语音对抗攻击与防御方法综述

徐东伟[1,2], 房若尘[1,2], 蒋 斌[1,2], 宣 琦[1,2]

[1]浙江工业大学网络安全研究院 杭州 310023
[2]浙江工业大学信息工程学院 杭州 310023

摘要 人工智能的不断发展, 使得人与机器的交互变得至关重要。语音是人与智能通讯设备之间通信的重要手段, 在近几年飞速发展, 说话人识别、情感识别、语音识别得到广泛地普及与应用。特别的, 随着深度学习的兴起, 基于深度学习的语音技术使机器理解语音内容、识别说话人方面达到近似人的水平, 无论是效率还是准确度都得到了前所未有的提升。例如手机语音助手、利用语音控制智能家电、银行业务, 以及来远程验证用户防止诈骗等。但是正是因为语音的广泛普及, 它的安全问题受到了公众的关注, 研究表明, 用于语音任务的深度神经网络(Deep neural network, DNN)容易受到对抗性攻击。即攻击者可以通过向原始语音中添加难以察觉的扰动, 欺骗 DNN 模型, 生成的对抗样本人耳听不出区别, 但是会被模型预测错误, 这种现象最初出现在视觉领域, 目前引起了音频领域的研究兴趣。基于此, 本文对近年来语音领域的对抗攻击、防御方法相关的研究和文献进行了详细地总结。首先我们按照应用场景对语音任务进行了划分, 介绍了主流任务及其发展背景。其次我们解释了语音对抗攻击的定义, 并根据其应用场景对数字攻击与物理攻击分别进行了介绍。然后我们又按照对抗防御, 对抗检测的划分总结了语音对抗样本的防御方法。最后我们对于该领域的不足、前景、以及发展方向进行了探讨。

关键词 深度神经网络; 语音识别; 对抗攻击; 对抗防御; 人工智能安全
中图法分类号 TP29 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.01.09

A Review of Speech Adversarial Attack and Defense Methods

XU Dongwei[1,2], FANG Ruochen[1,2], JIANG Bin[1,2], XUAN Qi[1,2]

[1]Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract With the development of artificial intelligence, the interaction between humans and machines has become more and more important. Speech is an important tool for communication between humans and smart communication devices, and has developed rapidly in recent years. Speaker recognition, emotion recognition, and speech recognition have been widely popularized and applied. In particular, with the rapid development of deep learning technology, speech technology that bases on deep learning enables machines to understand the content of speech and recognize the speaker at a level similar to that of humans. Both efficiency and accuracy have been unprecedentedly improved. For example, mobile phone speech assistant uses speech to control smart home appliances and banking, it can also be used to remotely verify user identity to prevent fraud, etc. But because of the widespread popularity of speech, its security issues have attracted public attention. Researches show that Deep Neural Network (DNN) for speech tasks is vulnerable to adversarial attacks. That is, the attacker can deceive the DNN model by adding imperceptible disturbances to original speech. The generated adversarial samples are indistinguishable by human ears, but they will be predicted by the model incorrectly. This phenomenon first appeared in the visual field, and now it has aroused research interest in the speech field. Based on this, this paper summarizes the research and literature related to adversarial attacks and defense methods in the speech field in recent years. First the speech tasks are divided according to application scenarios, we introduce the mainstream tasks and their general development background. Then we explain the definition of speech adversarial attacks and introduce digital attacks and physical attacks according to speech application scenarios. Later, for the defense methods of speech adversarial samples, we classify them into adversarial defense and adversarial detection, we introduce them separately. Finally, we further discuss the possible deficiencies, future prospects, and development directions of this research field.

Key words deep neural network; speech recognition; adversarial attack; adversarial defense; artificial intelligence security

通讯作者: 徐东伟, 副教授, 博士, Email: dongweixu@zjut.edu.cn
本课题得到国家自然科学基金 (No. 61903334), 浙江省自然科学基金(No. LY21F030016)资助
收稿日期: 2021-07-06; 修改日期: 2021-10-27; 定稿日期: 2021-11-17

1 引言

语音作为人与人间交流的主要方式, 自人工智能发展以来[1], 在过去几十年里引起人们的极大关注, 特别是随着深度学习[2-3]的兴起, 语音领域的各项任务如: 说话人识别、情感识别、语音识别等得到充分的发展。深度学习作为目前人工智能最常见的技术之一, 可以处理许多复杂的任务, 包括图像识别[4-6]、对象检测[7-8]、语音识别[9-12]、信号处理[13]等。但事实证明, 深度学习模型容易受到对抗攻击[14-15], 攻击者通过发现模型的弱点并制作出与原始样本不同的对抗样本, 从而使模型对生成的对抗样本产生错误的预测。

过去几年, 关于深度学习攻击的研究主要停留在视觉领域[16], 攻击者在输入图像中添加不可察觉的扰动, 产生的对抗样本足以误导深度神经网络模型, 各种攻击算法也陆续被提出。近来, 鉴于基于DNN 的智能语音系统的广泛应用, 研究的重点逐渐转向语音领域, 对于各种语音任务, 添加的对抗扰动会使任意的语音转录为攻击者指定的文本, 或者识别语音为规定的说话者, 不仅危害了个人财产与隐私, 还可能对人身安全造成损害。但一般情况下, 语音领域生成对抗样本在某种程度上要比在视觉领域更加困难, 添加在原样本上的扰动在实际中很容易被察觉, 用户拒绝该对抗样本输入模型, 从而导致攻击失败, 而且对于复杂的语音识别模型来说, 攻击算法有更苛刻的要求。

当前, 对语音数据的攻击可分为两大类: 语音到标签的攻击和语音到文本的攻击。语音到标签的攻击是为了找到与原始音频相近的对抗样本, 使模型产生不同(错误)的标签, 为此 Kreuk 等人[17]用图像上的快速梯度符号方法生成了对抗音频且取得了成功; Alzantot 等人[18]提出了一种基于遗传算法的黑盒攻击来欺骗指令分类模型; Chen 等人[19]提出用基于梯度估计的黑盒攻击来欺骗有阈值判定的说话人识别任务; Li 等人[20]则将攻击指向物理世界的实时语音流, 生成亚秒级的通用扰动等。语音到文本(语音识别)的攻击要求对抗音频的转录结果与原始文本不同, 如 Carlini 和 Wagner[21]直接对原始波形进行基于优化的目标攻击, 攻击 DeepSpeech 模型可以达到100%的成功率; Neekhara 等人[22]在少量语音样本上生成的通用对抗扰动对大多数语音样本有效, 作为非目标攻击, 节省了攻击的时间成本; Taori 等人[23]用改进的遗传算法来攻击黑盒语音识别模型; 还有一系列旨在探索物理世界中语音识别的攻击[24-28]等。这些对抗攻击对现实生活的算法安全有着重要意义。

因此为了抵御这些攻击, 实现算法安全, 研究者提出了一些防御方法, 如 Goodfellow 等人[29]最先提出的对抗训练, 对语音领域也有效果[30]; Yang 等人[31]提出的输入转换以及针对对抗音频的对抗检测[32-38]都是防御语音攻击的有效手段。目前, 随着语音的应用越来越广泛, 针对语音的攻击与防御方法仍在不断地探索中, 其主要目的是预防安全隐患, 为网络安全奠定基础。

本文的章节安排如下: 第 2 节介绍基于 DNN 语音任务的应用场景; 第 3 节介绍对抗攻击的定义以及语音领域主流对抗攻击的方法; 第 4 节介绍语音对抗攻击的防御手段; 第 5 节对目前语音攻击与防御方法进行总结, 并展望未来可能的研究方向。

2 语音任务及应用

语音通信是人工智能的一个重要产物, 改变着我们与智能设备的交互方式[39]。它的普及为语音控制技术创造了新的机会, 我们能够通过远程发送语音信号来控制移动设备, 语音信号根据其应用场景不同, 可分为说话人识别、音频分类(包括情感识别、语音指令分类、音乐流派分类等)以及语音识别。其中音频分类领域应用较多的是情感识别, 下面我们将介绍这 3 种语音任务的简单概念以及基本流程。

2.1 语音情感识别

情感作为人共有的情绪, 在人机交互中占重要组成部分, 智能机器根据语音的特征, 可以对情绪进行分类, 例如愤怒、无聊、恐惧、喜悦、幸福等。如果正确识别出情感, 则系统采取相应的行动。目前的研究表明可以在电话交流中感知说话者的情绪状态并进行分类[40-41], 以判断是否是值得信任的对象。还可以根据患者的声音来提供有关患者健康的信息[42], 这种识别可用于智能医疗系统, 计算机健康分类与心理状态评估系统。基于深度学习的情感分类精度得到进一步提高, Anvarjon[43]提出可以利用语音提取语谱图来进行情感分类, 通过构造一个卷积神经网络, 利用语谱图的特征进行提取、整合、分类。

2.2 说话人识别

说话人识别(声纹识别)是根据说话者的语音判断身份的任务, 已被广泛用作身份验证的生物特征[44]、个人数字助理[45]以及取证[46]等。它是银行和电子商务采用的最成熟的生物特征认证技术之一, 是在线和通过电话认证客户的主要手段。说话人识别主要包括以下 3 个步骤: 训练、注册、评估。在训

练阶段, 可以从一组话语中学习合适的说话者表示, 并建立简单的评分功能。在注册阶段, 说话者会提供一些语音, 保存在机器里用作后续的评估模板。在评估阶段, 通过对已注册的说话人模板与待确认的语音进行对比来执行任务。说话人识别大致分为两部分: 说话人辨认与说话人确认。前者是辨认给定话语是属于哪一个已注册说话者的过程, 是 N∶1 的关系, 这类主要用于公共场合, 如保险公司、或办公机构。而后者是验证语音是否来自所声明的说话人, 是 1∶1 的关系, 可用作电话业务的确认、私人信息控制以及数据访问权限。说话人辨认又可以分为开集辨认[47]与闭集辨认[48], 闭集辨认的输出结果一定是注册中的人, 开集辨认最后阶段需要进行阈值比较, 分数低于阈值的结果会被系统拒绝。

比较早的说话人识别技术利用高斯混合模型(Gaussian mixture models, GMM)建模进行研究。后来为了克服训练数据不多的情况, 又引入了高斯混合-通用背景模板(Gaussian mixture models-Universal background model, GMM-UBM), 使识别率得到提高。进入 21 世纪后, 基于身份向量(Identity vector, i-vector)[49-50]的说话人识别技术得到广泛应用, i-vector 结合概率线性判别分析(Probabilistic linear discriminant analysis, PLDA)可以实现信道补偿, 让说话人识别系统在复杂环境背景下也能提取良好的特征。后来 DNN 的引入极大地降低了识别错误率, 从这时开始的算法, 可以称为嵌入特征提取, 不过提取的是神经网络最后一个隐藏层的激活单元, 用来代替 i-vector 作为语音的特征表示[51]。2008 年, Snyder[52]提出了用 x-vector 作为特征, 它在训练的时候考虑了整段声音信号的信息, 把每一段声音的输出特征通过求均值与方差后再结合起来, 不像 d-vector 简单地取均值, 在网络结构上采用时延神经网络, 它与 DNN 只考虑某一帧不同, 时延神经网络考虑了某一帧的上下相关性作为输入, 能表达语音特征在时间上的关系, 此外该网络的权值具有时间不变性, 减少了计算成本。无论是 i-vector 还是 x-vector 都是先提取特征, 再用训练好的后端打分模型(如 PLDA)进行相似度比较。最新的端到端系统却将这两段整合到一个系统中, 一体化特征训练与分类打分, 提高了运算效率[53-54]。

2.3 语音识别

语音识别(Automatic speech recognition, ASR)即利用人工智能算法把语音信号转为文本的过程, 目前语音识别的用处最为广泛, 各种 APP 的语音转文本、智能家居等普遍应用。对语音识别数十年的深入研究开发了多种技术, 例如高斯混合模型和隐马尔可夫模型[55], 随着 DNN 技术的快速发展, 基于 DNN 的语音识别已成为主流技术。包括 Google[56]、Apple[57]和 Amazon[58]在内的众多公司已广泛采用基于 DNN 的 ASR 与物联网设备。

语音识别的流程如图 1 所示。可大致分为三个阶段: 语音预处理、语音特征提取、音频分类。语音预处理包括从原始语音中去除噪声、从背景噪声中分离语音信号; 特征提取阶段是针对某段语音信号, 生成唯一表示该语音的特征; 分类阶段是用不同的人工智能分类算法来识别给定语音。

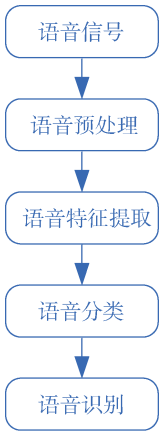

图 1 语音识别的流程图

Figure 1 The flow chart of speech recognition

传统的语音识别模型分为两个部分: 声学模型、语言模型。语音识别的目标是使语音与给定文字的匹配度更高, 匹配度用概率表示, 用 X 表示语音信号, S 表示文字序列, 需要求解以下公式:

$$S^* = \underset{S}{\operatorname{argmax}} P\left(S|X\right)$$

根据贝叶斯公式, 可以写成:

$$S^* = \underset{S}{\operatorname{argmax}} \frac{P\left(X|S\right)P(S)}{P(X)} = \underset{S}{\operatorname{argmax}} P\left(X|S\right)P(S)$$

此公式可以解释为最终期望的 S, 使 $P(X\,|\,S)$ 和 $P(X)$ 概率尽可能大, 这个任务由声学模型和语言学模型完成。其中 $P(S)$ 表示一串文本序列本身出现的概率, $P(X\,|\,S)$ 表示在目标文本下给定语音能转录成该文本的概率。语言模型用链式法则把一个句子的概率拆为每个词概率之积, 其中引入词典, 可以把文本串转化为音素串。声学模型是给定文本后, 计算发出这段语音的概率。此外, 要计算匹配程度, 需要找到一种合适表示信号的方法。梅尔频率倒谱系数(Mel frequency cepstral coefficients, MFCC)[59]提取技

术是语音特征提取中最流行的技术, 对说话人识别和语音识别均能取得很好的效果[60], 此外, 线性预测系数(Linear predictor coefficients, LPC)[61]、感知线性预测(Perceptual linear predictive, PLP)[62]也在广泛使用。为了计算语言和音素串的匹配度, 还需要知道每个音素的起止时间, 对于传统的声学模型的训练, 对于每一帧的数据, 需要知道对应的标签才能进行有效的训练, 在训练数据之前, 要做语音对齐的预处理, 本身是一件耗时的工作。

Grave 等人[63]提出了用 Connectionist temporal classification(CTC)作为损失函数的声学模型训练, 这是一种端到端的训练, 不需要数据预对齐, 只需要一个输入序列一个输出序列。CTC 损失只关心预测输入序列是否与真实序列对齐, 不关心预测结果在某个时间点上是否一致。最近, Hannun 等人[64]提出的端到端 DeepSpeech 模型引起了关注, 它把声学模型、语言模型结合一起, 不引入传统音素或词的概念, 直接训练音频到文本的模型, 可以有效地提高识别率。

3 语音对抗攻击

本节我们会先对语音对抗攻击的类型、对抗样本的定义以及扰动衡量方法做一个说明, 然后详细介绍近些年主流的语音攻击方法。我们旨在使本次的调查综述尽可能全面, 希望可以对以后这个领域的研究工作有所帮助。

按照对手先验知识不同, 语音的攻击可分为白盒攻击、黑盒攻击与灰盒攻击。其中白盒攻击假定对手全面了解目标神经网络, 包括模型类型、模型体系结构以及所有的参数和训练权重的值。黑盒模型假定对手不能访问神经网络, 只能了解模型的输出, 这对攻击者更具挑战性。而灰盒攻击介于白盒攻击与黑盒攻击之间, 例如攻击者仅了解模型的一部分。按照攻击场景、对手知识、是否可用于物理世界中、攻击的模型以及攻击成功率, 我们列出了不同的方法, 如表 1 所示, 其中特定 CNN、特定 RNN 表示针对某一任务的具体模型, ◆用来衡量攻击的效果(此处的效果指的是某攻击算法针对特定应用场景的实现程度, 不能用来单纯地衡量方法好坏)。

因语音应用场景不同, 实现的目标不同, 所以我们按照语音的三大任务来划分攻击, 此外, 我们总结了它们的优点、可能存在的不足以及未来可能的研究方向。其中, 对于一些重要的概念和方法我们会做详细的介绍, 而对其余一些相对复杂或相关性较弱的算法和知识点, 如物理世界的对抗攻击涉及到搭建物理平台等, 我们会大致介绍其方法并列出相关文献, 读者可以自行查阅。

3.1 对抗样本的定义

给定输入音频样本 x, 攻击者的目标是构建一个不易察觉的微小扰动 δ, 生成扰动音频信号即对抗音频:

$$x' = x + \delta \text{ such that } \delta_P < \epsilon$$

表 1 语音对抗攻击方法

Table 1 Methods of speech adversarial attack

攻击场景	方法	对手知识	目标/非目标	攻击形式	物理攻击	攻击模型	攻击效果
说话人识别	端到端的快速梯度符号法[65]	白盒	非目标	梯度	否	Wave CNN	◆
	基于声学特征的快速梯度符号法[17]	白盒	目标	梯度	否	特定 RNN	◆◆
	对抗转换网络[66]	白盒	目标/非目标	优化	否	SincNet	◆◆◆◆
	强鲁棒性的通用对抗扰动[67]	白盒	目标	优化	是	特定 CNN	◆◆◆
	FAKEBOB 黑盒攻击[19]	黑盒	目标/非目标	优化	是	开源模型	◆◆◆◆
音频分类	从频域到时域的攻击[68]	白盒	非目标	梯度	否	特定 CNN	◆◆
	基于遗传算法的黑盒攻击[18]	黑盒	目标	优化	否	特定 CNN	◆◆
	实时语音攻击[20]	白盒	目标	优化	是	特定 CNN	◆◆◆
	Carlini 和 Wagner 语音攻击[21]	白盒	目标	优化	否	DeepSpeech	◆◆◆◆
	加权采样音频攻击[69]	白盒	目标	优化	否	DeepSpeech	◆◆◆◆
	基于感知的语音攻击[70]	白盒	目标	优化	是	DeepSpeech	◆◆◆
语音识别	语音识别的通用对抗扰动[22]	白盒	非目标	优化	否	DeepSpeech	◆◆◆
	Taori 黑盒攻击[23]	黑盒	目标	优化	否	DeepSpeech	◆◆
	基于脉冲响应的物理攻击[24-26]	白盒	目标	优化	是	DeepSpeech、Kalid、Lingvo	◆◆◆◆
	Metamorph: 针对语音控制系统的攻击[27]	白盒	目标	优化	是	DeepSpeech	◆◆◆
	针对商业语音识别系统的黑盒攻击[28]	黑盒	目标	优化	是	商业系统	◆◆◆◆

(注: 文献[24]中使用了高斯白噪声、带通滤波器和房间脉冲响应, 文献[25]和[26]中用了房间脉冲响应和人耳掩蔽效应, 为了方便归纳, 我们将这三种方法统称为基于脉冲响应的物理攻击)

目的是针对语音对抗样本 x', 分类器会产生错误的输出。对于非目标攻击, 如果原始语音 x 的标签为 y, 添加扰动生成的对抗样本 x' 经过 DNN 后的输出为 $\overline{y}$, $\overline{y} \neq y$, 则攻击成功; 对于目标攻击, 若对抗样本 x' 经过 DNN 后的输出为 y_t, y_t 表示攻击者期望的目标, 则攻击成功。

在生成对抗样本的过程中, 不仅要保证攻击成功率, 还要维持扰动在一定范围内。在视觉领域中, 对抗扰动 δ 通常是微小的像素点, 对于整幅图像来说不容易察觉。但是在语音领域, 即使是添加很小的噪声, 人耳听到的语音可能与原始有较大差异。因此, 这是一个不小的挑战。对于音频攻击, 为了衡量扰动引起的失真, 常用分贝(dB)来衡量失真, 用来表示音频的相对响度:

$$\mathrm{dB}(x) = \max_i 20 \cdot \log_{10} |x_i|$$

以 dB 作为衡量时需要与其他参考点比较才有意义。在这里, 我们将失真与原始音频的分贝水平作比较, 公式如下:

$$\mathrm{dB}_x(\delta) = \mathrm{dB}(\delta) - \mathrm{dB}(x)$$

因为引入的扰动通常比原始音频更“安静”, 所以这里失真是负数, 越小的负值表示更安静的失真。除了用分贝衡量失真之外, 还可以用 L2 范数、信噪比、语音质量感知评估[71](Perceptual evaluation of speech quality, PESQ)来评估对抗音频。

本节中, 我们介绍了语音对抗攻击的相关概念与扰动衡量标准, 下面对语音领域的对抗攻击方法进行详细的介绍。

3.2 说话人识别的对抗攻击

3.2.1 快速梯度符号法

快速梯度符号法是对抗攻击领域一种经典的基于梯度的方法, 一般属于非目标攻击, 其通过沿着固定步长的梯度方向修改网络的输入来增加神经网络的损失函数, 以此达到错误分类的目的。该方法能够快速生成音频对抗样本, 它依赖于一次迭代来修改音频特征, 基本公式如下:

$$\tilde{x} = x + \epsilon\, sign(\nabla x J(g_\theta(\tilde{x}, x_k), y))$$

其中 ϵ 表示步长, 用来限制扰动的大小, $\nabla x J$ 表示加入扰动后损失函数相对于输入的梯度。Gong 等人[65]第一次提出了基于快速梯度符号的方案来产生音频对抗性例子, 为了避免将声学特征转换回波形而引入的感知损失, 他们提出了一种直接扰动原始波形而不是特定声学特征的端到端方法。该方法的关键是引入了一个端到端的机器学习模型: WaveRNN[72], WaveRNN 是第一个使用递归神经网络从原始语音波形映射到任务标签的模型。此外, 为了解决在递归神经网络上使用基于梯度攻击方法时的梯度消失。作者通过使用替换网络解决此问题, 替换网络用前馈卷积结构替换了循环结构, 通过用卷积层代替循环层, 有效地解决了梯度消失。快速梯度符号方法以往通常用于视觉领域, 作者用来生成音频对抗样本是语音对抗攻击方面的一次成功尝试。虽然 Gong 等人证明了此方法的有效性, 但是它依赖替换网络的训练且作者没有研究此攻击的可迁移性。

Kreuk 等人[17]后来提出了针对端到端说话人验证模型的攻击, 目标模型合并了说话人识别中的特征嵌入提取与打分阶段, 作者通过在声音特征(MFCC)上应用快速梯度符号法, 然后从对抗声音特征中重建波形。端到端的说话人验证系统输入的是一对语音, 待验证的未知语音与注册的对比语音, 经过攻击后的未知语音为 $\tilde{x}$, 攻击的目的是最大化对抗样本中分类器的预测与正确标签之间的损失函数。在说话人验证中, 模型的输出为指定语音的判定结果, 因此此处的快速梯度符号法属于目标攻击。此方法仅加入扰动到未知测试语音中, 因此要保证注册的语音不变。从对抗声音重建波形的方法大大提高了攻击性能, 他们用了两种黑盒攻击证明了该方法有一定的可迁移性, 优于 Gong 等人的方案。

3.2.2 对抗转换网络

在之前视觉的相关任务中, 用来生成攻击的方法大多是基于梯度或者基于优化的, 这些方法需要在测试的时候进行优化迭代来生成对抗样本, 从某种程度来说这样实用性不强, Li 等人[66]提出了一种新的攻击方法, 通过构建一个单独的对抗转换网络(Adversarial translation network, ATN), 主要用来针对非目标攻击, 可以直接将原始输入转换为对抗输入, 这种方法的优点是在测试阶段不需要梯度且转换速度很快。ATN 是一个可训练的网络, 用来把输入的原始语音转换为对抗音频, 此方法添加的扰动是不可察觉的, 可以欺骗闭集说话人识别模型。使用一个预训练的音素识别模型来帮助训练攻击网络, 但在训练攻击者模型时, 无法对预训练的模块进行微调。

如图 2 所示, 对抗转换网络内部是一个小型的全卷积残差网络, 训练网络的方法没有采用通常的梯度上升, 因为 softmax 层的原因, 训练好的说话人识别模型反向传播的梯度几乎为零。另一方面, 因为要保证扰动的大小, 作者引入了预先训练的音素识

别网络将音素信息考虑在内，以优化感知质量，优化算法如下:

$$L_{total} = L_{spk} + \lambda_{phn} L_{phn} + \lambda_{norm} L_{norm}$$

$$L_{spk} = \begin{cases} x_{spk}[I_{1st}] - x'_{spk}[I_{2nd}], & I_{1st} = y_{spk} \\ 0, & else \end{cases}$$

$$L_{phn} = KL(P_{phn} \| P'_{phn})$$

$$L_{norm} = [\max(s - s' - m, 0)]^2$$

上述优化算法中 x' 是使用对抗样本 s' 时说话人识别模型 softmax 层之前激活的输出，I_{1st}/I_{2nd} 分别是 x' 中最大值与第二大值的索引，$P_{Phn} \| P'_{Phn}$ 分别是 s/s' 作为输入时音素识别模型softmax层的输出分布，KLD 来衡量两个分布间的距离，用来约束噪声感知程度。在 L_{norm} 中，m 为超参数，用来提供扰动不可感知的一个边界。$\lambda_{Phn}, \lambda_{norm}$ 是用来结合三个损失项的超参数。

Li 等人使用 SincNet[73]作为被攻击模型，Li 等人提出的对抗转换网络主要应用于非目标攻击，可以用较高信噪比的对抗样本达到 99.2%的攻击成功率，对于目标攻击，平均成功率可以达到 72.1%。对抗转换网络在测试阶段生成对抗样本速度较快、效率较高，但是对此攻击在其他模型上的可迁移性没有进行研究。

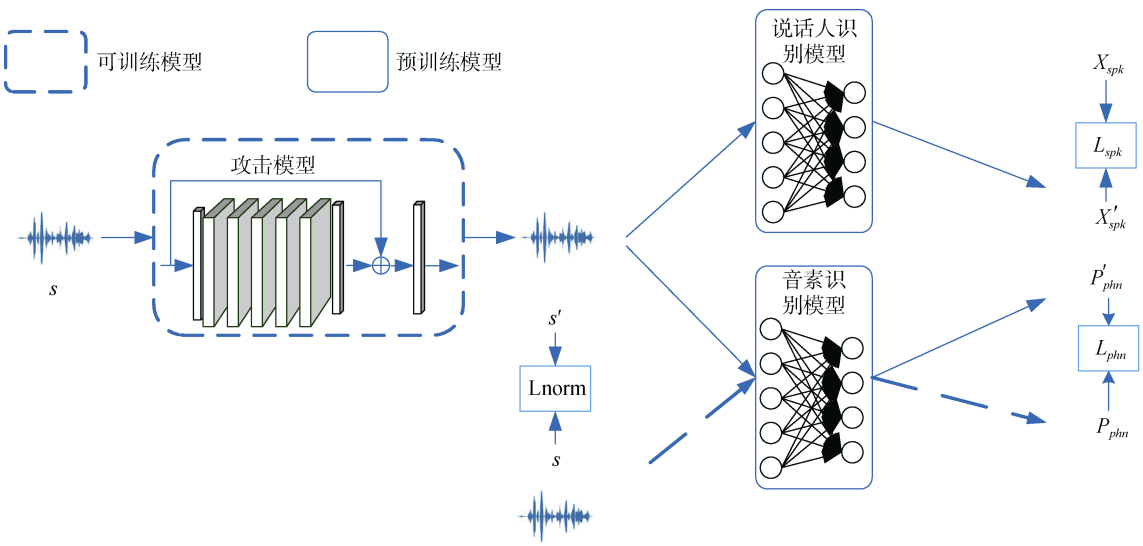

图 2　对抗转换网络框架图

Figure 2　The framework diagram of adversarial translation network

3.2.3　强鲁棒性的通用对抗攻击

目前对于说话人识别对抗攻击的研究，大多是直接将语音对抗样本输入给说话人识别系统。Xie 等人[67]提出了一个新的攻击形式，他们的研究向前迈了一步，通过估计语音在物理世界传播产生的声音失真来构建强鲁棒性的对抗攻击，此攻击不仅在空气信道传播后有效而且对于每个语音输入不需要单独产生扰动。作者提出的方法通过产生不可感知的通用对抗扰动来攻击多说话人辨认系统，使其输出目标说话者标签。

Xie 等人考虑的是白盒攻击，不仅对目标模型以及参数有完全的了解，还要假设攻击者可以进入房间的布局，了解房间里的失真情况。如图 3 所示，攻击生成的一个与音频无关的通用对抗扰动，攻击的目的旨在找到通用对抗扰动 δ 能够对任意 x，识别模型能把它分类为目标标签 t。生成的通用对抗扰动不仅对所有话语有效，而且通过重复对抗扰动来克服发声长度变化的问题。作者提出通过估计房间脉冲响应(Room Impulse Response, RIR)来表示房间环境的声传播，目标函数可表示为:

$$\text{Argmax}\left(P\left(x' * r\right)\right) = t$$

房间的声传播认为是线性且时不变的，因此麦克风接收的对抗样本信号可以表示为 $x' * r$，其中 r 是估计的房间脉冲响应(RIR)，*表示卷积运算。在训练阶段使用声学室模拟器[74]生成 RIR(代表从播放的声音到录制的声音的某种映射)，声学室模拟器可以模拟房间的大小、音频源与麦克风的位置以及混响率。用此方法进行训练优化，预先估计多个环境下的 RIR，每个输入语音更新通用扰动时随机选一个 RIR，用来增强对抗样本的鲁棒性，然后通过迭代优化找到对抗样本，该部分与针对图像攻击的通用对抗扰

动相似[75]。

Xie 等人提出的具有强鲁棒性的通用对抗扰动对任意已注册的说话人输入语音有效, 相比非通用攻击, 此攻击的发起时间大大缩短, 而且可在空气信道中传播的性质为探索物理世界的实时语音攻击提供了思路。

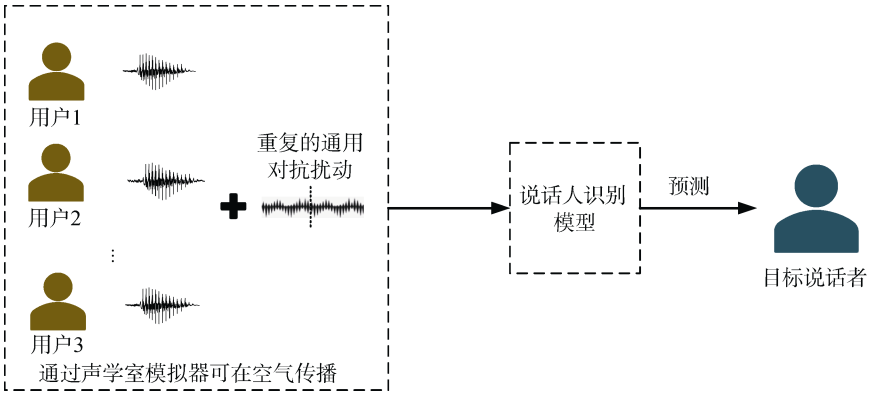

图 3 说话人识别的通用对抗扰动

Figure 3 Universal adversarial perturbations of speaker recognition

3.2.4 FAKEBOB 黑盒攻击

Chen 等人[19]提出了一种黑盒攻击来欺骗说话人识别模型, 算法考虑了开集说话人辨认的分数阈值, 可以针对目标攻击和非目标攻击, 是一种物理世界的攻击。FakeBob 攻击利用一种新的算法来估计阈值, 使用基于自然演化的策略(Natural evolution strategy, NES)来估计梯度[76], 最后经过多迭代来优化扰动。在开集说话人辨认中, 当目标说话人的分数高于已注册人的最大分数且不小于设定阈值即攻击成功, Chen 等人对损失函数的优化类似 Carlini 和 Wagner[77] 的做法, 并把阈值考虑进去有效地解决这个问题。因为黑盒攻击只能访问输出的分数以及判别结果, 所以用基于自然演化策略作为梯度估计技术, 得到梯度后再用梯度下降进行优化, 以找到对抗扰动, 为了保证攻击成功, 估计的阈值不应小于实际阈值, 又不能相对过大, 因此阈值估计算法首先初始化阈值为查询结果的最大分数, 每次迭代过程中更新阈值的大小, 再根据步长迭代构造对抗样本直至攻击成功。

虽然 Chen 等人提出的攻击方法属于黑盒算法, 但是此攻击对说话人识别系统有很高的成功率, 优于以往的黑盒攻击算法, 如梯度估计[78]、遗传算法[18]。实际上, 攻击者可以利用 FAKEBOB 进行手机声纹解锁, 移动应用声纹登录甚至银行交易声纹验证等, 从而对受害者的财产安全, 声誉等造成危害, FAKEBOB 在开源和商用声纹识别系统上均取得接近 100%攻击成功率, 并且能有效地迁移到其他声纹识别系统, 包括实际场景下的物理攻击。

3.3 音频分类的对抗攻击

3.3.1 从频域到时域的攻击

Koerich[68]提出将语音信号的时序数据转到频域的方法来攻击音乐流派分类模型。主要做法就是将语音信号利用短时傅里叶变换转换成频谱特征, 然后用 2D CNN 对其进行分类训练。对训练好的模型采用视觉领域中基于梯度的方法进行攻击。还可以进行攻击迁移性研究, 把由频谱特征生成的对抗样本进行重构, 将其转换到时域, 从扰动的频谱特征重建的音频波形也可以以较高的置信度欺骗原始音频训练的 1D CNN。下一步的研究可以利用此方法的可迁移性进行黑盒攻击。

3.3.2 基于遗传算法的黑盒攻击

Alzantot 等人[18]在研究中提出了遗传攻击算法, 这种算法属于黑盒攻击, 不需要提前了解受害者模型的内部参数, 只需要查询模型的输出就可以攻击成功, Alzantot 等人使用此攻击针对语音指令分类模型, 使模型错误识别指令, 如上、下、左、右、停、继续等。

遗传算法是自然选择模拟的算法, 使用遗传算法生成对抗样本的过程可分为三个步骤: 种群、选择、突变。以原始样本 x 为输入, 标签 t 为攻击目标, 首先将随机噪声添加到给定音频段的样本子集中来生成一组候选对抗样本(种群), 然后根据目标标签的预测分数来评估每个对抗样本的适应性得分, 若得分最高的对抗样本分类为标签 t, 则攻击成功, 否则进行下一步: 选择, 通过选择找到高适应度的对抗样本作为父代通过交叉生成子代, 此外还需要完整地保留适应度最高的对抗样本作为下一代的一部

分。为了使找到的解趋近最优解, 还需要对子代以很小的概率添加随机噪声(突变)再次得到一个种群, 以此迭代, 此算法会在求解过程中迭代到预期的次数或者攻击成功为止。

基于遗传算法的黑盒攻击是通过模拟自然选择来淘汰远离目标标签 t 的对抗样本, 选择目标标签分数比较高的对抗样本。作为黑盒攻击针对语音指令分类模型可以达到 87%的攻击成功率, 且产生的扰动对人耳来说处于较低范围, 对黑盒攻击的后续发展起到很大作用。

3.3.3 实时语音攻击

随着语音的广泛普及, 研究现实世界中的物理攻击更有意义。针对音频分类任务, Li 等人[20]考虑了一个物理世界攻击场景, 即对于实时语音(音频流输入), 可以通过在任意时刻播放对抗性扰动来欺骗识别系统。

目前对攻击算法的研究基本都需要扰动和输入音频完全对准, 需要攻击者了解发起攻击的确切时间, 这些问题限制了实时语音攻击的有效性。Li 等人总结成功的经验提出了 AdvPulse: 一种生成亚秒级对抗扰动的方法, 可以将其添加到音频流输入的任何点上, 以发起有目标的对抗攻击, 作者同时针对了语音指令分类模型与说话人识别模型进行了攻击, 由于此攻击方法对于分类模型的通用性, 所以归于音频分类对抗攻击部分。该方法的流程如图 4 所示, 这种攻击无需修改整个音频, 只需向音频加入 0.5 s 左右的对抗扰动即可, 对抗性扰动可以添加到流音频的任意位置, 不需要完全同步, 此方法生成的扰动为通用扰动, 会使任意音频被识别为目标标签。

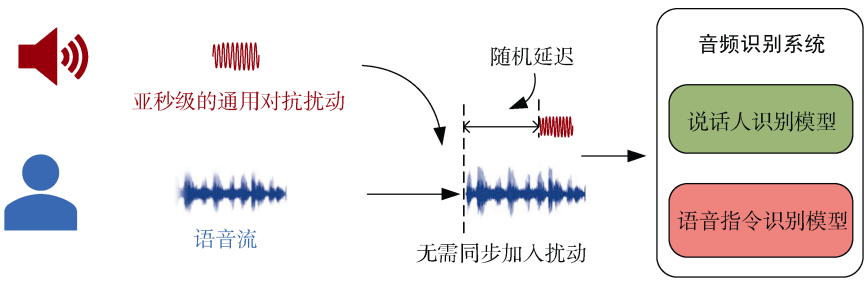

图 4 AdvPulse 攻击方法框架图

Figure 4 The framework diagram of AdvPulse attack method

具体而言, 为了达到攻击无需时间同步的要求, 作者利用了基于梯度的机器学习算法, 最大化目标标签在不同延迟条件下的输出概率, 使得可以在音频的任意输入时间戳上添加扰动。此外在一组输入样本上利用基于惩罚的通用训练来生成与输入无关的通用扰动, 为了降低攻击的可疑性, 根据环境声音调整对抗扰动, 使其听起来更像环境中的声音(如鸟叫、喇叭声等), 作者引入了环境声音模板进行优化。

实际上, 一般的对抗攻击直接用在物理世界中会在空中传播时遭到信号衰减与环境噪声引起的严重失真, 作者为了保持对抗扰动的有效性, 把扬声器与麦克风的限制、吸收、混响以及噪声的影响考虑进扰动生成过程。用带通滤波器将对抗扰动限制在有效频率范围; 用房间脉冲响应(RIR)模拟对抗扰动的空中失真, 克服吸收与混响; 用环境噪声来增强实际应用场景中噪声的鲁棒性。

Li 等人提出的方案是针对音频分类的语音流实时攻击, 有效地解决了以往对抗攻击在物理世界中鲁棒性弱的问题, 对物理世界的语音攻击和语音安全的研究具有参考价值。

3.4 语音识别的对抗攻击

3.4.1 Carlini 和 Wagner 语音攻击

在语音的对抗攻击领域, 除了对说话人识别的研究以外, 关注更多的是语音识别, 语音识别模型将指定语音转录为文本, 此时有目标的攻击才更有意义, 攻击者添加的扰动不仅要让模型有目标的转录, 且加入的扰动要不可感知。事实证明, 在语音识别中生成有针对性的对抗样本非常困难。

Carlini 和 Wagner[21]提出了一种直接修改原始音频的优化方法, 证明了语音识别中有目标对抗音频的存在。此方法用来攻击端到端的 DeepSpeech 模型, 用改进的损失函数来实现更快的收敛。该方法结合了语音识别中主流的 CTC 算法[79], 不仅攻击成功率极高, 而且作为语音识别对抗攻击中一种主流方法, 许多攻击与检测方法都使用它作为基线。

Carlini 和 Wagner 提出的攻击方法主要分为两个步骤, 第一个步骤使用 CTC 损失函数来优化对抗样本, 使语音识别模型把语音转录为目标语句:

$$\text{minmize } |\delta|_2^2 + c \cdot l(x+\delta,\ \pi_i)$$

l 表示 CTC 损失函数, 在优化攻击的同时限制扰动的大小。由于目标句子的复杂性, 很难在保持较低失真的情况下准确转录。为了解决这个问题, Carlini 和 Wagner 提出了一个新的损失函数可以更加准确的把对抗样本转录为目标句子, 第二步为:

$$\text{minmize } |\delta|_2^2 + c_i \cdot L_i(x+\delta,\ \pi_i)$$

$$\text{such that } dB_x(\delta) < \tau$$

其中 L 是改进的损失函数, 用来使攻击达到期望的转录结果, 此外为了解决某个字符难以被转录识别的问题, 对于每一帧都选择一个参数 c, c 若是足够大, 优化程序就会将重心放在降低这一帧损失函数上。总体而言, Carlini 和 Wagner 的攻击方法分为两步。第一步, 先用普通的 CTC 损失函数生成一个对抗样本 x_0, CTC 损失会在解码时构建一个对抗样本的标记序列 π, 提取这个标记序列。第二步, 用新改进的损失函数针对之前对抗样本的标记序列生成一个失真更小的对抗样本 x'。

Carlini 和 Wagner 针对语音识别提出的基于优化的攻击在 DeepSpeech 模型上可以达到 100%的成功率, 生成的对抗音频能达到与原音频 99%的相似度。除此之外他们还实现了将非语音攻击为目标文本以及通过攻击隐藏语音(转录空白), 并探讨了通用对抗扰动和攻击迁移的可能性。

3.4.2 加权采样对抗攻击

Liu 等人[69]提出加权采样音频对抗样本的概念, 利用样本的失真数量及权重来加强攻击, 文章主要提出了两种技术: 加权扰动技术(Weighted perturbation technology, WPT)与采样扰动技术(Sampling perturbation technology, SPT)。

SPT 可通过减少扰动点的数量来提高音频对抗样本的鲁棒性, 在传统音频上, 若让音频转录为目标 t, 在整段语音上添加细微的扰动即可, 但也可以只添加少量的扰动使攻击成功, SPT 是通过固定原音频的一部分点而只改变另外一部分点来添加扰动, 如可以将音频向量的扰动数目从 n 缩短到 m, 由此生成的对抗样本更接近原音频且鲁棒性较好。

在生成攻击过程中, 转录结果与目标短语越接近, 花费的时间越长, 此时大多数的点已不需要被扰动。WPT 可以通过调整不同位置的失真权重来减少时间成本, 该方法用 Levenshtein 距离[80]来解决此问题, 其中专注于 Levenshtein 距离比较小但是不为 0 的那些点, 这些点被称为关键点, 同时对这些点赋予更大的权重, 减少攻击时间, WPT 一方面可以减少攻击的生成时间, 另一方面可以通过迭代次数减小全局搜索步长, 避免过度扰动而错过算法的最优值。

加权采样音频攻击是对 Carlini 和 Wagner 攻击方法的优化, 后者也尝试为标记序列的字符设置权重来加快收敛, 但是必须先通过 CTC 损失函数获得一个对抗样本 x_0, 再进行改进, 这将花费大量的时间, 加权采样音频攻击可以在任意迭代过程中获取关键位置的信息, 随时更改权重来加速收敛, 提高了对抗音频的生成速率。

3.4.3 基于感知的语音对抗攻击

在音频领域, 可以利用人类听觉系统的心理声学特性来产生更有效但较不易察觉的攻击, 这些依赖于掩蔽效应。当音频同时呈现给听觉系统时, 某些音频的频率难以察觉, 因人耳系统处理声音的敏感性与离散性, 当多个频率同时呈现给听众时, 在关键频带周围会发生掩盖效果[81]。Szurley 等人[61]提出了一种基于心理声学的语音攻击的新方案, 利用了原始信号的心理声学掩蔽阈值, 将对抗性信号嵌入到特定的听觉阈值以下, 从而确保其不被察觉, 对抗音频可以完全在时域中生成, 并使用 PESQ 来衡量感知失真。

作者提出了一种寻找全局掩盖阈值的方法, 该全局掩盖阈值表示基于信号的强度和频率分量以及人类听觉系统的心理声学特性的感知加权。目标总损失函数由两部分构成, 一部分是对抗损失函数, 另一部分是基于感知的损失函数。此外为了语音攻击更加实用, 提高语音信号在物理世界的鲁棒性, 作者使用声学室模拟器, 通过估计房间脉冲响应进行模拟训练。

3.4.4 语音识别的通用对抗攻击

不仅视觉领域存在通用对抗扰动, Neekhara 等人[22]首次证明了语音识别中通用对抗扰动的存在, 利用迭代优化生成的通用对抗扰动可以使目标模型的转录结果出现错误, 通用对抗扰动的存在会给 ASR 系统带来更严重的威胁, 攻击模型如图 5 所示:

攻击的目标是找到一个扰动 v, 使其添加到任意原始语音后, 会导致语音识别模型的转录错误, 为了使攻击成功, 转录错误率应该足够高, 期望的攻击效果如下:

$$CER(C(x), C(x+v)) > t \ \text{ for "most" } x \in \mu$$

其中 μ 表示数据集中语音样本的分布, C 表示被攻击的语音识别模型, t 为设置的阈值, CER 表示字错误率, 当字错误率大于特定阈值时才会被认为攻击成功。通用对抗扰动的构成需要遍历一个样本子集并逐步构建扰动向量, 若遍历 x 中所有样本所添加的扰动达

不到期望成功率, 则再次遍历迭代直至达到要求。在每次优化过程中, 要解决当前数据点的优化问题, 为此 Neekhara 等人提出一个新的优化方法来找寻将数据点推至决策边界的最小扰动向量。

Neekhara 等人为了表明语音识别模型面对通用对抗扰动的脆弱性, 针对 DeepSpeech 模型, 在验证集上最多可以达到 89.06%的攻击成功率, 同时可以以很低的失真实现攻击。总体而言, 虽然通用对抗攻击的效果不如有针对性的攻击, 但是它可以针对较少的训练集就可以产生不错的成功率且具有一定的可迁移性。

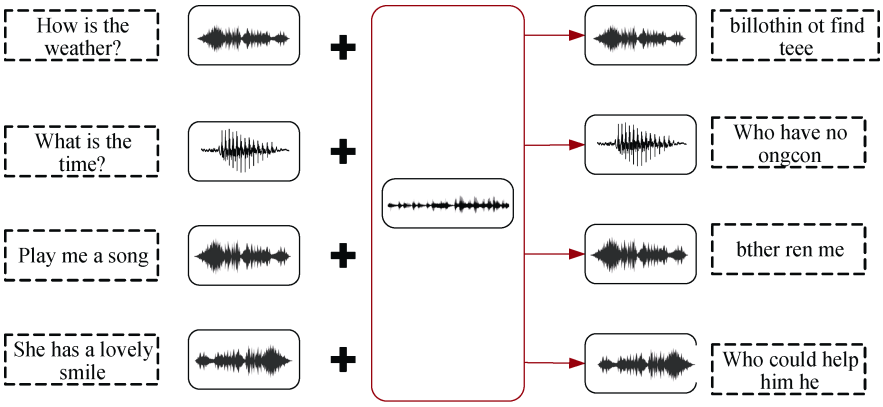

图 5　语音识别的通用对抗扰动

Figure 5　Universal adversarial perturbations of speech recognition

3.4.5　Taori 黑盒攻击

Alzantot 等人[18]提出的黑盒遗传攻击算法针对音频分类模型, Taori 等人[23]在此方法基础上进行了改进, 用来针对语音识别模型, 其难点在于用黑盒优化产生任意长度的目标短语。该方法由两部分构成, 第一部分应用遗传算法, 用 CTC 损失计算适应度, 但是在加入突变时会更新其突变概率, 受梯度下降动量更新的启发, 根据每一代适应性与上一代的差值自主改变突变概率:

$$p_{new} = \alpha \times p_{old} + \frac{\beta}{|currScore - prevScore|}$$

当两代适应性接近时, 此公式会为突变概率添加加速度, 当两代差别较大时, 会自动减小突变概率。既保证了收敛速度又防止陷入了局部最小点, 其中突变噪声用高通滤波器来减少人耳的感知程度。实际上遗传算法工作时, 能有效的搜索大量空间, 但当对抗音频接近目标时很难收敛, 此阶段只需在几个关键区域加扰动就可以得到成功的对抗音频, 因此在当前解码与目标解码的差异低于某个阈值时, 改用第二步, 用梯度估计[82]选择性的添加扰动来达到目标转录。

Taori 将遗传算法与梯度估计相结合, 利用两种方法的优点找到了高相似度的对抗样本, 尽管攻击 DeepSpeech 模型只有 35%的成功率, 但对语音识别黑盒攻击的研究有参考价值

3.4.6　基于脉冲响应的物理攻击

对语音识别的攻击大部分是数字攻击, 即直接将对抗样本输入给模型, 然而这些对抗样本应用在物理世界时会受到环境以及各种噪声的影响, 使得攻击无效。Yakura 和 Sakuma[24]提出用带通滤波器、脉冲响应(Impulse Response, IR)和高斯白噪声生成对抗音频来攻击 ASR 系统。首先他们考虑了人耳的感知频率, 用带通滤波器将扰动限制在规定范围。然后他们为了避免环境混响的影响, 利用来自不同环境的 IR 训练对抗样本以增强对混响的鲁棒性。最后为了减少噪声的影响, 他们又引入了高斯白噪声以增强对抗样本的鲁棒性。Yakura 和 Sakuma 测试了在物理世界中录制对抗语音以及通过无线电传输对抗语音, 两种情况在 DeepSpeech 模型中均有不错效果, 该方法生成对抗样本信噪比较低, 但最高可以达到 100%的攻击成功率。然而 Yakura 和 Sakuma 并没有评估在各种房间环境下攻击的成功率。

后来 Qin 等人[25]使用 RIR 来生成语音对抗样本, 目的是为了抵御环境干扰的影响, 具有一定的鲁棒性, 此外他们结合了听觉掩蔽的心理声学原理, 开发出了有效的且不易察觉的音频对抗样本, 针对 Lingvo[83]模型的实验表明, 对任意完整句子的有目标攻击可以达到 100%的成功率, 但他们的实验室只止步于模拟环境, 没能真正的应用在物理世界中以评估攻击的效果。

随着研究的发展, Schönherr 等人[26]利用 RIR 提出生成一种针对语音识别系统的一般性对抗音频。他们从不同的房间设置中采样 RIR, 将卷积与被采样的 RIR 作为 ASR 底层神经网络的附加层, 原始音

频在 RIR 的限制下生成可在空气信道传输的对抗音频。为了减少人对扰动的感知, 他们用基于心理学掩蔽的方法将扰动限制在原始音频的听觉阈值以下。Schönherr 等人针对 Kaldi[84]语音识别模型, 不需要先验知识的前提下就可以生成针对目标转录的对抗音频, 可适用于大部分房间设置, 具有很好的迁移性。

3.4.7 Metamorph: 针对语音控制系统的攻击

对语音进行物理世界攻击, 需要考虑音频信号在空气中的传播, 一般情况下, 经过信道失真的对抗样本不能成功地攻击语音系统, 之前对语音识别物理攻击的研究主要通过估计脉冲响应使生成的对抗音频具有鲁棒性。Chen 等人[27]用不一样的思路探索了利用语音攻击物理世界的语音识别系统。

他们认为经过信道失真后的音频 $H(x+\delta)$ 很难被模型识别为攻击者期望的目标, 如果知道失真 $H(.)$, 就可以恢复原始的对抗音频, 但是提前了解目标设备不现实。经实验表明无线传输对于语音攻击的影响主要是由多径传播和设备硬件造成的频率选择性引起的。虽然这两种频率选择源不能被区分以及精确评估, 但是作者认为在到达信道的频率选择性起主导作用的距离之前, 可以先提取聚合失真影响, 一旦这种主要影响被提取, 那么就可以将其加入到声音合成中。基于此, Chen 等人提出用两阶段法生成可在空气信道传播的对抗音频。首先从不同环境不同设备中收集一组 $H(.)$ 的测量结果来生成初始的扰动 δ, 这些测量结果捕捉到了频率选择性的主要影响成分, 此时生成的对抗音频具有一定的成功率, 为了增强攻击的效果, 他们又使用了域自适应算法来调整 δ, 以适应当前设备和环境的特征。

Chen 等人基于经验解释了无线环境下限制对抗音频的因素, 他们又进行了大量的真实实验来评估所提方法的性能, 针对 DeepSpeech 模型在 6 m 之内的攻击距离里可以实现 90%的成功率。

3.4.8 针对商业语音识别系统的黑盒攻击

Chen 等人[28]第一次提出了针对商业语音系统和智能语音设备的黑盒攻击, 他们生成一种“恶魔低语”, 可以攻击商业 ASR 系统(如 Google Home, Echo, Cortana 等)。这些“隐藏”的目标命令对人类来说是不可察觉的, 但可以被这些系统识别, 可以高效地生成能够欺骗商业 ASR 系统的对抗样本。

因为商业 ASR 系统的复杂以及攻击者所掌握信息的缺失, 以往的黑盒查询以及简单的模型替代很难保证有效性。Chen 等人提出了新的方法, 他们使用两个模型互补来生成更优秀的对抗样本。一个是使用先进的白盒模型 Kaldi 作为基模型, 另一个是近似目标模型的替代模型。具体而言, 先用制作的语料库来训练接近目标系统的替代模型, 再用基模型根据改进的算法生成具有一定迁移性的对抗样本, 将它传递给替代模型进行调整并更新对抗样本, 每间隔一段时间将它传入目标黑盒系统进行查询, 根据查询的结果再次通过基模型和替代模型增强对抗样本。最后使用有效的对抗样本攻击目标模型。

实验表明, 大约 1500 次查询所构建的替代模型可以很好的近似目标模型, 在 Kaldi 模型的帮助下对于大多数开源系统均可以实现极高的攻击成功率, 最后他们又根据用户的听觉反馈进行了对抗音频的可感知性评估, 证明了该方法的隐蔽性。

4 语音防御

在前面的章节中, 我们对语音领域的对抗攻击做了介绍, 这一节我们来研究语音对抗攻击的防御方法, 大体而言, 语音的防御方法和图像上的类似, 可简单分为两大类。一类是主动的防御: 通过强化或更改神经网络来增强其鲁棒性。另一类是对抗检测: 无需更改模型而是用各种手段在对抗样本输入到模型前检测出来。由于重新训练或者更改网络会增加成本, 所以对抗检测逐渐成为研究的一个热点。在视觉领域的防御方法有很多, 但据我们所知, 在现有的文献中, 语音领域的防御方法有限, 我们在表 2 中将语音对抗攻击的防御方法按照其防御的机理分为主动防御与对抗检测, 其中★表示防御效果。

4.1 主动防御

4.1.1 输入转换

语音识别中较为常用的防御方法是输入转换[31], 可以去除或者扭曲对抗噪声以减弱其影响, 且不影响原始语音的质量, 具体操作如下:

量化: 通过将音频采样数据的幅值取整为最接近 q 的整数倍, 因为扰动在输入空间中的振幅很小所以可以被破坏。通常选择 q=128, 256, 512, 1024 作为参数。

局部平滑: 用固定长度的滑动窗口进行局部平滑来减少对抗性扰动。对于某个音频样本 x_i, 通常考虑它之前与之后的 k–1 个样本, x_{i-k+1}, … , x_i, … , x_{i+k-1}, 把它们表示为参考序列, 并用参考序列的平均值或者中位数来代替 x_i。

下采样: 基于采样理论, 可以在不牺牲信号质量的情况下, 对带宽受限的音频文件进行下采样, 同时减轻重建过程中的对抗性扰动(如将原始 16 kHz 的音频数据下采样到 8 kHz)。

表 2 语音对抗攻击的防御方法

Table 2 Defense methods of speech adversarial attack

分类	方法	防御机制	应用场景	防御效果
主动防御	输入转换[31]	在数据输入模型前预处理, 保证语音质量的前提下去除对抗扰动	通用	★★
	对抗训练[85-86]	将攻击算法作为正则化项添加到损失函数中, 使模型对输入上的微小的变化具有鲁棒性, 平滑模型输出	说话人识别	★★
对抗检测	对抗子空间检测[32]	利用正常样本与对抗样本潜在子空间分布的差异进行检测	音频分类	★★★★
	噪声泛洪[33]	DNN 分类器相比于对抗输入, 对原始输入的自然噪声更具有鲁棒性	音频分类	★★★
	音频修改[34]	对抗音频对于音频修改的噪声具有敏感的变化, 细微的改变会导致转录结果发生大的变化	语音识别	★★★
	时间依赖性检测[35]	原始音频相对于对抗音频具有时间依赖性, 利用数据本身独特的属性进行防御	语音识别	★★★★
	激活量化检测[36]	模型量化后输出会产生量化误差, 根据量化误差检测对抗音频	语音识别	★★★
	多模型集成检测[37]	对抗音频的可迁移性较差, 利用多模型输出的相似度检测	语音识别	★★★★
	综合对抗检测[38]	声音混响与多段噪声填充相结合, 采用自适应步骤, 增加检测的复杂度	语音识别	★★★★

Yang 等人[31]提出的利用输入转换来防御对抗扰动, 最明显的优点是易于实现, 易于操作, 可以有效地减少对抗性干扰, 针对语音识别模型会使个别原始语音的转录结果出现错误。此外, 输入转换对于复杂的攻击防御效果不佳, 而且如果攻击者知道了预处理的操作细节(如量化参数、采样率), 很容易生成自适应攻击来绕过防御。

4.1.2 对抗训练

为了解决模型容易受到对抗攻击的问题, Arindam 等人[85]提出利用投影梯度下降(Projected gradient descent, PGD)攻击[30]、Carlini 和 Wagner(CW)攻击[77]生成的语音对抗样本来训练说话人识别模型, 对抗训练最初是由 Goodfellow 等人针对对抗图像提出来的, 可以提高模型抵御对抗攻击的能力。Arindam 等人将 PGD、CW(这些攻击通常被认为是强攻击)生成的对抗音频与原始音频混合一起来训练网络, 使网络对于音频对抗攻击具有鲁棒性, 以下是对抗训练的公式:

$$\arg\min_{\theta} E_{(x,y)\sim D}\ [\max_{\delta:\ \delta_p<\epsilon} L(x+\delta,\ y,\ \theta)]$$

此公式内部的最大化是利用攻击算法产生对抗样本, y 表示原始标签。函数外部的最小化是用对抗样本与原始样本对网络进行再训练, 优化损失函数, 增强模型防御攻击的能力。对抗训练的损失函数由两部分组成, 前半部分是模型对于原训练集的损失函数, 后半部分是模型相对于对抗样本的损失函数:

$$L_{AT}\left(x,\tilde{x},y,\theta\right)=\left(1-w_{AT}\right)\cdot L\left(x,y,\theta\right)+w_{AT}\cdot L(\tilde{x},y,\theta)$$

Wang 等人[86]针对说话人验证也提出了类似的想法, 他们提出了对抗正则化。他们认为如果生成的对抗性示例可以轻易使模型出错, 则意味着该模型的鲁棒性不足以抵抗对抗性扰动, 即该模型的输出对于输入不平滑, 因此用对抗正则化来增强模型鲁棒性。该方法试图找到当前数据点附近的最差点, 然后使用找到的最差数据点进行优化, 从而使模型对扰动具有鲁棒性, 并且输出分布更平滑。

对抗正则化原理与对抗训练相同, Wang 等人提出利用快速梯度符号法(FGSM)和局部分布平滑度(Local distributional smoothness, LDS)[87]来生成对抗样本, 其中 LDS 不需要真实标签就可以生成对抗样本, 它定义为模型分布 $p(x,\theta)$ 相对于输入扰动敏感度的负值。算法如下:

$$\Delta_{KL}(\delta,x,\theta)=KL[p(x,\theta)\|p(x+\delta,\theta)]$$

$$\delta_{L-adv}=\arg\max_{\delta}\left\{\Delta_{KL}\left(\delta,x,\theta\right);\delta_2\leqslant\epsilon\right\}$$

$$LDS(x,\theta)=-\Delta_{KL}\left(\delta_{L-adv},x,\theta\right)$$

KL 散度用来衡量扰动前后模型输出分布的差异, δ_{L-adv} 被称为虚拟对抗扰动, 用来破坏模型的分布, δ_{L-adv} 是 KL 散度上模型分布最敏感的方向, $\Delta_{KL}\left(\delta_{L-adv},x,\theta\right)$ 越小, 模型对于输入的分布越平滑。

目标是在所有观察到的输入附近提高模型的平滑度:

$$L_{LDS}(x;\theta) = L(x;\theta) - \alpha LDS(x,\theta)$$

这里 LDS 被作为正则化项以促进模型分布平滑, 它不使用标签信息就可以从模型中确定对抗方向, 也适用于半监督学习。

4.2 对抗检测

4.2.1 对抗子空间检测

为了探索原样本与对抗样本的潜在子空间的差异, Feinman 等人[88]在图像上通过估计核密度(Kernel density, KD)与贝叶斯不确定性(Bayesian uncertainty, BU)来检测对抗子空间, 他们认为对抗样本与真实数据的分布之间有一定的距离(对抗样本存在于稍远离真实数据的流形区域之外), 后来, MA 等人[89]用局部固有维数(Local intrinsic dimensionalit, LID)描述对抗样本所处区域的内在维度, 使用 LID 的估计检测对抗样本。但是经证明, 这些检测器在不利的条件下无法检测强大的对抗攻击。

Esmaeilpour 等人[32]在音频分类任务中提出了新的方法, 为了区分原始音频与对抗音频的向量空间, 他们研究了输入样本到广义 Schur 分解空间的映射(又称 QZ 分解)以及使用弦距离来识别其基础子空间。他们表明对任何原始音频与对抗音频来说, 语谱图特征值之间的弦距离一定满足约束条件, 可以用此来检测, 但实际上, 找到参考频谱图以及扰动的大小是不实际的, Feinman 等人提出比较原始音频与对抗音频的特征值并在它们间找到决策边界。从同类样本中随机选择一对语谱图, 使用 QZ 分解计算特征值, 因原样本与对抗样本 QZ 分解后特征值的内在分布不同, 所以用它们特征值训练一个二分类检测器, 测试阶段, 把音频转换为语谱图后再将 QZ 分解得到的特征值输入到检测器, 以此来检测对抗样本。

Feinman 等人将图像上对抗子空间以及流形分布的概念引入到音频上, 提出了一种新颖的方法来检测对抗样本, 与 KD、LID 等其他几种同类检测器相比拥有更好的效果。

4.2.2 噪声泛洪

Rajaratnam 等人[33]探索了利用随机噪声“淹没”音频信号的特定频段以检测对抗性示例, 这种方法不需要重新训练或者修改模型, 防御针对音频分类的对抗攻击可以取得不错的效果。

输入转换实际上是通过音频的预处理来消除对抗性扰动, 而噪声泛洪的思想是通过向信号添加噪声来抵御对抗性示例。作者认为相比于对抗输入, 神经网络分类器针对原始输入上的噪声拥有更好的鲁棒性。

实验表明, 与改变原始输入的类型相比, 改变对抗样本的预测类别通常需要更少的噪声, 因此可以通过观察模型预测产生变化之前需要向信号添加多少噪声来进行检测。此方法的核心是从音频信号中计算一个“泛洪分数”, 表示需要多少噪声来改变预测的结果。在训练过程中计算对抗样本与原始示例的“泛洪分数”, 从而找到一个理想阈值分数。此外, 作者利用带通滤波器对五个不同频段的噪声泛洪进行了测试, 并提出了一种集成防御的方法使评估阶段更复杂但比简单的噪声泛洪更有效。

针对 Alzantot 等人的黑盒攻击(相对大的扰动), 噪声泛洪可以达到 91.8%的检测率。但是对于其他复杂攻击的防御效果有待研究。噪声泛洪的方法可以和输入转换的方法结合, 进一步提高系统的防御能力。

4.2.3 音频修改

Kwon 等人[34]提出了一种利用音频修改来检测对抗样本的方法, 在音频中加入新的低失真, 原始样本的分类结果几乎没有影响, 但是对抗样本的分类结果会发生敏感的改变。该检测方法可以用来防御针对语音识别的攻击。

音频修改的检测方法分为两个步骤, 对于给定音频样本, 首先通过模型得出分类结果。其次生成经过音频修改后的原始样本后将它再次输入到模型进行识别。如果两次分类结果的差异较大就视为对抗样本, 差异较小视为原始样本。此检测方法与 Rajaratnam 等人提出的噪声泛洪有相似之处。事实上, 在生成音频对抗样本的过程中, 某些失真会添加到原始音频上直到机器开始误解信号, 因此对抗样本对于失真更具有敏感性, 若再添加由音频修改引起的失真, 则对抗样本在分类结果上的差异要大于原始示例。

实验表明, 对 CW 语音攻击[21]在 DeepSpeech 模型上生成的对抗样本进行音频修改, 对抗音频在 12db 情况下的转录精度只有 10%以下。此外, 音频修改的方法有许多, 可以用低通滤波器、高通滤波器、陷波滤波器等, 不同音频修改的集成组合可以提高对抗样本的检测效果。

4.2.4 时间依赖性检测

无论是图像还是音频, 基于深度神经网络的应用都取得了良好的效果, 但这两种不同类型的数据导致了神经网络的学习过程有明显的区别。图像中, 像素的空间连续性对应于相邻像素点间的联系和颜色描述, 卷积神经网络利用这些信息进行特征提取。在音频中, 时序数据具有明显的时间依赖性, 利用

递归神经网络可以很好地处理。受空间连续性在图像分割任务中可以提高鲁棒性的启发，Yang 等人[35]提出利用语音数据的时间依赖性来检测对抗样本，该方法可以有效地防御对抗攻击，且对原始样本的影响较小，Yang 等人还证明了即使是了解时间依赖性方法细节而生成的自适应攻击也不能绕过该防御。

时间依赖性检测是根据数据独特的属性开发的一种防御手段，如图 6 所示，给定一个语音序列，选择前 k 个部分(即长度为 k 的前缀)作为 ASR 的输入，获得的转录结果为 S_k。再将整个音频序列输入到 ASR 系统，并选择转录结果长度为 k 的前缀 S'_k，该长度与 S_k 相同。最后根据时间依赖距离比较 S_k 和 S'_k 之间的一致性。对于正常样本，因为语音识别模型由于它的时间依赖性而对给定序列的不同部分是连续的，所以两次转录结果是相似的。但对于对抗样本，增加的扰动旨在将 ASR 转录为目标文本，所以无法保持原始序列的时间信息，由于丧失了时间依赖性，所以 S_k 和 S'_k 无法产生相似的结果。

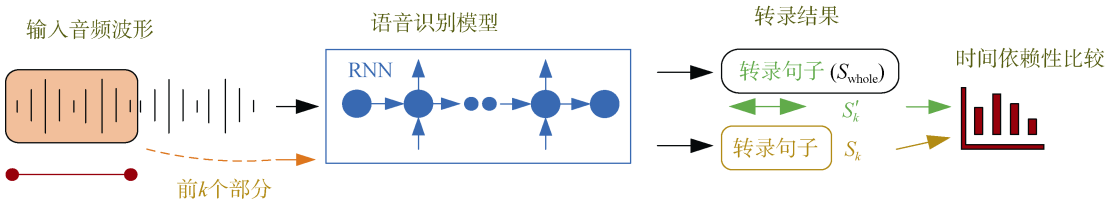

图 6　时间依赖性检测框架图

Figure 6　The framework diagram of temporal dependency detection

基于时间依赖性的检测方法不仅在防御对抗扰动方面起到了作用，还凸显了利用数据的独特属性来构建可靠的机器学习模型的重要性，对其他领域的数据安全方面提供了一个参考方向。

4.2.5　激活量化检测

DNN 量化是一种压缩模型大小并降低计算复杂度的技术。Liu 等人[36]提出了一种使用量化神经网络来检测对抗音频信号的方法。实际上，神经网络的权重和激活量化会显著地减少模型的大小且精度下降幅度有限[90]，这两种量化等效于使 DNN 损失函数的假设空间离散化，因此与全精度模型相比，量化的模型会导致量化误差。作者提出的方法就是用激活量化得到的量化误差来区分对抗音频。

量化神经网络激活的第一步是限制激活(如 ReLu 函数范围)，对于给定的量化位宽 k，激活范围被离散为 2^k 个区域，在音频数据输入到下一层之前，将激活映射到较低的精度值。

神经网络激活量化后，原始样本与对抗样本之间会存在明显的差异结果，与全精度模型相比，量化误差称为量化模型上的性能下降，结果可以用字符错误率来衡量，将量化误差高于(低于)阈值的音频视为对抗性(原始)音频。Liu 等人针对 DeepSpeech 模型，此模型由五个完全连接层和一个双向 RNN 层构成，因不同的量化位宽对网络有较大的影响，因此在量化网络激活时，需要对模型不同类型的层选择不同的量化位宽以及确定最佳的检测阈值。

Liu 等人首次将 DNN 的模型量化用到音频的对抗性检测上，对不同的攻击方法均取得 90%以上的检测效果，神经网络本身的冗余性以及对抗样本的敏感性为我们研究对抗检测提供了不一样的思路。

4.2.6　多模型集成检测

迁移性较差一直是对抗攻击的一个弱点，现存的大多数音频攻击方法都是针对特定模型产生的，即使是通用对抗扰动，迁移到其他模型后，成功率也会明显下降。基于此，给定一个音频对抗样本，用不同数据集训练的不同 ASR 系统转录结果应该也会有差异，Zeng 等人[37]受多版本编程的启发提出一种新颖的音频检测方法，利用各种现有的 ASR 模型来确定音频是否是对抗样本。

多版本编程(Multi-version programming，MVP)的主要思想是基于同一规范独立开发多个程序，多个程序同时运行并执行相同的任务，在每个检查点，会检查每个程序执行的一致性，最初用来作为防御软件缺陷的方法。而 Zeng 等人认为不同的 ASR 在 MVP 中可以被视为“独立开发的程序”，它们的任务都是把音频转录为文字，对于正常音频，它们的输出是非常相似的。然而，对抗音频因被视为“漏洞”，不能欺骗所有的 ASR 系统。

如图 7 所示，检测框架由目标 ASR、辅助 ASR、相似度计算模块、二分类器构成。音频输入之后，各个 ASR 系统会对它进行转录，根据结果的相似度预测音频是否为对抗样本。此检测器本质上是使用相

似度评分而非对抗样本来进行训练。Zeng 等人证明此方法对于不可见攻击的检测效果依然很好，即用来检测没有出现在训练集中的攻击。为了提高检测网络的鲁棒性，他们模拟出迁移性强的对抗音频进行训练(这些样本非音频，而是用特征向量来表示)，因此只要对抗音频无法欺骗检测系统中所有的模型，检测都可以保持有效，使得多模型集成检测方法领先于攻击者一步。实验表明，该检测系统针对 CW 白盒攻击[21]以及有目标的黑盒攻击[23]均能达到 98%以上的检测率。

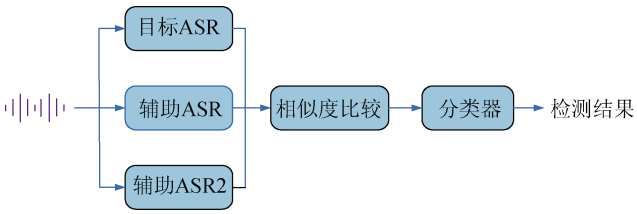

图 7　多模型检测框架

Figure 7　The framework diagram of Multi-model detection

4.2.7　综合对抗检测

Miyato 等人[38]提出了可以检测自适应音频对抗样本的综合对抗检测，该框架将噪声填充与声音混响相结合，利用算法构建具有多种随机性的操作，可以防止攻击者查询特定信息，提高了检测的准确率。

此检测方法由两部分操作构成，第一步通过房间脉冲响应(RIR)将输入的音频转换为混响音频，在修改对抗噪声的同时保证原始音频的结果不改变，第二步用多段噪声填充方法来破坏对抗样本的连续性，最后比较转录相似度来检测对抗音频。

检测方法如图 8 所示，具体而言，对于输入音频，先用语音活动检测器(Voice activity detection, VAD)[91]检测静默片段来确定干扰等级以自动修改防御的复杂性，根据复杂性将随机选择多个 RIR 来模拟不同环境下带混响的语音。事实上，对抗噪声经过混响后会失效，对于经过良好训练的模型，可以正常识别经过混响后的原始样本，这是此检测算法的第一道保障。为了进一步消除对抗性，作者提出了一种多片段噪声填充，用 VAD 检测出静默片段，将每个检测出来的静默片段作为剪切中心，把输入语音分为多个，在每个剪切后的语音间植入短时高斯噪声以此破坏对抗音频的连续性，这一步与 Yang 等人根据时间连续性检测有相似之处，但通过将噪声插入到音频的无声位置可以减小对正常样本的影响，此为第二道保障。

综合检测中的声音混响与多段噪声填充在功能上互补，不仅增加了防御的复杂性，还可以同时检测出白盒与黑盒攻击，实验表明了检测效果高于音频修改、噪声泛洪、时间依赖性检测等方法。

5　总结与展望

本文对语音领域对抗攻击与防御方法进行了一个比较详细的整理与介绍。本节我们对语音的攻击与防御方法进行总结，探讨它们的优缺点并对未来可能的研究方向进行展望。

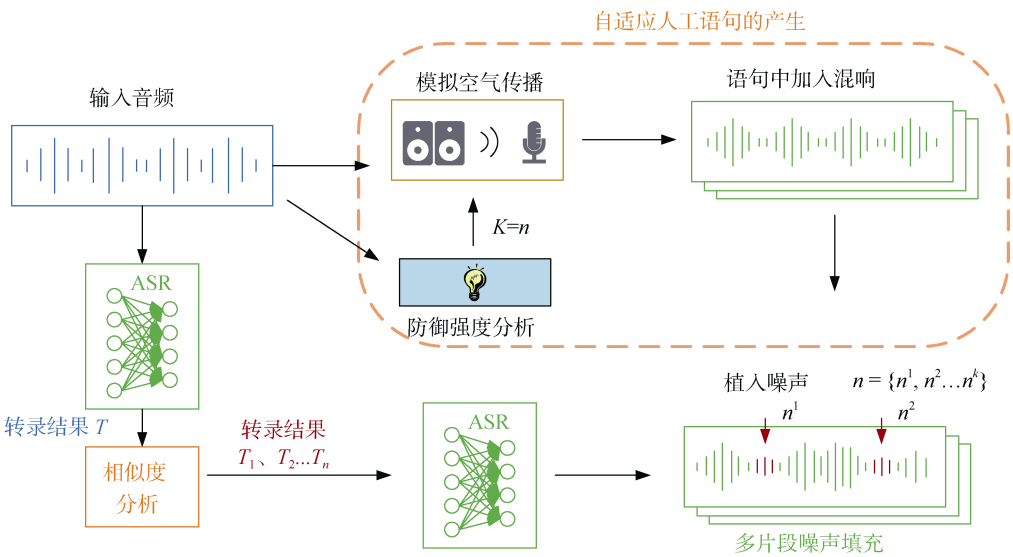

图 8　综合对抗检测的框架图

Figure 8　The framework diagram of comprehensive adversarial detection

5.1 攻击方法

语音对抗样本的生成一方面可以在时域或者频域利用视觉领域中常见的攻击方法, 也可以直接将其转化为声学特征并利用基于梯度或优化的方法。由于时间序列的特殊性, 语音任务在某一时刻的输出不仅依赖于样本在当前时刻的特征, 还依赖前后时刻特征的聚合, 所以可以只在音频的某一小段添加扰动从而干扰整个音频, 但此方法是否对长音频有效以及要攻击哪些帧是未来的研究重点。现阶段对抗攻击的研究主要集中在白盒, 但对于语音攻击来说, 迁移性强的黑盒攻击可以直接针对开源模型以及商用语音系统, 实现高成功率的同时确保其隐蔽性是需要研究的重点。此外, 许多语音攻击可以作用在物理世界, 成功的攻击了短单词或者说话人身份并在空气信道传播后有效, 因此语音对抗攻击下一步的研究可以放在对抗音频鲁棒性方面上, 在各种环境干扰下依然有效、独立于神经网络以及良好的可迁移性对研究对抗攻击的内在机理有重要意义。

5.2 防御方法

针对不同的语音攻击方法, 主要可以通过对抗防御、对抗检测这两个方面进行防御。输入转换是常用的防御方法, 但是对复杂攻击的防御效果明显下降而且会影响正常样本的精度。对抗训练通过再训练模型, 增强模型的识别能力与鲁棒性, 可以一定程度上防御对抗攻击, 但通常需要考虑成本的损耗。对抗检测可以提前筛选输入样本, 有效地降低识别错误率。了解不同对抗攻击的特点以及使用不同的防御方法能显著提高效果, 如噪声填充或音频修改可以有效地防御白盒攻击, 但对黑盒攻击有较差的检测效果, 因为黑盒攻击不需要模型特定的信息, 生成的扰动相对较大但同时有较好的鲁棒性, 而时间依赖性检测却可以很好地防御黑盒攻击。图像分类上对抗样本的防御与检测方法已相对成熟, 但是在语音中的效果如何是下一步可以研究的方向。攻击的方法始终领先防御一步, 后续的防御方向不仅要注重自适应攻击, 还要结合不同防御方法的优点, 构建集成防御方案来更好的抵御语音攻击。

5.3 安全性分析

目前语音领域的对抗攻击没有局限在数字攻击, 在物理世界的攻击也取得了一定进展, 深度学习虽然取得了相当好的效果, 但它的弊端不可忽视, 本文介绍攻击算法的根本目的是为了分析目前此领域的发展, 以此可以更全面的生成对抗样本, 进行对抗防御的研究。事实上, 除了语音的对抗攻击外, 还存在语音的隐藏语音攻击(攻击者将一些信息如指令嵌入到音频载体中, 在不引起用户注意的情况下使目标模型识别所需信息)、欺骗攻击(使用语音重放、合成、转换等方法获得一段与原说话者相同的语音, 可以误导说话人识别系统)、中毒攻击(通过在模型训练阶段中毒数据集或者直接中毒模型使得测试阶段模型的输出产生期望的结果)等, 这些都是影响语音安全的重要隐患。语音交互的使用愈加广泛, 安全性研究越来越重要, 攻击与防御的不断博弈促使着网络安全技术不断发展。

参考文献

[1] Reddy D R. Speech Recognition by Machine: A Review[J]. *Proceedings of the IEEE*, 1976, 64(4): 501-531.

[2] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling In Speech Recognition: The Shared Views of Four Research Groups[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 82-97.

[3] Goodfellow I, Bengio Y, Courville A. *Deep learning*[M]. MIT press, 2016.

[4] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. *Nature*, 2015, 521(7553): 436-444.

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.

[6] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[7] Huang G, Liu Z, van der Maaten L, et al. Densely Connected Convolutional Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261-2269.

[8] Held D, Thrun S, Savarese S. Learning to Track At 100 FPS with Deep Regression Networks[C]. *European conference on computer vision*, 2016: 749-765.

[9] Valmadre J, Bertinetto L, Henriques J, et al. End-to-End Representation Learning for Correlation Filter Based Tracking[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 5000-5008.

[10] Graves A, Mohamed A R, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[C]. *2013 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2013: 6645-6649.

[11] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]. *International conference on machine learning*, 2014: 1764-1772.

[12] Zhang Y, Pezeshki M, Brakel P, et al. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks[EB/OL]. ArXiv preprint ArXiv:1701.02720, 2017.

[13] Mousavi A, Baraniuk R G. Learning to Invert: Signal Recovery via Deep Convolutional Networks[C]. *2017 IEEE International Con-*

ference on Acoustics, *Speech and Signal Processing*, 2017: 2272-2276.

[14] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. ArXiv preprint ArXiv:1312.6199, 2013.

[15] Biggio B, Corona I, Maiorca D, et al. Evasion Attacks Against Machine Learning At Test Time[C]. *Joint European conference on machine learning and knowledge discovery in databases*, 2013: 387-402.

[16] Wiyatno R R, Xu A Q, Dia O, et al. Adversarial Examples In Modern Machine Learning: A Review[EB/OL]. ArXiv preprint ArXiv:1911.05268, 2019.

[17] 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.

[18] Alzantot M, Balaji B, Srivastava M. Did You Hear That? Adversarial Examples Against Automatic Speech Recognition[EB/OL]. ArXiv preprint ArXiv:1801.00554, 2018.

[19] 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.

[20] Li Z H, Wu Y, Liu J, et al. AdvPulse: Universal, Synchronization-Free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 1121-1134.

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

[22] Neekhara P, Hussain S, Pandey P, et al. Universal Adversarial Perturbations for Speech Recognition Systems[EB/OL]. ArXiv preprint ArXiv:1905.03828, 2019.

[23] Taori R, Kamsetty A, Chu B, et al. Targeted Adversarial Examples for Black Box Audio Systems[C]. *2019 IEEE Security and Privacy Workshops*, 2019: 15-20.

[24] Yakura H, Sakuma J. Robust Audio Adversarial Example for a Physical Attack[EB/OL]. ArXiv preprint ArXiv:1810.11793, 2018.

[25] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition[C]. *International conference on machine learning. PMLR*, 2019: 5231-5240.

[26] 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.

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

[28] 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.

[29] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. ArXiv preprint ArXiv:1412.6572, 2014.

[30] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[EB/OL]. 2017: arXiv: 1706.06083[stat.ML]. https://arxiv.org/abs/1706.06083

[31] Yang Z, Li B, Chen P Y, et al. Towards mitigating audio adversarial perturbations[J]. 2018.

[32] Esmaeilpour M, Cardinal P, Koerich A L. Detection of Adversarial Attacks and Characterization of Adversarial Subspace[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2020: 3097-3101.

[33] 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.

[34] 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.

[35] Yang Z L, Li B, Chen P Y, et al. Characterizing Audio Adversarial Examples Using Temporal Dependency[EB/OL]. ArXiv preprint ArXiv:1809.10875, 2018.

[36] Liu H, Ditzler G. Detecting Adversarial Audio via Activation Quantization Error[C]. *2020 International Joint Conference on Neural Networks*, 2020: 1-7.

[37] Zeng Q, Su J H, Fu C L, et al. A Multiversion Programming Inspired Approach to Detecting Audio Adversarial Examples[C]. *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2019: 39-51.

[38] Du X, Pun C M, Zhang Z. A Unified Framework for Detecting Audio Adversarial Examples[C]. *The 28th ACM International Conference on Multimedia*, 2020: 3986-3994.

[39] Schuller B W. Speech Emotion Recognition[J]. *Communications of the ACM*, 2018, 61(5): 90-99.

[40] Petrushin V A. Emotion recognition in speech signal: experimental study, development, and application[C]. *Sixth international conference on spoken language processing*. 2000: 222-225.

[41] Fragopanagos N, Taylor J G. Emotion Recognition In Human-Computer Interaction[J]. *Neural Networks*, 2005, 18(4): 389-405.

[42] Zhou G, Hansen J H L, Kaiser J F. Nonlinear Feature Based Classification of Speech under Stress[J]. *IEEE Transactions on Speech and Audio Processing*, 2001, 9(3): 201-216.

[43] Anvarjon T, Mustaqeem, Kwon S. Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features[J]. *Sensors* (*Basel*, *Switzerland*), 2020, 20(18): 5212.

[44] Wan L, Wang Q, Papir A, et al. Generalized End-to-End Loss for Speaker Verification[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 4879-4883.

[45] Nucci A, Keralapura R. Hierarchical Real-Time Speaker Recognition for Biometric VoIP Verification and Targeting: US8160877[P]. 2012-04-17.

[46] Becker T, Jessen M, Grigoras C. Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models[C]. *Ninth*

Annual Conference of the International Speech Communication Association. 2008: 1505-1508.

[47] Fortuna J, Sivakumaran P, Ariyaeeinia A, et al. Open-Set Speaker Identification Using Adapted Gaussian Mixture Models[C]. *Ninth European Conference on Speech Communication and Technology*. 2005: 1997-2000.

[48] Liu T T, Guan S X. Factor Analysis Method for Text-Independent Speaker Identification[J]. *Journal of Software*, 2014, 9(11): 2851-2860.

[49] Dehak N, Kenny P J, Dehak R, et al. Front-End Factor Analysis for Speaker Verification[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2011, 19(4): 788-798.

[50] Reynolds D A, Quatieri T F, Dunn R B. Speaker Verification Using Adapted Gaussian Mixture Models[J]. *Digital Signal Processing*, 2000, 10(1/2/3): 19-41.

[51] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]. *Interspeech 2017*, 2017: 999-1003.

[52] Snyder D, Garcia-Romero D, Sell G, et al. X-Vectors: Robust DNN Embeddings for Speaker Recognition[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 5329-5333.

[53] Li C, Ma X K, Jiang B, et al. Deep Speaker: An End-to-End Neural Speaker Embedding System[EB/OL]. ArXiv preprint ArXiv:1705.02304, 2017, 650.

[54] Wang D, Li L T, Tang Z Y, et al. Deep Speaker Verification: Do we Need End to End? [C]. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017: 177-181.

[55] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling In Speech Recognition[J]. *Ieee Signal Processing Magazine*, 2012(November): 82-97.

[56] Google Assistant. https://assistant.google.com.

[57] Aspire. https://github.com/kaldi-asr/kaldi/tree/master/egs/aspire.

[58] Amazon Alexa. https://developer.amazon.com/alexa.

[59] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[EB/OL]. 2010: arXiv: 1003.4083[cs.MM]. https://arxiv.org/abs/1003.4083

[60] Sonawane A, Inamdar M U, Bhangale K B. Sound Based Human Emotion Recognition Using MFCC & Multiple SVM[C]. *2017 International Conference on Information*, *Communication*, *Instrumentation and Control*, 2017: 1-4.

[61] Itakura F. Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals[J]. *The Journal of the Acoustical Society of America*, 1975, 57(S1): S35.

[62] Hermansky H. Perceptual Linear Predictive (PLP) Analysis of Speech[J]. *The Journal of the Acoustical Society of America*, 1990, 87(4): 1738-1752.

[63] Graves A, Fernández S, Gomez F, et al. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks[C]. *The 23rd international conference on Machine learning - ICML '06*, 2006: 369-376.

[64] Hannun A, Case C, Casper J, et al. Deep Speech: Scaling up End-to-End Speech Recognition[EB/OL]. 2014: arXiv: 1412.5567 [cs.CL]. https://arxiv.org/abs/1412.5567

[65] Gong Y, Poellabauer C. Crafting adversarial examples for speech paralinguistics applications[EB/OL]. ArXiv preprint ArXiv:1711.03280, 2017.

[66] Li J G, Zhang X F, Xu J Z, et al. Learning to Fool the Speaker Recognition[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2020: 2937-2941.

[67] 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.

[68] Michel Koerich K, Esmailpour M, Abdoli S, et al. Cross-Representation Transferability of Adversarial Attacks: From Spectrograms to Audio Waveforms[J]. *2020 International Joint Conference on Neural Networks*, 2020: 1-7.

[69] Liu X L, Wan K, Ding Y F, et al. Weighted-Sampling Audio Adversarial Example Attack[J]. *The AAAI Conference on Artificial Intelligence*, 2020, 34(4): 4908-4915.

[70] Szurley J, Kolter J Z. Perceptual Based Adversarial Audio Attacks[EB/OL]. ArXiv preprint ArXiv:1906.06355, 2019.

[71] Rix A W, Beerends J G, Hollier M P, et al. Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs[C]. *2001 IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*. *Proceedings*, 2001: 749-752.

[72] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network[C]. *2016 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2016: 5200-5204.

[73] Ravanelli M, Bengio Y. Speaker Recognition from Raw Waveform with SincNet[C]. *2018 IEEE Spoken Language Technology Workshop*, 2018: 1021-1028.

[74] Scheibler R, Bezzam E, Dokmanić I. Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 351-355.

[75] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[76] Ilyas A, Engstrom L, Athalye A, et al. Black-Box Adversarial Attacks with Limited Queries and Information[EB/OL]. ArXiv preprint ArXiv:1804.08598, 2018.

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

[78] Li Z H, Shi C, Xie Y, et al. Practical Adversarial Attacks Against Speaker Recognition Systems[C]. *The 21st International Workshop on Mobile Computing Systems and Applications*, 2020: 9-14.

[79] Hannun A. Sequence Modeling with CTC[J]. *Distill*, 2017, 2(11): e8.

[80] Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals[J]. *Soviet Physics Doklady*, 1966, 10(8):

707-710.

[81] Oxenham A J. Mechanisms and Mechanics of Auditory Masking[J]. *The Journal of Physiology*, 2013, 591(10): 2375.

[82] Bhagoji A N, He W, Li B, et al. Exploring the Space of Black-Box Attacks on Deep Neural Networks[EB/OL]. ArXiv preprint ArXiv:1712.09491, 2017.

[83] Shen J, Nguyen P, Wu Y H, et al. Lingvo: A Modular and Scalable Framework for Sequence-to-Sequence Modeling[EB/OL]. ArXiv preprint ArXiv:1902.08295, 2019.

[84] Kaldi ASR. http://kaldi-asr.org.

[85] 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: 101199.

[86] Wang Q, Guo P C, Sun S N, et al. Adversarial Regularization for End-to-End Robust Speaker Verification[C]. *Interspeech 2019*, 2019: 4010-4014.

[87] Miyato T, Maeda S, Koyama M, et al. Distributional smoothing with virtual adversarial training[EB/OL]. ArXiv preprint ArXiv:1507.00677, 2015.

[88] Feinman R, Curtin R R, Shintre S, et al. Detecting Adversarial Samples from Artifacts[EB/OL]. ArXiv preprint ArXiv: 1703.00410, 2017.

[89] Ma X J, Li B, Wang Y S, et al. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality[EB/OL]. ArXiv preprint ArXiv:1801.02613, 2018.

[90] Cai Z W, He X D, Sun J, et al. Deep Learning with Low Precision by Half-Wave Gaussian Quantization[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 5406-5414.

[91] Sohn J, Sung W. A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation[C]. *The 1998 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, *ICASSP '98*, 1998: 365-368.

徐东伟 于 2014 年在北京交通大学交通安全工程专业获得博士学位。现任浙江工业大学副教授。研究领域为交通信息处理、交通复杂网络、机器学习。研究兴趣包括: 人工智能、信号分析。Email: dongweixu@zjut.edu.cn

蒋斌 于 2018 年在陕西科技大学机电工程学院机械电子工程专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 对抗攻击、攻击检测与防御、深度学习。Email: jiangbin_1996@163.com

房若尘 于 2020 年在铜陵学院建筑电气与智能化专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 对抗攻击及防御、中毒攻击及防御、深度学习。Email: frc4045117@163.com

宣琦 于 2008 年在浙江大学控制科学与工程专业获得博士学位。现任浙江工业大学网络空间安全研究院教授。研究领域为人工智能安全、网络数据挖掘、信号智能。研究兴趣包括: 人工智能、信号分析、网络科学。Email: xuanqi@zjut.edu.cn