目录

摘要

基于个体独特的音色信息，说话人识别技术在多个领域得到了广泛应用，包括身份验证、金融交易、安全门禁系统、电话银行、犯罪调查等。这种技术能够通过分析声音信号的频率、音调、语速、音质等特征，来确定说话人的身份。然而，随着说话人识别技术的普及和应用，其安全性问题也逐渐凸显出来。一些攻击者利用对语音输入进行微小的改动就能够欺骗说话人识别系统，导致系统产生错误的输出，这为诸多场景带来了严重的安全隐患。

当前，说话人识别系统中的被动防御技术采用一系列方法，包括量化、平滑、压缩和降低采样率等，对说话人的音频进行处理。然而，这些方法常常会对正常的音频造成严重的失真，从而影响了说话人识别系统的准确性。与此同时，主动防御技术可以确保模型对正常样本的准确率，但却无法有效应对各类对抗攻击。这意味着模型在面对正常音频时可能表现良好，一旦受到对抗样本的干扰，模型的性能就会受到影响，容易产生误判。为了解决以上问题，本文从以下几个方面展开研究：

(1)基于生成对抗网络的被动防御方法研究。本研究基于生成对抗网络技术，提出了一种净化对抗样本的新型防御方法，直接消除对抗扰动。通过将带有扰动的声纹样本直接转化为干净样本，该方法对正常样本产生的影响极为微小。实验结果表明，在闭集识别中，正常样本经过处理过后能达到 99.9%的准确率；在开集识别中，能达到 97.7%的准确率。

(2)基于主动防御和被动防御的组合防御方法研究。本文首先探究了在说话人识别系统中主动防御的效果，通过对抗训练的方式，发现其仅对已训练过的对抗样本具有抵御能力。接着，本文尝试将对抗训练方法和净化方法进行叠加，以期获得更强大的防御效果。研究发现，这种组合防御方

法在提高模型鲁棒性和泛化能力方面取得了积极的效果。然而，本文也注意到，这种组合防御方法并非始终有效，有时会影响对特定攻击的防御效果。

关键词：对抗样本；对抗防御；说话人识别；对抗训练；生成对抗网络

Abstract

Based on the unique timbre information of an individual, speaker recognition technology is widely used in several fields, including identity verification, financial transactions, security access control systems, telephone banking, and criminal investigation. This technology is able to determine the identity of the speaker by analyzing the frequency, pitch, speech rate, tone quality and other characteristics of the voice signal. However, with the popularization and application of speaker identification technology, its security issues have gradually come to the fore. Some attackers are able to deceive the speaker recognition system by making minor changes to the voice input, resulting in the system producing incorrect outputs, which brings serious security risks to many scenarios.

Currently, passive defense techniques in speaker recognition systems mainly use a range of methods, including quantization, smoothing, compression, and sample rate reduction, to process the speaker's audio. However, these methods often cause severe distortion to normal audio, which affects the accuracy of the speaker recognition system. Meanwhile, active defense techniques can ensure the model's accuracy against normal samples, but they cannot effectively counter various types of adversarial attacks. This means that the model may perform well in the face of normal audio, and once it is interfered by adversarial examples, the performance of the model will be affected and prone to misjudgment. In order to solve the above problems, this paper investigates the following aspects:

(1) Research on passive defense method based on generative adversarial network. Based on generative adversarial network technology, this study proposes a novel defense method that purifies the adversarial samples and directly eliminates the adversarial perturbations. By directly transforming the

acoustic examples with perturbations into clean examples, the method produces extremely small impact on normal examples. The experimental results show that in closed-set identification, the normal examples can reach 99.9% accuracy after processing; in open-set identification, it can reach 97.7% accuracy.

(2) Research on combined defense methods based on active and passive defense. This paper first explore the effect of active defense in speaker recognition system by means of adversarial training, and find that it only has the ability to resist the trained adversarial examples. Then, this paper tries to superimpose the adversarial training method and the purification method, in order to obtain a stronger defense effect. It is found that this combined defense approach achieves positive results in improving model robustness and generalization ability. However, this paper also notes that this combined defense approach is not always effective and sometimes affects the effectiveness of defense against specific attacks.

Key Words: Adversarial Example, Adversarial Defense, Speaker Recognition, Adversarial Train, Generative Adversarial Nets

第一章 绪论

1.1 研究背景及意义

随着科技的迅猛发展，智能设备的普及程度逐渐加深，人们的生活方式也发生了翻天覆地的变革。语音已经成为人类与外界进行互动的重要媒介之一，与文本聊天一样，语音聊天已经成为日常生活中不可或缺的一部分。当前，语音技术不断进步，涵盖多个领域，包括说话人识别[1]、情感识别[2]、语音合成[3]等。这些技术的应用使得智能设备能够提供多样化的语音服务，例如智能语音助手[4]、电话交互[5]、语音搜索[6]、语音导航[7]、智能家居控制[8]等，从而为用户提供更智能、便利的体验。

说话人识别，也被称为声纹识别，是通过分析语音信号的具体特征确定不同身份说话人的一项生物特征识别技术。众所周知，每个人声音的音色都是独特而复杂的，是多个生理和环境因素相互作用的产物。当讲话者故意模仿他人声音和语调时，即使模仿得惟妙惟肖，其声纹却始终不变。因此，声纹识别被广泛应用于多个场景，如银行业务的身份验证[9]、司法认证[10]，以及智能设备上的个性化服务[11]等。

近年来神经网络模型应用广阔，其在多个领域的成功应用将改变人们生活的方方面面，例如微信的声纹解锁功能，对着微信录一段用户语音，做一把声音锁，可以直接用用户声音登录微信。然而，随着应用范围的扩大，对抗样本攻击[12]也成为一个不可忽视的隐患。对抗样本通过故意引入微小的扰动来欺骗神经网络模型，导致其错误分类。这种对抗攻击方法可能威胁到模型的鲁棒性和安全性，尤其在安全敏感领域，如金融、医疗和国防。因此，神经网络模型的发展需要更加注重对抗样本的防御机制，以确保其在实际应用中的可靠性和安全性。

如今，图像领域的安全性已受到广泛关注和研究。然而，说话人识别系统的安全问题尤为突出，抵御对抗样本方法的探索还有所欠缺。特别是当说话人识别系统用于提供与金融有关的服务或涉及个人隐私时，缺乏充分的安全保障将对个人的财产和名誉造成严重损害。因此，本文将研究说话人识别系统的防御方法，旨在寻找创新且高效的对抗样本防御策略，确保模型面对精心设计的对抗样本时仍然保持准确性和可靠性。这一研究的成果有望为说话人识别系统的安全性提供前沿的保障，推动说话人识别技术的可靠应用。

1.2 国内外研究现状

1.2.1 说话人识别系统研究现状

说话人识别技术经历从传统的模版匹配到基于统计学方法，再到基于深度神经网络的方法的转变。早期的说话人识别技术可以追溯到 20 世纪 60 年代，当时的方法主要依赖于声学特征提取和模板匹配的传统方法。尽管早期的声纹识别在一定程度上能够识别说话人，但也面临一些挑战。首先，它们对噪声和环境变化非常敏感，这可能导致识别准确性下降。其次，在大规模和真实世界的应用中，因为需要处理大量的音频数据，这对系统的计算资源和处理速度提出了挑战。

随着统计学方法的兴起，特别是引入高斯混合模型(Gaussian Mixture Model, GMM)[13]，声纹识别进入新的阶段。理论上，GMM 能够模拟任何数据的分布，从而也能模拟任何个体的声音。然而，GMM 模型的缺陷也很明显，模型越大，所表示的声音数量就越多，需要训练和调整的参数也越多。因此，为了构建更通用的、具有泛化能力的 GMM 模型，需要更庞大的数据集。不过，很多时候数据集并不能满足训练要求，导致模型的准确率受到限制。

为了解决目标用户数据不够的问题，Reynolds 等人[14]提出通用背景模型(Universal Background Model, UBM)的概念，即通过收集大量的非目标用户的声音作为背景数据，放入 GMM 模型中进行训练，使得模型能够学习到通用的特征表达。其实可以把这种模型看作是一个先验模型，或者是一个预训练模型[15]，训练目标用户数据时直接在 UBM 中进行参数的微调，就能实现很好的性能。

在实际应用中，说话人识别系统面临着说话人身份信息和各种干扰信息混杂在一起的问题，同时不同设备的收音效果不同，产生的信道干扰不同，会导致语音的质量也有所差异，进而影响系统的识别准确率。传统的 GMM-UBM 方法不能克服这一问题，导致系统性能不稳定。为了解决信道干扰的问题，Kenny 等人[16]提出联合因子分析(Joint Factor Analysis, JFA)，即对说话人身份信息和信道信息分别进行建模，从而提高语音识别的准确性和稳定性。然而，Dehak 等人[17]发现 JFA 不能很好地分离说话人身份信息和信道信息，信道因子中也会包含部分说话人信息。因此，提出基于 i-vector 的方法，将说话人差异空间和信道差异空间作为一个整体进行建模。i-vector 方法是从 GMM 的均值超向量中提取一个更紧凑的向量来代表说话人的向量。因为向量中包含说话人差异和信道差异信息，所以可以使用信道补偿技术(Probabilistic Linear Discriminant Analysis,

PLDA)[18]来消除信道干扰，将向量中的说话人差异信息和信道差异信息分开。

后来引入深度神经网络模型(Deep Neural Network, DNN)提高说话人识别系统的准确率，其中比较经典的是 d-vector[19]和 x-vector[20]。d-vector 模型训练一个前馈 DNN 用于说话人的分类，使用最后一个隐藏层的输出作为说话人的向量，实验结果显示，d-vector 系统的整体性能不如 i-vector 系统。x-vector 模型是对 d-vector 模型的改进，可与 i-vector 相提并论。然而，在数据集充足的情况下，x-vector 的效果要好于 i-vector。x-vector 模型直接处理 MFCC 特征，在 segment-level 层中提取向量，用来代替 i-vector 中的说话人向量，并作为 PLDA 处理的输入。

1.2.2 对抗样本研究现状

He 等人[21]在图像识别任务中首次利用残差网络，识别准确率显著提高。Szegedy 等人[22]在图像识别领域首次提出对抗样本的概念，他们发现深度神经网络在学习输入和输出之间存在一定程度的不连续性。通过向图像中添加难以察觉的扰动，可以导致神经网络分类错误，甚至可以任意改变神经网络的预测结果。此外，他们还在实验中发现，对抗样本具有迁移性，即针对一个模型生成的对抗样本，也可以对其他模型产生一定的攻击效果。对抗样本的存在将导致图像识别模型性能下降，增加图像识别系统的安全风险。接下来将对图像领域中常见的对抗样本攻击方法进行介绍。

Goodfellow 等人[23]提出快速梯度符号攻击(Fast Gradient Sign Method, FGSM)。FGSM 攻击是一种单步攻击，依赖于神经网络的线性假设，仅通过计算一次梯度就能生成对抗样本。由于梯度计算的次数少，FGSM 攻击的速度快，对抗图像的视觉效果差，攻击的成功率低。

Papernot 等人[24]提出基于雅可比的显著性图攻击(Jacobian-based Saliency Map Attack, JSMA)。JSMA 攻击通过学习输入到输出之间的映射关系，利用雅可比矩阵找出对神经网络误分类影响最大的像素，从而生成具有良好视觉效果的对抗样本。

Kurakin 等人[25]提出基于迭代的快速梯度符号攻击(Iterative Fast Gradient Sign Method, I-FGSM)。I-FGSM 是一种多步攻击，通过多次计算梯度生成对抗样本，但是梯度计算的过程中容易陷入局部最优，降低了对抗样本的迁移性。

Carlini 等人[26]提出 C&W 攻击(Carlini and Wagner, C&W)。C&W 是一种基于优化的攻击方式，根据三种不同范数约束扰动大小，生成对抗样本的效率较低，但对抗样本中添加的扰动几乎不可察觉。

Madry 等人[27]提出投影梯度下降攻击(Projected Gradient Descent, PGD)。PGD 攻击是 I-FGSM 攻击的变体，采用随机初始化和增加迭代次数的方法，每次迭代都会把梯度投影到指定范围内，被业内人士认为是最强大的一阶攻击。

Dong 等人[28]提出基于动量迭代的快速梯度符号攻击(Momentum Iterative Fast Gradient Sign Method, MI-FGSM)。MI-FGSM 在 I-FGSM 攻击的基础上加入动量。通过引入动量，可以在迭代过程中不断积累上一次迭代的梯度值，以帮助梯度跨越局部最优解，从而生成更完美的对抗样本。

Kong 等人[29]提出基于生成对抗网络(Generative Adversarial Nets, GAN)的 PhysGAN 方法。与以往只应用于数字世界不同，PhysGAN 生成的图像对抗样本应用于真实世界中，这些对抗样本具有极强的干扰能力。在自动驾驶汽车连续前行的过程中，不同时刻、不同角度观察到的对抗图像标志都能导致汽车产生错误的方向判断。

之前的研究主要集中在基于图像的对抗攻击，随着科学技术的发展，情感识别[30]、说话人识别[31]以及语音识别[32]等相关语音技术也逐渐应用于人们的日常生活，基于深度学习语音系统的攻击也受到越来越多的关注。接下来将对语音领域中常见的对抗样本攻击进行介绍。

Li 等人[33]提出对抗转换网络攻击（Adversarial Transformation Network, ATN）。攻击者通过创建并训练独立的网络模型，只需要输入原始音频就能得到对抗音频，从而达到欺骗说话人识别模型的目的。此外，Li 等人还提出一种对语音识别进行攻击的方法，此方法权衡攻击的成功率和语音质量，可以达到 99.2%的句子错误率。

Yuan 等人[34]针对现实世界的语音识别系统，提出 CommanderSong 的攻击。此方法将一些恶意的命令植入到歌曲中，在真实世界通过扬声器进行传播，而不被人类听到。为了确保对抗音频在真实世界中有效传播，Yuan 等人找出不同扬声器产生的电子噪声，并构建了通用的噪声模型用于生成对抗样本。

Zhang 等人[35]设计了海豚攻击（DolphinAttack）。在这项工作中，他们利用音频硬件的漏洞，即现在大多数移动设备上的麦克风允许信号超过 20 kHz，以超声波为载体制作恶意语音命令，从而实现在真实世界中人类完全听不见的攻击。但是此方法生成的扰动容易被低通滤波器过滤，导致攻击失败。

Neekhara 等人[36]证明存在通用的对抗音频扰动，即对任何语音信号添加某些固定的扰动都会导致语音识别模型错误分类。通过对音频数据的样本子集进行遍历，他们逐步形成一个通用的扰动。如果扰动未能达到预期的准确率，则继续遍历，直至达到要求。

在 DeepSpeech 模型中，验证集的攻击成功率可达到 89.06%。这项研究揭示了语音识别系统面临的安全风险，强调了对抗攻击具有普适性。

Chen 等人[37]提出 FAKEBOB 黑盒攻击，将对抗样本的生成转化为一个优化问题。在这个优化问题中，考虑对抗音频的置信度和最大失真，平衡对抗音频的扰动强度和不可感知性，并利用自然进化算法来估计梯度。研究结果表明，FAKEBOB 在开源系统和商业系统上都实现 99%的有目标攻击成功率，并对攻击的实用性进行验证，当生成的对抗音频在真实世界中播放时，对商业上说话人识别系统也具有攻击效果。

Zhang 等人[38]对 PGD 攻击进行改进，提出 ADA(A Highly Stealthy Adaptive Decay Attack)攻击。ADA 攻击的主要思想是在 PGD 攻击中加入一种动态衰减扰动强度的方法，以增加对抗样本的隐蔽性。具体来说，ADA 攻击使用余弦退火的方法来衰减扰动强度，这意味着在多次迭代的过程中，扰动强度会不断减小，以达到在不影响攻击成功率的情况下，对抗样本具有更高的隐蔽性。

Xie 等人[39]针对说话人识别系统提出一种实时、通用的对抗攻击方法。通过向任意注册说话人的语音输入添加一个与音频无关的通用扰动，使得说话人识别系统将说话人识别为任何目标。此外，由于空气传播而引起的声音失真，他们通过建模估计房间冲激响应，提高了攻击的稳定性。他们的实验使用一个包含 109 名说话人的公共数据集，证明了他们提出攻击方法的有效性，攻击成功率超过 90％。

Deng 等人[40]提出一种基于局部低频扰动的决策攻击。具体来说，与对整个音频样本施加扰动不同，他们寻找一个局部攻击区域，将扰动限制在一个局部区域内。其次，考虑到大多数能量集中在低频段，所提出的方法在低频域进行扰动生成。实验结果表明，在说话人识别系统中，他们的方法能够以更高的攻击成功率实现有目标攻击。

Zhang 等人[41]研究对说话人识别系统进行黑盒攻击的方法，提出一种名为 NMI-FGSM-Tri 的攻击。与 FGSM 攻击和 PGD 攻击设置固定步长不同，他们的方法在攻击时，计算梯度设置自适应步长，避免陷入局部最优解。在开集识别和闭集识别任务中，攻击成功率分别到达 97.8%和 61.9%。此外，生成的对抗音频具有高信噪比，达到 48 dB，与原始音频非常相似。

Yao 等人[42]提出一种名为 STA-MDCT 的通用框架，通过修改离散余弦变换（MDCT）来提高对抗音频在黑盒模型中的攻击能力。他们的方法在时域和频域中修改音频，改善攻击的可迁移性和感知性。此外，他们使用模型集成的方法，并通过类激活图可视化对抗音频的分布，为基于迁移的攻击提供可解释的依据。实验结果表明，黑盒攻击中有目

标攻击成功率可达 70.5%。

1.2.3 对抗防御研究现状

目前，对抗样本的防御方法主要分为两类：主动防御和被动防御。主动防御采用对抗训练的方式重新训练说话人识别模型，以使其在面对对抗样本时具备更强的鲁棒性。被动防御则通过添加新的组件，以抵御对抗样本。在被动防御中，可以根据新组件的功能进一步细分为检测方法和净化方法。检测方法只需要检测出对抗样本，并把对抗样本进行剔除；净化方法需要消除对抗样本中存在的扰动，进而使对抗样本变成良性样本。接下来将进行对抗防御研究的相关介绍。

Pal 等人[43]提出一种基于混合对抗训练（Hybrid Adversarial Training, HAT）的防御机制，通过利用有监督和无监督的方法来生成更多样化和更强大的对抗样本，以增强模型的鲁棒性。他们的实验结果表明，相对于使用 PGD 攻击的对抗训练，HAT 对模型的泛化能力影响极小，并且能抵御 PGD 和 C&W 攻击。

SUN 等人[44]使用对抗训练的方法来进行对抗样本的防御。他们使用 MFCC 特征作为输入，使用 FGSM 生成对抗样本，并选取每个小批次的数据，用对抗样本和正确标签制作用于对抗训练的数据，动态的将 FGSM 生成的对抗数据放入模型的训练集中，增强模型的分类能力。而且，他们使用教师-学生（T/S）训练方法[45]提高模型的鲁棒性。

Du 等人[46]使用对抗训练、音频下采样和平均滤波三种方法进行对抗样本的防御。他们认为对抗训练有很大的局限性，只能对固定参数的攻击进行有效的防御，如果攻击者增大扰动强度或者换另一种攻击方法，对抗训练可能会达不到良好的防御效果。在他们的实验中，音频下采样和平均滤波都可以降低攻击的成功率，达到一定的净化效果，但根据 Nyquist 采样定理，当采样率低于原始音频最高频率的两倍时，音频下采样会造成音频的失真。同样，平均滤波也会降低音频质量，从而降低模型对正常样本的准确率。

Yuan 等人[47]使用两种方法进行对抗样本的检测，分别是添加噪声的检测和降低采样率的检测。他们在实验结果中发现背景噪声降低对抗样本攻击的成功率，反而对原始音频的影响很小，因此他们提出添加噪声的防御方法。具体的方法是：在输入音频中添加噪声，如果识别模型对添加噪声后的音频输出结果和原始音频的输出结果不一致，则认为输入音频中存在对抗扰动。降低采样率的检测的方法是：降低采样率，得到音质低的音频，如果原始音频与低音质音频的识别结果不一致，则认为输入音频中存在对抗扰动。

Rajaratnam 等人[48]使用带通滤波器以及 AAC[49]、MP3[50]、OPUS[51]和 SPEEX[52]的音频压缩方法进行对抗样本防御的研究。他们发现 AAC 和 MP3 的防御方法的净化效果在所有音频压缩方法中表现最糟糕，使用带通滤波器的方法比音频压缩方法的净化效果好，但是对正常样本的影响要大于 OPUS 和 SPEEX 的音频压缩方法。

Zeng 等人[53]把多版本编程(Multi-version programming, MVP)[54]的主要思想应用于对抗样本的检测。他们基于对抗样本迁移性[55]的强弱，以及不同自动语音识别系统对正常语音的输出具有相同的输出结果，在不同自动语音识别系统中两两比较输出的相似度，对于相似度低于设定阈值的样本认为是对抗样本。

Yang 等人[56]提出利用时间依赖性的方法在说话人识别系统中进行对抗样本的检测。其主要流程就是对于某个音频样本，截取其中部分语音片段，放入识别模型中，进行输出结果的比较。如果是对抗样本，添加的扰动在不同的时间片段上分布的不均匀，则输出结果会有很大的差异。如果是良性样本，则不会受到影响，所有的输出结果一致。此外，他们还使用量化、平滑、下采样和自编码（Autoencoder）的净化方法进行实验。他们发现 Magnet 编码器[57]在图像领域能够有效抵御对抗样本，但在音频方面的防御效果有限。

Joshi 等人[58]基于 x-vector 的说话人识别系统进行白盒攻击的防御研究，并提出四种被动防御方法。他们的研究发现，其中一种净化方法——PWG[59]与随机平滑结合，对说话人识别系统提供非常好保护效果，其准确率平均达到 93%，相比之下，未加防御的系统准确率仅为 52%。

Esmaeilpour 等人[60]利用有条件的生成对抗网络(class-conditional generative adversarial network)[61]进行对抗样本的防御。他们最小化随机噪声和对抗样本之间的相对弦距，找到最优的输入向量放入生成器中生成频谱图，然后根据对抗样本的原始相位信息与生成频谱图重建一维信号，净化对抗样本。这种重构没有给信号添加任何额外的噪声，到达对扰动进行消除的目的。

Chen 等人[62]系统地研究主动防御方法和被动防御方法，以确保说话人识别系统的安全性。他们使用 22 种不同的被动防御方法，并用 7 种最常见的对抗攻击（4 种白盒攻击和 3 种黑盒攻击）对说话人识别系统进行彻底评估。此外，他们分析在不同防御参数下，被动防御方法抵御对抗样本的效果。并且，他们还分析被动防御和主动防御结合使用时，抵御各种对抗攻击的效果。实验结果表明，在白盒攻击下，与单纯的对抗训练相比，被动防御和主动防御的结合使用相当有效，模型的准确率提高了 13.62%。

1.3 本论文的主要研究内容

在实际应用中，说话人识别系统可能面临各种形式的对抗攻击，例如添加噪声或篡改声音等。这些对抗攻击可能会导致系统误判，从而危及系统的安全性。因此，音频对抗样本的防御成为一项重要课题，其防御效果直接影响身份验证、司法认证以及智能设备上的个性化服务的可靠性。

为了提高说话人识别系统的安全性，本文对被动防御技术和主动防御技术两个方面进行研究。其中，被动防御方法主要应用于测试阶段，无需修改说话人识别模型，可以直接整合到模型中作为前端处理对抗样本；主动防御方法主要应用于模型训练阶段，通过调整模型参数来提前预防可能的攻击。具体工作如下：

1.在被动防御技术研究中，本文提出了一种基于生成对抗网络的被动防御方法。这种方法直接消除对抗扰动，将带有扰动的声纹样本转化为干净样本，对正常样本的影响极其微小。因此，这种方法为说话人识别系统提供了更有效的被动防御手段，既提高系统的精准性，又增强系统的鲁棒性。通过这一创新，本文为声纹识别的安全性和可靠性带来了新的解决方案。

2.本文首先进行对抗训练，探索其在说话人识别系统中的防御效果。随后，根据被动防御和主动防御之间的关联，对组合防御方法进行了研究，并比较主动防御与被动防御相结合形成的组合防御方法的效果。结果显示，这种组合防御方法能够显著提升模型的鲁棒性，同时增强模型的泛化能力。然而，需要注意的是，这种组合防御方法并非总是有效的，在某些情况下，会略微降低模型对特定攻击的防御能力。因此，在设计防御方案时，需要综合考虑各种因素，以确保模型在面对不同类型的攻击时能够保持高效的防御能力。

1.4 本文的结构安排

本文主要分为五章：

第一章为绪论，该部分首先介绍说话人识别领域的研究背景和意义，其次说明说话人识别技术的发展历程、对抗样本的研究现状，以及对抗样本的防御现状，最后讲述本文的主要研究内容和结构安排。

第二章为相关技术介绍，首先讲述说话人识别的基本原理及其相关任务，其次对对抗攻击和防御的基础知识进行简要说明，最后对一些常用的数据集进行描述。

第三章是本文的首个研究点，主要介绍生成对抗网络的原理和特性，并针对其目前存在的不足提出了一种基于生成对抗网络的被动防御方法。该方法的训练流程和损失函数得到详细阐述，并介绍实验评估所采用的指标。随后，针对开集识别和闭集识别任务分别进行攻击与防御，并对实验结果进行比较和分析。

第四章是本文的第二个研究点，介绍主动防御中的对抗训练原理，并对组合防御方法的步骤进行解释说明。本文以 FGSM、C&W_2 和 PGD 攻击为基础进行对抗训练，比较训练模型的泛化性和鲁棒性。此外，本文还探讨主动防御与被动防御相叠加的实验，评估组合防御的效果。通过这些实验，本文研究能更全面地了解不同防御方法在对抗样本下的表现，并为提高模型的安全性提供参考。

第五章是总结与展望。首先对本文的研究内容进行全面总结，随后针对已完成工作中存在的问题和不足，提出未来的研究方向和展望。

第二章 相关技术

了解说话人识别技术是进行说话人识别对抗防御技术的基础。因此，本章首先介绍了说话人识别技术的基本原理和任务。随后，进一步探讨了对抗样本攻击算法和防御手段。最后，对用于说话人识别的常见数据集进行介绍。

2.1 说话人识别技术

2.1.1 说话人识别的基本原理

说话人识别，也被称为声纹识别，是一种根据说话人的声音特征来判断其身份的技术。这种技术依赖于声音信号的各种属性，如声音的频率、音调、语速、音质、音色等。与说话人识别容易混淆的是语音识别。语音识别是将音频信号转换为文本，强调理解说话内容而非说话人身份。相较之下，说话人识别任务的重点在于验证或辨别说话人的身份，关注音色信息而非语义。

一个典型的说话人识别系统流程如下图 2-1 所示。为了让计算机识别一个用户，说话人识别系统首先需要对语音信号进行预处理和特征提取，然后对模型进行训练，并将其存储在模型数据库中。在认证阶段，需要对待识别的声音进行预处理和特征提取，然后将提取到的特征与数据库中的模型进行相似度比较，进行打分并判断，从而确定当前声音的说话人。

说话人识别的预处理过程是将原始语音信号转换为适合后续处理的关键步骤。首先，需要采集语音信号，并将其分段。接着，进行预加重处理以减少高频噪声。然后，对语音信号进行分帧和加窗操作，以便于特征提取。在提取特征时，常用的特征包括梅尔倒谱系数(Mel-frcquency Cepstral Coefficients, MFCC)[63]、线性预测倒谱系统(Linear Predictive Cepstral Coefficient, LPCC)[64]等，这些特征能够捕获语音信号的频谱和时域特性。最后，使用提取到的特征对模型进行训练，常用的模型包括高斯混合模型（GMM）、支持向量机（SVM）、深度神经网络（DNN）等。

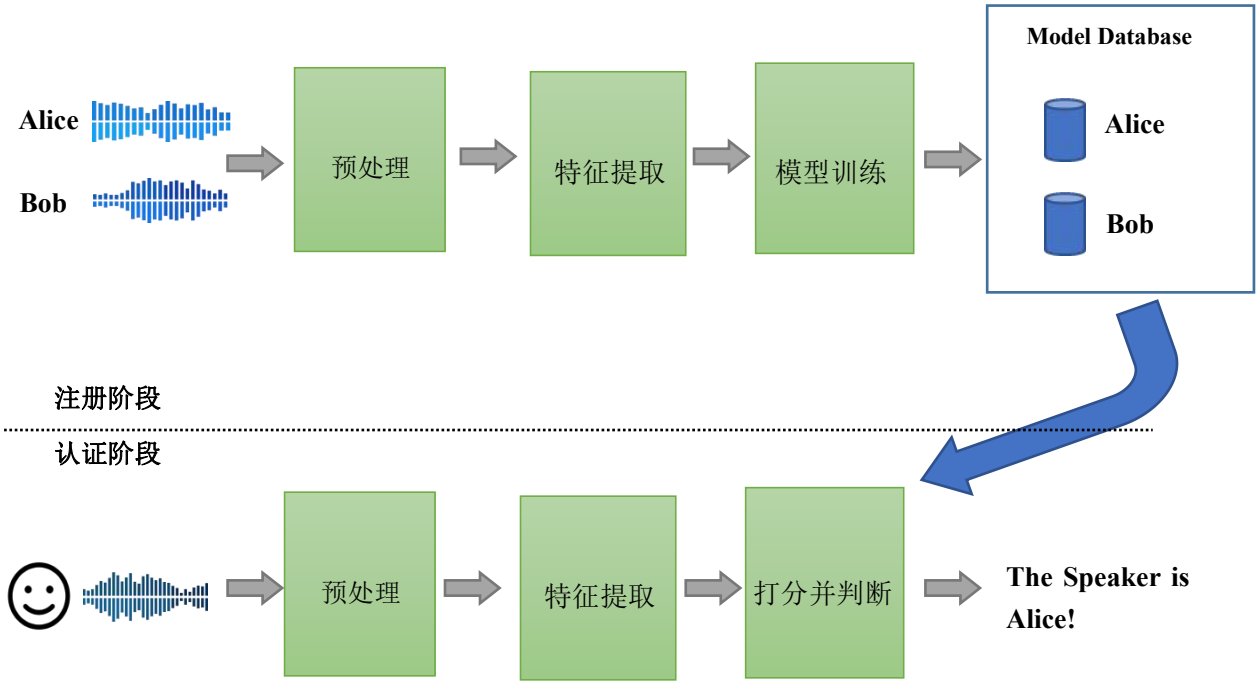

图 2-1 说话人识别的流程图

针对 MFCC 特征的提取过程如下：当语音信号被分帧和加窗后，接着通过傅里叶变换将信号转换到频域，再进行功率谱密度估计获取频谱信息。随后，利用滤波器组模拟人耳对声音的感知，并对滤波器组的输出进行对数运算和离散余弦变换，最终得到 MFCC 特征，整个流程如图 2-2 所示。MFCC 特征具有模拟人类听觉、减少冗余信息、抗噪声能力强等优点，因此被广泛应用于语音信号处理和说话人识别领域。

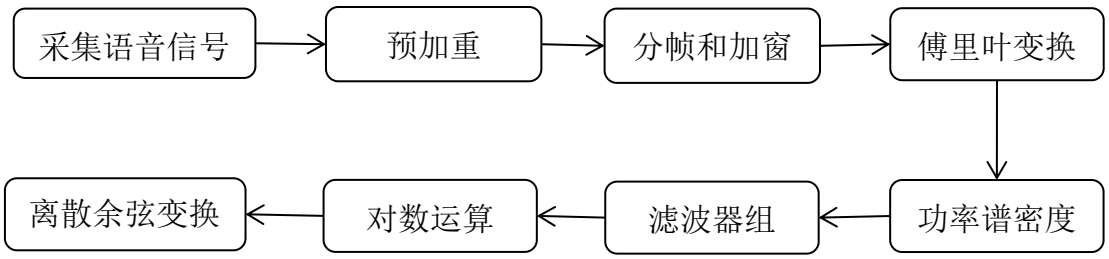

图 2-2 特征提取的流程图

2.1.2 说话人识别的内容和任务

说话人识别任务分为说话人确认（Speaker Verification, SV）和说话人辨认（Speaker Identification, SI）两种主要类型。说话人确认旨在验证特定声音是否属于特定说话人。

这一过程通常涉及将输入的声音样本与事先存储在数据库中的该说话人的声音模型进行比对，对输入声音进行打分得到一个分数，并与设定阈值进行比较，如果分数高于设定阈值，则给出肯定的答复，否则给出拒绝的答复，如公式 2-1 所示。在说话人确认中，可以看作是一个二分类的任务，被广泛应用于各种安全认证场景，例如电话客服中的用户身份验证、银行电话服务中的账户访问控制等。通过说话人确认，系统可以确保只有授权的个体才能访问敏感信息或执行特定操作，从而提高安全性。

$$D(x)=\begin{cases}\text{accept}, score \geq \theta \\ \text{reject}, \text{otherwise}\end{cases} \tag{2-1}$$

说话人辨认又可以分为开集识别（open-set identification, OSI）[65]和闭集识别（close-set identification, CSI）[66]。开集识别指的是在说话人识别任务中，识别出不在训练集中新说话人的身份。在训练阶段，通常会使用已知的说话人样本来建立模型，但在实际应用中，可能会遇到新的、未知的说话人，因此会出现某些说话人不在训练集中，这时就要对需要分辨的说话人与数据库中的说话人进行依次比较并打分，选出其中得分最高的与设置阈值进行比较，如果分数高于设定阈值，则辨认出说话人的身份，否则给出拒绝的答复，如公式 2-2 所示。

$$D(x)=\begin{cases}\underset{i\in G}{\text{argmax}}[S(x)]_i, & \text{if } \underset{i\in G}{max}[S(x)]_i \geq \theta \\ \text{reject}, & \text{otherwise.}\end{cases} \tag{2-2}$$

闭集识别指的是在一个已知的、有限的说话人集合内识别说话人的身份。在闭集识别中，系统需要识别说话人输入的语音信号，使其与说话人集合中一个已知的说话人匹配。这个任务通常在一个预定义的说话人数据库中进行，系统必须确定输入语音信号属于哪个已知说话人，而不能给出拒绝的答复，如公式 2-3 所示。

$$D(x)=\underset{i\in G}{\text{argmax}}[S(x)]_i \tag{2-3}$$

说话人识别按照识别内容可以分为两种类型：文本无关说话人识别（Text-Independent Speaker Recognition）和文本相关说话人识别（Text-Dependent Speaker Recognition）。文本无关说话人识别的目标是在不需要事先知道说话内容的情况下，仅通过说话人的声音特征来识别说话人的身份。这种方法不依赖于具体内容，而是专注于分析声音中的声学特征。通过对声音信号进行处理和分析，系统可以提取出与说话人身份相关的特征，并与存储在数据库中的模型进行比对，以识别说话人的身份。文本无关说话人识别通常应用于需要快速、自动化的身份验证场景。与依赖于特定文本提示的方法相比，文本无关说话人识别更加灵活，因为它不受特定文本内容的限制，可以适用于

各种语音交互场景。

文本相关说话人识别的目标是通过对说话内容的分析，结合声音特征来确定说话人的身份。这种方法利用说话人的语音特征以及与说话内容相关的语言特征，来识别说话人的身份。通过分析语音信号和文本内容之间的关系，系统可以提取出与特定说话人相关的特征，并将其与事先存储在数据库中的模型进行比对，以确定说话人的身份。文本相关说话人识别通常用于需要准确理解和识别说话内容的场景。在这些场景中，系统不仅需要识别说话人的身份，还需要理解其说话内容，以便进行后续的语义理解和处理。因此，文本相关说话人识别是一种综合利用语音和文本信息的方法，可以提高对说话人身份的准确性和可靠性。

2.2 对抗攻击与防御技术

2.2.1 对抗攻击

对抗样本是在原始语音中添加极其微小的扰动，而添加扰动后的对抗样本可以使目标模型分类错误，具体的对抗样本公式如下:

$$x' = x + \delta \text{ such that } \delta < \epsilon \tag{2-4}$$

其中x是原始正常样本，δ是添加的细微扰动，ϵ是扰动强度，x'是生成的对抗样本。对于无目标攻击，如果添加扰动后的x'对应的标签为y', x对应的标签为y, 要保证$\mathrm{y}' \neq y$，无目标攻击才算成功。对于有目标的攻击，如果想要x'的标签被识别为y'且$\mathrm{y}' \neq y$, x'经过说话人识别模型分类后为y'，则攻击成功。

在有目标的攻击中，按照攻击难度可以进一步分为简单标签攻击和困难标签攻击。简单标签攻击是指攻击者选择与原始标签相近的标签进行攻击，对原始数据的干扰相对较小。相比之下，困难标签攻击是指攻击者选择与原始标签相差较大的标签进行攻击，这种攻击更具挑战性，需要对原始数据进行大幅度修改。此外，对抗攻击按照基于攻击者对目标系统了解的程度，可以分为白盒攻击、黑盒攻击和灰盒攻击。

白盒攻击[67]是指攻击者对目标系统具有完全的了解，包括其结构、参数和内部工作原理等。在这种情况下，攻击者可以直接访问目标模型的所有信息，例如权重、激活函数和输入输出规范。由于攻击者拥有全面的知识，因此可以充分利用这些信息来生成对抗样本或执行其他形式的攻击。白盒攻击的典型示例包括 FGSM（Fast Gradient Sign Method）和 PGD（Projected Gradient Descent）等。

相比之下，黑盒攻击[68]是指攻击者对目标系统的了解程度有限，通常只能通过输入

输出来观察目标系统的行为。攻击者无法直接访问目标模型的内部信息，如权重和结构。因此，黑盒攻击者通常会通过试错和探索的方式来发现目标系统的弱点，并生成对抗样本或执行其他攻击。代表性的黑盒攻击包括基于迁移学习的攻击和基于优化的攻击。

灰盒攻击[69]是一种介于白盒攻击和黑盒攻击之间的攻击方式。在灰盒攻击中，攻击者拥有一定程度的关于目标模型的信息，但并非全部。通常情况下，攻击者可能知道目标模型的结构和参数，但可能无法直接访问或修改模型的内部细节，如梯度信息。因此，灰盒攻击旨在利用对目标模型的有限了解来生成对抗样本，以欺骗模型产生错误的分类结果或降低模型性能。灰盒攻击可以通过结合对目标模型的知识和常规的优化技术来实现，如迭代优化或梯度下降方法。这种攻击方式在现实世界的安全威胁和对抗性机器学习研究中都具有重要意义。

FGSM 是经典的攻击方法。基本思想是利用输入样本的梯度信息来生成对抗样本，使模型产生错误的分类结果。因为 FGSM 只对参数进行一次迭代，所以生成对抗样本的速度比较快。然而，由于这种方法的计算梯度的次数少，生成的对抗样本不够精细，其中一些样本会含有较大的噪音。其基本公式如下:

$$x^{'} = x + \varepsilon \cdot \text{sign}\left(\nabla_x L(x, y)\right) \tag{2-5}$$

其中，x是正常样本，$x^{'}$是对抗样本，ε是扰动强度的大小，$\nabla_x L(x, y)$是加入扰动后损失函数的梯度，$\text{sign}(*)$是对括号里面的数值取符号。如果梯度大于 0，则 $\text{sign}\left(\nabla_x L(x, y)\right) = 1$，如果梯度小于 0，则$\text{sign}\left(\nabla_x L(x, y)\right) = -1$。

MI-FGSM 攻击是对 FGSM 攻击的升级，引入动量的概念。与传统的 FGSM 相比，MI-FGSM 在每次迭代时考虑上一次迭代的梯度信息，以更好地引导对抗样本的生成方向。动量项的引入有助于在攻击过程中跳出局部最优解，使得对抗样本更具多样性，提高攻击的成功率和稳健性。MI-FGSM 攻击生成对抗样本的过程相对更加平滑，减少生成样本的噪声，同时能够更有效地攻击目标模型。动量的引入，MI-FGSM 攻击可以更好地适应不同的目标模型和数据分布，具有更高的成功率和更广泛的适用性。其基本公式如下:

$$g_{k+1} = \mu \cdot g_k + \frac{\nabla_x L(x_k^{'}, y)}{\left\| \nabla_x L(x_k^{'}, y) \right\|_1} \tag{2-6}$$

$$x_{k+1}^{'} = x_k^{'} + \alpha \cdot \text{sign}\left(g_{k+1}\right) \tag{2-7}$$

其中，μ是控制梯度的衰减，g_k是前几次保存的梯度，g_{k+1}是加上动量后当前的梯

度。α是扰动大小，$x^{'}$是对抗样本。

C&W$_2$攻击是一种基于优化的攻击方式，旨在生成扰动最小但能够欺骗神经网络模型的对抗样本。相比于其他对抗攻击方法，C&W$_2$攻击产生的扰动非常微小，几乎在音频中不可察觉，这使得它具有更高的实用性和隐蔽性。此外，攻击中用ω代替对抗样本把约束范围从[0,1]扩展到[-∞,+∞]，其公式如下：

$$x_i + \delta_i = \frac{1}{2}(tanh(\omega_i) + 1) \tag{2-8}$$

$$\text{minimize} \parallel \frac{1}{2}(\tanh(\omega) + 1) - x \parallel_2^2 + c \cdot f(\frac{1}{2}(\tanh(\omega) + 1) \tag{2-9}$$

$$f(x^{'}) = max(max\{Z(x^{'})_i : i \neq t\} - Z(x^{'})_t, -\kappa) \tag{2-10}$$

C&W$_2$攻击需要优化的有两个目标，第一个目标是生成的对抗样本$(x_i + \delta_i)$与原始样本x的差距越小越好，通过L_2进行差距之间的约束，第二个目标是生成的对抗样本$x^{'}$要使模型Z分类成指定的目标t。

PGD 攻击是一种迭代式的对抗攻击方法。它通过在输入数据的梯度方向上添加一定的扰动，然后对扰动进行投影，以确保对抗样本在允许的范围内，并且能最大化损失函数。PGD 攻击的基本思想是在每一次迭代中，沿着损失函数梯度的方向对输入数据进行微小的修改，使得模型在对抗样本上的输出误分类，同时确保对抗样本与原始样本之间的距离不超过预先设定的阈值。这种方法通过多次迭代不断优化对抗样本，以提高攻击的成功率，并且可以通过调整迭代次数和扰动大小来平衡对抗性和自然性之间的关系。其基本公式如下:

$$x_{t+1}^{'} = Clip_{x,\epsilon}(x_t^{'} + \alpha \cdot sign(\nabla_x L(x_t^{'}, y; \theta)) \tag{2-11}$$

其中，$Clip(*)$是对大于ϵ（最大扰动强度）的扰动进行裁剪，α是每次迭代的扰动强度。此外，PGD 攻击会在开始时从起始点的约束范围内随机找到一个点进行 PGD 攻击。

2.2.2 对抗防御

QT(Quantization)[56] 防御是把声音的振幅限制在参数 q 的整数倍，参数 q={128,256,512,1024}。通常来说，为了使对抗音频的隐蔽性更高，不被人们所察觉，对抗音频中添加的扰动较小，这也将会导致在输入空间中扰动的振幅较小。因此，QT 防御方法对音频的振幅进行修改，可以有效的消除对抗扰动。

$$\lfloor \frac{audio}{q} + 0.5 \rfloor * q \tag{2-12}$$

AS（Average Smoothing）[46]防御是通过固定某个采样点 x，将 x 前后 k 个采样点，共 2k 个采样点作为参考点，计算 2k 个参考点的均值，根据均值来替换采样点 x 的值。通过此方法依次修改采样点，以到达降低对抗音频的影响。MS(Median Smoothing)[56]防御则是把计算参考点的均值改为计算参考点的中位数。

QT、AS 和 MS 都是基于时域的防御方法。这些方法直接作用于原始的音频信号，在声音信号的时域上进行处理来增强系统的安全性和鲁棒性。QT 防御无法对振幅较大的对抗样本进行有效的防御，当振幅较大时，对抗样本中的扰动更加显著。此时，扰动的振幅超过 QT 方法可以处理的范围，从而导致对抗样本的扰动仍然能够干扰到说话人识别系统的性能。

AS 防御和 MS 防御可能会降低音频质量，因为在进行均值和中位数替换的过程中，会对声音信号施加一些噪音，这些改变削弱或模糊声音的一些特征，影响声音的清晰度、自然度和可听性，从而对模型的性能产生影响。

DS(Down Sampling)[34]防御是通过降低音频的采样率对音频进行修改，然后对音频进行恢复来消除对抗扰动。具体来说，设置下采样频率 τ (τ <1)，然后根据 τ 值对音频进行下采样，最后以上采样的方式对下采样后的音频进行恢复。

$$\tau = \frac{new\_audio\_sample}{ori\_audio\_sample} \tag{2-13}$$

LPF (Low Pass Filter)[70]防御是通过设置一个低通滤波器，过滤高频率声音中存在的对抗扰动。低通滤波器有两个参数：通带截止频率（f_p）和阻带截止频率（f_s）。

BPF (Band Pass Filter)[71]防御是设置一个滤波器，其既可以过滤高频率的声音，又可以过滤低频率的声音。BPF 有四个参数：通带截止频率的范围（f_{pl}和f_{pu}），低频的阻带截止频率为（f_{sl}）、高频的阻带截止频率（f_{su}）。

DS、LPF 和 BPF 都是基于频域的防御方法。这些方法通常涉及将音频信号转换到频域空间，然后对频域表示进行处理，以抵御对抗样本的攻击，保护音频系统的安全性和可靠性。DS 防御使用时也有很多限制，因为采样定理要求采样频率至少是信号中最高频率的两倍，以避免采样失真和混叠效应，所以 DS 方法的采样率不能小于信号最高频率的两倍。

LPF 无法防御低频率对抗音频的扰动，而且通带截止频率和阻带截止频率需要进行不断的测试，才能找到一个较好的平衡点，以保留音频的自然特性并抵御对抗样本的攻击。由于不同的对抗样本具有不同的频率特征，BPF 对于选择合适的截止频率是一项挑

战。因此需要根据具体情况来选择合适的过滤器参数，这需要进行大量的实验和调优，无法直接用于实际的应用。综上所述，这两种频域防御方法存在的挑战是需要在频域中精确调节参数，以确保对抗样本的扰动被有效地抑制，同时尽可能地保留音频的质量和可识别性。

OPUS[51]和 SPEEX[52]是两种音频编解码器，具有音频压缩的功能。这两种编解码器在对音频压缩时都会对噪声进行处理，从而破坏对抗样本中的扰动。OPUS 对噪声处理分为两个阶段，NSA(Noise Shaping Analysis)和 NSQ(Noise Shaping Quantization)。其中，在 NSA 阶段的目的是找到用于 NSQ 中补偿增益G和滤波器系数。在 NSQ 阶段，通过以下公式对噪音进行过滤：

$$Y(z) = G \cdot \frac{1-F_{\text{ana}}(z)}{1-F_{syn}(z)} \cdot X(z) + \frac{1}{1-F_{syn}(z)} \cdot Q(z) \tag{2-14}$$

其中，$F_{\text{ana}}(z)$和$F_{syn}(z)$分别是分析噪声整形滤波器（Analysis Noise Shaping Filter）和合成噪声整形滤波器（Synthesis Noise Shaping Filter），X和Q是量化器（Quantizer）。公式 2-14 的前半部分是对输入信号进行整形，后半部分是对噪声信号进行整形。

SPEEX 主要被设计用于实时的网络电话通信。不过，在 SPEEX 编码器中也对噪音进行抑制，以提高网络通话的质量，具有抵御对抗样本的能力。对噪音的处理如下公式：

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{2-15}$$

其中，$A(z)$是线性预测滤波器，γ_1和γ_2的参数值分别为 0.9 和 0.6。线性预测滤波器作用是允许声音在不同的频率有不同的噪声水平。具体来说，$W(z)$的输出在声音的低频率段有更少噪声，在高频率会添加或引入一些噪声。

OPUS 和 SPEEX 是基于语音压缩的防御。这些方法通过对原始语音信号进行压缩编码，将其转换为压缩的格式，然后再解码还原为语音信号。在这个过程中，由于压缩算法的处理，一些对抗样本中的细微扰动会被消除或减弱，从而使得对抗样本的攻击效果降低。然而，需要注意的是，OPUS 和 SPEEX 的压缩算法会导致语音信号的失真或质量降低，因此在选择压缩算法时需要权衡压缩率和语音质量之间的关系。此外，对于某些压缩算法，可能需要调整参数以优化防御效果。在实际应用中，基于语音压缩的防御方法对音频文件处理较慢。因此，基于语音压缩的防御方法需要考虑参数设置、时间效率和对语音信号的影响，以实现有效的防御效果。

2.3 常用数据集介绍

语音领域中，存在许多不同的数据集，研究人员可以根据其特定的研究目的和需求选择合适的数据集进行实验和评估。本文将介绍一些常用的数据集。

TIMIT[72]是一个常用于语音识别、说话人确认和说话人辨认等任务的研究和评估。该数据集包含 630 个说话人的美国英语语音样本，包括 6300 个句子，每个说话人约有 10 个句子。这些句子涵盖各种语音特征和语音内容，从而能够全面评估语音处理算法的性能。TIMIT 数据集的语音样本被分为训练集（Training set）、测试集（Test set）、开发集（Development set）和评估集（Evaluation set）等子集，以便进行不同类型任务的评估和测试。此数据集仅涵盖标准的英语文本，意味着数据集中的语音样本内容相对单一，无法覆盖到各种不同的语言和口音变化。

Aishell[73]是中文语音数据集，用于语音相关任务的研究。数据集由来自中国科学院自动化研究所的语音与语言技术中心发布。Aishell 数据集包含约 178 小时的普通话语音数据，包含 400 位说话人的语音，其中训练集 340 个人，测试集 20 个人，验证集 40 个人，每个人讲三百多句话。此数据集的数据量相对较小，限制了模型的训练和性能。

VoxCeleb1[74]是一个用于说话人识别研究的公开数据集，其中包含来自 YouTube 的大量视频剪辑，涵盖各种说话人的语音样本。该数据集由英国牛津大学语音研究组创建，并于 2017 年发布。VoxCeleb1 包含 1251 位说话人的 100000 多条话语。这些音频片段的说话人具有不同的口音、职业和年龄，覆盖不同的场景和环境。由于音频数据来源于 YouTube，音频的录音设备和环境也会各不相同，导致数据集中存在质量参差不齐的样本，影响模型的训练效果和性能评估。

LibriSpeech 数据集[75]是一个用于语音识别和说话人识别研究的庞大开源语料库，总时长约 1000 小时，包含超过 9000 个说话人的语音样本。这些语音数据来自于有声读物，经过分段对齐和过滤其他人声音的处理，保证语音质量。此外，该数据集包含丰富的说话人样本，跨越了不同年龄、性别和口音，覆盖多样化的文化背景和地理位置，因此具有很好的代表性。为方便研究和评估，数据集已经被划分为训练集、验证集和测试集，这样的划分在表 2-1 中清晰呈现。每个样本都与相应的文本转录和说话人 ID 相关联，方便语音识别和说话人识别任务的研究和评估。

表 2-1 LibriSpeech 数据集

子集	说话人数量和性别	总时长（小时）	存储空间
dev-clean	20 男、20 女	5.4	337M
test-clean	20 男、20 女	5.4	314M
dev-other	17 男、16 女	5.3	346M
test-other	16 男、17 女	5.1	328M
train-clean-100	126 男、125 女	100.6	6.3G
train-clean-360	482 男、439 女	363.6	23G
train-other-500	602 男、564 女	496.7	30G

2.4 本章小结

在本章中，主要介绍说话人识别、对抗攻击以及相关的防御技术。首先，对说话人识别的基本流程进行介绍，包括音频预处理、特征提取和模型训练的步骤，并详细描述了说话人识别任务的各种类型。其次，阐述了对抗样本的定义、生成方法以及防御方法。最后，介绍了在语音领域中广泛使用的四种数据集。

第三章 基于生成对抗网络的被动防御方法

本章将介绍基于生成对抗网络的被动防御方法。在 CycleGAN-VC2[76]模型的基础上进行改进，提出了 CycleGAN-L2 的被动防御方法。与其他防御方法相比，本章的方法能够消除对抗扰动，同时又不会造成正常样本严重失真。

3.1 生成对抗网络

生成对抗网络（GAN）[77]最早由 Goodfellow 等人在 2014 年提出，最初主要应用于图像领域。该网络架构包括两个主要组件：生成模型和判别模型。模型的训练过程通常分为两个阶段：首先是生成器的训练，然后是判别器的训练。在生成器的训练阶段，判别器的参数保持不变，生成器接收一个随机噪声矢量作为输入，尝试使用这个噪声生成一个样本。生成的样本随后被传递给判别器，判别器评估样本的真实性并将结果反馈给生成器。判别器的任务是成为一个二元分类器，检查生成器生成的虚假样本和来自任务的真实样本，并通过持续训练提高判别能力。例如，对于给定的样本，判别器对其真实性进行建模，并将判别结果的概率作为反馈传递给生成器。在判别器的训练阶段，生成器的参数保持不变，判别器通过真实样本和生成器生成的虚假样本进行训练，以提高对真实和虚假样本的区分能力。最终，通过各自的损失函数进行优化，生成器和判别器不断相互协作和竞争，以实现生成更高质量的数据。

Goodfellow 等人在提出 GAN 时引入了 Minimax 损失函数。生成器力图最小化此函数，而判别器则力图最大化它。损失函数如下：

$$\underset{G}{Min}\underset{D}{Max}\, L(D,G) = E_x[log(D(x))] + E_z[log\left(1 - D\big(G(z)\big)\right)] \tag{3-1}$$

其中E_x表示所有真实样本的期望，$D(x)$是判别器将x判断为真的概率。$G(z)$是生成器以z为输入向量生成的输出，$D\big(G(z)\big)$是判别器将生成器生成的样本判断为真的概率，E_z是生成器所有输入的期望。

随着时间的推移，生成对抗网络的应用领域逐渐扩展到语音领域。其中，CycleGAN-VC2 在语音转换领域扮演着重要角色。它是一种声音转换技术，通过使用生成对抗网络实现从一个说话人的语音转换成另一个说话人的语音，同时保留语音的内容和语调特征。具体的工作原理如图 3-1 所示，利用生成对抗网络结构进行特征转换，把 Alice 的声音和 Bob 的声音相互转换，并通过循环一致性损失函数和身份映射性损失函

数确保转换后的语音保持原有的语义特征。

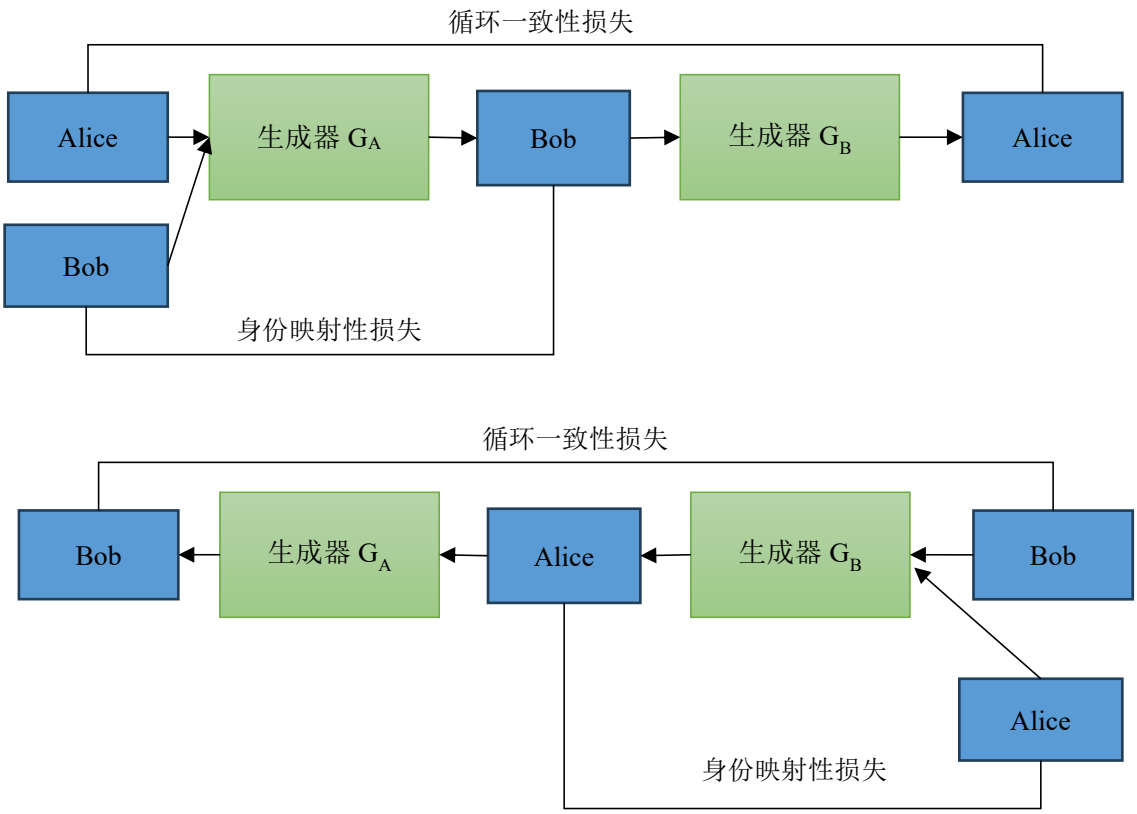

图 3-1 CycleGAN-VC2 的工作原理

但是本文研究的是说话人识别，其关注的是说话人的身份，若是用 CycleGAN-VC2，进行声音一对一的转换，把 Alice 的声音转变成 Bob 的声音，改变原始说话人的身份信息，说话人识别模型将会对说话人身份识别错误。因此，CycleGAN-VC2 无法直接应用到说话人识别的防御中。为了解决上述问题，本文把 CycleGAN-VC2 中说话人声音一对一转换改变成多对多的转换，即在训练的过程中不局限于仅使用单个说话人的正常样本，而是使用多个说话人的对抗样本和正常样本，并对损失函数进行适当的修改，保留说话人的身份信息，并提高模型的鲁棒性，使得说话人识别模型在面对对抗样本时更加稳健。

3.2 高鲁棒性的被动防御方法

3.2.1 总体思路

为了探索生成模型与说话人识别模型鲁棒性之间的关系，本文采用生成对抗网络来学习输入特征的分布。本文的方法受到了对抗防御 RL-VAEGAN[78]研究的启发。与

RL-VAEGAN 不同，其生成器仅学习对抗样本的分布和正常样本的分布，旨在消除对抗扰动，并未考虑生成器对正常样本的影响。然而，在实际应用中，进行对抗样本的防御时，必须同时处理正常样本。因此，本文生成器在考虑对抗样本的同时也会考虑正常样本。生成器的目标是将自然数据映射到自然数据，自然数据包括恶意数据和正常数据。

CycleGAN-VC2 中仅使用 L1 进行模型训练，并且在说话人识别防御领域，无人使用 L1 范数的 CycleGAN-VC2 用于对抗样本的防御。为了研究不同损失函数对防御效果的影响，本文使用 L1 范数和 L2 范数对损失函数进行约束，同时采用了类似于减量学习的方法来解决生成对抗网络训练难的问题，在训练过程中大大减少了训练时间。

此外，本文尝试使用 L1 进行对抗样本的防御时，发现其效果不如使用 L2。这是因为 CYCLEGAN-VC2 在训练模型时使用的数据是一个说话人对应一个说话人的情况，而在进行对抗防御时，面对的是多个说话人对应多个说话人的情况。L1 范数是求向量中各个元素绝对值之和，倾向于产生稀疏的系数向量，L2 范数是求向量元素的平方和的平方根，倾向于产生非稀疏的系数向量。因此，L2 范数可以鼓励模型筛选更多的特征，更适合本文的防御场景。为了检验本文方法的效果，在实验中，CycleGAN-L2 将防御不同的对抗样本攻击，并与其他的防御方法进行比较。

3.2.2 训练流程

首先，对音频进行预处理。在预处理过程中，得到四个主要特征：基频（F0）、频谱包络（SP）、振幅调整情况（AP）以及梅尔频率倒谱系数（MCEPs）。其中，F0 代表声音信号的音调，SP 代表声音信号的频谱分布，AP 表示声音信号的振幅在传输或处理过程中的调整或变化情况，而 MCEPs 是对梅尔频谱图的系数进行的特征提取。将 SP、F0 和 AP 组合起来，通过线性预测法估计声道的响应函数，然后对其取对数并应用离散余弦变换（DCT）来得到 MCEPs 系数。对 MCEPs 系数进行降维处理，选择合适数量的 MCEPs 系数作为最终的特征向量。

生成器和判别器的训练过程如图 3-2 和 3-3 所示。图 3-2 和图 3-3 这两个过程将交替进行训练。在图 3-2 中，首先训练生成器$G_{nat\to ori}$，然后训练生成器$G_{ori\to \mathrm{nat}}$。在这个过程中，$G_{nat\to ori}$的输入被分为两个阶段。在第一阶段，$G_{nat\to ori}$输入由自然数据组成，包含对抗样本和正常样本；在第二阶段，$G_{nat\to ori}$输入为正常样本。$G_{nat\to ori}$在这两个阶段的输出都是生成的正常样本，不过在第二阶段中，需要计算输入的正常样本和生成的正常样本之间的身份映射性损失。$G_{nat\to ori}$的输出的正常样本作为$G_{ori\to nat}$的输入，

$G_{ori\to nat}$的输出为生成的自然数据，包含对抗样本和正常样本，此时计算自然数据和生成的自然数据之间的循环一致性损失。

在图 3-3 中，首先训练生成器$G_{ori\to nat}$，然后训练生成器$G_{nat\to ori}$。在这个过程中，$G_{ori\to nat}$的输入被分为两个阶段。在第一阶段，$G_{ori\to nat}$输入由正常数据组成，只包含正常样本；在第二阶段，$G_{ori\to nat}$输入为自然数据，包含对抗样本和正常样本。$G_{ori\to nat}$在这两个阶段的输出都是生成的自然数据，包含对抗样本和正常样本，不过在第二阶段中，需要计算输入的自然数据和生成的自然数据之间的身份映射性损失。$G_{ori\to nat}$的输出的自然数据作为$G_{ori\to nat}$的输入，$G_{nat\to ori}$的输出为生成的正常样本，此时计算正常样本和生成的正常样本之间的循环一致性损失。

图 3-2 和图 3-3 中的判别器D_{nat}和D_{ori}负责区分生成的数据和真实数据，其中D_{ori}主要用于区分是否为生成的正常样本，而D_{nat}主要用于区分是否为生成的正常样本和对抗样本。对于生成器$G_{nat\to ori}$和$G_{ori\to nat}$，使用$G_{nat\to ori}$作为防御组件。

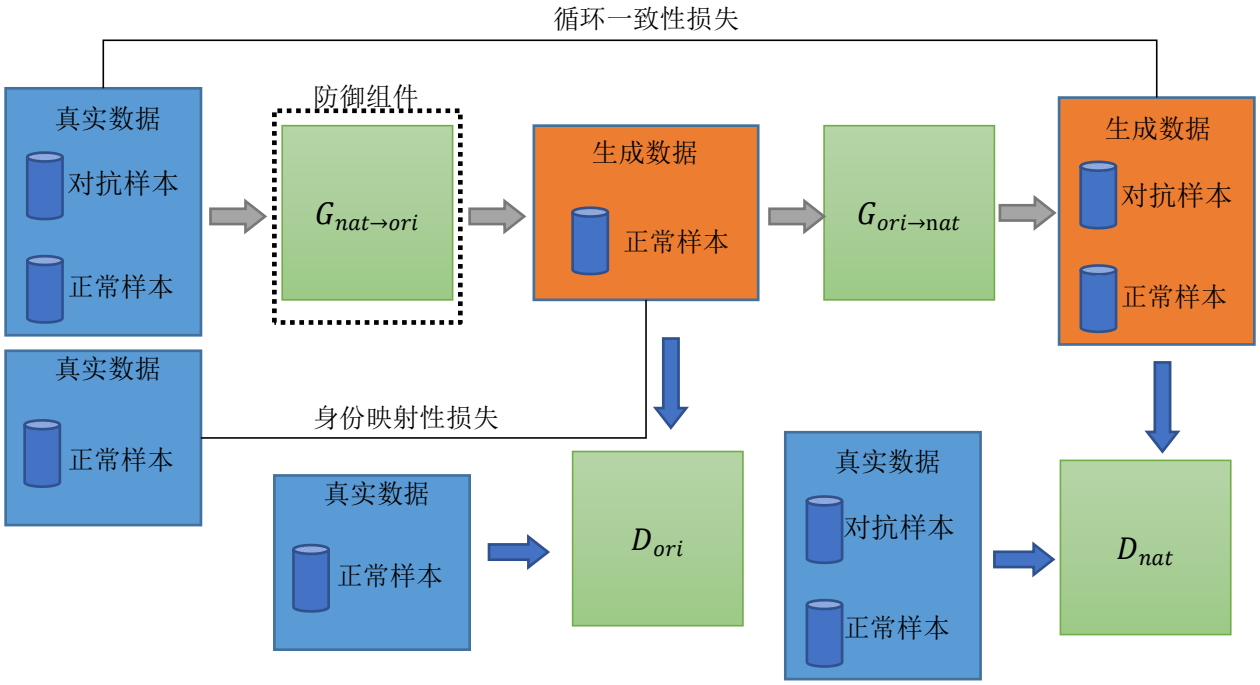

图 3-2 训练过程 a

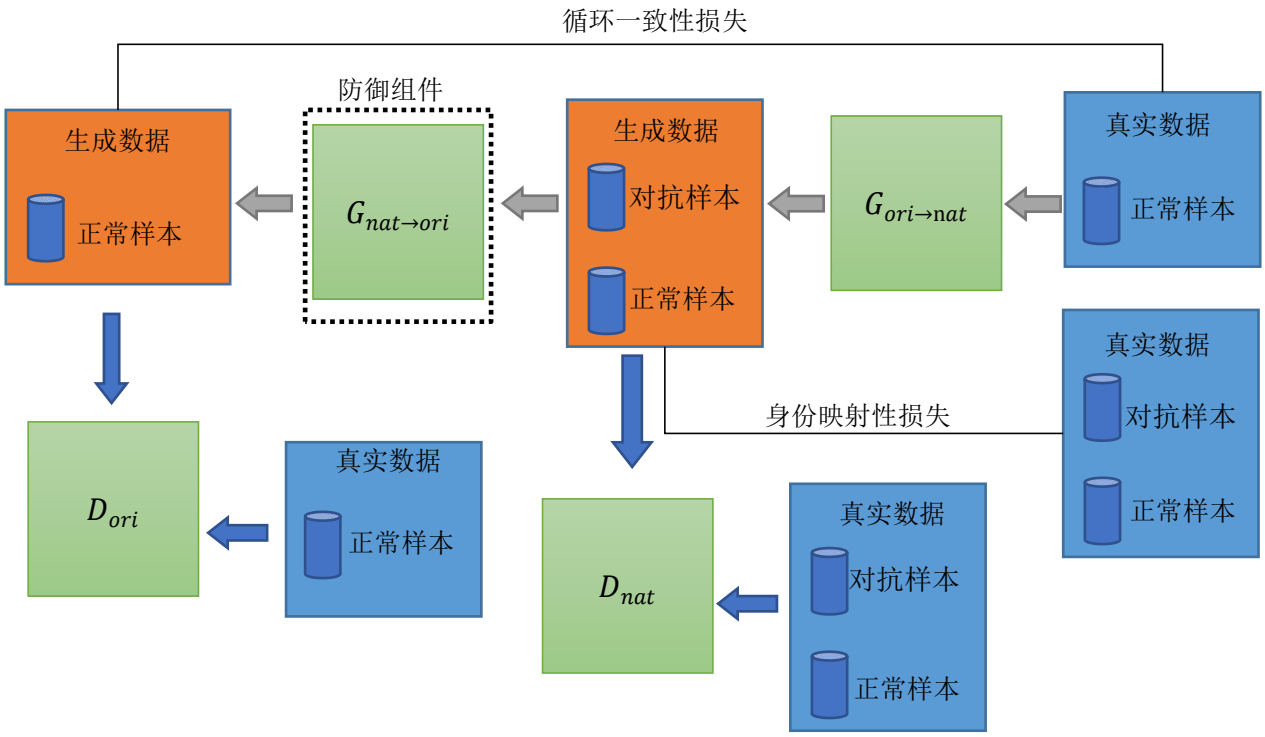

图 3-3 训练过程 b

最后，需要用 F0、SP 和 MCEPs 三个参数来合成一个正常音频。将自然数据中对抗样本和正常样本的 MCEPs 特征放入生成器$G_{nat\to ori}$中，产生正常样本的 MCEPs 特征。其次，对 MCEPs 进行维度扩展得到正常样本的 SP。然后，把自然数据的基频 F0 映射到正常数据中，得到正常的基频特征。

3.2.3 损失函数

生成对抗网络的目标优化函数是训练的关键。为此，本文分别为生成器和判别器设计了损失函数L_G和L_D。对于生成器，将其损失函数定义为：

$$L_G = L_{adv1}^G + \alpha L_{cyc} + \beta L_{id} + L_{adv2}^G \tag{3-2}$$

其中L_{adv1}^G和L_{adv2}^G为生成器的对抗性损失，L_{cyc}为循环一致性损失，L_{id}为身份映射性损失，α和β是权重参数，其初始值为 10 和 5。

生成器的对抗性损失：生成器的对抗性损失分为第一步对抗性损失L_{adv1}^G和第二步对抗性损失L_{adv2}^G。L_{adv1}^G的损失函数如下：

$$\begin{aligned} L_{adv1}^G(G_{nat\to ori}) = \mathbb{E}_{x\sim P_X(x)}\left[\log\left(1 - D_{ori}\left(G_{nat\to ori}(x)\right)\right)\right] + \\ \mathbb{E}_{y\sim P_Y(y)}\left[\log\left(1 - D_{nat}\left(G_{ori\to nat}(y)\right)\right)\right] \end{aligned} \tag{3-3}$$

其中，$x \in X$和$y \in Y$，X代表自然数据的集合（{adversarial, benign}$\subseteq X$），Y代表正常数据的集合（{benign}$\subseteq Y$），adversarial 和 benign 分别代表的是对抗样本和正常样本。D_{ori}是判别器，主要是对正常数据和通过$G_{nat\to ori}(x)$生成的正常数据进行判别，使得$G_{nat\to ori}$生成的数据更接近正常数据。同样，D_{nat}是判别器，D_{nat}是对$G_{ori\to nat}(y)$生成的数据进行判别。也就是说，L_{adv1}^{G}损失也就是衡量真实数据和生成数据之间的区别，L_{adv1}^{G}的值越小，代表着生成器的能力越强，或者判别器的能力有限，无法分辨出数据是否为生成的。L_{adv2}^{G}的损失函数如下：

$$\begin{aligned} L_{adv2}^{G}(G_{nat\to ori}, G_{ori\to nat}) = \mathbb{E}_{x\sim P_X(x)}[\log(1 - \\ D_{nat}(G_{ori\to nat}(G_{nat\to ori}(x))))] + \mathbb{E}_{y\sim P_Y(y)}[\log \\ (1 - D_{ori}(G_{nat\to ori}(G_{ori\to nat}(y))))] \end{aligned} \tag{3-4}$$

其中，$G_{ori\to nat}$用于生成自然数据，$G_{nat\to ori}$用于生成正常数据，然后D_{ori}对生成的正常数据进行判别，对D_{nat}对生成的自然数据进行判别，从而对$G_{nat\to ori}$和$G_{ori\to nat}$生成的数据进行监督，进行$G_{nat\to ori}$和$G_{ori\to nat}$训练。

循环一致性损失:本文把 CycleGAN-VC2 中 L1 范数换成 L2，用来约束的循环一致性损失，L_{cyc}损失函数如下：

$$\begin{aligned} L_{cyc}(G_{nat\to ori}, G_{ori\to nat}) = \mathbb{E}_{x\sim P_X(x)}\left[\| G_{ori\to nat}\left(G_{nat\to ori}(x)\right) - x \|_2\right] \\ + \mathbb{E}_{y\sim P_Y(y)}[\| G_{nat\to ori}(G_{ori\to nat}(y)) - y \|_2] \end{aligned} \tag{3-5}$$

CycleGAN-L2 模型与经典的 GAN 模型的不同之处在于它引入L_{cyc}来解决生成对抗网络中的模式崩溃问题。对于生成器$G_{nat\to ori}$来说，，其目标是生成的正常数据不能被D_{ori}分辨出来，如果$y \in Y$，Y具有 10 种标签，$G_{nat\to ori}(x)$认为只需要把x生成Y中的任意一个y就能很好的欺骗D_{ori}，那么$G_{nat\to ori}(x)$只会生成单一数据，发生模式溃散。这时就需要循环一致性损失，通过对输入的数据x和输出的数据x'($x' = G_{ori\to nat}\left(G_{nat\to ori}(x)\right)$)进行监督，使之间的差别尽可能小，迫使$G_{nat\to ori}(x)$与输入数据数据$x$之间的差别不大，防止生成器模式溃散。

身份映射性损失:同样，本文用 L2 范数来约束身份映射损失函数，L_{id}损失函数如下：

$$\begin{aligned} L_{id}(G_{nat\to ori}, G_{ori\to nat}) = \mathbb{E}_{x\sim P_X(x)}[\| G_{nat\to ori}(y) - y \|_2] \\ + \mathbb{E}_{y\sim P_Y(y)}[\| G_{ori\to nat}(x) - x \|_2] \end{aligned} \tag{3-6}$$

在图 3-2 的训练过程 a 的第二阶段，$G_{nat\to ori}$的输入是正常数据，输出是生成的正常数据，在图 3-3 的训练过程 b 的第二阶段，$G_{ori\to nat}$的输入是自然数据，输出是生成的自

然数据。数据使用 L2 范数对输入和输出进行约束。因为自然数据集中不是只有一种标签的数据，而是具有多种身份标签的数据，具有多种特征。使用 L1 范数约束L_{id}和L_{cyc}会趋向于产生少量的特征，不能很好的对多类特征进行选择，而使用 L2 范数约束会选择更多的特征，符合自然数据中具有多类数据的要求。

判别器的对抗性损失：对于判别器，本文将损失定义为：

$$L_D = L_{adv1}^D + L_{adv2}^D \tag{3-7}$$

L_{adv1}^D是判别器的第一步对抗性损失，L_{adv2}^D是判别器的第二步对抗性损失。第一步对抗性损失如下：

$$\begin{gathered} L_{adv1}^D(D_{ori}) = \mathbb{E}_{y\sim P_Y(y)}[logD_{ori}(y)] + \\ \mathbb{E}_{x\sim P_X(x)}\left[log\left(1 - D_{ori}\left(G_{nat\to ori}(x)\right)\right)\right] + \\ \mathbb{E}_{x\sim P_X(x)}[logD_{nat}(x)] + \\ \mathbb{E}_{y\sim P_Y(y)}\left[log\left(1 - D_{nat}\left(G_{ori\to nat}(y)\right)\right)\right] \end{gathered} \tag{3-8}$$

其中，判别器D_{ori}的作用是区分正常数据和生成器生成的正常数据；判别器D_{nat}的作用是区分自然数据和生成器生成的自然数据。$D_{ori}(*) \in [0,1]$, 当$D_{ori}(*)$=1 时，判别器认为输入的数据是真实的数据，当$D_{ori}(*)$=0 时，判别器认为输入的数据是生成的数据。一个完美的训练结果$D(*)$=0.5，即判别器无法区分数据是否为生成器生成的。第二步对抗性损失如下：

$$\begin{gathered} L_{adv2}^D(D_{ori}) = \mathbb{E}_{y\sim P_Y(y)}[logD_{ori}(y)] + \\ \mathbb{E}_{y\sim P_Y(y)}[log\left(1 - D_{ori}\left(G_{nat\to ori}\left(G_{ori\to nat}(y)\right)\right)\right) + \\ \mathbb{E}_{x\sim P_X(x)}[logD_{nat}(x)] + \\ \mathbb{E}_{x\sim P_X(x)}[log\left(1 - D_{nat}\left(G_{ori\to nat}\left(G_{nat\to ori}(x)\right)\right)\right) \end{gathered} \tag{3-9}$$

其中，判别器D_{ori}的作用是区分正常的数据和数据经过$G_{ori\to nat}$和$G_{nat\to ori}$两个生成器后生成的正常数据；判别器D_{nat}的作用是区分自然数据和数据经过$G_{nat\to ori}$和$G_{ori\to nat}$两个生成器后生成的自然数据。通过两个判别器分别监督两个生成器的生成数据，起到更好的训练效果。

3.2.4 训练方法

在训练开始时，会训练两个生成器和两个判别器，其中生成器分别是$G_{nat\to ori}$和$G_{ori\to nat}$，$G_{nat\to ori}$是把自然数据分布映射到正常数据分布，$G_{ori\to nat}$数把正常数据分布映射到自然数据。判别器分别是D_{ori}和D_{nat}，D_{ori}是区分正常数据和生成的正常数据，D_{nat}

是区分自然数据和生成的自然数据。

当训练进行到一半时，会对$G_{nat\to ori}$生成正常样本的能力进行测试。当$G_{nat\to ori}$输入为正常样本时，产生的输出使闭集识别中说话人识别模型能达到98%的准确率，而且继续对$G_{nat\to ori}$进行训练得到的输出结果使说话人识别模型的准确率不变或降低，运用减量学习的思想，去掉自然数据集合中的正常样本，并且使参数β置为0，取消身份映射性损失的计算。当$G_{nat\to ori}$输入为对抗样本时，输出的正常样本使分类器的准备率不变或降低结束训练。本文的方法简称为CYC-L2，其伪代码如表3-1所示。

表3-1 CYC-L2伪代码

算法 1 CYC-L2

输入： 自然样本 x，正常样本 y，对抗样本 $x^{'}$，迭代次数 E和K，生成器$\{G_{nat\to ori}, G_{ori\to nat}\}$，判别器$\{D_{ori}, D_{nat}\}$，参数 α和β，分类器 C

输出： 正常样本 y_{fake}

初始化

for e ← 1 to E **do**

  for k ← 1 to K **do**

    $y_{fake} \leftarrow G_{nat\to ori}(x)$

    $x_{cycle} \leftarrow G_{ori\to nat}(y_{fake})$

    $y_{identity} \leftarrow G_{nat\to ori}(y)$

    $x_{fake} \leftarrow G_{ori\to nat}(y)$

    $y_{cycle} \leftarrow G_{nat\to ori}(x_{fake})$

    $x_{identity} \leftarrow G_{ori\to nat}(x)$

Update L_D:

  $\nabla((0 - D_{ori}(y_{fake}))^2 + (1 - D_{ori}(y))^2 + (0 - D_{nat}(x_{cycle}))^2 + ((1 - D_{nat}(x)^2)$

  $\nabla((0 - D_{nat}(x_{fake}))^2 + (1 - D_{nat}(x))^2 + (0 - D_{ori}(y_{cycle}))^2 + ((1 - D_{ori}(y))^2)$

Update L_G:

$\nabla((1 - D_{ori}(y_{fake}))^2 + (1 - D_{nat}(x_{cycle}))^2 + (1 - D_{nat}(x_{fake}))^2 + (1 - D_{ori}(y_{cycle}))^2)$

$\nabla(\alpha(\sqrt[2]{(y - y_{cycle})^2} + \sqrt[2]{(x - x_{cycle})^2}) + \beta(\sqrt[2]{(y - y_{identity})^2} + \sqrt[2]{(x - x_{identity})^2}))$

  end for

  if e % 100 ← 0 **then**

    if $C(G_{nat\to ori}(y))_k \geq 0.98$ and $C(G_{nat\to ori}(y))_k - C(G_{nat\to ori}(y))_{k-1} \leq 0$ **then**

      $\beta \leftarrow 0, x \leftarrow x^{'}$

    end if

  end if

end for

Return y_{fake}

3.3 实验设置

3.3.1 实验环境和数据集

本文实验是在 CPU 主频为 3.6GHz 的 Intel i7-12700KF，显卡为显存是 12GB 的 NVIDIA GeForce RTX 3080Ti，内存为 64GB 的 Ubuntu20.04 系统上实现的。

正常数据来源于 Librispeech，经过筛选从中选取 10 个人（5 位男性，5 位女性），每个人有 110 条音频数据，共 1100 条音频数据用作实验。每个人选取 10 条音频，共 100 条音频文件用于对 x-vector 模型进行说话人注册；每个人选取取 100 条音频，共 1000 条音频文件用于生成开集识别中无目标攻击的对抗样本，其中 900 条对抗样本和 900 条正常样本组成自然数据集训练 CycleGAN-L2 模型；每个人剩下的 20 条音频，共 200 条音频文件（100 条正常样本，100 条对抗样本）用于测试训练效果。

此外，因为 PGD 攻击是最强的一阶攻击，如果 CYC-L2 防御方法能对 PGD 生成的对抗样本起到很好的防御效果，那么也能防御其他攻击生成的对抗样本，所以本文使用 PGD 攻击经过 30 次迭代产生开集识别中无目标攻击的对抗样本用作恶意数据，把恶意数据和良性数据合并成自然数据，用于生成器和判别器的训练。

3.3.2 模型介绍

本文使用 CycleGAN-VC2 中的生成器和判别器进行训练，来实现说话人识别系统中的防御，使用 x-vector 模型作为说话人识别模型。

x-vector 模型是一种用于说话人识别的深度神经网络模型，是当前说话人识别中主流的基线模型框架。它将语音信号映射到一个固定维度的向量空间中。这个向量通常被称为 x-vector，用于表示说话人的特征。x-vector 完整的系统参数如表 3-2 所示：前五层是时延神经网络（TDNN），用于捕获帧级特征。第六层是统计池化层，对第五层捕捉到的特征进行聚合。第七层和第八层是全连接层，第九层是 softmax 层。在七层的 segment6 中提取出说话人的特征向量 x-vector, softmax 层中的 N 对应于训练中说话人的数量。

表 3-2 x-vector 系统参数

Layer	Layer context	Total context	Input×output
frame1	{t-2,t+2}	5	120×512
frame2	{t-2,t,t+2}	9	1536×512
frame3	{t}	15	1536×512
frame4	{t}	15	512×512
frame5	{t}	15	120×1500
stats pooling	[0,T)	T	1500T×3000
segment6	{0}	T	3000×512
segment7	{0}	T	512×512
softmax	{0}	T	512×N

如图 3-4 所示，整个 x-vector 模型可以分为两个主要模块：帧级别模块(frame-level)和句子级别模块(segment-level)。在帧级别模块中，原始语音信号被分割成小的时间片段，称为帧。然后，对每个帧提取声学特征，如梅尔频率倒谱系数(MFCC)。接下来，经过 3 秒的滑动窗口的均值归一化和语音活动检测处理，去除没有说话人语音的帧，确保网络只处理有效的语音信息，而不是噪声或静音部分。随后，采用 TDNN 来提取帧级别的特征。因为每个人的声音特征存在于所有帧上，并且帧与帧之间具有时序性，所以使用 TDNN 来提取帧级别的特征。这个网络可以学习语音信号的时序结构信息。最终，该模块的输出是帧级别的说话人特征，这些特征将传递到句子级别模块进行进一步处理。

在句子级别模块中，对帧级别模块输出的特征向量在统计池化层进行聚合和汇总，将其转换为句子级别的特征。这一过程通过一系列统计操作完成，例如计算帧级别特征的均值和标准差，将不同长度的帧级别特征序列转换为固定长度的句子级别特征向量。接着，这些句子级别的特征向量通过全连接层和 softmax 输出层进行处理，以生成最终的 x-vector，代表说话人的身份特征。整个句子级别模块的目的是从帧级别的特征中提取出具有区分度的特征向量。