AUDIO AUGMENTATION MEETS AUDIO UNLEARNING: HELP OR HINDER?

Moulik Gupta[1]*, Achyut Mani Tripathi*[2]

[2]Department of Computer Science & Engineering, IIT Dharwad, India,
[1]Independent Researcher, New Delhi, India.
gupta.moulik@gmail.com [1], t.achyut@iitdh.ac.in [2]

ABSTRACT

Machine unlearning targets to eliminate the influence of specific data from a well-trained deep model while maintaining its performance on the remaining data. In the audio domain, this task is particularly challenging due to temporal dependencies and the sequential complexity of audio signals. The presence of memorized information further complicates unlearning, and although data augmentation can affect memorization, its role in audio unlearning has not been thoroughly explored. In this work, we investigate the impact of various audio augmentation techniques on unlearning audio models using state-of-the-art methods, including Amnesiac, Bad Teacher, Boundary, Fisher, and Scrub. Experiments conducted on three benchmark datasets, Viz. Speech Commands V1, ESC-10, and US8K demonstrate that appropriate audio augmentations significantly enhance audio unlearning performance and reduce the performance gap between retrained and unlearned audio models. Our results show that combining time masking and additive noise consistently outperforms other augmentation strategies, achieving the lowest average performance gap. These findings highlight that audio augmentation not only reduces memorization but also plays a vital role in enabling privacy-preserving and effective machine unlearning. This highlights its significance in creating robust and reliable methods for audio unlearning.

Index Terms— Audio Classification, Audio Unlearning, Data Augmentation

1. INTRODUCTION

Large datasets are essential for training machine learning models, but they often contain sensitive, copyrighted, or private information. In some cases, such data may even be collected without user consent. Regulations such as the General Data Protection Regulation (GDPR) [1] mandate the removal of this information, yet retraining models from scratch is often costly and impractical. Machine unlearning [2] addresses this challenge by enabling the removal of specific data from a trained model without full retraining, ensuring that the model behaves as if the deleted data were never part of its training set. Machine unlearning has been widely explored in the domain of image [3] classification. In general, unlearning techniques can be categorized as exact or approximate [4] methods. Exact methods aim to completely remove targeted data by selectively retraining parts of the model, though they remain computationally expensive. Approximate methods, on the other hand, reduce the influence of the data without guaranteeing full elimination, making them more efficient and better suited for frequent or large-scale unlearning scenarios. Several strategies have been proposed to weaken the influence of forgotten data. For example, Golatkar et al. [5] introduced a perturbation-based function that adjusts model weights to erase information about the target data. Influence-based methods [6] rely on one-shot updates to enforce unlearning of specific samples. Gradient ascent has also been extensively studied, particularly in the context of large language models [7]. Other notable approaches include logits-based unlearning [8], and mitigation of highly memorized samples [9].

Although numerous approaches have been proposed to evaluate machine unlearning in image classification, the domains of audio and speech processing have received far less attention. Only a limited number of studies address audio unlearning. For instance, [10] introduced the first benchmark study, evaluating eight machine unlearning methods by unlearning Wave2Vec and HuBERT models for spoken language understanding. Likewise, Cheng et al. [11] investigated several unlearning methods for speech tasks, revealing why unlearning in speech is inherently more difficult compared to image or text data. Label smoothing [12] and network pruning [13] are widely adopted machine unlearning techniques designed to reduce model memorization and complexity. Zhao et al. [14] showed that the presence of memorized information substantially increases the difficulty of unlearning. Although methods mentioned above have opened promising directions in the literature, the role of audio data augmentation in audio unlearning has remained largely unexplored. Despite the critical role of

audio data augmentation in controlling memorization and mitigating bias in audio machine learning [15], its systematic impact on audio unlearning remains largely unexplored, representing a significant research gap.

To bridge this gap, in this work, we present the first comprehensive empirical evaluation of the impact of different audio data augmentation strategies on unlearning performance. Our study spans multiple state-of-the-art class-wise unlearning methods and three diverse audio datasets, providing clear evidence that carefully chosen audio augmentation techniques can significantly enhance the effectiveness of audio unlearning. Our key contributions are as follows: we provide a comprehensive evaluation of five state-of-the-art machine unlearning methods in the audio domain across three benchmark datasets, complemented by a systematic study of audio data augmentation. Our results demonstrate that appropriate augmentations substantially reduce the performance gap between unlearned audio models and those retrained from scratch, while also enhancing unlearning efficiency. Additionally, through a detailed class-wise forgetting analysis, we reveal the sensitivity of unlearning methods to augmentation strategies, highlighting their pivotal role in building robust and reliable audio unlearning frameworks.

2. EXPERIMENTS

2.1. Datasets,Implementation & Evaluation Metrics

We conducted experiments on three audio datasets (Table 1), splitting the audio dataset D into retained D_r and forget D_f sets, where D_f contains samples from classes to be removed from the model. The experiments

Table 1: Dataset Specifications

Dataset	Sampling Rate	Classes	#Samples
GSCD V1 [16]	16 kHz	10	23,682
ESC-10 [17]	44.1 kHz	10	400
US8K [18]	16 kHz	10	8,732

were implemented in Python 3.10.16 on Ubuntu 24.04.1 LTS with CUDA 12.1, running on a single NVIDIA RTX 4060 GPU. Model training employed the AdamW optimizer with a learning rate of 2e-5. A batch size of 16 was used, along with early stopping (patience=7). Cross-entropy loss is employed while training the models. The value of the forgetting rate is set to 10% while concluding all the experiments. Following [10], we evaluate the model using unlearning accuracy (UNA, lower better), remaining accuracy (RMA, higher better), testing accuracy (TEA, higher better), and membership inference attack (MIA, lower better). Additionally, we introduce two metrics: average gap (AGP, lower better), which quantifies the mean difference across these four metrics between retrained and unlearned models, and runtime efficiency (RTE, lower better), which measures unlearning time (or retraining time for the retained dataset). We begin by fine-tuning an ImageNet-pretrained ResNet-18 model on each of the three datasets. Subsequently, a target class (in our study, it is selected as class 0 for all the datasets) is randomly selected, and experiments are conducted to unlearn all audio samples belonging to that class. The number of Mel filters, FFT size, and window overlap are set to 128, 1024, and 512, respectively, during spectrogram extraction.

2.2. Data Augmentations & Unlearning Methods

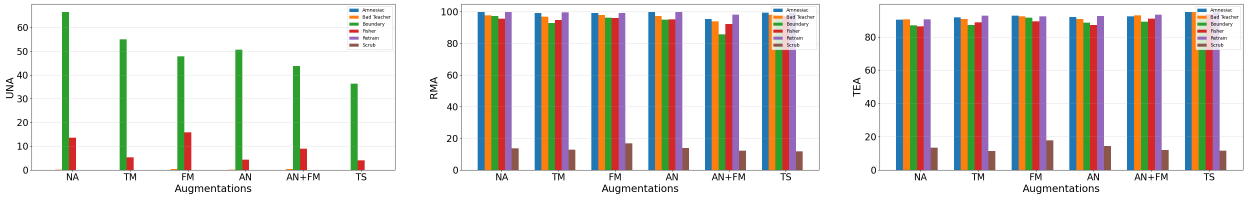
To investigate the effect of data augmentation on unlearning performance, we applied various audio data augmentation techniques [19] individually as well as in combination. Specifically, we employed five augmentation strategies: Time Masking (TM), Frequency Masking (FM), Additive Noise (AN), Time Stretch (TS), and a combined method of AN and FM. The class-wise unlearning performance of the audio model is evaluated using five state-of-the-art approaches, Viz. Amnesiac [20], Bad Teacher [21], Boundary [22], Fisher [23], and Scrub [24]. For fairness, the performance of these unlearned models is compared against the audio model retrained on the remaining data samples after excluding the target class.

3. RESULTS

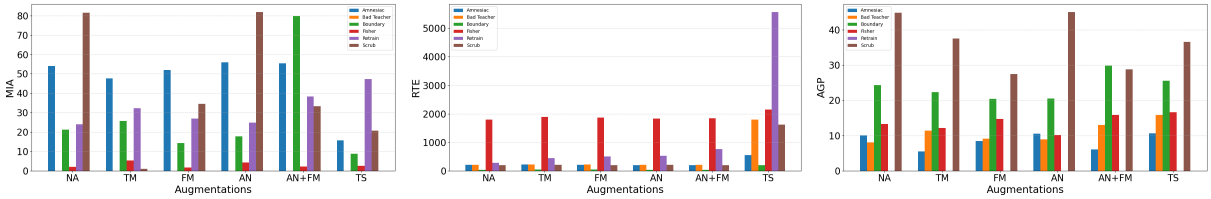
Owing to space constraints, the results are presented as bar graphs. Fig.1, 2, and 3 show the performance of the unlearned audio model under different unlearning methods and data augmentations for the GSCD, ESC-10, and US8K datasets, respectively. A detailed tabular version of the results can be accessed through the link [1]

Speech Commands: For the unlearning methods Amnesiac, Bad Teacher, Boundary, Fisher, and Scrub, the lowest average performance gap is obtained with the TM, NA, FM, AN, and FM augmentations, respectively. Conversely, the highest gaps for these methods occur with TS, TS, AN&FM, TS, AN, and FM, respectively. Incorporation of audio data augmentation generally lowers unlearning accuracy, thereby supporting more efficient unlearning. For instance, the best-performing augmentations are TS+TM for Bad Teacher, TS for Boundary, and TS for Fisher. In contrast, for Scrub and Amnesiac, all augmentations perform comparably. Notably, the worst unlearning accuracies for Bad Teacher and Boundary are observed with TS+TM and TS, respectively. Furthermore, applying specific

[1] https://github.com/SteadySurfdom/Machine-Unlearning-using-Data-Augmentation

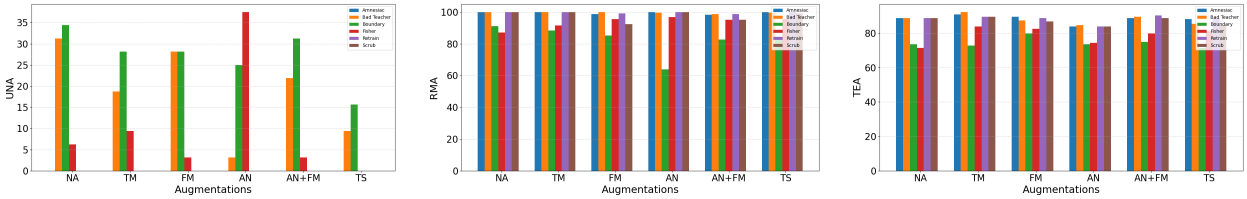


(a) UNA (b) RMA (c) TEA

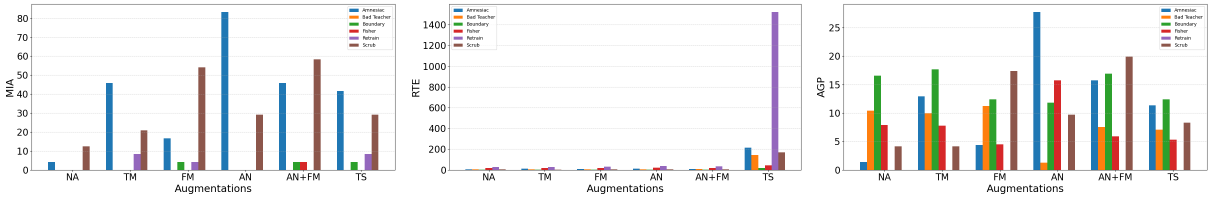


(d) MIA (e) RTE (f) AGP

Fig. 1: Impact of Audio Data Augmentations While Unlearning for Speech Command Dataset

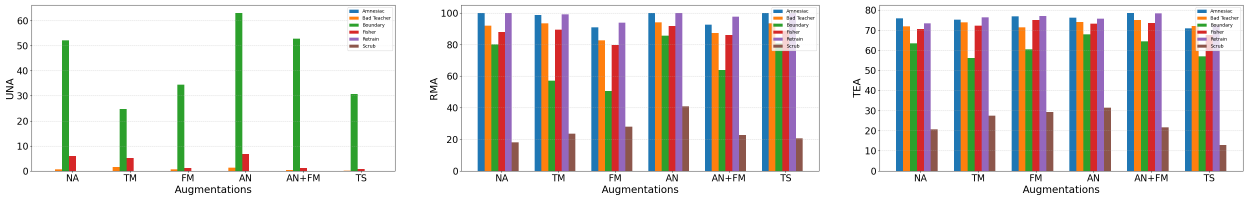


(a) UNA (b) RMA (c) TEA

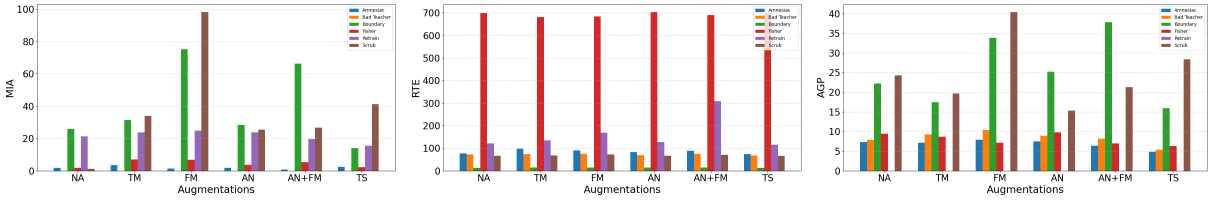


(d) MIA (e) RTE (f) AGP

Fig. 2: Impact of Audio Data Augmentations While Unlearning for ESC-10 Dataset



(a) UNA (b) RMA (c) TEA



(d) MIA (e) RTE (f) AGP

Fig. 3: Impact of Audio Data Augmentations While Unlearning for US8K Dataset

augmentations improves retain accuracy compared to the retrain baseline obtained without augmentation. Since machine unlearning methods are vulnerable to membership inference attacks (MIA), augmentation also influences robustness. For some methods (e.g., Bad Teacher, Boundary, and Fisher), augmentation lowers MIA attack accuracy, whereas for others (e.g., Scrub and Amnesiac), it increases vulnerability.

ESC-10: For the unlearning methods Amnesiac, Bad Teacher, Boundary, Fisher and Scrub, the lowest average performance gap is achieved with the No Aug, AN, AN, FM, and TM audio data augmentations, respectively. Conversely, the highest gaps for these methods are observed with AN, FM, TM, AN, and AM+FM, respectively. Overall, incorporating audio data augmentation reduces unlearning accuracy, thereby supporting more effective unlearning. For instance, the best-performing augmentations for Bad Teacher, Boundary, and Fisher are AN, TS, and TS, while for the remaining methods, all augmentations perform comparably. Notably, for Bad Teacher and Boundary, the worst unlearning accuracy occurs when no augmentation is applied during training. Moreover, the use of specific augmentations improves retain accuracy compared to the no-augmentation retrain baseline. Since machine unlearning methods are vulnerable to membership inference attacks (MIA), the role of augmentation is critical. For some methods (e.g., Bad Teacher, Boundary, and Fisher), augmentation helps lower MIA attack accuracy, while for others (e.g., Scrub and Amnesiac), it increases vulnerability.

US8K: For the Amnesiac, Bad Teacher, Boundary, Fisher, and Scrub, unlearning methods, the lowest average gap is achieved with TS, TS, TS, TS, and AN augmentations, respectively. Conversely, the highest gap for these methods is observed with FM, FM, AN&FM, AN and TS. Across all methods, the incorporation of audio data augmentation consistently reduces unlearning accuracy, indicating its effectiveness in supporting efficient unlearning. For instance, the best performing augmentations are TS for Bad Teacher, TM for Boundary, and TS for Fisher, while Scrub and Amnesiac show comparable performance across all augmentations. The worst unlearning accuracy is observed with TS for Bad Teacher and Fisher, and TM for Boundary. Similar to the ESC-10 and GSCD results, specific augmentations also improve retain accuracy compared to the no-augmentation retrain baseline. Furthermore, machine unlearning methods remain vulnerable to MIA attacks. While augmentations reduce attack accuracy for methods like Bad Teacher, Boundary, and Fisher, they increase it for Scrub and Amnesiac regardless of the augmentation used.

Across all three datasets, the results consistently demonstrate that the choice of audio data augmentation plays a critical role in shaping unlearning performance. Appropriate augmentations not only improve unlearning efficiency but also enhance the robustness of the resulting models. However, the impact of augmentation varies across different unlearning methods, making it essential to carefully tailor the selection of augmentation strategies to each method. Notably, the RTE analysis reveals that, across all unlearning methods, audio data augmentation exerts minimal influence on runtime efficiency while offering significant speed advantages compared to full retraining. These observations underscore that augmentation is not merely supportive but fundamental to achieving reliable and resilient audio unlearning.

4. CONCLUSION & FUTURE WORK

In this work, we present the first systematic empirical study on the role of audio data augmentation in audio unlearning. Using the Speech Command, ESC-10, and US8K datasets with multiple state-of-the-art unlearning methods, we show that carefully selected augmentation strategies narrow the performance gap between retrained and unlearned audio models. We further highlight their benefits for class-wise forgetting, demonstrating consistent improvements. These findings establish audio data augmentation as an essential component of machine unlearning pipelines. For future work, we aim to extend this study to larger audio and speech datasets and explore advanced models such as Audio Mamba and the Audio Spectrogram Transformer, thereby enhancing the robustness and generalizability of our conclusions.

5. REFERENCES

[1] Alessandro Mantelero, "The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'," *Computer Law & Security Review*, vol. 29, no. 3, pp. 229–235, 2013.

[2] Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu, "Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[3] Mark He Huang, Lin Geng Foo, and Jun Liu, "Learning to unlearn for robust machine unlearning," in *European Conference on Computer Vision*. Springer, 2024, pp. 202–219.

[4] Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu, "Machine unlearning: A comprehensive survey," *arXiv preprint arXiv:2405.07406*, 2024.

[5] Aditya Golatkar, Alessandro Achille, and Stefano Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9304–9312.

[6] Vinith Suriyakumar and Ashia C Wilson, "Algorithms that approximate data removal: New results and limitations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18892–18903, 2022.

[7] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter, "Tofu: A task of fictitious unlearning for llms," *arXiv preprint arXiv:2401.06121*, 2024.

[8] Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana R Kompella, Sijia Liu, and Shiyu Chang, "Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference," *Advances in Neural Information Processing Systems*, vol. 37, pp. 12581–12611, 2024.

[9] Aly M Kassem, Omer Ahmed Mohamed Mahmoud, and Sherif Saad, "Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models," 2023.

[10] Alkis Koudounas, Claudio Savelli, Flavio Giobergia, and Elena Baralis, ""Alexa, can you forget me?" Machine Unlearning Benchmark in Spoken Language Understanding," in *Interspeech 2025*, 2025, pp. 1768–1772.

[11] Jiali Cheng and Hadi Amiri, "Speech Unlearning," in *Interspeech 2025*, 2025, pp. 3209–3213.

[12] Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu, "Label smoothing improves machine unlearning," *arXiv preprint arXiv:2406.07698*, 2024.

[13] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu, "Model sparsity can simplify machine unlearning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51584–51605, 2023.

[14] Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou, "What makes unlearning hard and what to do about it," *Advances in Neural Information Processing Systems*, vol. 37, pp. 12293–12333, 2024.

[15] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar, "A survey of synthetic data augmentation methods in machine vision," *Machine Intelligence Research*, vol. 21, no. 5, pp. 831–869, 2024.

[16] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[17] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[20] Laura Graves, Vineel Nagisetty, and Vijay Ganesh, "Amnesiac machine learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 11516–11524.

[21] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli, "Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 7210–7217.

[22] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang, "Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7766–7775.

[23] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli, "Fast yet effective machine unlearning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 13046–13055, 2023.

[24] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou, "Towards unbounded machine unlearning," *Advances in neural information processing systems*, vol. 36, pp. 1957–1987, 2023.