An Opt-in Framework for Privacy Protection in Audio-based Applications

Wei-Cheng Wang
IDLab, Ghent University - imec, Ghent, Belgium

Sander De Coninck
IDLab, Ghent University - imec, Ghent, Belgium

Sam Leroux
IDLab, Ghent University - imec, Ghent, Belgium

Pieter Simoens
IDLab, Ghent University - imec, Ghent, Belgium

***Abstract*—Installing audio-based applications exposes users to the risk of the data processor extracting additional information beyond the task the user permitted. To solve these privacy concerns, we propose to integrate an on-edge data obfuscation between the audio sensor and the recognition algorithm. We introduce a novel privacy loss metric and use adversarial learning to train an obfuscator. Contrary to existing work, our technique does not require users to specify which sensitive attributes they want to protect (opt-out) but instead only provide permission for specific tasks (opt-in). Moreover, we do not require retraining of recognition algorithms, making the obfuscated data compatible with existing methods. We experimentally validate our approach on four voice datasets and show that we can protect several attributes of the speaker, including gender, identity, and emotional state with a minimal recognition accuracy degradation.**

Introduction

■ **AUDIO** data is used in an increasing number of pervasive IoT applications that are deployed in our private space. Microphones in our houses and smartphones have been proposed for acoustic event detection in ambient assisted living [1], speech processing by smart speakers of voice commands [2], or cough detection in telemedicine [3].

This rich palette of applications are all realized as processing algorithms on the same data stream. All applications request direct access to the microphone, but raw audio contains more data than strictly needed to perform the task. This problem of data bundling [4], [5] opens the door for the audio being used for other purposes than the one originally agreed upon. An acoustic system for fall detection of an older person might also reveal if other persons are present. From voice commands targeted to a smart speaker, many sensitive attributes can be derived about the user that go beyond the content of the spoken words, such as speaker identity, emotion, gender or ethnicity. Patent filings indeed show that companies consider these options as valuable sources of information for targeted advertising [6].

Information extraction from audio is typically realized by state-of-the-art deep neural networks (DNNs) with millions of parameters, such as CycleGAN-VC2 [7]. High-end edge devices, such as smartphones, have the necessary substantial computational resources to evaluate machine learning models of sound and speech applications locally. Although no raw audio data is transmitted to a cloud back-end, the risk of data misuse remains because the machine learning models are typically integrated in third-party apps. Operating systems such as Android or iOS require apps to ask permission to use the microphone, but once this permission is granted, there is no way for the user to restrict the type of information that the app can extract from the raw data. In applications like audio-based surveillance in smart cities or nursing homes, up to hundreds of edge devices need to be installed and maintained. In such cases, cloud-based audio processing reduces the installation and maintenance cost but provides data subjects with even less privacy guarantee.

To protect the privacy of the end user, we propose to obfuscate the audio on the edge device. The obfuscation is implemented as a deep neural network (DNN) with a small computational footprint that transforms the original audio in such a way that only selected sensitive attributes are retained, such as gender, identity, etc. We use the principle of adversarial training with a newly designed privacy loss metric to train the obfuscator. The downstream analysis model (running in the cloud or on the edge device) then only has access to the obfuscated audio instead of directly to the raw microphone data. In this paper, we refer to this model as the target task, with a task consisting of the extraction of one or more permitted attributes. The principle is illustrated in Figure 1.

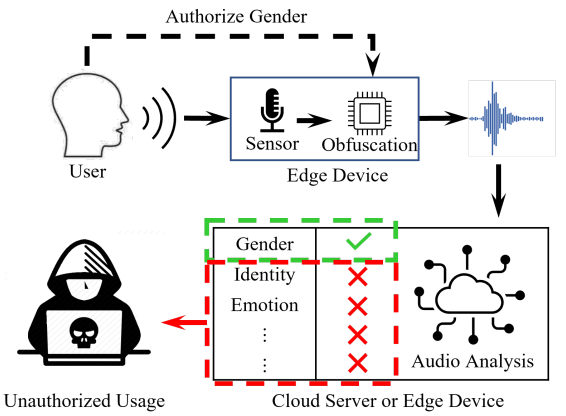


Figure 1. In our proposed system, users authorize one target task. Raw audio is obfuscated on the device, resulting in a transformed audio stream that is transmitted to the target task also on the edge device or in the cloud.

Crucially, the audio is transformed in such a way that it can still be processed by a pre-trained DNN. Our filtering approach could thus be offered as a virtual sensor to existing 3rd party applications, with the filter running in a protected hardware environment.

A second major benefit of our approach, and differentiating it from existing approaches, is that it provides an “opt-in” regime, meaning that the exposed data can only be used for authorized tasks. Alternative works are “opt-out”, requiring users to enumerate the attributes they do not want to provide permission for, which is less protective of privacy.

The main contributions of this paper are three-fold. First, to the best of our knowledge, this is the first work to consider the opt-in regime on audio analysis tasks.

Secondly, we propose a privacy loss function that uses latent space feature representations that capture higher level attributes than the commonly used metrics that work directly with the raw audio. Finally, our solution outputs an obfuscated audio stream that is still compatible with a pre-trained DNN for the target task, making it compatible with 3rd party applications.

The remainder of this paper is structured as follows. After discussing related prior work, we describe our obfuscator framework. This framework is evaluated in an experimental set-up involving four datasets and three attributes.

Finally, we discuss the limitations and scope of our work and provide pointers for future work.

Related Work

As minor clues can already reveal privacy sensitive personal information [8], there is an increasing interest in privacy-enhancing technologies for machine learning applications. Most of the existing works protect one specific attribute, such as gender [9], identity [10], [5], or emotion [11].

In the VoicePrivacy 2020 Challenge [12], the task is to protect speaker identity in automatic speech recognition (ASR) tasks. State-of-the-art in this competition is the Distribution-Preserving X-Vector Generation approach [10]. X-vectors are fixed length embeddings of audio fragments that capture all information on the speaker identity but not on the spoken content [13]. Speech is anonymized by generating synthetic audio, replacing the original x-vector with a fake x-vector sampled from a Gaussian Mixture Model that was fitted on the principal components of the x-vectors of speakers in a large public dataset.

Other works on privacy in audio-based applications focus on acoustic event classification instead of ASR as the target task. Nelus et al. observed that feature extractors designed for event classification often produce representations containing a significant amount of speaker-dependent data [5]. They first train a feature extractor for the target classification task, which they call the trust model. Through a hyperparameter, they control during the training process the balance between classification performance and the mutual information between the original input and the extracted feature vector. Afterwards, they train a threat model that interprets the extracted feature vectors as x-vectors and aims to extract speaker information. They experimentally demonstrate the trade-off between trust and threat model performance. While they use existing architectures for both trust and threat models, the main disadvantage of their approach is that the classifier of the target task has to be retrained with the modified loss function. Our approach, on the other hand, does not require retraining the model of the target task.

Other approaches rely on disentanglement to protect certain speaker attributes. Noé et al. propose an adversarial disentangling autoencoder to conceal the gender attribute from the speaker identification task [9]. Their framework consists of a pre-trained gender classifier, an encoder, a decoder, and a gender classifier. During training, the decoder tries to generate a gender-protected x-vector from the output of the encoder and the pre-trained classifier. During the inference phase, the decoder is fed with a randomly selected gender value to generate a gender protected x-vector. The major disadvantage of this work is that it is an opt-out approach which only protects pre-specified attributes.

Whereas these approaches modify feature representations to remove a predefined sensitive attribute, we generate a transformed audio signal. This allows us to use our model in combination with an off-the-shelf recognition model without retraining. Moreover, x-vector-based approaches are limited to applications with ASR as the target task. Our approach is also applicable to other tasks. In addition, we provide an opt-in framework where all task-irrelevant information is removed.

Opt-in Privacy Protection Framework

Our framework consists of a target model, an obfuscator and a deobfuscator. The relationship between these components is visualized in Figure 2. Note that the deobfuscator is only used during the training phase. The target model T is a function that represents the task the user wants to opt-in for. The input signal X is the log spectrum of the raw audio, a common feature representation used as input in many audio DNN processing applications, while still allowing to decode into raw audio when needed.

Given the original signal X, the target model outputs $T(X)$, for instance, a label indicating the recognized gender, emotion or speaker. We only assume read access to the model T and never change its parameters. In modern applications, T is a pre-trained DNN.

The obfuscator F_σ is a DNN with trainable parameters σ that transforms the original signal X to $\tilde{X} = F_\sigma(X)$. To be compatible with the target model, $\tilde{X}$ has the same dimension as the input signal X. We also introduce a deobfuscator F_μ DNN with trainable parameters μ, that

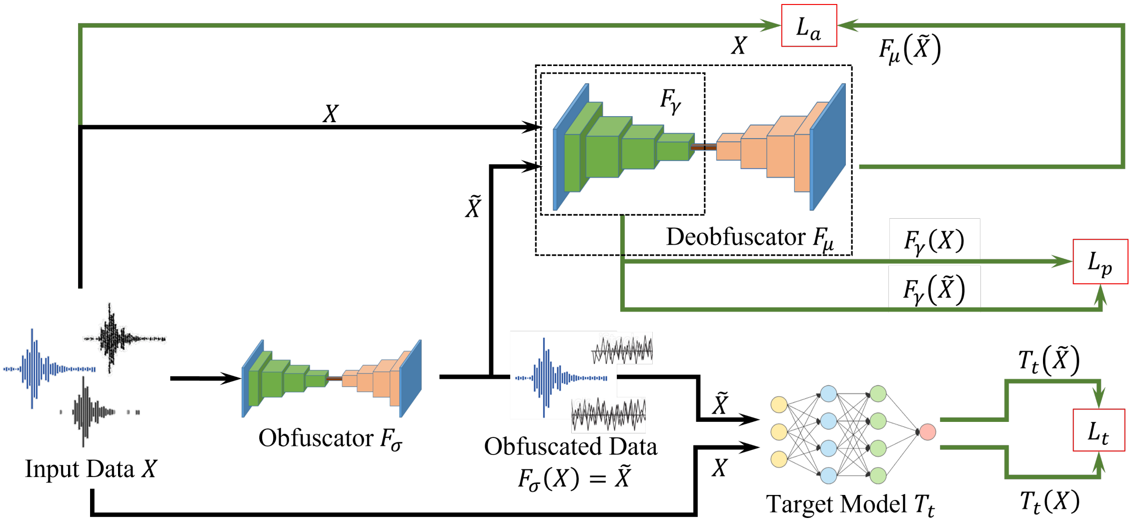


Figure 2. Overview of the proposed framework. During training, the raw data is fed to the obfuscator which generates a privacy-preserving version. The pre-trained target model is agnostic to the data obfuscation and can still perform the intended task. Meanwhile, the deobfuscator tries to reconstruct the original data from the privacy-preserved data. The deobfuscator is discarded for inference. L_a, L_t, and L_p are the adversarial loss, target loss, and privacy loss. The black, and green lines indicate the flow of data, and the calculation of loss, respectively.

tries to reconstruct the original signal from the obfuscated signal. Since we aim for an opt-in approach, we cannot train the deobfuscator on the performance achieved in particular tasks. Instead, the training objective is to minimize the Mean Square error (MSE):

$$L_a = MSE(F_\mu(\tilde{X}), X). \quad (1)$$

The training objective of the obfuscator consists of two (weighted) loss terms L_t and L_p, reflecting the opposing goals to sustain task performance after transforming X, while removing as much information as possible in order to prevent the reconstruction of X from the modified signal. In classification target tasks, as used in this paper, L_t is the cross-entropy loss H between the original and transformed signal:

$$L_t = H(T(\tilde{X}), T(X)). \quad (2)$$

Following the traditional adversarial approach with $L_p = -L_a$ did not provide satisfying results. The main reason is that the log spectrum of the audio X is a too sparse feature encoding. Instead, we use an intermediate distributed representation of dimension M that is the output after processing the first N layers of F_μ. We thus define F_γ as the sub-model of F_μ, with trainable parameters $\gamma \subset \mu$, and define the privacy loss L_p as the distance in each dimension of the latent representation:

$$L_p = \frac{1}{M}\sum_{m=1}^{M} D(F_\gamma(\tilde{X})_m, F_\gamma(X)_m). \quad (3)$$

Since the variance between latent dimensions might vary significantly, we compute the distance in each dimension with the following distance function D:

$$D(\alpha, \beta) = |\frac{1}{1+e^{-(\alpha-\beta)}} - 0.5|. \quad (4)$$

Experimental Setup

Our evaluation focuses on sensitive attributes that can be extracted from speech. Although the principle should extend to non-speech applications as well, the choice to focus on speech was made because of the availability of public datasets with labels for multiple sensitive attributes and existing opt-out algorithms for these attributes to benchmark against.

We compare our method with the recent Adversarial Disentanglement Representation (ADR) framework[9], which is to our knowledge the work that comes closest to our approach. This opt-out framework uses x-vector as audio feature encoding. These x-vectors are converted by an encoder, which is adversarially trained against a classifier for a pre-specified protected attribute.

Datasets

We evaluate our proposed method on four datasets: The Emotional Voices Database (EmoV) [14], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [15], Librispeech [16], and VoxCeleb2 [17]. The EmoV and RAVDESS datasets contain audio fragments of speakers with four labeled attributes: gender, speaker identity, emotion, and utterance. The five emotions considered in EmoV are neutral, amused, angry, sleepy and disgust. The eight emotions considered in RAVDESS are neutral, calm, happy, sad, angry, fearful, surprise and disgust. The other two datasets include more speakers and utterances, but have no labels for emotion.

In real applications, it is very unlikely that the end-user of an application is one of the users whose voice was included in the labelled datasets used to train the obfuscator and target models. To mimic this setting, we split Librispeech and VoxCeleb2 into training and test sets that have different speakers. For Librispeech, we use as training set the merger of the `train100` and `train360` subsets and the `testclean` subset as our testing set. As for the VoxCeleb2 [17] dataset, we followed the experimental set-up of ADR and consider the `V2D` subset as our training set and `V2T` subset as our testing set. For EmoV and RAVDESS datasets, limited by the number of speakers in each dataset, the training and testing datasets contain different fragments but the same speakers.

We pre-process the data by first re-sampling at 16000 Hz and compressing into mono-channel audio. After normalizing the volume, we follow the settings of deepspeech2 [18] to extract the spectrum.

Model Architecture

We use the DeepSpeech2 model [18] as the architecture of the target model T. This model takes raw audio as input and is widely used for speech-to-text recognition but can be easily adjusted to attribute classification tasks. This model exists in different variants, for our experiments we set the number of channels for convolution and BatchRNN to 16 and 64, respectively. To fit our target task of classifying sensitive attributes, we only adapted the dimension of the last fully-connected layer to the number of classes for the task at hand. For each segment of the input audio, the model produces a prediction and the final class is decided by vote counting over all segments. While ADR was evaluated against a different target model, the evaluation results presented in the next section indicate that performance of our target model on original audio fragments is similar to the performance of the ADR target model.

The goal of the obfuscator and deobfuscator is to transform the input data into an output with the same format. Since this task is very similar to voice conversion, we adopted for the obfuscator and deobfuscator the CycleGAN-VC2 [7] architecture, a popular state-of-the-art voice converter that maps the content and style of one speaker onto another. The obfuscator is downsized to fit on a resource-constrained edge device. We adjusted the number of residual blocks (3 instead of 6) and reduced the number of feature channels by a factor of 8. The deobfuscator was not downsized. The number of parameters of the obfuscator and deobfuscator are 0.759M and 6.847M, respectively. We train both obfuscator and deobfuscator with stochastic gradient descent with an initial learning rate of $1e-2$ and a momentum of 0.9. Obfuscator and deobfuscator were trained jointly on a Tesla V100-SXM2 model. Training converged after 8 hours on the smaller EmoV and RAVDESS datasets, and after 5 days on the larger VoxCeleb2 and Librispeech datasets.

Attacker model

We consider two types of attackers: an ignorant attacker and an informed attacker. The ignorant attacker is unaware of the existence of the obfuscation, but is capable of training his own classification model based on the same publicly available datasets that the target task was trained on. The informed attacker on the other hand is aware of the existence of the privacy protector. He has retrieved access to the trained obfuscator and was able to generate obfuscated versions of the fragments in the public datasets. He thus possesses a dataset containing obfuscated data with ground truth labels and can train a model specifically to undo the obfuscation. For a fair comparison, all the attackers share the same architecture of the target model of our method.

Results

In the following sections, we first discuss the target task performance and privacy protection for both systems. We then show how the opt-in system can protect other attributes that were not specified beforehand. Finally, we analyze the computational cost of our approach.

Privacy Protection

In the first experiment, we evaluate our opt-in system and the opt-out ADR in terms of classification accuracy on the target task and on the unauthorised attributes. We focus on the attributes of gender and speaker id, since these labels were present in all datasets.

The results of the first experiment are shown in Figure 3. The classification performance on the target task and on the unauthorised task are shown on the Y-axis and X-axis respectively. The random classification performance on the unauthorised task corresponds with optimal protection and is shown by a vertical dashed line. Good protection on the unauthorised task and good performance on the target task corresponds to the upper left corner of the graphs.

Since we use a different audio encoding than ADR, we first confirmed if both representations contain the same amount of information about the to-be protected attributes by training classifiers on these input representations, resulting in similar classification performance as indicated by the circles in Figure 3. Squares and stars indicate the results obtained by the ignorant attacker and the informed attacker, respectively.

Figure 3 (a) illustrates the results obtained by an attacker on gender while allowing for speaker recognition. Both our framework and ADR are able to protect gender recognition up to the level of random guessing against ignorant attackers. Informed attackers who were able to train specifically against ADR or our obfuscator manage to retrieve more information on the protected gender attribute. Our model provides a slight but consistently better protection on all datasets.

For both frameworks, this protection comes at the cost of a degradation in classification performance of 2-6% on the target task of speaker identity recognition. Arguably, there is mutual information between gender and speaker identity as they are correlated attributes; however a more in-depth analysis, e.g. as performed in [19] is needed to determine whether this correlation fully explains the performance degradation. On the larger datasets, our model is outperformed by ADR on the task performance. This could be caused by two reasons. First, the DeepSpeech2 model takes spectrogram as input, which preserves more information but is also more complex to reconstruct. The second possible reason is the correlations between the two attributes. It is logical that some attributes, e.g. gender and speaker identity, share some information. Thus it is impossible to completely remove the information from one another.

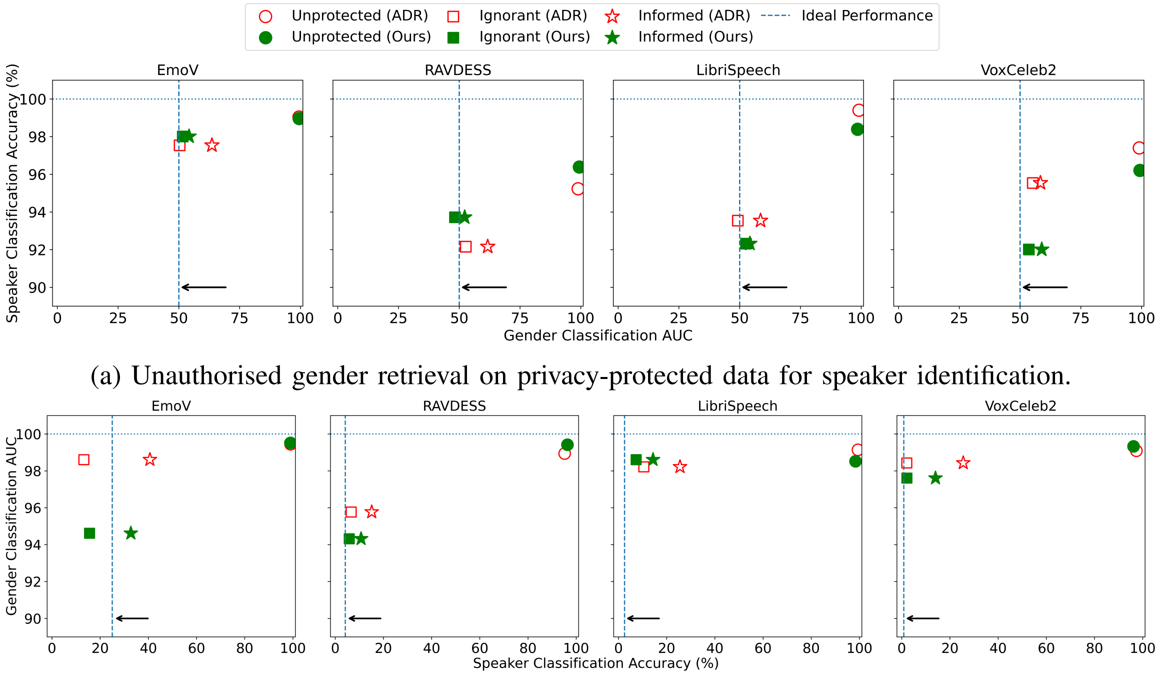
When we switch the target task and the unauthorised tasks, similar conclusions can be drawn, see Figure 3 (b). Ignorant attackers are not able to perform better than randomly guessing on data protected by both frameworks, but our framework provides consistently better protection against informed attackers than ADR. This improved protection comes at the cost of a small drop in target task performance.

Opt-out Versus Opt-in

In the second experiment, we aim to demonstrate the differences between an opt-out and an opt-in system, and the advantages of the latter. As mentioned before, an opt-out system requires explicitly specifying which attribute has to be protected, rather than which attributes are permitted. Thus, to demonstrate the difference between an opt-in system and an opt-out system, we perform an attack on the attributes that are not specified by the opt-out system.

Besides the speaker id and gender, we include emotion as a third sensitive attribute. Since only the EmoV and RAVDESS datasets provide labels for these three attributes, this experiment is only conducted on these two datasets.

Following the description of [9], we train ADR with speaker recognition as target task. Being an opt-out system, ADR also requires us to



(a) Unauthorised gender retrieval on privacy-protected data for speaker identification.

(b) Unauthorised speaker retrieval on privacy-protected data for gender recognition.

Figure 3. Results of sensitive information retrieval on privacy-protected data for different tasks. (a) shows the cases of allowing speaker identification and defending against attacks on gender retrieval. (b) shows the cases of allowing gender recognition and defending against attacks on speaker retrieval. The vertical dash lines indicate the results of random guessing, where the arrow indicate better privacy protection. The horizontal dash lines indicate the performance on the target task, resulting the crossing of the dash lines as the best performance. Note that the axes of target task performances are scaled to 90 to 100 for better visualization.

explicitly specify which attribute to be protected. We choose gender as protected attribute, as in the original paper. Emotion is thus the unspecified attribute that a user might inadvertently expose.

The results of this experiment are illustrated in Figure 4. The axes show the classification performance on the protected gender attribute, and the unspecified emotion attribute. Dashed lines indicate the performance of random guessing, the best possible protection level one can achieve.

Both frameworks provide protection against an uninformed attacker, as the classification on obfuscated data approaches those of a random classifier. Our framework is however much more robust against informed attacks, protecting both attributes while ADR only achieves reasonable protection against the pre-specified gender attribute. An informed attacker manages to retrieve emotion information with more than 70 % accuracy on the ADR-protected data.

This experiment shows that the opt-in regime provides better protection when the attribute of interest is not previously known. Although specifying specify multiple protected attributes could

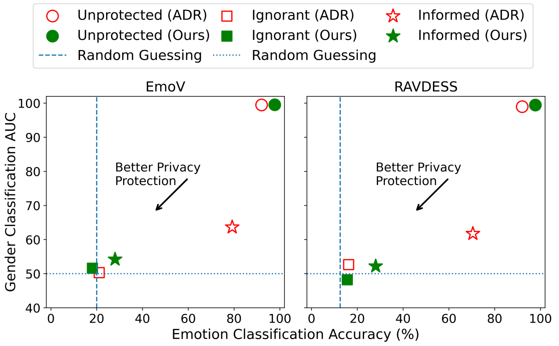


Figure 4. Results of sensitive information retrieval (gender and emotion) on privacy protected data for speaker identification. The horizontal and vertical dash lines indicate the results of sensitive information retrieval on hypothetically perfectly protected data, with their crossing indicating the best protection.

remedy this particular case, the fundamental challenges of opt-out systems remain.It is impossible to define all the possible attributes that may be interested by all parties. Thus, opt-out systems like ADR would still leak privacy information even with multiple predefined attributes.

Summarizing the results of both experiments, we conclude that our framework provides good protection against both ignorant and informed attackers. The major advantage of our framework is the opt-in aspect, which aims to only retain information in the obfuscated signal relevant for the authorized task. However, this improved protection comes at a limited cost in classification performance on the permitted attribute.

Computational Cost

We further measure the computational time of the framework (without deobfuscator) on edge devices to simulate how the proposed framework works in a real-world scenario. In Table 1, we show the average execution time of obfuscating one second of audio on different platforms. Only on devices with embedded GPU, the model can work in real-time.

Conclusion and Future Work

In this paper, we introduced a novel opt-in framework to preserve privacy while using audio applications. We use adversarial training and a novel privacy-preserving loss metric to train an obfuscator that removes all but the information needed for the authorised task. Unlike existing approaches, we do not require an adaptation of the target task classification models. This allows the obfuscator to be integrated in a pipeline with existing third-party audio services.

We validated our approach on four voice datasets and compared it against one state-of-the-art approach for privacy protection. We evaluated protection against two types of attacks and show that our method can protect privacy with only a small reduction in classification accuracy on the permitted task. We further showed the strength of the opt-in framework against unspecified attacks compared to the opt-out framework.

The proposed opt-in framework still has a few limitations that mandate future research before being applied in real-world scenarios. Firstly, we have evaluated our obfuscator architecture with the classification of one attribute as target task. How the current model performs on other task types such as speech-to-text recognition is yet to be investigated. In its current inception, having multiple permitted target tasks would require multiple obfuscators. Creating one obfuscator model with configurable target tasks would first require an in-depth study of the correlation between attributes. Secondly, although we do not modify the target task model, we require white-box access to back-propagate weight updates to the obfuscator. This makes our approach only compatible with third-party services with known architecture and parameter values. To overcome this limitation, one possible solution is to leverage transfer learning strategies. Finally, deployment of the obfuscator on low-end devices would require network compression techniques such as pruning and quantization [20].

Audio-based applications are very attractive, but pose significant privacy risks to the user. We hope that this paper will inspire other researchers to contribute to better protection mechanisms.

Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

■ REFERENCES

1. J. Navarro, E. Vidaña-Vila, R. M. Alsina-Pagès *et al.*, “Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios,” *Sensors*, vol. 18, no. 8, p. 2492, 2018.
2. F. Bentley, C. Luvogt, M. Silverman *et al.*, “Understanding the long-term use of smart speaker assistants,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–24, 2018.
3. I. D. Miranda, A. H. Diacon, and T. R. Niesler, “A comparative study of features for acoustic cough detection using deep architectures,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2019, pp. 2601–2605.
4. A. Trask, E. Bluemke, B. Garfinkel *et al.*, “Beyond privacy trade-offs with structured transparency,” *arXiv preprint arXiv:2012.08347*, 2020.

Table 1. Computational time measured on different devices

Device	CPU	GPU	Time (ms)
Raspberry pi 2B	ARM Quad-Core Cortex-A7	Not used	1923
NVIDIA Jetson TX1	ARM Quad-Core Cortex-A57	256-core NVIDIA Maxwell™	34
Server	Intel Xeon Silver 4108	NVIDIA GTX 1080 Ti	1

5. A. Nelus and R. Martin, "Privacy-preserving audio classification using variational information feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2864–2877, 2021.
6. H. Jin and S. Wang, "Voice-based determination of physical and emotional characteristics of users," Oct. 9 2018, uS Patent 10,096,319.
7. T. Kaneko, H. Kameoka, K. Tanaka *et al.*, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6820–6824.
8. A. Korolova, "Privacy violations using microtargeted ads: A case study," in *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 474–482.
9. P.-G. Noé, M. Mohammadamini, D. Matrouf *et al.*, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," *arXiv preprint arXiv:2012.04454*, 2020.
10. H. Turner, G. Lovisotto, and I. Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020," *arXiv preprint arXiv:2010.13457*, 2020.
11. R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: Privacy-preserving speech analysis for voice assistants," *arXiv preprint arXiv:1908.03632*, 2019.
12. N. Tomashenko, X. Wang, E. Vincent *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.
13. D. Snyder, D. Garcia-Romero, G. Sell *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2018, pp. 5329–5333.
14. A. Adigwe, N. Tits, K. E. Haddad *et al.*, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.
15. S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
16. V. Panayotov, G. Chen, D. Povey *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2015, pp. 5206–5210.
17. J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.
18. D. Amodei, S. Ananthanarayanan, R. Anubhai *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.
19. S. De Coninck, W.-C. Wang, S. Leroux *et al.*, "Selective manipulation of disentangled representations for privacy-aware facial image processing," *arXiv*, 2022.
20. T. Liang, J. Glossner, L. Wang *et al.*, "Pruning and quantization for deep neural network acceleration: A survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

Wei-Cheng Wang, is currently pursuing his PhD in Computer Science Engineering from Ghent University, Belgium since September 2019. His main research interests are machine learning, neural networks and deep learning with a focus on audio and video processing. Contact him at WeiCheng.Wang@ugent.be.

Sander De Coninck, is currently pursuing his PhD in Information Engineering Technology from Ghent University, Belgium since September 2021. His main research interests are machine learning, neural networks, privacy-preserving machine learning and computer vision. Contact him at Sander.DeConinck@ugent.be.

Sam Leroux, is currently active as a Postdoctoral researcher at Ghent University. His main research interests are machine learning, neural networks and deep learning with a focus on techniques to make neural networks more efficient. Contact him at Sam.Leroux@ugent.be.

Pieter Simoens, currently holds a position as Associate Professor at Ghent University, and is also

affiliated with imec. His main research interest is in the domain of intelligent distributed systems, with a specific focus on resource-efficiency, unsupervised learning and collective intelligence. Contact him at Pieter.Simoens@ugent.be.