能力, 并且在不同测试集上的结果也验证了其具有良好的泛化能力.

关键词　深度学习; 声纹识别; 稠密网络; 残差网络; 多尺度特征
分类号　TN912.34

Voiceprint recognition method based on SE-DR-Res2Block

LI Ping[1,2)], *GAO Qingyuan*[1,2)], *XIA Yu*[1,2)], *ZHANG Xiaoyong*[1,2)], *CAO Yi*[1,2)]✉

1) School of Mechanical Engineering, Jiangnan University, Wuxi 214122, China
2) Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, China
✉Corresponding author, E-mail: caoyi@jiangnan.edu.cn

ABSTRACT　Aiming at the problems of insufficient feature expression ability and weak generalization ability of the traditional Res2Net model in the field of voice print recognition, this paper proposes a feature extraction module known as the SE-DR-Res2Block, which combinedly uses dense connection and residual connection. The combination of low-semantic features with spatial information characteristics allows focusing more on detailed information and high-semantic information that concentrates on global information as well as abstract features. This can compensate for the loss of some detailed information caused by abstraction. First, the feature of each layer in the dense connection structure is derived from the feature output of all previous layers to realize feature reuse. Second, the structure and working principle of the ECAPA-TDNN network using traditional Res2Block is introduced. To achieve more efficient feature extraction, the dense connection is used to further realize full feature mining. Based on SE-block, a more efficient feature extraction module, SE-DR-Res2Net, is proposed by combining the residual join and dense links. As compared to the traditional SE-

收稿日期: 2022−09−19
基金项目: 国家自然科学基金资助项目 (51375209); 江苏省“六大人才高峰”计划项目 (ZBZZ-012); 江苏省研究生创新计划项目 (KYCX18_0630, KYCX18_1846); 高等学校学科创新引智计划资助项目 (B18027)

Block structures, the convolutional layers are used here instead of fully connected layers. Because they not only reduce the number of parameters needed for training but also allow weight sharing, thereby reducing overfitting. Therefore, effective extraction of feature information from different layers is essential for obtaining multiscale expression as well as maximizing the reuse of features. During the collection of more scale-specific feature information, a large number of dense structures can lead to a dramatic increase in parameters and computational complexity. By using partial residual structures instead of dense structures, we can effectively prevent the dramatic increase in parameter quantity while maintaining the performance to a certain extent. Finally, to verify the effectiveness of the module, SE-Res2block, Full-SE-Res2block, SE-DR-Res2block, and Full-SE-DR-Res2block are adopted based on the different network models. Voxceleb1 and SITW (speakers in the wild) datasets were used for Voxceleb1 and SITW, respectively. The performance comparison of Res2Net-50 models with different modules on the Voxceleb1 dataset shows that SE-DR-Res2Net-50 achieves the best equal error rate of 3.51%, which also validates the adaptability of this module on different networks. The usage of different modules on different networks, as well as experiments and analyses conducted on different datasets, were compared. The experimental results showed that the optimal equal error rates of the ECAPA-TDNN network model using SE-DR-Res2block had reached 2.24% and 3.65%, respectively. This verifies the feature expression ability of this module, and the corresponding results based on different test data sets also confirm its excellent generalization ability.

KEY WORDS　deep learning; voiceprint recognition; dense connection; residual connection; multiscale features

声纹识别是一种现代生物识别技术，其通过转换仪器将收集的声波特征转化成相应的波谱图形并与已经存储的波谱图形进行对比，从而辨别是否属于同一个体以实现身份验证的功能[1]. 声纹识别是语音处理领域的热点研究方向之一，其可实现计算机准确识别说话人的语音信息，进而分析语音中的声纹信息，进一步提高了计算机的语音处理能力. 声纹识别具有非接触式、便利性高、安全性高、识别成本低、可远程确认等优点，因此声纹识别技术被广泛应用于银行交易和远程支付的信息安全[2]、调查嫌疑人是否有罪[3−5]、自动身份标记[6] 等领域.

针对声纹识别技术，国内外诸多学者分别基于传统机器学习、深度学习两类方法开展了大量的实验与理论研究. 其中，基于传统机器学习，Burget 等[7] 提出特征信道自适应以降低信道干扰；鲍焕军与郑方[8] 提出高斯混合模型−通用背景模型（Gaussian mixture model−Universal background model, GMM−UBM），采用多个 GMM 模型来拟合不同的说话人；Kenny 等[9] 提出联合因子分析，采用 GMM 超矢量空间的子空间进行重新建模以消除信道差异的干扰；Cumani 等[10] 提出新的概率线性判别分析以用于短语音的识别. 近年来，随着深度学习方法的不断深入，声纹识别技术也取得了飞跃性的进展. 谷歌提出通过深度神经网络训练，提出了 d-vector 作为说话人特征，并对说话人在帧级别进行分类[11]；Snyder 等[12] 结合 d-vector 和时间延迟神经网络[13]，提出了能够有效表示包含上下文信息的语句级 x-vector 说话人特征；Okabe 等[14] 通过引入一种新的注意力机制以捕获声纹的长期变化；Jiang 等[15] 提出将稠密连接卷积网络[16] 与门控机制相融合的 DDB+Gate(Dilated dense blocks and gate blocks) 网络，其采用扩张滤波器以获取更多的时频上下文信息，并通过前馈方式的稠密连接来收集上下文信息；Zhou 等[17] 通过 ResNet-SE(Residual network with sequeeze and excitation) 和 AS-Softmax(Additive supervision softmax) 相结合的 ResNet-34-SE 系统，利用错误分类样本的先验知识提升分类能力；Li 等[18] 通过引入一种带典型相关分析约束的多特征学习策略，最大化相同说话人话语的相关性.

综上所述，尽管诸多学者对声纹识别开展了较为深入的研究，但必须指出的是：(1) 当前国内外对声纹识别技术的研究仍存在待解决的技术难点；(2) 已有模型缺乏对声纹低语义特征的关注，其导致模型特征表达能力不足、泛化能力不强. 文献 [19] 提出的 ECAPA-TDNN（Emphasized channel attention, propagation and aggregation in time delay neural network）神经网络采用的是 Res2Net[20] 中的 Res2Block，其具有更大的感受野，虽可获取不同尺度的特征，但仍缺乏对低语义特征的关注，其中低语义特征是指浅层网络中包含大量空间信息、更注重细节信息的特征，高语义特征更集中于全局信息. 文献 [16] 提出了稠密连接神经网络，其通过层与层之间的稠密连接来达到特征重用的目的，但增加了模型的大小和计算复杂度. 因此，本文在 Res2Block 基础上，首先将 DenseNet 中的特征重用应用于 Res2Block；其次为进一步提升泛化能力，本文基于通道特征响应 SE-Block (Squeeze and excitation

block) 模型的思想对 Res2Block 进行改进, 进而提出一种基于稠密连接、残差连接和通道特征响应的特征提取模块 SE-DR-Res2Block (Sequeeze and excitation with dense and residual connected Res2Block).

1　ECAPA-TDNN 网络和稠密连接结构

1.1　ECAPA-TDNN 网络

ECAPA-TDNN 是一种基于时间延迟神经网络(TDNN)的声纹提取器, 其工作原理是首先通过将 TDNN 和传统的残差模块 Res2Block 相结合, 形成一维的 Res2Block 以期获取时间上下文信息; 其次, 添加 SE-Block 来改善信道特征信息, 形成 SE-Res2-Block(Sequeeze and excitation Res2Block), 如图 1 所示, 其具体工作流程为:

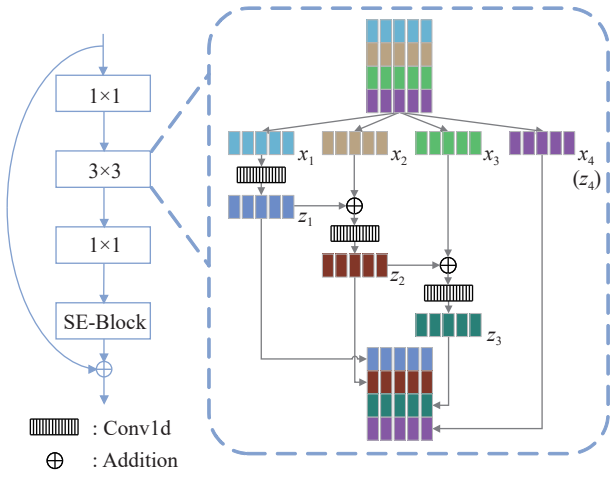

图 1　SE-Res2Block 结构示意图 (示例取 T 为 5, s 为 4)

Fig.1　Schematic diagram of the structure of SE-RES2Block (T is 5, and s is 4 in the example)

首先, 设输入二维特征(C×T), 其中 C 代表维数和 T 代表帧数;

其次, 将特征的维数等分为 s 组, 且每组分别进行卷积, 设 $\boldsymbol{x}_i(i \in \{1, 2, \cdots, s\})$为输入特征, 其输出 $\boldsymbol{z}_i$ 可表示为:

$$\boldsymbol{z}_i = \begin{cases} \boldsymbol{C}_i(\boldsymbol{x}_i) & (i = 1) \\ \boldsymbol{C}_i(\boldsymbol{x}_i \oplus \boldsymbol{z}_{i-1}) & (1 < i < s) \\ \boldsymbol{x}_i & (i = s) \end{cases} \tag{1}$$

其中, $\boldsymbol{C}_i$ 表示一维卷积、⊕表示特征相加, 一维卷积可有效的结合时间上下文信息, 处理不同时长的语音[13]. $\boldsymbol{z}_i = \boldsymbol{C}_i(\boldsymbol{x}_i \oplus \boldsymbol{z}_{i-1})$, 其将前一组的输出 $\boldsymbol{z}_{i-1}$ 与第 i 组的当前输入特征进行相加作为第 i 组新的输入再进行卷积, 这种层次残差连接的方式, 增加了输入特征的尺度数量进而扩大其感受野[20].

最后, 如图 1 所示将不同的输出 $\boldsymbol{z}_i$ 再重新进行聚合, 并经过一个 SE-Block[21] 以校准通道特征响应.

SE-Res2Block 中: (1) 特征拆分有助于提取全局和局部信息; (2) 通过对不同复杂程度的层特征进行聚合达到不同尺度的信息融合, 提升模型的特征提取能力. 该结构通过将不同感受野的特征进行聚合, 其虽有效地提升了性能, 但必须指出的是, 除第 s 组 $\boldsymbol{z}_s = \boldsymbol{x}_s$, 其余每组的输入特征都经过一维卷积处理, 再将处理后的特征进行聚合形成输出, 其输出包含 $\boldsymbol{x}_s$ 全部信息和 $\boldsymbol{x}_1$ 到 $\boldsymbol{x}_{s-1}$ 中高语义特征, 因此输出特征未包含原始输入 $\boldsymbol{x}_1$ 到 $\boldsymbol{x}_{s-1}$ 中的低语义特征, 因此导致原始输入特征 $\boldsymbol{x}_i(1 \leqslant i \leqslant s)$低语义特征信息未能充分利用.

1.2　稠密连接网络

DenseNet 网络的主要结构为 DenseBlock, 其每一层的输入均源于前面所有层的输出. 设一个 DenseBlock 结构中有 l 层, 故其包含 l×(l+1)/2 个连接. 因其每层特征均通过稠密连接的方式连接后续所有层, 记每一层的输入为 $\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_l$, 则第 l 层的输入与前 l−1 层的特征相关, 其可表示为 $\boldsymbol{x}_l = H_l([\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{l-1}])$, 其中 $[\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{l-1}]$为前 $l-1$层特征在通道维度上的合并, $H_l(\cdot)$代表非线性转化函数, 其为卷积操作、批量标准化、激活函数后的结果, 如图 2 所示. 该结构虽可实现特征重用, 提升效率, 并有效改善梯度消失的问题, 但由于其需通过增加每层的信道维度来增加网络宽度, 故不仅增加了模型的大小和计算复杂度, 且只能获取有限的性能提升[16].

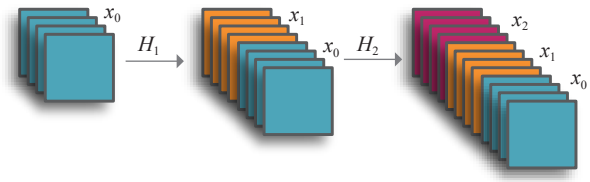

图 2　稠密连接结构示意图

Fig.2　Schematic diagram of the dense connection structure

上述研究表明: (1) SE-Res2Block 通过对不同感受野的特征聚合实现对多尺度特征信息的提取, 但其未能充分利用原始特征中的部分低语义特征, 导致特征信息的损失; (2) DenseBlock 通过对特征的重用来保证特征信息的完整性, 但特征的过多重复利用会导致特征冗余和效率降低.

2　SE-DR-Res2Block 模块结构

基于上述理论, 为保证特征信息的完整性并减少特征冗余, 论文通过将 DenseBlock 的稠密连接结构和 SE-Res2Block 相结合提出 SE-DR-Res2Block.

DenseBlock 中的稠密连接结构及 SE-Res2Block 中的残差连接结构分别以增加信道维度和堆叠更多的卷积层的方式来加深网络, 均可有效捕捉声纹信息, 将两者进行融合得到 SE-DR-Res2Block, 其将通道维度上每一层的特征映射进行连接作为下一层的输入, 同时堆叠更多的卷积层, 使不同层次特征信息进行融合, 更加充分地利用了多分辨率层的信息.

SE-DR-Res2Block 模型结构如图 3 所, 其工作流程如下所示:

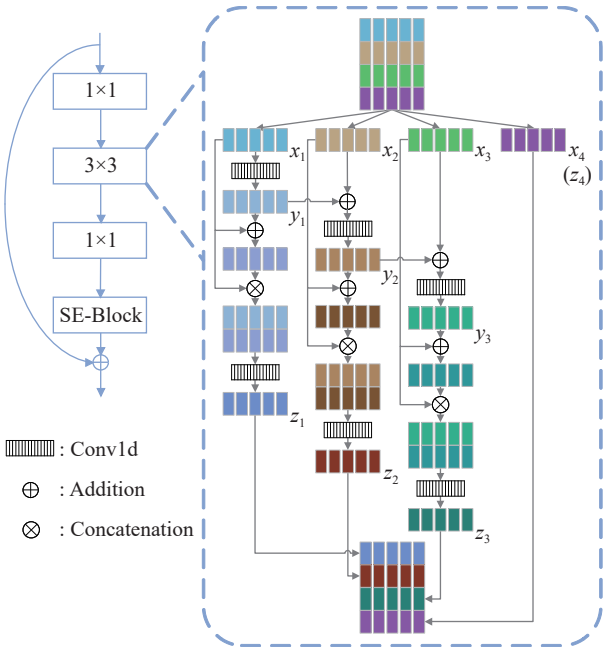

图 3　SE-DR-Res2Block 结构示意图 (示例取 T 为 5, s 为 4)

Fig.3　Schematic diagram of the structure of SE-DR-Res2Block (T is 5, and s is 4 in the example)

首先, 将输入特征切分为 s 组, 每组特征 $\boldsymbol{x}_i(1 \leqslant i \leqslant s)$分别进行卷积, 图中 $\boldsymbol{y}_i(i \in \{1, 2, \cdots, s-1\})$为中层特征, $\boldsymbol{y}_i$可表示为:

$$\boldsymbol{y}_i = \begin{cases} \boldsymbol{C}_i(\boldsymbol{x}_i) & (i = 1) \\ \boldsymbol{C}_i(\boldsymbol{y}_{i-1} \oplus \boldsymbol{x}_i) & (1 < i \leqslant s-1) \end{cases} \tag{2}$$

其中, $\boldsymbol{y}_i = (\boldsymbol{y}_{i-1} \oplus \boldsymbol{x}_i)$中将当前特征 $\boldsymbol{x}_i$与前一组的特征 $\boldsymbol{y}_{i-1}$进行相加后进行卷积, 获取当前组的中层特征 $\boldsymbol{y}_i$. 当前组中层特征 $\boldsymbol{y}_i$接收前一组特征 $\boldsymbol{y}_{i-1}$信息后, 相应信息感受野增大, 上述不同组特征相加, 使得每组中层特征实现对不同感受野特征的聚合, 不同感受野包含不同尺度信息, 当一个 $\boldsymbol{C}_i$接收来自前一个 $\boldsymbol{C}_i$的特征信息时, 相应的感受野会增大, 而在这种残差结构中有若干个卷积层, 这种操作经过层层作用, 最终使得网络的输出获得多种感受野大小的组合, 从而有效地以多尺度特征提取全局信息.

其次, 每一组特征分别进行稠密和残差连接, 其输出特征 $\boldsymbol{z}_i$可表示为:

$$\boldsymbol{z}_i = \begin{cases} \boldsymbol{C}_i((\boldsymbol{y}_i \oplus \boldsymbol{x}_i) \otimes \boldsymbol{x}_i) & (1 \leqslant i < s) \\ \boldsymbol{x}_i & (i = s) \end{cases} \tag{3}$$

其中, $\otimes$表示特征合并. 式(3)中, $\boldsymbol{y}_i \oplus \boldsymbol{x}_i$将同组中层特征 $\boldsymbol{y}_i$与相应原始特征 $\boldsymbol{x}_i$相加, 在同一组中增加感受野, 对不同尺度特征进行聚合, 该结构以多尺度特征提取本地信息.

然后, 由 $\boldsymbol{C}_i((\boldsymbol{y}_i \oplus \boldsymbol{x}_i) \otimes \boldsymbol{x}_i)$可知, 将上述聚合后的特征再与原始特征 $\boldsymbol{x}_i$进行合并, 实现原始特征的重用, 其既保证原始特征信息完整性, 又可获得高语义特征, 增强特征表达能力, 合并后的特征经过卷积从每组中获取不同感受野大小的特征, 将所有组输出特征 $\boldsymbol{z}_i$重新聚合以融合不同组的特征信息, 从而获取更多尺度的全局信息.

最后, 将聚合后的特征输入到 SE-Block, 其结构如图 4 所示, 相较于传统结构, 这里采用卷积层替代全连接层, 不仅降低了训练需要的参数, 同时权重共享, 可降低过拟合. SE-Block 通过建立通道间的相互依赖关系, 从而达到通道特征响应的目的, 其可获取不同特征通道的重要程度, 增强重要特征并抑制非重要特征.

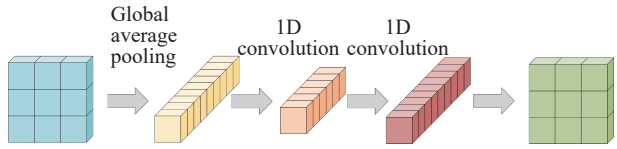

图 4　SE-Block 结构图

Fig.4　SE-Block structure diagram

综上所述, SE-DR-Res2Block 的工作原理为: (1) 基于残差结构将每个维度下声纹信息进行叠加以提高每个维度上的声纹信息, 实现不同尺度特征的聚合; (2) 基于稠密连接结构实现特征信息重用, 其通过对特征维度上的合并以提取整个特征所含的声纹信息; 二者结合使模型同时增加了维度数和每个维度的信息以实现对特征信息的充分提取, 从而增强特征表达能力; (3) SE-Block 增加了对通道信息的关注, 更有利于提取重要声纹特征, 提升泛化能力.

3　实验设置

3.1　实验数据集

论文实验采用文本无关说话人识别的开源数据集 Voxceleb1[22] 和 SITW (Speakers in the wild)数

据集[23]. 实验中训练集采用的 Voxceleb1 的训练集, 包含了 1211 名说话人共计 148642 条语音, 采样频率为 16 kHz, 单声道, 音频无静音段, 不需要进行语音活动检测处理[24]. Voxceleb1 的测试集包含 40 名说话人共计 4874 条语音, 这些语音数据被处理成 37720 个测试对用于注册和测试.

SITW 数据集是来自媒体的人工注释的语音样本, 该数据集包含 299 名说话人, 平均每人有 8 句语音. 实验使用 SITW 的评估集, 包含 180 名说话人共计 2883 条语音, 采用 core-core 测试场景进行测试, 其中 core 表示样本中只包含单个说话人.

3.2　系统设置

声纹识别都是基于 Pytorch 平台实现, 采用 Adam 优化器优化模型性能, batch size 设为 128, 初始学习速率设为 0.001, 并采用余弦衰减的学习速率策略来调整学习速率, 训练轮次设为 70, 使用批量标准化和 ReLU 激活函数加速收敛. 原始语音特征采用梅尔频率倒谱系数[25], 并对特征进行归一化处理. 所有系统采用 AAM-softmax(Angular additive margin softmax) 损失函数[26] 进行训练, 其中参照 ECAPA-TDNN 体系结构中, 卷积层采用 1024 通道, SE-Block 和注意力模块瓶颈维度设为 128. SE-DR-Res2Block 的参数 s 设置为 8, 采取最后一层 192 维向量作为说话人特征向量. 最终得分采用简单的余弦距离进行打分, 性能指标使用等错误率(Equal error rate, EER)和最小检测代价函数(Minimum normalized detection cost, minDCF).

3.3　实验结果

3.3.1　不同模块下 Res2Net-50 的性能比较

实验采用不同模块下的网络模型, 并分别在 Voxceleb1 数据集上进行性能测试. 实验采用 EER 和最小检测代价函数(DCF0.1、DCF0.01、DCF0.001)作为性能指标来评价其性能, 0.1、0.01、0.001 为真实说话人出现的先验概率, 以下用 p 值表示, 其中 x-vector 作为基线系统, 其他实验均在 Res2Net-50 网络上但采用不同的 Res2Block 进行, 其中 50 表示网络结构中包含 49 个卷积层和 1 个全连接层. 以下实验分别为原始 Res2Net-50 系统; 文献 [27] 提出的 FULL-Res2Block 的结构应用在 Res2Net-50 上的系统, 均简记为 FULL-Res2Net-50, 其中 FULL 表示文献 [27] 中全连接结构形式; 基于本文 SE-DR-Res2-Block 结构的 SE-DR-Res2Net-50 系统; 以及在 FULL-Res2Net 上的变体 FULL-SE-DR-Res2Net-50 系统.

由表 1 可知: (1) SE-DR-Res2Net-50 系统相较于 Res2Net-50 系统, 参数量增加 10.69×10^6, EER 下降了 5.9%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 3.9%、1.8%、2.5%; (2) FULL-SE-DR-Res2Net-50 系统相较于 FULL-Res2Net-50 系统, 参数量增加 10.69×10^6, EER 下降了 5.4%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 1.9%、1.6%、0.6%. 结果表明, 应用 SE-DR-Res2Net 的参数量增加, 系统性能均有所提升, 表明其具有更低的等错误率和最小检测代价函数, 也证明了稠密连接和残差连接结合的有效性.

3.3.2　不同模块下 ECAPA-TDNN 的性能比较

为体现 SE-DR-Res2Block 结构的适用性, 在 ECAPA-TDNN 上进行性能测试. 结果如表 2 所示, 其中 x-vector 作为基线系统, 其他实验均在 ECAPA-TDNN 系统上实现, 分别采用不同的 Res2Block 进行实验. 其中, 基于原始 Res2Block 模块的简记为 Res2-Block, 基于文献 [27] 中 FULL-Res2Block 模块的简记为 FULL-Res2Block, 基于本文结构的记为 SE-DR-Res2Block, 基于 FULL-Res2Block 的变体记为 FULL-SE-DR-Res2Block.

由表 2 可知: (1) 其中基于 SE-DR-Res2Block 模块的系统, 相于原始 Res2Block 模块下的 ECAPA-TDNN 系统, 在参数量仅增加 1.98M 的情况下, EER 下降了 10%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 9%、8.9%、3.8%; (2) FULL-SE-DR-Res2Block 模块下的系统相较于 FULL-Res2Block 模块下的系

表 1　Voxceleb1 测试集在不同 Res2Net-50 系统下的性能比较

Table 1　Performance comparison of the Voxceleb1 test set under different RES2Net-50 systems

System	Parameters/ 10^6	EER/%	minDCF		
			(p = 0.1)	(p = 0.01)	(p = 0.001)
x-vector[27]		4.19	0.212	0.391	0.512
Res2Net-50	24.50	3.73	0.205	0.381	0.485
FULL-Res2Net-50	24.50	3.91	0.210	0.385	0.481
SE-DR-Res2Net-50	35.19	3.51	0.197	0.374	0.473
FULL-SE-DR-Res2Net-50	35.19	3.70	0.206	0.379	0.478

表 2　Voxceleb1 测试集在不同 ECAPA-TDNN 系统下的性能比较

Table 2　Performance comparison of the Voxceleb1 test set under different ECAPA-TDNN systems

System	Parameters/10^6	EER/%	minDCF		
			(p = 0.1)	(p = 0.01)	(p = 0.001)
x-vector[27]		4.19	0.212	0.391	0.512
Res2Block	14.73	2.49	0.132	0.269	0.312
FULL-Res2Block	14.73	2.51	0.145	0.296	0.372
SE-DR-Res2Block	16.71	2.24	0.120	0.245	0.300
FULL-SE-DR-Res2Block	16.71	2.37	0.137	0.280	0.364

统, 参数量仅增加 1.98M, EER 下降了 5.5%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 5.6%、5.4%、2.2%. 实验结果表明, 论文提出的结构在不同网络模型下也具有良好的性能, 且在该模型下参数量增幅小, 对训练耗时影响小.

机器学习的目的是为了让训练后的模型能更好地适用于新鲜样本, 这种适应能力称为泛化能力. 为验证结构的泛化能力, 在数据集 SITW 中的 core-core 测试场景中进行测试, 其实验结果如表 3 所示: (1) SE-DR-Res2Block 相对于 Res2Block, 参数量仅增加 1.98M, EER 下降了 6.6%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 2.2%、4%、7.6%; (2) FULL-SE-DR-Res2Block 相对于 FULL-Res2Block, 参数量仅增加 1.98M, EER 下降了 4.3%, minDCF 在 p 值为 0.1、0.01、0.001 分别降低了 0、3.4%、3%. 由上述结果可知, SE-DR-Res2Block 在新鲜样本的测试中也具有良好的性能, 进一步证明了该模块具有良好的泛化能力.

3.3.3　不同时长的性能对比

为评估系统对不同时长的效果, 实验采用 core-core 测试集下的三个子测试集, 分别是小于 15 s 的语音, 大于 15 s 小于 25 s 的语音及大于 25 s 小于 40 s 的语音. 采用不同的 Res2Block 在 ECAPA-TDNN 网络系统上进行测试, x-vector 系统作为基线系统, 实验的具体结果如表 4 所示.

由表 4 可知: (1) 随着时长的增长, 因语音时长越长, 包含的声纹信息也越多, 所有系统的 EER 均降低, 表示其性能均有提高; (2) 在系统中采用 SE-DR-Res2Block 相较于 Res2Block, EER 在 0 ~ 15 s、15 ~25s、25 ~40s 分别下降了 11%、9.3%、3.8%;(3)FULL-SE-DR-Res2Block 相较于 FULL-Res2Block, EER 在0 ~15s、15 ~25s、25 ~40s 分别下降了 13.3%、5.3%、2.8%; (4) 其中时长越短, 性能提升的愈明显, 在所有时长中 SE-DR-Res2Block 的性能最佳. 实验结果表明, 论文提出的结构对不同时长也具有明显优势, 且其对短时语音的性能表现最好.

图 5 显示的不同时长下不同系统的检测错误权衡曲线. 由图 5 可知, SE-DR-Res2Block 的大部分曲线在其他系统曲线的下方, 表明较其他系统, 在大部分工作点, 即在 False positive rate 相同的条件下, 其 False negative rate 更低, 具有更好的性能.

3.3.4　不同模型下的性能对比

为评估系统的有效性, 在相同数据集下, 对不同模型的性能进行比较. 所有实验训练集均为 Voxceleb1 的开发部分, 测试集 Voxceleb1 和 SITW 分别为 Voxceleb1 的测试集和 SITW 的 core-core 测试场景.

结果如表 5 所示, 本文系统在 Voxceleb1 数据集上相较于文献 [14] 的新注意力机制、文献 [17]

表 3　SITW 测试集在不同 ECAPA-TDNN 系统下的性能比较

Table 3　Performance comparison of the SITW test set under different ECAPA-TDNN systems

System	Parameters/10^6	EER/%	minDCF		
			(p = 0.1)	(p = 0.01)	(p = 0.001)
x-vector		6.92	0.357	0.567	0.832
Res2Block	14.73	3.91	0.179	0.348	0.551
FULL-Res2Block	14.73	4.0	0.179	0.349	0.529
SE-DR-Res2Block	16.71	3.65	0.175	0.334	0.509
FULL-SE-DR-Res2Block	16.71	3.83	0.179	0.337	0.513

表 4　SITW 不同时长下的 EER

Table 4　EER at different SITW duration　　%

System	ERR		
	<15 s	15–25 s	25–40 s
x-vector	7.52	7.21	6.65
Res2Block	4.56	4.21	3.43
FULLRes2Block	4.75	4.30	3.52
SEDRRes2Block	4.06	3.82	3.30
FULLSEDRRes2Block	4.12	4.07	3.42

的错误样本提升先验概率、文献 [18] 的多特征学习策略, EER 分别下降了 42%、36%、23%. 在 SITW 数据集上相较于文献 [18] 中的多声学特征结构 (MTAF) 和多特征学习策略 LSTF-opt+S-CCA (Long-term and short-term features learning structure with canonical correlation analysis constraint), EER 分别下降了 29%、13%. 结果表明, 本文系统在相同数据集下, 相较于其他系统取得了更低的等错误率, 模型具有更好的性能.

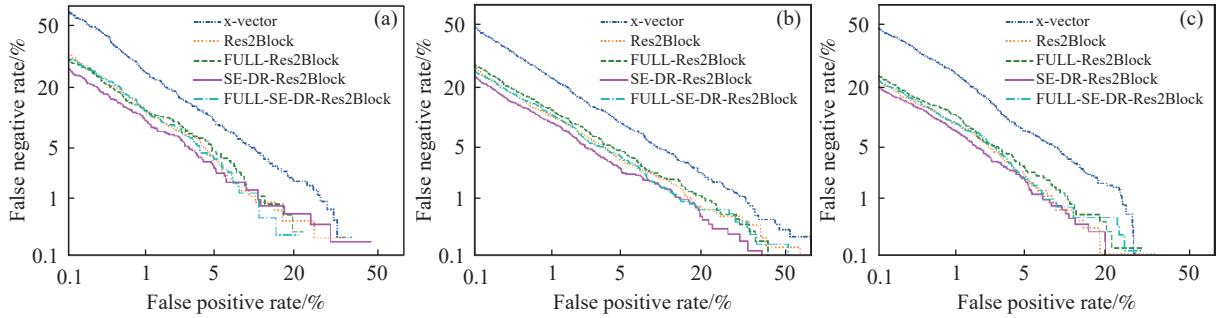

图 5　检测误差权衡曲线. (a) 0 ~ 15 s; (b) 15 ~ 25 s; (c) 25 ~ 40 s

Fig.5　Detection error tradeoff curve: (a) 0−15 s; (b) 15−25 s; (c) 25−40 s

表 5　不同模型下的性能比较

Table 5　Performance comparison under different models

System	Testing set	EER/%	System	Testing set	EER/%
Attentive statistics[14]	Voxceleb1	3.85	MTAF[18]	SITW	5.11
ResNet-34-SE[17]	Voxceleb1	3.52	LSTF-opt+S-CCA[18]	SITW	4.21
LSTF-opt+S-CCA[18]	Voxceleb1	2.92	Ours	SITW	3.65
Ours	Voxceleb1	2.24			

4　结论

针对声纹识别中传统的 Res2Net 模型的特征表达能力不足、泛化能力不强的问题, 本文提出一种基于稠密和残差连接的结构 SE-DR-Res2Block. 该结构中同时包含稠密和残差结构, 其残差连接对不同尺度特征进行融合, 增加了每个维度上的特征信息, 其稠密连接通过特征重用, 使部分低语义特征得以保留, 实现了对特征的有效提取, 且最大程度上保留了原始特征信息. 同时在 SE-DR-Res2-Block 结构中通过增加通道注意力模块, 提升对通道信息的关注, 强化对重要特征的权重并降低不必要特征的权重, 增强其泛化能力; 实验结果表明, SE-DR-Res2Block 模块相较于 Res2Block 模块, 不是单一地对特征进行堆叠和拼接, 而是对每一层的特征进行聚合, 同时保留低语义特征信息, 使得不同尺度特征和不同感受野信息进行互补, 对不同层的特征进行最大化利用. 其中应用 SE-DR-Res2-Block 的 ECAPA-TDNN 的网络模型在 Voxceleb1 和 SITW 数据集上的最佳等错误率 (EER) 分别为 2.24% 和 3.65%, 相较于 Res2Block 以及已有研究成果模型具有优异的特征表达能力及良好的泛化能力.

参　考　文　献

[1]　Zheng F, Li L T, Zhang H, et al. Overview of Voiceprint Recognition Technology and Applications. *J Inf Secur Res*, 2016, 2(1): 44.
(郑方, 李蓝天, 张慧等. 声纹识别技术及其应用现状. 信息安全研究, 2016, 2(1): 44)

[2]　Hayashi V T, Ruggiero W V. Hands-free authentication for virtual assistants with trusted IoT device and machine learning. *Sensors*, 2022, 22(4): 1325

[3]　Faundez-Zanuy M, Lucena-Molina J J, Hagmueller M. Speech watermarking: An approach for the forensic analysis of digital

telephonic recordings[J/OL]. *arXiv preprint*（2022-03-12）[2022-09-19]. https://arxiv.org/abs/2203.02275

[4] Garain A, Ray B, Giampaolo F, et al. GRaNN: Feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals. *Neural Comput Appl*, 2022, 34(17): 14463

[5] Waghmare K, Gawali B. Speaker recognition for forensic application: A review. *J Pos Sch Psychol*, 2022, 6(3): 984

[6] Mittal A, Dua M. Automatic speaker verification systems and spoof detection techniques: review and analysis. *Int J Speech Technol*, 2022, 25: 105

[7] Burget L, Matejka P, Schwarz P, et al. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. *IEEE Trans Audio Speech Lang Process*, 2007, 15(7): 1979

[8] Bao H J, Zheng F. Combined GMM-UBM and SVM speaker identification system. *J Tsinghua Univ Sci Technol*, 2008(S1): 693
(鲍焕军, 郑方. GMM-UBM 和 SVM 说话人辨认系统及融合的分析. 清华大学学报(自然科学版), 2008(S1): 693)

[9] Kenny P, Stafylakis T, Ouellet P, et al. JFA-based front ends for speaker recognition // 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, 2014: 1705

[10] Cumani S, Plchot O, Laface P. On the use of i–vector posterior distributions in probabilistic linear discriminant analysis. *IEEE/ACM Trans Audio Speech Lang Process*, 2014, 22(4): 846

[11] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification // 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, 2014: 4052

[12] Snyder D, Ghahremani P, Povey D, et al. Deep neural network-based speaker embeddings for end-to-end speaker verification // 2016 *IEEE Spoken Language Technology Workshop*. San Diego, 2016: 165

[13] Peddinti V, Povey D, Khudanpur S, et al. A time delay neural network architecture for efficient modeling of long tem-poral contexts // *Sixteenth Annual Conference of the International Speech Communication Association*. Dresden, 2015: 3214

[14] Okabe K, Koshinaka T, Shinoda K. Attentive statistics pooling for deep speaker embedding // *Interspeech*. Hyderabad, 2018: 2252

[15] Jiang Y H, Song Y, McLoughhlin I, et al. An effective deep embedding learning architecture for speaker verification // *Interspeech*. Graz, 2019: 4040

[16] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, 2017: 4700

[17] Zhou J F, Jiang T, Li Z, et al. Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function // *Interspeech*. Graz, 2019: 2883

[18] Li Z, Zhao M, Li L, et al. Multi-feature learning with canonical correlation analysis constraint for text-independent speaker verification // 2021 *IEEE Spoken Language Technology Workshop*. Shenzhen, 2021: 330

[19] Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: emphasized channel attention, propagation and aggregation in tdnn based speaker verification // *Interspeech*. Shanghai, 2020: 3830

[20] Gao S H, Cheng M M, Zhao K, et al. Res2net: A new multi-scale backbone architecture. *IEEE Trans Pattern Anal and Mach Intell*, 2019, 43(2): 652

[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, 2018: 7132

[22] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset[J/OL]. *arXiv preprint*（2018-05-30）[2022-09-19]. https://arxiv.org/abs/1706.08612

[23] McLaren M, Ferrer L, Castan D, et al. The speakers in the wild（SITW） speaker recognition database // *Interspeech*. San Francisco, 2016: 818

[24] Guo Z C, Yang Z, Ge Z R, et al. An endpoint detection method based on speech graph signal processing. *J Signal Process*, 2022, 38(4): 788
(郭振超, 杨震, 葛子瑞, 等. 一种基于语音图信号处理的端点检测方法. 信号处理, 2022, 38(04): 788)

[25] Zheng Y, Jiang Y X. Speaker clustering algorithm based on feature fusion. *J Northeast Univ Nat Sci*, 2021, 42(7): 952
(郑艳, 姜源祥. 基于特征融合的说话人聚类算法. 东北大学学报(自然科学版), 2021, 42(7): 952)

[26] Deng J K, Guo J, Xue N N, et al. Arcface: Additive angular margin loss for deep face recognition // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 4690

[27] Chen Z G, Li P, Xiao R Q, et al. A multiscale feature extraction method for text-independent speaker recognition. *J Electron Inf Technol*, 2021, 43(11): 3266
(陈志高, 李鹏, 肖润秋, 等. 文本无关说话人识别的一种多尺度特征提取方法. 电子与信息学报, 2021, 43(11): 3266)