说话人感知反欺骗

Xuechen Liu[1,2], *Md Sahidullah*[3],*Kong Aik Lee*[4],*Tomi Kinnunen*[1]

[1]东芬兰大学计算学院，约恩苏，芬兰[2]洛林大学、法国国家科学研究中心、法国国家信息与自动化研究所、洛林计算机科学与应用研究实验室，F-54000，南锡，法国〞智能发展研究院、TCG CREST，印度[3]资讯通信研究院，A[4]STAR，新加坡

xuechen.liu@uef.fi, sahidullahmd@gmail.com, lee kong aik@i2r.a-star.edu.sg, tkinnu@cs.uef.fi

arXiv:2303.01126v2 [cs.SD] 8 Jun 2023

摘要

我们研究了说话人感知反欺骗，即将预先获知的目标说话人信息纳入语音欺骗对策系统（反欺骗对策）。不同于常用的说话人无关解决方案，我们以说话人条件化的方式训练反欺骗对策系统。作为概念验证，我们针对最先进的 AASIST（使用集成频谱—时间图注意力网络的音频反欺骗）模型，研究其说话人感知扩展。为此，我们分别考虑在帧级和话语级纳入目标说话人信息的两种替代策略。基于 ASVspoof 2019 数据集所构建的自定义协议进行的实验结果表明，通过注册引入说话人信息是有效的：与说话人无关基线相比，等错误率（EER）和最小串联检测代价函数（t-DCF）分别取得了最高 25.1% 和 11.6% 的相对提升。

索引词：说话人验证、说话人感知反欺骗、ASVspoof、深度伪造、欺骗对策。

1. 引言

得益于近年来对原始语音波形进行神经声码器合成的进展 [1], , 现代文本转语音（TTS）能够灵活地生成听起来如同自然人类语音的人工语音 [2]–[4]。结合通过神经说话人嵌入 [5], [6]提取说话人信息并用于控制波形生成 [7], [8], 的并行进展，现代 TTS 原则上能够让目标人物以其声音“替任何人说话”。

尽管这类技术有诸多实用用途，其灵活性也引发了显而易见的担忧。首先，在生物特征认证领域，对手（攻击者）冒充他人（目标）来欺骗自动说话人验证（ASV）的可能性早已为人所知 [9]。其次，深度伪造的潜在负面影响——它结合了“深度学习”与“伪造”，并以对抗性机器学习为基础 [10]——近来已引起研究人员 [11] 和公众 [12]的关注。我们已经见到一些令人警觉的案例 [13], , 尽管与基于图像和视频的深度伪造相比，与语音相关的深度伪造受到的关注较少。用于恶意目的的深度伪造不仅会损害目标人物的声誉，还可能削弱公众对视听媒体和生物特征技术的普遍信任。要维护这种信任，就需要采取新型防护措施。

值得庆幸的是，人们很早就主动认识到区分“真实”输入与“伪造”输入的重要性——远早于“对抗性机器学习”或“深度伪造”概念的出现。尤其是，生物识别研究界一直在研究各种反欺骗方法，以保护生物识别系统，至今已有二十多年 [14]。呈现攻击检测（PAD）系统 [15], 也称为反欺骗对策（CM），指旨在检测欺骗输入的方法。

在本研究中，“反欺骗对策”指的是一种接收语音输入并输出二元真实/欺骗预测的分类器。自 2015 年以来，ASVspoof挑战系列 [16]一直引领着采用通用数据和性能指标 [17]对语音反欺骗对策进行基准测试。尽管名称如此，ASVspoof挑战聚焦于独立式、说话人无关的反欺骗对策，这些对策可与自动说话人验证系统或其他应用集成。得益于 ASVspoof挑战及其他近期类似项目 [18], [19], 提供的通用数据，已有多种独立式语音反欺骗对策问世，涵盖早期统计方法 [20], [21] 到近期的深度架构 [22]–[24]。

遗憾的是，现有的大多数反欺骗对策远非完美，尤其是在面对未知情况时——无论是未见过的声码器、文本转语音系统、数据域还是编解码器 [25]。独立式说话人无关反欺骗对策任务不受约束的性质，加之人工语音已难以与真实语音区分，使得反欺骗对策难以泛化至训练数据之外。要实现完全通用、说话人无关的反欺骗对策，就必须在先验知识有限的情况下，补偿说话人、内容和信道变化可能带来的混杂效应。

即使 ASVspoof 等挑战中未涉及这一点，在许多应用场景中，我们确实掌握目标人物的先验信息，而反欺骗对策可以直接加以利用：欺骗攻击通常针对特定个人——也就是同一位我们试图验证其身份、且自动说话人验证系统已根据此前收集的注册数据“认识”的个人。对于针对政治人物、新闻主播等公众人物的深度伪造，我们同样可以较为放心地假定，知道潜在深度伪造样本中被冒充的目标是谁。因此，在测试时向反欺骗对策提供基于注册样本所推断的说话人身份，似乎十分合理。为此，我们首次研究了在反欺骗中使用目标说话人信息的方法，并将其称为说话人感知反欺骗。

我们的研究并非首次探索这一总体思路。据作者所知，此前有两项相关研究 [26], [27] 分别聚焦于采用高斯混合模型（GMM）后端进行重放攻击检测 [26] ，或改进自动说话人验证系统的后端 [27]。在反欺骗对策方案（统计模型 [26]与深度学习）、假数据类型（重放攻击 [26]与合成媒体）、实验设置，以及协议设计和指标方面的评估上，我们的工作与这两项研究存在显著差异。我们工作的主要创新在于对说话人感知反欺骗问题提出了精确的表述。此外，我们还比较了将不同目标说话人信息整合到先进 AASIST 模型中的多种方案 [26], [27] 。具体而言，如图1所示，并将在下一节详细介绍，我们采用深度说话人嵌入来表示这些信息，并将其整合到 AASIST 的不同部分。

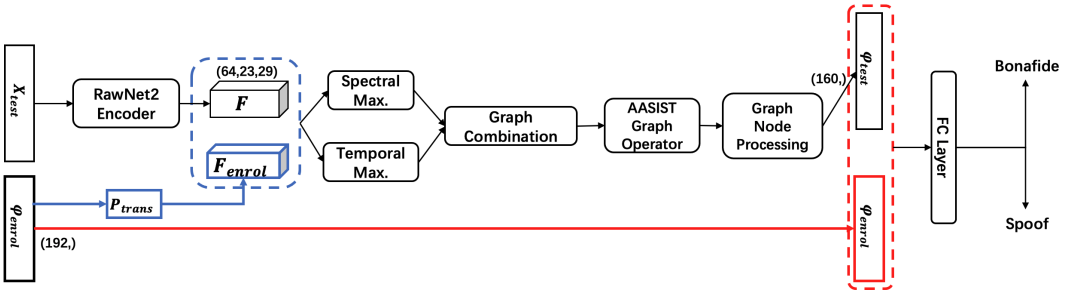

图1：通过注册向量集成说话人信息的示意图。蓝线和红线对应不同的方法，这些方法不会同时应用。虚线表示辅助附加操作。彩色效果最佳。

将目标说话人信息整合到先进的 AASIST 模型中的方式 [23]。具体而言，这些信息通过深度说话人嵌入来表示，并整合到 AASIST 的不同部分，如图1所示，详情见下一节。

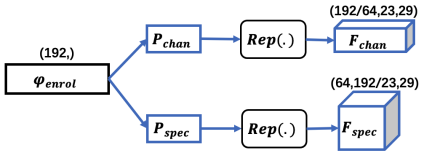

图2：将说话人注册向量转换为可按通道或按频谱附加的组成部分的示意图。$\mathbf{F}_{enr}$ 在图*1*中可以是$\mathbf{F}_{chan}$ 或$\mathbf{F}^{spec}$。变换矩阵$\mathbf{P}_{trans}$ 相应地可以是$\mathbf{P}_{chan}$ 或$\mathbf{P}^{spec}$。$Rep(.)$ 表示对其他两个维度执行重复操作。彩色效果最佳。

2. 对策系统的说话人感知训练

2.1. 问题定义

我们首先将说话人感知反欺骗问题定义为**二元分类任务，即在已注册说话人的条件下，区分真实语音和伪造语音**。更具体地说，它是一个定义如下的条件假设检验：

- H_0：测试样本是真实语音，且对应于目标说话人。
- H_1：测试样本是欺骗语音，且对应于目标说话人。

在实践中，我们通过引入目标说话人的额外真实语音来处理这一任务，具体内容如下。值得注意的是，这两个假设均以目标说话人为条件，因此不同于传统的反欺骗定义。

2.2. 说话人感知反欺骗

所提的说话人感知 CM 如图1所示。说话人信息可以通过多种方式表示，从原始音频到成熟的深度说话人嵌入均可。在本研究中，我们关注后者。我们将注册音频数据输入预训练的自动说话人验证模型（此处为 ECAPA-TDNN [28]）。对于每位注册说话人，我们相应地提取说话人嵌入，并对其取平均，得到一个注册嵌入：

$\varphi_{\text{enrol}} = \frac{1}{N}\sum_{i=1}^{N}\varphi_i$，其中 N 是注册时可用的话语数量。

在本研究中，我们采用近期提出的 AASIST [29]。该方法由以下部分组成：基于 RawNet2 [30] 的语音编码器；分别在频谱轴和时间轴上运行的两个异构图注意力层；以及一个图池化层。池化层之后连接节点堆叠层和用于二元决策的全连接（FC）层。如图1中的蓝线所示，我们提出将注册嵌入 φ_{enrol}融入训练过程，将其视为辅助条件信息。下文将介绍所提方法及其简称。

在编码器输出处进行集成：首先，我们关注编码器的输出，它是一个具有通道、频谱和时间轴的三维特征图。将该特征图的形状记为（d_{c}，d_{s}，d_{t}）。受早期通道级扩展 [31]–[33] 以及基于频谱特征的说话人自适应 [34], 的启发，我们按照图2所示，在通道级或频谱级扩展嵌入向量 $\mathbf{F}_{\text{enr}}$。因此，图1中的 $\mathbf{F}_{\text{enr}}$ 可以是$\mathbf{F}_{\text{chan}}$ 或$\mathbf{F}^{\text{spec}}$。$\mathbf{F}_{\text{chan}}$ 的形状为（d_{embed}，d_{s}，d_{t}），而$\mathbf{F}_{\text{chan}}$ 的形状为（d_{c}，d_{embed}，d_{t}）。其中，d_{embed} 是注册向量的维度。在通道轴或频谱轴上进行附加的这两种方法分别记为***enc-chan*** 和***enc-spec***。

在编码器输出处进行集成并降维：由于嵌入向量与原始特征图的维度分别不同（d_{embed} 与 d_{s} 或 d_{t}），其中一者可能会对模型预测产生更大的潜在影响。因此，作为替代方案，我们考虑引入变换矩阵$\mathbf{P}_{\text{trans}}$ 进行降维，如图1所示。根据通道级和频谱级的附加方式，$\mathbf{P}_{\text{trans}}$ 可以分别为$\mathbf{P}_{\text{chan}}$ 或$\mathbf{P}_{\text{spec}}$，如图2所示。当注册向量首先被降至 d_{c} 或 d_{s}时，$\mathbf{P}_{\text{trans}}$ 采用正态分布进行初始化，并与AASIST中的其他可学习组件联合优化，其形状分别为（d_{embed}，d_{c}）（对于$\mathbf{P}_{\text{chan}}$）或（d_{embed}，d_{s}）（对于$\mathbf{P}_{\text{spec}}$）；当不进行降维时，$\mathbf{P}_{\text{trans}}$ 为形状为（d_{embed}，d_{embed}）的单位矩阵。我们在所得特征图后添加后缀*-reduced*进行表示，因此相应的方法分别为***enc-chan-reduced*** 和***enc-spec-reduced***。

在全连接层输入处进行融合：到目前为止，我们已经介绍了在 AASIST 的早期层中使用注册嵌入。作为另一种策略，我们还考虑在全连接层之前进行融合，如图1所示。决策前全连接层的输入是一个话语级的 160 维向量，记作 φ_{test}。先前的研究 [35]已提取该向量，以便与自动说话人验证系统进行联合优化。这里，我们只需将 φ_{enrol} 拼接到 φ_{test}，此时

全连接层的输入维度为 $d_{\text{embed}} + 160$。在第4节展示结果时，我们将这种拼接方式称为话语 。

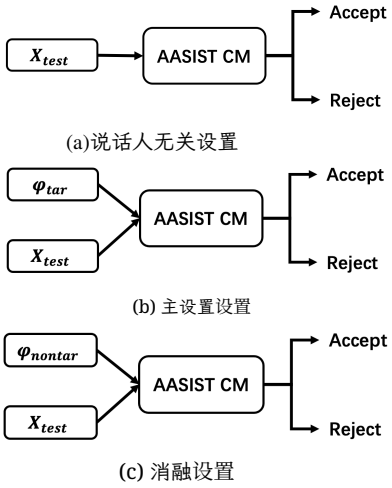

图3： （a）传统说话人无关反欺骗、（b）说话人感知反欺骗（主设置）和（c）消融研究设置（消融）的概念示意图。φ_{tar} 表示来自与输入音频$\mathbf{X}_{test}$中同一说话人的 φ_{enrol} 。φ_{nontar} 表示来自与$\mathbf{X}_{test}$中不同说话人的 φ_{enrol} 。主设置符合目标说话人已知这一假设，而消融设置用于评估违背该假设对反欺骗对策性能的影响。

表 1： 完整反欺骗对策和定制协议可用的开发集/评估集试验数量。

设置	dev	eval
原始设置 [36]	24844	71237
主设置 （我们的方法）	23780	69252

3. 实验设置

数据。我们基于 ASVspoof 2019 LA [36], 开展实验，该数据集涵盖 19 种欺骗攻击。反欺骗对策训练数据包含 20 名说话人（9 名男性，11 名女性），涵盖 6 种攻击（A01-A06）以及真实语音条件。反欺骗对策评估数据包含另外 13 种攻击（A07-A19）。训练反欺骗对策模型时，对于 20 名说话人中的每一位，我们通过对真实语音条件下随机选取的 N 条话语的说话人嵌入取平均，生成 φ_{enrol} 。我们为女性设置 $N = 11$ ，为男性设置 $N = 19$ 。

协议。图 3b 展示了该协议的设计逻辑，并与图 3a 所示的常规反欺骗设置进行了比较。我们再次说明我们的假设：测试话语源自已知的目标说话人，任务是确定该样本是真实语音还是伪造语音。因此，对于每个测试话语，其用于注册的关联说话人嵌入 φ_{enrol} 均来自同一目标说话人。在这种情况下，我们的评估协议基于原始的 ASVspoof 2019 反欺骗对策协议（表 1 中的“原始设置”），并移除了数据集中没有对应目标说话人的真实语音试验。我们使用原始元数据中 ASV 协议文件所提供的说话人信息。协议统计数据见表 1。我们将此协议设置称为主设置，它不同于第 4.2 节所述的消融设置。

表2： 以合并等错误率和最小化检测代价函数 串联检测代价函数。

	Main		消融	
方法	等错误率(%)	tDCF	等错误率(%)	tDCF
(基线)	1.51	0.043	1.51	0.043
enc-chan	1.48	0.049	1.80	0.061
enc-chan-reduced	2.27	0.077	2.20	0.073
enc-spec	**1.13**	**0.038**	**1.47**	**0.049**
enc-spec-reduced	1.65	0.055	1.88	0.061
话语	1.89	0.059	1.78	0.052

模型。对于 AASIST 模型，我们采用开源代码库中的方案，将其作为与说话人无关的基线[1]。模型训练和超参数设置遵循 [29], 中所述的配置，唯一例外是由于计算资源有限（使用单个 NVIDIA GeForce GTX 2080Ti），批量大小从 24 减少到 12。特征图的原始形状为（d_c， d_s， d_t）=（64， 23， 29），如图 1 所示。我们使用开源的预训练 ECAPA-TDNN[2] [28]作为预训练的自动说话人验证模型，通过 $d_{\text{embed}} = 192$提取说话人嵌入。对于每个输入句子，每个说话人嵌入均从池化层之后的第一个全连接层中提取。

评估。我们报告等错误率 (EER)和最小串联检测代价函数 (tDCF) [17]。由于与最小化检测代价函数相比，EER 更能反映 CM 性能本身 [17], [29], ，因此我们的结果分析主要基于 EER，包括针对各类攻击的分析。

4. 结果与分析

4.1. 结果

表2展示了汇总等错误率和串联检测代价函数方面的结果。尽管未经降维的通道级说话人整合仅略微改善了等错误率，但频谱级整合表现良好，在两项指标上均取得最低数值；与基线相比，等错误率和最小化检测代价函数分别相对降低了25.1%和11.6%。这表明，将目标说话人整合方法应用于频谱特征图是有效的。相比之下，通道级整合的表现相对较差，这或许是因为原始音频为单通道且噪声水平较低。降维会降低两种方法的反欺骗对策性能。将注册向量与瓶颈嵌入一起接入全连接层之前，并未带来性能提升。

表3按攻击类型详细列出了基线以及表现最佳的说话人感知反欺骗方法（按协议划分）的结果。除了前四行所示的反欺骗对策结果外，该表还列出了自动说话人验证系统在完整反欺骗对策协议上的等错误率。这些数值用于说明各类攻击欺骗自动说话人验证系统的有效程度（但不应与反欺骗对策结果进行比较）。从自动说话人验证等错误率可以看出，有些攻击无法有效欺骗自动说话人验证模型，例如 A09、A17 和 A18，这意味着这些算法未能充分建模说话人信息。

回到反欺骗对策（CM）性能，在五种攻击（A08、A09、A16、A17、A18）中，表现最佳的所提方法在两种协议下均优于（或达到与之相近的性能）基线。这些攻击中有四种的自动说话人验证等错误率（ASV EER）相对较低（低于 20%），但

[1]https://github.com/clovaai/aasist[2] https://github.com/TaoRuijie/ECAPA-TDNN/

表3： 基线系统和表现最佳系统在各类攻击下的等错误率（%）。根据 *[36]*，粗体标出的欺骗攻击表示在开发过程中获取了说话人信息。自动说话人验证等错误率由与本研究所用模型相同的预训练 *ECAPA-TDNN* 模型计算得出，详见 *[35]*。

	方法	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19
Main	（基线）	0.75	0.19	0.02	0.88	0.37	0.72	0.14	0.15	0.47	0.73	2.15	4.80	0.78
	enc-spec	1.18	**0.07**	**0.00**	1.38	0.41	0.98	0.22	0.28	0.98	**0.65**	**1.28**	**2.70**	**0.34**
消融	*enc-spec*	1.57	**0.08**	**0.00**	1.95	0.47	1.26	0.28	0.35	1.30	**0.71**	**1.79**	**3.13**	**0.30**
自动说话人验证等错误率 [35]		32.66	18.80	2.20	50.61	47.08	39.56	11.62	35.39	36.54	60.71	1.85	2.38	4.77

A16 除外；这表明，所提说话人信息集成方法进一步利用了欺骗算法无法良好编码目标说话人这一弱点。在 A16 上也可以观察到改进，该攻击对应于所有欺骗算法中最高的自动说话人验证等错误率。这可能利用了所提算法对能够良好建模说话人信息的强攻击的补偿能力。未来的工作可以进一步研究欺骗算法的说话人信息建模能力与通过此类集成由反欺骗对策（CM）进行补偿之间的关系。

4.2. 消融研究：说话人身份指定错误

上述评估设置和结果基于这样一个假设：输入音频来自目标说话人。由此自然引出一个问题：如果这一假设不成立，会发生什么？——也就是说，面对注册话语与测试话语之间说话人身份不匹配所导致的模型指定错误，反欺骗对策的鲁棒性如何。为此，在本消融研究中，我们假设真实输入音频并非来自其对应的说话人。在这种情况下，我们保留与主协议完全相同的测试话语，但将相应的注册话语替换为从另一位随机选定的说话人那里随机选取的一条注册话语。图3c展示了这一消融设置。

表2展示了所提方法在此设置下的整体结果。对于大多数所提方法，与主设置相比，两项指标的结果均有所下降，但降幅不大。表现最佳的*enc- spec* 的等错误率相对上升了25.1%，但即使在严重的模型——情况下，其准确率仍保持在说话人无关基线的水平。

建模假设设定错误／严重违背建模假设。对于 *enc-chan-reduced* 和话语，结果仍处于大致相同的水平。此设置下*enc-spec* 的各攻击结果如表3所示。对于在主设置下观察到性能提升的六种攻击，*enc-spec* 仍优于基线，但除 A09 和 A19 外，性能较主设置 略有下降。虽然这种下降表明，相较于来自另一位说话人的信息，目标说话人信息更有用，但非目标说话人信息的潜力仍值得进一步研究，并拓展至其他场景。

4.3. 消融研究：额外真实语音训练数据

注册向量不仅是说话人表征，也是额外真实语音信息的载体。说话人信息和真实语音信息都可作为训练反欺骗系统的先验条件。因此，我们研究了额外真实语音训练数据的影响。

我们通过汇集来自不同数据集的额外语音数据，在完整的反欺骗对策协议下实现了这一扩展。我们考虑了 VoxCeleb [37] 和 LibriSpeech [38] 语料库。对于每个数据集，我们改变用于反欺骗对策训练的话语数量。

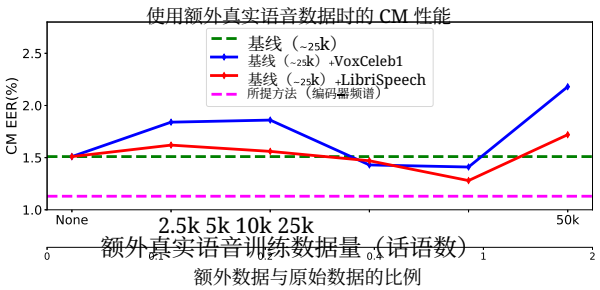

图 4： 不同常见语音处理数据集提供的额外数据与主设置下反欺骗对策性能之间的关系。绿色虚线表示基线性能，粉色线表示性能最佳的系统。建议在彩色模式下查看。

结果如图 4 所示，同时展示了基线和性能最佳的说话人感知 CM。图中揭示了两个有趣的现象。首先，来自 VoxCeleb1 或 LibriSpeech 的更多额外纯真实语音数据能够提升性能。其次，增加 25k 条额外话语后，基线性能有所提升；此时新增数据量几乎等于反欺骗对策训练数据总量（25380 条话语 [36]）。然而，增加更多数据不一定会带来更好的性能。结合上述结果，这表明额外的说话人信息带来的益处更为显著，但这也可能是因为 ASVspoof 数据集源自*VCTK*[3]，，这是一个在消声室中录制的非常干净的数据集。未来的工作可以进一步研究这一问题。

5. 结论

我们研究了采用最先进的 AASIST 欺骗对策进行合成欺骗攻击检测时，**说话人感知反欺骗**的可行性。研究结果表明，将目标说话人注册嵌入作为辅助信息，有助于使反欺骗等错误率相对降低，最高可达 25.1%。关于替代说话人信息的作用，以及利用辅助语料库扩充真实语音训练数据的进一步实验表明，所提出的说话人感知训练策略可能更为有效。本研究证实了两项早期研究中类似的发现——尽管它们采用了完全不同的分类器和数据集 [26], [27], ，并进一步为目标说话人先验信息的积极作用提供了证据。未来的工作可着眼于使用孪生网络编码说话人信息，并在 CM 模块训练期间提供这些信息，同时采用更先进的队列模型对说话人信息进行编码。

[3]https://datashare.is.ed.ac.uk/handle/10283/ 2651

6. 参考文献

[1] A. van den Oord、S. Dieleman、H. Zen 等，“Wavenet：一种用于原始音频的生成模型”，载于第九届 ISCA语音合成研讨会，2016 年，第 125–125 页。[2] Y. Jia、Y. Zhang、R. Weiss 等，“从说话人验证到多说话人文本转语音合成的迁移学习”，《神经信息处理系统进展》，第 31 卷，2018 年。[3] A. Oord、Y. Li、I. Babuschkin 等，“Parallel Wavenet：快速高保真语音合成”，载于国际机器学习大会，PMLR，2018 年，第 3918–3926 页。[4] J. Shen、R. Pang、R. J. Weiss 等，“通过以梅尔频谱图预测为条件的 Wavenet 实现自然 TTS 合成”，载于 ICASSP 论文集，IEEE，2018 年，第 4779–4783 页。[5] B. Desplanques、J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：基于 TDNN 的说话人验证中的强调通道注意力、传播与聚合”，载于 Interspeech 论文集，2020 年，第 3830–3834 页。[6] D. Snyder、D. Garcia-Romero、G. Sell 等，“X向量：用于说话人识别的鲁棒深度神经网络嵌入”，载于 ICASSP 论文集，IEEE，2018 年，第 5329–5333 页。[7] S. Li、B. Ouyang、L. Li 等，“Light-tts：轻量级多说话人、多语言文本转语音”，载于 ICASSP 论文集，IEEE，2021 年，第 8383–8387 页。[8] M. K. Baskar、L. Burget、S. Watanabe 等，“EAT：用于自监督语音识别的增强型自动语音识别-文本转语音”，载于 ICASSP 论文集，IEEE，2021 年，第 6753–6757 页。[9] X. Wang、J. Yamagishi、M. Todisco 等，“ASVspoof 2019：合成、转换和重放语音的大规模公开数据库”，《计算机语音与语言》，第 64 卷，第 101 114 页，2020 年。[10] I. Goodfellow、J. Pouget-Abadie、M. Mirza 等，“生成对抗网络”，载于《神经信息处理系统进展》，Z. Ghahramani、M. Welling、C. Cortes 等编，第 27 卷，Curran Associates, Inc.，2014 年。[11] Y. Mirsky 和 W. Lee，“深度伪造的制作与检测：综述”，《ACM计算调查》（CSUR），第 54 卷，第 1 期，第 1–41 页，2021 年。[12] N. Schick，《深度伪造：即将到来的信息末日》。Hachette UK，2020 年。[13] 今年席卷互联网的五大深度伪造诈骗，https : / / www . analyticsinsight . net / top - 5 - deepfake-scams-that-stormed-the-internet- this-year/，[在线；访问日期：10 月 10 日，2022], 2022。[14] N. K. Ratha、J. H. Connell 和 R. M. Bolle，“增强基于生物特征的身份验证系统的安全性与隐私性”，《IBM系统杂志》，第 40 卷，第 3 期，第 614–634 页，2001 年。[15] “ISO/IEC 30107。信息技术——生物特征呈现攻击检测”，标准，2016 年。[16] Z. Wu、J. Yamagishi、T. Kinnunen 等，“ASVspoof：自动说话人验证欺骗与对策挑战赛”，《IEEE信号处理选题期刊》，第 11 卷，第 4 期，第 588–604 页，2017 年。[17] T. Kinnunen、H. Delgado、N. Evans 等，“欺骗对策与自动说话人验证的串联评估：基础”，《IEEE/ACM音频、语音与语言处理汇刊》，第 28 卷，第 2195–2210 页，2020 年。[18] J. Yi、R. Fu、J. Tao 等，“ADD 2022：首届音频深度合成检测挑战赛”，载于 ICASSP 论文集，IEEE，2022 年，第 9216–9220 页。[19] Z. Zhang、Y. Gu、X. Yi 等，“FMFCC-a：一个具有挑战性的普通话合成语音检测数据集”，载于国际数字水印研讨会，Springer，2021 年，第 117–131 页。

[20] M. Todisco、H. Delgado 和 N. Evans，“恒定 Q 倒谱系数：一种用于自动说话人验证的欺骗对策”，《计算机语音与语言》，第 45 卷，535，2017 年 516 页。–。[21] M. Sahidullah、H. Delgado、M. Todisco 等，“集成欺骗对策与自动说话人验证：在 ASVspoof 2015 上的评估 ”，载于 Interspeech 会议论文集，2016 年，第 1700–1704 页。[22] X. Wang 和 J. Yamagishi，“用于合成语音检测的近期神经网络欺骗对策比较研究”，载于 Interspeech 会议论文集，2021 年，第 4259–4263 页。[23] J.-w. Jung、H.-S. Heo、H. Tak 等，“AASIST：使用集成时频图注意力网络的音频反欺骗”，载于 ICASSP 会议论文集，IEEE，2022 年，第 6367–6371 页。[24] Z. Teng、Q. Fu、J. White 等，“ARawNet：利用原始波形进行欺骗语音检测的轻量级方案”，载于国际模式识别会议（ICPR），2022 年，第 692–698 页。[25] T. Chen、A. Kumar、P. Nagarsheth 等，“音频深度伪造检测的泛化性”，载于 Odyssey 2020 说话人与语言识别研讨会会议论文集，2020 年，第 132–137 页。[26] G. Suthokumar、K. Sriskandaraja、V. Sethu 等，“重放检测中的说话人相关模型分析”，《APSIPA信号与信息处理学报》，第 9 卷，2020 年。[27] D. Castan、M. H. Rahman、S. Bakst 等，“针对说话人的合成语音检测”，载于说话人与语言识别研讨会（Odyssey 2022）会议论文集，2022 年，第 62–69 页。[28] B. Desplanques、J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：基于 TDNN 的说话人验证中的强化通道注意力、传播与聚合”，载于 Interspeech 会议论文集，2020 年，第 3830–3834 页。[29] J.-w. Jung、H.-S. Heo、H. Tak 等，“AASIST：使用集成时频图注意力网络的音频反欺骗”，载于 ICASSP 会议论文集，2022 年，第 6367–6371 页。[30] J.-w. Jung、S.-b. Kim、H.-j. Shim 等，“使用特征图缩放和原始波形改进 RawNet 的文本无关说话人验证”，载于 Interspeech 会议论文集，2020 年，第 1496–1500 页。[31] D. Cai、X. Qin 和 M. Li，“混响和噪声环境下端到端说话人识别的多通道训练”，载于 Interspeech 会议论文集，2019 年，第 4365–4369 页。[32] P. Swietojanski、A. Ghoshal 和 S. Renals，“用于远场语音识别的卷积神经网络”，《IEEE信号处理快报》，第 21 卷，第 9 期，第 1120–1124 页，2014 年。[33] V. Lostanlen、J. Salamon、M. Cartwright 等，“逐通道能量归一化：原因与方法”，《IEEE信号处理快报》，第 26 卷，第 1 期，第 39–43 页，2019 年。[34] V. Gupta、P. Kenny、P. Ouellet 等，“基于 i-vector 的深度神经网络说话人自适应用于法语广播音频转录”，载于 ICASSP 会议论文集，2014 年，第 6334–6338 页。[35] Y. Zhang、G. Zhu 和 Z. Duan，“一种面向欺骗感知说话人验证的概率融合框架”，载于说话人与语言识别研讨会（Odyssey 2022）会议论文集，2022 年，第 77–84 页。[36] A. Nautsch、X. Wang、N. Evans 等，“ASVspoof 2019：用于检测合成、转换和重放语音的欺骗对策”，《IEEE生物识别、行为与身份科学汇刊》，第 3 卷，第 2 期，第 252–265 页，2021 年。[37] J. Chung、A. Nagrani、E. Coto 等，“VoxSRC 2019：首届 VoxCeleb 说话人识别挑战赛”，载于 ISCA档案库，2019 年。[38] V. Panayotov、G. Chen、D. Povey 等，“LibriSpeech：基于公版有声读物音频的自动语音识别语料库”，载于 ICASSP 会议论文集，2015 年，第 5206–5210 页。