基于音频的生物特征认证系统面临的深度伪造语音合成威胁

MengzeHong[1], **Di Jiang**[1]*,**Zeying Xie**[2],**Weiwei Zhao**[2]**Guan Wang**[1],
Chen Jason Zhang[1]1香港理工大学， [2]AI Group，微众银行有限公司

arXiv:2601.02914v1 [cs.SD] 6 Jan 2026

摘要

随着音频深度伪造从研究伪影发展为广泛可用的商业工具，高风险行业中鲁棒的生物特征认证正面临紧迫的安全威胁。本文基于大规模语音合成数据集，对最先进的说话人认证系统进行了系统的实证评估，揭示了两大安全漏洞：1）基于极少样本训练的现代语音克隆模型可以轻易绕过商业说话人验证系统；2）反欺骗检测器难以泛化到不同的音频合成方法，导致域内性能与现实世界鲁棒性之间存在显著差距。这些发现促使我们重新审视安全措施，并强调了架构创新、自适应防御以及转向多因素认证的必要性。

1 引言

基于声纹的生物特征认证是一种至关重要的安全手段，被广泛用于保障金融交易、远程验证身份、控制安全访问，并防范电信系统中的欺诈行为（Li 等，2024年；Kamel 等，2025年）。预计全球声纹认证市场将从2025年的28.7亿美元增长至2032年的156.9亿美元（Al-sheavi 等，2025年）。然而，这种依赖正面临日益加剧的脆弱性。音频深度伪造已迅速从实验室里的新奇事物演变为现实世界中切实存在的安全威胁（Rabhi 等，2024年），造成了重大的社会和经济损失，其中包括冒充公众人物、拨打数百万通电话的人工智能生成的自动拨号电话，以及诈骗老年人、骗取超过20万美元的语音克隆诈骗（Alali 和 Theodorakopoulos，2025年；Mittal 等，2025年；Shapiro，2025年）。这一危机要求我们紧急评估现代防御系统抵御深度伪造技术的能力。

本文通过严格的实证评估，探讨了以下关键问题：**最先进的说话人认证系统能否抵御当前的开源语音克隆（深度伪造）模型**，这些模型仅凭几分钟的目标说话人数据即可合成语音（Li 等，2025年）。我们通过在普通话说话人数据上训练多个具有代表性的语音克隆系统，构建了一个大规模基准，并在多种设置下评估商业说话人验证平台和最先进的反欺骗检测器，以揭示其真实的安全水平和鲁棒性。我们的结果揭示了此前未有文献记录的两种安全漏洞：（1）说话人验证只能对现代语音克隆攻击提供部分防御；（2）反欺骗检测器无法有效泛化到未见过的合成模式。这两项发现都因语音合成技术的快速发展而更加突出，需要行业和学术界主动予以关注。本文的贡献总结如下：

- 我们首次对遭受深度伪造攻击的基于音频的身份验证系统进行了系统性评估，揭示了最先进模型中的关键系统漏洞。
- 我们识别并刻画了系统漏洞背后的关键失效模式，重点阐明了架构局限性、训练数据多样性在泛化中的作用，以及自监督预训练对跨语言迁移能力的影响。
- 我们概述了未来研究的具体方向，强调需要持续更新训练语料库，并设计能够捕捉内在合成特征的架构，从而实现真正有效的防御层。

*通讯作者

模型	开源	数据（分钟）	时间 / 成本
文本转语音			
GPT-SoVITS	✓	0.5 – 2	∼10 分钟
Bert-VITS2	✓	1 – 5	∼2 h
ElevenLabs	×	2 – 30	∼ $0。73
豆包	×	0.5 – 2	∼ $15
阿里云	×	20 – 30	∼ $645
语音转换			
RVC	✓	10 – 30	∼2 h

表 1：现代语音合成系统在开源可用性、所需目标说话人数据以及训练时间或每位说话人成本方面的比较。

2 相关工作

语音克隆。 现代语音克隆系统已取得长足进步，如今只需几分钟的样本语音，不再需要数小时的目标语音数据。表1 总结了主流系统，包括根据文本输入生成语音的文本转语音（TTS），以及将一位说话人的声音转换成另一位说话人声音的语音转换（VC）（Kaur and Singh，2023年）。GPT-SoVITS、Bert-VITS2 和 RVC 等开源模型只需几分钟的目标语音，并可在单张 V100 GPU 上于数小时内完成训练。相比之下，商业模型所需的训练数据更少，成本也合理，大幅降低了恶意使用的门槛；而早期 ASVspoof 时代的攻击则需要大量数据和计算资源（Todisco 等人，2019年）。

音频深度伪造检测。 音频深度伪造检测大致可分为两类：一类是将手工设计特征与分类器结合的流水线检测器，另一类是直接处理原始波形的端到端模型（Li 等人，2025年）。流水线方法通常采用 LFCC、梅尔频率倒谱系数（MFCC）或常数 Q 倒谱系数（CQCC）特征（Todisco 等人，2016年、2018年），而端到端模型则利用原始波形表示（Tak 等人，2021年；Hua 等人，2021年）。近期研究通过自监督学习（SSL）和混合策略提升鲁棒性： Ge 等人 (2025年)提出对 SSL 模型进行后训练，以增强其对未见攻击的泛化能力；而Tahaoglu 等人 (2025年)提出了一种基于 ResNeXt 并结合谱特征的架构，以提高检测可靠性。

鲁棒性与泛化。 现实世界中的部署要求身份认证系统兼具鲁棒性和泛化能力。以往研究探讨了系统的鲁棒性——例如系统对编解码器压缩、传输噪声和混响等环境因素的鲁棒性（Tak 等人，2022年），以及跨数据集泛化能力；在一个语料库上训练的模型往往在另一个语料库上出现显著的性能下降（Wang andYamagishi，2021年）。然而，针对多种合成架构开展的系统性评估仍然有限，这一关键缺口从根本上损害了系统安全性。

3 实验设置

为系统评估语音克隆攻击下身份验证的稳健性，我们提出了一个集成最先进的说话人验证模型、反欺骗检测器和多样化语音合成方法的框架[1]。

3.1 说话人验证模型

我们采用新兴的 ECAPA-TDNN 架构进行说话人验证（Serre 等人，2025年）。该模型以时延神经网络（TDNN）为骨干网络，结合通道级和上下文注意力机制，提取具有判别性的说话人嵌入。我们在 Vox-Celeb（Nagrani 等人，2017年）这一用于说话人识别的大规模数据集上训练系统，该数据集包含超过一百万条语句，时长达 2，000+小时。按照关键应用中的标准做法，我们在开发集上将检测阈值调整为误接受率 0.01%，以符合合规要求（Brydinskyi 等人，2024年）。

3.2 深度伪造检测模型

为检测深度伪造语音，我们采用一种最先进的架构，将 XLS-R（Zhang 等人，2024a）这一在 436k 小时多语言语音上预训练的多语言自监督语音表征模型，与 AASIST（Zhang 等人，2024b；Jung 等人，2022年）这一基于图注意力的欺骗检测器相结合。该系统能够捕捉丰富的语义特征和细粒度的欺骗伪迹，并已在各种基准测试集上展现出经过验证的性能（Yamagishi 等人，2021年；Tran 等人，2025年）。

3.3 基准数据集

我们从 AISHELL-3 中文普通话数据集中随机选取 50 名说话人（25 名男性、25 名女性）（Shi 等人，2021年）。每位说话人的 20 分钟真实语音用于训练三个开源语音合成系统[2]：GPT-

[1]代码和数据集将在录用后发布。[2]项目页面：GPT-SoVITS、Bert-VITS2和RVC。

	来源	# 说话人	总时长
真实	AISHELL-3	50	1000
合成	GPT-SoVITS	50	1000
	Bert-VITS2	50	1000
	RVC	50	1000

表 2：基准数据集概览，时长以分钟计。

合成模型	绕过率	平均相似度
GPT-SoVITS	56.2%	0.598
Bert-VITS2	82.7%	0.679
RVC	43.1%	0.558

表 3：声纹验证绕过率。

SoVITS 和 Bert-VITS2 文本转语音系统，以及 RVC 语音转换系统。每个系统为每位说话人生成 20 分钟的合成语音（见表2）。数据集按说话人划分：30 人用于训练，10 人用于开发，10 人用于测试，确保测试集在模型训练期间完全未被接触。

3.4 评估指标

在说话人验证中，绕过率表示被误分类为目标说话人的攻击所占比例。对于深度伪造检测，性能使用等错误率（EER）进行评估（Reis 等人，2016年），其定义为误接受率与误拒绝率相等时的值。EER 概括了将伪造语音当作真实语音接受，以及将真实语音当作伪造语音拒绝这两者之间的权衡。EER 值越低，区分能力越强；0% 表示性能完美。

4 结果与讨论

4.1 说话人验证漏洞

表3 显示，针对最先进的说话人验证模型，三种语音克隆系统的绕过率都很高，揭示了一项关键漏洞：尽管该系统对真实用户的误接受率极低（FAR = 0.01%），却无法区分高质量的合成语音与说话人的声纹特征之间的高度相似性。所有攻击的平均余弦相似度均超过 0.55，接近真实同一说话人话语通常达到的 0.6–0.8 范围。我们的研究揭示了一个令人担忧的事实：全球数百万用户所依赖的声纹认证系统，只需

模型	等错误率（%）
线性频率倒谱系数 +高斯混合模型 (Todisco 等人，2018年)	12.43
ResNet34 (He 等人，2016年)（声谱图输入）	3.21
RawNet2(Tak 等人，2021年)	2.14
AASIST（独立模型）(Jung 等人，2022年)	1.37
XLS-R + AASIST	**0.83**

表 4：域内测试集上的深度伪造检测模型比较。

10–30 分钟的目标语音就可能遭到攻破，而这些语音很容易从社交媒体、播客或公开演讲中获取。攻击门槛极低，只需一台消费级 GPU，就能在不到 2 小时内训练出所需模型。综合来看，这些结果揭示了严重且亟待应对的安全风险，凸显了迫切需要采取强有力的防御措施，以抵御合成语音攻击。

4.2 域内深度伪造检测

在域内测试集上，训练数据和测试数据均由同一组深度伪造模型生成，XLS-R + AASIST 的等错误率（EER）为 0.83%，显著优于传统方法，展现出其作为身份认证系统中鲁棒层的潜力（见表 4）。然而，我们认为，这样的性能并不能可靠地反映实际鲁棒性。在现实场景中，攻击者可以从广泛的深度伪造模型中进行选择，从而形成域外条件：攻击语音由训练期间未见过其合成模式的模型生成。因此，有必要进一步开展鲁棒性评估。但在现有研究中，这一步往往被忽视，导致部署期间感知到的鲁棒性与真实鲁棒性之间出现偏差。

4.3 鲁棒性分析

4.3.1 对未见语音合成模型的泛化

为了评估模型对未见过的攻击模型的鲁棒性，我们使用**八种前沿文本转语音系统** 生成的语音扩展测试集，这些系统不同于训练中使用的系统。扩展后的测试集包含326条话语（157条真实语音、169条合成语音），主要来源于闭源模型或仅提供演示的模型[3]。这些系统涵盖多种合成范式，包括基于流、基于扩散和提示条件架构，从而能够评估检测器学习到的是可泛化特征，还是特定于模型的特征。

[3]完整的模型列表请参见附录 C。

测试集	等错误率 (%)	性能差距
域内	0.83	-
域外	24.84	29.9×
ASVspoof 2021 LA	3.48	4.2×
ASVspoof 2021 DF	4.59	5.5×

表5：从域内到域外（模型变化）以及跨语言攻击时，深度伪造检测性能的下降情况。

如表 5所示， 30× 性能下降揭示了一项严峻的安全风险：尽管域内准确率近乎完美，当前最先进的端到端检测器主要记住了训练期间出现的攻击特定统计模式。面对不断演进的合成方法，尤其是会引入不同伪影的基于扩散和提示条件系统时，检测器的性能便会崩溃。人工检查还表明，域外检测失败主要集中在以下两种情形：

1. 高保真扩散式文本转语音能够生成自然的韵律、相位相干性和呼吸动态，这些特征与基于生成对抗网络和基于流的架构存在根本差异，从而暴露出检测器未能学习的特征。

2. 依赖基于声码器的训练系统会使检测器偏向于狭窄的频谱线索（例如相位不连续和共振峰失真），从而阻碍其对未见过的合成方法进行泛化。

这一泛化问题带来了严重的现实风险。攻击者可以利用公共训练语料库中未包含的合成方法绕过检测器，钻快速发展的文本转语音创新的空子。新架构每月不断涌现，而重新训练周期却需要数月甚至数年。攻击者还可能结合多个合成阶段，生成训练期间未见过的混合伪迹，这是数据集逐步扩展也无法解决的问题。由于当前检测器关注的是特定于攻击的模式，而不是普适的合成模式，因此应对这一挑战不仅需要持续更新的语料库，更重要的是，需要能够捕捉不变合成特征的模型架构。

4.3.2 跨语言评估

为评估超越训练语言（中文）的跨语言泛化能力，我们在两个大规模英语数据集上测试检测模型，

训练	干净	信噪比=10 dB	汇总
仅使用干净数据	0.83	16.24	8.54
+ RawBoost	**0.53**	**2.55**	**1.53**

表 6：采用数据增强时的噪声鲁棒性。

ASVspoof 2021 LA 和 DF 赛道 （Yamagishi 等人，2021年），且未对检测器重新训练。表5 显示性能仅有小幅下降，表明 XLS-R 多语言前端能够将普通话训练模型有效迁移至英语深度伪造音频，并为检测提供一定程度的鲁棒性。

4.3.3 环境噪声下的鲁棒性

最后，我们评估了环境噪声下的性能。如表 6 所示，当信噪比（SNR）为 10 dB 时，检测性能急剧下降，等错误率从 0.83% 升至 16.24%，凸显了模型对噪声的高度敏感性。使用 RawBoost 以真实环境噪声增强训练数据后，噪声条件下的等错误率降至 2.55%（提升了 6。4× ），而使用汇集后的干净 + 噪声数据集进行训练，则达到 1.53% 的等错误率。这些结果证实，噪声感知训练对于在真实世界中实现鲁棒部署至关重要。

5 结论

本文对基于音频的生物特征认证抵御当代语音克隆攻击的能力进行了系统评估，但遗憾的是，结果并不乐观，揭示了当前防御措施存在的重大漏洞。基于说话人验证的系统很容易被仅使用少量数据集训练的开源深度伪造模型绕过；而深度伪造检测模型尽管在域内性能出色，却无法泛化到未见过的攻击，凸显了高估系统安全性所带来的风险。鉴于认证系统保护着众多高风险领域中的数百万用户，语音合成技术的迅速发展要求我们从根本上转变音频安全的应对方式。未来研究应超越基于特征签名的检测，学习合成过程的不变属性，并认识到仅靠声纹认证不太可能提供可靠保护。结合稳健检测、多因素认证和自适应措施的纵深防御策略，是抵御下一代音频深度伪造的关键保障。

6 局限性

尽管本文进行了全面评估，但仍存在若干局限，也为进一步研究提供了机会：

• **训练数据规模：** 我们的评估聚焦于使用少量目标说话人数据训练的深度伪造模型。尽管现有结果已发出强烈而紧迫的警示，指出其中的安全风险，但研究这些风险如何随可用数据量增加而变化，或许能带来更多见解。这在实际应用中尤其值得关注，因为许多人有大量语音内容可在互联网上公开获取，这可能使攻击更加有效。

•**跨语言评估的复杂性：**我们的研究展示了从中文到英语的跨语言迁移，这两种语言在语音组成、声调结构和音节模式方面存在根本差异。尽管这些实验凸显了多语言自监督表征的潜力，但评估具有多样音系和韵律特征的其他语言，并考虑语码转换语音，或许能进一步揭示模型的泛化能力和鲁棒性，为设计真正语言无关的防御措施提供指导。

参考文献

Abdulazeez Alali 和GeorgeTheodorakopoulos。2025年。利用深度伪造音频实施的真实环境中的部分伪造语音攻击。《网络安全与隐私杂志》，5(1):6。

Amar N Alsheavi、Ammar Hawbani、Wajdy Othman、Xingfu Wang、Gamil Qaid、Liang Zhao、Ahmed Al-Dubai、Liu Zhi、AS Ismail、Rutvij Jhaveri 等。2025年。物联网认证协议：挑战与比较分析。《*ACM*计算调查》，57(5):1–43。

Vitalii Brydinskyi、Yuriy Khoma、Dmytro Sabodashko、Michal Podpora、Volodymyr Khoma、Alexander Konovalov 和 Maryna Kostiak。2024年。用于说话人验证的现代深度学习模型比较。*AppliedSciences*，14(4):1329。

Wanying Ge、Xin Wang、Xuechen Liu 和 Junichi Yamagishi。2025年。用于深度伪造语音检测的后训练。*arXiv*预印本 *arXiv:2506.21090*。

Zhifang Guo、Yichong Leng、Yihan Wu、Sheng Zhao 和 Xu Tan。2023年。PromptTTS：通过文本描述实现可控文本转语音。载于*ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE。

Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页。

Guang Hua、Andrew Beng Jin Teoh 和 Haijian Zhang。2021年。迈向端到端合成语音检测。《*IEEE*信号处理快报》，28：1265–1269。

Zeqian Ju、Yuancheng Wang、Kai Shen、Xu Tan、Detai Xin、Dongchao Yang、Yanqing Liu、Yichong Leng、Kaitao Song、Siliang Tang 等。2024年。Natural-speech 3：利用因式分解编解码器和扩散模型进行零样本语音合成。载于第*41*届国际机器学习大会论文集，第22605–22623页。

Jee-weon Jung、Hee-Soo Heo、Hemlata Tak、Hye-jin Shim、Joon Son Chung、Bong-Jin Lee、Ha-Jin Yu 和 Nicholas Evans。2022年。AASIST：利用集成时频图注意力网络进行音频反欺骗。载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6367–6371页。IEEE。

Kamel Kamel、Keshav Sood、Hridoy Sankar Dutta 和 Sunil Aryal。2025年。关于语音认证和反欺骗系统所面临威胁的综述。预印本，arXiv:2508.16843。

Navdeep Kaur 和 Parminder Singh。2023年。文本转语音合成中采用的传统与当代方法：综述。《人工智能评论》，56(7):5837–5880。

Yichong Leng、Zehua Chen、Junliang Guo、Haohe Liu、Jiawei Chen、Xu Tan、Danilo Mandic、Lei He、Xiangyang Li、Tao Qin 等。2022年。BinauralGrad：用于双耳音频合成的两阶段条件扩散概率模型。神经信息处理系统进展，35:23689–23700。

Yichong Leng、Zhifang Guo、Kai Shen、Zeqian Ju、Xu Tan、Eric Liu、Yufei Liu、Dongchao Yang、Kaitao Song、Lei He 等。2024年。PromptTTS 2：使用文本提示描述和生成声音。载于第十二届国际学习表征会议。

Menglu Li、Yasaman Ahmadiadli 和 Xiao-Ping Zhang。2024年。音频反欺骗检测：综述。*arXiv* 预印本 *arXiv:2404.13914*。

Menglu Li、Yasaman Ahmadiadli 和 Xiao-Ping Zhang。2025年。语音深度伪造检测综述。《*ACM*计算调查》，57(7):1–38。

Govind Mittal、Arthur Jakobsson、Kelly Marshall、Chinmay Hegde 和 Nasir Memon。2025年。Pitch：使用质询-响应对深度伪造音频通话进行人工智能辅助标记。载于第*20*届*ACM*亚洲计算机与通信安全会议论文集，第559–575页。

Arsha Nagrani、Joon Son Chung 和 Andrew Zisser- man。2017年。VoxCeleb：大规模说话人识别数据集。载于*Interspeech 2017*，第2616–2620页。

Mouna Rabhi、Spiridon Bakiras 和 Roberto Di Pietro。2024年。音频深度伪造检测：对抗攻击与应对措施。《*Expert Systems with Applications*》，250:123941。

PauloMax GI Reis、João Paulo CL da Costa、Ricardo K Miranda 和 Giovanni Del Galdo。2016年。利用 ESPRIT 电网频率估计值的峰度进行音频认证。载于 *2016*年第十届信号处理与通信系统国际会议（*ICSPCS*），第 1–6 页。IEEE。

Thomas Serre、Mathieu Fontaine、Éric Benhaim 和 Slim Essid。2025年。用于个性化语音增强中嵌入精炼的对比知识蒸馏。载于 *ICASSP 2025—2025*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 1–5 页。IEEE。

Lauren R Shapiro。2025年。美国针对老年人的网络赋能冒充诈骗。《Security Journal》，38(1):43。

Kai Shen、Zeqian Ju、Xu Tan、Eric Liu、Yichong Leng、Lei He、Tao Qin、Jiang Bian 等。Naturalspeech 2：潜在扩散模型是自然且支持零样本的语音与歌声合成器。载于第十二届国际学习表征会议。

Yao Shi、Hui Bu、Xin Xu、Shaoji Zhang 和 Ming Li。2021年。AISHELL-3：多说话人普通话文本转语音语料库。载于*Interspeech2021*，第 2756–2760 页。

Gul Tahaoglu、Daniele Baracchi、Dasara Shullani、Massimo Iuliani 和 Alessandro Piva。2025年。基于谱特征和 ResNeXt 架构的深度伪造音频检测。《*Knowledge-Based Systems*》，第 113726 页。

Hemlata Tak、Madhu Kamble、Jose Patino、Massimiliano Todisco 和 Nicholas Evans。2022年。RawBoost：一种用于自动说话人验证反欺骗的原始数据增强与扩充方法。载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6382–6386 页。IEEE。

Hemlata Tak、Jose Patino、Massimiliano Todisco、Andreas Nautsch、NicholasEvans 和AnthonyLarcher。2021年。基于原始音频的 RawNet2 端到端反欺骗。载于 *ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6369–6373 页。IEEE。

Xu Tan、Jiawei Chen、Haohe Liu、Jian Cong、Chen Zhang、Yanqing Liu、Xi Wang、Yichong Leng、Yuan-hao Yi、Lei He 等。2024年。Naturalspeech：具有人类级质量的端到端文本转语音合成。*IEEE* 模式分析与机器智能汇刊，46(6):4234–4245。

Massimiliano Todisco、Héctor Delgado 和 Nicholas WD Evans。2016年。一种用于自动说话人验证反欺骗的新特征：恒定Q倒谱系数。载于*Odyssey*，第2016卷，第283–290页。

Massimiliano Todisco、Héctor Delgado、Kong Aik Lee、Md Sahidullah、Nicholas Evans、Tomi Kinnunen 和 Junichi Yamagishi。2018年。集成式呈现攻击检测与自动说话人验证：通用特征与高斯后端融合。载于*Interspeech 2018*-国际语音通信协会第 *19* 届年会。ISCA。

Massimiliano Todisco、Xin Wang、Ville Vestman、Md Sahidullah、Hector Delgado、Andreas Nautsch、Junichi Yamagishi、Nicholas Evans、Tomi Kinnunen 和 Kong Aik Lee。2019年。ASVspoof 2019：欺骗音频与假音频检测的未来展望。载于*Interspeech 2019*，第 1008–1012 页。国际语音通信协会。

Hoan My Tran、Damien Lolive、Aghilas Sini、Arnaud Delhay、Pierre-François Marteau 和 David Guen- nec。2025年。用于音频深度伪造检测的多级自监督学习特征门控。载于第*33*届*ACM*国际多媒体会议论文集，第11766–11775页。

Xin Wang 和 Junichi Yamagishi。2021年。近期用于合成语音检测的神经网络反欺骗对策比较研究。载于*Interspeech 2021*论文集，第4259–4263页。

Yihan Wu、Xu Tan、Bohan Li、Lei He、Sheng Zhao、Ruihua Song、Tao Qin 和 Tie-Yan Liu。2022年。AdaSpeech 4：零样本场景下的自适应文本转语音。载于*Interspeech 2022*论文集，第2568–2572页。

Junichi Yamagishi、Xin Wang、Massimiliano Todisco、Md Sahidullah、Jose Patino、Andreas Nautsch、Xuechen Liu、Kong Aik Lee、Tomi Kinnunen、Nicholas Evans 等。2021年。ASVspoof 2021：加速推进欺骗语音与深度伪造语音检测。*arXiv* 预印本 *arXiv:2109.00537*。

Guangyan Zhang、Kaitao Song、Xu Tan、Daxin Tan、Yuzi Yan、Yanqing Liu、Gang Wang、Wei Zhou、Tao Qin、Tan Lee 等。2022年。混合音素 BERT：利用混合音素和超音素表示改进 BERT，用于文本到语音。*Interspeech 2022*年，第456–460页。

Qishan Zhang、Shuangbing Wen 和 Tao Hu。2024a。使用自监督 XLS-R 和 SLS 分类器进行音频深度伪造检测。载于 第*32*届 *ACM* 国际多媒体会议论文集，第6765–6773页。

Yuxiang Zhang、Jingze Lu、Zengqiang Shang、Wen- chao Wang 和 Pengyuan Zhang。2024b。使用 AASIST2 改进短语音反欺骗。载于 *ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 11636–11640 页。IEEE。

格式	等错误率 (%)
原始音频 (16 kHz)	0.83
WAV（未压缩）	0.91
MP3（245 kbps）	0.84
MP3（100 kbps）	0.94
OGG（160 kbps）	0.84
OGG（256 kbps）	0.86

表 7：编解码器压缩鲁棒性。

A 实施细节

该检测系统使用 AAM-softmax损失训练 100 个周期，批次大小为 32，以增强类间可分性和类内紧凑性；优化器采用 AdamW（初始学习率为 3×10^{-4}）和余弦退火调度，以稳定收敛。输入语音时长为 4 秒，采样率为 16 kHz，并使用 SpecAugment（时间掩蔽最多 40 帧，频率掩蔽最多 4 个频带）和 RawBoost（ISD信噪比 10–40 dB、LnL增益 0–30 dB、SSI信噪比 0–40 dB）进行增强，以提升模型在不同声学条件下的鲁棒性。将梯度裁剪阈值设为 5.0，以防止训练不稳定。使用 $4\times$ V100 GPU 进行训练约需 10 小时；在单张 V100 GPU 上，推理每分钟音频约需 3 秒。训练好的模型和代码将公开发布，以支持高效且可复现的评估。

为确保实现具有鲁棒性并反映实际使用情况，我们在多种音频格式和压缩级别下测试了检测器（表7）。结果显示，性能差异极小（等错误率为 0.83–0.94%），表明 RawBoost 增强能够有效减轻压缩伪影的影响。这说明该系统高度贴近可部署的真实环境，能够确保在电话银行等场景中实现可靠检测，因为音频可能会经过多个编解码阶段。

B 基线深度伪造检测模型

在表4中，我们比较了一系列具有代表性的基线深度伪造检测模型，涵盖音频反欺骗技术的发展历程，并纳入最先进的 XLS-R + AASIST 架构，以全面概述系统性能。下文将进一步说明基线模型的选择及其依据。

基于线性频率倒谱系数和高斯混合模型（LFCC + GMM）的方法采用手工设计的特征和生成式建模，提供了经典的统计基线，广泛应用于早期的 ASVspoof 挑战赛（Todisco 等人，2018年）。以声谱图为输入的 ResNet34 是一种标准的基于卷积神经网络的基线模型，利用残差学习捕捉具有判别性的时频模式（He 等人，2016年）。RawNet2 是一种直接处理原始波形并采用可学习滤波器组的端到端模型，展现了无需显式特征提取即可进行有效数据驱动特征学习的能力（Tak 等人，2021年）。

基于集成频谱-时序图注意力网络的音频反欺骗（AASIST）通过图注意力联合建模频谱-时序欺骗伪迹，对 RawNet2 进行了扩展，代表了一种无需自监督预训练的最先进独立反欺骗方法（Jung 等人，2022年）。与 XLS-R 结合后，该系统融入了多语言自监督表征，在域内测试集上取得了出色的性能。然而，它对未见攻击模型的泛化能力仍然有限，这是一个关键漏洞，在实际部署中会带来显著的安全风险。

C 深度伪造模型

为扩展基准测试集，我们纳入了由以下文本转语音系统生成的语音：

- **AdaSpeech4**(Wu 等人，2022年）：一种零样本自适应文本转语音系统，利用因子化说话人表征为未见过的说话人合成语音，实现了较高的自然度和相似度。

- **BinauralGrad**(Leng 等人，2022年）：一种用于双耳音频合成的两阶段条件扩散模型，可捕捉空间线索，从而生成逼真的空间化声音。

- **MPBert**（Zhang 等人，2022年）：通过整合混合音素和子音素嵌入增强文本转语音效果，从而提升语音音素细节和发音准确度。

- **NaturalSpeech1** (Tan 等人，2024年）：一种端到端文本转语音模型，通过联合建模整个合成流程，实现了媲美人类的质量和表现力。

- **NaturalSpeech 2** （Shen 等）：利用神经音频编解码器和潜在扩散技术生成自然语音和歌声，并支持为新说话人进行零样本生成。

• **NaturalSpeech 3**（Ju 等人，2024年）：利用神经编解码器和扩散模型，将语音分解为内容、韵律、音色和声学子空间，从而提升零样本质量和说话人相似度。

• **PromptTTS 1** (Guo 等人，2023年）：根据描述内容和风格的自然语言提示生成语音，从而实现可控且灵活的文本转语音。

• **PromptTTS 2** (Leng 等人，2024年）：在 PromptTTS 的基础上，采用提示驱动的风格建模并增强可变性，生成一致且富有表现力的合成语音。

这些模型总体涵盖了多种合成范式，包括自适应方法、基于扩散的方法和提示条件方法，确保域外评估能够捕捉现代语音合成中真实且具有挑战性的变化，同时揭示其真正的安全鲁棒性。