跨域注意力特征融合的说话人确认方法

杨震[1,2]，王天朗[1]，郭海燕[1]，王婷婷[1]
（1. 南京邮电大学通信与信息工程学院，江苏 南京 210003；
2. 南京邮电大学通信与网络技术国家地方联合工程研究中心，江苏 南京 210003）

摘 要：针对目前说话人确认系统中前端特征的语音信号样点间结构信息缺失问题，提出了跨域注意力特征融合的说话人确认方法。首先，提出了一种基于图信号处理的图频域特征提取方法来有效利用语音信号的结构信息，将语音信号帧的每个样点作为图节点，构建语音图信号，通过图傅里叶变换以及滤波器组提取图频域特征。其次，提出了一种由残差模块与挤压−激励模块构成的注意力特征融合网络，对传统时频域特征与图频域特征进行跨域融合，来提升说话人确认系统的性能。最后，在 VoxCeleb、SITW 和 CN-Celeb 数据集上进行实验。实验结果表明，所提方法在等错误率以及最小检测代价函数的评价指标上，优于基线模型 ECAPA-TDNN。

关键词：说话人确认；图信号处理；注意力特征融合
中图分类号：TN912.34
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023142

Speaker verification method based on cross-domain attentive feature fusion

YANG Zhen[1,2], WANG Tianlang[1], GUO Haiyan[1], WANG Tingting[1]
1. College of Communication & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. National Local Joint Engineering Research Center for Communications and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: Aiming at the problem that the lack of structure information among speech signal sample in the front-end acoustic features of speaker verification system, a speaker verification method based on cross-domain attentive feature fusion was proposed. Firstly, a feature extraction method based on the graph signal processing (GSP) was proposed to extract the structural information of speech signals, each sample point in a speech signal frame was regarded as a graph node to construct the speech graph signal and the graph frequency information of the speech signal was extracted through the graph Fourier transform and filter banks. Then, an attentive feature fusion network with the residual neural network and the squeeze-and- excitation block was proposed to fuse the features in the traditional time-frequency domain and those in the graph frequency domain to promote the speaker verification system performance. Finally, the experiment was carried out on the VoxCeleb, SITW, and CN-Celeb datasets. The experimental results show that the proposed method performs better than the baseline ECAPA-TDNN model in terms of equal error rate (EER) and minimum detection cost function (min-DCF).

Keywords: speaker verification, graph signal processing, attentive feature fusion

0 引言

随着深度学习的兴起，涌现了大量基于深度神经网络的说话人确认模型，这些模型的说话人特征提取过程主要分为两部分：前端声学特征提取与话语级说话人嵌入特征提取。目前，主流的说话人识别模型的前端声学特征使用梅尔频率倒谱系数（MFCC, mel-frequency cepstral coefficient）或对数梅

收稿日期：2023−05−11；**修回日期：**2023−07−27
通信作者：王天朗，1021010605@njupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62071242）
Foundation Item: The National Natural Science Foundation of China (No.62071242)

尔滤波器组能量（FBank, log-mel filter bank energy）等声学特征[1]。这些特征都是在短时傅里叶变换的幅度谱的基础上得到的，因此只利用了语音信号的时频信息，而忽略了反映样点间相关性的信号结构信息[2]。然而，前端声学特征中结构信息的丢失，会导致后续话语级说话人嵌入特征提取网络获得的信息不完整，进而制约了说话人确认方法的性能。

针对上述说话人身份特征提取中信号结构信息的缺失问题，文献[3]通过在一条语音中提取多个片段级说话人嵌入特征，在后端判别模型中将每个说话人嵌入作为一个图节点，利用图注意力网络（GAN, graph attention network）提取特征的结构信息进行判别。文献[4]将神经网络提取的帧级别特征作为图的节点，利用 GAN 与图池化替代原始的统计池化层，提取帧级别特征的结构信息，得到话语级说话人特征。上述这些工作利用的是高维特征间的关联性等结构信息，并未关注反映原始语音信号样点间关联性的结构信息。

同时，为了获得更多的说话人身份信息，一些研究者提出了特征融合方法。文献[5]在残差网络的基础上提出了通道注意力模块（CAM, channel attention module）以及并行注意力（CA, coordinate attention）来融合恒等映射特征与残差特征，在提取高维特征的同时，保留了低维特征。文献[6]提出一种多特征融合的说话人确认方法，分别将 MFCC 特征、频率域线性预测（FDLP, frequency domain linear prediction）特征以及原始语音信号输入各分支网络，在各分支经过池化层之后，通过一个共同的话语级特征提取网络，之后计算多种输入特征的交叉熵损失函数的和，将其作为最终的损失函数来更新网络参数。然而，上述方法主要针对同一个域的特征进行融合，并没有额外增加信号的结构信息。此外，在其他研究领域，也有通过融合多领域特征进行各种任务的方法。文献[7]将对数梅尔谱图和测度向量经过卷积神经网络后的输出进行拼接，得到了融合特征，用于后续的干扰语音评估；文献[8]通过 U-Net 提取 4 个不同尺度的视觉特征后，将归一化的特征进行拼接，得到了融合视觉特征。此外，其他融合方法通过各种算法赋予不同特征不同的权重后进行特征叠加[9]。然而，无论是特征的拼接还是叠加，都是线性操作，无法充分利用多领域特征之间的相关性。

为了克服说话人识别中前端特征提取的结构信息缺失问题，本文使用图信号处理（GSP, graph signal processing）技术[10]提取语音样点之间的图结构信息。相比于传统的数字信号处理方法，GSP 可以通过边和边权重充分利用信号点之间的关系。同时，理论上已经证明，离散傅里叶变换（DFT, discrete Fourier transform）是图傅里叶变换（GFT, graph Fourier transform）在有向周期循环图下的一个特例[10]。此外，已有研究表明，在语音增强以及语音分离等语音信号处理任务中，采用 GSP 技术提取语音信号的结构信息，有利于提升语音信号处理任务的性能[11-15]。因此，本文使用 GSP 技术，对语音信号在帧内构建图结构，通过图傅里叶变换得到语音的图频谱，进而通过滤波器组得到图对数梅尔滤波器组能量（GFBank, graph log-mel filter bank energy）特征，以此来表征语音信号样点之间的结构信息。在此基础上，本文对传统频域特征与图频域特征进行了特征融合。与其他传统常用的特征拼接或叠加方法不同，本文引入了残差网络（ResNet, residual network）[16]和挤压–激励网络（SE, squeeze-and-excitation network）[17]进行特征融合，其中 ResNet 将 FBank 特征和 GFBank 特征映射为多通道特征，增强特征的表示能力，并通过残差连接防止梯度消失，而 SE 在 ResNet 的基础上提供了注意力机制，根据不同特征通道的重要性赋予不同权重。

本文通过提取图频域特征，并与时频域特征融合，得到跨域信息融合特征，用于基线模型 ECAPA-TDNN（emphasized channel attention, propagation and aggregation in time delay neural network）[18]。本文工作主要包括以下几个方面。

1) 提出了一种基于 GSP 的新型图频域特征，能够提取传统时频特征无法包含的信号样点间的结构信息。

2) 引入了 ResNet [16]和 SE [17]对提出的图频域特征以及传统时频域特征进行跨域注意力特征融合，提升了特征提取的效果。

3) 在 VoxCeleb1&2[19-20]、SITW（speaker in the wild）[21]和 CN- Celeb[22]数据集上的实验结果表明，本文提出的图频率特征以及特征融合网络在 ECAPA-TDNN 模型[18]上的等错误率（EER, equal error rate）与最小检测代价函数（minDCF, minimum detection cost function）均优于使用传统时频域特征的基线模型。

1　相关工作

1.1　语音图信号处理

在 GSP 中，图信号可以定义为 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E},\boldsymbol{W})$，其中，$\boldsymbol{V}$、$\boldsymbol{E}$ 和 $\boldsymbol{W}$ 分别表示图信号的顶点集、边集和边权重矩阵。对于一帧语音 $\boldsymbol{s}=[s_0,s_1,\cdots,s_{N-1}]^{\mathrm{T}}\in\mathbf{R}^N$，通过将其每个样点 s_i 视为图的顶点 v_i，可以将其从时域映射到图域，即

$$p:\ \boldsymbol{s}\to\boldsymbol{s}_{\boldsymbol{G}}\in\mathbf{R}^N,\quad \boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E},\boldsymbol{W}) \tag{1}$$

其中，$\boldsymbol{V}=\{v_0,v_1,\cdots,v_{N-1}\}$表示图信号的顶点集，$\boldsymbol{E}=\{e_{ij}\}\in\mathbf{R}^{N\times N}$表示边集，$\boldsymbol{W}=\{w_{ij}\}\in\mathbf{R}^{N\times N}$表示边权重矩阵。通常，$\boldsymbol{W}$ 由图邻接矩阵 $\boldsymbol{A}$ 或图拉普拉斯矩阵 $\boldsymbol{L}$ 表示，代表图信号的拓扑结构[10]，为图信号提供结构信息。

此外，GFT 可以将信号从图域变换到图频域，其中的 GFT 基可以通过对边权重矩阵进行特征分解或奇异值分解得到[10]。由于 GFT 是对反映语音图信号结构的边权重矩阵进行分解得到的，因此由 GFT 得到的图频域特征一定程度上蕴含了语音信号的结构信息。

1.2　SE 模块

SE 模块[17]通过显式地构建不同特征通道间的相互关系，自适应地调整通道间的特征响应，从而提升模型的建模能力，共分为挤压与激励两步。挤压时，对输入 $\boldsymbol{Y}\in\mathbf{R}^{H\times W\times C}$ 的前 2 个维度进行全局池化，其中 C 为通道数，则第 c 个通道的输入 $\boldsymbol{Y}_c\in\mathbf{R}^{H\times W}$ 的输出为 z_c，表示为

$$z_c=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}y_c(i,j) \tag{2}$$

对于激励操作，其输出 $\boldsymbol{h}=[h_1,h_2,\cdots,h_C]\in\mathbf{R}^C$ 为

$$\boldsymbol{h}=\sigma(\boldsymbol{W}_2(\delta(\boldsymbol{W}_1\boldsymbol{z}+\boldsymbol{b}_1))+\boldsymbol{b}_2) \tag{3}$$

其中，$\sigma(\cdot)$为 sigmoid 激活函数，$\delta(\cdot)$为 ReLU 函数，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为 2 个全连接层的权重矩阵，$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 为 2 个全连接层的偏置，$\boldsymbol{z}=[z_1,z_2,\cdots,z_C]^{\mathrm{T}}\in\mathbf{R}^C$。$\boldsymbol{h}$ 中的元素的取值范围为 0～1，将其作用于最初的输入，可得 SE 模块的输出 $\boldsymbol{Y}^*\in\mathbf{R}^{H\times W\times C}$，其第 c 个通道的输出 $\boldsymbol{Y}_c^*\in\mathbf{R}^{H\times W}$ 为

$$\boldsymbol{Y}_c^*=h_c\boldsymbol{Y}_c \tag{4}$$

2　本文方法

2.1　模型结构

本文提出了一种跨域注意力特征融合的说话人确认方法，其模型结构如图 1 所示。模型由图结构特征提取、时频域特征提取、注意力特征融合、说话人嵌入特征提取以及损失函数五部分组成。其中，灰色为本文创新部分。在图结构特征提取模块，本文提出了一种新的基于 GSP 的图频域特征，即 GFBank 特征。在注意力特征融合模块，本文提出了使用 ResNet 和 SE 模块进行注意力特征融合的方法。说话人特征嵌入提取模块使用 ECAPA-TDNN 模型[18]。

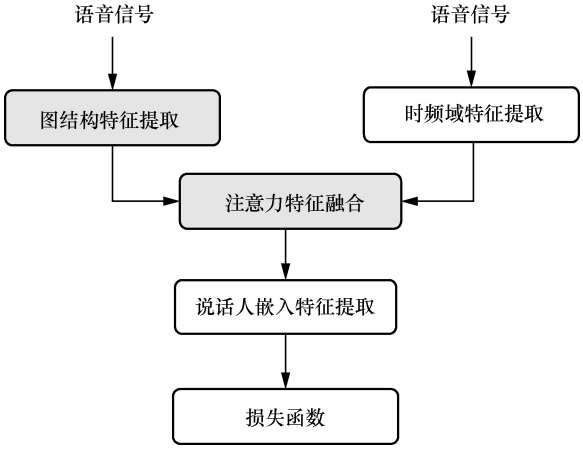


图 1　模型结构

2.2　图对数梅尔滤波器组能量特征

如图 2 所示，GFBank 特征的提取包括预加重、分帧、构建图信号、$\lg(|\mathrm{GFT}|^2)$ 以及滤波器组五部分。其中，灰色为本文创新部分。预加重通过增加语音信号的高频分量，可以有效补偿声音传输过程中高频分量的损失。鉴于语音信号的时变非平稳性，对语音进行分帧的短时处理，以有效减少语音非平稳性的影响。预加重与分帧过程与传统 FBank 特征提取[1]相同，这里省略。

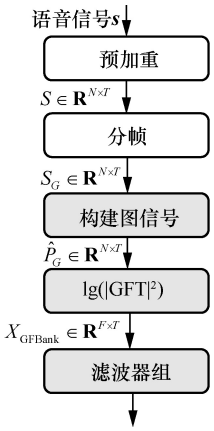


图 2　GFBank 特征提取过程

语音分帧之后，一段长语音被分成多段有重叠

的短语音，此时，语音信号帧内与帧间均存在相关性[13]，因此在语音信号的帧内和帧间均可构建图结构。考虑到说话人嵌入提取的 TDNN 通过计算帧间特征的卷积，可以获得语音信号帧间的相关性。因此，本文仅考虑语音信号帧内的相关性，具体而言，本文考虑语音信号帧内相邻 k 个样点之间的相关性，使用 k 阶移位（k-shift）图[11]$\boldsymbol{\Psi}_k \in \mathbf{R}^{N\times N}$ 作为图邻接矩阵，构建语音图信号，其图拓扑结构如图 3 所示，当前节点仅与本节点以及其后的 $k-1$ 个节点存在直接的边相连，且具有循环移位特性，图邻接矩阵 $\boldsymbol{\Psi}_k$ 第 i 行第 j 列元素为

$$\psi_{k,ij}=\begin{cases}1, & (j-i)\bmod N<k\\ 0, & \text{其他}\end{cases} \tag{5}$$

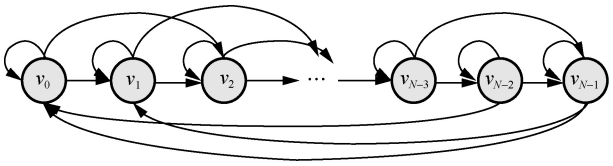


图 3 k 阶移位图结构

设预加重和分帧后的语音信号 $\boldsymbol{S}\in\mathbf{R}^{N\times T}$，其中，$N$ 为帧长，T 为帧数。根据式(1)，通过 k-shift 图将其映射到图域，得到语音图信号 $\boldsymbol{S}_G\in\mathbf{R}^{N\times T}$。时域语音信号映射到图域后，图节点的值与原语音信号样点值相同，但增加了节点之间的边连接。因此，需要对语音图信号进行图滤波或变换到图频域进一步处理。对于时域的语音信号，可以使用 DFT 得到其频谱；对于图信号，可以使用 GFT 得到其图频谱；对于有向图信号，通过对邻接矩阵 $\boldsymbol{\Psi}_k$ 进行奇异值分解，可以得到其图傅里叶变换基，即

$$\boldsymbol{\Psi}_k=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

其中，$\boldsymbol{\Sigma}=[\boldsymbol{\sigma}_0,\boldsymbol{\sigma}_1,\cdots,\boldsymbol{\sigma}_{N-1}]\in\mathbf{R}^{N\times N}$ 为奇异值矩阵，奇异值 $\boldsymbol{\sigma}_n\ (n=0,1,\cdots,N-1)\in\mathbf{R}^N$ 为图频率，左奇异矩阵 $\boldsymbol{U}=[\boldsymbol{u}_0,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_{N-1}]\in\mathbf{R}^{N\times N}$，$\boldsymbol{u}_n\in\mathbf{R}^N$（$n=0,1,\cdots,N-1$）为图频率 $\boldsymbol{\sigma}_n$ 对应的图频率分量，且 $\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{U}^{-1}=\boldsymbol{V}^{\mathrm{T}}$ 为图傅里叶变换基。由于一帧语音信号的点数过多，在图中难以观察，因此以包含 15 个顶点的 3-shift 图信号为例，其第 1 个～第 4 个图频率分量如图 4 所示，每个图频率分量 k 个顶点（即信号样点）之间有边连接，图频率越高，表示顶点的值沿着边的振荡越快，因此图频率特征表示了信号样点间的结构信息。

借助图傅里叶变换基，可以得到语音图信号 S_G 经 GFT 后的图频谱为

$$\hat{\boldsymbol{S}}_G=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{S}_G=[\hat{\boldsymbol{s}}_{G,0},\hat{\boldsymbol{s}}_{G,1},\cdots,\hat{\boldsymbol{s}}_{G,T-1}] \tag{7}$$

其中，$\hat{\boldsymbol{s}}_{G,t}(t=0,1,\cdots,T-1)$ 对应第 t 帧语音的图频率系数，对 $\hat{\boldsymbol{S}}_G$ 各元素 $\hat{S}_{G,ij}$ 求平方可以得到能量谱 $\hat{\boldsymbol{P}}_G$。

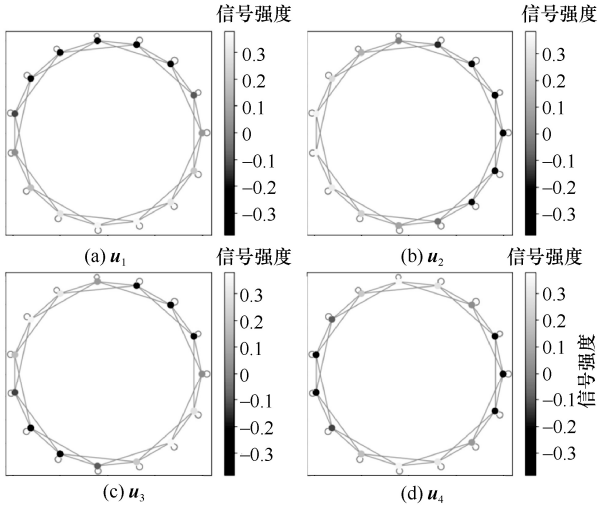


图 4 k-shift 图信号的图频率分量

在 FBank 特征提取过程中，利用人耳对低频信号敏感、高频信号不敏感的特点，设计了梅尔滤波器组，得到了符合人耳特性的声学特征。在图频率域处理时，为了实现与 FBank 特征对齐，同时减小特征参数，使用滤波器组 $\mathbf{FB}\in\mathbf{R}^{N\times F}$ 对图能量谱 $\hat{P}_G$ 进行滤波，即

$$\boldsymbol{X}_{\mathrm{GFBank}}=\left(\hat{\boldsymbol{P}}_G^{\mathrm{T}}\mathbf{FB}\right)^{\mathrm{T}}\in\mathbf{R}^{F\times T} \tag{8}$$

图 5 给出了 VoxCeleb2 数据集中 id00012/21Uxsk56VDQ/00001.wav 语音中提取的 FBank 与 GFBank 特征对比。从图 5 可以看出，FBank 特征谱的频率分布范围为−15～0 dB，GFBank 特征谱的频率分布范围为−15～−5 dB，GFBank 特征谱能量更加集中。

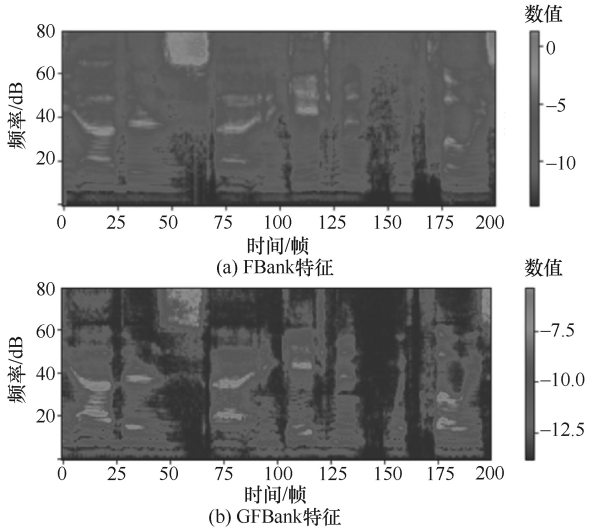


图 5 FBank 与 GFBank 特征对比

此外，本文分析了 VoxCeleb1&2 数据集中每个语音的 FBank 特征与 GFBank 特征的最大频率差的分布，如图 6 所示。从图 6 可以看出，GFBank 特征的最大幅度差主要分布在 5～17 dB，FBank 特征的最大幅度差主要分布在 10～22 dB。由图 5 与图 6 可知，相比 FBank 特征，语音信号的 GFBank 特征由于考虑了信号样点间的图结构，频谱的能量更加集中，也验证了图频率特征能够反映信号样点间的结构信息。因此，传统时频域的 FBank 特征与图域的 GFBank 特征存在较大差异，这使简单的线性叠加或者是拼接的特征融合方法都无法充分融合两者特征，需要一种非线性的自适应的融合方法来动态调整 2 种特征的权重分配。

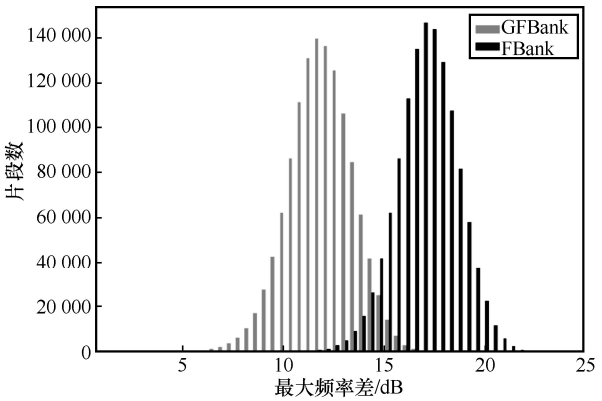


图 6　VoxCeleb1&2 数据集中 FBank 与 GFBank 最大频率差的分布

2.3　注意力特征融合

原始的 ECAPA-TDNN 模型中仅使用了 FBank 特征，未利用语音信号的结构信息，为此本文加入了图域特征 GFBank。由于 FBank 与 GFBank 是属于不同域的 2 种特征，关注语音的不同方面，因此不能通过简单的特征叠加或拼接来融合。无论是特征叠加还是拼接，都是线性操作，无法充分利用多领域特征之间的相关性，并且特征的拼接会改变输入特征维度，对后续网络的性能产生影响。

本文提出的注意力特征融合方法主要由 ResNet[16]和 SE[17]组成。具体而言，由 ResNet 组成的卷积层通过不同卷积核和非线性激活函数可以将 FBank 特征和 GFBank 特征映射为多通道特征，进一步提升特征的表示能力。然后，利用 SE 模块的挤压操作聚合每个特征通道，计算注意力系数，再经过激励操作，得到注意力权重分配后的特征，并与原始特征进行残差连接，以避免产生梯度消失问题。最后，经过一层卷积层将多通道特征聚合为单通道特征，得到最终的跨域融合特征。通过这种方式，不仅实现了注意力融合，同时还保持了输入特征维度的不变性，避免了由特征维度变化引起的影响。注意力特征融合网络结构如图 7 所示。

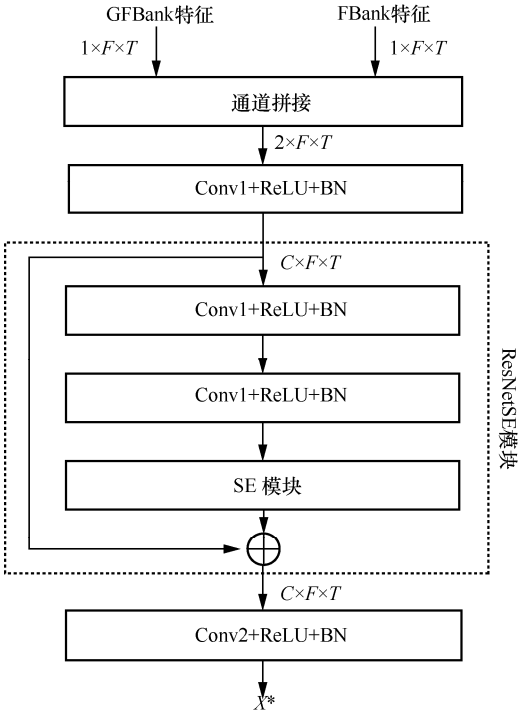


图 7　注意力特征融合网络结构

对于注意力特征融合网络，首先，将 2 种前端特征 X_{FBank} 与 X_{GFBank} 构成双通道特征，即注意力特征融合网络的输入

$$X=\mathrm{CAT}(X_{\mathrm{FBank}},X_{\mathrm{GFBank}})\in\mathbf{R}^{2\times F\times T} \tag{9}$$

然后，通过三层的二维卷积神经网络（2D CNN, two-dimension convolution neural network）将特征通道数扩大到 C，以获取更多的通道信息，再借助 SE 模块的注意力机制，自适应调整特征通道的特征响应，最后通过一个 2D CNN 聚合多通道特征信息，将特征通道数降为 1。此外，在第一层网络与最后一层网络之间加入了残差连接，其过程如下

$$X^{*}=\mathrm{AFF}(X)=\mathrm{Conv2}(\mathrm{ResNetSE}(\mathrm{Conv1}(X))) \tag{10}$$

$$\mathrm{ResNetSE}(X)=\mathrm{SE}(\mathrm{Conv1}(\mathrm{Conv1}(X)))+X \tag{11}$$

其中，Conv1 和 Conv2 为不同卷积核的 2D CNN，SE 为 SE 模块，每层网络都省略了批标准化（BN, batch normalization）和 ReLU 激活函数。注意力特征融合网络中每层网络的详细参数如表 1 所示。

表 1 注意力融合网络参数

模块名称	参数	输出
Conv1	核函数：$3\times3\times C$，步长：1×1	$C\times F\times T$
Conv2	核函数：$3\times3\times1$，步长：1×1	$1\times F\times T$
SE	衰减因子 r=8	$C\times F\times T$

2.4 融合特征应用

目前，说话人确认的主流模型包括基于 TDNN 的 ECAPA-TDNN 模型[18]、基于 ResNet 的 ResNet34 模型[23]，以及基于 Transformer 的模型[24-25]等。主流的说话人识别数据集包括 VoxCeleb1&2[19-20]、SITW[21]和 CN-Celeb[22]等数据集，其中 VoxCeleb 数据集的说话人数量最多，是大多数研究者使用的。而本文选取的 ECAPA-TDNN 模型在 VoxCeleb 数据集上的结果优于其他主流模型。因此，本文选择 ECAPA-TDNN 作为基线模型。为了验证所提出的融合特征的有效性，本文在 ECAPA-TDNN 模型上进行实验，模型结构如图 1 所示。将模型中的单一 FBank 特征替换为融合特征，作为网络的输入。

3 实验及结果

3.1 实验数据与评估指标

本文分别在 VoxCeleb1&2、SITW 和 CN-Celeb1 数据集上进行实验，以验证所提方法。实验使用 VoxCeleb2 的开发集作为训练集，其中包含 5 994 个说话人的 1 092 009 条语音。此外，模型训练过程中，使用了 MUSAN 数据集[26]、RIR 数据集[27]以及 SpecAugment[28]进行数据增强。实验使用 VoxCe-leb1、SITW 以及 CN-Celeb1 作为测试集，包括 Vox1-E cl.、Vox1-H cl.、SITW-dev、SITW-eval 以及 CN-Celeb1-eval。考虑到 VoxCeleb 和 SITW 包含重叠的说话人，本文在 SITW 中去除了重复的说话人语音数据。实验结果使用等错误率和先验目标概率为 0.01 的最小检测代价函数作为评估指标。

3.2 实验设置

实验语音使用 32 ms 窗函数分帧，帧移为 12.5 ms，每段语音截取 200 帧，得到 80 维的 F-Bank 特征和 GFBank 特征。损失函数使用边缘（margin）为 0.2、尺度因子（scale）为 30 的 AAM-softmax[29]损失。初始学习率设置为 0.001，每次 epoch 学习率下降 3%，数据批大小设置为 400。使用 Adam 优化器对网络参数进行优化。

在训练模型的基础上，将 AAM-softmax 损失函数的边缘和尺度因子分别设置为 0.4 与 60，每条语音的持续时间加长到 300 帧，对模型参数进行微调。其中，ET-FBank 模型为原始的使用 FBank 特征作为输入的 ECAPA-TDNN 模型，ET-AFF-CS*x* 为本文提出的基于跨域注意力的通道数为 *x* 的特征融合网络，融合 FBank 和 GFBank 后的特征作为输入的 ECAPA-TDNN 模型。

3.3 实验结果及分析

表 2～表 4 分别列出了本文提出的 ET-AFF-CS*x* 模型与基线模型 ET-FBank 在 VoxCeleb、SITW 和 CN-Celeb 数据集上的实验结果。值得注意的是，基线模型 ECAPA-TDNN 使用 FBank 特征作为输入，在表 2～表 4 中，本文用基线模型 ET-FBank 来表示原始的 ECAPA-TDNN 模型，以和本文提出的 ET-AFF-CSx 模型区分。

如表 2 所示，本文提出的 ET-AFF-CS*x* 模型的 EER 和 minDCF 均低于 ET-FBank 模型与 ResNet34 模型。其中，ET-AFF-CS128 模型取得了最低的 EER 与 minDCF，在 Vox1-E cl.上的 EER 与 minDCF 分别为 1.121%和 0.070，相比基线模型的 EER 与 minDCF 分别降低了 12.53%和 17.65%；在 Vox1-H cl.上的 EER 和 minDCF 分别为 2.010%和 0.124，与基线模型的 EER 和 minDCF 相比，分别降低了 16.63%和 16.78%。此外，从表 2 还可以发现，随着注意力特征融合网络的通道数增加，模型的性能也在不断提升。

表 2 不同模型在 VoxCeleb1 数据集上的结果对比

模型	参数量/个	Vox1-E cl.		Vox1-H cl.	
		EER	minDCF	EER	minDCF
ET-FBank（基线）	14.73×10^6	1.279%	0.085	2.411%	0.149
ET-AFF-CS32	14.75×10^6	1.195%	0.076	2.146%	0.133
ET-AFF-CS64	14.80×10^6	1.142%	0.073	2.084%	0.131
ET-AFF-CS128	15.03×10^6	1.121%	0.070	2.010%	0.124

如表 3 所示，本文提出的 ET-AFF-CS*x* 模型在 SITW 数据集上的 EER 与 minDCF 均优于其余模型。其中，ET-AFF-CS32 模型在 SITW-dev 上取得了最低的 EER，为 1.617%，相比基线模型降低了 16.09%；ET-AFF-CS128 模型在 SITW-dev 上的 minDCF 为 0.098，相比基线模型降低了 23.44%；在 SITW-eval 上的 EER 和 minDCF 分别为 1.725%和 0.108，相比基线模型分别降低了 15.85%和 18.80%。

如表 4 所示，本文模型在 CN-Celeb1 数据集上的各项评价指标均优于基线模型，其中

ET-AFF-CS64 模型取得了最低的 EER，相比基线模型降低了 9.87%；ET-AFF-CS32 与 ET-AFF-CS128 的 minDCF 最低，相比基线模型降低了 13.20%。

表 3　不同模型在 SITW 数据集上的结果对比

模型	SITW-dev		SITW-eval	
	EER	minDCF	EER	minDCF
ET-FBank（基线）	1.927%	0.128	2.050%	0.133
ET-AFF-CS32	1.617%	0.108	1.804%	0.110
ET-AFF-CS64	1.711%	0.107	1.804%	0.110
ET-AFF-CS128	1.733%	0.098	1.725%	0.108

表 4　不同模型在 CN-Celeb1 数据集上的结果对比

模型	CN-Celeb1-eval	
	EER	minDCF
ET-FBank（基线）	15.868%	0.568
ET-AFF-CS32	14.559%	0.493
ET-AFF-CS64	14.302%	0.500
ET-AFF-CS128	14.959%	0.493

总体而言，本文提出的基于不同通道数的注意力融合特征模型的性能在 VoxCeleb、SITW 以及 CN-Celeb 这 3 个数据集上均优于基线模型，同时，ET-AFF-CS128 模型在大多数数据集上实现了最好的性能。

3.3.1　不同特征融合方法对比

为了验证本文提出的注意力特征融合网络方法的有效性，实验比较了特征叠加、特征拼接与本文方法在 VoxCeleb1 数据集上的性能，如表 5 所示。其中，ET-CAT 为将 FBank 和 GFBank 沿频率维拼接作为输入特征的 ECAPA-TDNN 模型；ET-ADD 为使用 FBank 和 GFBank 的线性叠加特征作为输入特征的 ECAPA-TDNN 模型。从表 5 可以看出，拼接或线性叠加等融合方法无法充分利用 FBank 与 GFBank 特征，反而会造成模型性能的下降，而本文提出的注意力特征融合方法通过自适应分配特征权重，充分利用了 FBank 与 GFBank 特征，实现了模型性能的提升。

表 5　不同特征融合方法在 VoxCeleb1 数据集上的结果对比

模型	参数量/个	Vox1-E cl.		Vox1-H cl.	
		EER	minDCF	EER	minDCF
ET-ADD	14.73×10^6	1.388%	0.086	2.534%	0.154
ET-CAT	15.12×10^6	1.308%	0.082	2.357%	0.142
ET-AFF-CS128	15.03×10^6	1.121%	0.070	2.010%	0.124

3.3.2　与其他模型实验结果对比

表 6 列出了本文方法与当前的主流模型 ResNet34[23]、ECAPA-TDNN[18]以及其他新模型 ReaNet34-GAT[4]、ResNet34-ft-CBAM[30]、MFCC+FDLP+wav2vec[6]、SAEP[24]、GCSA[25]和 MLP-SVNet[31]在 VoxCeleb1 数据集上 EER 的实验结果对比。

表 6　不同模型在 VoxCeleb1 数据集上的 EER 对比

方法	模型	Vox1-O cl.	Vox1-E cl.	Vox1-H cl.
基于 ResNet	ResNet34-GAT	1.75%	—	—
	ResNet34	1.46%	1.55%	2.76%
	ResNet34-ft-CBAM	1.08%	1.43%	2.67%
基于 TDNN	ECAPA-TDNN	1.05%	1.28%	2.41%
	MFCC + FDLP + wav2vec	2.86%	—	—
基于 Transformer	SAEP	2.91%	2.87%	4.75%
	GCSA	1.96%	2.07%	3.65%
基于 MLP	MLP-SVNet	1.36%	1.46%	2.49%
本文方法	ET-AFF-CS128	**0.95%**	**1.12%**	**2.01%**

如表 6 所示，相比其他模型，本文方法的 EER 在 Vox1-O cl.测试集上提升了 9.52%～67.35%，在 Vox1-E cl.测试集上提升了 12.5%～60.98%，在 Vox1-H cl.测试集上提升了 16.60%～57.68%。

3.3.3　消融实验

本节设计消融实验，以验证本文提出的基于图信号处理的 GFBank 特征提取，以及 FBank 与 GFBank 的注意力特征融合网络的有效性，实验结果如表 7 所示。其中，FBank 和 GFBank 均为单一特征，未使用注意力特征融合网络。FBank + LFCC 为使用 FBank 与线性频率倒谱系数（LFCC, linear frequency cepstral coefficient）的融合特征，FBank + FBank 为使用 FBank 与自身融合的特征，ET-R-CS64 为仅使用 ResNet 进行特征融合的模型，ET-SE-CS64 为仅使用 SE 进行特征融合的模型（保留图 7 中第一层与最后一层卷积层）。从表 7 可以看出，单一的 GFBank 特征的模型性能略差于单一的 FBank 特征，但两者的融合特征的模型性能优于单一的 FBank 特征，这证实了跨域融合 FBank 和 GFBank 特征能有效地提升说话人确认的性能。因此基于图信号处理的 GFBank 特征为模型提供了信号之间的结构信息，从而实现了模型识别性能的提升。此外，从表 7 还可以看出，采用 FBank 与 LFCC 的融合特征，或 FBank 与自身融合的特征，相比于采用单一的 FBank 特征，

模型的性能更差，这说明采用本文提出的跨域融合特征能够提升模型的性能并不是因为网络参数的增加，而是因为 GFBank 特征提供了 FBank 特征以外的信息，这进一步证实了 GFBank 特征的有效性。最后，在单独使用 ResNet 或 SE 进行特征融合的消融实验中，ET-R-CS64 性能优于前 4 种方法，而 ET-SE-CS64 由于缺少残差连接而导致模型性能下降。通过对比 ET-R-CS64 和 ET-AFF-CS64 的结果可以发现，SE 网络提升了仅使用 ResNet 进行融合的方法。因此验证了本文方法的有效性。

表 7　　消融实验

方法	参数量/个	Vox1-E cl.		Vox1-H cl.	
		EER	minDCF	EER	minDCF
FBank	14.73×10^6	1.279%	0.085	2.411%	0.149
GFBank	14.73×10^6	1.340%	0.086	2.492%	0.150
FBank + LFCC	14.80×10^6	1.294%	0.084	2.447%	0.149
FBank + FBank	14.80×10^6	1.388%	0.085	2.424%	0.150
ET-R-CS64	14.79×10^6	1.231%	0.079	2.320%	0.142
ET-SE-CS64	14.73×10^6	1.325%	0.083	2.452%	0.153
ET-AFF-CS64	14.80×10^6	1.142%	0.073	2.084%	0.131

3.3.4 特征泛化性实验

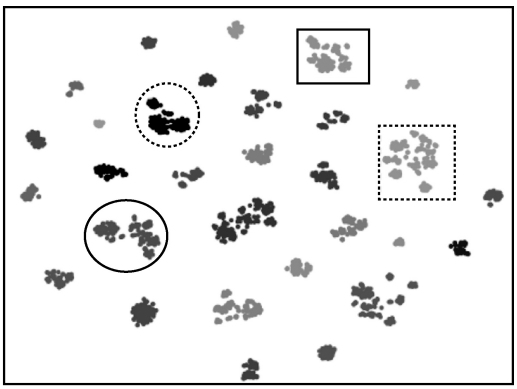
表 8 给出了使用 ResNet34 作为后端说话人特征提取网络的 EER 结果，其中 ResNet34 使用 FBank 特征，ResNet-AFF-CS64 使用融合特征。如表 8 所示，对于 ResNet34 模型，本文方法使 EER 在 Vox1-E cl.上降低了 5.69%，在 Vox1-H cl.上降低了 10.16%。由此可见，本文提出的特征融合方法不仅适用于 ECAPA-TDNN 模型，也适用于 ResNet34 模型，因此本文方法具有较好的泛用性。

表 8　　特征泛化性实验

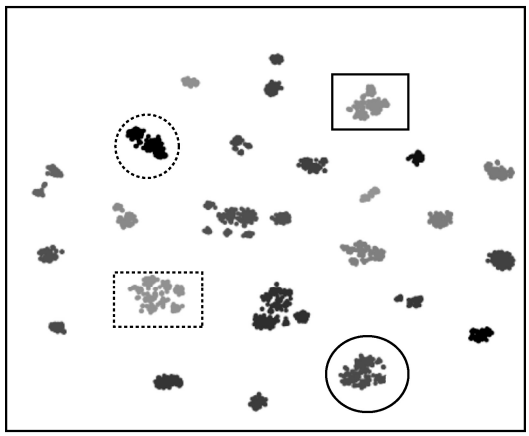
模型	Vox1-E cl.	Vox1-H cl.
ResNet34	1.406%	2.707%
ResNet34-AFF-CS64	1.326%	2.432%

3.3.5 说话人特征表示可视化图像对比

为了进一步验证本文方法的有效性，本文从 Vox1-O cl.数据集中随机选取了 23 个说话人的 2 500 条语音，分别使用 ECAPA-TDNN 和本文提出的 ET-AFF-CS128 模型提取了说话人特征表示，并采用 t 分布随机邻居嵌入（t-SNE, t-distributed stochastic neighbor embedding）[32]方法进行了可视化图像的对比，结果如图 8 所示，其中相同的线框表示同一说话人的特征。



(a) ECAPA-TDNN



(b) ET-AFF-CS128

图 8　说话人特征表示的可视化对比

从图 8 可以看出，与采用基线模型 ECAPA-TDNN 提取的说话人特征表示相比，采用 ET-AFF-CS128 模型提取的说话人特征表示对于相同说话人特征通常更加集中，有利于说话人确认任务，验证了本文提出的 ET-AFF-CS128 模型的有效性。

4 结束语

本文提出了一种基于图信号处理的 GFBank 特征，为说话人信息提取提供图结构信息，并使用注意力特征融合网络融合 FBank 与 GFBank 特征，得到跨域特征，应用于 ECAPA-TDNN 模型。在 VoxCeleb、SITW 和 CN-Celeb 数据集上的实验结果表明，与传统的单一特征相比，跨域融合特征提升了说话人识别模型的性能。此外，本文还研究了不同的特征融合方式以及不同的特征对最终的说话人识别模型性能的影响，并在 ResNet34 模型上进行了特征泛化性实验。

参考文献：

[1] ATAL B S. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460-475.

[2] 杨震, 王婷婷. 语音图信号处理理论与技术研究[J]. 南京邮电大学学报(自然科学版), 2020, 40(5): 43-51.
YANG Z, WANG T T. Research on speech graph signal processing theory and technology[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2020, 40(5): 43-51.

[3] JUNG J W, HEO H S, YU H J, et al. Graph attention networks for speaker verification[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6149-6153.

[4] SHIM H J, HEO J, PARK J H, et al. Graph attentive feature aggregation for text-independent speaker verification[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7972-7976.

[5] LIU B, CHEN Z Y, QIAN Y M. Attentive feature fusion for robust speaker verification[C]//Proceedings of Interspeech 2022. New York: ACM Press, 2022: 286-290.

[6] SANKALA S, RAFI B S M, K S R M. Multi-feature integration for speaker embedding extraction[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7957-7961.

[7] 林云, 徐怀韬, 王森, 等. 基于特征融合的通信语音干扰效果客观评估[J]. 通信学报, 2023, 44(3): 105-116.
LIN Y, XU H T, WANG S, et al. Objective assessment of communication speech interference effect based on feature fusion[J]. Journal on Communications, 2023, 44(3): 105-116.

[8] 郑金志, 汲如意, 张立波, 等. 基于 Transformer 解码的端到端场景文本检测与识别算法[J]. 通信学报, 2023, 44(5): 64-78.
ZHENG J Z, JI R Y, ZHANG L B, et al. End-to-end scene text detection and recognition algorithm based on Transformer decoders[J]. Journal on Communications, 2023, 44(5): 64-78.

[9] 秦志金, 赵茭茭, 李凡, 等. 多模态语义通信研究综述[J]. 通信学报, 2023, 44(5): 28-41.
QIN Z J, ZHAO T T, LI F, et al. Survey of research on multimodal semantic communication[J]. Journal on Communications, 2023, 44(5): 28-41.

[10] ORTEGA A, FROSSARD P, KOVAČEVIĆ J, et al. Graph signal processing: overview, challenges, and applications[J]. Proceedings of the IEEE, 2018, 106(5): 808-828.

[11] YAN X, YANG Z, WANG T, et al. An iterative graph spectral subtraction method for speech enhancement[J]. Speech Communication, 2020, 123: 35-42.

[12] WANG T T, GUO H Y, YAN X, et al. Speech signal processing on graphs: the graph frequency analysis and an improved graph Wiener filtering method[J]. Speech Communication, 2021, 127: 82-91.

[13] WANG T T, GUO H Y, ZHANG Q Q, et al. A new multilayer graph model for speech signals with graph learning[J]. Digital Signal Processing, 2022: doi.org/10.1016/j.dsp.2021.103360.

[14] WANG T T, PAN Z X, GE M, et al. Time-domain speech separation networks with graph encoding auxiliary[J]. IEEE Signal Processing Letters, 2023, 30: 110-114.

[15] ZHANG C H, PAN X. Single-channel speech enhancement using graph Fourier transform[C]//Proceedings of Interspeech 2022. New York: ACM Press, 2022: 946-950.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[18] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECA-PA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Proceedings of Interspeech 2020. New York: ACM Press, 2020: 3830-3834.

[19] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: a large-scale speaker identification dataset[J]. arXiv Preprint, arXiv: 1706.08612, 2017.

[20] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: deep speaker recognition[J]. arXiv Preprint, arXiv: 1806.05622, 2018.

[21] MCLAREN M, FERRER L, CASTAN D, et al. The speakers in the wild (SITW) speaker recognition database[C]//Proceedings of Interspeech 2016. New York: ACM Press, 2016: 818-822.

[22] FAN Y, KANG J W, LI L T, et al. CN-celeb: a challenging Chinese speaker recognition dataset[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 7604-7608.

[23] ZEINALI H, WANG S, SILNOVA A, et al. BUT system description to VoxCeleb speaker recognition challenge 2019[J]. arXiv Preprint, arXiv: 1910. 12592, 2019.

[24] SAFARI P, INDIA M, HERNANDO J. Self-attention encoding and pooling for speaker recognition[C]//Proceedings of Interspeech 2020. New York: ACM Press, 2020: 941-945.

[25] HAN B, CHEN Z Y, QIAN Y M. Local information modeling with self-attention for speaker verification[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6727-6731.

[26] SNYDER D, CHEN G, POVEY D. MUSAN: a music, speech, and noise corpus[J]. arXiv Preprint, arXiv:1510.08484, 2015.

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

[28] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[C]//Proceedings of Interspeech 2019. New York: ACM Press, 2019: 2613-2617.

[29] DENG J, GUO J, YANG J, et al. ArcFace: additive angular margin loss for deep face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 5962-5979.

[30] YADAV S, RAI A. Frequency and temporal convolutional attention for text-independent speaker recognition[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 6794-6798.

[31] HAN B, CHEN Z Y, LIU B, et al. MLP-SVNET: a multi-layer perceptrons based network for speaker verification[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7522-7526.

[32] MAATEN L V D, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11): 2579-2605.

[作者简介]

杨震（1961− ），男，江苏苏州人，博士，南京邮电大学教授、博士生导师，主要研究方向为语音信号处理与现代语音通信、无线通信中的通信与信号处理技术等。

王天朗（1998− ），男，江苏常州人，南京邮电大学硕士生，主要研究方向为语音图信号处理。

郭海燕（1983− ），女，湖北钟祥人，博士，南京邮电大学副教授、硕士生导师，主要研究方向为语音处理与现代语音通信、协作通信、无线安全传输等。

王婷婷（1992− ），女，安徽六安人，南京邮电大学博士生，主要研究方向为图信号处理、语音信号处理等。