arXiv:2611.30106v1 [cs.CR] 23 Nov 2026

超声波水印：用于防范智能手机语音诈骗的实时声学水印

Renzhi Hao
horange1024@outlook.com
彭博有限合伙企业，美国纽约州纽约市

Shirui C
aoshiruicao@cs.umass.edu 马萨诸塞大学阿默斯特分校，美国马萨诸塞州阿默斯特市

Yizhu Wenyizwen@iu.edu 印第安纳大学布卢明顿分校，美国印第安纳州布卢明顿市

Rishav Guptarishavg1@umbc.edu 马里兰大学巴尔的摩县分校，美国马里兰州巴尔的摩

Dharani Nadendladharann1@umbc.edu 马里兰大学巴尔的摩县分校，美国马里兰州巴尔的摩

Mehran Kafai
mkafai@amazon.com 亚马逊，美国加利福尼亚州圣迭戈

Hanqing Guoguohan@iu.edu 印第安纳大学布卢明顿分校，美国印第安纳州布卢明顿

Dong Lidli@umbc.edu 马里兰大学巴尔的摩县分校，美国马里兰州巴尔的摩

摘要

文本转语音（TTS）、语音转换（VC）以及人工智能驱动的语音合成的最新进展，使攻击者能够在电话通话中令人信服地冒充受信任的个人。现有防御往往依赖离线分析，在电话失真下性能下降，或仍易受自适应攻击影响。为应对这些局限，我们提出超声波水印，一种实时、智能手机原生的框架，通过声学水印将语音与呼叫方的物理发音动作绑定，从而保护实时通话。通话期间，智能手机发射超声信号，并仅使用其内置扬声器和麦克风捕获面部发音动作的反射。超声波水印从这些反射中提取运动特征，并将其与外发语音融合以生成动态水印。由于超声测量在物理上和时间上与语音产生耦合，若没有底层发音动作，则难以伪造。同步生成器-检测器架构强制跨模态语音-运动一致性，以抵御重放、合成语音和模仿发音攻击。超声波水印持续运行，延迟开销低于200毫秒，且无需额外硬件。我们在iOS和Android上实现了超声波水印，并使用来自22名说话人、跨越八部智能手机和四种真实环境的超过2.5小时同步语音-超声录音对其进行评估。结果表明，该方法能够抵御多种攻击和电话通信失真，同时保持感知音频质量。

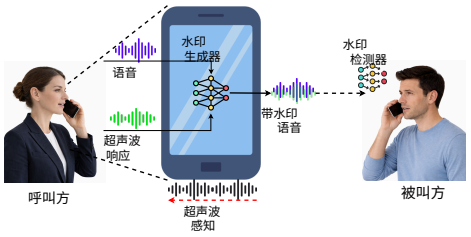


图1：超声波水印的使用场景。

1 引言

语音是人类交流的一种自然形式，承载着语言信息和说话者特有的生物特征。在移动环境中，语音越来越多地被用作版权保护、数据访问和安全通信的标识符。然而，生成式人工智能的发展带来了严重的安全威胁。现代文本转语音（TTS）和语音转换（VC）系统，包括基于WaveNet 和生成对抗网络的模型 [14, 15, 27, 31], ，能够仅使用极少的训练数据合成出高度仿真的个人语音。这些能力助长了语音网络钓鱼（“vishing”）等冒充攻击，导致美国消费者在 2024年损失超过 29.5 亿美元

仅此一项 [11, 26]。与此同时，人类往往难以区分合成语音克隆与真实语音，其检测性能已接近随机猜测水平 [12, 36]。

场景。 设想这样一个场景：用户接到一通电话，来电者听起来像是他认识的人。呼叫方可能冒充主管，要求立即采取行动，例如批准一笔付款。由于声音听起来很熟悉，被叫方可能会相信这通电话并迅速采取行动。攻击一旦得逞，可能造成经济损失、数据泄漏，并长期损害信任和职业关系。在这种情况下，被叫方必须实时判断语音来自合法人类还是合成冒充。

现有方案。 如表1所示，现有防御措施在这一场景中存在不足。深度伪造检测器（例如基于 ASVspoof 的系统 [40–42], ）会分析音频内容，以识别合成语音引入的声学伪影。这些方法通常离线应用或在服务器端运行。尽管它们能够防御重放和合成攻击，但其依赖的声学伪影在真实电话通话中往往会受到电话编解码器、回声和流媒体效应的抑制或扭曲，因此不适合在实时通话期间检测冒充。数字水印会将不可感知的比特嵌入合法语音中，之后可提取或检测这些比特以进行身份验证 [4, 21, 29, 32, 46]。然而，由于水印嵌入和验证通常离线进行，这些方法无法在实时通话中验证语音是否来自合法扬声器。活体检测用于验证音频是否来自真人扬声器，可能需要专用传感器或用户明确交互 [13, 16]。然而，仅凭活体检测无法确认身份：攻击者可以对着电话说话，通过活体检测，同时注入或重放合成语音，冒充受信任的人，例如家人或主管 [16, 30, 43]。最后，更广泛的来源追踪倡议和平台级政策 支持媒体来源追踪，但无法在通话进行期间以密码学或物理方式将音频与合法人类扬声器绑定 [9, 25]。总的来说，这些局限凸显了实用通话内防御的缺失：这种防御必须能够在实时条件下运行、确认活体状态并验证语音真实性。

为防御通话中冒充攻击，我们提出超声波水印：一种实时、原生于智能手机的呼叫者验证系统，通过声学数字水印将语音与呼叫方的生理发音动作绑定。超声波水印由呼叫方手机上的水印生成器和被叫方手机上的检测器组成。如图 1 所示，呼叫方的手机持续发射超声波，并利用其内置扬声器—麦克风对采集口面部运动（即嘴部、下颌和脸颊）产生的反射。这些信号与外发语音融合，以生成动态水印；水印将以人耳无法察觉的方式嵌入语音中，并传输给被叫方。检测器通过三个层验证呼叫方的合法性：**(i) 协议合规性**，确保接收信号符合预期的水印处理流程，并支持有效解码；**(ii)实时语音存在性**，检测发音动作产生的超声波反射，以确认呼叫方手机附近正在实时产生语音；**(iii) 语音-运动一致性**，验证语音与观测到的发音动作相匹配。

挑战。构建此类系统并非易事，主要面临四项挑战。（**C1：超声波特征提取与融合**。）系统必须仅使用普通智能手机硬件，可靠地从超声反射中提取发音动作线索，并将其与可听语音实时融合。这些特征必须紧凑、在时间上与语音对齐，并且具有足够的表达能力，以捕捉无法仅凭语音推断的细粒度发音动作。（**C2：对抗真实性**。）水印必须在各种攻击下保持有效，包括重放、合成语音，以及发音动作不匹配或同步不严密的实时真人冒充；同时还要应对真实通话中的失真，例如编解码器压缩、回声和噪声。这要求在语音与超声波之间强制实施因果对齐，而不能独立依赖任一模态。（**C3：实时性能**。）所有处理都必须以低延迟和有界计算量持续运行，逐帧生成和验证水印，同时将端到端延迟控制在难以察觉的程度。（**C4：跨平台部署**。）最后，系统必须完全运行于原生 iOS 和 Android 智能手机上，无需额外硬件或用户交互，尽管不同平台在信号处理、设备端模型执行和音频处理流程集成方面存在特定限制。

为应对这些挑战，我们设计了超声波水印，将其构建为流式双模态系统，并采用三项关键设计：(i) 轻量级超声处理流程提取基于相位和振幅的运动特征，并通过跨模态编码器-解码器将其与语音融合，生成与实时发音动作相关联的水印；(ii) 同步的生成器-检测器架构确保语音与动作之间实现细粒度的因果对齐，防止重放、合成或模仿发音攻击生成具有一致性的跨模态对应关系；(iii) 滑动窗口流式设计实现低延迟运行，同时通过设备感知校准、编解码器感知优化和冗余机制，提升系统在不同设备、环境和通话动态条件下的鲁棒性。上述设计共同实现了基于生理特征、面向智能手机的实时语音验证，同时保持通话质量。我们的贡献如下：

方法	通话中无需额外手机			重放攻击	人工智能合成（文本转语音/语音转换）	真人冒充	用户操作负担	参考文献
		HW	原生					
深度伪造检测器（ASVspoof、机器学习）	△	✓	△	△	✓	✗	△	[40–42]
音频水印	✗	✓	△	✗	△	✗	✓	[4, 21, 29, 32, 46]
活体检测（传感器、协议）	△	通常	△	✓	△	✗	△	[16, 30, 43]
溯源（平台级）	✗	✓	✓	✗	△	✗	✓	[9, 25]
我们的工作：超声波水印	✓	✓	✓	✓	✓	✓	**Low**	-

表1：商用智能手机实时通话中针对语音冒充的防御措施比较。通话中：**在通话期间运行；**无需额外硬件：**无需额外硬件；**手机原生：**使用手机原生硬件；**重放、人工智能合成**和真人：**分别表示对录制语音、合成/转换语音和真人冒充的鲁棒性；**用户操作负担：**交互负担。**符号：✓、✗和△ 分别表示支持、不支持和部分/有条件支持。

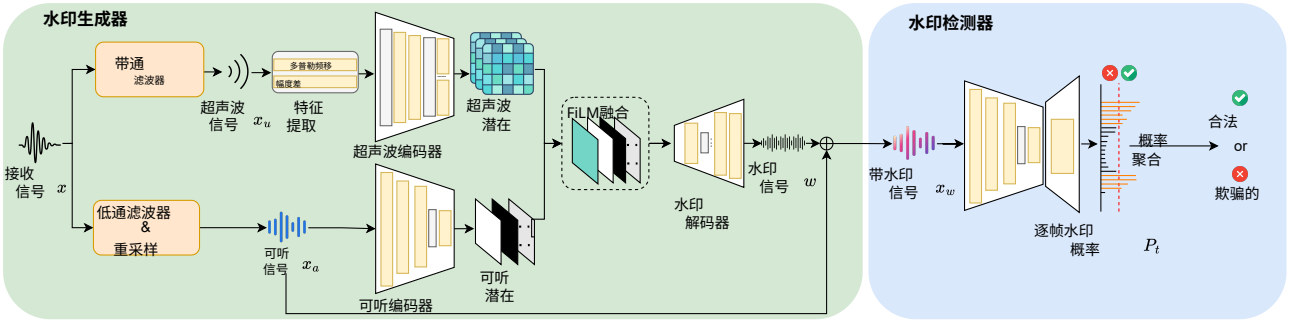


图2：我们提出的超声波水印框架概览。

• **实时超声波水印框架：**我们提出了超声波水印，这是一种实时声学数字水印框架，利用智能手机超声波将实时语音与真实的发音动作绑定。通过将语音和超声波反射共同编码为逐帧水印，并在检测器端确保二者的时间一致性，超声波水印能够实时验证语音，有效防范重放、合成和模仿发音攻击。

• **全面评估：**在经机构审查委员会批准的方案下，我们使用八款普通智能手机，在四种真实环境中，从 22 位说话人处采集了超过 2.5 小时的语音同步超声波录音。我们针对仅音频攻击、随机超声波攻击、语音同步超声波攻击和自适应对抗攻击，评估了超声波水印的表现，同时考察了重放语音和合成语音。我们还将其与近期的仅音频活体检测方法进行比较，并评估其在常见电话失真和编解码器条件下的鲁棒性。检测、攻击成功率、延迟和语音质量指标方面的结果表明，该方法能够提供鲁棒保护，同时对感知音频质量的影响有限。

• **跨平台移动系统：**我们实现了超声波水印的完整设备端原型，支持 iOS 和 Android，且无需额外硬件。该原型可在真实通话期间，针对实际攻击场景实时生成和检测水印。实时系统演示请见此处。源代码和数据集将在论文被接收后公开。

2 威胁模型

用户场景。受害者是一名具有安全意识的被叫方，会根据感知到的来电者身份做出敏感决策。我们假设受害者在涉及财务或敏感信息的通话中使用超声波水印，并期望对方使用同一应用。该场景具有现实可行性，因为安全通话通常由应用级服务提供，要求双方端点采用同一平台（例如 WhatsApp [37] 和 Apple FaceTime [1]）。

攻击者的目标。攻击者旨在冒充可信说话者（例如家庭成员或主管），诱使被叫方执行安全敏感操作。具体而言，攻击者试图让被叫方检测器将攻击者控制的语音识别为携带有效水印，从而使通话看起来合法。

攻击者能力。我们假设攻击者能够使用智能手机发起实时通话，并获取目标说话人的既有语音样本（例如来自社交媒体），用于重放或语音合成。攻击者可以在自己的设备上运行我们的应用，但无法远程访问目标通话的实时超声感知路径，也无法直接重现目标说话人的超声波响应与物理声学路径耦合。因此，重放或合成语音可能会模仿目标说话人的声音，但不会天然重现实时发音动作中同步的语音–超声波耦合。我们考虑四种能力逐渐增强的攻击场景：

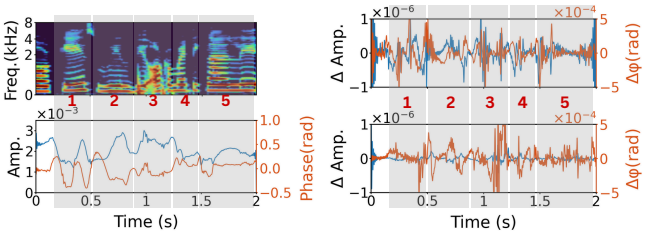
- **场景1：仅音频欺骗**。攻击者在未启用超声波水印的情况下呼叫受害者，并且只能控制语音。我们考虑两种攻击：（i）重放，即在通话过程中重放先前录制的目标语音 [5, 35]；以及（ii）合成，即使用 TTS [3]或 VC [17, 18]生成目标的语音。
- **场景2：随机超声欺骗**。攻击者知晓超声波水印，并可以在自己的设备上运行该应用，但无法生成与伪造语音正确耦合的超声波轨迹。在重放或合成目标语音的同时，攻击者在手机附近做出任意且不相关的口部动作，以诱发超声波反射。

- **场景3：语音同步超声欺骗**。攻击者采用更强的策略，使其口部动作与重放或合成的语音同步。通过在设备附近模仿相应的发音动作，攻击者试图产生与伪造语音一致的超声波反射，并诱骗超声波水印接受这些组合信号，将其判定为来自合法呼叫者。

- **场景4：自适应对抗攻击**。攻击者掌握系统知识，并针对生成器—检测器框架调整伪造信号。攻击者可以利用检测器查询次数、架构知识或替代模型来构造扰动，使伪造语音被接受。他们还可以在水印生成之前修改重放语音或合成语音，目的是使其被接受为来自合法呼叫者。

3 系统设计

概述： 我们提出了超声波水印，这是一种双模态框架，可将超声反射中的发音动作线索嵌入语音中，从而在智能手机上实时验证合法呼叫者。如图 2 所示，该系统由水印生成器 和水印检测器组成。生成器联合编码可听语音和超声信号，将语音潜表示与超声特征图融合；后者能够捕捉瞬时相位和振幅变化。这种设计使水印基于实时发音动作



(a) 语音与超声波对比。 **(b) 合法与欺骗对比。**

图3：语音期间的超声波特征。阴影区域表示口语词语：1：“Your”，2：“case”，3：“will”，4：“be”，5：“automatically”。(a) 带有超声波振幅和相位特征的语音梅尔频谱图。(b) 合法超声波（上）与语音同步伪造超声波（下）的相位和振幅差异。

而非声学重放。检测器执行轻量级设备端推理，估计帧级水印置信度，并随时间进行汇总，以实现完整性验证。所有组件均针对低延迟、鲁棒性和智能手机部署进行了优化，从而能够持续、无感地验证合法呼叫者。

3.1 特征提取与融合

我们提出了一种轻量级感知设计，可提取一组与语音产生具有因果关联、难以通过重放或合成伪造，并适用于持续通话中运行的最小运动线索。

设计原理。我们的设计基于三个方面的认识。首先，发音动作发生在毫米级和毫秒级尺度上，低速率的基于帧的特征难以捕捉这些动作。连续波超声波能够进行与语音事件 [33, 45]对齐的细粒度样本级感知。其次，抗伪造能力需要以物理机制为基础的特征。尽管绝对频谱模式可以被重放或合成，但由实时动作引起的相位和振幅的时间导数更难伪造。第三，移动设备限制排除了重量级和非因果处理，因此需要采用轻量级信号处理的增量式流水线。

通话期间，呼叫方的智能手机会持续发射多载波超声信号 [45]，其基频为 $f_0 = 20$ 千赫，包含六个等间隔载波（$\Delta f = 300$ 赫兹）：$u(t) = \sum_{i=0}^{5} \cos(2\pi(f_0 + i \times \Delta f)t)$，其中 $f_0 + i \times \Delta f$ 是第 i 个载波频率。同时，内置麦克风以 48 千赫采集 $x(t)$，其中既包含超声反射 $x_u(t)$，也包含可听语音 $x_a(t)$。二者均源自相同的发音动作，因此无需外部同步即可实现自然的时间对齐。

超声波特征提取。我们使用轻量级有限脉冲响应（FIR）带通滤波器分离各个超声波载波，并采用余弦–正弦（I/Q）混频，将连续波信号解调至基带。这样便能以采样级精度跟踪振幅和相位，从而捕捉与语音对齐的细微发音动作。对于每个载波 $x_{u_i}(t)$，同相分量和正交分量为：

$$I_i(t) = x_{u_i}(t)\cos(2\pi f_i t), \quad Q_i(t) = -x_{u_i}(t)\sin(2\pi f_i t). \quad (1)$$

应用移动平均抑制高频伪影后，我们恢复相位 $\phi_i(t)$ 和振幅 $A_i(t)$：

$$\phi_i(t) = \operatorname{atan}(Q_i(t), I_i(t)), \quad A_i(t) = \sqrt{I_i^2(t) + Q_i^2(t)}. \quad (2)$$

图 3a 展示了语音梅尔频谱图，以及第一个载波的相位和振幅，其变化反映了发音器官的运动。为提高鲁棒性，我们计算相位差异和振幅差异。相位差异通过多普勒效应捕捉发音器官速度，而振幅差异则捕捉反射体距离和方向的变化。图 3b 表明，这些特征能够区分合法超声波与语音同步伪造超声波，同时抑制静态信道效应。将各载波上的特征拼接后，可得到一个 12 通道特征图。紧凑型多分支时序网络在不同尺度上捕捉发音器官模式，随后通过扩张卷积建模更长范围的依赖关系。最终的投影层生成与音频潜在空间对齐的超声波潜在序列 $\mathbf{h}_u$。

可听语音特征提取与编码。与此同时，接收信号经过低通滤波，并使用 FIR 滤波和线性插值从 48 千赫重采样至 16 千赫，使音频与电话编解码器保持一致，同时最大限度地减少感知失真。音频处理流程有意设计得比超声波处理流程更简单，因为超声波提供主要的活体信号，而语音充当感知载体。重采样后的音频由受 AudioSeal [7] 和 Encodec [10], 启发的步进残差卷积网络进行编码，在提高通道维度的同时降低时间分辨率，从而生成紧凑的潜在表示 $\mathbf{h}_a$，并保留用于水印嵌入的语音音系和韵律结构。

通过超声波引导的 FiLM 实现跨模态融合。在保留语音作为主导载波，我们使用特征级线性调制（FiLM）通过超声序列对音频潜表示进行调制。给定 $\mathbf{h}_u$ 和 $\mathbf{h}_a$，两个时序 1×1 卷积 $g_{\text{scale}}(\cdot)$ 和 $g_{\text{shift}}(\cdot)$ 生成缩放和平移场：

$$\gamma = \sigma(g_{\text{scale}}(\mathbf{h}_u)), \quad \beta = g_{\text{shift}}(\mathbf{h}_u), \quad (3)$$

其中，$\sigma(\cdot)$ 表示逐元素 Sigmoid 函数。融合潜表示$\mathbf{z}$ 的计算方式如下：

$$\mathbf{z} = (1+\gamma)\odot\mathbf{h}_a + \beta, \quad (4)$$

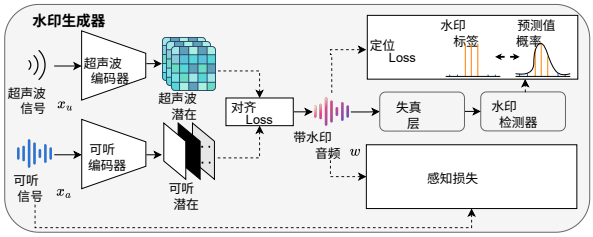


图 4：生成器-检测器训练流程。

其中，$\odot$ 表示哈达玛积。

水印解码器。融合潜表示由轻量级转置卷积网络解码，以重建音频载体并生成水印。

3.2 对抗对齐训练

我们设计了一种对抗对齐训练框架，迫使生成器—检测器组合学习语音与超声发音动作之间的同步性。

概述。图4总结了我们的方法。成对的可听语音和超声信号首先输入水印生成器，由其生成带水印音频。同步失真层会对带水印音频施加编解码器压缩、回声和混响，确保模型能够接触到真实的电话通话劣化和欺骗式扰动。随后，生成的带水印音频由水印检测器处理，并输出帧级水印概率。在训练过程中，语音被视为一种多模态事件，其中声波与感知到的超声信号保持同步。模型经过训练后，只有这种协同演变的模式才会触发水印激活，而由重放攻击、不匹配的动作、无声模仿或信道失真造成的任何同步性破坏都会受到明确惩罚。

水印检测器。检测器经过训练，可区分同步语音与超声波对和不匹配的配对，并输出帧级水印概率。具体而言，检测器使用 Encodec 风格的编码器提取紧凑的潜在表示。经过训练的上采样模块可恢复时间分辨率。1×1 卷积头会在每个时间步生成两个逻辑值。Softmax 将它们转换为帧级概率。训练过程中，对齐的语音与超声波对会产生较高的水印概率，反映出与语音相结合的真实发音动作。相比之下，欺骗样本会破坏这种结合，产生较低的水印概率。在推理时，检测器通过计算水印概率超过检测阈值的帧所占比例，将帧级输出汇总为全局置信度分数。

生成器–检测器联合优化。 我们以端到端的方式联合训练水印生成器和检测器。水印生成器同时处理真实音频（具有对齐的音频–超声波对）和伪造音频（具有未对齐的配对），而检测器则学习区分两者。这种联合优化确保水印编码能够捕捉时间同步信息，使检测器在推理时能够可靠地验证该信息。整体训练损失为：

$$\mathcal{L} = \lambda_{\text{loc}}\,\mathcal{L}_{\text{loc}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}} + \lambda_{\text{percept}}\,\mathcal{L}_{\text{percept}}. \tag{5}$$

• **定位损失：** 定位损失 Lloc 通过真实标签监督检测器的帧级预测。在训练集期间，真实的带水印音频和欺骗的带水印音频会被联合采样并并行优化，使模型能够学习刻画真实语音的跨模态对应关系，同时将其与不匹配的配对区分开来。令$\mathbf{p}^+$ 和$\mathbf{p}^-$ 分别表示检测器对真实样本和欺骗样本的 Softmax 输出。该损失通过负对数似然（NLL）计算：

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{NLL}}(\log \mathbf{p}^+, y^+) + \mathcal{L}_{\text{NLL}}(\log \mathbf{p}^-, y^-), \tag{6}$$

其中， y^+ 和 y^- 是真实标签： $y^+ = 1$ 表示真实的带水印音频（接受），而 $y^- = 0$表示欺骗的带水印音频（拒绝）。

• **对齐损失：** 不同于基于分类的方法，我们对学习到的隐藏表示采用对比目标。令$\mathbf{h}_a$、$\mathbf{h}_u^+$和$\mathbf{h}_u^-$分别表示由水印编码器生成的经 ℓ_2 归一化的隐藏特征。具体而言，它们分别从可听语音、对齐的超声波和未对齐的超声波中提取。对齐损失 Lalign 鼓励同步的音频—超声波配对具有相似的表示，同时将未对齐的配对相互推离：

$$\begin{aligned}\mathcal{L}_{\text{align}} = &\underbrace{(1 - \cos(\mathbf{h}_a, \mathbf{h}_u^+))}_{\text{alignment term}} \\ &+ \underbrace{\max(0, \cos(\mathbf{h}_a, \mathbf{h}_u^-) - \delta)}_{\text{misalignment term}},\end{aligned} \tag{7}$$

其中，$\cos(\cdot,\cdot)$ 表示余弦相似度， δ 是一个边界超参数。对齐项会最小化真实音频–超声波配对之间的角距离，而基于合页的未对齐项则强制未对齐配对的相似度低于 δ。这种非对称设计可防止模型坍塌，并确保语音–超声波同步。

• **感知损失：** 为确保不可感知性，我们采用多种语音质量损失。总感知损失 Lpercept 是互补的频谱

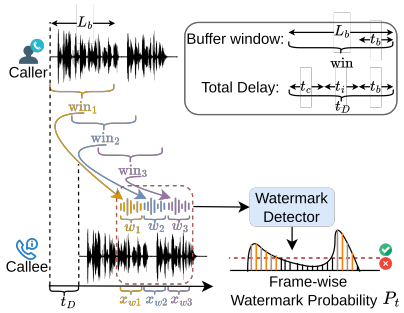


图 5：实时通话的滑动窗口水印生成与检测。

和基于能量的目标的加权和：

$$\begin{aligned}\mathcal{L}_{\text{percept}} = &\lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \lambda_{\text{msspec}}\mathcal{L}_{\text{msspec}} \\ &+ \lambda_{\text{sisnr}}\mathcal{L}_{\text{sisnr}} + \lambda_{\text{loud}}\mathcal{L}_{\text{loud}}.\end{aligned} \tag{8}$$

具体而言，我们纳入了：(i) 梅尔频谱图损失 L_{mel}，作用于按感知尺度表示的频率表征；(ii) 多分辨率短时傅里叶变换损失 L^{msspec}，用于捕捉多个时频分辨率下的频谱保真度；(iii) 尺度不变信噪比（SI-SNR）损失 L_{sisnr}，用于衡量与信号振幅无关且经过能量归一化的失真；以及 (iv) 时频响度比损失 L_{loud}，用于惩罚原始音频与带水印音频在各频带上感知显著的能量偏差。这种多方面的目标确保嵌入式水印保持不可听，同时保留原始语音的声学特征。

端到端训练与检测。 训练期间，生成器接收原始波形 x 并生成水印信号 w和提取出的语音 x_a。带水印的音频 x_w构造如下： $x_w = x_a + \alpha w$，其中 α 是用于控制水印强度的标量。定位、对齐和感知损失产生的梯度会通过生成器—检测器流水线反向传播。我们监测梯度范数，并采用带平衡机制的梯度裁剪，防止任何目标主导训练。这些技术共同实现了稳定而高效的训练，使系统在真实声学环境和对抗条件下兼具高准确率、鲁棒性和不可感知性。

3.3 实时实现

我们设计了一种实时流式架构，使生成器和检测器能够在实时通话期间持续运行，同时满足严格的延迟和计算资源约束。与离线水印不同，通话中水印必须在语音到达时立即进行处理，同时尽可能减少缓冲延迟，并对自然变化保持鲁棒。我们的解决方案

算法 1：实时水印生成

输入：可听信号流 $S(t)$、缓冲区长度 L_b、步长 t_b
输出：连续的带水印音频流 $W(t)$
初始化一个空缓冲区 $B \leftarrow []$;
当通话处于活动状态**时**
  将来自 $S(t)$ 的输入采样追加到 B;
  如果 $|B| \geq L_b$，**则**
    $A_{\text{wm}} \leftarrow$生成器(B);
    $A_{\text{send}} \leftarrow A_{\text{wm}}[L_b - t_b : L_b]$;
    将 A_{send} 传输至被叫设备;
    丢弃 B 中的前 t_b 个采样;
  end
end

在两端均采用滑动窗口流水线，协调缓冲、窗口步进以及基于均方根（RMS）能量的语音门控，以减少计算量，同时保持音频流畅传输。如图 5 所示，生成器以固定增量处理音频流，生成较短的带水印片段；检测器则将其滑动窗口与生成器的延迟预算对齐，从而提供低延迟的逐帧置信度估计。

滑动窗口水印生成。 如算法 1 所示，呼叫方设备会将输入的语音样本累积到长度为 L_b 的缓冲区中。在初始预热阶段，语音在水印生成开始前会直接传输，直至第一个缓冲区填满，从而为水印生成提供足够的上下文。随后将整个缓冲区输入生成器，但仅传输最后 t_b 个采样，即对应最新带水印语音的部分，其中t_b 表示步长。窗口随后向前移动 t_b，之后每个长度为 L_b个采样的窗口都会生成一个新的、长度为t_b个采样的带水印片段。

基于 RMS 的语音门控。 我们应用 0.05 的 RMS 阈值来检测语音并减少不必要的计算。RMS 低于 0.05 的片段被视为静音，并在不进行数字水印的情况下直接传递。这样可以提高效率，避免静音造成偏差，并保持可预测的延迟。**窗口对齐水印检测**。 在水印检测期间，我们采用相同的滑动窗口算法。当检测器缓冲区达到长度 L_b时，将对其进行处理，以生成逐帧置信度图。由于最早的部分对应于水印开始前传输的音频，因此会将这些帧排除在置信度计算之外。设 $n = \lfloor L_b/t_b \rfloor$ 表示缓冲区完全带水印前的滑动步数。在新通话会话的前 n 个步长中，只有最新的 t_b 带水印区域会对置信度图做出贡献，而长度为 $L_b - n \times t_b$ 的初始无水印片段将被排除。经过 n 个步长后，将在整个缓冲区上计算置信度。检测期间采用相同的 RMS 阈值，将静音帧排除在检测之外。

手机型号	OS	超声增益（分贝）
Motorola edge 5G UW 1	Android 13	0.47
Motorola edge 5G UW 2	Android 13	0.39
Google pixel 9	Android 16	3.64
Sony Xperia XA1 Plus	Android 7	3.72
iPhone 17 Pro	iOS 26.2.1	5.78
iPhone 16 Pro Max	iOS 18.6.2	4.94
iPhone 14 Pro Max	iOS 26.2.1	5.27
iPhone 13 Mini	iOS 26.2.1	4.20

表 2：用于数据收集的智能手机。

置信度估计。 对于每个新填充的窗口，检测器都会生成一个即时置信度 ，反映模型的输出。由于该值可能因瞬态噪声或失真而波动，我们在移除最大值和最小值后，基于最近的 N 个即时置信度计算平均置信度 。这种稳定化处理可以抑制离群值并减少误报，同时保留可表明攻击的持续置信度下降。检测器应用会实时显示这两种置信度，并跟踪通话期间最低的平均置信度，以指示操纵行为。**延迟建模与同步**。 为补偿水印生成和传输延迟，被叫方将播放延迟 $t_D = t_b + t_i + t_c$， 其中， t_i 为生成器推理时间， t_c 为通话传输延迟。在此延迟之后，检测将采用相同的窗口大小 L_b 和步长 t_b进行，从而与生成器保持同步。

3.4 跨平台部署

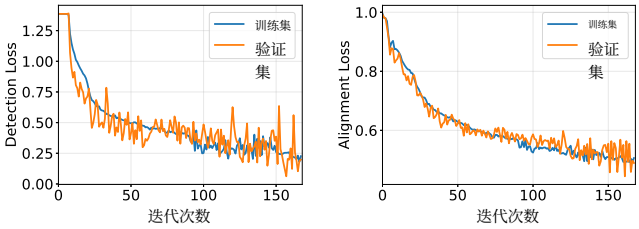
我们通过原生 iOS 和 Android 应用程序，在商用移动设备上端到端部署系统。生成器和检测器模型会导出为兼容平台的格式，从而无需外部服务器即可进行设备端推理。设备之间通过基于 IP 的通话建立连接，将带水印语音从呼叫方设备传输到被叫设备。

跨平台设备端推理。 为确保兼容性，我们将生成器和检测器转换为轻量级格式。Android 应用使用 NCNN [24] 进行模型加速，iOS 应用则使用针对实时音频处理优化的 Core ML。两款应用均在设备端执行推理，无需云服务，从而保护隐私。由于 torchaudio 等标准信号处理库在移动平台上不可用或效率低下，我们实现了兼容移动设备的替代方案，用于 FIR 滤波、线性重采样和模型导出，使其能够在 Android 和 iOS 上原生运行，无需额外硬件。为促进可复现性，我们将在论文发表后开源这些移动端实现。

4 评估

4.1 实验设置

数据集。 我们的数据集包含 2,000 多条话语，其中包括来自 22 位说话人的 1,700 多段真实录音、150 个文本转语音样本和 150 个语音转换样本。我们按 80%/10%/10% 的比例将数据划分为训练集、验证集和评估集。每个数据划分中的欺骗实例均独立构造，以防止数据泄漏。正样本由真实录音生成，包括干净的带水印语音和通话失真后的带水印语音。所有信号均切分为时长两秒、采样率为 48 千赫的采样片段。



（a）检测损失。 （b）对齐损失。

图 6：随迭代次数变化的训练损失。

真实数据集。我们开发了定制的 Android 和 iOS 应用程序，可同时发射超声感测信号，并录制可听语音和超声反射，从而无需外部硬件即可实现同步多模态采集。在数据收集过程中，参与者朗读一组固定的十句话，这些句子代表常见的电话诈骗场景，包括紧急账户警告、付款请求和法律威胁。我们从 22 名参与者处收集了 1,700 多段真实录音，使用的设备包括四部 iPhone 和四部 Android 智能手机，覆盖四种环境（家庭、办公室、实验室和室外环境）以及两种通话模式（扬声器和听筒）。如表 2 所示，这些设备涵盖不同的操作系统，超声增益范围为 0.39 分贝至 5.78 分贝。训练时将不同设备、环境和通话模式下的数据合并，以提升模型对设备异质性和使用差异的鲁棒性。

合成数据集。 为确保不同攻击类型的语言内容一致，我们对真实语音、文本转语音和语音转换样本使用相同的语音脚本。对于每位说话人，我们选取一段具有代表性的真实录音并进行重采样，作为文本转语音语音克隆和语音转换的说话人参考音频。我们使用 WaveNet [34], Tacotron 2 [31], 和 FastSpeech [28], 生成 150 个文本转语音样本，并使用基于 VAE 的 [22],基于生成对抗网络的 [19],和基于扩散的模型 [6] 生成 150 个语音转换样本。语音脚本详见附录A。

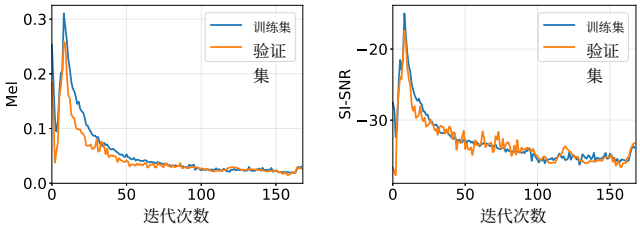
欺骗数据集。 数据划分后，我们在每个子集中独立构建欺骗实例。欺骗语音池包含 1,700 多条重放话语、150 条文本转语音话语和 150 条语音转换话语。我们将这些语音样本与超声波结合，具体方式遵循下文所述的攻击场景，以确保不同子集之间没有交叉
语音–超声配对发生。

攻击设置。 每个伪造语音由可听语音和超声波信号组成。场景*1*：仅音频欺骗。 提供不含超声波的伪造语音。场景*2*：随机超声欺骗。将伪造语音与不相关的超声波形配对。场景*3*：语音同步超声欺骗。 将伪造语音与来自另一位说出相同句子的扬声器的超声波配对。基于均方根的分段对齐话语起始点，从而产生时间同步但物理不匹配的语音—超声配对。场景*4*：自适应对手。 自适应攻击在训练后单独进行评估：通过模型推理或优化扰动现有的检测器输入，因此不需要额外的欺骗语料库。

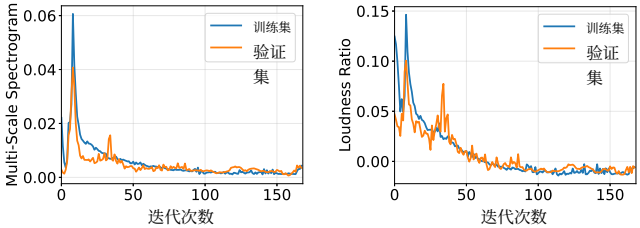
通话中失真。 为提升并评估模型在真实电话通话条件下的鲁棒性，我们在训练和评估过程中均加入通话中失真。对于每个带水印的语音样本，我们在检测前随机选择一种失真，包括低通滤波（5 千赫截止频率）、高通滤波（500 赫兹截止频率）、带通滤波（300 赫兹–8 千赫）、通过 32 千赫进行上采样和下采样、128 kbps 的 MP3 和 AAC 压缩、粉红色（$\sigma = 0.01$）噪声和高斯（$\sigma = 0.001$）噪声、± 20分贝增益变化、窗口大小为 2–10 个采样点的移动平均平滑、音量为 0.1–0.5 且延迟为 0.1–0.5 秒的回声，以及使用 16 个量化级别进行 Encodec压缩后再进行重采样。

训练设置。 我们使用配备 32 GB 内存的 NVIDIA RTX 5090 GPU，基于 AudioCraft 框架 [7], 进行训练，该框架已针对我们的数据加载和训练流程进行了适配。我们使用 Adam，参数为 $\beta = (0.5, 0.9)$，初始学习率为 5×10^{-4} 并采用自适应调度，最大范数为 1.0 的梯度裁剪，以及衰减率为 0.99 的指数移动平均。训练采用 FP32精度，批量大小为 16，共进行 700k 步。损失权重设置为 $\lambda_{\text{loc}} = 10.0$、$\lambda_{\text{align}} = 10.0$、$\lambda_{\text{mel}} = 1.0$、$\lambda_{\text{msspec}} = 2.0$、$\lambda_{\text{sisnr}} = 1.0$和 $\lambda_{\text{loud}} = 10.0$，以平衡检测准确率和感知质量。训练完成后，生成器和检测器分别部署在商用智能手机上，用于呼叫方和被叫方的设备端推理。

评估指标。我们从检测性能和语音质量两方面评估超声波水印。对于检测，我们报告准确率（Acc）、真正率（TPR）、假正率（FPR）、攻击成功率（ASR），以及受试者工作特征（ROC）曲线和曲线下面积（AUC）。准确率衡量整体分类正确性，TPR 衡量被正确检测出的带水印样本所占比例，FPR 衡量被错误分类为带水印的非水印样本所占比例，而 ASR 衡量攻击



(a) 梅尔频谱图。 **(b) SI-SNR。**



(c) 多尺度频谱图。 **(d) 响度比。**

图7：各迭代次数下的质量损失。

尝试被成功接受为合法操作的比例。ROC 曲线刻画不同检测阈值下 TPR 与 FPR 之间的权衡，而 AUC 则概括整体检测性能。对于语音质量，我们报告感知语音质量评价（PESQ）以衡量感知质量，短时客观可懂度（STOI）以衡量语音可懂度，以及尺度不变信噪比（SI-SNR）以衡量重建质量；数值越高，表示性能越好。

4.2 模型训练性能

我们以端到端方式联合训练水印生成器和检测器。正样本是由对齐的语音—超声配对生成的带水印音频，负样本则包括未加水印音频，以及由未对齐的配对生成的带水印音频。训练目标对水印可检测性、攻击鲁棒性和感知音频质量进行联合优化。

损失收敛。 图6显示，检测损失和对齐损失持续下降，且训练—验证差距较小，表明优化稳定且泛化能力良好。

音频质量。 图7显示，质量相关损失（梅尔频谱图、SI-SNR、多尺度频谱和响度）持续收敛。初期损失上升是因为检测和对齐目标占主导地位，导致扰动增大；随着可检测性提高，质量目标逐步降低感知失真。训练—验证差距较小，进一步表明泛化能力良好。

水印检测性能。 图8显示，真正率和真负率在训练集期间稳步提升，分别达到0.93和0.99。二者均远高于0.5的随机基线，表明能够有效区分正样本和负样本。

攻击类型	AUC	FPR
仅音频 +重放	1.00	0.01
仅音频 +合成	1.00	0.01
随机超声波 + 重放	1.00	0.00
随机超声波 + 合成	1.00	0.00
同步超声 + 重放	**0.84**	**0.37**
同步超声 + 合成	**0.88**	**0.30**
平均值	**0.95**	**0.12**

表 3：针对攻击类型的检测结果（$\alpha = 1.0$）。

4.3 欺骗攻击下的性能

我们评估了超声波水印在三种欺骗攻击场景下的表现。在每种场景中，可听输入要么被重放，要么使用文本转语音或语音转换进行合成，共形成六种攻击类型。表 3 总结了检测性能，图 9 展示了水印强度的影响 α。

- **仅音频攻击：**对于重放和合成攻击，超声波水印均实现了近乎完美的检测，两种情况下的曲线下面积均为 1.00，假正率均为 0.01。图 9 显示，提高 α 可增强可分性，其中 $\alpha \geq 0.8$ 时可实现近乎完美的检测。
- **随机超声攻击：**当无关超声波与伪造的可听语音结合时，无论重放还是合成，曲线下面积均为 1.00，且没有假阳性。如图 9 所示，即使在 $\alpha = 0.2$时，检测仍然稳健，这表明无关的超声运动不会产生有效水印特征。

- **同步超声攻击：**同步超声带来了最严峻的挑战，使重放和合成的曲线下面积分别降至 0.84 和 0.88，假阳性率分别为 0.37 和 0.30。图 9 显示，随着 α 增大，检测性能显著提升：在 $\alpha = 0.2$ 时，性能接近随机猜测；而在$\alpha = 0.8$ 和 1.0 时，语音音频质量仅有限退化，但可分性显著增强。这些结果表明，当水印强度足够高时，伪造语音与模仿的发音动作之间残留的不一致仍可被检测到。

4.4 自适应攻击下的性能

我们进一步评估 UltrasoundWatermark 对自适应对手的防御能力，这些对手会优化输入以规避水印检测。与欺骗攻击不同，这些对手会利用对检测目标的了解，通过结构化扰动降低水印置信度。

自适应对手模型。 我们考虑三种能力递增的攻击者：(i) 黑盒，具有查询访问权限

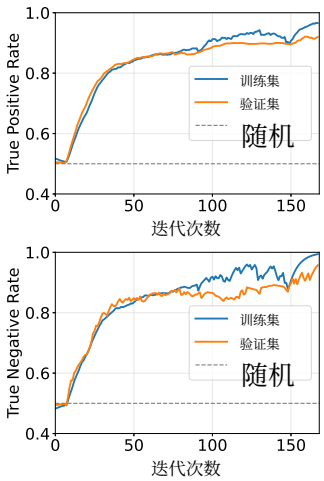


图 8：训练真正例率/真负例率。

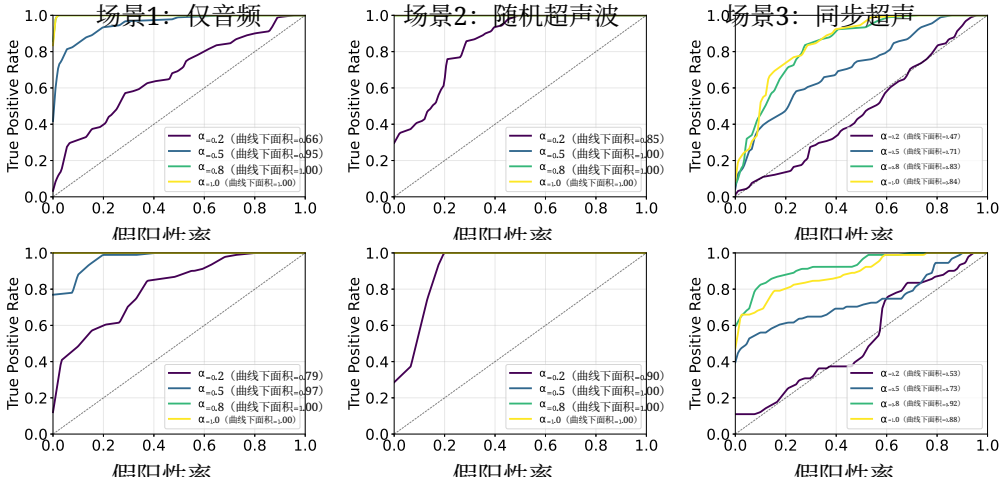


图 9：重放攻击（上）和合成攻击（下）的 ROC曲线。

攻击方法	能力 检测器级别	检测器查询次数	检测器架构	检测器权重	实时可行
高斯噪声	基线	✗	✗	✗	✓
查询迁移 黑盒		✓	✗	✗	✗
实时语音转换	黑盒	✗	✗	✗	✓
配方迁移 灰盒		–	✓	✗	✗
APGD	白盒	–	✓	✓	✗

表4：自适应攻击的攻击者能力。

但不知道模型参数或梯度；(ii)灰盒，了解架构、训练流程和损失函数，但不知道训练后的权重；以及 (iii) 白盒，可完全访问模型和梯度。我们评估了涵盖这些能力的五种攻击：高斯噪声、查询蒸馏迁移、带有真实发音动作的实时语音转换、配方感知迁移，以及白盒梯度优化。表4总结了它们的能力与实时可行性。我们使用0.5的检测阈值，并将攻击成功率定义为被接受为合法的欺骗性输入所占的比例。

- **高斯噪声：** 作为非自适应基线，我们添加标准差为 $\sigma \in [5\times10^{-4}, 2\times10^{-1}]$的零均值白高斯噪声。在所有噪声水平下，攻击成功率均保持为 0，表明无结构扰动不会导致误接受。
- **查询蒸馏迁移：**对于黑盒攻击，攻击者使用有限的检测器查询次数（$Q \in \{100, 200, 300, 500\}$）通过分数匹配训练替代模型，并采用投影梯度下降（PGD）生成可迁移扰动。在所有查询预算下，攻击成功率均低于 0.05。即使在 $Q = 500$时，替代模型与目标模型输出之间的皮尔逊相关系数仍低于 0.16，表明其对目标决策边界的近似效果较差，可迁移性有限。
- **保留真实发音动作的实时语音转换：**攻击者使用实时语音转换将自己的语音转换为目标语音，同时保留实时超声发音动作。表5显示，速度最快的模型（YourTTS&VC [3]）相对于水印检测器的推理延迟，产生了 3.6× 的延迟开销；而 FreeVC [18] 和 Seed-VC [17] 分别产生 6.9×和 13.4×。在实时通话中，这会带来显著开销以及语音—超声错位。
- **配方感知迁移：**对于灰盒攻击，攻击者使用不同的权重训练一个相同的替代模型，并应用 PGD 生成可迁移扰动。如表6所示，在较小的扰动预算下（对于 $\varepsilon \le 2\times10^{-3}$，攻击成功率分别为 0.06 和 0.16），迁移攻击成功率仍然较低；但在 $\varepsilon = 10^{-2}$ 时达到 0.81，并在更大的预算下超过 0.85。因此，配方级知识主要在更强、且可能可感知的扰动下提升迁移能力。
- **白盒攻击：** 在完全访问模型的情况下，攻击者直接优化有界波形扰动 $\delta \in [-\varepsilon, \varepsilon]^T$。为排除梯度掩蔽的影响，我们评估了 Adam、PGD [23], 和 Auto-PGD（APGD） [8]。表6显示，不同优化器的攻击成功率保持一致，在$\varepsilon = 10^{-3}$ 时最多相差 0.04，在 $\varepsilon \ge 2\times10^{-3}$时相差 0.01。在较小的预算（$\varepsilon \le 10^{-3}$）下，攻击成功率仍相对较低（≈ 0.3），但随着 ε 增加而接近 1.0，表明足够强的白盒扰动能够绕过检测器。
- **实时可行性：**图 10 对比了攻击成功率和运行时间，并将运行时间归一化为我们系统的端到端推理时间。高斯噪声和查询蒸馏迁移的计算开销较低，但攻击成功率也较低；而有效的配方感知攻击和白盒攻击则会带来 $>10\times$ 运行时间开销。这种权衡限制了实时通话延迟约束下有效自适应攻击的可行性。

方法	合成（毫秒）	检测器（毫秒）	开销
YourTTS&VC	36.1	15.3	**3.6×**
FreeVC	85.3	15.3	**6.9×**
Seed-VC	178.6	15.3	**13.4×**

表5：实时语音转换攻击的运行时开销。

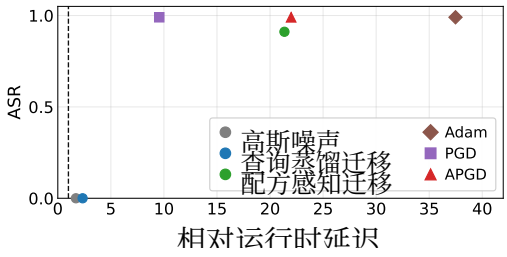


图10：对抗攻击的实时可行性。运行时间以端到端推理时间为基准进行归一化。

	白盒自动语音识别			灰盒自动语音识别	
ε	Adam	PGD	APGD	替代模型	迁移
5e-4	0.07	0.07	0.07	–	–
1e-3	**0.27**	**0.31**	**0.31**	0.08	**0.06**
2e-3	0.75	0.75	0.76	0.35	**0.16**
5e-3	0.97	0.98	0.98	0.88	0.61
1e-2	0.99	0.99	0.99	1.00	**0.81**
2e-2	0.99	0.98	0.99	0.99	0.86
5e-2	1.00	1.00	1.00	0.92	0.91

表6：白盒和灰盒攻击在不同扰动预算下的攻击成功率 ε。

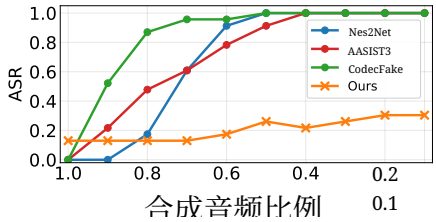


图11：部分合成攻击下的攻击成功率。

4.5 与活体检测的比较

我们将系统与近期的**仅音频活体检测** 方法进行了比较。先前的工作会从可听语音中检测欺骗伪迹 [38, 44], , 但难以应对包含真实和合成片段的部分合成话语。相比之下，超声波水印会验证语音与超声波发音动作之间的物理耦合。合成片段缺少相应的实时发音动作，因此会破坏这种跨模态一致性。

为构造部分合成攻击，我们从一段真实的 2 秒话语开始，其中语音与超声波同步。对于合成音频比例 $r \in \{0.1, 0.2、\ldots\ldots、1.0\}$，我们将一段连续的 $2r$ 秒可听语音替换为合成音频，同时保持超声波不变。例如，$r = 0.1$ 会替换 0.2 秒的语音，同时保留其余 1.8 秒的真实音频。任何包含 $r > 0$ 的话语都视为攻击，而攻击成功率衡量被错误地判定为真实语音的话语所占比例。图 11 显示，随着合成部分占比降低，仅音频方法变得愈发容易受到攻击。尽管这些方法能够可靠地拒绝完全合成的话语，但在 $r = 0.5$时，Nes2Net [20]和 CodecFake [39] 的攻击成功率达到 1.0，而 AASIST3 [2]达到 0.91。在 $r = 0.1$时，三种方法的攻击成功率均达到 1.0。相比之下，超声波水印在不同合成音频比例下均保持显著较低的攻击成功率，在 $r = 0.5$ 时为 0.26，在 $r = 0.1$时为 0.30。这些结果表明，语音–超声波耦合有助于检测部分合成攻击，因为合成语音无法重现相应的实时发音动作。

4.6 真实通话鲁棒性

我们在两种实际的信号退化情况下评估超声波水印：**通话中失真**和**VoIP 编解码器转码**。总体而言，超声波水印在常见失真和高保真编解码器下仍具有鲁棒性，但强力滤波和压缩会显著降低检测性能。

通话中失真。表 7 评估了常见通话中失真条件下的水印检测性能。在干净条件下，超声波水印的曲线下面积为 1.00，真正率为 0.93，假正率为 0.01。在重采样、MP3压缩、加性噪声、幅度缩放、平滑和低通滤波条件下，性能几乎没有变化。回声和 Encodec 会导致中等程度的性能下降，使真正率分别降至 0.80 和 0.58，同时曲线下面积仍保持在 0.95 以上。更强的带通滤波和高通滤波会显著降低检测性能，曲线下面积分别降至 0.76 和 0.54。AAC压缩是最具挑战性的失真，使曲线下面积降至 0.46，真正率降至 0.01。这些结果表明，超声波水印能够耐受常见的通话失真，但对于会显著改变低振幅水印分量的变换仍然较为敏感。

VoIP 编解码器转码。我们进一步评估了水印在具有代表性的 VoIP 编解码器下的鲁棒性，包括 G.711、G.722、Opus 和 Speex。编解码器实验采用了独立于通话中失真实验的评估流程和工作点，因此基线真正率和假正率有所不同。表 8 显示，G.711 和高比特率 Opus（96 kbps）基本保留了真正率，分别达到 0.98 和

组别	失真	曲线下面积	真正率	假正率
干净	None	1.00	0.93	0.01
鲁棒	重采样 / MP3	1.00	0.93	0.01
	加性噪声	1.00	0.93	0.01
	幅度缩放	1.00	0.93	0.01
	平滑 / 低通	1.00	0.92	0.01
中等	Echo	0.97	0.80	0.02
	Encodec	0.95	0.58	0.00
具有挑战性	带通滤波器	0.76	0.29	0.03
	高通滤波器	0.54	0.02	0.04
	AAC压缩	0.46	0.01	0.01

表 7：通话中失真下的鲁棒性。

编解码器	Rate	比特率	TPR	FPR
基线	–	–	0.98	0.19
G.711 μ/A-law	8 千赫	64 kbps	0.98	0.19
Opus（96 kbps）	16 千赫	96 kbps	0.97	0.20
G.722	16 千赫	64 kbps	0.45	0.06
Opus（48 kbps）	16 千赫	48 kbps	0.08	0.50
Opus（16 kbps）	16 千赫	16 kbps	0.07	0.45
Speex	16 千赫	16 kbps	0.06	0.01

表8：VoIP 编解码器转码下的鲁棒性。

0.97。相比之下，G.722 和较低比特率的 Opus 会大幅降低真正率，其中 48 kbps 的 Opus 真正率为 0.08，假正率为 0.50。16 kbps 的 Speex 同样会将真正率降至 0.06。这些结果表明，水印不可感知性与对编解码器压缩的鲁棒性之间存在权衡：超声波水印在保真度较高的编解码器下仍然有效，而强压缩则会大幅削弱水印。

4.7 感知质量

客观评估。 我们首先评估水印嵌入是否会降低音频质量。带水印语音的 PESQ 得分为 3.99，表明其具有较高的感知质量。SI-SNR 为 25.6 分贝，表明信号失真较低；STOI 得分为 0.922，则证实语音可懂度基本得以保留。

用户研究。 我们进一步开展用户研究，以评估在真实电话通话场景中进行数字水印的感知影响。我们采用双智能手机设置来模拟实时通话：一台设备生成干净或带水印语音，另一台设备用于聆听。水印强度保持固定，所有实验均在安静的室内环境中使用商用智能手机进行。八名参与者完成被试内研究。他们聆听句子，其中包含

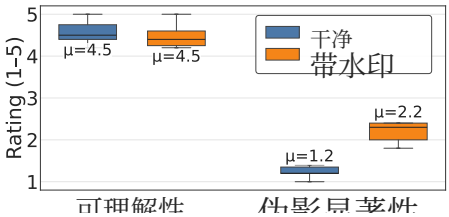


图12：水印的主观感知。

环境	Acc.	信噪比（分贝）
办公室	0.97	27.46 ± 6.4
室外环境	0.98	26.40 ± 3.4
Lab	1.00	29.15 ± 6.2
Home	0.95	28.99 ± 5.8
总计	**0.97**	**27.57 ± 6.1**

表 9：不同环境下的检测性能。

干净和带水印样本，这些样本以随机顺序交错呈现，且参与者事先不知道其条件。每听完一个句子，参与者便使用 1–5 李克特量表对可理解性和伪影显著性进行评分。如图 12 所示，干净语音和带水印语音的可理解性均保持较高水平且相当，而伪影显著性仍然较低。这些结果表明，我们的系统在保持语音可懂度的同时，仅引入了极小的感知退化。

4.8 自然变化下的性能

我们评估了超声波水印在日常通话中遇到自然变化时是否依然稳健，包括不同的声学环境、智能手机摆放位置、通话模式、用户和测量日期。

跨环境。 我们在四种具有代表性的环境中评估超声波水印：办公室、户外、实验室和家庭。这些场景涵盖了背景噪声、混响和设备摆放位置方面的自然变化。如表 9 所示，所有环境中的检测准确率均高于 0.95，总体准确率为 0.97。不同环境下的信噪比也相近，范围为 26.40 至 29.15 分贝。这些结果表明，超声波水印在多种真实世界声学环境中均能保持稳定的检测性能。

跨天与不同智能手机放置位置。 我们通过一项为期八天、由五名参与者参加的研究，进一步评估水印随时间的稳定性，以及智能手机摆放位置自然变化时的稳定性。每位参与者在不同日期各完成一次实验。端到端系统以原生 iOS 和 Android 应用的形式运行，生成器与检测器通过实时 IP 通话进行通信。每次实验中，参与者站在校准棋盘格前，

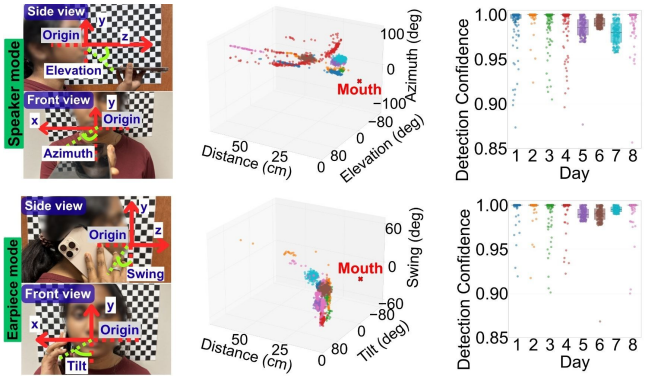


图13：扬声器和听筒模式下跨天及不同智能手机摆放位置的长期性能。

自然地以扬声器模式或听筒模式手持智能手机，如图 13（左）所示。生成器应用会实时嵌入水印，另一台设备则运行检测器应用。正面和侧面摄像头会记录每次实验。我们将智能手机位置定义为从嘴部中心到手机底边中心的三维向量；在扬声器模式下，该向量由距离、仰角和方位角表示，在听筒模式下则由距离、倾斜角和摆动角表示。

在所有参与者、日期、通话模式和观测到的手机位置下，均成功检测到水印。图13（中）展示了每100毫秒采样的位置，并按日期以不同颜色编码；图13（右）则显示了不同日期和模式下始终较高的批次检测置信度。结合环境结果，这些发现表明，该系统能够鲁棒地应对声学条件、智能手机摆放位置、用户行为和时间方面的自然变化。

4.9 设备端推理时长

表10报告了对于每个200毫秒音频片段，水印生成器和检测器在商用智能手机上的平均推理时长。生成器占据了主要计算开销，而检测器产生的开销相对较低。在较新的旗舰设备上，生成器需要15–24毫秒，检测器则少于8毫秒。即使在较旧的设备上，每个片段的总推理时长仍低于200毫秒，从而支持在商用智能手机上实时运行。

5 讨论

双水印攻击下的局限性。在这种攻击中，攻击者重放目标说话人已经带水印的语音，并使用伪造超声波进行第二次水印处理。随机超声波会被完全拒绝（攻击成功率 = 0，曲线下面积 = 1.00），而语音同步超声波的攻击成功率为 0.46，并将曲线下面积降至 0.89。

手机型号	生成器（毫秒）	检测器（毫秒）
Samsung Z Fold 7	19.27	1.28
Samsung Z Flip 7	62.35	4.32
Google Pixel 9	71.72	7.31
Motorola Edge（2020）	160.65	21.74
Motorola Edge 5G UW	148.48	19.39
ASUS Zenfone 8	134.15	10.58
iPhone 14 Pro Max	23.54	7.21
iPhone 16 Pro Max	18.41	5.88
iPhone 17 Pro	15.28	4.72

表10：不同智能手机上的设备端推理时间。

这揭示了一个局限：攻击者可能获取目标的带水印语音，并生成时间对齐的超声波。在每个水印中嵌入新的会话专用随机数可以缓解这一威胁。

失真情况下的局限性。尽管 Ultrasound Watermark 能够鲁棒地应对常见的通话失真和编解码器，但激进的频谱滤波和心理声学压缩会大幅降低检测效果。带通滤波和高通滤波会去除水印分量，而 AAC 和低比特率 Opus 压缩则会通过感知信号裁剪显著削弱水印。这些结果凸显了低能量水印固有的不可感知性与鲁棒性之间的权衡。未来的工作将探索多频带和自适应嵌入策略。

跨说话人干扰的局限性。在与语音同步的超声波攻击下（表3），假阳性率升高至 0.30–0.37，这表明不同说话人的超声波发音模式存在一定相似性，从而偶尔导致误接受。这说明说话人无关表征存在局限。潜在的缓解措施包括采用说话人条件检测、轻量级说话人嵌入以及更具区分性的特征学习。

6 结论

我们提出了超声波水印，这是一种实时、智能手机原生的音频水印系统，通过超声感测将语音与同步的生理发音动作绑定。通过联合利用可听语音和超声发音动作，超声波水印为实时通话中的语音真实性验证提供了物理依据。在不同用户、设备、环境、通话条件和攻击场景下进行的大量评估表明，该系统能够鲁棒地检测重放语音和合成语音，同时保持感知质量，并支持实时设备端运行。这些结果证明，语音—运动耦合可作为一种实用防御手段，在实时通信中抵御语音冒充攻击。

致谢
我们使用了生成式人工智能工具进行语言润色。

参考文献
[1] 苹果公司。2024年。FaceTime 要求。https://support.apple.com/guide/facetime/facetime-requirements-fctm35515/mac。访问日期：2026-01-29。[2] Kirill Borodin、Vasiliy Kudryavtsev、Dmitrii Korzh、Alexey Efimenko、Grach Mkrtchian、Mikhail Gorodnichev 和 Oleg Y. Rogov。2024年。AASIST3：使用 KAN增强型AASIST、基于自监督学习特征和额外正则化的语音深度伪造检测，用于 ASVspoof 2024挑战赛。arXiv: 2408 17352 [ ] 2408 17352 。cs.SD[3] https://arxiv.org/abs/ 。Edresson Casanova、Julian Weber、Christopher D Shulby、Arnaldo Can- dido Junior、Eren Gölge 和 Moacir A Ponti。2022年。Yourtts：面向所有人的零样本多说话人文本转语音和零样本声音转换。在 国际机器学习会议 上。PMLR，2709–2720。[4] Guangyu Chen、Yu Wu、Shujie Liu、Tao Liu、Xiaoyong Du 和 Furu Wei。2023年。Wavmark：用于音频生成的数字水印。*arXiv* 预印本 *arXiv:2308.12770*（2023年）。[5] Zhuxin Chen、Weibin Zhang、Zhifeng Xie、Xiangmin Xu 和 Dong- peng Chen。2018年。用于自动重放欺骗攻击检测的循环神经网络。在 *2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）（加拿大艾伯塔省卡尔加里）上。IEEE Press，2052–2056。doi:10.1109/ICASSP.2018.8462644[6] Ha-Yeong Choi、Sang-Hoon Lee 和 Seong-Whan Lee。2023年。Diff-HierVC：采用鲁棒基频生成和掩码先验的基于扩散的分层声音转换，用于零样本说话人适配。arXiv:2311.04693 [eess.AS] https://arxiv.org/abs/2311.04693[7] Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre Défossez。2023年。简单且可控的音乐生成。在 第三十七届神经信息处理系统大会 上。[8] Francesco Croce 和 Matthias Hein。2020年。通过多样化无参数攻击的集成，可靠地评估对抗鲁棒性。arXiv: 2003 01690 [ ] 2003 01690 。cs.LG[9] 2024 https://arxiv.org/abs/ 。DeepMind。。SynthID：人工智能生成图像的数字水印与识别。https://deepmind.google/technologies/synthid/ 访问日期：2024年9月10日。[10] Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。2022年。高保真神经音频压缩。*arXiv* 预印本 *arXiv:2210.13438*（2022年）。[11] 美国联邦贸易委员会。2025年。美国联邦贸易委员会重点介绍保护消费者免受冒充诈骗侵害的行动。https://www.ftc.gov/news-events/news/press-releases/2025/04/ftc-highlights-actions-protect-consumers-impersonation-scams。访问日期：2026-01-29。[12] João Figueiredo、Afonso Carvalho、Daniel Castro、Daniel Gon‚calves 和 Nuno Santos。2024年。关于完全由人工智能自动化的语音网络钓鱼攻击的可行性。*arXiv* 预印本 *arXiv:2409.13793*（2024年）。受控研究表明，人工智能代理能够令人信服地实施电话诈骗。[13] Hanqing Guo、Qiben Yan、Nikolay Ivanov、Ying Zhu、Li Xiao 和 Eric J Hunter。2022年。Supervoice：利用人类语音中的超声能量进行文本无关说话人验证。在 *2022*年*ACM*亚洲计算机与通信安全会议论文集中。1019–1033。[14] Hirokazu Kameoka 等。2018年。StarGAN-VC：利用星型生成对抗网络进行非并行多对多语音转换。在 *2018*年*IEEE*口语语言技术研讨会（*SLT*）上。[15] Jaehyeon Kong 等。2020年。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。在 *NeurIPS*上。[16] Gen Li、Huaili Zeng、Hanqing Guo、Yidong Ren、Aiden Dixon、Zhichao Cao 和 Tianxing Li。2024 年。PiezoBud：一种采用压电技术辅助的安全耳塞，具备实用的扬声器认证功能。载于第 *22* 届 *ACM* 嵌入式网络传感器系统会议论文集。564–577。[17] Jingyi li、Weiping tu 和 Li xiao。2022 年。FreeVC：迈向高质量的无文本单次语音转换。arXiv:2210.15418 [cs.SD] https: //arxiv.org/abs/ 2210 15418 。Jingyi Li、Weiping Tu 和 Li Xiao。[18] 2023 年。Freevc：迈向高质量的无文本单次语音转换。载于*ICASSP 2023-2023 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，1–5。[19] Yinghao Aaron Li、Ali Zare 和 Nima Mesgarani。2021 年。StarGANv2-VC：用于自然语音转换的多样化、无监督、非平行框架。arXiv:2107.10394 [cs.SD] https://arxiv.org/abs/ 2107 10394 。[20] Tianchi Liu、Duc-Tuan Truong、Rohan Kumar Das、Kong Aik Lee 和 Haizhou Li。2025 年。Nes2Net：一种由基础模型驱动的语音反欺骗轻量级嵌套架构。arXiv: 2504 05657 [ ] 2504 05657 。eess.AS[21] https://arxiv.org/abs/ 。Yixin Liu、Lie Lu、Jihui Jin、Lichao Sun 和 Andrea Fanelli。2025 年。XAt-tnMark：利用交叉注意力学习鲁棒音频水印。载于第四十二届国际机器学习会议。https: // openreview.net/forum?id=zuOKPHHfEr [22] Ziang Long、Yunling Zheng、Meng Yu 和 Jack Xin。2022 年。利用自注意力变分自编码器增强零样本多对多语音转换。arXiv:2203.16037 [cs.SD] https://arxiv.org/abs/2203.16037[23] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。2019 年。迈向能够抵抗对抗攻击的深度学习模型。arXiv:1706.06083 [stat.ML] https: //arxiv.org/abs/ 1706 06083 。Hui Ni 和 ncnn 贡献者。2017 年。[24] *ncnn*。https://github.com/ Tencent/ncnn[25] OpenAI。2024 年。在线了解我们所看到和听到内容的来源。https://openai.com/index/understanding-the-source-of-what-we-see-and-hear-online/ 访问日期：2024年9月10日。[26] Mike Pappas。2025 年。语音的隐秘威胁：为什么电话诈骗给消费者造成的损失比以往更多。博客文章。https://www.modulate.ai/blog/the-hidden-threat-of-voice-why-phone-scams-are-costing-consumers-more-than-ever Modulate.ai，2025 年 5 月 1 日。[27] Yang Qian 等。2019 年。AutoVC：仅使用自编码器损失实现零样本语音风格迁移。载于国际机器学习会议（*ICML*）。[28] Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu。2019 年。FastSpeech：快速、鲁棒且可控的文本转语音。神经信息处理系统进展32（2019）。[29] Robin San Roman、Pierre Fernandez、Alexandre Défossez、Teddy Furon、Tuan Tran 和 Hady Elsahar。2024 年。利用局部化水印主动检测语音克隆。*arXiv* 预印本 *arXiv:2401.17264*（2024）。[30] Irtaza Shahid 和 Nirupam Roy。2023 年。“这是我的总统在讲话吗？”实时录音中的语音防篡改。载于第 *21* 届年度移动系统、应用与服务国际会议论文集。219–232。[31] Jonathan Shen 等。2018 年。通过以梅尔频谱图预测结果为条件的 WaveNet 实现自然的 TTS 合成。载于*2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。[32] Mayank Kumar Singh、Naoya Takahashi、Weihsiang Liao 和 Yuki Mitsufuji。2024 年。SilentCipher：深度音频水印。载于*Interspeech*。

[33] Ke Sun 和 Xinyu Zhang。2021 年。UltraSE：使用超声波进行单通道语音增强。载于第 *27* 届年度移动计算与网络国际会议论文集。160–173。[34] Aaron Van Den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior、Koray Kavukcuoglu 等。2016 年。WaveNet：一种用于原始音频的生成模型。*arXiv* 预印本 *arXiv:1609.03499* 12（2016），1。[35] Xin Wang、Junichi Yamagishi、Massimiliano Todisco、Héctor Delgado、Andreas Nautsch、Nicholas Evans、Md Sahidullah、Ville Vestman、Tomi Kinnunen、Kong Aik Lee 等。2020 年。ASVspoof 2019：一个大规模的合成、转换和重放语音公共数据库。*Computer Speech & Language* 64（2020），101114。[36] Emily Warren 等。2025 年。人类难以有效检测由 AI 驱动的语音克隆。*Scientific Reports* 15（2025），94170。该研究表明，人类对合成语音的检测准确率较低。[37] WhatsApp LLC。2024 年。关于 WhatsApp 费用。https://faq.whatsapp. com/494964175875406。访问日期：2026-01-29。[38] Zhizheng Wu、Tomi Kinnunen、Nicholas Evans、Junichi Yamagishi、Cemal Hanilçi、Md. Sahidullah 和 Aleksandr Sizov。2015 年。ASVspoof 2015：首届自动说话人验证欺骗攻击与对策挑战赛。载于*Interspeech 2015*。2037–2041。doi:10.21437/Interspeech.2015-462[39] Yuankun Xie、Yi Lu、Ruibo Fu、Zhengqi Wen、Zhiyong Wang、Jianhua Tao、Xin Qi、Xiaopeng Wang、Yukun Liu、Haonan Cheng、Long Ye 和 Yi Sun。2024 年。Codecfake 数据集及其针对通用深度伪造音频检测的对策。arXiv:2405.04880 [cs.SD] https://arxiv.org/abs/ 2405 04880 。Junichi Yamagishi、Xin Wang、[40] Massimiliano Todisco、Md Sahidul-lah、Jose Patino、Andreas Nautsch、Xuechen Liu、Kong Aik Lee、Tomi Kinnunen、Nicholas Evans 等。2021 年。ASVspoof 2021：加速语音欺骗攻击与深度伪造检测的进展。*arXiv* 预印本 *arXiv:2109.00537*（2021）。[41] Jiangyan Yi、Ruibo Fu、Jianhua Tao、Shuai Nie、Haoxin Ma、Chenglong Wang、Tao Wang、Zhengkun Tian、Ye Bai、Cunhang Fan、Shan Liang、Shiming Wang、Shuai Zhang、Xinrui Yan、Le Xu、Zhengqi Wen 和 Haizhou Li。2022 年。ADD 2022：首届音频深度合成检测挑战赛。载于*ICASSP 2022 - 2022*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。9216–9220。doi:10. 1109/ICASSP43922.2022.9746939[42] Jiangyan Yi、Jianhua Tao、Ruibo Fu、Xinrui Yan、Chenglong Wang、Tao Wang、Chu Yuan Zhang、Xiaohui Zhang、Yan Zhao、Yong Ren、Le Xu、Junzuo Zhou、Hao Gu、Zhengqi Wen、Shan Liang、Zheng Lian、Shuai Nie 和 Haizhou Li。2023 年。ADD 2023：第二届音频深度伪造检测挑战赛。arXiv:2305.13774 [cs.SD] https://arxiv.org/abs/2305.13774[43] Linghan Zhang、Sheng Tan 和 Jie Yang。2017 年。听到你的声音还不够：一种基于发音动作的语音身份验证活体检测方法。载于 *2017* 年 *ACM SIGSAC* 计算机与通信安全会议论文集。57–71。[44] Lin Zhang、Xin Wang、Erica Cooper、Nicholas Evans 和 Junichi Yamagishi。2023 年。PartialSpoof 数据库及其用于检测话语中嵌入的短伪造语音片段的对策。*IEEE/ACM* 音频、语音与语言处理汇刊 31 2023 813 825 10 1109 2022 3233236（），–。[45] doi：。/taslp。。Qian Zhang、Dong Wang、Run Zhao、Yinggang Yu 和 Junjie Shen。2021 年。感知以实现听觉：使用声学信号为移动设备进行语音增强。*ACM* 交互式、移动、可穿戴与泛在技术论文集 5，3（2021），1–30。[46] X. Zong 等。2025 年。AudioMarkNet：用于深度伪造语音检测的音频水印。载于*USENIX* 安全研讨会论文集。

附录

A 语音脚本

我们的数据收集使用十个句子，代表常见的电话诈骗场景，包括紧急账户警告、付款请求和法律威胁，如图14所示。为了确保不同攻击类型中的语言内容一致，我们对真实语音、文本转语音语音和语音转换语音使用相同的脚本。每个句子均单独合成。

脚本1：这是关于您账户的一通紧急电话。**脚本2：**我们发现您的账户存在可疑活动。**脚本3：**如果您现在不采取行动，您的账户将被暂停。**脚本4：**我是美国国税局的Officer Paul。**脚本5：**您有一笔未偿余额，必须立即缴清。**脚本6：**不要挂断电话，否则您的案件将自动升级处理。**脚本7：**只要按照我的指示操作，我们现在就能解决这个问题。**脚本8：**出于安全考虑，请确认您的完整社会安全号码。**脚本9：**我们来电是因为您涉及一项法律事务。**脚本10：**已有针对您的逮捕令。

图14：用于数据生成的语音脚本。