多通道重放语音检测的声学仿真框架

Michael Neri, Tuomas Virtanen

Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
{michael.neri, tuomas.virtanen}@tuni.fi

ABSTRACT

重放语音攻击对语音控制的系统构成重大威胁，尤其是在广泛部署了语音助手的智能环境中。虽然多通道音频提供了可以增强回放检测鲁棒性的空间线索，但现有的数据集和方法主要依赖于单通道录音。在这项工作中，我们引入了一个声学仿真框架，该框架设计用于使用公开可用的资源来模拟多通道重放语音配置。我们的设置在不同的环境中建模真实的和伪造的语音，包括现实的麦克风和扬声器脉冲响应、房间声学以及噪声条件。该框架在回放攻击期间采用测量的扬声器方向性以提高仿真的真实性。我们定义了两种模拟设置，模拟在重放场景中使用混响语音还是无回声语音，并评估全向和扩散噪声对检测性能的影响。使用最先进的 M-ALRAD 模型进行重放语音检测，我们展示了合成数据可以支持检测器在未见过的环境中的泛化能力。

Index Terms— 重放攻击，物理访问，空间音频，语音欺骗，房间声学模拟

1. 介绍

cro:VAvoice assistants (VAs) 在人机交互中的使用越来越普遍，利用语音作为生物识别标识符 [1]。它们使用户能够通过 cro:IoTInternet of things (IoT) 网络控制智能设备并在线交换敏感信息。这种日益依赖使得基于音频的系统成为利用语音录音漏洞进行攻击的目标，并且攻击针对的是 cro:ASVautomatic speaker verification (ASV) [2]。cro:LALogical access (LA) 攻击使用 cro:TTStext-to-speech (TTS) 和 cro:VCvoice conversion (VC) 技术模仿说话人，而 cro:DFdeepfake (DF) 方法通过压缩和量化进一步模糊了这些特征。相比之下，cro:PAphysical access (PA) 攻击在麦克风级别欺骗 ASV 系统 [3]，包括冒充攻击 [1] 和重放攻击 [4]，其中播放录音以获得未经授权的访问。

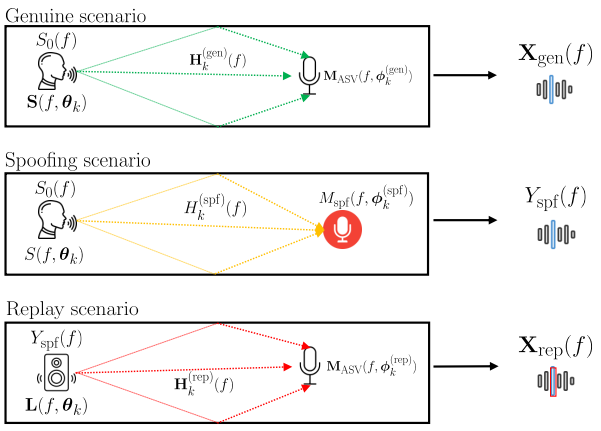

Fig. 1. 声学仿真框架在重放语音检测中的概述。

先前的研究为应对重放攻击主要依赖于单通道数据集，如 RedDots [3] 和 ASVSpoofPA 系列 [2,5,6]。这些资源支持了基于 cro:DNNdeep neural network (DNN) 架构 [7,8] 的重放检测模型的开发以及手工制作的时间频率特征 [9,10]。然而，麦克风阵列通常被整合到 ASV 系统中，以通过利用空间信息来改进语音增强和分离等任务 [11]。除了音频质量之外，多通道数据为回放攻击检测提供了关键优势。具体来说，(i) 多通道录音中的空间线索支持更稳健的检测 [4,12]，以及 (ii) 这些空间特征对于攻击者而言难以复制，不同于单通道数据中的时间或频谱线索 [13]。然而，大多数现有的回放检测器依赖于单通道输入，并未能利用空间信息 [4,14,15]。这一问题由于 cro:ReMASCRealistic Replay Attack Microphone Array Speech (ReMASC) 数据集 [12] 是唯一包含来自四个麦克风阵列和四个环境的真实录音的回放语音数据集而进一步加剧。事实上，硬件成本、同步与校准的需求、大量存储需求以及在各种环境中重现一致的空间设置难度使得收集多通道及空间音频数据集变得困难。此外，由于获取精确地面真实空间标签、维护隐私及实现数据多样性的工作强度较大，因此一般化仅限于未见过的情况。因此，在回放语音任务中评估每个单一组件的影响颇具挑战性。

为了解决上述问题，本工作的贡献如下：(i) 我们提出了一种声学模拟框架，可以生成多通道录音，使得能够对影响重放语音检测任务的不同因素进行受控实验，(ii) 我们模拟了两种欺骗配置，其中攻击者是否可以获得无回声语音。此外，我们分析了噪声对重放语音检测任务的影响，并且 (iii) 我们研究了使用最先进的重放语音检测器的环境无关性能，展示了在训练数据中不存在的封闭环境中使用合成数据进行重放检测的潜力。

2. 重播语音模拟

在一个典型的重放攻击条件下，说话人的语音既被验证系统的麦克风阵列捕获，也被攻击者的录音设备捕获，要么同时进行，要么独立进行。攻击者随后通过扬声器（或其他播放设备）重放捕获的信号，检测器必须确定接收到的信号是实时的还是已被重放的。

我们模拟了重放攻击，定义了三种场景，并使用表 1 中的频域表示法 [16] 。在真实场景中，ASV 系统记录说话人的语音，这可以被建模为不同传播路径的总和。特别是，源语音频谱 $S_0(f)$ 被说话者在房间中的每个声学传播路径 k 上的频率依赖性定向响应 $\boldsymbol{S}(f,\boldsymbol{\theta}_k)$ 过滤。该路径在离开说话者时具有路径向位和仰角 $\boldsymbol{\theta}_k = [\theta_k^{\text{azimuth}}, \theta_k^{\text{elevation}}]$。具体而言，$\boldsymbol{S}(f,\boldsymbol{\theta}_k)$ 是一个定义到达每个检测麦克风的路径上定向响应的向量。然后，源语音频谱通过路径依赖的声学传递函数 $\boldsymbol{H}_k^{(\text{gen})}(f)$ 进行滤波，该传递函数包括所有检测麦克风的传递函数向量和阵列麦克风的定向传递函数 $\boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{gen})})$ 以及到达角度 $\boldsymbol{\phi}_k^{(\text{gen})}$。从 ASV 阵列 $\boldsymbol{X}_{\text{gen}}(f)$ 记录的真实多通道语音可以表示为

$$\boldsymbol{X}_{\text{gen}}(f) = S_0(f) \sum_k \boldsymbol{S}(f,\boldsymbol{\theta}_k) \odot \boldsymbol{H}_k^{(\text{gen})}(f) \odot \boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{gen})}), \tag{1}$$

其中 $\odot$ 表示向量的逐元素乘法。

在攻击的欺骗场景阶段，对手将一个单通道的欺骗麦克风放置在说话者附近记录源信号通过其自身的 K_{spf} 传播路径。每条路径应用其自己的方向相关的传输函数 $S(f,\boldsymbol{\theta}_k)$ 来建模声源的方向性，路径相关的传输函数 $H_k^{(\text{spf})}(f)$ 来建模房间的声学特性，并且具有方向性的欺骗麦克风响应 $M_{\text{spf}}(f,\boldsymbol{\phi}_k^{(\text{spf})})$，从而产生单通道的欺骗信号。

$$Y_{\text{spf}}(f) = S_0(f)\sum_k S\big(f,\boldsymbol{\theta}_k\big)\, H_k^{(\text{spf})}(f)\, M_{\text{spf}}\big(f,\boldsymbol{\phi}_k^{(\text{spf})}\big). \quad (2)$$

最后，在重放场景中，攻击者通过具有方向性辐射图案 $\boldsymbol{L}(f,\boldsymbol{\theta}_k)$ 的扬声器重放 $Y_{\text{spf}}(f)$，该辐射图案应用于房间 k 中的每个声波传播路径，这些路径离开源时具有各自的方位角和仰角 $\boldsymbol{\theta}_k = [\theta_k^{\text{azimuth}}, \theta_k^{\text{elevation}}]$。类似地，对于语音信号，$\boldsymbol{L}(f,\boldsymbol{\theta}_k)$ 是包含每个检测麦克风的方向响应的向量。然后，重放信号沿着 K_{rep} 个重放路径传播，每个路径都有其房间的声学传输函数 $\boldsymbol{H}_k^{(\text{rep})}(f)$ 和方向性麦克风阵列响应 $\boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{rep})})$。同样，在求和之前，传播和麦克风响应在各通道之间逐元素结合，从而产生最终的多声道重放信号。

$$\boldsymbol{X}_{\text{rep}}(f) = Y_{\text{spf}}(f)\sum_k \boldsymbol{L}\big(f,\boldsymbol{\theta}_k\big)\odot\boldsymbol{H}_k^{(\text{rep})}(f)\odot\boldsymbol{M}_{\text{ASV}}\big(f,\boldsymbol{\phi}_k^{(\text{rep})}\big). \quad (3)$$

由于重放链导致的时间扩散或双倍回声，重放信号与真实信号不同，并产生可量化的伪影。因此，为了稳健地泛化，检测系统依赖于对重新录制的伪影、设备/房间特征和方向性敏感的特性。

Table 1. 重放语音攻击的描述所用的符号。

符号	意义、典型值/集和域
$K_{\text{gen}}, K_{\text{spf}}, K_{\text{rep}}$	$\mathbb{Z}_{\geq 1}$: Number of propagation paths in each scenario.
C	$\mathbb{Z}_{\geq 1}$: Number of ASV array microphones.
$S_0(f)$	$\mathbb{C}$: Source speech spectrum.
$\boldsymbol{\theta}_k$	$\mathbb{R}^2$: Source or loudspeaker radiation direction for path k.
$\boldsymbol{\phi}_k$	$\mathbb{R}^2$: Arrival direction to the microphones for path k.
$\boldsymbol{S}(f,\boldsymbol{\theta}_k)$	$\mathbb{C}^{C\times 1}$:Talker directivity for path k, $k = 1,\ldots,K_{\text{gen}}$ or K_{spf}.
$\boldsymbol{H}_k^{(\text{gen})}(f)$	$\mathbb{C}^{C\times 1}$: Acoustic transfer function of path k to ASV array.
$\boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{gen})})$	$\mathbb{C}^{C\times 1}$: ASV array directional response for path k.
$H_k^{(\text{spf})}(f)$	$\mathbb{C}$: Acoustic transfer function of path k to spoofing microphone.
$M_{\text{spf}}(f,\boldsymbol{\phi}_k^{(\text{spf})})$	$\mathbb{C}$: Spoofing microphone directional response for path k.
$\boldsymbol{L}(f,\boldsymbol{\theta}_k)$	$\mathbb{C}^{C\times 1}$: Loudspeaker directivity for replay path k, $k = 1,\ldots,K_{\text{rep}}$.
$\boldsymbol{H}_k^{(\text{rep})}(f)$	$\mathbb{C}^{C\times 1}$: Acoustic transfer function of path k from loudspeaker to ASV array.
$\boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{rep})})$	$\mathbb{C}^{C\times 1}$: ASV array directional response for replay path k.
$\boldsymbol{X}_{\text{gen}}(f)$	$\mathbb{C}^{C\times 1}$: Multichannel signal recorded by ASV array during genuine scenario.
$Y_{\text{spf}}(f)$	$\mathbb{C}$: Single-channel recording by spoofing microphone.
$\boldsymbol{X}_{\text{rep}}(f)$	$\mathbb{C}^{C\times 1}$: Multichannel replayed signal captured by ASV array.

3. 材料

本节描述了用于构建重播语音合成数据集的材料。整体生成框架如图 1 所示。

3.1. 数据集

以下开源数据集已在仿真框架中使用：

(i) EARS [17] 数据集是一系列无回声语音录音的集合，包含多位说话人的带有情感的发声。该数据集在控制下的无回声室中使用高保真麦克风录制而成，以确保信号不含有任何环境混响或背景噪音。该数据集包括广泛的情感状态（如：中性、快乐、愤怒、悲伤），并且采样率为 48 千赫兹，为 107 位说话人提供清晰的全频带语音，性别和年龄各异。此数据集用于在真实场景中采样源语音信号 $S_0(f)$。

(ii) Gallien 等。[18] 提供了存储在 SOFA（面向空间的声学格式）文件中的测量扬声器方向脉冲响应。每个文件捕捉了使用放置在多个方位角和仰角处的麦克风阵列获得的扬声器的方向辐射模式。这些测量是在无回声条件下进行的，且每个 SOFA 文件包括诸如麦克风和声源位置等元数据。该数据集已在重放场景中用于建模扬声器 $\boldsymbol{L}\big(f,\boldsymbol{\theta}_k\big)$ 的方向性。

(iii) 数据集 [19] 中包含了在受控条件下从多个方向和声源距离捕捉的各种商用麦克风的测量值 cro:IRimpulse responses (IRs)。该数据集包括全向、心形和双向麦克风。这些 IRs 模拟了由麦克风硬件引入的方向频率响应和色彩效果，可以与模拟信号进行卷积以生成真实的麦克风着色录音。在我们的实验中，我们仅使用全向麦克风，因为它们最常用于 VAs [12]。该数据集分别用于真实场景和重放场景中的欺骗和验证器麦克风阵列的脉冲响应 $M_{\text{spf}}(f,\boldsymbol{\phi}_k^{(\text{spf})})$ 和 $\boldsymbol{M}_{\text{ASV}}(f,\boldsymbol{\phi}_k^{(\text{gen})})$。

(iv) WHAMR! [20] 数据集用于在方程 (1)–(3) 中建模真实、欺骗和重放场景的背景噪声 $N(f)$。由于噪声的影响从未在重放语音检测的背景下被讨论过，我们考虑了两种情况：全向噪声和扩散噪声。

全方位噪声。噪声首先作为单声道信号从 WHAMR! [20] 数据集提供。为了将其与多通道录音结合，相同的噪声被复制到各个通道中。然后将每个通道的噪声缩放到相对于该通道的干净语音达到目标 cro:SNRsignal-to-noise ration (SNR) 再进行添加。

扩散噪声。来自 WHAMR! [20] 数据集直接在多个通道中呈现（例如，为麦克风阵列模拟漫反射噪声）。与复制不同，每个通道接收空间一致但不同的噪声信号。然后，噪声被全局缩放以实现相对于干净多通道信号的 SNR。在这项工作中，我们采用了 [21] 中描述的算法。

整个声学模拟框架使用 Pyroomacoustics [22] 进行了模拟，包括空间 cro:RIRroom impulse responses (RIRs) ($\boldsymbol{H}_k^{(\text{gen})}(f)$、$H_k^{(\text{spf})}(f)$ 和 $\boldsymbol{H}_k^{(\text{rep})}(f)$)、说话人方向性 $\boldsymbol{S}(f,\boldsymbol{\theta}_k)$、房间尺寸和路径衰减的模拟。模拟的参数集和约束条件如表 2 所示。

Table 2. 数据生成的参数和约束条件。

房间参数	**范围**
Room width, height, and length	$\mathcal{U}[3.0, 6.0]$ m
# of materials (wall, floor, ceiling)	$13, 7, 8$
Absorption coefficients	$\mathcal{U}[0.1, 0.6]$
SNR	$\mathcal{U}[-10, 40]$ dB
源参数	**约束条件**
$\|\|\boldsymbol{p}_{\text{tlk}} - \boldsymbol{p}_{\text{ASV}}\|\|$	> 1 m
$\|\|\boldsymbol{p}_{\text{tlk}} - \boldsymbol{p}_{\text{spf}}\|\|$	< 1 m
$\|\|\boldsymbol{p}_{\text{spk}} - \boldsymbol{p}_{\text{ASV}}\|\|$	> 1 m
Source-to-surface distance	> 1 m
$\boldsymbol{S}(f,\boldsymbol{\theta}_k)$	Cardioid
麦克风阵列参数	**值**
Number of microphones C	2
Sampling frequency f_s	48 kHz
Spacing distance d	50 mm

3.2. 研究中的配置

在这项工作中，我们研究了两种不同的配置：混响欺骗和无回声欺骗。

混响欺骗。这个条件反映了典型的重放攻击流程，其中攻击者首先使用欺骗麦克风录制语音，然后通过扬声器重播放录的信号。因此，重放信号包括了录音阶段引入的声音色彩以及来自重播环境的变换。具体来说，第一个场景是从 Env-A 收集 $\boldsymbol{X}_{\text{gen}}(f)$ 和 $Y_{\text{spf}}(f)$。然后，在 Env-B 中使用来自 [18] 的 4SOFA

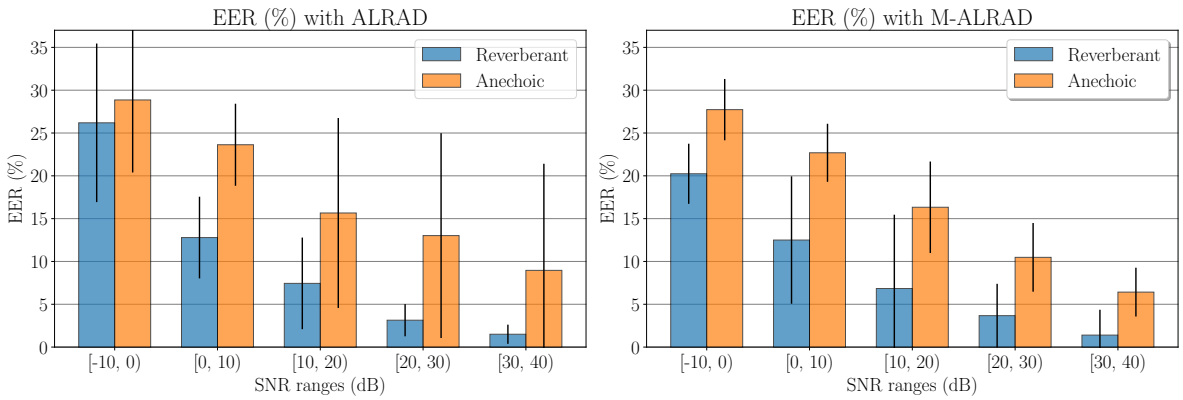

Fig. 2. 分别在合成配置测试集上以 48kHz 并加入扩散噪声的 ALRAD 和 M-ALRAD 的 EERs (%) 。

文件重放 $Y_{\text{spf}}(f)$。通过这样做，每个数据集版本分别包含了真实和重放录音，其比例分别为 1:4。

无回声欺骗。回放信号直接从干净的无回声语音信号生成，而不是由欺骗麦克风捕获的记录版本。这种设置假设攻击者可以访问原始无回声语音 $S_0(f)$ 并通过扬声器重播它。在这种情况下，$Y_{\text{spf}}(f) = S_0(f)$。

每个合成数据集的版本都是通过模拟来自 EARS 数据集的 107 说话者的 2 秒的 10 次演讲生成的，遵循 [12]。Pyroomacoustics 中房间和麦克风阵列模拟的参数详细信息见表 2。值得注意的是，麦克风的数量 C 及其排列可以自定义。在我们的案例中，我们模拟了一个间距为 $d = 50$ 毫米的双麦克风阵列，在 48 千赫兹下。

4. 评估

在本节中，我们描述了用于生成数据实验的学习型重播语音检测器。然后，我们在第 3 节所述的两种配置上评估该模型。最后，进行了噪声对重播语音检测的影响分析以及向未见过环境的泛化能力分析。

4.1. 检测模型

在我们的实验中，我们采用 M-ALRAD [14] 作为重播语音检测器，这是一种基于 cro:CRNNconvolutional recurrent neural network (CRNN) 的方法，联合处理 C 多通道复数 cro:STFTshort-time Fourier transforms (STFTs) $\{X_{\text{STFT}_{c,T,F}}, c = 1, \ldots, C\}$ 与 T 和 F 分别代表时间和频率的 bin 数量，以生成用于检测重播语音的单通道波束形成的频谱图。首先，一个 cro:CNNconvolutional neural network (CNN) $f_{BM} : \mathbb{C}^{C\times T\times F} \to \mathbb{C}^{T\times F}$ 通过一系列 Conv2D-BatchNorm-ELU-Conv2D 操作输出波束成形权重 $\boldsymbol{W} \in \mathbb{C}^{C\times T\times F}$，其中实部和虚部沿通道维度进行连接。波束形成的谱图 $\hat{X}_{\text{STFT}}$ 计算为 $\hat{X}_{\text{STFT}_{t,f}} = \sum_c X_{\text{STFT}_{c,t,f}} \cdot w_{c,t,f},$，其中 $t = 1, \ldots, T$ 和 $f = 1, \ldots, F$ 分别是时间 bins 和频率 bins 的索引。波束形成的频谱图然后由一个先前用于说话人距离估计的 CRNN 分类器进行处理 [23, 24]。提取了波束形成 STFT 的幅度和相位特征，并对相位应用了正弦和余弦变换。该模型被训练以最小化预测的二进制类别与真实的二进制类别之间的二元交叉熵损失，即真实或重放。STFT 是通过长度为 32 和 46 毫秒且重叠率为 50% 的汉宁窗计算得出的，频率为 48 千赫。提出的模型然后使用批量大小为 32 进行训练，采用余弦退火调度的学习率为 0.001，共进行 50 个周期。此外，正交性和稀疏性损失被采用，并使用与 [14] 中相同的超参数。

为了评估检测的性能，我们采用了 EER 指标。从 5 次独立训练中收集了结果以计算 95% 置信区间。为了评估多通道记录在重播语音检测任务中的重要性，我们也使用了模型的单通道版本（即 ALRAD），其中单个通道（阵列中的第一个）被复制到每个麦克风通道，并随后输入给检测器。

4.2. 结果

表 3 给出了在两个采样率（16 和 48kHz）及两种欺骗条件下的 ALRAD 和 M-ALRAD 的 EERs。M-ALRAD 在这两种配置和采样率下均优于 ALRAD，表明多通道录音可以简化重放片段的检测。值得注意的是，与**混响欺骗**相比，**无回声欺骗**导致了明显更高的 EERs。此外，在全向噪声情况下，M-ALRAD 实现了比扩散噪声情况更好的检测性能，这表明复制单通道噪声引入的信道相关性被模型用作区分特征。

Table 3. 不同采样率和使用通道数量在每个合成场景中 EERs 的表现。

方法	噪声	16 千赫兹		48 千赫兹	
		Reverberant	Anechoic	Reverberant	Anechoic
ALRAD	Diffuse	7.0 ± 1.4	30.8 ± 4.0	7.8 ± 2.2	32.4 ± 3.2
M-ALRAD	Diffuse	6.1 ± 0.6	24.5 ± 4.1	6.4 ± 1.7	26.2 ± 2.1
ALRAD	Omni	6.1 ± 1.9	29.4 ± 1.4	6.6 ± 1.3	18.2 ± 9.4
M-ALRAD	Omni	2.8 ± 0.6	23.6 ± 0.9	2.4 ± 0.4	14.9 ± 2.5

为了分析在不同噪声条件下重放检测的鲁棒性，我们评估了两种已使用变量 SNR 训练的方法：ALRAD 和 M-ALRAD，在从 −10dB 到 40dB 的五个 SNR 范围内。图 2 显示了两种模型在两种欺骗配置中的 EERs。结果显示 M-ALRAD 一致优于 ALRAD，特别是在低信噪比条件下，这证实了合成数据集中捕获的空间线索增强了检测的鲁棒性。

4.3. 推广到实际数据

我们通过在 ReMASC 数据集 [12] 上测试来评估基于合成数据训练的模型的一般化能力。表 4 和表 5 比较了使用合成和真实数据作为单通道和多通道噪声注入训练集时，三种环境下的 EERs。在扩散噪声的情况下，合成训练通常无法很好地迁移，在大多数环境中 EERs（$> 50\%$）都非常高。无论是 ALRAD 还是 M-ALRAD 在这些条件下表现都不佳，这表明增加的扩散噪声使训练分布与 ReMASC 的分布差异更大。

相反，当使用单通道噪声时，值得注意的是，在合成数据上训练的模型仅在 EnvC 中与仅使用真实数据的模型表现相似，而在该环境中说话者和 ASV 麦克风的位置都是固定的。相反，在 EnvA 和 EnvC（分别是公园和汽车环境）中，得到的结果

Table 4. EER(%) 在 ReMASC 的 D1 测试子集上，在环境独立场景 [12] 中使用不同训练数据及 48 千赫兹和扩散噪声注入的比较。

Methods	Training	EnvA	EnvC	EnvD
M-ALRAD	ReMASC	10.4 ± 9.4	22.8 ± 9.4	20.2 ± 3.1
ALRAD	Reverberant	53.8 ± 12.8	51.1 ± 29.5	53.6 ± 7.5
ALRAD	Anechoic	69.8 ± 14.3	67.1 ± 11.5	56.4 ± 4.7
M-ALRAD	Reverberant	64.3 ± 4.8	53.5 ± 27.5	54.3 ± 6.0
M-ALRAD	Anechoic	52.3 ± 25.8	73.9 ± 9.9	49.0 ± 9.1

环境-B* 由于数据收集期间的硬件故障，不包含 D1 真实陈述。

甚至比随机猜测还要差。这是可以预料的，因为合成数据是在房间内生成的，而室外环境和车辆环境在背景噪声和声学特性方面与房间差异很大。这表明全向噪声引入了有时有助于模型捕捉相关回放线索（例如，在 EnvC）但也在其他环境中导致严重不匹配的伪影。总体而言，EERs 在真实数据中的值远高于合成数据，突显出需要能够适应实际环境条件的方法。

Table 5. EER(%) 在 ReMASC 的 D1 子集上的比较，在环境独立场景 [12] 下，使用不同训练数据在 48kHz 和全向噪声注入情况下的表现。

Methods	Training	EnvA	EnvC	EnvD
M-ALRAD	ReMASC	10.4 ± 9.4	22.8 ± 9.4	20.2 ± 3.1
ALRAD	Reverberant	73.5 ± 6.1	23.8 ± 5.9	53.2 ± 8.4
ALRAD	Anechoic	65.1 ± 8.3	37.0 ± 7.8	59.1 ± 6.2
M-ALRAD	Reverberant	64.3 ± 4.8	23.4 ± 18.0	54.7 ± 4.6
M-ALRAD	Anechoic	62.9 ± 18.4	52.7 ± 25.6	51.9 ± 13.6

环境-B* 不包含由于数据收集期间硬件故障导致的 D1 真实言语。

5. 结论

在这项工作中，我们介绍了一个专门为多通道重放语音检测设计的声学仿真框架。通过利用公开的数据集和现实主义模拟工具，我们的框架能够创建反映多样化声学环境和对抗配置的空间丰富的重放案例。尽管将领域转移到像 ReMASC 这样的真实世界数据集中仍然具有挑战性，但我们的结果突显了合成数据在弥合这一差距和支持与环境无关的重放检测方面的潜力。未来的工作将使用此数据集（及其变体）来解决麦克风阵列不匹配的一般化问题。

6. REFERENCES

[1] W. Huang, W. Tang, H. Jiang, J. Luo, and Y. Zhang, "Stop deceiving! an effective defense scheme against voice impersonation attacks on smart devices," IEEE Internet of Things Journal, vol. 9, no. 7, pp. 5304–5314, 2022.

[2] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2507–2522, 2023.

[3] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in IEEE ICASSP, 2017.

[4] Y. Gong, J. Yang, and C. Poellabauer, "Detecting replay attacks using multi-channel audio: A neural network-based method," IEEE Signal Processing Letters, vol. 27, pp. 920–924, 2020.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Interspeech, 2017.

[6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Interspeech, 2019.

[7] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A Capsule Network Based Approach for Detection of Audio Spoofing Attacks," in IEEE ICASSP, 2021.

[8] J. Xue, C. Fan, J. Yi, J. Zhou, and Z. Lv, "Dynamic Ensemble Teacher-Student Distillation Framework for Light-Weight Fake Audio Detection," IEEE Signal Processing Letters, vol. 31, pp. 2305–2309, 2024.

[9] J. Boyd, M. Fahim, and O. Olukoya, "Voice spoofing detection for multiclass attack classification using deep learning," Machine Learning With Applications, vol. 14, pp. 100503, 2023.

[10] L. Xu, J. Yang, C. H. You, X. Qian, and D. Huang, "Device Features Based on Linear Transformation With Parallel Training Data for Replay Speech Detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 1574–1586, 2023.

[11] M. Omologo, M. Matassoni, and P. Svaizer, "Speech recognition with microphone arrays," in Microphone Arrays: Signal Processing Techniques and Applications, pp. 331–353. Springer, 2001.

[12] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," Interspeech, 2019.

[13] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in ACM SIGSAC, 2016, pp. 1080–1091.

[14] M. Neri and T. Virtanen, "Multi-channel replay speech detection using an adaptive learnable beamformer," IEEE Open Journal of Signal Processing, vol. 6, pp. 530–535, 2025.

[15] M. Neri and T. Virtanen, "Impact of Microphone Array Mismatches to Learning-based Replay Speech Detection," in EUSIPCO, 2025.

[16] J. Benesty, J. Chen, and Y. Huang, Microphone Array Signal Processing, Springer, 2008.

[17] J. Richter, Y. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation," in Interspeech, 2024.

[18] A. Gallien, K. Prawda, and S. J. Schlecht, "Matching early reflections of simulated and measured RIRs by applying sound-source directivity filters," in Audio Engineering Society Conference: AES, 2024.

[19] J. C. Franco Hernández, B. Bacila, T. Brookes, and E. De Sena, "A multi-angle, multi-distance dataset of microphone impulse responses," Journal of the Audio Engineering Society, vol. 70, no. 10, pp. 882–893, 2022.

[20] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in Interspeech, 2019.

[21] D. Mirabilii, S. J. Schlecht, and E. A. P. Habets, "Generating coherence-constrained multisensor signals using balanced mixing and spectrally smooth filters," The Journal of the Acoustical Society of America, vol. 149, no. 3, pp. 1425–1433, 2021.

[22] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in IEEE ICASSP, 2018.

[23] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Speaker Distance Estimation in Enclosures From Single-Channel Audio," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2242–2254, 2024.

[24] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Single-Channel Speaker Distance Estimation in Reverberant Environments," in IEEE WASPAA, 2023.