IEEE Open Journal of Signal Processing

Received 7 September 2023; accepted 3 December 2023. Date of publication 1 January 2024;
date of current version 17 January 2024. The review of this article was arranged by Associate Editor Lisandro Lovisolo.

Digital Object Identifier 10.1109/OJSP.2023.3349113

Adversarial Representation Learning for Robust Privacy Preservation in Audio

SHAYAN GHARIB [1], **MINH TRAN** [2] **(Graduate Student Member, IEEE), DIEP LUONG** [2], **KONSTANTINOS DROSSOS** [2,3] **(Member, IEEE), AND TUOMAS VIRTANEN** [2] **(Fellow, IEEE)**

[1] Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland
[2] Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland
[3] Nokia Tech, 02610 Espoo, Finland

CORRESPONDING AUTHOR: TUOMAS VIRTANEN (e-mail: tuomas.virtanen@tuni.fi)

The work of Minh Tran and Diep Luong was supported by Research Council of Finland under Grant 338502.

ABSTRACT Sound event detection systems are widely used in various applications such as surveillance and environmental monitoring where data is automatically collected, processed, and sent to a cloud for sound recognition. However, this process may inadvertently reveal sensitive information about users or their surroundings, hence raising privacy concerns. In this study, we propose a novel adversarial training method for learning representations of audio recordings that effectively prevents the detection of speech activity from the latent features of the recordings. The proposed method trains a model to generate invariant latent representations of speech-containing audio recordings that cannot be distinguished from non-speech recordings by a speech classifier. The novelty of our work is in the optimization algorithm, where the speech classifier's weights are regularly replaced with the weights of classifiers trained in a supervised manner. This increases the discrimination power of the speech classifier constantly during the adversarial training, motivating the model to generate latent representations in which speech is not distinguishable, even using new speech classifiers trained outside the adversarial training loop. The proposed method is evaluated against a baseline approach with no privacy measures and a prior adversarial training method, demonstrating a significant reduction in privacy violations compared to the baseline approach. Additionally, we show that the prior adversarial method is practically ineffective for this purpose.

INDEX TERMS Adversarial neural networks, adversarial representation learning, privacy preservation, sound event detection.

I. INTRODUCTION

The proliferation of ever-present devices equipped with sensors has led to an exponential increase in data availability. These devices continuously collect and process large amounts of data, facilitating remarkable advancements in various machine learning tasks [1], [2], [3]. However, this trend raises concerns regarding the privacy of users' personal information both during the data collection process and when the data is utilized by machine learning models [4], [5]. Speech interfaces and acoustic monitoring are prominent areas among those with active research focusing on preserving user data privacy. These systems record audio which may contain biometric information such as human voices that can be identified and attributed to individuals. Therefore new legislation has been enacted to safeguard users against the inherent risks associated with the exposure of personal information [6]. Acoustic pattern classification has numerous applications in smart cities, smart homes, and context-aware devices [7]. These systems aim to automatically detect targeted sound events such as sirens, birds chirp, and window breakage, among others. While the primary focus of these methods may not be on speech-related information, the environments where they operate often include speech. Human speech contains a significant amount of personal information, including speakers' identity, gender, accent, or sensitive content discussed during conversations [6], [8]. Many voice interface and acoustic monitoring systems, including daily consumer devices such as smartphones, locally extract features from

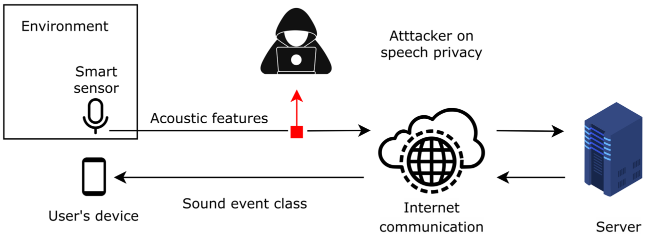

FIGURE 1. Illustration of the problem setup where speech privacy is compromised during the transmission of acoustic features to a cloud platform.

audio and transmit them to a cloud for recognition tasks [9]. Unauthorized access to this information by adversaries can have detrimental consequences for the individuals involved. Fig. 1 provides an illustration of a typical setup for this scenario, highlighting that the disclosure of such information should only occur with speakers' consent.

An example of this scenario, which served as the primary motivation for conducting this research, pertains to automatic sound recognition devices employed in home care settings. The objective was to develop a device capable of promptly notifying nurses when an elderly individual is in a dangerous situation requiring assistance. The system actively monitors the surrounding soundscape to trigger an alert in the event of an emergency. Therefore, it inevitably captures many speech signals. To ensure privacy, it becomes imperative for recording devices to hide the speech-related information within the encoded features of the signals to protect such information during the data transmission.

To address this challenge, the objective of this study is to integrate a privacy-preserving algorithm into the representation learning process in order to base final classification decisions on features that ensure users' privacy. To achieve this, we employ an adversarial learning setup based on deep neural networks (DNNs) to create latent representations of audio recordings that contain information required for the recognition of targeted sound events, while removing information that could be used for speech analysis.

Inspired by Ganin and Lempitsky [10], our adversarial setup includes two neural networks: a feature extractor and a speech classifier. The feature extractor is designed to manipulate the latent features in such a way that it confuses the speech classifier, thereby reducing its performance on speech classification tasks.

Although the general idea has been used previously in a different application, this approach is vulnerable to the retrieval of speech attributes [11], especially when the learned latent features are used to re-train a separate speech classifier outside the adversarial process. To address this vulnerability, we propose a straightforward solution that can be seamlessly integrated into the learning process. The speech classifier is frequently replaced by a new one, that is trained until convergence, outside of the adversarial learning process. This approach ensures that the speech classifier is not easily tricked by the feature extractor, creating a robust training process that ensures the speech-related information is not retrievable. Throughout this paper, we refer to this algorithm as *robust discriminative adversarial learning* (RDAL).

The main contribution of this paper lies in the introduction of the RDAL algorithm, which facilitates the generation of adversarial learning representations that effectively prevent the detection of sensitive information within the latent feature space derived from acoustic features, thereby enabling robust sound event classification. A preliminary version of this work was previously published as a conference paper [12]. This paper introduces the RDAL algorithm, which was not previously covered in the conference paper. While the conference paper primarily focused on evaluating the effectiveness of source separation in conjunction with RDAL, it did not delve into the detailed explanations, rationale, and vulnerabilities addressed by RDAL. In this extended version, we present comprehensive insights into the RDAL algorithm, including a thorough exploration of the algorithm itself, its underlying principles, and its efficacy in mitigating observed vulnerabilities in prior studies. Furthermore, to enhance the understanding of RDAL's generalizability across various sound event classes, we expand our dataset from our previous work in [12]. Lastly, we incorporate gender classification into our evaluation setup to demonstrate RDAL's performance on this novel task.

II. RELATED WORK

In this section, we discuss previous works on privacy preservation in two audio-related tasks: machine listening and speech recognition. While the application of adversarial training for learning privacy-preserving features in machine listening tasks has not been extensively explored, there have been successful attempts to use adversarial training to anonymize speech characteristics, such as speaker identity, in automatic speech recognition (ASR) systems.

A. PRIVACY-PRESERVATION IN MACHINE LISTENING TASKS

Larson et al. [13] developed a cough detection system for mobile phones, aiming to render speech unintelligible in the reconstructed audio. They used principal component analysis (PCA) analysis on cough sound spectrograms to select eigenvectors with the biggest eigenvalues for audio reconstruction. However, PCA's limited learning capacity reduced system performance, especially for polyphonic sound event detection. Additionally, unintelligible speech in the recovered audio does not guarantee that speech information cannot be extracted. Wang et al. [14] assert that human speech predominantly falls within the frequency range of 80 Hz to 3 kHz. To recognize indoor human activity, they suggest using a bandstop filter to remove speech from the audio signal. However, this filtering process also results in the loss of information related to other sound events. Consequently, machine listening performance is adversely affected, particularly for events that share a significant frequency range with human speech.

Nelus and Martin [15] propose privacy-preserving representation learning using DNNs to extract variational information. Their objective is to generate informative latent representations from log mel-band energy for the classification task while minimizing speaker information. This is achieved by minimizing the mutual information between mel-band energy and extracted features by DNNs, ensuring low dependency between these two features. Simultaneously, the model is trained to keep the extracted latent features informative for classification.

B. ADVERSARIAL PRIVACY-PRESERVING REPRESENTATIONS IN SPEECH RECOGNITION TASKS

In previous research on speech recognition tasks, adversarial representation learning was employed to enhance the robustness of predictive models by disregarding irrelevant information. Specifically, subsequent studies utilized adversarial learning to generate speaker-invariant features, as speaker variability can have a detrimental impact on the performance of acoustic modeling systems.

Meng et al. [16] minimize speaker information in audio representations for senone classification. They introduce a minimax speaker classification objective to generate bottleneck features that are speaker-invariant and yet discriminative for senone predictions. Similarly, Tsuchiya et al. [17] employ an adversarial training setup to develop a speaker-invariant representation of audio data for zero-resource language acoustic modeling. These studies share similarities with privacy-preserving representation learning as they eliminate speaker information from the extracted features.

Srivastava et al. [11] conducted one of the initial studies on privacy preservation in the audio domain using adversarial learning. Their objective was to conceal speaker identities in ASR systems by anonymizing the latent representations of an end-to-end ASR network. This was achieved through a minimax objective between the encoded representations from a speech encoder and a speaker classifier. While the system reduced performance in close-set speaker identification, residual information about speakers could still be recovered, leading to improved performance in open-set speaker verification.

Later studies focused on removing specific speaker attributes instead of general speaker identity information. For example, Noé et al. [18] used adversarial training to preserve the privacy of speakers' gender information in an automatic speaker verification system. While this study investigated a single attribute, in practice, there may be a need to conceal multiple types of information. Perero-Codosero et al. [6] reconstructed privacy-preserving x-vectors using multiple adversarial privacy domains related to different speaker attributes, such as ID, gender, and accent. They demonstrated that incorporating multiple adversarial privacy domains improved both utility tasks and privacy performance. However, the recoverability of sensitive speech information in audio features after adversarial learning, as observed in [11], has not been addressed by proposed adversarial learning systems. The privacy-preserving features obtained through adversarial training do not guarantee complete removal of sensitive information or prevention of recovery.

III. METHOD

A. PROBLEM SETUP

In this study, we focus on identifying a utility attribute $\mathbf{y}$ using a latent representation $\mathbf{z}$ derived from acoustic features $\mathbf{x}$. Our goal is to ensure that $\mathbf{z}$ contains minimal information related to a sensitive attribute $\mathbf{s}$. While our specific utility attribute is targeted sound event classes and the sensitive attribute is speech presence, our approach can be extended to other attributes, such as speaker identity, accent, or gender. Given that speech presence is a prominent type of speech information, our focus is primarily on speech presence estimation. We assume that a method capable of removing information about speech presence would also be effective in removing information about other speech characteristics, which are generally more challenging to recognize. Following the above problem setup, we assume that we have access to a labeled dataset $\mathbb{X}$, consisting of N data samples $\mathbf{x}$ accompanied by sound event labels $\mathbf{y}$ and speech labels $\mathbf{s}$, i.e. $\mathbb{X} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i)\}_{i=1}^{N}$.

Given an input $\mathbf{x}$, our goal is to compute a latent representation $\mathbf{z}$ using a feature extractor F, i.e. $\mathbf{z} = F(\mathbf{x})$, that enables a classifier C to perform multi-class classification for targeted sound events where the goal is to classify each input into one of the predefined sound event classes denoted as $\mathbf{y} \in \{1, 2, \ldots, Y\}$, resulting in an estimated class $\hat{\mathbf{y}} = C(\mathbf{z})$. To prevent disclosing speech-related information, the latent representation $\mathbf{z}$ should not reveal any indications that allow classification of $\mathbf{x}$ into its speech class $\mathbf{s}$ using a speech classifier D, i.e. $\hat{\mathbf{s}} = D(\mathbf{z})$. We formulate this problem such that both goals are met simultaneously. Fig. 2 illustrates each component of our method and their interconnection.

B. ROBUST DISCRIMINATIVE ADVERSARIAL LEARNING

We build upon a discriminative adversarial learning approach which was initially introduced by Ganin and Lempitsky [10] for the purpose of obtaining domain invariant representations in an unsupervised domain adaptation task.

To achieve a well performing sound event classification, the feature extractor F and the sound event classifier C are jointly trained to predict the present sound event in an input, i.e. $\hat{\mathbf{y}}_i = C(F(\mathbf{x}_i))$. As the first part of our algorithm, the objective function

$$\min_{F,C} \mathcal{L}_{cls} = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathbb{X}} \sum_{i=1}^{N} \mathbf{1}_{[i=\mathbf{y}]} \log(\hat{\mathbf{y}}_i), \tag{1}$$

is then minimized by optimizing the parameters of F and C in order to reduce the classification error between the true labels $\mathbf{y}_i$ of targeted sound events and their corresponding predictions $\hat{\mathbf{y}}_i$.

In order to prevent the recognition of any speech information, we use an adversarial training method consisting of two components: the feature extractor F and the speech classifier D. More specifically, D is employed to predict present speech in an input $\mathbf{x}_i$ based on latent features $\mathbf{z}_i$. This is achieved using

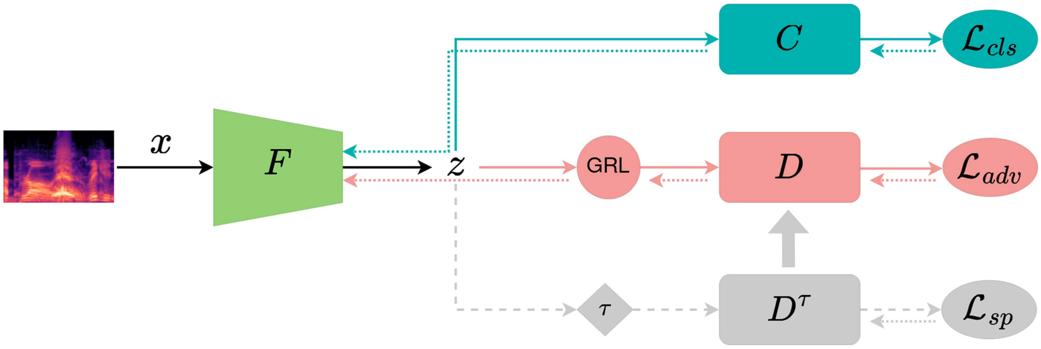

FIGURE 2. Schematic diagram of the proposed method. F, C, D, and D^τ are neural networks and $\mathcal{L}$ denotes different loss terms employed in our method. The solid lines illustrate the regular forward pass. The dashed line actives after τ epochs. Finally, the dotted lines represent the backpropagation of each specific error w.r.t the associated parameters.

the objective function

$$\max_F \min_D \mathcal{L}_{adv} = \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{X}} \sum_{i=1}^{N} \ell(\mathbf{s}_i, D(F(\mathbf{x}_i))). \tag{2}$$

We use binary cross entropy as the loss function ℓ in our experiments. The primary objective of the speech classifier D is to achieve optimal performance in speech classification. However, within the adversarial learning framework, the feature extractor F aims to obfuscate D by minimizing its ability to classify speech based on the internal representations $\mathbf{z}_i$. To facilitate this, a gradient reversal layer (GRL) module is introduced, connecting D and F in the network [10]. The GRL operates differently during forward and backward propagation. In the forward pass, it functions as an identity mapping, preserving the input as the output. However, during backward propagation, it multiplies the partial derivatives of the adversarial loss $\mathcal{L}_{adv}$ with respect to the feature extractor parameters θ_F, denoted as $\frac{\partial \mathcal{L}_{adv}}{\partial \theta_F}$, by a negative coefficient $-\lambda$, where $\lambda \geq 0$. This implies that when $\lambda = 0$, the feature extractor F is solely optimized based on the classification objective $\mathcal{L}_{cls}$. As λ increases, the contribution of $\mathcal{L}_{adv}$ in optimizing θ_F becomes more pronounced. Consequently, F is encouraged to generate invariant representations that discard speech-related information, thereby undermining the speech classification performance of D. Therefore, we can summarize the optimization process of F as

$$\theta_F \longleftarrow \theta_F - \mu \left(\frac{\partial \mathcal{L}_{cls}}{\partial \theta_F} - \lambda \frac{\partial \mathcal{L}_{adv}}{\partial \theta_F} \right) \tag{3}$$

where μ is the learning rate.

Although the minimax objective in (2) leads to the convergence of the adversarial training and subsequently reduces the speech classification performance of the speech classifier D, it does not guarantee that the resulting representations $\mathbf{z}$ are free from sensitive attribute information $\mathbf{s}$. This limitation becomes evident when training a new speech classifier solely on $\mathbf{z}$ outside the adversarial learning process. A similar issue was identified by Srivastava et al. [11], who aimed to anonymize speakers in an ASR system. Srivastava et al. [11] found that the method's generalization performance was hindered by limitations in the representation capacity of the adversarial branch. It is important to note that this problem extends beyond privacy-preserving representations learned within an adversarial framework.

In a related study conducted by Jin et al. [19], they address the same issue from a broader perspective by aiming to minimize distribution shift in the latent space through a discriminative adversarial setup, similar to our approach. As the adversarial training progresses, the alignment between the distributions of latent features for different speech classes increases, resulting in reduced discriminative ability of the speech classifier D in distinguishing between them. Consequently, the feature extractor F has less incentive to further align the latent representations, posing challenges for optimizing F and D to learn invariant representations within this adversarial setting.

To address such vulnerability in the optimization of adversarial branch, we propose a mechanism aimed at enhancing the discrimination power of the speech classifier D and ensuring the generation of generalizable and robust privacy-preserving latent features. RDAL introduces a supervised training step using the latent representations $\mathbf{z}$ after every τ epochs of adversarial training to train a new speech classifier, denoted as D^τ, in a supervised manner. Subsequently, the parameters of D are updated using those of D^τ before continuing adversarial process. This iterative process compels the feature extractor F to continuously modify its outputs, making the representations of speech classes indistinguishable so that the new experts/speech classifiers are not able to distinguish between them. This process is repeated until further training iterations no longer lead to improved performance of D^τ. In our study, speech labels can be represented using binary labels $\mathbf{s}$, and the training of D^τ is done by minimizing

$$\begin{aligned} \min_{D^\tau} \mathcal{L}_{sp} = -\mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{X}} \sum_{i=1}^{N} \mathbf{s}_i \log(D^\tau(F(\mathbf{x}_i))) \\ +(1-\mathbf{s}_i)\log(1-D^\tau(F(\mathbf{x}_i))). \end{aligned} \tag{4}$$

Notably, only the parameters of D^τ are optimized, while the parameters of F are kept fixed. The details of the RDAL

Algorithm 1: Robust Discriminative Adversarial Learning (RDAL).

```
Require: Labelled data $\mathbb{X}$, trainable networks
  parameters: $F_\theta, C_\theta, D_\theta, D_\theta^\tau$
Require: Learning rate $\mu$, batch size $B$, training $D_\theta^\tau$ every
  $\tau$ epochs
Require: GRL multiplier $\lambda(m)$ at epoch $m$
while NOT converged do
    Sample a mini-batch $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i)\}_{i=1}^B$
    $\mathcal{L}_{cls} = \mathrm{CE}(\{(C_\theta(F_\theta(\mathbf{x}_i)), \mathbf{y}_i)\}_{i=1}^B)$
    $\mathcal{L}_{adv} = \mathrm{BCE}(\{(D_\theta(F_\theta(\mathbf{x}_i)), \mathbf{s}_i)\}_{i=1}^B)$
    $\theta_C \longleftarrow \theta_C - \mu(\frac{\partial \mathcal{L}_{cls}}{\partial \theta_C})$,
    $\theta_D \longleftarrow \theta_D - \mu(\frac{\partial \mathcal{L}_{adv}}{\partial \theta_D})$,
    $\theta_F \longleftarrow \theta_F - \mu(\frac{\partial \mathcal{L}_{cls}}{\partial \theta_F} - \lambda \frac{\partial \mathcal{L}_{adv}}{\partial \theta_F})$
    if $m$ mod $\tau = 0$ then
        Initialize $D_\theta^\tau$
        while NOT converged do
            Sample a mini-batch $\{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^B$
            $\mathcal{L}_{sp} = \mathrm{BCE}(\{(D_\theta^\tau(F_\theta(\mathbf{x}_i)), \mathbf{s}_i)\}_{i=1}^B)$
            $\theta_{D^\tau} \longleftarrow \theta_{D^\tau} - \mu(\frac{\partial \mathcal{L}_{sp}}{\partial \theta_{D^\tau}})$
        end while
        $\theta_D \longleftarrow \theta_{D^\tau}$
    end if
end while
```

method are fully outlined in Algorithm 1. In this study, the capability of RDAL algorithm is enhanced by augmenting a masking U-net architecture prior to the feature extractor F, as outlined in [12].

IV. EVALUATION

In this section, we evaluate the performance of our method in obfuscating speech information in the latent feature space while simultaneously performing sound event recognition for targeted classes as the primary utility task.[1] We measure the effectiveness of preserving privacy in audio recordings by conducting speech presence classification and gender classification in one-second audio segments. Additionally, we assess the performance on the utility task, which involves classifying sound events within these one-second segments. In our evaluation the segments are isolated from each other and not processed as a part of continuous audio, but we term the tasks as speech activity detection (SAD), gender detection (GD), and sound event detection (SED), to indicate that in a realistic application scenario we would be processing a continuous stream of segments, and segment-wise classification would lead to detecting the temporal activities of classes.

We compare the performance of RDAL against *baseline* and naive adversarial methods. The *baseline* is defined as a sound event classification system where no privacy measures are taken, meaning F and C are optimized jointly using (1) in a supervised manner without the adversarial branch. In addition, the naive adversarial method does not include D^τ and therefore F, C, D are optimized only using (1), (2), and (3). We refer to this method in the rest of the paper as *NaiveAdv*. In order to augment the capabilities of the RDAL algorithm and further improve its performance, we incorporate the masking network, as detailed in our preliminary study [12]. This enhanced variant is denoted as RDAL+M in this paper. By integrating a U-Net architecture of DNNs prior to the feature extractor F, RDAL+M aims to separate the speech component from the magnitude spectrogram of each data sample and reconstruct a non-speech version of that sample. The masking network is pre-trained and remains unchanged throughout the adversarial training process. More details of the masking approach are provided in [12].

TABLE 1. Number of One-Second Sound Event Samples in Each Split of Our Dataset

Sound events	Train	Validation	Test
Dog barking	608	66	96
Glass breaking	480	52	62
Gun shot	469	52	179
Cough	384	42	132
Slam	383	42	93
Applause	425	46	62
Dishes, pot, and pan	298	36	73
Toilet flush	324	36	52
Cat meowing	208	22	94
Doorbell	174	18	49
Crying	171	18	40
Drill	268	28	61

A. DATASET

In order to address the problem formulation outlined in Section III, it is necessary to use audio recordings which include both targeted sound events and speech. To achieve this, we create a simulated dataset using real-world audio recordings to generate one-second mixtures containing speech and other sound events.

We collect the sound event data from the FSD50K dataset [20]. Among the 144 leaf nodes in the FSD50K dataset, we select 12 specific sound event classes that are potentially applicable in acoustic monitoring applications. These selected classes are listed in Table 1. For each sample belonging to the target sound event classes, we extract the two most energetic one-second segments to ensure an adequate number of audio segments for each sound event class. These segments are then normalized by subtracting their mean and dividing by their standard deviation. If a recording is shorter than one second, we pad it with zeros at the end after normalization. The processed samples are divided into the *development* and *test* splits, following the “development” and “evaluation” splits defined in the FSD50 K dataset.

The speech content for our dataset is sourced from the LibriSpeech corpus [21]. To match the sampling frequency

[1] [Online]. Available: https://github.com/lndip/RDAL

of the FSD50K dataset, we resample the recordings from 16 kHz to 44.1 kHz. Similar to the procedure used for the FSD50K samples, we extract the most energetic one-second speech segment from each audio recording and normalize these segments. The selected segments are obtained from the LibriSpeech "train-clean-100" and "dev-clean" sets. These segments are initially attenuated by 5 dB, then mixed with the processed one-second segments of sound events from the *development* and *test* sets respectively.

The *development* set comprises speech content from 126 male and 125 female speakers, while the *test* set consists of 20 speakers from both genders. In selecting speech recordings from the LibriSpeech corpus, we ensure an equal representation of male and female speakers across each targeted sound event class. Furthermore, the number of extracted one-second speech segments from LibriSpeech is half of the number of sound event segments. This approach allows us to create mixtures with a balanced number of samples for both the speech and non-speech classes.

To facilitate model selection during training, the *development* set is further divided randomly into the *train* and *validation* sets in a 9:1 ratio. This division ensures that half of the mixtures within each split, across sound event classes, contain speech. Specifically, the number of samples containing speech in *train*, *validation*, and *test* are 2094, 227, and 494, respectively. However, there are no specific constraints on speaker allocation. Therefore, the speakers in the *train* and *validation* sets may overlap, and the ratio of male to female speakers may not be balanced within the two splits.

We utilize log-mel spectrograms as the low-level features, denoted as $\mathbf{x}$, for the audio mixtures. The parameters for this transformation are derived from the work of Kong et al. [22] but are adjusted to account for the sampling frequencies of our audio recordings (44.1 kHz instead of 32 kHz as used in [22]). In computing the short-time Fourier transform (STFT), we apply a Hamming window of size 1411 with a hop length of 441. To obtain the log-mel spectrograms, we employ 64 mel filter banks. In the RDAL+M method, the magnitude of the STFTs serves as the input to the masking network, while the log-mel spectrogram is calculated on the masked spectrogram prior to being passed through the feature extractor F.

B. NETWORK ARCHITECTURE

For our feature extractor F, we utilize the "CNN6" architecture from Kong et al. [22] as a basis, making minimal adjustments to accommodate our data size. Our adapted architecture consists of 4 convolutional blocks, each containing a 2D convolutional layer with a kernel size of 3 × 3, ReLU activation, and batch normalization. The number of convolutional filters in these blocks is 64, 128, 256, and 512, respectively. Except for the last block, all blocks incorporate max pooling with a kernel size of 2 × 2. The final convolutional block employs max global pooling to transform the 2D features into a 1D vector representation. A linear layer is then utilized to generate the latent features $\mathbf{z}$ as a 64-element vector. The sound event classifier C consists of a single linear layer with softmax activation. As for the speech classifiers D and D^{τ}, it comprises 4 linear layers, with output dimensionality of 48, 32, and 16 in the first three layers, respectively, each followed by a LeakyReLU activation function. The output layer employs sigmoid as the activation function. For RDAL+M, we employ the exact architecture of masking network described in [12].

C. TRAINING

Initially, we train the entire networks in a supervised manner for the first 30 epochs ($\lambda = 0$). This helps address stability issues associated with adversarial training during the initial iterations [23] and ensures proper training of the speech classifier D to recognize the speech presence before the start of adversarial training. Following this, we gradually increase the value of λ to initiate adversarial training. Once the maximum value of 1 is reached, λ remains fixed. The schedule for increasing λ is adapted from [10] and is defined as:

$$\lambda_{\beta} = \frac{2}{1 + \exp(-\gamma.\beta)} - 1 \tag{5}$$

where γ is set to 100 in all experiments, and $\beta \in [0, 1]$ represents the progress of the training process. After the first 30 epochs, β starts at 0 and increases with a step size determined by dividing the range from 0 to 1 into equal intervals over the course of the epochs, with a maximum number of 5000 epochs. The initial epoch for starting adversarial training has not been optimized. The optimal value of τ, indicating the number of epochs for adversarial training before training a new D^{τ}, is selected from the values in the set of {10, 20, 30, 50, 70, 100} using the validation set. The models are trained with a batch size of 64, with half of the samples containing speech. Stochastic gradient descent (SGD) with a learning rate of 0.01 and momentum of 0.9 is employed for optimizing the networks parameters. Early stopping is determined by monitoring the best $\mathcal{L}_{sp}$ value on the validation set, and if there is no improvement after 10 repetitions, the training process is halted.

D. RESULTS

Table 2 presents the accuracy results for SED, as well as the accuracy and AUC scores for SAD and GD tasks on the test set. After completing the training phase, we proceed to train a new classifier, referred to as the attacker model, using the latent representations $\mathbf{z}$. This classifier is specifically designed for the task of identifying speech presence or gender. It simulates a scenario where an unauthorized individual attempts to recognize speech activities in one-second segments using latent features. The evaluation in Table 2 is based on the average values obtained from 10 separate runs.

The results presented in Table 2 yield several noteworthy observations. Firstly, the comparable SED accuracy scores across the three adversarial methods indicate that optimizing the adversarial branch to eliminate speech activity, while simultaneously training a supervised SED system, does not hinder the optimization of the SED task. Furthermore, RDAL

TABLE 2. **Results of *Baseline*, *NaiveAdv*, RDAL, RDAL+M**

Methods	$\text{SED}_{\text{accuracy}}$	$\text{SAD}_{\text{accuracy}}$	$\text{SAD}_{\text{AUC score}}$	$\text{GD}_{\text{accuracy}}$	$\text{GD}_{\text{AUC score}}$
Baseline	0.75 ± 0.01	0.76 ± 0.01	0.84 ± 0.01	0.63 ± 0.02	0.68 ± 0.03
NaiveAdv	0.75 ± 0.01	0.75 ± 0.02	0.83 ± 0.02	0.59 ± 0.02	0.63 ± 0.03
RDAL (τ=50)	0.75 ± 0.01	0.62 ± 0.03	0.68 ± 0.04	0.59 ± 0.03	0.63 ± 0.03
RDAL+M (τ=70)	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.52 \pm 0.02}$	$\mathbf{0.52 \pm 0.02}$	$\mathbf{0.50 \pm 0.02}$	$\mathbf{0.50 \pm 0.02}$

The best results for each metric are emphasized using the bold face font.

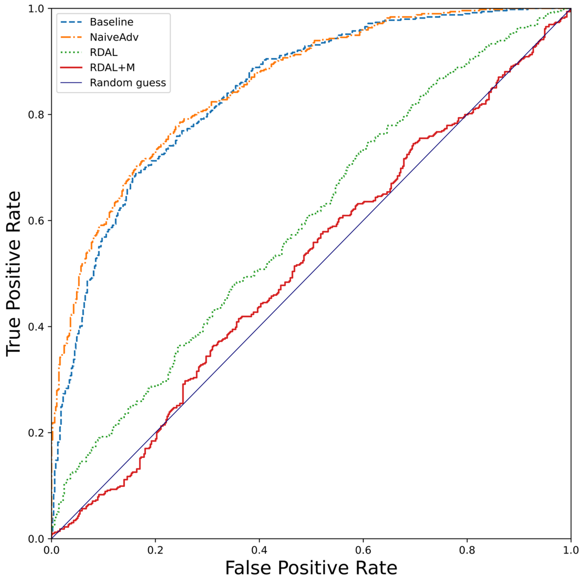

FIGURE 3. **ROC curves for each method are displayed, showcasing the privacy preservation results on the SAD task as outlined in Table 2.**

demonstrates a notable improvement in privacy preservation compared to the *baseline*, as evidenced by the SAD and GD metrics. Additionally, RDAL+M further enhances RDAL's privacy preservation performance, achieving a reduced accuracy and AUC score for both SAD and GD tasks, approaching the level of random guess scores in binary classification tasks. Thirdly, the SAD accuracy of *NaiveAdv* suggests that it does not offer enhanced privacy preservation compared to the *baseline* when evaluated against the attacker model. Therefore, *NaiveAdv* does not truly provide privacy-preserved features. Lastly, the lower GD performance across all methods indicates the inherent difficulty of this task compared to SAD. Given the absence of specific gender information during training, the scores for this task generally fall below those of SAD, making it relatively easier to obfuscate in the context of privacy preservation. Fig. 3 illustrates the receiver operating characteristic (ROC) curves of the tested methods for speaker activity detection. The curve of RDAL-M is closest to the random guess, demonstrating its best capability to preserve privacy.

E. DISCUSSION

To ensure a fair comparison among the methods presented in Table 2, we ensure the adoption of an identical architecture for DNNs models across all methods. In addition, the same scheduling of λ values is used for the *NaiveAdv*, RDAL, and RDAL+M methods.

The significant privacy preservation improvement achieved by RDAL in comparison to the *baseline* method substantiates the main claim of this paper. Furthermore, we emphasize that a naive implementation of adversarial training proves ineffective due to inherent limitations in optimizing the feature extractor to align speech and non-speech distributions. Consequently, the *NaiveAdv* method fails to effectively preserve privacy within our problem setup. Previous work by Srivastava et al. [11] has demonstrated that the privacy-preserving performance of *NaiveAdv* can even be worse than that of the *baseline* method, which lacks any privacy-preserving components, particularly in open-set classifications.

Fig. 4 showcases the 2D distributions of latent features $\mathbf{z}$ using t-SNE analysis [24]. We compare RDAL's feature distributions with a privacy-preserving lower bound system that incorporates supervised information of targeted sound events and speech presence without considering privacy preservation measures. The comparison reveals proper alignment of speech and non-speech class distributions within each targeted sound event class, providing further evidence of RDAL's improved privacy preservation performance.

Fig. 5 visualizes the kernel density estimates representing the predicted probabilities for the speech and non-speech classes. These probabilities are computed using the test set after training the attacker model. In the *baseline* method (left figure), the density curves clearly indicate the attacker's confident predictions regarding the presence or absence of speech. This highlights the information embedded in the latent representations of SED systems, even when speech is not the target sound event. In contrast, the density curves of RDAL (middle figure) exhibit significant overlap between the speech and non-speech classes, indicating an increase in model uncertainty and the attacker's challenge in distinguishing between the two. This increased uncertainty in the attacker model's predictions is a result of aligning speech and non-speech latent features through the optimized feature extractor from the RDAL method. Furthermore, RDAL+M (right figure) improves upon the results achieved by RDAL, and demonstrates a precise alignment between the density curves of predicted probabilities for the speech and non-speech classes. Given that the attacker model performs binary classification, uncertainty in its predictions can be quantified using the binary entropy function. Maximum

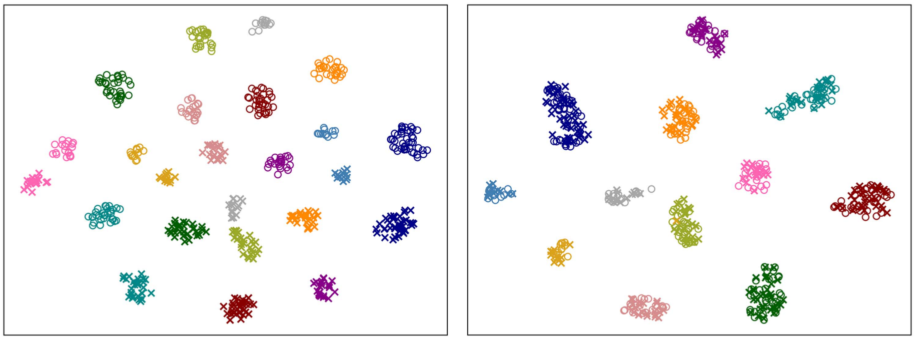

FIGURE 4. **Comparison of latent features obtained by RDAL's *F* (right) and supervised training of *F* for sound events and speech (left). Sound events are color-coded with 12 different colors, while speech and non-speech samples are marked with "o" and "x" respectively.**

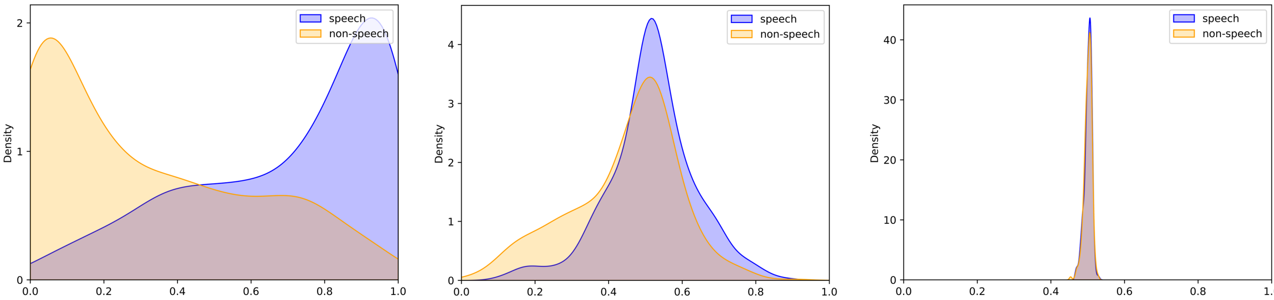

FIGURE 5. **Density curves using Gaussian kernel to represent predicted probability densities from the attacker model on the test data using the latent features of *baseline* (left), RDAL (middle), and RDAL+M (right) methods.**

uncertainty arises when the attacker predicts a sample with a probability of 0.5. As evidenced by the Fig. 5, a substantial portion of the density mass for both speech and non-speech classes is concentrated around this value.

V. CONCLUSION

Privacy breaches pose a significant threat to the confidentiality of user information and sensitive data in SED systems. To mitigate this risk, it is crucial to employ privacy-preserving algorithms that safeguard against the disclosure of private information. This study addresses the issue by formulating privacy preservation as the detection of speech activity in the latent features of audio mixtures. We introduce RDAL, an adversarial training approach, which learns robust and speech-invariant latent features. RDAL ensures agnosticism towards speech presence and gender identity, while preserving the targeted sound event information for SED.

The proposed method utilizes two neural networks: a feature extractor and a speech classifier, in a minimax game to ensure the privacy preservation of audio mixtures. The feature extractor generates invariant latent features of speech-containing audio signals that are indistinguishable from those of non-speech ones, while the speech classifier tries to distinguish between them. We also address the limitations of this approach by introducing a new speech classifier periodically into the adversarial training process to enforce the feature extractor to consistently improve the performance for aligning the distributions of speech and non-speech samples during the adversarial training process.

The empirical results indicate that the proposed RDAL approach significantly improves the privacy performance of SED systems. By effectively preserving privacy in latent features of audio mixtures, this approach can help prevent potential privacy violations and ensure the confidentiality of sensitive information. Furthermore, we demonstrated that the performance of RDAL can be further improved through its integration with a source separation method.

REFERENCES

[1] Y.-A. Chung et al., "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 244–250.

[2] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[3] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, Aug. 2022.

[4] S. Kumar, L. T. Nguyen, M. Zeng, K. Liu, and J. Zhang, "Sound shredding: Privacy preserved audio sensing," in *Proc. 16th Int. Workshop Mobile Comput. Syst. Appl.*, 2015, pp. 135–140.

[5] C. Glackin et al., "Privacy preserving encrypted phonetic search of speech data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 6414–6418.

[6] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hernández-Gómez, "X-vector anonymization using autoencoders and adversarial training for preserving speech privacy," *Comput. Speech Lang.*, vol. 74, 2022, Art. no. 101351.

[7] S. Krstulović, “Audio event recognition in the smart home,” in *Computational Analysis of Sound Scenes and Events*. Berlin, Germany: Springer, 2018, pp. 335–371.

[8] J. Williams, J. Yamagishi, P.-G. Noé, C. Valentini-Botinhao, and J.-F. Bonastre, “Revisiting speech content privacy,” in *Proc. ISCA Symp. Secur. Privacy Speech Commun. Int. Speech Commun. Assoc.*, 2021, pp. 42–46.

[9] M. S. Hossain and G. Muhammad, “An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework,” *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 62–68, Jun. 2019.

[10] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[11] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, “Privacy-preserving adversarial representation learning in ASR: Reality or illusion?,” in *Proc. Interspeech*, 2019, pp. 3700–3704.

[12] D. Luong, M. Tran, S. Gharib, K. Drossos, and T. Virtanen, “Representation learning for audio privacy preservation using source separation and robust adversarial learning,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[13] E. C. Larson, T. J. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, “Accurate and privacy preserving cough sensing using a low-cost microphone,” in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 375–384.

[14] W. Wang, F. Seraj, N. Meratnia, and P. J. M. Havinga, “Privacy-aware environmental sound classification for indoor human activity recognition,” in *Proc. 12th ACM Int. Conf. PErvasive Technol. Related Assistive Environ.*, 2019, pp. 36–44.

[15] A. Nelus and R. Martin, “Privacy-preserving audio classification using variational information feature extraction,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2864–2877, 2021.

[16] Z. Meng et al., “Speaker-invariant training via adversarial learning,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5969–5973.

[17] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, “Speaker invariant feature extraction for zero-resource languages with adversarial learning,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 2381–2385.

[18] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, “Adversarial disentanglement of speaker representation for attribute-driven privacy preservation,” in *Proc. Interspeech*, 2021, pp. 1902–1906.

[19] X. Jin, C. Lan, W. Zeng, and Z. Chen, “Re-energizing domain discriminator with sample relabeling for adversarial domain adaptation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9174–9183.

[20] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, “FSD50 K: An open dataset of human-labeled sound events,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “PANNs: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[23] D. Acuna, M. T. Law, G. Zhang, and S. Fidler, “Domain adversarial training: A game perspective,” in *Proc. Int. Conf. Learn. Representations*, 2022.

[24] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.