https://crad.ict.ac.cn
62(3): 751−764, 2025

针对音频识别的物理世界音素对抗攻击

王嘉凯[1,2] 孔宇升[1,3] 陈镇东[4,5] 胡 琎[2] 尹子鑫[2] 马宇晴[2] 杨晴虹[5] 刘祥龙[1,2,6]

[1](北京中关村实验室 北京 100194)

[2](复杂关键软件环境全国重点实验室(北京航空航天大学) 北京 100191)

[3](北京航空航天大学自动化科学与电气工程学院 北京 100191)

[4](南栖仙策(南京)科技有限公司 南京 211106)

[5](北京航空航天大学软件学院 北京 100191)

[6](合肥综合性国家科学中心数据空间研究院 合肥 231200)

(wangjiakai@zgclab.edu.cn)

Phonemic Adversarial Attack Against Audio Recognition in Physical World

Wang Jiakai[1,2], Kong Yusheng[1,3], Chen Zhendong[4,5], Hu Jin[2], Yin Zixin[2], Ma Yuqing[2], Yang Qinghong[5], and Liu Xianglong[1,2,6]

[1](*Beijing Zhongguancun Laboratory*, *Beijing* 100194)

[2](*State Key Laboratory of Complex and Critical Software Environment* (*Beihang University*), *Beijing* 100191)

[3](*School of Automation Science and Electric Engineering*, *Beihang University*, *Beijing* 100191)

[4](*Polixir Technologies*, *Nanjing* 211106)

[5](*School of Software*, *Beihang University*, *Beijing* 100191)

[6](*Institute of Data Space*, *Hefei Comprehensive National Science Center*, *Hefei* 231200)

Abstract Audio recognition has been widely applied in the typical scenarios, like auto-driving, Internet of things, and etc. In recent years, research on adversarial attacks in audio recognition has attracted extensive attention. However, most of the existing studies mainly rely on the coarse-grain audio features at the instance level, which leads to expensive generation time costs and weak universal attacking ability in real world. To address the problem, we propose a phonemic adversarial noise (PAN) generation paradigm, which exploits the audio features at the phoneme level to perform fast and universal adversarial attacks. Experiments are conducted using a variety of datasets commonly used in speech recognition tasks, such as LibriSpeech, to experimentally validate the effectiveness of the PAN proposed in this paper, its ability to generalize across datasets, its ability to migrate attacks across models, and its ability to migrate attacks across tasks, as well as further validating the effectiveness of the attack civilian-oriented Internet of things audio recognition application in the physical world devices. Extensive experiments demonstrate that the proposed PAN outperforms the comparative baselines by large margins (about 24 times speedup and 38% attacking ability improvement on average), and the sampling strategy and learning method proposed in this paper are significant in reducing the training time and improving the attack capability.

Key words security of artificial intelligence; adversarial attack; audio recognition; physical attack; phonemic adversarial noise

摘 要 语音识别等智能技术在自动驾驶、物联网等场景下得到了广泛的应用.近年来,针对语音识别的对抗攻击研究逐渐受到关注.然而,现有的大多数研究主要依赖于粗粒度的音频特征来在实例级别生成

收稿日期: 2023-06-05; **修回日期**: 2024-05-10

通信作者: 刘祥龙(xlliu@buaa.edu.cn)

对抗噪声,这导致生成时间成本高昂且攻击能力弱.考虑到所有语音可以被视为基本音素的不同组合,提出了一个基于音素的通用对抗攻击方法——音素对抗噪声(phonemic adversarial noise, PAN),该方法通过攻击在音频数据中普遍存在的音素级别的细粒度音频特征,以生成音素级对抗噪声,取得了更快的对抗噪声生成速度并具备更强的通用攻击能力.为了全面地评估所提出的PAN框架,在实验中基于LibriSpeech等多种语音识别任务中被广泛采用的公开数据集,对提出的音素对抗噪声的攻击有效性、跨数据集的泛化能力、跨模型迁移攻击能力和跨任务迁移攻击能力进行了验证,并进一步在物理世界设备中证实了对民用智能音频识别应用的攻击效果.实验结果表明,所提出的方法比其他对比方法的攻击能力提高了38%,生成速度快了24倍以上,且提出的采样策略和学习方法对降低训练时间和提升攻击能力具有重要作用.

关键词 人工智能安全;对抗攻击;音频识别;物理攻击;音素对抗噪声

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330445 **CSTR:** 32373.14.issn1000-1239.202330445

深度学习技术的繁荣推动了人工智能在语音识别领域的应用,例如自动语音识别(automatic speech recognition, ASR)、声纹识别(voiceprint recognition, VPR)等.与此同时,这些智能语音技术在自动驾驶、物联网等场景下有大量的应用,对推动社会发展、促进生产力提升发挥了重要的作用.然而,音频识别智能应用可能被用于不良的社会目的,例如通过窃取用户的音频信息并分析其行为.因此,近年来保护用户隐私的需求不断增加并广受重视.作为一种影响模型决策的攻击手段,对抗样本为保护用户隐私提供了新的选择.通过开展对抗攻击研究,可以生成具有通用攻击性的对抗噪声,从而误导窃听者用于分析语音的智能模型,能够为深度学习时代的隐私保护提供一种新的可行解决方案,吸引了研究人员的广泛关注.

截至目前,已经有一系列针对音频识别任务的对抗攻击策略的研究[1−11].已有工作中,一部分旨在为每个特定音频生成对抗噪声[1-4](对抗样本).这种对抗噪声由于在一些确定的数据点上进行了良好的训练,因此具有较强的攻击能力.此外,这些生成的对抗噪声可以实现误导模型输出特定结果(即目标攻击)或随机结果(即非目标攻击)的效果.另一部分工作[6-11]的目标是生成对任意输入音频都有效的对抗噪声,即通用对抗噪声.这种对抗噪声通常是非目标的,因为受影响的音频数据不可预测,开展目标攻击很难得到预期的攻击效果.

总体而言,尽管很多工作尝试了生成音频对抗噪声,但现有方法总是集中于粗粒度特征,忽略了音频语音的细粒度音素特征,因此仍表现出显著的不足.简言之,现有的对抗攻击方法具有2个局限性:1)执行对抗攻击耗时较久,这意味着必须花费大量时间生成对抗噪声,使其在实际场景中不适用;2)当前方法对多样化输入的通用攻击能力并不令人满意,即在使用一个噪声攻击多个音频数据时,其攻击成功率仍有很大可提升空间.

为了解决上述问题,本文提出一种音素对抗噪声(phonemic adversarial noise, PAN)生成框架.如图1所示,该框架通过全面考虑音频中的基本音素特征,以进行更快更高强度的通用音频攻击.语言学研究结果表明音素是组成语音语言的一种更基本的特征[12],例如一句语音可以被视为不同音素的组合,它们之间的关系就像“木屋”和“木料”一般.因此直觉上,如果可以对“木料”直接进行攻击,应能取得更理想的效果.通过充分利用关键音素特征优化对抗噪声,生成的对抗音频噪声便能够获得更好的攻击性能,实现更加快速、更加通用的攻击能力.具体而言,为了加快生成速度,本文提出了均衡化音素密度采样,以获得数量少但音素特征丰富的代表性实例,从而有效降低训练成本,在保持攻击性能的同时具有显著的生成加速效果;为了提高通用攻击能力,本文提出了多样化音素模式学习方法,更好地完成生成噪声的优化过程,本文方法使得有限的训练样本上的优化过程能够更充分地遍历包含不同音素特征的不同音频片段,从而更好地激发被攻击模型的偏差.在音素级别生成对抗噪声,以生成更好的音频识别攻击,这是一种防止音频隐私窃取的保护策略.本文的音素噪声具有更快的生成速度和更强的通用攻击能力.

为了全面评估提出的PAN方法,本文采用了包括DeepSpeech2, Wav2Vec等在内的多个性能出色的语音识别模型作为基线模型.实验结果表明,与其他先

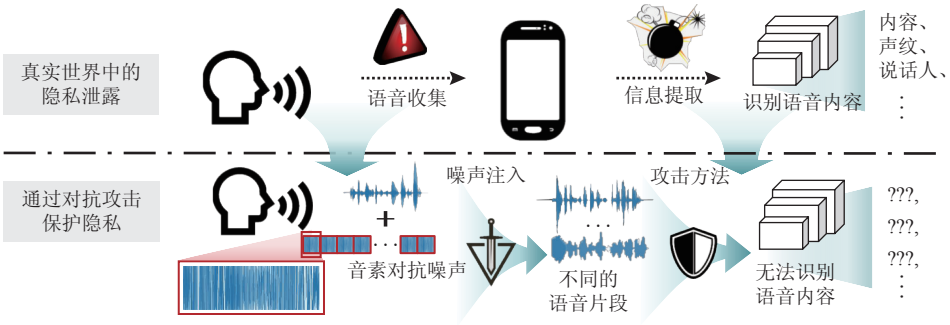


Fig. 1 Phonemic adversarial noise generation framework

图 1 音素对抗噪声生成框架

进的对抗噪声生成方法相比，本文提出方法的攻击能力具有显著优越性. 本文的主要贡献总结为 3 个方面：

1)受音素存在性的启发，本文提出了一种音素级别的对抗攻击方法，设计了针对音频识别模型进行较强攻击能力的 PAN 方法.

2)提出了音素对抗攻击框架，包括均衡化音素密度采样和多样化音素模式学习，以快速生成具有通用攻击能力的强有力对抗噪声.

3)在数字世界和物理世界基于多种音频识别任务进行了大量实验，证明本文提出的 PAN 方法在性能上比其他先进方法有着显著的优势. 例如，在 ASR 任务上节省了 96.33% 的时间并提升了 38% 的攻击性能.

本文围绕针对音频识别的物理世界音素对抗攻击方法展开研究和讨论.

1 相关工作

1.1 真实世界音频识别

近年来，智能音频识别技术吸引了很多关注，其中的典型代表包括 ASR 任务和 VPR 任务. ASR 任务主要旨在使用深度学习技术[13]直接将语音波形或特征序列转换为文本，DeepSpeech[14]及其改进版本 DeepSpeech2[15]是该子领域中的代表性工作. Wav2Vec[16]结合了基于预测编码的无监督预训练技术、音频离散化技术和 Transformer 架构，取得了显著优越的效果. SEW[17]对 Wav2Vec 模型作出了优化并提升了效率和准确率. 在 VPR 任务的研究中，典型的方法包括使用时延网络作为深度特征提取网络的 X-vectors[18]方法，以及使用深度残差 RNN 和三元组损失来提高识别准确性的 DeepSpeaker[19]方法.

1.2 音频对抗攻击

对抗样本最初在计算机视觉领域被提出，用于误导深度模型做出错误的预测[20]. 到目前为止，在不同任务上均有大量关于对抗样本的研究，例如交通标志识别、车辆识别等[21-24]. 近年来，这一研究方向也吸引了音频领域研究人员的注意[1-11]. 然而，尽管取得了一定的结果，但仍然存在缺陷，有较大的可以改进提升的余地. 例如，文献 [2] 可以通过足够的迭代使音频转录为对手设置的任何文本，但巨大的时间成本使其在实践中可行性较差. 文献 [3] 降低了计算成本，但缺乏通用的攻击能力，在攻击的实用性上不足. 文献 [6] 将通用对抗样本引入到语音攻击领域，提出了通用攻击扰动 (universal adversarial perturbations, UAP) 生成方法. 在其基础上，文献 [7] 基于连续形式的时序数据分类损失，改进了 UAP 模型并给出了定向通用对抗扰动 (targeted universal adversarial perturbations, TUAP)；文献 [8] 使用叠加和前缀扰动，提出一种定向通用扰动生成方法 APUP(additive and prepending universal perturbations)，可令一段音频被识别为特定文本. 文献 [9] 面向真实场景生成了针对窃听的通用对抗噪声 CUAP(crafting audio-based UAP). 然而，这些通用攻击噪声表现并不令人满意. 我们认为，目前制约语音对抗攻击方法通用攻击能力的瓶颈之一是缺乏有效的采样策略和学习方法，特别是在进行攻击时，对音频识别模型的细粒度理解和有效利用不足. 而上述工作在优化训练时主要针对整段音频实例生成实例尺度的粗粒度对抗噪声，难以实现对更加精准的攻击目标进行对抗优化. 文献 [4] 为了不依赖于特定音频，采用了根据语音对应的文字样本，通过深度生成模型直接生成对抗音频的方式进行攻击，但已知文字样本的假设使得应用场景受限. 本文旨在基于对音频的基本音素特征的深入理解，在音素划分的基础上构造细粒度的音素采样策略和学习方法，生成音素尺度的细粒度对抗噪声，并在现实世界中生成具有较强攻击能力的通用攻击对抗噪声.

2 方 法

在本节中，首先介绍问题定义，然后详细阐述提出的 PAN 方法，该方法可以快速生成具有强攻击力的通用对抗噪声.

2.1 定 义

给定音频数据集$\mathcal{X}$，对于一个干净的音频$\boldsymbol{x}$(其中$\boldsymbol{x}\in\mathcal{X}$)，一个音频对抗噪声$\boldsymbol{\delta}$可以通过扰动使给定深度模型$\mathcal{F}(\cdot)$产生错误预测，且噪声$\boldsymbol{\delta}$受$\varepsilon$-约束. 进一步地，一个通用的音频对抗噪声，其具备更强的对模型的欺骗误导能力，即对几乎所有$\boldsymbol{x}\in\mathcal{X}$都有效，可以表示为

$$\begin{aligned}&\mathcal{F}(\boldsymbol{x}+\boldsymbol{\delta})\neq\mathcal{F}(\boldsymbol{x})\ \text{s.t.}\ \|\boldsymbol{\delta}\|<\varepsilon,\\&\text{for almost all}\ \boldsymbol{x}\in\mathcal{X},\end{aligned}\tag{1}$$

其中$\|\cdot\|$是用∞-范数测量噪声强度的距离度量，ε为约束值. 然而，本文旨在提出 PAN 方法，与上述定义存在一定区别. 更确切地说，由于本文提出的 PAN 基于音素级特征信息，因此在上述定义的框架下将其表示为

$$\begin{aligned}&\mathcal{F}\left(\boldsymbol{x}+C\left(\boldsymbol{\delta}_{\mathrm{p}},n\right)\right)\neq\mathcal{F}(\boldsymbol{x})\ \text{s.t.}\ \left\|\boldsymbol{\delta}_{\mathrm{p}}\right\|<\varepsilon,\\&\text{for almost all}\ \boldsymbol{x}\in\mathcal{X},\end{aligned}\tag{2}$$

其中$C\left(\boldsymbol{\delta}_{\mathrm{p}},K\right)$是一个拼接算子，用于将$K$个音素对抗噪声$\boldsymbol{\delta}_{\mathrm{p}}$进行拼接. 进一步来讲，$C$可表示为

$$C\left(\boldsymbol{\delta}_{\mathrm{p}},n\right)=Concat\left(\left[\boldsymbol{\delta}_{\mathrm{p}},\boldsymbol{\delta}_{\mathrm{p}},\cdots,\boldsymbol{\delta}_{\mathrm{p}}\right]_{1\times K}\right).\tag{3}$$

在本文提出的 PAN 方法下，对抗噪声的持续时间在音素级别，这使得需要循环拼接数十个甚至上百的音素噪声，例如，对于持续时间为l_x的音频$\boldsymbol{x}$，本文需要拼接$K=\dfrac{l_x}{l_{\delta_{\mathrm{p}}}}$个噪声片段，其中$l_{\delta_{\mathrm{p}}}$是音素对抗噪声$\boldsymbol{\delta}_{\mathrm{p}}$的持续时间.

2.2 概 述

语言学研究[12]发现，音素在人类语言中扮演着基础性的角色，不同的语音可以被视为音素的不同组合. 这一发现驱使本文更深入地挖掘音频中的音素特征，并通过针对音素的攻击来生成更好的对抗样本. 为此，本文提出了 PAN 方法，如图 2 所示，通过均衡化音素密度采样和多样化音素模式学习来产生更有效的对抗攻击，即更快的生成速度和更强的通用攻击能力，并在物理世界中生成真实有效的噪声攻击.

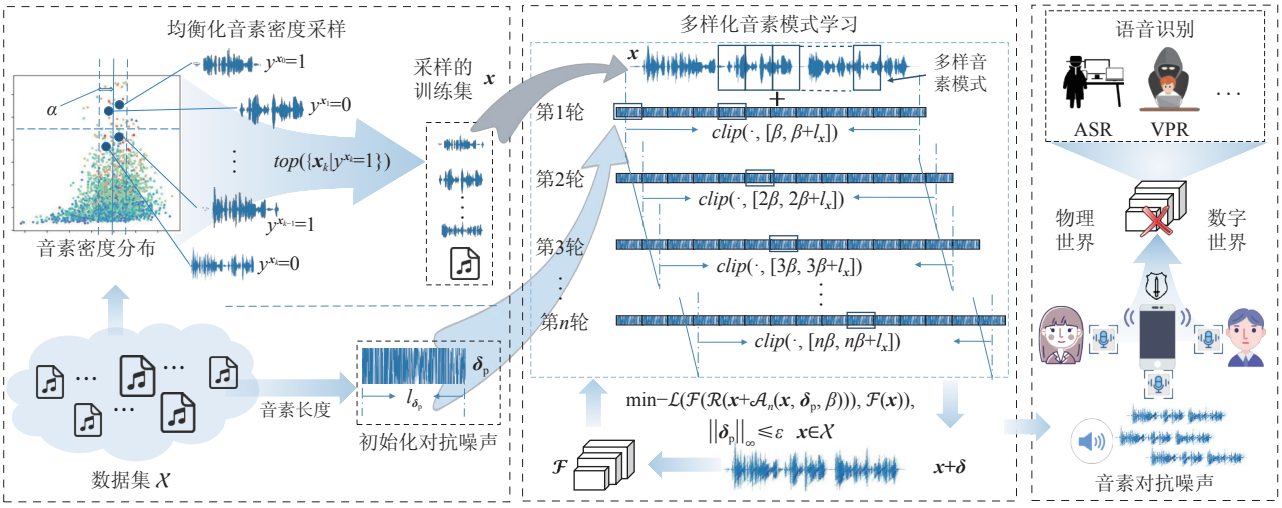


Fig. 2 The PAN generation framework

图 2 PAN 生成框架

在加速对抗噪声生成速度方面，本文主要旨在通过均衡化音素密度采样减少模型对大规模训练的依赖性，即需要大量的训练数据. 由于本文的攻击目标是音素特征，因此可以选择具有丰富音素模式的少量典型实例进行训练，并利用这些实例来优化音素对抗噪声，以实现相当的攻击性能，同时大幅度减少生成攻击的时间成本.

为提高针对各种输入音频的通用攻击能力，PAN 方法通过多样化音素模式学习，来提高生成噪声步骤中遍历到的音素模式的多样性. 由于音素噪声以异步滑动窗口的方式注入音频实例，因此对抗噪声更容易捕捉到多种不同音素特征中的关键模式，从而使生成的 PAN 可以更加准确地针对基本音素特征，更好地激发模型的不确定性，产生更强的通用攻击能力.

2.3 均衡化音素密度采样

通用对抗噪声生成方法严重依赖于训练数据集

规模，耗费大量时间. 在现有的对抗攻击方法研究中，大多采用在音频实例层面上生成对抗噪声，这导致需要大量音频数据实例参与训练，时间开销较大. 而在本文的 PAN 方法中，面向音素级别特征开展攻击允许用较少的训练实例来优化音素对抗噪声. 具体地说，由于音素特征与实例级数据相比极为简短，一个实例可能包含几十个不同的音素特征，这使得用较少的实例来优化音素对抗噪声具备较高的可行性. 因此，基于这一观点，本文在考虑音素密度的基础上，从给定的数据集中抽取更具代表性的实例，即均衡化音素密度采样.

为了选择更有代表性的采样实例，定义音素密度为单位时长内的音素数量，音素密度描述了某一数据点内音素的丰富程度. 音频$\boldsymbol{x}$的音素密度可以表述为 $\frac{1}{l_x}counter(\psi(\boldsymbol{x}))$，其中$\psi(\cdot)$是一个音素识别器[25]. 此外，考虑到实例和数据集之间的相关性，本文认为合适的采样数据点的音素密度可能与数据集的平均音素密度近似，即音素密度应该尽可能地与数据集层面的整体音素分布相一致. 这一观点的来源是在实践中，过高或过低的音素密度可能会导致过拟合或欠拟合.

具体来说，对于每个样本$\boldsymbol{x}_k \in \mathcal{X}$，本文通过估计其音素密度，决定其是否应该成为采样实例，如下所示：

$$y^{\boldsymbol{x}_k}=\begin{cases}1, & \text{if } \left|D-\frac{1}{l_{\boldsymbol{x}_k}}counter(\psi(\boldsymbol{x}_k))\right| \leqslant \alpha,\\ 0, & \text{其他},\end{cases} \tag{4}$$

其中D是给定数据集的平均音素密度，满足 $D=\frac{\sum_i counter(\psi(\boldsymbol{x}_i))}{\sum_i l_{\boldsymbol{x}_i}}$，其中$\boldsymbol{x}_i \in \mathcal{X}$，$\alpha$是控制密度间隔的阈值. 最后，所有的可接受实例都将通过一个简单的选择函数$top(\cdot)$决定其是否成为训练数据，即$\boldsymbol{x}=top(\{\boldsymbol{x}_k \in \mathcal{X} | y^{\boldsymbol{x}_k}=1\})$. 函数$top(\cdot)$选取具有更长持续时间和更大音频强度的样本，采样 10, 5, 1 个实例，如图 2 所示.

由于采样的训练实例具有相对丰富的音素特征，因此它们可以作为有效且有针对性的音素对抗噪声的训练样本. 此外，与正常的优化过程相比，采样音频数据的规模要低得多，从而降低了训练的计算成本.

2.4 多样化音素模式学习

在本文所提出的 PAN 框架中，决定攻击能力最关键的因素是生成的噪声对样本实例中音素模式的影响. 考虑到特征多样性可能对模型泛化至关重要[26-27]，本文认为在优化过程中提高所使用的音素特征的多样性，同样可以帮助待优化音素噪声获得更可观的对音素的攻击能力. 因此，本文提出了多样化音素模式学习策略，在 PAN 生成过程的每一步强制其异步地经历多个音频片段，而不是简单地重复并进行对齐. 这样一来，生成的对抗噪声可以更好地学习到不同音素模式中的关键表征，进而激活模型不确定性，获得更强的通用攻击能力.

具体而言，先前的对抗噪声优化过程总是通过在每个步骤中简单地重复，将噪声$\boldsymbol{\delta}$注入音频实例$\boldsymbol{x}$中，导致对不同音频样本实例的攻击适应性弱. 为此，本文引入了一个滑动窗口，以多样化的方式在优化过程中将音素对抗噪声注入样本实例. 对于每次迭代，音素噪声将以滑动步长β被可变地注入不同的音频片段. 例如，对于一个音素噪声$\boldsymbol{\delta}_{\mathrm{p}}$，如果它在第 1 次迭代中被注入到音频片段$[M, N]$中，在下一次迭代中，它将滑动$\beta$的持续时间，即是被注入持续时间间隔$[M+\beta, N+\beta]$的片段中. 直观地，待优化的音素噪声将在不同的迭代轮数中面对具有更多样音素模式的音频片段，因而能够高效地学习到对抗性表示，更好地在不同样本上激活模型不确定性.

在实践中，给定一个音素噪声$\boldsymbol{\delta}_{\mathrm{p}}$、一个实例$\boldsymbol{x}$和滑动步长$\beta$，可以通过以下噪声注入模块$\mathcal{A}$注入音素噪声：

$$\mathcal{A}_n\left(\boldsymbol{x}, \boldsymbol{\delta}_{\mathrm{p}}, \beta\right)=clip\left(C\left(\boldsymbol{\delta}_{\mathrm{p}}, \left\lfloor \frac{l_x}{l_{\boldsymbol{\delta}_{\mathrm{p}}}}+n \right\rfloor\right), \left[n\beta, n\beta+l_x\right]\right), \tag{5}$$

其中$clip(\cdot, [\text{start}, \text{end}])$是一个切分函数，将音频剪切到指定的区间[start, end]中，l_x是训练期间采样实例$\boldsymbol{x}$的长度，$l_{\boldsymbol{\delta}_{\mathrm{p}}}$是音素噪声$\boldsymbol{\delta}_{\mathrm{p}}$的长度，$C(\boldsymbol{\delta}_{\mathrm{p}}, k)$表示将音频$\boldsymbol{\delta}_{\mathrm{p}}$重复$k$次并拼接，$\beta$是滑动窗口的步长，$n$是迭代轮数.

借助多样化音素模式学习，音素噪声合成模块在每轮迭代中跨越采样实例中的多个音素片段. 因而在优化过程可以更有效地捕捉跨音素的关键音素模式，学习更有针对性的对抗表示，从而更好地激发原模型输出的不确定性，产生更强的通用攻击能力.

2.5 总体优化过程

本文的 PAN 框架针对语音中的音素特征进行快速和通用的对抗攻击. 为了加快生成速度，PAN 首先通过所提出的均衡化音素密度采样获得代表性音频实例. 为了提高通用攻击能力，PAN 通过进行多样化音素模式学习策略来增强训练中音素模式的多样性. 通过这种方式，生成的 PAN 将充分利用音素级优化，从而实现更好的对抗攻击性能.

除此之外, 还需要关注复杂物理环境中的实际效果. 已有研究在音频分析方面取得了一定成果, 本文引入了其中具有代表性的室内脉冲响应(room impulse response, RIR)[28]方法, 该方法通过编码声学通道状态信息(channel state information, CSI)来保证PAN在现实世界中的可用性.

具体而言, 给定一个音频数据集$\mathcal{X}$和一个目标模型(例如, 一个ASR模型)$\mathcal{F}$, 针对音频识别的PAN生成过程可以被表述为

$$\min -\mathcal{L}\left(\mathcal{F}\left(\mathcal{R}\left(\boldsymbol{x}+\mathcal{A}_n\left(\boldsymbol{x},\boldsymbol{\delta}_{\mathrm{p}},\beta\right)\right),\mathcal{F}\left(\boldsymbol{x}\right)\right)\right), \quad (6)$$

$$\text{s.t.}\ \left\|\boldsymbol{\delta}_{\mathrm{p}}\right\|_{\infty}\leqslant\varepsilon,\boldsymbol{x}\in\mathcal{X},$$

其中$\mathcal{R}(\cdot)$是RIR函数, $\|\cdot\|_{\infty}$是无穷范数. 总体的算法流程如算法1所示.

算法 1. PAN生成.

输入: 音频数据集$\mathcal{X}$, 目标模型$\mathcal{F}$;

输出: 音素对抗噪声$\boldsymbol{\delta}_{\mathrm{p}}$.

① for $\boldsymbol{x}_k\in\mathcal{X}$

② $y^{\boldsymbol{x}_k}\leftarrow$式(4); /*均衡化音素密度采样*/

③ end for

④ $\mathcal{X}'\leftarrow top(\{\boldsymbol{x}_k\in\mathcal{X}|y^{\boldsymbol{x}_k}=1\})$; /*选取训练集*/

⑤ for i in 1 to $iters$ /*重复$iters$次*/

⑥ for $\boldsymbol{x}\in\mathcal{X}'$

⑦ 根据式(6)优化$\boldsymbol{\delta}_{\mathrm{p}}$; /*音素模式学习*/

⑧ end for

⑨ end for

⑩ return $\boldsymbol{\delta}_{\mathrm{p}}$.

3 实 验

在本节中, 首先介绍了详细的实验设置; 然后从多个角度评估了提出的对抗噪声生成框架, 并报告了实验结果. 此外, 本节还提供了消融实验结果以及额外的讨论以更好地理解音频领域中的对抗噪声.

3.1 实验设置

数据集方面, 实验主要使用了著名的LibriSpeech数据集[29]用于语音识别任务. 为了验证生成的对抗噪声的通用性, 实验还使用了TED-LIUM数据集[30]和Mozilla Common Voice数据集[31]. 对于基线模型, 本节实验选择了多个具有不同体系结构和参数规模的先进模型以执行不同的任务. 具体而言, 对于ASR任务, 选择DeepSpeech2, Wav2Vec, SEW模型. 对于VPR任务, 选择DeepSpeaker模型. 所有的模型都是在开源数据集上进行预训练的.

为了准确评估性能, 本文引入了多个广泛使用的指标来评估不同任务. 对于ASR任务, 采用字符错误率(character error rate, CER)和攻击成功率(success rate, SR)作为评估指标. 此外, 噪声强度$I(\boldsymbol{\delta})$被引入以评估噪声$\boldsymbol{\delta}$的强度. 对于VPR任务, 仅引入准确性作为评估基础, 这也是该子领域中广泛使用的评估指标.

本节实验选择了几种最先进的音频对抗噪声生成方法, 特别是通用对抗音频噪声方法. 具体来说, 选择了UAP[6], TUAP[7], APUP[8]和CUAP[9]作为比较基线. 实验采用Adam优化器, 学习率为0.002, 训练轮数$E=3$, 迭代次数$iters=30$, 参数$l_{\delta_{\mathrm{p}}}=0.2$, $\alpha=0.2$, $\beta=0.77$. 所有数据集上的采样实例数量都被设定为10.所有实验都在NVIDIA RTX 3070 GPU上进行, 取3次重复实验平均值作为结果.

3.2 PAN的攻击能力

为了证明所提出的PAN方法的有效性, 首先在数字世界和物理世界进行实验. 对于数字世界攻击, 实验使用从LibriSpeech数据集中抽样的数据点训练对抗噪声, 并将DeepSpeech2作为被攻击模型. 除了基线模型之外, 实验还引入了来自Noise92的一些模拟噪声进行比较[32]. 实验结果如表1所示, 本文的方法生成的对抗噪声实现了具有显著优越性的攻击能力, 并总结得到3个结论:

1)本文提出的PAN由于其特殊设计而具有强大的通用攻击能力. 具体而言, 在LibriSpeech数据集上, PAN的SR有很大的提高, 即0.87, 而CER指标也达到了较高值, 即0.83. 与基线方法相比, PAN方法生成的噪声在SR上平均显示出46个百分点的提升, 在CER上平均显示出38个百分点的提升, 表现出强大的通用攻击能力. 此外, 与Noise92模拟的噪声相比, PAN的SR指标要显著优越, 表明对抗噪声的事实有效性.

2)本文提出的PAN方法能够以更低的时间成本生成通用的音频对抗噪声, 即更快的生成速度. 例如, LibriSpeech数据集上典型的TUAP攻击的CER值只能达到0.18, 但其时间成本为187 min(是PAN方法耗时的13倍). 目前最好的通用攻击模型CUAP具有强大的攻击能力, 但需要更多的计算成本, 即需要41倍以上的额外时间开销. 其他数据集上具有相同的趋势. 总体来看, PAN方法的平均时间成本是10.67 min, 而UAP, TUAP, CUAP, APUP的时间成本分别是308 min, 186.67 min, 512.33 min, 311.33 min.与基线方法相比, PAN方法平均实现了32.46倍的速

Table 1 Experimental Results on Different Datasets in the Digital World

表 1 数字世界不同数据集上的实验结果

方法	LibriSpeech				TED-LIUM				Mozilla Common Voice			
	I/dB	t/min↓ (加速比↑)	SR↑	CER↑	I/dB	t/min↓ (加速比↑)	SR↑	CER↑	I/dB	t/min↓ (加速比↑)	SR↑	CER↑
Raw				0.07				0.09				0.14
Noise[32]	−33		0.01	0.07	−35		0.01	0.09	−31		0.04	0.17
UAP[6]	−33	345(1 倍)	0.03	0.15	−35	345(1 倍)	0.08	0.25	−31	231(1 倍)	0.3	0.39
TUAP[7]	−33	187(1.84 倍)	0.06	0.18	−35	196(1.76 倍)	0.08	0.28	−31	177(1.31 倍)	0.59	0.56
APUP[8]	−33	350(0.99 倍)	0.02	0.16	−35	350(0.99 倍)	0.06	0.24	−31	234(0.99 倍)	0.34	0.42
CUAP[9]	−33	575(0.60 倍)	0.49	0.59	−35	577(0.60 倍)	0.38	0.47	−31	385(0.60 倍)	0.73	0.79
PAN(本文)	−33	14(24.64 倍)	0.87	0.83	−35	13(26.54 倍)	0.46	0.54	−31	5(46.20 倍)	0.84	0.89

注：上箭头表示指标数值越大越好，下箭头则相反.I表示对抗噪声的强度.Raw 表示非攻击方法.t 表示攻击过程的时间消耗. 加速比表示该方法的时间消耗与所选基线算法的时间消耗相比的加速倍数. 由于 UAP 是通用音频攻击的最典型方法之一，实验将 UAP 方法的生成速度作为基线速度计算加速比.

度提升.

3)除 CUAP 外，其他基线方法的攻击能力较弱，即在 LibriSpeech 上的 SR 指标，UAP 达到 0.03，TUAP 达到 0.06，APUP 达到 0.02.对于 UAP，尽管模型按照原论文中的方法充分训练，但结果似乎并不令人满意，表明其设计不具有实用性. 对于 TUAP 和 APUP，本文认为它们是针对个体目标攻击而设计的，因此执行通用攻击的能力有限. 由于非针对性的通用音频攻击工作很少，实验将这些针对性的攻击也纳入了比较范围.

对于物理世界的攻击，本节在物理世界中构建了一个模拟场景并报告了实验结果. 具体而言，实验中使用 3 个设备(即 2 个 iPad 和 1 个手机)作为主要播放器和接收器. 其中一个 iPad 用于播放噪声音频，另一个用于播放正常音频；手机用于接收混合物理音频，如图 3 所示. 为了准确评估，从 LibriSpeech 中随机选择 50 个实例进行测试，实验结果如表 2 所示. 结果表明与基线模型相比，本文的 PAN 方法具有最强的攻击能力，即本文所提模型的 SR 和 CER 指标分别为 0.68 和 0.74，高于 CUAP 模型的 0.32 和 0.48.此外，参考表 2，可以发现本文提出的 PAN 方法在跨数字域-物理域攻击时的性能损失小于其他基线，体现了基本音素特征对于提升对抗攻击性能的重要意义.

3.3 分析与讨论

本节进一步提供了有关本文提出的快速音素通用对抗攻击框架的更多原理探讨. 首先，试图回答一个问题，即相对于其他常见的对抗噪声，什么使 PAN 效果突出. 此外，调查了音素密度和音频长度之间的相关性，从而更好地理解它们对噪声训练的有效性. 最后还进行了跨模型的、任务的黑盒实验，以探索本文提出的 PAN 框架的可迁移性.

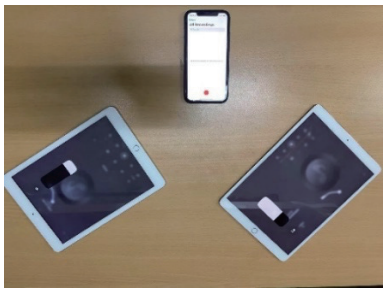


Fig. 3 Attack scenario in the physical world

图 3 物理世界攻击场景

Table 2 Experimental Results of Different Attack Methods in the Physical World

表 2 物理世界不同攻击方法的实验结果

指标	Noise	UAP[6]	TUAP[7]	CUAP[9]	APUP[8]	PAN (本文)
SR↑	0.02	0.04	0.04	0.32	0.04	0.68
CER↑	0.11	0.15	0.21	0.48	0.16	0.74

注：由于物理世界条件下不需要额外的时间开销，因此不展示时间消耗. “↑”表示指标数值越大越好.

3.3.1 音素噪声的作用分析

为了研究 PAN 框架表现良好的内在原因，本节从音频频谱的角度分析不同的对抗噪声.

具体而言，本文首先使用 UAP, APUP, TUAP, CUAP 和 PAN 方法生成对抗音频样例并绘制对应的频谱图，如图 4 所示. 进而，本文对不同对抗样本的频谱图进行模式分析. 不失一般性，本文对频谱图中表征噪声的频谱模式，与表征正常语音的频谱模式的共性和特性进行分析. 总体而言，可以得出一些有意义的定性结论：1)对于像 CUAP 等常见的对抗攻击，它们将音频数据扰动为在频谱图视图中无法区分的模式，因此使模型捕捉可识别模式更加困难. 例如，相比于干净样本，UAP, CUAP, TUAP, APUP 以及 PAN

的频谱图中,正常语音对应的频谱(黄绿色波形)都有所弱化. 2)本文的 PAN 方法不仅扰动了频谱图,而且使频谱图显示出强烈的周期性模式,进一步误导模型捕捉到错误模式. 具体地,相较于 UAP 等方法,PAN 方法能显著地在非正常语音对应的频谱(深色部分)观察到周期性的噪声频谱,说明其可能对模型的识别造成了更强烈的干扰. 因此,我们总结认为,PAN 通过同时混淆正确的频谱模式、注入错误的频谱图模式来攻击音频识别模型,从而具有更强的攻击能力.

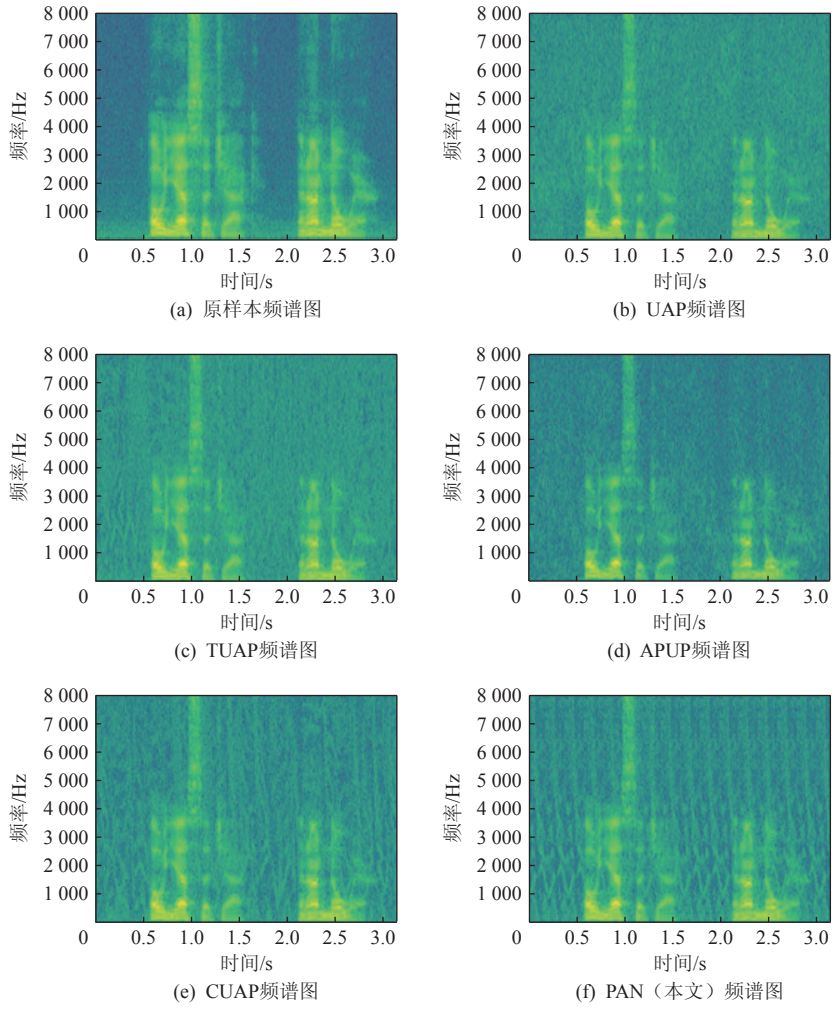


Fig. 4 Spectrogram of different adversarial instances

图 4 不同对抗样本的频谱图

3.3.2 跨数据集泛化

与在单一数据集上的通用攻击能力相比,评估对抗噪声在不同数据集上的泛化能力更具说服力. 因此,在本节中进行了额外的实验,以研究在基于一个数据集生成噪声的情况下,在另一个数据集进行测试时的攻击能力. 本节实验使用 LibriSpeech($\mathcal{X}_1$), TED-LIUM($\mathcal{X}_2$), Mozilla Common Voice Dataset($\mathcal{X}_3$)作为数据集,使用$\mathcal{X}_1\to\mathcal{X}_2$表示对抗噪声是在$\mathcal{X}_1$上生成并在$\boldsymbol{x}\in\mathcal{X}_2$上进行测试. 实验结果如表 3 所示.

根据表 3 可以得出结论,在大多数情况下,本文提出的 PAN 方法具有更强的攻击能力,特别是在跨数据集评估中. 例如,在设置$\mathcal{X}_1\to\mathcal{X}_2$下, PAN 的 CER 指标为 0.43, 而 UAP, CUAP, TUAP, APUP 攻击后的 CER 指标分别为 0.25, 0.36, 0.25, 0.24; 在$\mathcal{X}_1\to\mathcal{X}_3$的设置下, PAN 的 CER 指标为 0.67, 而作为对比方法的 UAP, CUAP, TUAP, APUP 的对应指标值为 0.40, 0.67, 0.42, 0.42.类似的观察可以在所有不同的跨数据及实验设置中找到,总体而言, PAN 的跨数据集攻击能力在 SR 指标上的提升最高可达 45 个百分点,在 CER 指标上的提升最高可达 36 个百分点. 因此,我们认为相关的实验结果能够有力地证明 PAN 的跨数据集泛化性.

Table 3 Cross Dataset Generalization Results

表 3 跨数据集泛化结果

数据集	指标	攻击方法				
		UAP[6]	CUAP[9]	TUAP[7]	APUP[8]	PAN (本文)
$X_1 \to X_2$	SR↑	0.06	0.20	0.07	0.06	0.33
	CER↑	0.25	0.36	0.25	0.24	0.43
$X_1 \to X_3$	SR↑	0.30	0.66	0.36	0.33	0.70
	CER↑	0.40	0.67	0.42	0.42	0.67
$X_2 \to X_1$	SR↑	0.03	0.14	0.03	0.02	0.23
	CER↑	0.13	0.30	0.14	0.14	0.39
$X_2 \to X_3$	SR↑	0.27	0.66	0.37	0.26	0.71
	CER↑	0.38	0.67	0.43	0.36	0.72
$X_3 \to X_1$	SR↑	0.04	0.13	0.07	0.06	0.24
	CER↑	0.14	0.28	0.23	0.17	0.39
$X_3 \to X_2$	SR↑	0.08	0.16	0.14	0.07	0.29
	CER↑	0.26	0.34	0.32	0.25	0.42

注: “↑”表示指标数值越大越好.

3.3.3 跨模型迁移攻击

本文的 PAN 方法攻击基本音素特征, 因此可能具有更好的跨模型迁移攻击性能. 为了验证这一观点, 本节使用了其他常用的模型, 即 Wav2Vec($\mathcal{F}_2$)和 SEW($\mathcal{F}_3$)以及 DeepSpeech2($\mathcal{F}_1$), 以评估本文提出的 PAN 框架的黑盒可转移性. 本节实验分别在$\mathcal{F}_1$和$\mathcal{F}_2$上优化对抗噪声(没有采用$\mathcal{F}_3$的原因是其训练不支持本文采用的对比方法的框架), 并在其他模型上进行测试, 如表 4、表 5 所示.

结果表明, 与基线方法相比, PAN 方法实现了更强的可迁移攻击能力, 在迁移攻击中相比于基线方法至少实现了 1 倍的 SR 指标提升和 0.5 倍的 CER 指标提升. 尽管这些模型具有不同的架构和参数, 但它们的识别性能也受到 PAN 方法生成的对抗噪声的影响, 表明关键音素特征有助于对抗攻击. 进一步地, 观察到对抗噪声在$\mathcal{F}_1$和$\mathcal{F}_2$模型之间的迁移能力较差,

Table 4 Results of Transfer Attack Trained on DeepSpeech2 ($\mathcal{F}_1$)

表 4 DeepSpeech2（$\mathcal{F}_1$）上训练的迁移攻击结果

攻击方法	$\mathcal{F}_1$上测试		$\mathcal{F}_2$上测试		$\mathcal{F}_3$上测试	
	SR↑	CER↑	SR↑	CER↑	SR↑	CER↑
UAP[6]	0.03	0.15	0	0.05	0	0.11
TUAP[7]	0.06	0.18	0	0.04	0.01	0.09
CUAP[9]	0.49	0.59	0.02	0.09	0.11	0.25
APUP[8]	0.02	0.16	0	0.06	0.01	0.11
PAN(本文)	0.87	0.83	0.09	0.19	0.31	0.44

注: “↑”表示指标数值越大越好.

Table 5 Results of Transfer Attack Trained on Wav2Vec ($\mathcal{F}_2$)

表 5 Wav2Vec（$\mathcal{F}_2$）上训练的迁移攻击结果

攻击方法	$\mathcal{F}_1$上测试		$\mathcal{F}_2$上测试		$\mathcal{F}_3$上测试	
	SR↑	CER↑	SR↑	CER↑	SR↑	CER↑
UAP[6]	0.03	0.14	0	0.05	0	0.1
TUAP[7]	0.04	0.16	0	0.06	0.02	0.12
CUAP[9]	0.07	0.2	0.11	0.25	0.14	0.3
APUP[8]	0.04	0.18	0	0.05	0.02	0.12
PAN(本文)	0.14	0.3	0.6	0.7	0.45	0.58

注: “↑”表示指标数值越大越好.

SR 和 CER 指标均不超过 0.03, 这可能由较大的模型结构差异导致.

3.3.4 跨任务迁移攻击

本文提出的 PAN 攻击方法由于利用了音素特征和对语音中这一基本且共有音素的影响, 我们认为其理论上具备对不同语音任务的攻击力. 因此, 本文在声纹识别任务上验证了 PAN 方法的有效性. 在实验设置上, 本文采用了完全黑盒设置, 也就是, 使用在 ASR 任务中生成的 PAN 并直接在 VPR 数据中开展测试. 本文引入了一个基于 LibriSpeech 数据集预训练的 DeepSpeaker 模型来验证 PAN 跨任务攻击能力, 该模型是声纹识别领域典型的且被广泛使用的模型. 在对比方法方面, 我们只利用攻击力较强的 CUAP 方法和 TUAP 方法作为对比, 其实验结果如表 6 所示. 可以看出, PAN 达到了更低的准确率, 这代表 PAN 可以在跨任务攻击中取得一定的效果, 且相比于其他的方法, PAN 的攻击性更强, 证明了其通用攻击能力.

Table 6 Attacking Results on VPR Task

表 6 在 VPR 任务上的攻击结果 %

攻击方法	TUAP[7]	CUAP[9]	PAN(本文)
准确率↓	30.0	30.0	21.0

注: “↓”表示指标数值越小越好.

3.3.5 音素噪声的可防御性分析

由于本文提出的 PAN 是一种对抗噪声生成方法, 可以攻击深度模型并进行误导, 所以对其滥用存在一些道德上的担忧, 例如该方法可能被利用来干扰在线学习应用和音频识别应用. 为此, 本文提供了一些额外的策略来规避这种社会环境中的潜在滥用威胁. 通过这种方式, PAN 可以最大程度地被用于公益性应用(例如语音隐私保护). 具体来说, 实验采用了对抗训练(adversarial training, AT)[20] 和对抗音频检测

(adversarial audio characterizing, AAC)[33]策略作为建议的防御措施,分别采用语音识别指标和检测精度指标进行评估.详细的实验结果如表 7 所示.实验结果表明,PAN 生成的对抗样本能够有效地被对抗音频检测防御,从而在一定程度上缓解了伦理问题,可以更多地被用于本文所倡导的对社会积极的目的.当然,这一实验结果也从另一角度说明 PAN 在隐蔽攻击方面的应用局限性.

Table 7 Results of Adversarial Defense Methods

表 7 对抗攻击方法的结果

防御方法	指标	攻击方法				
		UAP[6]	TUAP[7]	CUAP[9]	APUP[8]	PAN(本文)
AT	SR	0.02	0.02	0.01	0.02	0.01
	CER	0.10	0.09	0.10	0.11	0.10
AAC	AUC	0.59	0.66	0.73	0.61	0.80

3.4 消融实验

由于本文的 PAN 方法中有 2 个关键操作,因此有必要对它们进行详细的消融研究,以进一步证实其有效性.此外,本节还具体地研究了超参数l_{δ_p}, α, β的影响以及采样的训练样本量.

3.4.1 关键操作的有效性

为了研究本文的 PAN 方法内部关键操作的有效性,即均衡化音素密度采样(BPDS)和多样化音素模式学习(DPPL),本节通过控制变量设计了消融实验.

具体而言,为控制关键操作的影响因素,本文仅使用单一的操作:当采用 BPDS 时,DPPL 将不会被使用.为了使结果更有说服力,本文提供了多个具有不同实验设置的结果作为比较.实验结果如表 8 所示.实验得出下述结论:1)BPDS 显著降低了训练过程中的优化成本,训练时间从 557 min 降至 14 min;2)DPPL 在提高通用攻击能力方面表现更加有效,即与基线设置(无 BPDS 和 DPPL)相比,DPPL 的 SR 和 CER 指标分别从 0.14, 0.27 增加到 0.40, 0.52.此外,注意到整体 PAN 方法的性能远高于仅使用 BPDS 的性能,本文认为这一现象背后的原因是 BPDS 更加符合真实情况下音素分布的归纳偏置,最大化 DPPL 的效用.

Table 8 Ablations Experiments on the Key Operations

表 8 关键操作的消融实验

关键操作		指标		
BPDS	DPPL	t/min↓	SR↑	CER↑
×	×	557	0.14	0.27
✓	×	14	0.39	0.5
×	✓	557	0.4	0.52
✓	✓	14	0.87	0.83

注:“↓”表示指标数值越小越好;“↑”表示指标数值越大越好.

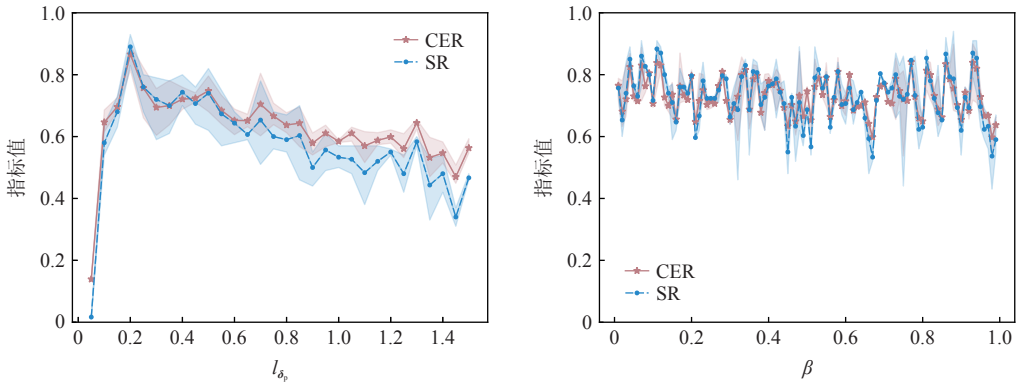
3.4.2 超参数的敏感性

在本文提出的 PAN 方法中,存在多个超参数,包括l_{δ_p}, α, β.本节具体探讨了这些超参数如何影响框架性能,对其进行了消融实验,并报告了实验结果,以进一步认识 PAN 方法.具体而言,对于l_{δ_p},其与音素持续时间高度相关.因此,在实验中将其采样范围限制为$l_{\delta_p} \in (0, 1.5]$,并将间隔设置为 0.05(即总共采样 30 个点),然后分别优化音频对抗噪声并测试其攻击能力.同样,对于α和β超参数也采用间隔采样策略,α和β的采样范围分别为(0,2](间隔为 0.2,总共采样 10 个点)和(0,1)(间隔为 0.01,总共采样 99 个点).结果如图 5 所示.对于l_{δ_p},可以发现,在l_{δ_p}=0.20 时,生成的对抗噪声的性能达到了一个相对较高的水平.本文推测其背后的原因是合适的持续时间使噪声生成训练能更有效地学习针对音频的对抗音素特征,即在对音素特征的过拟合和欠拟合之间达到平衡.对于α和β,发现其没有明确的规律,即整体的攻击力水平都高于对照方法,因此将其分别设置为α=0.2 和β=0.77,如此可以达到更可观的攻击力水平.值得注意的是,对于α而言,较低的值会产生更高的性能.此外,尽管β表现相对不稳定,但其最坏情况也比基线模型好得多,即 CER 指标为 0.6,高于 CUAP 模型的 0.59.

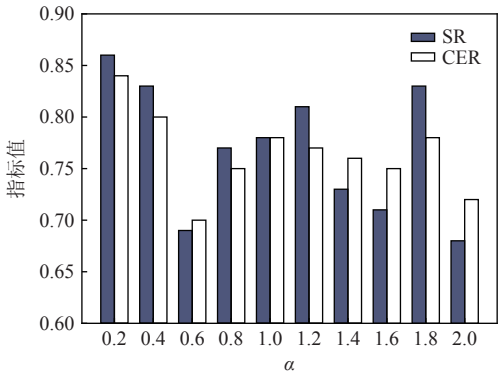
3.4.3 训练过程中采样的样本数量

除了l_{δ_p}, α, β外,由于所提出的 PAN 生成框架中包含训练样本采样改进(即所提出的均衡化音素密度采样策略),有必要探究采样数量变化对 PAN 的攻击效果带来的影响.为此,我们在本节探究了不同的采样样本数对实验结果的影响.

具体地,本文将采样的样本数分别设置为 10, 8, 6, 4, 2, 1,并在 LibriSpeech 上评估它们对 PAN 攻击性的影响程度.对于每种设置,进行了 10 次重复实验.细节上,本文报告了 SR 和 CER 指标在 10 次实验中的最大值、最小值、中位值和平均值,以全面了解采样数对 PAN 的有效性.实验结果如表 9 所示,通过此结果可以得到结论,采样的样本数对 PAN 方法的有效性存在一定影响.更准确地说,尽管采样 1 个训练样本可以达到更低的时间开销和可观的通用攻击能力,即最大 SR 为 0.79(这一结果可以超过采样样本数为 10 时最坏的 SR 值),但与较高的采样样本数所达到的实验结果相比,其通用攻击稳定性似乎更差,



(a) l_{δ_p}消融实验　　(b) β消融实验



(c) α消融实验

Fig. 5　Ablation results under the different hyperparameters

图 5　对不同超参数下的消融结果

Table 9　Ablation Study Under different Sampling Numbers on LibriSpeech Dataset

表 9　LibriSpeech 数据集上不同采样样本数下的消融实验

指标	统计量	采样样本数					
		10	8	6	4	2	1
SR	最大值	0.95	0.90	0.96	0.95	0.88	0.79
	中位值	0.87	0.85	0.80	0.84	0.70	0.53
	最小值	0.75	0.64	0.58	0.64	0.46	0.19
	平均值	0.85	0.79	0.80	0.84	0.70	0.50
CER	最大值	0.90	0.84	0.88	0.90	0.81	0.77
	中位值	0.83	0.81	0.79	0.83	0.70	0.60
	最小值	0.74	0.67	0.64	0.64	0.55	0.37
	平均值	0.83	0.77	0.78	0.82	0.69	0.59

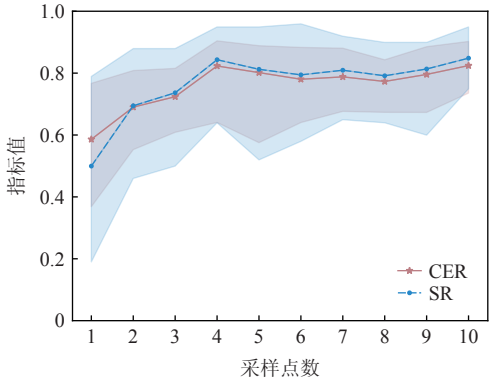


Fig. 6　Overall tendency of the attacking ability under different sampling numbers

图 6　不同采样样本数下攻击能力的总体趋势

最坏的 SR 和 CER 情况要低得多. 如图 6 所示的总体结果, 可以清楚地发现这种趋势. 因此, 考虑到 PAN 攻击性和稳定性之间的权衡, 本文默认主要实验结果中的采样样本数为 10.

4　结果与讨论

本文面向无人驾驶、物联网等典型领域的智能音频识别应用, 提出了一种基于 PAN 的生成框架, 该框架利用音频数据中的基本音素特征, 来实现更快速、通用的对抗攻击. 具体而言, 本文提出了均衡化音素密度采样来获取更少但具有代表性的实例, 以进行 PAN 优化, 从而加速生成速度. 此外, 本文介绍了多样化音素模式学习策略, 在优化过程中使用滑动噪声注入完成多样化音素模式学习, 有助于增强对抗噪声的通用攻击能力. 受益于音素级特征的利

用, 本文所提出的方法生成的 PAN 取得了良好的性能表现. 为了评估 PAN 的有效性, 本文进行了广泛的实验, 包括详细的消融实验和讨论. 结果表明, PAN 方法在速度和攻击成功率方面均优于目前其他最好方法, 相比于基线方法平均具有 24 倍以上的提速和 38% 的攻击成功率提升.

在这项工作中, 本文提出的 PAN 方法对音频识别执行了成功攻击, 并侧面揭示了音素特征在音频识别领域的重要性. 相信这种典型的音素特征可以被更深入地利用, 并激发更深入的研究. 例如, 训练数据与对抗样本可迁移性的关系仍旧是一个开放性的问题, 近期计算机视觉领域的实验研究表明, 数据增强对对抗样本的可迁移性具有负面影响[34], 而本文的研究表明基于音素构建恰当的数据清洗和数据增强策略, 能够生成具有高迁移特性的对抗样本, 推测其原因为恰当的训练样本能够更加精确地揭示存在于不同语音模型中的归纳偏置以及相应敏感模式. 进一步地, 基于音素的分析及对抗攻击能够为语音识别模型的可解释性和鲁棒性分析带来帮助. 针对通过音素特征来增强物联网中的智能音频模型的识别效果同样意义重大, 在未来可基于本文研究成果开展进一步的探索.

PAN 攻击的局限性主要集中在易于检测方面. 由于 PAN 噪声具有显著的频谱模式, 因此可以更容易地被特征检测技术识别. 这制约了真实场景中的潜在音频攻击的负面应用. 尽管进一步探索基于 PAN 的隐匿攻击能力会提升负面的社会影响, 但从研究的角度出发, 本文认为其仍然是一个有意义的方向, 即如何令噪声逃脱特征检测技术的识别, 并在语音领域生成具有更高隐匿特性的对抗样本.

作者贡献声明: 王嘉凯提出了算法思路、设计了实验方案并撰写了论文的主要内容; 孔宇升、陈镇东、胡琎、尹子鑫负责算法设计、实现和实验执行, 并参与了论文修改; 马宇晴、杨晴虹和刘祥龙对论文提出了指导性意见并参与了论文修改.

致谢: 本文得到了中关村实验室和 MindSpore 框架、CANN 和 Ascend AI Processor 的项目支持.

参 考 文 献

[1] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C/OL]//Proc of 2018 IEEE Security and Privacy Workshops (SPW). Piscataway, NJ: IEEE, 2018[2024-01-24]. https://ieeexplore.ieee.org/abstract/document/8424625

[2] Qin Yao, Carlini N, Cottrell G, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[C]// Proc of the 36th Int Conf on Machine Learning. Cambridge, MA: JMLR, 2019: 5231−5240

[3] Liu Xiaolei, Wan Kun, Ding Yufei, et al. Weighted-sampling audio adversarial example attack[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4908−4915

[4] Qu Xinghua, Wei Pengfei, Gao Mingyong, et al. Synthesising audio adversarial examples for automatic speech recognition[J]. Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 1430−1440

[5] Han Songshen, Guo Songhui, Xu Kaiyong, et al. Perturbation analysis of the vital region in speech adversarial example based on frame structure[J]. Journal of Computer Research and Development, 2024, 61(3): 685−700 (in Chinese)
(韩松莘, 郭松辉, 徐开勇, 等. 基于帧结构的语音对抗样本重点区域扰动分析[J]. 计算机研究与发展, 2024, 61(3): 685−700)

[6] Neekhara P, Hussain S, Pandey P, et al. Universal adversarial perturbations for speech recognition systems[J]. arXiv preprint, arXiv: 1905.03828, 2019

[7] Zong Wei, Chow Y W, Susilo W, et al. Targeted universal adversarial perturbations for automatic speech recognition[C]//Proc of the 24th Int Conf on Information Security. Berlin: Springer, 2021: 358−373

[8] Lu Zhiyun, Han Wei, Zhang Yu, et al. Exploring targeted universal adversarial perturbations to end-to-end ASR models[J]. arXiv preprint, arXiv: 2104.02757, 2021

[9] Mathov Y, Senior T B, Shabtai A, et al. Stop bugging me! evading modern-day wiretapping using adversarial perturbations[J]. Computers and Security, 2022, 121: 102841

[10] Li Jiguo, Zhang Xinfeng, Jia Chuanmin, et al. Universal adversarial perturbations generative network for speaker recognition[C/OL]//Proc of 2020 IEEE Int Conf on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2020 [2024-01-24]. https://ieeexplore.ieee.org/document/9102886

[11] Xie Yi, Li Zhuohang, Shi Cong, et al. Enabling fast and universal audio adversarial attack using generative model[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 14129−14137

[12] Twaddell W F. On defining the phoneme[J]. Language, 1935, 11(1): 5−62

[13] Malik M, Malik M K, Mehmood K, et al. Automatic speech recognition: A survey[J]. Multimedia Tools and Applications, 2020, 80: 9411−9457

[14] Hannun A, Case C, Casper J, et al. Deep speech: Scaling up end-to-end speech recognition[J]. arXiv preprint, arXiv: 1412.5567, 2014

[15] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin[C]//Proc of the 33rd Int Conf on Machine Learning. Cambridge, MA: JMLR,

2016: 173−182

[16] Baevski A, Henry Z, Mohamed A, et al. Wav2Vec 2.0: A framework for self-supervised learning of speech representations[C]//Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2020: 12449−12460

[17] Felix W, Kwangyoun K, Jing P, et al. Performance-efficiency trade-offs in unsupervised pre-training for speech recognition[C]//Proc of 2022 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 7667−7671

[18] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robust dnn embeddings for speaker recognition[C]//Proc of 2018 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2018: 5329−5333

[19] Li Chao, Ma Xiaokong, Jiang Bing, et al. DeepSpeaker: An end-to-end neural speaker embedding system[J]. arXiv preprint, arXiv: 1705.02304, 2017

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[21] Wang Jiakai. Adversarial examples in physical world[C]// Proc of the 30th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 4925−4926

[22] Liu Aishan, Wang Jiakai, Liu Xianglong, et al. Bias-based universal adversarial patch attack for automatic check-out[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 395−410

[23] Wang Jiakai, Yin Zixin, Hu Pengfei, et al. Defensive patches for robust recognition in the physical world[C]//Proc of 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 2456−2465

[24] Wang Jiakai, Liu Aishan, Yin Zixin, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world[C]//Proc of 2021 IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 8565−8574

[25] Kyubyong P, Jongseok K, Nicholas L. g2pE: A simple Python module for English grapheme to phoneme conversion[CP/OL]. 2019 [2024-01-24]. https://github.com/Kyubyong/g2p

[26] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6: 1−48

[27] Steven Y F, Varun G, Jason W, et al. A survey of data augmentation approaches for NLP[C]//Proc of 2021 Findings of the Association for Computational Linguistics: ACL/IJCNLP. Stroudsburg, PA: ACL, 2021: 968−988

[28] Peddinti V, Chen G, Povey D, et al. Reverberation robust acoustic modeling using i-vectors with time delay neural networks[C]//Proc of the 16th Annual Conf of the Int Speech Communication Association. Baixas, France: ISCA, 2015: 2440−2444

[29] Panayotov V, Chen G, Povey D, et al. LibriSpeech: An ASR corpus based on public domain audio books[C]//Proc of 2015 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 5206−5210

[30] Rousseau A, Deleglise P, Esteve Y. TED-LIUM: An automatic speech ´ recognition dedicated corpus[C]// Proc of the 8th Int Conf on Language Resources and Evaluation. Istanbul, Turkey: ELRA, 2012: 125−129

[31] Ardila R, Branson M, Davis K, et al. Common voice: A massively multilingual speech corpus[C]// Proc of the 12th Language Resources and Evaluation Conf. Istanbul, Turkey: ELRA, 2019: 4218–4222

[32] Johnson D H, Shami P N. The signal processing information base[J]. IEEE Signal Processing Magazine, 1993, 10(4): 36−42

[33] Yang Zhuolin, Li Bo, Pin-Yu C, et al. Characterizing audio adversarial examples using temporal dependency[J]. arXiv preprint, arXiv: 1809.10875, 2018

[34] Zhang Yechao, Hu Shengshan, Leo Y Z, et al. Why does little robustness help? Understanding adversarial transferability from surrogate training[J]. arXiv preprint, arXiv: 2307.07873, 2023



Wang Jiakai, born in 1995. PhD, associate professor. His mian research interests include physical adversarial attacks and defenses, model security evaluation, and trustworthy AI.

王嘉凯, 1995 年生. 博士, 副研究员. 主要研究方向为物理对抗攻击和防御、模型安全性评估、可信赖人工智能.



Kong Yusheng, born in 1995. PhD candidate, engineer. His main research interests include physical adversarial attacks and defenses, industrial artificial intelligence, and IIoT cyber security.

孔宇升, 1995 年生. 博士研究生, 工程师. 主要研究方向为物理对抗攻击和防御、工业人工智能、工业物联网安全.



Chen Zhendong, born in 1998. Master, assistant engineer. His main research interest includes the audio adversarial attack.

陈镇东, 1998 年生. 硕士, 助理工程师. 主要研究方向为音频对抗攻击.



Hu Jin, born in 2000. PhD candidate. His main research interests include physical adversarial attacks, AI attack and defense simulation, and trustworthy artificial intelligence.

胡　琎, 2000 年生. 博士研究生. 主要研究方向为物理对抗攻击、智能攻防推演、可信赖人工智能.



Yin Zixin, born in 1999. Master. His main research interests include physical adversarial attacks and defenses, and trustworthy artificial intelligence.

尹子鑫, 1999 年生. 硕士. 主要研究方向为物理对抗攻防、可信赖人工智能.



Ma Yuqing, born in 1992. PhD, associate professor. Her main research interests include open-world learning, model generalization, and trustworthy AI.

马宇晴, 1992 年生. 博士，副教授. 主要研究方向为开放世界学习、模型泛化性、可信赖人工智能.



Yang Qinghong, born in 1974. PhD, professor of engineering, PhD supervisor. Her main research interests include software engineering and artificial intelligence.

杨晴虹, 1974 年生. 博士，正高级工程师，博士生导师. 主要研究方向为软件工程、人工智能.



Liu Xianglong, born in 1985. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include machine learning, computer vision, and multimedia information retrieval.

刘祥龙, 1985 年生. 博士，教授，博士生导师. CCF 高级会员. 主要研究方向为机器学习、计算机视觉、多媒体信息检索.