通过扩散模型防御对抗音频

ShutongWu[1, 2*] **JiongxiaoWang**[1] **Wei Ping**[3] **Weili Nie**[3] **Chaowei Xiao**[1]1亚利桑那州立大学 [2]上海交通大学 [3]英伟达

arXiv:2303.01507v1 [cs.SD] 2 Mar 2023

摘要

近年来，深度学习模型已广泛应用于商业声学系统。然而，对抗性音频样本会导致这些声学系统出现异常行为，而且人类难以察觉。人们提出了多种方法来保护声学系统免受对抗攻击，例如基于变换的防御和对抗训练，但这些方法在面对自适应攻击时效果较差。此外，由于音频数据具有独特属性，直接套用图像领域的方法可能会导致结果不够理想。本文提出了一种基于对抗性净化的声学系统防御流程 AudioPure，利用现成的扩散模型。借助扩散模型强大的生成能力，AudioPure 首先向对抗音频添加少量噪声，然后执行反向采样步骤，以净化含噪音频并恢复干净音频。AudioPure 是一种即插即用方法，无需微调或重新训练，即可直接应用于任何预训练分类器。我们在语音命令识别任务上进行了大量实验，以评估AudioPure的鲁棒性。我们的方法能够有效应对多种对抗攻击（例如 $\mathcal{L}_2$ 或 $\mathcal{L}_\infty$-范数）。在由 $\mathcal{L}_2$ 或 $\mathcal{L}_\infty$-范数界定的强自适应白盒攻击和黑盒攻击下，我们的方法均优于现有方法，鲁棒准确率最高提升 +20%。此外，我们还通过随机平滑评估了 $\mathcal{L}_2$-范数界定的扰动下的认证鲁棒性。我们的处理流程取得了高于基线方法的认证准确率。代码见 https://github.com/cychomatica/AudioPure。

1 引言

深度神经网络（DNNs）已在音频领域的不同任务中取得了巨大成功，例如语音命令识别、关键词检测、说话人识别和自动语音识别。由深度神经网络构建的声学系统（Amodei 等人，2016；Shen 等人，2019）已应用于从拨打电话到控制家庭安全系统等安全关键型应用中。尽管基于 DNN 的模型已展现出显著的性能提升，但大量研究表明，它们容易受到对抗样本的攻击（Szegedy 等人，2014；Carlini & Wagner，2018；Qin 等人，2019；Du 等人，2020；Abdullah 等人，2021；Chen 等人，2021a）：攻击者向原始音频中加入难以察觉且经过精心设计的扰动，使系统产生错误预测。因此，设计能够抵御对抗样本的稳健型基于 DNN 的声学系统变得至关重要。

为了解决这一问题，现有研究（例如 Rajaratnam & Alshemali，2018；Yang 等人，2019）尝试利用音频的时间依赖特性来防御对抗样本。他们对对抗样本应用时域和频域变换，以提升鲁棒性。尽管这些方法在一定程度上缓解了这一问题，但面对强大的自适应攻击时仍然脆弱，因为攻击者可以获取整个声学系统的全部信息（Tramer 等，2020）。增强对抗样本鲁棒性的另一种方法是对抗训练（Goodfellow 等，2015；Madry 等人，2018），即在训练阶段加入对抗扰动。尽管对抗训练被公认为最有效的防御方法，但其训练过程需要耗费大量计算资源，而且模型仍然

*在 ASU 实习期间完成的工作。

容易受到其他类型对抗样本的攻击，而这些样本与训练过程中使用的样本并不相似（Tramer & Boneh，2019）。

对抗净化（Yoon 等人，2021；Shi 等人，2021；Nie 等人，2022）是另一类防御方法，它利用生成模型在将输入样本送入神经网络之前，净化其中的对抗扰动。这类方法的关键在于设计有效的生成模型来进行净化。近期研究表明，扩散模型在图像（Song 和 Ermon，2019；Ho 等人，2020；Nichol 和 Dhariwal，2021；Dhariwal 和 Nichol，2021）和音频合成（Kong 等人，2021；Chen 等人，2021b）方面已达到最先进水平。这促使研究社区将扩散模型用于净化。具体而言，在图像领域，DiffPure（Nie 等人，2022）将扩散模型用作净化器，在各种图像分类任务中，无论是干净样本准确率还是鲁棒准确率，都取得了良好表现。由于这类方法无需使用预先定义的对抗样本训练模型，因此能够泛化至多种不同的威胁。鉴于扩散模型在图像领域取得了显著进展，我们不禁要问：能否在音频领域取得类似的成功？

与图像领域不同，音频信号具有一些独特属性。音频表示有多种选择，包括原始波形和各种类型的时频表示（例如梅尔频谱图、MFCC）。设计声学系统时，可能会选择某些特定的音频表示作为目标特征，而在某些特征上表现良好的防御方法，在其他特征上可能效果不佳。此外，人们可能会考虑将二维时频表示（即频谱图）视为图像，其中频率轴设为高度，时间轴设为宽度，然后将图像领域中成功的 DiffPure（Nie 等人，2022）直接应用于频谱图。尽管这种做法简单直接，但仍存在两个主要问题：*i)* 声学系统可以接受时长可变的音频作为输入，而 DiffPure 中的底层扩散模型只能处理宽度和高度固定的输入。*ii)* 即使暂且以固定长度分段的方式应用该方法，其效果仍不理想，正如我们将在本文中展示的那样。这些独特的问题给音频领域防御系统的设计与评估带来了新的挑战。

在本文中，我们旨在无需对抗训练即可防御多种未见过的对抗样本。我们提出了一种名为 AudioPure 的即插即用净化流程，该流程基于预训练扩散模型，并利用音频的独特属性。具体而言，我们的方法由两个主要组件构成：（1）基于波形的扩散模型；（2）分类器。该方法以音频波形作为输入，并利用扩散模型净化对抗音频扰动。对于以波形格式表示的对抗输入，AudioPure 首先通过扩散过程加入少量噪声，以覆盖对抗扰动，然后使用截断的反向过程恢复干净样本。恢复后的样本将输入分类器。

我们在语音命令识别任务上开展了广泛实验，以评估我们方法的鲁棒性。我们精心设计了自适应攻击，使攻击者能够准确计算完整梯度，从而评估我们方法的有效性。此外，我们还全面评估了我们的方法对不同黑盒攻击以及变换期望（EOT）攻击的鲁棒性。在针对多种对抗样本的白盒攻击和黑盒攻击下，我们的方法均表现出更好的性能。此外，我们还通过随机平滑评估了 AudioPure 的认证鲁棒性，该方法为模型抵御基于 $\mathcal{L}_2$ 的扰动提供了可证明的鲁棒性保证。我们表明，与基线方法相比，我们的方法能够实现更好的认证鲁棒性。具体而言，与对抗训练相比，我们的方法在鲁棒准确率方面取得了显著提升（最高提升 +20 %），并且认证鲁棒准确率比基线方法高出 5% 以上。据我们所知，我们是首个利用扩散模型增强声学系统安全性，并研究防御方法所处不同工作域如何影响对抗鲁棒性的工作。

2 相关工作

对抗攻击与防御。Szegedy 等人（2014）提出了对抗样本，这些样本看起来与普通样本相似，却会诱使神经网络给出错误预测。通常，对抗样本受到 $\mathcal{L}_p$ 范数的约束，以确保其不可感知性。近年来，更强大的攻击方法不断涌现（Madry 等人，2018；Carlini & Wagner，2017；Andriushchenko 等人，2020；Croce 与 Hein，2020；Xiao 等人，2018a；b；2019；2022b；a；Cao 等人，2019b；a；2022a）。

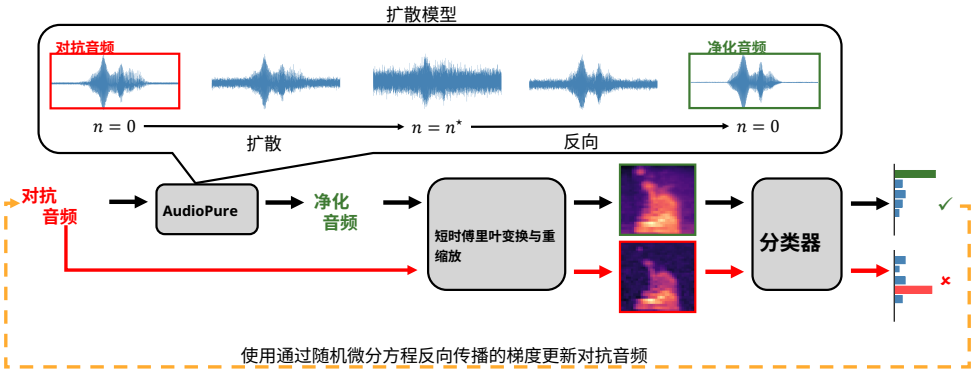


图 1：AudioPure 保护下的整个声学系统架构（图中黑线）以及自适应攻击（图中橙线）。AudioPure 首先向对抗音频添加噪声，然后运行反向过程以恢复净化音频。接着，将净化音频转换为频谱图，并将频谱图输入分类器以获得预测结果。攻击者通过随机微分方程反向传播的梯度来更新对抗音频。没有 AudioPure 时，对抗音频会直接转换为频谱图并输入分类器。

在音频领域，Carlini & Wagner（2018）提出了音频对抗样本，Qin 等人（2019）则设法使其更加难以察觉。此外，人们还开发了黑盒攻击（Du 等人，2020；Chen 等人，2021a），旨在误导端到端声学系统。

为了保护神经网络免受对抗攻击，人们提出了多种防御方法。应用最广泛的是对抗训练（Madry 等人，2018），即有意将对抗样本用作神经网络的训练数据。对抗训练的主要问题在于良性样本的准确率下降，以及高昂的计算成本。许多改进版的对抗训练旨在缓解这些问题（Wong 等人，2020；Shafahi 等人，2019；Zhang 等人，2019b;a；Sun 等人，2021；Cao 等人，2022b；Zhang 等人，2019c）。另一类研究是对抗净化（Yoon 等人，2021；Shi 等人，2021；Nie 等人，2022），即在分类之前使用生成模型去除对抗扰动。这两类防御主要针对计算机视觉任务开发，无法直接应用于音频领域。本文根据音频数据的特点，专门设计了一套防御流程。

语音处理。许多语音处理应用容易受到对抗攻击的影响，包括语音命令识别（Warden，2018）、关键词检测（Chen 等人，2014；Li 等人，2019）、说话人识别（Reynolds 等人，2000；Ravanelli & Bengio，2018；Snyder 等人，2018）以及语音识别（Amodei 等人，2016；Shen 等人，2019；Ravanelli 等人，2019）。具体而言，语音命令识别与关键词检测密切相关，可以视为词汇量有限的语音识别。在本研究中，我们选择语音命令识别作为所提出的 AudioPure 处理流程的测试平台。所提出的处理流程也适用于关键词检测和语音识别。

语音命令识别系统由特征提取器和分类器组成。特征提取器处理原始音频波形并输出声学特征，例如梅尔频谱图或梅尔频率倒谱系数（MFCC）。随后，这些特征被输入分类器，由分类器给出预测结果。对于二维频谱图特征，图像卷积神经网络可以直接应用（Simonyan & Zisserman，2015；He 等人，2016；Zagoruyko & Komodakis，2016；Xie 等人，2017；Huang 等人，2017）。

3 方法

3.1 扩散模型背景

LS

扩散模型通常由前向扩散过程和反向采样过程组成。前向扩散过程逐渐向输入数据添加高斯噪声，直到其分布

收敛到标准高斯分布。反向采样过程以标准高斯噪声为输入，逐渐对含噪数据进行去噪，从而恢复干净样本。目前，扩散模型可分为两种不同类型：基于序贯采样的离散时间扩散模型，例如 SMLD Song & Ermon (2019)、去噪扩散概率模型（Ho 等人，2020）和 DDIM（Song 等人，2021a）；以及基于随机微分方程的连续时间扩散模型（Song 等人，2021c）。Song 等人（2021c）还建立了这两种扩散模型之间的联系。

去噪扩散概率模型（DDPM）（Ho 等人，2020）是应用最广泛的扩散模型之一。此后提出的许多扩散模型，包括用于音频的 DiffWave（Kong 等人，2021），都基于 DDPM 的公式。在 DDPM 中，扩散过程和反向过程均由马尔可夫链定义。对于输入数据 $\mathbf{x}_0 \in \mathbb{R}^d$，我们将$\mathbf{x}_0 \sim q(\mathbf{x}_0)$记作原始数据分布，而$\mathbf{x}_1$、……、$\mathbf{x}_N$是来自分布$q(\mathbf{x}_1|\mathbf{x}_0)$……，$q(\mathbf{x}_N|\mathbf{x}_{N-1})$的中间潜变量，其中 N为总步数。通常，给定预定义或学习得到的方差调度 β_1、……、β_N（通常为线性递增的小常数），前向转移概率 $q(\mathbf{x}_n|\mathbf{x}_{n-1})$可表示为：

$$q(\mathbf{x}_n|\mathbf{x}_{n-1}) = \mathcal{N}(\mathbf{x}_n; \sqrt{1-\beta_n}\mathbf{x}_{n-1}, \beta_n\mathbf{I}), \tag{1}$$

基于方差调度 $\{\beta_n\}$，定义一组常数如下：

$$\alpha_n = 1-\beta_n, \quad \bar{\alpha}_n = \prod_{n=1}^{N}\alpha_n, \quad \tilde{\beta}_n = \begin{cases} \frac{1-\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n}\beta_n, & n>1 \\ \beta_1, & n=1 \end{cases}, \tag{2}$$

利用重参数化技巧，我们有：

$$q(\mathbf{x}_n|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_n; \sqrt{\bar{\alpha}_n}\mathbf{x}_0, (1-\bar{\alpha}_n)\mathbf{I}) \tag{3}$$

当 n 逐渐增大至无穷大时，$q(\mathbf{x}_n|\mathbf{x}_0)$将收敛到标准高斯分布。同时，对于反向过程，我们有：

$$\mathbf{x}_{n-1} \sim p_\theta(\mathbf{x}_{n-1}|\mathbf{x}_n) = \mathcal{N}(\mathbf{x}_{n-1}; \mu_\theta(\mathbf{x}_n, n), \sigma_\theta^2(\mathbf{x}_n, n)\mathbf{I}), \tag{4}$$

whe其中，均值项 $\mu_\theta(\mathbf{x}_n, n)$和方差项 $\sigma_\theta^2(\mathbf{x}_n, n)$由参数 θ实例化。H o 等人。（2020）；Kong 等人（2021）使用神经网络 $\boldsymbol{\epsilon}_\theta$ 来定义 μ_θ，并将 σ_θ 固定为常数 ：

$$\mu_\theta(\mathbf{x}_n, n) = \frac{1}{\sqrt{\alpha_n}}\left(\mathbf{x}_n - \frac{\beta_n}{\sqrt{1-\bar{\alpha}_n}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_n, n)\right), \quad \sigma_\theta(\mathbf{x}_n, n) = \sqrt{\tilde{\beta}_n}. \tag{5}$$

我们将 $\mathbf{x}_n(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_n}\overline{\mathbf{x}_0} + \sqrt{(1-\overline{\bar{\alpha}_n})\boldsymbol{\epsilon}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 记为，优化目标为：

$$\theta^\star = \arg\max_\theta \sum_{n=1}^{N}\lambda_n \mathbb{E}_{\mathbf{x}(0)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_n}\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_n)}\boldsymbol{\epsilon}, n)\right\|_2^2 \tag{6}$$

其中 λ_n 是权重系数（Ho 等人，2020）。

根据 Song 等人（2021c），当 $N \to \infty$ 时，DDPM 成为 VP-SDE，即扩散模型的连续时间形式。具体而言，前向SDE 表述为：

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}. \tag{7}$$

其中 $t \in [0,1]$，$\mathrm{d}t$ 是无穷小的正时间步长，$\mathbf{w}$ 是标准维纳过程，$\beta(t)$ 是连续时间噪声调度。类似地，反向SDE 可定义为：

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)[\mathbf{x} + 2\nabla_\mathbf{x}\log p_t(\mathbf{x})]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\bar{\mathbf{w}}, \tag{8}$$

其中 $\mathrm{d}t$ 是无穷小的负时间步长，$\bar{\mathbf{w}}$ 是反向时间的标准维纳过程。

3.2 AUDIOPURE：面向声学系统的即插即用防御

为统一防御的表述方式，按照 Nie 等人（2022）的建议，我们采用式 7 和式 8 定义的连续时间表述。需要注意的是，由于现有的预训练 DiffWave 模型（Kong 等人，2021）基于去噪扩散概率模型（DDPM），因此我们将使用与其等价的 VP-SDE。

如果我们使用 Euler–Maruyama方法求解 VP-SDE，且步长为 $\Delta t = \frac{1}{N}$，那么反向时间SDE 的采样将等价于去噪扩散概率模型（DDPM）的反向采样（详细证明参见 Song 等人（2021c））。在此前提下，我们有 $t = \frac{n}{N}$其中 $n \in \{1, \dots, N\}$。我们定义 $\beta(\frac{n}{N}) := \beta_n$, $\bar{\alpha}(\frac{n}{N}) := \bar{\alpha}_n$, $\beta(\frac{n}{N}) := \beta_n$，以及 $\mathbf{x}(\frac{n}{N}) := \mathbf{x}_n$给定一个对抗样本$x_{adv}$ 作为 $t =$处的输入，即 $\mathbf{x}_0 = \mathbf{x}_{adv}$，我们首先通过求解式 7，将前向SDE从 $t = 0$ 运行至 $t^\star = \frac{n^\star}{N}$（这等价于运行 n^* 个 DPPM 步骤），从而得到：

$$\mathbf{x}(t^\star) = \sqrt{\bar{\alpha}(t^\star)}\mathbf{x}_{adv} + \sqrt{1 - \bar{\alpha}(t^\star)}\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}), \tag{9}$$

接下来，我们通过求解式 8，从 $t = t^\star$ 到 $t = 0$ 运行截断的反向随机微分方程。与 Nie 等人（2022）类似，我们定义了一个使用 Euler–Maruyama方法的随机微分方程求解器 **sdeint**，并依次输入六个参数：初始值、漂移系数、扩散系数、维纳过程、初始时间和结束时间。在 $t = 0$处的反向输出 $\hat{\mathbf{x}}(0)$ 可表示为：

$$\hat{\mathbf{x}}(0) = \mathbf{sdeint}(\mathbf{x}(t^\star), f_{rev}, g_{rev}, \bar{\mathbf{w}}, t^\star, 0). \tag{10}$$

其中，漂移系数和扩散系数分别为：

$$f_{rev}(\mathbf{x}, t) := -\frac{1}{2}\beta(t)[\mathbf{x} + 2\mathbf{s}_\theta(\mathbf{x}, t)], \qquad g_{rev}(t) := \sqrt{\tilde{\beta}(t)}. \tag{11}$$

请注意，为了获得更干净的输出，我们使用了不同于 Nie 等人（2022）的扩散系数（详细说明见第3.3节）。接下来，我们使用离散时间噪声估计器 $\epsilon_\theta(\mathbf{x}_n, n)$ 来计算连续时间得分估计器 $s_\theta(\mathbf{x}\ t)$。定义 $\tilde{\epsilon}_\theta(\mathbf{x}(t)\ t) := \epsilon_\theta(\mathbf{x}(\frac{n}{N})\ n) = \epsilon_\theta(\mathbf{x}_n, n)$，其中 $t := \frac{n}{N}$，则反向 VP-SDE 中的得分函数可估计为：

$$\mathbf{s}_\theta(\mathbf{x}, t) = -\frac{\tilde{\epsilon}_\theta(\mathbf{x}, t)}{\sqrt{1 - \bar{\alpha}(t)}} \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x}). \tag{12}$$

因此，$\hat{\mathbf{x}}(0)$，即对抗输入$\mathbf{x}(0) = \mathbf{x}_{adv}$净化后的输出，会被送入声学系统的后续阶段进行预测。整个净化操作可以表示为函数**净化器**：$\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$:

$$\mathbf{Purifier}(\mathbf{x}_{adv}, n^\star) = \mathbf{sdeint}\left(\sqrt{\bar{\alpha}(\frac{n^\star}{N})}\mathbf{x}_{adv} + \sqrt{1 - \bar{\alpha}(\frac{n^\star}{N})}\mathbf{z}, f_{rev}, g_{rev}, \bar{\mathbf{w}}, \frac{n^\star}{N}, 0\right) \tag{13}$$

声学系统通常基于从原始音频中提取的特征构建。例如，系统可以提取梅尔频谱图作为特征：1）首先对时域波形应用短时傅里叶变换（STFT），得到线性尺度频谱图；2）然后将频带重新缩放至梅尔刻度。我们将这一过程表示为 Wave2Mel：$\mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^n$，它是一个可微函数。随后，分类器 F：$\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^c$（通常为卷积网络）以梅尔频谱图为输入并给出预测。

由于时域波形和时频域频谱图都会经过该处理流程，因此净化器既可以应用于时域，也可以应用于时频域。如果净化器应用于时域，则整个防御型声学系统**AS**：$\mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^c$ 可以表示为：

$$\mathbf{AS}(\mathbf{x}_{adv}, n^\star) = F(\text{Wave2Mel}(\mathbf{Purifier}(\mathbf{x}_{adv}, n^\star))) \tag{14}$$

其中，基于 DiffWave 的是波形

同时，如果我们想在时频域净化输入的对抗样本，可以选择用于图像合成的扩散模型，并将其应用于 Wave2Mel 的输出频谱图。我们将此净化器记作 **净化器**$_{spec}$：
$\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m \times \mathbb{R}^n$。在这种情况下，整个防御声学系统将为：

$$\mathbf{AS}(\mathbf{x}_{adv}, n^\star) = F(\mathbf{Purifier}_{spec}(\text{Wave2Mel}(\mathbf{x}_{adv})), n^\star) \tag{15}$$

整个处理流程的架构如图 1 所示。对于时频域频谱图的净化，我们使用在音频数据的梅尔频谱图上训练的改进型 DDPM（Nichol 和 Dhariwal，2021），并将其称为 *DiffSpec*。我们比较了这两种净化器，发现时域波形的净化更能有效抵御对抗音频。详细的实验结果见第4.2节。

3.3 迈向评估 AudioPure

自适应攻击 对于按公式 9 定义的前向扩散过程，输出$\mathbf{x}(t^\star)$ 相对于 输入$\mathbf{x}(0)$ 的梯度是常数。对于按公式 10 定义的反向过程，我们应用伴随方法（Li 等，2020），通过求解另一个增广随机微分方程来计算目标函数 $\mathcal{L}$ 相对于$x(t^\star)$ 的完整梯度，且不会出现内存溢出问题:

$$\begin{pmatrix} \mathbf{x}(t^\star) \\ \frac{\partial \mathcal{L}}{\partial \mathbf{x}(t^\star)} \end{pmatrix} = \mathbf{sdeint}\left(\begin{pmatrix} \mathbf{x}(0) \\ \frac{\partial \mathcal{L}}{\partial \mathbf{x}(0)} \end{pmatrix}, \begin{pmatrix} f_{rev} \\ \frac{\partial f_{rev}}{\partial \mathbf{x}}\mathbf{z} \end{pmatrix}, \begin{pmatrix} g_{rev}\mathbf{1} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} -\mathbf{w}(1-t) \\ -\mathbf{w}(1-t) \end{pmatrix}, 0, t^\star \right) \tag{16}$$

其中**1** 和**0** 分别表示全 1 向量和全 0 向量。

用于干净输出的随机微分方程修改 我们观察到，直接将 Nie 等人（2022）提出的框架应用于音频领域会导致性能下降。也就是说，当将 DiffWave（Kong 等人，2021）的离散时间反向过程转换为公式 8 中对应的反向 VP-SDE 时，输出音频仍然含有大量噪声，从而导致分类准确率降低。我们确定了两个影响因素，并通过修改随机微分方程的形式来解决这一问题。

第一个因素是扩散误差，这是由于离散情况与连续情况下的反向方差不匹配所导致的。Ho 等人（2020）在图像领域通过实验观察到， $\sigma_\theta^2 = \tilde{\beta}_t$ 和 $\sigma_\theta^2 = \beta_t$ 都能取得相似的结果。然而，我们发现，在使用扩散模型进行音频建模时，情况并非如此。对于使用 $\sigma_\theta^2 = \tilde{\beta}_t$训练的 DiffWave 进行音频合成，如果我们将反向方差调度切换为 $\sigma_\theta^2 = \beta_t$，输出音频就会产生噪声。因此，在第 3.2 节中，我们定义$\tilde{\beta}\left(\frac{n}{N}\right) = \tilde{\beta}_n$，并在公式 11 中使用扩散系数 $g_{rev} = \sqrt{\tilde{\beta}}(t)$ 而不是 $g_{rev} = \sqrt{\beta}(t)$，以匹配 DiffWave 中的方差 β_t。

第二个因素是 Nie 等人（2022）采用的连续时间噪声调度 $\beta(t) = \beta_0 + (\beta_N - \beta_0)t$和 $\tilde{\alpha}(t) = e^{-\int_0^t \beta_{(s)} \mathrm{d}s}$ 存在不准确性。 $\beta(t) = \beta_0 + (\beta_N - \beta_0)t$ 与 β_{Nt} 之间的差异所产生的影响不可忽略，尤其是在 N不够大时（例如 $N = 200$对于我们使用的预训练 DiffWave 模型）。此外，当 t接近 0 时， $\tilde{\alpha}(t) = e^{-\int_0^t \beta_{(s)} \mathrm{d}s}$不再是 $\bar{\alpha}_{Nt}$ 的良好近似。因此，为了获得更好的去噪输出和更准确的梯度计算，我们直接基于离散调度定义连续时间噪声调度，即 $\beta\left(\frac{n}{N}\right) := \beta_n$ 和 $\bar{\alpha}\left(\frac{n}{N}\right) := \bar{\alpha}_n$。

4 实验

本节首先介绍详细的实验设置。随后，我们在强白盒自适应攻击和黑盒攻击下，对比我们的方法与其他防御方法的性能；在白盒自适应攻击中，攻击者完全了解防御机制。为进一步展示我们方法的鲁棒性，我们还通过随机平滑评估认证准确率（Cohen 等人（2019）），该方法可对模型在 $\mathcal{L}_2$ 范数有界对抗扰动下的鲁棒性提供可证明的保证。

4.1 实验设置

数据集。我们在语音命令识别任务上评估我们的方法。我们使用 Speech Commands 数据集（Warden，2018），该数据集包含 85,511 条训练语音、10,102 条验证语音和 4,890 条测试语音。按照 Kong 等人（2021）的设置，我们选取表示数字的语音，并将此子集记为 SC09。

模型。我们使用 DiffWave（Kong 等人，2021）和 DiffSpec（基于改进型 DDPM（Nichol 和 Dhariwal，2021 ））作为防御净化器，它们分别是波形域和频谱域中具有代表性的扩散模型。我们使用 DiffWave 的无条件版本及官方提供的预训练检查点。由于改进型 DDPM 模型没有提供音频预训练检查点，我们从头开始使用 SC09 音频的梅尔频谱图进行训练。训练细节和超参数见附录 A。对于分类器，在消融研究之外的实验中，我们对频谱图表示使用 ResNeXt-29-8-64（Xie 等人，2017），对波形使用 M5 Net（Dai 等人，2017）。

表 1：不同方法应对自适应攻击的性能。

防御	干净样本	$\mathcal{L}_\infty$ 白盒						$\mathcal{L}_2$ 白盒					$\mathcal{L}_\infty$ 黑盒
		投影梯度下降$_{10}$	投影梯度下降$_{20}$	投影梯度下降$_{40}$	投影梯度下降$_{50}$	投影梯度下降$_{70}$	投影梯度下降$_{100}$	投影梯度下降$_{10}$	投影梯度下降$_{20}$	投影梯度下降$_{40}$	投影梯度下降$_{50}$	投影梯度下降$_{100}$	FAKEBOB
None	**100**	3	1	1	1	1	1	2	0	0	0	0	21
AS（Yang 等人，2019）	**100**	4	2	1	1	1	1	1	0	0	0	0	24
MS（Yang 等人，2019）	**100**	6	3	2	1	1	1	4	1	0	0	0	21
DS（Yang 等人，2019）	99	2	1	1	1	1	1	0	0	0	0	0	16
带通滤波器 (Rajaratnam & Alshemali, 2018)	**100**	5	2	1	1	1	1	2	0	0	0	0	20
带通滤波器 (Rajaratnam & Alshemali, 2018)	99	5	1	1	1	1	1	1	1	0	0	0	18
对抗训练 (Madry 等人, 2018)	**100**	86	79	78	74	72	71	73	70	68	65	65	**92**
AudioPure	97	**89**	**89**	**89**	**85**	**84**	**84**	**89**	**86**	**83**	**85**	**84**	86

攻击。对于白盒攻击，我们使用投影梯度下降（Madry 等人，2018），在 L_∞ 和 L_2 范数下采用 10 到 100 不等的迭代步数。除消融研究外，对于 $\mathcal{L}_\infty$-范数约束，攻击预算设为 $\epsilon = 0.002$；对于 L_2 范数约束，则设为 $\epsilon = 0.253$。对于黑盒攻击，我们采用基于查询的攻击 FAKEBOB（Chen 等人，2021a），将迭代步数设为 200、NES 样本数设为 200，并设置置信系数 $\kappa = 0.5$。

基线方法。我们将我们的方法与两类基线方法进行比较，包括：(1) 基于变换的防御方法（Yang 等人, 2019; Rajaratnam & Alshemali, 2018），包括平均平滑（AS）、中值平滑（MS）、下采样（DS）、低通滤波器（LPF）和带通滤波器（BPF）；以及 (2) 基于对抗训练的防御方法（对抗训练）（Madry 等人, 2018）。对于对抗训练，我们遵循 Chen 等人（2022）的设置，使用 $\mathcal{L}_\infty$ 投影梯度下降$_{10}$，并采用 $\epsilon = 0.002$ 和 $ratio = 0.5$。

4.2 主要结果

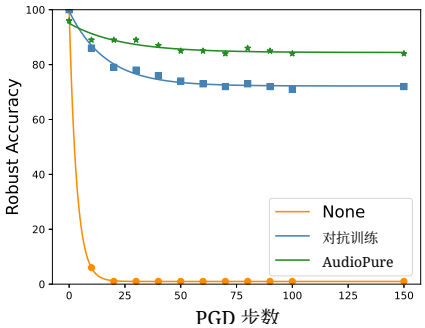
我们在自适应攻击下评估AudioPure （默认$n^\star$=3 ），并假设攻击者完全了解我们的防御。我们采用上一节所述的自适应攻击算法，使攻击者能够准确计算完整梯度以发起攻击。结果如表 1 所示。我们发现，基线的基于变换的防御方法（Yang 等人，2019；Rajaratnam & Alshemali，2018），包括平均平滑（AS）、中值平滑（MS）、降采样（DS）、低通滤波器（LPF）和带通滤波器（BPF），几乎都被白盒攻击攻破，鲁棒准确率最高仅为 4%。对于在 $\mathcal{L}_\infty$-范数对抗样本上训练的对抗训练方法（AdvTr），尽管其针对基于 $\mathcal{L}_\infty$的对抗样本取得了 71% 的鲁棒准确率，但该方法对其他类型的对抗样本（即基于 $\mathcal{L}_2$的方法）效果不佳，在基于 $\mathcal{L}_2$的 PDG$_{100}$ 攻击下鲁棒准确率为 65%。与所有基线方法相比，AudioPure 在基于 $\mathcal{L}_\infty$的对抗样本上能取得高得多的鲁棒准确率，平均提升约 10%；面对基于 $\mathcal{L}_2$的白盒攻击时同样有效，鲁棒准确率达到 84%。

我们还评估了AudioPure 在黑盒攻击下的表现，包括：FAKEBOB（Chen 等人，2021a）和基于可迁移性的攻击。FAKEBOB 的结果见表 1，表明我们的方法在基于查询的黑盒攻击下仍然有效。基于可迁移性的攻击结果见附录 B，得出的结论相同。这些结果进一步验证了我们方法的有效性。所有结果均表明，AudioPure 能够应对具有不同类型约束的多种攻击；而对抗训练必须采用不同的训练策略并重新训练模型，因此在面对未见过的攻击时，其效果不如我们的方法。我们在附录 J 中报告了实际推理时间，并在附录 F、G 和 H 中将我们的方法与更多现有方法进行了比较。此外，我们还在 Qualcomm 关键词语音数据集（Kim 等人，2019）上进行了实验，结果和详情见附录 E。在该数据集上，我们的方法仍能有效抵御对抗样本。

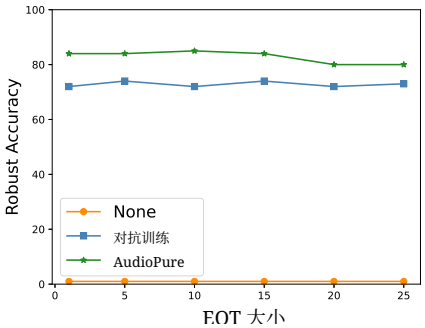
4.3 消融研究

PGD 步数为确保 PGD 攻击的有效性，我们测试了从 10 到 150 的不同迭代步数。如图 2a 所示，鲁棒准确率在经过迭代步数 $n \geq 70$后趋于收敛。

抵御变换期望（EOT）攻击的有效性。此外，由于AudioPure 的扩散和反向过程包含许多随机采样操作，我们采用变换期望（EOT）攻击来评估AudioPure 的有效性，具体如下



(a) 不同 PGD 步数下的鲁棒准确率



(b) 不同 EOT 大小下的鲁棒准确率

图 2：基线（无防御，记为 None）、对抗训练（记为 AdvTr）和 AudioPure 在不同迭代步数和 EOT 大小的攻击下的性能。(a) 表示收敛步数，迭代 70 步后攻击几乎达到最优。(b) 表明，增大 EOT 大小几乎不会影响我们方法的鲁棒性。

表2：使用不同反向步数 $n^\star$时，PGD_{10} 在不同攻击预算 ϵ 下的鲁棒准确率。较大的 ϵ 需要较大的 $n^\star$，才能确保更好的鲁棒性。

攻击预算	扩散步数 $n^\star=0$	$n^\star=1$	$n^\star=2$	$n^\star=3$	$n^\star=5$	$n^\star=7$	$n^\star=10$
$\epsilon=0.002$	3	**94**	90	89	84	77	67
$\epsilon=0.004$	0	76	**89**	86	83	74	66
$\epsilon=0.008$	0	27	70	**85**	84	74	68
$\epsilon=0.016$	0	0	21	53	**69**	57	63

表 3：不同模型架构的消融研究。在 $\mathcal{L}_\infty$- PGD_{70} 攻击下评估鲁棒准确率。我们的方法适用于具有不同架构的各种模型。

防御	ResNeXt-29-8-64		VGG-19-BN		WideResNet-28-10		DenseNet-BC-100-12		M5	
	干净样本	稳健	干净样本	稳健	干净样本	稳健	干净样本	稳健	干净样本	稳健
None	100	1	100	2	100	1	100	5	94	12
AudioPure	97	84	99	81	99	85	96	79	94	70

不同的 EOT 样本数量。图 2b 展示了结果。我们发现，AudioPure 在不同的 EOT 数量下均有效。

攻击预算 ϵ。我们在不同的 ϵ （包括$\epsilon = \{0.002, 0.004, 0.008, 0.016\}$）下评估了我们方法的有效性。由于扩散步数 $n^\star$ 是 AudioPure 的超参数，因此我们在不同的 $n^\star$下开展了实验。如表2所示，当 $n^\star$ 大于2时，AudioPure 在不同的 ϵ下均表现出较强的有效性。随着 ϵ 增加，要实现最佳鲁棒性，就需要更大的$n^\star$，因为更大的对抗扰动需要扩散模型前向过程产生更大的噪声来覆盖对抗扰动，并需要相应更大的步数来恢复净化音频。然而，如果 $n^\star$ 过大，它也会覆盖原始音频信息，导致从扩散模型中恢复的音频丢失原始音频信息，从而造成性能下降。此外，我们在附录 I 中探究了扩散模型用于净化的能力范围。

架构。此外，我们将 AudioPure 应用于不同的分类器，包括基于频谱图的分类器：VGG-19-BN（Simonyan & Zisserman，2015）、ResNeXt-29-8-64（Xie 等人，2017）、WideResNet-28-10（Zagoruyko & Komodakis，2016）、DenseNet-BC-100-12（Huang 等人，2017），以及基于波形的分类器：M5（Dai 等人，2017）。表 3 显示，我们的方法对各种神经网络分类器均有效。

音频表示 音频有不同类型的表示，包括原始波形或时频表示（例如梅尔频谱图）。我们通过使用不同的表示进行消融研究，以展示扩散模型的有效性，包括用于波形的扩散模型 *DiffWave*（Kong 等人，2021）以及基于原始图像模型、用于频谱图的扩散模型 *DiffSpec*（Nichol 和 Dhariwal，2021）。结果如表 4 所示。我们发现

表 4：不同音频表示之间的消融研究。我们使用两种不同的扩散模型作为净化器来实现 AudioPure：DiffWave 和 DiffSpec，它们分别处理时域和时频域中的表示。

防御	干净样本	$\mathcal{L}_\infty$ 白盒						$\mathcal{L}_2$ 白盒				
		PGD_{10}	PGD_{20}	PGD_{30}	PGD_{50}	PGD_{70}	PGD_{100}	PGD_{10}	PGD_{20}	PGD_{30}	PGD_{50}	PGD_{100}
DiffWave	97	89	**89**	**89**	**85**	**84**	**84**	**89**	**86**	**83**	**85**	**84**
DiffSpec	**99**	**92**	84	78	75	72	71	74	62	58	54	49

表 5：不同方法的认证准确率。对于每个噪声水平 σ，我们添加相同水平的噪声来训练分类器，并将其应用于 RS-Gaussian。

方法	噪声水平	认证半径 ($\mathcal{L}_2$)						
		0	0.25	0.50	0.75	1.0	1.25	1.50
RS-Vanilla	$\sigma = 0.5$	30	21	12	6	4	3	3
	$\sigma = 1.0$	8	8	8	7	4	3	3
RS-Gaussian	$\sigma = 0.5$	**49**	39	33	23	14	6	3
	$\sigma = 1.0$	18	15	11	10	5	5	4
AudioPure	$\sigma = 0.5$	45	**40**	**35**	**27**	**21**	**17**	**13**
	$\sigma = 1.0$	**27**	**22**	**16**	**15**	**12**	**11**	**8**

与基于 $\mathcal{L}_2$ 和 $\mathcal{L}_\infty$ 的对抗样本对抗时，*DiffWave* 始终优于*DiffSpec* 。此外，与*DiffWave*相比，尽管*DiffSpec* 的干净准确率更高，但其鲁棒准确率仅为 49%；面对基于 $\mathcal{L}_2$ 的对抗样本时，性能显著下降了 35%。我们认为，潜在原因是短时傅里叶变换（STFT）是一种信息压缩操作。频谱图包含的信息远少于原始音频波形。这项实验表明，领域差异会导致结果显著不同，直接将图像领域的方法应用于音频可能会导致性能不佳。这也验证了AudioPure 在实现对抗鲁棒性方面的关键设计。

4.4 认证鲁棒性

在本节中，我们通过随机平滑（Cohen 等，2019）评估AudioPure 的认证鲁棒性。这里，我们抽取 $N = 100{,}000$ 个噪声样本，并选择噪声水平 $\sigma \in \{0.5, 1.0\}$进行认证。请注意，我们沿用 Carlini 等人（2022）的设置，并选择使用单步去噪方法。我们方法的详细实现见附录 C。我们将结果与使用原始分类器和高斯增强分类器进行随机平滑的结果进行比较，分别记为 RS-Vanilla 和 RS-Gaussian。结果如表 5 所示。我们还在附录 C 的图 A 中展示了不同高斯噪声 $\sigma = \{0.5, 1.0\}$ 下、不同 $\mathcal{L}_2$ 扰动预算对应的认证鲁棒性。观察结果可知，除半径为 0 的情况外，我们的方法优于基线方法，获得了更高的认证准确率。我们还注意到，随着输入噪声增大，我们方法的性能会进一步提升。这可能是因为AudioPure 在面对较大的 $\mathcal{L}_2$-基扰动时仍能恢复干净音频，而高斯增强模型在使用如此大的噪声进行训练时甚至无法收敛。

5 结论

本文提出了一种用于声学系统的基于对抗净化的防御流程。为评估AudioPure 的有效性，我们设计了自适应攻击方法，并在自适应攻击、EOT 攻击和黑盒攻击下评估了该方法。全面的实验表明，在多种对抗样本中，我们的防御方法比现有方法（包括对抗训练）更加有效。我们还表明，与其他基线方法相比，AudioPure 通过随机平滑实现了更强的可认证鲁棒性。此外，我们的防御方法可作为一种通用的即插即用方法，适用于采用不同架构的分类器。

局限性。AudioPure 引入了扩散模型，从而增加了时间和计算成本。因此，如何提高时间和计算效率是一项重要的未来工作。例如，研究蒸馏技术（Salimans 和 Ho，2022）和快速采样方法（Kong 和 Ping，2021），以降低扩散模型带来的计算复杂度，是很有意义的。

伦理声明

我们的工作提出了一种防御流程，用于保护声学系统免受对抗音频样本的攻击。具体而言，我们的研究聚焦于语音命令识别，这与关键词检出系统密切相关。众所周知，此类系统容易受到对抗攻击。我们的处理流程将提升此类现实世界声学系统的安全性，并造福社会大众。本研究使用的 Speech Commands 数据集由其他研究者发布，且已公开多年。该数据集包含匿名说话者的各种语音。据我们所知，其中不包含任何与这些说话者隐私相关的信息。

参考文献

Hadi Abdullah、Muhammad Sajidur Rahman、Washington Garcia、Kevin Warren、Anurag S warnim Yadav、Tom Shrimpton 和 Patrick Traynor。Hear" no evil"，see" kenansville"*：针对语音识别和语音识别系统的高效、可迁移黑盒攻击。发表于*2021*年*IEEE*安全与隐私研讨会（*SP*），第 712–729 页。I EEE，2021 年。

Dario Amodei、Sundaram Ananthanarayanan、Rishita Anubhai、Jingliang Bai、Eric Battenberg、Carl Case、Jared Casper、Bryan Catanzaro、Qiang Cheng、Guoliang Chen 等。Deep speech 2：英语和普通话的端到端语音识别。发表于国际机器学习会议，第 173–182 页。PMLR，2016 年。

Maksym Andriushchenko、Francesco Croce、Nicolas Flammarion 和 Matthias Hein。Square attack：一种通过随机搜索实现的查询高效黑盒对抗攻击。发表于欧洲计算机视觉会议，第 484–501 页。Springer，2020 年。

Yulong Cao、Chaowei Xiao、Benjamin Cyr、Yimeng Zhou、Won Park、Sara Rampazzi、Qi Alfred Chen、Kevin Fu 和 Z Morley Mao。针对自动驾驶中基于激光雷达的感知的对抗性传感器攻击。*CCS*，2019a。

YulongCao、ChaoweiXiao、DaweiYang、JingFang、Ruigang Yang、MingyanLiu 和BoLi。针对基于激光雷达的自动驾驶系统的对抗性物体。*arXiv* 预印本 *arXiv:1907.05418*，2019b。

Yulong Cao、Chaowei Xiao、Anima Anandkumar、Danfei Xu 和 Marco Pavone。AdvDO：用于轨迹预测的逼真对抗攻击。载于欧洲计算机视觉会议，第 36–52 页。Springer，2022a。

Yulong Cao、Danfei Xu、Xinshuo Weng、Zhuoqing Mao、Anima Anandkumar、Chaowei Xiao 和 Marco Pavone。针对对抗攻击的鲁棒轨迹预测。*CORL*（口头报告），2022b。

Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于*2017*年*IEEE*安全与隐私研讨会 *(SP)*，第 39–57 页。IEEE，2017。

Nicholas Carlini 和 David 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)*，第 1–7 页。IEEE，2018。

Nicholas Carlini、Florian Tramer、J Zico Kolter 等。免费实现（经认证的！！）对抗鲁棒性！*arXiv* 预印本 *arXiv:2206.10550*，2022。

Guangke Chen、Sen Chenb、Lingling Fan、Xiaoning Du、Zhe Zhao、Fu Song 和 Yang 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)*，第 694–711 页。IEEE，2021a。

Guangke Chen、Zhe Zhao、Fu Song、Sen Chen、Lingling Fan、Feng Wang 和 Jiashui Wang。迈向理解和缓解说话人识别中的音频对抗样本。*IEEE*可靠与安全计算汇刊，2022。

Guoguo Chen、Carolina Parada 和 Georg Heigold。使用深度神经网络进行小型关键词检测。载于*2014 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 4087–4091 页。IEEE，2014。

Nanxin Chen、Yu Zhang、Heiga Zen、Ron J Weiss、Mohammad Norouzi 和 William Chan。Wave-grad：估计波形生成的梯度。载于国际学习表征会议，2021b。Jeremy Cohen、Elan Rosenfeld 和 Zico Kolter。通过随机平滑实现可认证的对抗鲁棒性。载于国际机器学习会议，第 1310–1320 页。PMLR，2019。Francesco Croce 和 Matthias Hein。使用多种无参数攻击组成的集成，可靠地评估对抗鲁棒性。载于国际机器学习会议，第 2206–2216 页。PMLR，2020。Wei Dai、Chia Dai、Shuhui Qu、Juncheng Li 和 Samarjit Das。用于原始波形的超深卷积神经网络。载于*2017 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第 421–425 页。IEEE，2017。Prafulla Dhariwal 和 Alexander Nichol。扩散模型在图像合成方面胜过生成对抗网络。神经信息处理系统进展，34:8780–8794，2021。Chris Donahue、Julian McAuley 和 Miller Puckette。对抗音频合成。载于国际学习表征会议，2018。Tianyu Du、Shouling Ji、Jinfeng Li、Qinchen Gu、Ting Wang 和 Raheem Beyah。Sirenattack：为端到端声学系统生成对抗音频。载于第*15*届*ACM*亚洲计算机与通信安全会议论文集，第 357–369 页，2020。Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。载于国际学习表征会议，2015。Shixiang Gu 和 Luca Rigazio。迈向能够抵御对抗样本的深度神经网络架构。*arXiv*预印本 *arXiv:1412.5068*，2014。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第 770–778 页，2016。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020。Judy Hoffman、Daniel A Roberts 和 Sho Yaida。采用雅可比正则化的稳健学习。*arXiv*预印本 *arXiv:1908.02729*，2019。Gao Huang、Zhuang Liu、Laurens Van Der Maaten 和 Kilian Q Weinberger。密集连接卷积网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 4700–4708 页，2017。Sonal Joshi、Saurabh Kataria、Yiwen Shao、Piotr Zelasko、Jesus Villalba、Sanjeev Khudanpur 和 Najim Dehak。使用去噪器联合对抗微调，防御混合语音识别中的对抗攻击。*arXiv*预印本 *arXiv:2204.03851*，2022。Byeonggeun Kim、Mingu Lee、Jinkyu Lee、Yeonseok Kim 和 Kyuwoong Hwang。基于查询的设备端关键词检测，2019。Zhifeng Kong 和 Wei Ping。关于扩散概率模型的快速采样。*arXiv*预印本 *arXiv:2106.00132*，2021。Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro。Diffwave：一种用于音频合成的通用扩散模型。载于国际学习表征会议，2021。Juncheng Li、Shuhui Qu、Xinjian Li、Joseph Szurley、J Zico Kolter 和 Florian Metze。对抗音乐：针对唤醒词检测系统的真实世界音频对抗样本。神经信息处理系统进展，32，2019。

Xuechen Li、Ting-Kam Leonard Wong、Ricky TQ Chen 和 David Duvenaud。随机微分方程的可扩展梯度。载于人工智能与统计国际会议，第 3870–3882 页。PMLR，2020。Yi Luo 和 Nima Mesgarani。Conv-tasnet：超越用于语音分离的理想时频幅度掩蔽。*IEEE/ACM*音频、语音与语言处理汇刊，27(8)：1256–1266，2019。Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian 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。Muzammal Naseer、Salman Khan、Munawar Hayat、Fahad Shahbaz Khan 和 Fatih Porikli。一种用于对抗鲁棒性的自监督方法。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 262–271 页，2020。Alexander Quinn Nichol 和 Prafulla Dhariwal。改进的去噪扩散概率模型。载于国际机器学习会议，第 8162–8171 页。PMLR，2021。Weili Nie、Brandon Guo、Yujia Huang、Chaowei Xiao、Arash Vahdat 和 Anima Anandkumar。用于对抗净化的扩散模型。载于国际机器学习会议。PMLR，2022。Yao Qin、Nicholas Carlini、Garrison Cottrell、Ian Goodfellow 和 Colin Raffel。用于自动语音识别的不可感知、稳健且有针对性的对抗样本。载于国际机器学习会议，第 5231–5240 页。PMLR，2019。Krishan Rajaratnam 和 Basemah Alshemali。用于缓解和检测自动语音识别中音频对抗样本的语音编码与音频预处理。http://cs.uccs. edu/~jkalita/work/reu/REU2018/07Rajaratnam.pdf，2018。Mirco Ravanelli 和 Yoshua Bengio。利用 SincNet 从原始波形进行说话人识别。载于*2018 IEEE*语音语言技术研讨会*(SLT)*，第 1021–1028 页。IEEE，2018。Mirco Ravanelli、Titouan Parcollet 和 Yoshua Bengio。PyTorch-Kaldi语音识别工具包。载于*ICASSP 2019-2019 IEEE*声学、语音与信号处理国际会议*(ICASSP)*，第 6465–6469 页。IEEE，2019。Douglas A Reynolds、Thomas F Quatieri 和 Robert B Dunn。使用自适应高斯混合模型进行说话人验证。数字信号处理，10(1-3)：19–41，2000。Tim Salimans 和 Jonathan Ho。通过渐进式蒸馏实现扩散模型的快速采样。*arXiv*预印本 *arXiv:2202.00512*，2022。Hadi Salman、Mingjie Sun、Greg Yang、Ashish Kapoor 和 J Zico Kolter。去噪平滑：针对预训练分类器的一种可证明防御。神经信息处理系统进展，33：21945–21957，2020。Pouya Samangouei、Maya Kabkab 和 Rama Chellappa。Defense-GAN：利用生成模型保护分类器免受对抗攻击。载于国际学习表征会议，2018。Simo S¨arkk¨a 和 Arno Solin。应用随机微分方程，第 10 卷。剑桥大学出版社，2019。Ali Shafahi、Mahyar Najibi、Mohammad Amin Ghiasi、Zheng Xu、John Dickerson、Christoph Studer、Larry S Davis、Gavin Taylor 和 Tom Goldstein。免费的对抗训练！神经信息处理系统进展，32，2019。Changhao Shan、Junbo Zhang、Yujun Wang 和 Lei Xie。用于小型关键词检测的基于注意力的端到端模型。*Interspeech 2018* 论文集，第 2037–2041 页，2018。Jonathan Shen、Patrick Nguyen、Yonghui Wu、Zhifeng Chen、Mia X Chen、Ye Jia、Anjuli Kannan、Tara Sainath、Yuan Cao、Chung-Cheng Chiu 等。Lingvo：用于序列到序列建模的模块化、可扩展框架。*arXiv*预印本 *arXiv:1902.08295*，2019。

Changhao Shi、Chester Holtz 和 Gal Mishne。基于自监督学习的在线对抗净化。载于国际学习表征会议，2021。Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。载于国际学习表征会议，2015。David Snyder、Daniel Garcia-Romero、Gregory Sell、Daniel Povey 和 Sanjeev Khudanpur。用于说话人识别的稳健深度神经网络嵌入 X-vectors。载于*2018 IEEE声学、语音与信号处理国际会议(ICASSP)*，第 5329–5333 页。IEEE，2018。Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。载于国际学习表征会议，2021a。Yang Song 和 Stefano Ermon。通过估计数据分布的梯度进行生成建模。神经信息处理系统进展，32，2019。Yang Song、Conor Durkan、Iain Murray 和 Stefano Ermon。基于分数的扩散模型的最大似然训练。神经信息处理系统进展，34：1415–1428，2021b。Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。载于国际学习表征会议，2021c。Jiachen Sun、Yulong Cao、Christopher B Choy、Zhiding Yu、Anima Anandkumar、Zhuoqing Morley Mao 和 Chaowei Xiao。利用自监督实现稳健的三维点云识别。*NeurIPS*，34：15498–15512，2021。Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的有趣特性。载于国际学习表征会议，2014。Florian Tramer 和 Dan Boneh。针对多种扰动的对抗训练与鲁棒性。神经信息处理系统进展，32，2019。Florian Tramer、Nicholas Carlini、Wieland Brendel 和 Aleksander Madry。关于针对对抗样本防御的自适应攻击。神经信息处理系统进展，33：1633–1645，2020。Pete Warden。Speech Commands：一个用于有限词汇语音识别的数据集。*arXiv*预印本 *arXiv:1804.03209*，2018。Eric Wong、Leslie Rice 和 J Zico Kolter。快速胜过免费：重新审视对抗训练。载于国际学习表征会议，2020。Chaowei Xiao、Bo Li、Jun-Yan Zhu、Warren He、Mingyan Liu 和 Dawn Song。利用对抗网络生成对抗样本。载于*IJCAI*，2018a。Chaowei Xiao、Jun-Yan Zhu、Bo Li、Warren He、Mingyan Liu 和 Dawn Song。空间变换对抗样本。载于国际学习表征会议，2018b。Chaowei Xiao、Dawei Yang、Bo Li、Jia Deng 和 Mingyan Liu。MeshAdv：用于视觉识别的对抗网格。载于*IEEE*计算机视觉与模式识别会议论文集，第 6898–6907 页，2019。Chaowei Xiao、Zhongzhu Chen、Kun Jin、Jiongxiao Wang、Weili Nie、Mingyan Liu、Anima Anandkumar、Bo Li 和 Dawn Song。DensePure：理解扩散模型以实现对抗鲁棒性。*arXiv*预印本 *arXiv:2211.00322*，2022a。Chaowei Xiao、Xinlei Pan、Warren He、Jian Peng、Mingjie Sun、Jinfeng Yi、Mingyan Liu、Bo Li 和 Dawn Song。深度强化学习攻击的特征分析。*AAMAS*，2022b。Saining Xie、Ross Girshick、Piotr Doll´ar、Zhuowen Tu 和 Kaiming He。用于深度神经网络的聚合残差变换。载于*IEEE*计算机视觉与模式识别会议论文集，第 1492–1500 页，2017。

Zhuolin Yang、Bo Li、Pin-Yu Chen 和 Dawn Song。使用时间依赖性刻画音频对抗样本。载于国际学习表征会议，2019 年。

Jongmin Yoon、Sung Ju Hwang 和 Juho Lee。使用基于分数的生成模型进行对抗净化。载于国际机器学习会议，第 12062–12072 页。PMLR，2021 年。

Sergey Zagoruyko 和 Nikos Komodakis。宽残差网络。载于英国机器视觉会议 *2016*。英国机器视觉协会，2016 年。

Dinghuai Zhang、Tianyuan Zhang、Yiping Lu、Zhanxing Zhu 和 Bin Dong。只传播一次：通过极大值原理加速对抗训练。神经信息处理系统进展，32，2019a。

Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。鲁棒性与准确率之间具有理论依据的权衡。载于国际机器学习会议，第 7472–7482 页。PMLR，2019b。

Huan Zhang、Hongge Chen、Chaowei Xiao、Bo Li、Duane Boning 和 Cho-Jui Hsieh。迈向可验证鲁棒神经网络的稳定高效训练。*ICLR*，2019c。

附录

A 改进型去噪扩散概率模型的训练细节。

我们使用官方代码库（https://github.com/openai/improved- diffusion）训练改进型 DDPM。对于 UNet 模型，我们设置 $image\ size\ =-\ 32$、 $num\ channels\ =\ 3$ 和 $num\ \bar{}res\ blocks\ =\ 128$。对于扩散参数，我们设置 $N\ =\ 200$、 $\beta_1\ =\ 0.0001$、 $\beta_N\ =\ 0.02$，并使用线性方差调度。对于模型训练，我们将学习率设为 $1e-4$，批量大小设为 230。经过 80,000 个训练步骤后，训练损失已收敛，我们使用此检查点来构建净化器。

B 基于迁移的攻击的补充实验

我们还在基于迁移的攻击下评估了我们的方法。在这种攻击中，我们假设攻击者只能获取声学系统的输出逻辑值，而不了解所采用的防御方法。

我们使用模型功能窃取来训练替代模型。具体而言，我们首先将输入样本送入由 DiffWave 和 ResNeXt分类器组成的声学系统，并获取输出逻辑值。随后，我们将声学系统的这些输出逻辑值用作标签，训练一个新的替代 ResNeXt 模型，其架构与声学系统中的分类器相同。结果见表 A。窃取准确率 表示替代分类器以受防御的声学系统的预测作为真实标签时的准确率。迁移至原始模型 和迁移至受防御模型 分别表示无防御原始分类器和受防御的声学系统。我们对替代分类器发起攻击以生成对抗样本，并将这些对抗样本迁移过去，以评估无防御原始分类器和受防御声学系统的鲁棒性。

表 A：通过模型功能窃取实施基于迁移的攻击。我们使用防御声学系统的输出作为标签来训练替代模型。然后，通过攻击替代模型生成对抗样本，并将其迁移至无防御原始分类器和防御声学系统。

窃取目标	窃取准确率	迁移至原始模型		迁移至防御系统	
		干净样本	稳健	干净样本	稳健
AudioPure($n^\star=1$)	100	100	22	100	99
AudioPure($n^\star=5$)	98	100	58	96	94

C 认证鲁棒性详情

随机平滑（Cohen 等，2019）通过在噪声下评估模型，为 $\mathcal{L}_2$-范数提供可证明的鲁棒性保证。通常，向原始分类器输入高斯扰动输入时，其性能会下降。为缓解这一问题，我们可以在高斯增强数据上重新训练一个新网络，或对预训练网络进行微调。不过，这两种方法都可能耗费大量训练时间。另一种方法是在原始分类器之前应用去噪器，称为去噪平滑（Salman 等，2020）。由于扩散模型的反向过程可视为一种优秀的去噪器，我们可以将预训练扩散模型作为即插即用方法，使任何模型都具备可认证的鲁棒性。对于给定的噪声水平 σ，我们可以计算出相应的扩散步数 $t^\star$，以便向输入样本添加相同水平的噪声。扩散过程可以重新表述为：

$$\mathbf{x}_n=\sqrt{\bar{\alpha}_n}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\mathbf{z}=\sqrt{\bar{\alpha}_n}(\mathbf{x}_0+\sqrt{\frac{1-\bar{\alpha}_n}{\bar{\alpha}_n}}\mathbf{z}),\qquad \mathbf{z}\sim\mathcal{N}(0,\mathbf{I}),\tag{S17}$$

而随机平滑的噪声输入 $\hat{}\mathbf{x}$ 是

$$\hat{\mathbf{x}}=\mathbf{x}_0+\sqrt{\sigma}\mathbf{z},\qquad \mathbf{z}\sim\mathcal{N}(0,\mathbf{I}).\tag{S18}$$

因此，在输入 $\hat{\mathbf{x}}$上乘以缩放系数 $\sqrt{\bar{\alpha}_n}$ 后，我们可以得到 $n^\star$ 使得 $\frac{1-\bar{\alpha}_n}{\bar{\alpha}_n} = \sigma$。

根据 Carlini 等人（2022）的研究，单个反向步骤就能恢复图像，并使分类器获得较高的准确率；通过直接经由 $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_n}}(\mathbf{x}_n - \sqrt{1-\bar{\alpha}_n}\epsilon_\theta(\sqrt{\bar{\alpha}_n}\hat{\mathbf{x}}\ n))$ 恢复数据，还能大幅节省计算时间。因此，在我们的反向过程中，我们可以直接采用单步去噪，而不必执行完整的步骤。

图 A 展示了 AudioPure 与 RS-Gaussian 和 RS-Vanilla 相比的认证准确率。结果表明，除 $\sigma = 0.50$时较小的认证半径外，我们的方法的认证鲁棒性始终优于基线方法。

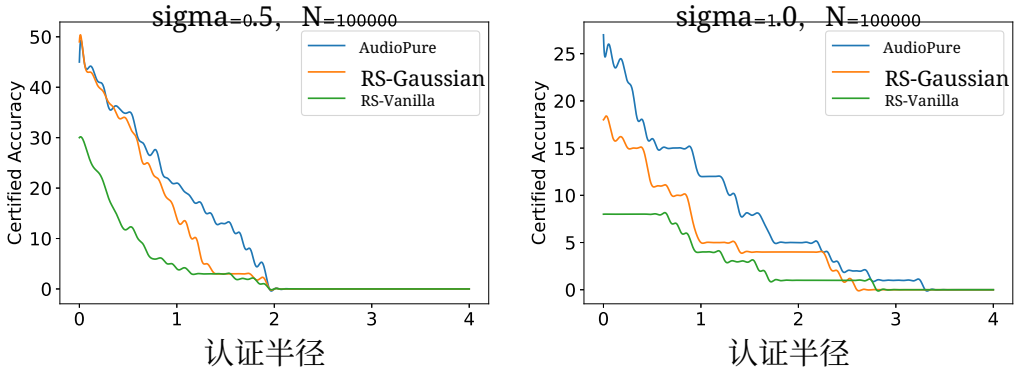


图 A：不同输入噪声水平 σ下的认证鲁棒性（$\mathcal{L}_2$）。更大的 σ 能够确保在更大扰动下具有更好的鲁棒性，但会降低对良性输入的性能。

D 净化能力的理论分析

定理 D. 1。假设 $p(x)$ 和 $q(x)$ 分别是干净样本的数据分布和对抗样本的数据分布。我们使用 p_t 和 q_t 分别表示当 $x(t) \sim p(x)$ 时 $x(t)$ 的分布，以及当 $x(t) \sim q(x)$ 时 $x(t)$ 的分布。则有

$$\frac{\partial D_{KL}(p_t||q_t)}{\partial t} \leq 0 \tag{S19}$$

其中，等号仅在 $p_t = q_t$时成立。该不等式表明，当 t 从 0 增加到 1 时，p_t 与 q_t 的 KL 散度单调减小。换言之，随着扩散步数 $n^\star$ 增加，更多的对抗扰动将被移除。考虑到当 $n^\star$ 过大时，原始语义信息也会被移除，从而影响干净准确率，因此，为扩散模型净化器设置 $n^\star$ 时应进行权衡。

证明：遵循 Nie 等人（2022）；Song 等人（2021b），我们首先将式 7 中前向SDE 的福克-普朗克方程（S¨arkk¨a & Solin，2019）表示为（其中定义 $f(x,t) := -\frac{1}{2}\beta(t)$ 和 $g(t) := \sqrt{\beta(t)}$ ：

$$\begin{aligned}\frac{\partial p_t(x)}{\partial t} &= -\nabla_x\left(f(x,t)p_t(x) - \frac{1}{2}g^2(t)\nabla_x p_t(x)\right)\\ &= -\nabla_x\left(f(x,t)p_t(x) - \frac{1}{2}g^2(t)p_t(x)\nabla_x \log p_t(x)\right)\\ &= \nabla_x \cdot (h_p(x,t)p_t(x))\end{aligned} \tag{S20}$$

其中 $h_p(x,t) := \frac{1}{2}g^2(t)\nabla_x \log\ p_t(x) - f(x,t)$。假设 p_t和 q_t均为光滑且快速衰减的函数，即 对于任意 $i = 1, \ldots, d$，我们有

$$\lim_{x_i\to\infty} p_t(x)\frac{\partial}{\partial x_i}\log p_t(x) = 0, \qquad \lim_{x_i\to\infty} q_t(x)\frac{\partial}{\partial x_i}\log q_t(x) = 0 \tag{S21}$$

对于 x_i， $x \in \mathbb{R}^d$ 的第 i 维。然后我们将 KL 散度重新表述为

$$\begin{aligned}
\frac{\partial D_{KL}(p_t||q_t)}{\partial t} &= -\frac{\partial}{\partial t}\int p_t(x)\log\frac{p_t(x)}{q_t(x)}\mathrm{d}x \\
&= -\nabla_x\left(f(x,t)p_t(x) - \frac{1}{2}g^2(t)p_t(x)\nabla_x\log p_t(x)\right) \\
&= \int \nabla_x\cdot(h_p(x,t)p_t(x))\log\frac{p_t(x)}{q_t(x)}\mathrm{d}x + \int\frac{p_t(x)}{q_t(x)}\nabla_x\cdot(h_p(x,t)p_t(x))\mathrm{d}x \\
&= -\int p_t(x)[h_p(x,t) - h_q(x,t)]^\top[\nabla_x\log p_t(x) - \nabla_x\log q_t(x)]\mathrm{d}x \\
&= -\frac{1}{2}g^2(t)\int p_t(x)\|\nabla_x\log p_t(x) - \nabla_x\log q_t(x)\|_2^2\mathrm{d}x \\
&= -\frac{1}{2}g^2(t)D_F(p_t||q_t)
\end{aligned} \tag{S22}$$

其中 $D_F(p_t||q_t)$ 是 Fisher 散度。考虑到 $g^2(t) = \beta(t) > 0$，且 Fisher 散度 $D_F(p_t||q_t) \geq 0$，并且只有在 $p_t = q_t$ 时等式才成立，因此我们得到公式 S19，其中等式仅在 $p_t = q_t$ 时成立。

E Qualcomm 关键词语音数据集上的实验

除了常用的 SC09，为了更全面地进行评估，我们还在 Qualcomm 关键词语音数据集（Kim 等人，2019）上进行了实验，以下称为 QKW。QKW 包含 4 个类别的 4270 条语音，时长不等，范围为 0.48 秒至 1.92 秒。我们将其划分为训练集（3770 条语音）、验证集（400 条语音）和测试集（100 条语音）。为处理尺寸可变的输入，我们训练了一个注意力循环卷积网络（Shan 等人，2018），并保存了在验证集上准确率最高的检查点。随后，我们在 QKW 上对 DiffWave 模型进行了 50,000 步微调，使用 $lr = 2e-4$和 $batch\ size\ per\ gpu = 2$，共 3 个 GPU。在 $\mathcal{L}_\infty$ 投影梯度下降$_{10}$ 下使用 $\epsilon = 0.002$-得到的结果如表 B 所示。我们可以看到，AudioPure 仍能实现不容忽视的鲁棒性，并能很好地处理时长可变的音频。

表 B：我们将 AudioPure 应用于 Qualcomm 关键词语音数据集。扩散步数 $n^\star$ 设为 2。

防御	干净样本	稳健
None	100	0
AudioPure	91	61

F 对抗样本上的微调

AudioPure 利用预训练扩散模型。我们想知道，若在对抗样本上进行微调，净化性能是否会提升。此外，我们通过加入自监督扰动（SSP）（Naseer 等，2020），进一步微调 DiffWave 模型。具体而言，我们使用短时傅里叶变换（重缩放至梅尔刻度）作为特征提取器，并最大化以下目标以生成扰动样本：

$$\arg\max_{x'}\Delta(x,x') = \|STFT(x), STFT(x')\|_\infty, \quad s.t.\|x - x'\|_\infty \tag{S23}$$

其中 x 是干净样本， x' 是扰动样本。然后，我们使用梯度下降对扰动样本进行优化：

$$x'_{t+1} = clip(x'_t + \alpha\cdot sign(\nabla_x\Delta(x,x'_t)), x-\epsilon, x+\epsilon), \tag{S24}$$

对于 $t = 1,\ldots,T$。这里我们使用 $T = 100$、 $\epsilon = 0.002$ 和 $\alpha = 0.0004$。接下来，我们在 SSP 样本上微调预训练 DiffWave 模型，使以下损失最小化：

$$\mathcal{L}_{tuning} = \mathcal{L}_{audio} + \lambda\mathcal{L}_{feat}, \tag{S25}$$

其中

$$\mathcal{L}_{audio} = MSE(x, \textbf{Purifier}(x'_t, n^\star)), \tag{S26}$$

$$\mathcal{L}_{feat} = MSE(STFT(x), STFT(\textbf{Purifier}(x'_t, n^\star)). \tag{S27}$$

我们选择 $\lambda = 0.1$，并使用随机梯度下降优化 $\mathcal{L}_{tuning}$，将学习率设为 $1e-5$。结果如表 C 所示。因此，在 $\mathcal{L}_\infty$ 投影梯度下降$_{10}$ 和使用 $\epsilon = 0.002$ 的投影梯度下降$_{70}$ 下，这并未提升AudioPure （使用 $n^\star = 3$）的性能。这些结果进一步验证了使用预训练模型的有效性。

表 C：我们使用 SSP 生成的对抗样本对预训练 DiffWave 模型进行微调。微调后，性能并未提升。

防御	干净样本	PGD$_{10}$	PGD$_{70}$
None	100	3	1
AudioPure	97	89	84
经 SSP 调优的 AudioPure	97	89	82

G 与其他基于去噪器的防御方法比较

我们将 AudioPure 与 DefenseGAN（Samangouei 等人，2018）和联合对抗微调（Joshi 等人，2022）进行比较。DefenseGAN 最初旨在防御对抗图像，其方法是寻找能生成与对应对抗图像最相似图像的最优噪声；我们将其应用于音频领域，并在此处理流程中选用 WaveGAN（Donahue 等人，2018）作为生成对抗网络模型。我们使用 Adam 优化器在 SC09 数据集上训练 WaveGAN 100 轮，设置为 $lr = 1e-3$、$\beta_1 = 0.5$ 和 $\beta_2 = 0.9$。对于联合对抗微调，我们遵循 Joshi 等人（2022）的设置，使用 Conv-TasNet（Luo 和 Mesgarani，2019）作为去噪器。与 Joshi 等人（2022）相同，我们使用 L-inf 投影梯度下降-100 攻击（参数为 $\epsilon = 0.002$ ）对预训练分类器生成离线对抗 SC09 数据集（记为 OffAdv-SC09）。随后，我们在 OffAdv-SC09 上训练 Conv-TasNet 模型 30 轮，得到预训练去噪器。我们将使用预训练 Conv-TasNet 的防御方法称为 CTN 基线。基于攻击整个声学系统生成的对抗样本，我们仅更新 Conv-TasNet 去噪器，同时冻结分类器，并将此方法记为 CTN Adv-Finetune-Joint-frozen。在对抗调优过程中，我们使用 $\mathcal{L}_\infty$ PGD$_{10}$ 攻击，参数为 $\epsilon = 0.002$。使用 $batch\ size = 20$ 调优 1000 步后，我们在本文所用的同一测试集上计算干净样本准确率和鲁棒准确率（分别针对 $\mathcal{L}_\infty$ PGD$_{10}$ 和 PGD$_{70}$，参数为 $\epsilon = 0.002$）。

我们在表 D 中报告了结果。我们发现，基于 WaveGAN 的 DefenseGAN 在音频领域表现不佳。这表明领域差异会影响最终结果，也验证了我们的流水线设计的重要性。此外，即使经过微调，Conv-TasNet 去噪器抵御自适应攻击的效果仍不如扩散模型。

表 D：我们将 AudioPure 与不同的基于去噪器的防御进行了比较。事实证明，DiffWave 是一种更有效的净化器。

防御	干净样本	PGD$_{10}$	PGD$_{70}$
None	**100**	3	1
AudioPure	97	**89**	**84**
DefenseGAN	8	0	0
CTN 基线	98	13	1
CTN Adv-Finetune-Joint-frozen	90	52	41

H 与基于正则化的防御进行比较

Gu & Rigazio (2014); Hoffman 等人 (2019) 将网络的输入输出雅可比矩阵作为正则化项引入优化目标，其形式如下:

$$\mathcal{L}_{reg} = \sum_i \left(\mathcal{L}(x_i, y_i) + \lambda \| \frac{\partial f(x_i)}{\partial x_i} \|_F \right), \tag{S28}$$

其中， $x_i \in \mathbb{R}^d$ 是输入数据， $y_i \in \mathbb{R}^n$ 是标签， $\mathcal{L}: \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ 是原始损失函数， $f: \mathbb{R}^d \to \mathbb{R}^n$ 是神经网络。通过最小化雅可比矩阵的弗罗贝尼乌斯范数，可以提升网络的对抗鲁棒性。为了进行更全面的研究，我们还使用不同的 λ将AudioPure 与这种基于正则化的方法进行比较。结果如表 E 所示，其中我们将基于正则化的防御称为雅可比正则化。

表 E：我们使用不同的 λ将AudioPure 与基于正则化的防御进行比较。

防御	干净样本	PGD_{10}	PGD_{70}
None	**100**	3	1
AudioPure	97	**89**	**84**
雅可比正则化 (λ=1e-8)	45	9	5
雅可比正则化 (λ=1e-9)	84	27	15
雅可比正则化 (λ=1e-10)	91	31	18
雅可比正则化 (λ=1e-11)	96	19	4

I 更大攻击预算下的实验。

除了表2中不同 ϵ 的结果外，我们还进行了额外实验，以探索扩散模型用于净化的潜力。我们选择 $\epsilon = \{0.01, 0.02, 0.03, 0.04, 0.05\}$，并将扩散步数设为 $n^\star$。结果如表F所示。我们发现，在 $\epsilon = 0.03$时，我们的方法仍能达到42%的准确率，而这会给音频带来严重失真。直到 $\epsilon = 0.05$，我们的方法仍能净化对抗扰动。我们还将不同 ϵ攻击下的音频波形可视化，如图 B 所示。可以明显观察到其中存在显著噪声。

表F：我们探索了 DiffWave 在更大攻击预算下的潜力。扩散步数 $n^\star$ 设为 5。

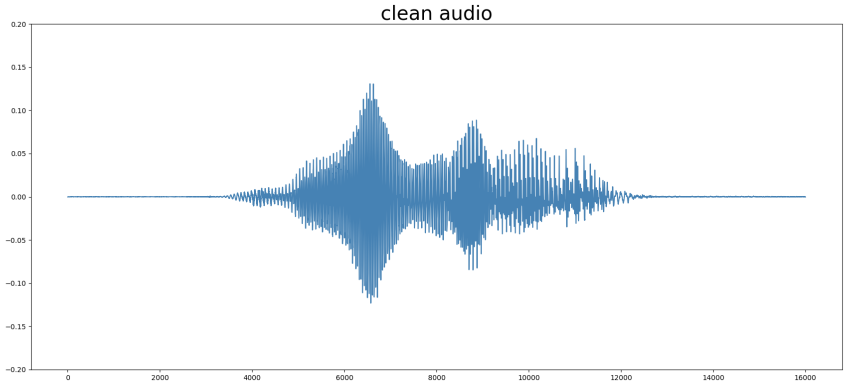
攻击预算	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.03$	$\epsilon = 0.04$	$\epsilon = 0.05$
鲁棒准确率	82	67	42	14	0

J 附加推理时间开销。

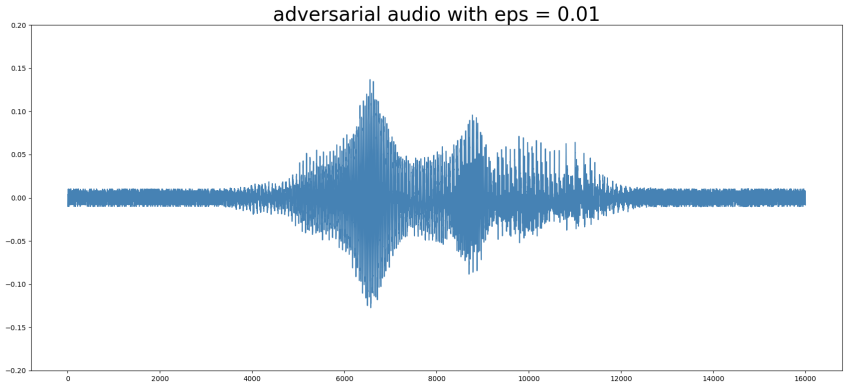
由于引入了扩散模型，AudioPure 在推理过程中会产生额外的时间开销。如表G所示，我们计算了每段音频的时间开销，取100个样本的平均值，每个样本的时长约为一秒。我们在配备 Intel® CoreTM i9-10920X CPU @ 3.50GHz 和 64 GB 随机存取存储器的 NVIDIA RTX 3090 GPU 上进行了评估。

表G：使用不同扩散步数时的推理时间开销 $n^\star$。

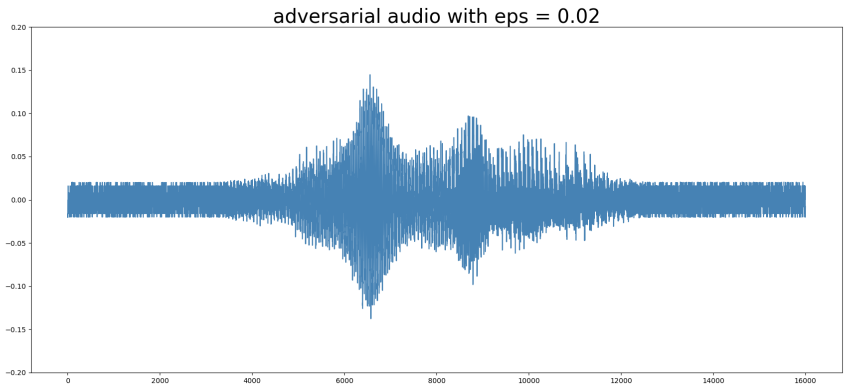
扩散步数	$n^\star = 0$	$n^\star = 1$	$n^\star = 2$	$n^\star = 3$	$n^\star = 5$	$n^\star = 7$	$n^\star = 10$
时间开销（秒）	0.0967	0.5522	0.7876	1.0162	1.4795	2.0125	2.6839



(a)

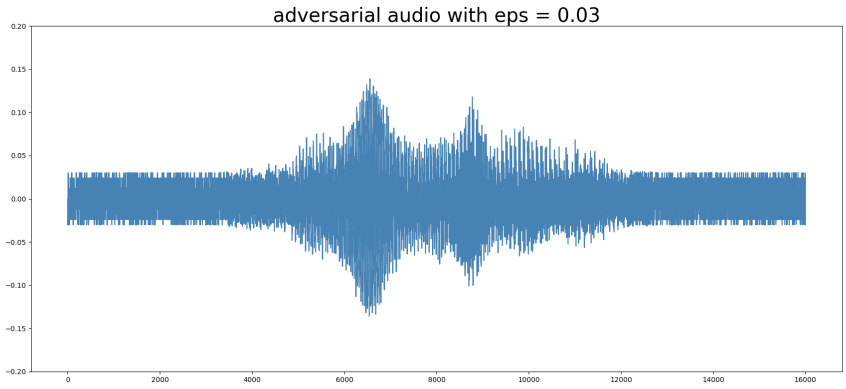


(b)

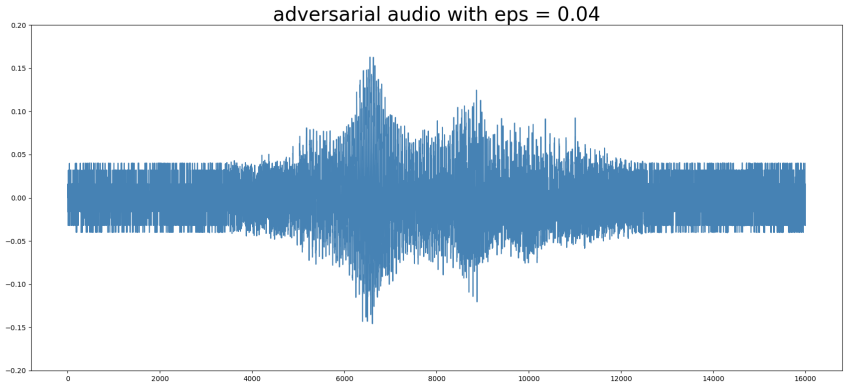


(c)

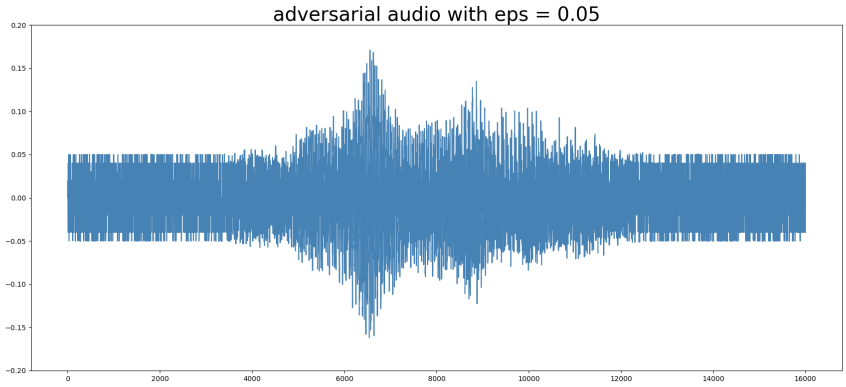
图 B：不同攻击预算下干净音频和对抗音频的可视化结果 s.



(d)



(e)



(f)

图 B：不同攻击预算下干净音频和对抗音频的可视化结果 s.