arXiv:2305.05736v1 [cs.SD] 9 May 2023

VSMask：通过实时预测扰动防御语音合成攻击

Yuanda Wang 密歇根州立大学 东兰辛，密歇根州，美国 wangy208@msu.edu

Hanqing Guo 密歇根州立大学 东兰辛，密歇根州，美国 guohanqi@msu.edu

Guangjing Wang 密歇根州立大学 东兰辛，密歇根州，美国 wanggu22@msu.edu

Bocheng Chen 密歇根州立大学 东兰辛，密歇根州，美国 chenboc1@msu.edu

Qiben Yan 密歇根州立大学 东兰辛，密歇根州，美国 qyan@msu.edu

摘要

基于深度学习的语音合成技术能够生成类似人类的人工语音，已被用于深度伪造或身份盗窃攻击。现有防御机制会向原始语音音频中注入细微的对抗扰动，以误导语音合成模型。然而，优化对抗扰动不仅会消耗大量计算时间，还要求获取完整语音。因此，这些机制不适合保护实时语音流，例如语音消息或在线会议中的语音。本文提出了 VSMask，一种针对语音合成攻击的实时保护机制。不同于离线保护方案，VSMask 利用预测型神经网络，预测即将到来的流式语音所需的最有效扰动。VSMask 引入了一种适用于任意语音输入的通用扰动，以保护完整的实时语音。为尽量减少受保护语音中的音频失真，我们设计了基于权重的扰动约束，以降低所添加扰动的可感知性。我们在不同场景下全面评估了 VSMask 的防护性能。实验结果表明，VSMask 能够有效防御 3 种广受欢迎的语音合成模型。在 VSMask 的保护下，合成语音均无法欺骗说话人验证模型或人耳。在真实环境实验中，我们通过空中传输注入扰动，证明了 VSMask 能够成功保护实时语音。

CCS 概念

• **安全与隐私**；• **计算方法** → **机- 器学习**；

关键词

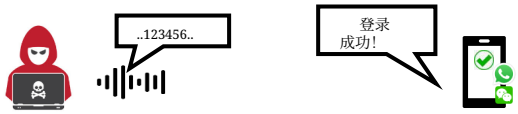
语音合成、实时对抗机器学习、隐私增强

1 引言

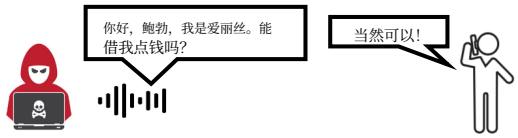
人工智能（AI）的最新发展推动了智能语音控制领域的研究。如今已经可以进-
几乎能完美模仿他人的语音，正如各种用于传播虚假信息和欺诈的“深度伪造”行动所展示的那样 [1]。基于深度学习的语音合成是深度伪造的核心技术，可在没有特定语音样本的情况下，生成任意目标说话人语音的演讲内容。语音合成有两种类型



(a) 使用合成语音命令绕过说话人验证流程，并非法控制语音助手。



(b) 合成语音被用于入侵私人账户。



(c) 对手可以利用语音合成窃取受害者身份。

图 1：在真实场景中，语音合成攻击能够欺骗自动说话人验证系统和人耳。

方法：文本转语音合成（TTS）和语音转换（VC）。TTS 使用脚本文本和来自说话人的单个语音样本生成合成语音。然而，它通常生成不自然、机械化的人类语音。另一方面，VC 会直接将原说话人的语音转换为另一位说话人的语音，而不改变语音中的语言信息。VC 能够生成生动的语音，呈现真实人类语音的特征，例如停顿、情绪，甚至口音。

先进的语音合成技术已被用于良性应用，例如用已故祖母的声音给孩子读故事 [2]，，或让你最喜欢的电影明星为你指引路线 [3]。然而，如图 1 所示，人耳或自动说话人验证系统（ASV）通过语音模式验证说话人身份，因此语音合成技术会带来潜在的安全威胁。首先，语音助手通常要求用户先通过经认证的语音激活 Siri 或 Google Assistant，才能操作已锁定的智能手机。攻击者可以轻易控制智能手机，并使用合成语音发送恶意命令。其次，许多热门应用程序都集成了自动说话人验证系统，以进行用户身份验证。例如，微信要求用户说出一个 6 位数

，才能在未使用过的设备上首次登录 [4]。此类用户验证容易受到语音合成攻击。第三，攻击者可以利用合成语音在诈骗电话中冒充受害者 [5]。人耳难以区分高质量的合成语音与自然人声 [6]。

针对语音合成攻击的威胁，研究人员开发了合成人类语音检测模型，以识别伪造的人类语音 [7–9]。然而，在所有移动设备上部署语音合成检测可能成本高昂，而且这些检测方法在真实场景中未必能达到其宣称的高准确率 [6]。为防御未经授权的语音合成攻击，Attack-VC [10]会在受害者的语音样本中添加对抗扰动，阻止对手根据受害者的语音生成合成语音。由受保护语音生成的合成语音听起来会像另一位预先指定的目标说话人，而非受害者说话人[1]。不过，Attack-VC 的应用受到一定限制。用于优化对抗扰动的 PGD方法要求能够获取完整语音，并且需要经过大量迭代才能生成对抗性语音样本。因此，优化过程无法跟上实时语音的处理速度。由此，在语音消息、在线会议等实时应用中传输的实时人类语音和语音信号仍容易受到语音合成威胁。

在本工作中，我们提出了首个实时语音保护方法 VSMask，用于保护数字世界和物理世界中的实时语音。我们不直接优化扰动本身，而是训练一个神经网络来预测未来最有效的扰动。因此，保护性扰动可以即时融入实时音频流，而不会造成额外的时间延迟。此外，为了保护完整语音，我们引入了通用扰动头，无论语音内容如何，都能屏蔽语音的开头部分。基于人类听觉感知特性，我们提出了一种补充性的基于权重的幅度约束方法，以进一步减轻受保护语音的音频失真。总而言之，本文作出以下贡献：

- 我们提出了 VSMask，一种针对以下内容的实时保护系统语音合成攻击。通过预测即将到来的流式语音所需的扰动，VSMask 可以在几乎不产生延迟的情况下保护用户的语音。
- 我们提出了一种优化方案，用于训练通用扰动，以保护语音开头暴露的音频。这使 VSMask 能够不加区分地保护不同长度和大小的语音数据。我们进一步提出了一种基于权重的振幅约束方法，以降低对抗扰动的可感知性。
- 我们在三种不同的语音合成模型上评估了 VSMask 的性能。实验结果表明，VSMask 能够在数字空间和物理空间中实时防御语音合成攻击，有效保护自动说话人验证系统和人类听觉系统免受恶意语音合成的侵害。

[1]“受害者说话人”指语音合成攻击的目标。“目标说话人”指由 VSMask 预先确定、用于干扰语音合成的另一位说话人。

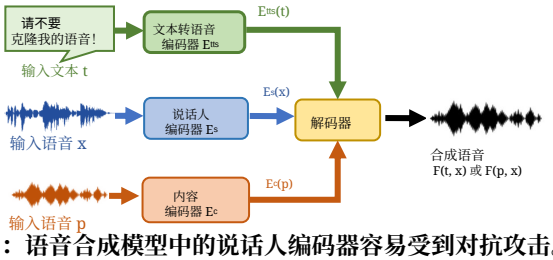


图 2：语音合成模型中的说话人编码器容易受到对抗攻击。

2 背景

2.1 语音合成

语音合成攻击旨在生成与真实人声相似的人工语音。图 2 展示了一个通用语音合成框架，其中包括文本转语音和语音转换。对于文本转语音和语音转换模型，内容编码器分别从文本 $\mathbf{t}$ 或语音 $\mathbf{p}$中提取语音内容嵌入 E_{tts} 或 E_c。说话人编码器 E_s 对受害者说话人的特征 E_s（$\mathbf{x}$）进行编码，其中 $\mathbf{x}$ 是来自受害者说话人x的语音样本。最后，解码器将生成合成语音 F（$\mathbf{x}, \mathbf{p}$）或F（$\mathbf{x}, \mathbf{t}$），其语音特征来自任意受害者说话人 x。

2.2 语音合成防御

Attack-VC [10] 引入了嵌入攻击，旨在攻击说话人编码器模型 E_s。如果我们输入与 $\mathbf{x}$ 差异很小的对抗性语音样本 $\mathbf{x}+\delta$，则编码向量 $E_s(\mathbf{x}+\delta)$会接近另一个目标说话人 y，而不是真实的受害者说话人 x。用于语音保护的对抗攻击可在问题（1）中表述如下：

$$\begin{aligned} \underset{\delta}{\text{minimize}} \quad & \mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{y})) - \lambda \mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{x})) \\ \text{subject to} \quad & \|\delta\|_\infty < \epsilon, \end{aligned} \tag{1}$$

其中，L 是均方误差（MSE）损失函数，$\mathbf{y}$ 是来自目标说话人 y的语音样本。问题（1）中的第一项最小化对抗性输入 $E_s(\mathbf{x}+\delta)$与目标说话人 $E_s(\mathbf{y})$的说话人嵌入向量之间的损失，迫使模型生成与目标说话人 y相似的语音；第二项则消除受害者说话人的语音 $E_s(\mathbf{x})$。最终，在输入受保护语音 $x+\delta$后，合成语音将无法再欺骗说话人验证或人耳。在问题（1）中，有必要预先设定一个目标说话人 y。我们进行了额外实验，不设定目标说话人（即，设置 $\lambda \to \infty$）。结果表明：尽管合成语音变得嘈杂，但听起来仍然像受害者说话人的语音。

2.3 实时对抗攻击

由于人类语音是一种流式音频信号，使用PGD方法生成实时对抗扰动并不切实际。VoiceCamo [11] 提出了一种针对语音识别（SR）模型的实时对抗攻击，可通过优化模型 g_θ来预测即将生成的对抗扰动。VoiceCamo中的预测模型优化过程如下所示：

$$\begin{aligned} \underset{\theta}{\text{maximize}} \ & \mathbb{E}_{(x_t, y_t)}[\mathcal{L}_{CTC}(\bar{y}_t, y_t)] \\ \textbf{s.t.} \quad & \bar{y}_t = f_\psi(x_t + g_\theta(x_{t-r-\delta})) \quad \text{and} \quad \|g_\theta(x_t)\|_\infty < \epsilon, \end{aligned} \tag{2}$$



(a) 实时通信应用。



(b) 社交媒体平台。



(c) 真实世界对话。

图 3：VSMask 可在在线和离线场景中提供保护。

其中，L_{CTC} 是连接时序分类（CTC）损失函数，x_t 是流式语音音频，y_t 是语音转录文本的真实标签。f_ψ 是参数固定为 *fixed*的语音转文本模型 ψ，g_θ是用于预测扰动的预测模型，而 $x_{t-r-\delta}$是在时间 $t-r-\delta$捕获的语音信号。预测出第一个扰动后，输入窗口会滑动到下一个时间槽，以预测第二个扰动，以此类推。最后，实时攻击启动时，语音识别准确率将大幅下降。

3 威胁模型

最先进的语音合成模型使攻击者只需一个受害者说话人的语音样本，就能生成其语音。本节将根据对手对受害者及潜在防御措施的了解，定义语音合成攻击的威胁模型。

3.1 对手知识

The attacker can use the victim's recorded voice to train a 模型生成受害者的语音，以此攻破说话人验证系统或欺骗人耳。模型可以是训练完善的模型，也可以是对手在本地训练的模型。他们还可以在说话人识别模型上测试合成语音，以评估语音相似度。如果他们发现样本受到保护，就可以采用降噪或其他信号处理方法来恢复清晰的音频。

3.2 对手能力

对手能够保存移动设备（例如智能手机或笔记本电脑）播放的音频信号，也能够使用麦克风录下现实环境中的语音。受害者说话人的语音样本可从多个来源获取。

实时通信应用。如图 3(a) 所示，人们通常会使用实时通信应用程序发送语音备忘录或进行语音通话。攻击者可能潜伏在聊天群组中，录制任何人的语音以进行语音合成。与此同时，自新冠疫情以来，许多人会在线开会。攻击者可以加入公开的在线会议，并录制受害者的语音，而无需征求许可。但现有的防御方法无法保护实时语音。

社交网络平台。人们乐于在社交媒体上分享日常生活。数十亿活跃用户将自己制作的视频上传到 TikTok、YouTube 和 Instagram [12]。通常，所有访客都可以下载并重复使用这些视频。即使只有一小段语音，恶意攻击者也能克隆你的语音并冒充你的身份。尽管现有的防御方法能够保护你在社交媒体上的语音，但上传前需要较长的处理时间 [10]。

真实场景中的对话。此外，现实生活中的人类语音也面临语音合成攻击的威胁。人们在演讲或交谈时，对手可以悄悄录下语音并发起语音合成攻击。然而，现有防御措施无法高效地掩蔽我们的语音，从而防范恶意语音合成。要将扰动实际注入真实语音中也颇具挑战。

3.3 防御者的知识

同时，我们考虑一种白盒场景：防御者完全了解对抗性语音合成模型，包括模型结构和模型参数。为了优化 VSMask，防御者还需要收集受保护的受害者说话人 x的少量清晰语音样本。此外，防御者拥有来自公开数据集的多个目标说话人候选者 y_i 的语音样本，并可为受害者说话人 x选出最合适的目标 y，以最大限度地提高语音差异。原始语音音频将在本地由 VSMask 处理，并上传至实时应用或社交媒体平台，不会造成额外延迟。

4 VSMASK系统设计 4.1 设计考量

一种针对自动语音识别系统的典型攻击方法，是将对抗扰动隐藏在人类语音的心理声学特征中 [13],，这可能成为防御语音合成攻击的一种潜在方案。然而，这些扰动需要获取完整语音才能对其进行掩蔽。因此，它们无法应用于实时场景。

我们还注意到，无论原始语音内容如何，通用扰动都能欺骗自动语音识别系统 [14]。但这类短扰动不足以保护长语音免受语音合成攻击。为了扩大防护范围，我们尝试利用周期性通用扰动来保护人类语音。遗憾的是，周期性扰动方法的效果并不理想。只有当扰动信号 δ 足够强时，防御才有效。通常，短通用扰动的自由度有限，这会限制防护性能。z 相比之下，较长的扰动周期通常会导致过拟合；此外，攻击者一旦在语音间隔期间捕获扰动，就可以对周期性扰动进行去噪。因此，开发一种稳健的实时保护机制以抵御语音合成攻击至关重要。

4.2 实时预测模型

预处理。图 4 展示了 VSMask 的系统框架。VSMask 的输入是流式人类语音信号，容易受到语音合成攻击。一旦 VSMask 检测到人类

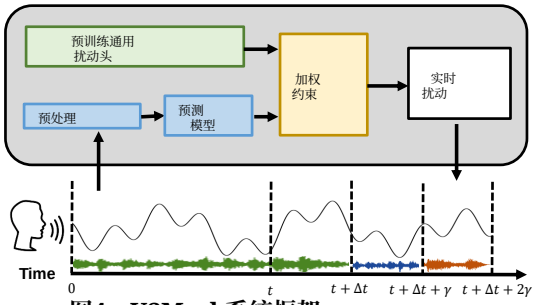


图4：VSMask系统框架。

语音开始，就会捕获接下来的音频信号。在预处理过程中，音频信号会转换为梅尔频谱图，并将频率转换为非线性的梅尔刻度，以匹配人类听觉。生成的梅尔频谱图与目标语音合成模型中说话人编码器 E_s 的输入具有相同的维度。

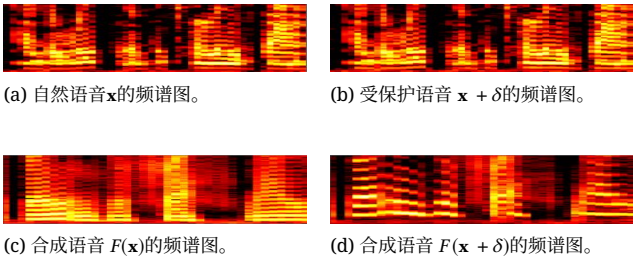
模型架构。 接下来，转换后的梅尔频谱图作为 VSMask 预测模型 f_θ 的输入。预测模型的输出同样是长度更短的梅尔频谱图。VSMask借鉴了VoiceCamo [11], , 选择了一个由7个下采样块和5个上采样块组成的实时预测模型，其性能优于层数更少的模型，并且与层数更多的模型保持相近的性能。每个下采样块包含反射填充，后接二维卷积层、二维批归一化和PReLU激活函数；每个上采样块由二维转置卷积层和带泄漏的ReLU激活函数组成。最后一个上采样块后接双曲正切函数，以约束 $\|\delta\|_\infty$。输出的梅尔频谱图将被转换为时域音频信号，并注入原始音频中。我们在附录A中说明了微调过程和模型架构的详细信息。

优化与工作流程。 假设目标语音合成模型中的说话人编码器函数为 E_s，我们可以将预测模型训练表述为以下优化问题：

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{y})) - \lambda \mathcal{L}(E_s(\mathbf{x}+\delta), E_s(\mathbf{x})), \\ \text{subject to} \quad & \delta^{\gamma}_{t+\Delta t} = f_\theta(\mathbf{x}_t) \quad \text{and} \quad \|\delta\|_\infty < \epsilon, \end{aligned} \tag{3}$$

其中，L 为均方误差损失函数，$\mathbf{x}$ 是来自受害者说话人 x的输入流式语音，$\mathbf{y}$ 是从选定的目标说话人 y处获取的语音样本，用于误导 E_s，而 δ是模型 f_θ预测的扰动。$\delta^{\gamma}_{t+\Delta t}$是在 $t+\Delta t$处添加的扰动，其长度为 γ。由于预测模型 f需要输入长度较短的信号 $\mathbf{x}_t$（长度为 t），VSMask 无需预先获取完整语音，从而实现实时保护。

如图4所示，VSMask会在经过短暂的 Δt延迟后，首先生成长度为 γ 的扰动（以蓝色波形表示）。随后，输入窗口从 γ 滑动到 $t+\gamma$所对应的下一个时间槽，以生成第二个扰动（以红色波形表示），并重复这一过程。与Attack-VC不同，VSMask微调的是预测模型参数θ，而非操纵扰动 δ。因此，一旦建立预测模型 f_θ，VSMask便可通过单个计算步骤识别出最佳扰动 δ，而无需进行大量迭代。



(a) 自然语音$\mathbf{x}$的频谱图。 (b) 受保护语音 $\mathbf{x}+\delta$的频谱图。

(c) 合成语音 $F(\mathbf{x})$的频谱图。 (d) 合成语音 $F(\mathbf{x}+\delta)$的频谱图。

图5：经VSMask保护的语音可有效降低语音合成模型的性能。

4.3 通用扰动头

然而，VSMask 预测模型存在一些不足，因为语音合成模型可以利用受害者说话人非常短的语音样本生成合成语音。例如，受害者说话人提供了一段语音示例——“请保护我的语音，免遭语音合成”。预测模型可以很好地保护后面的短语 “我的语音，免遭语音合成”。然而，语音开头的词语“请保护” 仍会暴露给攻击者，因为 VSMask 预测模型无法为语音开头生成扰动。因此，攻击者可能会截取语音音频，并使用一个音频片段合成语音。

为应对这一挑战，我们采用一种通用扰动，将其添加到流式语音信号的开头，以保护整段语音。近期研究 [14] 表明，无论话语内容如何，都存在能够迫使自动语音识别系统输出特定转录文本（例如“是”或“否”）的通用扰动。尤其在输入语音较短时，这种通用扰动仍能保持较高的成功率。幸运的是，预测模型的“盲区”也较小。因此，我们可以训练一种通用扰动，使说话人编码器无论语音内容如何，都输出所选目标说话人 y 的嵌入向量 $E_s(\mathbf{y})$，而非真实向量 $E_s(\mathbf{x})$。

为了训练这种通用扰动头 δ_h，我们将受害者说话人 x的训练语音样本裁剪为长度相同的音频片段$\mathbf{x}_i$ $t+\Delta t$。然后，我们使用PGD方法优化通用扰动头 δ_h，如下所示：

$$\begin{aligned} \underset{\delta_h}{\text{minimize}} \quad & \sum_{\mathbf{x}_i \in \mathcal{D}} \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}), \\ \text{subject to} \quad & \mathcal{L}_s(\delta_h, \mathbf{x}_i, \mathbf{y}) = \mathcal{L}(E_s(\mathbf{x}_i+\delta_h), E_s(\mathbf{y})) \\ & - \lambda \mathcal{L}(E_s(\mathbf{x}_i+\delta_h), E_s(\mathbf{x}_i)) \quad \text{and} \quad \|\delta_h\|_\infty < \epsilon. \end{aligned} \tag{4}$$

其中，D 是裁剪后的语音$\mathbf{x}_i$的集合。

在第4.2节的优化过程之前，我们将通用头添加到所有训练样本中，以增强防护性能。结合预测模型和通用头后，VSMask 成功地在实时场景中保护流式语音。图5(a)展示了攻击者可获取的一个自然语音样本$\mathbf{x}$。图5(b)是受保护语音$\mathbf{x}+\delta$的音频频谱图。扰动使这两个语音样本之间几乎没有差异。图5(c)和图5(d)分别是来自$\mathbf{x}$和$\mathbf{x}+\delta$且语音内容相同的合成语音频谱图。尽管$\mathbf{x}$和$\mathbf{x}+\delta$非常相似，但合成语音 $F(\mathbf{x})$呈现男性语音模式，而 $F(\mathbf{x}+\delta)$则呈现具有较高频率的女性语音。

4.4 基于权重的噪声缓解

不可避免地，注入的保护性扰动会导致音频失真。以往研究通常会限制扰动幅度，以尽量降低对人类感知的影响，或在优化过程中考虑扰动幅度，以在噪声水平和性能之间实现最佳权衡。然而，即使扰动经过充分优化，语音信号中仍存在显著的背景噪声。

为了降低 VSMask 扰动的可感知性，我们研究了人类听觉灵敏度。图 6 展示了等响度曲线 [15], ，它测量了人类对不同频率音调的听觉感知。在低频（< 1.6 kHz）或高频（> 4 kHz）下，要使人耳感受到相同的响度级，需要更高的声压。对于 1.6 kHz 到 4 kHz 之间的中频频段，人耳的灵敏度最高。基于这一现象，一种可能的方法是将所有扰动隐藏在低频频段中。然而，我们的实验表明，这种方法无法为我们的语音提供足够的防护。

因此，我们需要将全局扰动注入整个频谱图中。为了减轻人类的听觉感知，我们应当降低 1.6 kHz 到 4 kHz 之间的扰动幅度。然而，降低扰动幅度也会削弱防护性能。为了弥补中频频段的不足，我们提高其他人耳难以感知的频段中的扰动幅度。因此，我们将幅度约束改写如下：

$$\begin{aligned} &\delta = \begin{bmatrix} \delta_l & \delta_m & \delta_h \end{bmatrix}^T, \\ &\textbf{s.t.} \quad \|\delta_l\|_\infty < \epsilon_1,\ \|\delta_m\|_\infty < \epsilon_2,\ \|\delta_h\|_\infty < \epsilon_3, \end{aligned} \tag{5}$$

其中 $\epsilon_2 < \epsilon_3 < \epsilon_1$。通过对扰动幅度添加这种基于权重的约束，我们在不牺牲防护性能的前提下，限制了人类感知到的扰动，从而成功提升了受保护音频质量。

4.5 VSMask 应用场景

如第 3 节所述，攻击者有多种不同方式获取受害者说话人的语音音频。下面介绍 VSMask 的多种应用场景。

在线防护。 第一种应用场景是在实时通信和社交媒体应用中保护人类语音。对于这些应用程序，VSMask 可作为插件组件来保护用户的语音。VSMask 一旦检测到语音开始，便会自动添加通用扰动头，同时采集语音音频。注入扰动头后，VSMask 会预测下一个时间槽所需的扰动，然后迭代地保护后续语音流。基于预测机制，VSMask 不会给实时通信带来额外延迟。

真实世界物理防护。 在真实场景中，攻击者可能会偷偷录下受害者语音，并实施语音合成攻击。为了应对此类对手，用户可以在移动设备上安装 VSMask 来保护自己。交谈时，VSMask 会播放扰动噪声信号，并有效地将对抗性语音样本注入窃听到的音频中。最终，它将保护我们在真实世界中的语音免受语音合成攻击，正如第 6 节的实验所示。

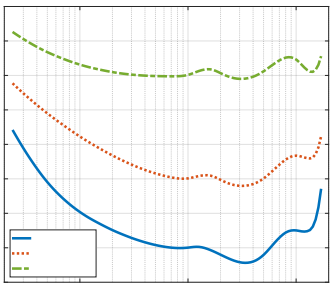


图6：等响度曲线。

5 评估

5.1 目标模型与实验设计

目标语音合成模型。 此外，我们选取以下三种具有代表性的语音合成模型来评估 VSMask 的性能：

- AdaIN-VC [16]: 一种基于实例归一化的无监督单样本语音转换模型；
- AutoVC [17]: 一种采用精心设计瓶颈的新型风格自编码器，是首个使用非平行语音数据实现零样本语音转换的工作；
- SV2TTS [18]: 一种将文本转录文本转换为目标说话人声音的人类语音的文本转语音模型。

数据集。 对于 AdaIN-VC 和 AutoVC 模型，我们使用 CSTR VCTK语料库 [19] 数据集（这是语音转换研究中广泛使用的语音数据集）来评估 VSMask 的性能。VCTK 包含 109 位具有不同口音的英语说话人的语音样本。每位说话人朗读最多 400 个随机选自报纸或引导材料的句子。对于 SV2TTS 模型，我们选择 LibriSpeech 数据集 [20] 进行评估。具体而言，由于某些语音样本过短，无法激活 VSMask 预测模型，因此我们仅使用时长超过 3 秒的语音音频进行训练和测试。

5.2 VSMask 在自动说话人验证系统上的有效性

用于评估的自动说话人验证系统。 为了评估 VSMask 降低合成语音与受害者说话人语音之间相似度的效果，我们使用 SpeechBrain [21], 这一先进模型来计算两个语音样本的语音相似度。SpeechBrain 说话人验证模型由增强型通道注意力、时延神经网络中的传播与聚合（ECAPA-TDNN）组成，用于提取说话人嵌入。相似度分数通过计算两个语音样本嵌入向量的余弦相似度得出。分数越高，两个语音样本来自同一说话人的可能性越大。如果分数高于阈值 k，语音便可通过验证。在我们的实验中，我们采用默认阈值 $k = 0.25$，该阈值对于清晰的人类语音样本可实现最佳准确率。**VSMask 与基线方法设置**。本节介绍 VSMask 的评估设置，并将其与另外 4 种基线方法进行比较。我们通过向数字音频信号注入扰动，在数字域中评估这些方法。

VSMask: 首先，对于 VSMask 方法，我们根据输入时间窗口长度的 λ 值，在问题 (3) 和 (4) 中设置 $\lambda = 1$ [6], t 为

表1：我们将 VSMask 的防御性能与其他在线或离线基准防御方法进行了比较。VSMask 的表现优于所有其他实时保护方法，并达到了与离线 PGD 方法相近的性能。

Type	方法	得分	准确率	STOI
非合成	原始语音	0.780	98。8%	1.000
	受保护语音	0.429	68.5%	0.839

Type	方法	M2M		F2F		M2F		F2M	
		得分	ASR	得分	ASR	得分	ASR	得分	ASR
合成的	原始语音	0.595	91。9%	0.612	93.2%	0.561	88.3%	0.546	86。0%
	随机噪声	0.516	86.6%	0.538	89。0%	0.505	84.0%	0.473	81。5%
	周期性扰动	0.192	11.0%	0.203	12.5%	0.177	9.8%	0.156	8.6%
	离线 PGD	0.064	0.0%	0.085	0.0%	0.049	0.0%	0.055	0.0%
	VSMask	**0.077**	**0.0%**	**0.104**	**0.0%**	**0.056**	**0.0%**	**0.073**	**0.0%**

M2M：男性到男性，F2F：女性到女性，M2F：男性到女性，F2M：女性到男性。

1.25 秒。为确保实时预测，应避免将时间延迟设得过短。另一方面，时间延迟过长会导致 VSMask 的性能下降。因此，我们将时间延迟（Δt）和输出长度（γ）均设为 0.4 秒。因此，扰动头长度为 $t + \Delta t = 1.65$ 秒。此外，我们设置了扰动幅度约束 $\epsilon = 0.10$。经过噪声缓解方法处理后的扰动幅度为：$\epsilon_3 = \epsilon$、$\epsilon_1 = 1.15\,\epsilon$ 和 $\epsilon_2 = 0.85\,\epsilon$。

此外，我们从数据集中随机选取 30 名男性说话人和 30 名女性说话人。我们使用 AdaIN-VC 模型为每位说话人生成 100 个合成语音样本。对于女性受害者说话人，我们选取一名男性说话人作为目标说话人，反之亦然。对于语音转换模型而言，说话人的性别很重要，因为男性和女性具有不同的语音模式，例如，女性说话人的语音通常频率更高。我们根据受害者说话人和语音源说话人的不同性别对结果进行划分。

随机噪声：降低语音合成模型性能的一种可能方法是向语音信号 $\mathbf{x}$ 中添加随机噪声。在实验中，我们将随机噪声幅度设为 $\epsilon_n = 0.15$。周期性扰动：由于通用头无论语音内容如何都能保护较短的语音，因此，周期性通用扰动是实时防御的一种潜在方案。在实验中，我们反复向原始语音音频中注入幅度为 $\epsilon_p = 0.12$ 的同一通用扰动头，并评估其防护性能。

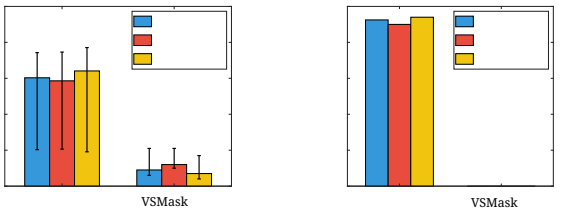
离线 *PGD* 方法：此外，我们还将我们的方法与 Attack-VC [10] 中的离线 PGD 方法进行了比较。我们采用问题（1）所示的嵌入攻击方法，迭代次数设为 1,500, $\epsilon_{off} = 0.10$。尽管离线 PGD 方法不适用于实时保护，但它有助于确定防护性能的上限。

VSMask 在自动说话人验证系统上的评估。我们使用非合成语音样本对自动说话人验证系统的性能进行了定量评估。对于未经过 VSMask 保护的清晰音频，自动说话人验证系统的准确率超过 98%。但对于受 VSMask 保护的语音音频，与 Attack-VC 中的结果类似，自动说话人验证准确率会下降，因为扰动会扭曲语音中的语音模式。通过这种方式，VSMask 可能会影响现实场景中的说话人识别功能，例如语音助手激活。幸运的是，移动设备可以使用回声消除抵消扬声器产生的扰动 [22]。因此，VSMask 不会干扰激活过程。此外，我们使用短时客观可懂度（STOI） [23]来衡量受保护音频的语音质量。结果表明，处于 VSMask 保护下的语音对人耳而言完全可懂。

接下来，我们以 AdaIN-VC 作为目标语音合成模型，将 VSMask 与不同的基线防御方法进行比较。我们使用攻击成功率（ASR）来评估语音合成攻击的效果。作为参考，当攻击者使用原始语音音频克隆语音时，超过 85% 的合成音频能够以较高的相似度分数通过自动说话人验证系统。当语音转换发生在同性别之间（男性到男性或女性到女性）时，ASR 甚至更高。如果受害者说话人是女性，语音转换的 ASR 略高于受害者说话人为男性的情况。这一结果证实了利用语音合成攻击攻破自动说话人验证系统的有效性。

接下来，我们评估基线实时防御方法。首先，我们在语音中加入白噪声来合成语音。当扰动幅度较小时（$\epsilon_n = 0.15$），随机噪声无法降低语音合成性能。事实上，当 δ为较弱的随机噪声时，嵌入的说话人向量 $E_s(\mathbf{x})$ 和 $E_s(\mathbf{x} + \delta)$几乎没有差异。如果加入更强的噪声，则会因语音质量较差而起作用。当噪声幅度$\epsilon > 0.4$时，没有合成语音样本能够危害自动说话人验证模型。但受保护音频 $\mathbf{x} + \delta$对人耳来说难以辨识。因此，在现实世界中，强噪声并非可行的防护方法。

其次，我们使用周期性通用扰动进行防御。如表1所示，周期性通用扰动显著降低了合成语音与受害者语音之间的相似度。但仍有超过10%的同性别合成音频



(a) 不同的相似度分数语音合成模型

(b) 不同语音合成的自动语音识别模型

图7：对于所有目标语音合成模型，VSMask 都能显著降低语音相似度。使用 VSMask 进行防护后，没有合成语音能够通过说话人验证系统。

能够通过自动说话人验证系统。原因在于，当扰动幅度较小时，通用扰动无法影响整段语音，尤其是较长的语音样本。我们可以增大扰动幅度或周期长度，以提升周期性方法的性能。然而，如果扰动过强，就会导致明显的音频失真；而对扰动采用较长周期，可能无法有效保护较短的语音片段。

最后，我们将离线 PGD 方法的性能与 VSMask 进行比较。从理论上讲，离线 PGD 方法在防御语音合成攻击方面能取得最佳性能。由离线 PGD 保护的语音所生成的合成音频均无法通过自动说话人验证系统。增加迭代次数还能进一步提升其性能。但在迭代 1,500 次时，其耗时约为音频时长的 $20\times$ 倍。如果模型部署在没有高性能图形处理器的设备上，计算时间可能会长得多。我们在表1中用**粗体** 标出 VSMask 的性能。与所有其他在线基线方法相比，VSMask 的防护效果最强。与离线 PGD 方法相比，VSMask 无需耗时即可实现相近的离线防护性能。对于同性别和异性别语音转换，合成语音样本均无法通过自动说话人验证系统。因此，结果表明，VSMask 能够成功保护人类语音免受未经授权的语音合成攻击。

5.3 VSMask 在不同语音合成模型上的评估

在本节中，我们评估 VSMask 在不同语音合成模型上的性能。从第 5.2 节可知，对于语音转换模型，同性别语音转换比异性别语音转换具有更高的相似度。对于 AutoVC 模型，我们为每位说话人生成 50 个同性别合成语音样本。

图7(a)和图7(b)展示了VSMask防护前合成语音的相似度和自动语音识别结果。对于语音转换模型，在使用VCTK语料库中的数据训练后，AdaIN-VC和AutoVC生成的合成语音样本均能通过自动说话人验证系统的验证，成功率超过90%。由于采用了精心设计的瓶颈机制，AutoVC的方差较小。不过，AdaIN-VC的整体相似度分数更高。我们还测试了文本转语音模型，并将其性能与语音转换模型进行比较。SV2TTS模型使用16千赫语音样本进行训练。为确保其达到最佳性能，我们从LibriSpeech数据集中选取60位受害者说话人（30名男性和30名女性）。此外，我们为每位说话人生成50个合成语音样本用于测试。随后，我们在自动说话人验证系统中测试所有生成的文本转语音样本。与语音转换模型相比，无论输入短语的长度如何，SV2TTS都能取得最高的相似度分数和最高的自动语音识别率。其潜在原因是，SV2TTS直接根据文本输入生成语音内容，因此不包含任何已有的语音模式。但对于语音转换模型，即使经过充分优化，源语音中仍会残留语音模式$\mathbf{p}$。这会降低合成语音与受害者说话人之间的语音相似度。

相比之下，我们还展示了经过 VSMask 保护后的相似度分数和自动语音识别结果。对于语音转换模型，VSMask 显著降低了合成语音与受害者说话人的原始语音之间的相似度分数。没有任何合成语音样本能够攻破自动说话人验证系统。由于 SV2TTS 中的合成语音模式完全由参考语音 $\mathbf{x}+\delta$ 决定，因此，与语音转换模型相比，该模型更容易受到 VSMask 保护的影响。总体而言，在 VSMask 的保护下，攻击者几乎不可能通过语音转换或文本转语音来欺骗说话人验证系统。

5.4 人类受试者研究

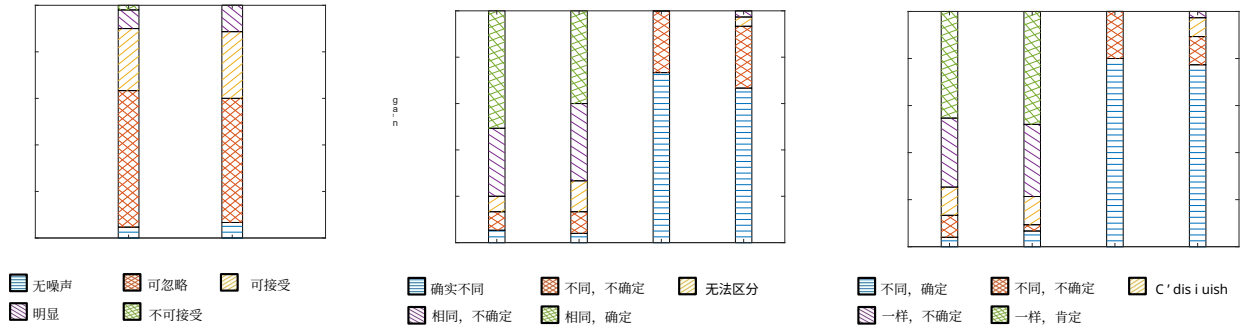
近期研究 [6]表明，语音合成攻击不仅会攻破自动说话人验证系统，还会欺骗人耳。在本节中，我们针对 VSMask[2]开展人类受试者研究。我们选择 AdaIN-VC 和 AutoVC 进行合成语音生成，因为与文本转语音相比，语音转换模型生成的话语更加自然。

在人类受试者研究评估中，我们招募了25名听力正常的匿名志愿者（14名男性和11名女性，年龄在20至38岁之间），并让他们通过同一台音量固定为（~60 dBA）的扬声器，聆听12组语音样本。其中，6组合成语音样本来自AdaIN-VC模型，另外6组来自AutoVC模型。对于每个模型，一半样本为男性到男性，另一半为女性到女性。每组包含4个音频样本：① 未经保护的原始语音音频；② 同一段语音音频，但经过VSMask保护；③ 根据原始语音生成的合成语音样本；④ 根据经过VSMask保护的语音生成的同一合成语音样本。所有语音样本的时长均为4至8秒。

首先，我们询问志愿者对VSMask保护语音中噪声水平的感受。听完原始语音和受保护语音后，听者需要评估受保护样本中扰动的可感知程度。选项从低到高依次为：（A）听不出任何噪声；（B）能听到噪声，但可以忽略；（C）能听到噪声，但可以接受；（D）能听到明显的噪声；（E）噪声非常强，无法接受。

我们收集了所有回答，并将结果呈现在图8(a)中。结果表明，对于 AdaIN-VC 和 AutoVC，受保护语音样本中的噪声水平对人耳来说是可以接受的。约 60% 的回答认为噪声「可以忽略」或「无法察觉」，这意味着 VSMask 保护后的语音音频与原始音频具有相近的清晰度。超过 90% 的回答认为受保护音频中的噪声对人耳来说是可以接受的。300 个回答中仅有 2 个指出受保护语音存在「无法接受」的噪声。

[2]该机构审查委员会申请已获大学委员会批准。



(a) 语音转换模型中受保护语音的噪声水平。(b) 对 VSMask 在 AdaIN-VC 模型上的防护性能进行的人类评估。(c) 对 VSMask 在 AutoVC 模型上的防护性能进行的人类评估。

图 8：对受保护语音噪声水平以及 VSMask 防护性能的主观人类评估结果。

接下来，我们请志愿者评估原始语音与合成语音样本之间的相似度。相似度选项从低到高依次为：（A）肯定是不同的语音；（B）是不同的语音，但不确定；（C）无法区分；（D）是相同的语音，但不确定；（E）肯定是相同的语音。考虑到不同说话人性别导致的语音转换效果差异，我们分别展示男性到男性和女性到女性语音样本的结果。

图 8(b) 和图 8(c) 展示了①原始音频与由原始语音生成的③合成音频之间，以及①原始音频与由受保护语音生成的④合成音频之间的语音相似度。对于 AdaIN-VC 和 AutoVC，大多数由原始语音生成的合成语音都能欺骗人耳。超过 75% 的回答表示，合成语音肯定或可能来自与原始音频相同的说话人。只有 4% 的回答表示，该语音绝对来自另一位说话人。因此，结果表明，在发起同性别语音转换时，合成语音与受害者说话人的语音具有足够的相似度。相比之下，由 VSMask 保护语音生成的合成语音样本（M2M-VM 和 F2F-VM）已无法再欺骗人耳。超过 75% 的回答认为，由 VSMask 保护语音生成的合成音频，其语音与受害者说话人的语音明显不同。总计超过 97% 的回答表示，他们会对语音背后的说话人身份产生怀疑。没有人认为合成语音与受害者语音非常相似。此外，研究表明，离线 PGD 方法在防止深度伪造语音欺骗人耳方面能够达到相当的防护性能 [10]。此外，我们在附录 B 中展示了基于权重的噪声缓解所产生的影响。

5.5 跨模型评估

Attack-VC [10] 通过训练替代模型，在黑盒场景下评估防御性能。然而，有时我们并不知道将要采用哪种语音合成模型。本节将在跨模型条件下评估 VSMask 的性能。

我们应用 VSMask 生成针对一个语音合成模型的对抗性语音样本，并将其应用于其他模型。跨模型自动语音识别结果列于表2。总体而言，VSMask 在跨模型场景中仍然有效，但性能有所下降。

表2：我们从源模型生成对抗样本，并将目标模型用于语音合成。结果表明，在跨模型设置下，VSMask 仍然能够降低自动语音识别率。

目标 \ 源	AdaIN-VC	AutoVC	SV2TTS
AdaIN-VC	-	15.0%	10.5%
AutoVC	12.8%	-	0.0%
SV2TTS	7.3%	15.2%	-

由于编码器模型、预处理方法和训练数据不同，结果也会有所差异。例如，AdaIN-VC 使用 512 维梅尔频谱图作为说话人编码器的输入，而 AutoVC 的说话人编码器则使用 80 维梅尔频谱图。此外，SV2TTS 使用 16千赫的音频进行训练，而 AdaIN-VC 和 AutoVC 使用 48 千赫的音频。当我们将 16 千赫的语音样本输入 AutoVC 时，它无法输出合格的音频。此外，针对 AdaIN-VC 生成的对抗样本也能够降低 AutoVC 合成语音的相似度。当我们将来自 AdaIN-VC 或 AutoVC 的对抗样本应用于 SV2TTS 模型时，少数合成样本仍然能够破坏自动说话人验证系统。我们还可以应用 AdaBelief 方法 [24]，通过逐步降低学习率来提高受保护样本的迁移能力，这将作为未来的研究工作。

5.6 自适应攻击评估

去噪器广泛应用于语音增强模型，以提升音频质量。一些音频信号变换方法也能减轻对抗扰动。为评估 VSMask 对自适应攻击者的鲁棒性，我们使用去噪器和 WaveGuard [25] 去除受保护音频中的扰动，并通过 AdaIN-VC 模型进行语音合成。所有合成语音样本均来自同性别语音合成。

简单的降噪器无法去除受保护语音中精心设计的扰动；而且，由于降噪器消除了一些语音中的低功率信号，合成语音的相似度甚至会进一步下降。通常，即使分辨率较低，语音转文本模型仍能准确转录语音。这使 WaveGuard 能够

表 3：自适应攻击者应用去噪器和 WaveGuard 时 VSMask 的性能。

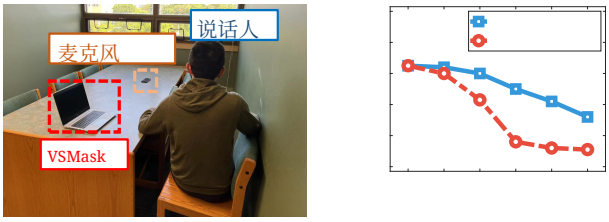
自适应方法	None	去噪器	WaveGuard			
			向上-向下 (f=24k)	数量。- 反数量。	Mel. (二进制=128)	LPC (序数=10)
分数	0.096	0.090	0.078	0.082	0.080	0.073

在缓解对抗扰动的同时降低语音音频质量。我们在经过 VSMask 保护的语音样本上评估了 WaveGuard 的四种信号变换方法，包括采用不同频率的降采样-升采样、量化-反量化、采用不同梅尔滤波器组数量的梅尔频谱图转换，以及采用不同阶数的线性预测编码（LPC）。我们在表 3 中列出了最佳相似度分数。与针对语音转文本模型的对抗样本不同，WaveGuard 方法无法削弱 VSMask 的防护性能，因为 VSMask 在梅尔频谱图中优化扰动，而这种扰动能够抵御常规信号变换技术。因此，对音频进行压缩或变换无法准确恢复最初的清晰语音。此外，WaveGuard 会降低合成语音的音质。因此，VSMask 对降噪和信号变换方法仍具有很强的鲁棒性。对于复杂的攻击者而言，如果他们完全了解 VSMask，就有可能通过逆向工程恢复原始音频。其他方法，例如随机平滑 [26], 对抗训练 [27] 和扩散模型 [28]，可能是规避 VS Mask 的潜在解决方案。

6 讨论与局限性

真实场景中的防护。图 9(a) 展示了 VSMask 在物理世界中的防护设置。说话人在安静的房间里朗读 VCTK 数据集中的文本。同时，VSMask 部署在距离说话人 50 cm 的笔记本电脑上。与此同时，窃听者在距离说话人和 VSMask 一米处，通过 iPhone 13 悄悄录制语音音频。随后，窃听者将利用 AdaIN-VC 模型，根据录制的语音合成说话人的语音。同时，我们在麦克风附近放置了声级计，以测量声音音量。房间内的背景噪声为 *40* dBA，我们将语音响度维持在约 *75* dBA。受害者在安静的房间里说话时，攻击者可以窃听到清晰的语音信号，并生成合格的深度伪造语音。

值得注意的是，在真实场景中，与网络空间不同，扰动 δ 在空中传输过程中通常会失真。为克服这一挑战，我们采用带通滤波器（500 ~ 4000 Hz）滤除易受影响的频率，并采用房间脉冲响应（RIR）滤波器来减轻信号在空中传输过程中的失真 [29]。在新环境中应用 VSMask 时，应重新设计 RIR 滤波器。图 9(b) 展示了其在真实场景中的防护效果。当 VSMask 的平均扰动响度达到 46 dBA 时，便能有效干扰目标语音合成模型，同时不影响语音的可懂度。相比之下，同等音量的随机白噪声无法显著降低相似度分数。当扰动音量达到 50 dBA 时，VSMask 能成功将自动语音识别



(a) 真实环境实验设置。

(b) VSMask 在真实场景中的防护性能。

图 9：我们在真实场景中实现了 VSMask，其保护效果远强于白噪声。

率降至 0%，而大多数由白噪声掩蔽语音生成的合成语音仍能绕过自动说话人验证模型。实验在安静环境中进行。如果环境中存在噪声（音量 > 50 dBA），即使不使用 VSMask，合成语音的质量也会受到影响。特别是，自适应攻击者可能会利用麦克风阵列，通过相位抵消来消除 VSMask 扰动 [30]。我们可以利用多通道扬声器增强空间复杂度，从而增加消除 VSMask 扰动的难度。

实时可行性。在 NVIDIA A6000 GPU 上为一位说话人训练预测模型，耗时不到 5 分钟。在推理过程中，预测模型在普通中央处理器上单次计算耗时为 ~ 13 毫秒。我们还测量了音频信号与梅尔频谱图之间双向转换造成的时间延迟，为 ~ 200 毫秒。因此，生成语音扰动总共约需 0.2 秒，而 0.4 秒的延迟足以让 VSMask 实现实时适用性。

对抗训练。对抗训练可以有效提升深度学习模型的鲁棒性 [27]。自适应攻击者可以采用对抗训练，提高合成语音与受 VSMask 保护的受害者说话人之间的相似度。然而，在处理清晰的语音音频时，对抗训练不可避免地会降低合成性能。此外，近期研究 [31] 表明，即使是一个小型数据集，使用高性能 GPU 集群进行对抗训练也需要长达 10 天。因此，鉴于 VSMask 的实时保护特性，对抗训练并非攻击者可用来攻破 VSMask 的现实解决方案。**局限性**。VSMask 存在若干局限性。首先，本文未考虑专门设计的自适应攻击，例如完全了解 VSMask 的对手。自适应攻击者可以通过精心设计的神经网络或扩散模型净化语音，我们将把对这些规避方法的探索留待未来工作。其次，我们仅在物理实验中于单个房间开展了真实场景评估。在不同环境中，VSMask 的性能可能有所不同。第三，无论语音合成模型如何，VSMask 都无法提供 100% 的成功防护。在纯黑盒设置下，仍有少量合成语音可以通过自动说话人验证系统。最后，所提出的方法无法为人类语音提供形式化的隐私保证。本文旨在使用 VSMask 开展实证隐私分析，并提供一种实用方案，以增强实时物理和数字场景中的语音隐私。

7 相关工作 7.1 语音合成

对于传统的语音转换方法，需要使用源说话人和受害者说话人的平行语音数据来训练模型。Chen 等人 [32] 提出了一种基于高斯混合模型（GMM）的语音转换模型。Erro 等人 [33] 采用加权频率扭曲方法来改变语音中的声音。然而，这些方法存在过度平滑 [34] 和语音不自然的问题。此外，收集不同说话人的平行语音数据也十分困难 [35, 36]。

为应对这些挑战，研究者引入了变分自编码器（VAE），将人类语音转换为语音内容 [37]。与传统的语音转换模型相比，VAE 能够学习独立于说话人身份的语音内容。因此，VAE 转换效率更高，且训练时无需平行语音数据。Hsu 等人 [38] 提出了一种语音转换模型，可逐帧生成转换后的语音。AdaIN-VC [16] 通过增加一个额外的对抗训练阶段，显著提升了转换后的语音质量。Cycle-GAN [39] 和 Star-GAN [40] 也将对抗学习方法应用于基于 VAE 的语音转换中。AutoVC [17], 是最先进的工作之一，它是一种将瓶颈机制引入 VAE 的语音转换模型。借助精心设计的瓶颈结构，该模型成功实现了零样本转换，并兼顾了说话人解耦与语音转换质量。此外，生成对抗网络（GAN）本身也可用于克隆人类语音 [41], 但其转换质量有限。文本转语音也能合成具有人类真实音色的语音。WaveNet [42] 引入了一种生成模型，可根据输入文本输出原始音频。Tacotron 2 [43] 进一步利用注意力机制，将 WaveNet 生成的频谱图反演为语音。但 Tacotron 2 只能生成具有一种默认说话人音色的语音。Deep Voice 2 [44] 是一种支持多说话人的语音生成模型，而 Deep Voice 3 [45] 使用基于编码器-解码器的神经网络。它可以利用 LibriSpeech 中的 2,400 多种不同音色进行文本转语音合成。SV2TTS [18] 提出了一种文本转语音模型，可为任意说话人合成语音。该模型采用了与语音转换模型相似的说话人编码器模块。Wenger 等人 [6] 对真实场景中的语音合成威胁进行了全面评估。

7.2 对抗性语音攻击与防御措施

现有研究表明，神经网络很容易被对抗样本欺骗。Carlini 等人 [46] 指出，对抗性语音样本可能危害自动语音识别系统。Schönherr 等人 [13] 证明，超出人类听觉范围的扰动也能危害自动语音识别。对抗性语音样本可以隐藏在语音 [47]或音乐 [48], 中，并经过精心设计，以欺骗语音识别系统。在某些情况下，这些对抗性样本可以做到人耳无法察觉 [49], , 从而在数字和物理环境中造成安全威胁。

同样，说话人识别模型也容易受到对抗攻击 [50, 51]。Chen 等人 [52] 提出了 FAKEBOB，这是一种对抗攻击，可在黑盒和空中传输攻击场景下成功攻破说话人识别系统。Zhang 等人 [53] 也通过精心设计的对抗语音样本成功欺骗了商用说话人验证系统。

离线对抗语音攻击只有在获得完整语音后才能奏效。为了针对流式语音信号发起实时攻击，Lu 等人 [14] 提出了通用对抗扰动，无论实际输入是什么，都能迫使自动语音识别模型输出目标转录文本。AdvPulse [54] 通过一个简短的对抗“脉冲”成功欺骗自动语音识别系统。这些对抗扰动无需信号同步，也能成功误导自动语音识别系统。但这些通用扰动只有在所构造的输出较短时才有效。Gong 等人 [55] 提出了一种针对流式语音信号的实时攻击。这种实时攻击适用于使用预定义词典的关键词检测应用。SpecPatch [56]提出了一种对抗攻击，可消除即将到来的语音命令，并攻破较长的语音样本。VoiceCamo [11]提出了一种基于预测模型的实时攻击。它将过去的语音信号作为输入，并预测最有效的扰动，从而与即将到来的语音一起生成对抗样本。

7.3 语音隐私增强

Qian 等人 [57] 利用语音转换隐藏说话人的身份。McAdams [58] 和 Vaidya 等人 [59] 通过信号处理来屏蔽人类语音，从而增强语音隐私。Han 等人 [60] 提出了一种基于差分隐私的快速语音去身份识别系统，可在不可信服务器上提供形式化的隐私保证。Emotionless [61] 是一种基于差分隐私、用于语音助手的隐私保护语音分析工具。相比之下，本文旨在开发一种实用的语音保护系统，为现实世界中的语音应用提供隐私增强。

与此同时，对抗样本可用于增强语音隐私。V-Cloak [62] 提出了一种保留音色的模型，以实现语音匿名化，但尚未评估其抵御语音合成攻击的防护性能。Attack-VC [10]提出了一种基于对抗样本的防御方法，可保护离线人类语音免受语音合成攻击。VSMask 则创新性地结合通用扰动与预测模型，以保护实时语音免受语音合成攻击。

8 结论

语音合成攻击是语音可控系统面临的重大威胁。本研究旨在阻止攻击者合成受保护语音。具体而言，我们探讨了实时防御语音合成攻击的可行性。通过实现实时预测模型和通用扰动头，VSMask 成功地在在线和离线的真实场景中保护实时语音信号。此外，VSMask 通过基于权重的幅度约束机制，减轻了受保护语音中的音频失真。我们在不同真实场景下评估了 VSMask 的语音保护性能，实验结果表明，VSMask 能够在实时语音场景中有效抵御语音合成攻击。

致谢

我们谨此感谢匿名审稿人和指导人对本研究提出的深刻意见与反馈。本研究部分由美国国家科学基金会资助，资助项目编号为 CNS-1950171，另获 CISCO 研究奖项资助。

参考文献

[1] B. U. Mahmud 和 A. Sharmin，“深度伪造技术的深入解析：综述”，arXiv 预印本 arXiv:2105.00192，2021年。[2] “亚马逊 Alexa 模仿奶奶的声音，吓坏网友”，https://www.bloomberg.com/news/articles/2022-06-23/amazon-s-alexa-impersonates-grandma-creeping-out-the-internet，2022年6月23日。[3] “语音引导与导航”，https://www.readspeaker.ai/applications/voice-guidance-and-navigation/，访问日期：2022年8月15日。[4] “声纹：微信的新密码”，https://blog.wechat.com/2015/05/21/vo iceprint-the-new-wechat-password/，2015年5月21日。[5] Y. Wang、H. Guo 和 Q. Yan，“GhostTalk：通过电力线对智能手机语音系统实施交互式攻击”，载于《网络与分布式系统安全（NDSS）研讨会》，2022年。[6] E. Wenger、M. Bronckers、C. Cianfarani、J. Cryan、A. Sha、H. Zheng 和 B. Y. Zhao，“‘你好，是我’：基于深度学习的语音合成攻击在现实世界中的应用”，载于《2021年 ACM SIGSAC计算机与通信安全会议论文集》，2021年，第235–251页。[7] M. E. Ahmed、I.-Y. Kwak、J. H. Huh、I. Kim、T. Oh 和 H. Kim，“VOID：一种快速轻量的语音活体检测系统”，载于第29届 USENIX安全研讨会（USENIX Security 20），2020年，第2685–2702页。[8] Y. Zhang、F. Jiang 和 Z. Duan，“面向合成语音欺骗检测的单类学习”，《IEEE信号处理快报》，第28卷，第937–941页，2021年。[9] X. Li、X. Wu、H. Lu、X. Liu 和 H. Meng，“通道门控 Res2Net：面向稳健的合成语音攻击检测”，arXiv 预印本 arXiv:2107 08803，2021年。[10] C.-y. Huang、Y. Y. Lin、H.-y. Lee 和 L.-s. Lee，“保护你的语音：针对语音转换的对抗攻击”，载于《2021年 IEEE口语技术研讨会（SLT）》。IEEE，2021年，第552–559页。[11] M. Chiquier、C. Mao 和 C. Vondrick，“实时神经语音伪装”，载于《国际学习表征会议》，2021年。[12] “71+ Instagram 2022年统计数据”，https://https://www.demandsage.com/instagram- statistics/，2022年11月。[13] L. Schönherr、K. Kohls、S. Zeiler、T. Holz 和 D. Kolossa，“通过心理声学隐藏攻击自动语音识别系统”，arXiv 预印本 arXiv:1808.05665，2018年。[14] Z. Lu、W. Han、Y. Zhang 和 L. Cao，“探索针对端到端自动语音识别模型的定向通用对抗扰动”，ar Xiv 预印本 arXiv:2104.02757，2021年。[15] “等响度曲线（ISO 226-2003）”，https://chart-studio.plotly.com/~mrlyu le/16/equal-loudness-contours-iso-226-2003/#/，访问日期：2022年8月10日。[16] J.-c. Chou、C.-c. Yeh 和 H.-y. Lee，“通过实例归一化分离说话人与内容表征，实现单样本语音转换”，arXiv 预印本 arXiv:1904.05742，2019年。[17] K. Qian、Y. Zhang、S. Chang、X. Yang 和 M. Hasegawa-Johnson，“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》。PMLR，2019年，第5210–5219页。[18] Y. Jia、Y. Zhang、R. Weiss、Q. Wang、J. Shen、F. Ren、P. Nguyen、R. Pang、I. Lopez Moreno、Y. Wu 等，“从说话人验证迁移学习到多说话人文本转语音合成”，《神经信息处理系统进展》，第31卷，2018年。[19] C. Veaux、J. Yamagishi、K. MacDonald 等，“CSTR VCTK语料库：供 CSTR语音克隆工具包使用的英语多说话人语料库”，爱丁堡大学语音技术研究中心（CSTR），2017年。[20] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“LibriSpeech：基于公有领域有声读物的自动语音识别语料库”，载于《2015年 IEEE国际声学、语音与信号处理会议（ICASSP）》。IEEE，2015年，第5206–5210页。[21] M. Ravanelli、T. Parcollet、P. Plantinga、A. Rouhe、S. Cornell、L. Lugosch、C. Sub- akan、N. Dawalatabad、A. Heba、J. Zhong、J.-C. Chou、S.-L. Yeh、S.-W. Fu、C.-F. Liao、E. Rastorgueva、F. Grondin、W. Aris、H. Na、Y. Gao、R. D. Mori 和 Y. Bengio，“SpeechBrain：通用语音工具包”，2021年，arXiv:2106.04624。[22] K. Murano、S. Unagami 和 F. Amano，“回声消除及其应用”，《IEEE通信杂志》，第28卷，第1期，第49–55页，1990年。[23] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，“一种用于预测加权时频噪声语音可懂度的算法”，《IEEE音频、语音与语言处理汇刊》，第19卷，第7期，第2125–2136页，2011年。[24] J. Zhuang、T. Tang、Y. Ding、S. C. Tatikonda、N. Dvornek、X. Papademetris 和 J. Duncan，“AdaBelief优化器：根据对观测梯度的信任自适应调整步长”，《神经信息处理系统进展》，第33卷，第18795–18806页，2020年。[25] S. Hussain、P. Neekhara、S. Dubnov、J. McAuley 和 F. Koushanfar，“WaveGuard：理解并缓解音频对抗样本”，载于第30届 USENIX安全研讨会（USENIX Security 21）。USENIX协会，2021年8月，第2273–2290页。[在线]。网址：https://www.usenix.org/conference/usenixsecu rity21/presentation/hussain [26] J. Cohen、E. Rosenfeld 和 Z. Kolter，“通过随机平滑实现认证对抗鲁棒性”，载于《国际机器学习会议》。PMLR，2019年，第1310–1320页。

[27] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，“迈向能够抵御对抗攻击的深度学习模型”，载于国际学习表征会议，2018年。[28] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao，“通过扩散模型防御对抗音频”，arXiv预印本 arXiv:2303.01507，2023年。[29] H. Yakura and J. Sakuma，“用于物理攻击的鲁棒音频对抗样本”，arXiv预印本 arXiv:1810.11793，2018年。[30] J. Meyer and C. Sydow，“用于麦克风阵列的噪声消除”，载于1997年IEEE国际声学、语音与信号处理会议，第1卷，1997年，第211–213卷第1部分。[31] B. Zheng, P. Jiang, Q. Wang, Q. Li, C. Shen, C. Wang, Y. Ge, Q. Teng, and S. Zhang，“以最少信息对商业语音平台实施黑盒对抗攻击”，载于2021年ACM SIGSAC计算机与通信安全会议论文集，2021年，第86–107页。[32] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu，“采用平滑高斯混合模型和最大后验概率自适应的语音转换”，载于第八届欧洲语音通信与技术会议，2003年。[33] D. Erro, A. Moreno, and A. Bonafonte，“基于加权频率扭曲的语音转换”，IEEE音频、语音与语言处理汇刊，第18卷，第5期，第922–931页，2009年。[34] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj，“使用偏最小二乘回归的语音转换”，IEEE音频、语音与语言处理汇刊，第18卷，第5期，第912–921页，2010年。[35] S. H. Mohammadi and A. Kain，“使用带有说话人无关预训练的深度神经网络进行语音转换”，载于2014年IEEE口语技术研讨会（SLT）。IEEE，2014年，第19–23页。[36] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura，“用于修改基于隐马尔可夫模型的语音合成中调制频谱的后滤波器”，载于2014年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2014年，第290–294页。[37] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang，“使用变分自编码器从非平行语料库进行语音转换”，载于2016年亚太信号与信息处理协会年度峰会暨会议（APSIPA）。IEEE，2016年，第1–6页。[38] C. C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang，“使用变分自编码 Wasserstein生成对抗网络从未对齐语料库进行语音转换”，2017年。[39] T. Kaneko and H. Kameoka，“使用循环一致对抗网络进行无平行数据语音转换”，arXiv预印本 arXiv:1711.11293，2017年。[40] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo，“StarGAN-VC：使用星形生成对抗网络进行非平行多对多语音转换”，载于2018年IEEE口语技术研讨会（SLT）。IEEE，2018年，第266–273页。[41] Y. Gao, R. Singh, and B. Raj，“使用生成对抗网络进行语音冒充”，载于2018年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2018年，第2506–2510页。[42] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu，“WaveNet：一种用于原始音频的生成模型”，arXiv预印本 arXiv:1609.03499，2016年。[43] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan et al.，“通过根据梅尔频谱图预测对WaveNet进行条件化实现自然文本转语音合成”，载于2018年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2018年，第4779–4783页。[44] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou，“Deep Voice 2：多说话人神经文本转语音”，《神经信息处理系统进展》，第30卷，2017年。[45] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller，“Deep Voice 3：通过卷积序列学习扩展文本转语音”，arXiv预印本 arXiv:1710.07654，2017年。[46] N. Carlini and D. Wagner，“音频对抗样本：针对语音转文本的定向攻击”，载于2018年IEEE安全与隐私研讨会（SPW）。IEEE，2018年，第1–7页。[47] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou，“隐藏语音命令”，载于第25届USENIX安全研讨会（USENIX Security 16）。美国得克萨斯州奥斯汀：USENIX协会，2016年8月，第513–530页。[在线]。可访问：https://www.usenix.org/conference/usenixsecurity16/tech nical-sessions/presentation/carlini [48] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter，“{CommanderSong}：实用对抗性语音识别的系统方法”，载于第27届USENIX安全研讨会（USENIX Security 18），2018年，第49–64页。[49] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel，“用于自动语音识别的不可感知、鲁棒且定向的对抗样本”，载于国际机器学习会议。PMLR，2019年，第5231–5240页。[50] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet，“使用对抗样本欺骗端到端说话人验证”，载于2018年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2018年，第1962–1966页。[51] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet，“为语音副语言学应用构造对抗样本”，arXiv预印本 arXiv:1711.03280，2017年。

[52] G. Chen、S. Chenb、L. Fan、X. Du、Z. Zhao、F. Song 和 Y. Liu，“谁是真正的鲍勃？针对说话人识别系统的对抗攻击”，载于*2021*年*IEEE*安全与隐私研讨会（*SP*）。IEEE，2021年，第694–711页。[53] L. Zhang、Y. Meng、J. Yu、C. Xiang、B. Falk 和 H. Zhu，“针对智能家居中说话人验证系统的声纹仿冒攻击”，载于*IEEE INFOCOM 2020-IEEE*计算机通信会议。IEEE，2020年，第377–386页。[54] Z. Li、Y. Wu、J. Liu、Y. Chen 和 B. Yuan，“Advpulse：通过亚秒级扰动实现通用、无需同步且有针对性的音频对抗攻击”，载于*2020*年*ACM SIGSAC*计算机与通信安全会议论文集，2020年，第1121–1134页。[55] Y. Gong、B. Li、C. Poellabauer 和 Y. Shi，“实时对抗攻击”，*arXiv*预印本 *arXiv:1905.13399*，2019年。[56] H. Guo、Y. Wang、N. Ivanov、L. Xiao 和 Q. Yan，“Specpatch：人在回路的针对语音识别的对抗性音频频谱图补丁攻击”，载于*2022*年*ACM SIGSAC*计算机与通信安全会议论文集，2022年，第1353–1366页。[57] J. Qian、H. Du、J. Hou、L. Chen、T. Jung 和 X.-Y. Li，“Hidebehind：享受具备声纹不可克隆性与匿名性的语音输入”，载于第*16*届*ACM*嵌入式网络传感器系统会议论文集，2018年，第82–94页。[58] J. Patino、N. Tomashenko、M. Todisco、A. Nautsch 和 N. Evans，“使用McAdams系数进行说话人匿名化”，*arXiv*预印本 *arXiv:2011 01130*，2020年。。[59] T. Vaidya 和 M. Sherr，“你说得太多了：通过语音输入限制隐私暴露”，载于*2019*年*IEEE*安全与隐私研讨会（*SPW*）。IEEE，2019年，第84–91页。[60] Y. Han、Y. Cao、S. Li、Q. Ma 和 M. Yoshikawa，“语音不可区分性——在不可信服务器环境下利用差分隐私保护声纹”，载于*2020*年*ACM SIGSAC*计算机与通信安全会议论文集，CCS ’20系列。美国纽约州纽约市：计算机协会，2020年，第2125–2127页。 [在线]。可访问：https: //doi.org/ 10 1145 3372297 3420025 。/ 。R. Aloufi、H. Haddadi 和 D. Boyle，“[61] Emotionless：面向语音助手的隐私保护语音分析”，*arXiv*预印本 *arXiv:1908.03632*，2019年。[62] J. Deng、F. Teng、Y. Chen、X. Chen、Z. Wang 和 W. Xu，“V-cloak：保留可懂度、自然度和音色的实时语音匿名化”，*arXiv*预印本 *arXiv:2210.15140* ，2022年。

附录

A 预测模型

根据第 4 节中的 VSMask 系统设计，预测模型的输入和输出应为维度相同、长度不同的二维向量。受 VoiceCamo [11]，中的模型架构启发，我们在 VSMask 预测模型中采用了类似的下采样和上采样架构。为确保最佳预测性能，我们测试了不同的卷积核大小、步幅和层数。我们从卷积核大小 (2, 2) 开始，并确定卷积核大小 (3, 3) 在防护性能与模型复杂度之间实现了最佳权衡。随后，我们仔细调整步幅和层数，以修正最终输出向量的维度。最后，我们使用 tanh 激活层对预测扰动进行归一化。详细的参数配置见表 4。架构参数会因不同语音合成模型所采用的预处理方法不同而有所变化。

B 权重对音频质量的影响

我们为不同频段仔细选择问题 (5) 中的权重，以确保它们具有相同的防护性能。本节将探讨基于权重的幅度约束如何降低受保护音频中的噪声可感知度。

在人类受试者研究中，我们还要求志愿者比较未受权重约束的清晰语音和受保护语音。图 10 展示了受保护音频之间的可感知性差异。

表 4：针对 AdaIN-VC 模型的 VSMask 实时预测模型的详细网络参数。

Type	输入尺寸	卷积核尺寸	步幅
下采样	1×512×100	3×3	(1,2)
	32×512×50 128×512×25 256×256×13 256×128×7 512×64×4 512×32×2	3×3	(2,2)
上采样	512×16×1 256×32×2 128×64×4 64×128×8 32×256×16	3×3	(2,2)
tanh	1×512×32	N/A	N/A

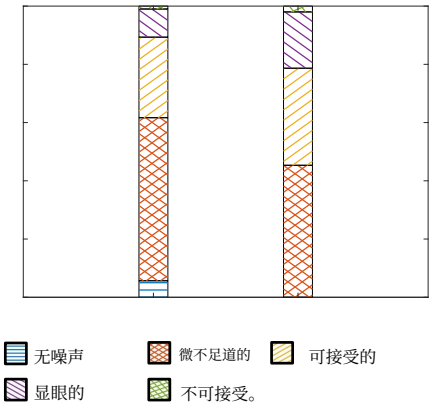


图 10：采用和未采用基于权重的约束时，人类对 VSMask 保护语音的可感知性。

以及不采用加权约束的情况。采用基于权重的约束后，中频扰动会减弱。这样一来，人耳对语音样本中的噪声就不那么敏感。超过 60% 的回答认为噪声可以忽略或无法察觉，这意味着 VSMask 扰动完全不会影响语音中的话语和语音。另有 30% 的回答表示，尽管他们意识到了扰动，但仍能接受语音中的噪声。

相比之下，若移除权重机制，扰动就更容易被察觉。不到 50% 的回答者表示他们能够忽略语音中的扰动，而超过 20% 的结果表明音频中存在明显噪声。少数回答指出噪声过强，影响了语音中的话语和语音。因此，人类受试者研究结果表明，VSMask 的加权约束能够显著降低受保护音频中的音频失真。