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

Authors: Guangke Chen[1], Yedi Zhang[2], Fu Song[3,4]
Presenter: Qifan Zhang[5]

1 上海科技大学 ShanghaiTech University

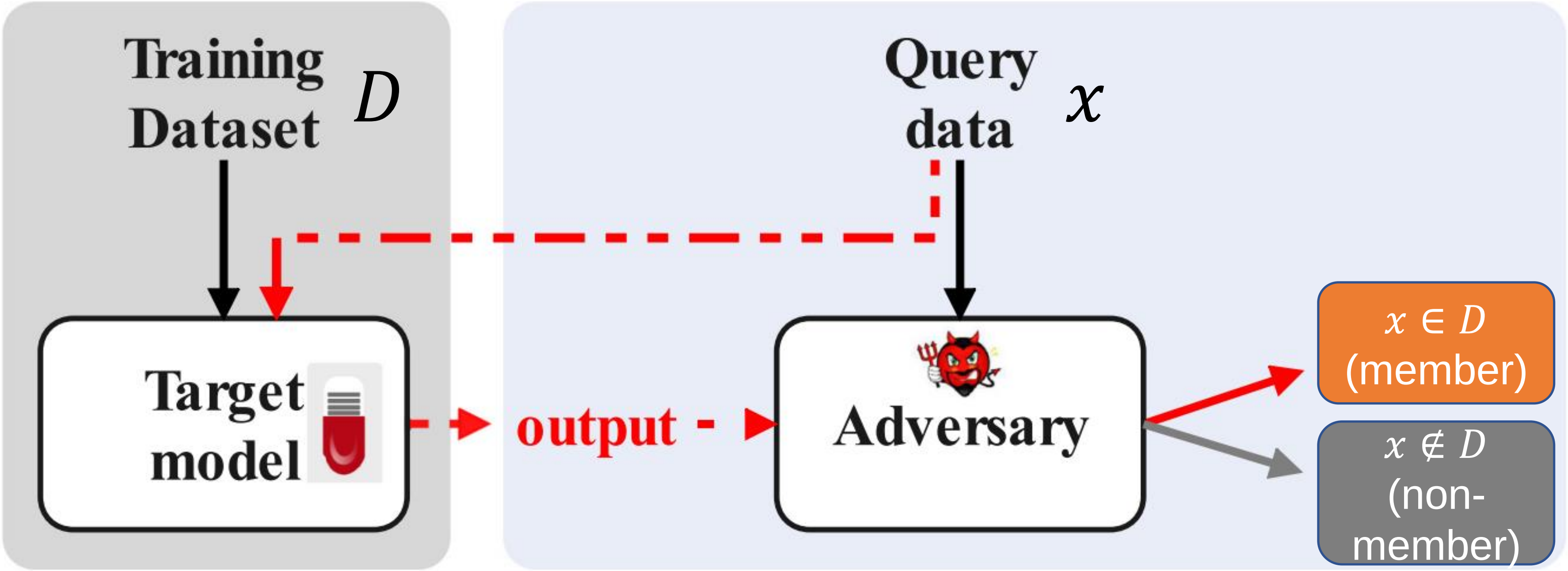
2 NUS National University of Singapore

3 中国科学院软件研究所 Institute of Software Chinese Academy of Sciences

4 中国科学院大学 University of Chinese Academy of Sciences

5 UCI University of California, Irvine

Membership Inference Attack (MIA)



[Li Hu et al. Defenses to Membership Inference Attacks: A Survey]

Speaker Recognition Systems (SRSs)

- Identify a person by his/her speeches

Speaker Recognition Systems (SRSs)

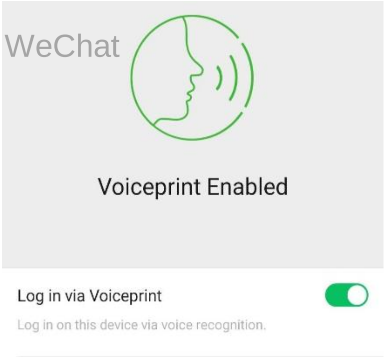
- Identify a person by his/her speeches

- Application:

voice assistant wake up

personalized service
on smart home

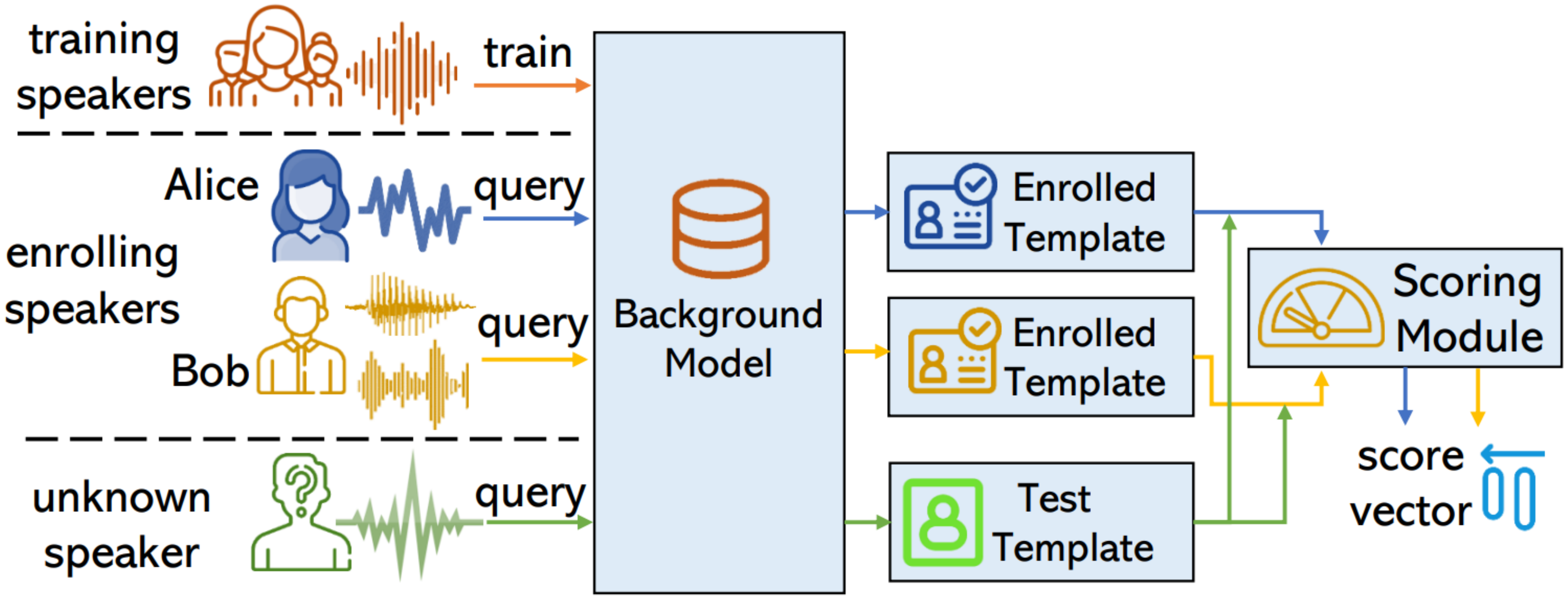
financial
transaction



app log in

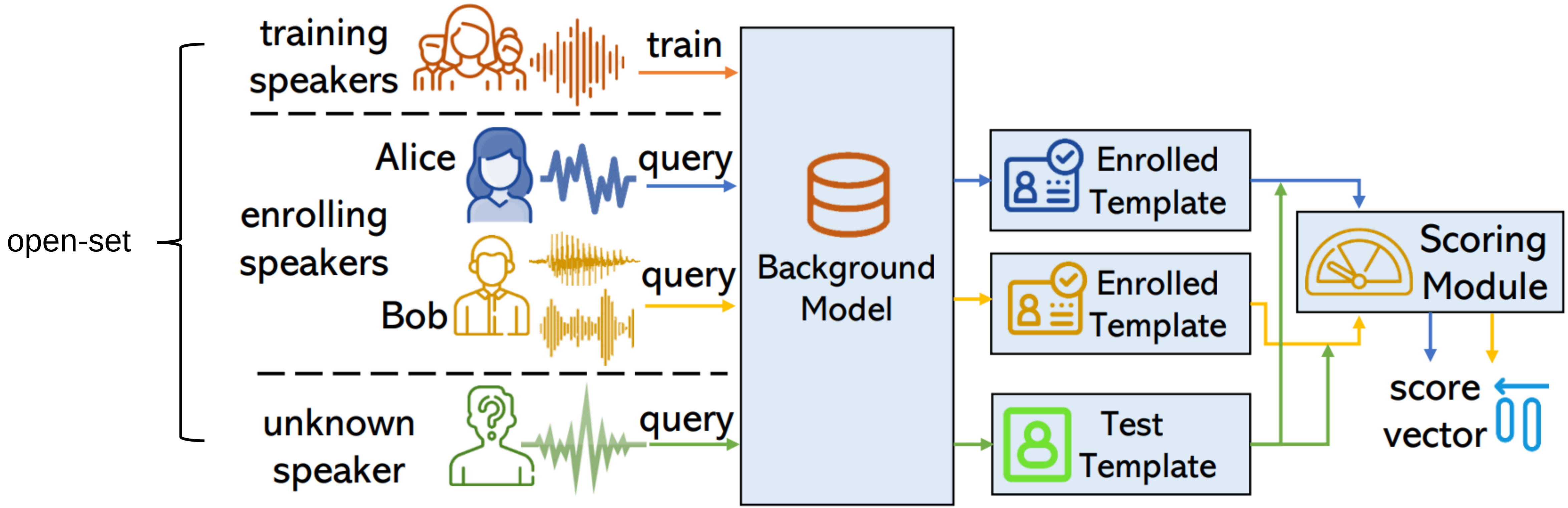
Speaker Recognition Systems (SRSs)

- Workflow:

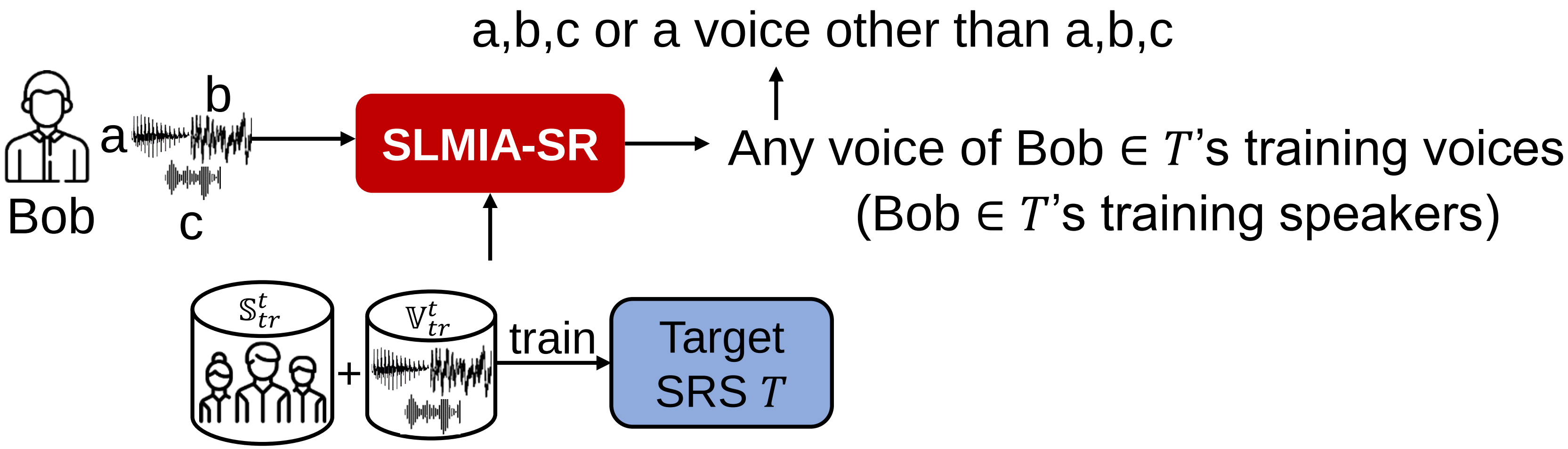


Speaker Recognition Systems (SRSs)

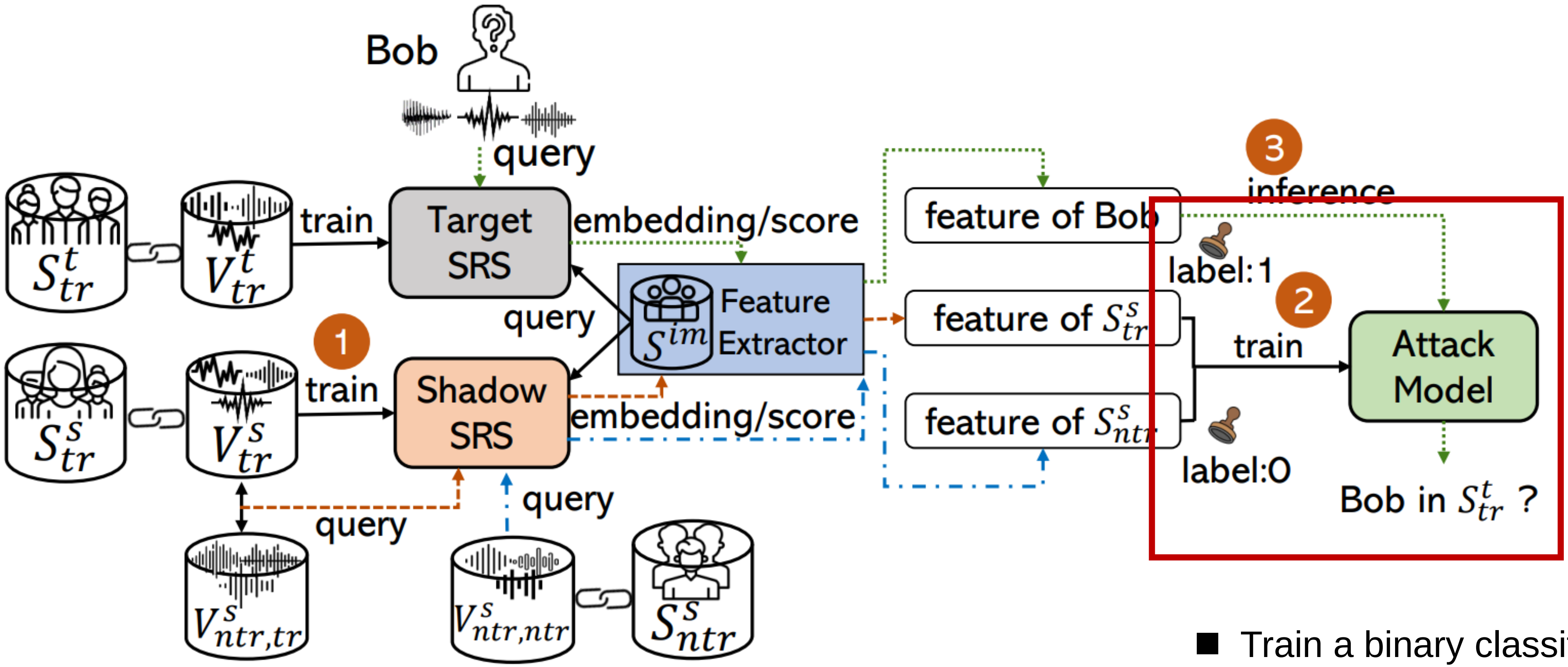
- Workflow:



Speaker-Level Membership Inference Attack (SLMIA-SR)

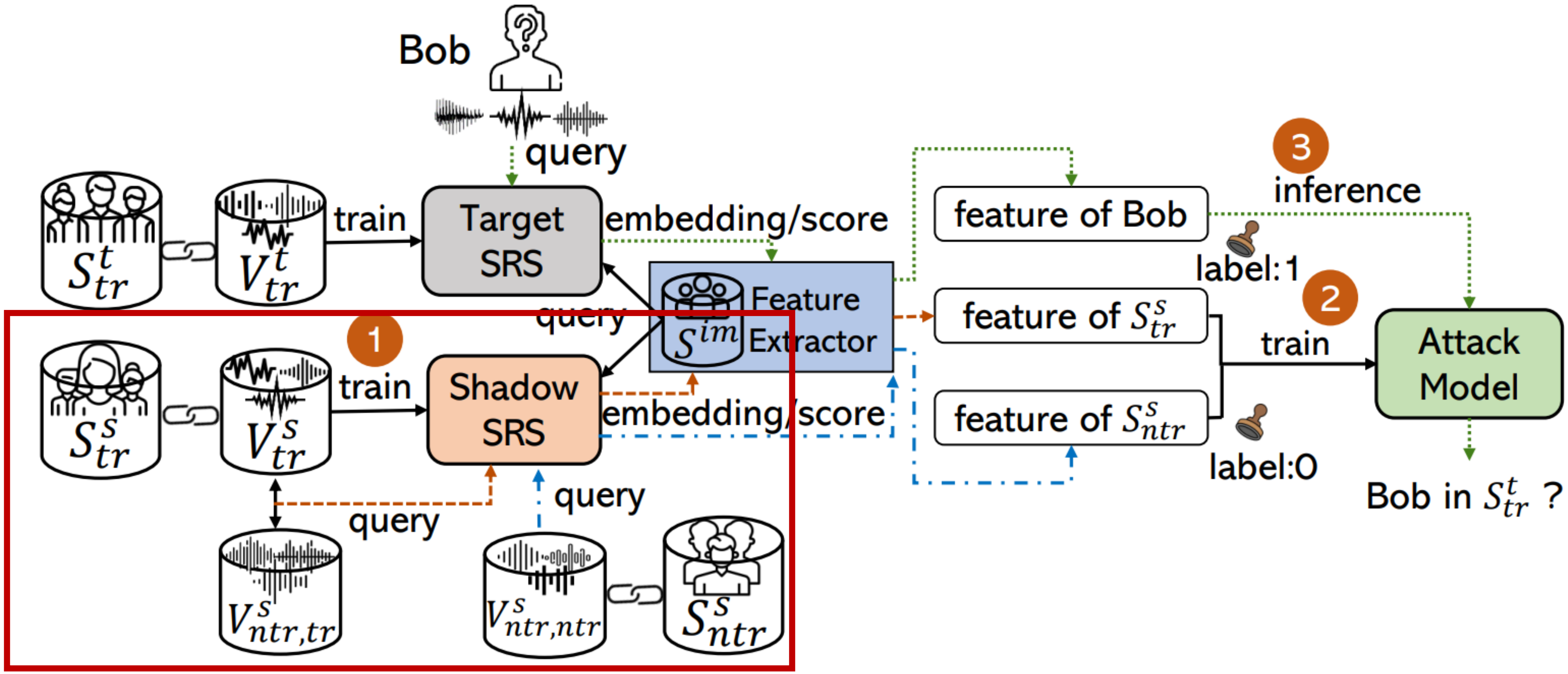


SLMIA-SR



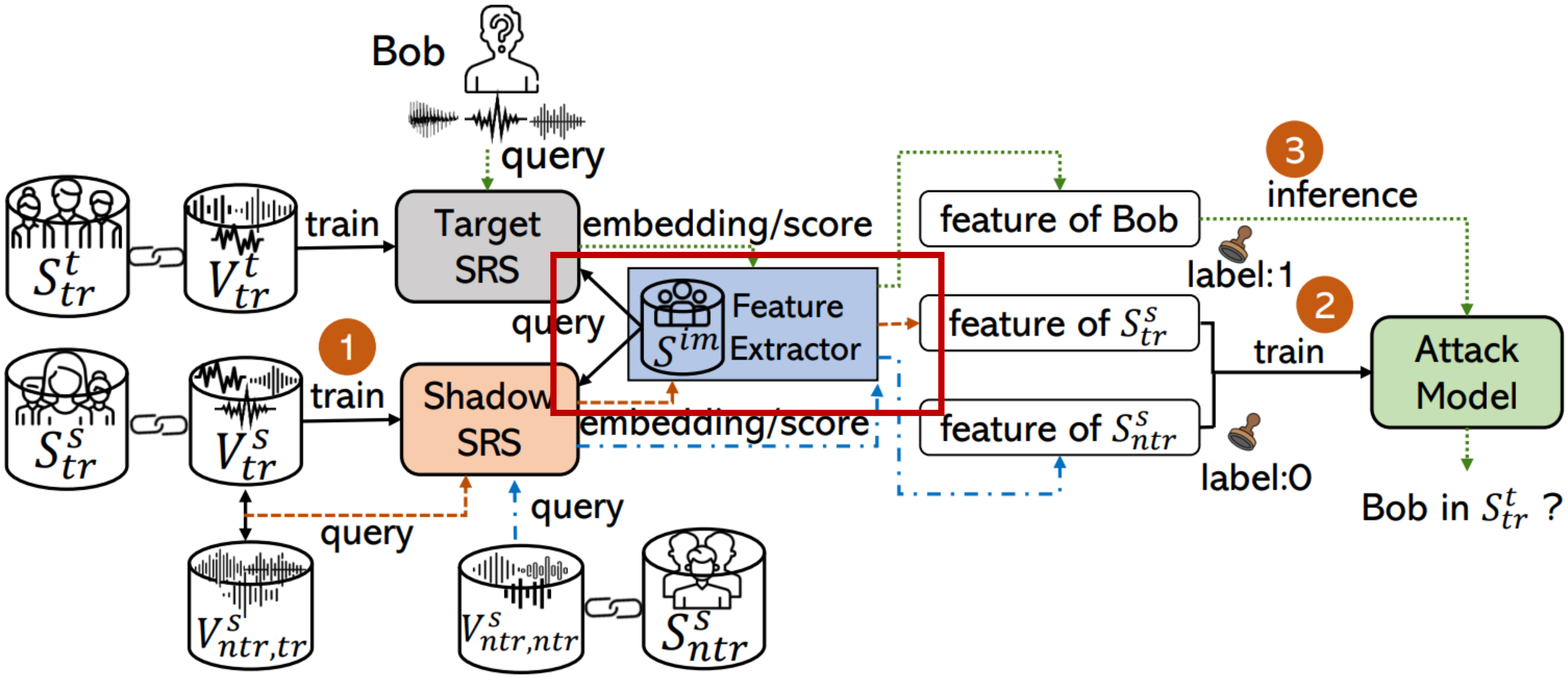
- Train a binary classifier

SLMIA-SR



- Supervised training: shadow SRS

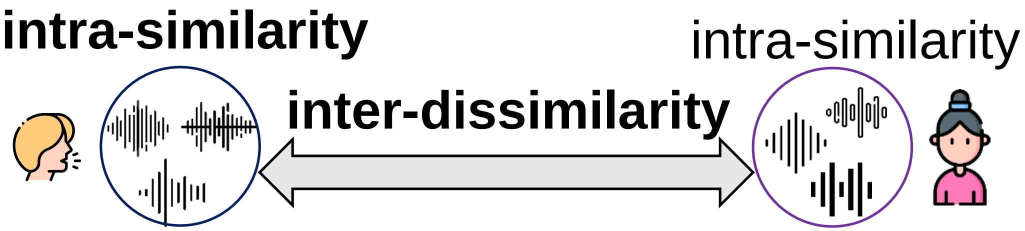
SLMIA-SR: Feature Engineering



- Feature engineering for binary classifier

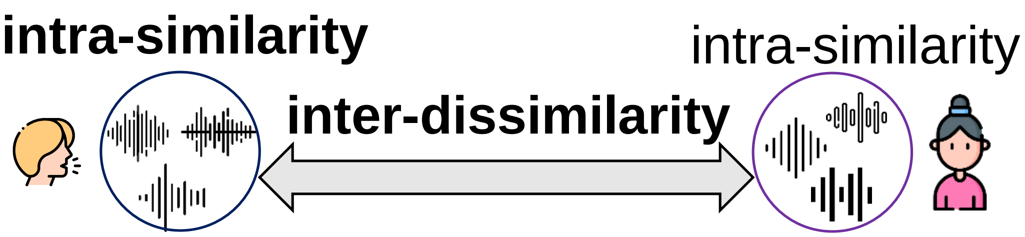
SLMIA-SR: Feature Engineering

- Training objectives of SRS



SLMIA-SR: Feature Engineering

- Training objectives of SRS

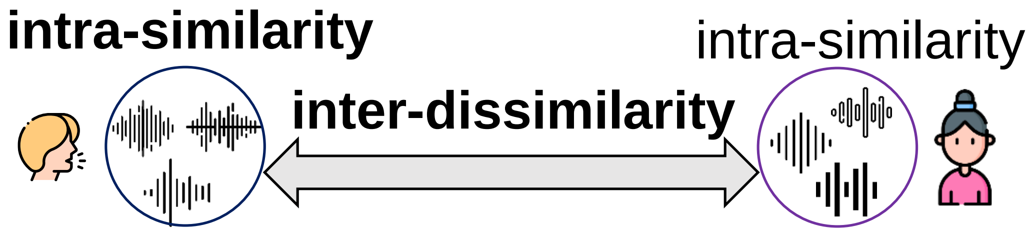


- hypothesis:

training speakers > non-training speakers

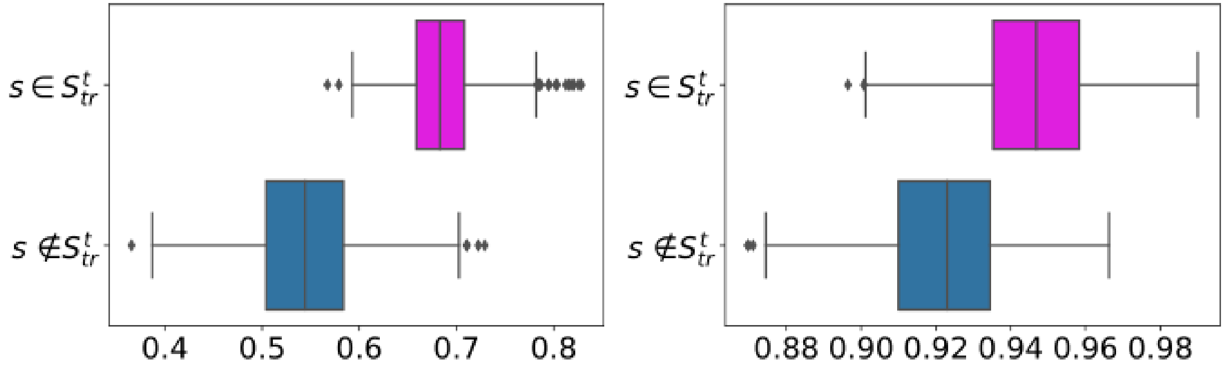
SLMIA-SR: Feature Engineering

- Training objectives of SRS



- hypothesis:

training speakers $>$ non-training speakers



(a) intra-similarity (b) inter-dissimilarity

Fig. 2: The comparison of intra-compactness and inter-farness between training and non-training speakers.

SLMIA-SR: Feature Engineering

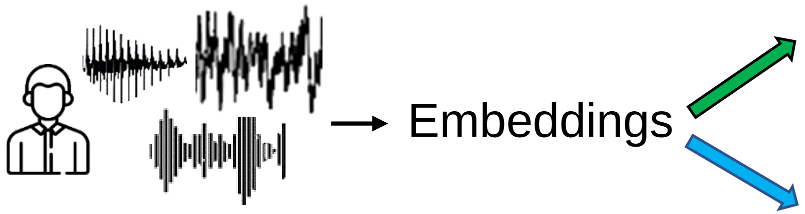
- Training objectives of SRS

intra-similarity intra-similarity

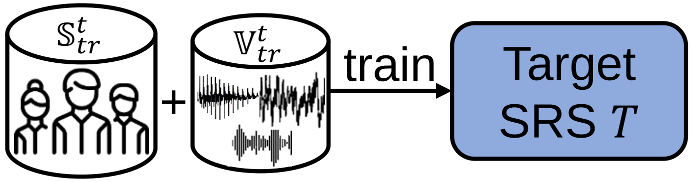
inter-dissimilarity

- hypothesis:

training speakers > non-training speakers

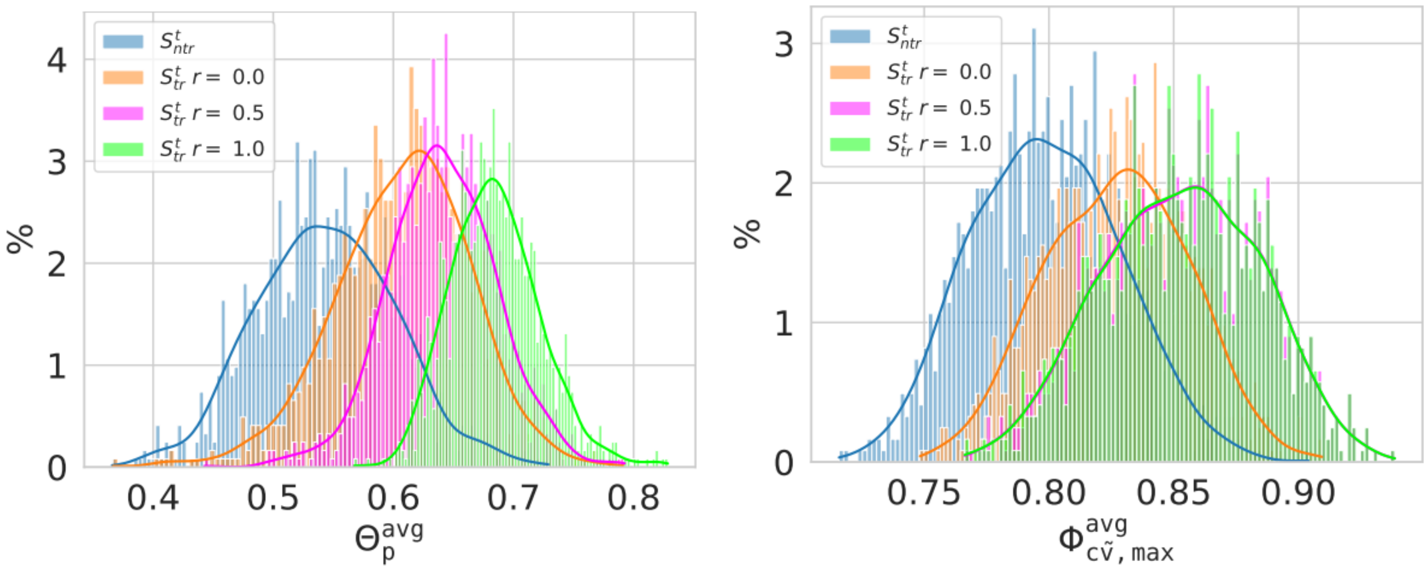


SLMIA-SR: Enhancement



$N: |S|$ $\quad$ $r: \dfrac{|S \cap V^t_{tr}|}{N}$

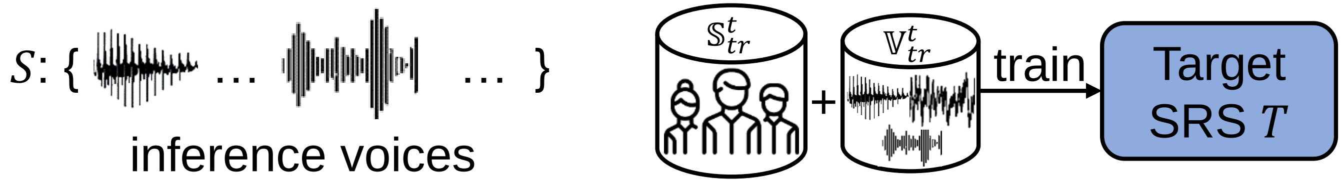
- Mixing ratio r training (of attack model)



(a) intra-similarity $\quad$ (b) inter-dissimilarity

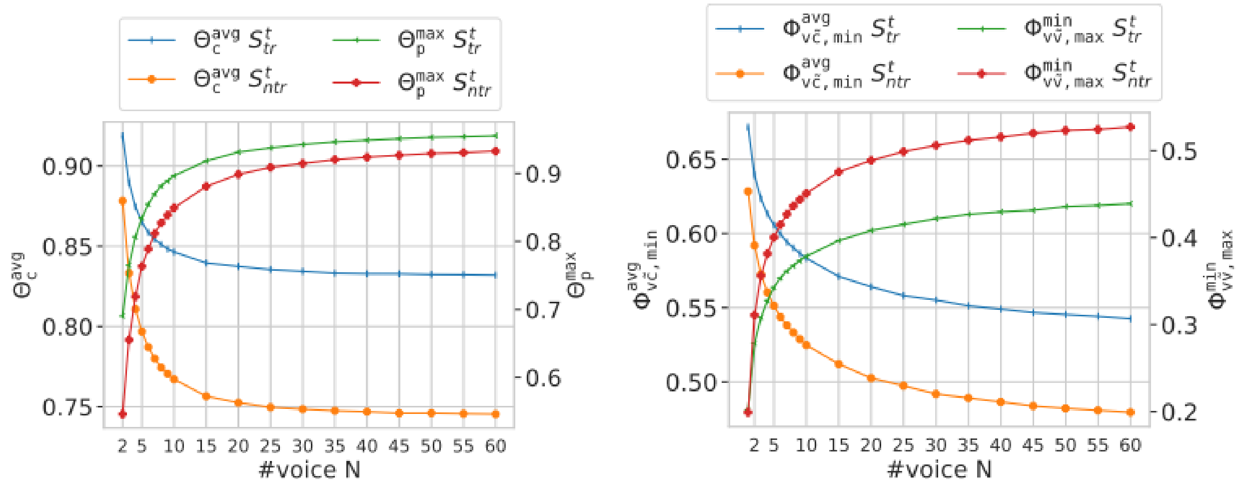
Fig. 4: Comparison of features with different r.

SLMIA-SR: Enhancement



$N: |S|$ $\quad r: \frac{|S \cap V^t_{tr}|}{N}$

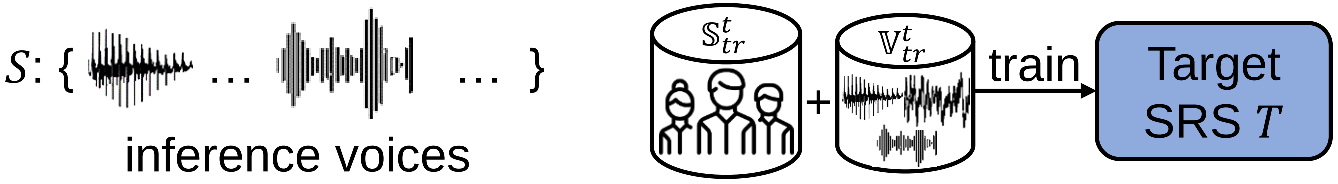
- N-dependent attack model



(a) intra-similarity (b) inter-dissimilarity

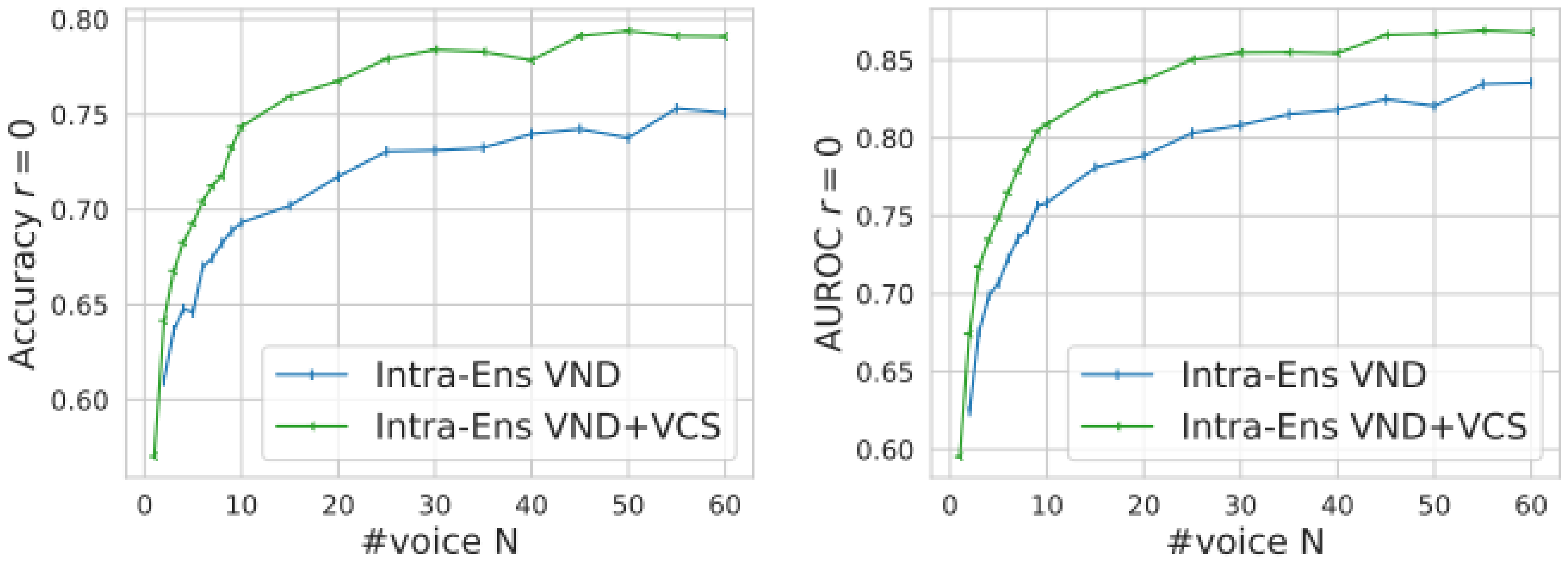
Fig. 5: Comparison of features with different N.

SLMIA-SR: Enhancement

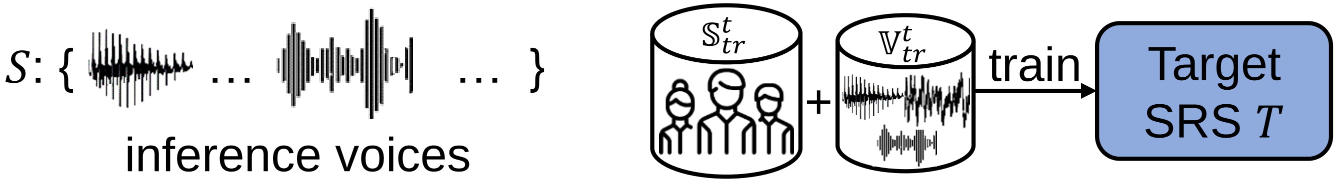


$N: |S|$ $\quad$ $r: \dfrac{|S \cap V^t_{tr}|}{N}$

- Voice Chunk Splitting

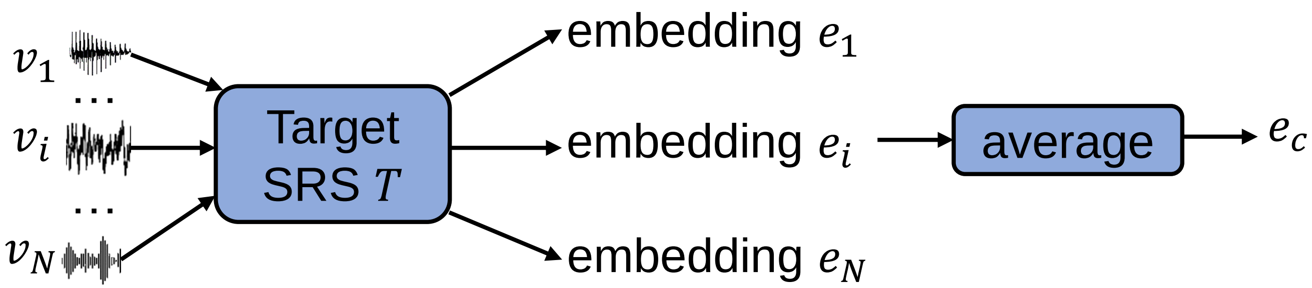


SLMIA-SR: Reducing #Query

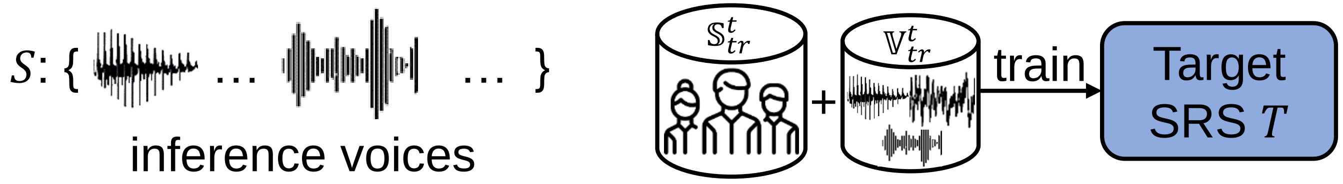


N: $|S|$ $\quad$ r: $\frac{|S \cap V^t_{tr}|}{N}$

- Enrollment Voice Concatenation

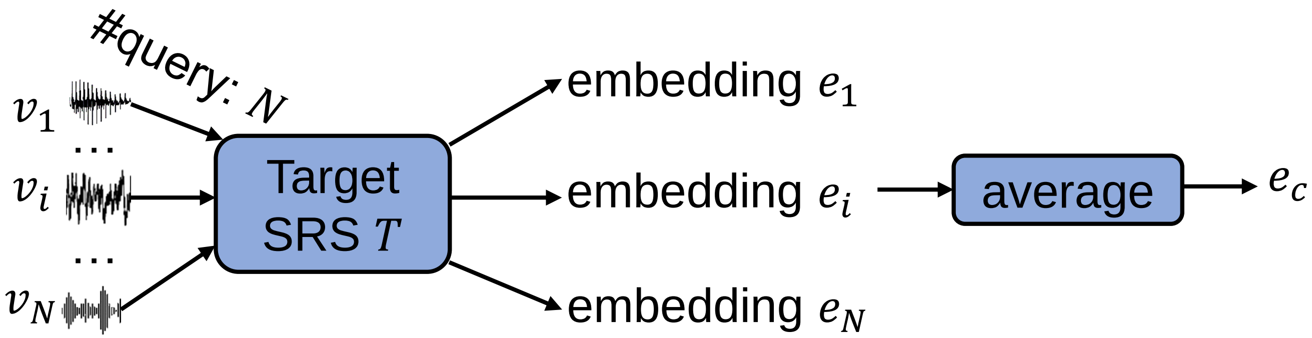


SLMIA-SR: Reducing #Query

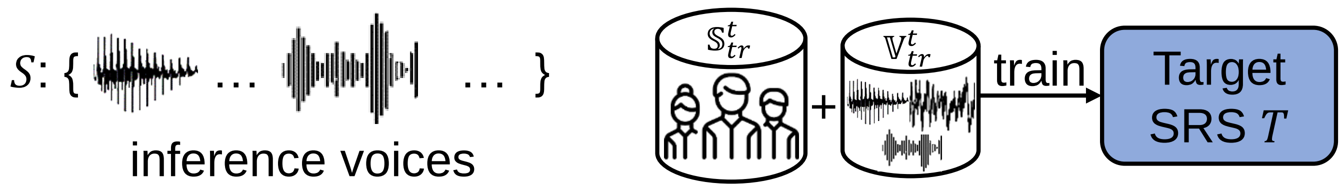


N: $|S|$ $\quad$ r: $\frac{|S \cap V_{tr}^{t}|}{N}$

- Enrollment Voice Concatenation

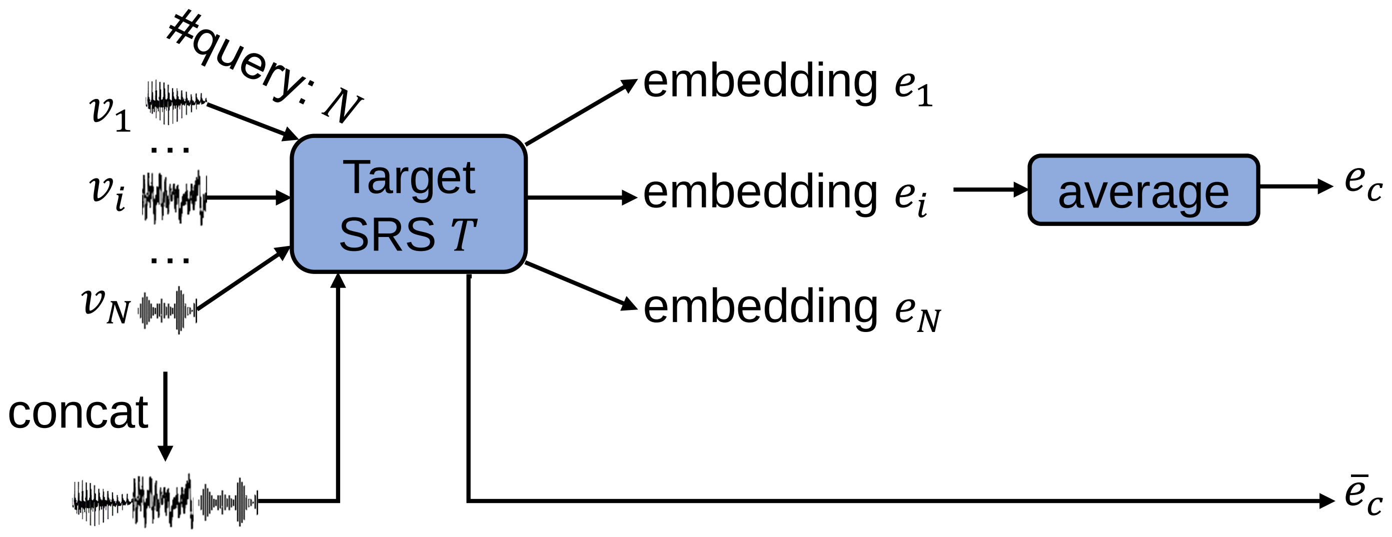


SLMIA-SR: Reducing #Query

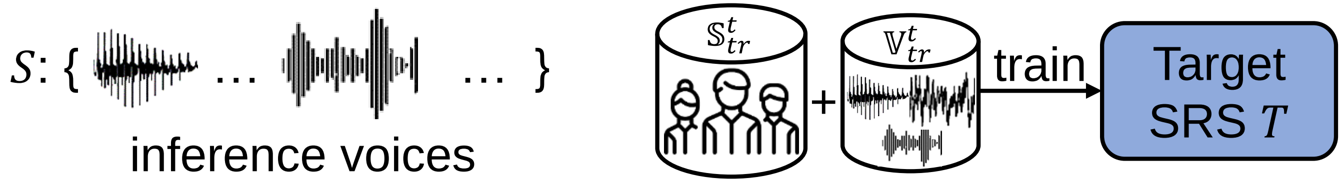


N: $|S|$ $\quad r$: $\dfrac{|S \cap V_{tr}^{t}|}{N}$

- Enrollment Voice Concatenation

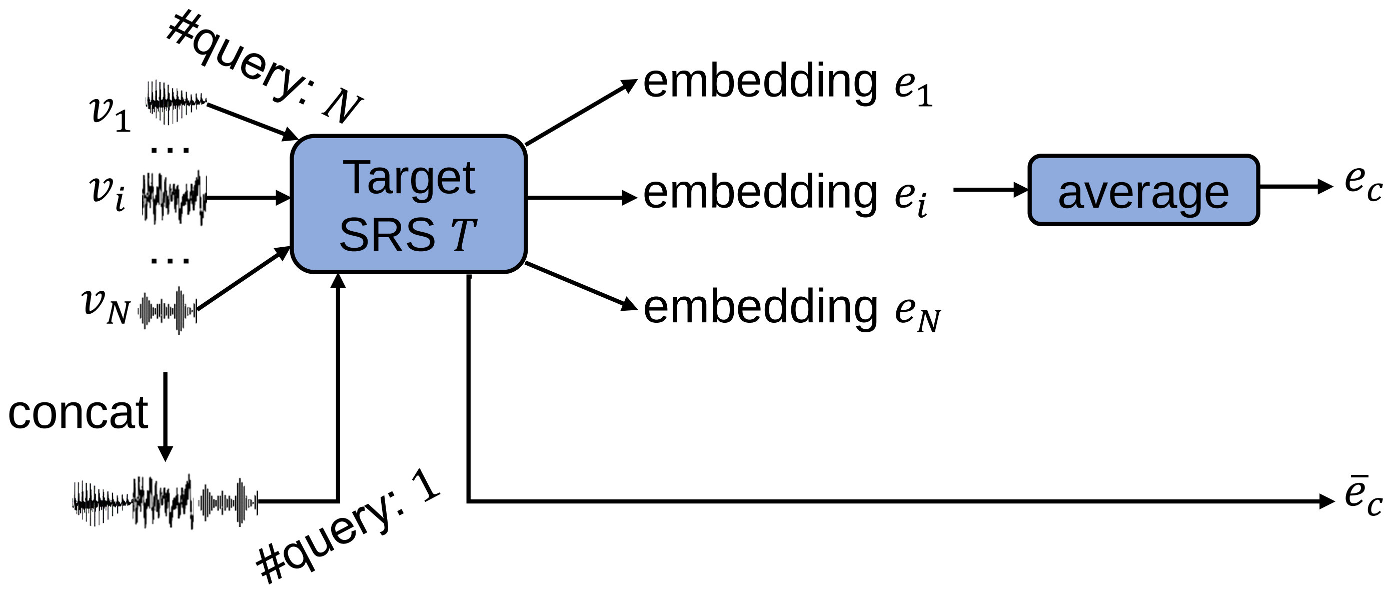


SLMIA-SR: Reducing #Query

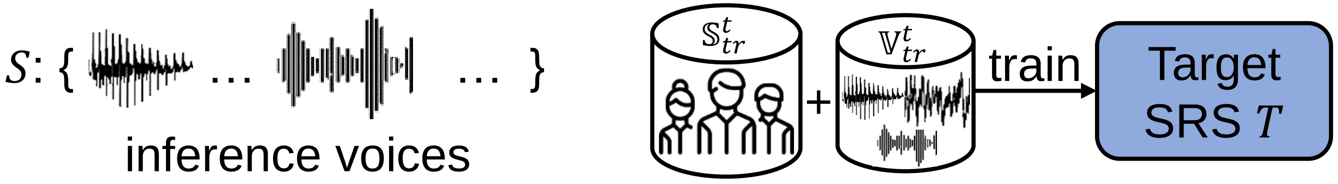


N: $|S|$ $\quad$ r: $\frac{|S \cap V_{tr}^t|}{N}$

■ Enrollment Voice Concatenation

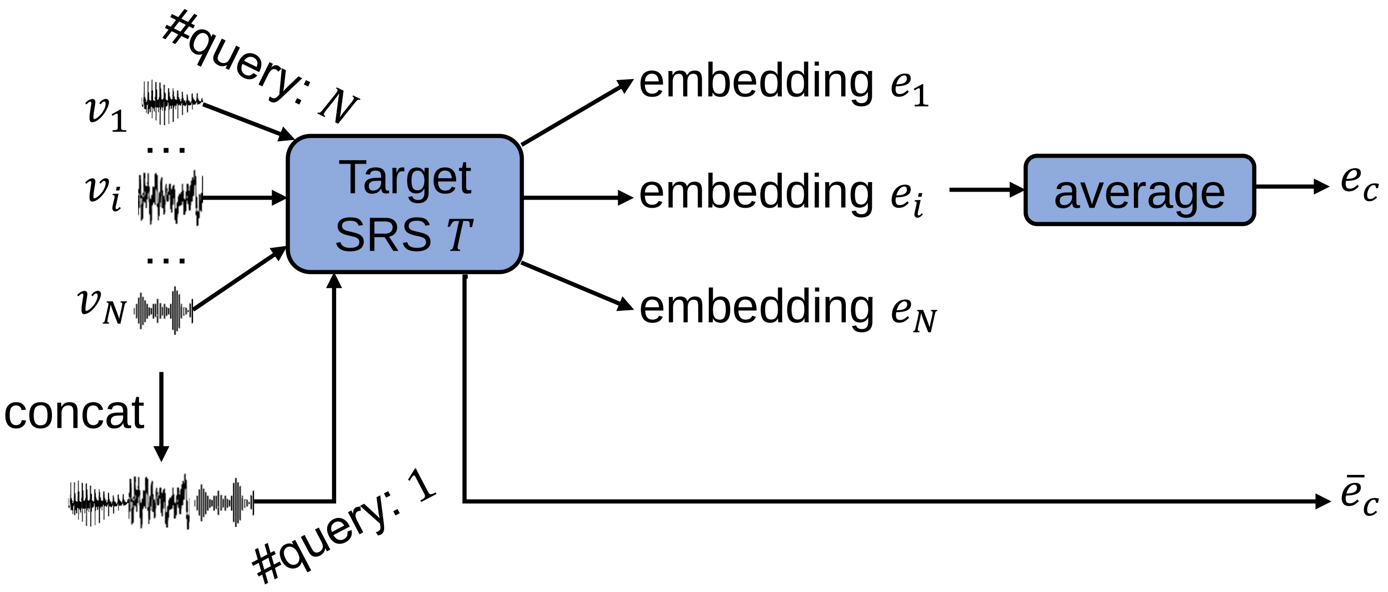


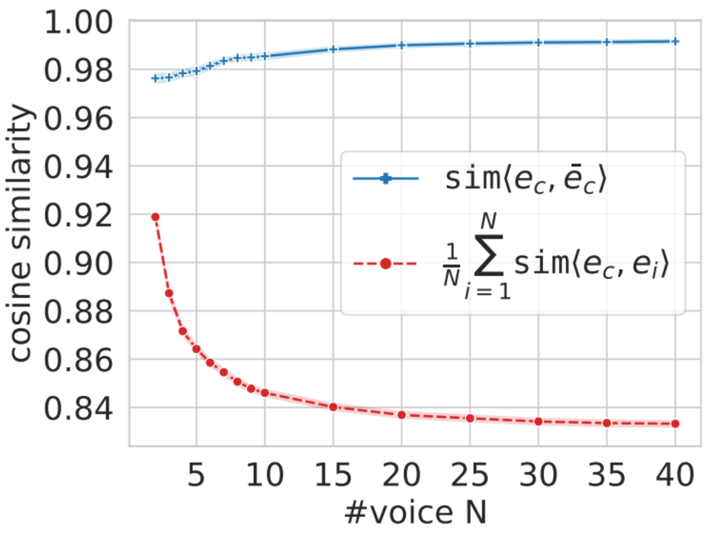
SLMIA-SR: Reducing #Query



$N: |S|$ $\quad r: \frac{|S \cap V^t_{tr}|}{N}$

- Enrollment Voice Concatenation





SLMIA-SR: Evaluation

- Model

Name	Architecture
LSTM-GE2E	LSTM [29], [53]
TDNN-CE	TDNN [54], [55]
Raw-AAM	RawNet3 [56], [57]
Res-AP	ResNetSE34V2 [58], [57]
VGG-GE2E	VGGVox40 [59], [57]

- Dataset: VoxCeleb-2 (en), LibriSpeech (en), KeSpeech (zh)

- Metric: Accuracy, AUROC, True Positive Rate (TPR) at x% False Positive Rate (FPR)

- Baseline: LRL-MIA [1], EncoderMI [2], TLK-MIA [3], FaceAuditor [4]

[1] User-level membership inference attack against metric embedding learning
[2] Membership inference attacks against self-supervised speech models
[3] EncoderMI: Membership inference against pre-trained encoders in contrastive learning
[4] FACEAUDITOR: data auditing in facial recognition systems

SLMIA-SR: Overall Performance

		Accuracy			AUROC			TPR @ x% FPR			TPR @ 1% FPR		
		VC-2	LS	KS	VC-2	LS	KS	VC-2 (x=0.1)	LS (x=0.2)	KS (x=0.1)	VC-2	LS	KS
LSTM-GE2E	LRL-MIA	0.698	0.895	0.669	0.789	0.952	0.767	1.5%	31.7%	1.7%	8.1%	41.7%	8.4%
	EncoderMI-T	0.7	0.877	0.592	0.792	0.952	0.768	1.6%	31.7%	1.8%	7.8%	41.7%	8.5%
	TLK-MIA	0.7	0.877	0.592	0.792	0.952	0.768	1.6%	31.7%	1.8%	7.8%	41.7%	8.5%
	SLMIA-SR	**0.894**	**0.974**	**0.785**	**0.958**	**0.994**	**0.880**	**33.5%**	**66.5%**	**10.5%**	**58.7%**	**83.5%**	**28.1%**
TDNN-CE	LRL-MIA	0.82	0.723	0.595	0.906	0.791	0.676	20.1%	0.6%	1.8%	46.6%	6.7%	4.6%
	EncoderMI-T	0.779	0.713	0.583	0.904	0.791	0.668	20.8%	0.6%	0.8%	45.9%	6.7%	4.0%
	TLK-MIA	0.779	0.713	0.583	0.904	0.791	0.668	20.8%	0.6%	0.8%	45.9%	6.7%	4.0%
	SLMIA-SR	**0.891**	**0.83**	**0.679**	**0.965**	**0.897**	**0.761**	**33.8%**	**11.1%**	**5.2%**	**64.4%**	**21.9%**	**11.1%**
Raw-AAM	LRL-MIA	0.705	0.679	0.622	0.786	0.732	0.676	1.6%	0.8%	0.4%	9.6%	2.7%	6.1%
	EncoderMI-T	0.703	0.661	0.601	0.785	0.732	0.656	1.9%	0.8%	0.2%	9.8%	2.7%	1.9%
	TLK-MIA	0.703	0.661	0.601	0.785	0.732	0.656	1.9%	0.8%	0.2%	9.8%	2.7%	1.9%
	SLMIA-SR	**0.749**	**0.783**	**0.689**	**0.856**	**0.856**	**0.754**	**5.6%**	**6.8%**	**3.0%**	**18.3%**	**12.7%**	**9.0%**
Res-AP	LRL-MIA	0.756	0.924	0.627	0.842	0.974	0.740	8.8%	6.6%	1.0%	24.4%	64.5%	7.4%
	EncoderMI-T	0.747	0.887	0.606	0.841	0.974	0.740	8.8%	6.6%	1.0%	24.3%	64.7%	7.4%
	TLK-MIA	0.747	0.887	0.606	0.841	0.974	0.740	8.8%	6.6%	1.0%	24.3%	64.7%	7.4%
	SLMIA-SR	**0.799**	**0.956**	**0.699**	**0.892**	**0.986**	**0.796**	**12.5%**	**14.4%**	**5.1%**	**40.2%**	**72.3%**	**11.0%**
VGG-GE2E	LRL-MIA	0.714	0.847	0.592	0.783	0.916	0.634	5.6%	9.8%	0.2%	17.2%	15.4%	2.8%
	EncoderMI-T	0.711	0.827	0.574	0.785	0.916	0.624	5.5%	9.8%	0.1%	17.4%	15.4%	2.2%
	TLK-MIA	0.711	0.827	0.574	0.785	0.916	0.624	5.5%	9.8%	0.1%	17.4%	15.4%	2.2%
	SLMIA-SR	**0.743**	**0.914**	**0.648**	**0.835**	**0.968**	**0.700**	**16.6%**	**22.1%**	**1.6%**	**26.4%**	**45.9%**	**5.0%**

		Accuracy			AUROC			TPR @ x% FPR					
								x=0.1	x=0.2	x=0.1	x=1	x=1	x=1
		VC-2	LS	KS	VC-2	LS	KS	VC-2	LS	KS	VC-2	LS	KS
LSTM-GE2E	EncoderMI-V	0.649	0.866	0.632	0.72	0.932	0.770	2.0%	19.2%	4.4%	6.6%	35.2%	10.5%
	FaceAuditor-S	0.655	0.842	0.698	0.714	0.932	0.768	1.2%	16.8%	1.6%	5.3%	33.0%	7.3%
	FaceAuditor-P/R	0.614	0.768	0.615	0.691	0.863	0.773	1.6%	3.8%	2.8%	6.2%	14.5%	9.9%
	SLMIA-SR	**0.785**	**0.976**	**0.794**	**0.861**	**0.994**	**0.885**	**7.2%**	**62.1%**	**13.6%**	**24.4%**	**82.7%**	**24.4%**
TDNN-CE	EncoderMI-V	0.724	0.703	0.603	0.81	0.772	0.681	19.6%	1.1%	1.0%	28.2%	5.6%	4.8%
	FaceAuditor-S	0.784	0.666	0.604	0.866	0.742	0.639	20.1%	2.1%	0.1%	34.2%	4.9%	1.6%
	FaceAuditor-P/R	0.772	0.578	0.512	0.866	0.628	0.562	11.9%	0.3%	0.2%	30.9%	1.4%	1.9%
	SLMIA-SR	**0.839**	**0.773**	**0.661**	**0.92**	**0.856**	**0.733**	**23.6%**	**2.8%**	**1.7%**	**42.9%**	**8.1%**	**6.6%**
Raw-AAM	EncoderMI-V	0.657	0.657	0.606	0.709	0.708	0.658	2.7%	0.9%	0.2%	8.7%	2.0%	2.3%
	FaceAuditor-S	0.636	0.64	0.598	0.686	0.702	0.640	0.2%	0.2%	0.3%	2.3%	2.0%	1.8%
	FaceAuditor-P/R	0.663	0.592	0.514	0.732	0.659	0.584	3.5%	0.4%	0.2%	6.7%	2.8%	1.3%
	SLMIA-SR	**0.697**	**0.764**	**0.650**	**0.774**	**0.827**	**0.701**	**3.8%**	**3.1%**	**1.0%**	**8.8%**	**5.1%**	**2.8%**
Res-AP	EncoderMI-V	0.712	0.87	0.599	0.789	0.948	0.730	4.9%	33.1%	3.0%	13.7%	41.3%	6.4%
	FaceAuditor-S	0.722	0.885	0.678	0.794	0.961	0.750	4.9%	28.3%	3.1%	14.9%	52.5%	8.8%
	FaceAuditor-P/R	0.672	0.697	0.549	0.744	0.771	0.630	4.2%	4.2%	0.6%	11.6%	8.7%	3.1%
	SLMIA-SR	**0.763**	**0.932**	**0.692**	**0.841**	**0.982**	**0.782**	**13.6%**	**43.0%**	**4.6%**	**29.0%**	**60.7%**	**10.9%**
VGG-GE2E	EncoderMI-V	0.692	0.797	0.584	0.756	0.878	0.634	1.0%	4.7%	0.5%	11.8%	9.9%	2.0%
	FaceAuditor-S	0.671	0.797	0.560	0.728	0.89	0.579	0.4%	4.8%	0.1%	5.9%	14.4%	0.6%
	FaceAuditor-P/R	0.605	0.648	0.527	0.685	0.71	0.566	2.1%	1.1%	0.5%	7.3%	2.8%	1.9%
	SLMIA-SR	**0.708**	**0.934**	**0.637**	**0.776**	**0.98**	**0.686**	**6.3%**	**46.7%**	**0.8%**	**15.5%**	**65.8%**	**3.5%**

Note: VC-2, LS, and KS denote the dataset VoxCeleb-2, LibriSpeech, and KeSpeech, respectively. The ratio $r = 0$.

SLMIA-SR: Ablation Studies

- Disjoint Architectures

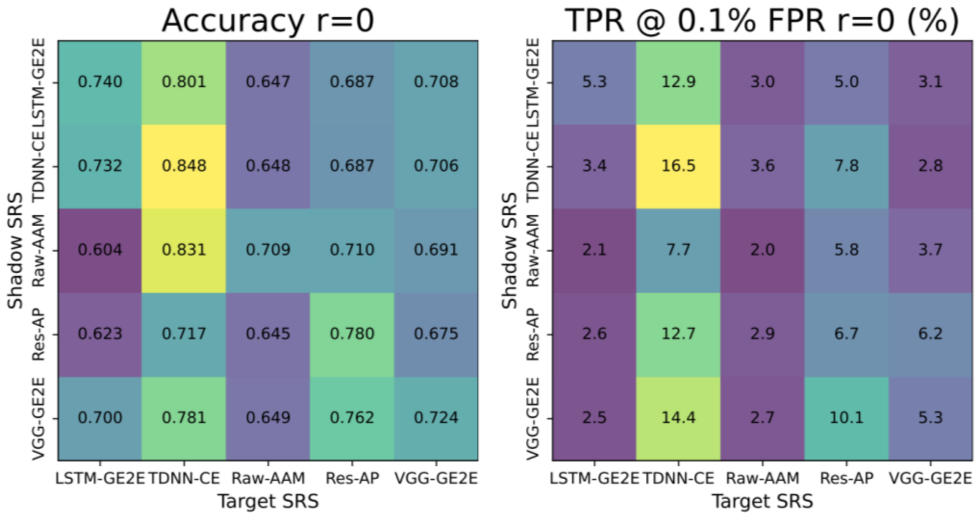


Fig. 16: Effect of the architectures.

- Disjoint Dataset Distribution

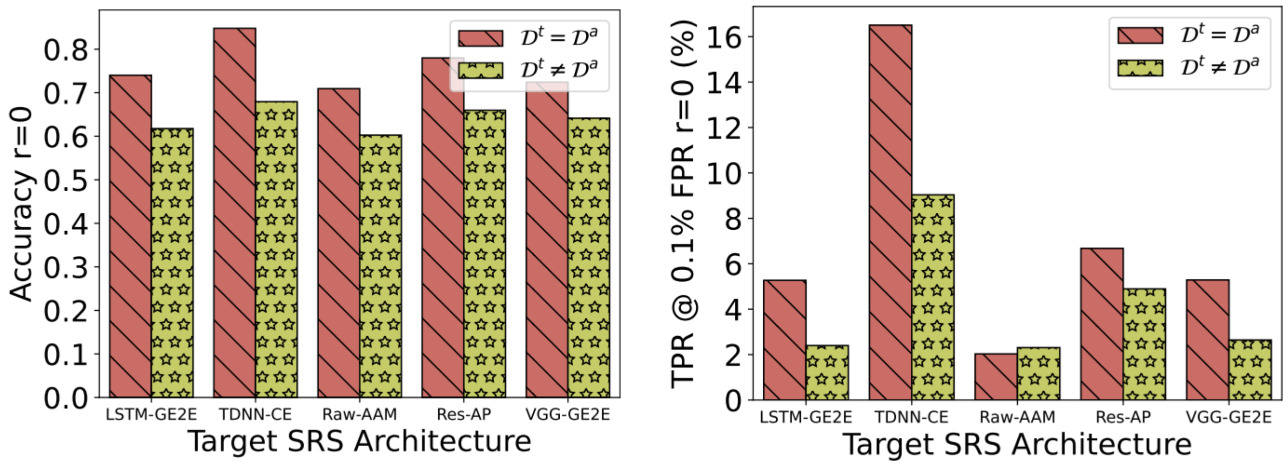


Fig. 15: Effect of the dataset distribution.

SLMIA-SR: Application & Potential Impact

- Ordinary users: Voice data auditing

Amazon sued over Alexa child recordings in US

13 June 2019

- System maintainers: Accessing privacy level of speaker recognition services before publishing

Take away

- Speaker-level membership inference attack against speaker recognition systems
- Distinguish training and non-training speakers by intra-similarity & inter-dissimilarity
- 103 features to launch attacks
- Three strategies to enhance attacks
- Strategy to reduce #queries
- Effective even for disjoint dataset distributions and architectures

Code: https://github.com/S3L-official/SLMIA-SR
Paper: https://www.ndss-symposium.org/ndss-paper/slmia-sr-speaker-level-membership-inference-attacks-against-speaker-recognition-systems

Any Question?
Thanks!

Guangke Chen

chengk@shanghaitech.edu.cn

https://guangkechen.site

Trustworthy Artificial Intelligence

- expected to graduate (Ph.D.) in June 2024
- on the job market
- Consider dropping emails for opportunities

- Publications:

[1] Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems

Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Fu Song, Yang Liu

S&P (Oakland) 2021. Citation>185

[2] QFA2SR: Query-Free Adversarial Transfer Attacks to Speaker Recognition Systems

Guangke Chen, Yedi Zhang, Zhe Zhao, Fu Song

USENIX Security 2023

[3] SLMIA-SR: Speaker-Level Membership Inference Attacks on Speaker Recognition
Guangke Chen, Yedi Zhang, Fu Song
NDSS 2024

[4] Towards Understanding and Mitigating Audio Adversarial Examples for Speaker Recognition

Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, Jiashui Wang

IEEE TDSC

[5] AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems

Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Yang Liu

IEEE TDSC