Towards Privacy-Preserving Audio Classification Systems

Bhawana Chhaglani
University of Massachusetts Amherst
Amherst, USA
bchhaglani@umass.edu

Jeremy Gummeson
University of Massachusetts Amherst
Amherst, USA
jgummeso@umass.edu

Prashant Shenoy
University of Massachusetts Amherst
Amherst, USA
shenoy@cs.umass.edu

Abstract

Audio signals can reveal intimate details about a person's life, including their conversations, health status, emotions, location, and personal preferences. Unauthorized access or misuse of this information can have profound personal and social implications. In an era increasingly populated by devices capable of audio recording, safeguarding user privacy is a critical obligation. This work studies the ethical and privacy concerns in current audio classification systems. We discuss the challenges and research directions in designing privacy-preserving audio sensing systems. We propose privacy-preserving audio features that can be used to classify wide range of audio classes, while being privacy preserving.

CCS Concepts: • **Human-centered computing** → **Personal digital assistants**; • **Security and privacy**;

Keywords: Audio classification, Privacy-preserving sensing, Acoustic features

ACM Reference Format:
Bhawana Chhaglani, Jeremy Gummeson, and Prashant Shenoy. 2024. Towards Privacy-Preserving Audio Classification Systems. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (HotEthics'24).* , 5 pages. https://doi.org/XXXXXXX.XXXXXXX

HotEthics'24, April 28, 2024, San Diego, California, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

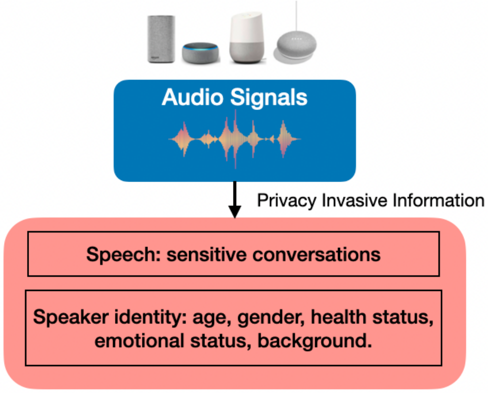
Audio Signals
Privacy Invasive Information
Speech: sensitive conversations
Speaker identity: age, gender, health status, emotional status, background.

Figure 1. Privacy Concerns in Audio Classification Systems

1 Introduction

Audio sensing has tremendous potential and can be used to infer a rich array of personal information about an individual [1, 2]. Kroger et al. [3] presents an overview of sensitive pieces of information that can, with the help of advanced data analysis methods, be derived from audio, including cues to a speaker's biometric identity, personality, physical traits, geographical origin, emotions, level of intoxication and sleepiness, age, gender, and health condition. These are all examples of privacy-sensitive information and its access should be regulated. Today, we are surrounded by listening devices including smart speakers, voice assistants, baby monitors, smoke alarms, and other emerging device categories. Some of them are always-on and listening to sensitive content, while others can be accidentally triggered [4]. Thus, privacy remains one of the core problems in audio sensing which needs to be addressed.

Existing audio classification systems use features like spectrograms that can potentially compromise user privacy. To avoid privacy concerns, prior work has primarily focused on eliminating or obfuscating speech to ensure privacy [5, 6]. However, speech is not the only privacy sensitive information present in the audio signal — it can also reveal information about the speaker's identity and whereabouts. The idea is to explore a more holistic definition of privacy and propose a mechanism to evaluate and quantify the privacy of an audio-based system. Through this research, we aim to explore the potential of audio sensing in enabling novel applications, while simultaneoulsy ensuring that user privacy is retained.

In this work, we discuss the key challenges in realizing privacy-aware audio classification systems. Using the example of environmental sound classification, we demonstrate the application of privacy-preserving audio features to safeguard privacy. We show that this approach is generalizable and can achieve comparable accuracy as the state-of-the-art techniques that do not consider privacy.

2 Need for Ensuring Privacy in Audio Classification

Most audio classification systems convert audio signals into Mel-Frequency Cepstral Coefficients (MFCCs), Mel spectrograms, and Short-Time Fourier Transform (STFT) and then feed them into advanced deep learning models as shown in Figure 2. Piczak et al. [7] explores the use of Mel spectrograms as input to convolutional neural networks (CNNs) for the classification of environmental sounds. Forsad et al. [8] compares MFCC and STFT for cough with CNNs for detection task. By effectively bridging the gap between raw audio signals and the sophisticated pattern recognition capabilities of deep learning (DL) architectures, these features enable the models to achieve remarkable accuracy and efficiency in variety of tasks. However, the same features that enhance the performance of audio classification models can also pose significant privacy risks. MFCCs, STFT, Mel spectrograms or other spectrograms, by their very nature, encapsulate detailed information about the audio source, potentially including sensitive personal information. For instance, MFCCs can retain distinctive speech characteristics that could be exploited to identify a speaker, thereby leaking personal identity information. Similarly, STFT and Mel spectrograms can inadvertently reveal background noises or conversations that were not meant to be shared, thereby compromising the privacy of individuals. The practice of transmitting spectrograms to cloud-based services for model prediction further heightens privacy concerns.

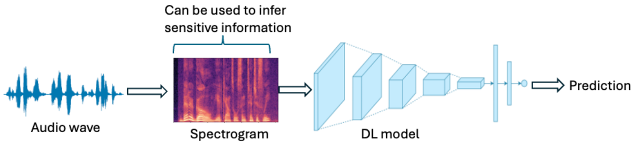

Figure 2. Typical Audio Classification Pipeline

This duality underscores the need for a careful consideration of privacy implications when designing and deploying audio classification systems.

3 Privacy-Aware Audio Classification: Challenges

In this section, we discuss the major hurdles in designing privacy-preserving audio-based systems.

- **Expanding Privacy Beyond Speech**: Prior work mainly considers speech as the sensitive content and tries to filter out [5], obfuscate [9], or replace [6] the speech segments. Audio contains more privacy invasive information than just speech. There is a need to look at a more holistic definition of privacy.
- **Complexity of Privacy Evaluation**: Evaluating privacy in audio sensing systems is a non-trivial problem as privacy varies with context. Most common techniques are to use automatic speech recognition using Google APIs [10] or perform human evaluation [11]. Boovaraghavan et al. [12] use word error rate (WER) and phoneme error rate (PER) as the privacy evaluation metrics. However, these techniques do not evaluate whether the systems are leaking speaker related information or their location. For example, the system could infer myriads of details about an individual even with high WER. Thus, there is a need for accurate privacy evaluation mechanism.
- **Towards Universal Privacy Preservation**: Some of the prior work has designed privacy preserving pipelines for different applications. Bourlard et al. [13] uses privacy sensitive features for speech/non-speech detection, while Parthasarathi et al. [14] uses privacy sensitive audio features to detect speaker change in conversations. Although, these techniques are useful, they are application specific and require manual effort to design hand-crafted features. There is a need to address the challenge of privacy from a more general perspective by

designing a set of generalizable privacy-aware features.

- **Privacy-Accuracy Trade-off**: The accuracy-privacy tradeoff in audio classification systems represents a critical challenge [11]. On one hand, achieving high accuracy in audio classification often necessitates the use of detailed and comprehensive audio features, such as MFCCs. These features capture nuanced aspects of the audio signal, allowing deep learning models to make precise inferences about the content, context, or identity associated with the audio data. On the other hand, the very richness of information that enables such accuracy can also lead to significant privacy concerns, as these features may contain or enable the reconstruction of sensitive information about individuals or their environment.

To address these challenges, we propose to redefine privacy in the context of audio classification to encompass more than just speech. We aim to use *privacy-preserving audio features that are universally applicable and maintain effectiveness of audio classification systems.*

4 Preliminary Experiments with ESC-50

4.1 Dataset Description

The ESC-50 [15] dataset is a comprehensive and well-curated collection designed for the task of environmental sound classification, highlighting the diversity and complexity of auditory scenes that can be encountered in everyday life. ESC-50 is a labeled collection of 2,000 environmental audio recordings (each 5 seconds long) with 50 classes. These categories encompass a wide range of sounds from natural environments (such as rain, thunderstorms, and animal sounds), human-made noises (like car horns, alarms, and mechanical tools), and human sounds (including laughing, clapping, and crying). The dataset is fully annotated, with labels indicating the category of each audio clip.

4.2 Generalizable Privacy-Preserving Features

We process the audio files using a sliding window of 500 ms with 100 ms overlap. We remove silent periods from the audio files using the top decibel threshold of 20. Next, each audio segment is converted into a list of time and frequency domain audio features. These features are as follows:

- Zero Crossing Rate (ZCR) indicates the number of times the signal changes sign. This can be related to the texture of the sound, which is very relevant in differentiating between diverse environmental sounds.
- Harmonic-to-Noise Ratio (HNR) measures the amount of harmonic content compared to noise within a signal. Since environmental sounds can be either more harmonic (e.g., bird singing) or more noisy (e.g., rain), this feature can help distinguish between them.
- Spectral Contrast refers to the difference in amplitude between peaks and valleys in the sound spectrum.
- Peak, RMS, Energy relate to the amplitude and power of the audio signal and are useful for distinguishing between loud and soft sounds, as well as the intensity of the sound source.
- Spectral Roll-off indicates the frequency below which a certain percentage of the total spectral energy is contained. This feature helps to separate sounds with high-frequency content from those with low-frequency content.
- Spectral Flatness and Bandwidth (BW) provide a measure of how noise-like a sound is, versus being tonal.
- Spectral Centroid is a measure of the 'brightness' of a sound, which can be useful to characterize sounds with high-frequency content like bird songs or bells.

We do not include features that contain privacy invasive information. For instance, fundamental frequency can reveal details about speaker or formants can reveal speech content as mentioned in source filter model of speech production [16].

4.3 Classification Results

We use these features to train a random forest classifier model. To evaluate the model, we split the dataset into train, test and validation set (70%, 15%, 15%) after shuffling. We achieve aggregate accuracy of 92.23% over all the classes. We observe the feature importances using gini importance aggregated over all the classes. We observe that spectral contrast, ZCR and HNR are the most important features for this task. This is because ZCR is particularly useful in distinguishing between percussive sounds (like footsteps or clapping) and more harmonic sounds (like animal calls or musical instruments). HNR is crucial for differentiating between sounds that

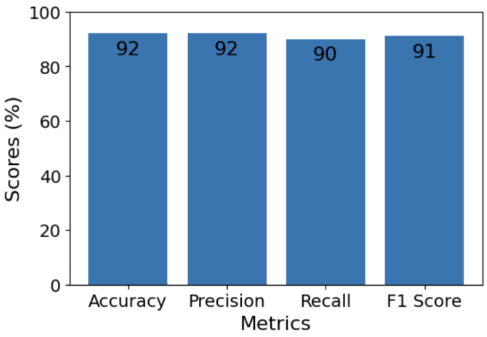

Figure 3. Classification metrics using privacy-preserving features

are more tonal and structured, such as human voices or music, versus those that are more stochastic and noise-like, such as wind or rushing water.

4.4 Comparison with Non Privacy-based approaches

Most of the prior work uses MFCCs or the Mel spectrogram for classifying ESC-50 classes. These spectrograms can be use to infer human speech [17] and speaker identity [18]. The accuracy results obtained from the ESC-50 dataset using spectrogram as feature input has a wide range of variation depending on the spectrogram type and the DL model. Mu et al. [19] use Log-Mel spectrogram and temporal-frequency attention based convolutional neural network model (TFCNN), and achieve accuracy of 84.4%. Wang et al. [20] propose parallel temporal-spectral attention mechanism for CNN to learn discriminative time-frequency representations of the spectrogram. This model achieve accuracy of 88.6%. Certain models [21] achieve very high accuracy of >90% using pre-training and data augmentation techniques. Thus, we can achieve comparable or better accuracy by taking the privacy-aware route. The proposed features are generalizable as they can be used for 50 different environment sounds

5 Conclusions

This paper explores the privacy concerns in audio classification systems that use features like MFCCs and spectrograms, which can inadvertently reveal sensitive personal information beyond just speech content, such as speaker identity, location, and other metadata. We identify key open challenges including developing robust privacy evaluation mechanisms and navigating the accuracy vs. privacy tradeoff. To address these issues, we propose a set of generalizable privacy-preserving audio features and demonstrate their ability to achieve comparable accuracy to prior work on the ESC-50 environmental sound dataset, while mitigating privacy risks. Overall, this work represents an important step towards enabling privacy-preserving audio sensing systems that can unlock the potential of this rich data modality while effectively safeguarding user privacy.

Acknowledgments

This research was supported in part by NSF grants 2211302, 2211888, 2213636, 2105494, and US Army contract W911NF-17-2-0196. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Hanbin Zhang, Chen Song, Aosen Wang, Chenhan Xu, Dongmei Li, and Wenyao Xu. Pdvocal: Towards privacy-preserving parkinson's disease detection using non-speech body sounds. In *The 25th annual international conference on mobile computing and networking*, pages 1–16, 2019.

[2] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21:93–120, 2018.

[3] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. Privacy implications of voice and speech analysis–information disclosure by inference. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pages 242–258, 2020.

[4] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Exploring accidental triggers of smart speakers. *Computer Speech & Language*, 73:101328, 2022.

[5] Stephen Xia and Xiaofan Jiang. Pams: Improving privacy in audio-based mobile systems. In *Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*, pages 41–47, 2020.

[6] Francine Chen, John Adcock, and Shruti Krishnagiri. Audio privacy: reducing speech intelligibility while preserving environmental sounds. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 733–736, 2008.

[7] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2015.

[8] Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–28,

2020.

[9] Daniyal Liaqat, Ebrahim Nemati, Mahbubur Rahman, and Jilong Kuang. A method for preserving privacy during audio recordings by filtering speech. In *2017 IEEE Life Sciences Conference (LSC)*, pages 79–82. IEEE, 2017.

[10] Wordless Sounds. Wordless sounds: Robust speaker diarization using privacy-preserving audio representations, sree hari krishnan parthasarathi, hervé bourlard and daniel gatica-perez, idiap-rr-28-2012.

[11] Bhawana Chhaglani, Camellia Zakaria, Adam Lechowicz, Jeremy Gummeson, and Prashant Shenoy. Flowsense: Monitoring airflow in building ventilation systems using audio sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–26, 2022.

[12] Sudershan Boovaraghavan, Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. Kirigami: Lightweight speech filtering for privacy-preserving activity recognition using audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–28, 2024.

[13] Herve Bourlard, Mathew Magimari, Daniel Gatica Perez, and Hari Krishna Parthasarathi. Privacy sensitive audio features for speech/non-speech detection. *IEEE transaction on audio, speech and language processing*.

[14] Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Daniel Gatica-Perez, and Hervé Bourlard. Speaker change detection with privacy-preserving audio cues. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 343–346, 2009.

[15] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[16] Gunnar Fant. The source filter concept in voice production. *STL-QPSR*, 1(1981):21–37, 1981.

[17] Chadawan Ittichaichareon, Siwat Suksri, and Thaweesak Yingthawornsuk. Speech recognition using mfcc. In *International conference on computer graphics, simulation and modeling*, volume 9, 2012.

[18] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, and Wang Lihao. Speaker recognition based on dynamic mfcc parameters. In *2009 International Conference on Image Analysis and Signal Processing*, pages 406–409. IEEE, 2009.

[19] Wenjie Mu, Bo Yin, Xianqing Huang, Jiali Xu, and Zehua Du. Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, 11(1):21552, 2021.

[20] Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang. Environmental sound classification with parallel temporal-spectral attention. *arXiv preprint arXiv:1912.06808*, 2019.

[21] Avi Gazneli, Gadi Zimerman, Tal Ridnik, Gilad Sharir, and Asaf Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv preprint arXiv:2204.11479*, 2022.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009