DOI:10.19651/j.cnki.emt.2416738

针对说话人识别对抗样本生成方法研究*

马栋林 宋佳佳 赵 宏 陈伟杰
(兰州理工大学计算机与通信学院 兰州 730050)

摘 要:针对基于生成式的对抗样本生成方法生成的对抗样本真实性较低和攻击效果欠佳的问题,提出一种基于 AdvGAN 和 CGAN 的对抗样本生成方法 ACGAN。首先,针对特定目标进行攻击,ACGAN 通过在训练和攻击阶段引入额外的目标标签,生成具有针对性的频域上的对抗样本。其次,在生成器和鉴别器中引入门控卷积神经网络,帮助 ACGAN 模型捕捉到更精确的数据特征,从而提高攻击成功率。最后,引入感知损失函数,最小化模型输出与目标输出在语音特征表示上的差异,提高生成样本的听觉质量。实验结果表明,在有目标攻击中相较于现有方法,ASR 提高了 1.5%,SNR 和 PESQ 分别提高了 10.5%和 11.1%,证明了 ACGAN 在对抗样本生成领域的有效性和潜力。

关键词:对抗样本;生成器;鉴别器;门控卷积神经网络;感知损失
中图分类号:TN391 **文献标识码**:A **国家标准学科分类代码**:510.4040

Research on adversarial sample generation methods for speaker recognition

Ma Donglin Song Jiajia Zhao Hong Chen Weijie
(School of Computer Science and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Aiming at the problem that adversarial samples generated by generative adversarial sample generation methods have low authenticity and poor attack effect, an adversarial sample generation method ACGAN based on AdvGAN and CGAN is proposed. First, attacking a specific target, ACGAN generates targeted adversarial samples in the frequency domain by introducing additional target labels in the training and attack stages. Secondly, the gated convolutional neural network is introduced in the generator and discriminator to help the ACGAN model capture more accurate data features, thereby improving the success rate of the attack. Finally, the perceptual loss function is introduced to minimize the difference in speech feature representation between the model output and the target output, thereby improving the auditory quality of the generated samples. Experimental results show that compared with the existing methods in targeted attacks, the ASR is improved by 1.5%, and the SNR and PESQ are improved by 10.5% and 11.1% respectively, which proves the effectiveness and potential of ACGAN in the field of adversarial sample generation.

Keywords: adversarial examples; generator; discriminator; gated convolutional neural network; perceptual loss

0 引 言

基于深度学习的说话人识别系统(speaker recognition systems, SRSs)[1]近年来取得了巨大成功,并广泛应用于移动支付和智能家居等领域,凸显了其在现代生活中的关键作用[2]。然而,随着对抗样本的出现,说话人识别系统的安全性面临着新的挑战。对抗样本(adversarial example, AE)的研究始于图像领域,随后扩展到语音识别范畴。Chen 等[3]发现,即便是在人耳难以辨别的微小扰动下,SRSs 也可能产生误判,这一发现对诸如金融交易、在线支付等依赖声音认证的系统构成了潜在威胁。攻击者通过模拟合法用户的声音特征,利用对抗样本绕过安全机制,实施欺诈行为,其后果不堪设想。因此,研究对抗攻击方法对于评估 SRSs 的鲁棒性和安全性至关重要[4]。

对抗攻击方法主要分为基于梯度的攻击、基于优化的攻击和基于生成式的攻击。Kreuk 等[5]利用基于梯度的快速梯度符号法(fast gradient sign method, FGSM)对梅尔频率倒谱系数(mel-frequency cepstral coefficients, MFCC)特征进行扰动,并重构成声波形式,在端到端说话人验证系统中实现了 90%的攻击成功率(attack success rate,

收稿日期:2024-08-26
* 基金项目:国家自然科学基金(62166025)项目资助

ASR),首次证明了 SRSs 的脆弱性。Chen 等[6]将(Carlini Wagner, CW)方法应用于说话人识别,达到了 100% 的 ASR,且支持指定目标攻击。Yao 等[7]提出的基于对称显著性的方法(symmetric saliency-based encoder-decoder, SSED),通过显著性映射和角度损失函数提高 ASR 和信噪比(signal-noise ratio, SNR)。Li 等[8]介绍的通用对抗扰动生成网络(universal adversarial perturbations, UAPs),通过低维正态分布到 UAPs 子空间的映射,成功欺骗了训练良好的 SRSs。

尽管上述方法取得了一定的成功,但它们也存在一些缺点。基于梯度的攻击方法 FGSM 根据目标模型损失函数的梯度方向添加扰动来生成 AE,生成的速度较快,但未考虑扰动幅度,且需要针对特定模型重新训练,难以在实际音频数据上实现高效攻击;基于优化的攻击方法 CW,生成 AE 的语音质量效果好,对目标模型的 ASR 高,但找到最优的攻击扰动所需时间成本较高,每次只能优化一个样本,且生成 AE 的通用性较差,不具备一定的迁移能力;基于生成式的攻击方法 SSED 和 UAPs,生成扰动所需时间成本低,但 SSED 方法利用编码器和解码器难以提取到语音的高维数据特征,语音中添加的扰动幅度较大,对超参数敏感,且未优化语音质量感知评估(perceptual evaluation of speech quality, PESQ),生成的对抗样本真实性较低。UAPs 方法利用生成网络生成对抗扰动,对扰动进行缩放来控制语音的失真,具有一定的通用性,但其难以对特定目标生成扰动,导致在有目标攻击中 ASR 较低。

这些缺点限制了对抗攻击在实际应用中的使用,需要进一步研究和改进。为此,本文提出了一种结合 AdvGAN (generating adversarial examples with adversarial networks)和 CGAN(conditional generative adversarial network)的高效 AE 生成方法 ACGAN,旨在生成特定目标的 AE,提高 ASR 和听觉质量。ACGAN 在训练和攻击阶段引入额外的标签信息,针对指定目标进行攻击,生成频域上的 AE。此外,在生成器和鉴别器中引入门控卷积神经网络(gated convolutional neural network, GCNN),以捕捉更精确的数据特征从而提高 ASR。通过最小化模型输出与目标输出在语音特征表示上的差异,引入感知损失函数,并与 AdvGAN 的其他损失函数联合优化,提升生成样本的听觉质量。实验结果表明,ACGAN 在 TIMIT 和 LibriSpeech 数据集上相比其他攻击方法,获得了更高的 ASR 和语音质量。

1　相关工作

1.1　对抗样本

AE[9]指在真实语音数据中添加微小且人耳难以察觉的扰动,导致目标模型以高置信度给出错误输出,AE 的定义如式(1)所示。

$$x' = x + \delta, s.t. \ \| \delta \|_p < \varepsilon \tag{1}$$

式中:x 为真实语音,δ 为对抗扰动,x' 为 AE。

根据攻击者对 SRSs 先验知识的掌握程度,攻击场景分为白盒攻击(white-box attack, WBA)[10]、灰盒攻击(gray-box attack, GBA)[11]和黑盒攻击(black-box attack, BBA)[12],攻击难度依此递增。本文选择 WBA,攻击者完全掌握目标模型的内部结构和参数,能充分利用梯度信息直接计算出误导模型的 AE,从而提高攻击效率和成功率。WBA 能最大限度地测试目标模型的鲁棒性,揭示其脆弱环节,为开发更健壮的 SRSs 提供参考。根据是否给定指定攻击标签,攻击类型分为有目标攻击(targeted attack, TA)[13]和无目标攻击(no targeted attack, NTA)[14]。前者以高置信度欺骗目标模型生成特定标签,挑战 SRSs 的整体分类能力和安全性,模拟实际应用中的恶意攻击。后者使目标模型输出任何非真实标签的错误结果,旨在最大限度扰乱目标模型的判断能力,评估模型的整体防御能力。扰动对象分为时域扰动(time-domain perturbation, TDP)[15]和频域扰动(frequency-domain perturbation, FDP)[16],将原始音频信号的时域采样值作为扰动对象称为 TDP。本文选用 FDP,通过操控语音信号的 MFCC 特征,更精细地控制 AE 的生成,避免对人耳可听范围内的语音质量造成影响,从而生成高效且具迷惑性的 AE。

1.2　AdvGAN

Xiao 等[17]图像领域提出了一种基于 GAN 的 AE 生成方法 AdvGAN,旨在生成更真实的 AE,以误导目标模型。AdvGAN 由生成器 G、鉴别器 D 和目标模型 f 组成,其中生成器 G 产生 AE,鉴别器 D 区分真实样本与 AE。对抗损失计算如式(2)所示。

$$L_{GAN} = \min_G \max_D E_x \log D(x) + E_x \log(1 - D(x + G(x))) \tag{2}$$

式中:D(x)表示 x 是真实图像的概率,x+G(x)表示 AE。在 TA 中,欺骗目标模型 f 的损失函数计算如式(3)所示。

$$L_{adv} = E_x l_f(x + G(x), t) \tag{3}$$

式中:t 表示攻击目标标签,l_f 表示训练目标模型的损失函数。

1.3　CGAN

CGAN 是 GAN 的变体,其中生成器 G 从条件 c 和随机噪声 z 中产生 G(c,z),鉴别器 D 将(x,c)分类为真,将(G(z,c),c)分类为假,其目标函数计算如式(4)所示。

$$\min_G \max_D V(G, D) = E_x[\log D(x, c)] + E_z[\log(1 - D(G(z, c), c))] \tag{4}$$

为了生成更具针对性和攻击性的 AE 并提高有目标攻击的 ASR,本文结合 AdvGAN 和 CGAN,在 AdvGAN 的训练和攻击阶段引入指定标签,将语音数据和标签信息输入进生成器,生成任意指定标签的 AE。

2　模型构建

2.1　ACGAN 模型设计

本文结合 AdvGAN 与 CGAN 的优点,提出一种 AE

生成方法 ACGAN。AdvGAN 生成真实且有效误导目标模型的 AE,CGAN 在特定条件下生成高质量样本。结合这两者的设计,ACGAN 在保持 AE 真实性的同时,通过条件约束生成更符合攻击目标的 AE。ACGAN 网络结构如图 1 所示,包括生成器 G、鉴别器 D 以及目标模型 f。首先,真实语音通过预加重、分帧、加窗等步骤提取声学特征 x。其次,将这些声学特征 x 与额外的攻击目标标签 y 一同输入到生成器 G 中,以产生对抗扰动 $G(x,y)$,该扰动被添加至原始声学特征中,生成 $x+G(x,y)$。AE 随后输入到说话人识别模型中计算对抗损失 L_{adv},同时,将 AE 和真实语音的声学特征输入到鉴别器中计算 L_{GAN} 损失。此外,在生成器和鉴别器中引入 GCNN 模块,利用其门控机制更精确地捕获和强调语音特征,从而帮助生成器生成在听觉上更为真实的 AE;GCNN 还可以帮助鉴别器识别更细微的语音特征,确保在区分真实语音与生成样本时更加准确,从而提高 ASR。最后,在 ACGAN 中引入感知损失 $L_{perceptual}$,通过最小化目标模型对真实语音特征和 AE 特征之间的差异,确保生成的 AE 在听觉上与真实语音相似。$L_{perceptual}$ 与 L_{adv}、L_{GAN} 和扰动约束损失函数 L_{pert} 进行联合优化,提高 AE 的听觉质量。

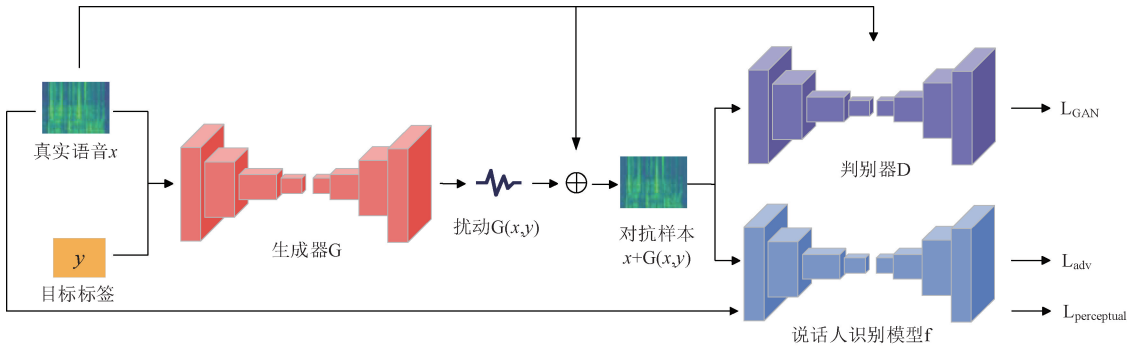

图 1 ACGAN 网络结构图

Fig. 1 ACGAN network structure

2.2 门控卷积神经网络

AdvGAN 的生成器存在特征表达不足的问题,难以模拟说话人的特有声学特征如语调、音色和语速等,影响 ASR。鉴别器在区分真实语音和生成语音时会过度依赖于训练数据的特定特征,为此,ACGAN 引入 GCNN,在生成器中,GCNN 的门控机制细致地控制声音特征的捕获和强调,生成在听觉上更为真实的 AE,特别是在音调、音色和语速等关键属性上。GCNN 自适应调节这些特征的表示,提高了 AE 欺骗 SRSs 的概率。对于鉴别器,GCNN 有助于更细致地理解和区分真实语音与生成语音的差异,提高其在各种声学环境下的判别能力,通过这种方式,鉴别器能够引导生成器生成更难以检测的 AE,整体提升对目标模型的 ASR。

GCNN 模块包含卷积和门控线性单元(gated linear unit, GLU)。如图 2 所示,GLU 是一个激活函数,隐藏层输出 $h_l(x)$,具体计算如式(5)所示。

$$h_l(x)=(x*\boldsymbol{W}+b)\times\sigma(x*\boldsymbol{V}+c) \tag{5}$$

式中:x 为第 h_l 层的输入,$\boldsymbol{W}$、$\boldsymbol{b}$、$\boldsymbol{V}$ 和 $\boldsymbol{c}$ 是卷积层的参数,σ 是 sigmoid 函数,$*$ 是卷积运算,$\times$ 是对应参数相乘。使用该门控机制,可以根据前一层的状态来控制下一层中传递的信息,同时 GLU 具有与 LSTM 类似的长距离依赖关系。

2.3 生成器模块

生成器负责生成难以被说话人识别模型辨认的 AE。在 ACGAN 模型的生成器中,将真实语音的声学特征和目标攻击标签作为输入,以生成特定目标的 AE,同时将

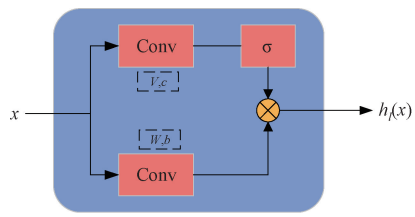

图 2 GCNN 结构

Fig. 2 GCNN structure

GCNN 加入到生成器中,更加细致地控制对声音特征的捕获和强调,从而生成在听觉上更为真实的 AE。生成器网络模块如图 3 所示,包括上采样层、残差块和下采样层,h、w 和 c 分别代表高度、宽度和通道数。在每个卷积层中 k、c 和 s 分别表示卷积核大小、通道数和步长。模型将真实语音特征和目标标签作为输入进行下采样,其中将目标标签变换到同等尺寸大小的目标特征,经过第一个卷积层,用于在时域和频域中发现更多的特征,随后通过 GCNN 提取特征,在提取到充分的语音特征和目标信息后经解码器进行逐步上采样,恢复原始语音大小,得到扰动信息,然后将扰动与真实语音相加得到 AE。

2.4 鉴别器模块

鉴别器的主要作用是区分输入样本是否由生成器产生,从而实现生成对抗网络的“博弈”机制。鉴别器网络模块如图 4 所示,卷积层每层均搭配特定的卷积核大小和步长,然后经过 GCNN,帮助模型集中注意力于对区分真实与生成样本最关键的特征。此外,最终的全连接层将高维

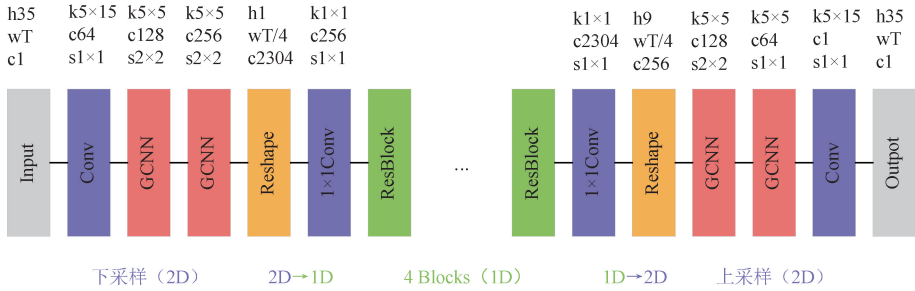

图 3　生成器网络模块

Fig. 3　Generator network module

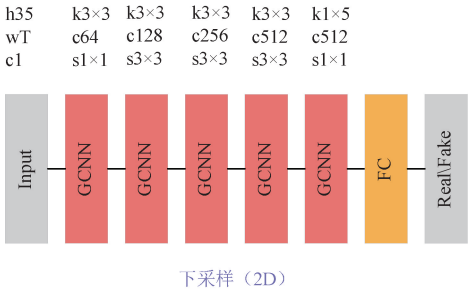

图 4　鉴别器网络模块

Fig. 4　Discriminator network module

特征向量映射到一个标量输出，表示样本为真实语音的概率。

2.5　损失函数设计

在训练过程中，生成器以真实语音的特征序列 x 和攻击目标标签 y 作为输入，生成对抗扰动 G(x,y)，然后将其与真实语音相加生成 x+G(x,y)。鉴别器 D 和生成器 G 通过联合训练，并使用 L_{GAN} 损失函数来优化模型，以提高模型的稳定性，L_{GAN} 损失函数计算如式(6)所示。

$$L_{GAN}=\min_{G}\max_{D}E_x[\log D(x)]+E_x[\log(1-D(x+G(x,y)))] \tag{6}$$

为成功欺骗目标模型，采用生成器 G 和目标模型 f 联合训练的方式，并使用 L_{adv} 损失函数指导生成具有攻击性的 AE，L_{adv} 损失函数计算如式(7)所示。

$$L_{adv}=E_x l_f(x+G(x,y),y) \tag{7}$$

L_{adv} 函数目的是鼓励对抗样本被目标模型错误的分类到目标 y 中。此外，使用 L_{pert} 损失来充分约束扰动的大小，L_{pert} 函数计算如式(8)所示。

$$L_{pert}=E_x\max(0,\|G(x,y)\|_2-c) \tag{8}$$

式中：c 是攻击者指定的约束，本文将 c 设置为 0.3。

为了增强生成的 AE 在听觉感知上与真实语音的相似性，本文引入感知损失 $L_{perceptual}$。首先，将真实语音 x 和对抗样本 x'输入到目标模型 F 中，提取中间特征表示 F(x)和 F(x')；然后，利用 L_2 范数计算两组特征的差异；最后，将 L_2 范数除以特征尺寸 CHW，以便消除维度大小的影响。将感知损失与 L_{GAN}、L_{adv} 和 L_{pert} 一起进行联合优化，目的是在不牺牲对抗性的同时，保证 AE 在感知上不会与真实语音产生显著的差异，从而维持 AE 的真实性。感知损失函数计算如式(9)所示。

$$L_{perceptual}=\frac{1}{C_jH_jW_j}\|F_j(x)-F_j(x')\|_2 \tag{9}$$

式中：F 表示目标网络模型，$F_j(\cdot)$表示第 j 层的特征，$C_jH_jW_j$ 表示特征尺寸。

ACGAN 模型总体损失函数计算如式(10)所示，α、β、γ 为各部分损失函数的平衡参数。为了增强模型对扰动的约束能力，上述参数默认设置为 1、10、1，以提高生成的 AE 的听觉质量。

$$L=L_{GAN}+\alpha L_{adv}+\beta L_{pert}+\gamma L_{perceptual} \tag{10}$$

训练过程保证了算法 ACGAN 生成的对抗语音在数据分布上接近其对应的真实语音，并且由目标模型 f 输出对抗语音标签更有可能是攻击者指定的攻击目标标签 y。

3　实验设计与结果分析

3.1　数据集采集与预处理

为了评估 ACGAN 模型在说话人识别任务上的性能，采用两个公认的数据集：TIMIT[18] 和 LibriSpeech[19]。TIMIT 数据集包含 462 名美国各地多种方言的说话人语音样本。LibriSpeech 数据集包含 2 484 名说话人的大量英语朗读材料。

如表 1 所示，为了验证本模型的有效性，本文从 TIMIT 和 LibriSpeech 中共选取 16 791 个样本进行训练，8 820 个样本用于测试。由于 LibriSpeech 样本时长不同，将每个样本裁剪为 3.5 s。语音文件经过预加重处理以增强高频信息，并采用窗长 25 ms 和步长为 10 ms 的汉明窗进行分帧处理，从每个语音中提取 34 个 MFCC 系数。所有特征提取后进行标准化处理，以消除不同录音设备或环境引入的噪声和偏差。

表 1　训练集与测试集

Table 1　Training set and test set

数据集	说话人数量	训练集	测试集	采样率/kHz
TIMIT	462	2 310	1 368	16
LibriSpeech	2 484	14 481	7 452	16
合计	2 946	16 791	8 820	16

3.2 评价指标

实验采用 ASR、SNR 和 PESQ 作为评价指标。

1)ASR 表示生成的 AE 被目标模型错误分类的数量占比,计算如式(11)所示。

$$ASR = \frac{N}{M} \times 100\% \tag{11}$$

式中:测试样本数为 M,攻击成功的样本数为 N。

2)SNR 是用于衡量信号强度相对于噪声水平的指标。SNR 计算如式(12)所示。

$$SNR(x, x') = 10\lg \frac{\| x \|_2}{\| x - x' \|_2} \tag{12}$$

3)PESQ 是一种国际电信联盟(ITU)推荐的标准,通过比较真实语音和 AE,来评估语音质量损失。该过程涉及模拟人耳的听觉特性,以量化的方式评估语音样本的质量。PESQ 提供一个质量分数,范围通常在 -0.5 到 4.5 之间,其中高分表示高质量的语音。

3.3 目标模型

本文选用 SincNet[20] 说话人识别系统作为目标模型,因其对音频信号中底层声学特征的高度敏感性,这在生成 AE 时尤为关键。使用 SincNet 能更精确地验证 ACGAN 生成 AE 的效果,确保其在实际应用中具备真实性和较高的 ASR。

3.4 实验设置

ACGAN 模型实验在 Linux 环境上进行,GPU 为 NVIDIA A100。使用 Python 和 PyTorch 实现网络模型搭建、训练和测试。本文对比了五种攻击方法:FGSM、PGD、CW、UAPs 和 SSED。FGSM 攻击中,采用 L_∞ 范式度量扰动,阈值为 0.002。PGD 攻击的学习率设为 0.000 4,迭代 100 次。CW 攻击学习率为 0.001,内循环 10 次,外循环最多 10 次。UAPs 和 SSED 均训练 10 个 epochs,学习率为 0.000 1,使用 Adam 优化器。

ACGAN 模型训练时,总训练轮数为 100 个 epochs,使用 Adam 优化器,学习率从 0.01 逐渐降至 0.000 1,并采用 L_2 范式限制扰动大小,阈值为 0.3。

3.5 实验结果分析

1)无防御攻击评估

NTA 的目的是评估 ACGAN 模型在未指定标签时对 SRSs 的欺骗能力。

如表 2 所示,FGSM 在 TIMIT 和 LibriSpeech 上的 ASR 分别为 55.2% 和 48.4%,且 SNR 和 PESQ 相对较低。由于 FGSM 是一种单步攻击方法,只在输入数据的梯度方向上执行单次扰动,难以精细调控扰动以保持语音质量。相比之下,PGD 和 CW 表现出更强的攻击能力,ASR 均超过了 95%,但 AE 质量较低,它们通过多次迭代找到误导目标模型的有效扰动,尽管提高了 ASR,但损害了 AE 的真实性。UAPs 作为一种通用扰动方法,获得了较高的 ASR 和 AE 质量。SSED 仅优化了 ASR 和 SNR,但未优化 PESQ,导致 PESQ 表现不佳。ACGAN 在 ASR、SNR 和 PESQ 上表现优异,这种综合性能的提升源于 ACGAN 引入了 GCNN 和感知损失函数。这使得 ACGAN 在攻击过程中更精确地模拟真实语音特征,并有效控制了生成 AE 的听觉差异。

表 2 在 TIMIT 和 LibriSpeech 数据集上的 NTA

Table 2 NTA on TIMIT and LibriSpeech datasets

数据集	攻击方法	ASR/%	SNR/dB	PESQ
TIMIT	FGSM	55.2	27.53	2.14
	PGD	98.3	30.12	2.53
	CW	98.7	31.14	2.67
	UAPs	96.6	49.65	3.00
	SSED	96.8	50.82	2.41
	ACGAN	97.7	52.56	3.19
LibriSpeech	FGSM	48.4	27.53	2.08
	PGD	95.3	30.09	2.44
	CW	95.6	31.11	2.57
	UAPs	95.9	31.16	2.32
	SSED	96.3	41.23	2.21
	ACGAN	97.5	43.74	2.66

TA 的目的是评估 ACGAN 模型在指定标签时对 SRSs 的欺骗能力,其难度高于 NTA。

如表 3 所示,在 TIMIT 数据集上,ACGAN 的 ASR 为 97.4%,优于其他方法;在 LibriSpeech 数据集上,ACGAN 的 ASR 为 78.6%,其他方法普遍较低,这表明 ACGAN 在复杂和大规模数据集上仍能保持较高的效果。ACGAN 在两个数据集上均获得了最高的 SNR 和 PESQ 评分,特别是在 TIMIT 数据集上的 PESQ 达到了 3.22,远高于其他方法。ACGAN 的优势来源于结合了 AdvGAN 和 CGAN

表 3 在 TIMIT 和 LibriSpeech 数据集上的 TA

Table 3 TA on TIMIT and LibriSpeech datasets

数据集	攻击方法	ASR/%	SNR/dB	PESQ
TIMIT	FGSM	51.5	27.52	2.14
	PGD	96.3	30.21	2.55
	CW	96.7	31.28	2.70
	UAPs	96.3	48.51	2.47
	SSED	97.2	50.07	2.35
	ACGAN	97.4	52.78	3.22
LibriSpeech	FGSM	43.2	27.34	2.07
	PGD	75.5	30.07	2.49
	CW	74.4	31.12	2.61
	UAPs	63.7	29.89	2.11
	SSED	76.4	37.63	2.02
	ACGAN	78.6	44.72	2.78

的设计。AdvGAN 生成欺骗目标模型的 AE，CGAN 通过条件约束生成精确的 AE，结合后 ACGAN 不仅维持了高 ASR，还能针对特定标签生成精确的 AE，提高了针对性和有效性。尽管 ACGAN 表现出色，但在 LibriSpeech 数据集上的 ASR 仍有提升空间。

2)有防御攻击评估

防御情况下的攻击是指在目标模型采取防御手段时，若攻击模型生成的 AE 仍能成功误导目标模型，表明攻击模型在防御情况下依然有效。

本文采用 FGSM 进行对抗训练，通过在训练过程中加入 AE 增强模型鲁棒性。假设在 WBA 场景中攻击者未知防御机制，若能成功攻击目标模型，则证明攻击方法具备强适应能力和鲁棒性。实验在 TIMIT 数据集上进行，对比 FGSM、PGD、CW、UAPs 和 SSED 五种方法。如表 4 所示，ACGAN 攻击 SincNet 模型的 ASR 为 64.2%，SNR 为 49.56 dB，PESQ 分数为 3.14，均高于其他方法，表明 ACGAN 在面对防御策略时依然能够保持较高的 ASR 和语音质量。

表 4　在防御方法下的各项指标

Table 4　Various indicators under defense methods

数据集	攻击方法	ASR/%	SNR/dB	PESQ
TIMIT	FGSM	19.6	27.31	2.03
	PGD	57.8	30.41	2.37
	CW	56.4	31.14	2.59
	UAPs	57.3	39.58	2.37
	SSED	61.8	40.63	2.21
	ACGAN	64.2	42.33	2.75

3)黑盒攻击场景迁移能力评估

使用迁移攻击来评估 ACGAN 在 BBA 场景中的性能表现。作为一种 BBA 手段，迁移攻击在替代模型上生成 AE，并在黑盒场景中测试其对目标模型的攻击效果。实验中，使用 SSED 和 ACGAN 在 SincNet 上生成 AE，然后在 ECAPA-TDNN 目标模型上对生成的 AE 进行 NTA 评估。

如表 5 所示，ACGAN 在黑盒场景下的 ASR 达到了 70%以上。值得注意的是，与同样基于生成式攻击方法的 SSED 相比，在 2 个数据集上的迁移攻击中，ACGAN 的 ASR 增益都超过了 4%。证明了 ACGAN 在黑盒场景中具有较强的迁移能力。

4)生成效率

实验结果展示了在 NTA 的情况下，各方法在 TIMIT 数据集上生成 500 个 AE，平均生成一个样本所需的时间。

如表 6 所示，FGSM 是一种单步攻击方法，计算梯度时，只需一次向前传播和一次向后传播，不需要多次迭代来寻找最佳扰动，计算复杂度相对较低，生成 AE 的效率较高，生成时间为 0.9 s；PGD 通过多次迭代逐步调整扰动，每次迭代都需计算梯度并更新扰动，且对于语音数据需要大量的浮点运算，所需时间成本相对较高，生成时间是 FGSM 的数十倍；CW 使用二阶优化方法来最小化损失函数，比基于梯度的一阶方法(FGSM)需要更多的计算资源。优化过程需要多次迭代，每次迭代都需要计算损失函数和梯度，因此生成 AE 所需的时间要大于 FGSM 和 PGD，生成时间最长，超过 130 s；UAPs、SSED 和 ACGAN 都是基于生成网络的方法，生成时间短，训练好模型参数后，只需一次简单的前向传播即可快速生成 AE，其中 UAPs 计算出一个扰动后，可以被重复应用于多个样本，减少了对计算资源的需求，生成时间最短。ACGAN 相比于 SSED，每个样本的生成是独立的，生成过程很容易并行化，可以同时生成大量 AE，因此生成效率高于 SSED，生成时间为 0.01 s。

表 5　黑盒场景中 ACGAN 的迁移攻击性能

Table 5　Transfer attack performance of ACGAN in black box scenario

数据集	替代模型	目标模型	攻击方法	ASR/%
TIMIT	SincNet	ECAPA-TDNN	SSED	70.3
			ACGAN	74.8
LibriSpeech			SSED	67.5
			ACGAN	71.6

表 6　生成效率

Table 6　Generation efficiency

方法	时间/s
FGSM	0.9
PGD	11
CW	131
UAPs	0.004
SSED	0.52
ACGAN	0.01

5)消融实验

本消融实验旨在验证 GCNN 和感知损失在 ACGAN 模型中的作用。实验涉及 3 个模型：(1)完整的 ACGAN 模型；(2)无 GCNN 的 ACGAN 模型；(3)无感知损失的 ACGAN 模型。3 个模型在 TIMIT 数据集上对 SincNet 模型进行 NTA，训练轮数均为 100 个 epochs 以确保可比性。

图 5 展示了引入 GCNN 的消融实验结果。其中，两个模型在第 84 轮训练后达到了最高的 ASR。对比可知，在引入 GCNN 后的模型 ASR 更高，表明 GCNN 通过门控机制提高了对关键声学特征的捕捉能力，生成更具欺骗性的 AE。

图 6 展示了引入感知损失的消融实验结果。引入感知损失后，模型优化了 AE 的 SNR 和 PESQ，使生成的语音在听觉上更难以区分，提高了 AE 的真实性。

6)对抗样本可视化

通过比较原始语音和 AE 的波形图和时频谱图，可以直观地观察到添加扰动后原始语音的变化。如图 7、8 所

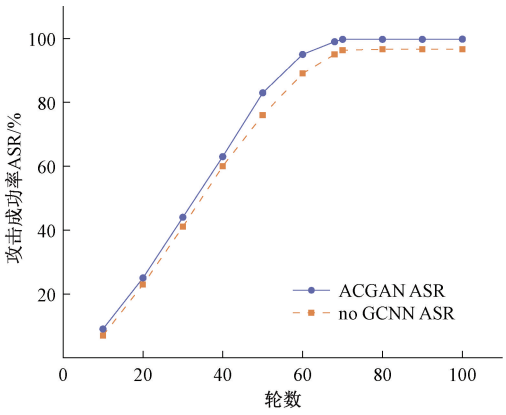

图 5 无 GCNN 对攻击成功率的影响

Fig. 5 Impact of no GCNN on attack success rate

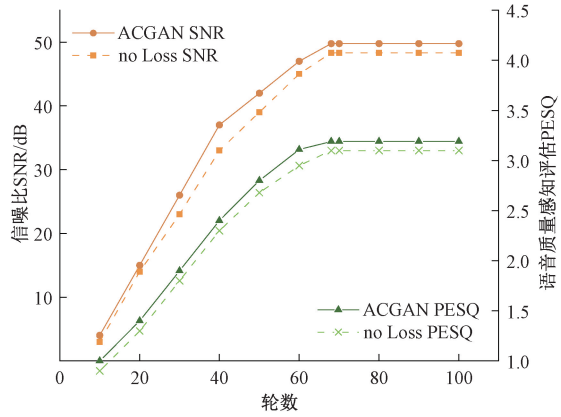

图 6 无感知损失对听觉质量的影响

Fig. 6 Impact of no perceived loss on hearing quality

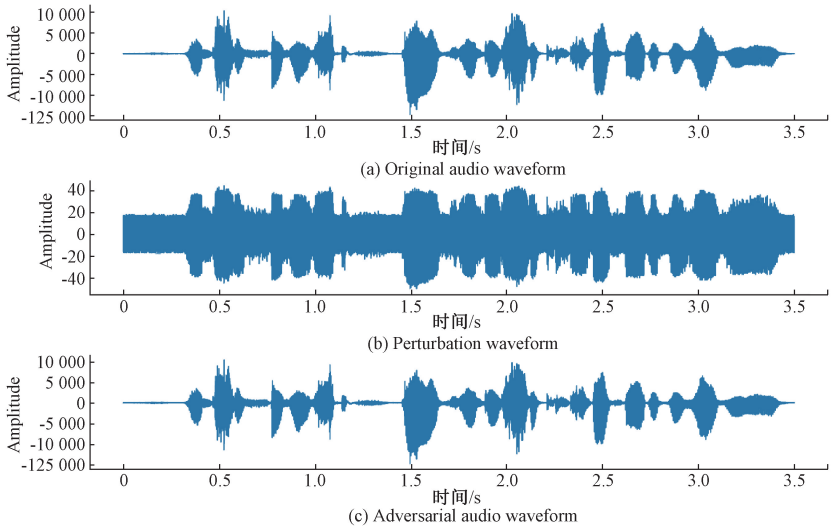

图 7 波形图

Fig. 7 Waveform

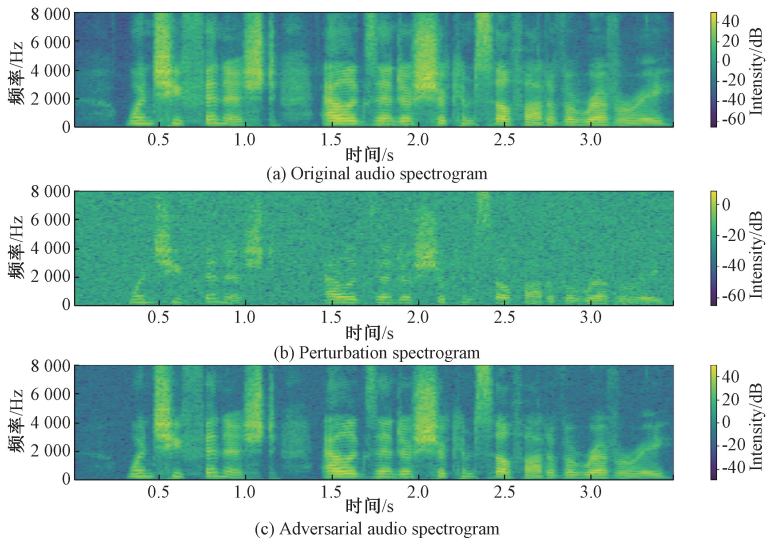

图 8 时频谱图

Fig. 8 Time-frequency spectrum

示，AE 的波形图和时频谱图更接近原始语音，生成的扰动显示出更低的全局能量，这表明生成的 AE 具有较高的真实性，且不易被人耳察觉。

4　结　　论

本文提出了一种基于 AdvGAN 和 CGAN 相结合的高效 AE 生成方法 ACGAN，以解决现有方法在听觉质量和 ASR 方面的不足。ACGAN 通过引入目标标签，实现针对特定目标的精确攻击，同时，引入 GCNN 和感知损失函数，优化听觉质量，提高 ASR。实验表明，ACGAN 在 TIMIT 和 LibriSpeech 数据集上的性能优于 FGSM、PGD、CW、UAPs 和 SSED。此外，ACGAN 仅需一次前向传播即可高效生成 AE，生成效率高。面对防御策略时，ACGAN 仍保持较高的 ASR，证明了其在 SRSs 对抗样本生成领域的有效性和潜力。

参考文献

[1] SHOME N, SARKAR A, GHOSH A K, et al. Speaker recognition through deep learning techniques: A comprehensive review and research challenges[J]. Periodica Polytechnica Electrical Engineering and Computer Science, 2023, 67(3): 300-336.

[2] LAMBAMO W, SRINIVASAGAN R, JIFARA W. Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition[J]. Applied Sciences, 2022, 13(1): 569.

[3] CHEN G K, CHEN S, FAN L L, et al. Who is real bob? Adversarial attacks on speaker recognition systems[C]. IEEE Symposium on Security and Privacy, 2021: 694-711.

[4] AYAZ F, ZAKARIYYA I, CANO J, et al. Improving robustness against adversarial attacks with deeply quantized neural networks[C]. International Joint Conference on Neural Networks, 2023: 1-8.

[5] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verification with adversarial examples[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 1962-1966.

[6] CHEN G, ZHAO ZH, SONG F, et al. SEC4SR: A security analysis platform for speaker recognition[J]. ArXiv preprint arXiv:2109.01766, 2021.

[7] YAO J D, CHEN X, ZHANG X L, et al. Symmetric saliency-based adversarial attack to speaker identification[J]. IEEE Signal Processing Letters, 2023, 30: 1-5.

[8] LI J G, ZHANG X F, JIA CH M, et al. Universal adversarial perturbations generative network for speaker recognition[C]. IEEE International Conference on Multimedia and Expo, 2020: 1-6.

[9] 陈亮，吴攀，刘韵婷，等. 生成对抗网络 GAN 的发展与最新应用[J]. 电子测量与仪器学报，2020，34(6)：70-78.
CHEN L, WU P, LIU Y T, et al. Development and latest application of generative adversarial network GAN[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(6): 70-78.

[10] 刘伟，王智豪，李卓，等. 基于 cGAN-SAE 的室内定位指纹生成方法[J/OL]. 电子测量技术，2024，1-8[2024-01-08]. http://kns.cnki.net/kcms/detail/11.2175.TN.20240927.1500.182.html.
LIU W, WANG ZH H, LI ZH, et al. Indoor location fingerprint generation method based on cGAN-SAE[J/OL]. Electronic Measurement Technology, 2024, 1-8[2024-01-08]. http://kns.cnki.net/kcms/detail/11.2175.TN.20240927.1500.182.html.

[11] LAPID R, SIPPER M. I see dead people: Gray-box adversarial attack on image-to-text models[J]. ArXiv preprint arXiv:2306.07591, 2023.

[12] LU S Y, WANG M Q, WANG D D, et al. Black-box attacks against log anomaly detection with adversarial examples[J]. Information Sciences, 2023, 619: 249-262.

[13] WEI ZH P, CHEN J J, WU Z X, et al. Enhancing the self-universality for transferable targeted attacks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12281-12290.

[14] WENG J J, LUO ZH M, ZHONG ZH, et al. Exploring non-target knowledge for improving ensemble universal adversarial attacks[C]. AAAI Conference on Artificial Intelligence, 2023, 37(3): 2768-2775.

[15] BARREIRO A, LIGA G, ALVARADO A. Data-driven enhancement of the time-domain first-order regular perturbation model[J]. Journal of Lightwave Technology, 2023, 41(9): 2691-2706.

[16] QI ZH Y, SHI ZH M, RASMUSSEN T C. Time-and frequency-domain determination of aquifer hydraulic properties using water-level responses to natural perturbations: A case study of the Rongchang Well,

Chongqing, southwestern China [J]. Journal of Hydrology, 2023, 617: 128820.

[17] XIAO CH W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[J]. ArXiv preprint arXiv:1801.02610, 2018.

[18] HUANG P S, AVRON H, SAINATH T N, et al. Kernel methods match deep neural networks on timit[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 205-209.

[19] TOMASELLO P, SHRIVASTAVA A, LAZAR D, et al. Stop: A dataset for spoken task oriented semantic parsing [C]. IEEE Spoken Language Technology Workshop, 2023: 991-998.

[20] RAVANELLI M, BENGIO Y. Speaker recognition from raw waveform with sincnet[C]. IEEE Spoken Language Technology Workshop, 2018: 1021-1028.

作者简介

马栋林,副教授,硕士生导师,主要研究方向为模式识别与人工智能。

E-mail: 592004890@qq.com

宋佳佳(通信作者),硕士研究生,主要研究方向为说话人识别对抗样本生成。

E-mail: 1150485908@qq.cpm

赵宏,教授,博士生导师,主要研究方向为计算机视觉,自然语言处理。

E-mail: 594286500@qq.com

陈伟杰,硕士研究生,主要研究方向为网络信息安全。

E-mail: 2900373335@qq.com