VocalBridge：用于击败基于扰动的声纹防御的潜空间扩散桥净化

Maryam Abbasihafshejani 计算机科学系 德克萨斯大学圣安东尼奥分校 圣安东尼奥，得克萨斯州

AHM Nazmus Sakib 计算机科学系 德克萨斯大学圣安东尼奥分校 圣安东尼奥，得克萨斯州

Murtuza Jadliwala 计算机科学系 德克萨斯大学圣安东尼奥分校 圣安东尼奥，得克萨斯州

arXiv:2601.02444v1 [cs.SD] 5 Jan 2026

摘要——**文本转语音（TTS）和语音转换（VC）等语音合成技术的快速发展，加剧了围绕语音克隆的安全与隐私担忧。近期的防御措施尝试通过向语音中嵌入保护性扰动，来防止未经授权的克隆，旨在模糊说话人身份，同时保持可懂度。然而，攻击者可以采用最先进的净化策略去除这些扰动，恢复真实的声学特征，并重新生成可供克隆的语音。尽管此类威胁日益成为现实，但在自适应净化情形下，这些保护机制的鲁棒性尚未得到充分研究。**

现有的净化与去噪方法大多旨在处理自动语音识别（ASR）或分类任务中的对抗噪声，却无法保留定义说话人声音的细粒度声学特征。因此，这些方法往往会降低感知质量，并在说话人嵌入空间中引入失真。近期基于扩散的净化框架能够更有效地去除针对自动说话人验证（ASV）系统的扰动，但其中许多方法直接处理波形，或在推理过程中需要依赖转录文本进行音素对齐。尽管这些方法优于面向自动语音识别的去噪方法，但仍难以彻底去除对抗扰动，从而限制了恢复保真度并降低整体性能。为解决这些局限，我们提出扩散桥（*VocalBridge*），一种在 EnCodec 潜空间中学习从受扰语音到干净语音的潜在映射的净化模型。它采用时间条件化一维 U-Net 去噪器，在余弦噪声调度下执行反向扩散，从而能够高效地净化语音，无需转录文本，同时保留说话人判别线索。我们还提出一种 Whisper 引导的音素变体，通过基于 Whisper 的音素对齐模块融入轻量级语言条件信息。不同于依赖转录文本或外部语言提示的既有文本条件扩散模型，我们的方法完全在声学域中运行，无需语言监督。

这些发现揭示了当前基于扰动的防御措施十分脆弱，也凸显了迫切需要采取更具韧性的防护措施，以应对不断演变的语音克隆威胁。

索引词—**净化器、格式、扩散桥、语音克隆、基于扰动的防御措施**

I. 引言

生成式人工智能的最新进展加剧了人们对安全、隐私和数据所有权的担忧。其中一个尤为令人担忧的领域是语音生成，包括文本转语音（TTS）和语音转换（VC）模型 [1]。这些技术如今能够生成高度逼真的音频深度伪造内容，从而助长身份冒充、虚假信息传播和身份盗窃 [2]——[4]例如，诈骗者最近克隆了一位公众人物的声音，诱骗其家人相信该人物遇到了法律麻烦 [5]。尤其值得关注的是，人工智能生成的假声音能够绕过最先进的自动说话人验证系统，而此类系统仍面临泛化能力和鲁棒性方面的挑战 [6]。除了针对个人的诈骗案件，合成语音已被用于大规模政治虚假信息宣传活动、欺诈性金融交易，以及针对企业和政府机构的社会工程攻击。2025年2月，意大利当局揭露了一起人工智能语音诈骗案：犯罪分子冒充意大利国防部长，利用语音克隆技术，诱骗一位知名企业家向海外账户汇入近100万欧元（≈ <style id='19">104万美元） [7]。为应对日益加剧的威胁，Ope- [8] nAI和美国联邦贸易委员会（FTC） [9]均已发布报告，警示语音转换和合成语音技术对安全、隐私及社会造成的影响，并强调亟须开发可靠的方法，检测并遏制现代语音生成系统的恶意用途。

为降低这些风险，研究人员开始探索主动语音保护方法，使用户的语音无法被合成模型学习。其核心思路是引入精心设计的扰动或定向变换，阻止文本转语音/语音转换系统提取说话人特有的特征（音色），同时确保音频自然且清晰易懂，以便用于通信或自动说话人验证等正当用途。这些扰动会降低合成模型准确模仿说话人身份的能力，确保根据受保护语音生成的任何伪造音频质量较低，或无法通过自动说话人验证系统[10], [11]。

然而，在攻击者能够先净化受保护语音再加以使用的场景中，现有主动防御的性能仍未得到充分研究。以往的净化研究主要关注去除针对

自动语音识别（ASR）系统的对抗噪声，因此，当前方法能否去除旨在防范语音克隆和说话人验证攻击的保护性扰动，仍是一个悬而未决的问题。为证明现有净化方法的不足，以及对更稳健语音保护技术的需求，我们提出*VocalBridge*，这是一种全面的净化攻击，能够有效去除保护性扰动，使攻击者得以绕过现代说话人验证系统。

VocalBridge 是一种扩散桥净化模型，可在紧凑的编码域中学习从受保护语音表示到干净语音表示的潜在映射。该模型采用时间条件化一维 U-Net 去噪器，在 E nCodec 潜空间中建模反向扩散过程，并通过余弦噪声调度训练，以估计残余扰动。这种设计能够高效地净化语音，无需转录文本，同时保留语音克隆和验证攻击所需的说话人特征。结果表明，VocalBridge 在攻击效果和性能方面均显著优于现有净化方法

我们的贡献总结如下：

1) **对主动防御进行全面评估**。我们对五种最先进的主动防御、六种现代语音合成工具和三种广泛使用的自动说话人验证编码器进行了大规模评估，全面呈现了当前保护机制在不同合成流程中的表现。我们的研究纳入了近期的合成模型、多种防御策略和净化工具，并使用由 VCTK 和 LibriSpeech 构建的大型数据集，从而全面评估了当代语音克隆场景中的主动保护。

2)**VocalBridge：语音潜在空间中的扩散桥净化**。我们提出了一种在语音潜在空间中运行的扩散桥净化模型，它能够学习将扩散后的防御性扰动表征反向映射回对应的干净潜在表征。我们还采用基于 Whisper 的轻量级条件机制来进一步提升性能，无需转录文本即可提供音素级引导。在验证恢复率（VR）、平均意见分（MOS）和词错误率（WER）等评估指标上，与近期净化方法相比，VocalBridge 持续提升了认证恢复效果、感知质量和可懂度，同时展现出良好的泛化能力，并能稳健应对自适应保护策略。

3) **用于社区评估的数据集创建**。我们生成了一个大型数据集，其中包含适用于评估主动防御措施的干净、受保护和净化后的语音样本，为未来研究提供了有价值的基准。

II. 背景与相关工作

A. 深度伪造语音合成技术

深度伪造语音合成涉及使用人工智能和机器学习技术，生成能够复现目标说话人声音特征的音频片段，包括音色、韵律和发音模式。深度伪造语音合成技术大致可分为：（i）文本转语音和（ii）语音转换工具。

文本转语音（TTS）工具可将书面文本转换为口语。一些最早期的技术（如拼接合成）能够生成清晰的语音，但其结果往往听起来不自然且机械 [12], [13]。与之相反，现代深度学习方法能够生成极其自然、近似人声的语音输出 [14], [15]。这类系统通常采用两阶段流程。首先，声学模型将输入文本转换为低级声学表征，例如梅尔频谱图，其中编码了语音信息、音高和时序信息。随后，神经声码器根据该表征合成高保真音频波形。在这两个阶段使用深度神经网络，使这些模型能够捕捉人类语音的复杂细微之处。这项技术的一项关键应用是语音克隆 [16], ，其目标是生成一种与特定目标说话人的声音难以区分的合成语音。实现这一目标的方法是，使用该个体的一组录音对 TTS 模型进行训练或微调。通过从这些样本中学习，模型能够捕捉目标说话人独特的声音特征，包括其音色、语调模式和口音 [17]–[19]。这使训练后的模型能够有效地生成克隆语音。

与文本转语音不同，语音转换（VC）会改变源说话人的发声属性，使其与某个目标说话人的发声属性相匹配，同时保留原始语言内容。其实现方式是将依赖说话人的声学特征与语音信息分离，再重新合成音频。

B. 语音隐私与保护机制 *ms*

深度伪造语音合成会严重损害用户隐私并带来安全风险，因此文献中已探讨了各种防御策略。尽管将说话人的声音匿名化是保护隐私的有效策略 [20], [21], ，但仅凭这一策略不足以构成防御；防范深度伪造音频还需要准确的检测与预防。检测主要侧重于识别合成音频（或克隆音频），所用技术包括利用梅尔频率倒谱系数（MFCC）特征进行分类 [22], 、情感识别 [23]、通过模拟人耳听觉效应进行声学信号分析 [24] 以及自动说话人验证系统 [25]。预防性防御技术则更具事前防范性质，旨在使音频合成工具的语音合成能力失效。为阻止伪造或合成语音的生成，已有若干技术被提出，它们利用对抗扰动来保护原始录音 [26]–[28]。通过在

数据层面引入细微扰动，这些技术可防止生成模型在推理过程中成功克隆声音，也就是说，最终生成的克隆音频与预期目标音频并不相似。Zhang 等人 [29] 提出了一种名为*POP* 的保护技术，该技术将不可感知的误差最小化噪声施加到原始语音样本上，防止文本转语音合成模型有效学习这些样本。在后续工作中，他们提出了*SPEC* [11], ，并声称该方法能够抵御先进的自适应对手。Dong 等人 [30] 提出了一种基于生成对抗网络（*GAN*）的扰动生成技术，以防范音频深度伪造生成。最后，诸如*AntiFake* [10] 等技术利用对抗样本扰乱合成音频中的人声音色，从而使输出与原说话人不相似，阻止零样本语音克隆和语音转换。

C. 对抗净化

由于保护性扰动的作用方式类似于添加到语音转换模型输入中的对抗噪声，一个重要问题是，现有的对抗净化技术能否去除这些扰动，从而绕过保护。对抗净化方法直接作用于输入音频，旨在通过消除音频到达下游语音转换系统之前的对抗性失真，恢复干净信号。

不同于会修改或重新训练模型本身的对抗训练，净化不要求访问或控制目标语音转换模型。这使净化成为我们所处场景中的一种颇具吸引力的策略，因为保护机制必须在不改变外部语音转换模型的情况下对其进行防御。此外，净化方法能够泛化到未见过的或自适应的扰动，因为它们侧重于恢复输入信号，而不是依赖训练期间遇到的一组固定对抗样本。

近期研究表明，基于生成模型的去噪方法，尤其是扩散模型，在去除音频中的扰动方面非常有效 [31], [32]。例如，Wu 等人 [33], 提出的技术先向对抗音频添加少量噪声，然后使用基于去噪扩散概率模型 [34]的扩散模型对其进行去噪，从而恢复干净音频。*Diffpure* [35]则将即插即用扩散模块用作预处理步骤，以去除扰动。Guo 等人 [36] 在此基础上，针对自动语音识别（ASR）构建了更稳健的解决方案，因为自动语音识别模型通常关注低频成分，而忽略高频成分。他们通过采用预训练引导扩散模型的分层扩散框架实现了这一目标。Tan 等人 [37] 采用单次无条件梅尔频谱扩散模型，并在时域和频域中去除扰动。首先，通过插值方法向时域信号中添加高斯扰动，然后使用扩散模型净化样本的梅尔频谱。以往的工作主要针对旨在破坏自动语音识别模型的对抗扰动，而 Fan 等人 [38] 提出了一种两步方法，用于去除保护性扰动，其目标是使语音无法被生成模型学习。第一步，使用无条件预训练扩散模型净化扰动音频。第二步，采用基于Ornstein–Uhlenbeck随机微分方程的细化模型 [39]，并以音素信息为引导，使语音音频更贴近原始信号。

在图像数据背景下，Li 等人 [40]研究了基于简单预训练扩散模型的净化方法有限的鲁棒性。具体而言，他们在基于梯度的PGD攻击 [41] 下评估了 DiffPure [35]，并证明干净准确率和对抗鲁棒性均显著下降。尽管后续已有若干工作 [42]–[46]试图提升 DiffPure 的鲁棒性，但由于这些方法所需的蒙特卡洛采样具有很高的计算成本，其实际适用性受到限制 [47]。为解决这些局限，Li 等人提出了 ADBM，该方法从扩散后的对抗图像数据分布构建一条通往干净数据分布的反向桥。他们的方法假设，在训练期间，对抗噪声 ϵ_a会在前向扩散过程开始时注入。模型经过训练后能够预测原始噪声 ϵ 以及经缩放的对抗噪声版本 ϵ_a，从而使反向扩散过程能够在对输入进行去噪的同时，有效去除对抗扰动。

III. 威胁模型与安全目标

我们考虑图 1 所示的威胁模型，其中 用户 希望防止其公开可用的语音录音遭到滥用，而攻击者 则试图生成深度伪造音频来冒充用户并欺骗自动说话人验证系统。

参与方与目标： 该系统由三个组件构成：用户、对手和自动说话人验证系统。

- **用户**。 用户会出于通信或身份验证的目的正常生成话语，并可能将其公开分享（例如在社交媒体或在线平台上）。为防止话语遭到滥用，用户会在发布每段干净话语 x 之前，使用由参数 ψ配置的防御机制P_ψ进行处理，从而生成受保护版本：

$$x_p = P_\psi(x).$$

  其目标是 P_ψ 防止恶意语音合成，同时保留话语的自然度和可验证性，以供自动说话人验证系统合法使用。

- **攻击者**。攻击者会收集公开可用的受保护话语 $\{x_p\}$，并尝试克隆或合成模仿用户音色的话语。对手的目标是生成克隆话语 $\tilde{x}$ ，使其被自动说话人验证系统判定为真实语音。
- **验证者（自动说话人验证系统）**。自动说话人验证系统会评估输入话语与合法用户的注册话语之间的相似度分数 $s(\cdot,\ \cdot)$。如果 $s(\cdot,\ \cdot) > \tau$，则验证成功；其中 τ 是系统的决策阈值。

用户目标： 防御机制 P_ψ 应确保发布的受保护话语 $\{x_p\}$ 保持可用性

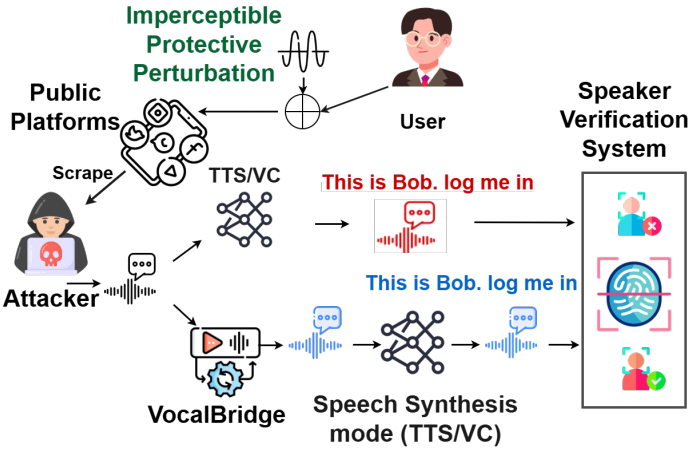

图 1：该图展示了威胁模型，说明攻击者如何利用 VocalBridge 绕过现有防御措施并实施语音克隆攻击。

同时抵御身份冒充。具体而言，它应满足以下属性：

- **不可感知性：** P_ψ 引入的扰动必须对人类听众不可感知，使 x_p 保持自然、清晰易懂且高质量，同时保留 x 的语言内容和意图。
- **可验证性：**经过保护后，合法用户仍应能通过自动说话人验证系统的验证：

$$s(x_p, x_e) \geq \tau,$$

其中，$s(\cdot,\ \cdot)$表示说话人嵌入向量之间的余弦相似度，τ 为自动说话人验证决策阈值。

- **不可仿冒性：**由受保护样本 x_p 生成的合成（或深度伪造）话语 $\tilde{x}$ 不应被自动说话人验证系统判定为真实话语：

$$s(\tilde{x}, x_e) < \tau.$$

攻击者目标：在我们的设定中，对手的目标是通过净化和合成使用户的保护目标失效。具体而言，攻击者旨在：

- **恢复可验证性：**净化旨在恢复自动说话人验证：将最初被拒绝的受保护话语（$s(x_p,\ x_e) < \tau$）转换为达到接受阈值的净化后的话语（$s(x_{\text{pur}}\ x_e) \geq \tau$）。
- **保持不可感知性：**确保净化后的音频对人类听众而言仍然自然且清晰易懂。
- **打破不可仿制性：**使由净化后的话语生成的合成语音或克隆语音能够被自动说话人验证系统判定为真实语音，从而表明用户的不可仿制防御不再有效。

综上，这些目标量化了攻击在逆转用户保护效果（恢复身份验证功能）的同时，保持感知质量并促进重新克隆的能力。

攻击者能力： 对手在现实且有限的能力范围内进行操作：

- 对手可以收集公开可用的受保护话语的短片段 $\{x_p\}$，并用它们训练或提示少样本或零样本文本转语音/语音转换模型。
- 为提升合成效果，对手可能会应用扰动移除机制 R_ϕ（由ϕ参数化），以近似干净话语：

$$x_r = R_\phi(x_p).$$

随后，将净化后的话语 x_r 输入生成模型 G_θ（由 θ 参数化），以生成仿制语音$\tilde{x} = G_\theta(x_r)$。

- 对手可以访问一个由非目标说话人配对样本构成的小型辅助数据集 $\{(x_{(i)}x_{(i)p})\}_{i=1}^{N}$，并据此调整或训练 R_ϕ，但这些配对样本不包含目标用户。
- 对手无法对白盒访问 P_ψ（即无法访问其参数、梯度或内部机制），也不了解自动说话人验证系统的内部参数或阈值。

IV. 音频扩散桥接模型

为与干净数据分布对齐，我们提出了音频扩散桥接模型VocalBridge。这是一种净化机制，通过直接构建从受扰动数据分布到干净数据分布的反向桥，从基于扰动的防御所保护的数据中恢复干净音频。

我们基于*Li* 等人提出的对抗扩散桥接模型（*ADBM*），[40],该模型最初是为图像数据开发的。由于音频的时间分辨率高得多、噪声特性不同，且波形级扩散不稳定，直接将 ADBM 应用于音频领域并非易事。为应对这些挑战，我们将 ADBM 的公式适配到语音，并在神经音频编解码器（EnCodec）的潜在空间中执行所有扩散操作。这一设计使扩散能够在更紧凑的表示上进行，并使模型能够利用配对训练样本学习受保护音频与干净音频之间的映射，而不必像原始 ADBM 那样依赖分类器引导。随后，解码器根据净化后的潜在表示重建音频。

A. 训练目标

该模型是一种基于扩散的净化框架，旨在抑制音频潜在表示中的保护性扰动。其前向过程遵循去噪扩散概率模型的公式，但初始状态并非从干净潜在样本开始，而是由防护噪声项 ε_a 进行扰动。对于每个干净潜在向量$\mathbf{z}_c$，初始加噪状态定义为

$$z_0^a = \mathbf{z}_c + \varepsilon_a.$$

前向扩散过程表示为

$$z_t^a = \sqrt{\bar{\alpha}_t}\, z_0^a + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \quad 0 \leq t \leq T,$$

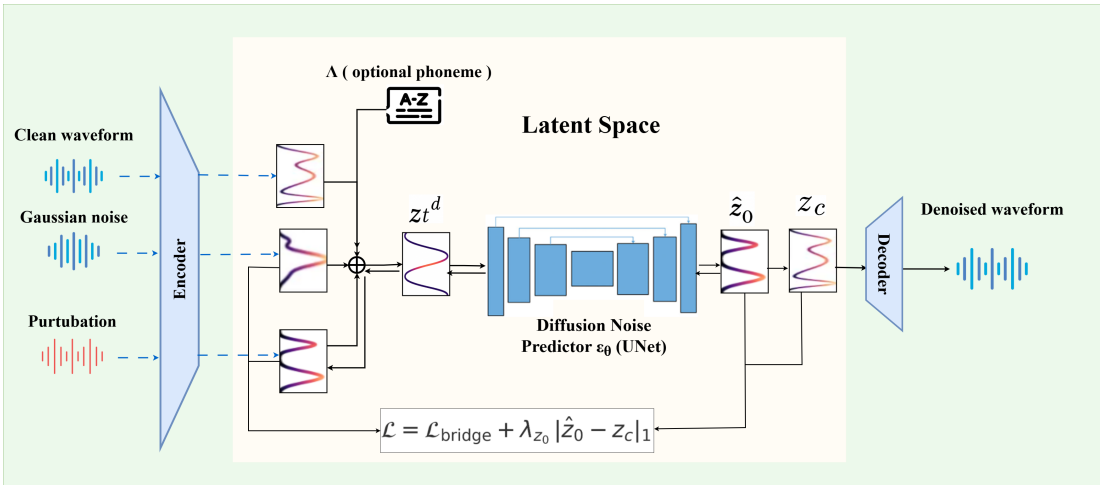

图 2：VocalBridge 训练

其中，$\bar{\alpha}_t$ 表示余弦噪声调度，T 表示用于净化的终止扩散步。

反向过程学习一个序列 $\{\hat{z}_t\}_{t:T\to 0}$，将经过扩散并添加保护噪声的潜变量分布 (z_T^a) 映射回干净分布$(\mathbf{z}_c)$。为使中间扩散轨迹对齐，缩放系数 $c_{\text{in}}(t)$和 $c_{\text{tgt}}(t)$按照 [40] 中的推导定义：

$$c_{\text{in}}(t) = \frac{\bar{\alpha}_T(1-\bar{\alpha}_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_T)}, \qquad c_{\text{tgt}}(t) = \frac{\bar{\alpha}_T\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_T)\sqrt{\bar{\alpha}_t}}.$$

桥接后的潜变量和相应的有效噪声目标定义如下

$$z_t^d = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_c + \sqrt{1-\bar{\alpha}_t}\,\varepsilon + c_{\text{in}}(t)\,\varepsilon_a, \tag{1}$$

$$\varepsilon_{\text{eff}} = \varepsilon + c_{\text{tgt}}(t)\,\varepsilon_a. \tag{2}$$

训练神经网络 $\varepsilon_\theta(z_t^d, t)$ 以估计有效噪声项 ε_{eff}。桥接损失定义为

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{t,\varepsilon}\left[\|\varepsilon_\theta(z_t^d, t) - \varepsilon_{\text{eff}}\|_2^2\right].$$

该公式与 Li 等人 [40], 提出的简化桥接损失等价；为清晰起见，其中省略了与 Ho 等人 [34]所述相同的常数缩放因子。

为增强潜变量空间的一致性，我们使用重建的干净潜变量估计引入一个辅助 L_1 正则化项：

$$\hat{z}_0 = \frac{z_t^d - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(z_t^d, t) - \big(c_{\text{in}}(t) - \sqrt{1-\bar{\alpha}_t}\,c_{\text{tgt}}(t)\big)\varepsilon_a}{\sqrt{\bar{\alpha}_t}}.$$

总目标函数变为

$$\mathcal{L} = \mathcal{L}_{\text{bridge}} + \lambda_{z_0}\,\|\hat{z}_0 - \mathbf{z}_c\|_1, \qquad \lambda_{z_0} \ll 1. \tag{3}$$

随着训练的进行，ε_a 的相对贡献会随着较小的 t而降低，使模型能够逐步抑制保护噪声并恢复干净的潜在表示。图2 展示了*VocalBridge*的整体训练过程。

B. 推理

在推理过程中，首先将受保护的波形编码到潜在域中，生成受保护的潜在表示$\mathbf{z}_a$。为启动反向扩散轨迹，我们在终止时间 T 构造一个带噪潜在状态，如下所示

$$z_T = \sqrt{\bar{\alpha}_T}\,\mathbf{z}_a + \sqrt{1-\bar{\alpha}_T}\,\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I).$$

随后应用多步 DDIM 风格的反向过程，将 $z_T \to \hat{z}_0$ 映射为目标。每个扩散步 t 中，去噪器 $\hat{\varepsilon}_\theta(z_t,\ t)$ 会预测当前潜变量中的噪声，并通过以下方式获得干净潜变量估计：

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\varepsilon}_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}.$$

按照预定的时间表 $t = T,\ t_{T-1},\ \ldots,\ 0$ 迭代执行此更新，可得到最终潜变量估计 $\hat{z}_0$，其近似于干净潜在表示 $\mathbf{z}_c$。通过 EnCodec 解码器对恢复的潜变量进行解码，即可得到净化后的波形。

C. Whisper 引导的音素条件化

我们在扩散桥净化器中加入了一种轻量级的音素引导的条件控制机制，为语音去噪和细化提供额外的时间结构。我们使用基于 Whisper 的对齐前端，直接从波形中估计近似的音素时间点，使模型能够在对齐净化后输出与干净语音分布的同时，更好地保留语音内容。不同于依赖真实转录文本的先前工作（Fan 等 [38]），我们的方法仅处理声学输入，无需转录文本监督。

令 y 表示音素引导信号，它是一个与时间对齐的序列，用于编码话语的音素结构。该信号由基于 Whisper 的对齐器生成的音素对齐图 Λ 推导而来，并与简单的声学先验结合，形成平滑的引导轨迹。在

训练期间，从干净波形中提取 y，进行标准化并通过有界非线性处理，然后将其均方根调整至与桥接潜变量 z_t^d 匹配，再进行通道拼接：

$$x_{\text{in}} = \left[z_t^d \parallel \text{RMSMatch}(y, z_t^d; \gamma) \right].$$

因此，去噪网络同时以含噪潜变量和音素引导为条件，使反向过程能够去除保护性扰动，同时保留语音结构。

在推理阶段，音素引导信号 y 直接从输入音频中提取，使用与前述相同的、基于 Whisper 的对齐模块，且无需访问干净参考音频或转录文本。由此得到的引导充当软时间先验，而非严格约束，因此该方法能够适应适度的对齐误差和残余噪声。若无法进行音素对齐，系统会自动切换到无条件模式，以零值引导通道代替，从而确保净化过程稳定且完全兼容。

D. 网络架构

净化网络采用带时间步条件化的一维 U-Net 架构。它完全在 EnCodec 潜变量域中运行，以桥接后的潜变量 z_t^d（以及可选的音素引导轨迹 y）为输入，并预测有效噪声项 $\hat{\varepsilon}_\theta(z_t^d,\ t,\ y)$。U-Net 采用分层编码器—解码器结构，并使用时延神经网络（TDNN）残差块，以捕捉音频表示中的局部和长程时间依赖关系。时间条件化通过正弦嵌入以及在各解码器阶段进行的特征级线性调制（FiLM）来实现，从而能够准确感知扩散步。

在音素引导变体中，额外的输入通道会将归一化的引导信号 $\Lambda(t)$与潜在张量拼接起来，使语言信息能够影响去噪，同时保持桥接一致性。网络使用 AdamW 优化器进行训练，并采用余弦学习率衰减和梯度裁剪来确保稳定性。

V. 实验与评估

我们从可验证性恢复和感知质量两方面评估所提出的净化框架，二者旨在共同破坏防御机制的不可模仿性，使受保护语音能够被克隆，并生成高质量的伪造语音。

A. 实验设置

所有实验均使用 PyTorch 实现，并在 NVIDIA A100 服务器（80 GB GPU、512GB 内存）上进行模型训练和净化。对于较轻量的推理任务，包括文本转语音和语音转换合成，我们还使用 NVIDIA Tesla V100 GPU（32 GB）。其余实验配置将在以下小节中介绍。

1）数据集：我们在两个广泛使用的基准语音数据集上进行了实验：**LibriSpeech** [48] 和**VCTK** [49]。LibriSpeech 包含多位说话人朗读不同转录文本的高质量有声读物录音；VCTK 则包含 110 位口音和录音条件各异的英语说话人。对于 LibriSpeech，我们选取 40 位说话人，并使用他们录制的第一章音频，其中包含适用于评估可懂度和说话人一致性的音素平衡句。对于 VCTK，我们使用每位说话人最先录制的 50 条话语。为了训练和评估我们提出的净化器 VocalBridge，我们将 VCTK 语料库中的 30 位说话人划分为训练集、80 位划分为测试集；将 LibriSpeech 语料库中的 27 位说话人划分为训练集、13 位划分为测试集。所有划分均保持性别均衡，以确保在不同数据集之间进行公平比较。两个数据集的评估共使用 4,526 个测试样本和 13,414 个训练样本。

为构建评估数据，我们对两个数据集应用保护性噪声防御，然后使用选定的文本转语音和语音转换工具合成克隆 音频。对于文本转语音克隆，我们从 VCTK 中抽取此前未见过的句子，并将每个句子指定为随机选取的一位受保护说话人的克隆目标。对于语音转换克隆，我们从源说话人中随机选取 50 个话语，并将每个话语转换成每位受保护说话人的声音。通过这一过程，每种防御措施使用文本转语音工具生成约 42,289 个合成样本（所有防御措施合计 209,936 个），使用三种语音转换工具生成 14,438 个样本（语音转换样本合计 80,404 个）。

为评估 VocalBridge 对抗基线净化方法的效果，我们在通过 TTS 和 VC 流程进行克隆之前，先对受保护样本应用每个系统。对于所评估的每种净化方法，我们生成 4,526 个净化测试样本，以及总计 290,340 个合成样本（其中 209,936 个来自 TTS，80,404 个来自 VC）。随后，我们使用这些合成输出评估可验证性恢复以及感知质量。

2) 保护工具：如前所述，基于扰动的语音防御方法旨在通过注入优化扰动来保护语音数据，使其无法被 TTS 和 VC 模型学习。在本工作中，我们依据以下两个标准选择了一组具有代表性的最先进防御方法：**(1)有效性：**每种方法都明确设计用于降低 TTS 和 VC 系统合成逼真语音的能力，同时保持语音的感知质量（质量防御），并保持用于自动说话人验证（ASV）的说话人身份（音色防御）。**(2) 可用性：**我们要求每种防御方法都提供公开发布的模型检查点以及相应的优化代码。因此，我们纳入了公开可用且报告性能优异的基于扰动的语音防御方法。具体而言，本工作选择了以下防御方法：

- **SafeSpeech：**SafeSpeech [11]提出了一种训练时防御方法，通过使音频样本无法用于文本转语音模型的学习，来应对零样本和微调语音克隆攻击。其核心机制——语音扰动隐蔽技术（SPEC）——利用一个替代生成模型来指导扰动的生成，

作者声称，这些扰动可以确保模型训练时无法从受保护数据中学到任何内容。SafeSpeech 不依赖推理时的对抗样本，而是引入通过梅尔频谱图损失和噪声分布的 KL 散度优化的扰动，旨在削弱由受保护音频生成的任何语音中的说话人身份（音色）和合成质量。

• **Attack-VC：** Attack-VC [50] 通过利用编码器-解码器结构来保护语音，其中编码器分为内容编码器和说话人编码器。负责捕捉语言内容的内容编码器保持不变，以保留话语所表达的内容；而该防御机制则专门针对提取说话人独特身份信息的说话人编码器。此机制会向输入话语添加精心设计的扰动。这些扰动旨在改变输出频谱图、说话人嵌入向量，或同时改变两者，因此即使语言内容仍然清晰，语音转换模型也无法准确克隆或识别说话人的声音。这些变化对人类而言几乎难以察觉，却能显著降低先进语音转换技术仿制语音或滥用声音的风险。嵌入攻击在有效性和速度之间实现了最佳权衡，并且在不同模型上具有稳健性；因此，我们在评估中使用嵌入攻击。

• **关键目标扰动（POP）：** POP [29] 是一种在发布音频前向其添加微小且难以察觉的对抗性扰动的方法，旨在使文本转语音语音克隆模型无法从数据中学习。作者声称，POP 仅通过优化重建损失（真实音频与合成音频之间的差异）来生成这些扰动；由于几乎所有文本转语音模型都使用这一损失，因此该方法能够普遍适用、高效且可迁移至不同架构。他们报告称，POP 生成的受保护音频听起来自然，人类听众难以察觉异常；但使用这些受保护数据训练文本转语音模型后，生成的合成语音会充满噪声，无法使用。

• **Anti-Fake：** AntiFake [10] 通过扰动说话人嵌入空间来破坏合成语音中的身份特征，确保生成的声音无论对人类还是机器而言，都不再像原始说话人。它采用两种优化策略：基于阈值的方法将嵌入向量推至与原始嵌入向量相距超过设定距离的位置；基于目标的方法则将嵌入向量移向另一位说话人。为确保能够迁移至未知的文本转语音模型，该方法使用多种不同的说话人编码器组成的集成来优化扰动。它还引入了一种基于人类听觉敏感度和信噪比的感知损失，以保持音频质量。最后，它采用人在回路流程来选择目标声音并验证感知上的差异。

•**通过生成对抗网络实现的语音转换主动防御（GAN-ADV）：** GAN- ADV [51]提出了一种对抗式防御框架，利用生成器-判别器架构（GAN）在梅尔频谱域生成扰动，旨在干扰语音转换系统，同时避免对音频质量造成可感知的影响。该系统包含一个仿真模块（SWCSM），用于模拟波形重建和特征重新提取过程中的有损操作，从而增强对真实推理管线的鲁棒性。训练期间使用替代语音转换模型提供梯度信号，扰动优化则兼顾三个目标：欺骗判别器（GAN 损失）、干扰语音转换输出（防御损失）以及保持音频保真度（质量损失）。推理时只需使用训练好的生成器，因此部署效率很高。

我们选择这些公开防御方法，是因为它们性能出色，且总体涵盖了在不同特征空间中施加扰动、并在合成模型间表现出不同程度可迁移性的方法。**SafeSpeech (SPEC)** 引入了由替代生成模型引导的训练时波形扰动。**Attack-VC**通过直接扭曲输出频谱图（端到端）、改变说话人嵌入表示，或通过反馈机制结合两者，对频谱图表示施加扰动；据报告，在这些方法中，说话人嵌入攻击变体最为有效，因此我们采用了该变体。**POP**同样在波形域中运行，但会针对梅尔频谱图重建损失优化扰动。**AntiFake** 通过在编码器集成上进行基于阈值和基于目标的优化来操纵说话人嵌入，从而实现跨合成系统的部分可迁移性。**GAN-ADV** 通过旨在经波形重建后仍保持鲁棒性的生成对抗网络框架，在梅尔频谱图域中生成扰动。

3）语音克隆工具：我们假设对手可以使用多种先进的开源文本转语音和语音转换模型来合成/克隆语音，这些模型分别代表了神经语音合成中不同的架构范式和最新进展。我们选取了六种具有代表性的模型

•**VALL-E-X：** VALL-E-X [52] 是一种神经编解码器语言模型，可根据源语言语音和目标语言文本预测离散声学令牌。它支持零样本、跨语言文本转语音和语音到语音翻译，同时保留说话人的音色、情感和声学环境线索。

• **Tortoise-TTS：** Tortoise-TTS [53] 将基于自回归 Transformer 的声学模型与扩散解码器和 UnivNet 声码器相结合。它首先预测经过压缩的声学令牌，然后将其细化为富有表现力的高质量波形。

•**StyleTTS2：** StyleTTS2 [54] 是一种文本转语音模型，它采用风格扩散和对抗训练，并结合大型语音语言模型（*SLM*）生成自然流畅的音频。它通过扩散将说话风格建模为潜变量，从而实现风格一致的合成

而无需参考语音。

• **VQMIVC：** VQMIVC [55] 是一种单样本语音转换模型，利用向量量化进行内容编码，并通过互信息最大化，以无监督方式解耦内容、说话人和基频表示。

• **HierSpeech++：** HierSpeech++ [56]是一种分层变分自编码器，融合了文本到语义单元建模、韵律控制和语音超分辨率。该模型集成了归一化流模块，以实现高保真零样本合成与转换。

•**DiffHierVC [57]：** 一种采用两个扩散模型 *DiffPitch* 和 *DiffVoice*进行序列转换的分层语音转换框架。该框架通过源声-滤波器编码器解耦语音表示，实现语音风格迁移；同时利用掩码先验提升说话人适应质量。

4）说话人验证系统：我们采用三种标准自动说话人验证系统来评估说话人身份保持与恢复能力： *x-vector*、*ECAPA* 和 *d-vector*，均使用来自 SpeechBrain [58]和 Resemblyzer [59]的预训练模型实现。每个系统都会将输入波形映射为固定维度的嵌入向量，用于表示说话人的声纹身份。测试语句与注册说话人质心之间的验证通过余弦相似度进行：

$$\mathrm{score}(\tilde{\mathbf{e}}, \mathbf{c}) = \frac{\tilde{\mathbf{e}}^\top \mathbf{c}}{\|\tilde{\mathbf{e}}\|_2 \, \|\mathbf{c}\|_2},$$

其中， $\tilde{\mathbf{e}}$ 表示测试嵌入向量，$\mathbf{c}$ 表示说话人注册嵌入向量的质心，通过对注册话语的嵌入向量取平均计算得出。

决策阈值依据等错误率（*EER*）准则确定，该准则将误接受率（*FAR*）与误拒绝率（*FRR*）相等时的工作点定义为：

$$\mathrm{EER} = \mathrm{FAR}(\tau_{\mathrm{eer}}) = \mathrm{FRR}(\tau_{\mathrm{eer}}),$$

其中， τ_{eer} 表示等错误阈值。所有自动说话人验证嵌入向量和决策阈值均使用我们的评估数据集计算。为防止数据泄漏，用于阈值校准的子集（开发集）与用于注册和净化评估的话语严格不重叠。每位说话人使用少量干净话语计算说话人质心，其余话语及其受保护或净化后的对应话语则用作测试试验。所有实验中，自动说话人验证系统始终保持冻结状态，仅作为说话人身份一致性的客观评估器

在包含评估数据集中149位说话人的开发集上，x-vector、ECAPA和d-vector系统的等错误率分别为0.0486、0.006和0.0297，对应的阈值为 τ_{eer}=0.951（x-vector）、0.419（ECAPA）和0.750（d-vector）。这些校准后的阈值在后续所有净化和恢复实验中保持固定。

5) 评估指标：为评估 VocalBridge 的有效性，我们根据第 III节中定义的攻击者目标对其进行评估。

• 可验证性恢复：遵循 AntiFake 的评估设计 [10]，（该设计通过认证规避率降低率（AERR）衡量防御效果），我们采用适用于净化的逆向视角。我们的目标是移除会抑制说话人验证的保护性扰动，从而恢复认证。为此，我们引入认证恢复率（ARR）作为主要指标，用于量化净化的效果。ARR 衡量先前未通过验证（低于阈值）的受保护话语在净化后成功通过验证的比例。对于每个身份，我们根据一组干净注册话语计算干净注册质心 c。令 s(e,c) 表示嵌入向量 e 与注册质心之间的余弦相似度。我们分别用 $s\mathrm{prot}_i$ = s(eprot_i , c) 和 $s\mathrm{pur}_i$ = s(epur_i , c) 表示同一话语的受保护版本和净化版本的相似度分数。给定第 V-A4 节中的等错误阈值 τeer，ARR 定义为

$$\mathrm{ARR}(\tau_{\mathrm{eer}}) = \frac{\sum_i \mathbf{1}\{s_i^{\mathrm{prot}} < \tau_{\mathrm{eer}} \ \wedge \ s_i^{\mathrm{pur}} \geq \tau_{\mathrm{eer}}\}}{\sum_i \mathbf{1}\{s_i^{\mathrm{prot}} < \tau_{\mathrm{eer}}\}},$$

ARR 值越高，表示身份验证恢复效果越强，且保护性扰动的去除效果越好。

• **不可感知性：**为评估感知保真度和自然度，我们采用客观的平均意见分（MOS），该分数由 NISQA 模型预测 [60]。NISQA 按照 1（较差）到 5（优秀）的等级评估语音质量和可懂度，并自动近似模拟人类的感知判断。MOS 值越高，表示语音越自然、越清晰易懂；分数高于 3 通常意味着音频质量良好 [10]。

我们还使用词错误率（WER） [61] 评估发音清晰度。由于预训练的 Whisper-small 模型 [46] 在处理我们的大型数据集时计算效率较高，因此我们采用该模型进行转录。WER 值越高，表示语音清晰度越低。

6）净化基线：本节简要介绍实验中使用的降噪工具。所选工具包括对抗性降噪器和保护性扰动去除器，代表了近期多种不同的方法。

• **De-Antifake：** De-AntiFake [38] 是一种语音克隆攻击评估与净化系统，旨在测试并击败依赖对抗扰动的现有语音保护机制。它模拟真实的攻击者：在执行语音克隆之前，先运用净化技术去除语音中的保护噪声。该工具提出了一种名为 *PhonePuRe* 的新型净化框架，分两个阶段运行：（1）净化阶段：使用基于扩散的模型清理经对抗性扰动的音频。（2）细化阶段：采用音素引导对齐，对净化后的语音进行微调，使其与自然、未受保护的语音高度相似。

• **WavePurifier：** WavePurifier [36] 是一种防御工具，旨在净化针对自动语音识别的音频对抗样本 ^{y}干扰成分。它采用分层扩散模型，

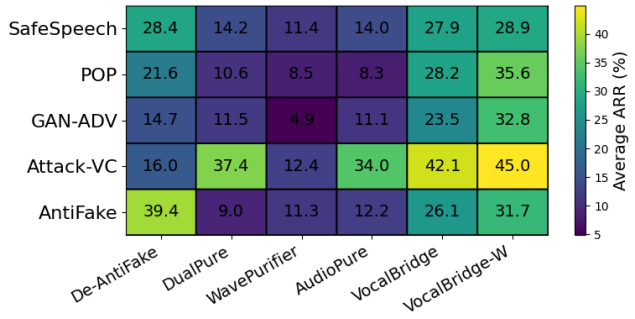

图 3：三个说话人验证后端（x-vector、ECAPA-TDNN 和 d-vector）的平均 ARR（%）。

通过正向扩散去除扰动，并通过反向扩散恢复干净语音。该工具将频谱图划分为低、中、高频带，并针对各频带优化净化强度，在去除攻击的同时保持语音质量。在多个自动语音识别模型和攻击上进行评估后，WavePurifier 的表现优于七种现有防御措施，在多种场景下实现了最低的字符错误率和词错误率，以及最高的净化成功率。

- **AudioPure：** AudioPure [33] 是一种基于对抗净化的防御流程，专为采用现成扩散模型的声学系统设计。它利用扩散模型生成噪声，并将少量噪声添加到对抗音频中。随后，通过反向采样步骤净化含噪音频并恢复干净音频。这是一种即插即用的方法，可应用于任何预训练分类器，无需额外微调或重新训练。

- **DualPure：** DualPure [37] 是一种针对对抗扰动的实时净化防御方法。首先，它会在波形层面扰乱样本中可能存在的恶意扰动。随后，使用无条件扩散模型在频域层面净化特征。具体而言，该方法首先使用时域净化器（*TDP*）净化波形信号，然后将波形转换为梅尔频谱图，并进行频域净化。该方法对白盒和黑盒攻击均具有良好的对抗鲁棒性。

B. 实验结果

1）恢复可验证性：*VocalBridge* 与基线方法：在本节中，我们将 VocalBridge 与攻击者可采用的四种净化机制进行比较：De-AntiFake（包含 Purification 和 Refinement 的完整模型）、AudioPure、DualPure 和 WavePurifier。其中，De-AntiFake 是唯一一项明确旨在通过移除保护性扰动来击败语音克隆防御措施的相关工作；据我们所知，针对这一场景，尚无其他基于净化的攻击被提出。其余净化器最初是为缓解针对自动语音识别系统的对抗攻击而开发，并非专为我们的保护移除威胁模型设计。尽管如此，我们仍将它们纳入基线方法，以确定这些净化技术是否会无意中成为有效的保护移除攻击。表 I 和表 II报告了由选定的文本转语音和语音转换模型生成的语音在经过净化的数据集上的 ARR，评估涵盖三种自动说话人验证系统：x-vector、ECAPA-TDNN 和 d-vector。如图 3所示，VocalBridge-W 在各种保护措施下总体恢复效果最佳。在 Attack-VC 保护下，VocalBridge-W 的 ARR 达到 45.0%，比现有最佳方法（DualPure，37.4%）高出 7.6 个百分点。在 GAN-ADV 和 POP 保护下，VocalBridge-W 的 ARR 分别达到 32.8% 和 35.6%，分别超过最强基线方法 18.1 和 14.0 个百分点。在 SafeSpeech 保护下，VocalBridge-W 同样实现了最高恢复率（28.9%），略高于先前工作。VocalBridge 也取得了类似的提升：Attack-VC 下的性能从 37.4% 提升至 42.1%，GAN-ADV 下从 14.7% 提升至 23.5%，POP 下从 21.6% 提升至 28.2%。唯一未由我们模型领先的情形是 AntiFake；在该情形下，专用方法 De-AntiFake 的表现仍然更高（39.4%，而 VocalBridge-W 为 31.7%）。值得注意的是，De-AntiFake 需要完整访问 AntiFake 检测器及其真实转录文本才能运行其细化器，而这一条件在实际应用中很少满足；相比之下，VocalBridge 和 VocalBridge-W 无需任何特权模型访问，因此实用性显著更强，适用范围也更广。

图 4 展示了语音克隆输入和输出样本的频谱图。视觉对比表明，与最先进的 De-AntiFake 基线方法相比，VocalBridge-W 能更有效地保留干净音频的结构和频谱细节。De-AntiFake 的一个显著局限是依赖干净转录文本；在 SafeSpeech 和 AntiFake 等嘈杂或失真条件下，这一假设往往不再成立。相比之下，VocalBridge 和 VocalBridge-W 无需真实转录文本，因此即使转录文本质量下降或不可用，仍能保持有效性，为鲁棒性评估提供了更实用的基础。

总体而言，这些结果表明，与现有方法相比，VocalBridge 和 VocalBridge-W 能提供更可靠的净化效果，有效去除保护性扰动，并在不同保护机制、自动说话人验证后端和合成架构下恢复说话人身份。

2）*VocalBridge* 跨扰动泛化能力评估：为了评估净化框架的泛化能力，我们使用一个*Mono* 模型，该模型仅使用一种单一 扰动模式进行训练。我们仅使用*Mono*在 *GAN-ADV* 下生成的扰动模式进行训练，然后将其应用于受到其他保护措施扰动的音频，且不进行任何形式的适应。通过这一设置，我们可以评估仅从一种扰动类型中学习是否足以去除多种未见扰动。表III显示，Mono 变体在所有保护类型下都与 Adaptive 模型保持高度一致。

总的来说，这些结果表明，VocalBridge 即使面对训练期间从未遇到过的扰动模式，仍能保持较高的身份认证恢复性能，

表 I：所选文本转语音模型的认证恢复率（%）。

保护	文本转语音模型	De-AntiFake			DualPure			WavePurifier			AudioPure			VocalBridge			VocalBridge-W		
		xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec
AntiFake	StyleTTS2	86.61	40.34	54.05	13.11	0.40	0.94	52.02	8.72	1.40	19.97	1.54	1.65	58.17	21.74	2.76	55.27	12.71	7.24
	Tortoise-TTS	62.05	53.75	59.90	19.57	1.11	0.98	24.37	9.66	0.84	25.56	4.55	2.07	47.86	35.86	8.37	42.83	30.34	11.57
	VALL-E-X	45.42	26.20	25.36	63.72	7.94	1.12	25.79	6.61	0.37	59.41	3.44	1.94	38.13	14.68	11.54	48.91	16.43	16.44
平均（文本转语音）		64.69				3.15	1.01	34.06	8.33	0.87	34.98	3.18	1.89	48.05		7.56	49.00		
Attack-VC	StyleTTS2	32.32	10.12	9.07	48.54	36.93	54.47	5.29	14.30	10.05	45.27	29.51	47.32	61.21	60.03	67.23	63.86	60.66	64.37
	Tortoise-TTS	32.52	9.60	3.94	36.56	23.89	28.29	3.45	16.26	2.53	41.74	19.90	20.52	38.07	25.91	15.67	38.30	24.40	10.03
	VALL-E-X	29.47	8.00	3.36	73.03	52.84	25.25	15.63	12.24	1.93	69.51	42.71	19.02	39.67	36.02	19.79	41.48	35.51	24.97
平均（文本转语音）		31.44	9.24	5.46	52.71			8.12	14.27	4.84	52.18								
GAN-ADV	StyleTTS2	39.82	14.93	6.09	17.37	2.44	4.05	6.24	7.56	1.87	14.41	2.43	5.04	49.01	36.92	31.31	57.58	40.24	34.22
	Tortoise-TTS	32.76	10.88	2.64	28.05	5.99	3.11	3.71	8.19	0.68	22.99	7.26	2.62	38.99	16.79	9.90	49.11	12.04	7.25
	VALL-E-X	29.80	8.92	3.27	58.56	8.68	4.73	12.94	4.67	0.95	61.95	5.94	4.69	40.86	19.98	17.71	47.99	20.83	18.45
平均（文本转语音）		34.13		4.00	34.66	5.70	3.96	7.63	6.81	1.17	33.12	5.21	4.12	42.95					
POP	StyleTTS2	38.91	30.21	43.38	14.24	0.00	0.00	10.82	14.77	5.82	12.35	0.09	0.00	40.00	50.04	58.76	45.49	52.78	60.50
	Tortoise-TTS	33.75	17.92	12.81	28.10	2.66	2.44	2.69	7.39	0.43	22.02	1.25	1.30	42.86	15.09	11.75	45.06	18.04	10.94
	VALL-E-X	31.02	20.98	17.77	59.92	7.59	7.19	8.49	10.12	1.11	54.92	6.59	5.11	35.16	31.03	40.87	35.26	29.25	43.35
平均（文本转语音）		34.56				3.42	3.21	7.33	10.76	2.45	29.76	2.64	2.14	39.34					
SafeSpeech	StyleTTS2	58.75	21.67	17.89	33.48	2.04	0.90	20.28	7.85	1.97	33.72	1.56	2.09	54.87	21.13	31.01	56.44	22.92	31.06
	Tortoise-TTS	49.30	20.54	13.38	34.66	5.04	3.84	14.21	7.79	1.04	28.83	7.49	4.00	39.44	5.88	5.25	33.91	4.90	5.78
	VALL-E-X	35.29	21.50	7.65	62.49	3.55	3.93	10.97	3.53	0.22	63.71	4.28	1.98	41.65	15.33	16.18	37.03	12.67	17.42
平均（文本转语音）		47.78				3.54	2.89	15.15	6.39	1.08	42.09	4.44	2.69	45.32					

表 II：所选语音转换模型的认证恢复率（%）。

保护	语音转换模型	De-AntiFake			DualPure			WavePurifier			AudioPure			VocalBridge			VocalBridge-W		
		xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec	xvec	ecapa	dvec
AntiFake	DiffHierVC	77.39	23.99	74.05	8.70	1.54	1.42	40.69	2.85	4.82	18.12	1.54	1.42	37.05	5.57	13.42	26.81	3.73	9.96
	HierSpeechpp	2.75	2.89	16.75	1.61	8.64	31.28	0.00	6.94	17.77	10.92	29.28	37.63	2.40	14.93	45.69	35.54	51.93	75.63
	VQMIVC	6.32	17.84	33.03	0.00	0.34	0.00	0.24	0.00	0.00	0.00	0.00	0.00	67.65	37.97	5.83	72.30	43.73	8.52
平均值（语音转换）		28.82			3.44	3.51	10.90		3.26	7.20	9.01	10.27							
Attack-VC	DiffHierVC	13.79	2.28	11.34	13.79	36.48	50.85	46.67	13.08	22.03	22.58	18.77	43.38	30.00	54.18	73.45	34.38	53.21	70.34
	HierSpeechpp	1.49	0.78	8.15	3.36	20.10	33.03	0.00	5.75	10.70	6.29	30.60	41.28	4.67	24.75	42.51	19.72	41.63	56.88
	VQMIVC	48.85	35.59	27.93	38.04	41.70	55.68	33.36	7.98	2.70	35.59	39.00	38.92	58.66	48.21	58.38	59.29	51.29	58.92
平均值（语音转换）		21.38							8.94	11.81									
GAN-ADV	DiffHierVC	25.91	3.60	2.66	13.64	1.46	3.98	8.65	2.57	4.31	13.84	1.51	5.27	22.07	11.95	25.27	35.14	17.43	38.34
	HierSpeechpp	17.83	2.77	2.22	3.55	0.82	2.77	1.40	1.21	4.10	9.66	1.35	2.36	10.24	7.99	17.09	33.80	14.53	29.59
	VQMIVC	20.06	25.19	15.86	23.33	8.39	15.65	14.39	2.11	2.04	24.74	6.21	6.80	24.67	25.00	17.69	59.39	43.17	30.61
平均值（语音转换）		21.27		6.91	13.51	3.56	7.47	8.15	1.96	3.48	16.08	3.02	4.81	19.00					
POP	DiffHierVC	6.25	5.09	36.13	5.26	0.47	0.27	60.00	9.29	2.73	0.00	0.69	0.27	10.00	22.38	28.42	14.29	25.93	36.61
	HierSpeechpp	0.83	2.00	29.86	0.00	0.00	0.00	0.74	0.54	1.10	0.79	0.27	0.00	0.76	7.26	17.63	29.01	38.07	49.59
	VQMIVC	21.43	32.17	7.94	27.52	15.38	19.30	10.89	5.26	0.00	21.43	14.29	7.94	37.61	29.33	28.07	38.74	35.06	33.33
Avg（VC）		9.50	13.09			5.28	6.52	23.88	5.03	1.28	7.41	5.08	2.74	16.13					
SafeSpeech	DiffHierVC	46.13	8.81	24.97	23.69	0.84	0.62	31.56	3.33	9.38	19.57	1.82	4.20	36.96	5.14	27.34	39.14	6.71	34.94
	HierSpeechpp	61.44	11.95	38.10	23.65	1.40	1.40	34.23	3.97	11.92	15.24	2.50	5.63	44.76	8.97	36.21	46.15	10.44	39.28
	VQMIVC	29.62	15.09	29.47	26.39	11.84	16.10	23.24	7.56	12.68	30.89	11.52	13.62	63.01	32.47	16.59	66.95	35.94	19.02
平均值（语音转换）		45.73				4.69	6.04	29.68	4.95	11.33		5.28	7.82	48.24					

表明其在异构净化防御措施中具有很强的稳健性和泛化能力。

C. *Whisper* 引导的音素条件化的有效性

将 Whisper 特征融入 VocalBridge（形成表格 *I* 和*II*中的 VocalBridge-W）可提高净化稳定性，并帮助模型保留与说话人相关的语音结构。在所有保护设置下，Whisper 引导的条件控制相较于简单的 VocalBridge 净化器都能带来稳定的提升。例如，在 GAN-ADV 下，VocalBridge-W 将语音转换模型的编码器平均 ARR 从 19/15/20% 提升至 43/25/33%；在 POP 下，ARR 则从 16/20/25% 提升至 27/33/40%。

1）语音质量与不可感知性的评估：表格IV 报告了所有语音转换/文本转语音系统的平均 NISQA-TTS MOS。受保护语音的感知质量接近原始语音（3.36 对 3.57）。现有净化防御措施的质量明显较低，平均 MOS 值介于 2.95 至 3.27 之间。我们的方法在质量上与受保护样本（3.36）相当，同时优于所有先前的净化方法。图6 报告了不同保护机制在合成语音上的平均词错误率。VocalBridge 达到了最低词错误率（0.258），优于所有其他方法。

这些结果表明，VocalBridge 能够保持较高的感知保真度，并避免过度平滑和频谱

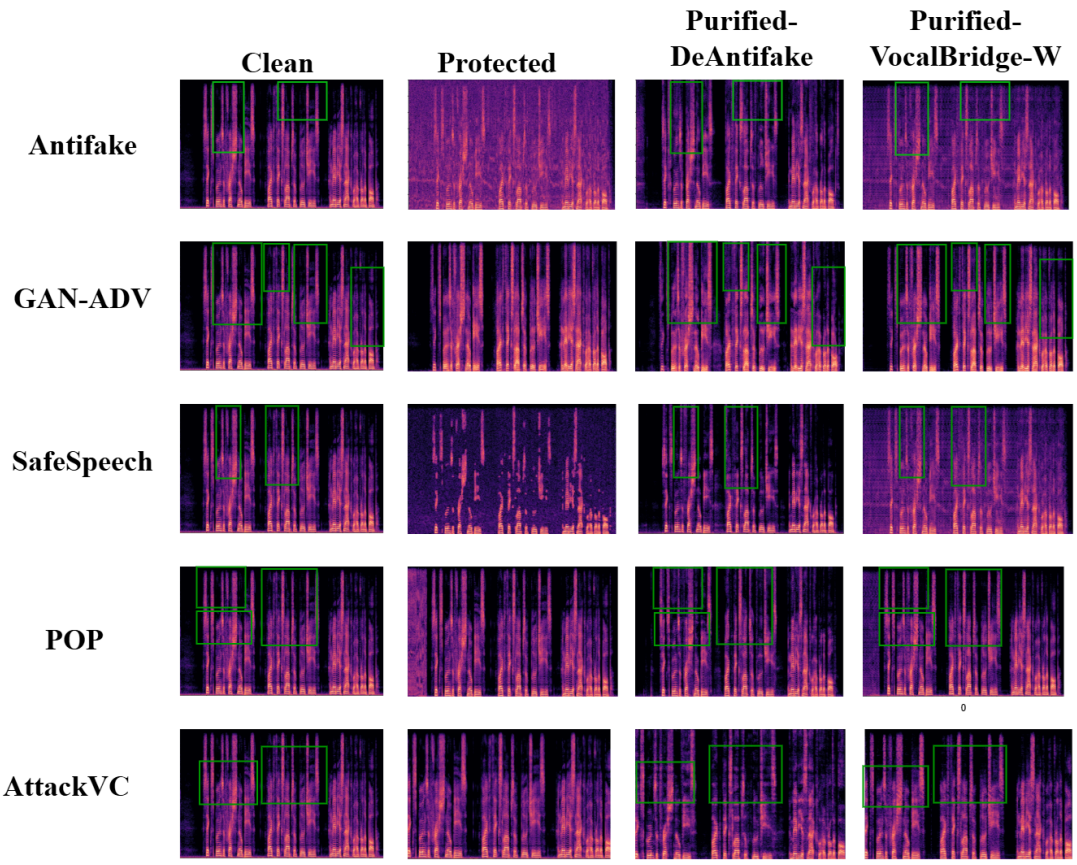

图4：频谱图对比显示，与领先的基线方法 De-AntiFake 相比，VocalBridge-W 更好地保留了干净样本的结构和细节。

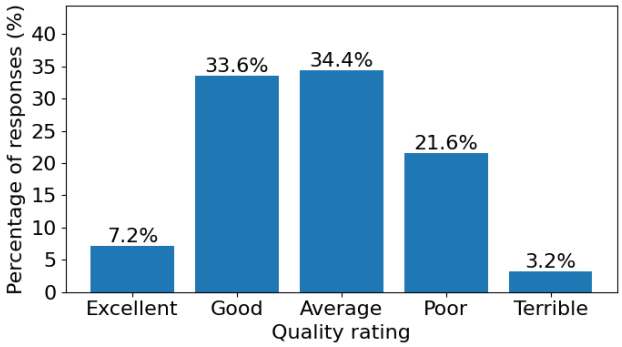

图5：不同自适应策略下的 ARR。

失真，这些失真存在于基线方法中；因此，攻击者使用的语音转换/文本转语音模型能够更准确地复现说话人特征。

D. 语音质量主观评估

我们开展了一项包含27名参与者的用户研究，以评估系统生成音频的主观质量。研究包括一份人口统计学问卷，随后是一组听辨问题，参与者在其中对通过克隆生成的合成音频的感知质量进行评分使用 VocalBridge 净化后的样本。参与者使用四级量表对每个音频片段进行评分：良好、一般、较差或糟糕。调查包括通过净化由 AntiFake、Attack-VC、GAN-ADV、POP 和 SafeSpeech 引入的扰动而获得的音频，随后使用 Tortoise-TTS 对净化后的音频进行克隆；每位参与者会看到这些样本中的随机子集。调查还包含一道带有明显白噪声损坏的注意力检查题，以确保参与者的可靠性。如图 5 所示，所有评分中有 75.2% 落在良好、一般或优秀类别，仅有很小一部分被评为较差或糟糕，这表明净化并克隆后的音频总体上被认为自然且清晰易懂。该调查使用 Qualtrics 创建，并通过亚马逊土耳其机器人开展。

E. 自适应保护

我们还评估了一种自适应保护场景：在该场景中，保护机制可以白盒访问我们的净化模型（包括其梯度），因此能够据此优化其扰动。我们采用 De-Antifake 中使用的自适应攻击方法 [38]。由于从输入波形到净化后输出的整体净化函数实际上不可微，这是由 EnCodec量化和随机扩散采样等组件导致的，因此我们采用其反向传播可微近似

表III：自适应和单一净化模型下 VC 和文本转语音的认证恢复率（%）

保护	模型	VocalBridge	VocalBridge（单声道）	
		xvec ecapa d向量	xvec ecapa	dvec
AntiFake	DiffHierVC	37.05 5.57 13.42	18.12 2.62	2.87
	HierSpeechpp	2.40 14.93 45.69	2.40 13.72	32.49
	VQMIVC	67.65 37.97 5.83	66.90 36.61	3.59
	StyleTTS2	58.17 21.74 2.76	33.13 4.83	2.15
	Tortoise-TTS	47.86 35.86 8.37	38.78 6.43	2.22
	VALL-E-X	38.13 14.68 11.54	34.00 12.46	12.04
平均值（全部）		**41.88**		**9.23**
POP	DiffHierVC	10.00 22.38 28.42	15.00 28.47	30.60
	HierSpeechpp	0.76 7.26 17.63	3.05 8.99	12.95
	VQMIVC	37.61 29.33 28.07	42.48 32.47	31.58
	StyleTTS2	40.00 50.04 58.76	49.17 49.68	53.05
	Tortoise-TTS	42.86 15.09 11.75	42.33 13.45	7.69
	VALL-E-X	35.16 31.03 40.87	34.86 32.07	37.61
平均值（全部）		**27.73**		**28.91**
Attack-VC	DiffHierVC	30.00 54.18 73.45	30.00 54.55	63.84
	HierSpeechpp	4.67 24.75 42.51	4.73 24.21	31.50
	VQMIVC	58.66 48.21 58.38	57.84 55.98	52.43
	StyleTTS2	61.21 60.03 67.23	61.31 59.97	63.85
	Tortoise-TTS	38.07 25.91 15.67	38.61 26.49	13.34
	VALL-E-X	39.67 36.02 19.79	39.82 35.48	19.43
平均值（全部）		**38.71**		**40.73**
SafeSpeech	DiffHierVC	36.96 5.14 27.34	28.96 4.14	14.48
	HierSpeechpp	44.76 8.97 36.21	44.33 7.22	15.87
	VQMIVC	63.01 32.47 16.59	57.42 30.77	10.73
	StyleTTS2	54.87 21.13 31.01	47.08 16.98	20.29
	Tortoise-TTS	39.44 5.88 5.25	28.63 4.44	2.45
	VALL-E-X	41.65 15.33 16.18	35.82 11.47	9.54
平均值（全部）		**46.78**		**12.23**

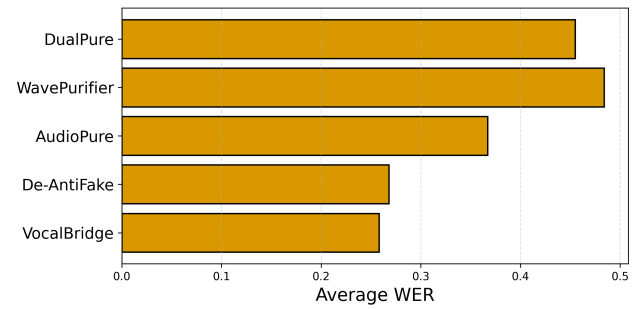

图 6：保护机制在合成语音上的平均词错误率。

（BPDA）策略。BPDA 在反向传播过程中将净化器视为恒等映射，从而生成替代梯度，使我们能够在这些不可微操作存在的情况下进行端到端扰动优化。与 De-Antifake 一样，我们还使用变换期望（EOT）来考虑扩散过程中的随机性。我们对 1、5、10 和 15 次随机运行所得的梯度估计取平均值，从而在随机采样下获得稳定的梯度近似。

为评估我们方法的有效性，我们首先使用 VocalBridge-W 净化经过自适应保护的语音，然后使用 StyleTTS2 克隆净化后的音频。我们使用 x-vector 模型测量攻击成功率，指标包括 *ARR*和说话人验证准确率（*SVA*）。

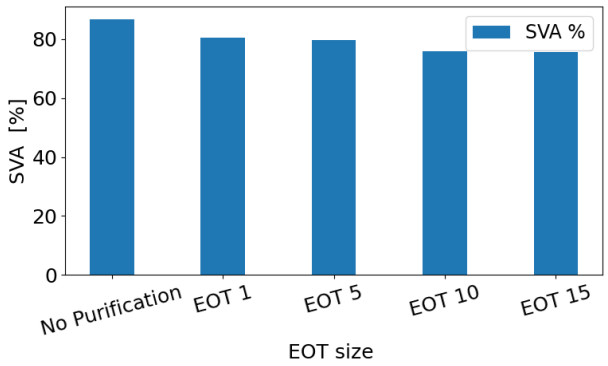

图 7：不同自适应策略下的 SVA。

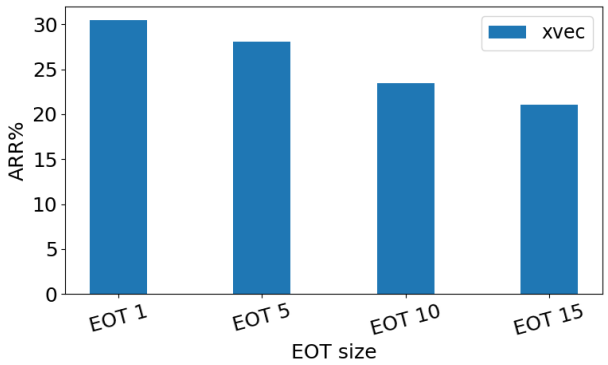

图 8：不同自适应策略下的 ARR es.

其中，SVA 定义为被自动说话人验证系统判定为真实语音的克隆样本所占比例，计算方式是对所有克隆样本的二元匹配判定取平均值。如图7 和图8所示，在这些自适应策略设置下，自动说话人验证准确率仍高于 75%，ARR 仍高于 20%。这些结果表明，即使掌握白盒知识，要针对我们的净化方法制定有效的自适应保护仍然困难，凸显了我们所指出的风险。

VI. 讨论与局限性

A. 音素引导细化

我们的方法使用 Whisper *small* 模型为 Λ引导的细化生成音素级对齐。由于计算资源有限，我们选择了这一模型，这降低了对齐准确率，尤其是在音频嘈杂或受到对抗性干扰时。因此，音素引导只能带来小幅提升。资源限制也使我们无法尝试更大型或更稳健的对齐模型。未来的工作可以探索更强大的 Whisper变体，以获得更可靠的 Λ 特征和更显著的去噪收益。

表 IV：NISQA-TTS 预测的平均意见分（MOS）（1–5，越高越好）

语音转换 / 文本转语音	原始	受保护	De-Anti	Fake	Dual	Pure	WavePurifier	AudioPure	VocalBridge（我们的方法）
StyleTTS2	3.68	3.61	3.13	2.88	3.31	2.86	3.59		
Tortoise-TTS	3.92	3.58	3.44	3.34	2.40	3.23	3.62		
VALL-E-X	3.59	3.27	3.05	3.20	2.74	3.19	3.29		
DiffHierVC	3.43	3.28	3.47	2.93	3.25	2.86	3.24		
HierSpeechpp	3.88	3.63	3.81	3.41	3.54	3.23	3.59		
VQMIVC	2.91	2.78	2.73	2.82	2.46	2.80	2.81		
平均值（全部）	**3.57**	3.36	3.27	3.10	2.95	3.03	**3.36**		

B. 伦理考量

语音合成技术的迅速发展，引发了语音克隆技术与保护性对策之间的竞赛。本研究通过揭示现有基于扰动的对策存在的缺陷，进一步加剧了这种对抗态势。我们认识到这项工作的影响是多方面的；虽然我们的主要动机是评估现有防护措施的稳健性，从而推动开发更具韧性的解决方案，但所提出的扩散桥净化模型也可作为有效方法，绕过这些防护措施。我们坚信，增强防护措施以应对未经授权的语音合成和语音克隆所带来的潜在益处，远大于滥用研究发现的风险，尤其是我们无意公开源代码。为了促进开放研究和科学进步，我们将在收到请求后，经过全面审查并确认请求意图，才会提供源代码。为确保所评估保护工具的开发者能够妥善应对并做出必要调整，我们计划在本文被接收后向他们披露研究发现，并且一定会在本研究发表之前完成披露。

我们的研究结果凸显出，业界迫切需要重新思考基于扰动的方法，并探索从根本上全新的语音数据保护策略。未来的保护机制必须在设计时充分考虑对先进预处理和净化操作的鲁棒性，确保其在对抗能力不断提升的情况下仍然有效。

VII. 结论

本文考察了当代语音保护机制对基于净化的攻击的脆弱性，并提出了 VocalBridge——一种用于重建说话人身份的桥接式潜在扩散模型。VocalBridge 能够在文本转语音和语音转换流程中实现有效的身份恢复，其 Whisper 引导扩展 VocalBridge-W 进一步稳定了语音结构，并进一步提升了重建质量。

我们的评估结果表明，VocalBridge 和 VocalBridge-W 的表现显著优于现有净化方法。我们还评估了模型的泛化能力，结果表明，攻击者无需明确了解防护系统的扰动模式，也能成功实施重构攻击。我们的自适应防护研究显示，即使拥有完全的白盒访问权限，要构建能够抵御我们净化方法的稳健防御措施仍然颇具挑战性，这凸显了我们所揭示风险的严重性。

总体而言，我们的结果表明，当前的语音保护技术仍然容易受到基于潜在空间扩散的净化攻击。通过发布我们的评估框架和桥接扩散模型，我们旨在推动开发更强大、更具原则性的语音隐私和合成媒体认证防御措施。

参考文献

[1] D. M. Ballesteros、Y. Rodriguez-Ortega、D. Renza 和 G. Arce，“Deep4snet：用于伪造语音分类的深度学习”，*Expert Systems with Applications*，第184卷，第115465页，2021年。 [在线]。可访问：https://doi.org/10 1016 2021 115465 。/j.eswa/。D. Nici。（二月）[2] 2024. 美国联邦通信委员会禁止在自动拨号电话中使用人工智能伪造语音。访问日期：2025年5月13日。 [在线]。可访问：https://www.forbes.com/advisor/personal-finance/fcc-bans-ai-voices-in-robocalls/ [3] 《纽约时报》。（2024年4月）欢迎来到诈骗世界。访问日期：2025年5月13日。 [在线]。可访问：https://www.nytimes.com/2024/04/21 /style/scams-identity-theft.html [4] P. Semansky。（2024年1月）拜登伪造自动拨号电话。访问日期：2025年3月27日。 [在线]。可访问：https://www.nbcnews.com/tech/misinformation/joe-biden-newhampshire-robocall-fake-voice-deep-ai-primary-rcna135 120 [5] A. Mitchell。（2024年10月）车祸人工智能诈骗。访问日期：2025年5月13日。[在线]。可访问：https://nypost.com/2024/10/08/tech/ [6] E. Jamdar 和 A. K. Belman，“Syntheticpop：利用合成语音弹出提示攻击说话人验证系统”，https://arxiv.org/abs/2502.09553，2025年2月，arXiv预印本 arXiv:2502.09553。[7] S. Angius 和 R. Staff，“意大利警方冻结人工智能语音诈骗所得资金”，路透社，2025年2月，访问日期：2025年11月13日。 [在线]。可访问：https: //www.reuters.com/technology/artificial-intelligence/italian-police-freeze-cash-ai-voice-scam-that-targeted-business-leaders-2025-02-12/ [8] OpenAI，“应对合成语音带来的挑战与机遇”，2024年，访问日期：2025年11月13日。 [在线]。可访问：https: //openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/ [9] 美国联邦贸易委员会，“美国联邦贸易委员会就美国联邦通信委员会保护消费者免受人工智能潜在有害影响的工作提交意见”，2024年，访问日期：2025年11月13日。 [在线]。可访问：https://www.ftc.gov/news-events/news/press-releases/2024/07/ftc-submits-comment-fcc-work-protect-consumers-potential-harmful-effects-ai [10] Z. Yu、S. Zhai 和 N. Zhang，“AntiFake：利用对抗音频防止未经授权的语音合成”，载于*2023*年 *ACM SIGSAC* 计算机与通信安全会议论文集，CCS ’23系列。美国纽约州纽约市：计算机协会，2023年，第460–474页。 [在线]。可访问：https: //doi.org/10 1145 3576915 3623209 。/。Z. Zhang、D. Wang、Q. Yang、P. [11] Huang、J. Pu、Y. Cao、K. Ye、J. Hao 和 Y. Yang，“SafeSpeech：针对恶意语音合成的稳健通用语音保护”，2025年。 [在线]。可访问：https://arxiv.org/abs/2504 09839 。Y. Tabet 和 M. Boughazi，“语音合成技术：综 [12] 述”，国际系统、信号处理及其应用研讨会，*WOSSPA*，第67–70页，2011年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:18113403 [13] B. Boashash，信号处理及其应用，阿尔及利亚蒂帕萨，9，2011年。

[14]Z. Almutairi 和 H. Elgibreen，“现代音频深度伪造检测方法综述：挑战与未来方向”，Algorithms，第15卷，第5期，2022年。 [在线]。可访问：https: // www.mdpi.com/ 1999 4893 15 5 155 - / / / M. Mcuba、A. Singh、R. A. Ikuesan 和 H. Venter，“深度学习方法对数字调查中音频深度伪造检测的影响”，Procedia Computer Science，第219卷，第211–219页，2023年，cENTERIS——企业信息系统国际会议 / ProjMAN——项目管理国际会议 / HCist——健康与社会护理信息系统及技术国际会议，2022年。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S 1877050923002910 [15] [16] J. Yamagishi、C. Veaux 和 K. MacDonald，“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（0.92版）”，2019年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 213060286[17] fishaudio，“Bert-vits2”，https://github.com/fishaudio/Bert-VITS2，2024年。[18] M. Kawamura、Y. Shirahata、R. Yamamoto 和 K. Tachibana，“采用多频段生成和逆短时傅里叶变换的轻量级高保真端到端文本转语音”，载于ICASSP 2023——IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2023年。[19] J. Kim、J. Kong 和 J. Son，“采用对抗学习的端到端文本转语音条件变分自编码器”，载于第38届国际机器学习会议论文集（ICML）。PMLR，2021年。[20] F. Fang、X. Wang、J. Yamagishi、I. Echizen、M. Todisco、N. Evans 和 J.-F. Bonastre，“利用x-vector和神经波形模型实现说话人匿名化”，载于第10届ISCA语音合成研讨会论文集（SSW 10），2019年，第155–160页。[21] J. Huang、C. Zhang、Y. Ren、Z. Jiang、Z. Ye、J. Liu、J. He、X. Yin 和 Z. Zhao，“MulliVC：基于循环一致性的多语言语音转换”，arXiv预印本，2024年。[22] I.-C. Yoo、K. Lee、S. Leem、H. Oh、B. Ko 和 D. Yook，“利用语音转换技术保护个人信息的说话人匿名化”，IEEE Access，第8卷，第198 637–198 645页，2020年。[23] A. Hamza、A. R. R. Javed、F. Iqbal、N. Kryvinska、A. S. Almadhor、Z. Jalil 和 R. Borghol，“利用MFCC特征和机器学习检测音频深度伪造”，IEEE Access，第10卷，134 028，2022年，134018页。–。[24] E. Conti、D. Salvi、C. Borrelli、B. Hosler、P. Bestagini、F. Antonacci、A. Sarti、M. C. Stamm 和 S. Tubaro，“通过情感识别检测深度伪造语音：一种语义方法”，载于IEEE声学、语音与信号处理国际会议论文集（ICASSP）。IEEE，2022年，第8962–8966页。[25] L. Blue、K. Warren、H. Abdullah、C. Gibson、L. Vargas、J. O’Dell、K. Butler 和 P. Traynor，“你是谁（我真的很想知道）？ 通过声道重建检测音频深度伪造”，载于第31届USENIX安全研讨会（USENIX Security 22）。USENIX协会，2022年，第2691–2708页。[26] Y. Wang、R. J. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio等，“Tacotron：迈向端到端语音合成”，arXiv预印本，2017年。[27] Z. Yu、S. Zhai 和 N. Zhang，“AntiFake：利用对抗音频防止未经授权的语音合成”，载于2023年 ACM SIGSAC 计算机与通信安全会议论文集（CCS）。ACM，2023年，第460–474页。[28] C. yu Huang、Y. Y. Lin、H. yi Lee 和 L. shan Lee，“保护你的声音：针对语音转换的对抗攻击”，载于2021年IEEE口语语言技术研讨会（SLT）。IEEE，2021年，第552–559页。[29] Z. Zhang、Q. Yang、D. Wang、P. Huang、Y. Cao、K. Ye 和 J. Hao，“缓解未经授权的语音合成以实现语音保护”，载于首届ACM大型人工智能系统与模型隐私及安全分析研讨会论文集，CCS ’24系列。ACM，2023年11月，第 13 24 [ ] 10 1145 3689217 3690615 –。在线。[30] 可访问：http://dx.doi.org/。/。S. Dong、B. Chen、K. Ma 和 G. Zhao，“通过生成对抗网络主动防御语音转换”，IEEE信号处理快报，第31卷，第706–710页，2024年。[31] Z. Kong、W. Ping、J. Huang、K. Zhao 和 B. Catanzaro，“DiffWave：一种多功能音频合成扩散模型”，2021年。 [在线]。可访问：https://arxiv.org/abs/ 2009 09761 。N. Chen、Y. Zhang、H. Zen、R. J. Weiss、M. Norouzi 和 W. Chan，“WaveGrad：用于波形生成的梯度估计”，2020年。[在线]。可访问：https://arxiv.org/abs/2009.00713 [32] [33] S. Wu、J. Wang、W. Ping、W. Nie 和 C. Xiao，《通过扩散模型防御对抗音频》，2023年。 [在线]。可访问：https://arxiv.org/abs/ 2303 01507 。J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，ArXiv，第 abs/2006.11239 卷，2020年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 219955663 [34] [35] W. Nie、B. Guo、Y. Huang、C. Xiao、A. Vahdat 和 A. Anandkumar，《用于对抗净化的扩散模型》，载于国际机器学习大会，2022年。 [在线]。可访问：https: // api.semanticscholar.org/CorpusID: 248811081 [36] H. Guo、G. Wang、B. Chen、Y. Wang、X. Zhang、X. Chen、Q. Yan 和 L. Xiao，《Wavepurifier：通过层次化扩散模型净化音频对抗样本》，第30届移动计算与网络国际年会论文集，2024年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 274090145[37]H. Tan、X. Liu、H. Zhang、J. Zhang、Y. Qian 和 Z. Gu，《Dualpure：一种用于语音命令识别的高效对抗净化方法》，载于 INTERSPEECH，2024年。 [在线]。可访问：https://doi.org/ 10 21437 2024 855 。/Interspeech。- [38] W. Fan、K. Chen、C. Liu、W. Zhang 和 N. H. Yu，《De-antifake：重新思考针对语音克隆攻击的保护性扰动》，ArXiv，第 abs/2507.02606 卷，2025年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 280129554 [39] A. Sokol，《奥恩斯坦-乌伦贝克随机微分方程中的干预》，2013年。 [在线]。可访问：https://arxiv.org/abs/ 1308 2152 。X. Li、W. Sun、H. Chen、Q. Li、Y. He、J. Shi 和 X. Hu，《ADBM：用于可靠对抗……的对抗扩散桥模型》，2024年10月。[在线]。可访问：https://openreview.net/forum?id=g0rnZeBguq&nes ting=2&sort=date-desc [41] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，《迈向能够抵抗对抗攻击的深度学习模型》，ArXiv，第 abs/1706.06083 卷，2017年。 [在线]。可访问：https://api.semanticsc holar.org/CorpusID: 3488815 [42] J. Wang、Z. Lyu、D. Lin、B. Dai 和 H. Fu，《用于对抗净化的引导扩散模型》，ArXiv，第 abs/2205.14969 卷，2022年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID: 249192338 [43] ++ B. Zhang、W. Luo 和 Z. Zhang，《Purify：利用先进扩散模型和随机性控制改进扩散净化》，ArXiv，第 abs/2310.18762 卷，2023年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:264306456 [44] ——，《通过基于分数的优化增强对抗鲁棒性》，ArXiv，第 abs/2307.04333 卷，2023年。 [在线]。可访问：https://api.sema nticscholar.org/CorpusID: 259501507 [45] ` N. Carlini、F. Tramer、K. D. Dvijotham 和 J. Z. Kolter，《（认证的！！）免费获得对抗鲁棒性！》，ArXiv，第 abs/2206.10550 卷，2022年。[在线]。可访问：https://api.semanticscholar.org/CorpusID:249889641 [46] C. Xiao、Z. Chen、K. Jin、J. Wang、W. Nie、M. Liu、A. Anandkumar、B. Li 和 D. X. Song，《Densepure：理解扩散模型以实现对抗鲁棒性》，ArXiv，第 abs/2211 00322，卷。。在线。可访问：2022 [ ] https://api.semanticscholar.org/CorpusID: 253244468[47] J. M. Cohen、E. Rosenfeld 和 J. Z. Kolter，《通过随机平滑实现认证的对抗鲁棒性》，ArXiv，第 abs/1902 02918，卷。。在线。可访问：2019 [ ] https://api.semanticscholar.org/CorpusID: 59842968[48] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《LibriSpeech：一个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），2015年，第5206–5210页。[49] J. Yamagishi、C. Veaux 和 K. MacDonald，《CSTR VCTK语料库：用于 CSTR语音克隆工具包的英语多说话人语料库（版本0.92）》，2019年。[50] C.-y. Huang、Y. Lin、H.-y. Lee 和 L.-S. Lee，《保护你的声音：针对语音转换的对抗攻击》，2021年1月，第552–559页。[51] C. Xiao、B. Li、J.-Y. Zhu、W. He、M. Liu 和 D. Song，《使用对抗网络生成对抗样本》，2019年。 [在线]。可访问：https://arxiv.org/abs/ 1801 02610 。《GitHub - Plachtaa/VALL-E-X：微软 VALL-E X 零样本文本转语音模型的开源实现。演示地址：https://plachtaa.github.io/vallex/ — github.com》，https://github.com/Plachtaa/VALL-E-X， [访问日期：14-11-2025]。 [52]

[53] 《GitHub - neonbjb/tortoise-tts：一个注重质量训练的多说话人文本转语音系统 — github.com》，https://github.com/neonbjb/tor toise-tts，[访问日期：14-11-2025]。
[54] Y. A. Li、C. Han、V. S. Raghavan、G. Mischler 和 N. Mesgarani，《StyleTTS 2：通过风格扩散、对抗训练和大型语音语言模型实现接近人类水平的文本转语音》，2023年。[ ] 2306 07691 在线。可访问：
[55] https://arxiv.org/abs/。D. Wang、L. Deng、Y. T. Yeung、X. Chen、X. Liu 和 H. Meng，《VQMIVC：基于向量量化和互信息的无监督语音表示解耦，用于单样本语音转换》，2021 [ ] 2106 10132 。在线。可访问：
[56] https://arxiv.org/abs/。S.-H. Lee、H.-Y. Choi、S.-B. Kim 和 S.-W. Lee++，《HierSpeech：通过分层变分推断弥合语音语义表征与声学表征之间的差距，实现零样本语音合成》，2023 [ ] 2311 12454 。在线。可访问：
[57] https://arxiv.org/abs/。H.-Y. Choi、S.-H. Lee 和 S.-W. Lee，《Diff-HierVC：具有稳健音高生成和掩码先验的基于扩散的分层语音转换，用于零样本说话人适应》，*ArXiv*，第 abs/ 2311 04693，卷。。在线。可访问：2023 [ ] https://api.semanticscholar.org/CorpusID: 260920394
[58] M. Ravanelli、T. Parcollet、P. W. V. Plantinga、A. Rouhe、S. Cornell、L. Lugosch、C. Subakan、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》，*ArXiv*，第 abs/ 2106 04624，卷。。在线。可访问：2021 [ ] https://api.semanticscholar.org/CorpusID: 235377273
[59] 《GitHub - resemble-ai/Resemblyzer：用于通过深度学习分析和比较声音的 Python软件包 — github.com》，https://github.com/r esemble-ai/Resemblyzer，[访问日期：14-11-2025]。[60] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，《NISQA：一种利用众包数据集进行多维语音质量预测的深度卷积神经网络自注意力模型》，2021年8月。[在线]。可访问：http://dx.doi.org/ 10 21437 2021 299 。/ Interspeech。
[61] - Z. Jiang、J. Liu、Y. Ren、J. He、Z. Ye、S. Ji、Q. Yang、C. Zhang、P. Wei、C. Wang 等，《Mega-TTS 2：通过提示机制增强零样本语音合成》，*arXiv*预印本 *arXiv:2307.07218*，2023年。