Blue Sky: Expert-in-the-Loop Representation Learning Framework for Audio Anti-Spoofing: Multimodal, Multilingual, Multi-speaker, Multi-attack ($4M$) Scenarios

Zahra Khanjani*, Vandana P. Janeja*†, Christine Mallinson*, Sanjay Purushotham*

Abstract

Audio spoofing has surged with the rise of generative artificial intelligence, posing a serious threat to online communication. Recent studies have shown promising avenues in detecting spoofed audio specifically those that use human expert knowledge in representation learning, but more work is needed to evaluate performance across various realistic scenarios that tend to pose challenges in spoofed audio detection. In this paper, we introduce a comprehensive framework for expert-in-the-loop representation learning for audio anti-spoofing that is robust enough to address four specific challenging scenarios. Multimodal, Multilingual, Multi-speaker, and Multi-attack ($4M$). Preliminary results demonstrate the framework's potential effectiveness in audio anti-spoofing.

1 Introduction

Spoofed media can be generated or altered with Artificial Intelligence (AI) (as in deepfakes), or without AI e.g., replay attacks and mimicry). Currently, with a surge in generative AI, deepfakes have emerged as a significant issue threatening the authenticity of the content humans encounter. In the study of deepfakes, the task of detecting fake videos has drawn much more interest from researchers than fake audio [1]. Our work focuses on fake audio detection as an understudied area [2]. Although deepfakes can have positive applications, as in AI voice assistants, such as for people who have lost their voices due to laryngectomies, there are vastly more challenges resulting from the use of deepfakes in cases of fraud and deception. For example, a CEO of a U.K. energy based firm lost 220,000 Euros based on a phone call with a presumed executive that was actually a deepfake voice [3]. In another vein, robocalls containing spoofed audio have spread misinformation regarding political elections [4]. These and other examples motivate the need for a serious and sustained investigation into deepfake detection, especially audio.

Although human-in-the-loop machine learning is commonly used to enhance models in areas such as computer vision and natural language processing [5], this helpful approach is rarely used in anti-spoofing and deception detection and remains an open challenge [1,6]. This paper presents a comprehensive framework to incorporate human expert knowledge in audio anti-spoofing, emphasizing four realistic and challenging evaluation scenarios that are currently overlooked in the literature. The proposed expert-in-the-loop representation learning framework can be used for all types of media anti-spoofing, although the focus of this paper and the evaluation scenarios are tailored to audio.

The rest of paper addresses these questions respectively: Section 2 (What is the Blue Sky Idea? What will success look like?), Section 3 (Why is it a Blue Sky Idea? Why should the community ponder over it? Why now?), Section 4 (Does the Blue Sky Idea push the frontier, does it challenge our current set of assumptions, or does it take a bold approach to solve a wicked problem?), Section 5 (What are the challenges?), Section 6 (Are there any primary experiments showing the framework is promising?), and Section 7, conclusion.

2 What is the Blue Sky Idea? What does success look like?

In this section, we explain our 4M expert-in-the-loop representation learning framework for audio anti-spoofing as depicted in Figure1. Past work using similar approaches have demonstrated promise. As shown in our prior study [8], sociolinguistics experts were involved in spoofed audio detection, by defining linguistic representations of audio data that were then auto-labeled, using machine learning models trained with the labeled data extracted by the experts, as shown in 6 and [9]. In our proposed blue sky framework, we define multiple model augmentations (Table 1) to involve the expert-in-the-loop representations for anti-spoofing as mentioned in Table 1. In this table, y_i^{pred} refers to the class (spoofed vs genuine) obtained from common baselines, or state-of-the-art methods such as models based on Self Supervised Learning (e.g. [10]); while $y_i^{pred-m-r}$ refers to the predicted class from automated expert-in-the-loop representations.

*University of Maryland Baltimore County
†Corresponding author: vjaneja@umbc.edu

Model Augmentation	Mathematical Definition
Ensemble Modeling	If $y_i^{pred} = spoofed \vee y_i^{pred-m-r} = spoofed$ then $\hat{y_i} = spoofed$.
Modifying Loss Function	Given the set of pairs: $(x_i = (x_{i,1}, .., x_{i,d}), y_i)_{i=1}^n$, $f : \mathbb{R}^d \to \mathbb{R}$, Such that the residual $(\|\|f(X) - y\|\|)^2$ is minimized, where $X \in \mathbb{R}^{n \times d}$ has entries $X_{i,j} = x_{i,j}$, we assume $f(X) = (f(x_i))_{i=1}^n \in \mathbb{R}^n$. Then, $\mathrm{L}_{\text{modified}} = \|\|f(X) - y\|\|^2 + \lambda\|\|f(X) - y_{NR}\|\|^2$
Feature Concatenation	Given the output of the last layer k-th as: $\mathrm{f(x)} = \mathrm{Layer}_k(\mathrm{Layer}_{k-1}(\ldots(\mathbf{x}))), and$ m-r, $\mathbf{z} = \begin{bmatrix} \mathbf{f(x)} \\ \text{m-r} \end{bmatrix}; then, \hat{y_i} = \mathrm{Classifier}(\mathrm{z}_i)$

Table 1: Proposed model augmentations to incorporate automatically extracted expert-in-the-loop representations (automated m-r). λ: A hyperparameter that controls the contribution of the knowledge term to the overall loss. It balances the trade-off between fitting the training data and adhering to the knowledge.

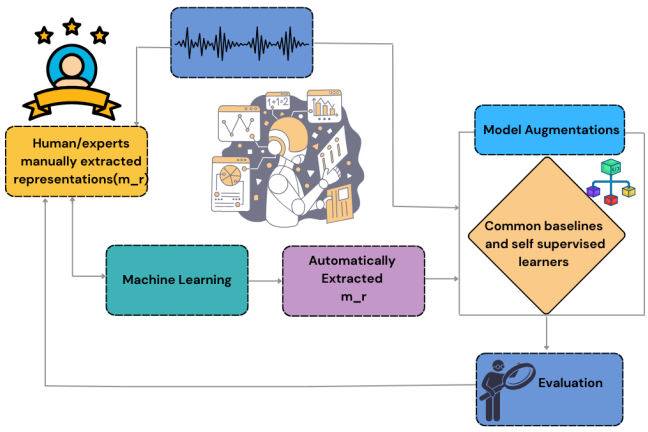

Figure 1: Human-in-the-loop representation learning framework for media anti-spoofing. Human experts' representations guide AI detectors to extract more relevant embedding from audio data. The model augmentation methods integrate these embeddings and steer the AI detectors based on human insights. m-r means representations obtained from experts.

The Ensemble Modeling presented above is applied to audio anti-spoofing [8]. The emphasis is on capturing spoofed audio; therefore, in the ensemble setup, if at least one of the expert-in-the-loop representations or the baselines labels content as spoofed, the final predicted class will be spoofed. For Modifying Loss Function with respect to the expert-in-the-loop representations, different approaches can be used, but most of them are for involving structured/ formalized knowledge such as physics equations [12]. When human experts are involved, unstructured knowledge is expected. In this situation, we encourage the use of the symbolic component of Neuro-Symbolic AI, which is capable of providing analytical equations (tractable mathematical expressions) purely from data [11]. After that, to use the knowledge-guided loss function, we modify the learning algorithm of ML models, to favor the selection of models that are consistent with obtained equations ((y_{NS}) as presented in Table 1). In a typical loss function, we compute the difference between the predicted values $f(X)$ and the true values (y). By adding y_{NR} we penalize deviations from both the observed data and the knowledge-driven equations. In the Feature Concatenation method, we combine automated m-r features and the output of one layer before the last layer of the neural network, to form a single input for the last dense layer and final classification. We aim to enhance the model's ability to make better predictions by integrating expert-in-the-loop representations. For the final phase of the framework we delve into evaluation under four highly significant and challenging scenarios: Multimodal, Multilingual, Multi-speaker, and Multi-attack $(4M)$. Multi-modality is emerging as a major threat when evaluating anti-spoofing models, as multimodal content is presented frequently on social media – e.g. a real video with spoofed audio [6]. Multilingualism is another important consideration, given the misalignment between current deepfake detection methods and the reality of global linguistic diversity. While content and data available on the internet are presented in various languages [6], most detection models are benchmarked on English datasets. Their capabilities in anti-spoofing in other languages are still understudied, causing great vulnerability for global misinformation and disinformation. The third "M"—Multi-speaker—emphasizes the need for approaches that can effectively evaluate anti-spoofing models on content that includes more than one audio signal. For instance, scenarios with audio that includes several speakers, have yet to be explored for audio deepfake detection. This scenario can encompass various sub-scenarios, such as audio that features multiple fake (AI-generated/synthetic) or multiple real speakers, or combinations of both. The fourth "M" is the Multi-attack scenario. This involves generalization for different types of attacks, both AI-generated (e.g., deepfakes) and non-AI-generated (e.g., replay attacks and mimicry). It is crucial to evaluate anti-spoofing systems using datasets that contain all types of attacks. For ex-

ample, in Automatic Speaker Verification and Spoofing Countermeasures Challenges [7] (ASVspoof), baseline models that perform well against certain types of attacks (e.g., logical access) often struggle against others (e.g., physical access). Since an attacker's choice of spoofing method is unpredictable, models must be robust and capable of generalizing well across various attack scenarios.

Given these challenges, we define success for our approach as a practical improvement in $4M$ scenarios, suggesting that the best performing models are robust and generalizable across modalities, languages, audio signals, and types of attacks. At the end of the loop of the framework, the evaluation outcome is brought to human experts, who may adjust and present additional relevant representations. This adjustment phase is crucial: with rapid advancements in generative AI, as expert-in-the-loop features may become less effective over time.

3 Why is it a Blue Sky Idea? Why now?

Incidents of deception and fraud that use generative AI are rapidly proliferating—and necessitating a greater investment by researchers into anti-spoofing and deepfake detection methods. When human experts are involved in anti-spoofing systems (such as [8]), there is improvement in the detectors' performance. However, the incorporation of expert knowledge has been very limited and applied in a few cases. Moreover, most methods for media anti-spoofing suffer from a lack of realistic scenarios in the evaluation phase, including focusing only on dominant languages, single modality, solo speaker audio, and solo attack scenarios—none of which are realistic and do not serve as a good simulation of what attackers can do. All of these issues yield models that are not sufficiently generalizable or reliable.

Our approach addresses these challenges (Figure 1). Regarding the critical need for greater attention to expert-in-the-loop representation frameworks, our 4M approach can serve as a comprehensive guide to incorporating human knowledge in anti-spoofing systems, with demonstrated ability to enhance the generalizability and robustness of the models. Our 4M framework is also centered on how to address challenging realistic scenarios.

4 Significance and Impact

Incorporating human expert knowledge into current anti-spoofing models is a novel and complex task that can push the frontier of the field. Traditional ML that only uses ML to enhance itself is a recursive process, with known limitations and dangers. Recently, for instance, researchers demonstrated how the performance of large language models is decreased when they are trained on AI generated data [13]. Anti-spoofing systems are prone to the same risks. As [14] notes, given the rapid progress in AI systems, training on data that are synthetically generated, often containing errors, can not only lead to diminished model performance but also amplify social inequality and weaken people's shared understanding of social reality. We therefore encourage expert-in-the-loop representation frameworks as a means of preserving human uniqueness in AI models and avoiding the risks of training on problematic data.

5 Challenges and Opportunities

In this section, we discuss several relevant challenges and definitions of success. First, we note the lack of appropriate datasets, which often do not contain realistic data or real-world scenarios. Available benchmarking datasets usually include data in English and contain a single type of attack. Most datasets also generally include single-stream audio data; to our knowledge, no multispeaker deepfake detection datasets currently exist. Second and relatedly, because spoofed media is a subcategory of deception, as [6] mentions, individuals and organizations are often hesitant to share sensitive information regarding targeted attacks. Researchers thus may face a lack of data from real-world attacks, as well as issues with imbalanced datasets.

Expert-in-the-loop representations may offer a more robust approach, as they thrive on a well-defined, multidisciplinary approach to achieve success. Experts from various fields may use different terminologies for similar concepts, highlighting the need for clear communication. It is crucial to share evaluation outcomes with human experts, providing the opportunity to adjust representations as needed. Scaling this with randomized evaluations for a closed loop is one way to ensure expert input. This collaborative process fosters trust between AI specialists and professionals in other domains, paving the way for reliable advancements.

6 Preliminary Results

In this section, we present preliminary results from our framework that demonstrate its promise as a pathway for media anti-spoofing, specifically audio anti-spoofing. This section is inspired by [8], which uses expert-in-the-loop representations—namely, linguistic cues called Expert Defined Linguistic Features (EDLFs)—to enhance a common audio anti-spoofing baseline. As a proof-of-concept, this section focuses on the Multiattack scenario. Through these results in one of the 4M scenarios we demonstrate the importance of the expert-in-the-loop approach to address for other M scenarios.

We trained machine learning models with five

EDLFs to automatically extract them from audio data [9]. The automatically extracted EDLFs are called predicted EDLFs or $EDLF^p$. These $EDLF^p$, then, serve as helper features in a model augmentation which is an ensemble modeling. We consider three best performing baselines from ASVspoof 2021 [7], as follows: 1) a pre-trained model of Light Convolutional Neural Network (LCNN) followed by Long Short-Term Memory (LSTM) and input with Linear Frequency Cepstral Coefficients (LFCCs); 2) LFCC-Gaussian Mixture Model (GMM); and 3) RawNet2. Our experiments demonstrate that all of the baselines are improved when we incorporate both manually and automatically extracted expert-in-the-loop representations, as shown in Figure 2. The figure indicates, both manual and automated expert-in-the-loop representations improved the performance of the baselines, in the multi-attack scenario. The benchmarking dataset contains multiple types of spoofed audio (mimicry, replay, and deepfakes).

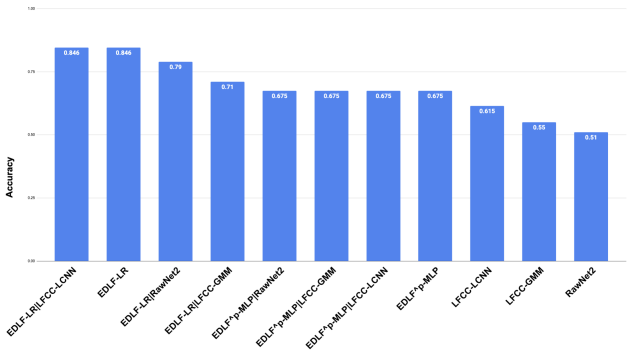

Figure 2: Accuracy for the unseen data. EDLFs are expert-in-the-loop representations and $EDLF^p$ are their auto-labeled version.

7 Conclusion

Deepfakes can occur in audio, video, text and image forms [2]; however, audio and video deepfakes may be potentially more harmful than other forms, even as audio forms are understudied in comparison to video and image. A longstanding question at the intersection of AI and different domains, and recently in the field of deception detection [6], is: can human experts enhance the performance of AI detectors? Our expert-in-the-loop representation learning framework begins to affirmatively answer this question, with preliminary results that show a promising pathway for detecting spoofed audio when sociolinguistic experts are involved. Further, our $4M$ framework introduces true-to-life scenarios to evaluate anti-spoofing systems in an approach that is comprehensive and more realistic—and one that is especially needed, given the prevalence of real-world deepfake fraud and deception.

Acknowledgments

Authors would like to acknowledge NSF award #2346473 and #2210011. Authors would like to acknowledge the contribution of Pragya Pandit for designing the infographic.

References

[1] Pham, L., Lam, P., Nguyen, T., Tang, H., Nguyen, H., Schindler, A., Vu, C. (2024). A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection. arXiv preprint arXiv:2409.15180.
[2] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," Frontiers in Big Data, vol. 5, 2022.
[3] Catherine Stupp. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. 2019.
[4] Verma, P., Kornfield, M., "Democratic operative admits to commissioning Biden AI robocall in New Hampshire" https://www.washingtonpost.com/technology/2024/02/26/ai-robocall-biden-new-hampshire/ (accessed Sep. 12, 2024).
[5] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 135, 364-381.
[6] Boumber, D., Verma, R. M., Qachfar, F. Z. (2024). Blue Sky: Multilingual, Multimodal Domain Independent Deception Detection. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM) (pp. 396-399). Society for Industrial and Applied Mathematics.
[7] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans et al.,"Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," arXiv preprint arXiv:2109.00537, 2021.
[8] Z. Khanjani, L. Davis, A. Tuz, K. Nwosu, C. Mallinson and V. P. Janeja, "Learning to Listen and Listening to Learn: Spoofed Audio Detection Through Linguistic Data Augmentation," 2023 IEEE International Conference on Intelligence and Security Informatics (ISI), Charlotte, NC, USA, 2023, pp. 01-06.
[9] Khanjani, Z., Mallinson, C., Janeja, V.P., Foulds, J. ALDAS: Audio-Linguistic Data Augmentation for Spoofed Audio Detection. arXiv:2410.15577
[10] Xie, Y., Cheng, H., Wang, Y., Ye, L. (2023). Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. In Proc. INTERSPEECH (Vol. 2023, pp. 2808-2812)
[11] Landajuela, M., Lee, C. S., Yang, J., Glatt, R., Santiago, C. P., Aravena, I., ... Petersen, B. K. (2022). A unified framework for deep symbolic regression. Advances in Neural Information Processing Systems, 35, 33985-33998.
[12] Karpatne, A., Jia, X., Kumar, V. (2024). Knowledge-guided Machine Learning: Current Trends and Future Prospects. arXiv preprint arXiv:2403.15989.
[13] Harrison Dupre, M. "AI Loses Its Mind After Being Trained on AI-Generated Data" https://futurism.com/ai-trained-ai-generated-data (accessed Sep. 12, 2024).
[14] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, Y. N. Harari, Y. Zhang et al. Managing extreme AI risks amid rapid progress, arXiv preprint arXiv:2310.17688 (2023).