计算机科学
COMPUTER SCIENCE

基于子频带前端模型和反向特征融合的说话人确认方法

王萌威, 杨哲

引用本文

王萌威, 杨哲. 基于子频带前端模型和反向特征融合的说话人确认方法[J]. 计算机科学, 2025, 52(3): 214-221.

WANG Mengwei, YANG Zhe. Speaker Verification Method Based on Sub-band Front-end Model and Inverse Feature Fusion [J]. Computer Science, 2025, 52(3): 214-221.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100222

基于子频带前端模型和反向特征融合的说话人确认方法

王萌威　杨　哲
苏州大学计算机科学与技术学院　江苏 苏州 215006
(mwwang@stu.suda.edu.cn)

摘　要　现有说话人确认方法中用于提取帧级特征的时延神经网络(TDNN)存在两个问题,一是缺少对局部频率特征的建模能力,二是多层特征融合方式无法对高层和低层特征之间的复杂关系进行有效建模。因此,提出一种新的前端模型以及一种新的多层特征融合方式。在前端模型中,通过将输入特征图划分为多个子频带,并逐层扩大子频带的频率范围,使TDNN可以渐进地对局部频率特征进行建模。同时,在主干模型中新增一条由高层向低层传递的反向路径,对相邻两层输出特征之间的关系进行建模,并将反向路径中每层的输出拼接后作为融合后的特征。此外,在主干模型中使用逆瓶颈层的设计,进一步提升模型的性能。在VoxCeleb1测试集上的实验结果表明,所提方法与目前的TDNN方法相比,等错误率和最小代价检测函数分别降低了9%和14%,而参数量仅为目前方法的52%。

关键词:声纹识别;说话人确认;时延神经网络;子频带特征提取;多层特征融合
中图分类号　TP183;TN912.34

Speaker Verification Method Based on Sub-band Front-end Model and Inverse Feature Fusion

WANG Mengwei and YANG Zhe
School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract　Two problems with time delay neural networks(TDNN) used to extract frame-level features in existing speaker confirmation methods are the lack of the ability to model local frequency features and the inability of the multilayer feature fusion approach to effectively model the complex relationships between high-level and low-level features. Therefore,a new front-end model as well as a new multilayer feature fusion approach are proposed. In the front-end model,by dividing the input feature map into multiple sub-bands and expanding the frequency range of the sub-bands layer by layer,the TDNN can model the local frequency features progressively. Meanwhile,a new inverse path passing from higher to lower layers is added to the backbone model to model the relationship between the output features of two adjacent layers,and the outputs of each layer in the inverse path are concatenated to serve as the fused features. In addition,the design of the inverse bottleneck layer is used in the backbone model to further improve the performance of the model. Experimental results on the VoxCeleb1 test set show that the proposed method has a relative reduction of 9% in the equal error rate and 14% in the minimum cost detection function,compared to the current TDNN method,while the number of parameters is only 52% of the current method.

Keywords　Speaker recognition,Speaker verification,Time delay neural network,Sub-band feature extraction,Multilayer feature fusion

1　引言

说话人确认是声纹识别中一个重要的研究任务,目的是判断两段音频是否属于同一个说话人。其在生物识别、智能家居等领域有着众多的应用[1]。自深度学习被应用于说话人确认任务后,其凭借深度神经网络强大的性能,逐渐成为研究的主流。一个典型的基于深度学习的说话人确认系统分为3部分[2]:帧级特征提取器、池化层和前馈网络。帧级特征提取器负责从输入音频中提取帧级特征,池化层负责将不定长的帧级特征映射为定长的话语级特征,前馈网络则负责从话语级特征中提取最终的说话人嵌入,最后通过比较不同说话人嵌入之间的距离来判断两段音频是否属于同一说话人。帧级特征提取是整个方法中的第一步,提取的特征好坏会直接影响整个方法的性能。因此,针对帧级特征提取器的改进是目前的一个研究热点[2]。

目前,基于一维卷积的模型[3-5]和基于二维卷积的

到稿日期:2024-01-31　返修日期:2024-06-12
基金项目:教育部产学合作协同育人项目(220606363154256)
This work was supported by the Ministry of Education University-Industry Collaborative Education Program(220606363154256).
通信作者:杨哲(yangzhe@suda.edu.cn)

模型[6-8]是两类常见的帧级特征提取器。其中基于一维卷积的模型又被称为时延神经网络(Time Delay Neural Network，TDNN)。模型以声谱图作为输入音频特征，对于 TDNN，通道维对应声谱图频率维，模型每一层都拥有覆盖全部频率范围的感受野。对于二维卷积模型，通道维独立于声谱图，随着模型由浅入深，每一层的频率维感受野逐渐增大，渐进地建模局部频率特征。为了结合两种模型的特点，常用的方法是在 TDNN 前添加一个二维卷积作为前端模型，将前端模型的输出特征作为 TDNN 输入。本文将帧级特征提取器中这个接收前端模型输出的模型称为主干模型。考虑到 TDNN 和二维卷积通道维上的差异，主干模型可能无法高效地从前端模型的输出中提取有效的说话人嵌入。

在早期的方法中[9-11]，考虑到帧级特征提取器(以下简称提取器)的深层特征会与说话人的辨识信息有着较强的相关性，因此只使用提取器最后一层的特征来计算说话人嵌入。然而，越来越多的研究表明，结合不同层的特征能够提高说话人嵌入的鲁棒性[12-13]，许多研究开始关注如何将不同层的特征进行更好的融合。目前常见的多层特征融合方法是将提取器各层的特征进行拼接，将拼接后的特征由全连接层进行特征融合。这会导致两个方面的问题[5]：首先，考虑到不同层特征所表征的内容随着层间距离增大会有较大的差异，直接对这些特征进行融合可能会导致大量的信息损失；其次，全连接层的参数量与参入融合的特征层数直接相关，为了减少提取器的参数量，往往只取有限层数(比如三层)的特征进行融合，这限制了融合特征的尺度多样性。

本文的主要工作包括两点：1)在基于 TDNN 的提取器中引入基于子频带划分的 TDNN 作为前端模型，模拟二维卷积中渐进地对局部频率特征进行建模，同时避免由 TDNN 和二维卷积结构上的差异导致的特征提取低效问题；2)提出了一种基于反向路径逐层融合的多层特征融合方法，既避免了直接对高层和低层特征进行融合，又减少了融合特征增加的参数量。此外，本文在主干模型部分采用逆瓶颈层的设计来进一步提升模型的建模能力。实验结果表明，改进后的方法相较于基线方法性能有所提升，并且以较少的参数量取得了优于主流的说话人确认方法的性能。

2 相关工作

在基于深度学习方法流行之前，i-vector[14]等传统方法是解决说话人确认问题的主要方案。d-vector[9]首次将深度学习引入说话人识别任务中，使用 TDNN 来提取帧级特征，并将帧级特征平均池化作为话语级特征。之后 x-vector[10-11]使用统计值池化来计算帧级特征的统计值并将其用作话语级特征，大幅提升了方法性能。自此，深度学习的方法凭借其出色的性能迅速取代传统的方法并成为研究的热点。如前文所述，基于深度学习的说话人确认系统通常可以分为帧级特征提取器、池化层和前馈网络 3 部分。帧级特征提取器负责从输入音频中提取帧级特征，主要可以分为：基于循环神经网络的提取器[15]、基于 TDNN 的提取器[5,10]、基于二维卷积的提取器[8,16]和基于 Transformer 的提取器[17]。池化层负责将不定长的帧级特征映射为定长的话语级特征，常用的池化层有平均池化[9,18]、统计值池化[5,10-11]和基于自注意力的池化[19]等。前馈网络负责从话语级特征中提取最终的说话人嵌入，绝大多数方法使用一层或者两层全连接层来提取说话人嵌入。考虑到帧级特征的重要性，许多研究提出了针对帧级特征提取器进行优化的方法：ECAPA-TDNN[5]在 TDNN 提取器中引入了 Res2Net 的模块设计并使用 SE 模块对每一层输出特征进行缩放；MFA-Conformer[17]使用 Conformer 模块在 Transformer 中引入局部特征信息以提升提取器的性能。

除了提取器结构上的改进，为了使基于 TDNN 的提取器能够像二维卷积那样渐进地建模局部频率特征，文献[20]在 TDNN 主干模型前添加一个四层二维 ResNet 作为前端模型，其表现出不错的性能提升。在此基础上，文献[21]在前端模型中使用二维 Res2Block 模块设计，同时引入注意力机制来强调重要的频率成分，进一步提升了提取器的性能。然而这些方法没有考虑到一维卷积和二维卷积结构上的差异，只是简单地将二维卷积的输出展平作为一维卷积的输入，可能会使主干模型无法高效地提取帧级特征。

由于帧级特征提取器中的浅层特征有助于得到更有区分度的说话人嵌入，因此许多研究关注如何有效地融合多层特征。常见的特征融合方法可以分为两类：帧级特征融合[12]和话语级特征融合[13]。文献[12]将后三层的特征沿通道维拼接起来经过一个全连接层得到融合后的特征；文献[13]将提取器各层的帧级特征统计值池化后得到的话语级特征进行拼接再经过一个全连接层得到融合后的特征。考虑到高层特征和低层特征之间较大的语义差距，使用全连接层直接融合高层和低层特征可能会造成较大的信息损失。此外，为了减少用于融合的全连接层参数量，这些方法只取有限层数的特征进行融合，限制了融合后特征的尺度多样性。

为了避免二维卷积和 TDNN 之间的结构差异带来的影响，本文提出的前端模型基于 TDNN 本身，并通过子频带划分的方法引入渐进局部频率特征建模的过程。为了避免直接对高层特征和低层特征进行融合，本文提出的多层特征融合方法采用逐层融合的方式，这种方法同时也减少了融合多层特征所增加的参数量。

3 方法介绍

整个说话人确认方法结果如图 1 左侧所示。

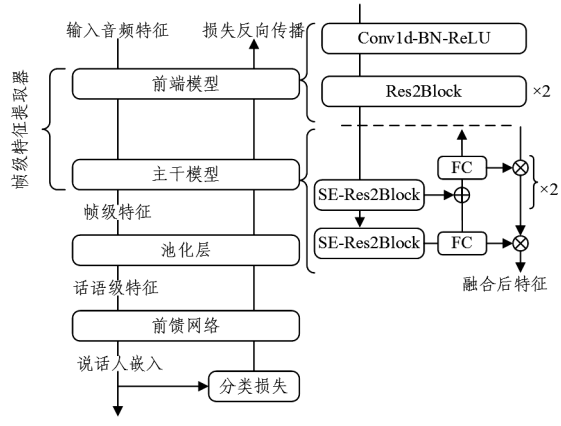

图 1 说话人确认方法

Fig. 1 Speaker verification method

3.1 节介绍在基于 TDNN 的提取器中引入渐进建模局部频率信息过程而提出的前端模型；3.2 节介绍主干模型中进行的改进以及提出的多层特征融合方法；3.3 节介绍引入改进后的说话人确认方法的具体实现。

3.1 子频带前端模型

在二维卷积模型中，模型的低层主要对神经元周围局部区域特征进行建模，高层则通过覆盖整个特征图的感受野来学习更抽象的特征表示。而在 TDNN 中，模型一开始便拥有覆盖整个频率维的感受野，不存在二维卷积中对频率维特征循序渐进的建模过程。为了在 TDNN 中，引入这一过程，文献[20]在 TDNN 主干模型前设置一个基于二维卷积的前端模型，将前端模型的输出展平后的输出作为主干模型的输入。假设前端模型的输出特征形状为 (B,C,F,T)，其中 B 为一个小批次中的样本个数，C 为二维卷积的输出通道数，F 和 T 分别为特征图频率维和时间维的长度。主干模型首先将特征的形状展平为 $(B,C\times F,T)$，之后将其作为一维卷积的输入。考虑到二维卷积的通道维独立于特征图，若直接将二维卷积的通道维和特征图的频率维合并，会导致展平后的特征通道维的信息既有可能位于同一特征空间(二维卷积特征图中的同一通道的值)，又有可能位于不同的特征空间(不同通道的值)，进而使得主干模型无法很好地建模展平后特征通道之间的关系。

因此，为了避免前端模型和主干模型结构上的差异，本文使用一个三层的 TDNN 作为前端模型，并借鉴文献[22]中主干模型的设计，在前端模型的每一层进行子频带划分，用不同的 TDNN 分别对子频带的特征建模。模型层次越深，子频带数量越少，TDNN 的频率感受野就越大。通过这种方式来模拟二维卷积中渐进建模局部频率特征的过程。具体来说，假设输入的音频特征频率维长度为 f，前端模型每一层的输出特征通道维长度为 c，前端模型的第一层为分组数为 g_1，卷积核大小为 k_1 的一维卷积，每一组的卷积核负责将输入音频特征中频率维长度为 f/g_1 的子频带映射为输出特征中的 c/g_1 个通道。第二层和第三层采用 Res2Block 的模块设计。Res2Block 如图 2 右下角所示。

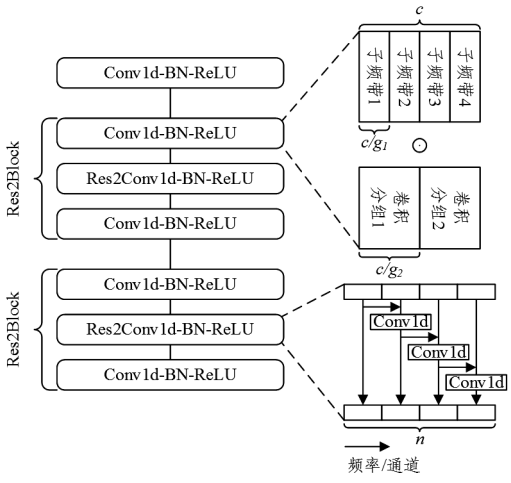

图 2 前端模型框架

Fig. 2 Framework of front-end model

Res2Conv1d 含有多组卷积核，每一组卷积核对应输入特征中的一部分通道。此外，在这些卷积核之间设置一条层内路径，第 $n-1$ 组卷积的输出除了传递给下一层外也会传递给第 n 组，这样路径开头的卷积核时间维感受野较小，路径末尾的卷积核时间维感受野较大，使得模型在同一层内能够考虑不同时间尺度的信息。Res2Block 前后两个一维卷积为分组数为 g_i、卷积核大小 $k_i=1$ 的分组卷积。Res2Conv1d 同样也分为 g_i 组，不同于 Res2Conv1d 内部划分 n 组用于建模不同时间尺度的特征，这里的 g_i 组分别对应不同的子频带。g_i 满足 $g_1>g_2>g_3$，在第 i 层第一个一维卷积中，由于 $c/g_i>c/g_{i-1}$，第 i 层中的一组的卷积核覆盖第 i-1 层输出特征中的 g_{i-1}/g_i 个子频带，即第 i 层的频率维感受野相较于第 $i-1$ 层增大 g_{i-1}/g_i，如图 1 右上角所示(⊙表示卷积操作)。在模型中每一个卷积操作之后都接着一个批量归一化层(Batch Normalization，BN)和一个 ReLU 非线性激活函数。

3.2 主干模型改进及反向特征融合

使用 ECAPA-TDNN[5] 的提取器作为本文方法提取器中的主干模型。ECAPA-TDNN 的提取器为一个四层结构，前三层均为 SE-Res2Block 模块，使用 SE 模块重新缩放 Res2Block 提取的帧级特征。假设 Res2Block 的输出特征形状为 (B,C,T)，SE 模块首先使用挤压操作得到时间维的全局特征 $\boldsymbol{z}$，如式(1)所示：

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

其中，$\boldsymbol{h}_t$ 为每一帧的特征。之后对 $\boldsymbol{z}$ 使用激励操作来得到输出特征中每个通道的缩放分数 $\boldsymbol{s}$，如式(2)所示：

$$\boldsymbol{s}=\sigma(\boldsymbol{W}_2 f(\boldsymbol{W}_1\boldsymbol{z}+\boldsymbol{b}_1)+\boldsymbol{b}_2) \tag{2}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别为第一和第二个全连接层的权重参数，$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 分别为第一和第二个全连接层的偏置项。f 为 ReLU 非线性激活函数，σ 为 sigmoid 非线性激活函数。最后使用 $\boldsymbol{s}$ 对输出特征进行缩放，如式(3)所示：

$$\boldsymbol{h}_c'=\boldsymbol{s}_c\,\boldsymbol{h}_c \tag{3}$$

其中 $\boldsymbol{h}_c'$ 为特征图中经过缩放后的一个通道。提取器最后一层为一个用于特征融合的全连接层。该层将前面三层输出特征沿通道维拼接后作为输入，输出融合后的特征。此外，为了更有效地利用浅层的特征，每一层都将之前所有层的输出特征逐元素相加作为输入。

在 ECAPA-TDNN 的 SE-Res2Block 中，Res2Conv1d 前后有两个卷积核大小 $k=1$ 的一维卷积，用来学习通道间的信息。假设第一个一维卷积输出通道与输入通道之比和第二个一维卷积输入通道与输出通道之比均为 r，考虑到逆瓶颈层在 MobileNetV2[23] 和 ConvNext[24] 中都有着良好的性能表现，本文设置 $r>1$，使主干模型的每一层都为逆瓶颈层结构。

本文借鉴特征金字塔网络(Feature Pyramids Networks，FPN)[25]，提出了一种新的多层特征融合机制，通过一条自顶向下的路径，向低层逐层传递高层的特征。FPN 通过逐元素相加来进行融合，而基于二维卷积的模型通常会在不同层之间添加下采样和增加通道数的操作，因此

高层的特征需要经过上采样操作，低层特征需要通过二维卷积提升通道数，之后才能够进行融合。文献[26]使用基于二维卷积的提取器，因此直接沿用 FPN 结构中融合特征的方法。如图 3 所示，假设提取器共 L 层，为方便起见图中仅展示单通道的特征图，⊕表示逐元素相加，⊗表示沿通道维拼接。

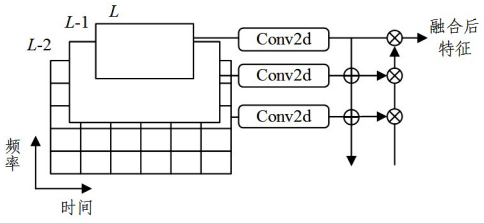

图 3 FPN 中的特征融合

Fig. 3 Feature fusion in FPN

不同于文献[26]，本文选择将低层特征与高层融合特征逐元素相加后再通过一维卷积进行特征融合，如图 4 所示，为方便起见图中仅展示单通道的特征图。将反向路径中第 L 层传递过来的特征和第 $L-1$ 层的输出特征逐元素相加后，使用全连接层(Fully Connected Layer，FC)进行特征融合，之后经过 BN 和 ReLU 得到反向路径中第 $L-1$ 层的特征，并将其传递给第 $L-2$ 层。最后将反向路径中各层的特征拼接起来作为融合后的帧级特征。另外，由于本文提取器采用了前端模型和主干模型两部分设计，而前端模型和主干模型分别建模不同的频率特征，因此本文只对提取器中主干模型的特征进行融合。本文主干模型层与层之间不设其他操作，因此在进行特征融合时，不需要对高层特征和低层特征进行额外操作。假设主干模型中共有 L 层参与特征融合，每一层输出特征的通道数为 C_1，融合后特征的通道数为 C_2。使用全连接层融合拼接后的多层特征为模型新增的参数量为 $L \cdot C_1 \cdot C_2$，而本文提出的多层特征融合方法新增的参数量为 $L \cdot C_1 \cdot (C_2/L) = C_1 C_2$，与参与融合的特征层数无关。此外，反向路径中每层只考虑当前层和高层融合后特征的融合，避免了直接对高层和低层特征进行融合。

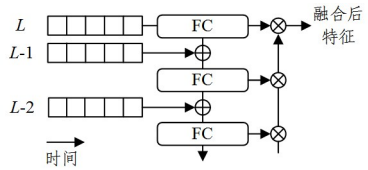

图 4 提出的特征融合方法

Fig. 4 Feature fusion method proposed in this paper

3.3 方法实现

同其他基于深度学习的方法一样，本文提出的方法也分为 3 部分：帧级特征特征提取器、池化层和前馈网络。帧级特征提取器由 3.1 节中的多频带前端模型和 3.2 节中改进后的主干模型构成。使用强调通道的注意力统计值池化(Attention Statistics Pooling，ASP)[5]来从帧级特征中计算话语级特征，具体过程如下：

$$\boldsymbol{e}_t = \boldsymbol{W}_4 f(\boldsymbol{W}_3 \boldsymbol{h}_t + \boldsymbol{b}_3) + \boldsymbol{b}_4 \tag{4}$$

其中，$\boldsymbol{e}_t$ 为每帧特征 $\boldsymbol{h}_t$ 的分数，由两个全连接层计算得到；$\boldsymbol{W}_3$ 和 $\boldsymbol{W}_4$ 分别为第一个和第二个全连接层的权重，$\boldsymbol{b}_3$ 和 $\boldsymbol{b}_4$ 分别为第一个和第二个全连接层的偏置项。之后沿时间维使用 Softmax 函数得到每一帧特征的注意力权重 $\boldsymbol{\alpha}_t$，如式(5)所示：

$$\boldsymbol{\alpha}_t = \frac{\exp(\boldsymbol{e}_t)}{\sum_{t'}^{T} \exp(\boldsymbol{e}_t')} \tag{5}$$

之后使用 $\boldsymbol{\alpha}_t$ 计算帧级特征的加权平均值以及标准差。

$$\bar{\boldsymbol{\mu}} = \sum_{t}^{T} \boldsymbol{\alpha}_t \boldsymbol{h}_t \tag{6}$$

$$\bar{\boldsymbol{\sigma}} = \sqrt{\sum_{t}^{T} \boldsymbol{\alpha}_t \boldsymbol{h}_t^2 - \bar{\boldsymbol{\mu}}^2} \tag{7}$$

将 $\bar{\boldsymbol{\mu}}$ 和 $\bar{\boldsymbol{\sigma}}$ 沿通道维拼接作为话语级特征。最后使用由一层全连接层组成的前馈网络从话语级特征提取最终的说话人嵌入。

训练整个说话人确认方法的常见方法有两种，一种是使用基于度量学习的方法，常用的损失函数为三元损失[27]。三元损失需要在数据集中构建三元组，每个三元组由一个锚样本、一个与锚样本属于同一说话人的正样本，以及一个与锚样本属于不同说话人的负样本组成。另一种方法则使用一个额外的分类器对提取的说话人嵌入进行分类，将说话人确认在训练时转换为闭集分类问题。由于说话人确认方法在实际应用中并不能确定具体需要区分的说话人数量，因此希望在训练时提取的嵌入的说话人判别性尽可能大，即每个类别的说话人嵌入之间的距离尽可能远，同类别的说话人嵌入之间的距离尽可能近。为了实现这一目标，常用的损失函数有 L-Softmax[28]，A-Softmax[29]，AM-Softmax[30]和 AAM-Softmax[31]。要通过三元损失训练得到一个高性能的说话人确认方法，对三元组构建有着较高的要求。考虑到实验可复现性以及与其他方法的公平比较，本文使用基于分类的 AAM-Softmax 损失函数来进行训练。为了减小同类说话人嵌入之间的距离，AAM-Softmax 引入了加性角间隔 m。

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s \cdot \cos(\theta_{y_i} + m))}{\exp(s \cdot \cos(\theta_{y_i} + m)) + \sum_{j=1, j \neq y_i}^{N} \exp(s \cdot \cos\theta_j)} \tag{8}$$

其中，N 为训练集中不同说话人的个数，θ_j 为归一化后的说话人嵌入与归一化后的分类器中 j 节点权重之间的余弦距离，s 为防止训练无法收敛的缩放因子。

在测试时，使用归一化后嵌入之间的余弦距离来衡量两个说话人嵌入之间的距离。若两个说话人嵌入之间的距离小于设定的阈值，则认为这两段嵌入对应的音频属于同一个说话人。

4 实验设置

4.1 数据集

本文实验使用的数据集是目前说话人确认领域中常用的两个英文数据集 VoxCeleb1[32]和 VoxCeleb2[33]。两个数据集中的音频均采集自 YouTube 视频网站，VoxCeleb1 包含来自 1 251 名说话人的 15 万条音频，平均每个说话人包含 116 条

音频，每条音频平均长约 8.2s。VoxCeleb2 包含来自 6112 个说话人的 110 万条音频，平均每个说话人包含 185 条音频，每段音频平均长约 7.8s。

本文使用 VoxCeleb2 的训练集来进行训练，训练集包含来自 5994 名说话人的 1092009 条音频。在官方划分的 3 个评估集 VoxCeleb1-O，VoxCeleb1-E，VoxCeleb1-H 中测试训练好的方法。其中 VoxCeleb1-O 为 VoxCeleb1 的原始测试集，包含由 40 个说话人的 4874 条音频组成的 37611 个音频对；VoxCeleb1-E 包含从整个 VoxCeleb1 数据集中随机采样的 579818 个音频对；VoxCeleb1-H 包含从整个 VoxCeleb1 数据集中采样的 550894 个音频对，每对音频都来自相同国家、相同性别的说话人。

4.2 特征提取

本文实验使用 80 维梅尔滤波器组（Mel Filter Bank，Fbank）特征作为输入音频特征。首先将所有音频重采样为 16KHz，并从每段音频中随机截取 3s 的片段，如果音频总长度小于 3s 则将其循环填充至 3s。使用 0.97 的预加重系数对音频进行预加重以补偿高频分量的损失。之后使用汉宁窗（Hanning Window）以 25ms 的窗长和 10ms 的帧移对片段进行分帧、加窗，并对每个窗进行 512 点的短时傅里叶变换。最后将得到的声谱图使用 80 个梅尔滤波器计算得到 80 维的 Fbank 特征，并将得到的特征沿着频率维使用均值方差归一化。对于每一段音频，最后提取的特征形状为(80,300)。

由于使用数据增强能够显著提升方法在测试时的鲁棒性，因此本文参照文献[8]的方法对数据集进行数据增强。在提取 Fbank 特征之前，对训练集中每条音频随机选择一种噪声或者混响进行加噪处理。其中加性噪声选取自 MUSAN 数据集[34]中的语音、噪声、音乐类别，混响随机选取自 RIR 数据集[35]中的中型和小型房间的冲激响应。在提取 Fbank 特征之后，对每条音频都应用 SpecAugment[36]数据增强，随机从频率维选择 0～20 个分量，从时间维选择 0～100 帧进行掩蔽。

4.3 实现细节

本文方法中的提取器使用 512 的通道数。对于形状为(80,300)的输入特征，在前端模型的第一层将其划分为 8 个独立的子频带，每个子频带包含 10 个频率分量，即将 g_1 设置为 8。为了在前端模型中完成渐进建模局部特征的过程，使得主干模型能够平滑地对全局频率特征进行建模，将前端模型中第二层和第三层的 g_2 和 g_3 分别设置为 4 和 2。在第二层中将 8 个子频带缩减为 4 个，在第三层将 4 个子频带缩减为 2 个。这样在主干模型第一层中第一个一维卷积将两个子频带合并为完整的频率维后，主干模型中的所有 SE-Res2Block 块都能够专注于建模全局频率特征。其他超参数设置同文献[5]。具体来说，在整个模型的第一层使用较大的卷积核，在后续层级中使用较小的卷积核，即在前端模型的第一层将 k_1 设置为 5，在后续的所有层级（包括主干模型）将 k 设置为 3。为了在保证模型整体参数量较少的同时增加主干模型每一层建模特征的时间上下文长度，在主干模型三层 SE-Res2Block 中使用膨胀卷积并将膨胀率 d 分别设置为 2,3,4。将 r 都设置 2 以实现逆瓶颈层设计。使用本文方法对主干模型所有层的特征进行融合，融合后的特征通道数为 1536。最后，将前馈网络输出通道数设置为 192，以提取 192 维的说话人嵌入。模型各层的具体超参数设置如表 1 所列。

表 1 本文模型的参数设置

Table 1 Parameter settings of the proposed model

	模型结构	k,g,d,r	输出形状
前端模型	Conv1-BN-ReLU	5,8,1,—	(512,300)
	Res2Block	3,4,1,1	(512,300)
	Res2Block	3,2,1,1	(512,300)
主干模型	SE-Res2Block	3,1,2,2	(512,300)
	SE-Res2Block	3,1,3,2	(512,300)
	SE-Res2Block	3,1,4,2	(512,300)
特征融合	FC-BN-ReLU	N/A	(512,300)
	FC-BN-ReLU	N/A	(512,300)
	FC-BN-ReLU	N/A	(512,300)
	拼接融合特征	N/A	(1536,300)
池化	ASP	N/A	(3072)
前馈网络	FC	N/A	(192)

训练时将每个批次的样本数量设为 400 并使用 Adam 优化器来更新模型参数。优化器的权重衰减系数设置为 2×10^{-5}。使用带有预热的余弦衰减学习率调度器进行学习率调度，学习率经过 3 轮训练由 0 预热为 1×10^{-3}，之后经过 77 轮训练衰减为 1×10^{-7}。对于损失函数，本文将 AAM-Softmax 中的 m 和 s 分别设置为 0.2 和 30。在测试时，使用整段音频提取特征，对于小于 3s 的音频则将其循环填充至 3s。所有方法均使用 2 张 40GB 显存的 NVIDIA A100 进行训练和测试。

4.4 评价指标

在实验中使用等错误率 EER（Equal Error Rate）/最小检测代价函数 minDCF（Minimum Detection Cost Function）两个指标来衡量本文方法的性能，并与其他方法的性能进行比较。EER 为检测误差权衡曲线中错误接受率和错误拒绝率相等时的错误率。minDCF 通过式(9)进行计算。

$$\min DCF = C_{FR} \cdot FR \cdot P_T + C_{FA} \cdot FA \times P_I \tag{9}$$

其中，FR 为错误拒绝率，FA 为错误接受率，C_{FR} 和 C_{FA} 分别为错误拒绝和错误接受的代价系数，P_T 和 P_I 分别表示真实说话人和冒名顶替者出现的先验概率。在本文的实验中，将 C_{FR} 和 C_{FA} 设置为 1，设置 P_T 和 P_I 为 0.01 和 0.99。

4.5 对比实验设计

在这部分实验中，使用 3.3 节中介绍的说话人方法（本文模型）与目前主流的方法进行对比，包括 ECAPA-TDNN[5]，ResNetSE34[8]以及 MFA-Conformer[17]。ECAPA-TDNN 的结构如 3.2 节中所述，这里将通道数设置为 1024。ResNet 是一个简单通用的模型，也是一种流行的二维卷积提取器。ResNetSE34 使用一个 4 阶段的 34 层 ResNet 作为提取器，使用 SE 模块对每个阶段的输出特征图进行缩放，使用 ASP 从提取器最后一层输出的帧级特征中计算得到话语级特征。MFA-Conformer 是一种流行的基于 Transformer 的说话人确认方法，其提取器的初始层为步长为 2 的跨步卷积，负责对特征进行下采样；提取器主干使用 Conformer 模块，最后提取器将不同层的特征拼接后通过一个全连接层进行特征融合，并使用 ASP 从融合后的帧级特征中计算得到话语级特征。总的来说，所有方法只在提取器上存在区别。

4.6 前端模型有效性实验设计

为了证明前端模型的有效性，本节基于512维的ECAPA-TDNN(基线模型)设计了5种提取器用来进行对比。首先根据文献[30]设计一个通道数为128的四层 $k=3$ 的ResNet作为前端模型，将基线模型的提取器作为主干模型与其组合，称之为Conv2d(128)-ECAPA。ResNet中第一层和最后一层在频率维使用步长为2的跨步卷积来降低计算量，中间两层的输入通过快捷连接相加到输出中，每个卷积层后都有一个ReLU和BN。最终前端模型输出的特征形状为(400，128，20，300)。为了将特征传递到主干模型中，需要先将其展平为(400，2560，300)，2560的通道数量使得主干模型的第一层拥有较大的参数量。为了进行公平的比较，另外设计了一个通道数为32的四层ResNet前端模型，将其与基线模型的提取器组合称为Conv2d(32)-ECAPA。除了通道数减少外，其余结构与Conv2d(128)-ECAPA相同。本节模型则使用与4.3节中相同的前端模型结构与基线模型的提取器进行组合。另外，由于提出前端模型的结构与主干模型相似，因此设计一个由一层 $k=5$ 的一维卷积初始层和两层与基线模型的提取器相同设置的SE-Res2Block组成的前端模型，与基线模型的提取器组成一个七层的ECAPA-TDNN，称之为ECAPA-TDNN-D，用以验证提出的前端模型并不是简单地加深主干模型深度。所有实验只在提取器中的前端模型上存在区别。

4.7 主干模型及反向特征融合有效性实验设计

为了证明本文对主干模型改进以及提出的反向特征融合方法的有效性，本节基于512通道数的ECAPA-TDNN设计了4种只包含主干模型的提取器进行对比实验。第一种提取器的SE-Res2Block块中不改变通道数，始终保持512的通道数，即 $r=1$；第二种提取器为 $r=2$ 的逆瓶颈层设计；另设一个 $r=0.5$ 的瓶颈层提取器。为了保证瓶颈层和逆瓶颈层的计算量和参数量相近，在 $r=0.5$ 的结构中将通道数设置为1024。最后，本节模型将 $r=2$ 结构中拼接所有层特征并使用全连接层进行融合的多层特征融合方法替换为3.2节中提出的多层特征融合方法。

5 实验结果分析

5.1 对比实验结果及分析

表2列出了本文方法与目前流行的几种方法的对比实验结果。可以看出，在3种流行的方法中，ECAPA-TDNN(1024)有着最好的性能表现。本文模型在3个测试集上的性能都优于ECAPA-TDNN(1024)，其中在VoxCeleb1-E上EER降低了9%，在VoxCeleb1-O上minDCF降低了14%，而参数量仅为ECAPA-TDNN(1024)的52%。这说明本文方法能够更高效地提取更有效的说话人嵌入。

表2 与其他方法的比较

Table 2 Comparison with other methods

模型	FLOPs	参数量	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-H	
			EER/%	minDCF	EER/%	minDCF	EER/%	minDCF
ResNetSE34	21.06×10^9	22.06×10^6	1.35	0.127	1.72	0.191	3.11	0.281
MFA-Conformer	1.90×10^9	19.76×10^6	1.07	0.112	1.35	0.153	2.59	0.247
ECAPA-TDNN(1024)	4.11×10^9	15.05×10^6	0.96	0.106	1.30	0.143	2.58	0.249
本文模型	1.38×10^9	7.85×10^6	**0.95**	**0.091**	**1.18**	**0.130**	**2.36**	**0.234**

5.2 前端模型有效性实验结果及分析

表3列出了前端模型有效性实验的结果。基线模型和Conv2d(128)-ECAPA的比较结果说明直接添加二维卷积来引入频率维的局部信息并不是高效的，虽然Conv2d(128)-ECAPA相较于基线模型性能有一定的提升，但却引入了大量的参数，而与4.6节模型参数量接近的Conv2d(32)-ECAPA，实际上模型性能却降低了。另一方面，ECAPA-TDNN-D的EER相较于基线模型降低了5%，说明直接加深模型深度确实能够带来一定的性能提升。而4.6节中所提模型的EER相较于基线模型降低了9%，这说明提出的前端模型带来的性能提升不仅仅是因为加深了模型的深度，更重要的是引入了模拟二维卷积渐进地建模局部频率特征的过程。

表3 前端模型有效性实验结果

Table 3 Results of the front-end model effectiveness

模型	FLOPs	参数量	VoxCeleb1-O EER/%
基线模型	1.04×10^9	6.19×10^6	1.17
Conv2d(128)-ECAPA	4.76×10^9	10.36×10^6	1.10
Conv2d(32)-ECAPA	1.39×10^9	7.00×10^6	1.19
ECAPA-TDNN-D	1.29×10^9	7.42×10^6	1.11
4.6节模型	1.22×10^9	7.06×10^6	1.06

5.3 主干模型及反向特征融合有效性实验分析

表4列出了为证明对主干模型改进的有效性而进行的实验的结果。其中瓶颈层($r=0.5$)与基线模型($r=1$)的EER相近，但是参数量和计算量却高于基线模型，这可能是由于瓶颈层Res2Block中的一维卷积压缩了特征通道数，导致了一定的信息损失，进而导致Res2Conv1d无法得到足以建模不同时间尺度的特征。相反，逆瓶颈层中一维卷积增加了特征的通道数，使得Res2Conv1d得到的信息更加充分，进而更加有效地建模不同时间尺度的特征，逆瓶颈层($r=2$)与基线模型实验结果的对比也印证了这一点。另一方面，比较本文提出的特征融合方法(4.7节模型)和原始的特征融合方法($r=2$)，本文模型在进一步降低了EER的同时还减少了模型整体的参数量，这足以说明本文提出的多层特征融合方法能够更高效地融合不同层的特征。

表4 改进后主干模型有效性实验结果

Table 4 Results of the improved backbone model effectiveness

模型	FLOPs	参数量	VoxCeleb1-O EER/%
$r=1.0$	1.04×10^9	6.19×10^6	1.17
$r=0.5$	1.87×10^9	10.73×10^6	1.16
$r=2.0$	1.51×10^9	8.55×10^6	1.11
4.7节模型	1.20×10^9	6.98×10^6	1.05

5.4 模型感受野可视化

图5展示了本文模型不同层感受野的可视化结果。图5(a)为主干模型第一层输出中的某几帧在输入特征图中的感受野,可以看出主干模型在第一层便对整个频率范围的特征进行建模,缺少针对局部特征建模的过程。图5(b)—图5(d)为前端模型中第一、二、三层输出中的某几帧在输入特征图中的感受野范围。可以看出,本文通过子频带划分有效地将前端模型的每一层感受野强制限制在局部的频率范围中,进而实现了引入对局部频率特征进行建模的目的。对于整个模型来说,低层用于对局部特征进行建模,高层用于对全局特征进行建模,模拟了二维卷积的建模过程。5.2节中的实验证明了这种建模过程能够有效增强TDNN的建模能力。

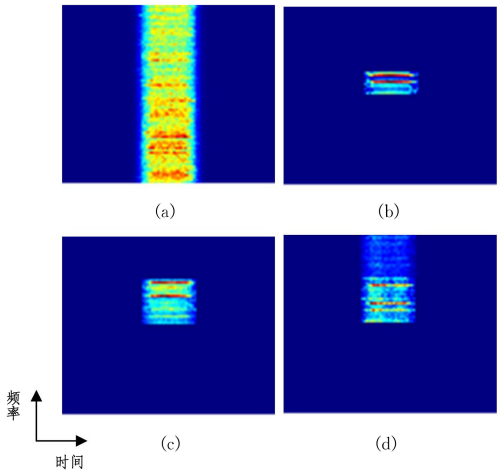

图5 不同层感受野可视化

Fig. 5 Visualization of receptive field in different layers

结束语 本文针对现有说话人确认方法中基于TDNN的帧级特征提取器存在的问题提出了两点优化:1)针对TDNN无法有效对局部频率特征进行建模的问题,提出了一种新的前端模型,利用子频带划分并逐层减少子频带数量使模型渐进地建模局部频率特征;2)针对TDNN现有的多层特征融合方法无法高效地建模高层特征和低层特征之间的复杂关系,提出了一种新的特征融合方式,在主干模型中设置一条反向路径,反向路径中每一层只专注于当前层特征和由上层传递来的特征之间的关系。实验表明,改进后的方法能够以更少的参数量取得超过目前主流方法的性能。

由于本文提出的前端模型与多层特征融合方法并未对主干模型部分进行修改,因此可以在对全部频率进行直接建模的模型(如TDNN,Transformer)中即插即用,以增强模型在语音处理任务中对局部频率特征的建模以及提取特征的层次多样性。在未来的工作中,将继续在其他语音处理任务中应用本文方法来进一步证明方法的通用性。

参 考 文 献

[1] SHOME N, SARKAR A, GHOSH A K, et al. Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges[J]. Periodica Polytechnica Electrical Engineering and Computer Science, 2023, 67(3): 300-336.

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

[3] WAN Z K, REN Q H, QIN Y C, et al. Statistical pyramid dense time delay neural network for speaker verification[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2022: 7532-7536.

[4] BENHAFID Z, SELOUANI S A, AMROUCHE A, et al. Attention-based factorized TDNN for a noise-robust and spoof-aware speaker verification system[J]. International Journal of Speech Technology, 2023, 26(4): 881-894.

[5] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification[C]// Proceedings Interspeech. 2020: 3830-3834.

[6] ZHANG X, LIU Q, GUO Q, et al. EIPFD-ResNet: Emphasized Information Propagation and Feature Distribution in ResNet Based Speaker Verification[J]. Journal of Chinese Computer Systems. 2023, 44(3): 463-470.

[7] KYNYCH F, ZDANSKY J, CERVA P, et al. Online Speaker Diarization Using Optimized SE-ResNet Architecture[C]// International Conference on Text, Speech, and Dialogue. Cham: Springer Nature Switzerland, 2023: 176-187.

[8] CHUNG J S, HUH J, MUN S, et al. In Defence of Metric Learning for Speaker Recognition[C]// Proceedings Interspeech. 2020: 2977-2981.

[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(ICASSP). New York: IEEE Press, 2014: 4052-4056.

[10] 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(ICASSP). New York: IEEE Press, 2018: 5329-5333.

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

[12] GAO Z, SONG Y, MCLOUGHLIN I, et al. Improving Aggregation and Loss Function for Better Embedding Learning in End-to-End Speaker Verification System[C]// Proceedings Interspeech. 2019: 361-365.

[13] TANG Y, DING G, HUANG J, et al. Deep speaker embedding learning with multi-level pooling for text-independent speaker verification[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). New York: IEEE Press, 2019: 6116-6120.

[14] 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, 2010, 19(4): 788-798.

[15] CHOWDHURY F A R R, WANG Q, MORENO I L, et al. Attention-based models for text-dependent speaker verification [C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). New York: IEEE Press, 2018: 5359-5363.

[16] WANG Z, YAO K, LI X, et al. Multi-resolution multi-head at-

tention in deep speaker embedding[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York:IEEE Press,2020:6464-6468.

[17] ZHANG Y,LV Z,WU H,et al. MFA-Conformer:Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification[C]// Proceedings Interspeech. 2022:306-310.

[18] LI C,MA X,JIANG B,et al. Deep speaker:an end-to-end neural speaker embedding system[J]. arXiv:1705.02304,2017.

[19] GU B,GUO W,DAI L,et al. An improved deep neural network for modeling speaker characteristics at different temporal scales [C]// 2020 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). New York:IEEE Press,2020:6814-6818.

[20] THIENPONDT J,DESPLANQUES B,DEMUYNCK K. Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification[C]// Proceedings Interspeech. 2021:2302-2306.

[21] LIU T,DAS R K,LEE K A,et al. MFA:TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York:IEEE Press,2022:7517-7521.

[22] ZHAO Z,LI Z,WANG W,et al. PCF:ECAPA-TDNN with Progressive Channel Fusion for Speaker Verification[C]// 2023 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). New York:IEEE Press,2023:1-5.

[23] SANDLER M,HOWARD A,ZHU M,et al. Mobilenetv2:Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4510-4520.

[24] LIU Z,MAO H,WU C Y,et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:11976-11986.

[25] LIN T Y,DOLLÁR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2117-2125.

[26] JUNG Y,KYE S. M,CHOI Y,et al. Improving Multi-Scale Aggregation Using Feature Pyramid Module for Robust Speaker Verification of Variable-Duration Utterances[C]// Proceedings Interspeech. 2020:1501-1505.

[27] SCHROFF F,KALENICHENKO D,PHILBIN J. Facenet:A unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:815-823.

[28] LIU W,WEN Y,YU Z,et al. Large-margin softmax loss for convolutional neural networks[C]// Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48. 2016:507-516.

[29] LIU W,WEN Y,YU Z,et al. Sphereface:Deep hypersphere embedding for face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:212-220.

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

[31] DENG J,GUO J,XUE N,et al. Arcface:Additive angular margin loss for deep face recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4690-4699.

[32] NAGRANI A,CHUNG J S,XIE W,et al. Voxceleb:Large-scale speaker verification in the wild[J]. Computer Speech & Language,2020,60:101027.

[33] CHUNG J S,NAGRANI A,ZISSERMAN A. VoxCeleb2:Deep Speaker Recognition[C]// Proceedings Interspeech. 2018:1086-1090.

[34] SNYDER D,CHEN G,POVEY D. MUSAN:A Music,Speech, and Noise Corpus[J]. arXiv:1510.08484,2015.

[35] KO T,PEDDINTI V,POVEY D,et al. A study on data augmentation of reverberant speech for robust speech recognition[C]// 2017 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). New York: IEEE Press, 2017:5220-5224.

[36] PARK D S,CHAN W,ZHANG Y,et al. SpecAugment:A Simple Data Augmentation Method for Automatic Speech Recognition[C]// Interspeech. 2019:2613-2617.

WANG Mengwei, born in 1998, postgraduate. His main research interests include speaker recognition and audio classification.

YANG Zhe, born in 1978, Ph. D, associate professor. His main research interests include artificial intelligence, machine learning and big data.

(责任编辑:何杨)