语音对抗样本的攻击与防御综述

魏春雨, 孙 蒙, 邹 霞, 张雄伟

陆军工程大学 指挥控制工程学院 智能信息处理实验室 南京 中国 210007

摘要 语音是人机交互的重要载体, 语音中既包含语义信息, 还包含性别、年龄、情感等附属信息。深度学习的发展使得各类语音处理任务的性能得到了显著提升, 智能语音处理的产品已应用于移动终端、车载设备以及智能家居等场景。语音信息被准确地识别是人与设备实现可信交互的重要基础, 语音传递过程中的安全问题也受到了广泛关注。对抗样本攻击是最近几年兴起的一个研究热点, 攻击者通过对样本进行微小的改动使深度学习模型预测错误, 从而带来潜在的安全风险。语音识别领域同样面临着来自对抗样本的安全威胁, 在对抗样本的攻击和防御方法上也与图像识别等领域存在显著差异。因此, 研究语音对抗样本的攻击和防御方法具有重要意义。本文在介绍对抗样本相关概念的基础上, 选取语音识别中的文本内容识别、声纹身份识别两个典型任务, 按照从白盒攻击到黑盒攻击、从数字攻击到物理攻击、从特定载体到通用载体的顺序, 采取从易到难、逐步贴近实际场景的方式, 系统地梳理了近年来比较典型的语音对抗样本的攻击方法。从分类边界构造的角度, 对语音对抗样本的防御方法进行分类论述, 揭示各类方法实现防御的机理。对现阶段语音对抗样本攻击与防御方法的技术难点进行了分析与总结, 并对语音对抗样本攻防未来的发展方向进行了展望。

关键词 对抗样本; 语音识别; 声纹识别; 攻击; 防御

中图法分类号 TP391.9 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.01.07

Reviews on the Attack and Defense Methods of Voice Adversarial Examples

WEI Chunyu, SUN Meng, ZOU Xia, ZHANG Xiongwei

Lab of Intelligent Information Processing, College of Command and Control Engineering, Army Engineering University, Nanjing 210007, China

Abstract Speech plays an important role in human-computer communications. It contains not only textual and semantic information, but also has additional information on the gender, age and emotion of the speaker. The development of deep neural network has significantly improved the performance of miscellaneous tasks on speech processing. Therefore, products based on intelligent speech processing using deep learning have been applied to mobile terminals, vehicle-mounted devices, smart home and so on. Accurate recognition of speech is an important basis for trusted interaction between human and device, so the security issues involved in speech transmission has attracted a lot of research. Fooling deep learning models using adversarial examples is a hot research topic in recent years. The attacker can mislead a deep neural network by just making slight changes to a data example, which brings potential security risks to the application of deep learning model. Voice recognition is also faced with security threats from adversarial examples, but there are significant differences from other fields (e.g., image recognition) in the methods of attack and defense using adversarial examples. Therefore, it is of great significance to study the attack and defense methods of voice adversarial examples. In this paper, based on the introduction of related concepts of adversarial examples, by taking automatic speech-to-text recognition and speaker recognition as two typical tasks, we summarized the typical attack methods of adversarial examples for voice recognition systems in recent years by following the ways from white-box to black-box, from digital attacks to physical attacks and from specific carrier voice to universal carrier voices. Furthermore, in the view of the configuration of classifiers boundaries, we categorized the defense methods proposed recently and investigated how those methods work. Finally, we summarized the technical difficulties of the attack and defense methods of adversarial examples for voice recognition at present, and the future directions of the attack and defense of adversarial examples for voice recognition were predicted.

Key words adversarial examples; speech recognition; speaker recognition; attack; defense

通讯作者: 孙蒙, 博士, 副教授, Email: sunmengccjs@163.com。

本课题得到江苏省优秀青年基金(No. BK20180080)和国家自然科学基金(No. 62071484)资助。

收稿日期: 2021-07-10; 修改日期: 2021-10-11; 定稿日期: 2021-11-11

1 引言

语音信号中含有丰富的信息, 其中文本内容(即说的什么)和说话人的身份(即谁说的)最为重要[1]。前者就是狭义概念上的自动语音识别(Automatic Speech Recognition, ASR), 它将语音信号转写为文本或指令符号[2], 本文将此类识别简称为语音识别; 后者被称为声纹识别, 它将语音信号映射为说话人的身份符号, 包括“从给定说话人集合中识别出语音来自哪个说话人”的自动说话人辨认(Automatic Speaker Identification, ASI)和“验证当前语音是否来自声称的说话人”的说话人确认(Automatic Speaker Verification, ASV)两个具体任务[3]。

随着深度学习和大数据技术的发展, 语音识别已经越来越广泛地应用于现实生活中, 如智能手机的语音输入、电商平台的智能客服、金融交易的声纹认证和智能家居的语音控制等, 其根本作用是以语音作为载体和媒介实现人与设备的互通互联[4]。然而, 随着机器学习理论和方法的发展, 出现了用于模仿特定说话人语音的深度伪造[5]、针对语音识别和声纹识别的对抗样本[6], 它们都为破坏语音载体的可信性和安全性提供了具体手段, 进而对各自应用场景的信息安全构成了挑战。

具体来说, 深度伪造是利用生成式对抗网络等方法, 通过构建特定的模型, 产生听起来像目标说话人的语音样本。之所以称之为伪造, 是因为目标说话人根本没有说过这些话。深度伪造的欺骗对象主要是人耳听觉, 也可以用于导致声纹识别系统出错[7]。与深度伪造不同, 语音对抗样本旨在通过对载体信号引入微小的扰动, 使语音识别或声纹识别系统出现特定的差错, 但并不影响人耳对该语音样本的听觉感知[8]。相对于深度伪造, 语音对抗样本的攻击具有很强的隐蔽性。本文选取语音的文本内容识别、声纹身份识别两个典型任务, 按照从白盒攻击到黑盒攻击、从数字攻击到物理攻击、从特定载体到通用载体的从易到难、逐步贴近实际场景的方式, 系统地梳理了语音对抗样本的攻击方法。

与此同时, 为了应对对抗样本带来的安全挑战[9], 研究者有针对性地提出了许多防御方法, 例如通过检测与识别对抗样本重新划分分类边界、借助数据增强加固语音或声纹识别器、利用对抗扰动带来的异常特征鉴别语音是否可信等。本文从统计学习中分类器边界构造的角度, 对语音对抗样本的防御方法进行分类论述, 揭示各类方法实现防御的深层机理。最后, 结合现阶段对抗样本攻击和防御方法存在的不足, 预测该领域未来的发展方向。

2 语音对抗样本攻击

语音对抗样本攻击可以分为无目标攻击和有目标攻击。无目标攻击迫使模型产生不同于其真实标签的预测输出, 而有目标攻击迫使模型预测输出攻击者指定的目标标签。根据攻击者对被攻击模型信息掌握的程度, 可以分为: 白盒攻击和黑盒攻击。根据对抗样本攻击方式的不同, 可以分为: 数字攻击和物理攻击。根据对抗扰动的适用范围, 还可以分为: 基于特定载体的攻击和基于通用载体的攻击。本文从白盒攻击与黑盒攻击、数字攻击与物理攻击和特定载体与通用载体三个方面, 介绍几种在语音对抗样本攻击发展史上比较有代表性的攻击方法, 并详细介绍各类攻击方法的原理, 如表 1 所示。

表 1 语音对抗样本的攻击方法对比

Table 1 Comparison of attack methods of voice adversarial examples

攻击方法	攻击类型	攻击场景	扰动对象	物理传播	被攻击模型	目标系统	适用范围
Gong [11]	无目标	白盒	波形	否	WaveCNN	ASI	单个
Kreuk [12]	有目标	白盒/黑盒	MFCC/Mel-Spectrum	否	LSTM	ASV	单个
Iter [13]	有目标	白盒	MFCC	否	WaveNet	ASR	单个
Cisse [15]	有目标/无目标	白盒/黑盒	波形	否	DeepSpeech2/Google Voice	ASR	单个
Li [16]	有目标/无目标	白盒/黑盒	LPMS/MFCC	否	i-vector GMM/x-vector	ASV	单个
Villalba [17]	有目标/无目标	白盒/黑盒	波形	否	x-vector ResNet/ResETDNN/ThinResNet34	ASV	单个
Zhang [18]	有目标	黑盒	波形	否	LCNN/AFNet/ResNet/SEResNet	ASV	单个

续表

攻击方法	攻击类型	攻击场景	扰动对象	物理传播	被攻击模型	目标系统	适用范围
Khare [19]	有目标/无目标	黑盒	MFCC	否	Deepspeech/Kaldi	ASR	单个
Schönherr[21]	有目标	白盒	波形	否	Kaldi	ASR	单个
Schönherr[22]	有目标	白盒	波形	是	Kaldi	ASR	单个
Qin [23]	有目标	白盒	波形	是	Lingvo	ASR	单个
Vaidya [24]	有目标	白盒	MFCC	否	Google Now app	ASR	单个
Carlini [25]	有目标	白盒/黑盒	MFCC	是	CMU Sphinx/Google Now app	ASR	单个
Yuan [26]	有目标	白盒/黑盒	波形	是	Kaldi/iFLYTEK	ASR	单个
Yakura [27]	有目标	白盒	波形	是	DeepSpeech	ASR	单个
Chen[28]	有目标	白盒	波形	是	DeepSpeech	ASR	单个
Chen[29]	有目标	黑盒	波形	是	Google Cloud/Amazon Transcribe/Microsoft Bing/IBM API; Google Home/Amazon Echo/Google Assistant/Microsoft Cortana/Apple Siri IVC	ASR	单个
Carlini [30]	有目标	白盒	波形	否	DeepSpeech	ASR	单个
Du [31]	有目标	白盒/黑盒	波形	否	DeepSpeech/CNN/VGG 19/DenseNet/ResNet18/ResNeXt/ WideResNet18/DPN92	ASR	单个
Neekhara [32]	无目标	白盒/黑盒	波形	否	DeepSpeech/WaveNet	ASR	通用
Li [33]	有目标/无目标	黑盒	波形	否	SincNet	ASI	通用
Xie [34]	有目标	白盒	波形	是	x-vector DNN	ASI	通用

2.1 白盒攻击与黑盒攻击

在白盒攻击中, 攻击者完全了解被攻击模型的结构、参数、损失函数和梯度等信息, 利用被攻击白盒模型的结构和参数信息构建对抗样本生成算法, 从而有指导性地修改原始样本, 以生成对抗样本。在黑盒攻击中, 攻击者不掌握被攻击模型的结构、参数等内部信息, 只能通过利用白盒模型对抗样本的迁移性, 或利用黑盒模型的输出结果训练替代模型等方式来生成对抗样本。在现实场景中, 攻击者难以获取被攻击模型的内部信息, 因此黑盒攻击相对于白盒攻击难度更高, 但也更符合实际。

比较常用的对抗样本生成方法是快速梯度符号法(Fast Gradient Sign Method, FGSM)[10]。FGSM 最早用于生成图像对抗样本, 后来的研究表明它也可以用于生成语音对抗样本。利用目标网络损失函数的梯度信息, 沿着梯度方向以固定步长逐渐增大损失函数, 使模型给出错误的分类结果, 通过迭代修改, 可以快速生成对抗样本。

Gong 等人[11]提出了基于 FGSM 产生语音对抗扰动的方法:

$$\delta = \varepsilon \text{sign}(\nabla_x J(\theta, x, t')) \tag{1}$$

其中, δ 表示对抗扰动, ε 用来控制扰动的幅度, θ 是模型参数, x 是模型输入, t' 是攻击者期望的分类标签, $J(q,x,t')$ 是神经网络的损失函数。该方法通过直接扰动原始波形, 避免了在特征域引入扰动再转换回波形域所带来的额外损失。他们用这种方法生成的对抗样本攻击一个基于卷积神经网络的声纹识别系统。实验结果显示, 当扰动系数 ε 为0.032时(相对于波形幅度值, 该值很小), 系统识别错误率从29%提高到了44%。然而, 该方法的效果依赖于白盒假设, 即假定攻击者完全知晓被攻击模型的网络结构。

Kreuk 等人[12]用 FGSM 生成对抗性的声学特征, 并从对抗性的声学特征中重建语音波形, 用于攻击一个端到端的声纹认证模型。他们的实验使用 YOHO 和 NTIMIT 两个数据集、梅尔频谱(Mel-Spectrum)和梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)两种声学特征。在白盒场景下, 使用不同特征和数据集的攻击都使系统的识别准确率降低到 37.5%以下, 其中利用 MFCC 特征的攻击比 Mel-Spectrum 特征下降得更严重。他们还设法摆脱白盒假设在黑盒场景中进行实验, 在 YOHO 数据集上的实验结果表明, 利用语音 Mel-Spectrum 特征生成的对抗样本, 攻击使用 MFCC 特征训练好的声纹认证模型, 使该声纹认证模型的准确率从

81.00%下降到了 62.25%。

除了上述针对声纹识别系统的攻击, Iter 等人[13]用 FGSM 和欺骗梯度法(Fooling Gradient Method)对 Kim 和 Park 设计的语音识别系统[14]进行了白盒攻击。欺骗梯度法是在训练过程中基于输入样本而不是网络模型参数来计算梯度, 利用梯度下降法以迭代的方式对输入样本施加小幅度的扰动, 使网络输出不正确的目标。他们在实验中使用 VCTK 语料库生成对抗样本, 多次迭代后的结果表明, 对单个词或整个句子的攻击会造成个别单词的拼写错误。该方法虽然使模型出现转录错误, 但并不能总是保证输出语法正确的目标单词或句子。而且经过多轮迭代后, 对原始单词或句子的预测也出现了很多拼写错误。

实际上, 利用对抗样本的可迁移性, 在白盒条件下得到的对抗样本是可以直接用于攻击黑盒模型的。Cisse 等人[15]提出了一种名为 Houdini 的方法来攻击语音识别系统。在 LibriSpeech 数据集上, 对 DeepSpeech2 模型进行白盒攻击, 对于不同扰动幅度下的无目标攻击, Houdini 都具有较好的攻击效果, 但在有目标攻击情况下效果不佳。该项研究还对 Houdini 的黑盒攻击效果进行了简单地验证, 用基于 DeepSpeech2 模型生成的对抗样本攻击 Google Voice, 发现无目标攻击的成功率较高, 表明攻击效果从白盒到黑盒具有一定的可迁移性。

对黑盒攻击来说, 不同白盒模型生成的对抗样本对黑盒模型的攻击效果有很大差异。当对特定黑盒模型进行攻击时, 设法找到迁移性最好的白盒模型用于生成对抗样本最为关键。Li 等人[16]研究了不同特征和模型结构之间攻击方法的可迁移性。实验设计了三种系统: 基于 MFCC 的高斯混合模型 i-vector(MFCC i-vector)、基于对数功率幅度谱(Log Power Magnitude Spectrum, LPMS)的高斯混合模型 i-vector(LPMS i-vector) 和基于 MFCC 的 x-vector(MFCC x-vector)。攻击场景包含三种: (1)用 LPMS i-vector 生成的对抗样本攻击 MFCC i-vector; (2)用 MFCC i-vector 生成的对抗样本攻击 MFCC x-vector; (3)用 LPMS i-vector 生成的对抗样本攻击 MFCC x-vector。实验结果表明, 三种攻击设置下的错误接受率(False Acceptance Rate, FAR)都显著增加, 最大增幅分别约为 60%、50%和 30%, 改变特征的增幅大于改变模型的增幅, 这表明模型改变后攻击效果会变得更差。

Villalba 等人[17]尝试将他们的攻击从小规模的白盒系统迁移到大规模的黑盒系统。他们分别在 ThinResNet34 和 ResETDNN x-vector 模型上用迭代快速梯度符号法(Iterative Fast Gradient Sign Method, I-FGSM)生成对抗样本, 然后攻击 ResNet34 模型。实验结果表明黑盒攻击的效果明显不如白盒, 只有在信号和扰动噪声的比值较低时才可实现有效的攻击。

Zhang 等人[18]将他们提出的一种迭代集成方法(Iterative Ensemble Method, IEM)与动量迭代快速梯度符号法(Momentum Iterative Fast Gradient Sign Method, MI-FGSM)相结合, 在白盒状态下生成可以同时欺骗多个模型的对抗样本, 增强了对抗样本攻击在黑盒模型上的可迁移性。实验使用 ASVspoof2019 数据集的 LA 部分, 评估了对 LCNN、SENet50、Resnet34 和 AFNet 4 个模型的黑盒攻击。对每个模型的黑盒攻击, 使用其他 3 个模型在白盒状态下集成生成对抗样本。实验发现, 基于 IEM 的 MI-FGSM 攻击成功率比单独利用 MI-FGSM 的攻击提高了 4%~30%。

Khare 等人[19]提出利用多目标进化的优化方法对语音识别系统进行黑盒攻击, 优化目标是使文本转录错误, 同时带有扰动的语音与原始语音仍然保持高度相似性。他们增大目标文本与原始文本的编辑距离, 同时最小化对抗样本和真实样本 MFCC 特征之间的欧几里得距离。基于进化的方法是对自然选择的模仿, 在创建了一群候选的对抗样本后, 具有较高适应度的对抗样本更有可能变异并成为下一代的一部分, 重复这个进化过程直到得到一个最佳的结果。通过向原始语音信号添加随机均匀噪声来初始化种群, 然后用每个候选种群的两个目标计算适合度分数, 那些得分较高的种群更倾向被选择进行交叉和突变。他们用这种方法对 Deepspeech 和 Kaldi[20]两个语音识别系统进行无目标攻击, 单词错误率(Word Error Rate, WER)分别升高了 980%和 368%, 同时对抗样本与原始音频保持了 98%的音频相似性, 取得了较好的黑盒攻击效果。

2.2 数字攻击与物理攻击

根据语音对抗样本是否经由空气播放, 可以将攻击方法分为数字攻击和物理攻击。数字攻击是指对抗样本以数字信号的形式进入识别系统, 而物理攻击是指对抗样本由扬声器播放、经空气传播、被麦克风接收进入识别系统。显然, 物理攻击过程中, 信号经过硬件设备的处理会出现失真, 声音在空气中传播也会受到空间中环境噪音的干扰, 攻击的难度比数字攻击更大。即使模拟真实环境实施攻击, 也存在与真实环境差异大、计算量大、无法实时等问题。

Schönherr 等人[21]设计了一种基于心理声学隐藏的针对语音识别系统的攻击。他们利用 MP3 的编码原理, 将对抗扰动隐藏在听力阈值以下, 使其几乎不被察觉。同时, 将音频预处理集成到梯度反向传播过程中, 以增强攻击效果。在对 Kaldi 识别系统的白盒攻击中, 可以在不到两分钟的时间为 10 s 的语音文件生成一个对抗样本。他们在实验中将迭代次数设置为小于 500, 听力阈值的允许偏差范围限定在 20 dB 以内, 缩短攻击时间的同时实现了较高的攻击成功率。然而, 这种攻击是以语音文件的形式进入到识别系统实施的数字攻击, 并不能在空间环境中播放实施物理攻击。

在上述数字攻击的基础上, Schönherr 等人[22]继续研究了适用于不同物理环境的对抗样本生成方法。他们将物理传输过程建模为原始音频信号与房间脉冲响应(Room Impulse Responses, RIR)的卷积, 利用不同的房间特性和麦克风与扬声器的不同位置, 对生成的对抗样本进行优化以使其具有更强的鲁棒性:

$$x' = \arg\max_{x} E_{h\sim H_\theta}[P(y' \mid x * h)] \tag{2}$$

其中, RIR 参数 h 服从 H_θ 分布, $*$ 代表卷积操作, x 为原始音频样本, x' 是生成的对抗样本, y' 是目标类别。在白盒场景下对 Kaldi 语音识别系统进行有目标攻击, 允许对信号的修改幅度超过标准听力阈值 20 dB。实验表明, 生成的对抗样本可以有效适应混响时间较长或扬声器和麦克风之间距离较大的场景, 可以在不同的房间和声音环境中较成功地进行物理攻击, 说明在没有任何先验知识的情况下他们的方法对复杂的声学环境具有一定的适应性。

Qin 等人[23]针对名为 Lingvo 的语音识别系统生成了可以经过物理传播的不可察觉的语音对抗样本。为了实现听觉上的不可察觉, 他们利用了听觉掩蔽原理。在音频域中每个频率附近存在一个“掩蔽阈值”, 任何低于这一阈值的信号实际上都是不可察觉的。利用该方法求解对抗样本需要同时优化两个部分: (1)优化扰动以欺骗待攻击的识别网络; (2)优化扰动以确保人类听觉无法感知。该优化问题用下面的公式表示:

$$\begin{aligned} &\min_{\delta} \quad l(x,\delta,y) = l_{net}(f(x+\delta),y) + \alpha \cdot l_\theta(x,\delta) \\ &\text{s.t.} \quad \|\delta\| < \varepsilon \end{aligned} \tag{3}$$

其中, l_{net} 是交叉熵损失函数, 保证语音对抗样本可以成功欺骗识别系统, l_θ 将扰动的归一化功率谱密度估计值限制在原始音频的频率掩蔽阈值之下, ε 控制扰动 δ 的最大幅度。为了提高对抗语音经过物理传播的攻击鲁棒性, 他们在实验中使用声学房间模拟器来生成反映物理传播规律的带有混响的语音。声学房间模拟器根据房间配置(房间尺寸、源语音和目标麦克风的位置以及混响)创建房间脉冲响应 r, 然后将生成的房间脉冲响应 r 与输入语音 x 卷积, 以获得具有混响的语音 $t(x)=x*r$。将模拟房间混响的变换 $t(x)$ 引入到损失函数中, 使 $f(t(x+\delta))$逼近 y, 生成经过物理传播后仍有效的对抗样本。实验使用 LibriSpeech 数据集, 在测试阶段采用与训练阶段具有相同房间混响概率分布的变换 $t(x)$, 利用上述方法生成对抗样本对 Lingvo 模型的有目标攻击成功率在 60%以上, 说明这种方法在模拟的物理环境中比较有效。需要说明的是, 这种攻击假定攻击者完全了解识别模型的信息, 仍是一种白盒攻击, 并且使用的是模拟的物理环境, 并不是在真实物理空间进行的攻击。

从数字攻击到物理攻击的发展过程中, 对声学特征参数的扰动也进行了尝试。Vaidya 等人[24]提出了一种生成语音对抗样本的方法, 通过不断微调提取的 MFCC 参数使语音识别系统识别错误, 然后将对抗性的 MFCC 特征重构为语音波形实施数字攻击, 但这种攻击方式无法进行物理攻击。在此基础上, Carlini 等人[25]对 Vaidya 的方法进行了完善, 使其可以经过物理传播, 提高了 Vaidya 攻击的实用性。他们通过增加一个对信号二阶导数的惩罚项, 来减缓由梯度下降产生的强烈的信号突变, 使扬声器膜的震动频率可以跟得上信号变化。在 MFCC 的计算公式 $\boldsymbol{C}\log(\boldsymbol{B}\|\boldsymbol{A}y\|^2)$ 中引入矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ 模拟扰动在物理传播过程中的变化量, 利用最小二乘法预测对抗样本 MFCC 的系数矩阵, 使其转换为波形后更适合通过扬声器播放。训练迭代过程中, 用扬声器播放语音并计算得到对抗样本的 MFCC, 对比该 MFCC 和目标 MFCC 之间的差异, 然后调整该 MFCC, 利用梯度下降法训练得到与目标 MFCC 更接近的 MFCC。这种方法求解的对抗样本考虑了语音播放的实际过程, 提高了物理攻击的成功率, 但迭代寻找对抗样本的时间成本较高。

在训练过程中人为引入噪声也可以提高物理攻击的成功率。Yuan 等人[26]设计了一种名为 CommanderSong 的针对语音识别工具 Kaldi 的对抗攻击, 它将语音命令嵌入到歌曲中播放, 使人耳听不出对抗扰动的存在。同时, 将随机噪声引入到对抗样本生成的过程中, 使该方法在物理环境中对不同的播放和接收设备具有鲁棒性。CommanderSong 利用一种概率密度函数标志符(Probability Density Function Identifier, PDF-ID)的序列匹配算法来生成对抗样本:

$$\begin{aligned}&\min_{\mu(t)} \quad \left\|g(x(t)+\mu(t)+n(t))-b\right\|_1\\&\text{s.t.} \quad \left|\mu(t)\right|\leqslant\varepsilon\end{aligned} \tag{4}$$

其中, $g(\cdot)$代表深度神经网络(Deep Neural Network, DNN)最有可能输出的 PDF-ID 预测, $x(t)$是原始歌曲, $\mu(t)$表示添加到原始歌曲中的扰动, $n(t)$表示引入的随机噪声, b 表示语音命令的 PDF-ID 序列, $\|\cdot\|_1$ 表示原始歌曲和对抗样本之间 PDF-ID 序列的 l_1 距离, 利用迭代优化来寻找合适的最小扰动 $\mu(t)$, ε 表示添加扰动的最大范围。实验结果显示, 这种攻击对不同语音命令和不同扬声器设备的效果有所差别, 在 Kaldi 平台上的攻击最高具有 96%的成功率, 当迁移到 iFLYTEK 上时个别语音命令的攻击成功率下降为 0。

为了实现可以通过物理播放的攻击, Yakura 等人[27]将物理世界中由回放和录音引起的变换集成到对抗样本生成过程中, 从而希望获得鲁棒性更强的对抗样本。他们通过在 3 个优化问题中分别引入带通滤波、脉冲响应和高斯白噪声, 减轻了设备噪声、环境混响和背景噪声带来的影响, 从而增强语音对抗样本在物理环境中播放的攻击效果。该实验在低信噪比条件下对他们设定的 3 条目标文本可以实现 100%的攻击成功率, 但对不同识别模型和任意目标文本的有效性尚不明确。

Chen 等人[28]通过考虑多径传播和硬件设备的频率选择性提出了一种可以适应不同播放设备和环境的对抗样本生成方法 Metamorph。该方法包括“生成”和“清除”两个阶段。在“生成”阶段, 收集少量不同信道和硬件设备产生的信号失真测量值作为先验数据集, 利用这些测量值对传播路径和频率选择性的影响, 生成初始扰动 δ; 在“清除”阶段, 利用域自适应算法对特定设备的特征进行补偿, 同时最小化来自这些失真测量的不可预测的相关环境特征, 以进一步提高攻击的成功率。实验结果表明, 在充斥多路径传播的办公室场景, Metamorph 在 6 m 直线距离内的攻击成功率在 90%以上; 即使在限定扰动幅度以使人类听觉不易察觉时, 该方法在 3 m 距离范围内的攻击成功率也超过了 90%。

Chen 等人[29]在黑盒条件下对商业智能语音控制(Intelligent Voice Control, IVC)设备进行了攻击。他们首先训练一个基于 Kaldi ASpIRE Chain 的语音识别通用基础模型, 然后用少量的数据训练一个替代模型 Mini Librispeech, 接下来用基础模型对替代模型进行改进, 使替换模型生成的对抗样本对被攻击的目标模型具有高度的可迁移性。实验结果表明, 对基于 Google Home Mini、Amazon Echo First Gen、Google Assistant App 和 Microsoft Cortana App 的 IVC 设备进行攻击, 平均攻击成功率可以达到 98%, 证实在物理场景下用该方法攻击真实的语音控制设备是有效的。

2.3 特定载体与通用载体

根据扰动的适用范围, 对抗样本攻击可以分为基于特定载体的攻击和基于通用载体的攻击。特定载体攻击的对抗扰动只适用单个语音样本, 不同的语音样本实现对抗攻击则需要添加不同的扰动, 例如以下 Carlini & Wagner 和 Du 等人的工作。

Carlini & Wagner[30]通过一种直接修改原始音频波形的优化方法生成特定载体对抗样本, 该优化问题可以用以下公式表示:

$$\begin{aligned}&\min_{\delta} \quad \left\|\delta\right\|_2^2+c\cdot l(x+\delta,t)\\&\text{s.t.} \quad dB(\delta)<\varepsilon\end{aligned} \tag{5}$$

其中, c 体现了攻击效果和扰动程度之间的权衡, t 表示目标文本, dB 用来度量扰动的强度, ε 是约束扰动 δ 的常数。实验结果表明, 对抗语音在 DeepSpeech 上有 100%的攻击成功率, 且与原始音频有 99%的相似性。然而, 这种攻击需要针对每一个样本求解一个优化问题, 也就是要为每个样本单独设计一个合适的对抗扰动, 当载体样本不断变化时实施攻击的计算量较大。

Du 等人[31]提出了一种名为 Siren 的攻击来为每一个语音样本单独生成对抗扰动。在白盒攻击中, 先利用修改过的粒子群优化(Particle Swarm Optimization, PSO)算法找到对抗扰动的粗粒度扰动, 再用欺骗梯度方法对粗粒度扰动进行修正来逼近精确的对抗扰动。在黑盒攻击中, 利用与白盒攻击不同的目标函数与终止条件, 用一步 PSO 算法求解对抗扰动。实验结果表明, 使用 Common Voice 数据集和 VCTK 语料库对 DeepSpeech 模型进行白盒攻击, 分别在平均不到 1600 s 和 1900 s 的时间生成具有 100%攻击成功率的语音对抗样本, 使用 Speech Commands 数据集、Synthesized Commands 数据集和 IEMOCAP 数据集对他们提出的几个卷积神经网络模型进行了黑盒攻击, 可以实现 83.60%以上的攻击成功率。

通用载体攻击的对抗扰动适应于所有语音样本, 即任何语音样本只要添加上这个扰动, 都会变成能达到攻击目的的对抗样本。通用载体的适用范围广、部署容易, 更有利于快速发起实时攻击, 但扰动的幅度通常较大。在载体选择的研究方面, 也是遵循了先从白盒攻击入手再迁移到黑盒攻击场景的原则, 如下所述。

为了缩短对抗样本生成时间, 使攻击可以实时进行, Neekhara 等人[32]经过研究发现, 在音频域存在通用的不可察觉的对抗性扰动, 对任意输入语音添加通用对抗性扰动(Universal Adversarial Perturbations, UAPs), 可以欺骗基于 DNN 的自动语音识别系统。满足要求的通用扰动 v 用下面的公式优化得到:

$$\begin{aligned} &\min_{v} \quad c\|r\|^2 - \mathrm{CTCLoss}(C(x_i+v+r), C(x_i)) \\ &\text{s.t.} \quad \|v+r\|_\infty < \varepsilon \end{aligned} \tag{6}$$

其中, C 表示语音识别模型, $C(x)$为输入语音 x 的识别结果, ε 是在 l_∞范数下的最大扰动范围, 在每个样本的初始扰动 v 上迭代地添加扰动 r, 最大化 $C(x_i+v+r)$与 $C(x_i)$之间的连接主义时序分类(Connectionist Temporal Classification, CTC)损失, 进而找到适用于所有样本的通用扰动。通过配置合适的参数 c 可以在给定扰动幅度下实现最大的攻击成功率。他们在白盒场景下得到的通用扰动, 在 DeepSpeech 上表现出良好的性能, 攻击成功率可达到 89.06%, 并在基于 WaveNet 的自动语音识别模型[35]上验证了攻击的可迁移性。但该实验主要是针对无目标攻击的, 即当识别的文本输出有超过自身长度 50%的字符删除(Deletion)、插入(Insertion)或替换(Deletion)错误时定义为攻击成功, 所以它在有目标攻击上的效果还不明确。

在 Neekhara 等人研究的基础上, Li 等人[33]提出了一种针对声纹识别系统的通用扰动生成方法, 可以生成不同的但具有相似攻击效果的通用扰动。该方法与输入样本无关, 适用于整个数据集, 通过生成模型学习从低维正态分布到通用扰动子空间的映射来产生通用扰动。在给定语音 s 及其说话人标签 y 的情况下, 对声纹识别模型的无目标攻击可以表示为:

$$\begin{aligned} &\min_{s'} \quad l(s, s+\delta) \\ &\text{s.t.} \quad f(s+\delta) = y' \\ &\qquad y' \neq y \end{aligned} \tag{7}$$

其中, $\delta=G_\theta(z)$, 利用生成模型 $G_\theta(z)$从噪声 z 生成通用扰动 δ, y' 是对抗样本 $s'=s+\delta$ 的预测标签, f 是声纹识别模型, $l(\cdot)$是度量原始信号和对抗样本之间差异的距离函数。对于有目标攻击, 将公式中的 $y' \neq y$ 修改为 $y'=y_t$ 即可, 其中 y_t 表示攻击的目标类别。他们在 TIMIT 和 LibriSpeech 数据集上用 SincNet 声纹识别模型进行了实验验证。用生成的通用扰动进行无目标攻击, 在 TIMIT(LibriSpeech)数据集上, 在 49.87dB(31.15dB)的信噪比和 3.00(2.33)的语音感知质量评估(Perceptual Evaluation of Speech Quality, PESQ)分数值下, 可以使错误率达到 97%(96%)以上。用生成的通用扰动进行有目标攻击, 在 TIMIT(LibriSpeech)数据集上, 在 48.53dB(29.94dB)的信噪比和 2.48(2.11)的 PESQ 分数值下, 可以使攻击成功率达到97.2%(64.1%)以上。虽然在LibriSpeech上的效果略差, 但该方法提供了一种有目标攻击条件下寻找通用对抗扰动的有效手段。

同样针对声纹识别系统的有目标攻击, Xie 等人[34]提出使用定长通用噪声的重复回放来适应不同长度的输入语音进而构造通用扰动, 通过估计物理传播带来的声音失真使对抗样本经过播放和接收过程仍然有效。他们通过以下目标函数迭代地求解对抗样本:

$$\arg\max_{i} \left(P_i\left((x+\delta)*r\right)\right) = t \tag{8}$$

其中, t 是目标说话人标签, *表示卷积运算, $(x+\delta)*r$ 是通过将对抗样本与环境变量 r 进行卷积来模拟麦克风录音得到的对抗语音, $P_i(x)$为将 x 分类为第 i 个说话人的概率, 扰动 δ 通过裁减操作被限制在区间 $(-\varepsilon, \varepsilon)$内以减少对抗样本和原始语音之间的失真。为了产生通用的扰动, 他们用整个训练数据集迭代修改用于估计 δ 的更新序列$\Delta\delta$, 直到满足期望的攻击成功率。对于每条训练语音, 如果目标类的预测概率大于其他类, 则在该训练语音上跳过针对扰动$\Delta\delta$ 的更新。他们在白盒场景下尝试攻击一个基于 x-vector 的声纹识别系统, 攻击实施的时间仅为 0.015 s 左右。在 CSTR 数据集上的实验表明, 当扰动与语音载体的功率比为−18.84 dB 时, 可以达到平均99.95%的攻击成功率; 当进一步降低扰动信号的幅度到−33.96 dB 的功率比时, 平均攻击成功率仍在80%以上。

3 语音对抗样本防御

为了消除对抗样本的影响, 使识别系统免遭欺骗, 人们开始研究并提出了一系列对抗样本的防御方法。本节从分类边界构造的角度, 对语音对抗样本的防御方法进行分类论述, 揭示各类方法实现防御的机理。图 1 归纳了三类典型的对抗样本防御方法: 1)在识别系统前施加额外的对抗样本检测器(如图 a 虚线所示); 2)在训练阶段通过对抗训练加固原分类器, 即将分类边界移动至可以区分出对抗样本的新区域(如图 b 所示); 3)根据对抗样本和真实样本经过变换处理后类别预测结果的异同, 使系统发现对抗样本, 减小对抗样本带来的影响(如图 c 所示)。其中, 前两类方法本质上是通过修改分类模型来增强对对

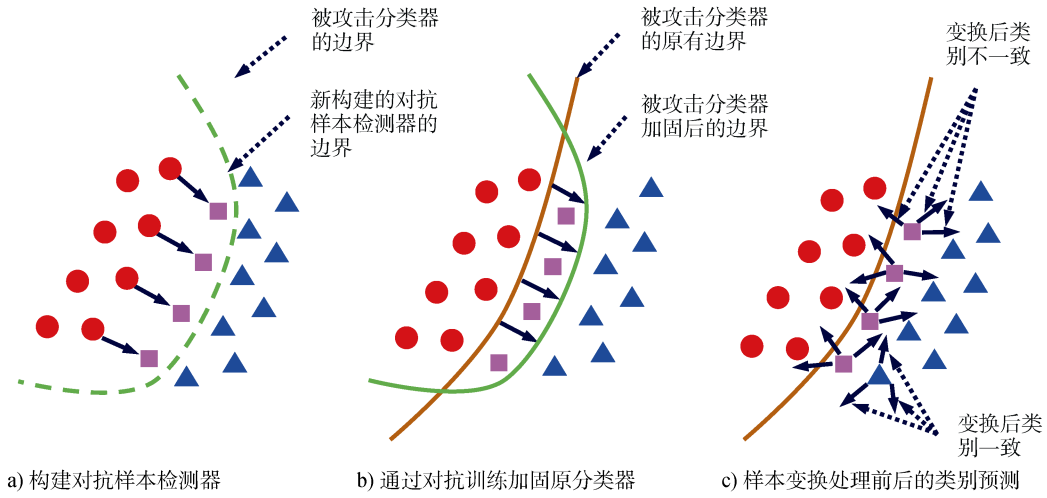

图 1　三类典型的对抗样本防御方法

Figure 1　Three typical methods of the defense against voice adversarial examples

抗样本的检测和识别能力, 因此我们下文统一表述。表 2 展示了本文对防御方法的分类以及不同防御方法的工作机理。

3.1　通过修改分类模型提高检测与识别能力

对抗样本是在真实样本上添加了扰动, 即使这个扰动很小以至于听起来难以分辨, 也会使对抗样本表现出与真实样本不同的特性。人们基于对抗样本的统计特性, 提出了构建对抗样本检测器和通过对抗训练加固原分类器的方法。构建对抗样本检测器利用深度神经网络学习真实样本和对抗样本之间的差异, 通过对输入的语音进行检测, 将对抗样本隔离在识别系统之外。通过对抗训练加固原分类器是在模型训练时引入对抗样本, 把对抗样本也一并加入到目标优化过程, 使系统具有将对抗样本归类为真实类别的能力, 提高系统鲁棒性。

表 2　语音对抗样本的防御方法

Table 2　Defense methods against voice adversarial examples

防御方法		特点
通过修改分类模型提高检测与识别能力	构建对抗样本检测器	对真实样本和对抗样本进行检测, 需构建额外的分类器
	通过对抗训练加固原分类器	训练过程中引入对抗样本, 对特定的攻击较为有效
借助样本变换前后识别结果的异同		在样本中添加噪声或对样本进行压缩、平滑、重构等处理, 根据识别结果的异同鉴别对抗样本, 降低对抗样本的效能

3.1.1　构建对抗样本检测器

构建对抗样本检测器可以通过训练一个单独的神经网络分类模型, 实现对抗样本和真实样本的检测。这是一种在对抗样本进入识别模型前对其进行检测的防御方法。

Samizade 等人[36]用一个卷积神经网络(Convolutional Neural Networks, CNN)检测对抗样本中的微小扰动, 其网络结构由三个核大小为 2*2 的卷积层、卷积层之间的池化层和一个全连接层组成。前两个卷积层的输出通道数为 64, 采用 Relu 激活函数; 第三个卷积层的输出通道数为 32, 采用 Selu 激活函数; 全连接层含有 128 个神经元, 最后通过 Softmax 激活函数得到识别结果。这个 CNN 结构对 MFCC 特征的微小变化较为敏感, 可以检测出对抗扰动。实验结果表明, 对 Carlini 攻击[30]和 Alzantot[37]攻击, 经过训练后的模型检测精度可以达到接近 100%, 但当使用上述任意一种方法生成的对抗样本攻击用另外一种方法生成的对抗样本训练的检测网络时, 检测性能会显著下降, 即对抗样本检测器的迁移能力较弱。

Li 等人[38]提出了一种对抗样本分类检测的方法, 通过引入一个类似 VGG 网络结构的检测网络, 利用卷积操作提取对抗样本和真实样本之间的细微偏差。具体的网络结构是: 底部有 4 个卷积层, 用于提取局部特征; 统计池化层聚集最后一个卷积层输出的平均值和偏差, 并将它们传递到全连接层; 两个全连接层将统计数据投影到一个二维输出空间中进行决策。实验结果表明, 对一些特定的攻击方法, 对

抗扰动的幅度越大被检测出来的概率越大。在小幅度扰动对抗样本上训练的检测网络可以很好地防御大幅度扰动的攻击; 反之, 效果较差。该检测网络可以较好的防护基于 i-vector 和 x-vector 的声纹识别系统; 然而, 当攻击者更换攻击方法时, 防御性能会严重下降。

3.1.2 通过对抗训练加固原分类器

基于构建对抗样本检测器的防御方式, 虽然在一定程度上将对抗样本隔离在识别系统之外, 但并没有提高识别系统的鲁棒性。基于对抗训练的防御方法可以提高识别模型本身的鲁棒性, 直接将对抗样本映射为其真实类别。

对某个特定的攻击方式, 使用训练数据和当前的模型参数生成对抗样本。对抗训练就是在以特定方式生成的对抗样本上训练模型, 使其对这种攻击方式具有更好的鲁棒性。Madry 等人[39]通过下式表示的 min-max 优化给出了对抗训练的一般定义:

$$\begin{aligned} &\min_{\theta} \quad E_{(x,y)\sim D}\left[\max_{\delta} \ l(x+\delta,y,\theta)\right] \\ &\text{s.t.} \quad \|\delta\|_p < \varepsilon \end{aligned} \tag{9}$$

其中, 目标函数内部的 max 优化问题利用对抗训练期间使用的攻击算法来求解, 外部的 min 优化目标则是通过重新调整模型参数 θ 以最小化总体损失函数。

Jati 等人[40]利用基于 FGSM 和投影梯度下降(Project Gradient Descent, PGD)[39]的对抗训练来防御针对声纹识别系统的对抗攻击, 用一步 FGSM 和 10 次迭代的 PGD(PGD-10)方法解决内部最大化问题。整个训练过程在干净样本 x 和对抗样本 $x+\delta$ 上实现, 损失函数由下式给出:

$$\begin{aligned} &\max_{\delta} \quad (1-\omega_{AT})\cdot l(x,y,\theta)+\omega_{AT}\cdot l(x+\delta,y,\theta) \\ &\text{s.t.} \quad \|\delta\|_p < \varepsilon \end{aligned} \tag{10}$$

其中, ω_{AT} 是对抗训练的权重。实验是在 Librispeech 数据集的 Train-Clean-100 子集上进行的, 识别系统由一个提取对数梅尔频谱的信号处理前端和一个带有 8 个卷积层的卷积神经网络组成, 最大扰动强度设置为 ε=0.002。实验结果显示, 在面对 FGSM、Carlini-Wagner(CW)[41]和 100 次迭代 PGD(PGD-100)的 l_∞攻击时, 相对于没有采取任何防御措施的情况, 基于 PGD-10 的对抗训练比基于 FGSM 的对抗训练更能显著改善该识别系统的性能, 分别获得了 73%、58%和 43%的识别准确率。但是, 在面对 CW 的 l_2 攻击时, FGSM 对抗训练和 PGD-10 对抗训练的防御效果都很差, 识别准确率几乎为 0。对抗训练也会对原始模型带来不利影响: 利用 FGSM 和 PGD-10 对抗训练加固后的模型, 在干净样本上的识别准确率都有所下降, 其中基于 FGSM 的对抗训练下降得更为明显。

对抗训练防御算法的成功与否很大程度上取决于引入的对抗扰动的质量, 用不同的、更强的对抗性扰动优化 min-max 方程中的内部极大值, 可以使语音识别模型变得更加鲁棒。Pal 等人[42]提出了一种基于混合对抗训练的防御机制, 与现有声纹识别系统中对抗攻击只使用交叉熵损失函数不同, 他们利用特征散射(Feature Scattering, FS)[43]生成对抗样本, 在对抗训练阶段, 结合交叉熵损失(Cross-Entropy Loss)、特征分散损失(Feature Scattering Loss)和边际损失(Margin Loss)的信息构成总的损失函数, 以便制造多样化的更强的扰动。其中, 用于内部 max 优化问题的损失函数如下:

$$\begin{aligned} &\max_{\|x_{adv}-x\|_\infty} \beta\cdot l_{CE}(x_{adv},y)+\gamma\cdot l_{FS}(x,x_{adv})+\varsigma\cdot l_M(x_{adv},y) \\ &\text{s.t.} \quad \|x_{adv}-x\|_\infty < \varepsilon \end{aligned} \tag{11}$$

其中, β, γ 和 ς 是可调节的超参数, 实验中均设置为 1。该实验训练和测试均为 l_∞攻击, 最大扰动强度 ε=0.002, 采用与 Jati 等人[40]的实验相同的数据集和识别模型结构。实验结果表明, 分别用 FGSM、PGD、CW 和 FS 方法生成的对抗样本发起无目标攻击, FGSM 对抗训练仍然是最弱的防御方式。与基于单一对抗样本生成方式的对抗训练相比, 引入多目标任务的混合对抗训练的防御效果更好。即使在增加迭代次数以获得更强攻击的情况下, 混合对抗训练的防御效果也优于其他对抗训练方法, 在 PGD-40, CW-40, FS-40 攻击下的识别准确率分别保持在 78.84%、78.52%和 96.67%。对于黑盒攻击, 他们分别用原始模型和 FS-10 对抗训练模型生成 PGD-40 和 CW-40 对抗样本, 攻击基于 PGD-10 对抗训练的防御模型和基于混合对抗训练的防御模型, 结果表明混合对抗训练的防御效果更好。对基于 l_1、l_2 范数攻击的效果尚不明确, 经过训练后的模型对干净样本的识别准确率同样有所下降。

3.2 借助样本变换前后识别结果的异同

基于构建对抗样本检测器的防御方法和通过对抗训练加固原识别器的方法本质上是一种有监督学习的方法, 它们利用已知类型的对抗样本构建检测边界或重新划定分类边界, 因此对已知类型攻击的防御非常有效, 但对未知类型对抗样本攻击的防御效果较差。

为了解决这些问题, 可以借助样本变换前后识别结果的异同来实施防御。该类方法的优点是它们不需要依赖有关攻击方法的先验信息。由于对抗样本与真实样本具有不一致的内在属性, 可以通过观察真实样本和对抗样本在噪声影响下或施加变换后识别结果的变化情况来检测对抗样本。此外, 还可以通过对输入语音进行平滑、下采样、压缩、重建等处理, 在一定程度上破坏对抗样本的功能, 使其达不到预设的输出, 降低其对识别系统的影响。

Yuan 等人[26]发现背景噪声会降低 CommanderSong 攻击的成功率, 但对真实语音命令的识别效果影响不大, 据此提出了音频湍流的防御方法。音频湍流的基本思想是在识别系统接收到输入语音 A_I 之前添加噪声(称为湍流噪声 A_N), 并检查产生的信号 $A_I \oplus A_N$ 是否可以识别为其他词语。具体来说, 如果 A_I 被识别系统解码为 text1, 然后在 A_I 中加入 A_N, 识别系统将 $A_I \otimes A_N$ 解码为 text2。如果 text1≠text2, 则推断 A_I 受到了对抗样本的攻击。实验表明, 添加噪声后语音的信噪比小于 15 dB 时, CommanderSong 在数字世界攻击的成功率几乎为 0。

在上述攻击场景下生成的语音对抗样本是以波形文件的形式直接输入语音识别系统的, 即数字攻击; 考虑到可能出现的环境噪声的影响, 这种检测方法对物理攻击未必有效, 因为物理环境中的噪声相比人为添加的噪音具有更大的随机性。为此, 语音压缩被用来作为检测对抗攻击的变换手段, 其基本思想是在识别输入语音 A_I 前对其进行下采样。将 A_I 进行下采样后的输出记为 $D(A_I)$, 识别系统分别对 A_I 和 $D(A_I)$进行解码, 得到 text1 和 text2。如果 text1≠text2, 就推断 A_I 为对抗样本。实验表明, 即使是在真实的物理环境中播放, 当下采样率与原始采样率的比值小于 0.7 时, CommanderSong 攻击的成功率下降到了 8%以下, 对真实语音命令的识别准确率保持在 91%以上。

Rajaratnam 等人[44]提出了通过观察在预测结果产生变化之前需要添加多少随机噪声来检测对抗样本的方法。由于识别系统在设计时对真实的含噪样本具有较强的鲁棒性, 通常情况下, 改变对抗样本的预测结果所需的噪声要比改变真实样本的预测结果所需的噪声更少。他们在 SpeechCommands 数据集上对 Alzantot 等人[37]提出的攻击进行防御, 结果表明, 这种防御方法检测对抗样本的准确率和召回率分别为 91.8%和 93.5%。该项研究仅讨论了从均匀分布中采样的随机噪声, 对于实验以外的其他攻击, 这种防御方法是否仍然有效, 有待研究。

Kwon 等人[45]也发现, 对语音样本添加一个较小的失真扰动, 对抗样本的分类结果会出现较为明显的变化, 而真实样本的分类结果变化不大。利用这一特性, 提出了一种检测语音对抗样本的方法。他们使用 Mozilla 通用语音数据集, 将 DeepSpeech 作为目标模型。实验结果表明, 这种方法可以成功地检测出 Carlini 等人[30]生成的对抗样本, 在大约 12dB 的信噪比条件下, 对抗样本的识别准确率下降到 6.21%, 而真实样本的识别准确率仍然维持在 80%以上。然而, 这种检测方法在寻找最佳的失真度时需要利用与原始样本相对应的语音对抗样本, 这在实践中对防御方而言是不现实的。

Andronic 等人提出[46]用 MP3 压缩来减轻对抗扰动的影响, 尝试通过对语音进行 MP3 压缩处理来抵消对抗扰动对识别结果产生的偏差。MP3 是一种音频压缩算法, 采用基于心理声学模型的有损、感知音频编码方案, 丢弃低于听力阈值的音频信息, 从而减小文件大小。该实验在一个结合连接主义时序分类和注意力机制的名为 ESPnet 的端到端语音识别系统上进行。首先对原始语音信号进行特征提取, 用 FGSM 生成对抗样本的特征, 并转化为波形, 在四种不同 MP3 压缩级别(未压缩、128kbps、64kbps 和 24kbps)的条件下进行了实验验证。结果表明, 用 MP3 压缩由特征域重构的对抗样本波形, 识别系统的字符错误率相比未进行 MPS 压缩的情况有所降低, 并在 64kbps 的压缩级别达到最高的降幅。这表明经过压缩处理后, 对抗样本的对抗功能被明显减弱。

Chen 等人[47]的研究表明局部平滑[48]可以减轻对声纹识别系统的 FakeBob[47]对抗样本攻击。局部平滑法在图像处理中被广泛应用于降噪, 它利用邻近的像素来平滑中心像素, 其中对邻近像素的平滑有不同的加权机制, 包括: 中值平滑、均值平滑、高斯平滑等。均值平滑是指在每一个移动窗口内的中心像素被这个窗口内的所有像素的均值所取代, 而中值平滑则是每一个滑动窗口内的中心像素被窗口内所有像素的中值所取代。它们都是通过使相邻像素的像素值更加接近来实现特征压缩, 进而降低对抗样本的攻击成功率。Chen 等人在实验中采用了中值平滑, 用一个窗口大小为 k(奇数)的中值滤波器将每个元素的取值 x_t 替换为相邻的 k 个元素 $\left[x_{t-\frac{k-1}{2}}, \ldots, x_t, \ldots, x_{t+\frac{k-1}{2}}\right]$ 的中值。结果显示, 随着 k 值的增大, 对于较低强度的对抗语音, 尽管对声纹识别的无目标攻击成功率从 99%下降到接近 0%, 但对真实说话人的错误拒绝率也升高到了 35%以上。当

对抗样本干扰强度增大时, 这种防御会失去效果。此外, 实验还发现, 在 i-vector 上进行中值平滑的防御效果要好于高斯混合模型(Gaussian Mixture Model, GMM)。对于有目标攻击, 中值平滑法可以在一定程度上增加 FakeBob 的攻击代价, 但当攻击强度和迭代次数增加时, 防御效果会明显下降, 例如, 在 k=7 的情况下针对 i-vector 和 GMM 的 FakeBob 攻击仍然可以在 250 次迭代后实现 90%的攻击成功率。

实际上, 上述防御方法还可以联合使用。Wu 等人[49]针对声纹认证的对抗样本使用局部平滑结合对抗训练的方法, 提高了模型的防御能力。他们利用 PGD 生成对抗样本, 使用 ASVSpoof2019 数据集的 LA 子集, 分别在 VGG 和 SENET 两个识别模型上进行实验。结果表明, 与单独使用对抗训练相比, 结合均值平滑或中值平滑的对抗训练使识别准确率有所提高, 结合高斯平滑的对抗训练会使识别准确率进一步降低。

在样本平滑的基础上, Wu 等人[50]将自监督学习引入对抗样本的防御, 提出了一种基于转换编码器变换表示法(Transformer Encoder Representations from Alteration, TERA)的级联自监督学习模型。TERA 模型经过训练, 具有通过消除对抗扰动将受损语音转化为干净语音的能力。TERA 模型是通过解决一个具有 l_1 重构损失函数的自监督变换预测任务来训练的。TERA 预训练任务要求模型将一部分随机选择的并施加随机变换的帧序列作为输入, 并尝试重建这些被改变的帧。经过预训练, 该模型学习了将受损语音映射为原始干净语音的能力。他们在实验中假设: 攻击者不知道级联于声纹认证模型前的 TERA 防御模型, 但知道声纹认证模型的参数。用基本迭代法(Basic Iterative Method, BIM)在 VoxCeleb1 数据集上生成对抗样本, 攻击基于 r-vector 的 ResNet 识别系统。用于预训练 TERA 模型的数据集为 VoxCeleb2, 特征为标准 Kaldi 脚本提取的 24 维 MFCC, 优化器为 Adam。级联的 TERA 模型作为滤波器来使用, 用以净化对抗样本。实验结果显示, 集成 TERA 模型可以将被攻击的声纹认证系统的等错误率(Equal Error Rate, EER)从 65%以上大幅降低到 20%左右, 但滤波过程产生的额外噪声会降低对真实语音的识别准确率。在与高斯滤波器、中值滤波器和均值滤波器比较时, TERA 模型在滤除输入的对抗性扰动和保持干净样本的识别准确率方面效果更好。

除自编码重建外, 语音合成技术也可以用来实施样本重建。WaveGAN 声码器[51]用来在给定语音波形的对数梅尔频谱图的情况下重建语音波形。生成器[52]可以有效地学习真实语音波形的分布。Zelasko 等人[53]发现利用 WaveGAN 声码器对输入语音信号进行预处理, 可以提高 DeepSpeech 和 Espresso Transformer 这两个语音识别系统的防御能力。实验设置为白盒场景, 训练过程使用 LibriSpeech 数据集, 引入了 FGSM、PGD 攻击和一种基于频率掩蔽的不可察觉的攻击方法。实验结果显示, 与基于随机平滑的防御相比, WaveGAN 在两个语音识别系统中都获得了更低的单词错误率。在面对不可察觉的有目标攻击时, WaveGAN 使得对两个系统的攻击成功率下降为 0, DeepSpeech 系统的单词错误率从没有防御时的 100%降至 48%, Espresso Transformer 系统的单词错误率从 100%降至 37.4%。与此同时, 引入 WaveGAN 也降低了未受攻击时系统的性能, 这是因为 WaveGAN 声码器在由对数梅尔频谱图重建语音波形时引入了额外的损失, 使系统对部分真实样本的识别也出现错误。在训练过程中对音频样本进行随机平滑增强处理, 可以进一步降低受到攻击时的单词错误率, 但在干净样本上测试的结果并没有得到改善。

4 面临的挑战和发展趋势

语音分类和识别领域的对抗样本攻防尚面临如何贴近真实场景、如何实施多轮博弈、如何融入现有系统等方面的挑战。

(1) 适用于黑盒物理场景的通用对抗

对抗样本攻击和防御的研究总是最先从容易实施的白盒和数字场景入手, 逐步向黑盒和物理场景过渡。商用语音识别系统的更新换代会影响业已见效的黑盒对抗效果, 需要针对升级后的系统重新设计攻防策略。利用机器学习解决物理场景下的攻击或防御问题时, 由于用来建模物理场景的数据匮乏, 往往不能有效地获取该场景的统计特征, 进而导致所得的机器学习模型与实际场景不匹配, 一些有关物理传播规律的先验知识(如麦克风/扬声器的性质、房间冲击响应等)可以被引入模型统筹考虑。此外, 迁移学习、强化学习等方法提供了新的更具通用性的问题求解思路, 并已取得了积极的效果。基于迁移学习的方法是借助目标场景的小样本对现有模型进行修正, 在一定程度上消除场景失配的影响。基于强化学习的方法则是将物理场景中目标系统的输出作为反馈引入迭代回路, 通过不断地现场修改攻防策略来实现攻防目标。总之, 如何摆脱对对手采用的攻击或防御模型中先验信息的依赖, 如何适用于与实

验环境不匹配的各种声学场景, 是实施通用的有效对抗所必须面临的问题。

(2) 攻防双方的多轮博弈

对抗样本的攻防是一个双方博弈的过程。当防御方对特定攻击方法采取检测、加固等措施时, 防御的效果会非常明显; 当攻击方针对特定防御方法生成对抗样本时, 攻击将突破对方的防线。Joshi 等人最新的研究展示了 WaveGAN 防御方法的博弈过程[54]。当攻击者不知道 WaveGAN 防御模型的结构时, WaveGAN 能检测出 90%以上的对抗样本; 当攻击者掌握了防御模型的信息时, WaveGAN 的性能显著下降。防御方如何及时地感知到攻击方已经有能力突破自己的防线、如何降低直至规避攻击方突破防线后的风险、如何再次提高自己的防御能力对抗升级后的攻击, 变得异常关键。对攻击方来说, 同样如此。

(3) 对现有系统的自然升级

对于生成对抗样本的攻击方, 攻击过程包括制作扰动、添加/播放音频扰动等步骤, 如何将语音传播环境的变化引入到对抗样本生成过程, 是对抗样本能在物理空间中发挥作用的关键。对于对抗样本防御方, 防御措施的引入通常会导致虚警、处理延迟等问题, 如何提高识别准确率和实时性, 并与现有的语音识别系统深度融合, 对加速防御方案的落地见效至关重要。

5 结束语

语音识别是人机交互的关键环节, 商业语音识别系统的实际部署不仅面临环境中信道噪声的干扰, 还面临包括对抗样本在内的各种形式攻击的威胁。本文立足实际场景, 由易到难详细地梳理了语音对抗样本的攻击和防御方法。对抗样本虽然对语音识别系统造成了一定的负面影响, 但从保护个人隐私的角度考虑, 却可能在未来发挥它的积极作用, 比如通过对抗样本躲避对特定词汇、特定说话人身份的监听。总之, 与其他新兴技术一样, 对抗样本是一把安全领域的双刃剑, 它对语音信息传递的影响, 取决于使用者的真实意图, 只有深入挖掘对抗样本产生作用的内在原理, 才能为保障语音信息安全提供重要的技术支持。

参考文献

[1] Hanilci C, Ertas F, Ertas T, et al. Recognition of Brand and Models of Cell-Phones from Recorded Speech Signals[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(2): 625-634.

[2] Malik M, Malik M K, Mehmood K, et al. Automatic Speech Recognition: A Survey[J]. *Multimedia Tools and Applications*, 2021, 80(6): 9411-9457.

[3] Kinnunen T, Karpov E, Franti P. Real-Time Speaker Identification and Verification[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2006, 14(1): 277-288.

[4] Rabiner L R. Voice Communication between Humans and Machines——an Introduction[J]. *The National Academy of Sciences of the United States of America*, 1995, 92(22): 9911-9913.

[5] Negi S, Jayachandran M, Upadhyay S. Deep Fake: An Understanding of Fake Images and Videos[J]. *International Journal of Scientific Research in Computer Science*, *Engineering and Information Technology*, 2021: 183-189.

[6] Abdullah H, Warren K, Bindschaedler V, et al. SoK: The Faults In our ASRs: An Overview of Attacks Against Automatic Speech Recognition and Speaker Identification Systems[C]. *2021 IEEE Symposium on Security and Privacy*, 2021: 730-747.

[7] Miao X K, Sun M, Zhang X W, et al. Deep Speech Forgery Based on Parameter Transformation and Threat Assessment to Voiceprint Authentication[J]. *Journal of Cyber Security*, 2020, 5(6): 53-59.
(苗晓孔, 孙蒙, 张雄伟, 等. 基于参数转换的语音深度伪造及其对声纹认证的威胁评估[J]. *信息安全学报*, 2020, 5(6): 53-59.)

[8] Das R K, Tian X H, Kinnunen T, et al. The Attacker's Perspective on Automatic Speaker Verification: An Overview[C]. *Interspeech 2020*, 2020: 4213-4217.

[9] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[C]. *2nd International Conference on Learning Representations*, 2014: 1-10.

[10] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[C]. *3rd International Conference on Learning Representations*, 2015: 1-11.

[11] Gong Y, Poellabauer C. Crafting adversarial examples for speech paralinguistics applications[C]. *Dynamic and Novel Advances in Machine Learning and Intelligent Cyber Security Workshop*, 2018: 1-8.

[12] Kreuk F, Adi Y, Cisse M, et al. Fooling End-to-End Speaker Verification with Adversarial Examples[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 1962-1966.

[13] Iter D, Huang J, Jermann M. Generating adversarial examples for speech recognition. https://web.stanford.edu/class/cs224s/project/reports_2017/Dan_Iter.pdf. Jun. 2017.

[14] Kim N, Park K. Speech-to-text-wavenet. https://github.com/ itzikgili/speech-to-text-wavenet. Nov. 2016.

[15] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples[C]. *31st International Conference on Neural Information Processing Systems*, 2017: 6980-6990.

[16] Li X, Zhong J H, Wu X X, et al. Adversarial Attacks on GMM I-Vector Based Speaker Verification Systems[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2020: 6579-6583.

[17] Villalba J, Zhang Y K, Dehak N. X-Vectors Meet Adversarial Attacks: Benchmarking Adversarial Robustness In Speaker Verifica-

tion[C]. *Interspeech 2020*, 2020: 4233-4237.

[18] Zhang Y K, Jiang Z Y, Villalba J, et al. Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples[C]. *Interspeech 2020*, 2020: 4238-4242.

[19] Khare S, Aralikatte R, Mani S. Adversarial Black-Box Attacks on Automatic Speech Recognition Systems Using Multi-Objective Evolutionary Optimization[C]. *Interspeech 2019*, 2019: 3208-3212.

[20] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi Speech Recognition Toolkit[C]. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011: 1-4.

[21] Schonherr L, Kohls K, Zeiler S, et al. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding[C]. *2019 Network and Distributed System Security Symposium*, 2019: 1-15.

[22] Schönherr L, Eisenhofer T, Zeiler S, et al. Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems[C]. *Annual Computer Security Applications Conference*, 2020: 843-855.

[23] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition[C]. *36th International Conference on Machine Learning*, 2019: 9141-9150.

[24] Vaidya T, Zhang Y, Sherr M, et al. Cocaine noodles: exploiting the gap between human and machine speech recognition[C]. *9th USENIX Conference on Offensive Technologies*, 2015: 16.

[25] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands[C]. *25th USENIX Security Symposium*, 2016: 513-530.

[26] Yuan X, Chen Y, Zhao Y, et al. CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition[C]. *27th USENIX Security Symposium*, 2018: 49-64.

[27] Yakura H, Sakuma J. Robust Audio Adversarial Example for a Physical Attack[C]. *The Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019: 5334-5341.

[28] Chen T, Shangguan L, Li Z J, et al. Metamorph: Injecting Inaudible Commands into Over-the-Air Voice Controlled Systems[C]. *2020 Network and Distributed System Security Symposium*, 2020: 1-17.

[29] Chen Y, Yuan X, Zhang J, et al. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices[C]. *29th USENIX Security Symposium*, 2020: 2667-2684.

[30] Carlini N, Wagner D. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text[C]. *2018 IEEE Security and Privacy Workshops*, 2018: 1-7.

[31] Du T, Ji S L, Li J F, et al. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems[C]. *The 15th ACM Asia Conference on Computer and Communications Security*, 2020: 357-369.

[32] Neekhara P, Hussain S, Pandey P, et al. Universal Adversarial Perturbations for Speech Recognition Systems[C]. *Interspeech 2019*, 2019: 481-485.

[33] Li J G, Zhang X F, Jia C M, et al. Universal Adversarial Perturbations Generative Network for Speaker Recognition[C]. *2020 IEEE International Conference on Multimedia and Expo*, 2020: 1-6.

[34] Xie Y, Shi C, Li Z H, et al. Real-Time, Universal, and Robust Adversarial Attacks Against Speaker Recognition Systems[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020: 1738-1742.

[35] Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio. https://deepmind.com/blog/article/wavenet- generative-model-raw-audio. Sept. 2016.

[36] Samizade S, Tan Z H, Shen C, et al. Adversarial Example Detection by Classification for Deep Speech Recognition[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020: 3102-3106.

[37] Alzantot M, Balaji B, Srivastava M. Did You Hear That? Adversarial Examples Against Automatic Speech Recognition[C]. *31st Conference on Neural Information Processing Systems*, 2017: 1-6.

[38] Li X, Li N, Zhong J H, et al. Investigating Robustness of Adversarial Samples Detection for Automatic Speaker Verification[C]. *Interspeech 2020*, 2020: 1540-1544.

[39] Irfan M M, Ali S, Yaqoob I, et al. Towards Deep Learning: A Review on Adversarial Attacks[C]. *2021 International Conference on Artificial Intelligence*, 2021: 91-96.

[40] Jati A, Hsu C C, Pal M, et al. Adversarial Attack and Defense Strategies for Deep Speaker Recognition Systems[J]. *Computer Speech & Language*, 2021, 68: 1-14.

[41] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[42] Pal M, Jati A, Peri R, et al. Adversarial Defense for Deep Speaker Recognition Using Hybrid Adversarial Training[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021: 6164-6168.

[43] Zhang H C, Wang J Y. Defense Against Adversarial Attacks Using Feature Scattering-Based Adversarial Training[C]. *33rd Conference on Neural Information Processing Systems*, 2019: 1-11.

[44] Rajaratnam K, Kalita J. Noise Flooding for Detecting Audio Adversarial Examples Against Automatic Speech Recognition[C]. *2018 IEEE International Symposium on Signal Processing and Information Technology*, 2018: 197-201.

[45] Kwon H, Yoon H, Park K W. POSTER: Detecting Audio Adversarial Example through Audio Modification[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 2521-2523.

[46] Andronic I, Kürzinger L, Chavez Rosas E R, et al. MP3 Compression to Diminish Adversarial Noise In End-to-End Speech Recognition[M]. Speech and Computer. Cham: Springer International Publishing, 2020: 22-34.

[47] Chen G K, Chenb S, Fan L L, et al. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems[C]. *2021 IEEE Symposium on Security and Privacy*, 2021: 694-711.

[48] Xu W L, Evans D, Qi Y J. Feature Squeezing: Detecting Adversarial Examples In Deep Neural Networks[C]. *2018 Network and Distributed System Security Symposium*, 2018: 1-15.

[49] Wu H B, Liu S X, Meng H, et al. Defense Against Adversarial At-

tacks on Spoofing Countermeasures of ASV[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020: 6564-6568.

[50] Wu H B, Li X, Liu A T, et al. Adversarial Defense for Automatic Speaker Verification by Cascaded Self-Supervised Learning Models[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021: 6718-6722.

[51] Yamamoto R, Song E, Kim J M. Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020: 6199-6203.

[52] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. *27th International Conference on Neural Information Processing Systems*, 2014: 2672-2680.

[53] Żelasko P, Joshi S, Shao Y W, et al. Adversarial Attacks and Defenses for Speech Recognition Systems[EB/OL]. 2021: ArXiv Preprint ArXiv:2103.17122.

[54] Joshi S, Villalba J, Żelasko P, et al. Adversarial Attacks and Defenses for Speaker Identification Systems[EB/OL]. 2021: ArXiv Preprint ArXiv:2101.08909.

魏春雨 于 2016 年在海军航空大学电子对抗指挥与工程专业获学士学位。现在陆军工程大学电子信息专业攻读硕士学位。研究领域为声纹识别、语音识别、语音伪装。Email: weichunyu2020@126.com.

孙蒙 于 2012 年在比利时鲁汶大学电子系获博士学位。现为陆军工程大学智能信息处理实验室副教授。研究领域为智能语音处理、机器学习。Email:sunmengccjs@163.com.

邹霞 现为陆军工程大学智能信息处理实验室副教授。研究领域为语音信号处理、人工智能和机器学习。

张雄伟 现为陆军工程大学智能信息处理实验室教授。研究领域为语音与图像处理、智能信息处理。