

基于可逆网络双嵌入和攻击层的鲁棒音频水印方法

张旭龙¹, 瞿晓阳¹, 李鹏程¹, 肖春光², 王健宗¹

1. 平安科技(深圳)有限公司, 广东 深圳 518063;

2. 深圳市宝安区教育事业发展中心, 广东 深圳 518101

摘要

数字音频水印技术旨在将信息嵌入音频, 并从含水印的音频中准确提取信息。传统方法依靠专家经验设计的算法将水印嵌入信号的时域或变换域。随着神经网络的发展, 基于深度学习的神经音频水印技术应运而生。与传统算法相比, 神经音频水印通过在训练过程中考虑各种攻击, 实现了更强的鲁棒性。然而, 当前的神经音频水印方法存在容量低、不可感知性差等问题。此外, 水印定位问题在神经音频水印中尤为重要, 却未得到充分研究。设计了一种基于可逆网络用于高效定位的双嵌入水印模型, 并在鲁棒性训练中考虑了攻击层对可逆神经网络的影响, 从而提高了模型的合理性和稳定性。实验表明, 与现有方法相比, 所提出的方法能够承受各种攻击, 具有更高的容量和更有效的定位能力。

关键词

神经音频水印; 可逆神经网络; 双嵌入; 同步码; 鲁棒性

中图分类号: TP391

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025054

Invertible networks-based method with dual-embedding and attack layer for robust audio watermarking

ZHANG Xulong¹, QU Xiaoyang¹, LI Pengcheng¹, XIAO Chunguang², WANG Jianzong¹

1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China

2. Shenzhen Baoan District Education Information Center, Shenzhen 518101, China

Abstract

Digital audio watermarking aims to embed information into audio and accurately extract it from the watermarked audio. Traditional methods rely on algorithms designed based on expert experience to embed watermarks into the time domain or transform domain of signals. With the development of deep neural networks, neural audio watermarking based on deep learning has emerged. Compared with traditional algorithms, neural audio watermarking achieves better robustness by considering various attacks during training. However, current neural audio watermarking methods suffer from low capacity and poor imperceptibility. Furthermore, the issue of watermark locating, which is extremely important and even more pronounced in neural audio watermarking, has not been adequately studied. In this paper, we design a dual-embedding watermarking model based on an invertible neural network for efficient localization. We also consider the impact of the attack layer on the invertible neural network in robustness training, improving the model to

enhance both its reasonableness and stability. Experiments show that the proposed model can withstand various attacks with higher capacity and more efficient locating ability compared with existing methods.

Key words
neural audio watermarking, invertible neural network, dual-embedding, synchronization code, robustness

0 引言

数字水印技术可将表示所有权或真实性的消息嵌入图像、视频和音频等多媒体中，并且难以被人察觉。这项技术广泛用于声明所有权和防伪。音频水印技术已经发展了几十年。数字水印最具挑战性的两个要求是不可感知性和鲁棒性，这意味着嵌入的水印难以被人感知，并且即使在水印媒体遭受无意损坏或恶意移除攻击时，水印仍可以被准确地提取。传统技术^[1-2]通过基于专家知识设计的算法将水印嵌入音频信号的时域或变换域中^[3]。传统水印

方法的鲁棒性通常源于主观设计，这导致了一定的局限性。深度学习的进步为隐写术^[4-5]和数字水印技术^[6-7]带来了新的解决方案。端到端神经音频水印模型在每次训练迭代中完成嵌入和提取过程，并通过设计的训练目标来约束不可感知性和提取水印的完整性。为了保证鲁棒性，在嵌入器-提取器（即编码器-解码器）结构中引入了模拟水印媒体常见损坏的攻击层。

神经音频水印技术目前仍处于早期发展阶段。人类的听觉有时比视觉更敏感^[8]，它可以很容易地分辨出噪声，从而使水印引起的细微变化被感知到。至于水印的鲁棒性，神经音频水印的嵌入-攻击-提取流程如图1所示，其中攻击层模拟了训练过程中对水印音频的各种移除攻击。为了增强适用性和抵抗使用剪裁和拼接的去同步攻击，相同的水印会被重复嵌入音频段的各个位置。然而，这种策略引发了定位问题。在实际场景中，提取过程中水印的嵌入位置是未知的。此外，修剪和拼接会导致水印位置发生变化。与传统方法相比，神经音频水印的提取依赖于神经网络的前向传播过程，这会导致不可忽略的时间成本，并且时间成本会随着网络复杂性的增加而增加。现有方法通常使用穷举法，逐步提取定位（也称同步码）和水印消息，如图2所示。定位效率是神经音频水印必须面对的一个问题。

水印嵌入和提取过程的对称性为可逆神经网络提供了充足的机会，但攻击层的引入破坏了原始的对称性^[9]。换句话说，

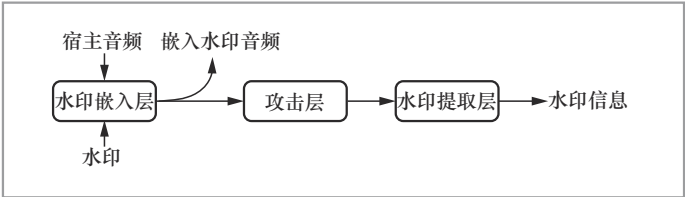


图1 神经音频水印的嵌入-攻击-提取流程

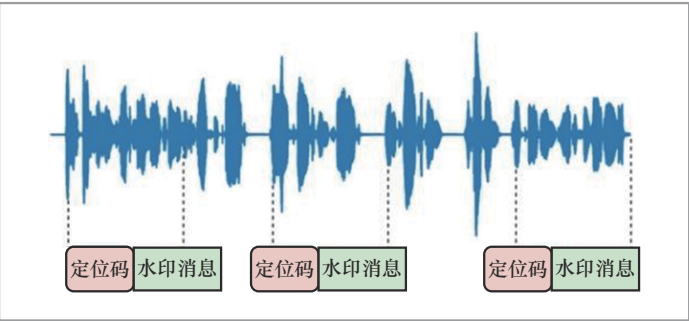


图2 现有方法的常见嵌入策略

由于攻击层的存在，编码器的输出（即水印音频）和解码器的输入（即经过攻击的水印音频）不一致，而编码器和解码器是相反的并且共享参数，这种不匹配影响了训练效果。

总而言之，传统音频水印技术主要依赖于专家设计的算法，将水印嵌入音频信号的时域或变换域中，如短时傅里叶变换（short-time Fourier transform, STFT）域或离散小波变换（discrete wavelet transform, DWT）域，对音频的嵌入容量相对较高，实现简单，但鲁棒性较差，容易受到各种攻击的影响，如噪声、滤波、压缩等。基于深度学习的神经音频水印技术利用深度神经网络进行端到端的训练，将水印的嵌入和提取过程整合到一个模型中，通过在训练过程中引入攻击层，可以显著提高水印的鲁棒性，抵抗各种攻击。但现有的神经音频水印技术普遍面临容量相对较低、音频的不可感知性较差的问题，并且水印定位问题尚未得到充分解决。基于可逆神经网络的音频水印技术利用可逆神经网络进行水印的嵌入和提取，保证了水印嵌入和提取过程的可逆性，在保持高鲁棒性的同时，提高了水印的嵌入容量，并且可以有效解决水印定位问题。但攻击层的引入会破坏可逆神经网络的对称性，影响训练效果。

本文提出了一种可逆双嵌入音频水印模型，该模型使用具有双嵌入策略的两阶段可逆神经网络分别嵌入水印消息和定位码，如图3所示。在提取过程中，首先提取计算成本较低的定位码，匹配成功后，再进行计算成本较高的水印消息提取。这使灵活地扩大水印容量成为可能。为了减轻攻击层造成的不对称影响，该模型引入了平衡块来增强训练稳定性，同时保留可逆神经网络的特性。

本文的贡献如下：①考虑到神经音频

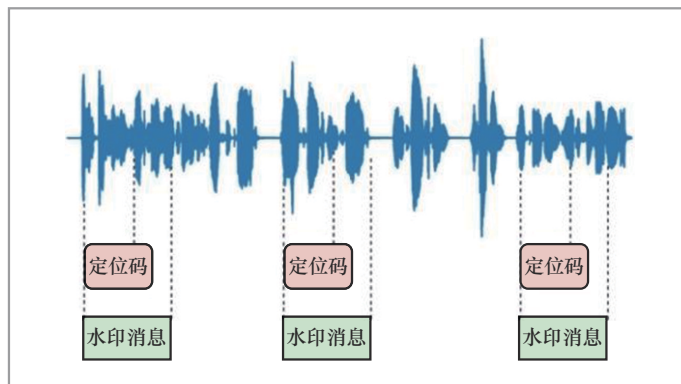


图3 本文提出方法的双嵌入策略

水印的特点，设计了一种双嵌入策略，将水印消息和定位码分别嵌入，以加快定位过程；②引入了平衡块来减轻攻击层对可逆网络造成的不对称性，并保持可逆神经网络的对称性；③本文提出的水印模型能够嵌入更多比特的水印，同时确保不可感知性和鲁棒性。

1 相关工作

1.1 神经音频水印

神经音频水印模型通常由两个神经网络组成，用于在短时傅里叶变换域或离散小波变换域中嵌入和提取水印。Pavlović等^[10]设计了一种用于语音水印的编码器-解码器架构，其中编码器和解码器形成对抗关系并一起训练。此后，他们在之前的工作中引入了一个攻击层^[11]，形成了一个编码器-攻击层-解码器结构，增强了基于DNN的语音水印的鲁棒性。WavMark^[12]将提取视为嵌入水印的逆过程，并利用可逆神经网络来执行音频水印的嵌入和提取。DeAR^[13]专注于重新录制攻击对音频水印的威胁，将重新录制建模为几个可微分过程。除了不可感知性和鲁棒性，容量和定

位效率也是评估音频水印方法的重要标准。与传统的水印方法相比,大多数神经音频水印方法的容量较低,并且没有考虑定位过程带来的高消耗。

1.2 可逆神经网络

NICE^[14]首先引入了可逆神经网络(invertible neural networks, INN)的概念,这是一个基于归一化流的框架,学习将遵循原始分布的数据转换为预定义分布的变换。Dinh等^[15]通过Real NVP中的卷积层提高了INN的性能。Ardizzone等^[16]引入了条件INN(cINN)来建立对生成的控制。Behrmann等^[17]将ResNet作为常微分方程的欧拉离散化,并证明了可以通过改变归一化机制来构建可逆的ResNet。INN已被广泛应用于图像生成任务^[14-15, 18-19]、图像增强^[20]、图像压缩^[21]、图像到图像的转换^[19]、数字隐写术^[22-23]等领域。

2 可逆音频水印网络架构

2.1 总体架构

图4展示了本文提出的音频水印模型

架构,它遵循嵌入器-攻击层-提取器的结构,其中嵌入器和提取器被精心设计为两阶段结构,分别在垂直维度嵌入消息和定位码。在嵌入过程中,一个 L_m 位的二进制水印消息 $m \in \{0, 1\}^{L_m}$ 通过第一阶段INN嵌入固定长度的音频块 X 的STFT域中;然后,第二阶段INN将二进制定位码 $c \in \{0, 1\}^{L_c}$ 嵌入前一步骤中包含 m 的音频中;最终的水印音频通过逆短时傅里叶变换(ISTFT)重建。鉴别器用于区分宿主音频和水印音频,以保证水印的不可感知性。在提取过程中,两阶段提取器首先从水印音频中提取 c ,然后按嵌入的相反顺序提取 m 。该模型引入了一个攻击层来增强对各种去除攻击的鲁棒性。提取器必须从经过攻击的水印音频中准确地提取 c 和 m 。平衡块旨在减轻攻击层引入的不对称性,保持INN的对称性。

2.2 用于双嵌入的两阶段INN

为了垂直分离定位码和消息,本文设计了一个两阶段INN:阶段1为 $INN_{\#1}$,用于嵌入和提取水印消息;阶段2为 $INN_{\#2}$,用于定位码嵌入和提取。每个INN由几个可逆块组成,这些块将变换后的音频和水印(即水印消息或定位码)作为输入,产

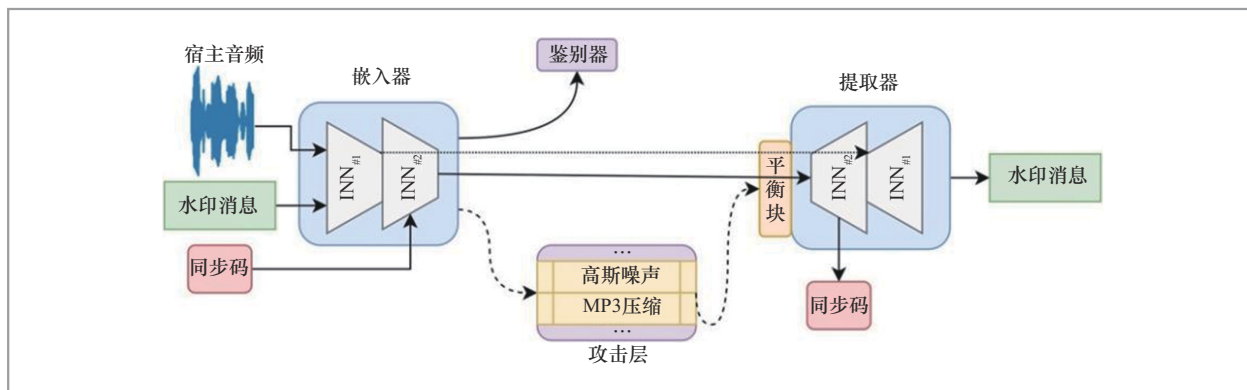


图4 本文提出的音频水印模型架构

生两个输出。将这些输入-输出对称两个数据流：音频流和水印流。音频流在嵌入过程中输出水印音频，水印流在提取过程中提供水印消息。可逆块中处理数据的方式如式（1）和式（2）所示。

$$x^{i+1} = x^i \odot \exp(\alpha(\psi(s^i))) + \phi(s^i) \quad (1)$$

$$s^{i+1} = s^i \odot \exp(\alpha(\rho(x^{i+1}))) + \eta(x^{i+1}) \quad (2)$$

其中， x 表示宿主数据，包括输入 $\text{INN}_{\#1}$ 的原始音频和输入 $\text{INN}_{\#2}$ 的嵌入消息的音频； s 表示秘密数据，包括输入 $\text{INN}_{\#1}$ 的消息和输入 $\text{INN}_{\#2}$ 的定位码； $\alpha(\cdot)$ 是一个Sigmoid函数，而 $\psi(\cdot)$ 、 $\phi(\cdot)$ 、 $\rho(\cdot)$ 和 $\eta(\cdot)$ 是由全连接块构成的子网络。

在第一个嵌入阶段， $\text{INN}_{\#1}$ 将水印消息嵌入宿主音频中，其中音频和比特序列首先分别通过STFT转换为时频域。然后 $\text{INN}_{\#2}$ 将定位码嵌入 $\text{INN}_{\#1}$ 的音频流输出中。双嵌入可以用式（3）来描述，其中 $\text{INN}(\cdot)$ 的下标 a 和 wm 分别表示INN的音频流和水印流的输出。

$$x_{wm} = \text{INN}_{\#2}(\text{INN}_{\#1}(x, m)_a, c)_a \quad (3)$$

本文提出的音频水印方法的提取过程与嵌入过程完全相反。 $\text{INN}_{\#2}$ 首先从水印音频中提取定位码，然后将 $\text{INN}_{\#2}$ 的音频流输出到 $\text{INN}_{\#1}$ 以提取水印消息。式（4）描述了获得嵌入消息的双重提取。

$$\hat{m} = \text{INN}_{\#1R}(\text{INN}_{\#2R}(x_{wm}, x_{aux_2})_a, x_{aux_1})_{wm} \quad (4)$$

其中， x_{aux} 表示随机采样的信号输入INN的消息流。下标R表示可逆网络的逆过程，提取过程采用与嵌入过程相同的网络和参数，只是顺序相反。

在训练过程中，消息在每个嵌入阶段后都由 $\text{INN}_{\#1R}$ 提取，然后与原始消息 m 进

行比较。定位码通过 $\text{INN}_{\#2R}$ 从最终的水印音频中提取，并与原始定位码进行比较。完整性损失见式（5）：

$$\mathcal{L}_{\text{integ}} = \|\hat{m} - m\|_2 + \left\| \text{INN}_{\#1R}(\text{INN}_{\#1}(x, m)_a, x_{aux_1})_{wm} - m \right\|_2 + \left\| \text{INN}_{\#2R}(x_{wm}, x_{aux_2})_{wm} - c \right\|_2 \quad (5)$$

在实际场景中，水印被嵌入音频的多个片段中，如图3所示。然而，水印音频可能会被修剪或拼接（称为去同步攻击^[24]），这使得确定水印位置变得具有挑战性。由于定位码的长度远低于水印消息的长度，使用更浅的神经网络即能保证其嵌入的质量，相较 $\text{INN}_{\#1}$ ， $\text{INN}_{\#2}$ 的设计使用较少的层数、更加轻量，计算成本更低，因此提取器可以快速提取和匹配定位码，只有在定位码匹配时才提取水印消息。此外，本文提出的模型中采用了类似于WavMark^[12]中模糊定位的训练方式，这有助于提取器从邻近位置提取水印。

2.3 不可感知性

作为评估数字水印的重要指标之一，不可感知性确保了水印的嵌入不会被人耳听觉区分。本文提出的音频水印方法采取两种措施来确保和提高水印的不可感知性。感知损失直观地要求水印音频和原始宿主音频之间的差异尽可能小，即减少水印对宿主音频的影响。感知损失如式（6）所示。

$$\mathcal{L}_{\text{percept}} = \|x_{wm} - x\|_2 \quad (6)$$

更重要的是，该模型利用鉴别器来区分水印音频和原始音频。本质上，鉴别器是一个二元分类器，将宿主音频和水

印音频分别标记为 0 和 1（表示为标签 y ）。鉴别器的目标是识别水印音频，而嵌入器的目标是隐藏水印，因此嵌入器和鉴别器形成对抗关系，并在训练过程中相互强化^[25]。鉴别损失和识别损失计算式如下。

$$\mathcal{L}_{\text{discr}} = -y \cdot \log(D(x)) - (1-y) \cdot \log(1-D(x)) \quad (7)$$

$$\mathcal{L}_{\text{ident}} = -\log(1-D(x_{\text{wm}})) \quad (8)$$

2.4 鲁棒性和对称性

水印方法的鲁棒性对于攻击者有效去除嵌入水印的能力至关重要。通常，攻击者会尝试通过几种去除技术从水印音频中去除水印，包括仅将音频通过滤波器以保留人类可以感知的信息，或者向水印音频添加噪声以干扰准确提取。有效的水印去除要求无法从去除水印的音频中正确提取水印，并且去除水印的音频应该是可用的，即收听质量不应该显著下降，否则这样的水印去除是没有意义的。

为了赋予神经音频水印抵抗各种水印去除攻击的鲁棒性，将攻击层纳入水印模型，并与嵌入和提取网络一起训练。攻击层对水印音频（即嵌入器的输出）进行各种攻击，随后，提取器尝试尽可能准确地从经过攻击的音频中提取定位码和水印消息。考虑到水印去除的有效性，预定义的攻击应确保受攻击的水印音频的质量不会下降太多。这些攻击包括高斯加性噪声、低通滤波器、MP3 压缩、量化、重采样、随机丢失、幅度修改和时间拉伸等。

然而，攻击层的引入破坏了整个嵌入-提取过程的对称性，影响了 INN 的训练。将完整的双嵌入过程和提取过程分别定义为 $f(\cdot)$ 和 $f^{-1}(\cdot)$ 。在嵌入过程中，提出的音频水印方法执行 $f(x, m, c) = w$ 。其

中， x, m, c, w 分别是前面提到的宿主音频、水印信息、定位码和水印音频，并且 w 遵循水印音频的分布 $P_w(w)$ ^[26]。在提取过程中，由于嵌入器和提取器之间的参数共享，可以根据 $P_w(w)$ 通过 $m, c = f^{-1}(w)$ 轻松采样 m 和 c 。但是攻击层的加入导致 $P_w(w)$ 发生变化，变成了 $P_{w'}(w')$ ，而提取过程仍然基于未改变的 $P_w(w)$ ，这影响了水印的性能。INN 的参数共享策略限制了提取器适应受攻击音频的学习能力，导致水印性能失真。

为了同时保持 INN 的参数共享和 INN 训练的对称性，本文采用平衡块来减轻攻击层引起的不对称性，并稳定模型的对称结构。平衡块由一组全连接层^[27]组成，这些全连接层将输入处理成大小相等的输出，为提取器提供额外的可训练参数。这种方式在上述完整性损失的指导下，将攻击后的水印音频分布 $P_{w'}(w')$ 转换为接近预期分布 $P_w(w)$ 的修正分布 $P_{\hat{w}(\hat{w})}$ ，从而抵消攻击层引入的偏移量的影响，而不会干扰嵌入器。

2.5 模型训练策略

同时确保准确性、不可感知性和鲁棒性是很困难的。因此，本文提出的音频水印方法的训练分为两个阶段：第一阶段只考虑不可感知性和提取的水印完整性，目的是构建一个可以无感知地嵌入水印并准确地提取水印的两阶段 INN；第二阶段引入了对水印鲁棒性的要求，将攻击层和平衡块加入模型中，并对整个模型进行整体训练。

两个训练阶段都使用相同的总损失函数，如式（9）所示，其中 λ_1 、 λ_2 和 λ_3 是每个分量的权重，但需要注意的是，第二阶段包含更多的可训练参数。鉴别器在每次

迭代中都与水印模型一起训练，损失函数如式（7）所示。训练过程的伪脚本以及总损失的获取见算法1。

$$L_{\text{total}} = \lambda_1 L_{\text{integ}} + \lambda_2 L_{\text{percept}} + \lambda_3 L_{\text{ident}} \quad (9)$$

算法1 训练过程以及总损失的获取

输入：原始音频片段 x ，水印信息 m ，定位码 c 模块：信息嵌入器 Emb_1 ，定位码嵌入器 Emb_2 ，信息提取器 Ext_1 ，定位码提取器 Ext_2 ，长度调节器 LR_1 、 LR_2 、 LR_3 、 LR_4 ，攻击层 Att ，平衡块 B 操作：短时傅里叶变换 $\text{STFT}(\cdot)$ ，逆短时傅里叶变换 $\text{ISTFT}(\cdot)$

输出：总损失 L_{total} 调节长度，转换到 STFT 域：

$$m' \leftarrow \text{STFT}(\text{LR}_1(m))$$

$$c' \leftarrow \text{STFT}(\text{LR}_2(c))$$

$x' \leftarrow \text{STFT}(x)$ 嵌入信息：

$$x'_{w1} \leftarrow \text{Emb}_1(x', m')$$

第一次提取：

$$\widehat{m}_1 \leftarrow \text{LR}_3(\text{ISTFT}(\text{Ext}_1(x'_{w1})))$$

嵌入定位码：

$$x'_{w2} \leftarrow \text{Emb}_2(x'_{w1}, c')$$

获取含水印的音频波形：

$$x_w \leftarrow \text{ISTFT}(x'_{w2})$$

训练鉴别器：

通过 x 以及 x_w 优化 $\mathcal{L}_{\text{discr}}$ 方向传递提升鉴别器

鲁棒性控制：

$$x_w \leftarrow B(\text{Att}(x_w))$$

提取位置编码：

$$x'_{\text{mid}}, \widehat{c}' \leftarrow \text{Ext}_2(x'_w)$$

$$\widehat{c} \leftarrow \text{LR}_4(\text{ISTFT}(\widehat{c}'))$$

提取消息：

$$\widehat{m} \leftarrow \text{LR}_3(\text{ISTFT}(\text{Ext}_1(x'_{\text{mid}})))$$

损失计算：

通过 $\{m, \widehat{m}_1, \widehat{m}, c, \widehat{c}\}$ 获取完整性损失 L_{integ}

通过 $\{x', x'_w\}$ 获取感知损失 L_{percept}

通过 $\{x, x_w\}$ 获取一致性损失 L_{ident}

总损失：

$$L_{\text{total}} = \lambda_1 L_{\text{integ}} + \lambda_2 L_{\text{percept}} + \lambda_3 L_{\text{ident}}$$

3 实验设置与评估

3.1 数据集和实验配置

本文提出的音频水印方法在 VCTK 语料库^[28]和 FMA 语料库^[29]上进行训练。VCTK 包含超过 100 h 的多说话人语音数据，而 FMA 语料库包含大量的音乐音频。这两种类型的数据在音频水印一般应用的场景中普遍存在。训练期间，所有音频都重新采样到 16 000 Hz，并分成 1 s 的片段。使用参数为 $n_{\text{fft}} = 1\,000$ 、 $\text{hop_length} = 250$ 、 $\text{win_length} = 1\,000$ 的 STFT 和 ISTFT 操作执行域间变换。

利用 Adam 优化器^[30]，参数为 $\beta_1 = 0.9$ 、 $\beta_2 = 0.99$ 、 $\epsilon = 10^{-8}$ 、 $\text{lr} = 10^{-5}$ （带有 StepLR 调度器），用于引入的两阶段训练。总损失中的超参数 λ_1 、 λ_2 和 λ_3 分别设置为 1、0.1 和 0.1。每个阶段包含 100 000 次迭代。攻击是样本级别的，批次中的每个音频段在训练期间都会经历不同类型的攻击。攻击层的配置见表1。

将定位码的长度设置为 10 位，因此与提取的非定位码位序列发生潜在冲突的概率约为 $1/2^{10}$ 。在每一批中，定位码和水印信息都是随机生成的，以确保训练后的模型能够处理任何 0~1 序列的排列。实验中

表1 攻击层参数设置

ID	攻击	描述
GN	高斯加性噪声	向含水印的音频中添加高斯噪声，并将信噪比保持在35 dB左右
LF	低通滤波器	使音频通过5 000 Hz的低通滤波器，滤波器的范围根据人类听觉范围设定
CP	MP3压缩	将波形压缩为64 000 bit/s的MP3格式,然后转换回wav格式，此转换过程会导致信息丢失
QZ	量化	将含水印音频波形的采样点量化为512个级别
RD	随机丢弃	将含水印音频中0.1%的采样点的值随机设置为零
RS	重采样	将音频重采样到新的采样率(原始采样率的200%)，然后重新采样回原始采样率
AM	幅度修改	将音频的整体幅度乘以90%的修改因子
TS	时间拉伸	在时域将音频压缩到原始长度的90%，然后拉伸以保持原始长度

选择了两个现有的神经音频水印工作 WavMark^[12]和 DeAR^[13]作为基线。

3.2 评价指标

信噪比（SNR）用于衡量水印的影响。它的计算方法是宿主音频功率与水印噪声功率之比的对数的10倍。准确率（ACC）用于衡量提取的信息与真实信息之间的差异。ACC通常用于衡量水印方法的鲁棒性。容量是指在确保水印的不可感知性和ACC的情况下，每秒音频可以嵌入的水印比特数。

表2 各种水印方法的基本指标比较(↑表示值越大越好)

方法	SNR/dB ↑	ACC ↑	容量/(bit·s ⁻¹) ↑
DeAR	26.18	99.61%	8.8
本文方法(10+10)	40.43	99.64%	20
WavMark	38.55	99.35%	32
本文方法(22+10)	37.72	99.52%	32
本文方法(46+10)	35.41	99.44%	56

3.3 实验结果

3.3.1 整体对比

各种水印方法的基本指标见表2。由于每个基线水印模型都针对不同的容量而设计，并且提出的音频水印方法拥有最大的容量，因此还训练了另外两个模型，它们的容量与其他方法相似或更大，以减轻容量对实验对比的影响。由于定位码和信息在提出的模型中是分开的，因此将定位码和信息的总长度作为提出的音频水印方法的容量，并且定位码的长度固定为10位。发现容量较低的水印模型具有更好的不可感知性和更高的准确性。当设计具有较大的有效载荷时，本文提出的音频水印方法获得了相当大的SNR和ACC。

3.3.2 鲁棒性对比

实验测量了每个模型在不同攻击下的水印提取准确率，以评估它们的鲁棒性。考虑了8种常见攻击，包括高斯加性噪声（GN）、低通滤波器（LF）、MP3压缩（CP）、量化（QZ）、随机丢弃（RD）、重

采样 (RS)、幅度修改 (AM) 和时间拉伸 (TS)。如上所述, 水印去除攻击需要考虑对音频的损坏程度, 每次攻击的强度设置延续之前的研究设置^[12]。鲁棒性评估结果见表3。本文提出的音频水印方法在具有更大容量的同时, 表现出与基线模型相当的鲁棒性。

3.3.3 定位测试

为了比较每种定位方法的时间成本, 在不同位置嵌入水印 (10 位定位码和 46 位信息), 并使用不同的方法进行定位。测试的方法包括两种。①与 WavMark^[12]相同, 定位码和信息连接在一起, 并通过单阶段 INN 模型嵌入音频段中。该模型使用移位策略进行训练。在定位过程中, 定位码和信息同时迭代提取。步长为块大小的 10%, 与基线方法相同。②本文提出的方法, 在定位过程中只提取定位码, 步长为块大小的 10%。方法①构建了一个单阶段 INN 水印模型, 其层数与提出的两阶段模型相同。为了消除定位过程中的误差影响, 因为无法保证方法①中容量为 56 位的模型的 ACC 只执行提取而无需进一步验证, 故遍历定位直到到达水印位置, 正如之前的实验表明的, 各模型在各自设置的承载能力下都可以达到理想的提取精度。

每种方法的定位时间测量在同一设备

上进行, 每个水印位置平均进行 100 次。在同一设备上每种方法的时间消耗如图 5 所示。结果表明, 该方法减少了大约 40%~50% 的时间开销, 特别是当水印远离头部时, 本文提出的双嵌入策略的优势更加明显。

3.3.4 消融实验

为了验证平衡块和判别器对本文提出的音频水印方法性能的积极影响, 笔者进行了消融实验, 构建了以下与提出的方法相关的模型。①M1, 从提出的方法中移除判别器; ②M2, 在鲁棒性训练中从提出的方法中移除平衡块。这些消融模型以及本文提出的模型在相同的数据集上以相同的

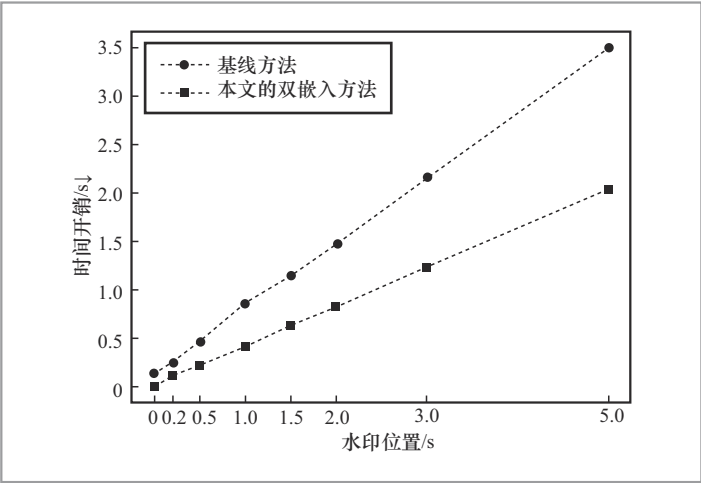


图5 在同一设备上每种方法的时间消耗比较

表3 鲁棒性评估结果

方法	GN (35 dB)	LF (5 000 Hz)	CP (64 000 bit/s)	QZ (29)	RD (0.1%)	RS (200%)	AM (90%)	TS (90%)
DeAR	99.61%	99.04%	99.55%	99.63%	99.61%	99.62%	99.61%	99.31%
本文方法(10+10)	99.47%	98.62%	99.41%	98.86%	99.6%	99.11%	99.49%	98.95%
WavMark	97.84%	98.54%	98.81%	96.6%	98.68%	98.42%	99.29%	95.35%
本文方法(22+10)	99.33%	98.53%	99.15%	98.73%	99.23%	99.02%	99.48%	98.82%
本文方法(46+10)	98.72%	98.12%	99%	98.61%	98.84%	98.87%	99.42%	98.66%

方式进行训练，测量每个模型的基本指标和鲁棒性。

消融实验结果见表 4 和表 5。结果表明，判别器有助于提高水印音频的质量和自然度，没有判别器的模型的信噪比有所下降，并且与本文提出的方法具有相当的鲁棒性。平衡块减轻了鲁棒性训练中攻击层引起的不对称性，使模型能够获得更好的鲁棒性，并实现更高的准确性。总的来说，判别器和平衡块的引入增强了水印质量和水印模型的稳定性。

3.4 讨论分析

本文重点关注神经音频水印定位的高开销问题，创新性地将定位码与水印信息分离，并分别进行嵌入/提取。然而，笔者发现双嵌入水印方法目前存在以下局限性：①训练模型的定位码和信息的长度不能灵活调整，因为两阶段可逆神经网络的设计固定了这两者的长度，不能像以前的方法那样将任意长度开始的比特指定为定位码；②通过本文提出的模型，嵌入在低能量音频中的水印不如嵌入在高能量音频中的水

印不可感知，原因可能是，除了选择两种类型的数据集进行训练外，模型的设计及其训练策略没有考虑载体的能量。

随着社交媒体的快速发展和版权意识的增强，数字水印技术具有广阔的应用前景。与传统方法相比，神经音频水印由于其鲁棒性、独立于专家干预和可扩展性，仍然是一个有价值的研究领域。然而，神经音频水印还没有完全超越传统的数字水印方法，特别是在容量方面。未来关于神经音频水印的工作可以考虑以下方面，但不限于此：①通过新颖的神经网络架构或嵌入策略来最大化容量；②探索自适应方法来缩小高能量音频和低能量音频之间的性能差距；③考虑直接在音频生成模型中实现水印嵌入，以获得带有水印的生成音频，而不是使用后处理水印方法。

4 结束语

本文提出了一种神经音频水印模型，该模型通过设计的两阶段 INN 分别嵌入定位码和水印信息。针对神经音频水印定位的挑战，双嵌入策略避免了同时提取定位码和信息的大量计算，从而加快了定位过程。另外，为了减轻攻击层引入的不对称因素，引入平衡块为提取器提供了训练空间，同时保持了可逆网络的优势。实验结果表明，本文提出的音频水印方法的不可感知性、鲁棒性和定位效率均取得了较好

表 4 基础指标消融对比

方法	SNR/dB ↑	ACC ↑
M1	32.3	99.4%
M2	35.27	98.7%
本文方法(46+10)	35.41	99.44%

表 5 消融实验中的鲁棒性比较

方法	GN (35dB)	LF (5 000 Hz)	CP (64 000 bit/s)	QZ (29)	RD (0.1%)	RS (200%)	AM (90%)	TS (90%)
M1	98.25%	98.18%	98.79%	98.73%	98.9%	98.46%	99.28%	98.58%
M2	98.14%	97.79%	98.55%	96.07%	98.24%	98.39%	97.62%	96.57%
提出方法(46+10)	98.72%	98.12%	99%	98.61%	98.84%	98.87%	99.42%	98.66%

的性能,在水印嵌入及提取质量理想时,本文提出的音频水印方法对比基线方法DeAR水印容量提升了11.2 bit/s,对比基线方法WavMark水印容量提升了24 bit/s,并且其双嵌入的策略带来了明显的定位效率提升。

参考文献:

- [1] SINGHA A, ULLAH M A. Development of an audio watermarking with decentralization of the watermarks[J]. Journal of King Saud University – Computer and Information Sciences, 2022, 34(6): 3055–3061.
- [2] ZHANG G F, ZHENG L L, SU Z P, et al. M-sequences and sliding window based audio watermarking robust against large-scale cropping attacks[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1182–1195.
- [3] POTDAR V M, HAN S, CHANG E. A survey of digital image watermarking techniques[C]//Proceedings of the 3rd IEEE International Conference on Industrial Informatics, 2005. Piscataway: IEEE Press, 2005: 709–716.
- [4] HUSSAIN I, ZENG J, QIN X, et al. A survey on deep convolutional neural networks for image steganography and steganalysis[J]. KSII Transactions on Internet and Information Systems, 2020, 14(3): 1228–1248.
- [5] CHANCHAL M, MALATHI P, GIREESH KUMAR T. A comprehensive survey on neural network based image data hiding scheme[C]//Proceedings of the 2020 Fourth International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud). Piscataway: IEEE Press, 2020: 1245–1249.
- [6] AMRIT P, SINGH A K. Survey on watermarking methods in the artificial intelligence domain and beyond[J]. Computer Communications, 2022, 188: 52–65.
- [7] SINGH R, SARASWAT M, ASHOK A, et al. From classical to soft computing based watermarking techniques: a comprehensive review[J]. Future Generation Computer Systems, 2023, 141: 738–754.
- [8] LEE H, LEE E, JUNG J, et al. Surface stickiness perception by auditory, tactile, and visual cues[J]. Frontiers in Psychology, 2019, 10: 2135.
- [9] LIU Y, GUO M X, ZHANG J, et al. A novel two-stage separable deep learning framework for practical blind watermarking[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1509–1517.
- [10] PAVLOVIC K, KOVACEVIC S, DUROVIC I. Speech watermarking using deep neural networks[C]//Proceedings of the 2020 28th Telecommunications Forum. Piscataway: IEEE Press, 2020: 1–4.
- [11] PAVLOVIĆ K, KOVAČEVIĆ S, DJUROVIĆ I, et al. Robust speech watermarking by a jointly trained embedder and detector using a DNN[J]. Digital Signal Processing, 2022, 122: 103381.
- [12] CHEN G Y, WU Y, LIU S J, et al. WavMark: watermarking for audio generation[EB]. arXiv preprint, 2023, arXiv: 2308.12770.
- [13] LIU C, ZHANG J, FANG H, et al. DeAR: a deep-learning-based audio re-recording resilient watermarking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(11): 13201–13209.
- [14] DINH L, KRUEGER D, BENGIO Y. NICE: non-linear independent components estimation[EB]. arXiv preprint, 2014,

- arXiv: 1410.8516.
- [15] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density estimation using real NVP [EB]. arXiv preprint, 2016, arXiv: 1605.08803.
- [16] ARDIZZONE L, LÜTH C, KRUSE J, et al. Guided image generation with conditional invertible neural networks[EB]. arXiv preprint, 2019, arXiv: 1907.02392.
- [17] BEHRMANN J, GRATHWOHL W, CHEN R T Q, et al. Invertible residual networks[C]//Proceedings of the International Conference on Machine Learning. [S.l.: s.n.], 2019: 573–582.
- [18] KINGMA D P, DHARIWAL P. Glow: generative flow with invertible 1x1 convolutions[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 10236–10245.
- [19] VAN DER OUDERAA T F A, WORRALL D E. Reversible GANs for memory-efficient image-to-image translation [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4715–4723.
- [20] LUGMAYR A, DANELLJAN M, VAN GOOL L, et al. SRFlow: learning the super-resolution space with normalizing flow[C]//Computer Vision – ECCV 2020. Cham: Springer, 2020: 715–732.
- [21] WANG Y L, XIAO M Q, LIU C, et al. Modeling lost information in lossy image compression[EB]. arXiv preprint, 2020, arXiv: 2006.11999.
- [22] LU S P, WANG R, ZHONG T, et al. Large-capacity image steganography based on invertible neural networks[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10816–10825.
- [23] MOU C, XU Y M, SONG J C, et al. Large-capacity and flexible video steganography *via* invertible neural network[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 22606–22615.
- [24] MUSHGIL B M, ADNAN W A, AL-HADAD S A R, et al. An efficient selective method for audio watermarking against de-synchronization attacks[J]. Journal of Electrical Engineering and Technology, 2018, 13(1): 476–484.
- [25] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139–144.
- [26] MA R, GUO M X, HOU Y, et al. Towards blind watermarking: combining invertible and non-invertible mechanisms[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1532–1542.
- [27] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261–2269.
- [28] VEAUX C, YAMAGISHI J, MACDONALD K. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92) [EB]. Semantic Scholar, 2016: 64303572.
- [29] DEFFERRARD M, BENZI K, VANDERGHEYNST P, et al. FMA: a dataset for music analysis[EB]. arXiv preprint, 2016, arXiv: 1612.01840.
- [30] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference on Learning Representations. [S. l.: s. n.], 2015.

作者简介



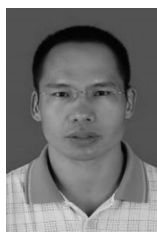
张旭龙（1988–），男，博士，平安科技（深圳）有限公司，高级算法研究员，复旦大学计算机理学博士，2023年入选上海市东方英才计划青年项目。兼任深圳清华大学研究院及中国科学技术大学先进技术研究院校外导师，中国自动化学会联邦数据与联邦智能专业委员会委员。目前是IEEE、中国自动化学会以及中国计算机学会会员。主要研究方向为大模型、具身智能、跨模态智能计算以及模型上下文协议（MCP）等。发表学术论文80余篇，申请国家发明专利授权20余项。



瞿晓阳（1988–），男，博士，平安科技（深圳）有限公司前沿机器学习算法分组负责人，清华大学深圳国际研究生院校外导师，中国科技大学校先进技术研究院校外导师，中佛罗里达大学访问学者，从事机器学习、大数据、体系结构方面的研究工作，在语音语义分析、自动化机器学习、零样本和小样本学习、高性能计算与存储等方面经验丰富。在体系结构方向（如INFOCOM、DAC、IPDPS、TPDS）和人工智能方向（如NeurIPS、IJCAI、ICASSP、Interspeech）等国际会议和期刊发表过近50篇文章，其中1篇论文荣获会议最佳学生论文奖提名；担任多个国际期刊的评委，已授权专利70篇，已出版专著2本。



李鹏程（1999–），男，中国科学技术大学硕士生，平安科技（深圳）有限公司算法工程师，中国自动化学会学生会员，主要研究方向为语音合成、语音转换和语音安全。



肖春光（1976–），男，深圳市宝安区教育事业发展中心高级教师，主要研究方向为教育数字化、网络安全、信息化教学创新。



王健宗（1983–），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

收稿日期: 2025-01-24

通信作者: 王健宗, jzwang@188.com; 肖春光, 511952519@qq.com

基金项目: 广东省重点领域研发计划“新一代人工智能”重大专项(No.2021B0101400003)

Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)