中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-13

论文引用格式： Su Zhaopin, Fang Zhen, Zhang Guofu, Wang Yaofei, Zang Huaijuan. Multi-scale feature fusion and dynamic enhancement for voice source identification[J/OL]. Journal of Image and Graphics, XXXX:1-13. DOI: 10.11834/jig.250570.(苏兆品, 方振, 张国富, 王垚飞, 臧怀娟. 多尺度特征融合与动态增强的语音来源识别[J/OL]. 中国图象图形学报, XXXX:1-13. DOI: 10.11834/jig.250570.)[DOI:10.11834/jig.250570]

多尺度特征融合与动态增强的语音来源识别

苏兆品[1,2,3],方振[1],张国富[1,3,4],王垚飞[1,2,3],臧怀娟[1,2,3]

1. 合肥工业大学 计算机与信息学院,合肥 230601; 2. 智能互联系统安徽省实验室 (合肥工业大学), 合肥 230009; 3. 工业安全与应急技术安徽省重点实验室(合肥工业大学),合肥 230009; 4. 音视频智能防识联合实验室,合肥 230000

摘 要：目的 在多媒体取证领域,语音证据的来源识别对于司法实践和信息安全至关重要。然而,现有方法大多仅能识别手机型号,而无法精确区分个体设备,导致语音证据多被视为辅助线索而非直接有效证据。**方法** 本文提出一种多尺度特征融合与动态增强的语音来源识别模型(Multi-scale feature fusion and dynamic enhancement for voice source identification, VSI)。首先,通过残差-注意力协同网络,增强模型对不同设备相关特征的捕捉能力,提取语音信号的硬件指纹特征;其次,设计基于表达增强TDNN的整体特征提取模块,能够更好地捕捉丰富的设备特征信息;并设计基于多级残差SE-Res2Net的局部特征提取模块,有效捕捉手机设备个体之间的细微特征差异;然后,设计基于特征重校准与动态全局滤波的特征增强模块,滤除与任务无关的信息,增强与设备个体相关的特征表示;最后,构建细粒度分类模型,实现从型号到个体的跨层级设备识别。**结果** 为了验证所提模型的有效性,论文构建了包含14个手机品牌、121个不同个体设备的语音数据集。所提VSI模型的等错误率(EER)、准确率(ACC)和最小检测代价函数(minDCF)指标分别为7.50%、89.97%和0.38,相较于文中对比的其他四种方法,EER分别降低了4.75%、4.71%、5.44%和3.46%,ACC分别提升了3.98%、2.20%、4.90%和2.52%,minDCF分别下降了0.04、0.04、0.30和0.03。而且,模型在改变语音时长、采样率、编码格式和幅值环境下具有一定的鲁棒性。**结论** 这表明该模型能够将语音数据作为电子证据,为司法取证、智能终端设备身份认证等领域提供有力技术支持。

关键词： 多媒体取证;语音来源识别;多尺度特征融合;动态增强;手机个体识别

Multi-scale feature fusion and dynamic enhancement for voice source identification

Su Zhaopin[1,2,3], Fang Zhen[1], Zhang Guofu[1,3,4], Wang Yaofei[1,2,3], Zang Huaijuan[1,2,3]

1. School of Computer and Information Technology, Hefei University of Technology, Hefei230601 China; 2. Intelligent Interconnected System Anhui Laboratory(Hefei University of Technology), Hefei230009 China; 3. Anhui Province Key Laboratory of Industry Safety and Emergency Technology(Hefei University of Technology), Hefei230009 China; 4. Joint Laboratory of Intelligent Prevention and Recognition of Voice and Video, Hefei 230000, China

Abstract: Objective In the field of multimedia forensics, identifying the source of voice evidence is critically important for judicial applications and information security. Voice recordings often serve as key evidence in legal investigations, yet

收稿日期: 2025-11-12;**修回日期:** 2026-02-11

基金项目: 教育部人文社会科学研究规划基金项目(24YJA870011); 国家自然科学基金项目(62302146);中央高校基本科研业务费专项资金资助(PA2025IISL0104,PA2025GDSK0078)

Supported by: MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant 24YJA870011; National Natural Science Foundation of China under Grant 62302146); the Fundamental Research Funds for the Central Universities of China (Grant No. PA2025IISL0104 and PA2025GDSK0078

their evidential value is limited by the capabilities of current analytical techniques. Most existing approaches focus on identifying the mobile phone model but lack the precision to differentiate between individual devices of the same model. This limitation significantly reduces the utility of voice evidence, as it cannot be uniquely linked to a specific device. As a result, such evidence is often treated as supplementary in court proceedings rather than being considered direct and legally valid proof. Fine-grained device identification inability hinders digital forensics effectiveness and highlights the need for advanced methodologies. **Methods** To address this challenge, this paper proposes a novel voice source identification (VSI) model based on multi-scale feature fusion and dynamic enhancement. The model is designed to capture and amplify the subtle hardware-specific fingerprints embedded in speech signals. These fingerprints arise from variations in microphone and circuit components, making them unique to each device. The proposed framework consists of several interconnected modules. First, a residual-attention collaborative network is employed to enhance feature representation. This network combines the strengths of residual learning and attention mechanisms, enabling the model to focus on device-related characteristics while suppressing irrelevant speech content and environmental noise. The residual connections facilitate training stability, while the attention mechanism dynamically weights important features, improving discriminability. Second, a dual-path feature extraction structure is implemented. The overall feature extraction module uses an expression-enhanced Time Delay Neural Network (ETDNN) to capture broad, context-rich device characteristics. This module processes long-term temporal dependencies, which are essential for identifying model-level attributes. Simultaneously, a local feature extraction module based on a multi-level residual SE-Res2Net architecture is used to capture fine-grained, high-frequency details. This component excels at identifying minute differences between individual devices by leveraging hierarchical residual connections and channel-wise attention, enabling multi-scale feature learning within a unified framework. Third, a feature enhancement module incorporating feature recalibration and dynamic global filtering is applied. This module refines the fused multi-scale features by emphasizing task-relevant information and filtering out noise. Feature recalibration adjusts channel-wise feature responses adaptively, while dynamic global filtering performs content-aware smoothing and enhancement, further strengthening the device-specific representations. Finally, a fine-grained classification model is constructed to achieve cross-level device identification. This classifier is trained to distinguish devices at multiple hierarchical levels, from brand and model down to the individual unit, improving both accuracy and generalization. The entire architecture is optimized end-to-end, ensuring seamless integration of all components. **Results** To comprehensively evaluate the effectiveness of the proposed VSI model, a large-scale voice dataset comprising 14 mobile phone brands and 121 different individual devices was constructed. This dataset is designed to cover a wide range of scenarios and conditions, thereby providing a rigorous testbed for assessing the model's performance. The VSI model was evaluated using three key metrics: equal error rate (EER), accuracy (ACC), and minimum detection cost function (minDCF). The results demonstrate the superior performance of the VSI model compared to existing methods. Specifically, the VSI model achieved an EER of 7.50%, an ACC of 89.97%, and a minDCF of 0.38. These results are significantly better than those obtained by four other state-of-the-art methods reported in the literature. Compared to these methods, the EER was reduced by 4.75%, 4.71%, 5.44% and 3.46%, respectively. The ACC was increased by 3.98%, 2.20%, 4.90%, and 2.52% respectively. The minDCF was decreased by 0.04, 0.04, 0.30, and 0.03, respectively. These improvements highlight the effectiveness of the VSI model in accurately identifying individual devices from voice signals. Moreover, the VSI model demonstrated robustness under varying conditions of voice duration, sampling rate, and encoding format. This robustness is crucial for practical applications, as voice signals in real-world scenarios often exhibit variations in these parameters. The ability of the VSI model to maintain high performance across different conditions indicates its potential for widespread adoption in forensic and security-related applications. **Conclusion** The findings demonstrate that the proposed VSI model significantly advances the field of voice source identification by enabling precise, individual-level device recognition. By effectively extracting and enhancing hardware-specific fingerprints from speech signals, the model provides a reliable means of linking voice recordings to specific mobile devices. This capability enhances the evidentiary value of voice data in judicial contexts, allowing it to serve as direct and valid evidence. Furthermore, the model's robustness to real-world variations ensures its practical applicability in diverse scenarios. These contributions offer strong technical support for judicial forensics, intelligent terminal authentication, and broader domains of multimedia security. Future work will focus on

extending the model to handle more complex audio manipulations and integrating it with complementary forensic techniques.

Key words: Multimedia Forensics; Voice Source Identification; Multi-scale Feature Fusion; Dynamic Enhancement; Individual Mobile Identification

0 引 言

语音来源识别作为语音真实性验证的重要手段，是多媒体取证中的一个重要分支（许裕雄等，2024），其核心任务是从语音信号中提取能够表征录制设备的特征指纹，通过分析、比对实现对语音数据录制设备的来源识别。已有基于语音的设备来源识别从利用传统的手工特征提取方法到基于深度学习的技术，均是将手机录音文件映射为特征向量，在开集或者闭集模式下实现具体手机型号的识别。

Hanilci 等人提取梅尔频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）作为手机设备类型识别特征，并利用支持向量机（Support Vector Machine，SVM）作为分类器进行模型训练和测试（Hanilci 等，2011）。Luo 等人利用不同生产商在语音采集管道上的不同带来的微小差异，提出一种新的带能量描述符特征，使用SVM 进行手机设备类型识别（Luo 等，2018）。Zou 等人采用高斯混合模型与通用背景模型，设计出一种基于梅尔频率倒谱系数和功率归一化倒谱系数的设备类型识别方法（Zou 等，2014）。裴安山等人将本底噪声作为手机设备型号的指纹用于识别，并通过使用自适应端点检测算法得到语音的静音段，然后将静音段中对数域的Mel滤波器组系数进行降维，作为分类特征（裴安山等，2017）。Qin 等人提出了一种基于常数 Q 变换域的语音特征，并配合使用卷积神经网络（Convolutional Neural Networks，CNN）进行手机设备类型识别模型的训练（Qin 等，2018）。苏兆品等人利用时间卷积网络（Temporal Convolutional Networks，TCN）对滤波器组特征（Filter Bank，Fbank）进行了一系列的处理得到高维特征，再用线性判别分析对特征进行降维处理，最后使用SVM 进行训练，实现了闭集上手机型号的识别（苏兆品等，2021）。岳峰等人基于 SE-Res2Block 和 Fastformer 对 Fbank 特征进行处理融合，设计基于注意力机制的确认网络在开集上实现了对未知手机型号的识别（岳峰等，2025）。

需要注意的是，已有研究均面向设备类型的识别，即通过语音信号识别出来源手机设备类型，但无法识别出其来自于该类型下的具体个体，导致语音证据多被视为辅助线索而非直接有效证据。为此，面向设备个体的语音来源识别已成为多媒体取证领域的热点问题之一。此问题要求通过分析语音信号的设备相关特征参数，实现语音设备个体的鉴别与认证，从而为司法取证提供更加直接的证据。

考虑到发展十分迅速的自动说话人验证（Automatic speaker verification system，ASV）系统（Joshi S 等，2024），其通过深度学习技术学习到不同说话人音色的细微差距来实现说话人识别，代表模型包括ECAPA-TDNN（Desplanques 等，2020）、DS-TDNN（Li 等，2024）、CAM++（Wang 等，2023）和 SERes2BiLSTM（Weng 等，2025）等。这与语音设备个体识别有一定的相似性。但是，相对于说话人验证问题中利用的说话人音色信息（王善敏等，2025），设备相关信息通常是一种弱信息，直接采用说话人识别模型很难达到识别效果。因此，如何有效提取语音中的设备个体信息，抑制设备个体无关信息就成为了语音设备个体识别的难点和重点。

基于此，本文提出一种面向设备个体的语音来源识别方法，实现从型号到个体的跨层级设备识别。主要工作如下：

1）. 构建了面向设备个体的语音来源识别模型，将任务聚焦于实现对未知设备个体的来源识别。

2）. 提出了一种多尺度特征融合与动态增强的语音来源识别方法（Multi-scale feature fusion and dynamic enhancement for voice source identification，VSI），首先通过基于残差块的特征预处理模块有效提取语音信号中不同层次的设备相关特征；然后利用基于表达增强 TDNN 的整体特征提取模块和基于多级残差 SE-Res2Net 的局部特征提取模块，捕捉丰富的设备特征信息及个体之间的细微差距；最后通过特征冲校准与动态全局滤波的模块，滤除与任务无关的信息，增强与设备个体相关的特征表示。

3). 为了验证所提方法的有效性，构建包含14个手机品牌、121个不同设备个体的语音数据集；利用等错误率(Equal Error Rate, EER)、准确率(Accuracy, ACC)和最小检测代价函数(minimum Detection Cost Function, minDCF)指标，从消融分析、对比实验和鲁棒性测试等多个方面进行全面评估。

1 面向设备个体的语音来源识别问题

如前所述，面向设备个体的语音来源识别是通过分析语音信号的设备相关特征参数，构建模型或模板实现设备个体的鉴别与认证。该任务聚焦于实现对未知设备个体的语音来源识别，可通过判断两条语音是否来自同一设备个体进行实现。求解过程可以描述为：

1)从原始语音信号$\boldsymbol{x}_t$中提取时频特征$\boldsymbol{f}\in\mathbf{R}^{\mathrm{T}\times\mathrm{d}}$，T表示每一帧语音的特征维度，d表示语音的帧数，如公式(1)。

$$\boldsymbol{f}=\mathrm{FEx}(\boldsymbol{x}_t). \tag{1}$$

2)利用数据集，训练深度神经网络模型Ψ，并利用训练好的模型Ψ对特征$\boldsymbol{f}$进行编码，得到一个固定维度的嵌入向量$\boldsymbol{e}$，如公式(2)，将$\boldsymbol{e}$作为设备个体特征。

$$\boldsymbol{e}=\Psi(\boldsymbol{f}),\boldsymbol{e}\in\boldsymbol{R}^{\mathrm{D}}. \tag{2}$$

3)对于待测语音$\boldsymbol{x}_{test}$，已知语音为$\boldsymbol{x}_{ref}$，分别经过上述流程得到嵌入向量$\boldsymbol{e}_{test}$和$\boldsymbol{e}_{ref}$，如公式(3)和公式(4)。

$$\boldsymbol{e}_{ref}=\Psi\left(\boldsymbol{f}_{ref}\right), \tag{3}$$

$$\boldsymbol{e}_{test}=\Psi\left(\boldsymbol{f}_{test}\right). \tag{4}$$

4)利用余弦相似度计算$\boldsymbol{e}_{test}$和$\boldsymbol{e}_{ref}$的相似性，如公式(5)。

$$\mathrm{S}(\boldsymbol{e}_{ref},\boldsymbol{e}_{test})=\frac{\boldsymbol{e}_{ref}\cdot\boldsymbol{e}_{test}}{\left\|\boldsymbol{e}_{ref}\right\|\left\|\boldsymbol{e}_{test}\right\|}. \tag{5}$$

5)如果$\mathrm{S}(\boldsymbol{e}_{ref},\boldsymbol{e}_{test})\geqslant th$，则待测语音$\boldsymbol{x}_{test}$与参考语音$\boldsymbol{x}_{ref}$来自同一个设备个体；否则，他们为不同设备语音。其中，th为预设阈值。

2 VSI方法

VSI方法旨在实现对未知手机个体设备的准确识别，整体框架如图1所示。

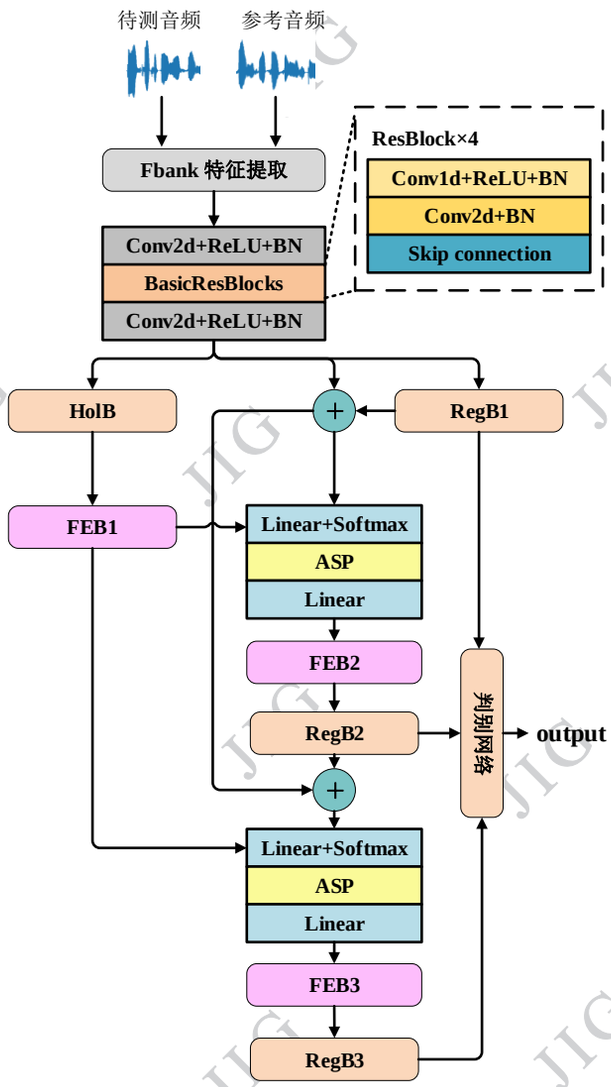


图1 VSI方法总体结构图

Fig. 1 Overall structure of the VSI method

首先，利用Librosa工具库对原始语音信号进行Fbank特征提取；其次，将提取到的Fbank特征输入特征预处理模块(Feature Preprocessing Block, FPB)，以增强设备相关特征的表达能力；接着，分别通过局部特征提取模块(RegionalBlocks, RegBs)、整体特征提取模块(HolisticBlock, HolB)和特征增强模块(Feature-enhanced Block, FEB)对特征进行有效融合，提升模型的整体表征能力；最后，通过基于注意力机制的判别网络将融合后的特征映射为固定维度的向量表示，并采用余弦距离度量两段语音是否来源于同一手机个体。

2.1 特征预处理模块

由于设备个体的特征信息可能分布于语音信号的各个维度，为了增强模型对设备相关特征的捕捉

能力，放大不同个体之间的细微差异性，本文设计了特征预处理模块FPB，主要由多个残差块（Residual Blocks）组成，如表1所示。该模块通过融合多种残差块结构，能够有效提取语音信号中不同层次的设备相关特征。

表1　特征预处理模块参数

Table 1　Parameters of FPB

Layer	Parameter
Conv2d	(1,32,3,1,1)
Resblock1	(32,32,3,(2,1),1)
Resblock2	(32,32,3,1,1)
Resblock3	(32,32,3,(2,1),1)
Resblock4	(32,32,3,1,1)
Conv2d	(32,32,3,(2,1),1)

注：表中各层参数分别代表（输入通道数，输出通道数，卷积核尺寸，步长，填充）。

每个残差块包含两个卷积层、两个批量归一化层（Batch Normalization，BN）以及一个跳跃连接（Skip Connection），通过非线性变换和特征融合提升设备个体特征的表达能力，同时缓解深层网络的梯度消失问题。模块输入首先通过reshape进行调整为频率和时间的二维特征图，以适应二维卷积操作。接着通过一个3×3卷积层和批量归一化层进行初步特征提取，并经过ReLU激活函数引入非线性。随后特征依次通过两个残差层，每个残差层包含两个残差块。每一个残差层中的残差块采用步长2降低时间维度，进一步压缩语音特征表示。这种分层设计能够逐步提取不同尺度的语音设备特征信息，从局部细节到全局上下文。然后，通过一个3×3卷积层和批量归一化层进一步细化设备特征表示，最后通过reshape操作将二维特征转换为一维特征序列，形成融合了通道和时间维度的语音设备特征表示，不仅保留了原始输入的关键信息，还通过多层次的特征提取引入了丰富的设备特性信息，有助于不同层次特征之间的互补，进一步提升模型的识别性能和泛化能力。

2.2　整体特征提取模块

语音信号中的设备个体信息贯穿于整个录音过程，捕捉语音样本的全局信息至关重要。整体特征提取模块HolB由3个表达增强TDNN层（Expression-enhanced Time Delay Neural Network，ETDNN）组成，每层输入道数逐步增加，通过密集连接将前一层的输出与当前层的输出拼接在一起，作为下一层的输入，逐步增强有关手机个体特征的表达能力，同时保留不同层次的全局特征信息。每一层ETDNN的网络结构如图2所示。

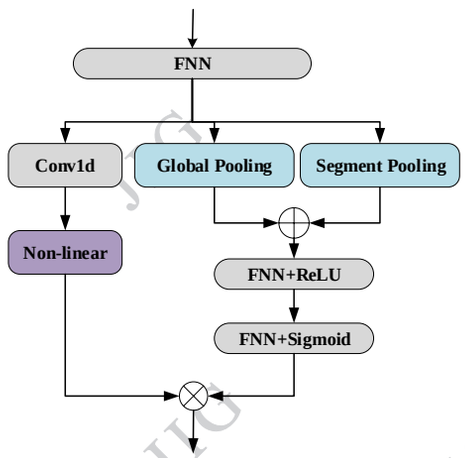


图2　表达增强TDNN网络结构

Fig. 2　Structure of the expression-enhanced TDNN

可以看出，ETDNN能够捕捉到设备在长时间跨度内的特征稳定性与一致性，增强与设备相关的特征通道，抑制无关特征通道。因此，整体特征提取模块可专注于全局特征的提取和差异增强，能够从宏观层面捕捉语音设备特征的全局差异性，为后续特征融合和设备个体识别提供坚实的全局特征基础。

2.3　局部特征提取模块

为了有效捕捉语音信号中不同尺度的设备相关特征，并显著扩大设备特征之间的差异性，本文设计了基于多级残差的局部特征提取模块RegBs，多级残差模块（RegionalBlock，RegB）如图3所示，将每一段的输出都进行叠加作为下一段的输入，如公式（6）。

$$\boldsymbol{y}_i = \mathrm{Conv}\left(\boldsymbol{x}_i + \sum_{j=1}^{i-1}\boldsymbol{y}_j\right), i = 2,3,\cdots,8. \tag{6}$$

这样可以保留更丰富的不同尺度的局部特征信息，充分挖掘设备个体之间的差异。优化后的残差网络每一层卷积的感受野都不同。压缩扩张网络（Squeeze-Excitation，SE）模块又能对特征向量进行压缩和扩展，从而增强有关设备信息的特征，放大设备个体之间的差异。由此，多级残差块可专注于捕

捉不同尺度下的局部特征差异性。

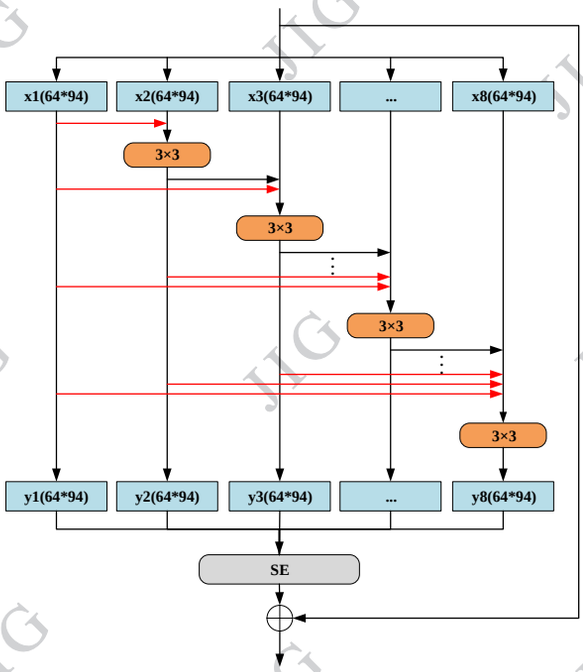


图 3　RegB 网络结构

Fig. 3　Structure of the RegB

局部特征提取模块由三层多级残差模块组成，具体参数如表 2 所示，感受野逐层递增。

表 2　局部特征提取模块参数

Table 2　Parameters of the RegBs

Layer	Parameters
RegB1	(512,512,3,2,8)
RegB2	(512,512,3,3,8)
RegB3	(512,512,3,4,8)
SE	(512,512)

注：表中 RegB 参数分别代表(输入通道数，输出通道数，卷积核尺寸，空洞率，分支数)，SE 参数分别代表(输入通道数，输出通道数)。

2.4　特征增强模块

由于初始特征、局部特征和全局特征具有不同的分布和特性，直接相加会存在信息融合不充分的问题，另外融合后特征中也会存在与设备信息不相关的其他冗余信息。因此，VSI 设计了基于 ScConv 和动态滤波器的特征增强模块 FEB，具体可以描述为：

1)对于输入的特征 $\boldsymbol{x}$，通过 reshape 进行维度重塑，得到与 ScConv 模块(Li 等，2023)相匹配的特征 $\boldsymbol{x}_{re} \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，B 表示批次大小，C 表示通道维度，H 表示高度维度，W 表示宽度维度。

2)利用组归一化(Group Normalization，GN)层中的缩放系数 $\boldsymbol{\beta}=[\beta_1,\beta_2\ldots\beta_{\mathrm{C}}]$ 计算权重 $\boldsymbol{W}_\beta$，用来量化特征 $\boldsymbol{x}_{re}$，如公式(7)；通过 Sigmoid 函数和门控函数 Gate 输出权重分量 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$，如公式(8)；将输入特征 $\boldsymbol{x}_{re}$ 解耦为富含信息的特征 $\boldsymbol{x}_1^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ 和信息匮乏的特征 $\boldsymbol{x}_2^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，如公式(9)，⊗表示逐元素相乘。

$$\boldsymbol{W}_\beta=\{w_i\}=\frac{\beta_i}{\sum_{j=1}^{\mathrm{C}}\beta_j},i=1,2,\cdots,\mathrm{C}, \tag{7}$$

$$[\boldsymbol{W}_1,\boldsymbol{W}_2]=\mathrm{Gate}\left(\mathrm{Sigmoid}\left(\boldsymbol{W}_\beta\left(\mathrm{GN}\left(\boldsymbol{x}_{re}\right)\right)\right)\right), \tag{8}$$

$$\begin{cases}\boldsymbol{x}_1^w=\boldsymbol{W}_1\otimes\boldsymbol{x}_{re}\\ \boldsymbol{x}_2^w=\boldsymbol{W}_2\otimes\boldsymbol{x}_{re}\end{cases}. \tag{9}$$

3)将 $\boldsymbol{x}_1^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ 和 $\boldsymbol{x}_2^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ 分别在通道维度均分为 $\boldsymbol{x}_{11}^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$、$\boldsymbol{x}_{12}^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$ 和 $\boldsymbol{x}_{21}^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$、$\boldsymbol{x}_{22}^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$，采用交叉重构操作融合不同信息密度的特征，增强信息交互，生成空间细化特征 $\boldsymbol{x}^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，如公式(10)。

$$\begin{cases}\boldsymbol{x}_{11}^w\oplus\boldsymbol{x}_{22}^w=\boldsymbol{x}^{w1}\\ \boldsymbol{x}_{21}^w\oplus\boldsymbol{x}_{12}^w=\boldsymbol{x}^{w2}\\ \boldsymbol{x}^{w1}\cup\boldsymbol{x}^{w2}=\boldsymbol{x}^w\end{cases}. \tag{10}$$

其中，⊕表示逐元素相加，∪表示拼接。

4)采用拆分-变换-融合策略，对空间细化特征 $\boldsymbol{x}^w$ 进行通道维度的冗余削减，拆分为 $\boldsymbol{x}_u^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$ 和 $\boldsymbol{x}_d^w \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}/2\times\mathrm{H}\times\mathrm{W}}$。对 $\boldsymbol{x}_u^w$ 同时使用分组卷积(Group-Wise Convolution，GWC)(Guo 等，2019)和点卷积(Point-Wise Convolution，PWC)(Hong 等，2024)变换得到 $\boldsymbol{x}_u \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，如公式(11)；对 $\boldsymbol{x}_d^w$ 只使用点卷积并进行拼接得到 $\boldsymbol{x}_d \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，如公式(12)；并对 $\boldsymbol{x}_u$ 和 $\boldsymbol{x}_d$ 进行拼接，再使用自适应平均池化(Okabe 等，2018)和 Softmax 获得注意力加权值，再进行加权得到滤除冗余后的特征 $\boldsymbol{x}_{ScConv} \in \mathbf{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$，如公式(13)和公式(14)。

$$\boldsymbol{x}_u=\mathrm{GWC}\left(\boldsymbol{x}_u^w\right)+\mathrm{PWC}\left(\boldsymbol{x}_u^w\right), \tag{11}$$

$$\boldsymbol{x}_d=\mathrm{Concat}\left(\mathrm{PWC}\left(\boldsymbol{x}_d^w\right)\right), \tag{12}$$

$$\boldsymbol{x}_{ud}=\mathrm{Concat}\left(x_u,x_d\right), \tag{13}$$

$$\boldsymbol{x}_{ScConv}=\mathrm{Soft\,max}\left(\boldsymbol{x}_{ud}\right)\cdot\boldsymbol{x}_{ud}. \tag{14}$$

5)将 $\boldsymbol{x}_{ScConv}$ 通过 reshape 进行维度重塑，得到特

征$\boldsymbol{x}_{DGF1}\in\mathbf{R}^{B\times C\times N}$，N表示序列长度；并利用动态稀疏全局滤波器SparseDGF（Wu等，2018），从不同的角度捕捉不同设备特征之间的全局关系，得到特征$\boldsymbol{x}_{DGF2}$，如公式（15）；最后，卷积核残差连接获取最终输出$\boldsymbol{x}_{DGF}\in\mathbf{R}^{B\times C\times N}$，如公式（16）。

$$\boldsymbol{x}_{DGF2}=\mathrm{SparseDGF}\left(\boldsymbol{x}_{DGF1}\right),\tag{15}$$

$$\boldsymbol{x}_{DGF}=\mathrm{BN}\left(\mathrm{ReLU}\left(\mathrm{Conv}\left(\boldsymbol{x}_{DGF2}\right)\right)\right)+\boldsymbol{x}_{DGF1}.\tag{16}$$

2.5 判别网络

如图1，将局部特征和整体特征融合后的特征输入判别网络，用以判断两段语音是否来源于同一手机个体。判别网络的整体结构如图4所示。

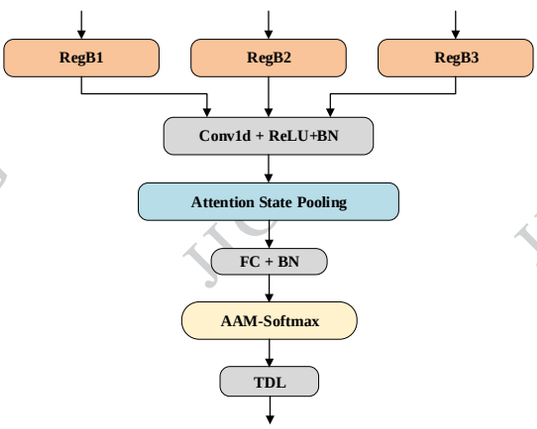


图4 判别网络的模块结构

Fig. 4 Structure of the discrimination network

具体来说，首先在时间维度上对融合特征进行拼接，得到不同尺度的特征信息；随后通过注意力池化层对特征表示进行降维处理，并通过全连接层将特征映射为192维的向量表示；然后使用AAM-Softmax损失函数（Deng等，2019）（Xu等，2019）优化模型的分类能力；最后在判别层（Type Discrimination Layer，TDL）采用余弦距离度量两段语音的特征相似性，从而判断其是否来源于同一手机个体。

3 实验结果与分析

为了验证本文所提VSI模型的有效性，首先构建了包含多个手机品牌和多个个体设备的数据集。其次，考虑到本文任务是通过输入两条设备语音判断是否为同一个语音设备个体，而说话人验证也是通过输入两条说话人语音判断是否为同一个说话人，两个任务具有高度相关性，因此选择当前效果出色的ECAPA-TDNN（Desplanques等，2020）、DS-TDNN（Li等，2024）、CAM++（Wang等，2023）和SERes2BiLSTM（Weng等，2025）作为对比模型，并从有效性和鲁棒性等多个方面将其与本文提出模型进行比较。

3.1 数据集与实验设置

本文在文献（苏兆品等，2021）基础上，构建了一个包含14个手机品牌、共计121部手机个体的语音数据集，其中每个手机型号不少于两个个体。数据集中的录音内容涵盖了静音、日常对话、背景噪声等多种场景，录音环境包括教室、寝室、高铁、实验室、田径场等多种真实场景。每个手机个体均包含不少于300条录音片段，每条录音时长为3秒，语音采样率为16000Hz。在数据集划分方面，将数据集随机划分为训练集和测试集，其中训练集包含106个手机个体，测试集包含15个手机个体，训练集与测试集中的手机个体互不重叠，以确保实验的开集识别设置。

所有模型代码均基于Python实现，在配备RTX 2070S显卡、32GB内存、Windows 10的计算机上进行。模型训练采用Adam优化器，并将margin和scale分别设置为0.2和30。选用EER、ACC和minDCF作为模型性能的评价标准。

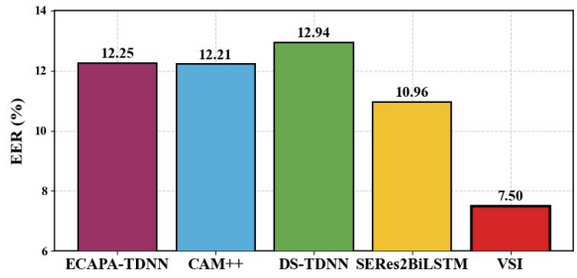
3.2 对比实验分析

图5为所提VSI方法与ECAPA-TDNN（Desplanques等，2020）、DS-TDNN（Li等，2024）、CAM++（Wang等，2023）和SERes2BiLSTM（Weng等，2025）的对比结果，所有模型均在本文数据集上进行训练和测试。

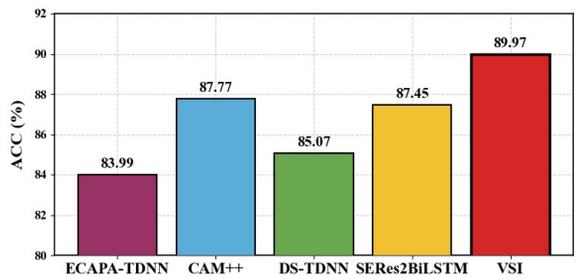
与已有方法相比，VSI方法的EER降低了4.75%、4.71%、5.44%和3.46%，ACC分别提升了3.98%、2.20%、4.90%和2.52%，minDCF分别下降了0.04、0.04、0.30和0.03，因此本文VSI方法在处理面向设备个体的语音来源识别任务时更具优势。这是因为VSI可以聚焦到不同手机个体的细微特征区别，增强不同手机个体的特征表示，抑制无关信息，因此可以提取到更具表现力的手机个体特征表示。

3.3 模块有效性分析

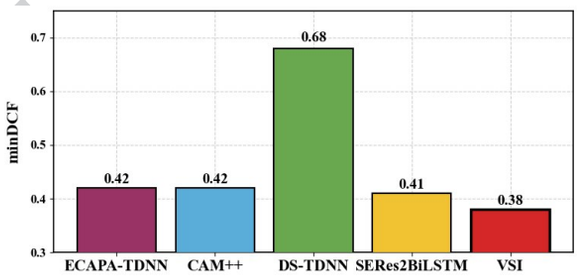
在VSI模型中，FPB、FEB、RegBs和HolB为核心模块。本小节将依次单独去除FPB、FEB、RegBs和HolB来验证每个模块的有效性，实验结果如图6所



（a） 等错误率



（b） 准确率

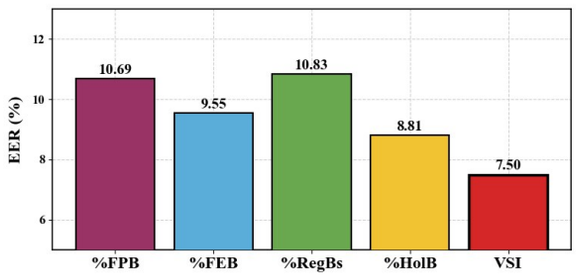


（c） 最小检测代价函数

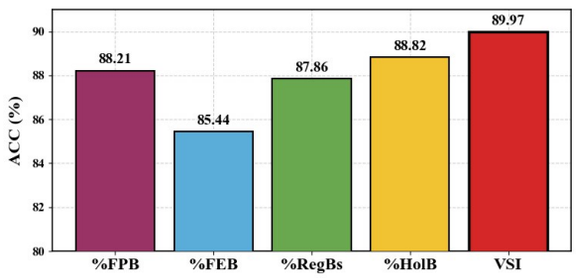
((a)EER;(b) ACC;(c)minDCF)

图 5 五种模型的对比实验结果

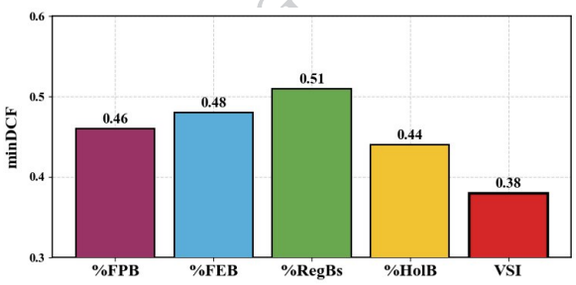
Fig. 5 Comparative results of five models



（a） 等错误率



（b） 准确率



（c） 最小检测代价函数

((a)EER;(b) ACC;(c)minDCF)

图 6 VSI模块有效性实验结果

Fig. 6 Validation results of the VSI blocks

示，其中%*表示去除*模块的模型。

由图 6 可以看出，相比于完整 VSI 模型，去除 FPB、FEB、RegBs 和 HolB 后，EER 分别提高了 3.19%、2.05%、3.33% 和 1.31%，ACC 分别降低了 1.76%、4.53%、2.11% 和 1.15%，minDCF 分别提高了 0.08、0.10、0.13 和 0.06。因此 FPB、FEB、RegBs 和 HolB 模块对 VSI 模型非常关键。FPB 能够从不同尺度对原始特征进行处理，获得更多设备特征信息，FEB 能够剔除特征中的冗余信息，RegBs 能够捕捉不同尺度下的局部特征的差异性，从而放大不同个体的特征差异，HolB 能够获取特征中任意两个位置的依赖情况，增强全局信息感知能力，从宏观层面捕捉设备特征的全局差异性。

3.4 消融分析

为了进一步验证各模块的性能，在已构建的数据集上进行消融实验，实验结果如表 3 所示，其中 FPB 表示只使用了特征预处理模块，FPB+HolB 表示在使用了特征预处理模块和整体特征提取模块。FPB+HolB+RegBs 表示使用了特征预处理模块、整体特征提取模块和局部特征提取模块。FPB+HolB+RegBs+FEB 表示完整的 VSI 方法。

由表 3，随着 HolB、RegBs 和 FEB 关键模块的加入，EER 分别降低至 9.37%、8.36% 和 7.50%，ACC 提升到 87.02%、87.93% 和 89.97%，minDCF 降低到 0.39、0.38 和 0.38。因此 VSI 各个模块对于模型性能均有提升效果。

表 3　消融实验结果

Table 3　Analysis of ablation experiments

网络模型	EER/%	ACC/%	minDCF
FPB	9.41	86.15	0.42
FPB+HolB	9.37	87.02	0.39
FPB+HolB+RegBs	8.36	87.93	**0.38**
FPB+HolB+RegBs+FEB	**7.50**	**89.97**	**0.38**

注：加粗字体为每列最优值。

3.5　不同语音因素的影响分析

为了评估模型的抗攻击能力，设计了在不同语音因素下的攻击实验，其中包括不同时长、不同采样率、不同编码格式和幅值调整。

3.5.1　语音时长的影响分析

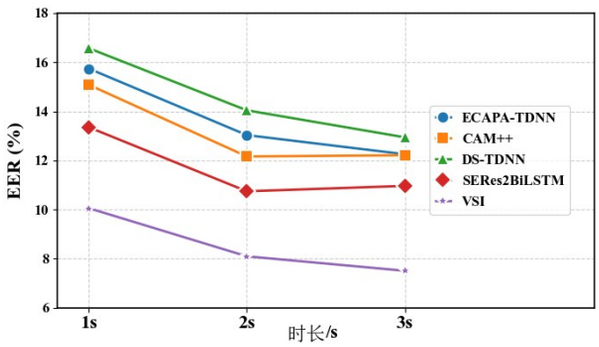
在实际应用过程中，手机录音时长可能小于3s，为了衡量语音时长对VSI方法性能的影响，分别采用1s，2s和3s的样本进行模型测试，实验结果如图7所示。

当测试语音时长减少时，ECAPA-TDNN、CAM++、DS-TDNN和SERes2BiLSTM模型性能均会降低。对于VSI来说，相较于3s时长，当时长为2s和1s时，EER分别提高了0.59%和2.55%，ACC分别降低了2.20%和5.52%，minDCF分别提高了0.05和0.12，这是因为语音时长较短时语音内部包含的设备个体信息就会不足，模型能够提取到的有效信息也就不够充分。在相同时长的测试集语音中，VSI模型均表现出明显优势，并且性能指标均保持相对稳定，说明相较于其他模型，VSI能够从较短的语音中提取更多的设备信息，即VSI能够适应不同时长的语音样本。并且VSI模型的性能指标随语音时长变化的波动幅度最小，在1s超短时极端场景下仍能保持较好性能，说明VSI模型鲁棒性也更好。

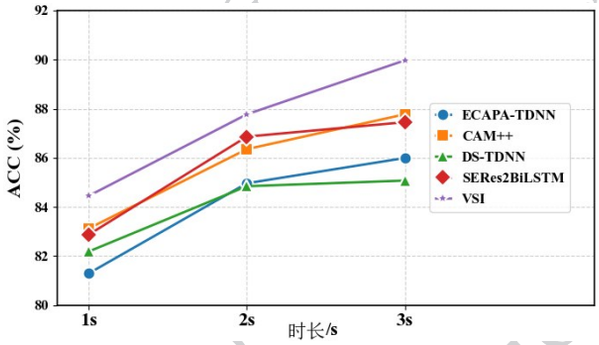
3.5.2　重采样攻击的影响分析

为了评估模型在不同采样率下的性能，分别使用22050Hz、32000Hz和48000Hz三种采样率对测试集中的16000Hz测试语音进行重采样，实验结果如图8所示。

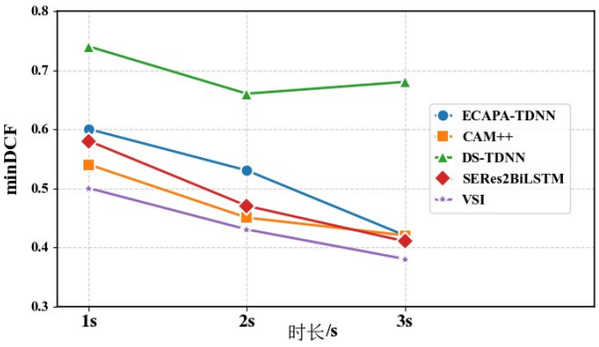
可以看出，当录音的采样率发生变化，会对模型的表现能力有一些影响，在EER指标上，VSI模型在各采样率下的数值远低于其他对比模型，平均EER较最优对比模型降低约30%，在ACC指标上，VSI模型在各种采样率下保持了89.00%~89.97%的高识别准确率，相比最优对比模型提升约1.5%，在minDCF指标上，VSI模型的数值始终维持在0.36~0.38之间，是所有模型最低的区间范围，说明VSI方法受采样率的影响较小，表现更加稳定，鲁棒性更好。



（a）　等错误率



（b）　准确率



（c）　最小检测代价函数

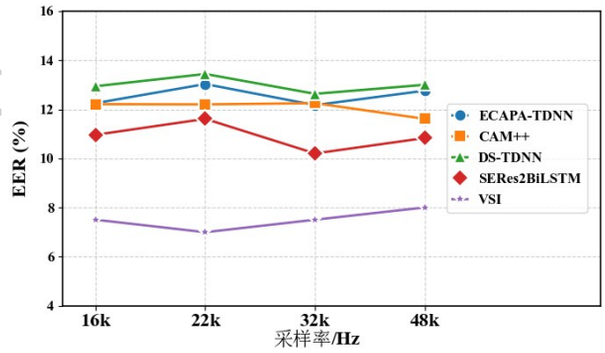
((a)EER；(b) ACC；(c)minDCF)

图 7　不同时长下模型指标对比实验结果

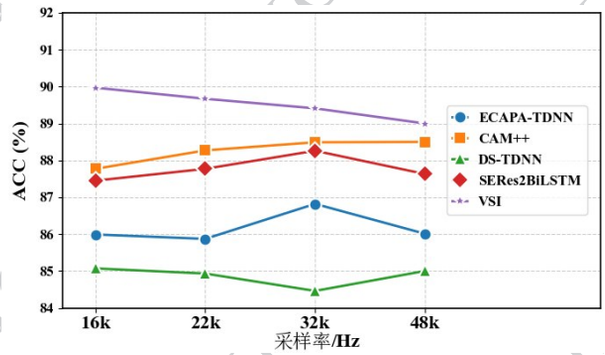
Fig. 7　Experimental results under different durations

3.5.3　语音编码的影响分析

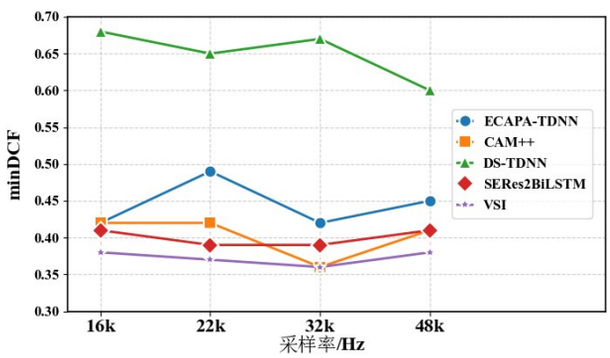
为了评估模型抗语音编码的攻击能力，将原有测试集的MP3编码分别转换为M4A、WAV、FLAC和AAC的编码，以模拟现实网络中的编码攻击。图9展示了不同编码格式下各模型的测试结果。



（a） 等错误率



（b） 准确率

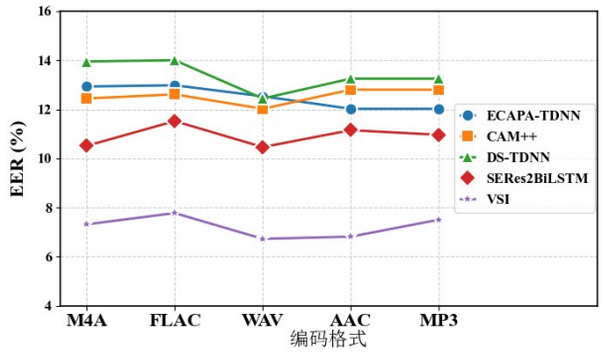


（c） 最小检测代价函数

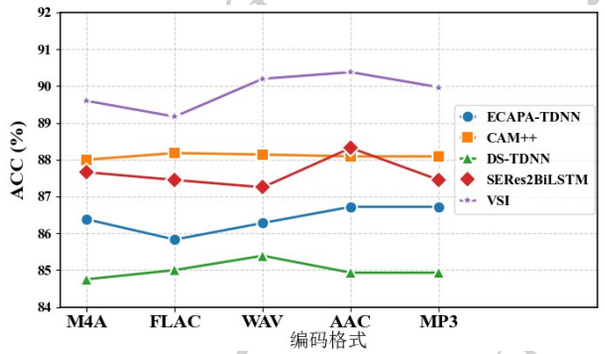
((a)EER;(b) ACC;(c)minDCF)

图 8 不同采样率下模型指标对比实验结果

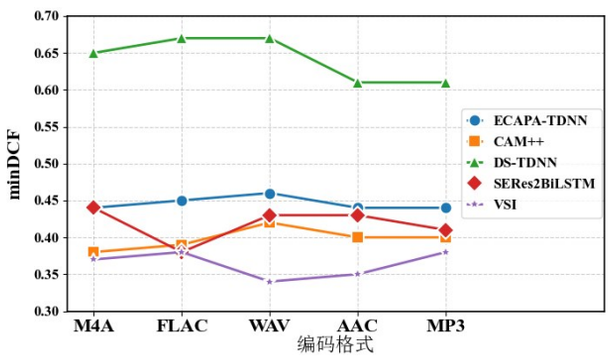
Fig. 8 Experimental results under different sampling rates



（a） 等错误率



（b） 准确率



（c） 最小检测代价函数

((a)EER;(b) ACC;(c)minDCF)

图 9 不同编码格式下模型指标对比实验结果

Fig. 9 Experimental results under different codecs

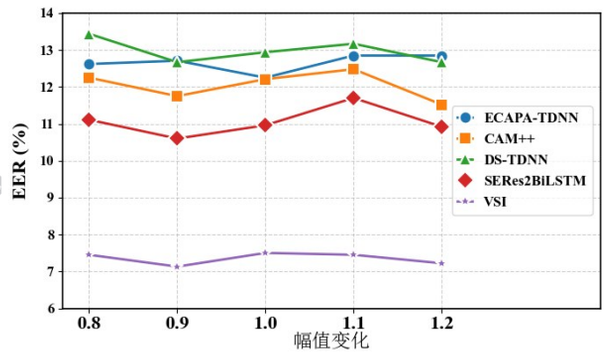
可以看出，当录音的编码格式发生变化，会对模型的表现能力有一些影响，但是 VSI 模型对于录音语音编码格式转变的攻击时，相较于其他的模型，VSI 模型在评价指标上仍均优于其他的模型。例如，在 ACC 指标上，VSI 模型在各种编码格式下保持了 89.17%~90.38% 的高识别准确率，是所有模型中唯一 ACC 稳定在 89% 以上的模型，说明 VSI 模型在应对语音编码攻击时依旧表现较好。

3.5.4 幅值量化攻击的影响分析

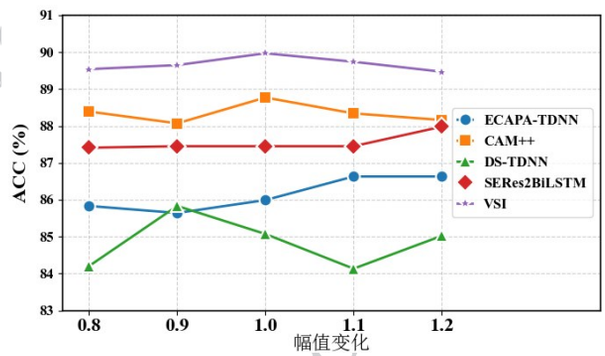
为了评估模型的抗幅值量化攻击能力，分别按照 0.8、0.9、1.1、1.2 比例调整测试语音的幅值，模拟实际环境中可能出现的幅值量化攻击，测试结果如图 10 所示。

当语音幅值被等比例压缩至原值的 80% 或扩展至 120% 时，音频的响度范围和动态对比度发生显著改变，VSI 模型的性能依旧保持稳定，各个指标的上下浮动变化非常小，EER 指标上始终维持在

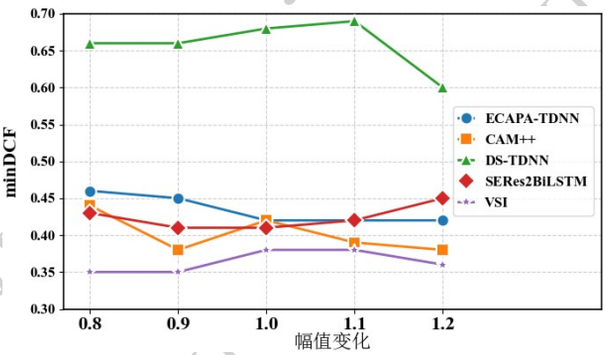
7.15%~7.50%的区间范围，ACC指标始终维持在89.47%~89.97%的区间范围，minDCF指标始终维持在0.35~0.38的区间范围，均是所有模型最低的区间范围，展现出了对于语音能量扰动攻击的鲁棒性。



（a） 等错误率



（b） 准确率



（c） 最小检测代价函数

((a)EER;(b) ACC;(c)minDCF)

图 10 不同幅值下模型指标对比实验结果

Fig. 10 Experimental results of model under different amplitude variations

4 结 论

为解决现有方法仅能识别设备型号而无法区分个体设备的关键问题，本文围绕面向设备个体的语音来源识别任务开展了深入研究，构建了面向设备个体的语音来源识别模型VSI，设计了一种多尺度特征融合与动态增强的识别框架。在包含14个手机品牌、121个手机个体数据集上的实验结果表明，本文方法在等错误率、准确率和最小检测代价函数指标上均显著优于现有方法，而且在不同时长、不同采样率、改变编码格式和幅值下均表现较好，验证了本文所提方法的有效性和优越性。这是因为，相较于现有方法，VSI通过多尺度特征融合策略有效整合了局部细节与全局上下文信息，通过动态增强机制自适应地强化设备相关特征、抑制语音内容和环境噪声干扰，从而能够聚焦于不同手机个体的细微硬件特征差异。这使得模型在极短语音和多种攻击场景下仍能保持稳定的识别性能，显示出较强的泛化能力，为提升语音证据在司法实践中的有效性提供技术支持，同时也为后续相关研究提供了借鉴思路。

由于在实际取证应用中可能会遇到各种各样的极端复杂情景，下一步研究将重点关注如何进一步提高模型在复杂社交场景下的抗攻击能力和鲁棒性。

参考文献(References)

Xu Y X, Li B, Tan S Q and Huang J W. 2024. Research progress on speech deepfake and its detection techniques. Journal of Image and Graphics, 29(08):2236-2268 (许裕雄, 李斌, 谭舜泉, 黄继武. 2024. 语音深度伪造及其检测技术研究进展. 中国图象图形学报, 29(08):2236-2268)[DOI:10.11834/jig.230476]

Hanilci C, Ertas F, Ertas T and Eskidere. 2011. Recognition of brand and models of cell-phones from recorded speech signals. IEEE Transactions on Information Forensics and Security, 625-634 [DOI: 10.1109/TIFS.2011.2178403]

Luo D, Korus P and Huang J. 2018. Band energy difference for source attribution in voice forensics. IEEE Transactions on Information Forensics and Security, 13(9): 2179-2189 [DOI: 10.1109/TIFS.2018.2812185].

ZOU L, YANG J and HUANG T. 2014. Automatic cell phone recognition from speech recordings. Proceedings of the 5th IEEE China

Summit and International Conference on Signal and Information Processing[C]. Xi'an, China: IEEE: 621—625 [DOI: 10.1109/ChinaSIP.2014.6889318].

Pei A S, Wang R D and Yan D Q. 2017. Cellphone origin identification based on spectral features of device self-noise.Telecommunications Science,33(1): 85-94 (裴安山,王让定,严迪群. 2017. 基于设备本底噪声频谱特征的手机来源识别. 电信科学,33(1): 85-94) [DOI: 10.11959/j.issn.1000-0801.2017019].

Pei A S, Wang R D and Yan D Q. 2017. Source cell-phone identification from recorded speech using non-speech segments. Telecommunications Science,33(1): 103-111 (裴安山,王让定,严迪群. 2017. 基于语音静音段特征的手机来源识别. 电信科学,33(1): 103-111) [DOI: 10.11959/j.issn.1000-0801).2017123].

Qin T, Wang R, Yan D and Lin L. 2018. Source cell-phone identification in the presence of additive noise from CQT domain. Information, 9(8): 205 [DOI: 10.3390/info9080205].

Su Z P, Wu Z Q, Yue F, Wu Q F and Zhang G F. 2021 Source Cell-Phone Identification Under Background Noise Based on Low-Dimensional Deep Features. Chinese Journal of Electronics, 49(4): 637 (苏兆品, 吴张倩, 岳峰, 武钦芳,张国富. 2021. 自然环境背景噪声下基于低维深度特征的手机来源识别. 电子学报, 49(4): 637 [DOI: 10.12263 /DZXB.20200658].

Yue F, Peng Y, Su Z P, Zhang G F, Lian C S, Yang B and Fang Z. 2025. Openset Source Cell-Phone Identification based on Feature Interaction and Representation Enhancement. Journal of Computer Applications, 1-9 (岳峰,彭洋,苏兆品,张国富,廉晨思,杨波,方振. 2025. 基于特征交互和表示增强的语音手机来源开集识别. 计算机应用, 1-9) [DOI: 10.11772/j. issn. 1001-9081. yyyymmnnnn].

Joshi S and Dua M. 2024.Noise robust automatic speaker verification systems: review and analysis. Telecommunication Systems ,87: 845-886 [DOI: 10.1007/s11235-024-01212-8].

Desplanques B, Thienpondt J and Demuynck K. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. INTERSPEECH 2020, 3830-3834 [DOI: 10.21437/Interspeech.2020-2650].

Li Y, Gan J, Lin X, Qiu Y, Zhan H and Tian H. 2024. DS-TDNN: Dual-stream time-delay neural network with global-aware filter for speaker verification. IEEE/ACM Transactions on Voice, Speech, and Language Processing. [DOI: 10.1109/talsp.2024.3402072].

Wang H, Zheng S, Chen Y, Cheng L and Chen Q. 2023. CAM++: A fast and efficient network for speaker verification using context-aware masking. INTERSPEECH, 5301-5305 [DOI: 10.21437/Interspeech.2023-1513].

Weng S L, Liu Y and Ji Mao. 2025. Effective Modeling of Critical Contextual Information for TDNN-based Speaker Verification. [DOI: arxiv preprint arxiv:2509.09932].

Wang S M, Liu C G, Chen S Y and Liu Q S. 2025. A survey of multimodal emotion recognition from facial expressions, audios, and language. Journal of Image and Graphics, 30(6):2120-2138 (王善敏, 刘成广, 陈胜宇, 刘青山. 2025. 面向表情、语音和语言的多模态情感识别综述. 中国图象图形学报, 30(6):2120-2138)[] 10.11834/jig.250168]

Qi J and Jiang Y. 2025. Enhancing Res2Net with cross-scale feature association and attention optimization for speaker recognition. Biomedical Signal Processing and Control, 110 (Part B): 108219 [DOI: 10.1016/j.bspc.2025.108219].

Li M, Zheng Y, Li D, Wu Y L, Wang Y X and Fei H J. 2024. MS-SENet: Enhancing Speech Emotion Recognition Through Multi-Scale Feature Fusion with Squeeze-and-Excitation Blocks//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 1-5 [DOI: 10.1109/ICASSP48485.2024.10447232].

Li J Wen Y and He L. 2023. ScConv: Spatial and channel reconstruction convolution for feature redundancy//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 6153-6162 [DOI: 10.1109/CVPR52729.2023.00596].

Wu H., Zheng S., Zhang J and Huang K. 2018. Fast End-to-End Trainable Guided Filter//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 1838-1847 [DOI: 10.1109/cvpr.2018.00197].

Guo X Y, Yang K, Yang W K and Wang X G. 2019. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3273-3282 [DOI: 10.1109/CVPR.2019.00339].

Hong H S and Kim H. 2024. Implementation of Tiled Point-Wise Convolution in MobileNet for Parallel Processing//2024 International Conference on Electronics, Information, and Communication (ICEIC). IEEE: 1-6 [DOI: 10.1109/ICEIC61013.2024.10457207].

Okabe K, Koshinaka T and Shinoda K. 2018. Attentive statistics pooling for deep speaker embedding. [DOI: arxiv preprint arxiv: 1803.10963].

Deng J, Guo J, Xue N and Stefanos Z. 2019. Arcface: Additive angular margin loss for deep face recognition//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 4690-4699 [DOI: 10.1109/CVPR.2019.00482].

Xu X, Wang S, Huang H, Huang H J and Qian Y M. 2019. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE: 1652-1656 [DOI: 10.1109/APSIPA ASC47483.2019.9023039].

作者简介

苏兆品,1983年生,女,副教授,主要研究领域为复杂智能系统、多媒体安全。E-mail: szp@hfut.edu.cn。

方振,2000年生,男,硕士研究生,主要研究领域是多媒体取证。E-mail:2023110474@mail.hfut.edu.cn。

张国富，1979年生，男，教授，主要研究为语音安全。E-mail：zgf@hfut.edu.cn。

王垚飞，1996年生，男，副教授，主要研究领域为多媒体内容安全。E-mail：wyf@hfut.edu.cn。

臧怀娟，1989年生，女，实验师，主要研究领域为多媒体内容安全。E-mail：zanghj@hfut.edu.cn。