基于噪声破坏和波形重建的声纹对抗样本防御方法

魏春雨[1], 孙 蒙[1], 张雄伟[1], 邹 霞[1], 印 杰[2]

[1]陆军工程大学 指挥控制工程学院 南京 中国 210007
[2]江苏警官学院 南京 中国 210031

摘要 语音是人类最重要的交流方式之一。语音信号中除了文本内容外, 还包含了说话人的身份、种族、年龄、性别和情感等丰富的信息, 其中说话人身份的识别也被称为声纹识别, 是一种生物特征识别技术。声纹具有获取方便、容易保存、使用简单等特点, 而深度学习技术的进步也极大地促进了识别准确率的提升, 因此, 声纹识别已被应用于智慧金融、智能家居、语音助手和司法调查等领域。另一方面, 针对深度学习模型的对抗样本攻击受到了广泛关注, 在输入信号中添加不可感知的微小扰动即可导致模型预测结果错误。对抗样本的出现对基于深度学习的声纹识别也将造成巨大的安全威胁。现有声纹对抗样本防御方法会不同程度地影响正常样本的识别, 并且局限于特定的攻击方法或识别模型, 鲁棒性较差。为了使对抗防御能够兼顾纠正错误输出和准确识别正常样本两个方面, 本文提出一种“破坏+重建”的两阶段对抗样本防御方法。第一阶段, 在对抗样本中添加具有一定信噪比幅度限制的高斯白噪声, 破坏对抗扰动的结构进而消除样本的对抗性。第二阶段, 利用提出的名为 SCAT-Wave-U-Net 的语音增强模型重建原始语音样本, 通过在 Wave-U-Net 模型结构中引入 Transformer 全局多头自注意力和层间交叉注意力机制, 使改进后的模型更有助于防御声纹对抗样本攻击。实验表明, 提出的防御方法不依赖于特定声纹识别系统和对抗样本攻击方式, 在两种典型的声纹识别系统下对多种类型对抗样本攻击的防御效果均优于其他预处理防御方法。

关键词 声纹识别; 噪声破坏; 语音增强; 对抗样本防御
中图法分类号 TP391.9 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2024.01.05

Defense of Speaker Recognition Against Adversarial Examples Based on Noise Destruction and Waveform Reconstruction

WEI Chunyu[1], SUN Meng[1], ZHANG Xiongwei[1], ZOU Xia[1], YIN Jie[2]

[1] College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China
[2] Jiangsu Police Institute, Nanjing 210031, China

Abstract Voice is one of the most import ways of human communications. Besides texts, voice signals also hold the information of the speaker's identity, race, age, gender, and emotion, where the recognition of speaker identity is also called speaker recognition which is a biometric technique. Given the fact that human voice is easy to be collected and saved, and that the development of deep learning improves the recognition accuracy, speaker recognition has been used in financial APP authentication, smart home, voice assistant and forensics. On the other hand, adversarial attacks against deep learning models have attracted great attention, which could make the models' predictions incorrect by adding imperceptible perturbations to input signals. Therefore, the emergence of adversarial examples also poses the same serious security threat to deep learning-based speaker recognition. In this paper, a two-stage method with "destructing" and "reconstructing" is proposed to defense against adversarial examples of speaker recognition by overcoming the shortcomings of existing defense methods, such as the inability to remove adversarial perturbations, the negative impacts on the recognition of normal examples, and the poor robustness to different models and attack methods. At the first stage, Gaussian noises with a certain range of SNR amplitudes are added to the input speech signal to destroy the structure of potential adversarial perturbations and to eliminate its adversarial function. At the second stage, the proposed speech enhancement model named SCAT-Wave-U-Net is used to reconstruct the original clean speech. Global multi-head self-attention of Transformer and interlayer cross-attention mechanisms are introduced into the Wave-U-Net structure, which is more useful for defending the speaker adversarial examples. Experimental results show that the effectiveness of the proposed defense method does not depend on the specific speaker recognition system and the adversarial example attack method. By conducting extensive experiments on two state-of-the-art speaker recognition systems, i.e., i-vector and x-vector, the performances of the defense against multiple types of adversarial examples are superior to other de-

通讯作者: 孙蒙, 博士, 副教授, Email: sunmeng@aeu.edu.cn。
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. 62371469, No. 62071484)资助。
收稿日期: 2022-05-08; 修改日期: 2022-07-06; 定稿日期: 2023-09-27

fense methods using preprocessing techniques.

Key words speaker recognition; noise destruction; speech enhancement; defense of adversarial examples

1 引言

近年来, 深度学习在语音、图像等识别任务中展现了优异的性能。然而, 研究表明, 深度学习模型容易受到在样本中添加小幅度扰动的影响, 这些受到扰动的非正常样本被称为“对抗样本”[1]。通过在音频中加入微小的扰动使声纹识别(Speaker Recognition)系统出错[2]的样本被称为声纹对抗样本。由于对抗样本具有很小的扰动失真, 人们从听觉上很难察觉到异常变化。对抗样本的出现对深度学习模型的安全性提出了严峻挑战。随着基于深度学习的声纹识别技术在金融、安防、智能家居等领域的广泛应用, 声纹识别系统中对抗样本的防御就成为亟待解决的重要课题。

现有的声纹对抗样本防御方法可分为对抗样本检测、对抗训练以及样本变换处理三种[3]。这些方法在不同程度上存在丢弃样本、泛化性能差、真实样本识别率降低等缺点。另一方面, 为了去除语音中的各种噪声, 近年来涌现出了大量的基于深度学习的语音增强方法[4-6]。从对抗样本的生成过程来分析, 对抗扰动也可以看成是一种幅度较小的加性噪声[7]。如何将对抗样本防御和语音增强有效结合, 使语音增强有助于去除对抗噪声, 进而减弱对抗样本带来的不利影响, 是一个非常有价值的研究方向。

为了解决这些问题, 本文借助语音增强从对抗样本中恢复出原始波形, 提出一种结合噪声破坏与波形重建的声纹对抗样本防御方法。该方法首先在对抗样本中加入高斯白噪声以破坏对抗扰动的结构, 然后利用改进的语音增强模型重建原始波形, 从而实现对对抗样本攻击的防御。

2 相关工作

本文以噪声破坏和波形重建相结合的方式来防御声纹对抗样本攻击, 通过语音增强重建原始音频样本。首先总结声纹对抗样本攻防和语音增强方面的相关工作如下:

2.1 声纹对抗样本的攻击与防御

2.1.1 声纹对抗样本攻击方法

根据攻击者是否了解被攻击模型的信息, 声纹对抗样本攻击可分为白盒攻击和黑盒攻击, 根据是否迫使声纹识别系统输出指定的目标标签又分为有目标攻击和非目标攻击。在声纹对抗样本攻击的发展历程中出现了一些具有代表性的研究。

1) FGSM

Gong 等[8]将快速梯度符号法(Fast Gradient Sign Method, FGSM)用于生成声纹对抗样本。FGSM 通过一步梯度上升在输入 x 中添加扰动以最大化损失函数, 计算公式如下:

$$\hat{x} = x + \varepsilon \operatorname{sign}(\nabla_x f(x, y)) \tag{1}$$

其中, ε 是梯度上升的步长, $f(x, y)$是将输入 x 分类为说话人标签 y 的损失函数。

2) PGD

Liu 等[9]将迭代梯度下降法(Projected Gradient Descent, PGD)应用于声纹识别系统。PGD 是 FGSM 的改进版本。在每次迭代中, PGD 以步长 α 应用 FGSM 并裁剪结果以确保其在原始输入 x 的 ε 邻域内, 第 i 次迭代后的样本为,

$$x^i = \operatorname{clip}_{x,\varepsilon}(x^{i-1} + \alpha \operatorname{sign}(\nabla_x f(x^{i-1}, y))) \tag{2}$$

在求解对抗样本之前, PGD 攻击为原始样本增加一个随机的扰动[10], 这有助于攻击方找到更好的损失函数局部最大值。

3) Carlini & Wagner(CW)

Carlini 和 Wagner[11]针对语音识别系统提出的 CW 攻击方法也被用于攻击声纹识别系统。CW 方法将对抗样本的求解定义为一个优化问题, 用一个权重因子调节目标函数中对抗样本的有效性与不可感知性之间的相对重要程度。用 $f(x, y)$度量有效性, 当且仅当攻击成功时损失函数 $f(x, y) \leqslant 0$ 。用对抗样本和原始样本之间的 L_2 和 L_∞距离来度量不可感知性, 由此产生了 CW 攻击的两个版本, 即 CW_2 和 CW_∞。CW 攻击使用参数 κ 度量扰动的强度, κ 越大, 对抗样本攻击性越强, 但同时也降低了对抗扰动的隐蔽性, 使人更容易察觉。

4) FakeBob

Chen 等[12]针对声纹识别系统提出了一种名为 FakeBob 的黑盒攻击方法。FakeBob 与 PGD 均以迭代方式生成对抗样本, 与 PGD 不同的是它作为一种黑盒攻击方法, 通过自然进化策略估计梯度, 并且攻击针对的是原始输入语音而不是添加了随机扰动的语音。FakeBob 采用早停策略来减少查询次数, 即一旦找到对抗样本就停止计算。与 CW 攻击类似, FakeBob 也可以通过参数 κ 控制对抗扰动的强度。

5) SirenAttack

Du 等[13]提出了一种名为 SirenAttack 的黑盒音频对抗样本攻击方法。他们利用粒子群优化(Particle Swarm Optimization, PSO)算法求解对抗扰动。PSO 算法不需要梯度信息, 通过迭代地使候选解(粒子)群体根据适应度在搜索空间中移动来求得全局最优解。当算法在设定的最大迭代次数内攻击成功, 即可获得满足要求的音频对抗样本。

上述攻击方法将作为本文的对抗样本生成手段来验证所提出的防御方法的有效性。

2.1.2 声纹对抗样本防御方法

对于声纹对抗样本的防御, Li 等[14]提出了对抗样本检测的方法, 有效避免了对抗样本被声纹识别系统验证通过, 但这种方法不能纠正由对抗样本造成的错误识别结果, 从而不得不丢弃这些被对抗扰动污染的语音样本。基于对抗训练[15]的防御方法虽然可以在一定程度上减轻对抗样本带来的负面影响, 但却严重依赖特定的模型以及特定的对抗样本生成方法, 迁移性较差。

最近, 一些基于样本变换的预处理方法被用于防御对抗样本的攻击, 在一定程度上纠正了对抗样本造成的错误识别结果, 但也会降低真实样本的识别准确率。这些基于样本变换处理的防御方法包括:

1) 时频变换

在时域和频域对语音进行变换, 变换方法包括量化(Quantization)[16]、音频湍流(Audio Turbulence, AT)[17]、均值平滑(Average Smoothing, AS)[13]、中值平滑(Median Smoothing, MS)[16]和低通滤波(Low Pass Filter, LPF)[18]。

量化是将每个语音采样点的幅值四舍五入到最接近量化因子的整数倍。音频湍流假设对抗性扰动对噪声敏感, 通过向输入语音添加特定信噪比的噪声以改变对抗样本的识别结果。均值平滑通过对输入语音波形进行平滑来减弱对抗样本带来的影响, 将每个样本点 x_k 替换为其 k 个相邻样本的平均值。中值平滑与均值平滑相似, 只是它用 x_k 的 k 个相邻样本点的中值进行替换。低通滤波[19-20]的方法认为人类语音处于较低的频率范围内, 应用低通滤波器可以在保留语音内容的同时, 去除许多高频的对抗扰动。

2) MP3 压缩

基于心理声学原理, 语音 MP3 压缩[21]旨在抑制语音中的冗余信息, 以提高存储或传输效率。当难以察觉的对抗性扰动是冗余信息时, 可以通过语音压缩来消除。

3) 特征压缩

特征压缩是一种在特征级别破坏对抗扰动的压缩方法[22]。对于具有 N 帧的特征矩阵 $\boldsymbol{M}$, 每帧由 d 个特征组成。将矩阵 $\boldsymbol{M}$ 视为 d 维空间中的 N 个数据点, 并在给定参数 $K < N$ 的情况下将 N 个数据点划分为 K 个簇, 同一个簇中的数据点由一个代表向量表示。将 K 个代表向量组合起来形成新的特征矩阵 $\boldsymbol{M_0}$。

上述基于样本变换的防御方法将作为基线系统与本文提出的方法进行对比。

2.2 语音增强模型与对抗样本防御

语音增强的任务之一是提高受噪声影响语音的质量[23]。基于深度神经网络的模型在非平稳噪声影响下的单通道语音增强任务中已经取得了比传统滤波方法更好的效果。例如, Wave-U-Net 模型是 Stoller 等由用于图像分割的 U-Net 模型[24]改进而来的, 在语音增强和语音分离任务中取得了良好的效果[25]。在对抗样本防御方面, Yang 等[26]提出了改进的 U-Net 模型, 用于防御针对语音内容识别(Speech Recognition)的对抗样本攻击, 在降低语音文本识别词错误率和语音感知质量的改善上都取得了不错的效果, 提高了语音识别系统对对抗扰动的鲁棒性。本文针对声纹对抗样本, 研究改进基于 Wave-U-Net 的深度学习语音增强模型, 提高声纹识别系统防御对抗样本攻击的能力。

相对于 2.1 和 2.2 的相关工作, 本文的贡献如下所述:

1) 提出了基于噪声破坏和波形重建的声纹对抗样本防御方法。

首先, 通过在语音样本中添加高斯白噪声破坏对抗扰动的结构; 然后, 用含噪语音数据集对语音增强模型进行训练; 最后, 将对抗样本输入训练所得的语音增强模型, 重建出的波形即为去除了对抗扰动的语音样本。实验发现, 相比 2.1.2 的几种基于样本变换处理的方法, 本文提出的方法可以显著提高声纹识别系统在对抗样本上的识别准确率, 且对正常样本识别的负面影响较小。

2) 设计了 SCAT-Wave-U-Net 语音增强模型。

通过引入 Transformer 全局多头自注意力(Self-Attention)[27]和层间交叉注意力(Cross-Attention)机制, 增强下采样层特征之间全局交互的能力, 同时减轻跳跃连接中来自下采样层不相关特征信息的影响。将 Self-Attention 和 Cross-Attention 注意力机制与 Wave-U-Net 相结合, 构建出本文的增强方法 SCAT-Wave-U-Net。实验发现, 相比包括原始 Wave-U-Net 模型在内的其他语音增强算法, 本文提

出的 SCAT-Wave-U-Net 模型可以进一步改善增强语音的质量, 提高了模型从含噪语音样本中重建原始波形的能力。

3 基于噪声破坏和波形重建的声纹对抗样本防御

3.1 “破坏+重建”的对抗样本防御方法

对抗样本攻击的目的是在尽可能不影响人耳听觉感知质量的同时, 使声纹识别系统出错。因此, 制作对抗样本时通常只在原始语音样本上添加微小幅度的扰动, 以保证人耳无法感知。语音增强的目的是最大程度地消除附加在干净语音上的背景噪声, 使语音听起来更清晰。然而, 对抗样本自身扰动幅度较小, 直接用训练好的语音增强模型对其进行处理并不能有效地缓解样本的对抗性, 防御效果并不理想。Yang 等[26]在白盒条件下的研究证实了这一点。

实际上, 语音增强通常处理的含噪语音听起来更嘈杂并严重影响到人耳的听觉感知, 涉及的噪声通常比声纹对抗样本中的噪声具有更大的幅度和更强的随机性。相对于一些环境背景噪声, 对抗样本中的扰动则是经过精心构造的。为了产生具有对抗性的效果, 往往会经过大量的迭代训练, 以得到结构相对固定的对抗扰动[28], 从而使得对抗样本的识别结果比真实样本的识别结果对环境噪声更加敏感。当在样本中添加相同幅度的随机噪声时, 对抗样本的识别结果更容易被改变[18, 29]。

在上述研究的基础上, 本文首先在对抗样本中添加比对抗扰动幅度更大的高斯白噪声, 从而改变对抗扰动的原有结构, 破坏其对抗性; 然后, 利用语音增强模型处理添加了噪声的对抗样本, 重构出与真实样本近似的语音波形, 提高声纹识别的准确率, 实现对对抗样本攻击的防御。如图 1 所示, 提出的防御方法分为两个阶段: 第一阶段, 在输入的语音样本中添加不同信噪比的高斯白噪声。添加噪声的过程如算法 1 所示。

算法 1 在音频样本中植入高斯噪声

输入: 音频样本 X, 信噪比最小值 SNR_{min}, 信噪比最大值 SNR_{max}

输出: 带有高斯噪声的音频样本 X_{noise}

1) 从均匀分布 U(SNR_{min}, SNR_{max})中随机选择一个数值 SNR 作为当前样本 X 添加噪声的信噪比。

2) 计算音频样本 X 的均方根 RMS_X。

3) 根据信噪比 SNR 计算需要添加的噪声的均方根 RMS_{noise}。

4) 生成与输入音频 X 具有相同维度且满足 N(0, RMS_{noise})高斯分布的噪声 $Noise$。

5) 得到添加噪声的样本 $X_{noise} = X + Noise$。

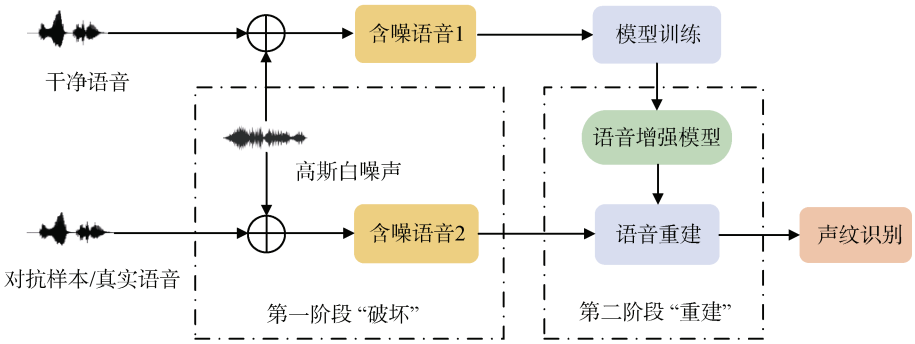

图 1 基于噪声破坏和波形重建的两阶段声纹对抗样本防御方法

Figure 1 A two-stage defense method against speaker adversarial examples based on noise destruction and waveform reconstruction

相比其他噪声, 高斯白噪声是一种比较常见且比较容易仿真实现的随机噪声, 本文将这种噪声添加在输入样本中用于覆盖样本中的对抗扰动, 进而破坏对抗扰动的固有结构。对于以不同方法生成的对抗样本, 由于其扰动幅度并不相同, 因此为实现最佳破坏效果而添加的适宜噪声幅度也可能并不一致。概括来说, 强度太小的噪声不能有效地破坏对抗扰动, 而强度太大的噪声会使原始音频难以修复。因此, 本文将添加噪声的信噪比范围限制在 0~25 dB (以 5 dB 为步长), 添加方法是以加性噪声的形式将噪声信号直接加在样本上, 这与对抗样本制作过程中在原始样本中添加对抗扰动的过程是一样的。

第二阶段, 在干净语音数据集上, 以同样的方式, 添加与第一阶段具有相同信噪比范围的噪声, 制作含噪语音数据集, 并在这个数据集上训练语音增强模型。然后, 用训练好的模型对第一阶段添加了

噪声的声纹样本进行处理，重构原始语音信号。这是在第一阶段破坏对抗扰动结构之后，进行语音波形的重建，以清除添加在样本中的噪声，最大程度地恢复原始语音。

由于第一阶段加入的噪声会破坏对抗扰动，第二阶段的重建有望在提高语音质量的同时，提高声纹识别的准确率。

3.2 SCAT-Wave-U-Net 模型结构

在波形重建阶段，为了进一步提高模型对声纹对抗样本的语音重建能力，本文在性能优异的 Wave-U-Net 模型基础上进行了改进，设计了一种名为 SCAT-Wave-U-Net 的模型结构，如图 2 所示。

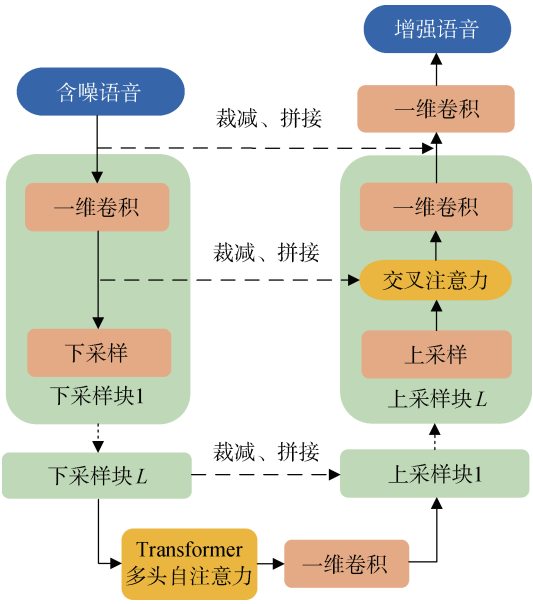

图 2 SCAT-Wave-U-Net 语音增强模型

Figure 2 Speech enhancement using SCAT-Wave-U-Net

Wave-U-Net 是由 U-Net 模型改进而来的，其中 U-Net 因其网络结构类似于字母“U”而得名。U-Net 包括下采样层的卷积加池化，上采样层的反卷积和相同上、下采样层的特征拼接等模块。为了适应对语音信号的处理，Wave-U-Net 将 U-Net 上采样层的反卷积操作变成了线性插值，同时在下采样层使用了居中裁减。该网络结构在音源分离和语音增强方面具有明显的优势。

语音信号具有明显的时间相关性。Yang 等[16]的研究表明，在语音样本中添加的对抗扰动会破坏掉这种时序信息，他们利用这一性质有效地区分了对抗样本和正常语音。在针对含噪对抗样本开展的语音重建任务中，借助原始语音中的时序依赖信息，可以更好地修复原始波形，从而恢复被对抗扰动破坏的时间相关性。然而，这种时序依赖性在 Wave-U-Net 模型中并没有得到充分体现。为使模型能更好地表示语音序列之间的相关关系，本文提出的 SCAT-Wave-U-Net 模型利用 Transformer 多头自注意力机制对最后一个下采样层的语音序列特征进行全局编码，使模型充分学习语音完整上下文信息之间的依赖关系。同时，在 Wave-U-Net 上、下采样层之间的跳跃连接中引入交叉注意力机制，使模型能更有效地利用来自下采样层的有价值的特征。

在 SCAT-Wave-U-Net 模型的网络结构中，最后一个下采样层末端的 Transformer 多头自注意力模块可访问包含整个音频序列的接收域，与原始 Wave-U-Net 的有限接收域形成对比。跳跃连接中的交叉注意力模块可以从对应上、下采样层的时序依赖关系中获取更有价值的特征信息用于构建拼接特征。对 Wave-U-Net 的这些改进可以更加有效地从含噪对抗样本中重建原始语音波形。

3.2.1 Transformer 多头自注意力

在最后一个下采样块之后，利用 Transformer 多头自注意力机制获取编码特征序列的全局依赖关系。本文的 Transformer 多头自注意力模块由 6 个顺序相连的相同子层组成，每个子层包括位置编码、具有残差连接的多头自注意力和前馈神经网络。每个子层的结构如图 3 所示。

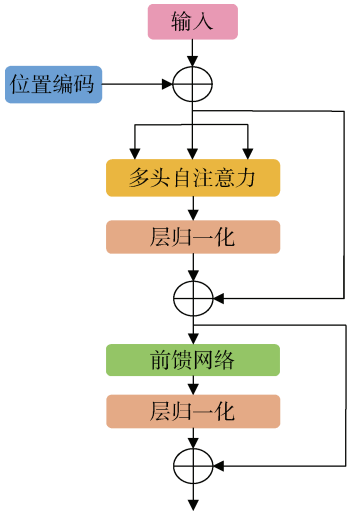

图 3 Transformer 多头自注意力

Figure 3 Multi-head self-attention in Transformer

在 Transformer 的自注意力机制中并没有输入特征序列的位置信息[30]，即序列中的矢量处于不同位置时对于自注意力的计算并没有区别，这在针对含噪语音的波形重建任务中显然是不合理的。因此，在输入特征中以文献[27]的方式添加位置编码[31]，以

获取含噪声纹样本编码特征序列中每个矢量在整个矢量序列中所处的相对位置关系。

自注意力作为模块最重要的部分, 旨在对具有不同维度、不同表示的序列特征进行加权融合, 从而实现对输入编码特征的全局访问。为了更好地利用具有不同维度、不同表示的子空间的信息, 本文使用了多头自注意力机制。本文将自注意力头个数设为 8, 每一个头的自注意力包括三个输入, 即查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$[27]。在针对含有“破坏”噪声的语音进行的波形重建任务中, $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 具有相同的大小, 对应于图 2 中最后一个下采样块 L 的输出。注意力计算公式如下:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d^{\text{model}}}})\boldsymbol{V}=\boldsymbol{A}\boldsymbol{V} \quad (3)$$

其中, $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{n\times d^{\text{model}}}$, n 是 $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ 输入特征的通道数, d^{model} 代表每个通道特征的维度, 矩阵 $\boldsymbol{A}$ 中的一行对应 $\boldsymbol{Q}$ 中特定通道某一维度特征相对于 $\boldsymbol{K}$ 中所有维度特征的相似性。

为了实现多头自注意力的并行计算, 在每个多头自注意力层中执行以下计算:

$$\text{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=[H_1,H_2,\dots,H_{d^{\text{head}}}]\boldsymbol{W}^{\text{head}} \quad (4)$$

$$H_h=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_h^q,\boldsymbol{K}\boldsymbol{W}_h^k,\boldsymbol{V}\boldsymbol{W}_h^v) \quad (5)$$

其中, $H_h\in\mathbf{R}^{n\times d^{\text{model}}/d^{\text{head}}}(h=1,\dots,d^{\text{head}})$ 是第 h 个注意力头的输出, d^{head} 是注意力头的个数, $\boldsymbol{W}_h^q,\boldsymbol{W}_h^k,\boldsymbol{W}_h^v\in\mathbf{R}^{d^{\text{model}}\times d^{\text{model}}/d^{\text{head}}}$ 和 $\boldsymbol{W}^{\text{head}}\in\mathbf{R}^{d^{\text{model}}\times d^{\text{model}}}$ 是可学习的权重矩阵。将多个自注意力头的计算结果拼接, 利用 $\boldsymbol{W}^{\text{head}}$ 进行一次线性变换得到的值作为多头自注意力的输出。

前馈神经网络具有 2048 维的单个隐藏层, 输入、输出层的神经元个数等于模块输入的编码特征的维度。网络的激活函数为 relu, 为了防止出现过拟合, 在训练过程中以 0.1 的概率应用 dropout。

在每个子层的多头自注意力和前馈网络之后均进行了层归一化处理。同时, 为了适应 Transformer 多头自注意力的计算, 将模块输入特征的形状从(批次, 特征, 通道)转换为(通道, 批次, 特征), 并在输出时重置特征的形状。

3.2.2 层间交叉注意力

与 Wave-U-Net 中直接将相同上、下采样层特征进行拼接不同, SCAT-Wave-U-Net 模型在跳跃连接中引入注意力门, 将下采样层的特征与注意力掩码相乘来识别其中的相关特征, 如图 4 所示。

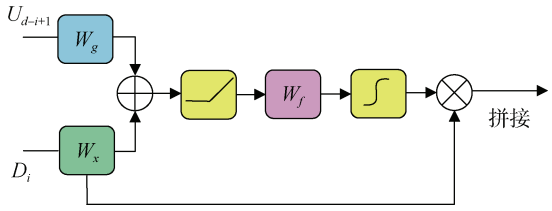

图 4 层间交叉注意力模块

Figure 4 Interlayer cross-attention module

为了实现注意力机制, 利用两个具有 u 个卷积核且核大小为 1 的一维卷积 W_g 和 W_x, 分别对上采样层特征 U_{d-i+1} 和下采样层特征 D_i 进行卷积操作并将结果相加。与 Giri 等人[6]的方法不同, 为了减少计算量并防止梯度消失, 本文在相加之后实施 relu 激活, 而不是 sigmoid 激活。激活后得到一个输出维度为 u 的中间特征 M_i,

$$M_i=\text{relu}(W_xD_i+W_gU_{d-i+1}+b_i) \quad (6)$$

其中, U_{d-i+1} 和 D_i 分别代表上下采样层的特征, 下标中的 i 为层序号, d 代表模型的总层数, 在本文中为 12, b_i 代表可学习的偏置。将中间特征输入到核大小为 1 的单个卷积 W_f, 输出进行批归一化处理, 再经过 sigmoid 激活得到注意力掩码 A_i,

$$A_i=\text{sigmoid}(W_fM_i) \quad (7)$$

最后, 将注意力掩码与下采样层特征 D_i 相乘后与上采样层特征 U_{d-i+1} 拼接。

4 实验评估

本文首先通过实验与其他语音增强算法进行对比, 验证提出的 SCAT-Wave-U-Net 模型对一般背景噪声的过滤效果。然后利用 SCAT-Wave-U-Net 模型, 针对典型的四种白盒、两种黑盒对抗样本攻击, 在两种不同声纹识别系统下进行实验, 验证提出的对抗样本防御方法的效果①。实验平台为 Ubuntu 20.04, 处理器为 Intel Xeon E5-2670 v3, 具有 62.8 GiB 内存、48 核 2.30 GHz 的 CPU 和一个 GeForce RTX 2080Ti GPU。

4.1 数据集

4.1.1 VCTK

本文使用 VCTK 数据集[32]验证 SCAT-Wave-U-Net 模型对一般背景噪声的语音增强效果, 同时在

① https://github.com/meisanhai/audios

该数据集的干净语音中添加高斯噪声对模型重新训练。数据集中的干净语音来自 30 位母语为英语的人, 其中 28 个说话人的语音用于训练, 剩余 2 个说话人的语音用于测试。将干净语音与各种噪声数据集混合生成含噪语音。训练集包含 40 种不同的噪声条件, 由 10 种类型噪声的各 4 种信噪比(0 dB、5 dB、10 dB 和 15 dB)组成, 每个说话人在每种条件下大约有 10 个不同的语句, 共有 11572 个训练样本。测试集与训练集在说话人、噪声类型和信噪比分布上都不同, 包含 20 种不同的噪声条件, 由 5 种类型噪声的各 4 种信噪比(2.5 dB、7.5 dB、12.5 dB 和 17.5 dB)组成, 每个测试说话人在每种噪声条件下大约有 20 个不同的语句, 共有 824 个测试样本。

4.1.2 Spk10

在声纹识别任务中选择 Chen 等人[22]公开的数据集, 包括 Spk10-enroll 注册集和 Spk10-test 测试集。说话人是从语音处理领域广泛采用的数据集 Librispeech 的“test-other”和“dev-other”子集中随机选择的。Spk10-enroll 包括 10 名说话人(5 名男性和 5 名女性), 每名说话人 10 个语句。Spk10-test 具有与 Spk10-enroll 相同的说话人, 但讲话内容不同, 每个说话人有 100 个语句。

4.2 实验设置

本文使用了两种典型的声纹识别系统, 分别是基于高斯混合模型(Gaussian Mixed Model, GMM)的 i-vector 系统[33]和基于时延神经网络(Time Delay Neural Network, TDNN)的 x-vector 系统[34], 这两种声纹识别系统均使用说话人嵌入(Embedding)将说话人的声学特性表示为固定维度的向量, 实验中使用基于语音识别平台 Kaldi[35]预训练的开源模型。本文进行的是闭集说话人鉴别, 即从一组注册的说话人中识别出测试语音来自哪个说话人, 在机器学习领域是一个多分类问题。两种系统均在 Spk10-enroll 上进行了注册, 将注册说话人发出的语音映射到注册嵌入特征, 作为注册说话人的唯一身份标识。测试过程中使用 Spk10-test 的数据进行测试。

在对抗攻击设置中, 对抗样本均是在 Spk10-test 上生成, 选择 Spk10-test 中 10 个说话人的各 20 条语音用来生成对抗样本。本文选择非目标攻击任务, 攻击的目标标签是从真实目标说话人之外的标签中随机选择的。FGSM 攻击步长设为 0.002。PGD 攻击的最大迭代次数设为 10, 步长设为 0.0004, 扰动幅度限制设为 0.002。CW_∞攻击的最大迭代次数设为 10, 步长设为 0.001, 扰动幅度限制设为 0.002。CW_2 攻击使用 9 步二进制搜索寻找对抗性扰动, 最大迭代次数设为 1 000, 参数 κ 设为 0。FakeBob 攻击的迭代次数设为 500, κ 设为 0, 扰动幅度限制设为 0.002。Siren 攻击粒子数为 50, PSO 最大迭代次数为 300。

对于基线对抗防御方法的设置, 在防御效果最佳的条件下, 将量化方法因子 q 的值设为 512, 音频湍流中的信噪比设为 15 dB, 中值平滑和均值平滑近似计算的样本点数设为 5, 低通滤波的截止频率设为 8 000 Hz, MP3 压缩的压缩级别设为 64 kbps, 特征压缩中 K 与 N 的比值设为 0.5, 特征聚类的方法使用 k 均值(k-means)聚类。

训练 SCAT-Wave-U-Net 模型时使用 Adam 优化器, 学习率为 0.0001, 批大小为 32。随机选取 1%的训练数据作为验证集, 如验证集上的效果在连续训练 20 个 epoch 时没有改进, 则停止训练。然后, 对训练参数进行微调, 批大小增加一倍, 学习率降至 0.00001, 同样在连续训练 20 个 epoch 验证集上的效果没有改进时, 停止训练。

4.3 评价指标

本文用声纹识别准确率衡量提出的方法对声纹对抗样本攻击的防御性能, 即计算能够准确识别目标说话人的语音数目与输入语音总数的比值。

使用语音质量的感知评价(Perceptual Evaluation of Speech Quality, PESQ)[36]、短时客观可懂度(Short-Term Objective Intelligibility, STOI)[37]和语音信噪比(Signal-to-Noise Ratio, SNR)三个主要的语音质量度量指标来评估从输入样本中重建原始音频的效果:

SNR: 信号和噪声的平均功率之比: S/N, 用分贝(dB)作为度量单位: $SNR = 10\lg(S/N)$。

PESQ: 语音质量感知评价分数是平均干扰 d_{sym} 和平均不对称干扰 d_{asym} 的总和, 在 0.5~4.5 之间, 输出信号和参照信号的差异性越大值越低。计算公式为: $PESQ = 4.5 - 0.1d_{sym} - 0.030\,9d_{asym}$。

STOI: 短时客观可懂度作为含噪语音非线性处理的稳健度量指标, 反映语音降噪后的清晰度, 范围在 0 到 1 之间, 值越大, 可懂度越高。

4.4 实验结果

4.4.1 SCAT-Wave-U-Net 模型的语音增强效果

为了评估 SCAT-Wave-U-Net 模型的有效性, 与经典滤波方法和包括原始 Wave-U-Net 模型在内的几种基于深度学习的语音增强方法进行了比较, 这些方法为: 维纳滤波(Wiener Filter)[38]、SEGAN[4]、Wave-U-Net[5]和 Attention Wave-U-Net[6]。本文使用与其他方法相同的 VCTK 数据集, 并引用他们公开的实验结果, 对比情况见表 1。

为了在不同方法之间进行公平比较, 本文使用

与其他语音增强算法相同的语音质量度量指标, 这些指标除PESQ外, 还包括与人类听觉感知相关的评分。CSIG: 关注语音信号失真的平均意见评分(Mean Opinion Score, MOS)预测。CBAK: 背景噪声侵入性的 MOS 预测。COVL: 整体处理后的语音质量 MOS 预测。此外, 还包括分段信噪比(Segment Signal-to-Noise Ratio, SSNR)。从表 1 中可以看出, 通过在 Wave-U-Net 模型下采样层后添加多头自注意力模块, 同时在上、下采样层间引入交叉注意力, 可进一步提高模型的去噪能力。SCAT-Wave-U-Net 模型在 PESQ、CSIG 和 COVL 三个指标上的结果均高于其他语音增强方法, 指标 CBAK 上的结果与文献[6]相同且高于其他三种方法, 在指标 SSNR 上的结果仅次于文献[5]和[6]。

表 1 不同语音增强方法增强后的语音质量对比

Table 1 Comparison of the quality of enhanced speech by different speech enhancement methods

指标	含噪语音	Wiener	SEGAN	Wave-U-Net	Attention Wave-U-Net	SCAT-Wave-U-Net
PESQ	1.97	2.22	2.16	2.40	2.62	**2.67**
CSIG	3.35	3.23	3.48	3.52	3.91	**3.94**
CBAK	2.44	2.68	2.94	3.24	3.35	**3.35**
COVL	2.63	2.67	2.80	2.96	3.27	**3.30**
SSNR	1.68	5.07	7.73	9.97	**10.05**	9.78
STOI	0.92	—	—	—	—	0.94
SNR	8.44	—	—	—	—	18.03

4.4.2 对不同对抗样本攻击的防御效果

本文首先在 VCTK 数据集的干净语音中添加具有不同信噪比范围限制的高斯白噪声, 用来训练语音增强模型。然后, 利用训练好的语音增强模型对添加了噪声的声纹样本进行重建。选取时域语音增强方法 Wave-U-Net[5]、Attention Wave-U-Net[6]和频域语音增强方法最小均方误差(Minimum Mean Square Error, MMSE)[39], 用来与本文提出的 SCAT-Wave-U-Net 模型在声纹对抗样本防御效果上进行对比。同时, 将本文方法防御对抗样本攻击的效果与基于样本变换的基线防御方法进行了对比, 实验结果见图 5。

从图 5 可以看出, 对于 i-vector 声纹识别系统, 时域量化方法使面对 FGSM 和 Siren 之外的其他攻击时的识别准确率得到提升, 但对真实样本的识别准确率却降低到了 35.5%。均值平滑、低通滤波和 MP3 压缩虽然能保证对真实样本的识别准确率在 99.5%以上, 但却不能防御 PGD 和 CW_∞对抗样本攻击。中值平滑的防御效果比均值平滑稍好, 但真实样本的识别准确率降低到了 88.5%。特征压缩可以在有效防御 FGSM、CW_2 和 FakeBob 攻击的同时, 保证真实样本的识别准确率在 98.5%, 但对 PGD、CW_∞和 Siren 攻击的防御效果不佳。音频湍流的防御效果相对更均衡, 在显著提高声纹识别准确率的同时保证真实样本的识别准确率在 89%。

本文方法在添加噪声信噪比为 10~15 dB 时的防御效果优于以上基线防御方法, 对 FGSM、PGD 和 CW_∞攻击的防御效果在添加噪声信噪比为 10~15 dB 时达到最佳, 对真实样本的识别准确率保持在约 94%。对于 CW_2、FakeBob 和 Siren 攻击, 在添加噪声信噪比为 20~25 dB 时, 效果最佳, 对真实样本的识别准确率依然保持在约 98.5%。与其他语音增强方法相比, SCAT-Wave-U-Net 模型对不同攻击的防御也在前述两个信噪比条件下取得了更好的效果。

对于 x-vector 声纹识别系统, 在几种基于样本变换的基线防御方法中, 相比在 i-vector 系统中的表现, 特征压缩的防御能力显著降低, 音频湍流获得了更好的防御效果, 其他几种防御方法与在 i-vector 系统中的表现类似。通过对比不同信噪比条件下的防御效果可以发现: 本文提出的方法在添加噪声信噪比为 10~15 dB 时, 对 FGSM、PGD、CW_2、CW_∞和 Siren 的攻击都具有比其他基线防御方法更好的效果, 在 FakeBob 攻击下的识别准确率仅比音频湍流防御低 1.5%, 同时对真实样本的识别准确率仍然保持在约 98.5%。在防御 CW_2、FakeBob 和 Siren 3 种对抗攻击时, 随着信噪比的增加, 防御的效果也逐渐提升。当信噪比为 20~25 dB 时, 在这 3 种攻击下的声纹识别准确率分别达到了 100%、99.5%和 99.5%。

以上实验结果表明, 本文提出的防御方法具有一定的通用性, 即对不同声纹识别系统和对抗样本攻击方式均有较好的防御效果, 且对正常样本的识别影响较小。值得一提的是, 即使是使用常见的 MMSE 方法进行波形重建, 在添加噪声信噪比为 10~15 dB 时的绝大多数防御中也获得了比基线防御方法更好的效果, 从而验证了“破坏+重建”防御框架的有效性。

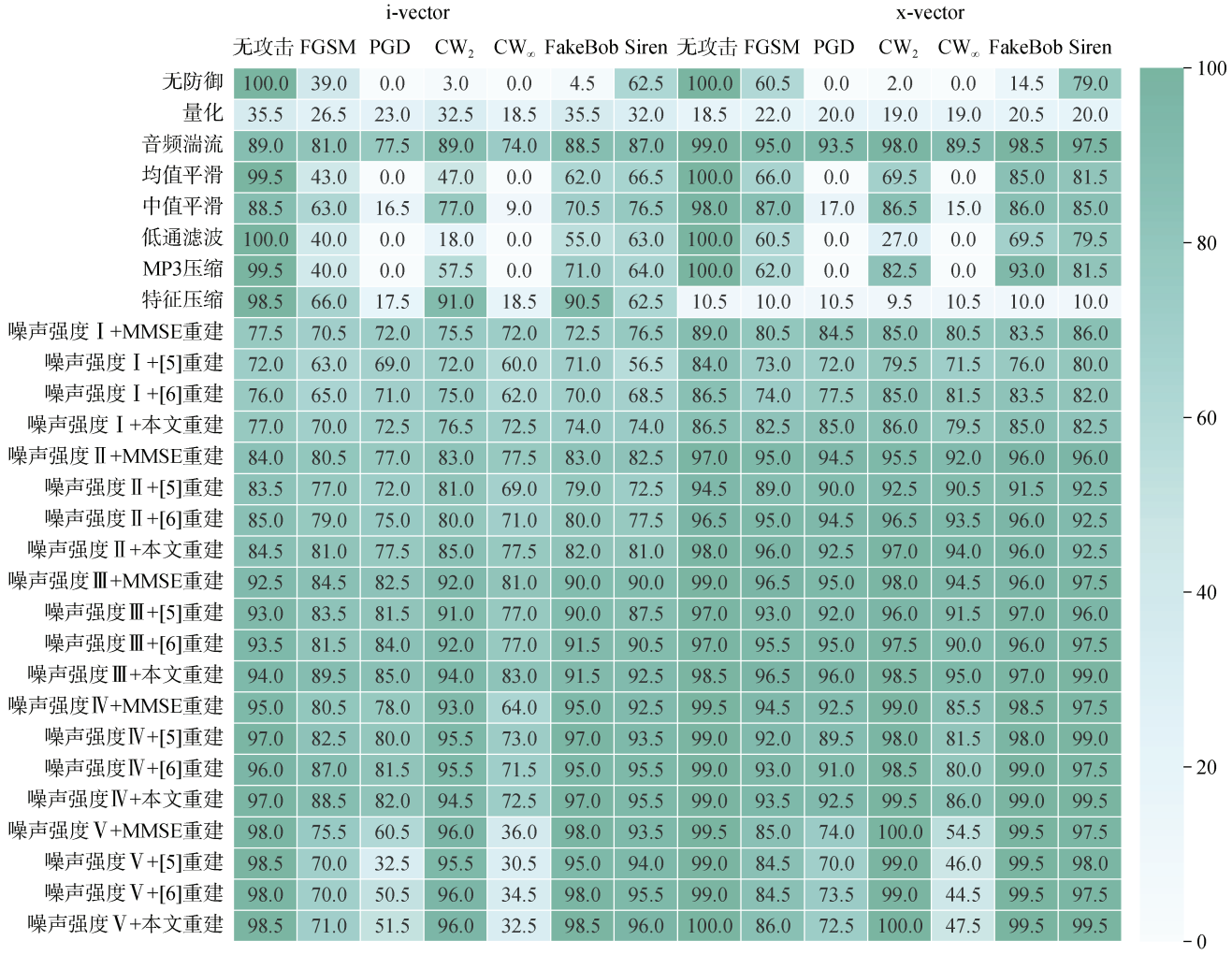

图 5 i-vector/x-vector 声纹识别系统的防御效果(噪声强度Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ分别表示含噪语音样本中信噪比范围为 0~5 dB、5~10 dB、10~15 dB、15~20 dB、20~25 dB)

Figure 5 Results on defending speaker recognition systems based on i-vector/x-vector (Noise intensities Ⅰ, Ⅱ, Ⅲ, Ⅳ and Ⅴ indicate that the SNR ranges of the speech examples with noises are 0~5 dB, 5~10 dB, 10~15 dB, 15~20 dB and 20~25 dB, respectively)

表 2 和表 3 是在两种声纹识别系统下, 当添加噪声信噪比为 10~15 dB 时以本文方法实施波形重建前后的语音质量, 表格中的数值是所有语音指标值的平均。

表 2 i-vector 声纹识别系统实施防御时的语音质量增强效果

Table 2 Performance on speech enhancement when defending speaker recognition system based on i-vector

	FGSM /增强后	PGD /增强后	CW_2 /增强后	CW_∞ /增强后	FakeBob /增强后	Siren /增强后
SNR	12.29 /18.58	12.44 /18.61	12.41 /18.52	12.31 /18.67	12.44 /18.61	12.49 /18.65
PESQ	1.21 /2.14	1.23 /2.16	1.22 /2.16	1.21 /2.17	1.22 /2.16	1.22 /2.16
STOI	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95

表 3 x-vector 声纹识别系统实施防御时的语音质量增强效果

Table 3 Performance on speech enhancement when defending speaker recognition system based on x-vector

	FGSM /增强后	PGD /增强后	CW_2 /增强后	CW_∞ /增强后	FakeBob /增强后	Siren /增强后
SNR	12.26 /18.55	12.45 /18.52	12.49 /18.37	12.22 /18.59	12.32 /18.69	12.42 /18.60
PESQ	1.21 /2.14	1.22 /2.17	1.23 /2.20	1.21 /2.16	1.21 /2.18	1.22 /2.16
STOI	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95	0.92 /0.95

从表中可以看出, 在添加噪声后, 不同对抗样本在 SNR、PESQ 和 STOI 三个指标上的数值相差不大, 这说明此时高斯白噪声已经淹没了音频样本中

的对抗性扰动。经过 SCAT-Wave-U-Net 模型的波形重建处理后, 语音质量得到了显著恢复。从实际的听觉感受来看, 经过重建的语音样本由于去除了额外的杂音, 相比原始语音更加清晰。

图 6 展示了语音样本在不同阶段的语谱图变化情况。可以观察到: 相比图 6(a)中的原始波形, 由于添加了对抗扰动, 图 6(b)中对抗样本的语谱图各语音帧高频部分具有更高的能量(如红色椭圆形圈出来的部分所示)。在图 6(c)中, 添加的随机噪声淹没并破坏掉了对抗样本中的对抗扰动部分。在图 6(d)中, 最后增强重建出来的语音消除了对抗噪声, 从而使得语谱图能量分布与原始语音谱更加接近。

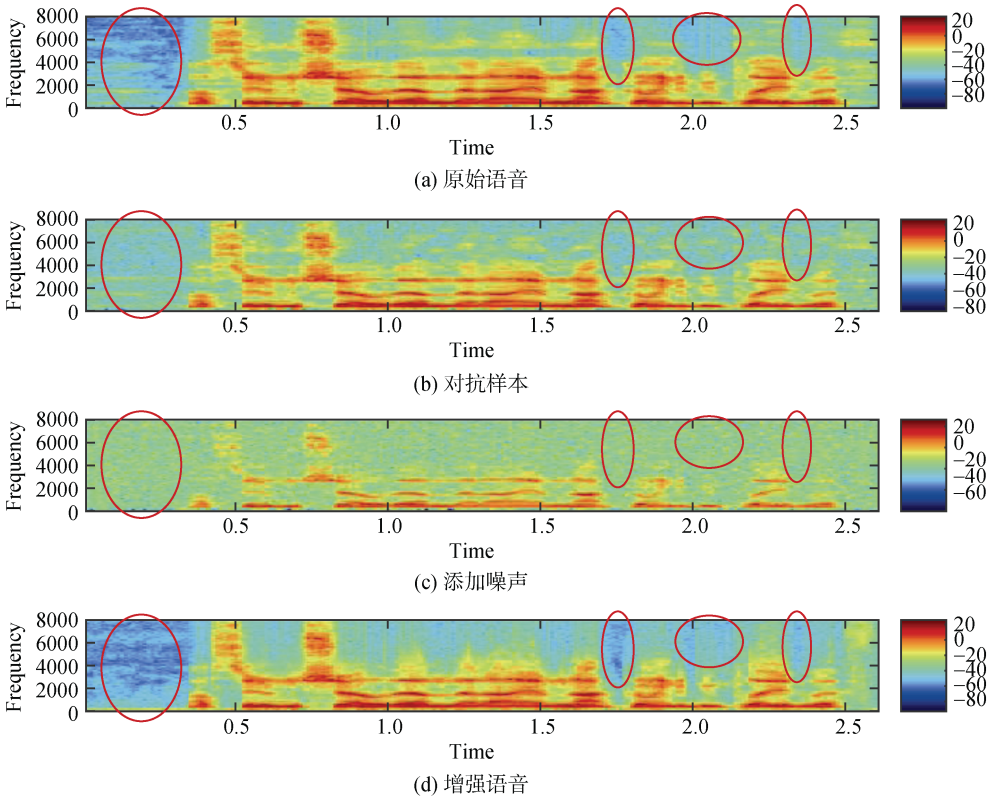

图 6 语音样本在不同防御阶段的语谱图 (a)原始语音 (b)对抗样本 (c)添加噪声 (d)增强语音

Figure 6 The spectrograms at different stages of defense (a) Original speech (b) Speech with adversarial perturbations (c) Speech with added noises (d) Enhanced speech

4.4.3 不同对抗扰动时长的防御效果分析

通常情况下, 在对抗样本生成过程中, 对抗扰动是以与原始音频样本相同的长度进行构造的。在本文提出的基于噪声破坏和波形重建的对抗样本防御方法中, 由于在输入语音样本中加入了随机噪声, 当对抗扰动的长度小于原始语音的长度时, 在音频完整时长范围内添加的随机噪声可能会对最终的防御效果带来负面影响。在这一小节中, 通过实验验证不同对抗扰动时长对防御性能的影响。针对前文所述的每一种对抗样本攻击方法, 将对抗扰动的时长分别设置为原始语音样本长度的 1/4、1/2 和 3/4, 并添加在原始语音完整时长范围内的某一处随机位置。防御时在样本中添加的噪声幅度限制为 10~15 dB。不同扰动时长下的攻防效果见图 7。

观察图 7(a)和(b)可以发现, 对抗扰动时长越短, 在相同攻击设置下的攻击效果越差, 当对抗扰动的时长与原始语音样本相同时攻击效果最佳。同时, 从图 7(c)和(d)中可以观察到, 在 i-vector 系统中, 随着对抗扰动时长增加, 对 CW_2 和 FakeBob 攻击的防御效果逐渐提高, 对 PGD 和 CW_∞攻击的防御效果逐渐降低, 对 FGSM 和 Siren 攻击的防御效果变化幅度较小。在 x-vector 系统中, 随着对抗扰动时长增加, 对 CW_2 和 FakeBob 攻击的防御效果逐渐提高, 对其他几种攻击的防御效果变化幅度较小。当在 i-vector 系统中防御具有原始音频 1/4 时长的 FGSM 和 Siren 攻击时, 声纹识别准确率相比没有防御时有小幅度的下降。相比在 i-vector 系统中的结果, 本文方法对 x-vector 系统的防御效果更好, 由不同对抗扰动时长引起的防御性能变化情况也比 i-vector 系统更加稳定。总体来看, 本文方法在面对具有不同扰动时长的对抗样本攻击时同样具有较稳定的防御效果。

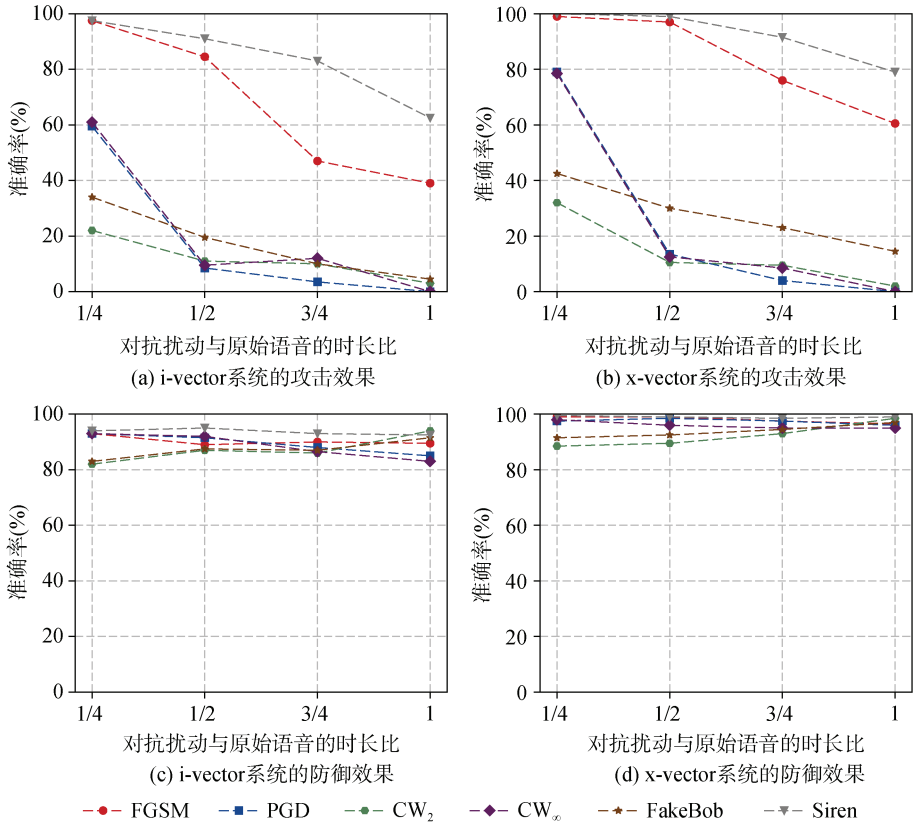

图 7　不同对抗扰动时长的攻防效果; (a)、(b)分别为对 i-vector 和 x-vector 系统的攻击效果, (c)、(d)分别为对 i-vector 和 x-vector 系统的防御效果

Figure 7　Adversarial perturbations with different durations; (a) and (b) are the attack results on i-vector and x-vector systems, respectively, (c) and (d) are the defense results on i-vector and x-vector systems, respectively

4.4.4　防御方法的实时性分析

语音识别、声纹识别均对实时性有较高的要求。因此, 在防御对抗样本攻击时, 实施防御所付出的时间成本非常重要。实时因子(Real Time Factor, RTF)定义为语音识别系统处理所有音频的耗时与输入音频总时长的比值。本文通过计算声纹样本处理过程中的实时因子, 对比实施防御前后的实时因子变化情况, 对本文方法的实时效果进行分析。实验结果见表 4 和表 5。

防御前后的音频总时长保持不变, 本文对抗样本的总时长为 1067.67 s。与无防御时的声纹识别相比, 在实施防御过程中新增了添加噪声和语音重建的时间成本。在添加噪声信噪比为 10~15 dB 条件下, 计算了用本文方法防御每一种攻击时, 添加噪声、语音重建和声纹识别的时间。实施防御前后的实时因子计算公式为:

$$\mathrm{RTF}_{\text{无防御}}=\frac{T_{\text{声纹识别}}}{T_{\text{全部音频}}} \tag{8}$$

$$\mathrm{RTF}_{\text{有防御}}=\frac{T_{\text{添加噪声}}+T_{\text{语音重建}}+T_{\text{声纹识别}}}{T_{\text{全部音频}}} \tag{9}$$

表 4　i-vector 声纹识别系统的实时因子

Table 4　The real-time factors of the speaker recognition system based on i-vector

	FGSM	PGD	CW_2	CW_∞	FakeBob	Siren
无防御	0.0243	0.0251	0.0252	0.0251	0.0258	0.0228
有防御	0.0520	0.0528	0.0530	0.0529	0.0536	0.0505

表 5　x-vector 声纹识别系统的实时因子

Table 5　The real-time factors of the speaker recognition system based on x-vector

	FGSM	PGD	CW_2	CW_∞	FakeBob	Siren
无防御	0.0064	0.0065	0.0065	0.0066	0.0068	0.0063
有防御	0.0342	0.0359	0.0343	0.0343	0.0346	0.0341

由于 i-vector 系统中包含了比较耗时的高斯混合模型, 而 x-vector 的前馈推理便于显卡加速处理, 因此 x-vector 系统声纹识别的时间少于 i-vector。从表 4、5 中可以看出, 不同对抗攻击方法之间实时因子值的变化不大。在实施防御后, i-vector 和 x-vector 系统中的实时因子大致分别变为原来的 2 倍和 5 倍。实时因子数值小于 1, 说明本文提出的防御方法在增加一部分时间成本后仍可以满足声纹识别的实时

性要求。

5 结束语

针对传统声纹对抗样本防御方法鲁棒性差、纠正错误输出的同时影响真实样本的识别等缺点, 提出了一种基于噪声破坏和波形重建的声纹对抗样本防御方法。通过在对抗样本中添加噪声破坏对抗扰动的结构, 使其失去对抗性; 然后利用语音增强模型重建语音波形。提出的 SCAT-Wave-U-Net 模型通过引入 Transformer 多头自注意力和层间交叉注意力机制增强了对含噪声纹对抗样本的波形重建能力, 同时相比原始 Wave-U-Net 模型也提高了在一般环境噪声条件下的语音增强能力。实验结果表明, 本文提出的方法在保证对真实声纹样本识别准确率影响较小的情况下, 对四种白盒、两种黑盒对抗样本攻击的防御效果优于其他基于样本变换的防御方法, 同时显著恢复了输入语音的音频质量。进一步地, 在本文的防御方法中, 添加“破坏”噪声的过程不可微, 很难将这一过程引入自适应攻击的梯度求解中。因此, 本文提出的方法对于防御自适应攻击也具有一定优势。

参考文献

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

[2] 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.

[3] Hu S S, Shang X C, Qin Z, et al. Adversarial Examples for Automatic Speech Recognition: Attacks and Countermeasures[J]. *IEEE Communications Magazine*, 2019, 57(10): 120-126.

[4] Pascual S, Bonafonte A, Serrà J. SEGAN: Speech Enhancement Generative Adversarial Network[C]. *Interspeech 2017*, 2017: 3642-3646.

[5] Macartney C, Weyde T. Improved speech enhancement with the wave-U-Net[EB/OL]. 2018: ArXiv Preprint ArXiv: 1811.11307.

[6] Giri R, Isik U, Krishnaswamy A. Attention Wave-U-Net for Speech Enhancement[C]. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019: 249-253.

[7] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]. *2015 International Conference on Learning Representations*, 2015: 1-11.

[8] Gong Y, Poellabauer C. Crafting Adversarial Examples for Speech Paralinguistics Applications[EB/OL]. 2017: arXiv: 1711.03280. https://arxiv.org/abs/1711.03280.pdf.

[9] Liu S X, Wu H B, Lee H Y, et al. Adversarial Attacks on Spoofing Countermeasures of Automatic Speaker Verification[C]. *2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2020: 312-319.

[10] 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.

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

[12] 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.

[13] Du T Y, 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.

[14] 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.

[15] 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.

[16] Yang Z L, Li B, Chen P Y, et al. Characterizing Audio Adversarial Examples Using Temporal Dependency[EB/OL]. 2018: arXiv: 1809.10875. https://arxiv.org/abs/1809.10875.pdf.

[17] Yuan X J, Chen Y X, Zhao Y, et al. Commandersong: A Systematic Approach for Practical Adversarial Voice Recognition[C]. *The 27th USENIX Conference on Security Symposium*, 2018: 49-64.

[18] 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.

[19] Hossen I, Hei X L. AaeCAPTCHA: The Design and Implementation of Audio Adversarial CAPTCHA[C]. *2022 IEEE 7th European Symposium on Security and Privacy*, 2022: 430-447.

[20] Abdullah H, Garcia W, Peeters C, et al. Practical Hidden Voice Attacks Against Speech and Speaker Recognition Systems[C]. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019: 1-15.

[21] Andronic I, Kürzinger L, Chavez Rosas E R, et al. MP3 Compression to Diminish Adversarial Noise in End-to-End Speech Recognition[C]. *Karpov A, Potapova R. International Conference on Speech and Computer*, 2020: 22-34.

[22] Chen G K, Zhao Z, Song F, et al. SEC4SR: A Security Analysis Platform for Speaker Recognition[EB/OL]. 2021: arXiv: 2109.01766. https://arxiv.org/abs/2109.01766.pdf.

[23] Xu Y, Du J, Dai L R, et al. An Experimental Study on Speech Enhancement Based on Deep Neural Networks[J]. *IEEE Signal Processing Letters*, 2014, 21(1): 65-68.

[24] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015: 234-241.

[25] Stoller D, Ewert S, Dixon S. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation[EB/OL]. 2018:

arXiv: 1806.03185. https://arxiv.org/abs/1806.03185.pdf.

[26] Yang C H, Qi J, Chen P Y, et al. Characterizing Speech Adversarial Examples Using Self-Attention U-Net Enhancement[C]. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2020: 3107-3111.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You Need[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 6000-6010.

[28] Zhang Y A, Xu H, Pei C F, et al. Adversarial Example Defense Based on Image Reconstruction[J]. *PeerJ Computer Science*, 2021, 7: e811.

[29] 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*, 2019: 197-201.

[30] Ahmed S, Nielsen I E, Tripathi A, et al. Transformers in Time-Series Analysis: A Tutorial[EB/OL]. 2022: arXiv: 2205.01138. https://arxiv.org/abs/2205.01138.pdf.

[31] Subakan C, Ravanelli M, Cornell S, et al. Attention is all You Need in Speech Separation[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 21-25.

[32] Valentini C. Noisy speech database for training speech enhancement algorithms and TTS models[J]. *University of Edinburgh, School of Informatics, Centre for Speech Technology Research*, 2016.

[33] Dehak N, Dehak R, Kenny P, et al. Support Vector Machines Versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification[C]. *Interspeech 2009*, 2009: 1559-1562.

[34] 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.

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

[36] 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* (*Cat. No.01CH37221*), 2002: 749-752.

[37] Taal C H, Hendriks R C, Heusdens R, et al. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech[C]. *2010 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2010: 4214-4217.

[38] Scalart P, Filho J V. Speech Enhancement Based on a Priori Signal to Noise Estimation[C]. *1996 IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing Conference Proceedings*, 2002: 629-632.

[39] Ephraim Y, Malah D. Speech Enhancement Using a Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator[J]. *IEEE Transactions on Acoustics*, *Speech*, *and Signal Processing*, 1984, 32(6): 1109-1121.

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: sunmeng@aeu.edu.cn

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。Email: xwzhang9898@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。Email: zlc1997@163.com

印杰 现为江苏警官学院高级工程师。研究领域为机器学习、大数据、网络安全。Email: yinjie@jspi.cn