2022 年全国大学生信息安全竞赛

作品报告

作品名称： SRS-Guard：面向语音伪造的阻断系统

电子邮箱： 987883395@qq.com

提交日期： 2022-6-20

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目录

摘要

随着信息化时代的蓬勃发展，语音识别等应用正深刻影响着人类社会的方方面面，但随之而来的是严峻的安全挑战，目前语音方面的 deepfake 攻击正严重威胁着相关领域的安全问题，无数用户的语言特征被提取用于合成新音频进而对语言识别产生困扰，在此背景下，本项目提出一种生成对抗样本的基于黑盒攻击的防御手段。

本项目的应用场景如下：针对黑客对目标用户语音进行 deepfake 伪造。我们采取了一种对目标语音加对抗样本干扰的方式对 deepfake 方法进行干扰攻击从而使其无法伪造的方式。本项目计划从某个说话人发出的声音中制作一个对抗性样本，使其具有除说话人以外音频的特征，在听觉上仍然表现为原说话人特征。这个对抗样本可以对常见的 deepfake 攻击模型进行干扰，从而使被处理过的语音无法被 deepfake。

本质上，我们的工作是对要保护的用户语音加上对抗干扰，使其具有其它用户的声纹特质，但是本身人耳却分辨不出区别。然后当攻击者进行伪造时自然就无法伪造出具有要保护用户的特征，使其伪造失败。对应在 SRS 系统上，即攻击者伪造的语音无法通过 SRS 的认证，其伪造效果对比之前未加干扰保护时效果大大下降。

具体做法，我们基于黑盒扰动的对抗样本生成方法对需要进行保护的语音添加噪声从而生成具有干扰性的对抗样本，使 deepfake 无法对该语音进行操纵来达到一些目的。该系统可以在实际生活中防止在说话者不知情的情况下使用其语音进行伪造。由于我们的对抗样本具有高隐蔽性特点，所以攻击者很难做到辨别一段语音是否已经添加了对抗性扰动。除此之外，因为我们采用的是一种黑盒攻击，所以我们可以对多种不同的 deepfake 模型进行防御，无需考虑攻击者的伪造方式，充分发挥对抗样本具有迁移性的特点以及黑盒攻击的特性。

实用性方面，该项目在测试 deepfake 攻击的结果优秀，生成的对抗样本有极高鲁棒性和应用价值；创新性方面，该项目将被攻击方的被动地位为主动，提出了极具前瞻性的攻击抵抗方案，而且攻击者很难识别出目标语音以及被保护，隐蔽性较强。

第一章　作品概述

1.1 背景分析

进入二十一世纪之后计算机技术迎来了长足发展，而计算机技术赖以生存的数据信息也变得十分多样，不仅包括各类语言、数字、图片，还包括各种的视频音频。在此之前，十九世纪，照相技术初步发展，人们对“眼见为不一定实，耳听为虚”的理解只停留在静态的理解上，但实际上人们很难跨越自己对自己亲眼见到的和亲耳听到的事物很难怀疑它是假的。在这种世界文化的背景下，二十一世纪的计算机技术已经能完成许多内容的伪造了，这其中就包括对视频、音频的伪造。

语音承载着人类语言和说话人身份信息,通过语音伪造技术可以精确模仿目标说话人的声音以达到欺骗人或机器听觉的目的。目前,深度伪造(Deepfake)正在对全球的政治经济及社会稳定带来极大的威胁,其中语音伪造是 Deepfake 实现舆论操控的核心技术之一，语音伪造的目的是生成目标说话人的声音，以欺骗人类听觉系统（HAS)或自动说话人验证系统 (ASV)。目前的语音伪造技术主要包括语音模仿语音合成 、语音转换 、重放攻击以及对抗攻击。语音模仿是指通过人类模拟产生目标说话人风格的 语音，属于非自动化的语音伪造手段，其实施难度较高，可模仿的目标说话人范围受限，抗检测能力弱。语音合成是指根据给定的语言内容合成目标说话人风格的语音，实现文本到声音的映射；语音转换是指将源说话人的语音转换为目标说话人风格的语音，实现声音到声音的映射。由于语音合成和转换技术可以实现任意语言内容的目标说话人风格语音伪造，因此是主要的语音伪造手段，在深度伪造中得到广泛使用。重放攻击是指对目标说话人的语音通过设备录制后进行编辑和回放以产生高度逼真的目标说话人语音。重放攻击容易实施，无需特定的专业知识和复杂的设备，因此常用于攻 击 ASV 系统。重放语音与目标说话人的声学特征高度相似，但语音内容的可控性弱。对抗攻击是指 通过对抗样本技术，在语音信号上添加微量扰动实现对 ASV 系统的攻击，该攻击手段是当前的新型攻 击手段，利用了深度神经网络所存在的缺陷 ，在不影响人类听觉感知的情况下，欺骗 ASV 系统使之产生误判。图 1 对以上五类语音伪造技术的攻击对象和可能带来的危 害进行了总结。

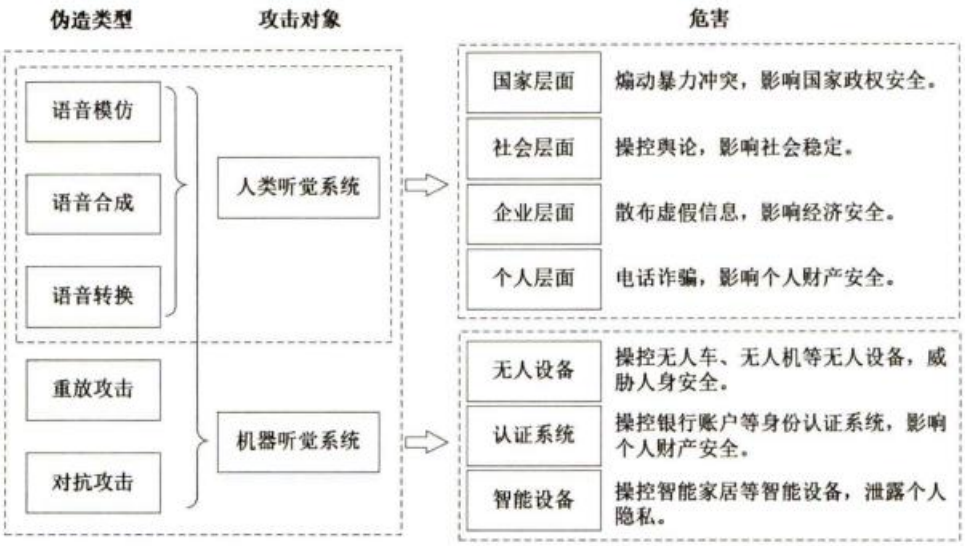


图 1　五种伪造语音攻击类型的攻击对象及危害

近年来语音伪造技术在拟人度、自然度方面有了显著进步,使得语音伪造检测技术面临着更大的挑战。随着深度学习方法在语音识别系统中的广泛应用，尤其是在自动驾驶、身份认证等安全等级较高的应用，语音识别系统的安全问题至关重要。深度学习给语音识别系统带来更便捷的训练步骤、更高的识别准确率的同时，也给系统的安全性带来了潜在风险.最近的研究表明深度神经网络容易受到对输入数据添加细微扰动的对抗攻击，导致模型输出错误的预测结果.当基于深度学习的语音识别系统被外加的细微扰动所攻击，自动驾驶汽车将会被恶意语音攻击执行危险操作，给自动驾驶系统带来了严重的安全隐患，针对语音识别系统的安全性，我们需要一种与这些攻击对抗的方法来使我们的系统更安全。

在此之前我们先从语音合成技术、语音转换、语音对抗样本的基本概念三个方面进行一个简要的介绍，并且主要从传统的技术与现流行的深度学习两方面进行对比介绍。

首先是语音合成技术的长足发展，语音合成将指定的语言文本生成目标说话人声音，实现文本到声音的映射。典型的语音合成系统如图 2 所示,包括前端文本分析和后端语音波形生成两部分。文本分析将输入文本通过规范化、分词、词性标注等步骤生成对应的音素序列、时长预测等信息；语音波形生成根据文本分析生成的语言规范合成目标说话人的语音波形。随着深度学习的发展,语音合成技术逐渐从传统语音合成发展为基于深度学习机制的语音合成。目前基于深度学习的语音合成技术已经逐渐

采用端到端语音合成机制，即将文本分析和波形生成过程连接，直接输人文本或者注音字符,输出语音波形。

图2 语音合成技术基础框架

传统语音合成主要包括波形拼接法和参数生成法。波形拼接语音合成将自然语音数据中的语音单元按照一定的规则拼接，合成与目标说话人高度相似且自然的语音，包括语料库收集、声学单元选取、拼接伪造等步骤。简单的波形拼接语音合成的方法是利用编辑软件直接对音频信号进行裁剪、插人、复制粘贴等修改操作，即复制粘贴篡改。更复杂的拼接方法会调整控制各拼接单元的韵律，以得到更自然流畅的合成语音，代表性工作包括基音同步叠加的PSOLA 技术和利用隐马尔可夫模型(HMM)限制目标单元韵律参数的单元选择系统等。近年来，波形拼接语音合成大多采用深度学习技术，如 2017年 Google提出的基于序列到序列LSTM自编码器的实时单元选择系统等。波形拼接式语音合成适用于某些特定领域，如天气预报、报时、金融业务等。该方法使用真实的语音片段，可以最大限度保留语音音质，可以合成高自然度的语音。然而，其需要大量目标说话人语料，且对于不同领域的文本合成稳定性不强，容易被人或机器识别。基于参数生成的语音合 成通过声学模型预测声学参数，将声学参数通过声 码器合成出目标说话人语音。参数生成语音合成技术的传统代表性工作包括基于H M M的统计参数 合成方法以及基于DNN 的参数合成方法等。参数语音合成可以输出稳定流畅的语音，但受限于参数合成器的缺陷以及统计建模的损失，如生成参数不够平滑、HMM 建模不够准确等问题，合成的语音通常不够自然。

随着深度学习技术的发展，近年来的语音合成技术基本采用深度学习方法，主要包括管道式(Pipeline)语音合成和端到端式语音合成两类。管道式语音合成整体上可分为文本分析、声学模型、声码器三个模块。文本分析模块根据输人文本进行韵律预测和每个音素的时长预测；声学模型建立文本特征和声学特征之间的联系，根据文本分析的输出经由 DNN 映射到声学特征;声码器模块实现声学参数到语音波形的转换。2017 年,百度人工智能实验室使用神经网络模型替换传统参数语音合成的子模块，结合改进后的 WaveNet 声码器提出 DeepVoice 系统。2018 年，百度人工智能实验室进一步提出基于注意力机制的全卷积语音合成系统 Deep Voice3,该系统使用编码器将文本转换为高级特征表示，使用自回归解码器生成梅尔尺度谱图，同时引入非自回归

的后处理转换网络 converter 根据解码器隐状态预测声码器参数，提高了语音合成的速度。管道式语音合成利用深度学习的强大学习能力，一定程度上弥补了传统统计建模式语音合成的不足，但多个模块之间会产生误差累积，且需要高昂成本的文本标注以及文本特征和声学特征的强制对齐，上述问题限制了其语音合成的效果。

接下来我们介绍语音转换的相关技术，语音转换是指将源说话人的语音转换为目标 说话人语音，实现声音到声音的映射。

典型的语音 转换模型如图3 所示，包括语音分析、映射和波形重 构三个主要环节。语音分析将源说话人的语音提 取出中间特征表示（超节段、分段信息），映射模块 将源说话人特征转换为目标说话人特征，重建模块 将目标说话人特征重构成语音波形信号。

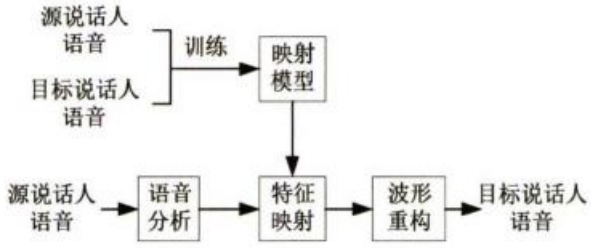


图 3　语音转换技术基础框架

针对所需训练数据形式和数据对齐任务的不同，语音转换可分为平行语料和非平行语料的语音转换两类。平行语料即源说话人和目标说话人的训练数据成对且语音内容相同。基于平行语料的语音转换需要对源和目标说话人语音特征通过时间对齐操作后进行特征映射。帧对齐常用的方法有动态时间规整（DTW)，音素对齐的常用方法有自动语音识别（ASR)等。基于平行数据的统计建模方法包括基于参数统计的高斯混合模型 (GMM）、最小二乘回归、方向核偏最小二乘法 （DKPLS),以及基于非参数统计的非负矩阵分解（NMF)方法等。非平行语料即源说话人和目标说话人的训练数据语言内容不相同，因此在非平行预料的语音转换中，建立源说话人和目标说话人之间的映射更为复杂。基于非平行数据 的统计建模方法包括基于 INCA 算法的对齐技术 ，以及基于音素后映射图 （PPG)方法等。近年来，随着深度学习的发展，借助深度神经网络的学习能力，可以从大量的语音数据集中学习映射关系，提高了语音转换的语音质量和相似度。

传统的语音转换技术通常包括特征提取，声学特征映射和波形重构三个主要环节。

特征提取通过分析语音信号得到声码器参数，包括共振峰频率、共振峰带宽、频谱倾斜等声道谱参数，以及基频、时长、能量等韵律参数。其中 ，声道谱包含 了说话人的个性化特征。经典的特征提取算法有 谐波-噪声模型（HNM ） 、STRAIGHT 模型等。声学特征映射从源说话人声音中提取的说话人语音特征转换成逼近目标说话人的语音特征。其中，韵律参数的转换主要集中在基频包络的转换，而声道谱参数的转换建模相对 复杂，目前仍是制约语音转换质量的瓶颈问题。声道谱参数转换最早由A be 等人提出使用矢量量化码本映射实现，但该方法在量化时会带来特征空间不连续的问题。Chen 等人提出使用 GMM 克服上述问题，但 GMM 存在非一一映射的情况，无法从根本上解决过平滑问题。随着神经网络的发展，Desai 等人提出使用人工神经网络（ANN ）实现源说话人和目标说话人的特征参数映射。上述模型通常需要使用大量的训练数据,且普遍存在过平滑所引入的音质不佳问题。波形合成采用声码器，基于特征映射获得的特征参数合成出接近目标说话人风格的语音。常用的声码器包括传统的参数声码器以及基于神经网络的声码器，经典的参数声码器有STRAIGHT、WORLD 等，基于神经网络的声码器有WaveNetm 、WaveRNN 、 WaveGlow 、ParallelWaveGAN等。传统的语音转换方法只能实现一对一的转换，也没有摆脱对训练数据的强依赖。

近年随着深度学习技术的快速发展，自动编码 器（AutoEncoder）、生成对抗网络（GAN) 等可实现序列到序列高精度转换的神经网络技术在语音转换领域取得了良好的 应用效果。目前主流的语音转换技术都基于深度学习方法，根据转换任务的不同，可分为一对一转 换 、多对多转换和少样本转换。一对一语音转换实现从单个固定源说话人语音到单个固定目标说话人语音的转换。2018 年 ，Donahue 等人提出使用 WaveGAN 来实现语音转换 。2018 年 Kaneko 等人提出一对一转换方法 CycleGAN- C ,使用周期一致的对抗网络 CycleGAN 进行平行数据的语音转换。与逐帧对齐方法相比，CycleGAN-VC 使用具有门控卷积神经网络（ CNN)和身份映射损失的 Cy cleGAN 映射函数，允许在保留语言信息的同时捕获语音顺序和分层结构，该方法打破了需要平行训练数据的限制。多对多语音转换实现从任意源说话人到任意目标说话人 的语音转换，主要实现思路基于语音内容与说话人 风格信息的分离。文献[3-5]均采用 AutoEncod-er，通过编码器将数据的表层特征转换为隐表示，通过解码器从隐层表示中恢复出表层特征，采用解纠 缠机制将语音中的内容和身份信息分开，以实现任意人之间的语音转换。2018 年,Kameoka 等人提出的 StarGAN-VC 采用星形

生成对抗网络实现非平行多对多语音转换。StarGAN-VC 使用对抗性损失训练生成器，确保每对属性域之间的映射保留语言内容信息，在预测时无需任何关于输入语音属性的信息。2019 年 ，SERRA 等人提出基于流模型 (flow)的语音转换技术 Blow ，采用超网络限制的单 尺度归一化流程实现多对多的语音转换。2021 年， Zhang 等人提出了一种可实现多对多非平行语音转换的 TTS-VC 迁移学习框架，该方案利用 TTS 系统将输人文本映射到说话人无关的上下文向量，从而监督基于 AutoEncoder 的语音转换系统潜在表示的训练。2021 年,Lin 等人提出了一种可实现非平行多对多语音转换的端到端系统 FragmentC。该系统从对数梅尔谱图中提取目标说话人频谱特征表示，使用 wav2vec2 .0 提取源说话人风格的潜在 特征表示，通过对两个不同特征空间的隐表示进行训练提取目标说话人的细粒度语音,进而融合到转 换语音中。该方法无需考虑语音内容和说话人身 份特征之间的解纠缠。由于大量可用的说话人语 料难以获取，因此少样本语音转换成为目前的研究 趋势。2019 年 Chou 等人 通过实例归一化技术 进行音色和内容分离，然后再重组音色和内容，可 实现仅一句原始语音和一句目标音色语音就能进 行语音转换。2 019 年，Qian 等人基于 StyleGAN 的思路，提出零样本（Zero -Shot)语音转换方案 Au-toVC ,使用一个具有瓶颈（bottleneck)的 AutoEncoder，在语音自重构质量与说话人身份解纠缠之间进 行瓶颈调谐，以实现非平行数据中的多对多少样本 语音转换。2021 年 ，Zhang 等人 提出基于 GAN 的零样本非平行语音转换方法 GAZEV ，在 GAN 框 架中采用说话人嵌人损失，同时引入自适应实例规 范化策略，解决现有方案中说话人身份转换的局限 性。该方法生成的转换语音在质量和相似度上明 显优于 AutoVC 方法。

最后我们对“对抗攻击”进行介绍，对抗样本是针对深度神经网络所存在的缺陷而产生的一类新型攻击手段，在计算机视觉领域被广泛研究。

在语音领域目前针对 ASR 的对抗攻击的研究较为广泛，代表性工作包括 2018 年 Yuan 等人提出的在音乐中植人命令的 Commandersong 系统,2019 年提出的不易感知且鲁棒的 对抗样本生成算法，以及 2020 年提出 的可攻击黑盒商业设备的 Devil 'swhisper 系统等。近两年，针对 ASV 系统的对抗攻击手段也在逐渐发展，揭露了该类攻击对 A SV 系统存在威胁。ASV 对抗攻击方法的基本原理如图 4 所示，即利用 ASV 深度神经网络的特点，在原始语音信号中添加微小扰动，使得人耳无法察觉的情

况下欺骗目标 ASV 系统产生错误的判决。对抗攻击作为伪造语音攻击中的主动式攻击，对其进行研究有助于发现 ASV 系统的缺陷，提高 ASV 系统鲁棒性。目前针对 ASV 系统对抗攻击的工作可分为白盒攻击和黑盒攻击两类。

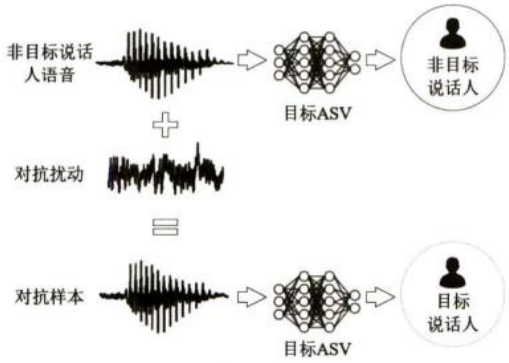


图 4　ASV 对抗攻击示意图

白盒攻击假定攻击者完全掌握目标 ASV 系统的模型结构和参数设置。2020 年，W a ng 等人参考 ASR 对抗样本领域工作，将对抗扰动限制在原始音频的全局掩蔽阈值以下来构建不可感知的 ASV 对抗样本。2020 年 提出了一种快速且通用的对抗样本生成技术，在任意说话人语音中加入音频无关的通用对抗扰动，可欺骗基于 x-vector 的目标 ASV 系统。该方法同时具有良好的鲁棒性， 生成的对抗音频在重放场景下成功率约为 80% 。2021 年提出了一种使用两阶段迭代优化的文本无关的通用对抗扰动生成方法。该方法在重放场景下对目标系统的攻击成功率为 100 %，且语音识别的词错误率（WER)仅增加 3.55 %。虽然白盒攻击可以取得良好的攻击效果，但在实际应用场景中，攻击者很难完全掌握商业 ASV 系统的详细信息，因此相关研究离实际应用还有一定的距离。

黑盒攻击假定攻击者无法获得目标ASV系统的具体模型信息，仅可获取特定输入所产生的输出结果，更接近实际中的攻击场景。2019 年 ，Li u 等 人使用快速梯度符号法 （FGSM ）和投影梯度下降法（PGD ）生成对抗样本,在白盒和黑盒条件下均可攻击基于轻量卷积神经网络（LC-NN)的抗欺骗ASV系统。2 0 2 0年,L i等人研究了GMMi-vector系统和x-vector系统面对对抗攻击的脆弱性，使用针对G M Mi-vector系统生成的对抗样本在 黑盒条件下成功攻击了目标x-vector系统。2020年 ，Villalba 等人使用 FGSM 和 Carlini-Wagne r 算法构造对抗样本，攻击三种基于x-vector的ASV系统。实验证明，即使是使用FGSM 等简单方法生成的对抗样本，也可成功利用其传递性，使小型白盒网络生成的对抗样本攻击大型黑盒 ASV系统。2 020年 ，Zhang 等人基于MI-FGSM 算法提出一种针对黑盒ASV 的攻击策略，采用迭代集成方法 (IEM)

提升对抗样本的可传递性。2020年，Chen 等人提出了一种对抗攻 击方法FAKEBOB ，结合对抗样本的置信度和最大失真,在对抗扰动强度和不可感知性之间取得平衡。实验证明,FAKEBOB 在开源和商业系统上都能达到 99 % 的目标攻击成功率。2020年 ，Das 等 人总结了现有工作的攻击手段、目标系统、数据 集和评价指标。2021年 ，Gomez-Alanis等人提出了一种针对伪造语音检测系统的对抗攻击方法，提出对抗生物特征识别转换网络（ABTN ）联合处理 ASV 系 统和伪造语音检测系统的损失，该方案可在不被 A SV 检测到的情况下对伪造语音检测系统进行白 盒或黑盒攻击。研究表明，该领域现有工作分散在不同数据集、不同攻击和防御手段上，难以进行直接对比，且较多工作集中于不具有实际意义的白盒攻击上，但以上工作的实验均证明，即使是集成了伪造语音检测系统进行辅助判决的ASV系统也容易受到黑盒对抗攻击的影响。

本作品主要应用了黑盒攻击的相关原理，近年来大量实验已经证明,微小的对抗性扰动可以欺骗深度神经网络,使其错误地输出攻击者所指定的目标。目前对抗ASR系统的工作主要集中于白盒攻击,而在黑盒环境下针对语音识别系统生成对抗样本的方法很少。在黑盒环境中模型架构和参数是未知的,这使得生成对抗样本相对困难,但黑盒攻击方法的优势在于其不依赖于模型结构,因此这类方法对语音识别系统具有更大威胁。

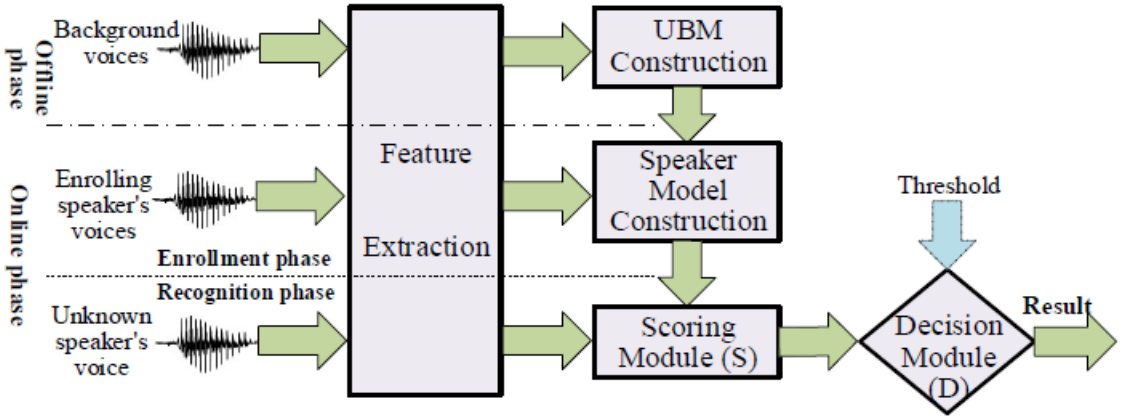
以上这种人工智能合成技术的进步已经显示出了它在创造高度逼真的声音方面的强大能力。这些很容易欺骗人的眼睛和耳朵，因此它给每个人带来了安全和隐私问题。相关技术导致了对隐私和身份验证的担忧。一个人的身份可能会通过语音合成被伪造并以不正当的方式被利用，这只是当今观察到的由深度学习产生的众多deepfake问题之一，例如合成的假照片或假语音。

针对此类活动的检测或防御变得越来越重要。神经网络在存在某些特定噪声的情况下是脆弱的，如果输入人类无法察觉的细微扰动的干扰，则神经网络容易产生不同或不正确的结果。对抗性攻击是产生可以欺骗神经网络的微妙扰动。它在一些判别模型上取得了成功，但在生成模型上的报道较少。在本项目中，我们对语音克隆进行对抗性攻击，以防止一个说话者的语音特征被不恰当地用于语音克隆，其中人类难以察觉的扰动被添加到要防御的说话者产生的话语中。

1.2 相关工作

1.2.1 说话人识别系统（SRS）

说话人识别是一种自动识别技术，它允许机器根据说话人的特点，根据他/她的话语来识别一个人的身份。这一项技术已经经历了长时间的研究，目前已经有一些开源平台和商业平台（例如，微软Azure。亚马逊Alexa[37]、Google home[38]、Talentedsoft 和 SpeechPro VoiceKey[39]）提供相关的应用。图中是一个典型的SRS模型，包括五个关键模块：特征提取、通用背景模型（UBM）构建、说话人模型构建、评分模块和决策模块。上半部分是离线阶段，下半部分是在线阶段，由说话人注册和识别阶段组成。



在离线阶段，使用特征提取模块从背景语音（即语音训练数据集）中提取的声学特征向量来训练 UBM。UBM 旨在创建一个数据集中每个人的平均特征模型，被广泛用于最先进的 SRS 中，以增强鲁棒性和提高效率。在说话人注册阶段，使用 UBM 和每个说话人的注册声音的特征向量建立一个说话人模型。在说话人识别阶段，给定一个输入语音 x，使用说话人模型计算出所有注册说话人的分数 S(x)，这些分数将与决定 D(x)一起作为识别结果发出。特征提取模块将原始语音信号转换成携带信号特征的声学特征向量。各种声学特征提取算法已被提出，如梅尔-频谱系数（MFCC）、频谱子带中心点（SSC）和感知线性预测（PLP）。其中，MFCC 是实践中最流行的一种。

对于说话人识别任务，说话人识别系统有三种常见的识别任务：开放集识别（OSI）、封闭集识别（CSI）和说话人验证（SV）。一个 OSI 系统允许在注册阶段注册多个说话人，在注册阶段，形成一个说话人群体 G。对于一个任意的输入语音 x，系统会确定 x 是否是由一个已注册的说话人说的，还是由一个没有注册的说话人说的。根据所

有注册的扬声器的分数和预设的阈值。形式上，假设说话人组 G 有 n 个发言人{1，2，3……,n}，决策模块输出 D(x):

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}} \, [S(x)]_i, & \text{if } \max\limits_{i \in G} [S(x)]_i \geq \theta; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

其中[S(x)]i 表示为说话人 i 说出的语音 x 的得分，该语音是由说话人 i 发出的。直观地说，系统将输入的声音 x 归类为说话人 i。当且仅当说话人 i 的得分[S(x)]i 是所有说话人中最大的，系统才会将输入的声音 x 归类为说话人 i。说话人 i 是所有注册说话人中最大的一个且不低于阈值。如果最大分值小于阈值系统直接拒绝该声音，即该声音不是由任何一个注册的说话人发出的。

CSI 和 SV 系统完成的任务与 OSI 系统类似，但有一些特殊的设置。CSI 系统从不拒绝任何输入的声音，也就是说，输入的声音总是被归类为登记的说话人之一。而 SV 系统可以有一个注册的说话人，并检查输入的声音是否是由注册的说话人发出的，即接受或拒绝。

对于 SRS 的实现。ivector-PLDA 是在学术界和工业界实现 SRS 的主流方法。它在所有的说话人识别任务中都达到了最先进的性能。另一种是基于 GMM-UBM 的方法，它训练高斯混合模型（GMM）作为 UBM。基本上，GMM-UBM 倾向于在短语料上提供比较高的准确性。

近期深度神经网络（DNN）开始用于语音和说话人语音识别。其中，语音识别的目的是确定语音信号的基础文本或命令。基于 DNN 的方法在语音识别方面取得了重大突破；对于说话人识别，基于 ivector 的方法仍然表现出最先进的性能。此外，基于 DNN 的方法通常依赖于大量的训练数据，与基于 ivector 和 GMM 的方法相比，这可能会大大增加计算的复杂性，因此不适合在客户端设备上进行离线注册。

1.2.3 对抗样本生成模型

我们假设对手打算从某个说话人发出的声音中制作一个对抗性样本，这样它就被攻击的 SRS 归类为注册说话人之一（非目标攻击）或目标说话人（目标攻击），但仍被普通用户识别为原来的说话人。

为了故意攻击目标受害者的认证，我们可以生成对抗性的声音，从 SRS 的角度模仿受害者的声纹。基于此对抗者可以解锁智能手机，登录应用程序，并进行非法金融

交易。例如，我们可以绕过基于语音的访问控制，其中有多个说话人被注册。在绕过认证后，可以发起后续的隐藏语音命令攻击。这些攻击场景实际上是可行的，当受害者不在对抗性声音的可听距离内，或者由于其他语音源（人或扩音器）的存在，攻击声音不会提高受害者的警觉性。

对于实际的黑盒设置，对手只能获得目标 SRS 对每个测试输入的识别结果即决策结果和分数，但不能获得内部配置或训练/注册的声音。这种黑盒设置在实践中是可行的，例如商业系统 Talentedsoft、iFLYTEK、SinoVoice 和 SpeakIn。如果分数不能被访问，可以利用可转移性攻击。我们假设对手有一些目标说话人的声音来建立一个代用模型，而这些声音不一定是注册时的声音。这在实践中也是可行的，因为人们可以录制目标发言人的讲话。

具体来说，在对抗语音生成模型中，考虑了五个参数：攻击类型（有针对性的攻击与无针对性的攻击）、说话人的性别、攻击渠道、说话人识别任务以及目标 SRS 的输出（决策和分数与仅有决策）。API 攻击假定目标 SRS 提供一个 API 接口来查询，而空中攻击是指攻击应该在物理世界中通过空中播放。纯决策攻击是指目标 SRS 只输出决策结果（即对手可以获得决策结果 D(x)），但不包括已注册的发言人的分数。

1.2.4 PyTorch

pytorch 是一个 python 优先的深度学习框架，是一个和 tensorflow、Caffe、MXnet 一样，非常底层的框架。Torch 自称为神经网络界的 Numpy,因为他能将 torch 产生的 tensor 放在 GPU 中加速运算,就像 Numpy 会把 array 放在 CPU 中加速运算。所以神经网络的话, 当然是用 Torch 的 tensor 形式数据最好。

Pytorch 的使用有很多的优势，Pytorch 主推的特性之一就是 Python 优先支持策略，因为直接构建自 Python C API，Pytorch 从细粒度上直接支持 python 的访问。相比于原生 Python 实现，引入的新概念很少，这不仅降低了 python 用户理解的门槛，也能保证代码基本跟原生的 python 实现一致。事实上，开发者可以直接用原生 python 代码扩展 Pytorch 的 operation。Pytorch 有很好的动态图的良好支持，Tensorflow 运行必须提前建好静态计算图，然后通过 feed 和 run 重复执行建好的图。但是 Pytorch 却不需要这么麻烦。PyTorch 的程序可以在执行时动态构建/调整计算图。相对来说，pytorch 具有更好的灵活性。这得益于 Pytorch 直接基于 python C API 构建的 python 接

口。Pytorch 的使用是易于 Debug 的，Pytorch 在运行时可以生成动态图，开发者就可以在堆栈跟踪中看到哪一行代码导致了错误。

1.2.5 神经网络声码器

声码器是语音分析和合成的一种工具，目前主要用来将将声学参数转换成语音波形，即合成。常见的传统声码器有 WORLD、STRAIGHT 及其变种等。还有目前较火的神经网络声码器，如 WaveNet，一种可训练的基于深度神经网络的声码器，可生成高质量的语音波形。

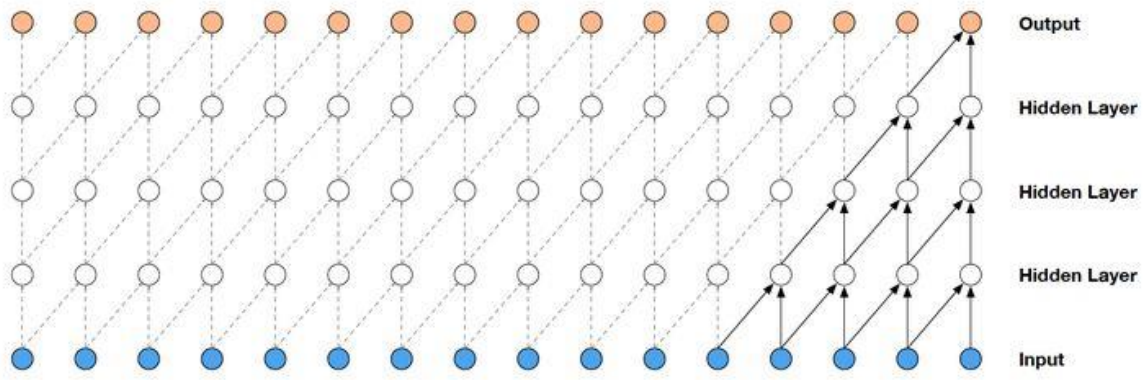
近年来，TTS 技术的发展受益于机器学习的发展，尤其神经网络声码器的提出，使 TTS 的合成质量提升了不止一个档次。虽然目前的神经网络声码器已经获得很大的突破，但或多或少的存在一些问题，如高复杂度或者合成质量不高。目前的声码器主要是把低维度的声学特征进行上采样生成时域波形。从高维度时域波形提取低维度的声学特征，然后通过声码器把声学特征恢复成波形。原始的 speech 预处理后进行 STFT 转换，然后进行 mel 刻度表示，该步骤造成了相位信息的丢失，而且不可逆。声码器的任务就是把 mel 谱等特征上采样恢复时域波形，整个过程都是近似过程，不可能恢复到原始的 speech。声码器按照采用的恢复方案主要分为两类：基于神经网络声码器和相位重构声码器。

Wavenet 模型是一种序列生成模型，可以用于语音生成建模。在语音合成的声学模型建模中，Wavenet 可以直接学习到采样值序列的映射，因此具有很好的合成效果。目前 wavenet 在语音合成声学模型建模，vocoder 方面都有应用，在语音合成领域有很大的潜力。

Wavenet 模型可以根据一个序列的前 t-1 个点预测第 t 个点的结果，因此可以用来预测语音中的采样点数值。基本公式如下：

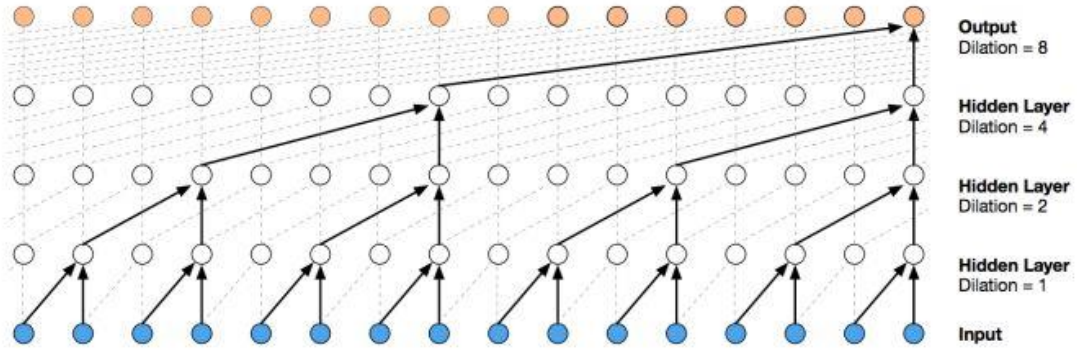
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

在该模型中，我们使用 softmax 层作为输出层，把采样值的预测作为分类任务进行。**Dilated Casual Convolutions：**



Wavenet 模型主要成分是这种卷积网络，每个卷积层都对前一层进行卷积，卷积核越大，层数越多，时域上的感知能力越强，感知范围越大。在生成过程中，每生成一个点，把该点放到输入层最后一个点继续迭代生成即可。

由于语音的采样率高，时域上对感知范围要求大，采用了 Dilated convolutions 这种模型。Dilated convolutions 加入了 dilation 这个概念，根据 dilation 大小选择连接的节点。比如 dilation=1 的时候，第二层只会使用第 t，t - 2，t - 4.....这些点。



Wavenet 在输出层使用了 softmax，求取每个采样点的概率。由于 16 位的采样点就有 65536 种采样结果，所以我们使用律对采样值进行转换。其公式如下：

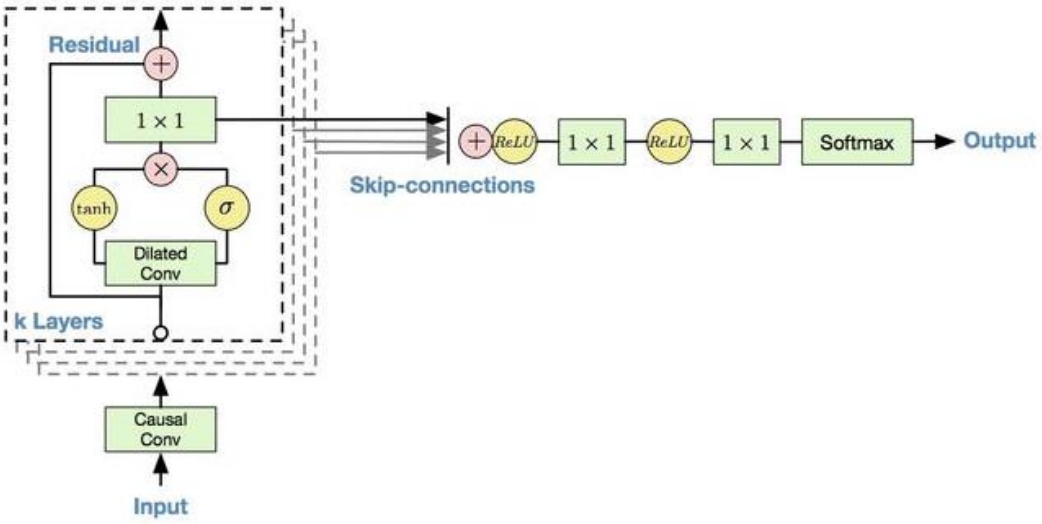
$$f(x_t) = sign(x_t)\frac{ln(1+\mu|x_t|)}{ln(1+\mu)}$$

转换后，65536 个采样值会转换成 256 个值，而且实验证明该转换方法没有对原始音频造成明显损失。在激活函数中使用了门单元:

$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x)$$

隐层中每一层的节点都会把该原来的值和通过激活函数的值相加后传递给下一

层，其中 1x1 的卷积核用来实现降通道数的操作。然后每一个隐层的过激活函数后的结果相加做一系列操作后传给输出层。



加条件特征主要是在激活函数处增加，分为两种形式，global condition 和 local condition。两者公式一致，但 local 的特征需要升采样。

$$z = tanh(W_{f,k} * x + V_{f,k}^T h) \odot \sigma(W_{g,k} * x + V_{g,k}^T h)$$

$$z = tanh(W_{f,k} * x + V_{f,k}^T y) \odot \sigma(W_{g,k} * x + V_{g,k}^T y)$$

升采样有两种方式，第一种是自己学习升采样的模型，可在模型中添加。另一种就是手动升采样，自己将特征复制多次。

1.3 特色描述

（1）扰动难以感知

在语音经过防御系统的处理过后所产生的变化是细微的使用人耳很难感受到。

（2）主动防御

相比于一般检测方法，本方案提出的主动防御从根源上减少了音频伪造身份事故发生的概率，增加一般情况下计算机音频身份识别感知能力。如今市面上的音频身份识别 deepfake 防御重点在于如何识别 deepfake 攻击的音频身份，注重在攻击完成后的末端检测，而本项目重点在于前期对于目标音频的修改，通过对音频添加微噪声，攻击者利用该音频生成的新音频不能通过身份鉴别机制，即攻击失败。

（3）可迁移性

可以对不同target model的具有实用性，对多个deepfake模型进行攻击，其有效性（准确度）也实现了迁移。

1.4 应用前景

Deepfakes 目前的一个重要的攻击目标就是绕过生物识别身份验证，创建内容并用于绕过生物特征验证。当前，面部和语音识别等生物识别技术提供了额外的安全层，可用于根据某人的独特特征自动验证某人的身份。然而，对于可以准确重现一个人外表或声音的 Deepfakes 技术，可以成功规避了这种身份验证技术，这为依赖生物识别特征进行身份和访问管理策略的企业组织带来重大风险。目前，在因为疫情的原因地广泛的远程工作环境中，犯罪分子正在积极发展这一技术。新冠肺炎疫情的大流行和远程工作时代的到来，催生了大量音频和视频数据，这些数据可以输入机器学习系统以创建引人注目的复制品。

Deepfakes 确实对基于生物识别的身份验证构成了显著风险。任何利用身份验证来开展业务，并保护自身免受网络犯罪分子侵害的企业组织都可能受到 Deepfakes 攻击。一些免费的开源应用程序也允许技术知识有限的欺诈者更轻松地生成 Deepfakes 视频和照片。

防御 Deepfakes 网络威胁，无论是通过文本、语音还是视频操作，欺诈者都会投资 Deepfakes 技术来扭曲数字现实以获取非法收益，而且，这种技术正在混乱和不确定的环境因素中蓬勃发展。虽然借助 Deepfakes 技术的网络攻击所构成的威胁看起来很严重，但企业组织仍然可以采取多种措施来抵御它们，所有这些都旨在应对恶意的 Deepfakes 活动。

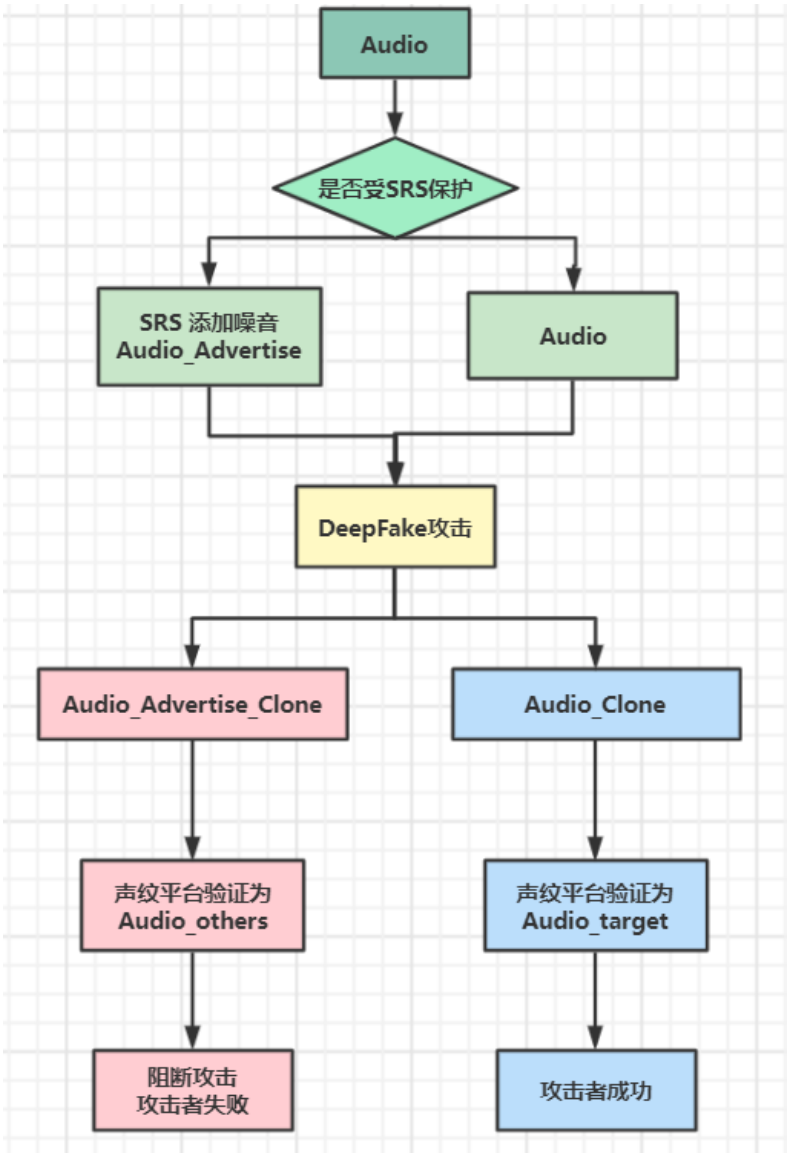
在基于声纹识别技术的应用场景，攻击者可以利用语音克隆进行手机声纹解锁，移动应用声纹登录甚至银行交易声纹验证等，从而对受害者的财产安全，声誉等造成危害。因此我们的防御系统可以进行个人语音隐私的保护，在个人生活层面上，能够在大部分需要语音认证的场景中发挥作用，有效保护公民个人语音隐私信息。同样的，一些攻击者和犯罪分子使用语音克隆模仿政客来发表不真实的声明，这可能会引起地区危机。只需使用一些受害者真实声音的剪辑，所有这些都可以轻松执行。因此，本项目的面向语音伪造的阻断系统是十分重要的。

第二章 作品设计与实现

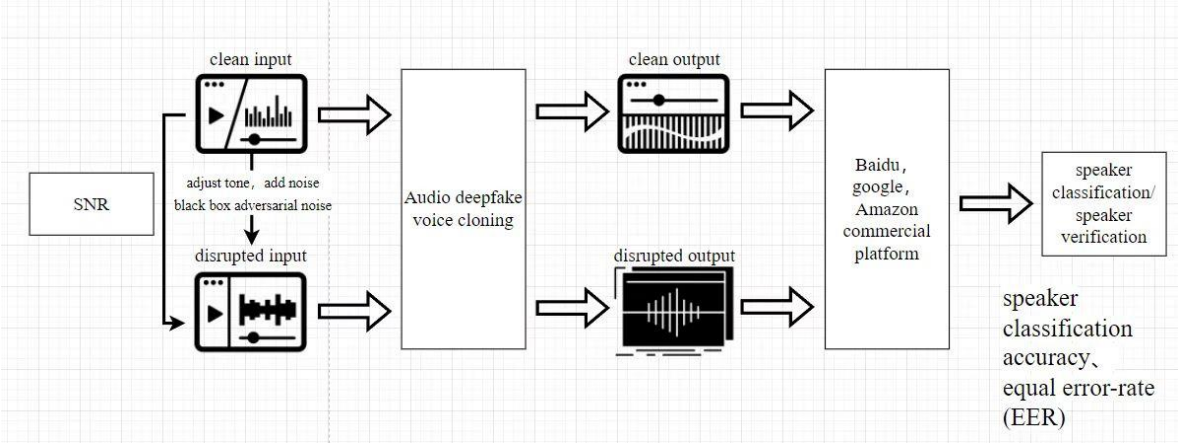
2.1 系统方案

为了针对语音 deepfake 生成进行防御，我们决定采用生成对抗样本的方式，去干扰语音 deepfake 生成模型的正常工作。具体的，我们对目标音频进行加噪声处理，从而生成对抗样本，使其特征发生改变，但是人耳无法识别。但是在投入 deepfake 模型后，则可以明显的干扰到 deepfake 模型的生成过程，从而无法进行声音伪造。

具体 SRS 阻断伪造攻击如下所示：



具体的我们分为三种方法进行实现和横向对比。分别为（1）调整音频音调（2）加背景噪音（3）使用深度模型（黑盒扰动的对抗样本生成方法）加噪音，分别生成对抗生产样本，然后分别将个声音样本送入 deepfake 模型，观察其干扰程度，从而凸显出我们的深度学习模型（黑盒扰动的对抗样本生成方法）的高效性，具体流程如下图：



如图中所示，我们首先将一个 clean input 输入到语音伪造模型中，得到一个 clean output，最后经过 SRS（说话人确认系统）发现可以通过，并且有相应得分，发现可以通过 SRS API。然后我们将 clean input 加上三种噪声，然后再经过相同的流程生成 disrupted output，送入 SRS API 后发现不可以通过 API，从而实现模型攻击。

对于前两种方式，其原理是白盒攻击，即我们得知了语音伪造的基本原理，并且针对这些原理来进行模型攻击。而第三种方案则是具有一定迁移性的黑盒攻击，我们采用 SRS 模型作为 source model 进行对抗样本生成，然后将生成的对抗样本音频直接迁移到 deepfake 模型中，从而干扰生成模型的生成，实现模型攻击。

2.2 实现原理

（1） 调整音调

现实生活中，人在不同情绪下的音调不同，且在不同情境下说话的语调不同。因此对音频音调进行调整实际上是模仿现实场景中的噪音，从而对 deepfake 进行对抗干扰攻击。具体的，我们对音频的声调使用软件进行调整，并且设置多个不同的数值。

（2） 添加背景噪声

同理，现实生活中很可能出现背景噪声，例如背景的车辆鸣笛声，键盘敲击的声

音等等，这些噪声在现实生活中是不可避免的。因此对于想要进行语音模仿的攻击者而言，其很难辨别出背景的噪声是天然的还是人为添加的。而且，环境噪声可以对deepfake 过程进行干扰，根据调研的数据可以证明，不同信噪比的语音 deepfake 的生成效果也不同，因此我们测试了多种噪声和多种不同的信噪比的语音分别对 SRS API 进行攻击，对比出其攻击成功率，选出最好的组合进行。

具体的我们选取了 breathing，footsteps，laughing，mouth-click，keyboard-type 和 clock-tick 六种噪音，分别使用 25，30，35，40，45 五种信噪比进行分别进行测试。统计针对 SRS API 的攻击成功率，最终确定最优的环境噪声组合。

（3） 黑盒攻击

说话人识别(SR)作为一种生物特征认证或识别机制在我们的日常生活中得到了广泛的应用。SR 的流行带来了严重的安全问题，最近的对抗攻击证明了这一点。然而，这种威胁在实际的黑盒场景中的影响仍然是未探索的，因为当前的攻击只考虑白盒场景。

我们对 SR 系统(SRSs)进行了全面和系统的对抗攻击，并且在实际黑盒场景下的利用这些安全弱点，使用黑盒扰动的对抗样本生成方法进行黑盒攻击，来制作对抗样本。具体地说，我们将对抗样本生成作为一个优化问题，结合对抗样本的置信度和最大失真来平衡对抗语音的强度和不可感知性。采用新的算法来估计分数阈值，这是 SRSs 中的一个特征，并将其用于优化问题来解决优化问题。从而实现出当在现实世界中通过空气播放时，次方法在开源和商业系统上也都是有效的，因此具有良好的迁移性。除此之外，人很难区分说话者的原始声音和对抗声音。因此这一方法生成的对抗样本具有隐蔽性。

黑盒扰动的对抗样本生成方法的主要思想是迭代地在一段语音上加入人耳无法感知的扰动，生成对抗语音，从而使得系统将对抗语音识别为来自说话人组中的某个（指定的）说话人。黑盒扰动的对抗样本生成方法的对抗语音生成被建模为一个带约束的优化问题，约束的存在保证添加的扰动不能被人耳感知。

2.3 主要特点包括

1、对三大不同的说话人识别任务，即开集说话人辨认，闭集说话人辨认 (Close-

set Identification, CSI)，说话人确认 (Speaker Verification, SV) 均有效。黑盒扰动的对抗样本生成方法对不同任务采用了不同的损失函数，以适应不同说话人识别任务打分和决策的差异。

2、和图像识别不同，开集说话人辨认以及说话人确认的决策机制基于一个预设的阈值，只有对抗语音的得分超过阈值，攻击才能成功。但在黑盒攻击模型下，攻击者无法提前获得阈值。为了解决这个问题，采用了阈值估计算法，该算法能很好地估计实际阈值，即保证估计阈值大于实际阈值但两者差距很小。

3、与白盒攻击不同，黑盒扰动的对抗样本生成方法不要求攻击者知道系统的内部结构及参数，数据集等，只需要能够访问受害者的说话人模型（即提供输入语音，获取得分及决策结果）。这一黑盒攻击模型比白盒攻击模型更具现实性。根据调研，多数商用声纹识别系统满足黑盒模型。在黑盒攻击模型下，为了能够利用有效的梯度信息进行梯度下降解决上述优化问题，黑盒扰动的对抗样本生成方法使用了基于自然进化策略 (Natural Evolution Strategy, NES) 的梯度估计算法，基于自然进化策略的梯度估计算法比有限差分梯度估计算法更高效。

2.4 方案设计

对于黑盒扰动的对抗样本生成方法的实现，具体地说，我们将对抗样本生成定义为一个优化问题。优化目标由置信参数和噪声振幅最大失真范数来参数化，以平衡对抗声音的强度和不可感知性，而不是使用噪声模型，由于其设备和背景依赖性。我们还将分数阈值(SRSs 中的一个关键特性)纳入优化问题。为了解决优化问题，我们利用了一种有效的梯度估计算法，即自然进化策略(NES)。然而，即使有估计的梯度，现有的基于梯度的白盒方法，都不能直接用于攻击 SRSs。这是由于分数阈值机制，如果预测分数小于阈值，攻击就失败。为此，我们提出了一种估计阈值的新算法，在此基础上，我们利用基本迭代法(BIM)估计梯度来解决优化问题。

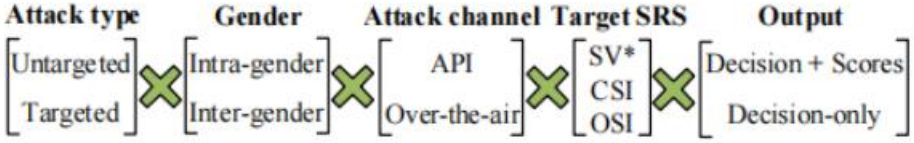
具体做法为：我们假设攻击者打算从某个源说话人发出的声音中制作一个对抗样本，以便被攻击的 SRS 分类为已登记的说话人之一(非目标攻击)或目标说话人(目标攻击)，但仍被普通用户视为源说话人。

为了故意攻击目标受害者的身份验证，我们可以编写对抗音频，从 SRSs 的角度模仿受害者的声纹。合理地说，攻击者可以解锁智能手机，登录应用程序，进行非法

金融交易。在没有目标的攻击下，我们可以操纵声音来模仿任何一个注册的说话人的声纹。例如，我们可以绕过基于语音的访问控制，如 iFLYTEK，其中登记了多个说话人。在绕过认证后，可以发起后续的隐藏语音命令攻击。这些攻击场景实际上是可行的，例如当受害者不在对抗音频的可听范围内，或者攻击声音由于存在其他声源(包括人或扬声器)而不能提高受害者的警惕性。

我们主要关注实用的黑盒设置，在该设置中，对手只能访问每个测试输入的目标 SRS 的识别结果(决策结果和分数)，而不能访问内部配置或训练/注册声音。这种黑盒设置在实际应用中是可行的，如 Talentedsoft、科大讯飞、SinoVoice、SpeakIn 等商用系统。如果分数不可访问，我们可以利用可迁移性攻击。我们假设对手有目标说话人的一些声音来构建代理模型，而这些声音不一定是注册的声音。这在实践中也是可行的，因为人们可以记录目标说话人的音频。有针对性的黑盒设置使所有以前的对抗攻击对 SRSs 不切实际。实际上，所有对 SRSs 的对抗攻击都是白盒，除了并发工作，它只执行无目标的攻击。

具体来说，在我们的攻击模型中，我们考虑了五个参数：攻击类型(有针对性攻击 vs.无针对性攻击)、说话者性别(性别间攻击 vs.性别内攻击)、攻击通道(API vs. 空气传播)、说话者识别任务(OSI vs. CSI vs. SV)和目标 SRS 输出如图：



性别内(性别间)是指来源和目标说话人的性别是相同的(不同的)。API 攻击假设目标 SRS 提供了一个 API 接口来进行查询，而空中传播意味着攻击应该在物理世界中通过空中播放。仅决定攻击是指目标 SRS(如 Microsoft Azure)只输出决策结果(即攻击者可以获得决策结果 $D(x)$，而不输出被登记说话人的分数。因此，有针对性的、跨性别的、空中传播的、只做决定的攻击是最实用的，也是最具挑战性的。综上所述，通过计算图 2 中所有参数的可能组合，有 48（ $= 2 \times 2 \times 2 \times 3 \times 2$）种攻击场景。由于 SV 任务中有针对性的攻击和无针对性的攻击是相同的，因此有 $40 = 48 - 2 \times 2 \times 2$ 种。

2.5 方法原理

假定给一个原始的声音 x，由某个源说话人发出，攻击者旨在制定一个对抗音频 x'=x+b，通过找到一个扰动 b，使 x'是一个有效声音。b 是人类尽可能难以察觉的，受到攻击的 SRS 将声音 x 分类为注册的说话人或目标说话人之一。为了保证对抗声音 x 是一个有效的声音，它依赖于音频文件格式(例如 WAV，MP3 和 AAC)。我们的攻击黑盒扰动的对抗样本生成方法首先将语音 x xx 在每个采样点 i 的振幅值 x(i)归一化到范围[−1,1]，然后构造扰动 b，最后将 x'转换回音频文件格式，并将其提供给目标 SRS。此后，我们设振幅值的范围为 [−1,1]。为了使人尽可能不被察觉，我们的攻击黑盒扰动的对抗样本生成方法采用一定规范来衡量原始声音和对抗声音之间的相似性，并确保距离小于阈值。为了成功地欺骗目标 SRS，我们将为语音 x 找到一个对抗语音 x':

$$\begin{aligned} &argmin_{\delta}\ f(x+\delta) \\ &\text{such that } ||x+\delta, x||_{\infty} < \epsilon \text{ and } x+\delta \in [-1,1]^n \end{aligned}$$

F 为损失函数，要成功地对 OSI 系统发起有针对性的攻击，需要同时满足以下两个条件：目标说话人 t 的得分[S(x)]t 应为所有登记说话人中的最高分，且不小于设定阈值 θ 。因此，目标说话人 t 的损失函数 f 应该定义为：

$$f(x) = max\{(max\{\theta, \max_{i \in G/\{t\}} [S(x)]_i\} - [S(x)]_t), -k\}$$

求解优化问题，我们采用 NES 作为梯度估计技术，并使用带有估计梯度的 BIM 方法来制作对抗样本。具体来说，BIM 方法首先设置 x0'=x，然后进行迭代。

$$x'_i = clip_{x,\epsilon}\{x'_{i-1} - \eta \cdot sign(\nabla_x f(x'_{i-1}))\}$$

我们利用 NES 计算梯度$\nabla_x f(x_{i-1}')$,接下来计算梯度$\nabla_x f(x_{i-1}')$的近似值：

$$\frac{1}{m \times \sigma} \sum_{j=1}^{m} f(\acute{x}_{i-1}^{j}) \times u_j$$

然而，仅使用估计梯度的 BIM 方法不足以在黑盒环境中构建对抗样本，因为攻

击者无法获得损失函数 f 中使用的阈值 θ 。为了解决这一问题，我们提出了一种新的 θ 估计算法。要估阈值 θ，估计的阈值 θ'应该不小于 θ，但也不应该超过 θ 太多，否则，攻击成本可能会变得过高。因此，我们的目标是计算一个小的 θ'

。具体步骤如下：

Algorithm 1 Threshold Estimation Algorithm

Input: The target OSI system with scoring S and decision D modules
An arbitrary voice x such that $D(x) = \texttt{reject}$

Output: Estimated threshold $\acute{\theta}$

$\acute{\theta} \leftarrow \max_{i \in G}[S(x)]_i$; ▷ *initial threshold*
$\Delta \leftarrow |\frac{\acute{\theta}}{10}|$; ▷ *the search step*
$\acute{x} \leftarrow x$;
while True **do**
  $\acute{\theta} \leftarrow \acute{\theta} + \Delta$;
  $f' \leftarrow \lambda x. \max\{\acute{\theta} - \max_{i \in G}[S(x)]_i, -\kappa\}$; ▷ *loss function*
  while True **do**
    $\acute{x} \leftarrow \texttt{clip}_{x,\epsilon}\{\acute{x} - \eta \cdot \texttt{sign}(\nabla_x f'(\acute{x}))\}$; ▷ *craft sample using* f'
    if $D(\acute{x}) \neq \texttt{reject}$ **then**; ▷ $\max_{i \in G}[S(\acute{x})]_i \geq \theta$
      return $\max_{i \in G}[S(\acute{x})]_i$;
    if $\max_{i \in G}[S(\acute{x})]_i \geq \acute{\theta}$ **then break**;

2.6 攻击评估方案

我们基于以下五个方面评估黑盒扰动的对抗样本生成方法的攻击能力：有效性/效率、可迁移性、实用性、不可感知性和鲁棒性。分别如下：

1）Evaluation ：

分别将三种方法的防御成功率（或 deepfake 攻击失败率）进行对比，对比出三种方法的防御效果，对三个方法进行横向对比，突出黑盒攻击的高效性。

具体的成功率，可以借助科大讯飞的声纹识别系统，去衡量对选用数据集的防御效果。

2）Comparison

我们可以从以下几个方面进行对比：有效性，迁移性，鲁棒性，隐蔽性以及实用性。具体的：

有效性可以根据 AUC，EEQ 或 ROC 曲线等衡量指标去衡量每个方法的有效性。

迁移性，可以对比每个方法对于不同 target model 的实用性，例如对多个 deepfake 模型进行攻击，查看其有效性（准确度）是否也实现了迁移，也可以对 deepfake 后的音频换取多种声纹识别平台进行对比。

鲁棒性，可以选取多性别，多人群的数据集，采用多种数据集进行验证。也可以结合论文中的一些常用防御方法，进行综合验证，这部分应该在后续实验中进行探索。

隐蔽性，可以进行人耳识别，对比前后音频的不同。也可以采用信噪比等指标衡量语音的改变程度。

实用性，在之后的实验再想 idea，可以推广一下 over-the-air 方式的语音。

3）Visualization

主要通过表格进行横向对比。配合上多曲线折线图进行对比。

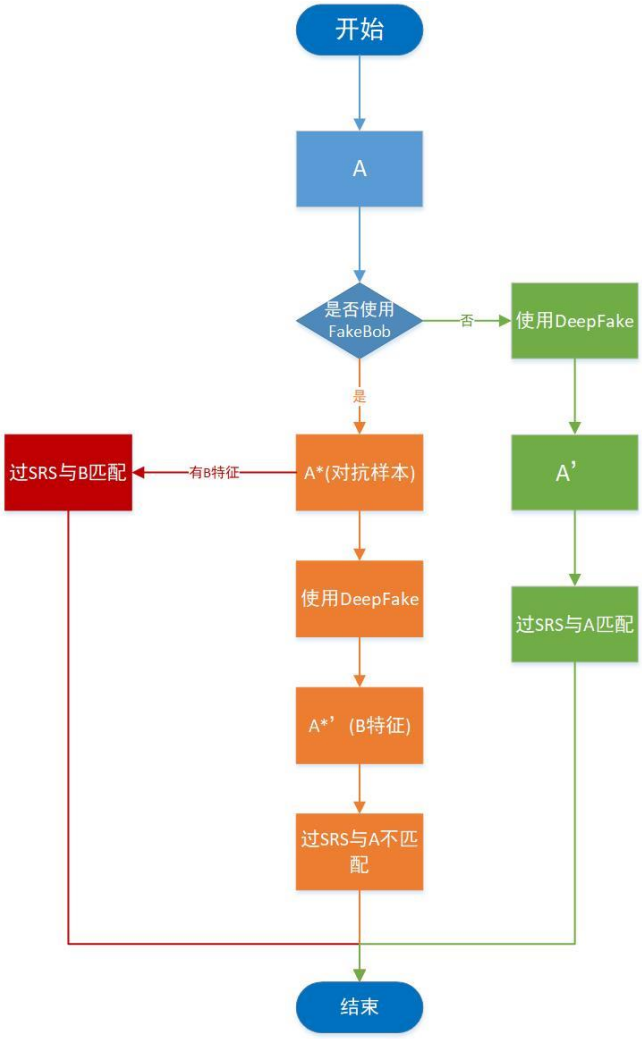
对于实验结果的演示，最好具有当堂录音，生成对抗样本，然后现场延时 deepfake 失败的场景，增强展示冲击力。

4） Transferability

对于 target model，选取多种 deepfake 模型，进行攻击，从而验证体现出我们的对抗样本具有迁移性，阻断具有迁移性。

对于结果验证，我们可以采用除了科大讯飞以外的多种平台，结合人耳识别，从而验证结果。体现出迁移性。

2.7 模式使用流程



将我们的 source voice 记作 A 进行黑盒扰动的对抗样本生成方法生成对抗样本 A*（和 A 人耳听起来无差别），然后让 A*去攻击 deepfake 模型，使其被干扰而生成了 A*'音频（人耳听起来有较大差别），同时也无法被 API 成功确认。这与不使用黑盒扰动的对抗样本生成方法加噪声的音频相比，经 SRS 时成功率有明显的下降。

第三章 作品测试与分析

3.1 测试方案

首先，登记说话人音频身份信息。在科大讯飞声纹识别系统中，注册目标说话人的音频声纹信息，标记目标说话人音频为 A_{target}。

其次，模拟现实的声纹克隆攻击。基于音频 $A_{illegal}$ 进行声纹克隆，指定说话文本，克隆成目标说话人 A_{target} 的声纹特征，得到 A_{clone}。现实场景中，原始陌生音频未进行黑盒扰动，攻击者可以直接克隆音频，使得 A_{clone} 在人耳听感上与 A_{target} 相似，即伪装成已注册的说话人声纹，来表达其他的文本信息，来实现伪装。

接着，主动向攻击音频添加黑盒扰动。为了对基于音频的 Deepfake 进行防御，保证声纹识别系统的安全有效性，我们通过主动添加黑盒扰动的作为防御手段，在上述第二章中已详细阐述，简言之：对原始攻击音频添加黑盒噪音扰动，该扰动无法被人耳觉察，但在音频中留下了“后门”，当攻击者尝试进行音频克隆时，便会触发“后门”机制，隐藏在音频中的黑盒扰动将破坏音频克隆机制，极大程度地破坏克隆效果，使克隆所得音频无法通过声纹识别，从而达到防御。基于 Atarget 原始音频，添加黑盒扰动，得到对抗音频 Aadverse，若攻击者基于 Aadverse 声纹克隆成目标说话人 Atarget 的声纹特征，将得到 Aadvertise_clone。

最后，对比攻击音频黑盒扰动前后的声纹识别结果。添加黑盒噪声之前的克隆音频 Aclone，在人耳上与 Atarget 相似，且能通过声纹识别平台；添加黑盒噪音后的对抗音频 Aadverse，黑盒扰动在人耳上难以觉察，但攻击者对其克隆得到的 Aadvertise_clone，该组音频的克隆效果被严重破坏，无法通过声纹识别测试。

我们将使分别使用 Aclone 和 Aadvertise_clone 对目标说话人声纹 Atarget 进行声纹识别测试，对比二者的识别结果。

3.2 测试准备

对于数据集，我们选用 VCTK(Centre for Speech Technology Voice Cloning Toolkit)，单组说话人相同的声纹特征包含 25 个音频，选用其中的一组声纹 Atarget，作为平台

注册声纹；另一组 Btarget，作为黑盒扰动改变目标的说话人声纹。

对于测试平台，我们选用了科大讯飞声纹识别 Voiceprint Recognition 平台，在平台上首先登记注册目标说话人音频身份信息，对声纹库中的音频进行 1：1 验证匹配，平台将返回一个两位小数的验证得分，当得分大于 0.6 时，可认为声纹特征匹配成功，即通过了声纹测试。

对于模拟克隆，我们选用 Real-Time Voice Cloning SV2TTS 声音克隆，这是一个分三个阶段的深度学习框架。在第一阶段，一个人从几秒钟的音频中创建一个声音的数字编码。在第二和第三阶段，将基于原始音频编码，在给定任意文本的情况下生成其他语音。

对于黑盒噪音扰动，我们采用的思路是，通过多次迭代添加高斯噪音，控制声纹特征的梯度损失函数，将原本 Atarget 的声纹特征扰动改变为 Btarget 的声纹特征（另一个说话人），具体方法在第二章中描述。

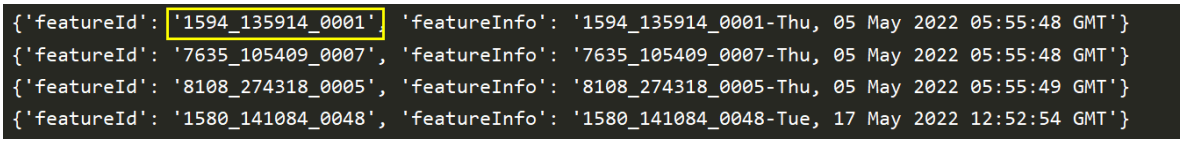
3.3 测试环境

Python3.7

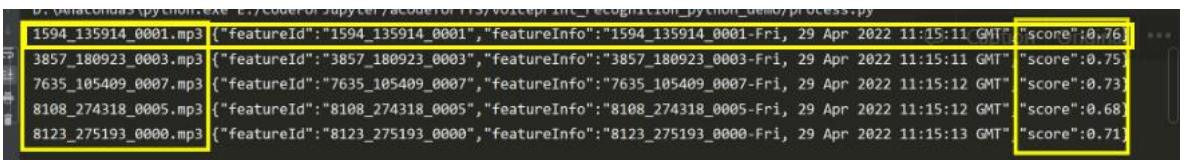
Intel(R) Core(TM) i5-9300H CPU @2.40GHz

3.4 测试过程

注册Atarget目标说话人声纹，标记目标声纹特征为“1594_135914_0001”

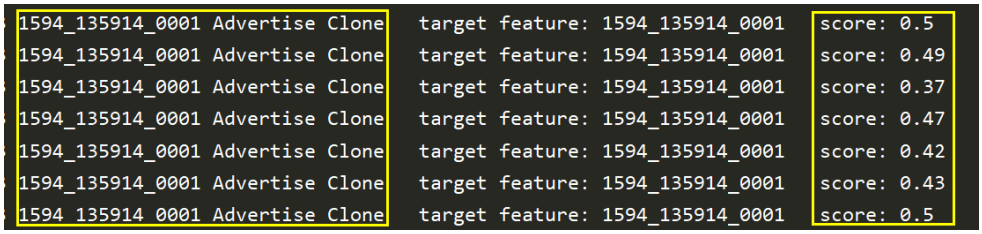


模拟攻击者克隆攻击，基于目标说话人Atarget进行声纹克隆，指定说话文本为："Nice to meet you, this is a synthesizing voice and I would like to test the system."，得到Aclone音频，与已注册的Atarget进行声纹匹配：



对Atarget目标说话人音频，添加黑盒噪音，留下“后门”，得到Aadverse，人耳上

无法觉察与Atarget之间的差别，但蓄意破坏攻击者对Aadverse的声纹克隆效果，与已注册的Atarget进行声纹匹配：



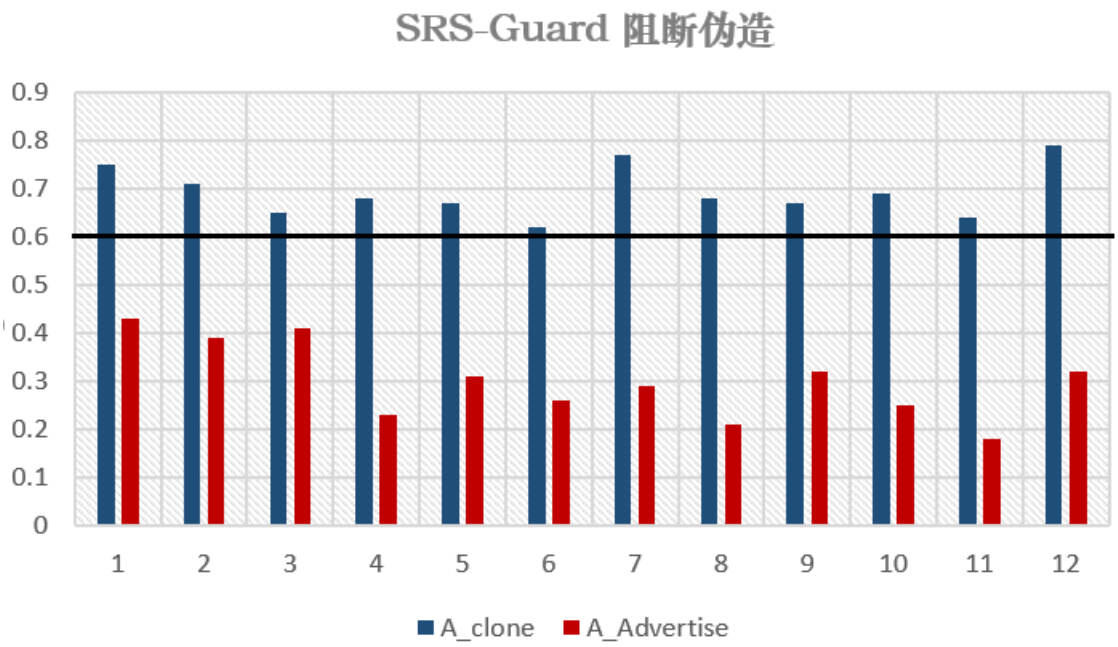
3.5 测试分析

上述测试过程中，基于Atarget进行声纹克隆，得到25条Aclone克隆音频，与已注册的Atarget进行声纹匹配，所得验证得分均大于0.6，平均分数达到0.75，攻击者模拟攻击科大讯飞声纹验证平台成功。

对说话人声纹Atarget的25条原始音频，添加黑盒噪音，暗中改变声纹特征，留下“后门”，得到**25条受保护的Aadvertise对抗音频**，人耳无法觉察Aadvertise和Atarget的区别。

“1594_135914_0001”	A_clone	A_advertise_clone
Atarget – Average Score	0.75	0.43

当攻击者基于“原声”Aadvertise克隆，将触发“后门”机制，得到Aadvertise_clone，该组音频的克隆效果被极大破坏，与已注册的Atarget进行声纹匹配，所得验证得分均小于0.6，平均分数0.43，与受保护前的克隆音频相比，得分大程度减小，无法通过声纹认证，即成功保护声纹识别平台。



对 100 组目标说话人，测试 clone 攻击和添加 advertise 噪音防御前后得分对比，部分测试，如上图柱状图所示，黑色表示攻击得分，红色表示阻断伪造后的得分，均成功将攻击效果减弱，阻断伪造后的得分**均小于 0.6**，攻击者无法通过平台验证，保证了声纹验证的安全性。

第四章 创新性说明

4.1 作品的创新性

本项目构建了国内首个通过生成对抗样本方式抵抗 deepfake 攻击的音频防御系统，自主提出了一种新的防御思路，同时通过多种白盒噪声攻击与黑盒噪声攻击验证该防御措施的有效性与高效性。

（1）减小感知扰动失真

该项目根据 HAS（Human Auditory System）将扰动添加到 Lab 空间，减少了人为因素的主观影响，为实验提供了高精度且具有信服力的实验空间。

该作品的训练音频首先在科大讯飞提供的声纹识别平台生成每位用户独一无二的声纹（同人类指纹唯一一个道理），生成声纹的过程采取同一人多种背景下的音频，将不同背景下的扰动添加到同一 lab 空间，避免因空间唯一而造成的自适宜性降低。

（2）主动防御

相比于一般检测方法，本方案提出的主动防御从根源上减少了音频伪造身份事故发生的概率，增加一般情况下计算机音频身份识别感知能力。

如今市面上的音频身份识别 deepfake 防御重点在于如何识别 deepfake 攻击的音频身份，注重在攻击完成后的末端检测，而本项目重点在于前期对于目标音频的修改，通过对音频添加微噪声，攻击者利用该音频生成的新音频不能通过身份鉴别机制，即攻击失败。

该项目转换了传统攻防领域攻击者和防御者的主动-被动角色，主动防御，为防御领域提出了新的思路。

4.2 作品的实用性

作品的实用性在于对于音频的噪声叠加、身份鉴别、针对语音的 deepfake 防御等方面提供了全新思路与高效措施。不仅减轻了安全从业者的工作压力，提升样本分析的效率，而且有利于企业降低样本分析成本、减小人才培养的成本，获取更大的经济效益。

此外，该项目结合市面上多种先进技术平台（如科大讯飞等），具有与实际结合的优秀落地测试效果。最后，该项目在语音身份认证方面有着得天独厚的天时地利人和，当下的互联网时代蓬勃发展，随之而来的安全危机引起全民抵制的浪潮，其中音频和视频的 deepfake 攻击更是危机中的核心难点，该项目在一定程度上解决了该问题，有一定的切实可行性。

第五章 总结

在这个数字时代，数据正越来越彰显其重要价值，而隐私数据是数据中的最具有价值的一种。随着数据的地位越来越重要，问题也随之而来。当前一些应用开发者通过收集到的海量的数据来进行一些违背道德乃至法律的事情，数据滥用是在收集了海量数据后必然出现的问题之一。deepfake就是数据滥用衍生出的问题。Deepfakes即深度伪造技术，对于社会和个人都是一种不断升级的网络安全威胁。如今诸多网络犯罪分子利用人工智能和机器学习等Deepfakes技术，以创建、合成或操纵包括图像、视频、音频和文本在内的数据内容，进行网络攻击和欺诈。这种技术可以真实地复制或改变一个人的外观、声音，其目的就是进行欺诈，让受害者相信他们所看到、听到或阅读的内容真实可信。

我们基于黑盒扰动的对抗样本生成方法对需要进行保护的语音添加噪声从而生成具有干扰性的对抗样本，使deepfake无法对该语音进行操纵来达到一些目的。该系统可以在实际生活中防止在说话者不知情的情况下使用其语音进行伪造。由于我们的对抗样本具有高隐蔽性特点，所以攻击者很难做到辨别一段语音是否已经添加了对抗性扰动。除此之外，因为我们采用的是一种黑盒攻击，所以我们可以对多种不同的deepfake模型进行防御。

参考文献

[1] 李建华.网络空间威胁情报感知、共享与分析技术综述[J] .网络与信息安全学报，2016, Vol. 2(2): 16-29. （样例，参考国标 GB/T7714-2015）