研究

通过集成方法实现可迁移的音频对抗攻击

郭峰[1,2], 孙正[1,2],陈宇轩[1], 2* 和鞠磊[1,2]

arXiv:2304.08811v1 [cs.CR] 18 Apr 2023

摘要

近年来，深度学习（DL）模型在自动驾驶、人脸识别和语音识别等多个领域取得了显著进展。然而，深度学习模型易受对抗攻击影响，这引发了学界的严重担忧，因为这些模型的鲁棒性和泛化能力不足。此外，可迁移攻击已成为黑盒攻击的重要方法。在本研究中，我们探讨了影响基于深度学习的语音识别中对抗样本（AEs）可迁移性的潜在因素。我们还讨论了不同深度学习系统的脆弱性以及决策边界的不规则性。结果显示，语音与图像中的对抗样本可迁移性存在显著差异：图像中的数据相关性较低，而语音识别中的情况恰恰相反。受基于丢弃法的集成方法启发，我们提出了随机梯度集成和动态梯度加权集成，并评估了集成方法对对抗样本可迁移性的影响。结果表明，这两种方法生成的对抗样本均可有效迁移至黑盒应用程序接口。

关键词： 对抗攻击；动态梯度加权；可迁移性；集成方法

引言

近期研究 [1–6] 表明，基于深度学习的任务（如图像识别和语音识别）容易受到对抗攻击的影响。Szegedy 等人 [7]首次提出了对抗攻击和对抗样本（AEs）的概念。对抗样本是通过有意向输入中注入精心构造的细微扰动而生成的。这些对抗样本可能会阻碍深度神经网络（DNNs）在面部识别和智能家居等领域的发展。

根据攻击者对目标模型的了解程度，对抗攻击可分为白盒攻击和黑盒攻击。在白盒攻击中，攻击者可以访问目标模型的详细信息，包括其结构、参数，甚至训练数据集，因此能够利用反向传播等技术构造对抗样本。另一方面，在黑盒攻击中，攻击者只能向目标模型发送输入数据并接收预测结果，而无法了解模型的内部运行机制。这类攻击尤其具有挑战性，也更贴近现实，特别是针对商业 API 模型。因此，近年来黑盒攻击受到了广泛关注。

尽管一些深度学习任务在训练数据集上的准确率可达 99.99%，但模型鲁棒性和泛化能力较差，可能会严重影响其在现实世界中的精度。攻击者已利用这些漏洞，对选定的黑盒模型发起攻击。学术界和企业界的研究人员开展的多项研究表明，针对特定深度学习（DL）模型设计的对抗样本，也可以攻击其他深度学习模型，这种现象称为“迁移攻击”。这意味着，攻击者可以基于一些公开可用的深度学习模型构建替代目标模型，再根据目标模型的查询反馈，定制对抗样本的扰动和生成算法，以攻击目标模型。这种方法已广泛用于攻击黑盒模型（例如商业黑盒应用程序接口）。许多研究 [5, 8–11] 提出了不同的算法，以提升迁移攻击能力，并提供了相关实验依据。

大多数关于迁移攻击的研究都聚焦于图像分类。语音识别模型的鲁棒性和泛化能力尚未得到充分关注。一些研究 [12–15] 在攻击黑盒 API 方面取得了积极成果，但

*通信作者： chenyuxuan@sdu.edu.cn[1]山东大学网络空间科学与技术学院，中国青岛[2]泉城实验室，QCL，中国济南。完整的作者信息列表见文章末尾

关于语音识别中对抗样本可迁移性的关键影响因素，目前证据有限。理解语音迁移攻击的方法，主要是与图像相关研究进行对比。然而，语音识别比图像识别更为复杂，许多适用于图像、却不适用于语音的技术差异都值得关注。针对语音的对抗攻击具有许多独特属性。

在本研究中，我们深入分析了影响对抗样本（AEs）可迁移性的潜在因素。具体而言，我们开展了大量实验研究和数据分析，探究噪声、尺度不变性及其他影响对抗样本可迁移性的因素，并对观察到的现象给出了可能的解释。此外，我们尝试描述深度学习模型的几何性质，并揭示了模型决策边界的波动性。

此外，我们发现，应用 dropout 自集成技术能够提升音频对抗样本的可迁移性。基于这一发现，我们提出了两种通过组合多个模型来生成可迁移对抗样本的策略，即随机梯度集成和动态梯度加权集成。这些方法旨在优化对抗样本在不同模型间的可迁移性，从而能够生成可在多种系统间迁移的对抗样本。

贡献。本文的贡献如下：

•我们观察到，图像与语音中的对抗样本在可迁移性方面存在显著差异。图像中像素之间的关联较弱，而数据上下文关联对于语音识别至关重要。我们首次尝试刻画语音深度学习模型的几何特性。我们认为，语音识别模型的决策边界不像图像模型中的那样平滑连续，而是呈现不规则且不稳定的状态。

•我们的实验揭示了噪声、静音帧、尺度不变性、丢弃法及其他因素对对抗样本的显著影响，并表明不同因素对迁移率和完美性的贡献程度各不相同。

•受基于丢弃法的集成模型取得的成功启发，我们提出了随机梯度集成和动态梯度加权集成，用于生成对抗样本。实验表明，采用这两种策略生成的对抗样本均可应用于黑盒应用程序接口（API）。此外，我们发现，p值为 0.5 时，对抗样本的生成和可迁移性达到最佳。

•我们在此发布本研究的源代码：xxx

背景

在此，我们将介绍有关对抗攻击和提升可迁移性的相关研究，并分析总结有关可迁移性的基本结论。

对抗攻击与对抗样本（AEs）

本文聚焦于对抗任务。在这种设定下，深度神经网络表示为 f，且f： $\boldsymbol{X} \rightarrow \boldsymbol{C}$ 表示给定输入 $x(x \in \boldsymbol{X})$ 被映射到类别集合 $\boldsymbol{C}$中的一个类别，其中 $f(x) = c \in \boldsymbol{C}$。深度神经网络模型容易受到对抗输入攻击，从而导致模型作出错误判断。针对深度神经网络的攻击可分为定向攻击和非定向攻击。本文将重点讨论定向攻击的情形。具体而言，对抗样本 x^* 通常通过对 x 和 $x^* = x+\delta$进行轻微修改来生成。求解 δ 可转化为一个最小化优化问题，即 $arg\ min\ \mathcal{L}(f(x+\delta)\ c^*)$。攻击者的目标是迫使 f 将 x^* 误分类为目标 c^*，即$f(x^*) = c^*$, $c^* \neq c$。为确保 x^* 在听觉上与 x相似，扰动需要限制在有限范围 $g(x^* - x) \leq \varepsilon$内，其中 g 是衡量听觉差异的测量函数。

在音频对抗攻击领域，Carlini & Wagner[8]于2018年首次成功攻击了 Deepspeech 模型。当时，Deepspeech 是最受欢迎的开源、基于神经网络的端到端语音识别模型。同年，Yuan 等人 [10]也成功攻击了 Kaldi ASpIRE Chain 模型[1]。Kaldi 是一种语音识别系统，微软、小米等公司广泛应用了该系统。次年，Qin 等人[16] 使用听觉掩蔽模型攻击 Lingvo[2] ，并生成了更稳健的对抗样本，这些样本可以在现实世界中回放。Lingvo 是谷歌开发的一种语音识别系统，采用深度学习识别语音并将其转换为文本，具有很高的精度和计算效率。Chen 等人[10] 通过本地替代模型，成功攻击了 Facebook、Bing、谷歌、AWS 等商业实体的应用程序接口。Zheng 和 Jiang [14] 通过协同进化算法，成功攻击了科大讯飞、阿里巴巴和腾讯等公司的语音识别应用程序接口。这些攻击针对商业 API 的成功率接近 100%，并取得了令人振奋的结果。

[1]Kaldi。http://kaldi-asr.org。[2] https://github.com/tensorflow/lingvo。

此后，针对开源深度学习模型的攻击研究几乎停滞不前。相反，研究人员将注意力转向了针对商业语音识别 API 的攻击。他们怀疑这些商业 API 可能采用深度学习模型，因此可能容易受到与开源深度学习模型相同的漏洞影响。

对抗样本的可迁移性

对抗样本的可迁移性是一项重要属性，指一个对抗样本能够攻击不同的模型。换言之，对抗样本并非只能欺骗某个特定模型，也能欺骗其他模型。从几何角度来看，论文 [17] 表明，不同模型之间的决策边界非常相似，这解释了对抗样本为何能够在不同模型之间迁移。该研究定量分析了基于优化的攻击与基于梯度的攻击在可迁移性方面的差异，结果表明，基于梯度的攻击具有更高的可迁移性。受数据增强策略的启发，论文 [8]提出通过输入多样性来提高对抗攻击的可迁移性。这种方法降低了对抗样本过拟合的风险，后续研究还提出了不同的输入变换，以进一步缓解过拟合，例如 [9, 10, 16]。一些人认为，深度神经网络最后一层的决策边界越线性，对抗样本的可迁移性就越好，[18]也提出了这一观点。论文 [19] 提出扰动特征空间而非输入以提高可迁移性，并采用注意力机制实现这一目标。最后，[11] 提出了对抗分布增强方法，通过调整对抗样本的分布来提高可迁移性。论文指出，对抗样本的可迁移性受多种因素影响，并提出了有效的增强方法。未来研究应致力于开发更稳健、泛化能力更强的方法，以防御对抗攻击。

另一项研究中，[20] 分析了不同模型生成的对抗样本在安静和嘈杂环境下的可迁移性。他们发现可迁移性与加性噪声之间可能存在关联，并提出了一种在梯度上升过程中注入噪声以增强可迁移性的方法。另一方面，[21] 提出了一种算法，用于生成具有高可迁移性的通用对抗扰动。该扰动通过梯度信息进行更新。在一项平行研究中，[22] 提出了 Houdini 攻击方法，该方法利用了人类难以感知对抗样本这一特点。结果表明，Houdini 具有良好的可迁移性和隐蔽性，甚至能够成功欺骗未知模型。

语音信号高度非线性且复杂，因此在对抗性语音识别攻击中，其特征具有很高的可变性和不确定性。同时，语音信号的传输环境非常复杂，涉及噪声、室内声学、说话人变化等因素。因此，攻击者可以利用这些特征生成具有高可迁移性的对抗样本，成功欺骗多个模型。

研究指出，对抗样本并非随机出现在小空间中，而是存在于大范围的连续子空间中。这些子空间的维度是影响可迁移性的关键因素，因为维度越高，不同模型的子空间越有可能重叠。有证据表明，这些扰动会影响不同模型用于做出决策的特定特征。因此，对抗样本可能在不同模型之间迁移。此外，结构、参数权重或特征提取方法相似的模型可能具有共同点，从而促进对抗样本在它们之间迁移。同样，如果源模型和目标模型的结构与参数相似，那么在这两个模型之间迁移对抗样本可能会更容易，相关研究也展示了这一点。

可迁移性的性质

基于上述分析及相关研究 [1–11, 16–19, 23]，我们总结了可迁移性的一些性质。

给定网络结构不同但使用相同训练数据集的模型 A、B 和 C，且参数数量满足关系 $A \gg B \approx C$，可以观察到以下性质：

•对抗样本的可迁移性在很大程度上取决于模型架构。具体来说，在模型 A 上生成的对抗样本，其可迁移性不如在模型 B 上生成的对抗样本。虽然规模更大、层数更深的模型往往性能更好，但其迁移能力也可能较弱。较小的模型往往具有更好的迁移能力，也较不容易过拟合。我们认为，降低对抗扰动对特定模型的拟合程度有利于提升可迁移性。

•当模型 B 和模型 C 的分类准确率较高但鲁棒性较低时，它们的决策边界通常较为接近。这意味着这些模型对输入会做出相似的分类决策。在这种情况下，对抗样本更容易从模型 B 迁移到模型 C，因为这些样本中的扰动方向通常沿着决策边界的切线方向。我们认为，如果模型 B 和模型 C 的决策边界相似，那么这些对抗样本对二者的影响也会相同。

•对抗样本可以在不同模型之间迁移，因为它们共享相似的输入空间，也就是说，它们可以

处理相似的输入。然而，由于决策边界具有波动性，对抗样本在模型之间的可迁移性可能有所不同。随着不同模型的决策边界发生变化，对抗样本的可迁移性可能会受到影响。因此，我们认为，理解不同模型的决策边界及其与对抗样本可迁移性的关系至关重要。

•在深度学习模型中，中间层对于提升鲁棒性和泛化能力至关重要。对抗样本与良性样本在深度神经网络（DNN）的中间层中表现出显著差异，而中间层特征的这些差异对于判断输入是否具有恶意至关重要。我们认为，利用中间层特征的差异，或许能够更好地区分对抗样本与良性样本，最终提升对抗样本的鲁棒性和可迁移性。

音频对抗样本的可迁移性特性I

n th音频对抗样本的可迁移性，e 并探索语音识别模型决策边界的几何性质。

影响可迁移性的潜在因素

在我们的研究中，我们获取了用于攻击 Deep-Speech[3] 和 Lingvo[4] 的代码，并复现了作者报告的类似结果。此外，我们联系了 ASpIRE 模型的作者，以获取攻击 Kaldi[15] 的代码，并获得了一些代码和样本。基于此，我们对语音识别中的对抗攻击进行了进一步研究，并观察到以下特征：

1) 我们在 Kaldi ASpIRE 模型上使用丢弃法的自集成技术来生成对抗样本。在生成过程中，我们每迭代 10 次就在各应用程序接口上查询一次解码结果，并发现查询约 80 次后，样本便可在百度语音识别 API 上解码为目标命令 [5]，阿里云语音识别 API[6]，和讯飞 API [7]。结果见表1。这表明，自集成对抗攻击是一种很有前景的方法。我们进一步研究了一种更全面的集成对抗攻击，通过结合多个基础模型来生成迁移性更强的对抗样本。

[3]https://github.com/carlini/音频对抗样本.[4] https://github.com/cleverhans-lab/cleverhans/tree / master/cleverhans˙v3.1.0/examples/对抗自动语音识别.[5] cloud.baidu.com[6]cn.aliyun.com[7]www.xfyun.cn

对抗样本的生成可以视为一个训练过程，且训练时只有一个输入数据集。采用丢弃法时，训练过程中每个神经元的输出都以概率 p 置为 0。为了保持训练和测试期间输出的一致性，通常需要将输出除以 $1-p$。但对抗样本的训练不涉及测试，因此可以将这一步简化为公式 1：

$$\begin{cases} m \sim Bernouli(p) \\ \tilde{x} = m \odot x \end{cases} \tag{1}$$

~其中 $\odot$ 表示乘法运算， x表示输出；我们在每次迭代之前对数据执行等式 1 运算。

2）在对抗攻击中，添加噪声是提高对抗样本鲁棒性的常用技术，使其能够在各种设置下成功攻击目标模型。多篇论文 [24–27] 证明了噪声在提升模型或对抗样本鲁棒性方面的有效性。噪声具有微妙的作用，包括防止训练过程中过拟合、生成高度鲁棒且抗扰动的样本，以及用于攻击和防御。在实验中，我们持续提高噪声水平，在生成对抗样本的每次迭代中向数据添加随机噪声，然后根据梯度进行优化。我们发现，随着噪声水平上升，对抗样本的表现显著改善。尽管迭代次数和耗时有所增加，但与未添加噪声的样本相比，经过噪声优化的样本更容易发起攻击，也更加稳定。噪声水平越高，生成对抗样本就越困难，但一旦生成成功，其对抗攻击效果就会强得多。

表1：API 服务上的迁移攻击结果。

噪声	百度 API	科大讯飞 API	阿里巴巴 API
±0	2/20	1/20	2/20
±4000	10/20	8/20	9/20
±8000	9/20	5/20	8/20
±12500	0/20	0/20	0/20
±17500	4/20	2/20	5/20
±20000	12/20	12/20	11/20
±25000	2/20	0/20	1/20
丢弃法 (p=0.5)	13/20	12/20	11/20
尺度不变 (m=4)	14/20	13/20	12/20

注： “±N”表示噪声范围为 $[-N, N]$。攻击成功率“ A/B”表示有 A 个对抗样本（共 B 个）在黑盒平台上触发了该命令。

在我们的实验测试中，我们将目标命令 as“关灯”并使用公式2 生成均匀分布的随机噪声，重复实验——

20 次（使用 20 个输入样本）。由此得到 20 个对抗样本。在生成每个对抗样本的过程中，我们每迭代 10 次就查询一次应用程序接口，总计查询 200 次。如果应用程序接口能够将对抗样本解码为目标命令，我们就认为这是一次成功的迁移攻击。如表 1所示，不添加噪声时，对抗样本的可迁移性非常低，几乎为零。我们还评估了样本的感知质量，发现几乎没有可听扰动。随着噪声水平逐渐增加至 20,000，可迁移性先逐渐上升后下降，而样本的感知质量则显著降低。当噪声水平达到 20,000 时，可迁移性最高，攻击效果也最好，但噪声完全覆盖了基础载体。随着噪声水平继续增加，攻击成功率急剧下降，样本的可迁移性也随之降低。

$$f(x) = \begin{cases} \frac{1}{2A}, -A < x < A \\ 0, else \end{cases} \quad (2)$$

语音和图像识别系统对上下文扰动表现出不同的敏感性。语音识别高度依赖上下文线索，细微波动也会造成显著的解码错误；而图像识别对类似变化的鲁棒性更强。这种差异可能源于噪声水平与可迁移性之间的非线性关系。噪声水平初步升高可能会增强对抗样本的韧性，但持续增大则会削弱上下文连贯性，直至信息完全模糊。当噪声水平超过 20,000 这一阈值时，语音信号的上下文结构会被彻底破坏，转而促成对抗性上下文关系的出现。然而，随着噪声水平进一步升高，这些适应不良的关系也越来越难以建立。

3) 通过大量测试，我们发现，尽管采用相同的算法，不同模型对不同对抗样本的敏感性仍各不相同。音频载体的选择会直接影响对抗样本的性能和可迁移性，导致扰动影响和攻击成功率存在差异。有些深度学习模型似乎天生对某些输入比其他输入更敏感。例如，模型往往更容易接受由雪景图像制作的对抗样本

相比狗的图像，风景图像呈现出更明显的效果。对于语音识别，这种效应更加突出，甚至十分显著：听觉扰动和攻击成功率因音频载体不同而存在显著差异。对话语音载体往往更

能抵御攻击，而音乐载体则较弱；与包含多个元素的音乐载体相比，简单的音乐元素更容易受到扰动。为了选择适合生成对抗样本的载体并降低人类对扰动的感知， [3, 28], 和 [25]等研究采用了自动选择和多角度评估等载体选择方法。

4）通过实验测试，我们发现，插入静音帧可以提高目标命令的隐蔽性，并提升对抗样本的攻击成功率。在适当位置插入静音帧会显著影响音频的解码结果，因此可以通过这种方式增强对抗攻击的鲁棒性。

在语音识别中，输入音频必须划分为帧。对于时长约为 4 秒的音频，通常可将其划分为约 500 帧。这意味着，时长为 4 秒的音频有 501 个可能的插入位置。同时，静音帧会干扰整个窗口的解码输出，并使原始上下文脱离关联。静音帧不会引起不适感，只会造成轻微停顿。我们在扰动叠加过程中随机选取了 10 个插入位置来研究这一效应。我们发现，插入静音帧在高频域的影响更为显著（见图1），而在低频域的影响较弱。关于载体对对抗样本的影响，参见《CommanderSong [15]》：通过比较频谱可知，载体模态会影响对抗样本的可行性；某些载体能够实现不易察觉的对抗样本，而另一些则不能。我们发现，插入静音帧更接近会话语音载体音频，因此能够轻易生成不易察觉的对抗样本。这可能是因为静音帧不仅使原始上下文脱离关联，还会引入空白片段，使信号更接近会话语音，从而提高对抗样本的可行性。

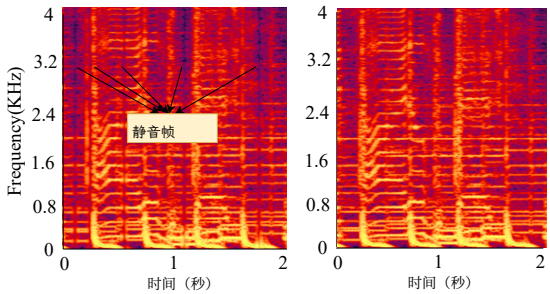

图 1：比较加入静音帧后的频域特征。

5) 尺度不变对抗攻击是一种针对对标量变换不敏感模型的方法。尺度不变对抗攻击基于

生成能够在不同尺度下欺骗模型的特征。 [26] 通过在不同尺度下处理样本来生成尺度可变的对抗样本，从而提高可迁移性。 [27] 提出了尺度不变卷积层，通过保持卷积核的尺度不变性来增强鲁棒性，使模型能够识别不同尺度下具有相似特征的输入。其他研究（例如 [29, 30]）通过应用尺度变换和平移变换，在不同程度上提高对抗鲁棒性。近期研究表明，对尺度变换具有不变性的对抗攻击在破坏机器学习模型方面可能非常有效。尺度不变攻击试图通过在多个尺度上训练对抗样本，利用模型有限的尺度不变性。我们通过公式 3 研究了一种特定的尺度不变对抗攻击方法，并评估其对语音识别模型的有效性。结果表明，这种尺度不变方法的攻击效果与其他对抗攻击方法相当。此外，生成的对抗样本对其他模型表现出很强的可迁移性。对于 $\mathrm{m} = 4$的最优超参数设置，该攻击在欺骗迁移模型上的平均成功率达到 65%。总体而言，我们的结果凸显了尺度不变对抗攻击的潜在影响，也表明需要进一步研究以缓解此类威胁对模型鲁棒性造成的影响。

$$g_t = \frac{1}{m}\sum_{i=1}^{m}\nabla_x l(S_i(x^{adv}), y) \tag{3}$$

其中， $S_i(x)$表示输入 x 的缩放副本，缩放因子为 $1/2^i$ ， m 表示缩放副本的数量， y 为真实标签。

自动语音识别模型的几何性质

简而言之，“迁移攻击”是指在模型 A 中生成的对抗样本能够成功攻击目标模型（应用程序接口）。在我们的实验中，我们分析了音素从良性样本到音频对抗样本的变化，发现某些音素比其他音素更难转变为目标音素，而且不同模型中的变化模式各异。我们推断，语音识别模型之间的决策边界可能不规则且不稳定，这是语音识别迁移攻击难以实现的原因之一。

在图 2中，绿色点表示模型 A 决策边界上方的数据分布点，属于良性样本的数据点，输入 x 可以被正确转录为 y。对抗样本（AEs，图中的红点）由跨越决策边界的良性样本点（绿色点）生成。对抗样本是在模型 A 上生成的，且目标是攻击目标模型。虽然所有绿色点都已跨越模型 A 的决策边界，但部分数据点（q）尚未跨越目标模型的决策边界，另一些则已跨越（s、p）。对于语音识别，通常采用三音素关联解码，如图 2所示，其中三个具有前后关联的数据点构成一个解码集合。对于可迁移攻击者而言，要在模型 A 上生成能够攻击目标模型的对抗样本，就必须使所有数据点都跨越目标模型的决策边界。Qab、pab 和 sab 是三个解码单元；在模型 A 上，pab、qab 和 sab 都可以解码为目标指令，但在目标模型上，只有 pab 和 sab 可以解码为目标指令。pab 与 sab 的主要区别在于 p 和 s 的位置不同，这表明扰动程度有所不同。（扰动越大，位置越低。）

例如，在“end”和“old”这两个词中，“end”的音素为‘IY-N-D’，而“old”的音素为‘UW-I-D’。我们的目标是将“old”转换为“end”。语音识别对抗攻击的本质是使音素跨越目标模型的决策边界，从而实现攻击。音素‘I’和‘D’分别对应图中的‘a’和‘b’点，而‘N’和‘D’则对应‘a’和‘b’上方的绿色点。点 q 在模型 A 中表示音素“UW”，但对于目标模型，只有点“s”和“p”才被视为“UW”。在目标模型上，qab无法解码为“end”，还可能被解码成其他词。即使大多数音素（‘N’和‘D’）已跨越决策边界，也无法确保“end”和“old”得到目标解码结果。这也是迁移攻击难以实现的根本原因。

在图像或音频对抗攻击中，“硬裁剪”是一种将扰动限制在一定范围内的技术。由于听觉扰动的测量方式，如图 2 所示，“s”点和“p”点位置的扰动程度不同，因此难以针对某个特定点（例如 s 点）进行优化，通常优化的是 p 点。硬裁剪是一种粗略的方法，这种裁剪后的扰动可能会导致听感不适，这也是许多音频对抗样本带有噪声的原因。尽管如此，优化到 s 点始终是期望的目标。此外，解码规则的选择也会影响解码结果。

使用基于集成的方法增强对抗样本的可迁移性

根据我们的发现，随机失活策略意外地提高了音频对抗样本的可迁移性。这

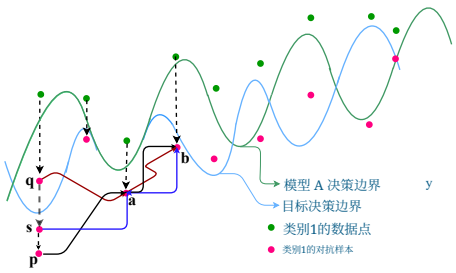

图 2：针对语音识别模型的可迁移攻击。

促使我们进一步考察集成方法对音频对抗样本可迁移性的影响。我们选取 Deepspeech、Kaldi 和 Lingvo 模型作为基础模型，并评估了不同集成策略对可迁移性的影响。

重新思考集成学习

在监督机器学习中，我们希望训练出一个稳定且各方面表现良好的模型。实际上，这一目标往往难以实现，我们可能只能得到几个表现良好但仍存在偏差的模型。集成学习是一种将多个弱监督模型结合起来，从而获得更准确、更稳健的整体预测结果的技术。通过组合这些模型，集成模型可以纠正各个模型中的错误或偏差。

集成学习有三种方法：装袋法、提升法和堆叠法。装袋法通过有放回随机抽样，反复从训练集中抽取不同的子集，用于训练一组基础模型。最后，使用投票法等方法组合这些基线模型的预测结果，以降低方差。提升法通过关注先前基础模型错误分类的样本来训练下一个基学习器。随后对这些结果加权并取平均，生成最终模型。堆叠法使用训练集训练多个基础模型，并获取这些基础模型的输出结果。然后将这些输出结果作为新的训练集，用于训练一个模型。

同样，集成学习在攻击深度学习模型，尤其是黑盒模型方面具有潜在应用价值。我们可以结合一组基础模型来构造稳健的对抗样本，从而有效攻击黑盒或灰盒模型。随后，使用特定方法训练这些样本，以获得嵌入式对抗样本。利用从不同模型收集的梯度信息，可以基于指定的优化方法生成对抗样本，以攻击目标模型。预计不同的集成方法对梯度信息会产生不同影响。

集成攻击

不同的集成策略对攻击结果有不同影响。以下介绍一些常见的集成攻击方法。 [17] 首先针对 k 个基础模型实现了多模型集成方法，以获得损失 $J_i(x$（经过 Softmax 后）以及一组权重 α_i 以$_k$集成优化过程 $\sum \alpha_i J_i(x)$。这种方法$_{i=1}$很快便成为基准方法。此外，研究者将各种模型用作集成的基础模型，也有人探索自集成策略。其中一种策略由 Yingwei Li 等人提出，[31] 采用幽灵网络，将丢弃法、跳跃连接及其他操作结合起来，生成多个候选模型，再对这些模型进行集成以生成对抗样本。类似地，Dongx-ian Wu 等人 [32] 提出了跳跃梯度法，该方法采用带有跳跃连接的网络，为浅层网络赋予更高权重，并尝试生成迁移率更强的对抗样本。另一种提出的方法是长期梯度记忆集成对抗攻击 [33], ，它基于两个假设：一是模型迁移等同于网络泛化；二是边界相似性比扰动限度更重要。SVRE-MI-FGSM[34] 也试图通过降低模型间的方差来提升对抗样本的可迁移性，最终获得平滑的决策边界曲线，从而提高可迁移性。文献 [35], 分析了 ImageNet 的几何特性，发现不同模型的梯度方向彼此正交，而各模型的决策边界相对相似，这有助于非目标攻击的迁移率。随后，文献 [36] 提出了两种基于集成的黑盒攻击策略，即选择性级联集成策略和堆叠并行集成策略，以便对深度学习模型实施更强大的黑盒攻击。集成中替代模型的多样性和数量是影响对抗样本迁移率的两个重要因素，这对于选择有效的集成替代模型至关重要。

针对单个模型的对抗攻击，只需跨越该特定模型的决策边界即可。然而，不同模型的决策边界之间存在一定差异。如上所示，语音识别模型的决策边界呈现出尖锐、波动且不规则的形态。不同模型跨越分类边界的能力并不相同，而集成方法有助于克服这一不足。如图3所示，由于决策边界存在差异，基于单个模型 1 生成的对抗样本可能无法成功迁移到其他模型。基于其他模型生成对抗样本时，通常也会遇到同样的问题。

如果将这些基础模型连接起来，就有可能获得能够跨越更多模型决策边界的对抗样本，从而形成更强的攻击能力。

可以观察到，基于集成的策略生成的对抗样本（红点）穿过了替代模型的 k 决策边界，这表明其具有更强的泛化能力，也更有可能成功欺骗目标黑盒模型应用程序接口，即穿过目标模型的决策边界。

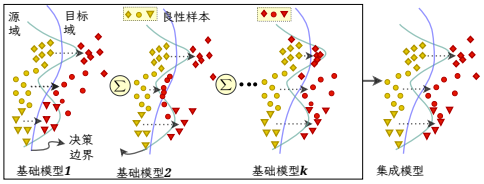

图 3：针对语音识别模型的可迁移攻击。黄色图标表示良性样本，红色图标表示对抗样本。

基于集成的黑盒 API 攻击策略

0.0.1 模型与攻击方法

如前所述，自集成丢弃技术在一定程度上提升了对抗样本的可迁移性。尽管源模型与目标模型之间存在不规则且不稳定的决策边界，对抗样本的迁移仍然已被证明颇具挑战。然而，集成方法可以继续逼近重叠子空间的最佳区域以及“pab”攻击路径。尽管如此，值得注意的是，自集成丢弃技术可能会使“sab”攻击路径收敛，从而轻微损害听觉感知。针对语音识别模型的对抗攻击比针对图像的攻击更为复杂，相关技术也可能存在显著差异。然而，尚无人尝试通过组合不同的语音识别模型来生成对抗样本。这部分是由于开源语音识别模型数量有限，也因为这类模型更为复杂，生成对抗样本需要耗费更多时间和资源。此外，语音识别模型决策边界的非光滑性也给实现预期的可迁移性带来了挑战。为进一步研究集成对对抗样本迁移的影响，我们考察了 Deepspeech、Kaldi 和 Lingvo 模型，并设计了一种组合算法，实验细节将在下文介绍。目前，Deepspeech、Kaldi 和 Lingvo 模型是最受欢迎的基于神经网络的语音识别模型，在工业界和学术界都得到了广泛认可。

在本研究中，我们选择这三个模型作为集成研究的基础模型。

为了攻击 Deepspeech 模型，我们将 Deepspeech 模型的 CTC 损失用作损失函数，并通过 Adam 对其进行优化。获取对抗样本的方法如公式 4。

$$x_t^{adv} = x_{t-1}^{adv} + clip_x^{\varepsilon}(Adam(\nabla_x loss_{ctc}(x_{t-1}^{adv}, y))) \quad (4)$$

可以通过计算输入样本 x 与目标 y 之间的 pdf-id损失函数，并使用 Adam优化器执行梯度下降来生成对抗样本，以攻击 Kaldi的Aspire模型。同时还会对扰动的最大值进行约束。具体实现如以下公式所示：5。

$$x_t^{adv} = x_{t-1}^{adv} + clip_x^{\varepsilon}(Adam(\nabla_x loss_{\mathrm{pdf-ids}}(x_{t-1}^{adv}, y))) \quad (5)$$

针对 Lingvo模型的对抗攻击是一种两步攻击算法。第一阶段使用 Adam优化器和梯度下降，根据梯度生成能够欺骗 Lingvo模型的对抗样本。在第二阶段，利用心理声学听觉掩蔽原理持续优化对抗样本。对抗扰动被注入人类无法感知的范围内。该原理如公式 6所示。

$$\begin{cases} x_{1t}^{adv} = x_{1(t-1)}^{adv} + clip_x^{\varepsilon}(Adam(\nabla_x loss_{\mathrm{mod\ }el}(x_{1(t-1)}^{adv}, y))) \\ x_{2t}^{adv} = x_{1t}^{adv} + Adam(\nabla_x loss_{masking}(x_{1t}^{adv}, y_{masking})) \end{cases} \quad (6)$$

0.0.2 集成攻击方法

在图像领域中，对于每个模型，集成攻击采用相同的攻击方法（通常采用快速梯度符号法）。在此基础上，集成可分为三个不同层级：预测、逻辑值和损失函数。相关方法如公式 7所示。

$$\begin{aligned} &l(x_t^{adv}, y) = -1_y * \log(\textstyle\sum_{k=1}^{n} \alpha_k p_k(x_t^{adv}))l(x_t^{adv}, y) = -1_y * \\ &\log(soft\max(\textstyle\sum_{k=1}^{n} \alpha_k \log\ its_k(x_t^{adv})))l(x_t^{adv}, y) \\ &= \textstyle\sum_{k=1}^{n} \alpha_k l_k(x_t^{adv}, y)) \end{aligned} \quad (7)$$

其中，l 是模型的损失，1_y 是 x的真实标签 y 的独热编码，p_k 是第 k 个模型的预测结果，$\alpha_k \geq 0$ 是受 $\sum_{k=1}^{n} \alpha_k = 1$约束的集成权重。

在音频对抗攻击领域，人们已经开发出大量针对特定模型的方法，这些方法通常依赖不同的损失函数来产生不同的攻击策略和结果。然而，这些攻击的核心仍然是使用梯度下降，这意味着它们都需要计算梯度，因此为基于梯度的嵌入提供了机会。本研究探讨了两种梯度集成策略，即串行梯度集成和并行梯度集成，以生成迁移性强的对抗样本。具体而言，我们采用两种集成算法，基于基础模型为给定输入计算不同的梯度，并根据集成策略将这些梯度结合起来，再利用梯度动量和 Adam算法更新对抗样本。概念框架如图 4 所示。

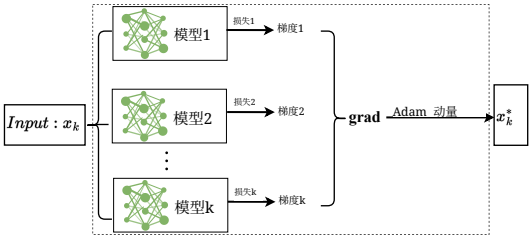

图4：基于集成攻击生成对抗样本。

此外，由于语音识别中的决策边界具有更强的振荡性，其对抗样本的可迁移性明显不如图像领域。这种不规则且不稳定的边界会阻碍对抗样本的可迁移性。不过，局部平滑的优势在于能够减少决策边界的剧烈振荡。在 [37] 和 [38], 中，作者提出在输入数据中添加少量噪声，并反复应用该方法，以平滑噪声对模型的影响。另一种方法是， [39] 对输入图像进行平滑处理，并使用梯度上升算法最大化预测分类器的错误率，从而对分类器发起更有效且迁移性更强的攻击。

在2017年NIPS竞赛中，Dong等人 [40]证明，梯度动量能显著提升样本的可迁移性，使其更好地避免陷入局部极小值。具体而言，梯度动量算法会将当前梯度向量与前一梯度向量相加，并在每次迭代更新参数之前乘以一个动量因子。这种方法保留了先前的梯度信息，使模型在下降过程中保持一定的方向和速度，从而提高收敛速度。

在本研究中，我们提出了一种新颖的梯度平滑方法，以提升集成攻击的对抗可迁移性，并解决上述问题。我们的方法受到高斯噪声梯度平滑方法的启发 [37——39], 该方法通过考虑修改点周围其他点的梯度来设计梯度，并将经过多次高斯模糊后得到的梯度平均值作为最终梯度。我们的方法在每次迭代中对多个模型的梯度进行平滑。首先，在输入上叠加白噪声；声音容易受到白噪声而非高斯噪声的干扰。接着，分别汇总每个模型的梯度信息，最后对汇总后的梯度进行归一化，得到最终的平滑梯度。为增加模型数量，我们结合自集成丢弃法，以增强模型多样性。我们的方法的计算过程如公式8 和公式9所示。

$$\tilde{g}_t = \sum_{i=1}^{m} \nabla_x l(D(x_t^{adv}) + \varsigma_i, y), \varsigma_i \sim U(-A, A) \quad (8)$$

$$g_t = \tilde{g}_t / means(abs(\tilde{g}_t)) \quad (9)$$

其中 D 是 Dropout函数。

获得平滑梯度后 g_t ，我们采用公式 10 来加快收敛速度，并克服梯度下降过程中收敛缓慢且不稳定的问题，从而促进更平滑的优化过程，并更快地生成高质量的对抗样本。

$$g_{t+1} = \mu g_{t-1} + g_t \quad (10)$$

0.0.3 集成策略

1) 随机梯度集成（RGE）。对于梯度集成，有串行集成和并行集成。串行集成对抗攻击依赖多个不同的基础模型，每次迭代随机选取一个模型的梯度作为集成梯度，经过 T 次迭代后得到最终对抗样本。确保模型多样性是集成攻击的关键因素，而随机梯度能够增加模型多样性。随机选择模型的梯度可以增加不确定性，并摆脱对特定模型的过拟合，从而提高可迁移性。此外，不同模型的鲁棒性可能有所不同，通过随机选择模型的梯度，可以在不同模型之间进行扩展并识别鲁棒模型，这可能有助于提高对抗样本的鲁棒性。随机梯度集成方法也已用于 [41] 中的集成

对抗训练，并取得了积极成果。算法的流程概述如下：算法 1。

2)动态梯度加权集成（DGWE）。一般而言，梯度值（L 范数）越大，其在模型参数更新中的作用越重要，也越能主导这一过程。然而，在集成训练中，如果某个模型的梯度值和方向与其他模型存在显著差异，集成模型就更容易受到大梯度模型的牵引，导致模型的预测偏向这些模型，无法实现预期的集成效果。出现这一问题的原因可能是，大梯度中含有噪声或出现在不规则区域，因而对参数更新的影响更大。较大的梯度也可能表明输入数据的可靠性较低，而较小的梯度则意味着可靠性较高。在集成对抗攻击中也存在同样的问题：降低梯度与其他模型不同的模型的权重，或许能够避免这些模型对最终输出产生过大的影响，从而提升对抗样本的多样性和鲁棒性。在我们的 DeepSpeech、Kaldi 和 Lingvo 代码库中，我们测得这三个模型的梯度存在相当大的差异。因此，避免在集成中对个别模型过度依赖十分重要。我们建议降低梯度较大的模型的权重，以限制其对集成结果的不当影响。我们发现，这种方法还可以降低集成的方差，最终减小用于参数更新的梯度，并避免较大梯度可能造成的破坏性偏差。

算法1： 音频集成攻击算-法

输入： 一个良性音频样本 x 及其标签 y，一组 K 个基础模型及其对应的梯度 $\{g_1, g_2 ... g_k\}$，以及集成梯度 g^{ens}。
扰动界限 ε，迭代次数 T，
内部更新频率 M。

输出： 满足条件的对抗样本 x^{adv}，$|x^{adv} - x||_2 \leq \varepsilon$

。初始化 $x^{adv} = x$ ，

对于 $t = 0\ to\ T-1$ **执行**

  选择模型索引 $k \in \{1, 2 ... K\}$

  梯度平滑，

  对于 $m = 0\ to\ M-1$ **执行**

    初始化 $g_{0k} = 0$

    计算基础模型的梯度：

    $\tilde{g}_{mk} = \nabla_x l(D(x_t^{adv}) + \varsigma_i, y)\ \varsigma_i \sim U(-A, A)$

    $g_{mk} = g_{(m-1)k} + \tilde{g}_{mk}$

  end

  归一化

  ，$g_{tk} = g_{mk} / means(abs(g_{mk}))$,

  重置 $g_{mk} \leftarrow 0$;

  动量梯度，

  $g_{tk} = \mu g_{(t-1)k} + g_{tk}$。

  # RGE 方法。

  生成一个随机索引 r $\in \{1, 2 ... K\}$,

  制作 $g^{ens} = g_{rk}$。

  # DGWE 方法。

  根据公式获取权重。11。

  $g^{ens} = \sum_{k=1}^{K} w_k g_{tk}$。

  # 更新外层对抗样本，

  $x_{t+1}^{adv} = x_t^{adv} + clip_x^{\varepsilon}(Adam(g^{ens}))$。

end

返回 $x^{adv} = x_T^{adv}$。

这里，我们提出了基于梯度 $||*||_2$ 范数的 DGWE。梯度 $||*||_2$ 范数提供了一种输入加权公式，权重取决于输入对结果的重要性。具体设置方法如下，见公式 11：

$$\begin{cases} w_i = \exp(-||g_{tk}||_2{}^{1/\sigma^2}) \\ w_i = w_i / \sum_{i=1}^{K} w_i \end{cases} \tag{11}$$

在该方法中，$||g_{tk}||_2$ 表示模型 k的梯度欧几里得范数，而 σ 是用于控制权重平滑度的参数（本文取 σ=1）。当 σ 较大时，权重会更加平滑，不同数据源之间的权重差异也会更小。相反，当 σ 较小时，权重会更加动态，不同数据源之间的权重差异也会更大。这是一种先进的数据融合技术，可根据各模型梯度对结果的贡献动态调整权重，从而得到更精细的整体梯度，使其更适合优化过程。算法流程见算法的 DGWE 部分1。

集成攻击的实验与评估 评估指标

可迁移对抗攻击是一种能够减少生成对抗样本所需时间和资源开销，并有望提高攻击成功率的方法。然而，评估攻击的可迁移性至关重要，因为目标模型的变化可能会削弱可迁移对抗攻击的效果。为了评估 RGE 和 DGWE 方法生成的对抗样本的可迁移性，我们测量了这些对抗样本的迁移率 (TR) [42, 43]。TR 是一种指标，用于衡量本地模型生成的对抗样本在目标模型上的误分类率。迁移率越高，表明可迁移性越强。

在我们的实验中，我们在集成模型上训练了 20 个对抗样本（单次输入，需要手动运行 20 次实验

），每 5 个样本对应一个目标命令（共 4 个目标命令）。在本研究中，尽管不同目标命令可能具有不同的可迁移性，但我们不考虑目标命令对可迁移性的影响 [44], 。我们的主要目标是评估通过集成方法生成的对抗样本的可迁移性，并侧重于整体可迁移性。我们将成功迁移至目标模型的、包含不同目标命令的对抗样本数量与对抗样本总数之比定义为迁移率，以评估我们方法的可迁移性表现，如式12所示。

$$TR = \frac{\sum_{i \in T} \sum_{j \in N} F(x_{ij})}{X} \tag{12}$$

其中 T 是目标命令组， N 是特定命令的样本组， x_{ij} 表示第 i 个命令的第 j 个样本， X 是样本总数，而 F (.) 是目标模型的输出。在表 2中，我们通常以“A/B”的形式展示此迁移率（也可以转换为小数）。

实验设置

为了评估我们提出的组合式对抗攻击策略的可迁移性，我们在商业语音识别 API 上进行了实验，例如科大讯飞、阿里巴巴和百度提供的 API。这些 API 提供高质量的英语语音识别服务，直接影响数百万用户的体验。针对这些商业 API 的攻击更接近真实世界的攻击模式，也更加复杂。在选择攻击目标时，我们选取了一些具有代表性的命令作为攻击目标， [8]。在此背景下，可迁移性越高，意味着潜在危险越大。

此外，如前述实验所示，对抗样本的载体对可迁移性有着重要影响。载体的选择会直接影响对抗样本的生成及其可迁移性。在 CommanderSong[15], 中，程宇轩等人首次在攻击中使用音乐片段作为对抗样本的载体，并指出音乐具有大众消费品的特性，凭借其广泛的普及度和覆盖范围，在攻击中具有天然优势。针对音乐片段的攻击可能会引发公众关注。在 [12, 13] 中，音乐片段也被证明可以作为载体。在我们的研究中，我们同样将音乐片段作为对抗样本的载体，包括流行音乐、古典音乐、摇滚音乐、以及轻音乐，涵盖韩语、英语、日语、中文、俄语和阿拉伯语等多种语言。每段音乐片段的时长约为 4 秒。

攻击评估

在本节中，我们通过比较几种最先进的攻击方法，对 RGE 和 DWGE 之间的迁移率差异进行定量分析，包括针对白盒攻击的 DS [45], ITRA [5], 和 CS [15] ，以及针对商业 API 的黑盒攻击 DW [12], Occam [14] 和 DC [13]。这一比较为研究模型的可解释性和鲁棒性提供了额外依据。这些语音识别对抗攻击均属于单模型迁移攻击，且未利用集成模型。这可能导致成本过高，而且可供组合使用的模型也不足。在探索基于集成模型训练的对抗样本的可迁移性方面，仍存在许多空白。本研究旨在弥补这些不足：我们收集了三个可用于集成攻击的语音识别模型，并提出了一种适用于语音识别模型集成的攻击算法。表2 展示了各类攻击对不同 API 的迁移率。

DS 攻击。DeepSpeech攻击（DS） [45]是一种针对 DeepSpeech 模型的白盒攻击。训练对抗音频有两种方法，即梯度下降和遗传算法。这两种方法都非常耗时。DS 攻击能够生成与原始音频相似度达 99.9 % 的音频波形，并可将其转录为任意目标短语。DS 攻击最初在 DeepSpeech-0.1.0 中实现，后来发现它对 0.2.0 和 0.3 版本也具有良好的可迁移性。0 [44] 这可归因于这些版本在模型结构和优化算法上的相似性。然而，DS 攻击对应用程序接口的可迁移性较差，平均迁移率仅为 0.1。这一局限使得该攻击难以扩展到其他模型。

IRTA 攻击。IRTA 攻击算法 [5]是一种两阶段方法，旨在攻击 Lingvo 白盒模型。第一阶段会生成一个样本，用来欺骗 Lingvo 模型。随后，在第二阶段应用心理声学掩蔽，在音频中注入人类无法察觉的对抗扰动。与其他方法相比，IRTA 的声音非常轻微，其扰动几乎无法被人耳察觉。这一特性与人类听觉范围密切相关。遗憾的是，我们的研究发现，IRTA 对抗样本的可迁移性受到严重限制，平均迁移率仅为 0.04。这意味着该方法

[8]包括关灯、”打开网站、我的车在哪里，以及天气怎么样

表2：商业云语音转文本 API 上对抗样本的迁移率（TR）。

API 服务	音频长度	DS	ITRA	CS	DW	Occam	DC	RGE	DGWE
阿里云 API	4s	1/20	0/20	4/20	6/20	20/20	-	14/20	8/20
讯飞 API	4s	0/20	0/20	8/20	14/20	20/20	13.8/20	12/20	14/20
百度云 API	4s	2/10	0/20	2/20	-	-	14.9/20	15/20	13/20
腾讯 API	4s	1/20	1/20	2/20	8/20	-	16.5/20	-	
Microsoft API	4s	6/20	3/20	6/20	16/20	20/20	-	-	
平均值	4s	2/20	0.8/20	4.4/20	11/20	20/20	15.1/20	13.7/20	11.7/20

注：缩略语 DS、ITRA、CS 和 DW 分别指 [45], [5], [15], 和 [12], 中介绍的不同攻击方法。Occam 和 DC 分别指 [14] 和 [13], 中提出的攻击方法。“-”表示该攻击方法尚未在应用程序接口上进行测试。在 DW [12], 中，作者评估了 10 个对抗样本。为与本文的数字格式保持一致，我们将该数字乘以二。在 DC [13], 中，作者将攻击成功率作为结果报告；通过计算将其转换为本文所用的格式。

仅对特定模型有效。造成这一限制的主要原因是，自动语音识别（ASR）模型只关注人类可听范围，而不关注不可听范围。因此，ASR 模型也会忽略该范围内的声音。

CS攻击。CommanderSong攻击（CS） [15]是一种白盒攻击，可将命令嵌入歌曲中。随后播放这首经过处理的歌曲，即可激活现实世界中的自动语音识别系统并执行目标命令。在 Kaldi 中，WTA攻击的成功率可达 100%，而 WAA攻击的成功率可达 96%。我们的研究发现，CS 对抗样本的平均迁移率（TR $= 0.22$）高于 DS 和 ITRA，这可能表明其扰动程度更高。通过分析对抗样本和代码，我们观察到，CS攻击中的音频片段扰动程度高于 DS 和 ITRA（CS 约为 4000，而 DS 和 ITRA 约为 2000）。因此，CS攻击的听觉可感知性不如 DS 和 ITRA。

DW攻击。Devil’s Whisper（DW）攻击 [12]旨在攻击商业黑盒模型，这类模型会部署替代模型来估计目标模型。对抗样本能够成功攻击各种商用设备，包括 Google Assistant、Google Home、Amazon Echo 和 Microsoft Cortana，执行目标指令的平均成功率为 98%。然而，DW 对科大讯飞的迁移率相对较高（TR $= 0.55$），对腾讯和阿里巴巴 API 的迁移率则较低。DW 是一种针对 API 的攻击，模型过拟合风险较低。可根据来自不同 API 的反馈调整对抗样本的训练，使其能更有效地攻击这些 API，从而获得比其他攻击更高的迁移率。然而，由于模型之间存在差异，基于特定 API 反馈生成对抗样本，可能会增强其对某些 API 的攻击能力，同时削弱其对其他 API 的攻击能力。

DC攻击。消失命令（DC）攻击 [13]依赖声学掩蔽模型进行攻击自动语音识别系统。通过注入噪声或插入额外音频来掩盖目标命令，防止其被识别为正常语音。在研究中，作者评估了使用 DC 攻击生成的对抗样本对科大讯飞、腾讯和百度 API 的攻击效果，并取得了令人鼓舞的结果。然而，作者没有公开源代码或样本，我们也尚未验证其准确性，或在其他 API 平台上进行进一步评估。

奥卡姆攻击。奥卡姆攻击 [14] 通过协同进化优化算法生成对抗样本，该算法能够以最少的信息识别最优解，并有效欺骗语音识别系统。虽然该攻击针对 API 的成功率可达 100%，但作者尚未发布源代码，也未提供对抗样本，因而限制了对所报告结果的进一步评估和验证。

RGE 和 DGWE 攻击的评估。采用 RGE 的集成攻击是一种串行方法，它会随机选择梯度作为集成模型梯度来进行更新 x^*。如表 2所示，RGE 的集成对抗方法对百度、阿里巴巴和 iFlytek 的应用程序接口表现出显著的可迁移性，平均迁移率为 0.69，高于 DS、ITRA、CS 和 DW，但低于 Occam 和 DC。这一发现凸显了集成攻击在增强样本鲁棒性并促进其向其他模型泛化方面的潜力。此外，我们的分析表明，语音识别模型决策边界的相似性对样本迁移率的影响有限。正如 [44],样本特征的鲁棒性可能比模型相似性对可迁移性更为关键，而随机化可以降低过拟合风险并提高对抗样本的鲁棒性。总体而言，结果表明，采用随机梯度集成的集成攻击能够有效提升对抗样本向普通黑盒 API 的可迁移性。

表中2，基于 DGWE 的并行集成攻击方法的迁移率低于

基于 RGE 的串行集成方法，其平均迁移率为 0.59，比 RGE 低约 0.1。值得注意的是，与科大讯飞和百度相比，RGE 和 DGWE 在阿里巴巴的应用程序接口上的表现都较差，这表明不同的应用程序接口对对抗样本具有不同的易感性。实验表明，我们研究中提出的 RGE 和 DGWE 对准确率较高的黑盒模型影响显著，而且任何集成方法都能对对抗样本的可迁移性产生积极影响。我们还观察到，精心构造且成功率较高的对抗样本具有更强的可迁移性。RGE 比 DGWE 更有效，能够生成可迁移性更强的对抗样本。

可迁移性因素探索

为解决基于深度学习的开源语音识别模型多样性不足的问题，我们提出一种结合 Dropout自集成和模型集成的方法，以提高模型多样性。具体而言，丢弃法以概率 p将部分输入随机置零，从而在传播过程中使部分神经元失活，进而形成多个网络。在对抗攻击中，输入 x 的值会被更新，既作为输入，也用于参数更新。我们以概率x_i 将部分 p随机置零，随后根据新输入的梯度更新x ，如式（13）所示。

$$\begin{cases} r_i \sim Bernoulli(p) \\ x_0^{adv} = x \\ x_t^{adv} = x_{t-1}^{adv} + clip_x^{\varepsilon}(Adam(\nabla_x loss_{ctc}(r_i * x_{t-1}^{adv}, y))) \end{cases} \tag{13}$$

其中， r_i 表示与 x维度相同的伯努利分布，概率参数 p在Dropout函数中起着关键作用。如果 p 设置过低，集成中的模型数量可能不足，从而削弱集成效果。反之，如果 p 的值过高，样本训练过程可能发散，导致无法收敛或收敛成本过高，最终训练出的样本可迁移性较弱。

实验结果显示了p 数值与可迁移性之间的关系，见图5。随着 p数值增加，两种集成方法的可迁移性都呈现先升后降的趋势。当$p = 0.5$时，可迁移性达到峰值，RGE 的可迁移性为 0.69，DGWE 为 0.59，差值仅为 0.1。起初，可迁移性与 p 数值p 值越高集成模型的数量增加，从而略微降低样本对特定模型的拟合程度，并提高对抗样本的鲁棒性。然而，达到某个点后，可迁移性会随着 p 值的增加而开始下降。当 $p = 1$、 r_i 为 0 时，输入变为 0，参数不会更新，导致可迁移性为 0。在可迁移性下降的过程中，RGE 的下降速度快于 DGWE，并且更早降至 0。 p 值持续增加会使损失函数更难收敛，并提高获取样本的成本，从而导致结果变差。

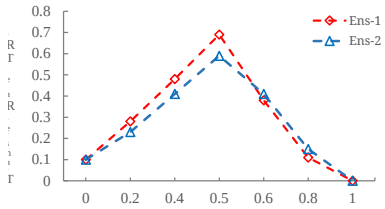

图5：迁移率（TR）随丢弃率（p值）的变化情况。随着 p 值增大，迁移率呈上升趋势，直至p 值达到 0.5。随后，随着 p 值继续增大，迁移率开始下降。

讨论：防御

自动语音识别系统是一种复杂模型，由多个组件构成，包括预处理、特征提取、声学处理和语言处理。这些组件都容易受到攻击，防御机制也不例外。对抗训练、输入预处理、模型适配、检测和随机化是常用的防御机制。

先前的研究 [46] 和 [47] 表明，使用对抗音频训练自动语音识别模型有助于防御某些攻击。然而，制作此类对抗音频的成本很高。此外，这种方法只能抵御已识别的攻击，无法防范新的攻击。此外，由于大多数对抗音频是由音乐片段生成的，因此难以评估对抗训练对自动语音识别模型的有效性。因此，对抗训练的效果评估面临着重大挑战。

在现有文献中，已有多种防御方法被提出，用于应对自动语音识别系统中的对抗扰动。随机平滑由 [11], 提出，而[1]指出，WaveGAN声码器可以重建

波形并去除扰动。其他方法，如标签平滑 [2]和音频压缩[3, 4] ，也得到了推荐。此外，下采样方法 [5]和附加失真信号 [6]也已得到探索。

在本研究中，我们评估了两种防御方法的有效性：下采样防御 [12] 和噪声防御 [15], 在三种自动语音识别接口（百度、阿里巴巴和科大讯飞）的情境下，我们针对 20 个对抗样本测试了 5200 Hz、5600 Hz 和 6000 Hz 的下采样重采样率，以及 500、1000 和 2000 的噪声幅度。结果见表3，结果表明下采样防御优于噪声防御。下采样方法之所以有效，是因为它能够压缩数据并通过插值进行重建，从而使精心注入的扰动难以恢复，并过滤掉高频信号。另一方面，噪声防御方法使用随机噪声掩盖对抗扰动，但这种方法也可能扭曲听觉感知。

表 3：防御结果。

方法	数值	百度 API	阿里巴巴 API	科大讯飞 API
降采样	5200	0/20	0/20	0/20
	5600	0/20	0/20	1/20
	6000	1/20	2/20	2/20
添加噪声	500	2/20	2/20	2/20
	1000	1/20	1/20	1/20
	2000	0/20	1/20	0/20

局限性

在本研究中，我们采用了三种语音识别模型作为集成对抗攻击的基础模型。然而，由于可用的基础模型数量有限，我们没有进一步探究基础模型数量对对抗样本训练的影响。此外，载体的选择对对抗样本能否成功至关重要。具体而言，与多元素音乐载体相比，单一音乐元素更容易在较少扰动下遭受成功攻击。然而，这一现象背后的具体音频特征仍不清楚。未来研究中，我们计划增加基础模型的数量，并探究多个模型之间的内在联系。我们还将探索为对抗样本挑选合适载体的算法，以提升深度学习算法的鲁棒性和可解释性。

结论

本研究探讨了影响音频对抗样本可迁移性的潜在因素。我们揭示了噪声在攻击及这些样本可迁移性方面所起的微妙作用。具体而言，不同程度的噪声可能促进或限制对抗样本的可迁移性。此外，我们发现，尺度不变性以及静音帧的存在与可迁移性呈正相关。关于用于生成对抗样本的载体，我们发现，对话载体比摇滚音乐载体具有更好的隐蔽性和更高的成功率。这从侧面证实了静音帧可能带来的积极影响，因为对话载体包含许多静音片段。总体而言，本研究揭示了影响音频对抗样本可迁移性的多种因素，并为开发更稳健的语音识别系统提供了见解。通过大量音素分析，我们认为，基于深度学习的语音识别模型的决策边界并不规则，这给对抗样本的迁移带来了挑战。但我们相信，存在一条最优攻击路径，能够在降低这些模型间听觉感知差异的同时提高攻击成功率。通过实验，我们观察到集成方法能够增强对抗样本的可迁移性。因此，我们提出了两种集成迁移攻击方法：RGE 和 DGWE，它们在科大讯飞、腾讯和百度的应用程序接口上取得了出色的攻击成功率。我们还注意到，自集成中的 p 值是一个会影响对抗样本可迁移性的关键参数。

缩略语

DL 深度学习 AEs 对抗样本 DNNs 深度神经网络 API 应用程序接口 RGE 随机梯度集成 DGWE 动态梯度加权集成 TR 迁移率 DS DeepSpeech攻击 IRTA 《用于自动语音识别的不可感知、鲁棒且有针对性的对抗样本》中的攻击方法 ASR 自动语音识别 CS CommanderSong攻击 DW Devil's Whisper攻击 DC Disappeared Command攻击

致谢

不适用 .

资助

作者的研究工作部分得到了国家自然科学基金项目（编号：62202275）和山东省自然科学基金项目（编号：ZR2022QF012）的资助。

数据和材料的可用性

代码可在此处获取：xxxx

作者详情

[1]山东大学网络空间科学与技术学院，中国青岛。 [2]泉城实验室（QCL），中国济南。

参考文献。

1 Athalye, A., Carlini, N., Wagner, D.：混淆梯度带来虚假的安全感：规避针对对抗样本的防御。在：国际机器学习会议，第 274–283 页（2018年）。PMLR 2. Carlini, N., Wagner, D.：迈向神经网络鲁棒性的评估。在：2017年 IEEE安全与隐私研讨会（sp），第 39–57 页（2017年）。Ieee 3. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.：深度学习在对抗环境中的局限性。在：2016年 IEEE欧洲安全与隐私研讨会（EuroS&P），第 372–387 页（2016年）。IEEE 4. Carlini, N., Wagner, D.：音频对抗样本：针对语音转文本的定向攻击。在：2018年 IEEE安全与隐私研讨会（SPW），第 1–7 页（2018年）。IEEE 5. Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C.：用于自动语音识别的不可感知、鲁棒且有针对性的对抗样本。在：国际机器学习会议，第 5231–5240 页（2019年）。PMLR 6. Taori, R., Kamsetty, A., Chu, B., Vemuri, N.：针对黑盒音频系统的定向对抗样本。在：2019年 IEEE安全与隐私研讨会（SPW），第 15–20 页（2019年）。IEEE 7. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.：神经网络的有趣特性。arXiv 预印本 arXiv:1312.6199（2013年） 8. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.：通过输入多样性提升对抗样本的可迁移性。在：IEEE/CVF计算机视觉与模式识别会议论文集，第 2730–2739 页（2019年） 9. Dong, Y., Pang, T., Su, H., Zhu, J.：通过平移不变攻击规避针对可迁移对抗样本的防御。在：IEEE/CVF计算机视觉与模式识别会议论文集，第 4312–4321 页（2019年） 10. Long, Y., Zhang, Q., Zeng, B., Gao, L., Liu, X., Zhang, J., Song, J.：用于对抗攻击的频域模型增强。在：计算机视觉—ECCV 2022：第 17 届欧洲会议，以色列特拉维夫，2022年10月23–27日，论文集，第 IV 部分，第 549–566 页（2022年）。施普林格 11. Zhu, Y., Chen, Y., Li, X., Chen, K., He, Y., Tian, X., Zheng, B., Chen, Y., Huang, Q.：从分布视角理解并提升对抗可迁移性。IEEE图像处理汇刊 **31**，6487–6501（2022年） 12. Chen, Y., Yuan, X., Zhang, J., Zhao, Y., Zhang, S., Chen, K., Wang, X.：Devil's Whisper：一种针对商用黑盒语音识别设备的通用物理对抗攻击方法。在：USENIX安全研讨会，第 2667–2684 页（2020年） 13. Xu, J., Zhang, J., Zhu, J., Yang, Y.：消失命令：利用声音掩蔽攻击自动语音识别系统的欺骗攻击。arXiv 预印本 arXiv:2204.08977（2022年） 14. Zheng, B., Jiang, P., Wang, Q., Li, Q., Shen, C., Wang, C., Ge, Y., Teng, Q., Zhang, S.：利用最少信息对商用语音平台发起黑盒对抗攻击。在：2021年 ACM SIGSAC计算机与通信安全会议论文集，第 86–107 页（2021年） 15. Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., Zhang, S., Huang, H., Wang, X., Gunter, C.A.：Commandersong：一种实用对抗性语音识别的系统化方法。在：第 27 届 {USENIX} 安全研讨会（{USENIX} Security 18），第 49–64 页（2018年） 16. Wu, W., Su, Y., Lyu, M.R., King, I.：通过对抗变换提升对抗样本的可迁移性。在：IEEE/CVF计算机视觉与模式识别会议论文集，第 9024–9033 页（2021年） 17. Liu, Y., Chen, X., Liu, C., Song, D.：深入探究可迁移对抗样本与黑盒攻击。arXiv 预印本 arXiv:1611.02770（2016年） 18. Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., Lim, S.-N.：利用中间层攻击增强对抗样本的可迁移性。在：IEEE/CVF国际计算机视觉会议论文集，第 4733–4742 页（2019年） 19. Kim, W.J., Hong, S., Yoon, S.-E.：注意力空间中的多样化生成式扰动，用于可迁移对抗攻击。在：2022年 IEEE 国际图像处理会议 (ICIP)，第 281 页 285 – (2022)。IEEE 20. Kim, H., Park, J., Lee, J.：为语音分类生成可迁移对抗样本。Pattern Recognition **137**，109286 (2023) 21. Neekhara, P., Hussain, S., Pandey, P., Dubnov, S., McAuley, J., Koushanfar, F.：用于语音识别系统的通用对抗扰动。arXiv 预印本 arXiv:1905.03828 (2019) 22. Cisse, M.M., Adi, Y., Neverova, N., Keshet, J.：Houdini：利用对抗样本欺骗深度结构化视觉与语音识别模型。载于：Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.（编）《神经信息处理系统进展》，第 30 卷。Curran Associates, Inc.，???（2017）。 https://proceedings.neurips.cc/paper/2017/file/d494020ff8ec181ef98ed 97ac3f25453-Paper.pdf 23. Tram`er, F., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.：可迁移对抗样本空间。arXiv预印本 arXiv:1704.03453（2017）24. Wu, L., Zhu, Z., Tai, C., 等：理解并增强对抗样本的可迁移性。arXiv预印本 arXiv:1802.09707（2018）25. Serban, A., Poll, E., Visser, J.：目标识别中的对抗样本：全面综述。ACM计算调查（CSUR）**53**(3), 1–38（2020）26. Lin, Z., Peng, A.,Wei,R., Yu,W., Zeng,H.：一种增强的尺度不变方法可迁移对抗攻击。载于：2022年IEEE国际图像处理会议（ICIP），第 3788–3792页（2022）。IEEE 27. Li, Z., Bhojanapalli, S., Zaheer, M., Reddi, S., Kumar, S.：使用尺度不变架构训练鲁棒神经网络。载于：国际机器学习会议，第12656–12684页（2022）。PMLR 28. Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C.：用于自动语音识别的不可感知、鲁棒且有针对性的对抗样本。载于：国际机器学习会议，第5231–5240页（2019）。PMLR 29. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M.L., Stolcke, A., Yu, D., Zweig, G.：微软2016年会话语音识别系统。2017年IEEE国际声学、语音与信号处理会议（ICASSP），5255–5259（2016）30. Xu, M., Zhang, T., Li, Z., Zhang, D.：用于评估和增强对抗防御的尺度不变对抗攻击。arXiv预印本 arXiv:2201.12527（2022）31. Li, Y., Bai, S., Zhou, Y., Xie, C., Zhang, Z., Yuille, A.：通过幽灵网络学习可迁移对抗样本。载于：AAAI人工智能会议论文集，第34卷，第11458–11465页（2020）32. Wu, D., Wang, Y., Xia, S.-T., Bailey, J., Ma, X.：跳跃连接至关重要：关于使用残差网络生成的对抗样本的可迁移性。arXiv预印本 arXiv:2002.05990（2020）33. Che, Z., Borji, A., Zhai, G., Ling, S., Li, J., Le Callet, P.：一种由长期梯度记忆驱动的新型集成对抗攻击。载于：AAAI人工智能会议论文集，第34卷，第3405–3413页（2020）34. Xiong, Y., Lin, J., Zhang, M., Hopcroft, J.E., He, K.：用于提升对抗可迁移性的随机方差缩减集成对抗攻击。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第14983–14992页（2022）35. Liu, Y., Chen, X., Liu, C., Song, D.：深入探究可迁移对抗样本与黑盒攻击。arXiv预印本 arXiv:1611.02770（2016）36. Hang, J., Han, K., Chen, H., Li, Y.：针对深度学习系统的集成对抗黑盒攻击。Pattern Recognition **101**, 107184（2020）37. Balduzzi, D., Frean, M., Leary, L., Lewis, J., Ma, K.W.-D., McWilliams, B.：梯度破碎问题：如果残差网络是答案，那么问题是什么？载于：国际机器学习会议，第 342–350页（2017）。PMLR 38. Smilkov, D., Thorat, N., Kim, B., Vi´egas, F., Wattenberg, M:

Smoothgrad：通过添加噪声来消除噪声。arXiv预印本 arXiv:1706.03825（2017）39. Wu, L., Zhu, Z., Tai, C., 等：理解并增强对抗样本的可迁移性。arXiv预印本 arXiv:1802.09707（2018）40. Dong, Y., Liao, F., Pang, T., Su, H., Zhu,J., Hu, X.,Li,J.：利用动量提升对抗攻击。载于：IEEE计算机视觉与模式识别会议论文集，第9185–9193页（2018）41. Tram`er, F., Kurakin, A.,Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.：集成对抗训练：攻击与防御。arXiv预印本 arXiv:1705.07204（2017）42. Hang, J., Han, K., Chen, H.,Li,Y.：针对深度学习系统的集成对抗黑盒攻击。Pattern Recognition **101**, 107184（2020）43. Papernot, N., McDaniel, P., Goodfellow, I., Jha,S., Celik, Z. B., Swami, A.：针对机器学习的实用黑盒攻击。载于：2017年ACM亚洲计算机与通信安全会议论文集，第506–519页（2017）44. Chen, Y., Zhang, J.,Yuan, X., Zhang, S.,Chen, K., Wang, X., Guo, S.：Sok：研究自动语音识别系统安全性的模块化方法。ACM隐私与安全汇刊 **25**(3), 1–31（2022）45. Carlini, N., Wagner, D.：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（2018）。IEEE 46. He, Z., Rakin, A.S., Fan, D.：参数化噪声注入：通过可训练随机性提升深度神经网络抵御对抗攻击的鲁棒性。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第588–597页（2019）47. Du, H., Xie, L., Li, H.：通过域对抗训练实现抗噪声语音转换。Neural Networks **148**, 74–84（2022）