Accepted by the 31st Network and Distributed System Security (NDSS) Symposium, 2024
(This is the version submitted for review; Final version under preparation)

SLMIA-SR: Speaker-Level Membership Inference Attacks against Speaker Recognition Systems

Guangke Chen[1], Yedi Zhang[1], Fu Song[1,2,3] (✉)
[1] ShanghaiTech University [2] Automotive Software Innovation Center
[3] Institute of Software, Chinese Academy of Sciences & University of Chinese Academy of Sciences

***Abstract*—Membership inference attacks allow adversaries to determine whether a particular example was contained in the model's training dataset. While previous works have confirmed the feasibility of such attacks in various applications, none has focused on speaker recognition (SR), a promising voice-based biometric recognition technique. In this work, we propose SLMIA-SR, the first membership inference attack tailored to SR. In contrast to conventional *example*-level attack, our attack features *speaker*-level membership inference, i.e., determining if *any* voices of a given speaker, either the same as or different from the given inference voices, have been involved in the training of a model. It is particularly useful and practical since the training and inference voices are usually distinct, and it is also meaningful considering the open-set nature of SR, namely, the recognition speakers were often not present in the training data. We utilize intra-closeness and inter-farness, two training objectives of SR, to characterize the differences between training and non-training speakers and quantify them with two groups of features driven by carefully-established feature engineering to mount the attack. To improve the generalizability of our attack, we propose a novel mixing ratio training strategy to train attack models. To enhance the attack performance, we introduce voice chunk splitting to cope with the limited number of inference voices and propose to train attack models dependent on the number of inference voices. Our attack is versatile and can work in both white-box and black-box scenarios. Additionally, we propose two novel techniques to reduce the number of black-box queries while maintaining the attack performance. Extensive experiments demonstrate the effectiveness of SLMIA-SR.**

I. Introduction

Speaker recognition (SR) is a promising biometric recognition technique that recognizes the identity of a person based on her/his voices [1]. SR is already deployed in a wide variety of realistic applications, such as identity verification of bank customers during telephone-communication [2], password-free and voice-based payment [3], device access control in smart home [4], and service personalization in voice assistants [5].

Modern and state-of-the-art speaker recognition systems (SRSs) are often built upon deep neural networks (DNNs) [6], which are trained on increasingly sensitive data. However, it is known that DNNs tend to memorize their training data because of overfitting [7], [8], [9]. As a result, an adversary given access to trained DNNs is able to recover representative views of a subset of training data [7], [8] or even reconstruct verbatim training data [9], leading to privacy leakage. Thus, it is crucial to evaluate the privacy risks of DNNs prior to deployment so that actions can be taken to enhance the privacy level based on the evaluation results. Nowadays, membership inference attacks (MIA) that are able to determine if a sample is involved in the training of a DNN are the de facto standard for assessing DNNs' privacy risks [10], [11], [12], [13].

While numerous studies have confirmed the feasibility of MIA in various applications of DNNs, including image classification [10], [14], [15], [11], [16], [12], [17], speech recognition [18], [19], language models [20], and generative models [21], [22], MIA against SRSs has not been considered yet. Considering the wide spread of voices across social media platforms, online meetings, and voice-enabled smart devices, users' voice data may be collected and used for training SRSs without their consent. For instance, Amazon was sued as Alexa recorded children's voiceprint without their parents' permission [23], probably used to improve its Voice ID [5] that recognizes users' identity to provide personalized services. It violates data protection regulations, e.g., GDPR [24]. Hence, ordinary users are increasingly eager to know if their voices were used for training SRSs without their permission. There are also regulations in place regarding artificial intelligence (AI) systems, e.g., the Blueprint for an AI Bill of Rights introduced by the White House [25]. This regulation requires companies to continuously inspect their systems to mitigate any unsafe outcomes that exceed their intended use. Given that many companies offer speaker recognition as a machine learning as a service (MLaaS), e.g., Microsoft [26] and Nuance [27], they need to evaluate the privacy level of their SRSs before making them publicly available. This evaluation is necessary to prevent queries to their systems from potentially disclosing sensitive information about the training data. These urge us to design MIA against SRSs. Also, understanding such attacks benefits further studies towards building more secure and privacy-preserving SRSs.

We first study the applicability of prior MIA to SRSs, which targeted conventional classification[1], generative, regression, or embedding models (cf. [14] for a survey). We find that: (1) The first three own distinct training paradigms and architectures from SRSs, so their MIA cannot be easily ported to SRSs. Take conventional classification models as examples on which most prior MIA focused [10], [14], [15], [11]. They are typically trained by minimizing the cross entropy loss on training samples, which requires appending a final fully connected layer to models. Thus, the outputs of the final layer are utilized to mount MIA [10], [14], [15], [11], [16], [12]. In contrast, the training of SR models aimed at deriving a voice embedding extractor utilizes either verification- or classification-based loss functions [6]. Verification-based loss functions (e.g., angular prototypical loss [28] and generalized end-to-end loss [29]) are computed on voice embeddings, and when minimized, the embeddings of two voices are close to each other (called *intra-closeness*) if they are uttered by the same speaker, otherwise far from each other (called *inter-*

[1]By conventional classification, we refer to the supervised learning paradigm that defines a set of target classes and directly trains a model to recognize them using labeled examples, e.g., MNIST. Fine-tuning after pretraining and few-shot learning-based facial recognition are out of the scope.

farness). SR models trained in this paradigm do not have a final fully connected layer. Though SRSs may be trained by classification-based loss (e.g., the cross entropy loss) in a similar paradigm as conventional classification, the final fully connected layer will be dropped after training due to the open-set nature, i.e., recognized speakers are enrolled speakers instead of training speakers. In contrast, conventional classification is of closed-set nature, as classes to be predicted are predefined and should be involved in training. (2) When prior MIA targeting embedding models [30], [31], [32], [33] are applied to speaker recognition, they lead to unsatisfactory performance, e.g., no more than 2% True Positive Rate (TPR) at 0.1% False Positive Rate (FPR), because these MIA are not tailored to speaker recognition. Details refer to Appendix A.

Motivated by the above study results, in this work, we propose and design the *first* membership inference attack tailored to speaker recognition. Our attack is *speaker-level* while most prior MIA are *example*-level, namely, determining if a given sample was contained in the training dataset or not. In the context of speaker recognition, example-level MIA is indeed a *voice*-level MIA, which becomes less practical, less useful, and thus less interesting. We argue that in the real world, it is of low probability that the voices of a training speaker provided for membership inference were used in training, hence it is more important to determine if *any* voices of a given speaker, either the same as or different from the voices provided for membership inference were used in training.

To perform speaker-level MIA against SRSs, we build an attack model that takes as input a few voices of a target speaker and outputs a binary decision indicating training (i.e., member) or non-training (i.e., non-member) speakers, by leverage the widely-adopted shadow training method [10], [14], [15], [11], i.e., training a shadow SRS to approximate the behavior of the target SRS. To train the attack model in a supervised manner, the following fundamental problem should be tackled: *how to characterize the differences between training and non-training speakers?* Observing that the training objectives of SRSs are *intra-closeness* and *inter-farness*, we hypothesize that training speakers enjoy better intra-closeness and inter-farness than non-training speakers. We then design features to quantify intra-closeness and inter-farness by using two types of similarities, four types of distances, four different statistics, and different arrangements of similarities and distances. Through this carefully-established feature engineering process, we totally design a set of 103 features, which are expected to comprehensively characterize the differences between training and non-training speakers in a complementary way.

However, there are still some challenges that need to be addressed. First, the speaker-level MIA should reliably reach the "member" decision for a training speaker when the ratio of her/his voices provided for membership inference that are used in training varies from 0 to 1. To realize such generalizability, we split the voices of the shadow SRS's training speakers and propose the mixing ratio training strategy to train the attack model. Regarding the number of voices provided for membership inference, to improve the attack performance, we propose to train voice-number-dependent attack models and propose a voice chunk splitting approach to artificially increase the number of inference voices. To reduce the number of queries probed to the target SRS in the black-box scenario, we propose two novel techniques, namely, group enrollment and enrollment voice concatenation.

We implement our approach in a tool, called SLMIA-SR, and thoroughly evaluate the performance of SLMIA-SR on two voice datasets and five SRSs under two settings regarding the number of inference voices. SLMIA-SR can achieve an average TPR of 10.2% at an extremely low FPR of 0.1% when there are only ten inference voices and none of them were used in training for training speakers. SLMIA-SR also outperforms previous MIA targeting embedding models, e.g., increasing the TPR at 0.2% FPR from 4.8% to 46.7%. We also conduct experiments to confirm the effectiveness of the approches to improve the attack performance and generalizability, and reduce the number of queries. For instance, our voice chunk splitting can boost the TPR at 0.1% FPR by 13%, and group enrollment and enrollment voice concatenation can reduce the queries from 400 to 30 without sacrificing accuracy. Finally, we perform ablation studies to study the effect of dataset distribution and architecture shift on SLMIA-SR. Unsurprisingly, performance may degrade, but SLMIA-SR still remains effective with more than 2% TPR at 0.1% FPR in the worst case.

To summarize, we make the following major contributions:

- We propose the first speaker-level membership inference attack SLMIA-SR for auditing privacy risks of speaker recognition systems.
- Through carefully-established feature engineering, we design 103 diverse features to quantify intra-closeness and inter-farness, and characterize the differences between training and non-training speakers in a comprehensive manner.
- We propose a mixing ratio training strategy to improve the generalizability, enabling SLMIA-SR to determine if any voices of a speaker were used in training regardless of the ratio of provided inference voices that were used in training.
- To enhance the attack performance, we propose to build voice-number-dependent attack models and propose a voice chunk splitting approach to cope with the limited number of inference voices.
- We propose two techniques, group enrollment and enrollment voice concatenation, to significantly reduce the number of queries probed to the target SRS in the black-box scenario, with no or little impact on the attack performance.

For convenient reference, we summarize the main notations used in this work in TABLE I. Our code is available at [34].

II. Background & Related Works

A. Speaker Recognition Systems

The overview of generic speaker recognition systems (SRSs) is shown in Fig. 1, comprising three phases: *training*, *enrollment*, and *recognition*. The training phase trains a background model using lots of voices from numerous training speakers [6] by minimizing a either classification- or verification-based loss function [6]. The background model learns a mapping $E(\cdot)$ from voices v to embeddings $E(v)$ such that the voice embeddings of the same speaker are pulled together, while the voice embeddings of distinct speakers are pushed away. Classification-based losses, e.g., cross entropy (CE) loss [6], aim to maximize the classification accuracy of the training speakers and require appending a fully connected

TABLE I: Main Notations.

Category	Notation	Meaning	Category	Notation	Meaning
Target	SR^t	target SRS	Shadow	SR^s	shadow SRS
	$\mathcal{S}^t_{tr}$	training speakers		$\mathcal{S}^s_{tr}$	training speakers
	$\mathcal{V}^t_{tr}$	training voices		$\mathcal{V}^s_{tr}$	training voices
	$\mathcal{V}^t_{ntr,tr}$	non-training voices of training speakers		$\mathcal{V}^s_{ntr,tr}$	non-training voices of training speakers
	$\mathcal{S}^t_{ntr}$	non-training speakers		$\mathcal{S}^s_{ntr}$	non-training speakers
	$\mathcal{V}^t_{ntr,ntr}$	non-training voices of non-training speakers		$\mathcal{V}^s_{ntr,ntr}$	non-training voices of non-training speakers
Imposter	$\mathcal{S}^{im}$	imposters	Auxiliary	$\mathcal{S}^a = \mathcal{S}^s_{tr} \cup \mathcal{S}^s_{ntr} \cup \mathcal{S}^{im}$	auxiliary set of speakers
	$\mathcal{V}^{im}$	imposters voices			
Approach	VCS	voice chunk splitting	Approach	Intra-Ens	ensemble of intra-features
	VND	voice-number-dependent attack models		Inter-Ens	ensemble of inter-features
	VNID	voice-number-independent attack models		SLMIA-SR	ensemble of intra- and inter-features

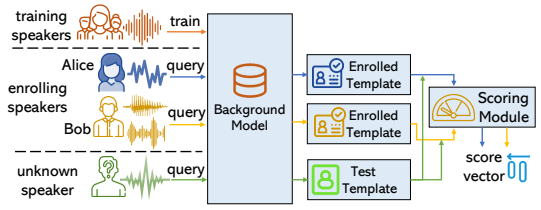


Fig. 1: Overview of SRSs.

layer such that the number of neurons is the same as the number of training speakers. But, different from conventional classification, the fully connected layer will be dropped after training and not used in later phases, since speakers in later phases are not necessarily involved in the training phase. Verification-based losses, e.g., angular prototypical (AP) loss [28] and generalized end-to-end loss (GE2E) [29], aim to minimize the Equal Error Rate (EER) of the training trials[2]. As such losses are directly computed on the voice embeddings, the fully connected layer is not needed. Typically, verification-based losses explicitly penalizes the distance among voice embeddings of the same speaker and the similarity between voice embeddings of distinct speakers. In contrast, cross entropy loss only explicitly separates different speakers, and there are some variants of cross entropy loss by explicitly regularizing the distance among voice embeddings of the same speaker, e.g., additive angular margin softmax loss (AAM) [35].

In the enrollment phase, an *enrolled template* $E(\mathbf{v}^e)$ is registered using the enrollment voice(s) $\mathbf{v}^e$ of an enrolling speaker. Note that the same speaker may be enrolled repeatedly to better characterize the speaker, so $\mathbf{v}^e$ can be multiple voices. When $|\mathbf{v}^e| > 1$, $E(\mathbf{v}^e)$ is the centroid of the embeddings $E(v)$ for $v \in \mathbf{v}$. An SRS may allow only one speaker (speaker verification) or multiple speakers (speaker identification) to enroll, leading to one or multiple enrolled templates, respectively. In the recognition phase, the *test template* $E(v)$ of a given voice v is first retrieved, and then the scoring module measures the similarity between each enrolled template $E(\mathbf{v}^e)$ and the test template $E(v)$, producing a (recognition) score $S(v|\mathbf{v}^e)$.

B. Security of Speaker Recognition Systems

Various security implications of SRSs have been unveiled. Audio adversarial example attacks [36], [37], [38], [39], [40], [41], [42], [43], [44] craft an adversarial voice from a voice uttered by a *source* speaker such that the SRS misclassifies it as a *target* speaker, but ordinary users do not. Hidden voice attacks [45] perturb a voice uttered by a target speaker such that the resulting voice is perceived as mere noise by humans, but is still correctly classified as the target speaker by the SRS. Audio deepfake attacks [46], including speech synthesis [47] and voice conversion [47], create a voice such that it is recognized as the target speaker by both SRSs and humans. Dictionary attacks [48] create a master voice that matches the identity of a large population instead of one specific target speaker. All these attacks are aimed at bypassing the authentication of target speakers using crafted voices. In contrast, this work reveals the privacy implication of SRSs. We show that the adversary can infer whether any voices of a given speaker were contained in the training of an SRS by querying the SRS and leveraging the feedback, thus obtaining extra information about the training speakers.

C. Membership Inference Attack

Membership inference attack (MIA) aims to determine whether a given example is contained in the training of a model [10], [14], posing privacy risks since it provides the adversary with extra information about the training data. The feasibility of MIA lies in the overfitting to the training data. The key to an effective MIA is the characterization of differences between training and non-training data via designated features, which should be tailored to the architecture and training paradigm of the specific task.

MIA on embedding models. This type of MIA targeted contrastive learning [31], speech self-supervised learning [32], metric learning-based person re-identification [30], and few-shot learning-based facial recognition [33]. The first attack is example-level while the others are user- or speaker-level. [32] utilized the average pairwise cosine similarity among embeddings of examples from the target speaker as the feature, with the assumption that this similarity is higher for the training speaker than non-training speakers. It also improved the attack by replacing the predefined cosine similarity with a similarity metric learned by neural networks. The same feature is also used in the example-level MIA EncoderMI [31] which computed the feature on the embeddings of the given example and its augmented versions, denoted by EncoderMI-T(hreshold). EncoderMI also utilized the set of similarities contributing to the feature and the sorted set (i.e., vector) as features, denoted by EncoderMI-S(et) and EncoderMI-V(ector), respectively. In addition to the pairwise similarity, [30] also utilize the average similarity between the centroid embedding and embeddings contributing to the centroid embedding.

The closest work to ours is FaceAuditor [33] which was available online when we were preparing this manuscript. Different from above attacks that assumed the accessibility of embeddings, FaceAuditor only relies on the recognition score derived from embeddings. It designed different features for different facial recognition networks. For SiameseNet, it utilized the same feature as EncoderMI-V, i.e., the vector of pairwise similarities among the facial images of the target user. For ProtoNet and RelationNet, it utilized the set of scores in which each score is the similarity between one facial image of the target user and the "prototype" of the target user or the "prototype" of other supplemented users.

While these MIA could be applied to SRSs, besides the recognition task, our attack SLMIA-SR differs from them in

[2]A trail is a pair of voices uttered by the same or distinct speakers. When the ground truth is the same (resp. distinct), but SRS concludes the opposite, SRS commits false rejection (resp. false acceptance). EER is the case where the false rejection rate is equal to the false acceptance rate on all trials.

the following aspects: (1) Our threat model is more systematic. They assumed that the adversary has access to either the embeddings or recognition scores but not both, while our attack applies to both (cf. III-B). (2) Our designed features used for membership inference are more comprehensive. We design the features from both intra-closeness and inter-farness, quantified by using two types of similarities, four types of distances, four different statistics, and different arrangements of similarities and distances (cf. IV-A). Through this carefully-established feature engineering, we design in total 103 features which strictly cover the features used in these previous MIA. Our designed features are found to complement each other, leading to a more expressive characterization of the differences between training and non-training speakers. (3) We propose a training strategy to improve attack generalizability. For speaker-/user-level MIA, the ratio of examples of the training speaker/user provided for membership inference that are used in training is unknown and may vary with the target speaker. [30], [33] used only the non-training examples of training speakers/users (i.e., the ratio is 0) to train an attack model, leading to low generalizability to different ratios. Hence, we propose a mixing ratio training strategy to enhance the generalizability (cf. IV-B2). (4) We propose four approaches that are tailored to speaker recognition, which either effectively enhance the attack performance or significantly reduce the number of queries (cf. IV-B3, IV-C, and IV-D). (5) Our attack outperforms the previous MIA for speaker-level MIA on speaker recognition under all the two datasets and five models (cf. V-B1).

MIA on other models. Other MIA targeted conventional classification [10], [15], [11], [16], [12], [17], [13], speech recognition [18], [19], generative [21], [22], regression models [14], etc. These MIA cannot be ported to speaker recognition due to the distinct training paradigm and architecture of speaker recognition. For instance, the training of speech recognition models minimizes Word Error Rate (WER) of the training utterances, so WER is used as a characterization feature. However, the training of SRSs focuses on speaker characteristics rather than the speech text or command. The training of conventional classification models minimizes the cross entropy loss of the training examples, so probability vector, class confidence, and entropy, derived from the final fully connected layer, are used as characterization features. However, SRSs may be trained in the paradigm of using verification-based loss functions without a final fully connected layer. Even if it is trained in a similar manner, the fully connected layer will be dropped after training. Thus, all the above features cannot be used to characterize the differences between training and non-training speakers.

III. Overview of SLMIA-SR

A. Problem Formulation

Assume the target SRS SR^t is trained on a set of training speakers $\mathcal{S}^t_{tr}$ (drawn from some underlying distribution $\mathbb{S}$) and their voices $\mathcal{V}^t_{tr}$. Fix a set of voices $\mathbf{v} = \{v_1, \cdots, v_N\}$ ($N \geq 1$) of a target speaker s. The *speaker-level membership inference attack* is defined as: $\mathcal{A} : SR^t, s, \mathbf{v} \rightarrow \{0, 1\}$, where "1" (resp. "0") means $s \in \mathcal{S}^t_{tr}$ (resp. $s \notin \mathcal{S}^t_{tr}$). Intuitively, a speaker-level MIA takes as input a set of voices from one speaker s and determines whether *any* voices uttered by the speaker s is contained in the training of the model.

We emphasize that for the "member", i.e., $s \in \mathcal{S}^t_{tr}$, we do not require that the inference voices $\mathbf{v}$ are involved in training, i.e., $\mathbf{v} \cap \mathcal{V}^t_{tr}$ may be $\emptyset$. This makes the attack more practical since the inference voices of training speakers are less likely to be used in training. Thus, a speaker-level MIA should be effective even if none of the inference voices of a training speaker has been used for training.

B. Threat Model

Adversarial purpose. The adversary of speaker-level MIA may be users, regulators (e.g., government), SRS developers, and adversarial attackers. Users may want to identify whether their voices have been used for training SRSs without their permission given the wide spread of voices across social media platforms and online meetings. Regulators can check if SRSs are compliant with their published data protection rules, e.g., GDPR [24]. SRS developers can evaluate the privacy risk of their SRSs before publishing in case of being punished or sued due to privacy violations. Adversarial attackers can train a better shadow SRS using common training speakers to improve transferability of adversarial examples [38].

Adversarial capacity. We consider both white- and black-box scenarios with different capacities. The white-box adversary has access to the background model of the target SRS, and thus can obtain the embedding of any given voice. Note that the parameters of the target SRS are still unknown. The black-box adversary only has access to the enrollment and recognition APIs exposed by the target SRS and has to invoke them sequentially to obtain similarity scores. In practice, white-box adversaries (e.g., regulators or SRS developers) are constrained by the number of queries to background models, while box-box adversaries (e.g., users and adversarial attackers) attempt to perform speaker-level MIA using as few queries as possible due to the charge or cost of queries.

Adversarial knowledge. We assume the adversary has an auxiliary speaker dataset $\mathcal{S}^a$ that is sampled from the same distribution $\mathbb{S}$ as the training speaker dataset $\mathcal{S}^t_{tr}$ of the target SRS. In addition, we assume the adversary knows the architecture of the target SRS. These two assumptions follow the standard setting of most previous MIAs [10], [14], [15], [11], [16], [12]. In ablation studies (cf. § VI), we show that our attack remains effective, although the performance may degrade, when these two assumptions are relaxed. However, different from [32] which assumed that the adversary is aware of a dataset in which each data is the non-member of the target SRS, we do not assume such knowledge. The reason is that although this assumption can free the adversary from training shadow SRSs, it also makes the attack less practical.

C. Pipeline of SLMIA-SR

Membership inference is indeed a binary classification task (*member* or *non-member*), so we can build an attack model to do the classification. Following the common pipeline of prior MIAs, the working pipeline of SLMIA-SR is illustrated in Fig. 2, consisting of three stages: *shadow SRS training*, *attack model building*, and *membership inference*. We first briefly describe *feature extractor*, a key module used in two stages.

Feature extractor. During the attack model building (resp. membership inference), feature extractor queries the shadow

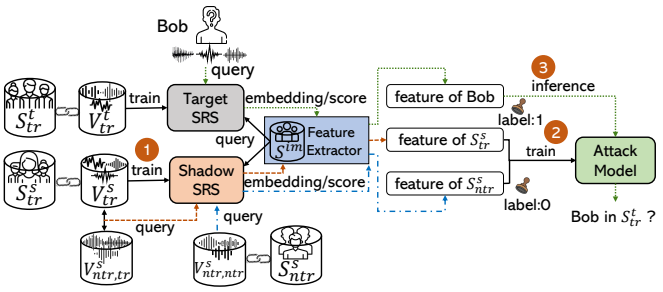


Fig. 2: The working pipeline of SLMIA-SR.

(resp. target) SRS with a set of voices to obtain their outputs (embeddings for white-box scenario and scores for black-box scenario), based on which it produces features for the characterizing differences between training and non-training speakers.

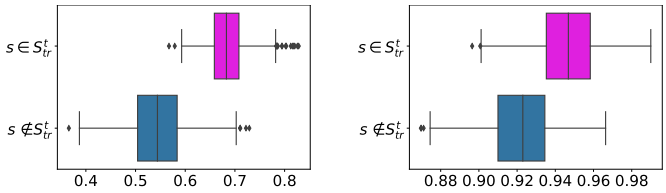
Shadow SRS training. Due to the lack of knowledge about the training and non-training speakers of the target SRS, it is impossible for the adversary to directly build the attack model in a supervised manner. To tackle this problem, we train a shadow SRS as a proxy of the target SRS upon which we build the attack model. We first sample an auxiliary speaker dataset $\mathcal{S}^a$ and their corresponding voices from some underlying distribution $\mathbb{S}$ and partition $\mathcal{S}^a$ into two sets $\mathcal{S}^s$ and $\mathcal{S}^{im}$. The speakers in the set $\mathcal{S}^{im}$, called *imposters*, will be used in attack model building and membership inference. The speaker set $\mathcal{S}^s$ is further partitioned into two sets with the same number of speakers: the set of training speakers ($\mathcal{S}^s_{tr}$) and the set of non-training speakers ($\mathcal{S}^s_{ntr}$). Since the inference voices of a training speaker are not necessarily used in training, we also partition the voices of the training speakers $\mathcal{S}^s_{tr}$ into two sets, $\mathcal{V}^s_{tr}$ and $\mathcal{V}^s_{ntr,tr}$, each of which has the same number of voices per speaker and belongs to "member". Finally, we train a shadow SRS using the training speakers $\mathcal{S}^s_{tr}$ and their voices $\mathcal{V}^s_{tr}$ with some chosen architecture and training algorithm.

Attack model building. To build an attack model, we query the shadow SRS using the voices $\mathcal{V}^s_{tr} \cup \mathcal{V}^s_{ntr,tr}$ of the shadow SRS's training speakers $\mathcal{S}^s_{tr}$ and obtain their outputs (embeddings for white-box scenario and scores for black-box scenario). Note that the voices $\mathcal{V}^s_{tr}$ are used in training the shadow SRS while the voices $\mathcal{V}^s_{ntr,tr}$ are not. The outputs are then fed into the feature extractor to extract the features of the training speakers $\mathcal{S}^s_{tr}$, which are labeled as "member". The same is done for the voices $\mathcal{V}^s_{ntr,ntr}$ of the shadow SRS's the non-training speakers $\mathcal{S}^s_{ntr}$, except that the extracted features are labeled as "non-member". Finally, the features of $\mathcal{S}^s_{tr}$ and $\mathcal{S}^s_{ntr}$ are used to build an attack model (cf. § IV-B1).

Membership inference. To determine the speaker-level membership of a given speaker, we first query the target SRS with the available inference voices of the given speaker, then forward the output of the target SRS to the feature extractor, and finally feed the extracted features to the attack model, which makes a "member" or "non-member" decision.

IV. Methodology of SLMIA-SR

In this section, we first elaborate in detail the feature extractor, then present the attack model, and finally propose an approach to boost the performance of SLMIA-SR when the target speaker provides a limited number of voices as well as two techniques to reduce the number of queries to the target SRS in the black-box scenario.



(a) intra-closeness (b) inter-farness

Fig. 3: The comparison of intra-closeness and inter-farness between training and non-training speakers.

A. Feature Extractor

Feature extractor takes the embeddings or scores of voices as input and outputs features that can effectively characterize the differences between training and non-training speakers. As mentioned in § II-C, existing MIA features are not applicable to SR due to its unique architecture and training paradigm. Hence, we turn to analyze and understand the training objectives of SR. Since SR is used to recognize the identity of individual speakers from their voices, a model is trained with the following two objectives regardless of the training paradigms (classification- and verification-based losses): (1) intra-closeness: the embeddings of two voices from the same speaker are close to each other, and (2) inter-farness: embeddings of two voices from two distinct speakers are far enough from each other. Thus, we hypothesize that training speakers enjoy better intra-closeness and inter-farness than non-training speakers, i.e., the embeddings of a training speaker's voices are closer to each other and farther from the embeddings of the other speakers' voices, than that of a non-training speaker. To give a first impression, we quantify the intra-closeness and inter-farness by the average pairwise cosine *similarity* among the embeddings of the same speaker's voices and by the average of maximal cosine *distance* from distinct speakers' voices, respectively. The box charts of intra-closeness and inter-farness are depicted in Fig. 3a and Fig. 3b, respectively. We can observe a significant statistic difference between the training and non-training speakers, supporting our hypothesis.

Based on the above hypothesis, we design two groups of features: intra-features and inter-features. During the membership inference for a target speaker, the intra-features quantify the similarity of voice embeddings of a target speaker (i.e., intra-closeness), while the inter-features quantify the distance of voice embeddings between the target speaker and other speakers (i.e., inter-farness).

To introduce the designed features, we first define the following function which takes as input two sets of voices ($\{v_1, \cdots, v_m\}$ and $\{v^e_1, \cdots, v^e_n\}$) and produces the similarity between the centroid embeddings of the two sets:

$$\omega\langle v_1, \cdots, v_m | v^e_1, \cdots, v^e_n \rangle = \texttt{sim}\langle \frac{1}{m}\sum_{i=1}^{m} E(v_i), \frac{1}{n}\sum_{i=1}^{n} E(v^e_i) \rangle$$

where $\texttt{sim}\langle a, b\rangle$ is the similarity between two embeddings a and b and $m, n \geq 1$. However, in the black-box scenario, the adversary has no access to embeddings $E(\cdot)$. To solve

this issue, we propose an alternative approach to compute the similarity ω. We first register an enrolled template with voices $v_1^e, \cdots, v_n^e$ in the enrollment phase and then obtain the recognition score $S(v_1|v_1^e, \cdots, v_n^e)$ of the test voice v_1 w.r.t. the enrolled template in the recognition phase (cf. § II-A). Then $\omega\langle v_1|v_1^e, \cdots, v_n^e\rangle$ is computed as $S(v_1|v_1^e, \cdots, v_n^e)$. Note that m should be 1 in the black-box scenario, since an SRS accepts only one voice per query in the recognition phase.

1) Intra-Features: We design intra-features by leveraging centroid-based similarity and pairwise similarity. Fix a set of voices $\{v_1, \cdots, v_N\}$ of the target speaker and let $[n]$ denote the set $\{1, \cdots, n\}$ for an integer n.

Centroid-based similarity measures the closeness between the embeddings and their centroid, namely, the set of centroid-based similarities is $\mathtt{F_c} = \{\omega\langle v_i|v_1, \cdots, v_N\rangle \mid i \in [N]\}$.

Pairwise similarity measures the closeness between each pair of embeddings. The set of pairwise similarities is defined as $\mathtt{F_p} = \{\omega\langle v_j|v_i\rangle \mid i, j \in [N], i < j\}$. The set $\mathtt{F_p}$ can be refined as follows: for each voice v_i, we first compute the similarities between v_i and other voices as $\mathtt{F}^i_{\tilde{\mathtt{p}}} = \{\omega\langle v_j|v_i\rangle \mid j \in [N], j \neq i\}$, then compute the statistic of $\mathtt{F}^i_{\tilde{\mathtt{p}}}$ as $\mathtt{stat}(\mathtt{F}^i_{\tilde{\mathtt{p}}})$, and finally define the refined set of similarities as $\mathtt{F}_{\tilde{\mathtt{p}}} = \{\mathtt{stat}(\mathtt{F}^i_{\tilde{\mathtt{p}}}) \mid i \in [N]\}$. We will instantiate `stat` by average, negative standard derivative, maximum, and minimum statistics, leading to the sets $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{avg}}$, $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{std}}$, $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{max}}$, and $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{min}}$, respectively.

The intra-features are defined as the statistics (i.e., average, negative standard derivative, maximum, and minimum) of the above six sets of similarities ($\mathtt{F_c}$, $\mathtt{F_p}$, $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{avg}}$, $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{std}}$, $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{max}}$, and $\mathtt{F}_{\tilde{\mathtt{p}},\mathtt{min}}$). Let $\Theta^{\mathtt{y}}_{\mathtt{x}}$ denote the intra-feature which is defined as the statistic y of the set $\mathtt{F_x}$. After de-duplicating three pairs of equivalent intra-features ($\Theta^{\mathtt{avg}}_{\mathtt{p}}$ and $\Theta^{\mathtt{avg}}_{\tilde{\mathtt{p}},\mathtt{avg}}$, $\Theta^{\mathtt{max}}_{\mathtt{p}}$ and $\Theta^{\mathtt{max}}_{\tilde{\mathtt{p}},\mathtt{max}}$, $\Theta^{\mathtt{min}}_{\mathtt{p}}$ and $\Theta^{\mathtt{min}}_{\tilde{\mathtt{p}},\mathtt{min}}$), there are $(6 \times 4 - 3 = 21)$ unique intra-features. Note that we use negative standard derivative instead of standard derivative since we expect the features of training speakers to be larger than that of non-training speakers.

2) Inter-Features: Inter-features make use of an additional set of M imposters $\mathcal{S}^{im} = \{s_1^{im}, \cdots, s_M^{im}\}$ and their voices $\mathcal{V}^{im} = \bigcup_{j=1}^{M} \mathcal{V}_j^{im}$, where $\mathcal{V}_j^{im} = \{v_1^{im,j}, \cdots, v_{K_j}^{im,j}\}$ is the set of voices of the imposter s_j^{im}. Also let $\mathcal{V}^{im} = \{v_1^{im} \cdots, v_Q^{im}\}$ with $Q = \sum_{j=1}^{M} K_j$. We design four types of distances.

Centroid-centroid distance measures the distance between the centroid of the voice embeddings of the target speaker and the centroid of the voice embeddings of all the imposters. Thus, we define the set of distances $\mathtt{F_{cc}} = \{-\omega\langle v_1^{im,j}, \cdots, v_{K_j}^{im,j}|v_1, \cdots, v_N\rangle \mid j \in [M]\}$. Recall that K_j should be 1 for this distance in the black-box scenario. In § IV-D, we propose an enrollment voice concatenation technique, allowing the adversary to obtain *one* concatenated and longer voice for both the target speaker and each imposter with multiple voices (i.e., $K_j \geq 1$). It does not lead to obvious performance gap between the two scenarios (cf. Appendix B).

Centroid-voice distance measures the distance between the centroid of the voice embeddings of the target speaker and the voice embeddings of all the imposters. The set of distances is defined as $\mathtt{F_{cv}} = \{-\omega\langle v^{im}|v_1, \cdots, v_N\rangle \mid v^{im} \in \mathcal{V}^{im}\}$. The set $\mathtt{F_{cv}}$ can be refined as follows: for each imposter s_j^{im}, we first compute the set of distances between the centroid of the voice embeddings of the target speaker and the voice embeddings of this imposter, i.e., $\mathtt{F}^j_{\mathtt{c}\tilde{\mathtt{v}}} = \{-\omega\langle v^{im}|v_1, \cdots, v_N\rangle \mid v^{im} \in \mathcal{V}_j^{im}\}$, then compute the statistic of $\mathtt{F}^j_{\mathtt{c}\tilde{\mathtt{v}}}$ as $\mathtt{stat}(\mathtt{F}^j_{\mathtt{c}\tilde{\mathtt{v}}})$, and finally define the refined set as $\mathtt{F}_{\mathtt{c}\tilde{\mathtt{v}}} = \{\mathtt{stat}(\mathtt{F}^j_{\mathtt{c}\tilde{\mathtt{v}}}) \mid j \in [M]\}$.

Voice-centroid distance measures the distance between the voice embeddings of the target speaker and the centroid of the voice embeddings of all the imposters. The set of distances is $\mathtt{F_{vc}} = \{-\omega\langle v_i|v_1^{im,j}, \cdots, v_{K_j}^{im,j}\rangle \mid i \in [N], j \in [M]\}$. The set $\mathtt{F_{vc}}$ can also be refined in two ways: (i) for each voice v_i of the target speaker, we first compute the distance between the embedding of v_i and the centroid of the voice embeddings of all the imposters as $\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{c}} = \{-\omega\langle v_i|v_1^{im,j}, \cdots, v_{K_j}^{im,j}\rangle \mid j \in [M]\}$, then compute the statistic of $\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{c}}$ as $\mathtt{stat}(\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{c}})$, and finally define the refined set $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{c}} = \{\mathtt{stat}(\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{c}}) \mid i \in [N]\}$; (ii) for each imposter s_j^{im}, we first compute the distance between the centroid of voice embeddings of this imposter and the voice embeddings of the target speaker, i.e., $\mathtt{F}^j_{\mathtt{v}\tilde{\mathtt{c}}} = \{-\omega\langle v_i|v_1^{im,j}, \cdots, v_{K_j}^{im,j}\rangle \mid i \in [N]\}$, then compute the statistic of $\mathtt{F}^j_{\mathtt{v}\tilde{\mathtt{c}}}$ as $\mathtt{stat}(\mathtt{F}^j_{\mathtt{v}\tilde{\mathtt{c}}})$, and finally define the refined set $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{c}}} = \{\mathtt{stat}(\mathtt{F}^j_{\mathtt{v}\tilde{\mathtt{c}}}) \mid j \in [M]\}$.

Voice-voice distance measures the distance between the voice embeddings of the target speaker and the voice embeddings of all the imposters. Formally, the set of distances is defined as $\mathtt{F_{vv}} = \{-\omega\langle v_i|v^{im}\rangle \mid i \in [N], v^{im} \in \mathcal{V}^{im}\}$. Similarly, two refined sets of distances can be defined accordingly, namely, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{v}} = \{\mathtt{stat}(\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{v}}) \mid i \in [N]\}$ and $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{v}}} = \{\mathtt{stat}(\mathtt{F}^k_{\mathtt{v}\tilde{\mathtt{v}}}) \mid k \in [Q]\}$ where $\mathtt{F}^i_{\tilde{\mathtt{v}}\mathtt{v}} = \{-\omega\langle v_i|v^{im}\rangle \mid v^{im} \in \mathcal{V}^{im}\}$ and $\mathtt{F}^k_{\mathtt{v}\tilde{\mathtt{v}}} = \{-\omega\langle v_i|v_k^{im}\rangle \mid i \in [N]\}$.

The same to intra-features, we instantiate `stat` by average, negative standard derivative, maximum, and minimum statistics, leading to 24 sets of distances, namely, $\mathtt{F_{cc}}$, $\mathtt{F_{cv}}$, $\mathtt{F}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{avg}}$, $\mathtt{F}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{std}}$, $\mathtt{F}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{max}}$, $\mathtt{F}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{min}}$, $\mathtt{F_{vc}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{avg}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{std}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{max}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{min}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{avg}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{std}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{max}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{min}}$, $\mathtt{F_{vv}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{avg}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{std}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{max}}$, $\mathtt{F}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{min}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{avg}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{std}}$, $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{max}}$, and $\mathtt{F}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{min}}$.

The inter-features are defined as the statistics of the above 24 sets of distances. Let $\Phi^{\mathtt{y}}_{\mathtt{x}}$ denotes the inter-feature which is defined as the statistic y of the set $\mathtt{F_x}$. After de-duplicating two pairs and six triples of equivalent features ($\Phi^{\mathtt{max}}_{\mathtt{cv}}$ and $\Phi^{\mathtt{max}}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{max}}$, $\Phi^{\mathtt{min}}_{\mathtt{cv}}$ and $\Phi^{\mathtt{min}}_{\mathtt{c}\tilde{\mathtt{v}},\mathtt{min}}$, $\Phi^{\mathtt{avg}}_{\mathtt{vc}}$ and $\Phi^{\mathtt{avg}}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{avg}}$ and $\Phi^{\mathtt{avg}}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{avg}}$, $\Phi^{\mathtt{max}}_{\mathtt{vc}}$ and $\Phi^{\mathtt{max}}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{max}}$ and $\Phi^{\mathtt{max}}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{max}}$, $\Phi^{\mathtt{min}}_{\mathtt{vc}}$ and $\Phi^{\mathtt{min}}_{\tilde{\mathtt{v}}\mathtt{c},\mathtt{min}}$ and $\Phi^{\mathtt{min}}_{\mathtt{v}\tilde{\mathtt{c}},\mathtt{min}}$, $\Phi^{\mathtt{avg}}_{\mathtt{vv}}$ and $\Phi^{\mathtt{avg}}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{avg}}$ and $\Phi^{\mathtt{avg}}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{avg}}$, $\Phi^{\mathtt{max}}_{\mathtt{vv}}$ and $\Phi^{\mathtt{max}}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{max}}$ and $\Phi^{\mathtt{max}}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{max}}$, $\Phi^{\mathtt{min}}_{\mathtt{vv}}$ and $\Phi^{\mathtt{min}}_{\tilde{\mathtt{v}}\mathtt{v},\mathtt{min}}$ and $\Phi^{\mathtt{min}}_{\mathtt{v}\tilde{\mathtt{v}},\mathtt{min}}$), there are $(24 \times 4 - 2 \times 1 - 6 \times 2 = 82)$ unique inter-features.

We remark that inter-features can exploit a set of imposters which is fully controllable by the adversary, thus are applicable when the target speaker only has one voice (i.e., $N = 1$). In contrast, intra-features requires $N \geq 2$.

3) Property of Feature Extractor: Here we discuss some notable properties of the proposed feature extractor.

Unity under different scenarios. The features defined above work in both white- and black-box scenarios, only differing in the way of obtaining the similarity ω. Consequently, in the white-box scenario, the embedding $E(v)$ of each voice v can be computed by querying the background model once, based on which all the intra- and inter-features can be computed. Thus, the total number of queries is $N + Q$. However, in

TABLE II: The number of queries in the black-box scenario.

Feature		Feature Computation	#Query		
			Enrollment	Recognition	Total
Intra	Centroid-based	Baseline	N	N	2N
		Concat	1	N	1+N
	Pairwise	Baseline	N-1	N(N-1)/2	(N-1)(N+2)/2
Inter	Centroid-centroid	Baseline	N	M	N+M
		Concat	1	M	1+M
	Centroid-voice	Baseline	N	Q	N+Q
		Concat	1	Q	1+Q
	Voice-centroid	Baseline	Q	NM	Q+NM
		Group	Q	N	Q+N
		Concat	M	NM	M+NM
		Group+Concat	M	N	M+N
	Voice-voice	W_1	N	QN	(1+Q)N
		W_2	Q	NQ	(1+N)Q

Note: (i) "Concat" and "Group" are short for "Enrollment Voice Concatenation" and "Group Enrollment", respectively. (ii) Baseline does not utilize the two techniques. (iii) For voice-voice inter-features, there are two ways of enrollment and recognition (W_1 and W_2), the adversary can choose the one with the less number of queries. W_1 use the voices of the target speaker to register enrolled templates in the enrollment phase and regards the voices of imposters as test voices in the recognition phase, which is opposite to W_2.

the black-box scenario, as shown in TABLE II (Baseline), the enrollment and recognition APIs have to be queried multiple times to obtain the required scores for computing a feature, although these scores can be reused for computing the features in the same group. To reduce the number of queries to the target SRS, we will propose two techniques, group enrollment and enrollment voice concatenation (cf. § IV-D). The reduced numbers of queries are also shown in TABLE II.

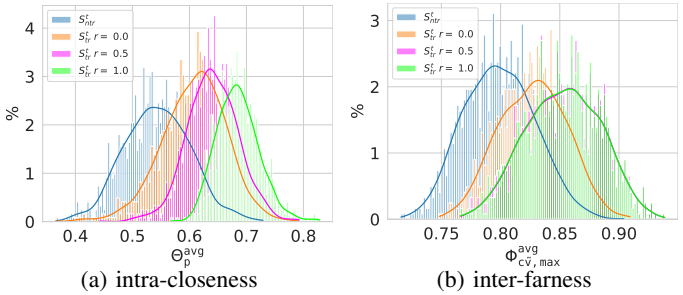
Comprehensiveness. In total, we design 103 features from two different aspects (intra-closeness and inter-farness) using two types of similarities, four types of distances, four statistics, and different arrangements of similarities or distances (original v.s. refined sets of similarities or distances), aiming to effectively quantify the differences between training and training speakers from different perspectives in a complementary way.

B. Attack Model

1) Attack Model Configuration: To utilize multiple features, SLMIA-SR adopts a classifier-based attack model. Specifically, we train a binary classifier f as the attack model in a supervised fashion based on the shadow SRS using the features of the voices of the training speakers $\mathcal{S}^s_{tr}$ and non-training speakers $\mathcal{S}^s_{ntr}$. Then, $\mathcal{A}(SR^t, \mathbf{v}, s)$ is implemented as $\mathbb{I}[f(\Psi(\mathbf{v})) > 0.5]$ where $\Psi(\mathbf{v})$ denotes the features of the inference voices $\mathbf{v}$ of a target speaker produced by the feature extractor via querying the target SRS SR^t and f gives the probability that the speaker s is one of the training speakers of the target SRS SR^t.

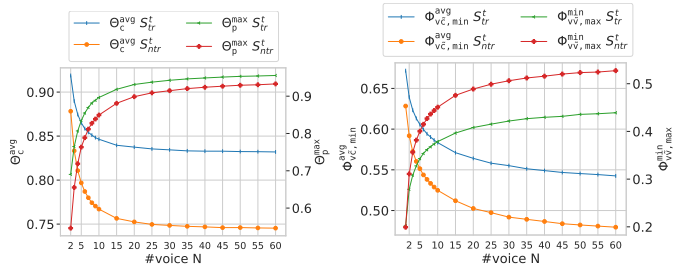
To demonstrate the complementarity of features, we also include a threshold-based attack model in our experiments, which makes decisions by thresholding a single feature. The attack $\mathcal{A}(SR^t, \mathbf{v}, s)$ is implemented as $\mathbb{I}[\psi(\mathbf{v}) > \tau]$ where $\mathbb{I}$ is the indicator function, τ is a threshold tuned on the shadow SRS by maximizing the classification accuracy between the training speakers $\mathcal{S}^s_{tr}$ and non-training speakers $\mathcal{S}^s_{ntr}$, and $\psi(\mathbf{v})$ is the feature of the inference voices $\mathbf{v}$ produced by the feature extractor via querying the target SRS SR^t. Namely, a speaker s is regarded as a training speaker of SR^t if the feature of the provided voices $\mathbf{v}$ is larger than the threshold τ.

2) Attack Model Generalization: For a target speaker s, let r denote the ratio of the target speaker's inference voices $\mathbf{v}$ that are included in the training of the target SRS. Since the ratio r is unknown in practice, it is expected that a speaker-level MIA should be effective for unknown r even if $r = 0$, i.e., none of the inference voices is used in the training of the target SRS. However, as depicted in Fig. 4, the distribution of features of training speakers varies with the ratio r. Thus, the attack model trained with one fixed r is more likely to generalize poorly on other r. To address this issue, we propose a *mixing ratio training strategy* which utilizes features of the training voices $\mathcal{V}^s_{tr}$ ($r = 1$) and non-training voices $\mathcal{V}^s_{ntr,tr}$ ($r = 0$) of the training speakers as "member", and the non-training speakers' voices $\mathcal{V}^s_{ntr,ntr}$ as "non-member", to train the attack model. Though the attack model is only trained with $r = 1$ and $r = 0$, it generalizes for membership inference with different r. We could include more diverse r when training the attack model. However, it will introduce more overhead to compute the features for different r and train the attack model on a larger dataset, thus not considered in this work.



(a) intra-closeness (b) inter-farness

Fig. 4: Comparison of features with different r.



(a) intra-closeness (b) inter-farness

Fig. 5: Comparison of features with different N.

3) Voice-Number-Dependent Attack Model: As shown in Fig. 5, features vary significantly with the number N of the inference voices, building one voice-number-independent (VNID) attack model will achieve sub-optimal performance. Thus, we propose to build a voice-number-dependent (VND) attack model for each number N. However, there is no upper bound of N, and it is impossible to build infinite attack models. Based on the observation that features converge after some N' in Fig. 5, we can set an upper bound N' and build N' VND attack models. During membership inference, the adversary simply discards some voices when $N > N'$. Note that this bound can be decided by the adversary on the shadow SRS with the auxiliary dataset, and we propose a T-test [49] based algorithm in Alg. 1 in Appendix C.

C. Voice Chunk Splitting

The performance of an attack model decreases with the decrease in the number of inference voices, because features become less precise. It is possible to create a sufficient number of inference voices for some adversaries (e.g., users), but may not

TABLE III: Details of dataset partition

Dataset		Total	Imposters	Shadow / Target SRS		
			Speakers $\mathcal{S}^{im}$	Training speakers $\mathcal{S}^s_{tr}$ / $\mathcal{S}^t_{tr}$		Non-training speakers $\mathcal{S}^s_{ntr}$ / $\mathcal{S}^t_{ntr}$
			Voices $\mathcal{V}^{im}$	Training Voices $\mathcal{V}^s_{tr}$ / $\mathcal{V}^t_{tr}$	Non-training voices $\mathcal{V}^s_{ntr,tr}$ / $\mathcal{V}^t_{ntr,tr}$	Non-training voices $\mathcal{V}^s_{ntr,ntr}$ / $\mathcal{V}^t_{ntr,ntr}$
VoxCeleb-2	#speakers	6112	1222	1222 / 1222	1222 / 1222	1222 / 1222
	#voices	>1 million	101,572	113,617 / 115,780	113,055 / 115,212	111,083 / 108,879
LibriSpeech	#speakers	2484	400	521 / 521	521 / 521	521 / 521
	#voices	≈ 0.3 million	46,140	30,928 / 30,734	30,648 / 30,746	30,892 / 30,992

Note: (i) For each dataset, we first partition the speakers into five approximately equal and *disjoint* parts, denoted by $\mathcal{S}^{im}$, $\mathcal{S}^s_{tr}$, $\mathcal{S}^s_{ntr}$, and $\mathcal{S}^t_{tr}$, and $\mathcal{S}^t_{ntr}$, respectively. (ii) $\mathcal{S}^{im}$ contains the imposters used to compute inter-features. (iii) $\mathcal{S}^s_{tr}$ and $\mathcal{S}^s_{ntr}$ are the training and non-training speakers for the shadow SRS. $\mathcal{S}^t_{tr}$ and $\mathcal{S}^t_{ntr}$ are the training and non-training speakers for the target SRS. (iv) For each speaker in $\mathcal{S}^s_{tr}$, we partition his/her voices into two nearly equal and *disjoint* parts: $\mathcal{V}^s_{tr}$ and $\mathcal{V}^s_{ntr,tr}$. The same is applied to $\mathcal{S}^t_{tr}$, leading to $\mathcal{V}^t_{tr}$ and $\mathcal{V}^t_{ntr,tr}$. The shadow SRS and target SRS are trained on $\mathcal{V}^s_{tr}$ and $\mathcal{V}^t_{tr}$, respectively. (v) The training dataset of the attack model is derived from $\mathcal{V}^s_{tr}$ (label "member"), $\mathcal{V}^s_{ntr,tr}$ (label "member"), and $\mathcal{S}^s_{ntr}$ (label "non-member"). (vi) The testing dataset used to evaluate the membership inference attack includes $\mathcal{V}^t_{tr}$ (label "member"), $\mathcal{V}^t_{ntr,tr}$ (label "member"), and $\mathcal{S}^t_{ntr}$ (label "non-member").

be feasible for other adversaries (e.g., regulators). A straightforward solution to address this issue is to create augmented voices by applying voice augmentation (e.g., Gaussian white noise). However, we find that this solution cannot effectively improve and even sometimes worsen the performance, since the target SRS does not necessarily utilize augmentation during training. Instead, we propose an approach, named voice chunk splitting, based on the observation that due to diverse voice duration, training voices are always divided into short segments with fixed duration to form a training batch. We apply a sliding window with certain window size w and window step s to split a voice into multiple overlapped chunks, thus increasing the number of voices and improving the precision of features. We note that voice chunk splitting is applied in both the attack model building and the membership inference.

D. Reducing Queries in the Black-Box Scenario

The black-box adversary often attempts to perform membership inference using as few queries as possible. We present two techniques to reduce the number of queries, thus reducing testing time and resources. The comparison of the number of queries with/without the two techniques is shown in TABLE II.

Group enrollment. Speaker identification allows multiple speakers to be enrolled each of which has an enrolled template, and a vector of the scores w.r.t. these enrolled templates can be obtained via *one* recognition query. In contrast, speaker verification allows only one speaker to be enrolled, so to obtain the scores of multiple speakers, these speakers have to be enrolled individually and multiple recognition queries are required. To reduce the number of recognition queries, we utilize speaker identification to compute the voice-centroid distance based on inter-features. Specifically, after enrolling all the imposters, we compute the scores of a voice w.r.t. all the imposters' enrolled templates via one query. It reduces the number of recognition queries from $N \times M$ to N.

Enrollment voice concatenation. To compute the features that rely on the centroid embedding of the voices $\mathbf{a} = \{a_1, \cdots, a_p\}$ of a speaker, the enrolled template of the speaker has to be registered via p enrollment queries. Observing that the embedding is extracted from the *frame-wise* acoustic feature[3], we conjecture that the centroid embedding e_c computed as the average of embeddings of the voices $\mathbf{a}$ is similar to the embedding $\bar{e}_c$ of the concatenation $\texttt{concat}(\mathbf{a})$ of the voices $\mathbf{a}$, for which one enrollment query is sufficient. Fig. 6 shows the cosine similarity between e_c and $\bar{e}_c$, which is very close to 1.0 and much higher than the similarity between the centroid embedding and embeddings of the voices $\mathbf{a}$, regardless of the number N of enrollment voices, confirming our conjecture. Hence, we apply enrollment voice concatenation to reduce the number of enrollment queries when computing centroid-related intra-features and inter-features. Specifically, we register an enrolled template using the voice $\texttt{concat}(\mathbf{a})$, and approximate the recognition score of a voice v w.r.t. the centroid embedding $S(v|a_1, \cdots, a_p)$ by the one w.r.t. $S(v|\texttt{concat}(\mathbf{a}))$. This reduces the number of enrollment queries from p to 1.

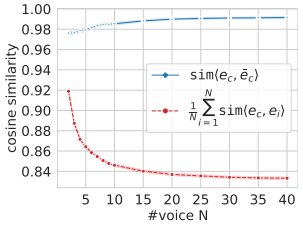


Fig. 6: The similarity between e_c and $\bar{e}_c$ varying N.

V. Evaluation of SLMIA-SR

A. Experimental Setting

Datasets. We use two widely-used voice datasets: VoxCeleb-2 [50] and LibriSpeech [51]. VoxCeleb-2 contains more than 1 million voices from 6,112 speakers, while LibriSpeech contains approximately 0.3 million voices from 2,484 speakers. The partition of these datasets is summarized in TABLE III.

SRSs. We use five SRSs: LSTM-GE2E, TDNN-CE, Raw-AAM, Res-AP, and VGG-GE2E. These SRSs are the combinations of five model architectures and four training losses, as shown in TABLE IV. We trained them in their default settings with the number of training epochs ranging from 50 to 1500 depending on the number of trainable parameters.

TABLE IV: The information and performance of SRSs

Name	Archi	Training		Performance in terms of EER			
				VoxCeleb-2		LibriSpeech	
		paradigm	loss	Training	Testing	Training	Testing
LSTM-GE2E	LSTM [29], [52]	verification	GE2E [29]	9.66%	18.19%	0.31%	8.08%
TDNN-CE	TDNN [53], [54]	classification	CE [6]	1.43%	6.70%	0.36%	2.39%
Raw-AAM	RawNet3 [55], [56]	classification	AAM [35]	2.5%	6.23%	0.20%	2.56%
Res-AP	ResNetSE34V2 [57], [56]	verification	AP [28]	2.11%	6.70%	0.15%	6.84%
VGG-GE2E	VGGVox40 [58], [56]	verification	GE2E [29]	4.18%	9.43%	0.24%	5.66%

Note: Training EER and testing EER are calculated on 4,000 randomly chosen trials from $\mathcal{V}^t_{tr}$ and $\mathcal{V}^t_{ntr,ntr}$, respectively.

Classifier-based attack model. The classifier-based attack model is a multilayer perceptron with one hidden layer comprising 64 neurons and ReLU as the activation function. It is trained by the Adam optimizer with 1e-3 learning rate for 1,000 epochs. We utilize the cosine similarity to measure the similarity between two embeddings. To reduce randomness, the training is repeated independently ten times and the average results are reported.

[3]Due to the time-varying non-stationary property, voices are not resilient enough to noises and other variations, and waveforms fail to effectively represent speaker characteristics. Hence, to achieve better performance, a raw voice is often transformed into a two-dimensional time-frequency representation via frequency analysis, called acoustic feature, where one coordinate at the time-axis represents a frame, a short segment of the voice.

TABLE V: The effectiveness of SLMIA-SR in the Setting-1.

		Accuracy				AUROC				TPR @ x% FPR				TPR @ 1% FPR			
		VoxCeleb-2		LibriSpeech		VoxCeleb-2		LibriSpeech		VoxCeleb-2 (x=0.1)		LibriSpeech (x=0.2)		VoxCeleb-2		LibriSpeech	
		r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0
LSTM-GE2E	**LRL-MIA**	0.938	0.698	0.934	0.895	0.98	0.789	0.979	0.952	14.3%	1.5%	48.9%	31.7%	51.0%	8.1%	59.9%	41.7%
	EncoderMI-T	0.927	0.7	0.925	0.877	0.98	0.792	0.979	0.952	15.0%	1.6%	48.9%	31.7%	52.0%	7.8%	59.9%	41.7%
	TLK-MIA	0.927	0.7	0.925	0.877	0.98	0.792	0.979	0.952	15.0%	1.6%	48.9%	31.7%	52.0%	7.8%	59.9%	41.7%
	SLMIA-SR	**0.962**	**0.894**	**0.979**	**0.974**	**0.998**	**0.958**	**0.995**	**0.994**	**84.1%**	**33.5%**	**68.6%**	**66.5%**	**96.1%**	**58.7%**	**86.8%**	**83.5%**
TDNN-CE	**LRL-MIA**	0.963	0.82	0.748	0.723	0.998	0.906	0.814	0.791	69.7%	20.1%	0.6%	0.6%	97.0%	46.6%	7.5%	6.7%
	EncoderMI-T	**0.983**	0.779	0.732	0.713	0.998	0.904	0.814	0.791	71.0%	20.8%	0.6%	0.6%	96.9%	45.9%	7.7%	6.7%
	TLK-MIA	**0.983**	0.779	0.732	0.713	0.998	0.904	0.814	0.791	71.0%	20.8%	0.6%	0.6%	96.9%	45.9%	7.7%	6.7%
	SLMIA-SR	0.972	**0.891**	**0.85**	**0.83**	**1.0**	**0.965**	**0.914**	**0.897**	**97.4%**	**33.8%**	**12.2%**	**11.1%**	**99.8%**	**64.4%**	**23.2%**	**21.9%**
Raw-AAM	**LRL-MIA**	0.875	0.705	0.693	0.679	0.941	0.786	0.753	0.732	4.6%	1.6%	0.6%	0.8%	23.5%	9.6%	3.1%	2.7%
	EncoderMI-T	0.868	0.703	0.675	0.661	0.939	0.785	0.753	0.732	5.1%	1.9%	0.6%	0.8%	22.9%	9.8%	3.1%	2.7%
	TLK-MIA	0.868	0.703	0.675	0.661	0.939	0.785	0.753	0.732	5.1%	1.9%	0.6%	0.8%	22.9%	9.8%	3.1%	2.7%
	SLMIA-SR	**0.936**	**0.749**	**0.819**	**0.783**	**0.99**	**0.856**	**0.889**	**0.856**	**60.6%**	**5.6%**	**9.7%**	**6.8%**	**82.2%**	**18.3%**	**15.3%**	**12.7%**
Res-AP	**LRL-MIA**	0.915	0.756	0.948	0.924	0.975	0.842	0.985	0.974	26.7%	8.8%	5.9%	6.6%	60.5%	24.4%	72.0%	64.5%
	EncoderMI-T	**0.923**	0.747	0.921	0.887	0.974	0.841	0.985	0.974	27.0%	8.8%	5.9%	6.6%	59.8%	24.3%	72.0%	64.7%
	TLK-MIA	**0.923**	0.747	0.921	0.887	0.974	0.841	0.985	0.974	27.0%	8.8%	5.9%	6.6%	59.8%	24.3%	72.0%	64.7%
	SLMIA-SR	0.919	**0.799**	**0.964**	**0.956**	**0.982**	**0.892**	**0.989**	**0.986**	**35.2%**	**12.5%**	**19.6%**	**14.4%**	**78.5%**	**40.2%**	**76.0%**	**72.3%**
VGG-GE2E	**LRL-MIA**	0.828	0.714	0.879	0.847	0.908	0.783	0.944	0.916	16.7%	5.6%	14.8%	9.8%	35.7%	17.2%	22.0%	15.4%
	EncoderMI-T	0.821	0.711	0.869	0.827	0.908	0.785	0.944	0.916	16.4%	5.5%	15.0%	9.8%	36.2%	17.4%	22.1%	15.4%
	TLK-MIA	0.821	0.711	0.869	0.827	0.908	0.785	0.944	0.916	16.4%	5.5%	15.0%	9.8%	36.2%	17.4%	22.1%	15.4%
	SLMIA-SR	**0.854**	**0.743**	**0.945**	**0.914**	**0.931**	**0.835**	**0.983**	**0.968**	**30.7%**	**16.6%**	**29.5%**	**22.1%**	**45.7%**	**26.4%**	**57.4%**	**45.9%**

Evaluation metrics. Following most prior work on MIA, we adopt two aggregate metrics: accuracy and AUROC. Since it is highlighted in [11] that a practical MIA should yield a high True Positive Rate (TPR) at sufficiently low False Positive Rate (FPR), we also consider TPR at 0.1% (for VoxCeleb-2), 0.2% (for LibriSpeech, each subset of which contains less than 1,000 speakers), and 1% FPR. When calculating these metrics, the number of training speakers is set to be the same as that of non-training speakers, as accuracy is highly sensitive to the ratio between the numbers of positive and negative examples.

Baselines. We consider recent promising baselines designed for embedding models: Li et al. [30] (named LRL-MIA), Tseng et al. [32] (named TKL-MIA), EncoderMI [31], and FaceAuditor [33], where EncoderMI has two variants EncoderMI-T and EncoderMI-V (the lest effective variant EncoderMI-S is not considered), and FaceAuditor has two variants FaceAuditor-S and FaceAuditor-P/R. Thus, there are six baselines. Details of these attacks refer to § II-C and Appendix A.

Experimental designs. Following the setting in recent works [11], [33], [12], [13], we will first assume in § V that the adversary can adopt the same architecture as the target SRS for the shadow SRS and owns an auxiliary dataset following the same distribution as the target SRS's training dataset, and then conduct ablation studies to investigate the effect of the dataset distribution shift and the architecture shift in § VI.

B. Experimental Results

1) Overall Performance: We evaluate the effectiveness of SLMIA-SR by comparing with the baselines in two settings:

Setting-1. We use all the voices of a target speaker in either $\mathcal{S}^t_{tr}$ or $\mathcal{S}^t_{ntr}$ for computing features, simulating the scenario where the target speaker provides sufficient voices for MIA. Also, all the imposters in $\mathcal{S}^{im}$ and all their voices in $\mathcal{V}^{im}$ are used to compute the inter-features, assuming the number of queries to the target SRS is unconstrained (e.g., in the white-box scenario). We do not build voice-number-dependent attack models in this setting since each target speaker has a different number of voices, and do not utilize the voice chunk splitting strategy since the number of voices is sufficient.

Setting-2. We use a voice-number-dependent (VND) attack-model, randomly choose 10 voices per target speaker, 20 imposters, and 10 voices per imposter to compute the inter-features, and apply the voice chunk splitting strategy.

SLMIA-SR utilizes the mixing ratio training strategy in both settings. We compare with LRL-MIA, EncoderMI-T, and TKL-MIA in Setting-1, because the other three baselines use classifier-based attack models requiring fixed number of voices per target speaker and fixed number of imposters, thus only compared in Setting-2. Notice that the following results apply for both white-box and black-box scenarios, due to the unity of the feature extractor (cf. IV-A3).

Results in Setting-1. The results are reported in TABLE V. Overall, SLMIA-SR achieves the best performance across the five target SRSs and the two datasets in terms of all the four metrics, e.g., 83.5%-99.4% AUROC and 5.6%-33.8% TPR at 0.1% FPR when $r = 0$, i.e., all the voices used for membership inference of training speakers are different from their training voices. SLMIA-SR performs even better when $r = 1$.

We observe that the performance of SLMIA-SR varies with the target SRSs and datasets. TDNN-CE and Raw-AAM are generally the least vulnerable to SLMIA-SR, in particular, on the dataset LibriSpeech. It is because they were trained with the classification-based loss functions CE and AAM, respectively, while the CE loss do not explicitly constrain the intra-closeness and the constraint of AAM may be weaker than that of verification-based losses. Regarding datasets, SLMIA-SR generally performs better on LibriSpeech than on VoxCeleb-2 when $r = 0$. It is probably because LibriSpeech contains fewer speakers than VoxCeleb-2, hence the target SRSs trained with LibriSpeech are more likely to memorize the training speakers.

Comparing with the baselines, SLMIA-SR outperforms them on all target SRSs and all datasets in terms of all the metrics. For example, on the target SRS LSTM-GE2E and the dataset VoxCeleb-2, when $r = 0$, SLMIA-SR improves the Accuracy, AUROC, TPR at 0.1% FPR, TPR at 1% FPR by 19.4%, 16.6%, 31.9%, and 50.6%, respectively, compared to the most effective baseline. The improvement mainly comes from two aspects. First, SLMIA-SR utilizes features that characterize both intra-closeness and inter-farness, while these baselines only consider intra-closeness. Second, SLMIA-SR utilizes much more features and these features driven by carefully-established feature engineering are comprehensive and complementary, providing a better characterization of the differences between training and non-training speakers.

TABLE VI: The effectiveness of SLMIA-SR in Setting-2.

		Accuracy		AUROC		TPR @ x% FPR			
						x=0.1	x=0.2	x=1	x=1
		VC-2	LS	VC-2	LS	VC-2	LS	VC-2	LS
LSTM-GE2E	**EncoderMI-V**	0.649	0.866	0.72	0.932	2.0%	19.2%	6.6%	35.2%
	FaceAuditor-S	0.655	0.842	0.714	0.932	1.2%	16.8%	5.3%	33.0%
	FaceAuditor-P/R	0.614	0.768	0.691	0.863	1.6%	3.8%	6.2%	14.5%
	SLMIA-SR	**0.785**	**0.976**	**0.861**	**0.994**	**7.2%**	**62.1%**	**24.4%**	**82.7%**
TDNN-CE	**EncoderMI-V**	0.724	0.703	0.81	0.772	19.6%	1.1%	28.2%	5.6%
	FaceAuditor-S	0.784	0.666	0.866	0.742	20.1%	2.1%	34.2%	4.9%
	FaceAuditor-P/R	0.772	0.578	0.866	0.628	11.9%	0.3%	30.9%	1.4%
	SLMIA-SR	**0.839**	**0.773**	**0.92**	**0.856**	**23.6%**	**2.8%**	**42.9%**	**8.1%**
Raw-AAM	**EncoderMI-V**	0.657	0.657	0.709	0.708	2.7%	0.9%	8.7%	2.0%
	FaceAuditor-S	0.636	0.64	0.686	0.702	0.2%	0.2%	2.3%	2.0%
	FaceAuditor-P/R	0.663	0.592	0.732	0.659	3.5%	0.4%	6.7%	2.8%
	SLMIA-SR	**0.697**	**0.764**	**0.774**	**0.827**	**3.8%**	**3.1%**	**8.8%**	**5.1%**
Res-AP	**EncoderMI-V**	0.712	0.87	0.789	0.948	4.9%	33.1%	13.7%	41.3%
	FaceAuditor-S	0.722	0.885	0.794	0.961	4.9%	28.3%	14.9%	52.5%
	FaceAuditor-P/R	0.672	0.697	0.744	0.771	4.2%	4.2%	11.6%	8.7%
	SLMIA-SR	**0.763**	**0.932**	**0.841**	**0.982**	**13.6%**	**43.0%**	**29.0%**	**60.7%**
VGG-GE2E	**EncoderMI-V**	0.692	0.797	0.756	0.878	1.0%	4.7%	11.8%	9.9%
	FaceAuditor-S	0.671	0.797	0.728	0.89	0.4%	4.8%	5.9%	14.4%
	FaceAuditor-P/R	0.605	0.648	0.685	0.71	2.1%	1.1%	7.3%	2.8%
	SLMIA-SR	**0.708**	**0.934**	**0.776**	**0.98**	**6.3%**	**46.7%**	**15.5%**	**65.8%**

Note: VC-2 and LS denote the dataset VoxCeleb-2 and LibriSpeech, respectively.

Results in Setting-2. The results are reported in TABLE VI, where the results of $r = 1$ is omitted since it is less challenging than $r = 0$ according to the results in Setting-1. Overall, SLMIA-SR outperforms the baselines on all target SRSs and all datasets in Setting-2 as well, especially in terms of TPR. For example, it improves the TPR at 0.2% FPR from 4.8% to 46.7% on the target SRS VGG-GE2E and dataset LibriSpeech. We conjecture that the improvement is mainly brought by our diverse features and the voice chunk splitting strategy.

We find that in most cases, the performance of SLMIA-SR degrades compared to Setting-1, which is possibly due to the reduced number of voices used for membership inference of target speakers in Setting-2. More voices of a target speaker contribute to a more precise approximation of the centroid embedding for that speaker, more precise statistics, and hence more precise features. However, on the target SRSs Res-AP (for both VoxCeleb-2 and LibriSpeech) and VGG-GE2E (for LibriSpeech only), we find that SLMIA-SR achieves much higher TPR at 0.1% or 0.2% FPR in Setting-2 than in Setting-1. This is probably attributed to the two strategies used in Setting-2, voice-number-dependent attack models and voice chunk splitting strategy which artificially increases the number of voices used for membership inference of target speakers.

2) Effectiveness of Individual Components: The success of SLMIA-SR may be attributed to the combination of several components, including the feature extractor with numerous features, the mixing training strategy, the voice-number dependent attack models, the voice chunk splitting strategy, and the group enrollment and the voice concatenation techniques. Therefore,we perform additional experiments to understand the effectiveness and necessity of each component.

Settings. Since we have already considered all the combinations of the five target SRSs and the two datasets in the previous experiments, here we target the VoxCeleb-2 dataset since it contains much more speakers than LibriSpeech and the target SRSs trained on this dataset are generally less vulnerable to SLMIA-SR than those trained on LibriSpeech. Also, we consider the target SRS LSTM-GE2E since it is lightweight and therefore computationally efficient, but is not the most vulnerable to SLMIA-SR among the five target SRSs.

For comprehensive evaluation, we will also consider the following three variants of SLMIA-SR: the attack using the threshold-based attack model with each single feature (denoted by the corresponding feature notation), the attack using the classifier-based attack model with all the intra-features (denoted by Intra-Ens), and the attack using the classifier-based attack model with all the inter-features (denoted by Inter-Ens).

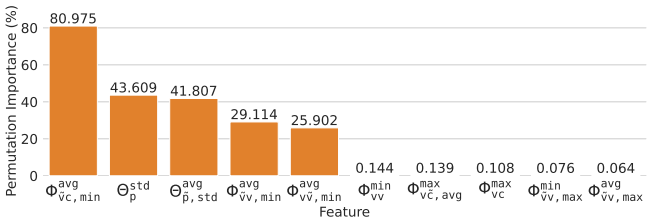


Fig. 7: Features with the top-5 highest and lowest PI

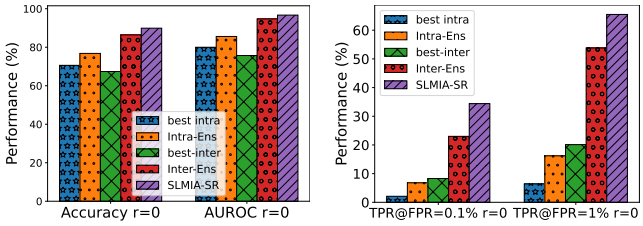


Fig. 8: Comparison of SLMIA-SR with its variants.

Contribution of single feature. We conduct two experiments to understand the contribution of each feature to SLMIA-SR.

First, we utilize and report the permutation importance (PI) [59] of each feature. The PI of a feature to the classifier is the change of a metric on the test data by permuting this feature across the test data. We randomly permutate each feature 10 times and report the average change of all the metrics over 10 permutations. The top-5 highest/lowest ones are depicted in Fig. 7. We find that all the features have positive PIs, indicating that they all contribute to SLMIA-SR, and the top-5 highest features have never been used in prior MIA, which is one of the reasons why SLMIA-SR outperforms them.

Second, we compare SLMIA-SR with its variants, and the results are shown in Fig. 8. We find that Intra-Ens outperforms the best sole intra-feature, Inter-Ens outperforms the best sole inter-feature, and SLMIA-SR outperforms both Intra-Ens and Inter-Ens, in terms of all the metrics. This further justifies the contribution of each feature and demonstrates that these features can complement each other.

Effectiveness of mixing ratio training. Let r_t (resp. r_m) denote the ratio r used in attack model building (resp. membership inference). To understand the effect of the ratio r_m and the effectiveness of our mixing training strategy, we vary r_m from 0 to 1 with step 0.1. For each training speaker, $N \times r_m$ voices are randomly sampled from $\mathcal{V}_{tr}^t$ and $N \times (1-r_m)$ voices are randomly sampled from $\mathcal{V}_{ntr,tr}^t$, forming the inference voices of the speaker. The results are depicted in Fig. 9, where SLMIA-SR-r_t=0&1, SLMIA-SR-r_t=0 and SLMIA-SR-r_t=1 denote SLMIA-SR with the attack models trained with both $r_t = 0$ and $r_t = 1$ (i.e., our mixing ratio training strategy), with only $r_t = 0$, and with only $r_t = 1$, respectively. We observe that the performance of SLMIA-SR increases with the ratio r_m. Though SLMIA-SR-r_t=0&1 performs worse than SLMIA-SR-r_t=0 (resp. SLMIA-SR-r_t=1) at $r_m = 0$ (resp. $r_m = 1$), it performs much better at the other values of r_m. This demonstrates the effectiveness of our mixing training strategy in enhancing the generalizability of SLMIA-SR to different ratios.

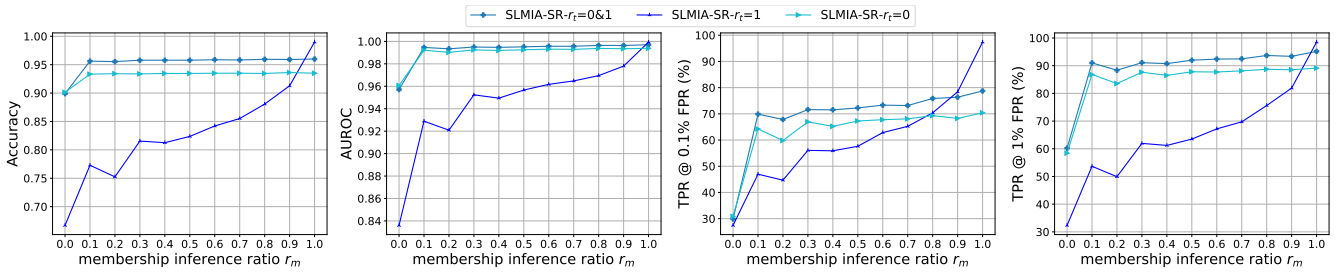


Fig. 9: Effectiveness of SLMIA-SR w.r.t. the ratio r.

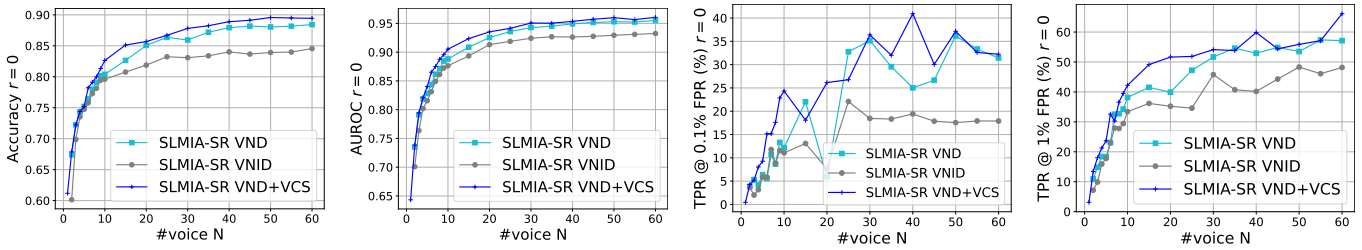


Fig. 10: Effectiveness of SLMIA-SR w.r.t. the number of voices N.

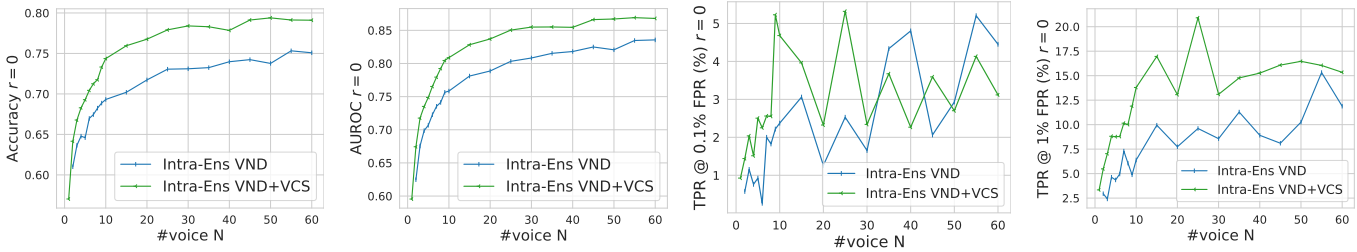


Fig. 11: Effectiveness of our voice chunk splitting.

Effectiveness of voice-number-dependent attack models. To understand the effect of the number N of voices provided for membership inference and the effectiveness of the voice-number-dependent attack models, we vary N from 1 to 10 with step 1 and from 15 to 60 with step 5. We train both VND and VNID attack models (denoted by SLMIA-SR VND and SLMIA-SR VNID). The results are depicted in Fig. 10. We observe that the performance of attack models increases with N, probably because more voices help compute more precise features. As expected, the VND attack models (i.e., SLMIA-SR VND) perform better than the single VNID attack model (i.e., SLMIA-SR VNID), because features vary significantly with N (cf. Fig. 5). This indicates the effectiveness and necessity of building voice-number-dependent attack models.

Effectiveness of voice chunk splitting. In our voice chunk splitting approach, we set the window size w to 3,200 milliseconds and the window step s to $\frac{w}{2}$, following the common practice in speech signal processing [60]. During splitting, when a chunk is shorted than the window size, it is padded with zero values to fulfill the window size if its length is no shorter than 70% of the window size, otherwise omitted. The results are also depicted in Fig. 10 (curve SLMIA-SR VND+VCS). Compared with SLMIA-SR VND, we observe that our voice chunk splitting improves the performance of SLMIA-SR VND with a little exception on TPR at 0.1% FPR when $N > 10$, indicating that our voice chunk splitting can more reliably improve TPR at low FPR when N is small. For example, our voice chunk splitting improves the TPR at 0.1% FPR from 12% to 25% when $N = 10$. To further understand the effectiveness of our voice chunk splitting, we also apply it to Intra-Ens, which is less effective than SLMIA-SR according to Fig. 8. As shown in Fig. 11, the improvement brought by our voice chunk splitting is more significant.

Effectiveness of enrollment voice concatenation and group enrollment. While both enrollment voice concatenation and group enrollment can effectively reduce the number of queries to the target SRS in the black-box scenario, they may affect the performance of SLMIA-SR. To understand the effect, we set $N = 10$, $M = 20$, and $K = 10$, the same as Setting-2. The effect of the number M of imposters and the number K of voices per imposter refers to Appendix D. We partition the features into fixed groups according to their similarity and distance types. We train one VND attack model with our mixing training strategy for each group of features. Voice chunk splitting is not utilized since it will increase the number of queries in the black-box scenario. The results are reported in TABLE VII. We can observe that our group enrollment technique reduces the number of queries by nearly half with no effect on the performance of SLMIA-SR, which is not surprising, as group enrollment does not change the recognition scores. Though our enrollment voice concatenation technique changes the recognition scores, the performance of SLMIA-SR decreases slightly and even increases in some cases, indicating that the embedding of the concatenated voice can well approximate the centroid of voice embeddings.

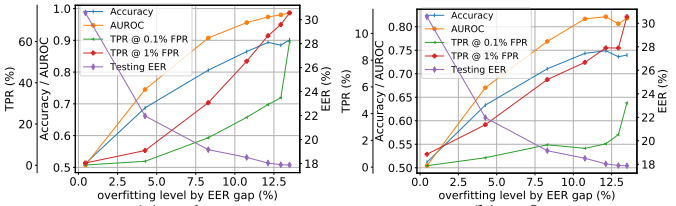
VI. Ablation Study

In this section, we perform ablation studies to understand the effects of the overfitting level of the target SRS and the

TABLE VII: The effectiveness of enrollment voice concatenation and group enrollment. "Recog" denotes recognition.

			Accuracy		AUROC		TPR @ x% FPR x=0.1		TPR @ x% FPR x=1		#Query (Target)		
			r=1	r=0	r=1	r=0	r=1	r=0	r=1	r=0	Enroll	Recog.	Total
Intra	Centroid-based	Baseline	0.797	0.684	0.959	0.747	9.9%	2.4%	38.9%	7.6%	10	10	20
		Concat	0.786	0.681	0.950	0.742	11.6%	2.4%	31.1%	6.1%	1	10	11
	Pairwise	Baseline	0.798	0.695	0.961	0.759	8.5%	2.0%	37.2%	6.5%	9	45	54
Inter	Centroid-centroid	Baseline	0.722	0.636	0.805	0.682	14.1%	2.5%	28.4%	10.4%	10	20	30
		Concat	0.738	0.648	0.821	0.714	9.3%	2.3%	23.1%	9.3%	1	20	21
	Centroid-voice	Baseline	0.769	0.651	0.852	0.715	17.1%	4.0%	26.5%	8.5%	10	200	210
		Concat	0.730	0.643	0.809	0.702	11.1%	2.8%	21.4%	8.2%	1	200	201
	Voice-centroid	Baseline	0.782	0.645	0.870	0.693	19.2%	4.9%	27.8%	8.2%	200	200	400
		Group	0.782	0.645	0.870	0.693	19.2%	4.9%	27.8%	8.2%	200	10	210
		Concat	0.794	0.651	0.878	0.696	17.4%	4.0%	29.6%	8.4%	20	200	220
		Group+Concat	0.794	0.651	0.878	0.696	17.4%	4.0%	29.6%	8.4%	20	10	30
	Voice-voice		0.830	0.675	0.918	0.731	12.3%	2.2%	36.0%	9.9%	10	2000	2010

Note: (i) Refer to TABLE II for the computation of the number of queries. (ii) The adversary has white-box access to the shadow SRS including the embeddings, so the total number of queries to the shadow SRS is $N + M * K = 210$.



(a) r=1 (b) r=0

Fig. 12: The effectiveness of SLMIA-SR w.r.t. the overfitting level of the target SRS.

effects of the differences in dataset distribution and model architecture between the target and shadow SRSs. Here we only incorporate all intra-features into SLMIA-SR (i.e., Intra-Ens) since it requires fewer queries than incorporating all the inter-features (i.e., Inter-Ens), according to the results in TABLE VII. We set the number N of voices per target speaker to 40 since according to Fig. 10, the accuracy and AUROC almost converge at $N \geq 40$.

Overfitting level of the target SRS. The same as in § V-B2, we use the LSTM-GE2E SRS and the VoxCeleb-2 dataset. We measure the overfitting level of the target SRS by analyzing the gap between training EER and testing EER. Specifically, we randomly sample 50,000 trials from the training voices $\mathcal{V}^t_{tr}$ of the training speakers $\mathcal{S}^t_{tr}$ and 50,000 trails from the non-training voices $\mathcal{V}^t_{ntr,ntr}$ of non-training speakers $\mathcal{S}^t_{ntr}$, on which the training and testing EER are computed, respectively. We control the EER gap by varying the number of training epochs from 40 to 1800. The results are reported in Fig. 12 under both $r = 1$ and $r = 0$.

We find that the effectiveness of SLMIA-SR generally increases with the overfitting level. This is not surprising since the success of membership inference attacks lies in the overfitting of target models. We also find that the testing EER decreases with the increase of the overfitting level, indicating that the target SRS should have a large overfitting level to achieve low EER, which in turn benefits SLMIA-SR. For example, to achieve about 18% testing EER, which is still a moderate performance, the target SRS has an overfitting level of about 12% where SLMIA-SR achieves 0.85 AUROC and 2% TPR at 0.1% FPR even when $r = 0$. The results demonstrate that it is nontrivial to balance the recognition performance and the resilience against SLMIA-SR.

Disjoint datasets. We relax the assumption in § III-B that

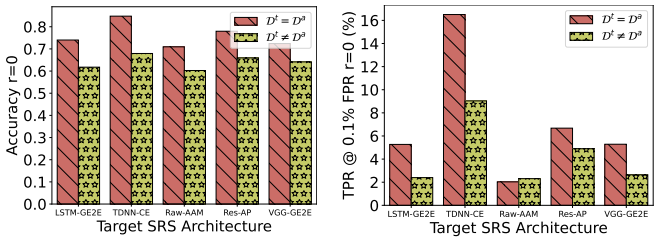


Fig. 13: Effect of the dataset distribution.

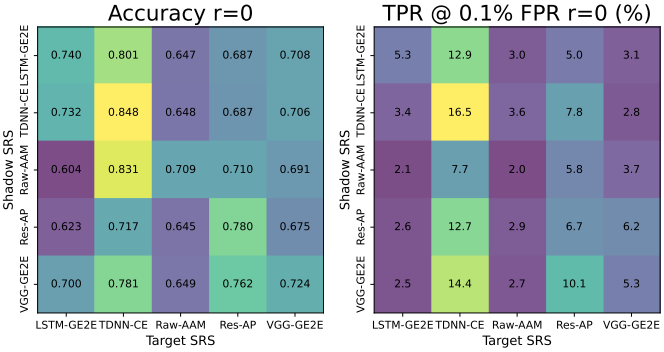


Fig. 14: Effect of the architectures.

the adversary has an auxiliary speaker dataset $\mathcal{S}^a$ which is sampled from the same distribution as the target SRS's training speaker dataset $\mathcal{S}^t_{tr}$. We denote by $\mathcal{D}^a$ and $\mathcal{D}^t_{tr}$ the distributions of $\mathcal{S}^a$ and $\mathcal{S}^t_{tr}$, respectively. We compare the effectiveness of SLMIA-SR between when $\mathcal{D}^a = \mathcal{D}^t_{tr}$ and when $\mathcal{D}^a \neq \mathcal{D}^t_{tr}$. For $\mathcal{D}^a = \mathcal{D}^t_{tr}$, both $\mathcal{S}^a$ and $\mathcal{S}^t_{tr}$ are sampled from the VoxCeleb-2 dataset, but are disjoint from each other in the speakers and voices. For $\mathcal{D}^a \neq \mathcal{D}^t_{tr}$, $\mathcal{S}^a$ and $\mathcal{S}^t_{tr}$ are respectively sampled from the LibriSpeech and VoxCeleb-2 datasets. Recall that the training dataset of the shadow SRS is a subset of the auxiliary dataset S^a. The accuracy and TPR at 0.1% FPR when $r = 0$ are shown in Fig. 13, while other results refer to Appendix E. We observe that SLMIA-SR still achieves good performance with at least 2% TPR at 0.1% FPR and 60% accuracy when $r = 0$, a more challenging setting than $r = 1$. Aligning with prior works [12], [10], the effectiveness of SLMIA-SR decreases with a dataset distribution shift in most cases, probably because training datasets with different distributions make the target and shadow SRSs learn different speaker embedding mappings.

Disjoint SRS architectures. We relax the assumption in § III-B that the adversary knows the architecture of the target SRS and adopt the same architecture for the shadow SRS. We consider all the 5×5 pairs of the five architectures given in TABLE IV, all of which are trained using the dataset VoxCeleb-2. The results are shown in Fig. 14 and Appendix F. Aligning with prior works [12], [10], [33], the architectures of the target and shadow SRSs do affect the effectiveness of SLMIA-SR. SLMIA-SR often achieves the best performance when the shadow SRS shares the same architecture with the target SRS, especially in terms of accuracy. However, interestingly, we find that the attack using a different shadow SRS architecture from the target SRS may achieve higher TPR at 0.1% FPR than the attack using the same shadow SRS architecture as the target SRS, e.g., when the architecture of the target SRS is RawNet3, ResNetSE34V2, and VGGVox40. Nevertheless, SLMIA-SR always achieves at least 60% accuracy and 2% TPR at 0.1% FPR when $r = 0$ which is more challenging than $r = 1$.

TABLE VIII: Results of DP-SGD. "Standard" denotes training using SGD without differential privacy.

σ	ϵ	testing EER	Accuracy	TPR @ 0.1% FPR
Standard	∞	6.56%	0.906	26.3%
0	∞	7.88%	0.842	17.6%
0.2	398	17.31%	0.521	0.4%
0.5	3.87	28.03%	0.506	0.1%
0.8	0.614	32.61%	0.502	0.1%

VII. Discussion

We discuss possible countermeasures against SLMIA-SR as well as improvements and extensions of SLMIA-SR.

Countermeasures. Both training and inference phase defenses could be utilized to prevent our attack SLMIA-SR.

Training phase defenses. Since the effectiveness of SLMIA-SR lies in the overfitting of the target SRS, techniques that alleviate overfitting could be used to defeat our attack, e.g., early-stopping [31]. As shown in Fig. 12, with the decrease in the number of training epochs of the target SRS, the overfitting level and attack effectiveness decrease, consistent with the finding in [31], while the testing EER increases. This suggests that early-stopping, i.e., training the target SRS with less number of epochs, can be used to prevent SLMIA-SR, at the cost of sacrificing the speaker recognition utility.

Another line of training phase defenses is differential privacy [11], [12], [31]. In contrast to early-stopping which is an empirical defense, differential privacy can provide provable membership privacy guarantees [31]. Differential privacy features a parameter called privacy budget ϵ, the smaller, the more private. To check feasibility, we adopt the differentially private stochastic gradient descent (DP-SGD) [61] to train a differential private SRS using the TDNN-CE and VoxCeleb-2 dataset, by first clipping the per-example gradient to a pre-set maximal gradient norm C and then adding noise to the aggregated gradient within a training batch where the magnitude of the noise is positively correlated with a parameter σ. We fix $C = 5$ after tuning, and vary σ from 0, 0.2, 0.5 to 0.8. The privacy budget ϵ, performance of SLMIA-SR, and testing EER are reported in TABLE VIII. We find that only clipping the per-example gradient with $C = 5$ without adding noise ($\sigma = 0$) cannot prevent our attack. While SLMIA-SR becomes almost a random guesser using DP-SGD with $\sigma = 0.2$, the testing EER increases by more than 10%. These results demonstrate that DP-SGD can prevent SLMIA-SR at the cost of sacrificing the speaker recognition utility.

Inference phase defenses. Since SLMIA-SR in the black-box scenario requires recognition scores from a target SRS to compute features, a straightforward inference phase defense is to hide scores and only output the recognition results, i.e., "accept" or "reject" for speaker verification, and the identified speaker or "reject" for speaker identification. However, such defense reduces the useful information for third-party developers, e.g., without scores, they cannot fine-tune the threshold of SRSs to achieve better recognition performance in their specific applications. It may also be circumvented by decision-only membership inference attacks [62].

Removing shadow SRS training. Due to the lack of knowledge about the training and non-training speakers of the target SRS, SLMIA-SR trains a shadow SRS based on which an attack model is built in a supervised manner. When some non-training speakers of the target SRS are available to the adversary, the adversary can either tune the threshold to satisfy a specific false positive rate or training an out-of-distribution classifier, thus removing the requirement of shadow SRS training, such as the attack in [32]. A possible way to obtain such non-training speakers is artificial speaker generation with text-to-speech [52], similar to the random image generation [63]. The adversary first randomly samples an embedding in the voice embedding space, and then synthesizes a voice from that embedding. Due to the remarkably large voice embedding space, the synthesized speaker is highly likely to be a non-training speaker of the target SRS.

Decision-only MIA. SLMIA-SR in the black-box scenario requires recognition scores from a target SRS to compute features. However, in practice, some SRSs may only output recognition results instead of recognition scores, e.g., "accept" or "reject" for speaker verification. The adversary can utilize adversarial attacks to achieve decision-only membership inference [62]. The assumption is that the voices of training speakers are more robust against adversarial perturbation than that of non-training speakers. Hence, the adversary can exploit the average minimal perturbation to change the recognition result of each voice of the target speaker as the feature to perform speaker-level membership inference. We leave this interesting problem for future work.

Extending to other biometric recognition. Speaker recognition is one kind of biometric recognition that recognizes a person's identity from biometric data. Other biometric recognition, e.g., face and fingerprint recognition, share similar pipelines (training, enrollment, and recognition phases), architecture (the final fully connected layer is removed after training), training paradigms (classification-based or verification-based losses), and training objectives (pulling the data of the same subject together and pushing different subjects away) with speaker recognition [1]. Therefore, SLMIA-SR may be extended to other biometric recognition, which is worthy to explore in the future.

VIII. Conclusion

We proposed SLMIA-SR, the *first* membership inference attack against speaker recognition. Instead of considering voice-level membership inference that determines whether *some* given voices were contained in the training of a target SRS, SLMIA-SR features speaker-level membership inference to determine whether *any* voices of a target speaker were contained in the training of a target SRS, where the voices of training speakers used for membership inference are not required to get involved in the training of the target SRS. We showed that prior MIA designed for embedding models are unsatisfactory to SRSs. Thus, we designed and studied a large number of features to characterize the differences between training and non-training speakers, introduced a mixing ratio training strategy to improve the generalizability of attack models, voice-number-dependent attack models and voice chunk splitting to enhance attack effectiveness, group enrollment and enrollment voice concatenation techniques to reduce queries probed to the target SRS in the black-box scenario. Extensive experiments demonstrate the effectiveness of SLMIA-SR, the proposed techniques and features. Our work sheds light on future research in the area of private speaker recognition.

REFERENCES

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: a survey," *Artifcial Intelligence Review*, 2023.

[2] Citi Uses Voice Prints To Authenticate Customers Quickly And Effortlessly. https://www.citibank.com.hk/english/info/pdf/VoiceBiometrics_PressRelease_Eng_final_online.pdf.

[3] Starbucks and Alibaba Launch Voice Ordering and Delivery. https://stories.starbucks.com/press/2019/starbucks-and-alibaba-launch-voice-ordering-and-delivery.

[4] H. Ren, Y. Song, S. Yang, and F. Situ, "Secure smart home: A voiceprint and internet based authentication system for remote accessing," in *ICCSE*, 2016.

[5] Alexa voice ID helps Alexa recognize you when you speak and provide a personalized experience. https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X.

[6] Z. Bai and X. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021. [Online]. Available: https://doi.org/10.1016/j.neunet.2021.03.004

[7] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *CCS*, 2015.

[8] K. Pizzi, F. Boenisch, U. Sahin, and K. Böttinger, "Introducing model inversion attacks on automatic speaker recognition," in *Proceedings of the 2nd Symposium on Security and Privacy in Speech Communication*, 2023.

[9] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *USENIX Security Symposium*, 2021.

[10] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy, SP*, 2017.

[11] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *IEEE Symposium on Security and Privacy, SP*, 2022.

[12] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *ACM SIGSAC Conference on Computer and Communications Security, CCS*, 2022.

[13] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, "Auditing membership leakages of multi-exit networks," in *CCS*, 2022.

[14] H. Hu, Z. Salcic, G. Dobbie, and X. Zhang, "Membership inference attacks on machine learning: A survey," *CoRR*, vol. abs/2103.07853, 2021.

[15] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *ACM SIGSAC Conference on Computer and Communications Security, CCS*, 2022.

[16] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *USENIX Security Symposium*, 2021.

[17] X. He, H. Liu, N. Z. Gong, and Y. Zhang, "Semi-leak: Membership inference attacks against semi-supervised learning," in *ECCV*, 2022.

[18] Y. Miao, M. Xue, C. Chen, L. Pan, J. Zhang, B. Z. H. Zhao, D. Kaafar, and Y. Xiang, "The audio auditor: User-level membership inference in internet of things voice services," *Proc. Priv. Enhancing Technol.*, 2021.

[19] M. A. Shah, J. Szurley, M. Müller, A. Mouchtaris, and J. Droppo, "Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks," in *Interspeech*, 2021.

[20] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying privacy risks of masked language models using membership inference attacks," in *EMNLP*, 2022.

[21] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proc. Priv. Enhancing Technol.*, 2019.

[22] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against generative models," in *CCS*, 2020.

[23] Amazon is being sued over its smart assistant's recordings of children's voiceprint. https://www.bbc.com/news/technology-48623914.

[24] J. Scherer and G. Kiparski, "Buchbesprechungen. feiler, lukas / forgó, nikolaus / weigl, michaela: The eu general data protection regulation (gdpr): A commentary," *Comput. und Recht*, 2018.

[25] Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[26] "Microsoft azure speaker recognition," https://docs.microsoft.com/en-us/rest/api/speakerrecognition/, 2022.

[27] Nuance Voice Biometric. https://www.nuance.com/omni-channel-customer-engagement/authentication-and-fraud-prevention/gatekeeper/what-is-voiceprint.html.

[28] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[29] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[30] G. Li, S. Rezaei, and X. Liu, "User-level membership inference attack against metric embedding learning," *CoRR*, vol. abs/2203.02077, 2022.

[31] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "Encodermi: Membership inference against pre-trained encoders in contrastive learning," in *CCS*, 2021.

[32] W. Tseng, W. Kao, and H. Lee, "Membership inference attacks against self-supervised speech models," in *Interspeech*, 2022.

[33] M. Chen, Z. Zhang, T. Wang, M. Backes, and Y. Zhang, "FACE-AUDITOR: data auditing in facial recognition systems," *CoRR*, vol. abs/2304.02782, 2023.

[34] Open-source code for SLMIA-SR. https://github.com/S3L-official/SLMIA-SR.

[35] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[36] G. Chen, Y. Zhang, Z. Zhao, and F. Song, "QFA2SR: query-free adversarial transfer attacks to speaker recognition systems," in *USENIX Security*, 2023.

[37] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real Bob? adversarial attacks on speaker recognition systems," in *S&P*, 2021.

[38] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[39] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, F. Wang, and J. Wang, "Towards understanding and mitigating audio adversarial examples for speaker recognition," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[40] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on gmm i-vector based speaker verification systems," in *ICASSP*, 2020.

[41] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *ASIACCS*, 2020.

[42] H. Abdullah, M. S. Rahman, W. Garcia, L. Blue, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, "Hear "no evil", see "kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems," in *IEEE S&P*, 2021.

[43] B. Zheng, P. Jiang, Q. Wang, Q. Li, C. Shen, C. Wang, Y. Ge, Q. Teng, and S. Zhang, "Black-box adversarial attacks on commercial speech platforms with minimal information," in *CCS*, 2021.

[44] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *CCS*, 2020.

[45] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. B. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," in *NDSS*, 2019.

[46] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, 2015.

[47] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, ""hello, it's me": Deep learning-based speech synthesis attacks in the real world," in *CCS*, 2021.

[48] M. Marras, P. Korus, A. Jain, and N. D. Memon, "Dictionary attacks on speaker verification," *IEEE Trans. Inf. Forensics Secur.*, 2023.

[49] N. Cressie and H. Whitford, "How to use the two sample t-test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[50] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[51] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[52] Implementation of SV2TTS. https://github.com/CorentinJ/Real-Time-Voice-Cloning.

[53] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020.

[54] "Ecapa-tdnn model implemented by speechbrain," https://github.com/speechbrain/speechbrain/tree/develop/recipes/VoxCeleb/SpeakerRec.

[55] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.

[56] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[58] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017, pp. 2616–2620.

[59] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinform.*, vol. 26, no. 10, pp. 1340–1347, 2010.

[60] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-Hall, Inc., 1978.

[61] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *CCS*, 2016.

[62] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *CCS*, 2021.

[63] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *NDSS*, 2019.

APPENDIX

A. SLMIA-SR *vs. Prior MIA on SR embedding models*

In TABLE IX, we compare our attack SLMIA-SR with four recent promising MIA designed for embedding models: Li et al. [30], Tseng et al. [32], EncoderMI [31], and FaceAuditor [33], regarding the focused tasks, MIA level, feature, adversarial capacity, attack model, etc. We also apply and generalize those MIA to speaker-level membership inference against speaker recognition and compare their performance with SLMIA-SR.

- **Li et al. [30]:** it utilized the intra-features $\Theta_{\mathrm{c}}^{\mathtt{avg}}$ and $\Theta_{\mathrm{p}}^{\mathtt{avg}}$ as inputs to a binary classifier-based attack model.
- **Tseng et al. [32]:** it utilized the intra-feature $\Theta_{\mathrm{p}}^{\mathtt{avg}}$ for a threshold-based attack model and improved the attack by replacing the pre-defined cosine similarity in $\Theta_{\mathrm{p}}^{\mathtt{avg}}$ with the similarity produced by a neural network. We consider the more effective improved attack.
- **EncoderMI [31]:** it proposed three different attack models, EncoderMI-T, EncoderMI-V, and EncoderMI-S. We consider the two most effective variants: EncoderMI-T and EncoderMI-V. EncoderMI-T is a threshold-based attack model that utilizes the intra-feature $\Theta_{\mathrm{p}}^{\mathtt{avg}}$, and EncoderMI-V is a classifier-based attack model that utilizes the sorted set of similarities contributing to $\Theta_{\mathrm{p}}^{\mathtt{avg}}$.
- **FaceAuditor [33]:** it proposed classifier-based attack models that utilize either the similarities contributing to feature $\Theta_{\mathrm{p}}^{\mathtt{avg}}$ (denoted by FaceAuditor-S) or the similarities contributing to $\Theta_{\mathrm{c}}^{\mathtt{avg}}$ and $\Phi_{\mathrm{vc}}^{\mathtt{avg}}$ (denoted by FaceAuditor-P/R).

B. Comparison between the white-box and black-box scenarios regarding the centroid-centroid inter-features

As mentioned in § IV-A, our feature extractor demonstrates unity in white-box and black-box scenarios. The minor difference lies in the computation of the centroid-centroid inter-features. In the black-box scenario, to compute such group of features, we apply enrollment voice concatenation to obtain one longer and concatenated voice for each imposter with multiple voices. Here we check whether this difference leads to a significant performance gap between the two scenarios. We adopt the LSTM-GE2E target SRS and the dataset VoxCeleb-2, set the number of voices of target speakers $N = 10$, the number of imposters $M = 20$ and the number of voices per imposter $K = 10$, the same as Setting-1 in § V. We consider two cases, namely, all features are used and only the centroid-centroid inter-features are used. The results are shown in TABLE X. We find that there is no obvious performance gap between white-box and black-box scenarios, and sometimes the attack even performs better in the black-box scenario than in the white-box scenario. It is because for each imposter, the embedding of the concatenated voice (used in the black-box scenario) can well approximate the centroid embedding of the voice embeddings (used in the white-box scenario).

Algorithm 1: Determining an upper bound N'

Input: shadow SRS SR^s; datasets $\boldsymbol{\mathcal{V}} = \{\mathcal{V}^s_{tr}, \mathcal{V}^s_{ntr,tr}, \mathcal{V}^s_{ntr,ntr}\}$; features $\Psi = \{\cdots, \psi_i, \cdots\}$; #voice step s; T-test significance value α
Output: an upper bound N'

$\mathcal{C} \leftarrow \emptyset$
for $\psi \in \Psi$ **do**
  for $\mathcal{V} \in \boldsymbol{\mathcal{V}}$ **do**
    $n_1 \leftarrow 1$; $n_2 \leftarrow 1 + s$
    while *True* **do**
      $\mathcal{D}_1 \leftarrow \psi$ of $\mathcal{V}$ on SR^s with n_1 voices per speaker
      $\mathcal{D}_2 \leftarrow \psi$ of $\mathcal{V}$ on SR^s with n_2 voices per speaker
      $\eta \leftarrow$ T-test($\mathcal{D}_1$, $\mathcal{D}_2$) ▷ compute p-value
      if $\eta \geq \alpha$ **then** ▷ n_1 attack models suffice for ψ on $\mathcal{V}$
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{n_1\}$
        break
      $n_1 \leftarrow n_1 + 1$; $n_2 \leftarrow n_2 + 1$
return *maximum of* $\mathcal{C}$

C. Algorithm to bound the number of attack models

Alg. 1 iterates the given set Ψ of features (single feature for threshold-based attack model and multiple features for classifier-based attack model) and three datasets ($\mathcal{V}^s_{tr}$, $\mathcal{V}^s_{ntr,tr}$, and $\mathcal{V}^s_{ntr,ntr}$). For each feature $\psi \in \Phi$ and dataset $\mathcal{V}$, it first computes two sets of features D_1 and D_2 from n_1 and n_2 randomly sampled voices per speaker from the dataset $\mathcal{V}$, where n_1 and n_2 are initialized by 1 and $1 + s$ with the given step size s. Then, it computes the p-value η by performing a two-samples T-test [49] on D_1 and D_2 with the null hypothesis H_0: *D_1 and D_2 have the same mean*. If the p-value is no smaller than the pre-set significance value α, we accept H_0,

TABLE IX: Comparison of SLMIA-SR with the membership inference attacks on embedding models.

	Task	Level	Feature		Adversarial Capacity	Attack Model	Shadow Model	Non-Member Set of Target	TPR@ 0.1% FPR
			Aspect	Num					
LRL-MIA [30]	person re-identification	user	intra-closeness	2	embedding	classifier	✓	✗	1.5%
EncoderMI-T [31]†	contrastive learning	example	intra-closeness	1♯	embedding	threshold	✓	✗	1.6%
EncoderMI-V [31]†						classifier			2.0%
TLK-MIA [32]§	self-supervised speech	utterance speaker♮	intra-closeness	1	embedding	threshold	✗	✓	1.6%
FaceAuditor-S [33]‡	face recognition	user	intra-closeness	1♯	score	classifier	✓	✗	1.2%
FaceAuditor-P/R [33]‡			intra-closeness & inter-farness	2♯					1.6%
SLMIA-SR (Ours)	speaker recognition	speaker	intra-closeness & inter-farness	103	embedding & score	classifier	✓	✗	33.5%

Note: (1) †: We compared with the most two effective versions of EncoderMI [31], excluding EncoderMI-S. (2) §: We compared with the best one between the basic attack and the improved attack of [32]. (3) ‡: FaceAuditor [33] was available online when we were preparing this manuscript. (4) "Level": Example- and utterance-level predicts whether a given example is used for training, while user- and speaker-level determines whether *any* example of a given user/speaker is involved in training. ♮: Although [32] designed speaker-level MIA, in their evaluation, the voices provided for MI of training speakers come from training voices, thus deviating from our definition of speaker-level MIA which should reliably reach the "member" decision for a training speaker even if all of her/his voices provided for MI are different from training voices. (5) "Aspect": the properties that features quantify. Prior works covered three unique features ($\Theta_{\mathrm{c}}^{\mathrm{avg}}$, $\Theta_{\mathrm{p}}^{\mathrm{avg}}$, $\Phi_{\mathrm{vc}}^{\mathrm{avg}}$), which are the strict subset of our 103 features. ♯: EncoderMI-V and FaceAuditor used the raw set of similarities contributing to the features as input to attack models. (6) "embedding": the adversary can obtain the embedding of any input. "score": the adversary can only obtain the output recognition score. (7) "classifier": binary classifier-based attack models. "threshold": attack models predicting by thresholding a pre-set threshold. (8) "Shadow Model": whether the adversary has to train a shadow model to build the attack model. (9) "Non-Member Set of Target": whether the adversary is aware of a dataset in which each data is the non-member of the target model. With this knowledge, [32] did not need to train a shadow model. (10) "TPR@0.1% FPR" denotes the true positive rate when the false positive rate is 0.1%. The reported TPR are obtained on target SRS GE2E-LSTM and dataset VoxCeleb-2. More results refer to TABLE V and TABLE VI.

TABLE X: Performance comparison between white-box and black-box scenarios regarding the centroid-centroid inter-features.

		Accuracy	AUROC	TPR @ x% FPR x=0.1	x=1
All features	**White-box**	0.753	0.834	4.0%	14.7%
	Black-box	0.749	0.819	5.1%	14.8%
Centroid-centroid inter-features	**White-box**	0.636	0.682	2.5%	10.4%
	Black-box	0.648	0.714	2.3%	9.3%

indicating that there is no obvious statistic difference of the feature ψ between $N = n_1$ and $N = n_2$, hence it suffices to build n_1 attack models for the feature ψ on the dataset $\mathcal{V}$. In this case, n_1 is recorded in the set $\mathcal{C}$. Otherwise, we increase n_1 and n_2 by 1 and repeat the above T-test until H_0 is accepted. Finally, Alg. 1 returns the maximal number in the set $\mathcal{C}$ as the upper bound N'.

D. Results of the Effect of the Number of Imposters and Imposter Voices

Fig. 15 demonstrates the effects of the number M of imposters and the number K of voices per imposter on the effectiveness of SLMIA-SR with the attack models trained using all inter-features (i.e., Inter-Ens). Intra-features are omitted as they do not require imposters.

Effect of the number of imposters. We evaluate the effect of the number M of imposters by varying M from 20 to 100 with step 20. The number N of voices per target speaker is set to 40 in attack model training and membership inference, as SLMIA-SR almost converges at $N = 40$ in terms of accuracy and AUROC (cf. Fig. 10). We only use VND attack models trained with our mixing training strategy.

We find that accuracy (as well as AUROC) generally increases with the number M of imposters, because more imposters make the statistics of the sets of distances more precise, leading to more discriminative inter-features. However, TPR at 0.1% FPR does not monotonically increase with the number M of imposters, indicating that using more imposters improves accuracy and AUROC at the cost of more queries, but is not necessarily helpful for TPR.

Effect of the number of imposters' voices. To understand the effect of the number of imposters' voices, we vary the number K of voices per imposter from 10 to 90 with step 20, while the other settings are the same as above. We surprisingly find that both accuracy and TPR at 0.1% FPR do not monotonically increase with the number K. For example, when $M = 20$, $K = 10$ yields higher accuracy and TPR at 0.1% FPR than $K > 10$.

E. More Results of Ablation Study on the Disjoint Datasets

The results are shown in Fig. 16. We observe that SLMIA-SR still achieves good performance with at least 2% TPR at 0.1% FPR and 60% accuracy when $r = 0$, a more challenging setting than $r = 1$. Aligning with previous works [12], [10], the effectiveness of SLMIA-SR decreases with a dataset distribution shift in most cases, probably because training datasets with different distributions make the target and shadow SRSs learn different speaker embedding mappings.

F. More Results of Ablation Study on Disjoint Architectures

The results are reported in Fig. 17. Aligning with previous works [12], [10], [33], the architectures of the target and shadow SRSs do affect the effectiveness of SLMIA-SR. In general, SLMIA-SR achieves the best attack performance when the shadow SRS shares the same architecture with the target SRS, especially in terms of the average accuracy and AUROC. However, interestingly, we find that the attack using a different shadow SRS architecture from the target SRS may achieve higher TPR at 0.1% FPR and 1% FPR than the attack using the same shadow SRS architecture as the target SRS, e.g., when the architecture of the target SRS is Raw-AAM, Res-AP, and VGG-GE2E. Nevertheless, SLMIA-SR always achieves at least 60% accuracy and 2% TPR at 0.1% FPR when $r = 0$ which is more challenging than $r = 1$.

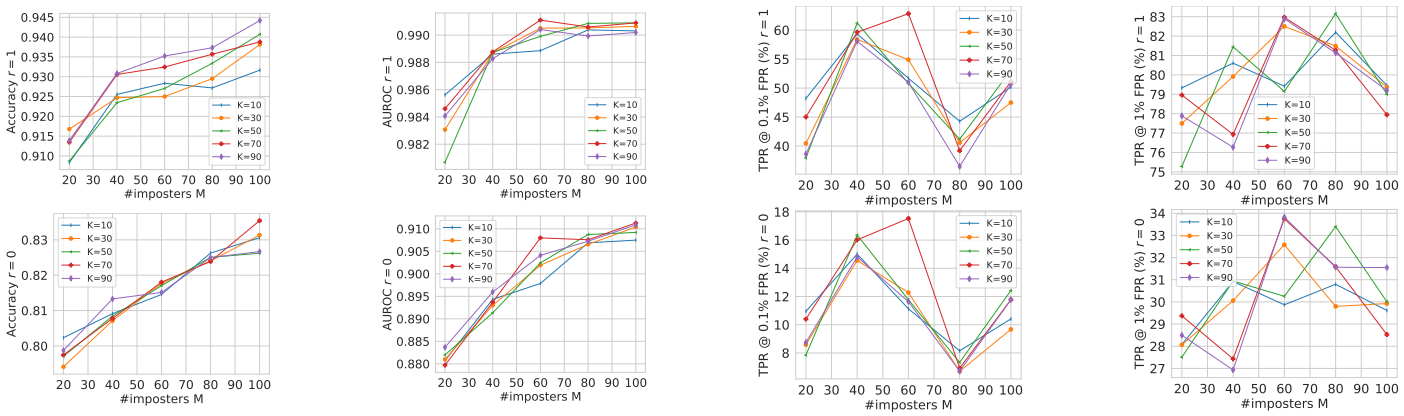


Fig. 15: Effects of the number M of imposters and the number K of voices per imposter on the performance of Inter-Ens.

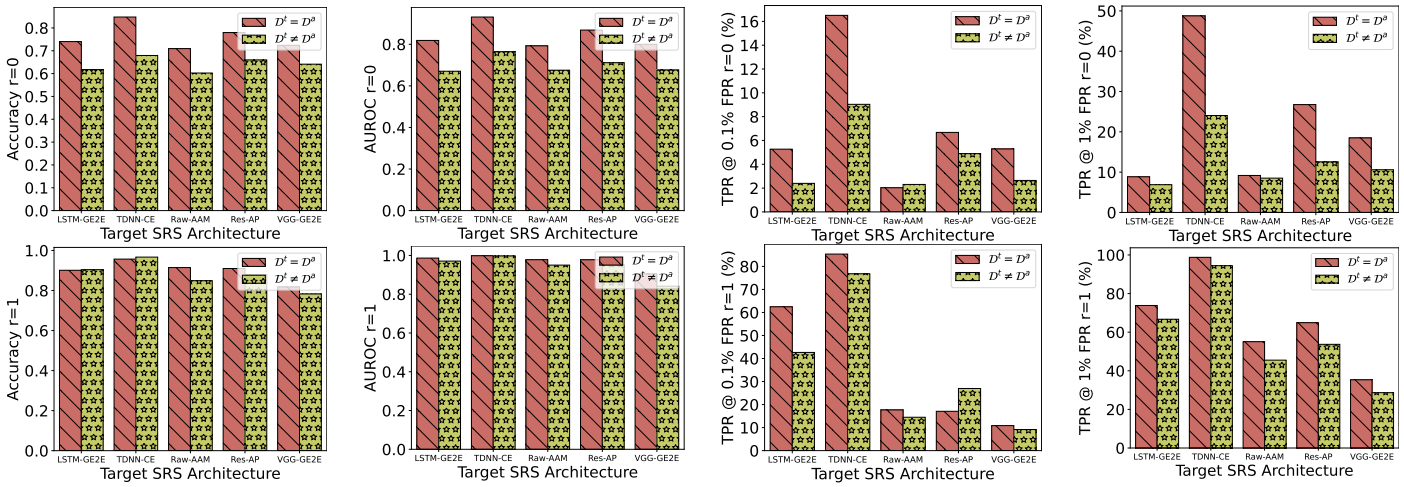


Fig. 16: Effect of the dataset distribution of the target and shadow SRSs on the effectiveness of SLMIA-SR.

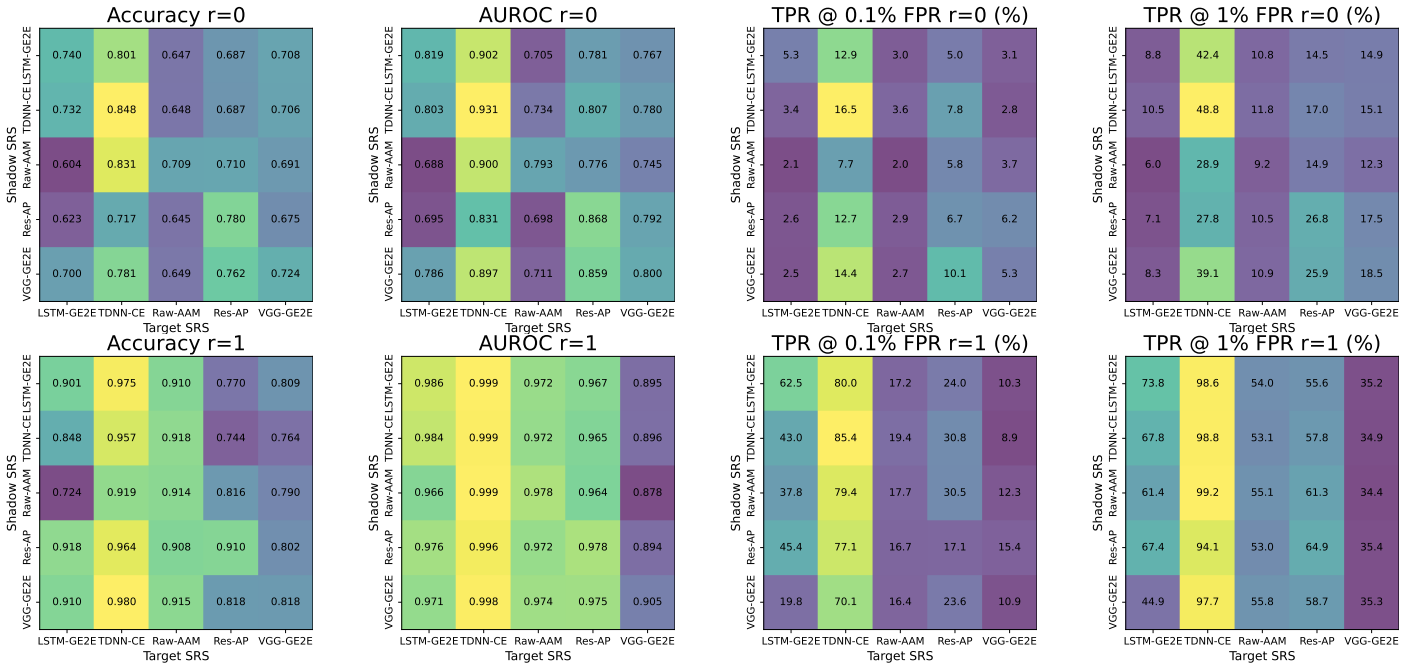


Fig. 17: Effect of the architectures of the target and shadow SRSs on the effectiveness of SLMIA-SR.