Gibberish is All You Need for Membership Inference Detection in Contrastive Language-Audio Pretraining

Ruoxi Cheng[1,*], **Yizhong Ding**[1,*], **Shuirong Cao**[2], **Zhiqiang Wang**[1,†], **Shitong Shao**[3]

Abstract

Audio can disclose PII, particularly when combined with related text data. Therefore, it is essential to develop tools to detect privacy leakage in Contrastive Language-Audio Pretraining(CLAP). Existing MIAs need audio as input, risking exposure of voiceprint and requiring costly shadow models. We first propose PRMID, a membership inference detector based probability ranking given by CLAP, which does not require training shadow models but still requires both audio and text of the individual as input. To address these limitations, we then propose USMID, a textual unimodal speaker-level membership inference detector, querying the target model using only text data. We randomly generate textual gibberish that are clearly not in training dataset. Then we extract feature vectors from these texts using the CLAP model and train a set of anomaly detectors on them. During inference, the feature vector of each test text is input into the anomaly detector to determine if the speaker is in the training set (anomalous) or not (normal). If available, USMID can further enhance detection by integrating real audio of the tested speaker. Extensive experiments on various CLAP model architectures and datasets demonstrate that USMID outperforms baseline methods using only text data.

1 Introduction

Microphones in Internet of Things (IoT) devices (Abdul-Qawy et al., 2015) like phones can lead to unintended inferences from audio (Shah et al., 2021; Feng et al., 2022; Zhao et al., 2023a; Li and Zhao, 2023). Vocal features and linguistic content can reveal personally identifiable information (PII) (Schwartz and Solove, 2011) like biometric identity and socioeconomic status. Combining audio with text data increases susceptibility to inference attacks. Thus, developing tools to detect privacy leakage in text-audio models like contrastive language-audio pre-training(CLAP) (Elizalde et al., 2023; Zhao et al., 2023b; Wu et al., 2023a) is essential.

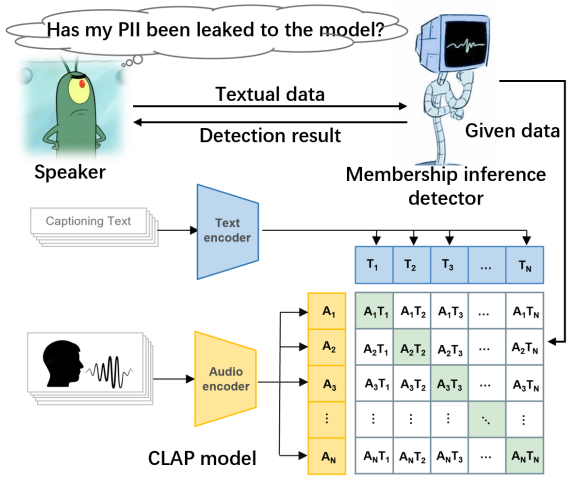

Figure 1: Current MIAs on MCL always query with dual-modal data of the tested individual for inference, while our goal is to avoid this.

Traditional methods like membership inference attacks (MIAs) (Shokri et al., 2017) focus on determining whether a specific data sample was used for model training. Research on MIAs for multimodal contrastive learning (MCL) (Yuan et al., 2021) like Contrastive Language-Image Pretraining(CLIP) (Radford et al., 2021) is extensive (Ko et al., 2023; Li et al., 2024a; Hintersdorf et al., 2024), but little attention is given to CLAP.

Traditional MIAs train shadow models to simulate target model's behavior (Abdullah et al., 2021; Chen et al., 2023; Tseng et al., 2021), which re-

*Contributed equally to this work. [1]Beijing Electronic Science and Technology Institute, Beijing, China. [2]AVIC Nanjing Engineering Institute of Aircraft Systems, Nanjing, Jiangsu, China. [3]The Hong Kong University of Science and Technology, Guangzhou, Guangdong, China. †Corresponding authors: wangzq@besti.edu.cn. Supported by the Fundamental Research Funds for the Central Universities (Grant No. 3282024050,3282024021); the key field science and technology plan project of Yunnan Province Science and Technology Department (Grant No.202402AD080004).

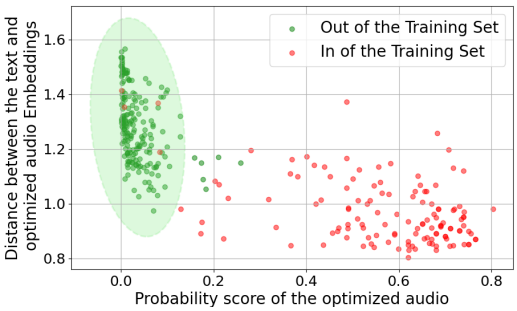

Figure 2: Optimization of audio is guided by a CLAP model trained on LibriSpeech dataset where each person has 50 audios. Distance between the embeddings of optimized audio and tested text, and probability score of the tested text among gibberish, can clearly distinguish between samples within and outside the training set of target CLAP model.

quires high computational costs, particularly for multimodal models like CLAP. We first propose PRMID, which uses the probability ranking provided by CLAP for membership inference detection, thereby avoiding the computational costs of shadow models.

However, current MIAs for MCL as well as PRMID often rely on dual-modal data inputs (Hu et al., 2022), which may lead to new leakage, as one modal of the pair might not have been exposed to the risky target model. Therefore, a detector that does not query CLAP with explicitly matched audio-text pair of speaker (see an example in Figure 1) is desirable. This concept is known as multimodal data protection (Liu et al., 2024).

To address these limitations, we propose USMID, a textual unimodal speaker-level membership inference (Miao et al., 2022) detector for CLAP models, which queries the target model with only text data. Specifically, we introduce a feature extractor that maps text data to feature vectors through CLAP-guided audio optimization. We then generate sufficient text gibberish that clearly does not match any text description in training dataset.

As shown in Figure 2, we observe a distinct separation between the features of gibberish and members in training set.

Based on this observation, we train multiple anomaly detectors using the feature vectors of generated text gibberish, creating an anomaly detection voting system. During testing, USMID inputs the feature vectors of test text into the voting system to determine if the corresponding speaker is in(anomalous) or out(normal) of the training set.

Our contributions are summarized as follows:

- We are the first to study membership inference detection in CLAP, constructing several audio-text pair datasets and trained various architectures of CLAP models.
- We introduce USMID, the first speaker-level membership inference detector for CLAP, which avoids exposing audio data to risky target model and the high cost for training shadow models in traditional MIAs.
- Extensive experiments show that USMID outperforms all baselines even using only text PII for query.

2 Related Work

2.1 Contrastive Language-Audio Pretraining

Contrastive language-audio pretraining(CLAP) has significantly improved multimodal representation learning (Wu et al., 2023b; Zhao et al., 2023b). Techniques like DSCLAP and T-CLAP enhance domain-specific applications and temporal alignment, showcasing the effectiveness of integrating language and audio (Li et al., 2024b).

2.2 Membership Inference in Automatic Speech Recognition

Recent studies show that automatic speech recognition (ASR) systems are vulnerable to MIAs(Li and Zhao, 2023; Shah et al., 2021).These MIAs typically rely on costly shadow models(Chen et al., 2023) and require real audio as input to target model(Abdullah et al., 2021), which may lead to new leakage.

3 Threat Model

Consider a CLAP model M trained on a dataset D_{train}. Each sample $s_i = (t_i, x_i)$ in D_{train} contains the PII of a speaker, consisting of a textual description t_i and its corresponding audio x_i. For distinct indices $i \neq j$, it is possible for $t_i = t_j$ while $x_i \neq x_j$, indicating that multiple non-identical audio samples may exist for the same speaker.

Detector's Goal. The detector aims to probe potential leakage of a speaker's PII through the target CLAP model M, seeking to determine whether any PII of the speaker were included in the training set D_{train}. For a speaker with textual description t, the detector aims to determine whether there exists a PII sample $(t_i, x_i) \in D_{\text{train}}$ such that $t_i = t$.

Note that our goal is not to detect a specific text-audio pair (t, x), but rather to identify the existence of any pair with textual description t. This is because that multiple audio samples of the same

speaker may be used for training, any of which could contribute to potential PII leakage.

Detector's Knowledge and Capability. The detector can query M and observe the output, including extracted audio and text embeddings as well as their matching score. For the target textual description t, depending on the application scenarios, the detector may or may not have actual audios corresponding to t. However, if the detector does have the corresponding audio samples, it cannot include them in its queries to M due to privacy concerns. Additionally, the detector is unable to modify M or access its internal state.

4 Methodology

4.1 Probability Ranking Membership Inference Detector

CLAP is trained to maximize cosine similarity between audio and text features of members. Thus, if one modality of a member is provided to target model, the corresponding other modality data typically yields a higher probability score in the calculated distribution when input alongside other samples.

Based on this, we propose PRMID (Probability Ranking Membership Inference Detector) as shown in Figure 3.

Probability Distribution Evaluated by CLAP. We first match the tested audio x with tested text t and a set of textual gibberish $\mathcal{G} = \{g_1, g_2, \ldots, g_\ell\}$. We use CLAP to obtain the probability distribution $\mathcal{P} = \{P(t), P(g_1), P(g_2), \ldots, P(g_\ell)\}$, where $P(t) + P(g_1) + P(g_2) + \ldots + P(g_\ell) = 1$.

Membership Inference through Ranking. We define the rank of the tested text t within the probability distribution $\mathcal{P}$ as $r_t = P(t)$. We conduct N repeated experiments, generating ℓ gibberish samples in each trial. Each experiment yields a probability distribution $\mathcal{P}$, which enables us to analyze r_t.

We set thresholds T_1 and T_2 for top $k\%$ and bottom $k\%$, where $k\%$ is a specified percentage (for example, 1%).

We consider three scenarios below:

- If count of r_t in top $k\%$ exceeds T_1 across N experiments, we infer that both t and x are present in D_{train}.
- If count of r_t in bottom $k\%$ exceeds T_2 across N experiments, t is outside of D_{train}, while x remains within.
- A sample is classified as random if r_t exhibits a uniform distribution across all $\ell + 1$ options. Specifically, the expected probability for any rank is $\frac{1}{\ell+1}$. If the observed frequencies for each rank fall within the expected range of $\frac{N}{\ell+1}$, we conclude that t is outside of D_{train}, with the status of x remaining undetermined.

Membership inference for Audio. In reverse inference, we can swap the roles of audio and text and repeat the inference process above as illustrated in Figure 4, allowing membership inference for both modalities.

4.2 Unimodal Speaker-Level Membership Inference Detector

While PRMID requires both audio and text inputs from the individual as input for the target model, this can introduce new privacy risks, as the target model may not have previously encountered dual-modal PII of that individual.

To address this limitation, we propose USMID (unimodal detector for membership inference detection). This detector is designed to ascertain whether the PII of a speaker is included in the training set of target CLAP model M, under the condition that only the speaker's textual description is provided to M.

An overview of USMID is illustrated in Figure 5. Firstly, for a textual description t, we develop a feature extractor to map t to a feature vector, through audio optimization guided by CLAP. Then, we make the key observation that *textual gibberish like "dv3*4l-XT0"—random combinations of numbers and symbols clearly do not match any textual descriptions in training set*, and hence the detector can generate large amount of textual gibberish that are known out of D_{train}. Using feature vectors extracted from these gibberish, detector can train multiple anomaly detectors to form an anomaly detection voting system. Finally, during inference phase, the features of the target textual description are fed into the system, and the inference result is determined through voting. Furthermore, when actual audio samples corresponding to the textual description are available, the detector can leverage them to perform clustering on feature vectors of the test samples to enhance detection performance.

Feature Extraction through CLAP-guided Audio Optimization. The feature extraction for a textual description t involves iterative optimization of an audio x, to maximize the correlation between

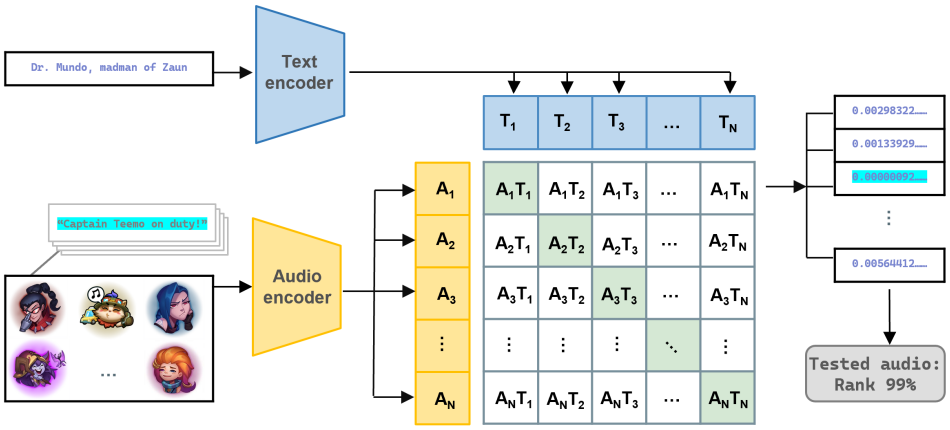

Figure 3: To determine whether a person's text is in the training set, we input his audio alongside a collection of other individuals' audios into the CLAP model. The model then generates a probability distribution based on the matching scores, which we use to conduct inference.

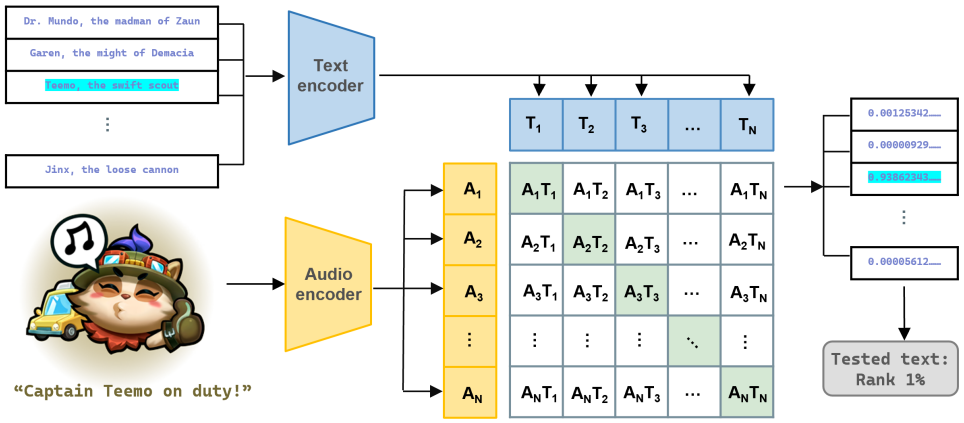

Figure 4: To determine whether a person's audio is in the training set, we input his text alongside a collection of texts from other individuals.

the embeddings of t and x produced by the target CLAP model. The extraction process, described in Algorithm 1, iterates for n epochs; and within each epoch, an audio is optimized for m iterations, to maximize the cosine similarity between its embedding of CLAP and that of target textual description. The average optimized cosine similarity S and standard deviation of optimized audio embeddings D are extracted as the features of t from model M.

Generation of Textual Gibberish. USMID starts the detection process with generating a set of ℓ gibberish strings $\mathcal{G} = \{g_1, g_2, \ldots, g_\ell\}$, which are random combinations of digits and symbols with certain length. As these gibberish texts are randomly generated at the inference time, with overwhelming probability that they did not appear in the training set. Applying the proposed feature extraction algorithm on $\mathcal{G}$, we obtain ℓ feature vectors $\mathcal{F} = \{f_1, f_2, \ldots, f_\ell\}$ of the gibberish texts.

Training Anomaly Detectors. Motivated by the observations in Figure3 that feature vectors of the texts in and out of the training set of M are well separated, we propose to train an anomaly detector using $\mathcal{F}$, such that texts out of D_{train} are considered "normal", and the problem of membership inference on t is converted to anomaly detection on its feature vector. More specifically, t is classified as part of D_{train}, if its feature vector is detected "abnormal" by the trained anomaly detector. Specifically in USMID, we train several anomaly detection models on $\mathcal{F}$, such as Isolation Forest (Liu et al., 2008), LocalOutlierFactor (Cheng et al., 2019) and AutoEncoder (Chandola et al., 2009). These models constitute an anomaly detection voting system that will be used for membership inference on the test textual descriptions.

Textual Membership Inference through Voting. For each textual description t in the test set, USMID first extracts its feature vector f using Algorithm 1, and then feeds f to each of the obtained

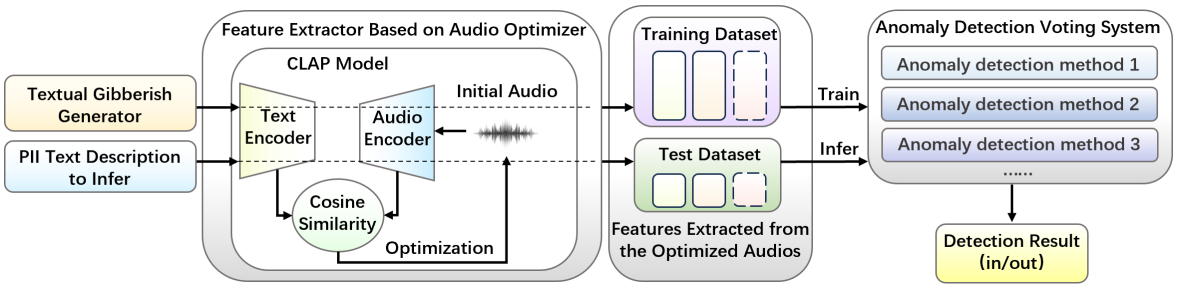

Figure 5: Overview of USMID.

Algorithm 1 CLAP-guided Feature Extraction

Input: Target CLAP model M, textual description t

Output: Mean optimized cosine similarity S, standard deviation of optimized audio embeddings D

$n \leftarrow$ number of epochs
$m \leftarrow$ number of optimization iterations per epoch
$\mathcal{S} \leftarrow \emptyset, \mathcal{V} \leftarrow \emptyset$
$v_t \leftarrow M(t)$ ▷ Obtain text embedding from M
for $i = 1$ **to** n **do**
  $x_0 \leftarrow \text{Rand}()$ ▷ Randomly generate an initial audio
  for $j = 0$ **to** $m - 1$ **do**
    $v_{x_j} \leftarrow M(x_j)$ ▷ Obtain audio embedding from M
    $x_{j+1} \leftarrow \arg\max_{x_j} \frac{v_t \cdot v_{x_j}}{\|v_t\| \, \|v_{x_j}\|}$ ▷ Update audio to maximize cosine similarity
  end for
  $S_i \leftarrow \frac{v_t \cdot v_{x_m}}{\|v_t\| \, \|v_{x_m}\|}$ ▷ Optimized similarity for epoch i
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{S_i\}, \mathcal{V} \leftarrow \mathcal{V} \cup \{v_{x_m}\}$
end for
$S \leftarrow \frac{1}{n} \sum_{S_i \in \mathcal{S}} S_i$
$\bar{v} \leftarrow \frac{1}{n} \sum_{v \in \mathcal{V}} v$
$D \leftarrow \sqrt{\frac{1}{n} \sum_{v \in \mathcal{V}} \|v - \bar{v}\|^2}$
return S, D

anomaly detectors to cast a vote on whether t is an anomaly. When the total number of votes exceeds a predefined detetion threshold N, t is determined as an anomaly, i.e., PII with textual description t is used to train the CLAP model M; otherwise, t is considered normal and no PII with t is leaked through training of M.

Enhancement with Real audios. At inference time, if real audios of the test texts are available at the detector (e.g., audios of a person), they can be used to extract an additional feature measuring the average distance between the embeddings of real audios and those of optimized audios using the CLAP model, using which the feature vectors of the test texts can be clustered into two partitions with one in D_{train} and another one out of D_{train}. This adds an additional vote for each test text to the above described anomaly detection voting system, potentially facilitating the detection accuracy.

Specifically, for each test text t, the detector is equipped with a set of c real audios $\{x^1_{\text{real}}, x^2_{\text{real}}, \ldots, x^c_{\text{real}}\}$. Similar to the feature extraction process in Algorithm 1, over k epochs with independent initializations, k optimized audios $\{x^1_{\text{opt}}, x^2_{\text{opt}}, \ldots, x^k_{\text{opt}}\}$ for t are obtained under the guidance of the CLAP model. Then, we apply a pre-trained feature extraction model F (e.g.,DeepFace for face audios) to the real and optimized audios to obtain real embeddings $\{v^1_{\text{real}}, v^2_{\text{real}}, \ldots, v^c_{\text{real}}\}$ and optimized embeddings $\{v^1_{\text{opt}}, v^2_{\text{opt}}, \ldots, v^k_{\text{opt}}\}$. Finally, we compute the average pair-wise ℓ_2 distance between the real and optimized embeddings, denoted by R, over $c \cdot k$ pairs, and use R as an additional feature of the text t.

For a batch of B test texts $(t_1, t_2, \ldots, t_B)$, we extract their features $((S_1, D_1, R_1), (S_2, D_2, R_2), \ldots, (S_B, D_B, R_B))$ first. Feeding the first two features S_i and D_i into a trained anomaly detection system, each text t_i obtains an anomaly score based on the number of detectors that classify it as abnormal. Additionally, the K-means algorithm with $K = 2$ partitions the feature vectors $\{(S_i, D_i, R_i)\}_{i=1}^B$ into "normal" cluster and an "abnormal" clusters, contributing another vote to the anomaly score of each instance. Finally, membership inference is performed by comparing the total votes received to a detection threshold N'.

5 Evaluations

We evaluate the performance of USMID, for speaker-level membership inference using only text

Table 1: Comparison with baseline methods.

Architecture	Number of Audios per person in training set	Method	Precision	Recall	Accuracy
LibriSpeech	1	Audio Auditor	63.38 ± 0.24	73.24 ± 0.33	65.19 ± 0.27
		SLMIA-SR	75.21 ± 0.18	88.64 ± 0.14	83.42 ± 0.21
		AuditMI	82.57 ± 0.21	95.26 ± 0.26	87.91 ± 0.24
		PRMID	85.32 ± 0.18	95.58 ± 0.22	89.75 ± 0.17
		USMID	**86.49 ± 0.19**	**96.49 ± 0.23**	**91.27 ± 0.15**
	50	Audio Auditor	65.59 ± 0.23	80.13 ± 0.16	66.59 ± 0.29
		SLMIA-SR	76.19 ± 0.31	90.07 ± 0.18	84.33 ± 0.25
		AuditMI	83.41 ± 0.14	98.04 ± 0.09	88.16 ± 0.13
		PRMID	86.15 ± 0.16	95.87 ± 0.24	90.12 ± 0.19
		USMID	**88.12 ± 0.26**	**98.76 ± 0.12**	**93.07 ± 0.16**
CommonVoice	1	Audio Auditor	54.85 ± 0.23	68.22 ± 0.19	60.52 ± 0.21
		SLMIA-SR	65.39 ± 0.36	76.91 ± 0.27	70.47 ± 0.24
		AuditMI	71.43 ± 0.28	81.45 ± 0.41	74.36 ± 0.18
		PRMID	72.35 ± 0.23	84.52 ± 0.20	78.43 ± 0.18
		USMID	**74.96 ± 0.25**	**86.01 ± 0.22**	**81.79 ± 0.15**
	50	Audio Auditor	56.11 ± 0.33	73.58 ± 0.27	61.35 ± 0.25
		SLMIA-SR	66.28 ± 0.21	79.27 ± 0.34	72.18 ± 0.22
		AuditMI	73.52 ± 0.17	84.81 ± 0.28	75.64 ± 0.23
		PRMID	75.12 ± 0.19	88.26 ± 0.18	80.98 ± 0.14
		USMID	**76.47 ± 0.12**	**89.46 ± 0.32**	**82.33 ± 0.19**

Table 2: Samples of randomly generated gibberish.

±dhu!f9dew	53e(s=pnI<S	fe3_;fw/
d3l%5G_	4teh<E{43ter	5gtb-hgF
#4c3rdg	’2_:gt6[45gb	g*l<trgtl3/

Table 3: Comparison of training time, GPU memory consumption, and inference time per sample with baselines on LibriSpeech dataset.

Method	Train Time	GPU Memory	Inference Time
Audio Auditor	7.5h	11.3GB	0.359s
SLMIA-SR	9h	13.7GB	0.406s
AuditMI	80h	49.5GB	2.375s
USMID	3.7h	24.3GB	0.628s

PII of the individual.

Dataset Construction. In addition to LibriSpeech (Panayotov et al., 2015), we built a speaker recognition dataset based on CommonVoice18.0 (Ardila et al., 2019), which covers various social groups and has richer background information. Specifically, 3,000 speakers (1,500 for training and 1,500 for verification) were selected from CommonVoice, and their audio files were accompanied by unique user PII like ID, age, gender, and region information; then for each user ID, we used GPT-4o to generate detailed background description based on their PII; finally, these expanded background descriptions and audio files corresponding to each user ID constituted the training set of CLAP.

By doing this, we obtained basic facts about who is in the training set and who is not. For each type of content, we created two datasets: one with 1 audio clip per person and another with 50 audio clips per person.

Models. In our CLAP model, audio encoder uses HTSAT(Chen et al., 2022), which is transformer with 4 groups of swin-transformer blocks(Liu et al., 2021). We use the output of its penultimate layer (a 768-dimensional vector) as the output sent to the projection MLP layer. Text encoder uses RoBERTa (Liu et al., 1907), which converts input text into a 768-dimensional feature vector. We apply a 2-layer MLP with ReLU activation (Agarap, 2018) to map the audio and text outputs to 512 dimensions for final representation.

Evaluation Metrics. USMID’s effectiveness is assessed using Precision, Recall, and Accuracy metrics, measuring anomaly prediction accuracy, correct anomaly identification, and overall prediction correctness, respectively.

Baselines. Current speaker-level membership inference detection methods require detector to query target model with real audio. Most MIAs involve training shadow models, which can be particularly costly for multimodal LLMs. We empirically compare the performance of USMID with PRMID and the following SOTA inference methods. The audio encoders for Audio Auditor and SLMIA-SR are LSTM, for AuditMI they are Transformer, and for

Table 4: Comparison of performance with a given audio.

Architecture	Number of audios per person in training set	USMID	Precision	Recall	Accuracy
LibriSpeech	1	Text only	86.49 ± 0.19	96.49 ± 0.23	91.27 ± 0.15
		With 1 audio	**89.21 ± 0.14**	**98.68 ± 0.18**	**93.54 ± 0.13**
	50	Text only	88.12 ± 0.26	98.76 ± 0.12	93.07 ± 0.16
		With 1 audio	**91.63 ± 0.21**	**99.57 ± 0.08**	**95.24 ± 0.23**
CommonVoice	1	Text only	74.96 ± 0.25	86.01 ± 0.22	81.79 ± 0.15
		With 1 audio	**76.02 ± 0.17**	**89.55 ± 0.31**	**83.56 ± 0.21**
	50	Text only	76.47 ± 0.12	89.46 ± 0.32	82.33 ± 0.19
		With 1 audio	**79.34 ± 0.23**	**91.13 ± 0.16**	**85.69 ± 0.24**

PRMID and USMID, they are CLAP.

- **Audio Auditor** (Miao et al., 2022) trains shadow models and extracts audio features for inference.
- **SLMIA-SR** (Chen et al., 2023) employs a shadow speaker recognition system to train attack model.
- **AuditMI** (Teixeira et al., 2024) trains shadow model using input utterances and features from model outputs.

All experiments are performed using four NVIDIA GeForce RTX 4090 GPUs. Each experiment is repeated for 10 times, and the average values and the standard deviations are reported.

5.1 Results

On training anomaly detectors, we randomly generated $\ell = 100$ textual gibberish (some of them are shown in Table 2).

The audio optimization was performed for $n = 100$ epochs; and in each epoch, $m = 100$ Gradient Descent (GD) iterations with a learning rate of 3×10^{-2}. Four anomaly detection models, i.e., LocalOutlierFactor (Cheng et al., 2019), IsolationForest (Liu et al., 2008), OneClassSVM (Li et al., 2003; Khan and Madden, 2014), and AutoEncoder (Chen et al., 2018) were trained, and $N = 3$ was chosen as the detection threshold.

As shown in Table 1, USMID consistently outperforms all baselines even with only text PII, achieving a precision of 88.12% on LibriSpeech with 50 audio clips per person.

Additionally, USMID demonstrates notable advantages in training time and resource efficiency compared to baseline methods as shown in Table 3. It requires only 3.7 hours of training, much less than AuditMI's 80 hours, while maintaining competitive inference times.

We also evaluate the effect of providing USMID with a real audio of the tested person. In this case, the embedding distances between the real and optimized audios of the test samples are used to perform a 2-means clustering, adding another vote to the inference. We accordingly raise the detection threshold N' to 4. As illustrated in Table 4, the given audio helps to improve the performance of USMID across all tested CLAP models, showing an increase of 3.36% on CommonVoice with 1 audio clip per person.

5.2 Ablation Study

We further explore the impacts of different system parameters on the detection accuracy.

Optimization parameters. Figure 6 and 7 show that during feature extraction, optimizing for $n = 100$ epochs, each with $m = 1,000$ iterations, offers the optimal performance. Additional epochs and optimization iterations yield minimal improvements despite increased computational costs.

Detection threshold. Figure 8 and 9 show that the system achieves higher accuracy with a threshold of three votes for text-only inputs and four votes when real audio is included. A high threshold may miss anomalies, while a low threshold may incorrectly classify normal inputs as anomalies.

Number of textual gibberish. As shown in Figure 10, for different target models, the detection accuracies initially improve as the number of gibberish texts increases, and converge after using more than 50 gibberish strings.

Number of real audios. As shown in Figure 11, integrating real audios can enhance the detection accuracy; however, the improvements of using more than 1 audio are rather marginal.

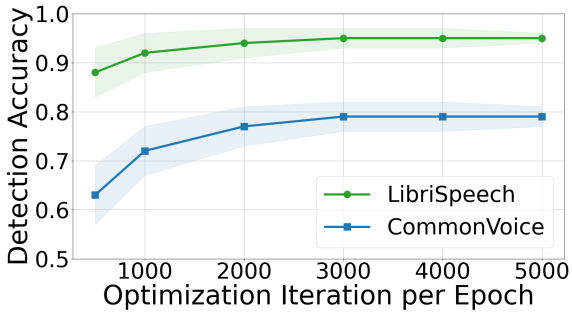

Figure 6: Detection accuracy for different numbers of optimization iterations per epoch.

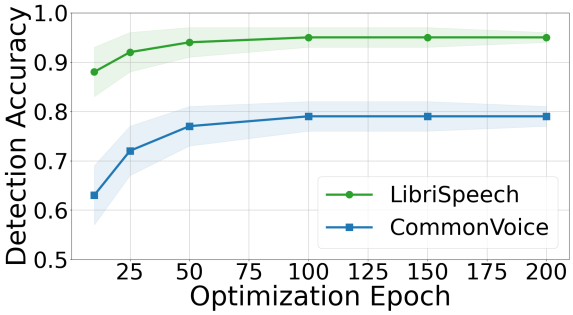

Figure 7: Detection accuracy for different numbers of epochs.

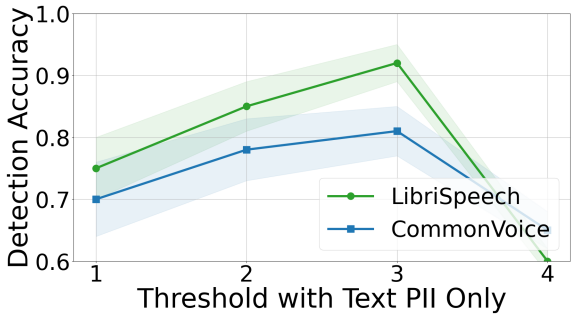

Figure 8: Detection accuracy with text PII only.

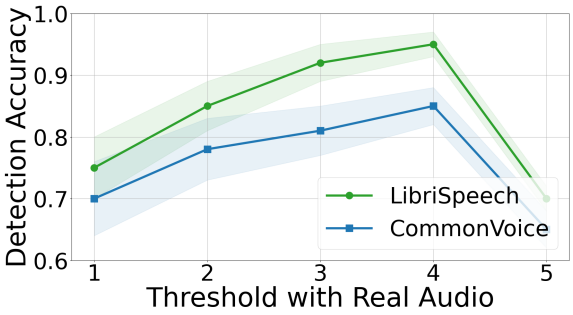

Figure 9: Detection accuracy with real audio for enhancement.

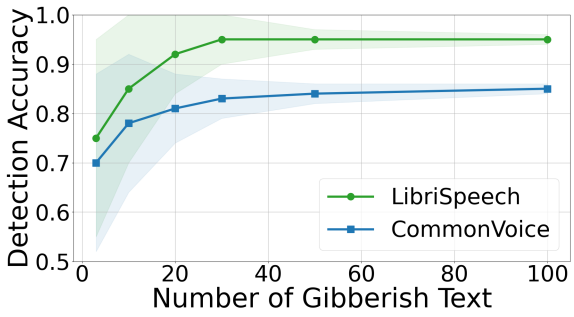

Figure 10: Detection accuracy for different numbers of gibberish.

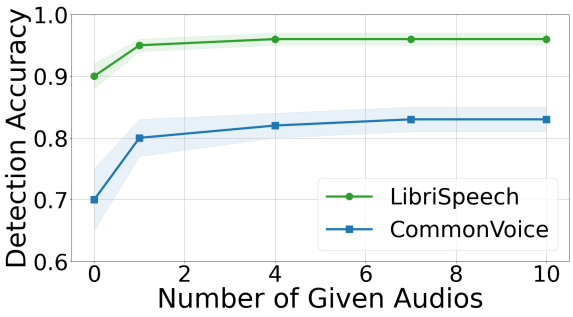

Figure 11: Detection accuracy for different number of real audio.

6 Defense and Covert Gibberish Generation

In real-world scenarios, target models may implement defense mechanisms to detect anomalous inputs like gibberish, potentially leading to misleading outputs that cause USMID to misidentify the inclusion of PII. To address this, we prompted GPT-3.5-turbo to generate fictional character backgrounds rather than mere gibberish as shown in Table 5.

Name	Occupation	Hometown
Jaston Spark	Alien Biologist	Martian Oasis
Carl Thunder	Climate Manipulator	Stormhaven
Vega Quasar	Cosmic Navigator	Starfall Galaxy

Table 5: Covert gibberish that seem to be real PII.

7 Conclusion

This paper presents the first focused study on membership inference detection in contrastive language-audio pre-training models. We introduce PRMID and USMID, both of which avoid the need for computationally expensive shadow models required in traditional MIAs. Additionally, USMID is the first approach to conduct membership inference without exposing real audio samples to target CLAP models. Evaluations across various CLAP model architectures and dataset demonstrate the consistent superiority of USMID across baseline methods.

References

Antar Shaddad Abdul-Qawy, PJ Pramod, E Magesh, and T Srinivasulu. 2015. The internet of things (iot): An overview. *International Journal of Engineering Research and Applications*, 5(12):71–82.

Hadi Abdullah, Muhammad Sajidur Rahman, Washington Garcia, Kevin Warren, Anurag Swarnim Yadav, Tom Shrimpton, and Patrick Traynor. 2021. Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 712–729. IEEE.

AF Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Guangke Chen, Yedi Zhang, and Fu Song. 2023. Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems. *arXiv preprint arXiv:2309.07983*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE.

Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. 2018. Autoencoder-based network anomaly detection. In *2018 Wireless telecommunications symposium (WTS)*, pages 1–5. IEEE.

Zhangyu Cheng, Chengming Zou, and Jianwei Dong. 2019. Outlier detection using isolation forest and local outlier factor. In *Proceedings of the conference on research in adaptive and convergent systems*, pages 161–168.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tiantian Feng, Raghuveer Peri, and Shrikanth Narayanan. 2022. User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning. *arXiv preprint arXiv:2204.02500*.

Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. 2024. Does clip know my face? *Journal of Artificial Intelligence Research*, 80:1033–1062.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. 2022. M^4i: Multi-modal models membership inference. *Advances in Neural Information Processing Systems*, 35:1867–1882.

Shehroz S Khan and Michael G Madden. 2014. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374.

Minseon Ko, Minseok Jin, Chen Wang, et al. 2023. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881.

Honglei Li and Xinlong Zhao. 2023. Membership information leakage in well-generalized auto speech recognition systems. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–7. IEEE.

Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. 2003. Improving one-class svm for anomaly detection. In *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, volume 5, pages 3077–3081. IEEE.

Songze Li, Ruoxi Cheng, and Xiaojun Jia. 2024a. Identity inference from clip models using only textual data. *arXiv preprint arXiv:2405.14517*.

Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. 2024b. Advancing multi-grained alignment for contrastive language-audio pre-training. *arXiv preprint arXiv:2408.07919*.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2024. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. *arXiv preprint arXiv:2407.16307*.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 1907. Roberta: a robustly optimized bert pretraining approach. corr 2019. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Yuantian Miao, Chao Chen, Lei Pan, Shigang Liu, Seyit Camtepe, Jun Zhang, and Yang Xiang. 2022. No-label user-level membership inference for asr model auditing. In *European Symposium on Research in Computer Security*, pages 610–628. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Paul M Schwartz and Daniel J Solove. 2011. The pii problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, 86:1814.

Muhammad A Shah, Joseph Szurley, Markus Mueller, Thanasis Mouchtaris, and Jasha Droppo. 2021. Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Francisco Teixeira et al. 2024. Exploring features for membership inference in asr model auditing. *Available at SSRN 4937232*.

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung-yi Lee. 2021. Membership inference attacks against self-supervised speech models. *arXiv preprint arXiv:2111.05113*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023a. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Huan Zhao, Haijiao Chen, Yufeng Xiao, and Zixing Zhang. 2023a. Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Tianqi Zhao, Ming Kong, Tian Liang, Qiang Zhu, Kun Kuang, and Fei Wu. 2023b. Clap: Contrastive language-audio pre-training model for multi-modal sentiment analysis. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 622–626.