中图法分类号:TP181; TP309　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-14

论文引用格式:Liu Shijia, Yan Xiaojin, Ren Haojie, Su Zhaopin. Audio deepfake attribution under low-resource data conditions[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260150. (刘石佳, 闫晓金, 任浩杰, 苏兆品. 少样本数据下的语音深度伪造归因方法[J/OL]. 中国图象图形学报,XXXX:1-14. DOI: 10.11834/jig.260150.)[DOI:10.11834/jig.260150]

少样本数据下的语音深度伪造归因方法

刘石佳[1],闫晓金[1],任浩杰[1],苏兆品[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

摘　要: 目的　语音深度伪造归因是指通过捕捉不同语音合成模型生成语音时留下的独特模型特征,精准识别并确认伪造语音来源的技术,不仅能为人工智能治理提供可解释性证据,也能推动AI语音行业的规范发展,倒逼平台落实内容溯源与监管义务,已成为人工智能安全的研究热点之一。然而,已有方法存在特征提取能力不足、泛化性能有限的问题,难以满足实际应用需求。**方法**　为此,本文提出一种少样本数据下的语音深度伪造归因方法(Low-resource Audio Deepfake Attribution,LADAR),利用多层特征融合的特征提取策略与多原型学习机制,实现少样本数据下的语音深度伪造精准归因。具体来说,首先构建基于多层特征融合的特征提取方法,通过可学习注意力权重聚合预训练Wav2Vec2-BERT 2.0模型的各层隐藏状态,并引入浅层偏置因子,将全局特征动态融合,生成强判别性的模型嵌入表示;其次,设计多原型学习模块,为每类伪造方法生成多个原型向量以丰富类内多样性,提高识别准确率;最后,分别在已知和未知语音伪造场景下验证LADAR方法的归因性能。**结果**　与已有方法相比,针对已知语音伪造方法场景,LADAR方法的准确率分别提升35.25%、26.26%、9.22%和5.65%,F1分数分别提升38.82%、27.42%、7.23%和5.74%;针对未知语音伪造方法场景,LADAR方法的准确率分别提升了15.80%、37.51%、10.73%和10.42%,F1分数分别提升了20.29%、34.17%、11.40%和12.33%;**结论**　对比实验结果表明,LADAR方法可有效解决语音深度伪造归因问题,具有较强的准确性和泛化性,可为司法取证、溯源追踪等实际场景提供了有效的技术支撑。

关键词: 语音深度伪造归因;少样本学习;特征融合;多原型网络;音频鉴伪

Audio deepfake attribution under low-resource data conditions

Liu Shijia[1], Yan Xiaojin[1], Ren Haojie[1], Su Zhaopin[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

Abstract: Objective Driven by rapid advances in deep learning, text-to-speech (TTS) and voice conversion (VC) technologies have achieved unprecedented improvements in audio naturalness, fidelity and diversity, enabling large-scale and

收稿日期:2026-03-25;**修回日期**:2026-06-25

*通信作者: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。

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

high-quality audio forgery. However, the widespread abuse of such deepfake technologies has posed severe and persistent threats to personal privacy, social credibility and even national security, triggering urgent demands for effective audio deepfake countermeasures. Existing research in this field mostly focuses on binary deepfake detection, which only judges whether an audio clip is forged but fails to identify the specific generation model or method behind the forgery. In contrast, audio deepfake attribution (ADA) targets tracing the exact source of forged audio by extracting unique model-specific artifacts, offering interpretable evidence for AI content governance, supporting complete evidence chains for judicial forensics, and promoting standardized supervision of the AI speech industry, thus emerging as a pivotal research hotspot in artificial intelligence security. Nevertheless, current ADA approaches are plagued by prominent defects that hinder practical deployment: on one hand, most methods rely on single-layer or shallow handcrafted features, lacking sufficient feature extraction capability to mine deep semantic information and fine-grained differences between diverse forgery models; on the other hand, they exhibit extremely limited generalization ability, especially in low-resource few-shot scenarios with scarce annotated training samples and open-set scenarios involving emerging unknown forgery models, suffering from drastic performance degradation and failing to adapt to the continuous iteration of speech synthesis technologies. Therefore, this study aims to develop a high-performance audio deepfake attribution method suitable for low-resource conditions, with strong discriminative ability for known forgery methods and reliable generalization for unknown forgery scenarios. **Method** To bridge the aforementioned research gaps, this paper proposes a novel Low-resource Audio Deepfake Attribution (LADAR) method dedicated to few-shot learning scenarios, which integrates innovative multi-layer feature fusion and multi-prototype learning mechanisms to break through the bottlenecks of existing methods. The core technical design is refined based on in-depth analysis of current mainstream attribution frameworks and their limitations. First, a multi-layer adaptive feature fusion module is built to leverage the powerful pre-training representation capability of Wav2Vec2-BERT 2.0, a state-of-the-art self-supervised audio model. Instead of using single-layer hidden states as in traditional methods, this module assigns learnable attention weights to aggregate multi-layer hidden features, and introduces a shallow bias factor to enhance the sensitivity to subtle low-level forgery traces that are easily ignored by deep layers; meanwhile, a learnable gating mechanism is adopted to dynamically fuse attention-weighted pooling and average pooling, generating compact and highly discriminative model embedding representations with optimized intra-class compactness and inter-class separability. Second, a K-means clustering driven multi-prototype learning module is designed to address the limitation of traditional single-prototype representation, which cannot cover the complex intra-class feature distribution of each forgery method. By generating multiple representative prototype vectors for each category, this module fully captures the feature variation patterns of the same synthesis model under different audio lengths, speakers and content conditions, strengthening the model's robustness to sample variability. Finally, the attribution performance of LADAR is systematically evaluated under two critical experimental setups: closed-set scenario with known forgery methods and open-set scenario with unknown forgery methods, to fully verify its effectiveness and generalization. **Results** Comprehensive comparative experiments are conducted against several representative state-of-the-art audio deepfake attribution methods, and the quantitative results fully validate the superiority of the proposed LADAR method. In the closed-set scenario targeting known forgery models, the LADAR method achieves significant performance gains, with overall accuracy improved by 35.25%, 26.26%, 9.22% and 5.65% respectively compared with baseline methods, and the F1-score, a comprehensive metric balancing precision and recall, increased by 38.82%, 27.42%, 7.23% and 5.74% correspondingly, reflecting its outstanding ability to accurately distinguish different known forgery methods. In the more challenging open-set scenario involving unknown forgery models, which is highly consistent with real-world application environments, LADAR still maintains stable and efficient performance: its accuracy is enhanced by 15.80%, 37.51%, 10.73% and 10.42% respectively, and the F1-score is elevated by 20.29%, 34.17%, 11.40% and 12.33% respectively compared with contrast methods. The experimental results demonstrate that LADAR effectively alleviates performance deterioration caused by insufficient training samples and inconsistent feature distribution of unknown samples, solving the core pain points of weak generalization and poor fine-grained recognition in existing methods. **Conclusion** The systematic experimental analysis confirms that the proposed LADAR method effectively addresses the key technical challenges in audio deepfake attribution, with remarkable advantages in accuracy, robustness and cross-scenario generalization. By optimizing the feature extraction pipeline and innovating the prototype learning mechanism, it

overcomes the inherent defects of insufficient feature mining and limited adaptability to unknown samples in traditional attribution methods. This method provides reliable, interpretable and efficient technical support for practical high-demand scenarios including judicial forensics, deepfake content source tracing, network audio supervision and national security defense, and also offers a feasible technical reference for other low-resource audio security tasks. In the context of continuous evolution of deepfake technologies, the LADAR framework can extend to new speech synthesis models, enabling sustainable AI audio security governance.

Key words: Audio deepfake attribution; Few-shot learning; Multi-layer feature fusion; Multi-prototype network; Anti-spoofing

论文引用格式：Liu S J, Yan X J, Ren H J and Su Z P. 2026. Audio deepfake attribution under low-resource data conditions. Journal of Image and Graphics, XXXX: 1-13 (刘石佳,闫晓金,任浩杰,苏兆品. 2026. 少样本数据下的语音深度伪造归因方法。中国图象图形学报, XXXX: 1-13) [DOI: 10.11834/jig.260150]

0 引　言

以文本转语音(Text-to-Speech,TTS)和语音转换(voice conversion,VC)为代表的深度伪造技术在自然度、逼真度和多样性等方面显著提升(Li 等,2025)(Xu 等,2024),而这些技术的滥用带来了严重的安全隐患,对个人隐私安全、社会公信力、国家安全构成了严峻威胁(Whittaker 等,2023)。

为应对语音深度伪造威胁,学术界已开展了大量反欺骗研究,多聚焦在检测语音中是否存在伪造内容(Yu 等,2024)(张国富等,2025)(刘斯鸿等,2026)(张宇翔等 2025)(于佳祺等,2022)。然而,上述方法虽然可判断语音是否存在伪造,却无法回答"由哪种方法伪造"这一关键问题,限制了其在司法取证、溯源等实际场景中的应用效能。

相比于深伪检测,语音深度伪造归因(Audio Deepfake Attribution,ADA)可以识别虚假语音的伪造方式,能够增强反欺骗系统的可解释性,满足司法取证对证据链完整性的要求,具有更重要的实际价值。

早期 ADA 研究假设测试样本来自训练阶段已见过的模型,探索如何从伪造语音中提取独特的模型特征。Müller 等人(2022)提出基于神经嵌入的攻击者签名方法,利用循环神经网络进行聚类分析,用于语音伪造系统的归因。Neri 等人(2022)提出 ParalMGC 方法,融合 Mel 频率倒谱系数和伽马通滤波器倒谱系数特征。Klein 等人(2024)提出两阶段学习框架,可分类输入类型、声学模型和声码器。Chhibber 等人(2024)提出基于可解释概率属性嵌入的伪造语音表征方法(Explainable Probabilistic Attribute Embedding,EPAE),设计了包含声学特征预测、波形生成和说话人建模等维度的概率属性,可以实现伪造检测和攻击归因任务。Phukan 等人(2025)探究了多种语音预训练模型捕获韵律签名的能力,发现 x-vector 表现最为优异。

然而,随着语音伪造新模型的不断涌现,上述方法却难以取得有效的结果。为此,Borrelli 等人(2021)率先开展未知语音伪造方法识别研究,提出一套基于短期与长期预测痕迹的语音描述符,通过提取不同阶数线性预测分析下的预测误差能量与预测增益等手工特征,并可选地结合双相干特征,利用随机森林或一类支持向量机进行分类。Salvi 等人(2022)将多个语音伪造检测模型适配为归因分类器,提出两种未知类处理策略:一是基于置信概率比阈值,通过设定比值阈值拒绝未知样本;二是基于一类支持向量机,利用已知类样本的置信度分布训练判别模型,用于判断测试样本是否属于已知类别。Bhagtani 等人(2023)提出细粒度合成语音归因 Transformer 方法(Fine-Grain Synthetic Speech Attribution Transformer,FGSSAT),利用 Transformer 网络提取 768 维嵌入表示,结合 t-SNE 降维与 HDBSCAN 进行无监督聚类。Zhang 等人(2025)提出基于 Transformer 的 t-vector 和 s-vector 表示方法,以子带对数梅尔频谱图为输入,通过微调预训练 Transformer 生成嵌入向量 t-vector,再经线性层与 Sigmoid 激活投影为置信度评分向量 s-vector,结合类别依赖阈值机制处理未知语音伪造归因场景。Chhibber 等人(2025)提出零样本未知语音伪造归因框架,借鉴说

话人验证的思想将语音嵌入与预注册的攻击样本进行比对，采用SSL-AASIST架构与加性角度边际损失函数训练攻击嵌入提取器，实现未知语音伪造场景下的归因。Stan等人(2025)提出无需训练的语音深度伪造归因与域外检测方法(Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes，TADA)，依托k最近邻与自监督学习模型的浅层特征，实现高效的模型归因与检查点归因，同时完成域外检测，在多语言语音深度伪造数据集上的模型归因表现优异，检查点归因效果突出，域外样本检测也取得了良好性能。

虽然上述工作取得了一定的进展，但仍存在如下关键问题：

1)模型特征提取能力不足，难以充分挖掘伪造模型中的深层语义信息无法有效表征不同伪造方法间的细粒度差异。

2)对未知伪造模型的识别能力有限，难以应对层出不穷的新型语音伪造技术。尽管近期研究尝试引入自监督通用语音表征或借鉴声纹识别的零样本验证框架(Stan 等，2025)(Chhibber 等，2025)，但依赖固定的预训练特征与统一距离阈值，在域外样本特征分布差异较大时性能波动明显；后者则严重受限于注册样本的数量与质量，难以同时兼顾已知类的准确识别与未知类的有效拒识(Garg 等，2025)。

针对上述问题，本文提出一种少样本数据下的语音深度伪造归因方法(Low-resource Audio Deepfake Attribution，LADAR)，主要贡献如下：

1)在特征提取层面，设计多层特征融合模块，通过可学习注意力权重聚合预训练模型各层隐藏状态，并引入浅层偏置因子增强细粒度伪造痕迹感知能力；构建全局特征聚合模块，以可学习门控参数动态融合注意力池化与均值池化，克服单一池化策略的表达局限；同时引入模型嵌入表示的生成模块，结合SELU激活函数与残差连接生成强判别性嵌入表示，提升类内紧凑性与类间可分性。

2)在识别层面，提出多原型学习模块，为每类伪造方法生成多个原型向量建模其类内多样性，捕捉同一生成方法在不同条件下的特征变化模式，克服传统单原型表示难以覆盖复杂分布的局限，增强未知语音伪造场景下归因鲁棒性。

1 语音深度伪造归因问题

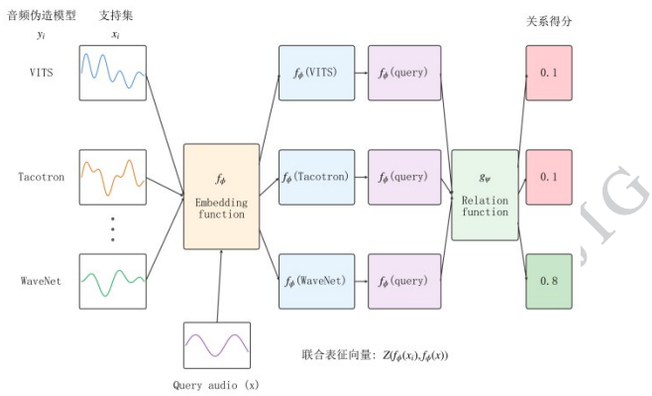


图1 语音深度伪造归因方法

Fig. 1 Audio Deepfake Attribution Method

语音深度伪造归因问题旨在通过捕捉不同语音合成模型生成语音时留下的独特模型特征，以精准识别并确认伪造语音模型，如图1所示。

首先，对于一个包含C种已知伪造方法的支持集$S=\left\{\left(\boldsymbol{x}_i,\boldsymbol{y}_i\right)|i=1,2.\ldots\ldots,m\right\}$(通常$m=C\times K$，$K$为每类样本数)，每个语音样本$\boldsymbol{x}_i$及其对应的语音伪造模型标签$\boldsymbol{y}_i$被输入嵌入网络，通过学习和训练获得伪造模型的嵌入表示$\boldsymbol{f}_\phi\left(C_i\right)$。通常会对同一方法的所有支持样本特征进行逐元素求和或平均，得到该方法的类原型特征$\boldsymbol{p}_c$，以聚合该类别的共性信息。

对于待归因的查询语音$\boldsymbol{x}$，同样通过嵌入网络提取其查询特征$\boldsymbol{f}_\phi(\boldsymbol{x})$。随后，将查询特征$\boldsymbol{f}_\phi(\boldsymbol{x})$与每个已知伪造模型的嵌入表示$\boldsymbol{f}_\phi\left(C_i\right)$通过关系网络$g_\psi(\cdot)$度量相似性，以获取查询语音与该伪造模型的置信度。通过比较所有置信度，得分最高的类别即被判定为查询语音的伪造来源。

值得注意的是，随着深度伪造技术的快速发展，伪造模型迭代更新快，很难获取新出现模型的大量样本，难以得到其准确的模型嵌入。为此，本文利用已知伪造模型的大量样本训练特征提取网络，以准确发现语音中的伪造模型细微差异；利用多原型网络构建未知模型少量语音的原型空间，提高方法对未知伪造模型的归因能力。

2 LADAR方法

LADAR方法整体框架如图2所示。

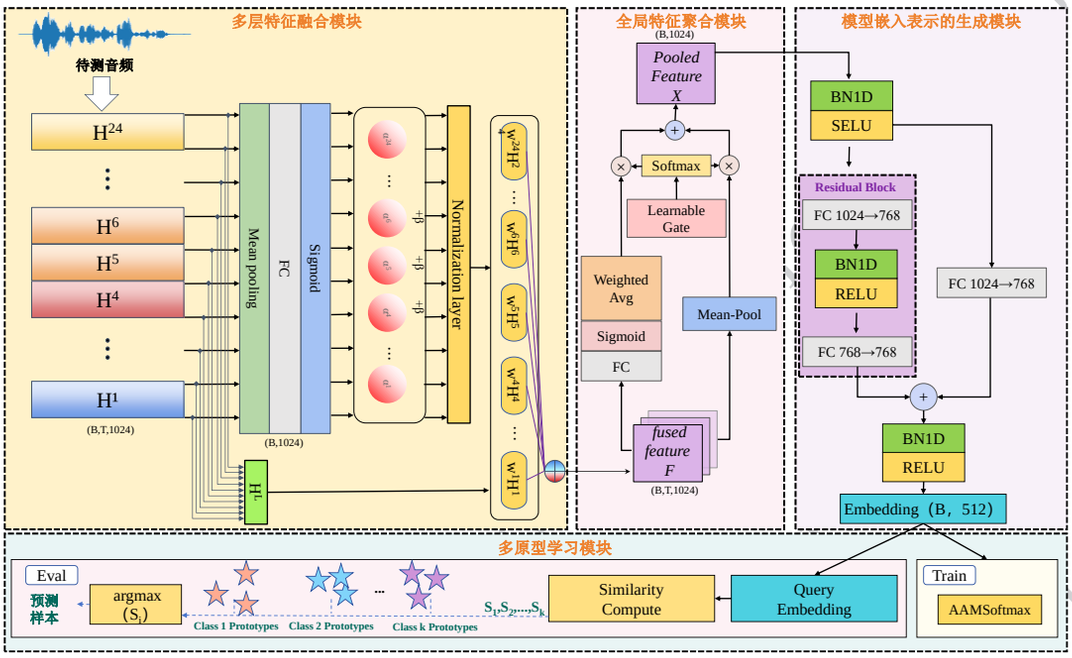


图2 LADAR方法总体结构图

Fig. 2 Architecture of LADAR

首先，基于Wav2Vec2-BERT 2.0提取原始语音的24层大模型特征，利用多层注意力融合模块对每一层Transformer输出进行加权融合，将原始层特征按归一化后的权重加权求和得到融合特征；然后，利用全局特征聚合模块将融合特征进行池化，获取更细微的模型特征，并进一步利用生成模块得到模型的嵌入表示。在归因时，通过K-means聚类为每个类别生成多个原型向量，通过计算正余弦相似度来确定伪造模型，实现语音伪造归因。

2.1 多层特征融合模块

自监督预训练语音模型的多层Transformer结构编码了不同粒度的声学与语义信息：浅层特征侧重于细粒度的时频细节与局部波形伪影，而深层特征则捕获全局结构及高级语义信息（Phukan 等，2025）。然而，现有语音深伪检测与归因方法或仅依赖单一层的输出（Stan 等，2025），或采用均匀加权等简单聚合策略（Zhang 等，2024），未能充分挖掘并利用不同层之间的特征互补性，限制了模型性能的提升。针对上述问题，本文提出多层特征融合模块（Multi-layer Feature Fusion，MFF）。

设预训练模型包含L层Transformer编码器，本文采用Wav2Vec2-BERT 2.0作为骨干网络，$L=24$。对于输入语音信号$\boldsymbol{x}$，模型输出各层隐藏状态$\boldsymbol{H}=\{\boldsymbol{H}^{(1)},\ldots,\boldsymbol{H}^{(L)}\}$，式中$\boldsymbol{H}^{(l)}\in\mathbf{R}^{T\times D}$表示第$l$层的输出，$T$为时序长度，$D=1024$为隐藏维度。

基于上述分析，本模块引入一种基于注意力机制的层级加权融合策略，旨在端到端地学习各层特征对深伪检测任务的贡献权重。具体而言：

首先，为获得每层特征的全局表示，如公式（1）所示，对第l层隐藏状态$\boldsymbol{H}^{(l)}\in\mathbf{R}^{T\times D}$沿时间维度进行平均池化。

$$\boldsymbol{h}^{(l)}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{H}_t^{(l)}。\tag{1}$$

随后，通过全连接层计算初始注意力分数，如公式（2）所示，以捕捉各层特征与检测任务之间的相关性。

$$\alpha^{(l)}=\sigma(\boldsymbol{W}_a\boldsymbol{h}^{(l)}+\boldsymbol{b}_a)\tag{2}$$

式中，$\boldsymbol{W}_a\in\mathbf{R}^{1\times D}$和$\boldsymbol{b}_a$为可学习参数，$\sigma(\cdot)$表示Sig-

moid函数，将分数映射至[0，1]区间。

考虑到浅层特征对细粒度伪造伪影的敏感性已被广泛验证（Phukan 等，2025），为引导模型更关注这些具有先验重要性的层级，本文提出一种可学习的浅层偏置机制，在初始注意力分数基础上，为预定义的浅层集合$S = \{4, 5, 6\}$引入偏置因子$\beta^{(l)}$，以增强其对最终决策的影响，如公式（3）所示。

$$\tilde{\alpha}^{(l)} = \alpha^{(l)} + \beta^{(l)} \tag{3}$$

式中，$\beta^{(l)}$初始化为0.3并参与端到端训练；对于其他层，$\beta^{(l)} = 0$。此举不仅将领域先验知识融入模型，也为浅层特征提供了更灵活的贡献空间。

为保证各层权重构成有效的概率分布，如公式（4）所示，对增强后的注意力分数进行归一化。

$$w^{(l)} = \frac{\tilde{\alpha}^{(l)}}{\sum_{j=1}^{L} \tilde{\alpha}^{(j)}} \tag{4}$$

最终，各层隐藏状态按归一化权重加权融合，如公式（5）所示，得到保留完整时序信息的联合表征。

$$\boldsymbol{F} = \sum_{l=1}^{L} w^{(l)} \boldsymbol{H}^{(l)} \tag{5}$$

式中，$\boldsymbol{F} \in \mathbf{R}^{T \times D}$作为后续时序池化模块的输入，用于捕获帧级依赖关系。

本模块的设计优势体现在自适应性、特征互补性与先验引导的有机结合。具体而言，通过端到端学习的注意力权重，模型能够根据输入样本的特征分布动态调整各层的重要性，从而避免了人工选择特定层的主观性与局限性；同时，该机制融合了浅层的细粒度声学特征与深层的全局语义特征，充分利用了预训练模型多层特征的异构信息，有效增强了对多样化伪造模式的覆盖能力。此外，浅层偏置机制将领域先验知识显式融入模型，在训练初期提供有效的梯度引导，在加速收敛的同时提升了检测性能，实现了数据驱动与知识驱动的协同优化。

2.2 全局特征聚合模块

基于自监督学习的预训练语音模型输出的特征天然携带时序维度，如何将变长时序特征压缩为固定维度的表征向量，同时完整保留对伪造归因至关重要的细粒度鉴别性信息，是亟待解决的问题。传统方法通常采用单一的池化策略，如均值池化（Li 等，2024）或注意力池化（Zhang 等，2024），但这些方法各有局限性：均值池化虽然计算简单且对所有时间步一视同仁，但可能淹没局部关键特征；注意力池化虽能自适应地聚焦于重要时间步，但可能忽略全局统计信息。针对上述问题，本文提出一种全局特征聚合模块（Global Feature Aggregation，GFA）。该模块融合注意力池化与均值池化的优势，并通过可学习的门控机制自适应地调节两种池化策略的贡献比例，从而实现局部显著性特征与全局统计特征的有效融合。

GFA模块包含两条并行的池化分支。第一条分支为基于Sigmoid的加权注意力池化分支，通过学习每个时间步的重要性权重实现对关键帧的自适应聚焦。给定融合后的时序特征$\boldsymbol{X} \in \mathbf{R}^{B \times T \times D}$（式中$B$为批处理大小，$T$为时序长度，$D = 1024$为特征维度），首先通过一个全连接层计算每个时间步的注意力分数$\boldsymbol{s} = \boldsymbol{W}_a \boldsymbol{X} + \boldsymbol{b}_a$，式中$\boldsymbol{W}_a \in \mathbf{R}^{D \times 1}$为可学习权重矩阵，$\boldsymbol{b}_a \in \mathbf{R}^1$为偏置项。随后通过Sigmoid激活函数将注意力分数归一化至（0，1）区间，得到注意力权重$\alpha = \sigma(\boldsymbol{s})$。最终，注意力池化结果通过加权平均计算得到，如公式（6）所示。

$$\boldsymbol{h}_{attn} = \frac{\sum_{t=1}^{T} \alpha_t \odot \boldsymbol{x}_t}{\sum_{t=1}^{T} \alpha_t + \epsilon} \tag{6}$$

式中，$\odot$表示逐元素乘法，$\epsilon = 10^{-8}$为数值稳定系数。

第二条分支为均值池化分支，采用全局平均池化策略[1，3]对时序维度进行简单平均操作，以捕获输入特征的全局统计信息，如公式（7）所示。

$$\boldsymbol{h}_{mean} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \tag{7}$$

均值池化能够有效保留特征的整体分布特性，对于捕获语音信号的长程依赖具有重要作用。

为了自适应地融合上述两条分支的输出，本文设计了一种基于Softmax的可学习门控融合机制。具体而言，定义可学习的门控参数$\boldsymbol{g} = [g_1, g_2]^T$，其初始值设为$[0.6, 0.4]^T$，表示初始时偏向注意力池化。通过Softmax函数对门控参数进行归一化处理，得到归一化门控权重，如公式（8）所示。

$$[\hat{g}_1, \hat{g}_2]^T = \mathrm{Softmax}(\boldsymbol{g}) = \left[\frac{e^{g_1}}{e^{g_1} + e^{g_2}}, \frac{e^{g_2}}{e^{g_1} + e^{g_2}}\right]^T \tag{8}$$

式中$\hat{g}_1 + \hat{g}_2 = 1$，保证了两种池化结果的权重之和为1。最终的融合池化结果计算如公式（9）所示。

$$\boldsymbol{h}_{pooled} = \hat{g}_1 \cdot \boldsymbol{h}_{attn} + \hat{g}_2 \cdot \boldsymbol{h}_{mean} \tag{9}$$

式中$\boldsymbol{h}_{pooled} \in \mathrm{R}^{B \times D}$为最终的池化输出。门控参数$\boldsymbol{g}$在训练过程中与网络其他参数联合优化，使模型能

够根据具体任务自动学习两种池化策略的最优组合比例。

注意力分支仅引入$D+1$个参数(全连接层权重和偏置)，门控机制仅引入2个可学习参数，整体参数量增加可忽略不计，但显著提升了时序特征聚合的灵活性和有效性。通过门控机制的自适应学习，模型能够在不同检测场景下动态调整局部显著性特征与全局统计特征的融合比例，从而获得更具鉴别力的时序表示。

2.3 模型嵌入表示的生成模块

池化后的高维特征$\boldsymbol{X}$需映射至低维嵌入空间以用于分类。为实现这一目标，设计了一种模型嵌入表示的生成模块(Embedding Generation，EG)，以混合时序池化输出的特征$\boldsymbol{X}\in \mathrm{R}^{B\times 1024}$为输入，首先通过批归一化与全连接层将维度压缩至768，以降低后续计算复杂度。随后，模块引入残差学习机制，通过一个两层残差块对特征进行非线性变换，同时使用残差投影将原始输入直接映射至同一维度，二者相加后经全连接层输出最终的512维嵌入特征$\boldsymbol{E}\in \mathrm{R}^{B\times 512}$。残差连接的引入使得模块在保持原始特征有效信息的同时，能够学习更丰富的特征表示，并有效缓解梯度消失问题，为后续的AAMSoftmax分类提供高质量的输入嵌入。

2.4 多原型学习模块

在未知语音伪造归因场景中，即使来自同一伪造方法生成的语音样本，也可能因说话人变化、语义内容不同或生成参数差异而表现出显著的类内多样性。传统的单原型表示方法通常仅依赖类别中心进行匹配，难以充分建模这种多样性分布，导致归因性能受到限制。为解决这一问题，本文提出多原型学习模块(Multi-prototype Learning，MPL)，通过引入聚类机制为每个类别生成多个原型向量，从而更精细地刻画类内特征分布。

MPL模块的处理流程包括以下步骤：首先，对于每个伪造类别的参考样本，通过预训练特征提取器和残差增强投影网络获取归一化的嵌入表示；接着，对每类样本的嵌入向量执行K-means聚类，将其划分为K个子簇，并计算各子簇的聚类中心；最后，对所有聚类中心进行L2归一化，构成该类别的多原型表示。在推理阶段，测试样本的嵌入向量与所有类别的原型计算余弦相似度，并选取相似度最高的原型所对应的类别作为最终的归因结果。

在基于聚类的原型选取任务中，每类样本的原型个数K是影响聚类结构合理性与泛化能力的关键参数。若K取值过小，则难以刻画类内样本的潜在子结构；若K取值过大，则易产生单点簇或使聚类退化为异常值检测，导致轮廓系数等内部评估指标失真。

为科学确定K值，本文综合运用聚类有效性理论与肘部法则。根据聚类有效性理论，最佳聚类数的搜索范围应满足$K\leqslant\sqrt{n}$，式中n为单类样本量(于剑等，2002)。该约束可有效避免因K过大而导致的过分割问题。

在候选K值范围内，采用肘部法则进一步确定最优值(刘畅等，2021)。设某类别包含n个样本，将其划分为K个子簇$C_1, C_2, \cdots, C_K$，子簇C_j的聚类中心为μ_j，x_i为子簇C_j内的第i个样本点，则误差平方和(Sum of Squared Errors，SSE)定义为：

$$SSE=\sum_{j=1}^{K}\sum_{x_i\in C_j}\left\| x_i-\mu_j\right\|^2 \tag{10}$$

当K从a增加至$b=a+1$时，相邻K值的SSE下降率(SSE Decreasing rate)计算如下：

$$SSE\ Decreasing\ rate=\frac{SSE(a)-SSE(b)}{SSE(a)}\times 100\% \tag{11}$$

该指标反映了增加一个原型对聚类误差的边际改善程度。随着K增加，SSE通常呈下降趋势，但下降率逐渐减小。当相邻K值间的SSE下降率出现显著衰减时，表明继续增加原型数的边际收益已有限，此时的K即为最优原型个数。

3 实验结果与分析

为全面评估LADAR模型的归因性能，本文设计了两组递进式实验：1)已知语音伪造模型归因实验，测试集中的所有伪造类型均在训练集中出现过。本实验旨在评估LADAR模型在类别完备条件下的归因性能，作为归因任务的基础性能基准。2)未知语音伪造模型归因实验，训练集与测试集的伪造语音类别互不相交，测试集中的所有伪造类型均未在训练集中出现。本实验旨在评估LADAR模型在面对未见伪造模型时的归因泛化能力。

将LADAR方法与领域内效果出色的EPAE(Chhibber 等，2024)、TADA(Stan 等，2025)、SLS(Zhang 等，2024)、SSL(Tak 等，2022)方法进行对比。

其中，EPAE（Chhibber 等，2024）和 TADA（Stan 等，2025）是专为欺骗语音归因任务设计的方法，EPAE 构建可解释概率属性嵌入结合决策树实现归因，TADA 基于预训练 SSL 模型与 kNN 实现无训练的跨数据集归因及域外检测，除此以外，可实现难度更高的检查点归因；对于 SLS（Zhang 等，2024）和 SSL（Tak 等，2022），二者原为二分类声伪检测模型，因基于预训练自监督模型的特征提取机制具备极强泛化性，对未知攻击算法和领域失配场景均表现出优异的检测性能，所以选择它们作对比。为适配多分类归因任务，将其后端分类器替换为对应类别数的全连接层，并采用 AAMSoftmax 损失函数进行优化，分类决策基于余弦相似度计算。

3.1 数据集与实验设置

从 Codecfake 数据集（Xie 等，2025）中随机抽取真实语音（real）及 F01 至 F06 六类伪造语音各 10，000 条，并采用 f5、fishspeech、gptsovits 三种 TTS 方法以及 rvc、svc 两种 VC 方法合成的数据各 5，000 条数据，构建包含真实语音与多种伪造类型的综合数据集。按 3∶1 的比例将上述数据划分为训练集与测试集。

针对已知语音伪造方法的测试，选取 f5、svc、fishspeech 三种类型，各收集与训练集不重叠的 1，000 条数据。从每类的 1，000 条中随机抽取 10 条作为支持集，其余 990 条作为测试集。

针对未知语音伪造方法的测试，数据由三部分构成：其一为 cosyv2、seedvc 两种方法合成的 4990 条语音数据，其二为从 Codecfake 数据集（Xie 等，2025）的 F07 类别中随机抽取 4，990 条；其三为从 Speech-Fake 数据集（Huang 等，2025）的 hifiGAN、StarGAN、VITS 三种伪造类型中随机抽取的 4，990 条语音。为支持未知伪造语音归因判别，分别从上述三类未知伪造数据中额外抽取 10 条语音构建支持集。需说明的是，第一类未知伪造数据（cosyv2 与 seedvc）为独立合成生成，与测试数据中的同源样本无任何交叉；第二类与第三类未知伪造数据（Codecfake F07 及 SpeechFake）亦确保参考集与测试集之间无样本重叠。

以上数据集，语音采样率为 64600HZ，所有方法代码均基于 python 实验，在配备 RTX4090D、60GB、ubuntu20.04 服务器上进行。选用 ACC、precision、F1-score、recall 作为模型性能的评价指标。

3.2 参数选择

在本文语音伪造模型归因实验的支持集中，共包含六种语音伪造模型：cosyv2、seedvc、hifiGAN、StarGAN、VITS、F07。每类样本量 $n = 10$。根据聚类有效性理论，最佳聚类数的搜索范围应满足 $K \leqslant \sqrt{n}$（于剑等，2002），代入得 $K \leqslant \sqrt{10} \approx 3$。因此，本文将有效 K 值范围限定为{1，2，3}。

在此基础上，采用肘部法则确定最优 K 值（刘畅等，2021）。对每个候选 K，按公式（10）计算各类别内嵌入向量的 SSE，并按公式（11）计算相邻 K 值的 SSE 下降率，结果如图 3 和图 4 所示。

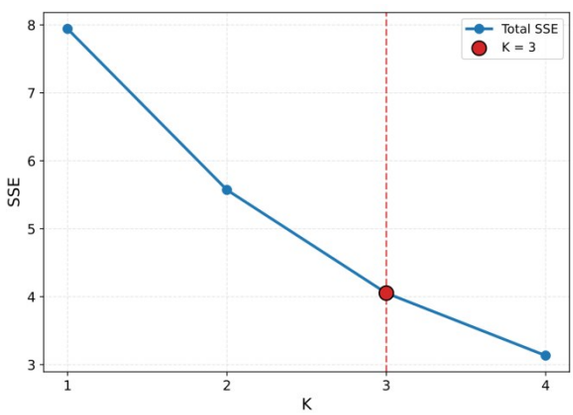


图 3　不同原型个数 K 对应的总误差平方和

Fig. 3　Total SSE for different numbers of prototypes K

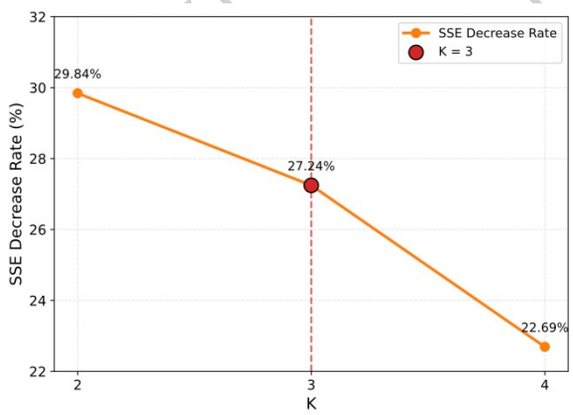


图 4　不同原型数量下的 SSE 下降率曲线

Fig. 4　SSE Decrease Rate Curve

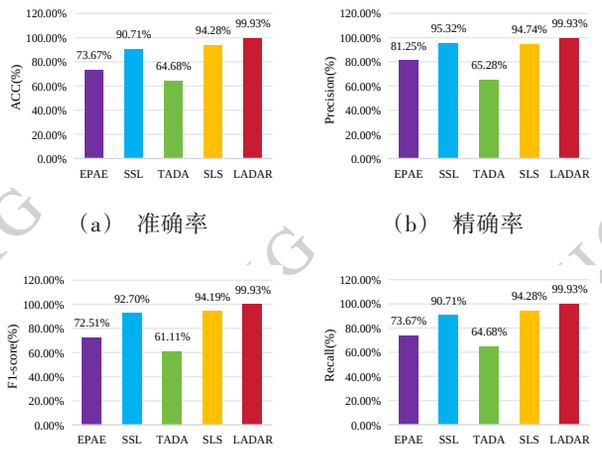
实验结果表明，SSE 随 K 增加呈下降趋势，但相邻 K 值之间的下降收益逐渐减弱。具体而言，$K = 1 \rightarrow 2$、$K = 2 \rightarrow 3$ 和 $K = 3 \rightarrow 4$ 的 SSE 下降率分别为 29.84%、27.24% 和 22.69%。其中，$K = 3 \rightarrow 4$ 的下降率较 $K = 2 \rightarrow 3$ 进一步降低了约 4.55 个百

分点，下降收益的衰减幅度明显增大，表明当K超过3后，继续增加原型数所带来的边际收益显著递减。综合样本量约束与肘部法则分析，本文最终选取$K=3$作为每类样本的原型个数。

3.3 对比实验分析

3.3.1 已知语音伪造模型归因实验

已知语音伪造模型归因实验结果如图5所示。在已知伪造场景下，所有类别在训练阶段均已见过，本文LADAR方法在四个评价指标上均取得了最优性能。



（a） 准确率　（b） 精确率

（c） F1分数　（d） 召回率

((a)ACC;(b)Precision;(c)F1-score;(d)Recall)

图5 已知语音伪造模型归因结果

Fig. 5 Attribution results on known audio

与已有方法相比，LADAR方法的ACC分别比TADA(Stan等，2025)、EPAE(Chhibber等，2024)、SSL(Tak等，2022)、SLS(Zhang等，2024)提升了35.25%、26.26%、9.22%和5.65%，Precision分别提升了34.65%、18.68%、4.61%和5.19%，F1-score分别提升了38.82%、27.42%、7.23%和5.74%，Recall分别提升了35.25%、26.26%、9.22%和5.65%。

值得注意的是，EPAE(Chhibber等，2024)和TADA(Stan等，2025)方法在本数据集的已知语音伪造方法场景下表现较差，准确率仅为73.67%和64.68%。这主要是因为EPAE方法依赖于人工设计的属性嵌入，其属性定义和分类器训练均基于特定的伪造模型结构，比如：声学特征预测器、声码器类型等，具有较强的先验依赖性，不适用数据集里面的f5、fishspeech和svc伪造方法。TADA(Stan等，2025)需要大量的支持集数据支撑，在支持集样本少，仅为10条的情况下，归因效果大幅下降。而SSL(Tak等，2022)和SLS(Zhang等，2024)两种方法均取得了较好的性能，其中SLS(Zhang等，2024)通过敏感层选择机制达到了94.28%的准确率，但仍低于本文方法5.65%，说明单纯的层加权聚合策略难以充分挖掘多层特征的互补性。

本文LADAR方法在已知语音伪造方法归因任务中表现出显著优势，图6和表1分别为LADAR在f5、fishspeech和svc三种类别上的混淆矩阵和具体检查结果。

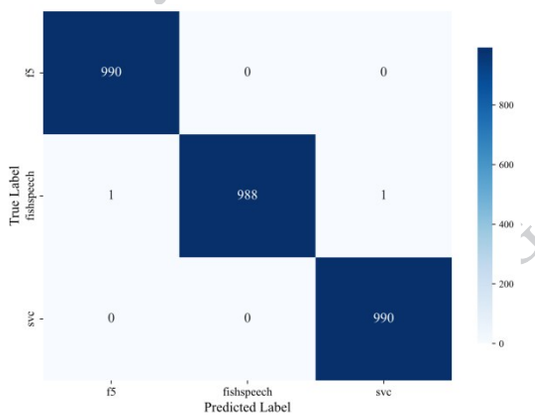


图6 已知场景下LADAR的混淆矩阵

Fig. 6 Confusion matrix of LADAR in known scenarios

表1 已知伪造方法归因的分类性能

Table 1 Known forgery classification performance

Method	ACC	Precision	Recall	f1-score
f5	100.000%	99.899%	100.000%	99.950%
fishspeech	99.798%	100.000%	99.798%	99.899%
svc	100.000%	99.899%	100.000%	99.950%
macro avg	99.933%	99.933%	99.933%	99.933%

每类各990条测试样本，共计2970条。其中，f5和svc的类别准确率与Recall均达到100%，表明模型对这两个类别实现了完全正确的识别；fishspeech的Precision为100%，即所有被预测为fishspeech的样本均判别正确，其2条误分类样本被分别归入f5和svc。整体来看，LADAR在已知伪造方法之间表现出较强的类间区分能力，其少量误差主要来自fishspeech类中的个别边界样本。

这是因为LADAR通过多层特征融合模块能够有效聚合预训练模型各层的特征表示，结合浅层偏

置机制增强对细粒度伪造痕迹的捕获能力;同时,全局特征聚合模块能够自适应地融合注意力加权池化与均值池化的优势,充分挖掘时序特征中的判别性信息;模型嵌入表示的生成模块进一步提升了嵌入表示的类内紧凑性和类间可分性,从而实现了更精准的伪造识别。

进一步,已知语音伪造模型归因的支持集样本在特征空间中的分布如图7所示。通过t-SNE降维将高维嵌入特征可视化到二维平面,可以观察到各类别样本形成了明显的聚类结构。每个类别通过K-Means算法提取3个原型中心(大星号标记),体现了多原型聚类表示模块捕获类内多样性的能力。不同类别之间边界清晰,表明模型学习到了具有良好区分性的特征表示,已知语音伪造模型归因准确率达到99.933%。

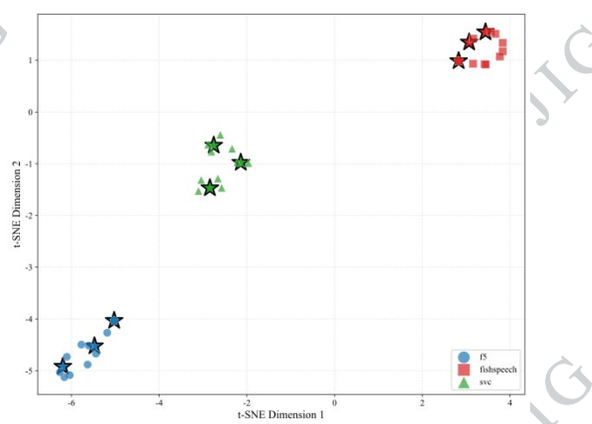


图7　已知语音少样本特征的t-SNE可视化

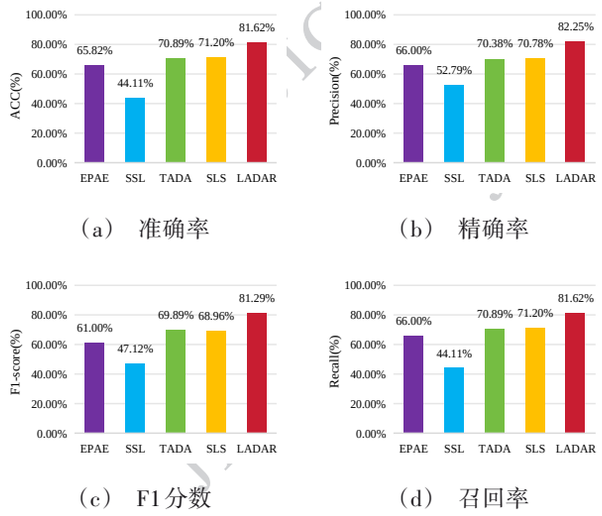
Fig. 7　t-SNE of Few-Shot Features on Known Audio

3.3.2　未知语音伪造模型归因实验

未知语音伪造模型归因实验结果如图8所示。在未知伪造场景下,测试集中包含训练阶段未见过的伪造类型(cosyv2、seedvc、hifiGAN、StarGAN、VITS、F07),用于评估模型对同类别中未见伪造方法的泛化能力。

与已有方法相比,LADAR方法的ACC分别比EPAE、SSL、TADA、SLS提升了15.80%、37.51%、10.73%和10.42%,Precision分别提升了16.25%、29.46%、11.87%和11.47%,F1-score分别提升了20.29%、34.17%、11.40%和12.33%,Recall分别提升了15.62%、37.51%、10.73%和10.42%。

本文LADAR方法在未知语音伪造方法归因场景下具有更强的泛化能力,混淆矩阵如图9所示。



(a)　准确率　(b)　精确率

(c)　F1分数　(d)　召回率

((a)ACC;(b)Precision;(c)F1-score;(d)Recall)

图8　未知语音伪造模型归因结果

Fig. 8　Attribution results on unknown audio

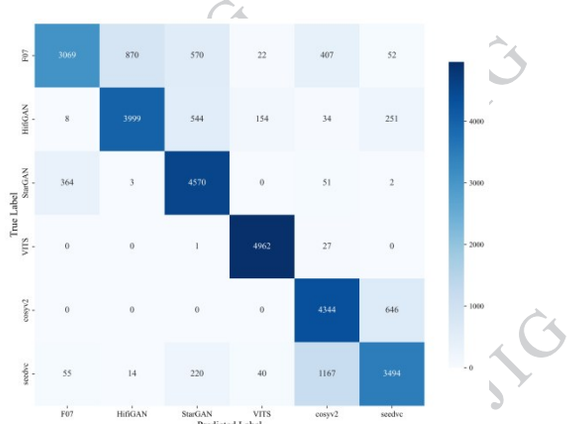


图9　未知场景下LADAR的混淆矩阵

Fig. 9　Confusion matrix of LADAR in unknown scenarios

这是因为LADAR基于嵌入相似度的未知语音伪造方法归因框架通过与参考原型的余弦相似度进行最近邻分类,能够有效捕获不同伪造方法之间的语义相似性,从而对同类别中的未见伪造类型实现准确归因;同时,多层特征融合模块学习到的特征表示具有更强的泛化性,能够提取到不同伪造方法共有的类别级特征。

相较于已知语音伪造归因场景,所有方法在未知场景下的性能均有所下降,这符合预期,因为未见伪造类型的特征分布与训练数据存在差异。然而,本文方法的性能下降幅度最小,ACC从99.93%降至81.62%,下降18.31个百分点,而SSL(Tak等,2022)方法的性能下降最为显著(ACC从90.71%降

至44.11%，下降46.60个百分点），表明SSL（Tak等，2022）方法对训练数据的过拟合较为严重。TADA（Stan等，2025）和SLS（Zhang等，2024）方法在未知语音伪造方法归因场景下的表现相近（ACC分别为70.89%和71.20%），但均明显低于本文方法，说明仅依赖预训练特征或简单的层选择机制难以有效应对分布偏移问题。本文方法通过全局特征聚合策略和模型嵌入表示的生成，有效提升了模型对未见类别的泛化能力。

未知场景下6类语音支持集样本的特征分布如图10所示，其中包含训练时已见类别和未见类别。通过t-SNE可视化可见，多原型聚类有助于改善类间可分性并压缩类内散布，例如在seedvc与cosyv2边缘区域、F07与StarGAN边缘区域，传统单一原型聚类方法易导致类内距离较大、类间边界模糊，从而影响归因准确率。每个类别的3个原型中心（大星号）大致分布于对应聚类区域，表明多原型表示能够在一定程度上增强对类内特征变化的覆盖能力。上述分布反映出该模型在未知语音伪造方法归因场景下展现出一定的特征区分能力。

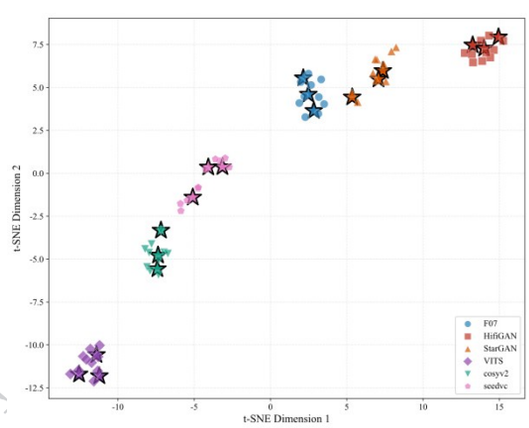


图10　未知语音少样本特征的t-SNE可视化

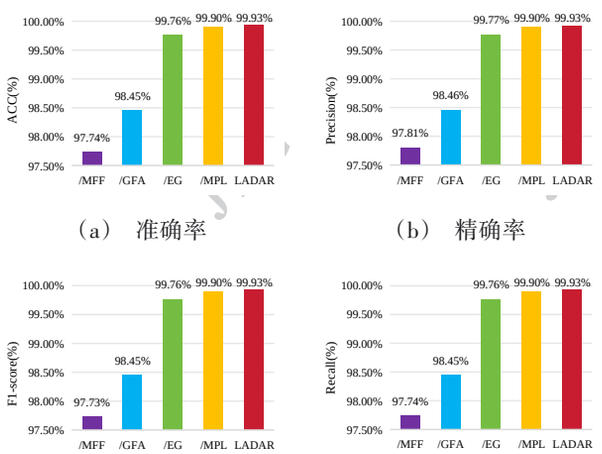
Fig. 10　t-SNE of Few-Shot Features on Unknown Audio

3.4　模块有效性实验

在本文模型中，MFF、GFA、EG和MPL为核心模块。本小节采用同一模型在已知和未知语音伪造情景下分别验证模块有效性，将依次单独去除MFF、GFA、EG和MPL来验证每个模块的有效性，实验结果如图11所示，其中/表示去除对应模块的模型。

在已知语音伪造情景下，由图11可以看出，相比于完整模型LADAR，去除MFF后，ACC降低了2.19%，Precision降低了2.12%，F1-score降低了2.20%，Recall降低了2.19%；去除GFA后，ACC降低了1.48%，Precision降低了1.47%，F1-score降低了1.48%，Recall降低了1.48%；去除EG后，ACC降低了0.17%，Precision降低了0.16%，F1-score降低了0.17%，Recall降低了0.17%；去除MPL后，ACC降低了0.03%，Precision降低了0.03%，F1-score降低了0.03%，Recall降低了0.03%。

MFF能够自适应地融合预训练模型不同层的特征表示，通过可学习的层权重和浅层偏好机制，有效捕获从低级声学特征到高级语义特征的多层次伪造痕迹；GFA通过均值池化捕获全局统计信息，通过注意力池化关注关键时间步特征，门控机制动态调节两者贡献比例，实现优势互补，对模型性能影响最为显著；EG通过残差连接和非线性变换，有效精炼特征表示，增强特征的判别性和鲁棒性；MPL通过多个聚类中心捕获类内多样性，提升已知语音伪造场景下的鲁棒性与准确率。因此，MFF、GFA、EG和MPL模块对本文模型的已知语音伪造方法归因性能均至关重要。



（a）　准确率　　（b）　精确率

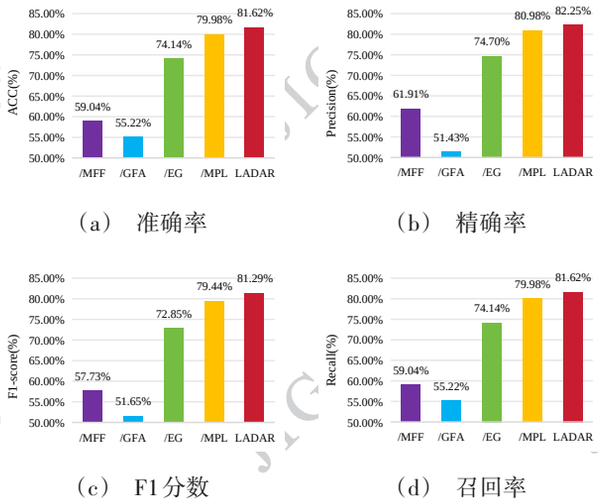
（c）　F1分数　　（d）　召回率

（(a)ACC；(b)Precision；(c)F1-score；(d)Recall）

图11　已知情景下模块有效性结果

Fig. 11　Results of module effectiveness in known scenarios

在未知语音伪造情景下，由图12可以看出，相比于完整模型LADAR，去除MFF后，ACC降低了22.58%，Precision降低了20.34%，F1-score降低了23.56%，Recall降低了22.58%；去除GFA后，ACC降低了26.40%，Precision降低了30.82%，F1-score降低了29.64%，Recall降低了26.40%；去除EG后，



(a) 准确率 (b) 精确率 (c) F1分数 (d) 召回率

((a)ACC;(b)Precision;(c)F1-score;(d)Recall)

图12 未知情景下模块有效性结果

Fig. 12 Results of module effectiveness in unknown scenarios

ACC降低了7.48%,Precision降低了7.55%,F1-score降低了8.44%,Recall降低了7.48%;去除MPL后,ACC降低了1.64%,Precision降低了1.27%,F1-score降低了1.85%,Recall降低了1.64%。

在未知语音伪造场景下,各模块的影响程度与已知语音伪造场景存在显著差异,GFA在未知语音伪造场景下影响最为显著,去除后性能骤降超过26%,表明全局统计信息与关键时间步特征的动态融合对于识别未见过的伪造方法尤为关键;MFF同样表现出极强的重要性,去除后性能下降超过22%,说明多层次伪造痕迹的自适应融合在面对新型伪造技术时能够提供更全面的判别依据;EG去除后性能下降约7.5%,残差连接和非线性变换所带来的特征精炼能力对未知语音伪造方法归因泛化同样不可或缺;MPL通过多聚类中心捕获类内多样性的机制为未知语音伪造方法归因泛化提供了贡献,在原有单中心的方法基础上提升稳定。

4 结 论

为解决现有语音深度伪造归因方法特征提取能力不足、未知伪造模型识别能力有限,难以适配少样本场景的关键问题,本文提出了一种LADAR的端到端归因模型。首先,构建了多层特征融合、全局特征聚合与多原型学习相结合的归因框架,通过可学习注意力权重融合Wav2Vec2-BERT 2.0不同层隐藏状态,引入浅层偏置因子增强底层伪造痕迹感知能力,利用门控机制融合注意力池化与均值池化生成具有较强判别性的模型嵌入表示,并通过多原型学习刻画同一伪造方法的类内差异。对比实验结果表明,相比依赖固定特征、统一阈值或单原型表示的方法,LADAR方法能够更充分地提取细粒度伪造源特征,更好地覆盖类别内部复杂分布,在少样本条件下具有更好的归因准确性和泛化能力。

需要注意的是,本文LADAR方法仍存在一定局限。当不同伪造方法采用相似声学模型、声码器或训练数据时,其伪造痕迹可能较为接近,导致模型对同源或近源方法的细粒度区分能力仍有限。此外,少样本场景下原型质量易受参考样本数量和代表性影响,可能造成部分未知类别归因结果波动。这将是我们下一步研究的重点。

参考文献

Li M L, Ahmadiadli Y and Zhang X P. 2025. A survey on speech deepfake detection. ACM Comput. Surv., 57(7): 165[DOI: 10.1145/3714458]

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(8):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(8):2236-2268)[DOI:10.11834/jig.230476]

Whittaker L, Mulcahy R, Letheren K, Kietzmann J and Russell-Bennett R. 2023. Mapping the deepfake landscape for innovation: a multidisciplinary systematic review and future research agenda. Technovation, 125: 102784[DOI: 10.1016/j. technovation. 2023. 102784]

Yu N, Chen L, Leng T, Chen Z G and Yi X Y. 2024. An explainable deepfake of speech detection method with spectrograms and waveforms. Journal of Information Security and Applications, 81: 103720[DOI:10.1016/j.jisa.2024.103720]

Zhang G F, Wang R, Su Z P, Yue F, Lian C S and Yang B. 2025. Multi-stage detection and multimodal localization for audio deletion tampering. Computer Engineering & Science, 47(11):1964-1973(张国富, 王茹, 苏兆品, 岳峰, 廉晨思, 杨波. 2025. 音频删除篡改的多阶段检测与多模态定位. 计算机工程与科学, 47(11):1964-1973)[DOI:10.3969/j.issn.1007-130X.2025.11.00]

Liu S H, Lei Z C, Qian G Y, Zhou Y and Liu C H. 2026. Speech forgery detection method based on cross-layer attention injection mechanism of pre-trained models. Journal of Information Security, 10(6):163-177(刘斯鸿, 雷震春, 钱广源, 周勇, 刘长红. 2026. 基于预训练模型跨层注意力注入机制的语音伪造检测方

法 . 信息安全学报 , 10 (6) : 163-177)[DOI: 10.19363/J.cnki.cn10-1380/tn.2025.11.13]

Zhang Y X, Li Z, Lu J Z, Shang Z Q, Chen S L, Wang W C and Zhang P Y. 2025. Spoof speech detection based on speaker features. ACTA ACUSTICA, 50 (1): 201-210 (张宇翔, 李茁, 陆镜泽, 尚增强, 陈树丽, 王文超, 张鹏远 . 2025. 基于声纹特征的伪造语音检测 . 声学学报 , 50 (1): 201-210)[DOI: 10.12395/0371-0025.2023278]

Yu J Q, Jian Z H, Xu J, You L, Wang Y L and Wu C. 2022. Spoofing speech detection algorithm based on joint feature and random forest. Telecommunications Science, 38 (6): 91-99 (于佳祺, 简志华, 徐嘉, 游林, 汪云路, 吴超 . 2022. 基于联合特征与随机森林的伪装语音检测 . 电信科学 , 38 (6): 91-99)[DOI: 10.11959/j.issn.1000-0801.2022089]

Müller N M, Dieckmann F and Williams J. 2022. Attacker attribution of audio deepfakes[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2203.15563

Neri M, Ferrarotti A, Luca De L, Salimbeni A and Carli M. 2022. Paralmgc: multiple audio representations for synthetic human speech attribution//2022 10th European Workshop on Visual Information Processing (EUVIP). Lisbon, Portugal: IEEE: 1-6[DOI: 10.1109/EUVIP53989.2022.9922861]

Klein N, Chen T X, Tak H, Casal R and Khoury E. 2024. Source tracing of audio deepfake systems//Interspeech 2024. Kos, Greece: ISCA: 1100-1104[DOI: 10.21437/Interspeech.2024-1283]

Chhibber M, Mishra J, Shim H J and Kinnunen T H. 2024. An explainable probabilistic attribute embedding approach for spoofed speech characterization[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2409.11027

Phukan O C, Singh D, Behera S R, Buduru A B and Sharma R. 2025. Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics: 4206-4214[DOI: 10.18653/v1/2025.findings-acl.218]

Borrelli C, Bestagini P, Antonacci F, Sarti A and Tubaro S. 2021. Synthetic speech detection through short-term and long-term prediction traces. EURASIP Journal on Information Security, 2021 (1): 2 [DOI: 10.1186/s13635-021-00116-3]

Salvi D, Bestagini P and Tubaro S. 2022. Exploring the synthetic speech attribution problem through data-driven detectors//2022 IEEE International Workshop on Information Forensics and Security (WIFS). Shanghai, China: IEEE: 1-6 [DOI: 10.1109/WIFS55849.2022.9975440]

Bhagtani K, Yadav A K S, Xiang Z Y, Bestagini P and Delp E J. 2023. Fgssat : unsupervised fine-grain attribution of unknown speech synthesizers using transformer networks//2023 57th Asilomar Conference on Signals, Systems, and Computers. Pacific Grove, CA, USA: IEEE: 1135-1140 [DOI: 10.1109/IEEECONF59524.2023.10476982]

Zhang Q, Zhang X W, Sun M and Yang J B. 2025. A transformer-based deep learning approach for recognition of forgery methods in spoofing speech attribution. Applied Soft Computing, 171: 112798[DOI: 10.1016/j.asoc.2025.112798]

Chhibber M, Mishra J and Kinnunen T H. 2025. Advancing zero-shot open-set speech deepfake source tracing[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2509.24674

Stan A, Combei D, Oneata D and Cucu H. 2025. Tada: training-free attribution and out-of-domain detection of audio deepfakes//Interspeech 2025. Rotterdam, The Netherlands: ISCA: 1543-1547 [DOI: 10.21437/Interspeech.2025-472]

Garg A, Cai Z X, Li Xinyuan H, García-Perera L P, Duh K, Khudanpur S, Wiesner M and Andrews N. 2025. Rapidly adapting to new voice spoofing: few-shot detection of synthesized speech under distribution shifts[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2508.13320

Zhang Q S, Wen S B and Hu T. 2024. Audio deepfake detection with self-supervised xls-r and sls classifier//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne VIC, Australia: Association for Computing Machinery: 6765-6773[DOI: 10.1145/3664647.3681345]

Li M L and Zhang X P. 2024. Interpretable temporal class activation representation for audio spoofing detection//Interspeech 2024. Kos, Greece: ISCA: 1120-1124 [DOI: 10.21437/Interspeech. 2024-2156]

Yu J and Cheng Q S. 2002. Search range of optimal cluster number in fuzzy clustering method. Science in China (Series E), 32 (2): 275-280 (于剑, 程乾生 . 2002. 模糊聚类方法中的最佳聚类数的搜索范围。中国科学 (E 辑), 32 (2): 275-280)[DOI: 10.1360/ze2002-32-2-274]

Liu C, Xiao B, Jiang T J, Su K, He P X and Wang C Y. 2021. Improved K-means small sample clustering algorithm. Journal of Ordnance Equipment Engineering, 42 (S1): 266-270 (刘畅, 肖斌, 蒋铁军, 苏凯, 何鹏翔, 王成宇 . 2021. 一种改进 K 均值的小样本聚类算法。兵器装备工程学报, 42 (S1): 266-270)[DOI: 10.11809/bqzbgcxb2021.S1.057]

Tak H, Todisco M, Wang X, Jung J W, Yamagishi J and Evans N. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation[EB/OL]. [2026-03-07]. https://arxiv.org/abs/2202.12233

Xie Y K, Lu Y, Fu R B, Wen Z Q, Wang Z Y, Tao J H, Qi X, Wang X P, Liu Y K, Cheng H N, Ye L and Sun Y. 2025. The codecfake dataset and countermeasures for the universally detection of deepfake audio. IEEE Transactions on Audio, Speech and Language Processing, 33: 386-400[DOI: 10.1109/TASLPRO.2025.3525966]

Huang W, Gu Y M, Wang Z M, Zhu H J and Qian Y M. 2025. Speechfake: a large-scale multilingual speech deepfake dataset incorporat-

ing cutting-edge generation methods//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics: 9985-9998 [DOI: 10.18653/v1/2025.acl-long.493]