中图分类号：　　单位代号：**10280**

密　　级：　　学　　号：**18721691**

上海大学

专业学位硕士学位论文

题目	面向语音信号的神经网络模型水印技术研究

作　　者　王聿敏

学科专业　电子与通信工程

导　　师　吴汉舟

完成日期　2022 年 5 月

姓　　名：王聿敏　　学号：18721691

论文题目：面向语音信号的神经网络模型水印技术研究

上海大学

本论文经答辩委员会全体委员审查,确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导　　师：

答辩日期：

姓　　名：王聿敏　　学号：18721691

论文题目：面向语音信号的神经网络模型水印技术研究

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____日 期：_____

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密的论文在解密后应遵守此规定**）

签 名：_____导师签名：_____日期：_____

上海大学工学硕士学位论文

面向语音信号的神经网络模型水印技术研究

姓　　名：　王聿敏
导　　师：　吴汉舟
学科专业：　电子与通信工程

上海大学通信与信息工程学院

2022 年 5 月

A Dissertation Submitted to Shanghai University for the Degree

of Master in Engineering

Neural Network Model Watermarking for Speech Signals

M.A. Candidate：Yumin Wang

Supervisor：Hanzhou Wu

Major：Electronic and Communication Engineering

School of Communication and Information Engineering,

Shanghai University

May, 2022

摘　要

以神经网络为代表的人工智能模型凝结了设计者的智慧，需要消耗大量的训练数据和计算资源，可部署在本地以供个人使用，也可部署在云端以提供公共服务。然而，作为一种数字产品，人工智能模型易于被复制、调整和篡改。在人工智能技术迅速发展的同时，研究其产权保护具有显著意义。在此背景下，本文研究面向语音信号的神经网络模型水印技术，取得的主要研究成果如下：

(1)针对语音分类模型，提出了一种基于频域扰动的“黑盒”水印算法，该算法允许产权保护者在无法知晓模型内部细节的情况下实现产权保护。水印嵌入者通过在原始音频样本的频域中添加触发信号构造出触发音频样本，所得到的触发音频样本不仅可以抵抗恶意攻击，还不会引入明显的失真。同时，为了不影响目标模型在原始任务上的性能，还为所有触发音频样本分配了一个新标签。水印嵌入者将原始音频样本和触发音频样本结合起来对目标模型进行训练，从而实现水印的嵌入。水印验证者只需将精心制作的触发音频样本输入到目标模型，从而得到相应的预测结果，并利用对应标签进行一致性分析，即可认证模型的知识产权。实验结果表明，该方法不仅能够成功实现水印的嵌入和验证，还能够很好地保持目标模型在其原始任务上的性能。

(2)针对语音生成模型，提出了一种输出结果带水印的“无盒”水印算法，该算法通过在输出语音中检测水印，可实现神经网络模型的产权保护。具体而言，本文以语音对抗样本生成模型的知识产权保护为目标，所提出的方法中包含载体网络和水印网络两个部分，前者是待保护的神经网络，后者用于水印信息的嵌入和提取。通过在模型训练的过程中联合优化载体网络和水印网络的损失函数，使得训练好的载体网络不仅能够完成其原始任务，还允许产权保护者从输出语音中检测出水印，实现模型的产权保护。实验结果表明，用该方法优化的语音对抗样本生成模型在原始的定向攻击任务上的平均攻击成功率达到90%以上，在产权认证时的最高误码率低于 0.5%。

关键词：语音信号，神经网络模型水印，数字水印，版权保护，对抗样本

ABSTRACT

Artificial intelligence models represented by neural networks condense the wisdom of designers and consume a large amount of training data and computing resources. They can be deployed locally for personal use or on the cloud to provide public services. However, as a digital product, artificial intelligence models are prone to being copied, tweaked and tampered with. With the rapid development of artificial intelligence technology, it is of great significance to study its property rights protection. In this context, this thesis studies the neural network model watermarking technology for speech signals, and the main research results are as follows:

(1) For the speech classification model, a "black-box" watermarking algorithm based on frequency domain perturbation is proposed, which allows property rights protectors to realize property rights protection without knowing the internal details of the model. The watermark embedder constructs trigger audio samples by adding trigger signals in the frequency domain of the original audio samples. The obtained trigger audio samples can not only resist malicious attacks, but also does not introduce obvious distortion. At the same time, in order not to affect the performance of the target model on the original task, a new label is assigned to all the trigger audio samples. The watermark embedder combines the original audio samples with the trigger audio samples to train the target model so as to embed the watermark. The watermark verifier only needs to input the carefully crafted trigger audio samples into the target model to obtain the corresponding prediction results, and use the corresponding labels to conduct consistency analysis to authenticate the intellectual property rights of the model. The experimental results show that this method can not only successfully implement watermark embedding and verification, but also well maintain the performance of the target model on its original task.

(2) For the speech generation model, a "non-box" watermarking algorithm with

a watermark in the output result is proposed, which can realize the property rights protection of the neural network model by detecting the watermark in the output speech. Specifically, this thesis aims at the intellectual property protection of the speech adversarial example generation model. The proposed method includes two parts, the carrier network and the watermark network. The former is the neural network to be protected, and the latter is used for embedding and extracting the watermark information. By jointly optimizing the loss function of the carrier network and the watermark network in the process of model training, the trained carrier network can not only complete its original task, but also allow the property rights protector to detect the watermark from the output speech to realize the property rights protection of the model. The experimental results show that the speech adversarial example generation model optimized by this method has an average attack success rate of more than 90% in the original targeted attack task, and the highest bit error rate in property rights authentication is lower than 0.5%.

Keywords: Speech Signal, Neural Network Model Watermarking, Digital Watermarking, Copyright Protection, Adversarial Example

目　录

第一章 绪论

1.1 课题来源

本课题来源于国家自然科学基金青年项目“社交网络多用户协同的行为隐写”（项目编号：61902235）。

1.2 研究背景及意义

近年来，深度神经网络（Deep Neural Network, DNN）凭借其高效的性能在计算机视觉[1-3]、语音识别[4-7]和自然语言处理[8]等领域取得了巨大的成功，许多科技公司都将其所设计的性能优越的神经网络模型部署在商业产品中，这显著提升了服务质量并增加了效益。可见，神经网络模型具有很高的商业价值。然而，高性能的神经网络模型[9]离不开大量高质量的训练数据和精心设计的模型结构等，无论是训练数据的获取还是模型结构的设计，都非轻而易举之事。训练数据的收集、清理、处理和存储的过程甚至在某些情况下需要手动标记的过程都是极其耗时耗力的。同时，模型结构的设计、模型参数的初始化方式以及模型算法的选择等都需要由精通该领域专业知识的专家来完成。此外，在大型神经网络模型的训练过程中，往往需要多块图形处理器共同执行计算任务以满足对计算资源的需求。综上，神经网络模型可以被视为昂贵的数字资产。

现如今，计算机网络快速发展，神经网络模型也如其他多媒体内容一样被发布和传播，它们或被用于学术交流，或被用于商业盈利。而在神经网络模型投入使用的过程中，无论其内部细节公开与否，都很容易成为恶意攻击者有利可图的攻击目标[10]。假设攻击者能够知晓模型的网络结构和超参数设置等内部细节，他们可以通过微调或模型剪枝等操作得到新的模型，并声称该模型是他们的，进而对模型进行非法使用或分发。假设攻击者无法知晓模型的内部细节，仅能利用公开的应用程序编程接口（Application Programming Interface, API）执行查询操作，他们可以通过多次查询得到足够多的输入输出数据对，并利用这些数据对完成监督学习任务，进而训练出一个替代模型（又称为模型副本），该

攻击方式被称为“模型替代攻击”或“模型窃取攻击”[11]。综上，保护神经网络模型使其免受盗窃、非法再分配和未经授权的使用成为了迫切需要，可以将经过训练的神经网络模型视为合法所有者的知识产权，并进行相应的保护。

神经网络模型被广泛应用于多媒体数据相关的任务中，多媒体数据包括图像、语音、视频和文本等。其中，语音作为语言的物质外壳，包含一定的语义信息，具有重要的社会属性。面向语音信号的神经网络模型作为神经网络模型的重要组成部分，经常被应用在语音转换、机器翻译、语音合成和语音识别等多个任务中，例如语音合成模型 WaveNet 作为第一个能够生成类似于人类自然语音的神经网络模型，已经商用于谷歌助手。与此同时，包括腾讯、百度、科大讯飞、亚马逊、谷歌和微软等在内的国内外公司都在研发面向语音信号的神经网络模型，将来势必会有越来越多的模型被投入商用。综上，研究面向语音信号的神经网络模型知识产权保护具有显著意义。

考虑到神经网络模型易于被盗窃的事实，完全防止盗窃行为可能是不现实的。既然无法事先预防盗窃行为，合法的模型所有者至少希望能够对已经造成的侵害作出反应，并能够声明神经网络模型的版权以便采取进一步的措施。这就需要在神经网络模型中嵌入适当的标识，以此追溯被盗神经网络模型的合法所有者。用于对数字财产进行标识的方法被称为数字水印，它是指在不影响数据使用的情况下将标识信息嵌入到原始数据(如图像)中来声明版权的技术。在过去的二十年中，数字水印技术被广泛地应用在保护多媒体内容的知识产权[12-14]上。受到经典的数字水印技术的启发，有学者提出可以将数字水印技术从保护多媒体内容的知识产权推广到保护神经网络模型的知识产权上来。然而，由于适用对象不同，并不能将现有的水印算法直接迁移用于神经网络模型的知识产权保护，这促使人们对神经网络模型水印技术进行全面且深入的研究。

目前，已经有一些神经网络模型水印技术[15, 16]相关的研究成果被发表出来。但是，这些成果大多都是针对图像相关的神经网络模型的，面向语音信号的神经网络模型水印技术的研究成果还较少。同时，由于语音和图像的处理方式不尽相同，无法将面向图像的神经网络模型水印技术直接迁移到面向语音信号的神经网络模型中，这促使本文对面向语音信号的神经网络模型水印技术进行研

究，以对面向语音信号的神经网络模型的知识产权进行保护，从而在必要的时候检测出是否发生了知识产权侵权行为，以此来维护神经网络模型合法所有者的经济利益，同时可以降低相应的经济风险。

1.3 神经网络模型水印

1.3.1 模型水印的基本概念

在不影响数字载体(如图像)使用价值的情况下，通过将标识信息嵌入到数字载体中，来实现版权保护的技术被称为数字水印。数字水印技术中，主要是利用数字载体中存在的某种形式的冗余来实现水印的嵌入，以此保证数字载体的信息或感知意义不被损害。与数字载体相类似，神经网络模型也存在一定的冗余。神经网络模型是依靠大量的权重参数来学习输入和输出之间的映射关系的，通常而言，这些参数拥有的能力不止于完成模型自身的任务，因而，对于参数具体数值的选择存在着一定的自由度。就分类模型而言，当对模型参数进行细微地修改时，分类模型的预测结果不一定会发生变化，此时，同一个预测结果下的模型参数的变化范围就可以看作是冗余的空间。类似的，可以利用这些冗余空间来承载水印，实现对神经网络模型的版权保护。

然而，现有的数字水印技术无法直接迁移用于保护神经网络模型的知识产权，主要是因为传统的数字水印技术是通过直接修改数字载体的具体数值来嵌入水印的，对于神经网络模型，如果采用类似的方式直接修改模型参数来嵌入水印，会导致无法估计水印对模型自身性能带来的影响。同时，传统的数字水印技术中水印是直接从数字载体的具体参数中提取的，而神经网络模型中的水印应该不仅能够从模型的输出或中间层的输出中提取，还能够通过观察模型对特定输入的预测结果与预设结果是否一致来进行验证。综上，需要研究能够适用于神经网络模型知识产权保护的水印技术。

受传统的数字水印技术的启发，有研究者将数字水印技术的思想从多媒体所有权认证推广到神经网络模型所有权认证上来，提出了神经网络模型水印(又称为模型水印)的概念。模型水印是一种在不影响神经网络模型原始任务性能的

情况下，通过某种通用框架将水印信息嵌入神经网络模型中，从而实现版权保护的技术，合法的模型所有者可以在必要的时候从目标模型中提取出水印用于版权的认证。与传统的数字水印技术[17-22]相类似，模型水印技术也包括水印嵌入阶段和水印提取(或验证)阶段，图 1-1 展示了模型水印的基本框架。在水印嵌入阶段，通过模型训练过程中损失函数的不断优化来实现水印信息(如一个字符串)的嵌入。在水印提取(或验证)阶段，神经网络模型的合法所有者可以从目标模型的权重参数中或从模型中间隐藏层的输出激活中提取水印，并将提取到的水印信息与所嵌入的水印信息进行对比，以进行所有权的认证。在某些情况下，是通过向目标模型输入特定的数据，并分析模型的输出结果与预期的结果是否一致来验证水印，从而实现模型所有权的认证。

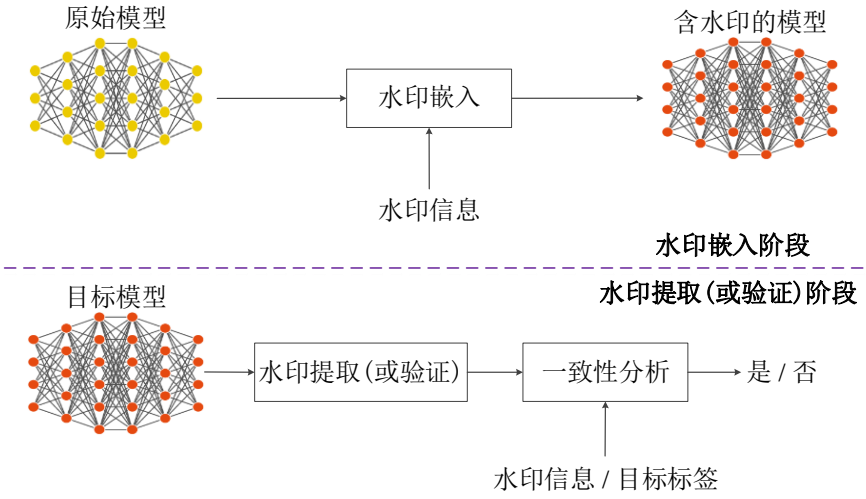


图1-1 模型水印的基本框架

1.3.2 模型水印技术的评价指标

虽然不同的应用场景下、结构不同的神经网络模型都有着各自适用的模型水印技术，但是这些技术所需要满足的要求都大致相同，对于它们的优劣有着统一的评价标准，其中，最常用的模型水印技术的评价指标[23]主要有以下十种：

(1)嵌入量：是指水印所承载的信息量，在多比特水印算法中，主要体现为水印信息所包含的比特数。为了让目标模型能够包含可能较长的水印，例如合法的模型所有者的签名，要求多比特水印算法能够做到在模型中嵌入尽可能多的水印信息。在零比特水印算法中，由于水印不用于承载任何有效的信息，因

而嵌入量不适用于零比特水印算法。

(2)安全性：是指在未经授权方蓄意攻击(如水印覆盖攻击和代理模型攻击)的情况下，要求水印的存在是保密的、水印是无法被检测到的，以防止水印被读取或修改。假设攻击者部分或完全了解水印算法，并试图破坏该算法，进而非法伪造或重构水印内容。此时，对于一个安全的模型水印算法，需要满足水印的丢失只能以模型自身任务性能的显著降低为代价。

(3)保真度：是指被嵌入水印的模型完成其原始任务的能力。对于任意的模型水印算法,均要求水印的嵌入不会显著降低目标模型在其原始任务上的性能。

(4)鲁棒性：是指从受干扰的神经网络模型中恢复水印的可能性。其中，常见的干扰主要包括微调或模型剪枝等操作。为了防止攻击者移除水印而使得模型的合法所有者无法进行版权的声明,要求所嵌入的水印能够抵抗不同的攻击，足够牢固不易被去除。

(5)可靠性：是指合法的模型所有者能够进行所有权认证的可能性。可靠的模型水印算法应该能够让合法的模型所有者以较高的概率识别出模型的知识产权，这就要求模型水印的假阴性率越小越好。其中，假阴性率可以理解为实际存在水印，但是却无法提取水印的概率。

(6)准确性：是指在不对模型进行修改或攻击的情况下，从模型中提取到的水印与所嵌入的水印相一致的可能性。对于多比特水印算法，使用误码率来评估其准确性，要求误码率无限接近于 0。对于零比特水印算法，使用错误检测率和漏检率来评估其准确性，前者是指在未嵌入水印的模型中检测到水印的概率，后者是指从嵌入水印的模型中无法检测到水印的概率，要求错误检测率和漏检率都尽可能的小。

(7)计算复杂度：是指对目标模型进行水印嵌入和水印提取时的计算开销。对于任意的模型水印算法，均希望嵌入水印的过程和提取水印的过程足够快，尽可能满足这两个过程所需要的计算开销可以忽略不计的要求。

(8)隐蔽性：是指水印不易被察觉的能力。为了使得攻击者既不能向模型中添加额外的水印，又不能声明现有水印的所有权，要求水印是不易察觉的。

(9)易于认证性：是指合法的模型所有者对模型所有权的认证能力。为了能

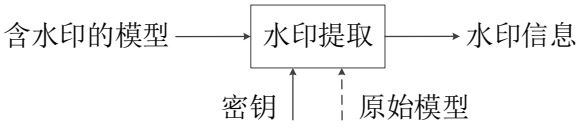
够让合法的模型所有者证明其身份，要求模型所有者和水印之间具有可供验证的强关联性。

(10)普适性：是指模型水印算法不仅限于用在特定的数据集和特定的神经网络模型上，还能够应用在其他各种数据集和神经网络模型上的能力。为了让模型水印算法被广泛地使用，要求其不仅能适用于最初测试和开发的模型，还能适用于执行不同任务的各种神经网络架构。

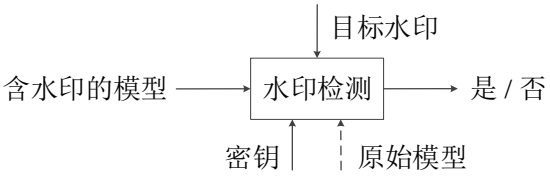
1.3.3 模型水印技术的分类

目前，越来越多的学者致力于研究模型水印技术，已经有一些研究成果被发表出来。这些模型水印技术依据不同的分类准则，有着不同的分类方法。

根据水印提取阶段水印的具体表现形式以及从模型中恢复水印的方式，可以将模型水印技术分为多比特水印技术[24, 25]和零比特水印技术[26]。如图 1-2 所示，在多比特水印技术中，水印信息对应于一个 N 比特的序列，因而从神经网络中提取到的水印是具体的比特信息。在零比特水印技术中，水印提取对应于检测任务，要求检测器确定所检测的内容中是否存在目标水印，因而从神经网络中提取到的是确切的水印存在情况。



(a) 多比特水印恢复

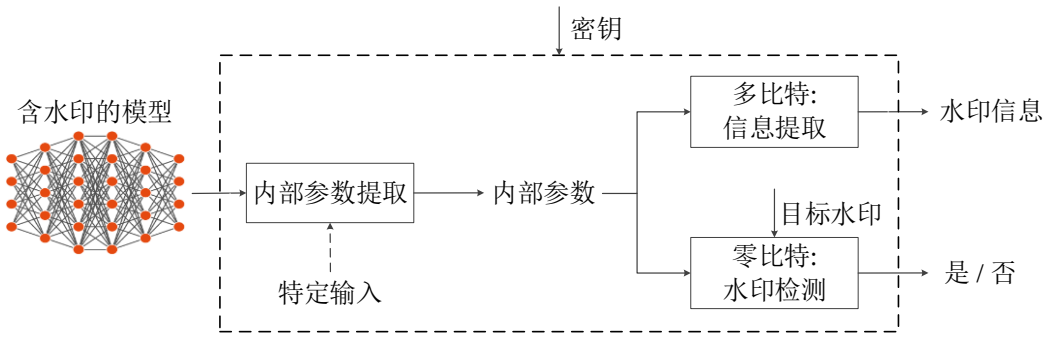


(b) 零比特水印恢复

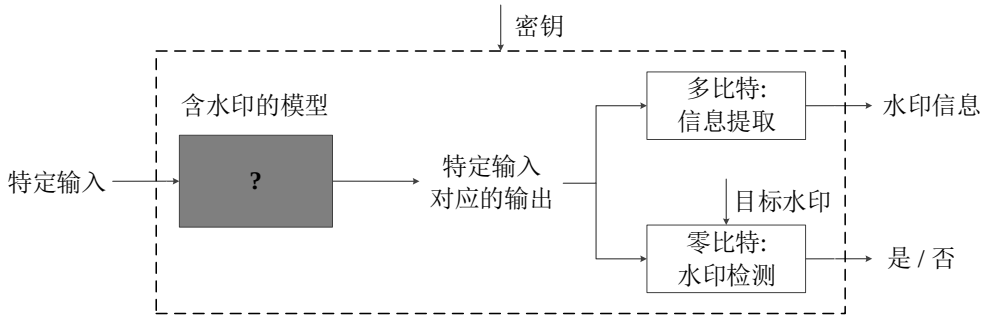
图1-2 水印恢复示意图

根据水印提取阶段可访问的数据的情况，可以将模型水印技术分为白盒水印技术[27-29]和黑盒水印技术[30, 31]。如图 1-3 所示，在白盒水印技术中，水印提取者能够知晓模型结构和内部参数等细节，内部参数可能直接对应于模型权重，

也可能对应于神经元对特定输入的输出激活。对于多比特白盒水印，水印提取器提取水印包含的具体比特信息。对于零比特白盒水印，水印检测器只需要判断是否检测到水印。在黑盒水印技术中，神经网络模型的结构和内部参数是全盲的，水印提取者只能选择用于查询模型的输入，并通过分析模型对于特定输入的输出结果来恢复水印。其中，特定的输入可能是秘密的也可能是公开的。



(a) 白盒水印提取

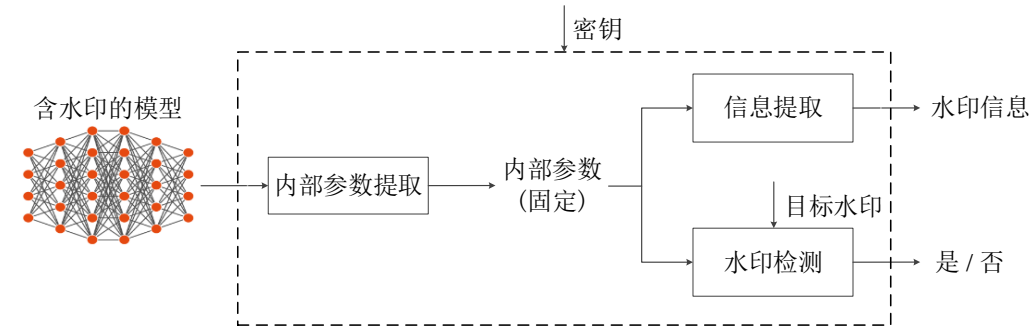


(b) 黑盒水印提取

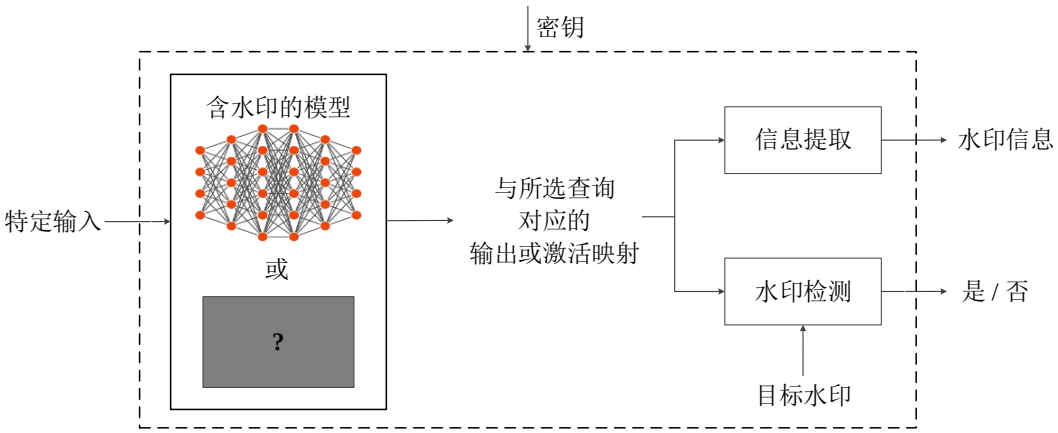
图1-3 水印提取示意图

此外，从嵌入水印的技术角度来看，模型水印技术可以被分为静态水印技术和动态水印技术。如图 1-4 所示，静态水印框架中，水印是被嵌入到模型的权重参数中的，这些权重参数在训练阶段确定，并假设其是不依赖于模型输入的固定值。动态水印框架中，水印则与对应于特定输入(又称为触发输入)的模型行为相关联。需要注意的是，即使某个方法也是通过选择适当的模型权重来嵌入水印的，只要它是通过观察水印对模型行为的影响来间接检索水印的，就属于动态水印技术。因此，可以推断出动态水印技术中水印的提取既可以在白盒模式下实现又可以在黑盒模式下实现。如果水印是通过查看模型的最终输出来恢复的，由于不需要访问模型的内部状态，因而是以黑盒方式提取的。当水

印与特定输入所对应的神经元的激活图相关联时，需要进行白盒提取。



(a) 静态水印框架



(b) 动态水印框架

图1-4 水印框架示意图

1.3.4 模型水印技术的研究概况

本文以静态水印和动态水印这一分类方式介绍国内外已有的部分研究成果。

(1) 静态水印技术

Uchida 等[27]首次提出了一种使用参数正则化器将水印嵌入到深度神经网络的通用框架。该方法中水印被定义成一个长度为 T 的字符串 $\{0,1\}^T$，通过优化组合损失函数 $L(\boldsymbol{\theta}) = L_O + \lambda L_R(\boldsymbol{\theta})$ 将水印嵌入到模型的参数中，其中 L_O 为原始任务所对应的损失，L_R 为水印嵌入任务所对应的正则化项，$\boldsymbol{\theta}$ 为模型的参数，系数 λ 是一个标量，用于影响所嵌入的水印的质量。该方法中水印的嵌入既可以在模型微调的过程中实现，又可以在模型从头训练的过程中实现。对嵌入水印的模型进行微调或模型剪枝操作后，所嵌入的水印依然能够存在。

Wang 等[32]扩展了 Uchida 等的工作，他们提出在目标模型之外，使用一个额外的、独立的且非公开的神经网络，该神经网络既可以用于在目标模型的特定参数上嵌入水印，又可以用于后续水印的验证。他们同样通过优化组合损失函数 $L(\boldsymbol{\theta}) = L_O + \lambda L_R(\boldsymbol{\theta})$ 对目标模型进行训练，其中 L_O 为原始任务所对应的损失，L_R 为水印嵌入任务所对应的正则化项，该项为神经网络的输出向量与目标水印之间的交叉熵损失函数值，$\boldsymbol{\theta}$ 为模型的权重参数，λ 为平衡系数，用于调整水印嵌入任务的重要程度。此外，为了保证模型原始任务的高保真度，他们还提出在早期收敛的模型参数中嵌入水印。但是，Wang 等[33]提出以上这两种方法均无法满足水印的安全性要求，因为这些方法容易导致模型参数的统计分布发生可检测到的变化。

Wang 等[29]提出了首个不会被属性推断攻击检测到的白盒水印算法，该算法基于生成式对抗网络（Generative Adversarial Networks, GAN）[34]的思想。具体的，将嵌入水印的模型 h_O 视为生成器，将用于检测模型参数统计分布变化的水印检测器视为鉴别器 h_D。训练过程中，在 h_D 努力区分嵌入水印的模型和未嵌入水印的模型的同时也促使 h_O 生成不易察觉的水印。该方法希望目标模型在被嵌入水印之后，模型参数的分布可以保持不变。

Fan 等[35]提出了一种基于护照的神经网络模型所有权认证方法，该方法中的护照层包含数字签名，将护照层嵌到卷积层的后面，由于护照函数需要学习网络层中的隐藏参数，这就使得模型的输出需要依赖于护照，如果给定的护照是错误的，就无法和隐藏参数匹配，模型的性能就会显著降低。对于所有权的认证，他们提出了三种方案，第一种是把护照分发给授权用户，但这样做存在护照被泄露的风险，会导致认证难度加大。第二种是基于多任务学习的思想，当不提供护照时，只对模型的原始任务进行训练；当提供护照时，就对组合损失函数进行优化。第三种是只在训练的过程中添加护照层，在所有权认证的时候则需要使用到额外的触发集。

Feng 等[36]提出了一种带有补偿机制的神经网络模型水印方案，他们首先通过密钥来选择待嵌入水印的权重，其次对所选择的权重进行正交变换得到系数矩阵，然后通过二值化方法将水印信息嵌入到系数矩阵中，最后对加载了水印

的系数矩阵进行逆正交变换，来恢复加载了水印的权重。此外，为了避免权重的变化对模型的原始性能造成过大的影响，他们对未嵌入水印的权重进行补偿微调，从而得到嵌有水印的模型。实验结果表明，该方法不但对模型的改动量很小，而且在保证理想的嵌入量的同时还具有鲁棒性。

(2)动态水印技术

Le 等[37]提出了一种基于对抗再训练的零比特水印算法，该算法通过对抗再训练来稍微移动决策边界，从而直接标记模型本身。他们首先确定了无限接近决策边界的对抗样本和原始样本；然后选择 50%的对抗样本和 50%的不会导致错误分类但是无限接近决策边界的原始样本来组成触发样本；最后利用触发样本对训练好的分类器进行微调，以使触发样本被预测为指定的类别。但是，Namba 等[38]认为该方法的一致性较差。

Adi 等[39]提出了一种基于触发集的黑盒水印技术，他们生成了抽象的彩色图片，并将其指定为随机的类别，从而构成触发集。为了验证所有权，他们在将水印嵌入到模型之前将一组图片指定为特定的标签，在验证的时候，他们就可以通过选择性地展示这些图片标签对来证明其所有权。

Rouhani 等[40]提出了一种通用的水印方法，该方法中将水印信息嵌入到不同网络层的激活图的概率密度函数中，其中，网络层的激活图大致遵循高斯分布。合法的模型所有者事先指定嵌入的水印字符串，并对模型进行训练以使得水印信息被融合进所选择分布的平均值中。该方法中投影矩阵用于将所选择的分布中心映射为二进制水印向量。在白盒设置下，该矩阵用于验证。在黑盒设置下，可以利用特征位于模型未使用区域的样本(即概率密度函数尾部区域的样本)构建触发数据集，他们认为这些样本比一般的对抗样本更加稳定。与 Uchida 等所提出的方法不同，该方法改变了模型的动态内容，即依赖于数据和模型的激活，因此更加灵活且不易察觉。

Chen 等[41]提出了一种端到端的多比特黑盒水印框架，他们将模型的合法所有者的二进制签名视为水印，为了将水印嵌入到模型中，他们构建了一个依赖于模型的编码方案，并根据相似性将模型的输出激活分为两组，一组被划分到类别 0，一组被划分到类别 1。同时，他们将包含在模型输出激活中的对应于任

意类别的签名视为签名位。在水印验证阶段，只需将特定的触发样本输入到模型中来检索签名即可。

除了上述的动态水印技术之外，一些动态水印技术中的触发样本是通过在原始训练样本中插入数字水印构造而成的。例如，Zhang 等[31]提出了一种远程黑盒验证机制下的图像分类模型水印算法，他们认为神经网络模型具有记忆能力且能够从训练样本中自动学习所嵌入水印的模式。具体的，他们将有意义的内容视为水印，在训练集中的图像样本上嵌入特定的字符串(如公司名称)，并将修改后的样本指定为与原始标签不同的标签。除了嵌入有意义的字符串，他们还尝试将噪声嵌入到原始训练样本中。

与 Zhang 等所提出的方法相类似，Li 等[42]提出了一种基于盲水印的神经网络模型水印框架，他们将独家标志嵌到原始样本中来构建触发样本，并训练模型，使得模型能够将这些触发样本预测为特定的标签，以此将水印嵌入到神经网络模型中。该方法中使用了自编码器，使得触发样本尽可能地接近原始样本。

Zhu 等[43]提出了一种能够抵抗伪造攻击的模型水印方法，该方法中运用单向哈希函数来构造触发样本，这些样本后续被用于模型的所有权认证。与此同时，触发样本的标签也被指定为样本的哈希值。这种情况下，如果攻击者没有训练神经网络模型的权限就不能构建伪造的水印。实验结果表明，该方法在不牺牲原始分类任务性能的同时，还能够抵抗水印伪造攻击。

上述的依赖于触发样本引发模型特殊行为的水印方法存在一个问题，这些方法实际上是在两个不同且独立的数据分布上进行训练的，第一个分布代表原始任务的数据分布，第二个分布代表水印触发器的数据分布。研究表明，由于原始任务和水印嵌入任务或多或少是不相关的，因而可以通过模型再训练或者微调的方式来去除水印。例如，Yang 等[44]提出，根据水印信息是冗余的且独立于模型原始任务的特点，可以采用蒸馏[45]的方法有效去除水印。为了解决这一问题，他们提出了一种称为“ingrain”水印的方法，该方法中使用了一个额外的包含水印的模型 ingrainer 来正则化原始模型，以将水印嵌入到原始模型中。

Jia 等[46]提出了一个类似的方法，该方法依赖于“纠缠的水印嵌入”，利用纠缠度使模型提取出能够代表原始任务数据和编码水印数据的共同特征。他们

采用软最近邻损失(Soft Nearest Neighbor Loss, SNNL)[47]测量来自同一类的点与来自不同类的点之间的距离，即标记数据的纠缠度，相对于两点之间平均距离更近的不同组中的点称为纠缠点[48]。在嵌入水印时，使用纠缠点可以保证水印嵌入任务与原始任务由相同的子模型完成，而不是由提取过程中可能受到损害的不同子模型完成。因此，攻击者在没有水印的情况下很难对模型进行提取。同时，通过纠缠，去除水印会降低模型在其原始任务上的性能。

Namba 等[38]提出了一种称为“指数加权”的方法，他们从训练分布中随机抽样，并在抽中的样本上签署错误的标签，从而构造水印触发集。他们提出通过指数加权来嵌入触发样本，使模型对其进行深刻的学习。该方法的依据在于，如果一个预测结果涉及到大量的绝对值较小的模型参数，则可以通过剪枝来改变预测结果，而如果一个预测结果只涉及少数的绝对值大的模型参数，则无法通过剪枝来改变预测结果。指数加权的方法就是通过以指数方式增加对模型预测有显著贡献的权重参数来达到这种状态，因而可以抵抗剪枝等攻击。

Li 等[49]提出了一种称为“空嵌入”的模型水印算法，该算法在模型的初始训练中就进行水印的嵌入，从而使得攻击者无法删除水印且无法再嵌入自己的水印。他们生成了一个由黑白像素组成的滤波模式 p，并将其放置在额外的训练图像上，白色图案下的图像像素改变为非常大的负数，黑色图案下的图像像素改变为非常大的正数，灰色图案下的像素保持不变。其中，变换后图像的预测类别需要保持与原始图像相同。在学习过程中使用极值并对优化设置强确定性的约束从而得到强水印。为了在所有者和像素模式之间建立绑定关系，他们还提出利用所有者的签名和确定性哈希函数来生成这一模式。

上述的大多数动态水印算法中，水印是模型的通用整体实例，这会导致当模型的被盗副本出现在某处时，合法的模型所有者无法确定哪些方可以访问该模型。因此，需要重点研究如何为每个用户创建独特的水印。为此，Chen 等[50]提出了一种白盒设置下的端到端的共谋安全水印框架，他们将针对个人用户的反共谋码本视作独特的水印，将这些水印嵌入到模型权重参数的概率密度函数中，通过在训练过程中使用特定于水印的正则化损失来实现合并。

Xu 等[51]提出了一种基于触发集的水印框架，该框架通过在神经网络模型

中嵌入序列号来进行模型的所有权认证。他们生成了一个唯一的序列号作为水印，并在序列号上创建了一个认证机构的背书。序列号由所有者通过数字签名算法生成，可以将序列号看作是所有者的私钥。在模型训练过程中，序列号会与原始任务一起通过损失函数拟合到模型中，训练完成后，可以通过发送触发输入、提取序列号和向认证机构进行验证这三个步骤，来实现所有权的认证。

Zhang 等[52]提出了一种用于保护计算机视觉模型知识产权的水印框架，他们精心设计了一个额外神经网络模型，称为水印模块，并利用其进行水印的嵌入。需要注意的是，水印是被嵌入到目标模型的输出中，且水印是统一的、不可见的。为了提高水印对模型替代攻击的鲁棒性，他们通过对抗训练对水印提取网络进行了增强。此外，他们还将目标模型的原始任务与水印嵌入任务联合起来进行训练，以此将水印嵌入的过程融合到目标模型中，从而能够对目标模型进行追踪溯源。

(3)其他水印技术

除了上述的静态水印技术和动态水印技术外，还有一些水印技术没有显式地将水印信息嵌入到神经网络模型中，而是使用模型自身现有的特征来识别潜在的被盗实例。这样做既不会给训练任务增加额外的计算开销，又不会影响模型原始任务的性能。例如 Zhao 等[53]使用对抗样本作为含水印的模型的触发集，他们在神经网络模型中识别了一些特殊的对抗样本，将其称为对抗标记，这些对抗标记在可迁移性上不同于传统的对抗样本，它们在相似的模型之间表现出较高的可迁移能力，而在不同的模型之间表现出较低的可迁移能力。由于对抗样本的数量和类型实际上是无限的，很难被移除，所以该方法中将对抗标记视为一种合适的水印触发器。

类似的，Lukas 等[54]也利用了对抗样本的可迁移性来验证神经网络模型的所有权。他们定义了可赠予的对抗样本类，这些样本只能迁移到目标模型的代理模型即潜在的非法副本上，而不能迁移到针对相关任务在相似数据上训练的相似模型(又称为参考模型)上。通过向模型查询这些样本，可以确定该模型是否是目标模型的副本。他们还针对这类对抗样本提出了一种生成方法，并证明了该方法对蒸馏和一些较弱的攻击都具有鲁棒性。

1.4 论文的主要研究内容

本文旨在研究面向语音信号的神经网络模型水印技术，在充分调研了国内外已有的神经网络模型水印算法的基础上，结合了语音分类模型和语音生成模型各自的特点，有针对性地提出了两种神经网络模型水印算法。针对语音分类模型，提出了一种基于频域扰动的“黑盒”水印算法；针对语音生成模型，提出了一种输出结果带水印的“无盒”水印算法。本文的各章内容安排如下：

第一章为绪论，首先介绍了课题的来源，其次介绍了本文的研究背景及意义，然后详细介绍了神经网络模型水印的基本概念、模型水印技术的评价指标和分类方法，并对国内外已有的部分研究成果进行了分类归纳和逐个介绍，最后介绍了本文各章的内容安排。

第二章主要对后面章节需要用到的面向语音信号的神经网络模型进行了介绍。对于语音分类模型，以说话人识别模型为例，先介绍了已有的说话人识别模型各个发展阶段的特点，再介绍了目前流行的基于 SincNet 的说话人识别模型的设计思路。对于语音生成模型，以语音对抗样本生成模型为例，先介绍了生成模型中最常使用的 GAN，再介绍了最新的基于 GAN 的语音对抗样本生成模型的设计思路。

第三章以说话人识别模型为例，提出了一种通用的语音分类模型水印算法，该算法利用触发音频样本实现水印的嵌入和验证。该章首先详细介绍了触发音频样本的设计思想和生成过程，主要是将传统的频域水印技术的思想拓展到触发音频样本的构造上，通过在原始音频样本的频域上以特定模式添加触发信号来构造触发音频样本。同时，为了保证模型的高保真度，还将所有触发音频样本指定为新增标签。其次介绍了水印的嵌入过程，通过将原始音频样本和触发音频样本结合起来对目标模型进行从头训练以嵌入水印。然后介绍了水印的验证过程，只需将触发音频样本输入到目标模型，得到相对应的预测结果，并利用对应标签进行一致性分析，即可完成所有权认证。最后从保真度、水印验证能力和鲁棒性等方面对所提出的方法进行了评估，实验结果表明，该方法对模型原始任务的性能影响很小，且可以成功地对模型所有权进行认证，此外，还

能在一定程度上抵抗噪声攻击。

第四章以语音对抗样本生成模型为例，提出了一种通用的语音生成模型水印算法，该算法在已有的载体网络之外，构造了一个额外的、独立的且非公开的水印网络，并利用该网络进行水印的嵌入和检测。该章首先介绍了所提出方法的总体架构，由载体网络和水印网络两部分组成。其次详细介绍了载体网络和水印网络的网络结构，载体网络中的生成器使用了编解码器结构，水印网络则使用了 inception-residual 模块，主要是想借助其多尺度学习和跨连接操作的能力。然后详细介绍了参与训练的各个模型的损失函数的设计，通过在模型训练的过程中联合优化载体网络和水印网络的损失函数，可以将水印嵌入到载体网络中的生成器的输出语音中，产权保护者只需要分析从输出语音中检测到的水印信息即可认证模型的所有权。最后从保真度、水印检测能力和鲁棒性等方面对所提出的方法进行了评估，实验结果表明，用该方法优化的语音对抗样本生成模型能够很好地保持模型原始任务的性能，且可以成功地对模型的所有权进行认证，此外，还能在一定程度上抵抗噪声攻击。

第五章对本文的工作进行了总结，并对工作中的可改进之处进行了展望。

第二章 面向语音信号的神经网络模型

本文旨在研究面向语音信号的神经网络模型水印技术，面向语音信号的神经网络模型主要包括语音分类模型和语音生成模型这两大类，其中，语音分类模型主要应用在说话人识别和情感识别等分类任务中，语音生成模型主要应用在语音合成和语音转换等生成任务中。具体的，在语音分类模型水印技术的研究中，本文以说话人识别模型为例，有针对性地提出一种适用于语音分类模型的通用水印算法；在语音生成模型水印技术的研究中，本文以语音对抗样本生成模型为例，有针对性地提出一种适用于语音生成模型的通用水印算法。为此，本章将详细介绍说话人识别模型的相关技术和语音对抗样本生成模型的相关技术，为后面章节所提出的方法提供理论依据。

2.1 说话人识别模型的相关技术

近年来，基于深度学习的说话人识别模型逐渐涌现出来，可以从输入特征的角度将这些模型大致分为几个发展阶段。总的来说，已有的研究成果展现出了一种以原始语音波形作为模型输入，同时根据对应任务的特点设计特定滤波器结构的发展趋势。本节将简要地介绍说话人识别模型的各个发展阶段，并详细地介绍目前流行的基于 SincNet 的说话人识别模型，从而为后面章节所提出的语音分类模型水印算法提供理论依据。

2.1.1 说话人识别模型的技术概要

基于深度神经网络在多个任务上的出色表现，目前被提出也可以用于说话人识别任务，基于神经网络的说话人识别模型逐渐涌现出来。起初，大多数的说话人识别模型都是使用手工制作的特征作为输入，常用的特征包括滤波器组特征（FBANK）和梅尔频率倒谱系数特征（Mel-Frequency Ceptral Coefficients, MFCC）等。例如 Ehsan 等[55]提出了一种用于小尺寸文本相关的说话人验证神经网络，其中，从给定帧及其上下文中所提取到的 40 维对数 FBANK 特征被用作

神经网络的输入。类似的，Gautam 等[56]也提出了一种用于短时说话人验证的深层说话人嵌入方法，该方法中提出了两种前馈网络结构，这两个网络都是使用 40 维对数 FBANK 进行训练的。Fred 等[57]提出了一种用于说话人识别的通用神经网络，其中，称为神经网络瓶颈的特征被用作输入，该特征是通过对神经网络隐藏层的输出激活进行采样得到的。David 等[58]提出了一种用于文本无关的说话人确认神经网络，其中，20 维 MFCC 特征被用作神经网络的输入。

就上述方法中所使用的特征而言，虽然它们都是依赖于人类的感知特性而设计的，但是无法保证在所有的语音相关任务中，使用这些人工提取的特征就可以得到最好的结果。需要承认的是，这些特征的确有其优势所在，但同时也必然存在着相应的不足。例如，标准的语音特征的确可以使得语音信号的频谱变得平滑，但对于音高和共振峰等说话人特征的提取可能会受到影响。为了应对这一不足，一些学者尝试将幅度频谱特征如语谱图作为神经网络的输入。例如，Zhang 等[59]提出了一种基于三重态损失的文本无关说话人验证框架和一种非常深的卷积神经网络结构 Inception-Resnet-v1，其中固定长度的说话人鉴别嵌入从稀疏的语音特征中学习，并用作说话人验证任务的特征表示。该方法中以线性尺度语谱图作为卷积神经网络（Convolutional Neural Network, CNN）的输入，并对其进行评估。Arsha 等[60]提出了一种基于两流同步卷积神经网络的说话人验证方法，该方法同样是从原始音频文件中直接提取出语谱图，并将其用作神经网络的输入。

与标准的手工提取的特征相比，语谱图的确可以保留更多的信息，但需要注意的是，语谱图也是经过手工提取的，这意味着在语谱图的绘制过程中，需要由具备专业知识的人员对一些重要的超参数进行选择和设置，如窗长、窗的类型、帧重叠的大小和离散傅里叶变换的点数等。为了避免人为的操作对提取到的特征产生过多的影响，有学者提出可以利用神经网络模型直接对原始语音信号进行特征的提取。例如，Hannah 等[61]提出了一种直接从原始语音信号中学习说话人鉴别信息的说话人识别方法，该方法中训练了一个基于卷积神经网络的说话人识别系统，系统的输入为原始语音信号，该系统以端到端的方式对未知说话人进行分类。该方法中还将说话人识别系统的输出层替换为真实的类别

和冒名顶替的类别这两类，并利用说话人的注册语音数据和冒名顶替者的语音数据来调整系统，从而在说话人识别系统中为每个说话人构建检测器。

上述的说话人识别模型中所使用的卷积层都是常规的卷积层，卷积层中的滤波器并没有被增加任何的约束，无法做到根据对应任务的特点在特定的频段上起作用。相反，Hiroshi 等[62]针对语音识别任务提出了一种将滤波器组学习与神经网络相结合的方法，他们将伪滤波器组引入到神经网络的底部，并将高斯函数应用在伪滤波器组中，联合训练整个滤波器组层和神经网络。大多数的系统采用预先提取的梅尔 FBANK 声学特征作为神经网络的输入，该方法则采用高斯函数代替梅尔 FBANK，这样可以使滤波器组层的频域保持平滑。美中不足的是，该方法是将功率谱串联并馈入到伪滤波器组中的，并不能实现真正意义上的端到端。而 Mirco 等[63]借鉴了 Hiroshi 等提出的滤波器设计思想，他们提出了一种可解释的卷积滤波器结构 SincNet，该结构以原始语音波形作为输入，实验结果表明，该结构可以成功地应用在说话人识别任务中。

2.1.2 基于 SincNet 的说话人识别模型

Mirco 等[63]提出了一种可以用于说话人识别任务的卷积滤波器结构，称为 SincNet。SincNet 可以直接将原始语音的时域波形作为输入，同时，它的第一层卷积层拥有更多有意义的滤波器。

卷积神经网络的第一个卷积层需要对高维的输入进行特征的提取，然而在深度神经网络模型中，第一个卷积层又很容易因为梯度消失的问题而无法进行参数的更新，可以看出，对于整个卷积神经网络来说，第一个卷积层是至关重要的。通常而言，在不对卷积层进行设计的情况下，将神经网络模型学习到的卷积核（也称为滤波器）可视化之后，可以发现滤波器是嘈杂的，且每个滤波器中会存在多个波段。这些滤波器所学习到的特征的确可以帮助神经网络模型完成相应的任务，但是对人类来说，是很难理解的，且人类很难找出学习到的特征与输入的高维向量之间的联系。为了应对这一挑战，Mirco 等提出对第一个卷积层进行精心地设计，他们通过使用基于参数化的 *sinc* 函数，来对滤波器的形状进行约束，这样做就等同于首先对高维的语音信号进行了带通滤波。在常

规的卷积层中，卷积核中的每一个权重参数都是需要进行学习的；而在 Mirco 等所设计的卷积滤波器结构 SincNet 中，卷积核中只有低截止频率和高截止频率需要进行学习，这促使神经网络模型更多地关注到后面网络层中权重参数的学习。可以看出，该方法紧凑且高效，同时，还具有相当大的灵活性。

标准的卷积神经网络的第一层是在输入波形和一些有限脉冲响应滤波器（Finite Impulse Response, FIR）[64]之间执行一组时域卷积，公式为：

$$\boldsymbol{y}(n) = \boldsymbol{x}(n) * \boldsymbol{h}(n) = \sum_{l=0}^{L-1} x(l) \cdot h(n-l) \tag{2-1}$$

其中，$\boldsymbol{x}(n)$表示语音片段，$*$是卷积符号，$\boldsymbol{h}(n)$表示滤波器，其长度为 L，$\boldsymbol{y}(n)$表示滤波操作后的输出。从式(2-1)可以看出，在标准的卷积层中，滤波器 $\boldsymbol{h}(n)$所包含的 L 个元素都是需要进行学习的。

与标准的卷积层不同，Mirco 等所提出的卷积滤波器结构 SincNet 中，输入波形是和预定义的函数 g 执行卷积操作的，公式为：

$$\boldsymbol{y}(n) = \boldsymbol{x}(n) * g(n, \boldsymbol{\theta}) \tag{2-2}$$

其中，$\boldsymbol{\theta}$ 表示函数 g 中可学习的参数。

该方法参考了数字信号处理中与滤波器设计相关的知识，将函数 g 定义成一个滤波器组，该滤波器组由多个矩形带通滤波器组成。通用的带通滤波器的幅值在频域中可以表示为：

$$G(f, f_1, f_2) = rect(\frac{f}{2f_2}) - rect(\frac{f}{2f_1}) \tag{2-3}$$

其中，f_1和 f_2 分别表示低截止频率和高截止频率，$rect(\cdot)$ 表示矩形函数：

$$rect(t) = \begin{cases} 0, & |t| > 0.5; \\ 0.5, & |t| = 0.5; \\ 1, & |t| < 0.5. \end{cases} \tag{2-4}$$

经过逆傅里叶变换[64]后，函数 g 的时域表达式为：

$$g(n, f_1, f_2) = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{2-5}$$

其中，$sinc$ 函数被定义为：$sinc(x) = sin(x)/x$。

根据奈奎斯特采样定理，截止频率 f_1和 f_2 应该在$[0, f_s/2]$区间内随机初始

化，其中，f_s表示信号的采样频率。考虑到与说话人身份相关的很多关键线索都在低频区域，而梅尔滤波器组恰好具有频域越低，滤波器分布就越密集的特点，因此，该方法中采用梅尔滤波器组的截止频率对低截止频率 f_1和高截止频率 f_2进行初始化。此外，为了确保 $f_1 \geq 0$且 $f_2 \geq f_1$，还对 f_1和 f_2进行了约束：

$$f_1^{abs} = |f_1| \tag{2-6}$$

$$f_2^{abs} = f_1 + |f_2 - f_1| \tag{2-7}$$

需要注意的是，Mirco 等发现在训练过程中 f_2始终小于奈奎斯特频率，因此没有给 f_2施加这一限制。

理想的带通滤波器具有通带完全平坦，且阻带衰减无限的特征，要想获得理想的带通滤波器需要借助无限多的元件，这在实际应用过程中几乎是不可能实现的。可以推测出，函数 g 实际上是通带中有波纹，且阻带中有有限衰减的滤波器。为此，Mirco 等通过流行的加窗[64]操作来缓解这一问题，他们将函数 g 与窗函数 w 相乘，以此来平滑函数 g 的截断特性，公式为：

$$g_w(n, f_1, f_2) = g(n, f_1, f_2) \cdot w(n) \tag{2-8}$$

考虑到汉明窗[65]特别适合实现高频选择性，因此，选用其作为窗函数 w：

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{L}) \tag{2-9}$$

其中，L 表示窗长。

卷积滤波器结构 SincNet 所涉及的上述操作都是可微的，因而可以像常规的神经网络模型一样进行训练，训练过程中，可以选用任意的优化算法，如随机梯度下降（Stochastic Gradient Descent, SGD）[66]算法等，且卷积滤波器结构 SincNet 中的截止频率是与其他层的模型参数一起进行迭代更新的。

图 2-1 展示了基于 SincNet 的说话人识别模型的架构，在第一层基于 *sinc* 的卷积之后，可以使用标准的卷积神经网络操作，如 Pooling 层、Layer Norm 层、Leaky ReLU 层、Dropout 层和 CNN/DNN 层等。然后，可以将多个标准的卷积层或全连接层连接在一起，最终使用 softmax 执行说话人分类任务。具体而言，Pooling[67]层主要包括 Max Pooling 和 Average Pooling 两种，Pooling 层主要用于减少特征的数量，既能降低过拟合的风险，又能提高运算的效率。此外，

还能保持空间变换的不变性，其中，空间变换包括平移、旋转和缩放等。Layer Norm[68]是一种标准化（又称为归一化）方法，用于对单个样本的所有特征进行归一化操作，使用 Layer Norm 可以保留不同特征之间的大小关系。Leaky ReLU 是一种激活函数，它不仅可以实现非线性映射，还能解决特定情况下某些神经元无法被激活的问题。Dropout[69]是一种正则化技术，通过丢弃部分神经元，可以防止神经网络过拟合。

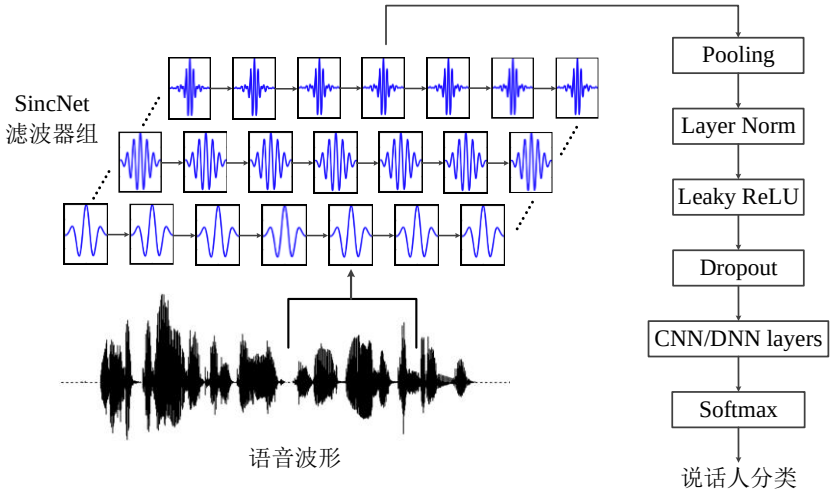


图2-1 基于 SincNet 的说话人识别模型的架构

Mirco 等所提出的卷积滤波器结构 SincNet 具有以下优势：(1)参数数量少—SincNet 中需要学习的参数只有低截止频率和高截止频率；(2)收敛速度快—SincNet 促使模型更多地关注对模型性能影响较大的参数；(3)计算效率高—所使用的函数 g 是对称的，在执行 SincNet 中的卷积操作时可以更加高效；(4)可解释性强—SincNet 中得到的特征图的形状可以被解释，且容易被理解。

Mirco 等率先提出了运用卷积滤波器结构 SincNet 对原始语音信号进行特征的提取，并将该结构成功地应用在了说话人识别任务中。他们通过实验证明了 SincNet 所学习到的滤波器紧凑且高效，且基于 SincNet 的说话人识别模型可以取得不错的性能，尽管在现实场景中，每个说话人所对应的句子可能只有几秒钟，基于 SincNet 的说话人识别模型也依然能够取得很好的性能。

2.2 语音对抗样本生成模型的相关技术

语音对抗样本生成模型中必然存在生成模型，目前，主流的生成模型主要

包括基于自回归的模型、变分自编码器、GAN 和流模型，其中，GAN 的发展最为迅猛。为此，本文对基于 GAN 的语音对抗样本生成模型水印技术进行研究。为了给后面章节中所提出的语音生成模型水印算法提供理论支持，本节将对 GAN 的基本概念和基于 GAN 的语音对抗样本生成模型进行详细的介绍。

2.2.1 GAN 的基本概念

深度学习本质上是学习数据集与目标之间的映射函数，主要包括回归、分类和结构化学习这三大类任务。其中，回归任务的输出是一个标量；分类任务的输出是一个类别，即 one-hot 向量；结构化学习的输出是一个序列、矩阵、图像或树等，且输出中的各个组成部分(如图像中的每一个像素点)之间都是有若干联系的。可以推断出，结构化学习通常被应用在相对复杂的任务中，如机器翻译、语音识别和图像转换等。

目前，传统的结构化学习的方法主要包括自上而下(Top Down)的方法和自下而上(Bottom Up)的方法这两大类。前者是先产生完整的对象，再从整体上去评估所产生的对象，从而找到最好的那个对象。后者则是一个部分一个部分地分开去生成所要产生的对象。这两类方法都存在相应的不足，导致结构化学习主要面临两个方面的挑战，一个是 One-shot/Zero-shot Learning，另一个是神经网络模型需要学会规划并具有大局观。One-shot/Zero-shot Learning 是 Top Down 这类方法所面临的挑战，以分类任务为例，该挑战是指在某一类别的训练样本数量很少或只有一个甚至没有的情况下，模型仍然可以学习到准确的映射关系，并能够在测试过程中对没有(或只有很少)训练样本的类别进行成功的预测。类似的，在结构化学习中，可以把每一个可能的输出看作一个类别，由于输出的可能性范围很广，大多数的类别都是没有任何训练样本与之对应的，这种情况下，该挑战则希望模型能够在测试阶段智能地创造出新的对象。神经网络模型需要学会规划并具有大局观是 Bottom Up 这类方法所面临的挑战，该挑战在结构化学习中是指，当模型生成一个很复杂的输出时，不应该片面地关注输出中各个部分的具体情况，而应该做到让输出的各个部分之间都具有很强的关联性。

GAN 是一种结构化学习的技术[70]，它包含生成器和鉴别器这两部分。其

中，生成器可以被视为一种 Bottom Up 的方法，它很容易利用神经网络模型完成生成任务，当给定一个输入向量时，它可以输出与之相对应的高维向量(如图像、文本或语音等)。生成器的性能是通过目标对象与生成对象之间计算得到的差异值来衡量的，然而差异值小并不能代表生成对象和目标对象从宏观上看是接近的，可见生成器在学习的时候只是片面地模仿目标对象，忽视了生成对象各个部分之间的联系。鉴别器可以被视为一种 Top Down 的方法，它很容易利用卷积核去建模生成对象各个部分之间的依赖关系。给定一个输入，鉴别器可以输出与之相对应的标量，该标量用于衡量输入的质量，且值越大表示输入越真实。可以推测出，如果让鉴别器完成生成任务，它则需要遍历所有的输入以得到相应的分值，从而确定出最高得分所对应的输入，将其作为生成对象。而要想让鉴别器求解出最优样本，需要具有足够真实的负样本，由于足够真实的负样本是很难得到的，因此鉴别器很难完成生成任务。GAN 则将生成器和鉴别器结合起来，通过取长补短，成为了结构化学习的一种比较好的解决方案。

生成器和鉴别器之间的联系在于,将生成器的输出(如图像、语音或文本等)输入到鉴别器后，鉴别器会分析该输入是否是真实的样本，如果生成器的输出不够真实没能骗过鉴别器，则生成器会进行进化并生成第二代的更加真实的输出，此时鉴别器也经过了进化且有着比第一代更强的判断能力，将第二代的生成器的输出输入到鉴别器后，鉴别器会再次对输入的真实性进行判断。上述过程将继续进行下去，生成器和鉴别器也在互相抗衡中不断进化，直到生成器生成的输出可以骗过鉴别器或该过程被提前结束为止。GAN 结合了生成器和鉴别器的优势，就鉴别器而言，可以利用生成器来完成生成任务，生成器可以生成负样本，完美地解决了鉴别器缺乏负样本的问题，且随着生成器的不断进化，所生成的负样本质量越来越高，这促使鉴别器的判别能力越来越强。就生成器而言，可以利用鉴别器学习到全局信息，从而生成足够好的输出，虽然生成器依然是一部分一部分的生成对象的，但是它不再通过求解 L_1 或 L_2 损失来衡量生成对象与目标对象之间的相似度,它的损失函数依赖于有着大局观的鉴别器。

对于 GAN 的训练，需要先初始化生成器(简称为 G)的权重参数和鉴别器(简称为 D)的权重参数，再进行每一轮生成器和鉴别器的交替训练，当达到预

先设定的训练周期时，训练结束。在每一轮的训练过程中，先固定生成器的网络参数并更新鉴别器的网络参数，使鉴别器学习给真实的输入样本打一个较高的分值，给生成器生成的样本打一个较低的分值。再固定鉴别器的网络参数并更新生成器的网络参数，使生成器学习生成一个能够误导鉴别器打高分的足够真实的样本。具体的，每一轮的训练过程如算法 1 所示：

算法 1 GAN 的训练策略

1. 从数据集中随机采样 m 个样本 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$，其中 m 为 minibatch 的大小；

2. 从一个正态分布中随机采样 m 个向量 $\{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m\}$；

3. 将这 m 个噪声样本输入到 G 中，利用公式 $\tilde{\boldsymbol{x}}_i = G(\boldsymbol{z}_i)$ 得到生成的数据 $\{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_m\}$；

4. 更新 D 的网络参数 $\boldsymbol{\theta}_d$ 来最大化：

$$\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} \log D(\boldsymbol{x}_i) + \frac{1}{m}\sum_{i=1}^{m} \log(1 - D(\tilde{\boldsymbol{x}}_i))$$

$$\boldsymbol{\theta}_d \leftarrow \boldsymbol{\theta}_d + \eta \nabla \tilde{V}(\boldsymbol{\theta}_d)$$

5. 再从同一个正态分布中随机采样 m 个向量 $\{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m\}$；

6. 更新 G 的网络参数 $\boldsymbol{\theta}_g$ 来最大化：

$$\tilde{V} = \frac{1}{m}\sum_{i=1}^{m} \log(D(G(\boldsymbol{z}_i)))$$

$$\boldsymbol{\theta}_g \leftarrow \boldsymbol{\theta}_g + \eta \nabla \tilde{V}(\boldsymbol{\theta}_g)$$

其中，η 表示学习率，步骤 1 至 4 为鉴别器的学习过程，步骤 5 至 6 为生成器的学习过程。

2.2.2 基于 GAN 的语音对抗样本生成模型

Wang 等[71]首次提出了基于 GAN 的语音对抗样本生成模型，他们将语音分类模型与 GAN 框架相结合，构成一个由生成器、鉴别器和语音分类模型组成的三方博弈。其中，生成器的作用是产生对抗扰动，进而可以得到语音对抗样本，所得到的语音对抗样本不仅可以误导语音分类模型，还能够骗过鉴别器。鉴别器的作用是判断生成器生成的对抗样本是否接近真实的样本。语音分类模型的作用是完成分类任务，得到分类误差，通过最小化该分类误差可以促使生

成器的参数进行更新，语音分类模型全程不参与训练。

Wang 等对模型的网络架构、损失函数和训练策略进行了精心设计，训练出了一个能够针对指定的语音样本和目标标签生成特定对抗扰动的生成器。所得到的语音对抗样本不仅可以成功地让最先进的语音分类模型分类错误，还能够具有可接受的听觉感知质量。此外，与流行的基于优化的语音对抗样本生成算法相比，该方法具有更快的生成速度。为了清晰且详尽地介绍该方法的设计思想，接下来将依次介绍该方法中的网络架构、损失函数和训练策略。

该方法的总体网络架构如图 2-2 所示，主要由生成器 G、鉴别器 D 和目标语音分类模型 f 这三个部分组成。该方法的具体流程如下，首先将原始语音样本 $\boldsymbol{x}$ 送入生成器 G，输出对抗扰动 $G(\boldsymbol{x})$，进而得到语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 即 $\boldsymbol{x}+G(\boldsymbol{x})$。然后，将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分别送入鉴别器 D 和目标语音分类模型 f。鉴别器 D 通过不断增强对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 和原始语音样本 $\boldsymbol{x}$ 的鉴别能力，促使生成器 G 产生能让鉴别器 D 无法察觉的对抗扰动 $G(\boldsymbol{x})$。目标语音分类模型 f 则指导生成器 G 生成能让其错误分类为预先指定的目标标签的对抗扰动 $G(\boldsymbol{x})$。总的来说，该架构的关键思想是让生成器 G 在小扰动的约束下，针对不同的原始语音样本 $\boldsymbol{x}$ 生成与之相对应的能够误导语音分类模型 f 的对抗扰动 $G(\boldsymbol{x})$。

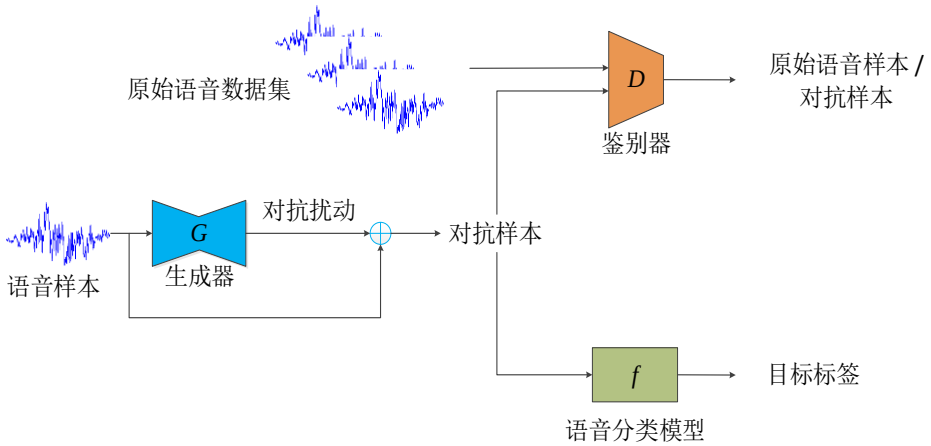


图2-2 基于 GAN 的语音对抗样本生成模型的总体网络架构

在总体网络架构中，目标语音分类模型 f 中的参数是被固定住的，参与训练的网络只有生成器 G 和鉴别器 D。对于生成器 G，它的目标是既能够误导目标语音分类模型 f 又能够愚弄鉴别器 D。其损失函数被定义为：

$$L_G = L_{\text{adv}}^f + \alpha L_{\text{fool}} + \beta L_{\text{hinge}} + \gamma L_2 \tag{2-10}$$

其中，L_{adv}^f 是对抗损失，表示生成器 G 对目标语音分类模型 f 的攻击能力。L_{fool} 表示生成器 G 对鉴别器 D 的欺骗损失。L_{hinge} 和 L_2 表示语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的铰链损失和 L_2 范数损失，主要是为了让语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的听觉质量在可接受范围之内。参数 α 、β 和 γ 是用于平衡各损失项重要性的权重。

式(2-10)中，对抗损失 L_{adv}^f 需要衡量目标语音分类模型 f 对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的预测结果与目标标签之间的差异，数学上被定义为：

$$L_{\text{adv}}^f = \text{E}[l_{\text{ce}}(f(\boldsymbol{x}_{\text{wa}}), \boldsymbol{y}_{\text{t}})] \tag{2-11}$$

其中，E[·]表示损失函数的期望，$\boldsymbol{y}_{\text{t}}$ 表示预先指定的目标标签，是一个 one-hot 向量，$f(\boldsymbol{x}_{\text{wa}})$ 表示目标语音分类模型 f 对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的预测结果，$l_{\text{ce}}(\cdot)$ 表示交叉熵损失函数。可以看出，最小化 L_{adv}^f 损失可以促使目标语音分类模型 f 将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分类为目标标签 $\boldsymbol{y}_{\text{t}}$。

式(2-10)中，欺骗损失 L_{fool} 被定义为：

$$L_{\text{fool}} = \text{E}[\log(1 - D(\boldsymbol{x}_{\text{wa}}))] \tag{2-12}$$

其中，$D(\boldsymbol{x}_{\text{wa}})$ 表示鉴别器 D 将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 判别为原始语音样本 $\boldsymbol{x}$ 的概率。可以看出，最小化 L_{fool} 等价于让 $D(\boldsymbol{x}_{\text{wa}})$ 足够接近 1，本质上是促使语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 足够接近原始语音样本 $\boldsymbol{x}$。

式(2-10)中，正则化项 L_{hinge} 和 L_2 的目标都是为了保证生成的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 有可接受的听觉质量。其中，L_{hinge} 被定义为：

$$L_{\text{hinge}} = \text{E}[\max(0, \|G(\boldsymbol{x})\|_2 - c)] \tag{2-13}$$

其中，c 表示边界值。该损失促使生成器 G 在采样位置上产生更稀疏的扰动，即尽可能修改较少的采样值。

L_2 损失被定义为：

$$L_2 = \|\boldsymbol{x}_{\text{wa}} - \boldsymbol{x}\|_2 \tag{2-14}$$

该损失用于控制扰动的总能量，即尽可能的让扰动的幅度较小，避免扰动过大。

对于鉴别器 D，它的目标是将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 与原始语音样本 $\boldsymbol{x}$ 区分开来，其损失函数被定义为：

$$L_D = \frac{1}{2}\{\mathrm{E}[\log(1-D(\boldsymbol{x}))] + \mathrm{E}[\log(D(\boldsymbol{x}_{\text{wa}}))]\} \tag{2-15}$$

其中，$D(\boldsymbol{x}_{\text{wa}})$ 表示语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 被鉴别器 D 判定为原始语音样本 $\boldsymbol{x}$ 的概率。

最后，结合生成器损失函数 L_G 和鉴别器损失函数 L_D，通过求解以下最小最大优化问题得到 G^*：

$$G^* = \arg\min_G \max_D (L_G + L_D) \tag{2-16}$$

其中，G^* 表示训练得到的生成器 G。一旦训练完成，G^* 可以针对指定的输入和目标标签高效地生成 $\boldsymbol{x}_{\text{wa}}$。

在训练过程中，为了确保鉴别器 D 能够有较好的区分语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 和原始语音样本 $\boldsymbol{x}$ 的能力，该方法一开始只对生成器 G 和鉴别器 D 进行训练，先让训练过程预热起来，之后再加入目标语音分类模型 f 进行剩余的训练。需要注意的是，生成器 G 和鉴别器 D 的训练是交替进行的，即生成器 G 的权重参数进行更新的时候，鉴别器 D 的权重参数是固定不变的，反之相反。目标语音分类模型 f 的权重参数在整个训练过程中都是固定的。

2.3 本章小结

本章对面向语音信号的神经网络模型的相关技术进行了介绍，具体的，将面向语音信号的神经网络模型分为语音分类模型和语音生成模型这两大类，针对语音分类模型，以说话人识别模型为例，对说话人识别模型各个发展阶段中的部分技术进行了概述，并对流行的基于 SincNet 的说话人识别模型进行了详细的介绍，为后面章节针对性地提出适用于语音分类模型的通用水印算法提供了理论支持。针对语音生成模型，以语音对抗样本生成模型为例，对生成模型中最常使用的 GAN 的基本概念进行了介绍，并对最新的基于 GAN 的语音对抗样本生成模型进行了详细的介绍，为后面章节针对性地提出适用于语音生成模型的通用水印算法提供了理论依据。

第三章 语音分类模型水印算法

本章旨在研究适用于语音分类模型的模型水印技术，鉴于语音分类模型主要应用在说话人识别和情感识别等分类任务中，本章以说话人识别模型为例，提出了一种通用的语音分类模型水印算法。

3.1 引言

模型水印是一种允许模型的合法所有者在训练阶段隐藏其身份在神经网络模型中以便后续进行身份认证的技术。鉴于神经网络模型知识产权保护的重要性和紧迫性，越来越多的学者致力于研究神经网络模型水印技术，并提出了一系列主流的模型水印方法，这极大地推动了神经网络模型水印技术的发展。然而，目前已有的研究大多都是针对图像相关的神经网络模型，关于说话人识别模型知识产权保护的研究还很少。此外，作为语音信号处理领域的一个研究方向，说话人识别常被应用在说话人核对、司法鉴定、语音检索、医学应用和军事领域等多个场景下，且越来越多的说话人识别模型都是基于神经网络的，为此，为基于神经网络的说话人识别模型设计一种水印方案，以保护这些模型的知识产权显得尤为重要。由于在实际应用场景中，说话人识别模型很容易被攻击者窃取并封装成 API 以牟取利益，此时，只能通过查询 API 进行交互，而无法获取某一商用产品中部署的说话人识别模型的内部细节。可见，保护部署在云服务器上的远程说话人识别 API 是一个重要需求。考虑到研究面向说话人识别模型的黑盒水印技术更符合实际应用场景，这促使本章为说话人识别模型设计一种黑盒设置下的神经网络模型水印方案。

本章提出了一种基于频域扰动的“黑盒”水印算法，该方法主要通过将构造的触发音频样本输入到目标模型中，得到对应的预测结果，并通过分析预测结果与预先指定的标签是否一致来对目标模型的所有权进行认证。该方法的思想已经成功地应用在了计算机视觉领域，由于本章的应用场景是语音领域，且图像和语音在处理方式上存在一定的差异，因而很难将已有的方法直接迁移用

于说话人识别模型的版权保护。此外，许多现有的方法中设计的触发信号具有容易被感知、容易被去除和容易受攻击等特点，当合法的模型所有者使用基于这些触发信号构造的触发样本进行查询时，很容易引起攻击者的警觉，攻击者可能会对这些触发样本进行拦截，从而使得查询无法进行下去，进而导致模型所有权认证失败，而本章所提出的方法可以有效应对这一问题。

具体的，要设计一个基于频域扰动的“黑盒”水印算法需要解决两个重要的问题，一是如何构造触发样本，二是如何高度保持说话人识别模型原始任务的性能。对于第一个问题，如果只是在原始音频样本上简单地添加一个明显的模式，例如预定义的标记或有意义的内容等，不仅可能会影响水印的不易察觉性，导致水印容易被去除或遭受攻击，还可能会导致攻击者伪造虚假的触发样本以混淆模型的所有权。针对这一问题，本章在频域上精心设计了触发音频样本，在不易察觉性和鲁棒性方面都取得了良好的性能。对于第二个问题，如果只是将构造的触发样本直接指定为正确标签以外的任意错误标签，很可能导致模型原始任务的性能显著降低。针对这一问题，本章在数据集现有标签的基础上添加了一个新标签，并将触发音频样本指定为该新标签，以尽可能地保持模型原始任务的性能。

3.2 基于频域扰动的“黑盒”水印算法

3.2.1 总体架构

本章所提出的方法包括三个阶段，触发样本生成阶段、水印嵌入阶段和水印验证阶段，图 3-1 展示了该方法的总体架构。其中，触发样本生成阶段的目标是设计一种通用的方法来构造两组触发音频样本，分别用于后续的水印嵌入阶段和水印验证阶段。需要注意的是，这两组触发音频样本互不相交。在触发样本生成阶段，首先对原始音频样本作离散余弦变换(Discrete Cosine Transform, DCT)，得到对应的频域系数；然后在频域系数上插入精心设计的触发信号，得到加载了水印信息的频域系数；最后对加载了水印信息的频域系数作逆离散余弦变换(Inverse Discrete Cosine Transform, IDCT)，得到触发音频样本。水印嵌

入阶段的目标是将水印信息嵌入到原始说话人识别模型中，从而得到含水印的说话人识别模型。在水印嵌入阶段，首先选取一组触发样本生成阶段生成的触发音频样本，并将这些样本指定为不包含在原始标签集中的新增标签，从而构造出用于训练的触发音频数据集；然后将该触发音频数据集与原始的数据集结合起来构成新的数据集，并利用新的数据集对原始说话人识别模型进行从头训练来嵌入水印；最后训练完成的模型被视为含水印的说话人识别模型，可以投入使用。水印验证阶段的目标是在含水印的模型发生泄漏的时候，模型的合法所有者能够在不知晓模型内部细节的情况下对模型的所有权进行认证。在水印验证阶段，首先选取另一组触发音频样本，并将这些样本指定为新增的标签类别，从而构造出用于验证的触发音频数据集；然后将这些样本输入到目标模型中，并查看模型的预测结果；最后通过分析目标模型的预测结果与预先指定的标签是否一致来认证模型的所有权。

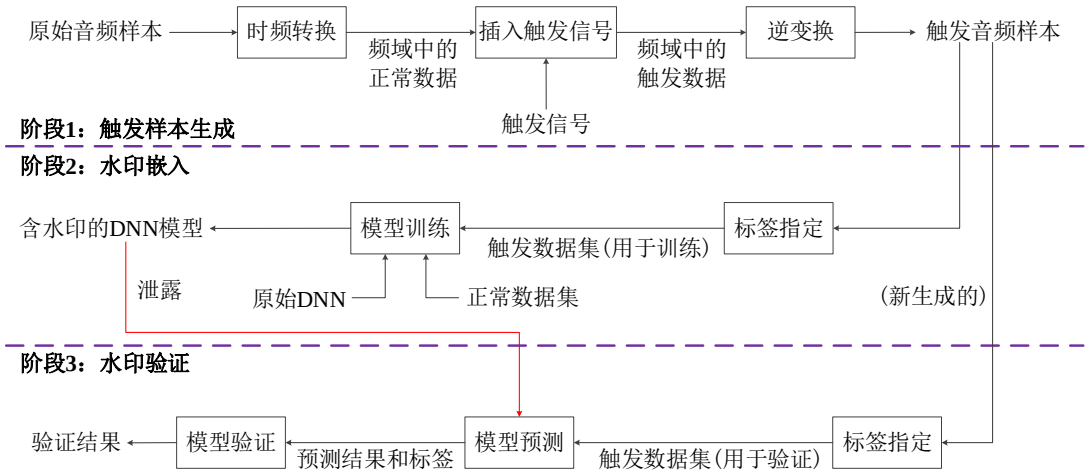


图3-1 基于频域扰动的“黑盒”水印算法的总体架构

3.2.2 触发音频样本的生成方法

传统的媒体水印技术[72, 73]主要包括空间域水印技术和变换域水印技术这两大类，变换域水印技术也被称为频域水印技术。在空间域水印技术中，水印信息是被直接加载到载体数据中的。以语音信号为例，直接操纵载体语音时域波形各个时刻的幅值，如直接在载体语音时域波形中加入高斯噪声，就是一种空间域水印方法。在变换域水印技术中，水印信息是被加载到载体数据变换域

的系数上的，其中，变换域主要包括傅立叶变换域、余弦变换域和小波变换域等。以语音信号为例，利用数字信号处理的相关知识对载体语音进行离散傅里叶变换，并将水印信息(如文本信息或图片信息等)嵌入到变换后所得到的频域系数中的方法就是一种变换域水印方法。近年来，有学者提出可以对传统的媒体水印技术的思想进行拓展，并运用到触发样本的构造上。

例如，在目前已有的部分基于触发样本的模型水印技术中，触发样本的构造是通过在原始样本(也称为干净样本)的空间域中添加一个明显的模式(也称为触发信号)来完成的。很显然，这是将空间域水印的技术拓展到触发样本的构造上的体现。然而，正如上文所提到的，该方法下构造的触发样本存在一定的被攻击的风险，主要包括：(1)这种明显的模式可能会引起攻击者的怀疑，而使得攻击者以类似的方式伪造触发样本，从而混淆模型的所有权；(2)攻击者一旦识别出触发模式，很有可能会对触发模式进行攻击，例如在触发样本中加入噪声，从而使得模型的合法所有者无法成功进行认证。可见，在空间域水印的启发下构造的触发样本的鲁棒性较差。

考虑到频域水印技术在多个方面均优于空间域水印技术[74]，主要包括：(1)频域水印技术是在载体内容的频域系数上加载水印的，经过逆变换操作后，频域系数上所加载的水印信息会相对应的分布在载体内容时域空间的各个位置上，这一特点保证了水印的不易察觉性；(2)人类感知系统只对特定频率范围内的变化比较敏感，因此在人类感知系统不太敏感的频段加载水印信息可以相对容易地躲避人类感知系统对变化或异常的捕捉；(3)使用频域水印技术能够在压缩域中实现水印算法的同时，还能抵抗相应的有损压缩。为此，本章选择拓展频域水印的思想来构造更具鲁棒性的触发样本。

具体的，本章中的触发音频样本主要通过以下步骤来进行构造，首先对原始音频样本作离散余弦变换，得到频域系数；然后，在频域系数上添加基于分段的触发信号，得到加载了水印信息的频域系数；最后，对加载了水印信息的频域系数作逆离散余弦变换，得到触发音频样本。在上述的过程中，触发信号被预先指定为仅包含元素 1、0 和 –1 的长度为 l（$l>0$）的随机序列 $\boldsymbol{t}$。同时，对于触发信号的加载方式，作出了以下思考。通常而言，在语音相关任务中，会对音频样本进行分帧处理，本章中的任务也不例外。需要注意的是，对本章所

构造的触发音频样本进行分帧操作后，无论其各帧是否携带触发信号，神经网络模型对于不同帧所进行的后续训练和测试的操作都是相互独立的。考虑到本章任务中，需要收集神经网络模型在触发音频样本每一帧上的预测结果以便后续进行所有权认证，为此，使得触发音频样本的每一帧都携带触发信号就显得尤为重要。如果只将 $\boldsymbol{t}$ 嵌入到原始音频样本频域的某些片段中，则触发信号只会影响触发音频样本的某些帧而不是所有的帧，此时不携带触发信号的帧的总数很可能显著高于携带触发信号的帧的总数，从而导致认证结果被不携带触发信号的帧所误导，最终使得所有权认证失败。如果只在原始音频样本频域的特定系数中嵌入一次 $\boldsymbol{t}$，则无法保证不同的触发音频样本上的触发信号的特征模式是一致的，从而使得神经网络模型很难学习到准确的触发信号的模式，最终也会导致所有权认证失败。为此，本章提出在原始音频样本频域的每个选定片段内嵌入 $\boldsymbol{t}$，以使神经网络模型在训练的过程中可靠地学习到触发信号与对应标签之间的映射关系，进而在模型所有权认证的时候作出准确的预测。

触发音频样本的生成过程在数学上表示如下，假设已知一个原始音频样本表示为 $\boldsymbol{x}$，且 $\boldsymbol{x}=\{x(0),x(1),...,x(L-1)\}\in\mathbb{R}^L$。首先将其平均分割成 $R=\lceil L/l \rceil$ 个片段，$\lceil\cdot\rceil$ 表示向上取整，l 为每一个小片段的长度，如果 l 不能整除 L，则最后一个片段的长度为 $L-(R-1)\cdot l$。分段后，$\boldsymbol{x}$ 就变成了 $\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_R\}$。

其次，对 $\boldsymbol{x}_r$（$1\le r\le R$）作离散余弦变换，得到频域系数 $\boldsymbol{y}_r$[75, 76]，公式为：

$$\begin{aligned}\boldsymbol{y}_r(u)&=\mathrm{c}(u)\sum_{i=0}^{l-1}\boldsymbol{x}_r(i)\cos\left[\frac{(i+0.5)\pi}{l}u\right]\quad(0\le u<l)\\ \mathrm{c}(u)&=\begin{cases}\sqrt{1/l}, & u=0;\\ \sqrt{2/l}, & u\ne 0.\end{cases}\end{aligned}\tag{3-1}$$

再次，将 $\boldsymbol{t}$ 嵌入到 $\boldsymbol{y}_r$ 中，公式为：

$$\boldsymbol{y}_r'=\boldsymbol{y}_r+\lambda\cdot\boldsymbol{t}\tag{3-2}$$

其中，λ 是一个标量，表示所嵌入的 $\boldsymbol{t}$ 的强度。本章中，λ 被选定为 0.001，这是经过多次实验并比较实验结果后所选出的最优系数。同时，为了避免人为改变 $\boldsymbol{y}_r$ 中的直流分量，将 $\boldsymbol{t}$ 中的第一个元素指定为 0。此外，需要注意的是，如果 l 不能整除 L，则只需在 $\boldsymbol{y}_R$ 上加载 $L-(R-1)\cdot l$ 个比特的触发信号。

然后，对 $\boldsymbol{y}'_r$ 作逆离散余弦变换，得到与原始音频片段 $\boldsymbol{x}_r$ 相对应的触发音频片段 $\boldsymbol{x}'_r$，公式为：

$$\boldsymbol{x}'_r(i)=\sqrt{\frac{1}{2}}\mathrm{c}(i)\sum_{u=0}^{l-1}\boldsymbol{y}'_r(u)\cos[\frac{(i+0.5)\pi}{l}u]\ \ (0\le i<l)$$
$$\mathrm{c}(i)=\begin{cases}\sqrt{1/l}, & i=0;\\ \sqrt{2/l}, & i\neq 0.\end{cases} \tag{3-3}$$

最后，将所有的触发音频片段连接起来，构造出与原始音频样本 $\boldsymbol{x}$ 相对应的触发音频样本 $\boldsymbol{x}'$。

由于触发音频样本在后续阶段将被用于训练神经网络模型，因而有必要为其分配一个合理的标签。如果将触发音频样本指定为现有的说话人类别集中的除正确的说话人标签以外的任意标签[77]，这显然是违反人类直觉的。同时，由于触发音频样本的音色与其所对应的原始音频样本的音色是难以区分的，这样做还有可能会让神经网络模型原始任务的性能下降。为了保持神经网络模型原始任务的性能，且让水印验证过程足够方便且合理，本章为触发音频样本分配了一个新的标签。假定原来的标签集为 $\boldsymbol{C}=\{0,1,2,...,c-1\}$，其中 $c\ge 1$，则新的标签集更新为 $\boldsymbol{C}'=\{0,1,2,...,c\}$。

3.2.3 水印的嵌入过程

水印嵌入的目标是将水印嵌入到原始的说话人识别模型中，得到一个跟原始模型足够接近的含水印的说话人识别模型。假设原始模型为 M，含水印的模型为 M^*，则满足 $M^*\approx M$。同时要求 M^*不仅在原始的说话人识别任务上保持很好的表现，还能学习到触发音频样本与对应标签之间的映射关系。具体的，本章中将原始音频样本和一组触发音频样本相结合，对 M 进行从头训练，从而得到 M^*。其中，每个原始音频样本与正确的标签相关联，每个触发音频样本与上文所提到的新增标签 c 相关联。在模型训练期间，对于每一次的迭代，都是将一小批随机选取的原始音频样本与触发音频样本输入到模型中，进行每一轮的优化。需要注意的是，在对 M 进行训练之前，需要稍微修改 M 的 softmax 层，在该层中添加一个与触发音频样本相对应的新标签类别。当 M 训练完成后，得

到了新的模型 M^*，该模型即为含水印的模型，并可投入使用。

3.2.4 水印的验证过程

水印验证的目标是使得模型的合法所有者将一组新的触发音频样本输入到目标模型中之后，能够通过分析模型的预测结果与目标标签是否一致来认证目标模型的所有权。假设总共有 n 个用于验证的触发音频样本$\{\boldsymbol{x}_0'', \boldsymbol{x}_1'', ..., \boldsymbol{x}_{n-1}''\}$，它们所对应的预测结果为$\{M^*(\boldsymbol{x}_0''), M^*(\boldsymbol{x}_1''), ..., M^*(\boldsymbol{x}_{n-1}'')\}$，需要注意的是，对于任意的$0 \le i \le n-1$，总存在$M^*(\boldsymbol{x}_i'') \in \boldsymbol{C}'$。为此，模型的所有权认证可以表示为：

$$\sum_{i=0}^{n-1} \delta(M^*(\boldsymbol{x}_i''), c) \ge \theta \cdot n \tag{3-4}$$

其中，c 为新增的说话人标签；n 为用于验证的触发音频样本的个数；θ 为预先指定的阈值，取值范围在[0,1]区间内，且θ越接近于 1，说明本章的方法对水印验证任务的能力要求越高。对于式(3-4)中的$\delta(x, y)$函数，预先指定以下的规则，当$x = y$时，$\delta(x, y) = 1$；当$x \ne y$时，$\delta(x, y) = 0$，其中，x 和 y 表示形式参数。综上，当式(3-4)成立时，视为所有权认证成功；否则，视为认证失败。

3.3 实验结果与分析

在本节中，将对实验的设置进行介绍，对实验的结果进行展示，并从原始的说话人识别任务的性能、水印验证任务的能力和鲁棒性这三个方面对本章所提出的方法进行评估。

3.3.1 实验的设置

本章的全部实验均是在流行的 TIMIT[78]数据集的基础上进行的，实验中使用了该数据集中的 462 个说话人的 3696 个音频样本。对于这些音频样本，80%用于模型的训练，20%用于模型的测试。其中，在用于训练的样本中，划分出 1/8 充当验证集，用于选择最佳的模型。在此基础上，本章还随机选取了 TIMIT 数据集中的对应于 16 个说话人的 36 个原始音频样本，并在这 36 个原始音频样

本的基础上构造出了 36 个新的触发音频样本。其中，16 个触发音频样本以 7:1 的比例被划分进原始的训练集和验证集中用于完成水印的嵌入任务。另外 20 个触发音频样本被全部划分进原始的测试集中用于完成水印的验证任务。至此，新的数据集构造完成。本章的全部实验均是在 Mirco 等所提出的流行的基于 SincNet 的说话人识别模型[63]的基础上进行的，需要指出的是，本章所提出的“黑盒”水印算法是可以扩展到其他语音分类模型上的，因而，实验中可以自行选择想要保护的神经网络模型。

实验过程中，为了保证实验数据的有效性，对包括触发音频样本在内的所有音频样本都进行了预处理操作，主要是剔除了每个音频样本开头和结尾处的无声片段。之后对预处理后的音频样本按照帧长为 200 毫秒，帧重叠为 10 毫秒的设置进行分帧处理，分帧后得到的音频片段被用作模型的输入。实验过程中不改变基于 SincNet 的说话人识别模型的网络结构和超参数等设置，以便将本章所提出的方法的实验结果与文献中方法的实验结果进行对比。其中，对于参数的优化选用的是 RMSprop 优化器，学习率的大小为 0.001，batch 的大小为 128，epoch 的大小为 360。选择 Leaky ReLU[79]作为模型所有隐藏层中所使用的激活函数，主要是因为 Leaky ReLU 不仅可以实现非线性映射，还能解决特定情况下某些神经元无法被激活的问题。

对于说话人识别任务，本章采用帧级错误率（Frame-Level Error Rate, FER）和句子级错误率（Sentence-Level Error Rate, SER）这两个评价指标。其中，帧级说话人分类的结果是通过 softmax 层获得的，该层为音频样本的每一帧提供一组说话人标签的后验概率。句子级说话人分类的结果是通过投票的方式获得的，主要是将所有帧所对应的预测标签中出现频率最高的那一个预测标签指定为某一音频样本的最终预测标签。可以看出，FER 和 SER 的值越低，表示说话人识别任务的性能越好。

3.3.2 说话人识别性能评估

为了评估所提出的方法在原始说话人识别任务上的性能，实验中将原始的测试样本集分别输入到不含水印的模型 M 和含水印的模型 M^*中，并获得了对

应的帧级错误率 FER 和句子级错误率 SER，实验结果如表 3-1 所示。从实验结果中可以看出，从不含水印的模型 M 处获得的 FER 和 SER 分别为 47.97%和 0.58%，从含水印的模型 M^*处获得的 FER 和 SER 分别为 48.83%和 0.65%。从 M^*处获得的 FER 和 SER 会比从 M 处获得的 FER 和 SER 略高，这是符合常理的，主要是因为水印的嵌入会使得神经网络模型中的参数不仅要完成说话人识别任务还要学习触发信号与对应标签之间的映射关系，那么水印的嵌入必然要以说话人识别任务性能的下降为代价。由于含水印的模型 M^*在说话人识别任务上的准确率为 99.35%，只比不含水印的模型 M 所对应的准确率低了 0.07%。可见，本章所提出的方法能够很好地保持原始说话人识别任务的性能。

表3-1 在原始说话人识别任务上的性能比较

	不含水印的模型 M	含水印的模型 M^*
FER	47.97%	**48.83%**
SER	0.58%	**0.65%**

3.3.3 水印评估

为了评估所提出的方法在水印验证任务上的性能，实验中将用于验证的触发音频样本输入到 M^*中，并分析 M^*的预测结果与指定的标签是否一致，实验结果如表 3-2 所示。从实验结果中可以看出，用于验证的触发音频样本从 M^*处获得的 FER 和 SER 分别为 18.85%和 5.00%，经过分析可知，20 个用于验证的触发音频样本中只有一个无法成功地认证模型的所有权，其余的均能成功地认证模型的所有权，水印验证的成功率为 95.00%，由此证明本章所提出的方法具有很强的水印验证能力。

表3-2 在水印验证任务上的性能评估

	用于验证的触发音频样本
FER	18.85%
SER	5.00%
成功率	95.00%

3.3.4 鲁棒性分析

考虑到攻击者很有可能会攻击用于水印验证的触发音频样本，使得这些样本丧失对神经网络模型的版权认证能力，为此，本节中对所提出的方法的鲁棒性进行了评估。实验中通过插入噪声来模拟现实世界中的攻击场景，所选用的噪声为高斯噪声，其概率密度函数服从正态分布，并通过指定不同的信噪比（Signal-Noise Ratios, SNRs），来获得不同强度的高斯噪声。通过在触发音频样本中加入不同强度的高斯噪声，并将其输入到含水印的模型 M^*中，可以获得不同攻击强度下的触发音频样本的水印验证成功率。需要注意的是，本节选取了上文中能够成功验证水印的触发音频样本进行鲁棒性实验，实验结果如表 3-3 所示。从实验结果中可以看出，当 SNR 为 20dB 或低于 20dB 时，触发音频样本从 M^*处获得的 SER 均为 0%，即水印验证的成功率均为 100%。可见，即使在攻击程度较强的情况下，本章所提出的方法依然能够成功地完成水印验证任务，说明该方法具有一定的鲁棒性。

表3-3 在水印验证任务上的鲁棒性评估

SNR（dB）	SER	成功率
5	0%	100%
10	0%	100%
15	0%	100%
20	0%	100%

3.3.5 对比分析

本节简单地尝试了将空间域水印的策略运用到触发音频样本的构造上，生成了基于时域扰动的触发音频样本（也称为时域触发音频样本）。具体而言，这些时域触发音频样本是通过在原始音频样本的时域波形中直接加载信噪比为 20dB 的高斯噪声构造而成的。将时域触发音频样本和原始音频样本结合起来，对不含水印的模型 M 进行从头训练，从而得到含水印的模型 M^{**}，将该方法称为基于时域扰动的水印算法。本节将从原始说话人识别任务的性能和水印验证任务的能力这两个方面对该方法与本章所提出的方法进行对比。

表 3-4 展示了基于时域扰动的水印算法与本章所提出的基于频域扰动的水印算法在原始说话人识别任务上的性能，其中第二列所示的实验结果是从含水印的模型 M^*处获得的，FER 和 SER 分别为 48.83%和 0.65%，第三列所示的实验结果是从含水印的模型 M^{**}处获得的，FER 和 SER 分别为 48.89%和 0.72%。从实验结果中可以看出，基于时域扰动的水印算法虽然同样能够很好地保持说话人识别任务的性能，说话人识别准确率为 99.28%，但是与本章所提出的基于频域扰动的水印算法相比，该方法使原始说话人识别任务的性能下降更多。具体来说，与不含水印的模型 M 的性能相比，本章所提出的方法在帧级错误率上提高了 0.86%，在句子级错误率上提高了 0.07%；而该方法在帧级错误率上提高了 0.92%，在句子级错误率上提高了 0.14%。

表3-4 在原始说话人识别任务上的性能比较

	含水印的模型 M^*	含水印的模型 M^{**}
FER	48.83%	**48.89%**
SER	0.65%	**0.72%**

表 3-5 展示了基于时域扰动的水印算法与本章所提出的基于频域扰动的水印算法在水印验证任务上的性能，其中第二列所示的实验结果是从含水印的模型 M^*处获得的，FER 和 SER 分别为 18.85%和 5.00%，第三列所示的实验结果是从含水印的模型 M^{**}处获得的，FER 和 SER 分别为 20.78%和 5.00%。从实验结果中可以看出，基于时域扰动的水印算法虽然同样具有很强的水印验证能力，水印验证成功率为 95.00%，但是该方法对应的帧级错误率比本章所提出的方法对应的帧级错误率高大约 10 个百分点。总的来说，本章所提出的基于频域扰动的水印算法在水印验证任务上更具优势。

表3-5 在水印验证任务上的性能比较

	含水印的模型 M^*	含水印的模型 M^{**}
FER	18.85%	**20.78%**
SER	5.00%	**5.00%**
成功率	95.00%	**95.00%**

3.4 本章小结

本章提出了一种基于频域扰动的“黑盒”水印算法，该算法能够用于保护任意的语音分类模型的知识产权。本章首先介绍了该方法的设计思想和总体架构，然后对总体架构中的触发样本生成阶段、水印嵌入阶段和水印验证阶段的具体步骤和详细过程进行了介绍，最后从原始的说话人识别能力、水印验证能力和鲁棒性这几个方面对所提出的方法进行了评估，同时，还将该方法与基于时域扰动的水印算法进行了对比。实验结果表明，该方法不但能够很好地保持原始的说话人识别任务的性能，而且具有很强的模型所有权认证能力。此外，该方法还具有一定的抗噪声攻击的能力。

第四章 语音生成模型水印算法

本章旨在研究适用于语音生成模型的模型水印技术，鉴于语音生成模型主要应用在语音合成、语音转换和语音对抗样本生成等任务中，本章以语音对抗样本生成模型为例，提出了一种通用的语音生成模型水印算法。

4.1 引言

近年来，计算机硬件计算能力得到了较大的提升，大规模数据集也逐渐涌现出来，这促使深度学习技术在与图像、语音和文本等相关的多个具有挑战性的任务上实现了巨大的突破。人们从基于深度学习的应用中受益良多，然而，深度学习技术的快速发展也让神经网络模型面临着比较严峻的安全性问题。最近有研究表明，神经网络模型易于受到对抗样本的攻击[80, 81]。其中，对抗样本指的是在原始样本(也称为干净样本)中有目的地添加微小的且不易察觉的对抗扰动后所得到的样本，对抗样本能够导致神经网络模型原始任务的性能显著降低。以语音识别为例，攻击者可以针对任意的原始语音样本构造出相对应的语音对抗样本，将该语音对抗样本输入到自动语音识别模型中，模型会输出攻击者预先指定的任意句子[82]。可见，语音对抗样本会威胁面向语音信号的神经网络模型的安全。鉴于面向语音信号的任务范围广泛，有必要对语音对抗样本进行研究，这对于保障面向语音信号的神经网络模型的安全具有显著意义。

需要注意的是，攻击和防御是相对应的，对具有更强攻击能力的语音对抗样本进行研究也推动着更多高效的防御方法的诞生，看似负面的语音对抗样本也有积极的应用，可以被用于提升面向语音信号的神经网络模型的鲁棒性。例如，有学者提出可以将语音对抗样本加入原始的训练集中[83]，通过对模型进行微调来提升模型的鲁棒性。从这一角度看，语音对抗样本及其所对应的语音对抗样本生成模型都是积极的且应该受到保护的。然而，目前的语音对抗样本生成算法主要致力于研究如何生成具有良好攻击能力且不易被察觉的语音对抗样本，并没有考虑到对语音对抗样本生成模型进行保护，这促使本章为语音对抗

样本生成模型提出一种水印算法，以保护语音对抗样本生成模型的知识产权。

对语音对抗样本生成模型水印技术进行研究需要借助已有的模型，目前基于深度学习的语音对抗样本生成模型大多存在以下缺点：(1)语音对抗样本生成的过程中对计算能力的要求过高，不利于实际的应用；(2)语音对抗样本中的对抗扰动过大，不具备良好的隐蔽性。而 Wang 等[71]针对以上不足，设计了一个基于 GAN[84]的语音对抗样本生成模型，他们将语音分类模型与 GAN 相结合，精心设计了网络结构、损失函数和训练策略，最终训练出了一个生成器，该生成器可以快速地生成与原始语音样本一一对应的对抗扰动，且对抗扰动不易使人察觉。鉴于此，本章以 Wang 等所设计的基于 GAN 的语音对抗样本生成模型为例，提出了一种通用的水印算法，该算法也可以推广到其他语音生成模型上。

4.2 输出结果带水印的“无盒”水印算法

4.2.1 总体架构

本章提出了一种输出结果带水印的“无盒”水印算法，该算法旨在训练出一个嵌有水印的生成器，且该生成器具有生成扰动信号的能力，进而可以得到带水印的语音对抗样本，使得模型的合法所有者可以从带水印的语音对抗样本中提取出水印信息，并用于生成器所有权的认证。该方法的总体架构如图 4-1 所示，主要包括四个不同的神经网络模型，生成器(用 G 表示)、鉴别器(用 D 表示)、水印网络(用 E 表示)和语音分类模型(用 F 表示)。其中，生成器 G 用于生成扰动信号，进而得到语音对抗样本；鉴别器 D 用于区分语音对抗样本与原始语音信号；水印网络 E 用于向生成器 G 中嵌入水印，同时用于从生成器 G 间接生成的语音对抗样本中提取水印；语音分类模型 F 用于对语音对抗样本进行分类。该架构的总体流程如下，首先将原始语音信号 $\boldsymbol{x}$ 输入到生成器 G 中，得到扰动信号 $\boldsymbol{\delta}=G(\boldsymbol{x})$；其次，将扰动信号 $\boldsymbol{\delta}$ 添加到原始语音信号 $\boldsymbol{x}$ 中，从而得到语音对抗样本 $\boldsymbol{x}_{\text{wa}}=\boldsymbol{x}+\boldsymbol{\delta}$；然后，将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分别输入到鉴别器 D、水印网络 E 和语音分类模型 F 中，从而得到相对应的结果，并计算出相对应的损失函数；最后，将各个损失函数进行组合，通过对组合损失函数不断进行优

化，从而得到比较好的生成器G'。其中，优化鉴别器 D 所对应的损失函数可以促使语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 与原始语音信号 $\boldsymbol{x}$ 足够接近，即让语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的失真足够小；优化水印网络 E 所对应的损失函数可以促使生成器 G 中被嵌入水印，同时促使生成器 G 间接生成的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中携带水印；优化语音分类模型 F 所对应的损失函数可以促使语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 具有攻击能力。

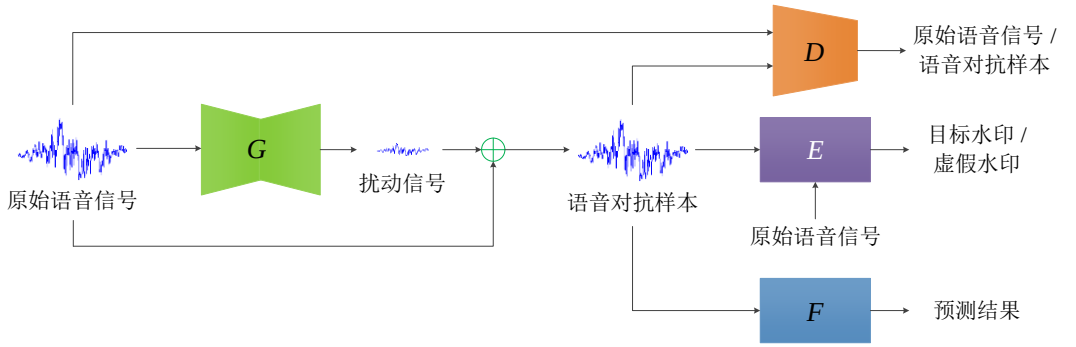


图4-1 输出结果带水印的“无盒”水印算法的总体架构

4.2.2 网络结构的设计

生成器 G 的目标是从原始语音信号 $\boldsymbol{x}$ 中自适应地生成扰动信号 $\boldsymbol{\delta}$，且需要满足将扰动信号 $\boldsymbol{\delta}$ 添加到原始语音信号 $\boldsymbol{x}$ 中之后所得到的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 在听觉感知质量上接近原始语音信号 $\boldsymbol{x}$，否则，很容易引起攻击者的怀疑。考虑到编解码器结构已经成功应用在多个任务中，例如编解码器结构 SEGAN[85]用在了语音增强任务中，编解码器结构 SegNet[86]用在了语义分割任务中，为此，本章中的生成器 G[71]也采用类似的编解码器结构。具体而言，编码器部分包含 8 个卷积层，每个卷积层中均有多个卷积核，且核尺寸为 1×32、步长为 2。解码器部分包含 8 个反卷积层，每个反卷积层中均有多个反卷积核，且核尺寸为 1×32、步长为 2。生成器 G 中的编解码器部分均使用 PReLU[87]激活函数，主要是为了训练的稳定性；生成器 G 的输出层中使用的激活函数为 tanh，主要是为了将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的幅值范围限制在有效语音样本的幅值区间$[-1,1]$内。受 SEGAN 的启发，生成器 G 中也使用了跳跃连接(Skip Connection)，跳跃连接可以将编码器中的卷积层与其所对应的解码器中的反卷积层连接起来，使得反卷积层直接共享由卷积层提取的特征。同时，使用跳跃连接还能够缓解深层网络

中容易出现的梯度消失或梯度爆炸问题，使得梯度能够在神经网络中流动。

鉴别器 D 的目标是促使生成器 G 生成人类听觉系统不易察觉的扰动信号 $\boldsymbol{\delta}$，从而使得加载了扰动信号 $\boldsymbol{\delta}$ 的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 足够自然。鉴别器 D 的主要任务是对原始语音信号 $\boldsymbol{x}$ 和语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 进行区分。为此，可以选择任意的能够用于语音信号分类的神经网络模型作为鉴别器 D，本章中使用的鉴别器 D 由 12 个卷积模块和 1 个全连接层[71]组成。具体而言，前 11 个卷积模块中依次是卷积(conv)层、批归一化(Batch Normalization)层和 Leaky ReLU 激活层[88]，且每个卷积层中均有多个卷积核，核尺寸为 1×31。最后 1 个卷积模块中依次是卷积层和 Leaky ReLU 激活函数，且卷积层中只有 1 个卷积核，核尺寸为 1×31，这主要是为了将所有的特征图压缩成一个一维的向量。全连接层所使用的激活函数为 softmax，输出[0,1]区间内的结果，表示输入的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 被分类为原始语音信号 $\boldsymbol{x}$ 的概率。

水印网络 E 的目标是在训练的时候向生成器 G 中嵌入水印，且在实际应用的时候用于从语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中提取水印。本章中所使用的水印网络 E 的网络结构如图 4-2 所示，包括 2 个卷积模块和 16 个 inception-residual 模块[89]，其中，每一个卷积模块均由卷积层、批归一化层、激活层和最大池化层组成。选择使用 inception-residual 模块是因为其不仅能够在不降低模型性能的前提下加速训练，还能够为自由选择更好的特征提供更广阔的网络结构；此外，有研究成果[90]表明 inception-residual 模块已经能够成功地将生成的图像中的模式划分到不同的通道中，鉴于本章中需要处理类似的场景，为此本章也将该模块包含在水印网络 E 中，inception-residual 模块的详细结构信息如图 4-3 所示。为了确定水印网络 E 中具体包含的 inception-residual 模块的数量，本章进行了多次实验，实验结果表明，当使用较少的 inception-residual 模块时，水印网络 E 从语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中提取到的水印信息与目标水印之间差异很大，计算得到的误码率很高，说明此时水印嵌入失败。当使用更多的 inception-residual 模块时，虽然误码率很低，但是需要耗费更多的计算资源。经过权衡之后，本章最终确定在水印网络 E 中设置 16 个 inception-residual 模块。需要注意的是，水印网络 E 的输入是一个长度为 l_{sequence} 的语音序列，实验过程中需要将该尺寸调整为

$h_{\text{input}}\times(l_{\text{sequence}}/h_{\text{input}})\times 1$。水印网络 E 的输出是一个尺寸为 $h_w\times w_w\times 1$ 的二进制矩阵，表示水印图像。本章中，h_w 和 w_w 均被设置为 32。

语音分类模型 F 的目标是促使生成器 G 间接生成的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 具有攻击能力。由于语音分类模型 F 在训练过程中只负责提供其所对应的损失函数，其本身的参数是不参与更新的，为此，本章中直接将已经训练好的语音分类模型 SampleCNN[91]作为 F。

需要指出的是，对于上述的生成器 G、鉴别器 D 和水印网络 E，它们结构的选择不是唯一的，而是可以自由地进行设计的。

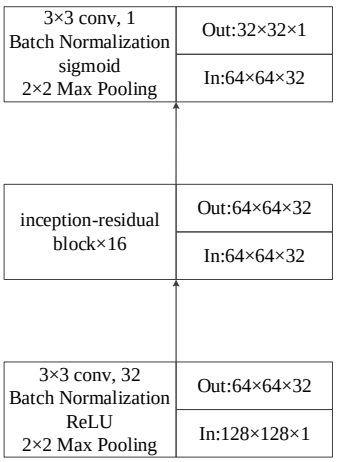


图4-2 水印网络 E 的网络结构

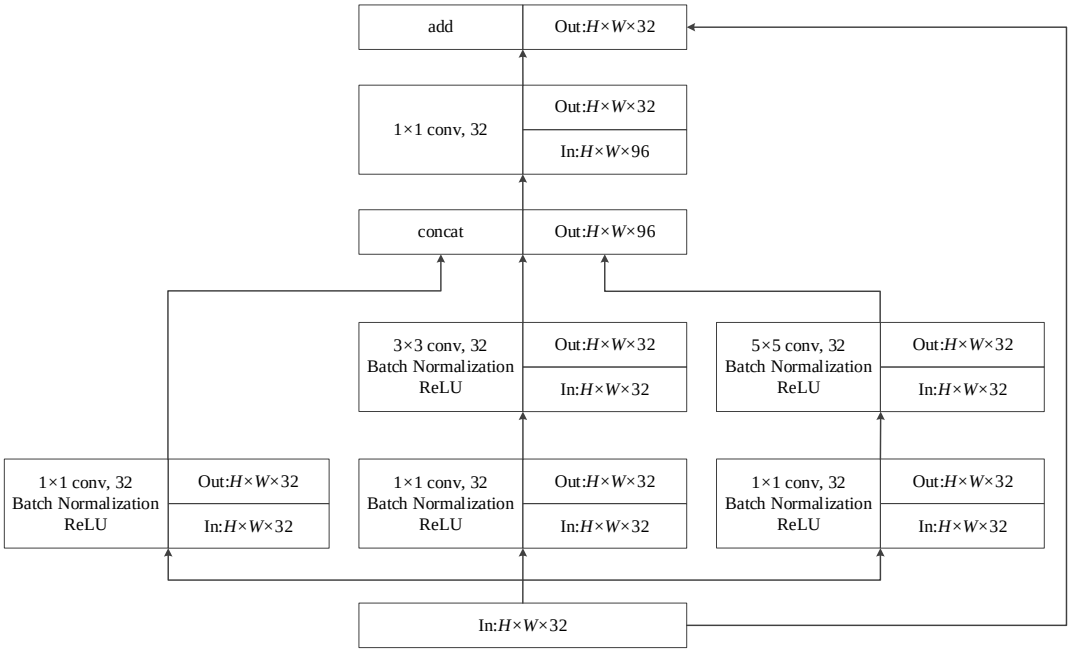


图4-3 inception-residual 模块的详细结构信息

4.2.3 损失函数的设计

对于生成器 G，主要有三个约束条件，包括：(1)误导语音分类模型 F 将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分类为指定的目标标签；(2)欺骗鉴别器 D 使其在原始语音信号 $\boldsymbol{x}$ 和语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 之间无法区分；(3)生成携带水印信息的扰动信号 $\boldsymbol{\delta}$。

第一个约束是通过优化攻击损失 L_{attack} 来实现的，L_{attack} 用于衡量语音分类模型 F 对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的预测结果与目标标签之间的差异，被定义为：

$$L_{\text{attack}} = \text{E}[l_{\text{ce}}(F(\boldsymbol{x}_{\text{wa}}), \boldsymbol{y}_{\text{t}})] \tag{4-1}$$

其中，E[·]表示总体期望，$F(\boldsymbol{x}_{\text{wa}})$ 表示语音分类模型 F 对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的预测结果，$\boldsymbol{y}_{\text{t}}$ 表示 $\boldsymbol{x}_{\text{wa}}$ 被事先指定的目标标签，是一个 one-hot 向量，$l_{\text{ce}}(\cdot)$ 表示交叉熵损失函数。通过对式(4-1)进行优化可以最小化语音分类模型 F 对语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的预测结果与目标标签 $\boldsymbol{y}_{\text{t}}$ 之间的差异，进而促使生成器 G 生成一个合适的扰动信号 $\boldsymbol{\delta}$，使得加载了该扰动信号 $\boldsymbol{\delta}$ 的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 具有攻击能力。

第二个约束是通过优化欺骗损失 L_{fool} 来实现的，L_{fool} 用于衡量鉴别器 D 将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分类为原始语音信号 $\boldsymbol{x}$ 的成本，被定义为：

$$L_{\text{fool}} = \text{E}[\log(1 - D(\boldsymbol{x}_{\text{wa}}))] \tag{4-2}$$

其中，$D(\boldsymbol{x}_{\text{wa}})$ 表示鉴别器 D 将语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 分类为原始语音信号 $\boldsymbol{x}$ 的概率。通过对式(4-2)进行优化可以最小化语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 与原始语音信号 $\boldsymbol{x}$ 之间的差异，进而促使生成器 G 生成一个合适的扰动信号 $\boldsymbol{\delta}$，使得加载了该扰动信号 $\boldsymbol{\delta}$ 的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的分布和原始语音样本 $\boldsymbol{x}$ 的分布趋于一致。

第三个约束是通过优化水印提取损失 L_E 来实现的，L_E 用于衡量水印网络 E 从语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中提取到的水印与目标水印之间的差异以及水印网络 E 从其他语音信号 $\boldsymbol{x}_{\text{other}}$ 中提取到的水印与虚假水印之间的差异，被定义为：

$$L_E = \text{E}[l_{\text{mse}}(E(\boldsymbol{x}_{\text{wa}}), \boldsymbol{w})] + \text{E}[l_{\text{mse}}(E(\boldsymbol{x}_{\text{other}}), \boldsymbol{w}_{\text{fake}})] \tag{4-3}$$

其中，$\boldsymbol{x}_{\text{other}}$ 表示未携带水印的任意语音信号，$E(\boldsymbol{x}_{\text{wa}})$ 和 $E(\boldsymbol{x}_{\text{other}})$ 分别表示水印网络 E 从语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中提取到的水印和水印网络 E 从其他任意语音信号

$\boldsymbol{x}_{\text{other}}$ 中提取到的水印，$\boldsymbol{w}$ 表示目标水印，$\boldsymbol{w}_{\text{fake}}$ 表示虚假水印，在本章中被指定成一个尺寸为32×32的全零矩阵，$l_{\text{mse}}(\cdot)$ 表示均方误差（Mean Squared Error, MSE）损失函数。通过对式(4-3)进行优化可以最小化水印网络 E 从语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中提取到的水印与目标水印 $\boldsymbol{w}$ 之间的差异，进而促使生成器 G 生成一个合适的扰动信号 $\boldsymbol{\delta}$，使得加载了该扰动信号 $\boldsymbol{\delta}$ 的语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 中携带水印信息。

对于生成器 G，除了上述三个约束条件外，还需要对生成器 G 生成的扰动信号 $\boldsymbol{\delta}$ 进行约束。为了使得扰动信号 $\boldsymbol{\delta}$ 具有良好的隐蔽性，使用 L_2 损失函数来约束扰动信号 $\boldsymbol{\delta}$ 的能量大小，公式为：

$$L_{\delta} = \left\| \boldsymbol{\delta} \right\|_2 \tag{4-4}$$

为了使得扰动信号 $\boldsymbol{\delta}$ 在原始语音信号 $\boldsymbol{x}$ 的采样点上是稀疏存在的，使用铰链损失 L_{hinge} [85]来约束扰动信号 $\boldsymbol{\delta}$ 的能量分布，公式为：

$$L_{\text{hinge}} = \text{E}[\max(0, \left\| \boldsymbol{\delta} \right\|_2 - c)] \tag{4-5}$$

其中，c 表示为 $\left\| \boldsymbol{\delta} \right\|_2$ 设置的阈值，以确保 $\left\| \boldsymbol{\delta} \right\|_2$ 的最大波动范围不会超过该阈值。

综上，生成器 G 所对应的损失函数 L_G 为：

$$L_G = L_{\text{attack}} + \alpha L_{\text{fool}} + \beta L_{\text{hinge}} + \gamma L_{\delta} + \varepsilon L_E \tag{4-6}$$

其中，α、β、γ 和 ε 均为标量，用于调节各个子项的重要程度。

考虑到鉴别器 D 主要是用来从原始语音信号 $\boldsymbol{x}$ 中区分出语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的，则鉴别器 D 所对应的损失函数 L_D 被定义为：

$$L_D = (\text{E}[\log(1 - D(\boldsymbol{x}))] + \text{E}[\log(D(\boldsymbol{x}_{\text{wa}}))]) / 2 \tag{4-7}$$

其中，$D(\boldsymbol{x})$ 表示原始语音信号 $\boldsymbol{x}$ 被鉴别器 D 分类为真实样本的概率，$D(\boldsymbol{x}_{\text{wa}})$ 表示语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 被鉴别器 D 分类为真实样本的概率，$D(\boldsymbol{x})$ 和 $D(\boldsymbol{x}_{\text{wa}})$ 的取值都在[0,1]区间内。通过对式(4-7)进行优化可以促使鉴别器 D 具有很强的区分原始语音信号 $\boldsymbol{x}$ 和语音对抗样本 $\boldsymbol{x}_{\text{wa}}$ 的能力。

4.3 实验结果与分析

4.3.1 实验的设置

本章的全部实验均是在流行的音乐类型集 GTZAN[92]上进行的，该数据集包含 1000 个音乐录制文件，每个文件的持续时间大约为 30 秒，这些音乐录制文件分别属于 blues、classical、country、disco、hiphop、jazz、metal、pop、reggae 和 rock 这 10 种音乐类型之一。将每个音乐录制文件分割成多个持续时间为 1 秒的音乐片段(由于是人声, 也可称为语音片段)，从而得到了 29000 个音乐片段，将这些音乐片段以 8:1:1 的比例划分到训练集、验证集和测试集中。考虑到 SampleCNN 作为一种基于卷积神经网络的音乐类型分类器，在 GTZAN 数据集上的分类准确率可以达到 92.90%，在实际应用中非常流行。为此，本章中选用 SampleCNN 作为总体架构中的语音分类模型 F。

实验进行之前，需要对语音片段进行归一化处理，以将语音片段 $\boldsymbol{x}$ 的幅值变换到[−1,1]区间内，归一化操作被定义为：

$$\mathrm{Norm}(\boldsymbol{x}) = \frac{2}{65535} \times (\boldsymbol{x} - 32767) + 1 \tag{4-8}$$

实验过程中，学习率被设置为 1×10^{-5}，epoch 的大小被设置为 60，batch 的大小被设置为 128，使用 Adam 优化算法[93]用于参数的优化。除此之外，式(4-5)中的阈值 c 根据经验被设置为 0。式(4-6)中的平衡参数 α 、 β 和 γ 在默认情况下被分别设置为 1、1 和 100，平衡参数 ε 则根据不同的目标标签被设置为不同的值，当目标标签是 blues、classical、country、hiphop、jazz、metal 和 reggae 时， ε 被设置为 1；当目标标签是 disco 时， ε 被设置为 10；当目标标签是 pop 时， ε 被设置为 15；当目标标签为 rock 时， ε 被设置为 20。为不同的目标标签设置的平衡参数 ε 的值是经过大量的实验后，通过对比实验结果确定的。实验发现，当设置相同的平衡参数 ε 时，从语音对抗样本 $\boldsymbol{x}_{\mathrm{wa}}$ 中提取到的水印的质量参差不齐。为了使得每一个目标标签下所提取到的水印都足够接近目标水印 $\boldsymbol{w}$，有必要针对特定的目标标签设置相对应的平衡参数 ε。

实验完成之后，对所得到的语音对抗样本 $\boldsymbol{x}_{\mathrm{wa}}$，还需要进行去归一化操作：

$$\mathrm{Inverse}(\boldsymbol{x}_{\mathrm{wa}}) = 32767 + (\boldsymbol{x}_{\mathrm{wa}} - 1) \times \frac{65535}{2} \tag{4-9}$$

为了评估所提出方法的攻击性能，本章使用了攻击成功率(SR)，公式为：

$$\mathrm{SR}=\frac{n_{\mathrm{attack}}}{n_{\mathrm{test}}} \tag{4-10}$$

其中，n_{attack} 表示成功误导目标语音分类模型 SampleCNN 的语音对抗样本的数量，n_{test} 表示用于测试的语音对抗样本的总数。

为了评估所提出方法的水印检测能力，本章使用了误码率(Bit Error Rate, BER)，公式为：

$$\mathrm{BER}(\boldsymbol{w}_{\mathrm{out}})=\frac{n_{\mathrm{error}}}{n_{\mathrm{all}}} \tag{4-11}$$

其中，n_{error} 表示提取到的水印 $\boldsymbol{w}_{\mathrm{out}}$ 中的错误比特数，n_{all} 表示水印 $\boldsymbol{w}_{\mathrm{out}}$ 中包含的总比特数。误码率越低，表明水印检测能力越强。需要注意的是，本章中目标水印被定义一个宽和高均为 32 的随机二值图像，由 1024 个比特位进行存储。

4.3.2 对抗性能评估

为了对本章所提出的方法的对抗性能进行评估，将本章方法下生成的所有语音对抗样本输入到 SampleCNN 中，得到各个目标标签下的攻击成功率，攻击成功率的混淆矩阵如图 4-4 所示。从图 4-4 中可以看出，混淆矩阵主对角线上的结果均为零，这是因为这些位置处目标标签与真实标签相同，这不属于攻击的范畴，因而这种情况是不用考虑的。观察混淆矩阵中的结果，可以发现所提出的方法对 SampleCNN 的最大攻击成功率为 100%，且几乎每种情况下的攻击成功率都高于 90%，平均攻击成功率为 93.07%。换句话说，SampleCNN 对语音对抗样本进行分类的平均准确率仅为 6.93%，这意味着本章方法下生成的语音对抗样本几乎使得 SampleCNN 分类模型完全瘫痪。可见，本章所提出的方法能够对基于神经网络的语音分类模型进行有效的定向攻击。

为了更好地说明攻击成功率的结果，进一步使用盒型图来描述所提出的方法对 SampleCNN 的攻击错误率的范围，如图 4-5 所示。从中可以看出，攻击错误率的范围在[0%,33.45%]区间，平均攻击错误率小于 7%，且只有 4 个异常值超过 17%，仅占整个测试用例的 4.44%。总的来说，本章所提出的方法在大多数情况下都能有效地攻击语音分类模型 SampleCNN，说明了该方法的适用性。

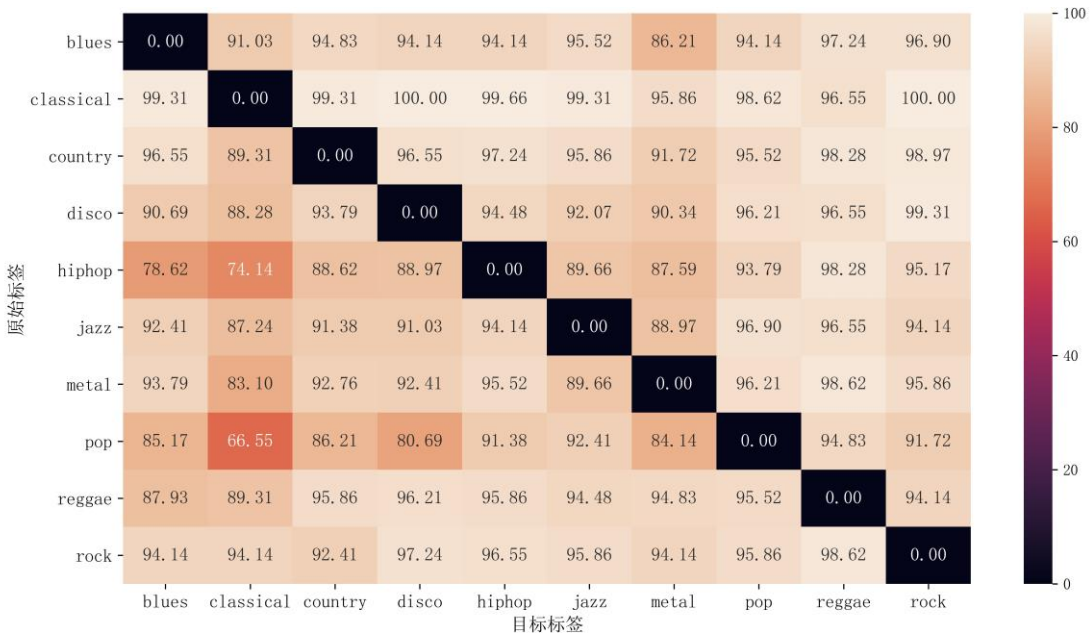


图4-4 攻击成功率的混淆矩阵

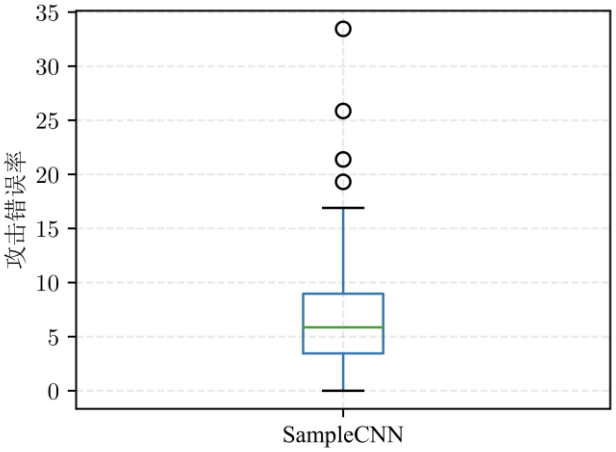


图4-5 攻击错误率的盒型图

4.3.3 水印评估

为了对本章所提出的方法的水印检测性能进行评估，使用本章中设计的水印网络对生成的语音对抗样本进行水印的提取，得到各个目标标签下提取到的水印信息，并将其与目标水印进行对比，从而计算出水印的误码率，提取到的水印的平均误码率如表 4-1 所示。从实验结果中可以看出，对于任意的目标标签，平均误码率都很低。其中，最高的平均误码率仅为 0.41%，换言之，提取

到的水印中最多只有大约 4 个错误位，这说明提取到的水印与目标水印非常接近。综上，本章的方法可以成功地进行模型所有权认证。

表4-1 从语音对抗样本中提取的水印的平均误码率

目标标签	平均误码率
blues	0.09%
classical	0.04%
country	0.03%
disco	0.33%
hiphop	0.05%
jazz	0.12%
metal	0.03%
pop	0.41%
reggae	0.27%
rock	0.12%

通常而言，水印应该只能从含水印的神经网络模型中提取出来，而无法从不含水印的神经网络模型中提取出来。由于本章所提出的算法是基于输出结果带水印的，水印是从输出的语音对抗样本中提取出来的。类似的，需要满足一定的要求，即只能从本章方法下生成的语音对抗样本中提取出目标水印，从其他语音信号中只能提取出虚假水印，其中，其他语音信号包括原始语音信号和其他方法下生成的语音对抗样本。将这一要求称作水印的唯一性，并通过实验对本章所提出的方法的水印唯一性进行评估，使用本章中设计的水印网络对其他语音信号进行水印的提取，得到各个目标标签下提取到的水印信息，并将其与目标水印进行对比，从而计算出水印的误码率，提取到的水印的平均误码率如表 4-2 所示。从实验结果中可以看出，大多数目标标签所对应的平均误码率都接近 0.5，由于误码率无限接近于 0 和无限接近于 1 时所对应的水印二值图像从感官上来说是一样的图案，则误码率接近于 0.5 时表明从其他语音信号中提取到的水印与目标水印的感官差距足够大，且误码率越接近 0.5，表示与目标水印之间的感官差距越大。综上所述，本章所提出的方法满足水印的唯一性要求。需要注意的是，为了满足水印的唯一性要求，需要让其他语音信号参与到模型的训练过程中。

表4-2 从不含水印的语音信号中提取的水印的平均误码率

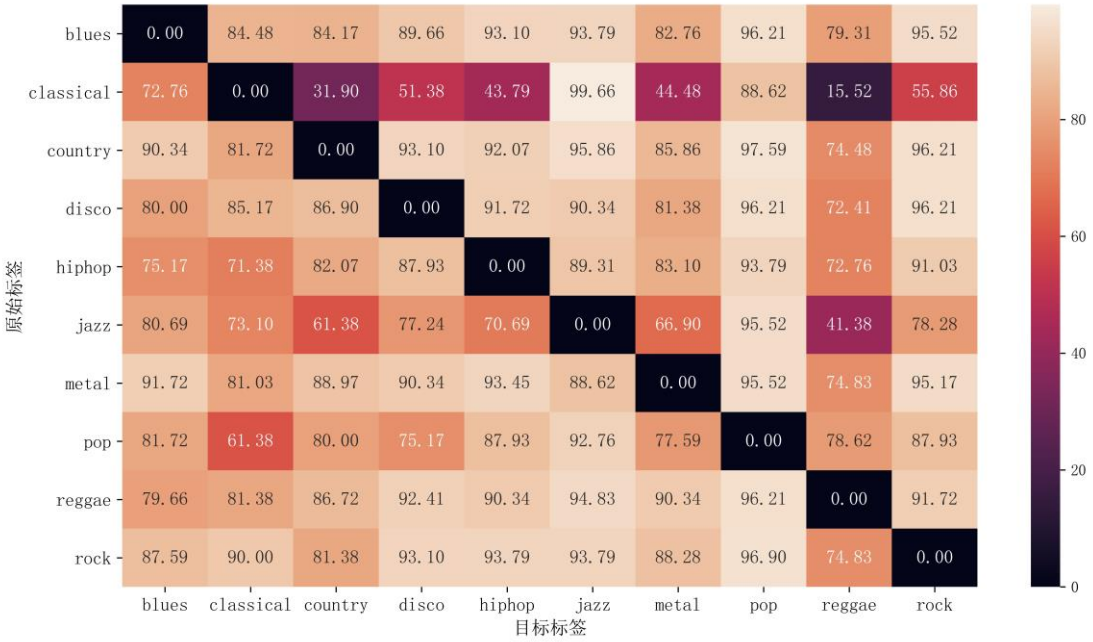
目标标签	平均误码率
blues	52.67%
classical	63.53%
country	53.31%
disco	65.42%
hiphop	49.75%
jazz	49.76%
metal	53.52%
pop	60.28%
reggae	49.95%
rock	56.32%

4.3.4 鲁棒性分析

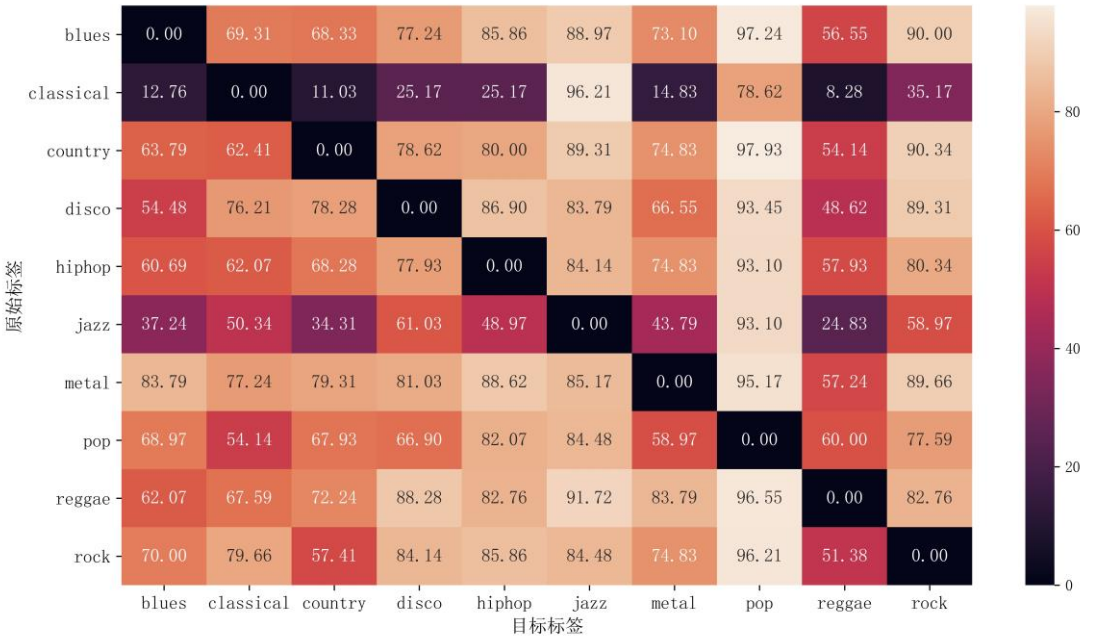
考虑到攻击者很有可能会攻击带水印的语音对抗样本，使得语音对抗样本丧失原有的对抗性能和对神经网络模型的版权认证能力，为此，本节中对所提出的方法的鲁棒性进行了评估。实验中通过插入噪声来模拟现实世界中的攻击场景，所选用的噪声为概率密度函数服从正态分布的高斯噪声，并通过指定不同的强度系数α，来获得不同强度的高斯噪声。通过在语音对抗样本中加载不同强度的高斯噪声，并将其输入到 SampleCNN 中，可以获得不同噪声强度下的攻击成功率的混淆矩阵；通过使用水印网络对加载了不同强度噪声的语音对抗样本进行水印的提取，可以获得不同噪声强度下的水印平均误码率。

图 4-6 展示了不同噪声强度下的攻击成功率的混淆矩阵，从图 4-6 中可以看出，当强度系数$\alpha = 0.01$时，所提出的方法对 SampleCNN 的最大攻击成功率为 99.66%，且在大多数情况下，攻击成功率都高于 80%，平均攻击成功率为 82.31%。换句话说，SampleCNN 对语音对抗样本进行分类的平均准确率仅为 17.69%，这意味着强度系数$\alpha = 0.01$时绝大多数嘈杂的语音对抗样本仍然可以使得 SampleCNN 分类模型失效。当强度系数$\alpha = 0.02$时，所提出的方法对 SampleCNN 的最大攻击成功率为 97.93%，且在大多数情况下，攻击成功率都高于 70%，平均攻击成功率为 69.61%，这意味着强度系数$\alpha = 0.02$时超过一半

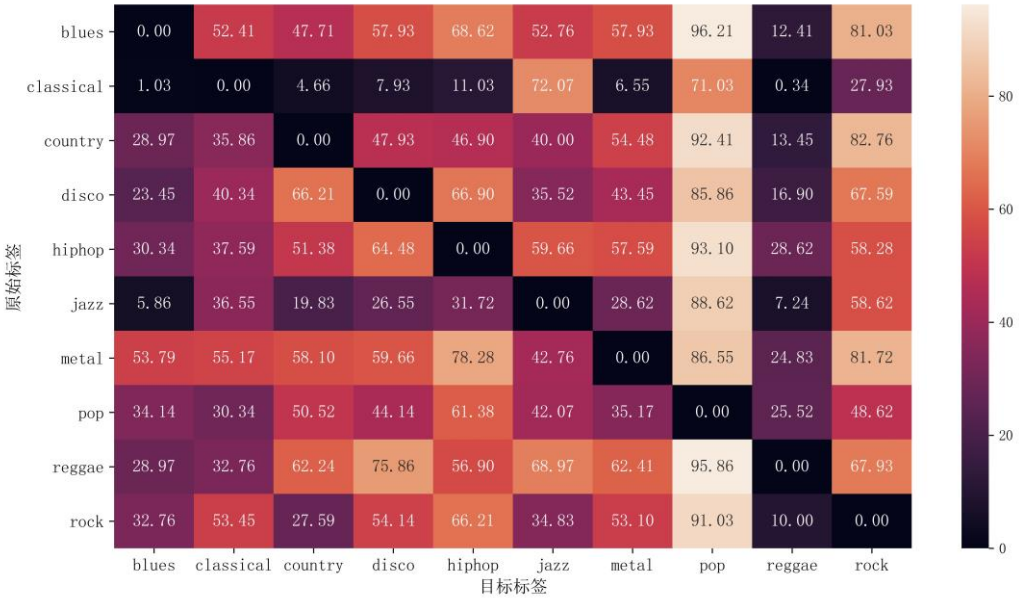
的嘈杂的语音对抗样本仍然可以使得 SampleCNN 分类模型失效。当强度系数 $\alpha = 0.04$ 时，虽然所提出的方法对 SampleCNN 的最大攻击成功率可以达到 96.21%，但是平均攻击成功率仅为 47.68%，这意味着强度系数 $\alpha = 0.04$ 时只剩下不到一半的嘈杂的语音对抗样本能够使得 SampleCNN 分类模型失效。综上，当强度系数 α 不超过 0.02 时，所提出的方法依旧能够很好地保持攻击性能。



(a) $\alpha = 0.01$



(b) $\alpha = 0.02$



(c) $\alpha = 0.04$

图4-6 不同噪声强度系数α下的攻击成功率的混淆矩阵

表 4-3 展示了从加载了不同强度噪声的语音对抗样本中提取到的水印的平均误码率，从实验结果中可以看出，当强度系数$\alpha \le 0.06$时，提取到的水印的平均误码率都相对较低，这意味着在一定的噪声强度下，水印依然可以可靠地重建。当强度系数$\alpha > 0.06$时，水印重建的能力变得相对较差。综上，在噪声攻击程度不强的情况下，所提出的方法在水印检测方面具有良好的鲁棒性。

表4-3 从加载了不同强度噪声的语音对抗样本中提取到的水印的平均误码率

目标标签	$\alpha = 0.02$	$\alpha = 0.04$	$\alpha = 0.06$	$\alpha = 0.08$	$\alpha = 0.1$
blues	0.83%	4.98%	8.80%	11.92%	14.44%
classical	0.25%	2.74%	10.10%	19.11%	25.39%
country	2.23%	11.17%	20.13%	26.46%	31.39%
disco	4.98%	17.61%	29.31%	37.89%	43.25%
hiphop	0.87%	5.80%	12.63%	19.22%	24.58%
jazz	1.50%	10.87%	24.65%	36.71%	46.17%
metal	0.87%	3.89%	8.79%	16.28%	24.88%
pop	2.27%	9.84%	19.77%	28.10%	33.51%
reggae	7.91%	25.07%	38.74%	48.04%	53.65%
rock	2.03%	16.06%	32.02%	42.93%	49.91%
average	2.37%	10.80%	20.49%	28.67%	34.72%

4.4 本章小结

本章提出了一种输出结果带水印的“无盒”水印算法，该算法能够用于保护任意的语音生成模型的知识产权。本章首先介绍了该方法的总体架构，其次介绍了生成器的网络结构、鉴别器的网络结构和水印网络的网络结构，然后介绍了生成器和鉴别器所对应的损失函数，最后从原始的攻击性能、水印检测能力、水印唯一性和鲁棒性这几个方面对所提出的方法进行了评估。实验结果表明，该方法不但能够很好地保持原始的攻击任务的性能，而且具有很强的模型所有权认证能力。此外，该方法还具有一定的抗噪声攻击的能力。

第五章 总结与展望

5.1 总结

面向语音信号的神经网络模型因其卓越的性能常被部署在语音转换、语音合成、语音识别和机器翻译等任务所对应的商业产品中，多家著名的科技公司也不断的在为构建性能更加强大的面向语音信号的神经网络模型投入大量的资金和人力，而恶意的攻击者却觊觎这些神经网络模型，并伺机窃取以牟利。因此，面向语音信号的神经网络模型的知识产权保护显得至关重要。本文将面向语音信号的神经网络模型分为语音分类模型和语音生成模型这两大类，针对语音分类模型，以说话人识别模型为例，提出了一种通用的语音分类模型水印算法。针对语音生成模型，以语音对抗样本生成模型为例，提出了一种通用的语音生成模型水印算法。本文的主要工作总结如下：

(1)针对语音分类模型，提出了一种基于频域扰动的“黑盒”水印算法，该算法中精心设计了触发音频样本。为了使触发音频样本能够抵抗恶意攻击，通过在原始音频样本的频域上添加基于片段的触发信号，对其进行构造；为了不影响原始任务的性能，指定触发音频样本为新增的标签。通过将触发音频样本与原始音频样本结合起来，对目标模型进行从头训练，使得水印信息被嵌入到目标模型中。当目标模型遭遇泄露时，模型的合法所有者只需要将触发音频样本输入到目标模型中，通过分析目标模型的预测结果与新增标签是否一致，即可对模型的产权进行认证，该认证过程在模型内部细节不公开的情况下也可以成功进行。通过实验对所提出方法的相关指标进行了评估，包括原始说话人识别任务的性能、水印验证的能力和鲁棒性。实验结果表明，该方法在原始说话人识别任务上的准确率达到了 99.35%，在水印验证任务上的成功率达到了 95%，此外，该方法还能在一定程度上抵抗噪声攻击。

(2)针对语音生成模型，提出了一种输出结果带水印的“无盒”水印算法，该算法在已有的载体网络之外，还设计了一个额外的、非公开的深度神经网络

模型，称为水印网络。训练过程中，通过联合优化载体网络和水印网络的损失函数，不仅可以使得载体网络完成其原始任务，还可以使得水印信息被加载到载体网络的输出语音中。当需要进行产权认证时，模型的合法所有者只需要利用水印网络从输出语音中检测水印信息，并通过分析检测到的水印与目标水印是否一致，即可完成认证，而不需要借助模型本身。通过实验对所提出方法的相关指标进行了评估，包括原始攻击任务的性能、水印检测的能力、水印的唯一性和鲁棒性。实验结果表明，该方法在定向攻击任务上的平均攻击成功率达到了 93.07%，在水印检测任务上的最高误码率只有 0.41%，且该方法能够满足水印的唯一性要求。此外，该方法还能在一定程度上抵抗噪声攻击。

5.2 展望

本文以面向语音信号的神经网络模型水印技术为研究目标，在已有的模型水印理论的基础上，提出了两种面向语音信号的神经网络模型水印算法，实验结果表明了所提出的方法是有效的，但是这些方法依然有可以改进的地方。因此，对面向语音信号的神经网络模型水印技术研究提出以下几点展望：

(1)构造触发模式更加隐蔽的触发音频样本。本文的触发音频样本是通过在原始音频样本的整个频域上添加触发信号构造而成的，这样做的优势是很容易让神经网络模型学习到触发信号的模式,但不足之处在于还不能完全做到隐蔽。未来可以尝试只在原始音频样本的中高频区域添加触发信号，来构造触发音频样本，并做好不易察觉性和触发模式学习能力这两方面的权衡。

(2)生成具有更小扰动的语音对抗样本。通过对本文生成的语音对抗样本进行分析，发现其所包含的扰动大于没有嵌入水印的语音对抗样本中的扰动。考虑到本文在训练生成模型的过程中给定了具有攻击能力和嵌入水印这两个优化目标，即独立地完成两个数据分布不一致的任务，因而这个现象是合理的。但是，未来可以将这两个优化目标合二为一来进行研究，即希望扰动既满足攻击的要求又可以包含水印信息,换句话说就是赋予对抗扰动以有意义的水印信息。

参考文献

[1] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. Piscataway: IEEE, 2015: 1-9.

[2] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, October 5-9, 2015, Munich, Germany. Berlin: Springer, 2015: 234-241.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, USA. Piscataway: IEEE, 2016: 770-778.

[4] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[5] MORGAN N. Deep and wide: Multiple layers in automatic speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20(1): 7-13.

[6] ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, March 25-30, 2012, Kyoto, Japan. Piscataway: IEEE, 2012: 4277-4280.

[7] ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545.

[8] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.

[9] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[10] ATENIESE G, MANCINI L V, SPOGNARDI A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers[J].

International Journal of Security and Networks, 2015, 10(3): 137-150.

[11] 任奎, 孟泉润, 闫守琨, 等. 人工智能模型数据泄露的攻击与防御研究综述[J]. 网络与信息安全学报, 2021, 7(1): 1-10.

[12] LUO M, BORS A G. Surface-preserving robust watermarking of 3-D shapes[J]. IEEE Transactions on Image Processing, 2011, 20(10): 2813-2826.

[13] SHEHAB M, BERTINO E, GHAFOOR A. Watermarking relational databases using optimization-based techniques[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 20(1): 116-129.

[14] OHBUCHI R, UEDA H, ENDOH S. Robust watermarking of vector digital maps[C]// Proceedings of the IEEE International Conference on Multimedia and Expo, August 26-29, 2002, Lausanne, Switzerland. Piscataway: IEEE, 2002: 577-580.

[15] 张颖君, 陈恺, 周赪, 等. 神经网络水印技术研究进展[J]. 计算机研究与发展, 2021, 58(5): 964.

[16] 冯乐, 朱仁杰, 吴汉舟, 等. 神经网络水印综述[J]. 应用科学学报, 2021, 39(6): 12.

[17] BARNI M, BARTOLINI F. Watermarking systems engineering: enabling digital assets security and other applications[M]. Boca Raton: Crc Press, 2004.

[18] YEUNG M M. Digital watermarking[J]. Communications of the ACM, 1998, 41(7): 31-33.

[19] PODILCHUK C I, DELP E J. Digital watermarking: algorithms and applications[J]. IEEE Signal Processing Magazine, 2001, 18(4): 33-46.

[20] COX I, MILLER M, BLOOM J, et al. Digital watermarking and steganography[M]. San Francisco: Morgan Kaufmann, 2007.

[21] SINGH N, JAIN M, SHARMA S. A survey of digital watermarking techniques[J]. International Journal of Modern Communication Technologies and Research, 2013, 1(6): 265852.

[22] ARNOLD M, SCHMUCKER M, WOLTHUSEN S D. Techniques and applications of digital watermarking and content protection[M]. Fitchburg: Artech House, 2003.

[23] BOENISCH F. A survey on model watermarking neural networks[J]. arXiv preprint

arXiv:2009.12153, 2020.

[24] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: Blackbox multibit watermarking for deep neural networks[J]. arXiv preprint arXiv:1904.00344, 2019.

[25] GUO J, POTKONJAK M. Watermarking deep neural networks for embedded systems[C]// Proceedings of the International Conference on Computer-Aided Design, November 05-08, 2018, San Diego, USA. New York: ACM, 2018: 1-8.

[26] LI Y, WANG H, BARNI M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171-193.

[27] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, June 6-9, 2017, Bucharest, Romania. New York: ACM, 2017: 269-277.

[28] DARVISH R B, CHEN H, KOUSHANFAR F. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, April 13-17, 2019, Providence, USA. New York: ACM, 2019: 485-497.

[29] WANG T, KERSCHBAUM F. Robust and undetectable white-box watermarks for deep neural networks[J]. arXiv preprint arXiv:1910.14268, 2019, 1(2).

[30] 谢宸琪，张保稳，易平. 人工智能模型水印研究综述[J]. 计算机科学，2021，48(7): 9-16.

[31] ZHANG J, GU Z, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security, June 04-08, 2018, Incheon, Republic of Korea. New York: ACM, 2018: 159-172.

[32] WANG J, WU H, ZHANG X, et al. Watermarking in deep neural networks via error back-propagation[J]. Electronic Imaging, 2020, 2020(4): 22-1-22-9.

[33] WANG T, KERSCHBAUM F. Attacks on digital watermarks for deep neural networks[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and

Signal Processing, May 12-17, 2019, Brighton, United Kingdom. Piscataway: IEEE, 2019: 2622-2626.

[34] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[35] FAN L, NG K W, CHAN C S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks[J]. arXiv preprint arXiv:1909.07830, 2019.

[36] FENG L, ZHANG X. Watermarking neural network with compensation mechanism[C]// Proceedings of the 13th International Conference on Knowledge Science, Engineering and Management, August 28-30, 2020, Hangzhou, China. Berlin: Springer, 2020: 363-375.

[37] LE Merrer E, PEREZ P, TRÉDAN G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233-9244.

[38] NAMBA R, SAKUMA J. Robust watermarking of neural network with exponential weighting[C]//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, July 09-12, 2019, Auckland, New Zealand. New York: ACM, 2019: 228-240.

[39] ADI Y, BAUM C, CISSE M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]//Proceedings of the 27th USENIX Security Symposium, August 15-17, 2018, Baltimore, USA. Berkeley: USENIX, 2018: 1615-1631.

[40] ROUHANI B D, CHEN H, KOUSHANFAR F. Deepsigns: A generic watermarking framework for ip protection of deep learning models[J]. arXiv preprint arXiv:1804.00750, 2018.

[41] CHEN H, ROUHANI B D, KOUSHANFAR F. Blackmarks: Blackbox multibit watermarking for deep neural networks[J]. arXiv preprint arXiv:1904.00344, 2019.

[42] LI Z, HU C, ZHANG Y, et al. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of DNN[C]//Proceedings of the 35th Annual Computer Security Applications Conference, December 09-13, 2019, San Juan, USA. New York: ACM, 2019: 126-137.

[43] ZHU R, ZHANG X, SHI M, et al. Secure neural network watermarking protocol against forging attack[J]. EURASIP Journal on Image and Video Processing, 2020, 2020(1): 1-12.

[44] YANG Z, DANG H, CHANG E C. Effectiveness of distillation attack and countermeasure on neural network watermarking[J]. arXiv preprint arXiv:1906.06046, 2019.

[45] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the 2016 IEEE Symposium on Security and Privacy, May 22-26, 2016, San Jose, USA. Piscataway: IEEE, 2016: 582-597.

[46] JIA H, CHOQUETTE-CHOO C A, CHANDRASEKARAN V, et al. Entangled watermarks as a defense against model extraction[C]//Proceedings of the 30th USENIX Security Symposium, August 11-13, 2021, Boston, USA. Berkeley: USENIX, 2021: 1937-1954.

[47] SALAKHUTDINOV R, HINTON G. Learning a nonlinear embedding by preserving class neighbourhood structure[C]//Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, March 21-24, 2007, San Juan, Puerto Rico. Cambridge: JMLR, 2007: 412-419.

[48] FROSST N, PAPERNOT N, HINTON G. Analyzing and improving representations with the soft nearest neighbor loss[C]//Proceedings of the 36th International Conference on Machine Learning, June 9-15, 2019, Long Beach, USA. New York: PMLR, 2019: 2012-2020.

[49] LI H, WENGER E, SHAN S, et al. Piracy resistant watermarks for deep neural networks[J]. arXiv preprint arXiv:1910.01226, 2019.

[50] CHEN H, ROUHANI B D, FU C, et al. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval, June 10-13, 2019, Ottawa, Canada. New York: ACM, 2019: 105-113.

[51] XU X R, LI Y Q, YUAN C. A novel method for identifying the deep neural network model with the Serial Number[J]. arXiv preprint arXiv:1911.08053, 2019.

[52] ZHANG J, CHEN D, LIAO J, et al. Deep model intellectual property protection via deep watermarking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, PP(99): 1-1.

[53] ZHAO J, HU Q, LIU G, et al. AFA: Adversarial fingerprinting authentication for deep neural networks[J]. Computer Communications, 2020, 150: 488-497.

[54] LUKAS N, ZHANG Y, KERSCHBAUM F. Deep neural network fingerprinting by conferrable adversarial examples[J]. arXiv preprint arXiv:1912.00888, 2019.

[55] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, 2014, Florence, Italy. Piscataway: IEEE, 2014: 4052-4056.

[56] BHATTACHARYA G, ALAM M J, KENNY P. Deep Speaker Embeddings for Short-Duration Speaker Verification[C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association, August 20-24, 2017, Stockholm, Sweden. New York: ISCA, 2017: 1517-1521.

[57] RICHARDSON F, REYNOLDS D, DEHAK N. A unified deep neural network for speaker and language recognition[J]. arXiv preprint arXiv:1504.00923, 2015.

[58] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association, August 20-24, 2017, Stockholm, Sweden. New York: ISCA, 2017: 999-1003.

[59] ZHANG C, KOISHIDA K, HANSEN J H L. Text-independent speaker verification based on triplet convolutional neural network embeddings[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1633-1644.

[60] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: a large-scale speaker identification dataset[J]. arXiv preprint arXiv:1706.08612, 2017.

[61] MUCKENHIRN H, DOSS M M, MARCELL S. Towards directly modeling raw speech signal for speaker verification using CNNs[C]//Proceedings of the 2018 IEEE International

Conference on Acoustics, Speech and Signal Processing, April 15-20, 2018, Calgary, Canada. Piscataway: IEEE, 2018: 4884-4888.

[62] SEKI H, YAMAMOTO K, NAKAGAWA S. A deep neural network integrated with filterbank learning for speech recognition[C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, 2017, New Orleans, USA. Piscataway: IEEE, 2017: 5480-5484.

[63] RAVANELLI M, BENGIO Y. Speaker recognition from raw waveform with sincnet[C]// Proceedings of the 2018 IEEE Spoken Language Technology Workshop, December 18-21, 2018, Athens, Greece. Piscataway: IEEE, 2018: 1021-1028.

[64] RABINER L, SCHAFER R. Theory and applications of digital speech processing[M]. Upper Saddle River: Prentice Hall Press, 2010.

[65] HU G S. Introduction to digital signal processing[M]. Beijing: Tsinghua University Press, 2005.

[66] AMARI S. Backpropagation and stochastic gradient descent method[J]. Neurocomputing, 1993, 5(4-5): 185-196.

[67] SUN M, SONG Z, JIANG X, et al. Learning pooling for convolutional neural network[J]. Neurocomputing, 2017, 224: 96-104.

[68] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[69] SRIVASTAVA N. Improving neural networks with dropout[J]. University of Toronto, 2013, 182(566): 7.

[70] GAO Y, SHEN L, XIA S T. DAG-GAN: Causal Structure Learning with Generative Adversarial Nets[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, June 6-11, 2021, Toronto, Canada. Piscataway: IEEE, 2021: 3320-3324.

[71] WANG D, DONG L, WANG R, et al. Targeted speech adversarial example generation with generative adversarial network[J]. IEEE Access, 2020, 8: 124503-124513.

[72] DABAS P, KHANNA K. A study on spatial and transform domain watermarking

techniques[J]. International Journal of Computer Applications, 2013, 71(14): 38-41.

[73] KONG Y, ZHANG J. Adversarial audio: A new information hiding method and backdoor for dnn-based speech recognition models[J]. arXiv preprint arXiv:1904.03829, 2019.

[74] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach[C]//Proceedings of the 19th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, December 29, 2020-January 1, 2021, Guangzhou, China. Piscataway: IEEE, 2020: 402-409.

[75] AHMED N, NATARAJAN T, RAO K R. Discrete cosine transform[J]. IEEE transactions on Computers, 1974, 100(1): 90-93.

[76] JEYHOON M, ASGARI M, EHSAN L, et al. Blind audio watermarking algorithm based on DCT, linear regression and standard deviation[J]. Multimedia Tools and Applications, 2017, 76(3): 3343-3359.

[77] ZHONG Q, ZHANG L Y, ZHANG J, et al. Protecting IP of deep neural networks with watermarking: A new label helps[C]//Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 11-14, 2020, Suntec Singapore, Singapore. Berlin: Springer, 2020: 462-474.

[78] ZUE V, SENEFF S, GLASS J. Speech database development at MIT: TIMIT and beyond[J]. Speech Communication, 1990, 9(4): 351-356.

[79] XU J, LI Z, DU B, et al. Reluplex made more practical: Leaky ReLU[C]//Proceedings of the 2020 IEEE Symposium on Computers and Communications, July 7-10, 2020, Rennes, France. Piscataway: IEEE, 2020: 1-7.

[80] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[81] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[82] CARLINI N, WAGNER D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//Proceedings of the 2018 IEEE Security and Privacy Workshops, May

24, 2018, San Francisco, USA. Piscataway: IEEE, 2018: 1-7.

[83] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[84] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[85] PASCUAL S, BONAFONTE A, SERRA J. SEGAN: Speech enhancement generative adversarial network[J]. arXiv preprint arXiv:1703.09452, 2017.

[86] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[87] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. Piscataway: IEEE, 2015: 1026-1034.

[88] NAYEF B H, ABDULLAH S N H S, SULAIMAN R, et al. Optimized leaky ReLU for handwritten Arabic character recognition using convolution neural networks[J]. Multimedia Tools and Applications, 2022, 81(2): 2065-2094.

[89] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, USA. Menlo Park: AAAI, 2017: 4278--4284.

[90] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2591-2601.

[91] LEE J, PARK J, KIM K L, et al. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms[J]. arXiv preprint arXiv:1703.01789, 2017.

[92] TZANETAKIS G, COOK P. Musical genre classification of audio signals[J]. IEEE Transactions on Speech and Audio Processing, 2002, 10(5): 293-302.

[93] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

作者在攻读硕士学位期间公开发表的论文

[1] **WANG Y**, YE J, WU H. Generating watermarked speech adversarial examples [C]//Proceedings of ACM Turing Award Celebration Conference, July 30-August 1, 2021, Hefei, China. New York: ACM, 2021: 254-260. (**EI:** 20214311050889)

[2] **WANG Y**, WU H. Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain[J]. Symmetry, 2022, 14(3): 619. (**SCI:** 000776303000001)

作者在攻读硕士学位期间所参与的项目

[1] 国家自然科学基金青年项目“社交网络多用户协同的行为隐写”(项目编号：61902235).

致　　谢

时光荏苒，弹指一挥间我的硕士生涯即将逝去。回想起在上海大学发生的点点滴滴，有遇到困难一时难以解决时的迷茫、有与同学挑灯夜战组队参赛时的坚持、有论文发表取得成绩时的喜悦，更有对导师、家人及朋友教导、鼓励与陪伴时的感激。正是这些宝贵的经历，让我成长为一个敢于大胆尝试、喜欢团队合作和充满自信的人。借此机会，我想向诸多在硕士期间悉心指导我的老师、尽力帮助我的同学和一直支持着我的家人表达内心最真挚的感谢。

首先，我要感谢我的导师吴汉舟老师，吴老师勇攀知识高峰的态度、严谨肯干的工作作风以及乐于助人的处事原则值得我终身学习。每当我对实验课题一筹莫展时，吴老师总是尽心尽力地指导我；每当我生活上遇到难关时，吴老师总是设身处地地替我想办法。吴老师对我而言，是良师，也是益友。尽管我即将离开校园踏入社会，我依然会谨记吴老师对我的教导，也会始终记得吴老师对我的帮助。在此，我谨向吴老师致以崇高的敬意和衷心的感谢！

其次，感谢课题组的张新鹏、冯国瑞、任艳丽、吕东辉和侯丽敏等老师为我们提供了良好的学习平台，从他们身上，我学习到了研究学问时的执着、对待工作时的勤勤恳恳和对待学生时的包容。感谢实验室的郑晓燕、陈诗怡、柳琦云、唐雄和徐超等同学在我学习或生活上遇到困难时帮助我；感谢我的室友营造了温馨、和谐且积极向上的寝室环境。

然后，我还要感谢我的家人，他们对我的爱是最无私、最毫无保留的。尽管身处异地，他们也总是每天关注着我所在城市的动向，我一切安好的时候他们总是在默默守护着我，我遇到挫折的时候他们总是毫不犹豫地挺身而出，感谢他们对我的付出。

最后，感谢在百忙之中抽出时间来评审我论文的各位老师！