第47卷 第6期 计 算 机 工 程 2021年6月
Vol.47 No.6 Computer Engineering June 2021

·网络空间安全· 文章编号：1000-3428(2021)06-0132-10 文献标志码：A 中图分类号：TP391

端到端说话人辨认的对抗样本应用比较研究

廖俊帆[1]，顾益军[1]，张培晶[2]，廖 茜[1]

（1.中国人民公安大学 信息网络安全学院，北京 102600；2.中国人民公安大学 网络信息中心，北京 100038）

摘 要：为探究对抗样本对端到端说话人辨认系统的安全威胁与攻击效果，比较现有对抗样本生成算法在语音环境下的性能优劣势，分析FGSM、JSMA、BIM、C&W、PGD 5种白盒算法和ZOO、HSJA 2种黑盒算法。将7种对抗样本生成算法在ResCNN和GRU两种网络结构的端到端说话人辨认模型中实现有目标和无目标攻击，并制作音频对抗样本，通过攻击成功率和信噪比等性能指标评估攻击效果并进行人工隐蔽性测试。实验结果表明，现有对抗样本生成算法可在端到端说话人辨认模型中进行实现，白盒算法中的BIM、PGD具有较好的性能表现，黑盒算法的无目标攻击能达到白盒算法的攻击效果，但其有目标攻击性能有待进一步提升。

关键词：说话人辨认；对抗样本；鲁棒性；对抗攻击；信噪比

开放科学（资源服务）标志码（OSID）：

中文引用格式：廖俊帆，顾益军，张培晶，等.端到端说话人辨认的对抗样本应用比较研究[J].计算机工程，2021，47(6)：132-141.

英文引用格式：LIAO Junfan，GU Yijun，ZHANG Peijing，et al.Comparative research on application of adversarial samples for end-to-end speaker identification[J].Computer Engineering，2021，47(6)：132-141.

Comparative Research on Application of Adversarial Samples for End-to-End Speaker Identification

LIAO Junfan[1]，GU Yijun[1]，ZHANG Peijing[2]，LIAO Qian[1]

（1.College of Information Network Security，People's Public Security University of China，Beijing 102600，China；
2.Network Information Center，People's Public Security University of China，Beijing 100038，China）

【Abstract】In order to explore the security threats and attack effects of the adversarial samples on the end-to-end speaker identification system，this paper analyzes five white box algorithms（FGSM，JSMA，BIM，C&W，PGD）and two black box algorithms（ZOO，HSJA）to compare the advantages and disadvantages of the existing adversarial sample generation algorithms in a phonetic context.Each generation algorithm implements targeted and non-targeted attacks in the end-to-end speaker identification model of ResCNN and GRU，and creates effective audio adversarial samples.Then the attack effects are evaluated by using the performance indicators such as Attack Success Rate（ASR）and Signal to Noise Ratio（SNR）. Finally，a manual concealment test is performed. Experimental results show that the existing adversarial sample generation algorithms can be implemented in the end-to-end speaker identification model.The BIM and PGD in the white box generation algorithm have excellent performance.The black box generation algorithm gets non-targeted attacks that are on par with that of the white box generation algorithm，while its targeted attack effect still needs improvement.

【Key words】speaker identification；adversarial sample；robustness；adversarial attack；Signal to Noise Ratio（SNR）

DOI：10. 19678/j. issn. 1000-3428. 0058239

0 概述

语音是人与人之间最自然直接的交流方式，也是具有最大信息容量的信息载体。目前，说话人识别技术已在人们日常生活中得到了广泛的应用，说话人辨认技术作为其重要分支在公安司法等领域具有较好的发展前景。随着人工智能和大数据时代的到来，同时得益于计算机计算能力的不断提高，深度学习技术已

基金项目：公安部技术研究计划竞争性遴选项目（2019JZX009）；中国人民公安大学公共安全行为科学研究与技术创新专项。

作者简介：廖俊帆（1995—），男，硕士研究生，主研方向为对抗样本攻击与防御；顾益军（通信作者），教授、博士；张培晶，副研究员、硕士；廖 茜，硕士研究生。

收稿日期：2020-05-03 **修回日期：**2020-06-17 **E-mail：**754605668@qq.com

经成为各界研究的热点，其可应用于说话人辨认系统的后端，使声学特征更具区分性，从而更有利于区分说话人，而端到端网络架构使用一个神经网络连接输入端和输出端，能将特征训练和分类打分进行联合优化[1-3]。因此，结合基于深度学习的端到端网络的说话人辨认技术能克服复杂环境干扰，具有易构建、强泛化的特点。机器学习算法是人工智能中的重要部分，给人们带来便利的同时也带来了诸多安全问题。机器学习模型的攻击方式一般为破坏其机密性、完整性和可用性，主要包括隐私攻击、针对训练数据的攻击以及针对算法模型的攻击[4-5]三类方式。对抗样本是能轻易地引发模型分类错误的针对算法模型的攻击方式[6-7]，随着对抗样本在图像、自动驾驶等领域被证实可使攻击者逃避模型检测，研究人员发现机器学习模型面对对抗样本表现出的脆弱性问题是普遍存在的，而基于深度学习的端到端说话人辨认模型也可能受到对抗样本的攻击。

为准确全面地评估端到端说话人识别技术面临的安全问题，本文系统地分析端到端说话人辨认系统和目前多种经典的白盒算法和黑盒算法，以基于卷积结构的端到端说话人辨认模型作为实验对象，通过实验比较评估这些对抗样本对端到端说话人辨认系统的攻击性能。

1　端到端说话人辨认

1.1　基于深度学习的端到端说话人辨认

说话人辨认是多分类问题[8]，即判断某段语音是由若干人中哪个人所说。端到端说话人辨认系统由深度神经网络组成，深度神经网络将不同长度的语段映射为一定维度的特征向量，即深度嵌入，再将不同说话人的语音特征映射到超球面的不同区域，最终通过各区域之间的差异实现分类。在识别过程中需要先在语音数据中提取声学特征，使用$\boldsymbol{X}\subset\mathbb{R}^d$表示声学特征向量的域，声学特征表示为向量序列$\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$，其中$\boldsymbol{x}_i\in\boldsymbol{X}$且$1\leqslant i\leqslant T$，由于输入信号长度不固定，因此$T$值也不固定。将特征向量$\boldsymbol{x}$输入深度神经网络生成帧级别的特征，帧级别的特征被激活后输入平均池化层得到话语级别的特征，再利用仿射层进行维度转换得到固定维度的深度说话人嵌入，最终输出层将固定维度的深度说话人嵌入映射到训练说话人类别。

1.2　针对端到端说话人辨认的攻击模型

针对端到端说话人辨认系统的对抗攻击，需要运用对抗样本生成算法制作针对端到端说话人辨认模型的对抗样本。对抗样本可以诱导模型算法出现误判或漏判，从而躲避系统的识别实现攻击。本文将在白盒和黑盒设置下对端到端说话人辨认模型进行攻击。在白盒设置下，攻击者可以完全访问说话人辨认系统，根据获取到的梯度信息制作噪声，并且能最大程度地减少扰动提高成功率。在黑盒设置下，攻击者只能有限制地访问模型，并且仅获得端到端说话人辨认模型的输出，无法直接获取输入与输出之间的梯度。与在声学特征上生成对抗样本的方法[9-10]不同，本文是在音频上直接制作对抗样本，具备更好的隐蔽性。如图1所示，一段音频经攻击者添加噪声后被输入目标说话人辨认系统中，攻击者根据模型反馈信息反复对噪声进行修改，最终制作出对抗样本，实现端到端说话人辨认系统的错误识别。

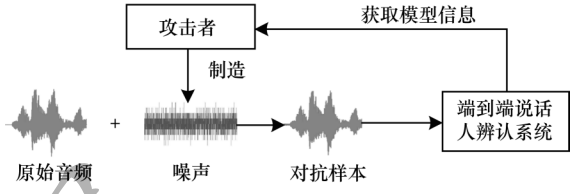

图1　攻击步骤

Fig.1　Attack steps

2　对抗样本生成算法

利用深度神经网络训练得到的模型在输入和输出之间的映射通常为非线性，因此在输入数据中通过故意添加不易察觉的细微扰动来生成的对抗样本，能够导致模型以高置信度给出一个错误的输出。对抗样本能够找出机器学习模型的弱点，在网络安全领域主要用于模型安全评估和对抗鲁棒性强化。

目前，关于攻击的分类有很多种，按照是否获得目标模型的具体结构和参数可分为白盒攻击和黑盒攻击。白盒攻击指攻击者能获取目标模型的所有信息，对抗样本较多，如FGSM[11]、JSMA[12]、BIM[13]、C&W[14]、PGD[15]等；黑盒攻击指攻击者无法直接获取模型的任何信息，只能通过访问模型来获取反馈信息对黑盒模型进行估计，从而使得攻击成功，如ZOO[16]、HSJA[17]等。此外，按照是否需要指定攻击类目可分为无目标攻击和有目标攻击。无目标攻击不指定具体类目，只需使识别模型出现错误，如Deepfool[18]等。有目标攻击比无目标攻击更困难，不仅需要识别模型出现错误，还需模型输出指定的结果，如C&W等。现有的对抗样本生成算法并不都能适应音频数据中复杂的时间域信息和计算复杂度，因此难以在端到端说话人辨认系统中进行实现，如Deepfool。本文仅选取可用于端到端说话人辨认系统的FGSM、JSMA、BIM、C&W、PGD这5种白盒算法和ZOO、HSJA这2种黑盒算法进行对抗样本攻击实验。

2.1　白盒算法

2.1.1　FSGM算法

在一般情况下，给定分类网络F和输入$\boldsymbol{x}$，通过求优化问题式(1)生成对抗样本，即在允许的最大扰动量ε的约束下，扰动δ的p范数能实现最大化网络预测$F(\boldsymbol{x}+\delta)$和真实标签y的损失函数L。

$$\delta=\underset{\|\delta\|_p\leqslant\varepsilon}{\operatorname{argmax}}\,L(F(\boldsymbol{x}+\delta),y) \tag{1}$$

FSGM[8]是根据高维空间下深度神经网络的线性行为会导致对抗样本的产生而设计得到，并利用损失函数梯度解决优化问题式(1)，计算公式如下：

$$\boldsymbol{x}'=\boldsymbol{x}+\varepsilon\cdot\operatorname{sign}(\nabla L(F(\boldsymbol{x}),y)) \tag{2}$$

其中，$\nabla L(F(\boldsymbol{x}),y))$表示损失函数的偏导数。若是目标攻击，则将$y$换成目标标签$t$。FSGM攻击需要考虑损失函数相对于输入梯度的符号，适用于端到端说话人辨认的非线性模型。本文采用的分类模型F包含特征提取模块，对应输入音频$\boldsymbol{x}$无需进行过多预处理，仅将扰动噪声添加到测试音频中。FGSM对抗样本生成速度快，但攻击性较弱，对模型防御能力提升小。

2.1.2 JSMA算法

JSMA[12]算法利用显著性映射，能够表征分类器的输出与输入之间的关联，仅在样本$\boldsymbol{x}$的关键分量上添加扰动，能够得到使分类器输出指定类目的对抗样本。因为分类器的结果受输入样本$\boldsymbol{x}$某些分量的影响较大，不同于FGSM的梯度通过对损失函数求导获得，JSMA算法的前向导数是神经网络的logit层的输出$Z(\cdot)$对输入特征的偏导，所以在端到端说话人辨认网络中实现分类器对样本x的显著性映射如下：

$$S(\boldsymbol{x},t)[i]\begin{cases}0, \dfrac{\partial Z_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}>0 \text{ 且 } \dfrac{\partial Z_t(\boldsymbol{x})}{\partial \boldsymbol{x}_i}<0\\ \dfrac{\partial Z_t(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\left|\dfrac{\partial Z_j(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\right|, \text{其他}\end{cases} \tag{3}$$

其中，i表示对应的输入分量，t表示分类器对应目标标签的输出分量，j表示输出的其他分量。根据最大化显著性效果获得输入的关键分量k，因此在迭代过程中对其添加扰动：

$$k_n=\underset{i}{\operatorname{argmax}}\, S(\boldsymbol{x}_{n-1},t)[i],\ \boldsymbol{x}_0=\boldsymbol{x}$$
$$\boldsymbol{x}_n=\begin{cases}\boldsymbol{x}_{n-1,i}+\varepsilon, i=k_n\\ \boldsymbol{x}_{n-1,i}, i\neq k_n\end{cases} \tag{4}$$

在获得的特征上添加扰动获得对抗样本，扰动方式分为正向扰动和反向扰动。不同于图像数值全为正值，音频的波形数值是正负值并存，实现结果可能有所差异。JSMA是基于梯度的迭代算法，仅对样本的部分分量进行修改，与原样本的相似度高，但是每次迭代均需要重新计算显著图，因此生成速度较慢，不适用于部分大规模数据集。

2.1.3 BIM算法

由于FGSM算法仅涉及单次梯度更新，对于大规模数据出错概率较高，因此KURAKIN等人[13]提出快速梯度符号法的改进迭代算法。迭代梯度符号法的对抗样本生成算法如下：

$$\boldsymbol{x}'_i=\boldsymbol{x}'_{i-1}+\text{clip}_{\alpha}(\varepsilon\cdot\text{sign}(\nabla L(F(\boldsymbol{x}),y)))$$
$$\boldsymbol{x}'_0=\boldsymbol{x} \tag{5}$$

其中，clip表示将溢出的数值用边界值代替，这是因为在迭代更新中，随着迭代次数的增加，部分元素可能会溢出，只有代替这些数值原有的边界值，才能生成有效的对抗样本。相比FGSM，BIM能够在音频信号中寻找更精准有效的噪声点，实现性能更优的对抗音频。

2.1.4 C&W算法

C&W[14]算法在式(1)的优化问题上添加欧几里得距离来量化对抗样本$\boldsymbol{x}'$和原始样本$\boldsymbol{x}$之间的差异$\|\boldsymbol{x}'-\boldsymbol{x}\|_2^2$。为消除$\boldsymbol{x}'\in[0,1]^p$区间约束，将$\boldsymbol{x}'$替换为$\frac{1}{2}(\tanh\omega+1),\omega\in\mathbb{R}^p$，由此将优化问题转化为无约束的最小化问题，如式(6)所示：

$$\text{minimize}_{\omega}\left\|\frac{1}{2}(\tanh\omega+1)-\boldsymbol{x}\right\|_2^2+\varepsilon\cdot f\left(\frac{1}{2}(\tanh\omega+1),t\right) \tag{6}$$

通过映射到tanh空间，对抗样本能在$(-\infty,+\infty)$上进行变换，其中$f(\boldsymbol{x},t)$表示损失函数，反映了对抗攻击的不成功概率，t表示目标类别。损失函数一般表示为：

$$L(\boldsymbol{x}',t)=\max\left\{\max_{i\neq t}\boldsymbol{x}'_i-\boldsymbol{x}'_t,-k\right\} \tag{7}$$

其中：$k\geqslant 0$表示攻击传递性的调整参数，k确保了$\max\limits_{i\neq t}[Z(\boldsymbol{x})]_i$和$\max\limits_{i\neq t}[Z(\boldsymbol{x})]_i$的恒定距离，随着$k$值的增大，攻击成功率越高；$Z(\cdot)$表示logit层的输出。C&W算法生成的扰动极小，但消耗时间较长。CARLINI等人[19]将C&W算法应用在语音识别模型中，并使语音识别模型能将任意音频输出为特定目标句子，因此C&W算法也可应用在说话人辨认模型中。

2.1.5 PGD算法

PGD[15]算法是一种迭代算法，可看作是在BIM的基础上添加一层随机化处理，其允许在范数球内的随机点上初始化，然后进行基本迭代，每次迭代均会将扰动投影到规定范围内，但能产生比BIM更好的攻击效果。在迭代过程中，将对抗音频进行如下操作：

$$\boldsymbol{x}_{t+1}=\Pi_{x+S}\boldsymbol{x}_t+\frac{\alpha\cdot g(\boldsymbol{x}_t)}{\|g(\boldsymbol{x}_t)\|_2}$$
$$g(\boldsymbol{x}_t)=\nabla L(F(\boldsymbol{x}_t),y) \tag{8}$$

其中，$S=r\in\mathbb{R}^d(\|r\|_2\leqslant\varepsilon)$表示扰动的约束空间，$\alpha$表示扰动修改的步长，$\Pi_{x+S}$表示在范数球上进行投影。在迭代过程中，若添加的扰动幅度过大，则将其拉回范数球的边界。通过一阶梯度得到的样本被称为一阶对抗样本，而PGD是一阶对抗样本中最优的对抗样本生成算法。PGD可看作是FGSM的拓展，能够在端到端说话人辨认模型上进行实现。

2.2 黑盒算法

2.2.1 ZOO算法

ZOO[16]算法基于C&W算法并修改其损失函数实现黑盒设置下的攻击，而无需替代模型[20]，其使用有限差分法获取近似梯度来解决黑盒设置下无法获取模型梯度的问题。受C&W算法启发，CHEN[16]等人提出一种新的类似铰链的损失函数，具体为：

$$L(\boldsymbol{x},t)=\begin{cases}\max\left\{\max\limits_{i\neq t}\ln[F(\boldsymbol{x})]_i-\ln[F(\boldsymbol{x})]_t,-k\right\}, \text{有目标攻击}\\ \max\left\{\ln[F(\boldsymbol{x})]_{t_0}-\max\limits_{i\neq t_0}\ln[F(\boldsymbol{x})]_i,-k\right\}, \text{无目标攻击}\end{cases} \tag{9}$$

其中，t_0表示$\boldsymbol{x}$的原始标签，$\max\limits_{i\neq t_0}\ln[F(\boldsymbol{x})]_i$表示除$t_0$之外最可能的预测类别。

对数运算符对黑盒攻击至关重要，因为DNN通常会在输出F上产生偏斜的概率分布，此类的置信度得分显著地支配另一类的置信度得分。因此，使用对数运算可减少主导效应，并保留由于单调性而导致的置信度得分顺序，同时采用对称差商[21]或Hessian估计来估计梯度：

$$\hat{g}_t=\frac{\partial L(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\approx\frac{L(\boldsymbol{x}+he_i)-L(\boldsymbol{x}-he_i)}{2h} \tag{10}$$

$$\hat{h}_t=\frac{\partial L(\boldsymbol{x})}{\partial \boldsymbol{x}_{ii}^2}\approx\frac{L(\boldsymbol{x}+he_i)-2L(\boldsymbol{x})+L(\boldsymbol{x}-he_i)}{h^2} \tag{11}$$

梯度评估是将黑盒转化为白盒的过程。两种估计方式分别对应ZOO的两种变体，即ZOO-ADAM和ZOO-Newton，并对应ADAM和Newton求解器以找到最佳的坐标进行更新。ZOO采用随机坐标下降来替代梯度下降方法，在每次迭代中随机选择一个变量（坐标），通过沿该坐标近似最小化目标函数进行更新，实现更快速有效的更新过程。ZOO适用于端到端说话人辨认模型，但对目标模型的访问次数较多，查询效率较低。

2.2.2　HSJA算法

HSJA[17]算法在决策边界使用二进制信息对目标模型的梯度方向进行预估，利用L_2和L_∞的相似性指标进行优化的无目标和有目标攻击。与边界攻击[22]相比，HSJA需要的模型查询更少，在攻击多种广泛使用的防御机制时，具有一定优势。HSJA引入布尔值函数$\phi_{x^*}:[0,1]^n\rightarrow\{-1,1\}$作为成功扰动的指标，对抗样本的目标是生成对抗样本$\boldsymbol{x}'$，使得$\phi_{x^*}(\boldsymbol{x}')=1$，同时保持$\boldsymbol{x}'$接近原始样本$\boldsymbol{x}$，从而将对抗样本制作问题转化为最优化问题，如式（12）所示：

$$\min_{x'} d(\boldsymbol{x}',\boldsymbol{x}),\ \phi_x(\boldsymbol{x}')=1 \tag{12}$$

其中，d是量化相似度的距离函数，HSJA为迭代算法，每次迭代均涉及梯度方向估计、通过几何级数进行步长搜索以及利用二分搜索将最后一次迭代推向边界这3个步骤。HSJA查询效率高，具有收敛性分析，适用于端到端说话人辨认模型，但对于限制边界查询的目标模型的攻击效果较差。

3　实验设置与结果分析

3.1　实验目标模型

本文选用百度的DeepSpeaker[23]作为目标模型，包括ResCNN和GRU两种模型，它们是目前最具代表性的基于深度学习的端到端说话人识别模型。在声学特征提取阶段，为保留更丰富的原始音频信息，将语音信号利用帧长25 ms、帧移10 ms的滑动窗口转化为64维FBank（FilterBank）特征。每个样本随机截取多个约1.5 s的语音段，生成160×64的特征矩阵。ResCNN和GRU网络结构见表1和表2，其中，“—”表示该层网络不涉及相应参数。

表1　ResCNN网络结构

Table 1　ResCNN network structure

网络层名称	网络结构	步长	输出尺寸
Conv64-s	5×5, 64	2×2	80×32×64
Res64	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	1×1	80×32×64
Conv128-s	5×5, 128	2×2	40×16×128
Res128	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times3$	1×1	40×16×128
Conv256-s	5×5, 256	2×2	20×8×256
Res256	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times3$	1×1	20×8×256
Conv512-s	5×5, 512	2×2	10×4×512
Res512	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	1×1	10×4×512
average	—	—	2 048
affine	2 048×512	—	512
ln	—	—	512

表2　GRU网络结构

Table 2　GRU network structure

网络层名称	网络结构	步长	输出尺寸
Conv64-s	5×5, 64	2×2	80×2 048
GRU	1 024cells	1	80×1 024
GRU	1 024cells	1	80×1 024
GRU	1 024cells	1	80×1 024
average	—	—	1 024
affine	1 024×512	—	512
ln	—	—	512

ResCNN网络中两个卷积核为3×3、步长为1×1的卷积层组成1个残差块，实现低层输出到高层输入的直接连接。ResCNN网络具有4种残差块，每种残差块有3个。同时，残差块后的一个卷积核为5×5、步长为2×2的卷积层使频域的维度在输出通道数增加时保持不变。经过多个卷积层和残差块提取到的帧级别特征进入时间平均池化层（average）。GRU网络使用和ResCNN网络相同的卷积层来降低时域和频域的维度。卷积层之后是3个前向的GRU层。时间平均池化层对特征在时域上整体取均值，得到话语级别的特征，使得构建的网络在时间位置上具有不变性，再经过仿射层（affine）将语音级别的特征映射成512维的深度说话人嵌入。最后输入Softmax层进行分类。

3.2　实验数据集及环境设置

实验使用中文语音数据库AISHELL-1（简记为AISHELL）[24]和英文语音数据库LIBRISPEECH（简记为LIBRI）[25]。AISHELL的录音文本涉及智能家居、无人驾驶和工业生产等，并且在安静室内同时使用3种不同设备总共录制178 h，其中包含400个说话人。LIBRI数据集包含1 000 h的16 kHz英语语

料。实验训练了400个说话人和10个说话人的端到端说话人识别模型，分别用于无目标的对抗攻击和有目标的对抗攻击。

实验平台及环境：Intel® Xeon™ Gold 5118 CPU@2.30 GHz（CPU），Tesla-V100-SXM2-32 GB（GPU），32 GB memory，Ubuntu 18.04.3 LTS（OS），Python 3.6，Tensorflow 2.10。

3.3 评价指标

本文使用攻击成功率（Attack Success Rate，ASR）、扰动大小、置信度、对抗样本生成时间来评价各生成算法对端到端说话人识别模型的性能。

攻击成功率：成功逃避模型识别的样本数占测试样本总数的比例，计算公式如下：

$$A_{\mathrm{ASR}}=\frac{s_{\mathrm{sumNum}}\left(l_{\mathrm{label}}\left(\boldsymbol{x}'\right)\neq y_0\right)}{s_{\mathrm{sumNum}}\left(l_{\mathrm{label}}\left(\boldsymbol{x}\right)=y_0\right)} \tag{13}$$

其中，$s_{\mathrm{sumNum}}(\cdot)$表示样本数量，$\boldsymbol{x}$表示原音频，$\boldsymbol{x}'$表示对抗样本，$l_{\mathrm{label}}(\cdot)$表示模型输出标签，$y_0$表示真实说话人标签；若有目标攻击时，分母改为$s_{\mathrm{sumNum}}\left(l_{\mathrm{label}}\left(x'\right)=y_t\right)$，$y_t$是目标说话人标签。

生成时间：生成一定数量的对抗样本所需的时间。为了准确地评估各算法的生成速度，实验设置的算法生成批次大小均为1，即每批次只生成一个对抗样本。

扰动大小：样本修改前后的变化量，衡量样本被处理前后的变化程度，计算公式如下：

$$\delta=\frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\boldsymbol{x}_i'-\boldsymbol{x}_i\right\|_1}{\left\|\boldsymbol{x}_i\right\|_1} \tag{14}$$

其中，N为样本个数，$\|\cdot\|_1$为1范数。

信噪比（Signal to Noise Ratio，SNR）：信号功率与噪声功率的比值，通常用来评估音频质量，计算公式如下：

$$S_{\mathrm{SNR}}=10\lg\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}=20\lg\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}} \tag{15}$$

其中，P_{signal}为信号功率，P_{noise}为噪声功率，A_{signal}为信号幅度，A_{noise}为噪声幅度。较大的SNR值表示较小的噪声等级。在本文实验中，SNR用来衡量对抗音频相对于原始音频的失真，比较生成算法生成的对抗性音频的差异。

置信度：在无目标攻击实验中，样本鲁棒性使用原类标置信度表示，对抗样本被识别为原类标的置信度越低，表示该样本越鲁棒。在有目标攻击的实验中，样本鲁棒性使用目标类标置信度表示，对抗样本被识别成目标类别的置信度越高，表示该样本越鲁棒。

3.4 算法参数设置

表3和表4表明FGSM、BIM、PGD的ASR和扰动随参数ε增加而增大，C&W在范数L_2和L_∞下的ASR随k变化不大，而扰动随之增大。但是，JSMA、ZOO和HSJA参数多样，难以统一比较。为在相似的攻击强度下对生成算法进行比较，在后续实验中：FGSM、BIM、PGD的度量单位均为L_∞且$\varepsilon=0.001$（描述可修改的L_∞范围大小）；JSMA的度量单位为L_2；C&W和ZOO使用置信度参数k来描述扰动大小且设置为0.0，其中C&W分别使用L_2和L_∞两种度量单位进行实验；JSMA设置每步修改的扰动量为0.1，最大特征分数为1.0。HSJA的初次和最大评估次数分别设置为100和1 000。

表3 不同ε下FGSM、BIM和PGD算法的ASR和扰动大小

Table 3 The ASR and perturbation size of FGSM, BIM and PGD algorithms under different ε

ε	FGSM		BIM		PGD	
	ASR/%	δ	ASR/%	δ	ASR/%	δ
0.000 1	52	0.000 1	84	0.000 1	84	0.000 1
0.000 5	89	0.000 5	100	0.000 4	100	0.000 4
0.001 0	91	0.001 0	100	0.000 7	100	0.000 7
0.005 0	95	0.005 0	100	0.003 1	100	0.003 1
0.010 0	98	0.010 0	100	0.006 0	100	0.006 0
0.050 0	52	0.050 0	84	0.029 2	84	0.029 2

表4 不同k和范数下C&W算法的ASR和扰动大小

Table 4 The ASR and and perturbation size of C&W algorithm under different k and norms

k	C&W(L_2)		C&W(L_∞)	
	ASR/%	δ	ASR/%	δ
0.0	100	2.83E-05	100	0.001 8
0.2	100	3.09E-05	100	0.002 9
0.4	100	3.28E-05	100	0.004 2
0.6	100	3.45E-05	100	0.004 7
0.8	100	3.66E-05	100	0.005 5

3.5 实验结果分析

3.5.1 无目标攻击实验结果分析

在无目标攻击的实验中，对于不同的生成算法，使用相同的100段音频，各自分别对不同网络结构和数据库训练的模型生成100个对抗样本。

表5给出了无目标攻击时各生成对抗样本算法的ASR、扰动大小和生成时间。对于说话人辨认的无目标攻击，8种算法均能躲避系统识别。平均扰动的值越小，噪声越小，这样能使对抗音频对人类的听力更加难以察觉，各算法均具有较小的扰动。FGSM无需进行迭代，生成速度最快，但ASR劣于其他算法。从生成时间而言，黑盒攻击明显比白盒攻击花费更多的生成时间。

表6给出了无目标攻击时各生成对抗样本算法

的信噪比，各算法得到的对抗样本都有较好的平均信噪比，但JSMA、C&W(L_∞)和ZOO的最低信噪比接近0，甚至负值。这说明音频信息完全丢失，无法完成攻击，C&W(L_2)和HSJA的平均信噪比在白盒和黑盒攻击时均最高，几乎能够躲避人听力的察觉。

表7给出了无目标攻击中对抗样本被端到端说话人辨认模型识别为真实类目的置信度。可以看出，面对端到端说话人辨认模型，每种算法均能使对抗样本偏离真实类目，但C&W(L_2)、C&W(L_∞)和ZOO高低差异较大，稳定性较差。PGD、BIM真实类目的置信度最低，对抗样本最具鲁棒性且稳定性较强。

表5　无目标攻击时各生成对抗样本算法的ASR、扰动大小和生成时间

Table 5　The ASR, perturbation size and generation time of each algorithm for generating adversarial samples with non-targeted attacks

数据集	生成算法	ResCNN			GRU		
		ASR/%	$\bar{\delta}$	生成时间/s	ASR/%	$\bar{\delta}$	生成时间/s
AISHELL	FGSM	92	0.001 0	51	41	0.001 0	62
	JSMA	100	0.000 4	1 843	100	0.027 6	207 373
	BIM	100	0.000 7	3 058	97	0.000 6	6 873
	C&W(L_2)	100	0.000 0	7 967	97	0.000 2	16 040
	C&W(L_∞)	100	0.001 8	171	93	0.005 2	451
	PGD	100	0.000 7	3 011	97	0.000 6	6 875
	ZOO	100	0.000 3	27 759	99	0.000 2	31 297
	HSJA	100	0.000 4	14 876	100	0.000 9	44 953
LIBRI	FGSM	91	0.001 0	68	49	0.001 0	60
	JSMA	100	0.000 8	3 729	100	0.019 3	413 536
	BIM	100	0.000 7	3 023	80	0.000 6	5 472
	C&W(L_2)	100	0.000 0	7 555	97	0.000 2	13 102
	C&W(L_∞)	100	0.003 0	163	95	0.013 3	328
	PGD	100	0.000 7	2 217	80	0.000 6	5 324
	ZOO	100	0.000 2	43 119	100	0.000 1	44 579
	HSJA	100	0.000 4	13 413	100	0.001 0	20 494

表6　无目标攻击时各生成对抗样本算法的信噪比

Table 6　The SNR of each algorithm for generating adversarial samples with non-targeted attacks　dB

数据集	生成算法	ResCNN			GRU		
		平均值	最大值	最小值	平均值	最大值	最小值
AISHELL	FGSM	28.238 1	38.961 7	18.264 1	28.436 1	38.961 3	18.264 1
	JSMA	12.016 9	21.017 3	0.048 7	−7.735 7	15.461 8	−35.175 2
	BIM	30.512 2	39.842 0	21.367 6	31.286 9	41.667 8	21.122 1
	C&W(L_2)	59.632 2	84.366 2	36.089 9	52.076 8	91.561 1	13.025 4
	C&W(L_∞)	25.048 8	53.749 0	−2.145 0	10.571 6	38.072 2	−13.414 6
	PGD	30.512 0	39.842 0	21.367 6	31.286 9	41.667 8	21.122 1
	ZOO	17.108 4	34.785 2	−8.048 5	21.083 3	63.364 9	−2.087 4
	HSJA	38.193 9	64.963 3	21.574 5	34.295 7	97.795 9	15.202 4
LIBRI	FGSM	35.809 0	44.934 0	26.148 0	35.948 7	50.733 3	26.694 4
	JSMA	16.858 1	27.617 6	4.975 5	−4.987 4	13.974 1	−29.798 7
	BIM	46.783 4	37.813 7	28.541 9	38.822 4	53.624 7	29.678 5
	C&W(L_2)	68.199 7	91.955 7	38.516 2	63.372 4	98.440 8	16.524 8
	C&W(L_∞)	28.337 2	58.454 2	3.165 4	−10.478 2	53.113 2	9.594 7
	PGD	37.813 7	46.783 4	28.541 9	38.814 7	53.624 7	29.678 5
	ZOO	26.495 8	48.525 3	−1.123 0	37.246 4	83.364 4	1.669 1
	HSJA	48.859 1	93.025 6	23.644 1	48.356 3	127.004 5	4.997 8

表7 无目标攻击时各生成对抗样本算法的置信度

Table 7 The confidence of each algorithms for generating adversarial samples with non-targeted attacks

数据集	生成算法	ResCNN			GRU		
		平均值	最大值	最小值	平均值	最大值	最小值
AISHELL	无攻击	0.958 1	0.998 9	0.615 1	0.828 0	0.995 9	0.362 5
	FGSM	0.024 0	0.364 4	0.000 0	0.086 7	0.445 8	0.000 0
	JSMA	0.066 1	0.321 2	0.000 0	0.006 8	0.158 5	0.000 0
	BIM	0.000 2	0.007 4	0.000 0	0.033 7	0.355 5	0.000 0
	C&W(L_2)	0.256 8	0.474 0	0.007 4	0.320 1	0.485 9	0.011 9
	C&W(L_∞)	0.126 8	0.484 9	0.000 0	0.137 6	0.476 4	0.000 0
	PGD	0.000 2	0.007 4	0.000 0	0.033 7	0.355 5	0.000 0
	ZOO	0.289 7	0.464 1	0.100 3	0.257 0	0.454 9	0.035 3
	HSJA	0.168 7	0.376 7	0.010 8	0.259 7	0.474 9	0.013 8
LIBRI	无攻击	0.924 2	0.999 9	0.347 0	0.954 3	0.999 9	0.529 2
	FGSM	0.044 4	0.462 3	0.000 0	0.081 6	0.295 7	0.000 0
	JSMA	0.044 5	0.369 4	0.000 0	0.004 3	0.152 4	0.000 0
	BIM	0.000 0	0.004 3	0.000 0	0.061 1	0.382 1	0.000 0
	C&W(L_2)	0.275 8	0.497 9	0.002 8	0.249 5	0.488 4	0.004 2
	C&W(L_∞)	0.122 4	0.464 5	0.000 0	0.110 9	0.474 5	0.000 0
	PGD	0.000 0	0.004 3	0.000 0	0.059 4	0.382 1	0.000 0
	ZOO	0.292 1	0.459 0	0.053 6	0.265 7	0.475 8	0.076 6
	HSJA	0.209 1	0.446 0	0.035 9	0.253 2	0.447 8	0.099 0

3.5.2 有目标攻击实验结果分析

在有目标攻击的实验中，随机抽取10段不同说话人的音频，每段音频以与该音频的真实标签不同的说话人为目标，生成9个对抗样本。

表8给出了有目标攻击中对抗样本的攻击成功率以及成功对抗样本的平均信噪比、置信度、扰动大小和生成时间。可以看出，JSMA、BIM和PGD的ASR较高，但JSMA的SNR和置信度较低，表现劣于BIM和PGD。在黑盒攻击中，ZOO和HSJA表现较差，但HSJA在信噪比、置信度和扰动三方面优于ZOO。图2给出了对抗样本对目标说话人的置信度的热力图，其中，横坐标Source Speaker表示真实说话人，纵坐标Target Speaker表示目标说话人，置信度从高到低进行分布。

表8 有目标攻击时各生成对抗样本算法的ASR以及平均SNR、置信度、扰动大小和生成时间

Table 8 The ASR and average SNR, confidence, perturbation size and generation time of each algorithm for generating adversarial samples with targeted attacks

数据集	生成算法	ResCNN					GRU				
		ASR/%	SNR/dB	置信度	δ	生成时间/s	ASR/%	SNR/dB	置信度	δ	生成时间/s
AISHELL	FGSM	6.7	26.29	0.95	0.001 0	0.35	4.4	27.09	0.88	0.001 0	0.78
	JSMA	100.0	12.62	0.54	0.000 4	25.16	100.0	7.82	0.46	0.001 6	112.27
	BIM	100.0	30.73	1.00	0.000 7	26.71	53.3	31.07	0.84	0.000 7	60.27
	C&W(L_2)	41.1	29.53	0.80	0.016 2	75.75	25.6	37.99	0.68	0.000 7	152.64
	C&W(L_∞)	26.7	10.64	0.67	0.099 3	7.40	14.4	14.65	0.76	0.009 7	14.96
	PGD	100.0	30.74	1.00	0.000 7	25.24	66.7	30.34	0.95	0.000 7	54.08
	ZOO	30.0	−0.45	0.50	0.003 5	354.59	20.0	4.02	0.59	0.001 0	330.76
	HSJA	11.1	22.58	0.68	0.002 3	5.72	31.1	19.02	0.84	0.004 4	8.00
LIBRI	FGSM	16.7	33.90	0.79	0.001 0	0.41	5.6	33.64	0.79	0.001 0	0.67
	JSMA	100.0	17.75	0.49	0.000 4	18.03	100.0	11.54	0.46	0.002 7	65.48
	BIM	77.8	35.38	0.99	0.000 8	27.11	60.0	35.55	0.96	0.000 7	51.96
	C&W(L_2)	25.6	44.13	0.86	0.001 4	88.10	17.8	39.59	0.71	0.002 8	135.28
	C&W(L_∞)	15.6	31.10	0.79	0.005 4	8.68	12.2	13.29	0.70	0.038 5	14.42
	PGD	76.7	35.42	0.99	0.000 8	32.13	60.0	35.55	0.96	0.000 7	55.40
	ZOO	23.3	6.59	0.53	0.002 9	341.95	15.6	11.64	0.59	0.000 8	328.80
	HSJA	10.0	23.89	0.50	0.007 5	5.68	20.0	32.55	0.79	0.001 3	7.62

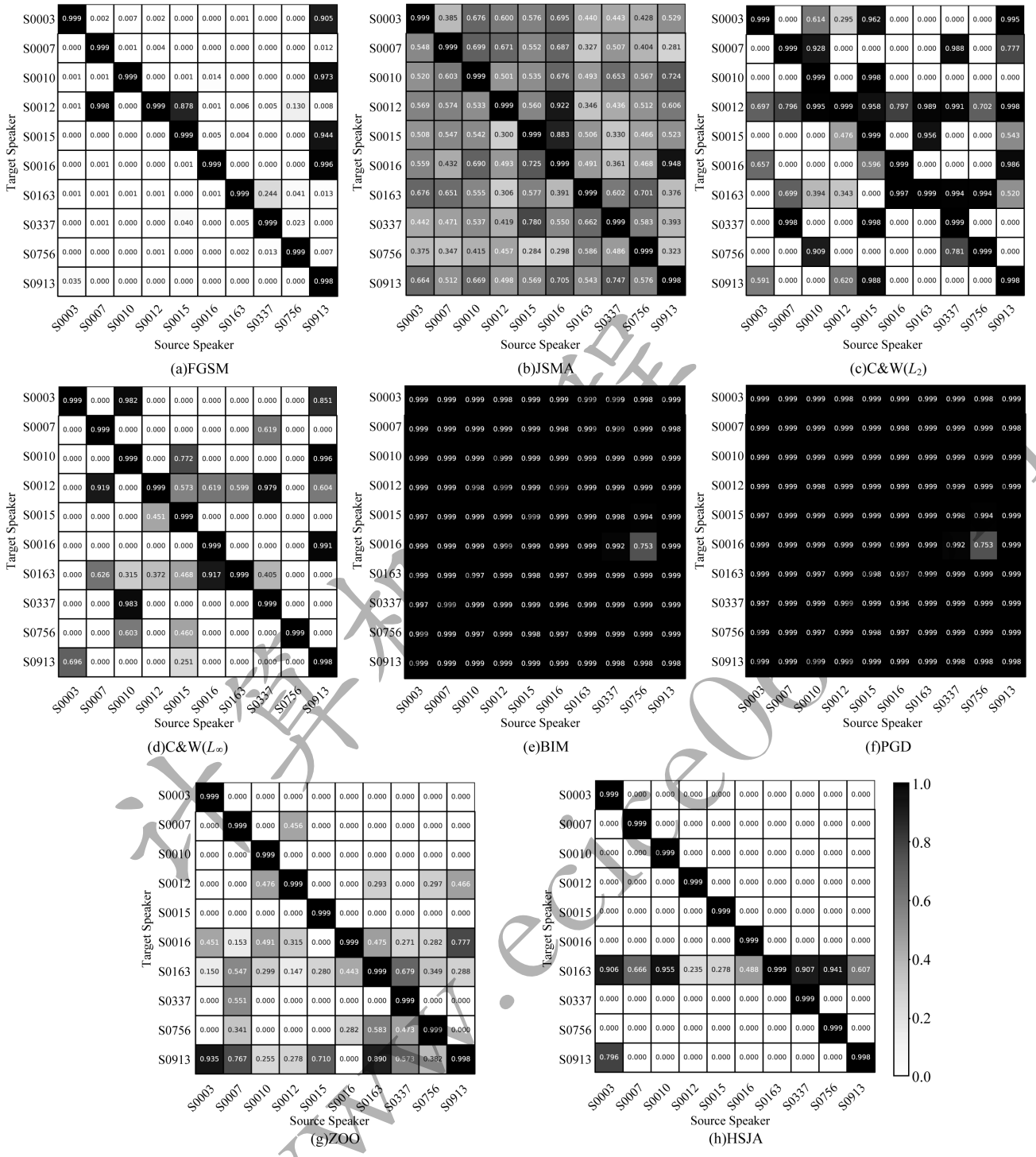

图2　有目标攻击时各算法置信度的矩阵热力图

Fig.2　The matrix heat map of the confidence of each algorithms with target attack

BIM和PGD将10个音频都生成相应目标的鲁棒性对抗样本，表现最优。在ZOO和HSJA的热力图上可以看出，以说话人S0163为目标的不同对抗样本的置信度都较高，推测模型存在部分薄弱的类目，较容易被算法估计出特征。

3.5.3　不同网络结构下的生成算法实验结果分析

在ResCNN和GRU网络结构模型的测试结果中，大部分算法在GRU模型测试的ASR较低、生成时间较长。这表明对GRU模型进行无目标攻击较为困难，其中JSMA的生成难度最大。而ResCNN和GRU网络结构的平均信噪比和真实类目的平均置信度相差不大。在有目标攻击时，其他算法对GRU模型的ASR较低（除了JSMA和HSJA之外），生成时间较长（除ZOO之外）。由此得出，对抗样本生成算法的性能会受端到端说话人辨认系统的网络结构限制，并且生成算法对GRU的攻击效果较差。

3.5.4 不同语种下的生成算法实验结果分析

上述实验结果显示，在相同的网络结构下，JSMA和ZOO在LIBRI英文数据集训练的模型和AISHELL中文数据集训练的模型上的生成时间差异较大，其他指标相近，这可能是因为模型训练差异，而其他算法的各项指标测试结果差异不大。由此得出，各对抗样本生成算法对模型攻击效果受不同语种的影响较小。

3.5.5 隐蔽性测试结果分析

为验证对抗音频与原始音频的区别，本文对30个听众进行3项测试：1)判断每种对抗音频是否为噪声(每种随机抽取1个)；2)确认能否听清对抗音频的内容(每种随机抽取1个)；3)听1对音频(原始音频和相应的对抗音频)，找出对抗音频，属于ABX测试。每项都设置对照组，测试结果见表9，其中，测试结果A表明感觉音频没有噪声的听众比例，测试结果B表明能听清音频内容的听众比例，测试结果C表明能正确找出对抗音频的听众比例。测试1的实验结果表明大部分听众认为JMSA和ZOO的对抗音频有明显的噪声，测试2的实验结果表明听众基本都能听清音频的内容，测试3的实验结果表明ABX测试中BIM、C&W(L_2)和PGD正确找出对抗音频的听众比例接近50%，可以认为其对抗音频与原始音频无法被人耳区分。

表9 隐蔽性测试结果

Table 9 Concealment test results %

生成算法	测试结果A	测试结果B	测试结果C
FGSM	40.00	83.33	26.67
JSMA	6.67	86.67	16.67
BIM	73.33	90.00	43.33
C&W(L_2)	80.00	63.33	46.67
C&W(L_∞)	30.00	73.33	30.00
PGD	70.00	86.67	46.67
ZOO	6.67	93.33	16.67
HSJA	66.67	90.00	36.67
对照组	83.33	90.00	43.33

上述实验结果表明，FGSM、JSMA、BIM、C&W、PGD、ZOO和HSJA这6种生成算法都能生成针对端到端说话人辨认模型识别的对抗样本，实现逃避攻击，但只有BIM、C&W(L_2)、PGD能实现无法被人耳察觉的对抗音频。在无目标攻击时，HSJA黑盒算法能达到白盒攻击的较好水平。在有目标攻击时，BIM和PGD白盒算法面对不同说话人音频都能很好地生成高置信度的目标对抗样本，ZOO和HSJA黑盒算法只能对模型的薄弱目标生成对抗样本，但质量不高，对抗样本生成算法的实现会受网络结构的限制。

4 结束语

为探究语音领域的对抗样本，本文基于端到端说话人辨认系统对现有经典的对抗样本生成算法在音频领域进行实现与比较研究。实验结果表明：在无目标攻击时，各类对抗样本在白盒和黑盒设置下均能逃避说话人辨认系统的识别，在整体性能表现上，BIM和PGD在白盒设置下表现最佳，在黑盒设置下HSJA表现较好；在有目标攻击时，BIM和PGD同样具有很好的性能表现，但在黑盒攻击方面，ZOO和HSJA在有目标攻击时均未能达到其作用在图像数据上的攻击性能表现。由于端到端说话人辨认模型存在安全脆弱性、实验数据局限于较短音频等问题，因此下一阶段将探索更具实际意义的语音对抗样本以及端到端说话人辨认的安全学习机制，提高深度学习模型防御对抗攻击的能力。

参考文献

[1] JUNG J W, HEO H S, YANG I H, et al. A complete end-to-end speaker verification system using deep neural networks: from raw signals to verification result[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 5349-5353.

[2] ZHANG C L, KOISHIDA K. End-to-end text-independent speaker verification with triplet loss on short utterances [EB/OL]. [2020-04-05]. http://m. isca-speech. org/archive/Interspeech_2017/pdfs/1608. PDF.

[3] KINNUNEN T, LI H Z. An overview of text-independent speaker recognition: from features to supervectors[J]. Speech Communication, 2010, 52(1): 12-40.

[4] VILLALBA J, CHEN N X, SNYDER D, et al. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations[J]. Computer Speech & Language, 2020, 60: 101026.

[5] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2015: 1322-1333.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1312. 6199. pdf.

[7] YUAN Xiaoyong, HE Pan, ZHU Qili, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[8] LUO Yuan, WANG Boyu, CHEN Xu. Research progresses of target detection technology based on deep learning[J]. Semiconductor Optoelectronics, 2020, 41(1): 1-10. (in Chinese)
罗元，王薄宇，陈旭. 基于深度学习的目标检测技术的研究综述[J]. 半导体光电，2020，41(1)：1-10.

[9] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verification with adversarial examples [C]// Proceedings of 2018 IEEE International Conference on Acoustics,Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2018: 1962-1966.

[10] LI Xu, ZHONG Jinghua, WU Xixin, et al. Adversarial attacks on GMM I-vector based speaker verification systems [C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2020: 6579-6583.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1412. 6572. pdf.

[12] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings [C]// Proceedings of 2016 IEEE European Symposium on Security and Privacy. Washington D. C., USA: IEEE Press, 2016: 372-387.

[13] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1607. 02533. pdf? source=post_page.

[14] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Washington D. C., USA: IEEE Press, 2017: 39-57.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1706. 06083.

[16] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models [C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York, USA: ACM Press, 2017: 15-26.

[17] CHEN J B, JORDAN M I, WAINWRIGHT M J. HopSkipJumpAttack: a query-efficient decision-based adversarial attack[EB/OL]. [2020-04-05]. https://arxiv. org/abs/1904. 02144v1.

[18] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2574-2582.

[19] CARLINI N, WAGNER D. Audio adversarial examples: targeted attacks on speech-to-text [C]//Proceedings of 2018 IEEE Security and Privacy Workshops. Washington D. C., USA: IEEE Press, 2018: 1-7.

[20] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning [C]// Proceedings of 2017 ACM on Asia Conference on Computer and Communications Security. New York, USA: ACM Press, 2017: 506-519.

[21] LAX P D, TERRELL M S. Calculus with applications [M]. Berlin, Germany: Springer, 2014.

[22] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models [EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1712. 04248. pdf.

[23] LI Chao, MA Xiaokong, JIANG Bing, et al. Deep speaker: an end-to-end neural speaker embedding system [EB/OL]. [2020-04-05]. https://blog. csdn. net/qq_34755941/article/details/109247992.

[24] BU Hui, DU Jiayu, NA Xingyu, et al. AISHELL-1: an open-source mandarin speech corpus and a speech, recognition baseline [EB/OL]. [2020-04-05]. https://arxiv. org/pdf/1709. 05522. pdf.

[25] PANAYOTOV V, CHEN G G, POVEY D, et al. LIBRISPEECH: an ASR corpus based on public domain audio books [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2015: 19-24.

编辑　陆燕菲