32ND USENIX SECURITY SYMPOSIUM

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

Guangke Chen, Yedi Zhang, Zhe Zhao, Fu Song ✉

上海科技大学
ShanghaiTech University

Guangke Chen: https://guangkechen.site
Fu Song: songfu1983@gmail.com

Voiceprint Recognition (VPR)

Identify a person by his/her speeches, a.k.a., speaker recognition

- identity verification in banks' telephone-communication

Citi Uses Voice Prints To Authenticate Customers Quickly And Effortlessly

Citi Bank

What is TD VoicePrint and how do I enroll?

TD VoicePrint is a voice recognition security technology we can use to verify your identity whenever you call us. Your voiceprint, like your fingerprint, is unique to you – no one else has a voice just like you.

Enroll today

- Call Live Customer Service 1-888-751-9000
- Request to enroll in TD VoicePrint
- The customer service representative will get you set up

TD Bank

Voiceprint Recognition (VPR)

Identify a person by his/her speeches, a.k.a., speaker recognition

- password-free payment

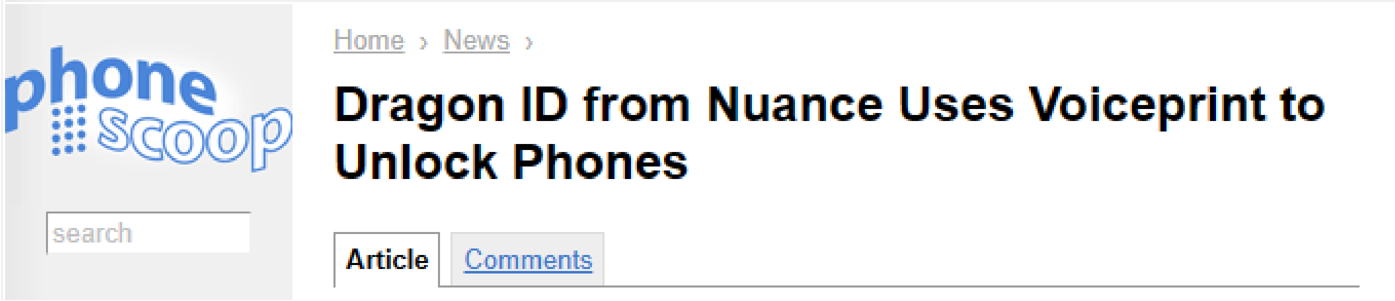
Customer	"Tmall Genie, I'd like to order a mobile refill card."
Genie	"Master, I would recommend China Mobile's refill card. The total price is 100 RMB. It will be delivered to (address). May I place the order for you?"
Customer	"Yes, please place the order."
Genie	"Sure! In order to proceed, let us do voice authentication first. Please keep quiet around, and after the 'beep', say 'Tmall Genie, 2065.'" (Here 2065 is the authentication code randomly generated by the system.)
Customer	"2065".
Genie	"Alipay discount is applied. If you want to know the delivery status, you can let me know by saying 'Tmall Genie, tracking information.'"

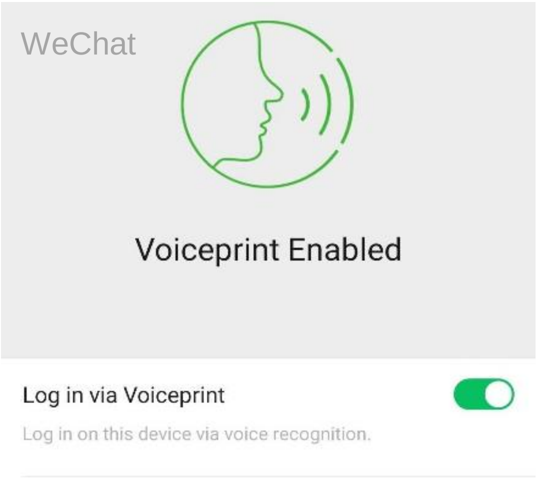
Voiceprint Recognition (VPR)

Identify a person by his/her speeches, a.k.a., speaker recognition

- access control in smart home, smartphones, and mobile applications

The applications of Junlin's voiceprint recognition solutions on Smart Household Appliances can realize user permission management to distinguish different family members' permission to different appliances. For example, the parents could be able to control all appliances, while the children can only control the appliances inliving room and children's room, which makes it easy, convenient, safe and comfortable to control the whole house by voice.





Voiceprint Recognition (VPR)

Identify a person by his/her speeches, a.k.a., speaker recognition

- key-word detection of voice assistants

Teach Google Assistant to recognize your voice with Voice Match

When you turn on Voice Match, you can teach Google Assistant to recognize your voice so it can verify who you are before it gives you personal results. You can turn on Voice Match for a

What Is Alexa Voice ID?

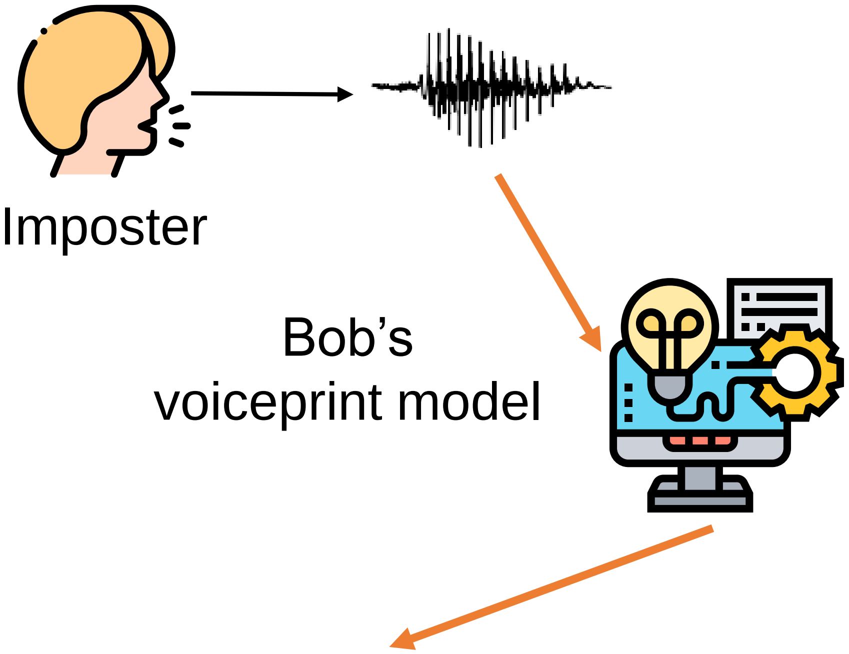
Alexa voice ID helps Alexa recognize you when you speak and provide a personalized experience.

Set up voice recognition and Personal Requests

When you set up voice recognition, Siri can recognize multiple voices, so that everyone in your home can enjoy personalized music and media. When you set up Personal Requests, you can do even more with voice recognition—like send and read messages, check your calendar, make phone calls, and more.

Speech Adversarial Examples against VPR

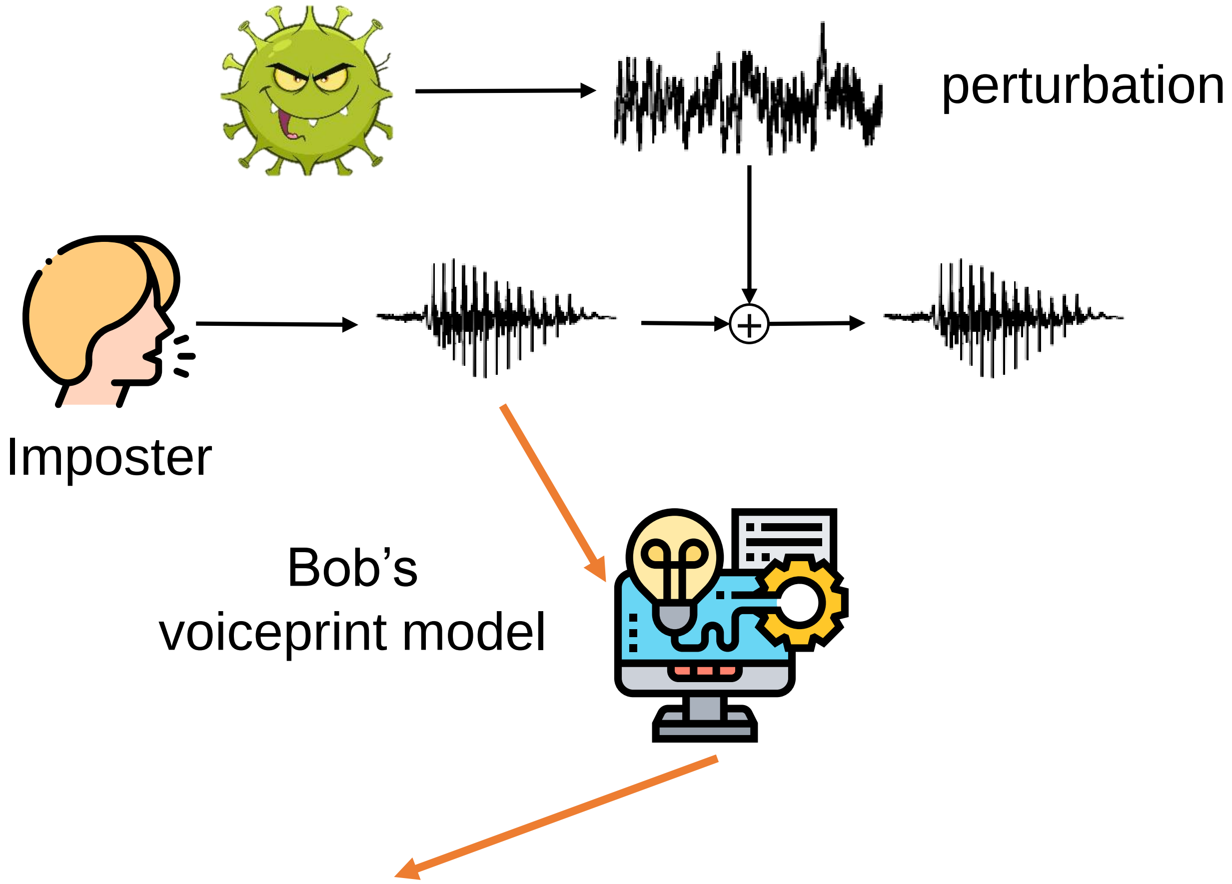
Attack vectors: replay, voice synthesis/conversion,
hidden speech, backdoor, *adversarial attacks*



“It is *not* uttered by Bob”

Speech Adversarial Examples against VPR

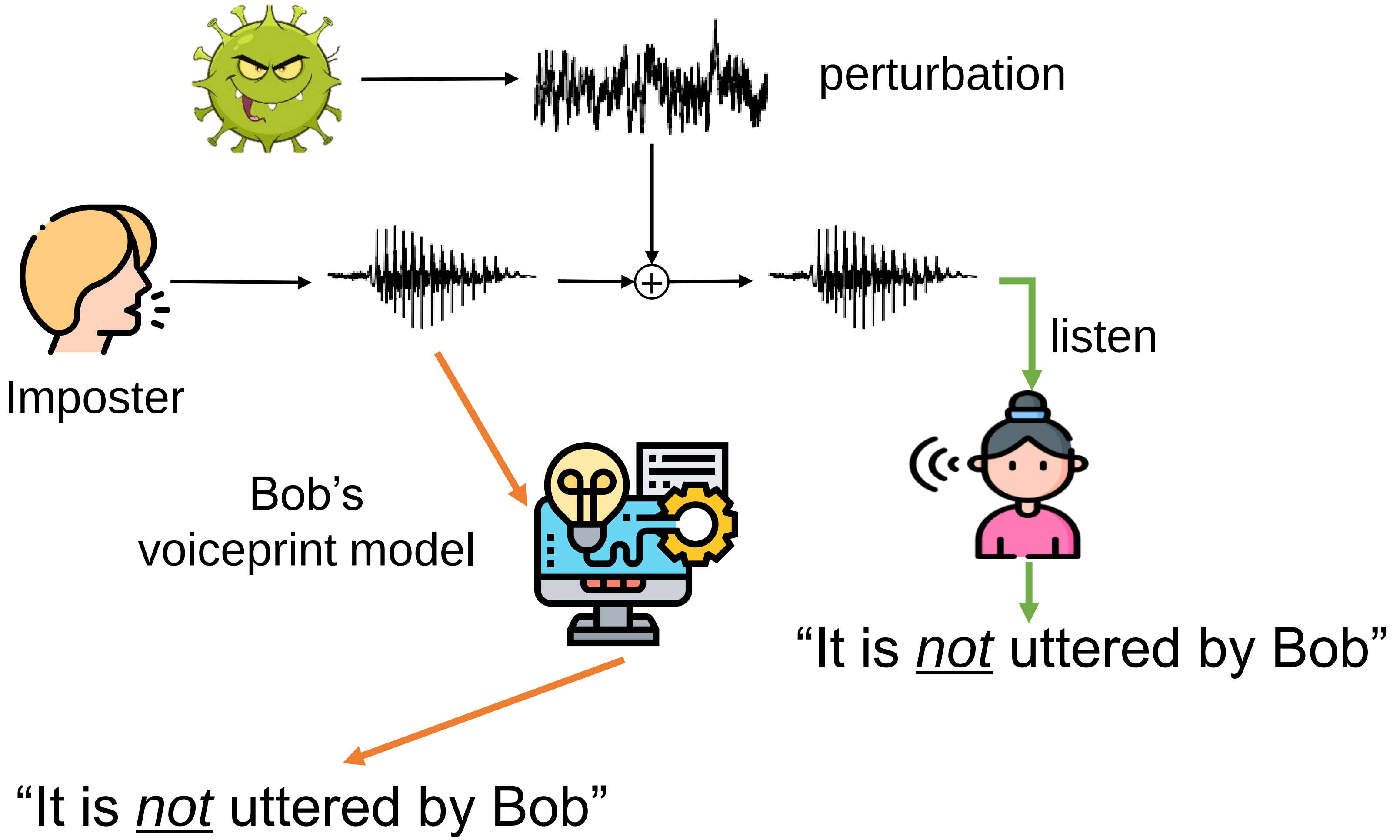
Attack vectors: replay, voice synthesis/conversion,
hidden speech, backdoor, *adversarial attacks*



"It is *not* uttered by Bob"

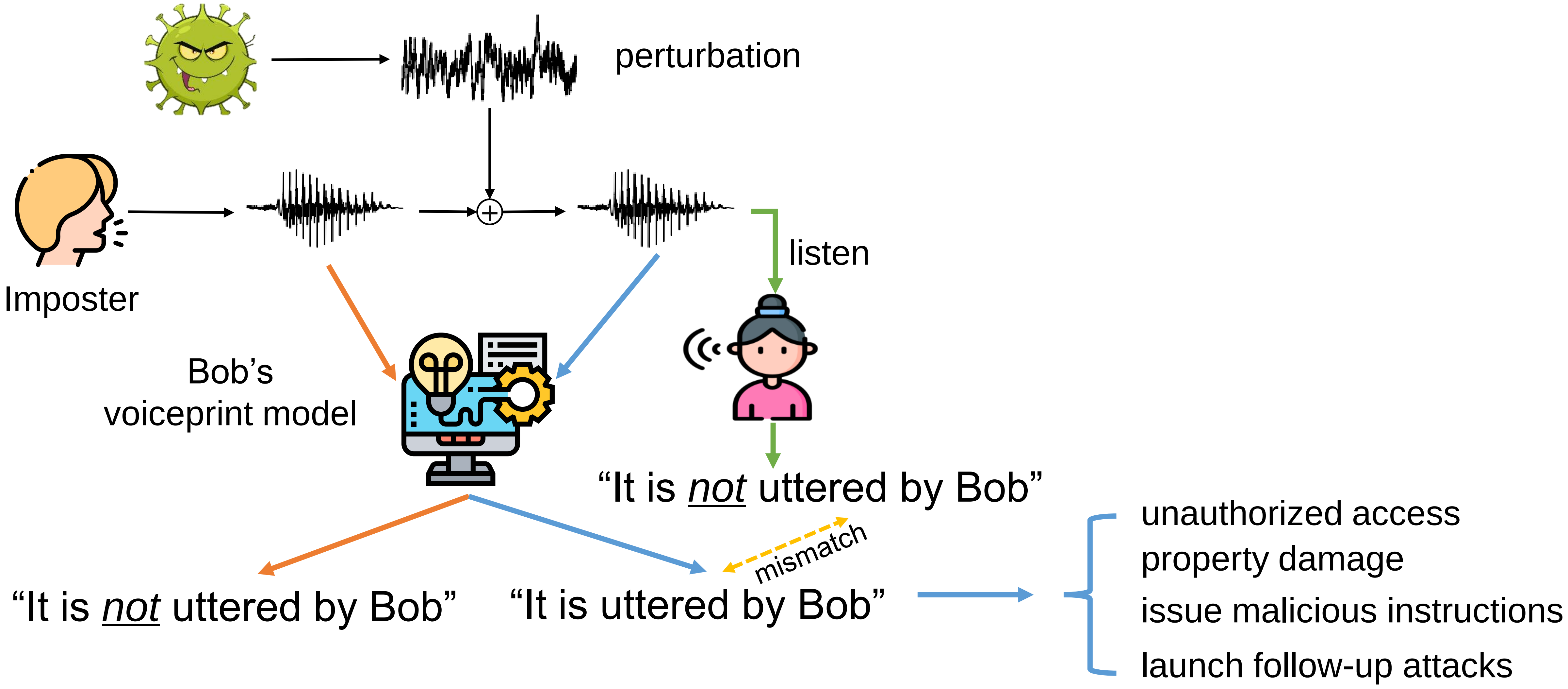
Speech Adversarial Examples against VPR

Attack vectors: replay, voice synthesis/conversion,
hidden speech, backdoor, _*adversarial attacks*_



Speech Adversarial Examples against VPR

Attack vectors: replay, voice synthesis/conversion,
hidden speech, backdoor, *adversarial attacks*



Query-free black-box attack: QFA2SR

Motivation:
White-box: unpractical
Query-based black-box: charges; frequency limit; no exposed query APIs

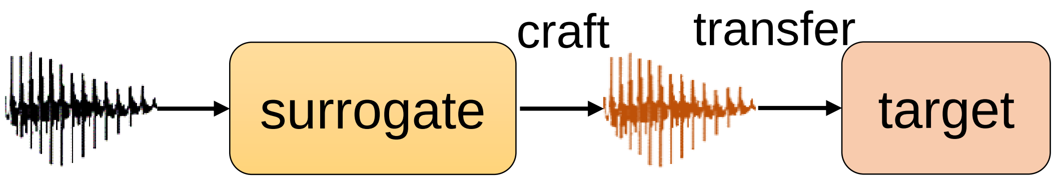
Threat model: Black-box & Query-free

Query-free black-box attack: QFA2SR

Motivation:
White-box: unpractical
Query-based black-box: charges; frequency limit; no exposed query APIs

Threat model: Black-box & Query-free

Solution:
Leverage transferability



Query-free black-box attack: QFA2SR

Motivation:
White-box: unpractical
Query-based black-box: charges; frequency limit; no exposed query APIs

Threat model: Black-box & Query-free

Solution:
Leverage transferability

Challenge:
Transferability of speech adversarial examples is extremely low

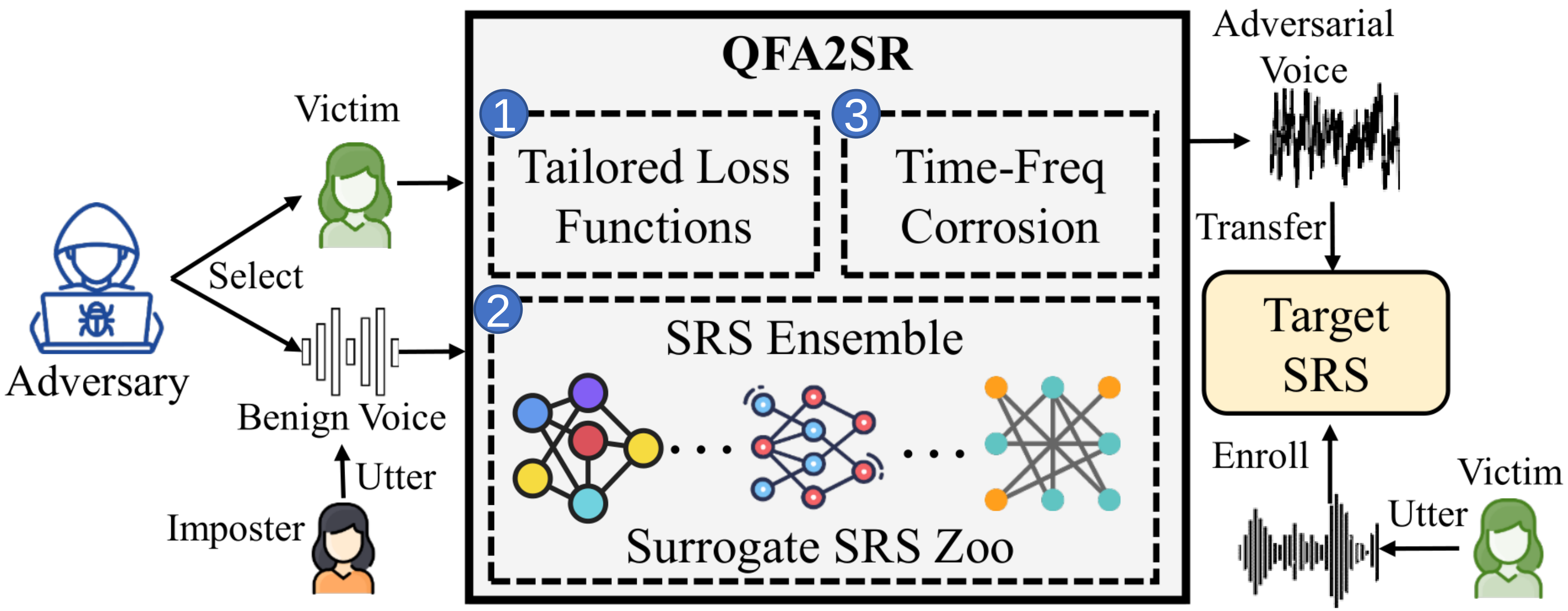
Transfer rate

-	-	0.3	0.5	1.6	1.5	0.9	0.2
4.8	1.2	-	-	3.9	1.5	2.2	0.5
0.6	0	0	0	-	-	0.9	0.3
0.9	0.1	0.2	0	1.8	0.2	-	-

same architecture, training dataset, acoustic feature, scoring method

Query-free black-box attack: QFA2SR

Our attack:
three approaches to enhance transferability



QFA2SR: Tailored Loss Functions

Design loss functions tailored to different attack scenarios and VPR

- targeted attack on open-set identification:

cross entropy → $f_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_t$

$\boxed{f_1(x) = -[S(x)]_t}$

margin loss → $f_{\mathrm{M}}(x) = \mathtt{max}_{i \in G, i \neq t}[S(x)]_i - [S(x)]_t$

$\boxed{f_2(x) = \mathtt{max}\{\theta, \mathtt{max}_{i \in G, i \neq t}[S(x)]_i\} - [S(x)]_t}$

threshold-based decision-making

x: voice
S(x): score vector
t: target speaker
G: group of enrolled speakers

QFA2SR: Tailored Loss Functions

Design loss functions tailored to different attack scenarios and VPR

- targeted attack on open-set identification:

cross entropy ⟶ $f_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_t$

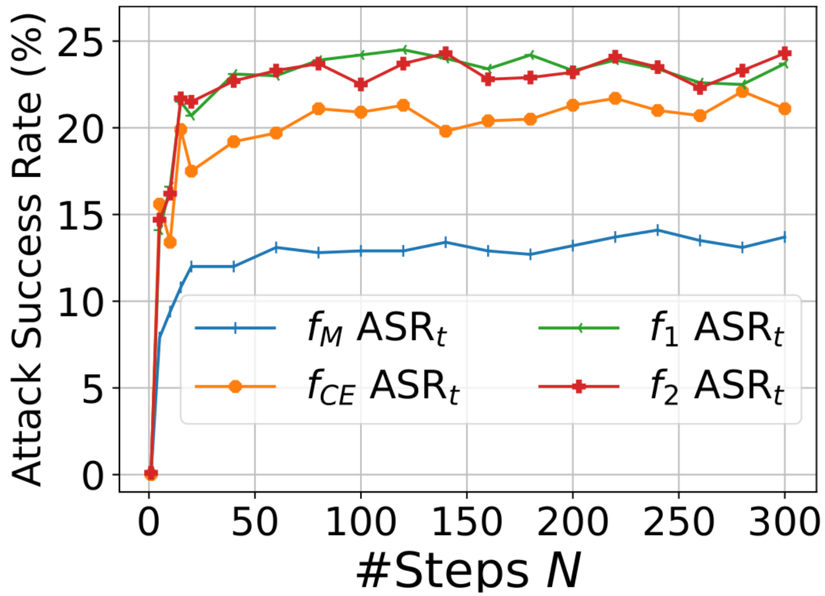
$\boxed{f_1(x) = -[S(x)]_t}$

margin loss ⟶ $f_{\mathrm{M}}(x) = \mathtt{max}_{i\in G, i\neq t}[S(x)]_i - [S(x)]_t$

$\boxed{f_2(x) = \mathtt{max}\{\theta, \mathtt{max}_{i\in G, i\neq t}[S(x)]_i\} - [S(x)]_t}$

threshold-based decision-making

x: voice
S(x): score vector
t: target speaker
G: group of enrolled speakers



!!! Cross entropy and margin loss are sub-optimal

QFA2SR: Tailored Loss Functions

Design loss functions tailored to different attack scenarios and VPR

- untargeted attack on open-set identification:

cross entropy → $f^s_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_s$

$\boxed{f^s_2(x) = \mathtt{max}\{\theta, \mathtt{max}_{i \in G, i \neq s}[S(x)]_i\} - [S(x)]_s}$

margin loss → $f^s_{\mathrm{M}}(x) = \mathtt{max}_{i \in G, i \neq s}[S(x)]_i - [S(x)]_s$

$\boxed{f^s_1(x) = -[S(x)]_s}$ $\boxed{f_3(x) = \theta - \mathtt{max}_{i \in G}[S(x)]_i}$

x: voice
S(x): score vector
s: $argmax_i\ [S(x_0)]_i$
G: group of enrolled speakers

QFA2SR: Tailored Loss Functions

Design loss functions tailored to different attack scenarios and VPR

- untargeted attack on open-set identification:

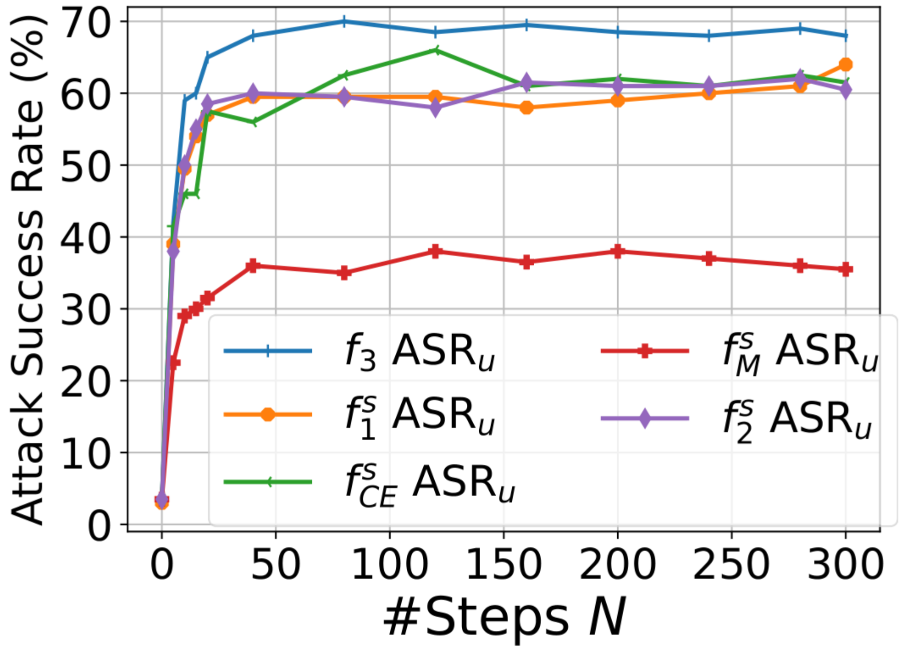
cross entropy → $f^s_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_s$

$\boxed{f^s_2(x) = \max\{\theta, \max_{i \in G, i \neq s}[S(x)]_i\} - [S(x)]_s}$

margin loss → $f^s_{\mathrm{M}}(x) = \max_{i \in G, i \neq s}[S(x)]_i - [S(x)]_s$

$\boxed{f^s_1(x) = -[S(x)]_s}$ $\boxed{f_3(x) = \theta - \max_{i \in G}[S(x)]_i}$

x: voice
S(x): score vector
s: $argmax_i\ [S(x_0)]_i$
G: group of enrolled speakers



!!! F3 is the best;
margin loss is the worst

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\texttt{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\mathtt{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

Two ensemble strategies:

- dynamic weights selection

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\mathtt{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

Two ensemble strategies:

- dynamic weights selection

  uniform weight $w_k = \frac{1}{K}$ for $k = 1, \cdots, K$

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\text{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

Two ensemble strategies:

- dynamic weights selection

uniform weight $\quad w_k = \frac{1}{K} \text{ for } k = 1, \cdots, K$

VPR 1 (PLDA; $[-\infty, +\infty]$) $\rightarrow$

VPR 2 (cosine; $[-1, +1]$) $\rightarrow$

$$\begin{bmatrix} f_1 = 22 \\ \cdots \\ f_K = 0.5 \end{bmatrix}$$

bias towards

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\text{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

Two ensemble strategies:

- dynamic weights selection

uniform weight $w_k = \frac{1}{K}$ for $k = 1, \cdots, K$

VPR 1 (PLDA; $[-\infty, +\infty]$) $\rightarrow$ $f_1 = 22$

$\cdots$

VPR 2 (cosine; $[-1, +1]$) $\rightarrow$ $f_K = 0.5$

bias towards

dynamic weight:

surrogate-specific; iteratively updated

$$f_{\text{ens}} \leftarrow f_{\text{ens}} + \frac{f_k - \mu_k}{\sqrt{\sigma_k}} \qquad \mu_k \leftarrow \mu_k + \frac{f_k - \mu_k}{n}; \;\; \sigma_k \leftarrow \sigma_k + \frac{1}{n}\left((f_k - \mu_k)^2 - \sigma_k\right)$$

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

$$f_{\mathrm{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$$

Two ensemble strategies:

- dynamic weights selection

✗ uniform weight $w_k = \frac{1}{K} \text{ for } k = 1, \cdots, K$

✓ dynamic weight:

surrogate-specific; iteratively updated

$$f_{\mathrm{ens}} \leftarrow f_{\mathrm{ens}} + \frac{f_k - \mu_k}{\sqrt{\sigma_k}} \qquad \mu_k \leftarrow \mu_k + \frac{f_k - \mu_k}{n}; \;\; \sigma_k \leftarrow \sigma_k + \frac{1}{n}\left((f_k - \mu_k)^2 - \sigma_k\right)$$

Table 22: The effectiveness of SRS ensemble for $\mathcal{A}^{\mathrm{T}}_{\mathrm{OSI}}$.

T \ S	IV		ECAPA		XV-P		XV-C		Res18-I		Res18-V		Res34-I		Res34-V		Auto	
	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d	ASR_t-s	ASR_t-d
Best-single	11.9	6.7	47.1	39.8	39.1	23.7	5.8	3.4	4.8	1.2	0.6	0.5	3.9	1.5	2.2	0.5	2.2	3.8
Uniform-Ens (w/o T)	**21.7**	**15**	58	52.7	47.5	27.2	**13.4**	**8.1**	3.3	0	0	0	7.8	4.6	6.7	3.2	6.5	4.6
Dynamic-Ens (w/o T)	19.7	14	**66.6**	**60**	**64.6**	**49.4**	12.3	6.8	**24.3**	**11.5**	**13.8**	**6.5**	**30.2**	**22.1**	**34.5**	**21.3**	**23.9**	**18.1**

Note: (1) S and T denote the surrogate and target SRSs, respectively. (2) Best single denotes the surrogate SRS that leads to the largest ASR_t, which varies with the target. (3) "W/o T" means that all the SRSs except the target are used as surrogate.

dynamic weight dominates uniform weight

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

Two ensemble strategies:

- dynamic weights selection
- Global score ranking

untargeted attack on open-set identification:

$$f_3(x) = \theta - \mathtt{max}_{i \in G}[S(x)]_i$$

x: voice

S(x): score vector

G: group of enrolled speakers

local score rank differs → i differs → inconsistent optimize directions

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

Two ensemble strategies:

- dynamic weights selection
- Global score ranking

untargeted attack on open-set identification:

$$f_3(x) = \theta - \mathtt{max}_{i \in G}[S(x)]_i$$

x: voice
S(x): score vector
G: group of enrolled speakers

local score rank differs → i differs → inconsistent optimize directions

Define global score rank to aggregate local ranks by voting or summation

QFA2SR: SRS Ensemble

combine diverse surrogates → more likely to transfer to unknown target

Two ensemble strategies:

- dynamic weights selection
- Global score ranking

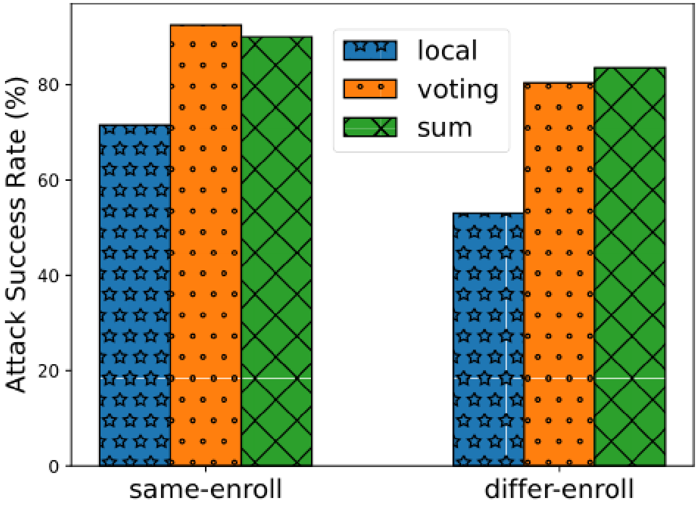
untargeted attack on open-set identification:

$$f_3(x) = \theta - \mathtt{max}_{i \in G}[S(x)]_i$$

x: voice
S(x): score vector
G: group of enrolled speakers

local score rank differs → i differs → inconsistent optimize directions

Define global score rank to aggregate local ranks by voting or summation



local rank

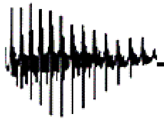
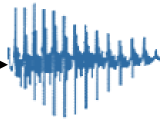
global rank

QFA2SR: Time-Freq Corrosion

use randomized modifications functions to simulate and approximate the decision boundary of unknow target

randomized modifications functions

- Time-domain

x  → time-domain function →  →

Reverberation-distortion (RD): convolve x with Room Impulse Response

Noise-flooding (NF): add Gaussian noise to x

Speed-alteration (SA): ↑↑ or ↓↓ speed of x

Chunk-dropping (CD): drop partial chunks of x

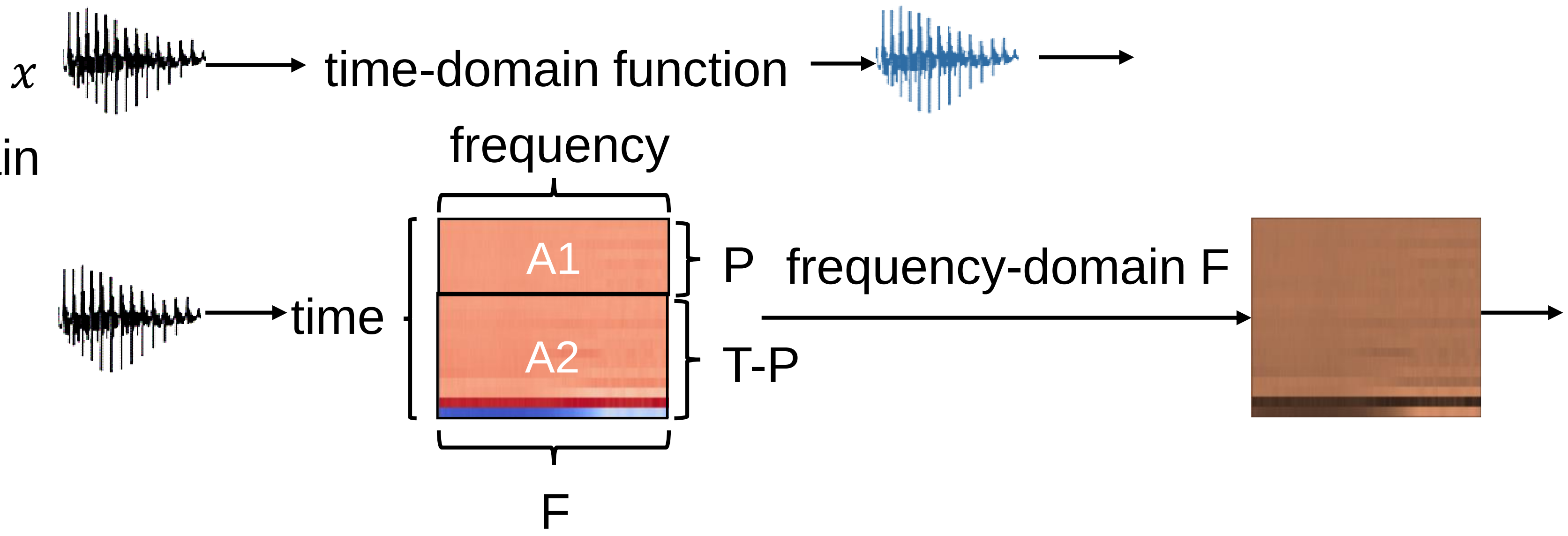
Frequency-dropping (FD): drop some frequency components of x

QFA2SR: Time-Freq Corrosion

use randomized modifications functions to simulate and approximate the decision boundary of unknow target

randomized modifications functions

- Time-domain
- Frequency-domain



Time-warping (TW): scale "image" A1 from $P \times F$ to $\mathrm{w} \times F$, scale A2 from $(T-P) \times F$ to $(\mathrm{T}-\mathrm{w}) \times F$, w is randomly chosen

Time-masking (TM): zero mask random consecutive frames along the time-axis

Frequency-masking (FM): zero mask random consecutive channels along the frequency-axis

QFA2SR: Time-Freq Corrosion

use randomized modifications functions to simulate and approximate the decision boundary of unknow target

randomized modifications functions

- Time-domain
- Frequency-domain
- Serial or parallel combinations

Table 19: ASR_t of time-freq corrosion in $\mathcal{A}^{T}_{\mathtt{OSI}}$, where Para denotes RD+NF||SA+CD+FD||TW+TM+FM.

	Baseline	Single								Serial			Parallel
		RD	**NF**	**SA**	**CD**	**FD**	**TW**	**TM**	**FM**	**RD+NF**	**SA+CD+FD**	**TW+TM+FM**	**Para**
Same-enroll	39.1	52.2	59	53.9	57.7	46.8	40.8	43	52.7	62	72.6	57.6	**78.4**
Differ-enroll	23.7	36.3	40.6	38.1	36.1	31	26.3	27.8	35.6	45.5	53.9	37.4	**64.1**

Each single function: ↑ serial combination: ↑↑ parallel combination: ↑↑↑

QFA2SR: experiments on commercial APIs

APIs: Microsoft Azure, iFlytek, TalentedSoft, Jingdong

- targeted attack on open-set identification

	Microsoft Azure				TalentedSoft				IFlytek			
	ASR_t-s	ASR_t-d	SNR	PESQ	ASR_t-s	ASR_t-d	SNR	PESQ	ASR_t-s	ASR_t-d	SNR	PESQ
SirenAttack	1	2.1	8.02	1.12	1.4	1.3	10.07	1.18	0	0	8	1.12
Kenansville	0	0	**16.23**	**1.75**	0	0	**16.23**	**1.75**	0	0	**16.23**	**1.75**
FakeBob	4.2	3.1	12.23	1.22	5.0	2.4	12.50	1.23	0	0	12.16	1.24
FakeBob + 1	6.2	4.1	12.23	1.23	5.6	2.7	12.51	1.24	1.9	1.9	12.16	1.23
FakeBob + 1 2	17.5	17.2	12.22	1.24	9.3	4.7	12.22	1.24	9.1	8.8	12.22	1.24
FakeBob + 1 2 3	3.8	2.7	12.71	1.28	4.0	2.5	12.71	1.28	0.6	0.6	12.71	1.28
BIM	18.9	12.7	11.49	1.18	8.9	6.5	11.28	1.19	16	15.5	11.50	1.18
BIM + 1	27.2	21.8	11.50	1.18	9.3	6.6	11.28	1.19	24	17.5	11.52	1.19
BIM + 1 2	42.8	34.2	11.29	1.18	16.9	12.5	11.29	1.18	25.9	21.6	11.29	1.18
BIM + 1 2 3 (QFA2SR)	**89.6** ↑ 70.7	**82.8** ↑ 70.1	10.85	1.18	**40.1** ↑ 31.2	**27.4** ↑ 20.9	10.85	1.18	**46.1** ↑ 30.1	**39.5** ↑ 24	10.85	1.18

- untargeted attack on open-set identification

	Microsoft Azure				TalentedSoft				IFlytek			
	ASR_u-s	ASR_u-d	SNR	PESQ	ASR_u-s	ASR_u-d	SNR	PESQ	ASR_u-s	ASR_u-d	SNR	PESQ
SirenAttack	16.67	8.25	8.16	1.12	23.9	18.7	10.07	1.18	0	0	8.07	1.12
Kenansville	0	0	**16.97**	**1.8**	7	4	**17.58**	**1.84**	0	0	**16.66**	**1.77**
Hidden	21.4	23	-2.84	1.14	22.9	21.9	-2.9	1.18	0	0	-2.95	1.15
FakeBob	33.33	15.46	12.24	1.23	26.8	24	12.41	1.24	11.5	5.8	12.12	1.23
FakeBob + 1	33.33	15.46	12.24	1.23	26.8	24	12.41	1.24	11.5	5.8	12.12	1.23
FakeBob + 1 2	47.92	37.11	12.22	1.22	31	26.7	12.22	1.22	19.2	13.5	12.22	1.22
FakeBob + 1 2 3	15.42	6.41	12.55	1.27	11.7	7.2	12.55	1.27	5.0	2.7	12.55	1.27
BIM	61.22	47.21	11.55	1.18	17.8	16.2	11.37	1.18	60	58	11.53	1.17
BIM + 1	68.4	50.8	11.54	1.18	22.7	19.9	11.37	1.19	64	61.9	11.54	1.18
BIM + 1 2	80.62	66.53	11.37	1.19	30.1	23.5	11.37	1.19	69	62.9	11.37	1.19
BIM + 1 2 3 (QFA2SR)	**99.49** ↑ 38.27	**92.39** ↑ 45.18	11.01	1.19	**55** ↑ 28.2	**39.6** ↑ 15.6	11.01	1.19	**70** ↑ 10	**68** ↑ 10	11.01	1.19

- targeted attack on text-dependent verification

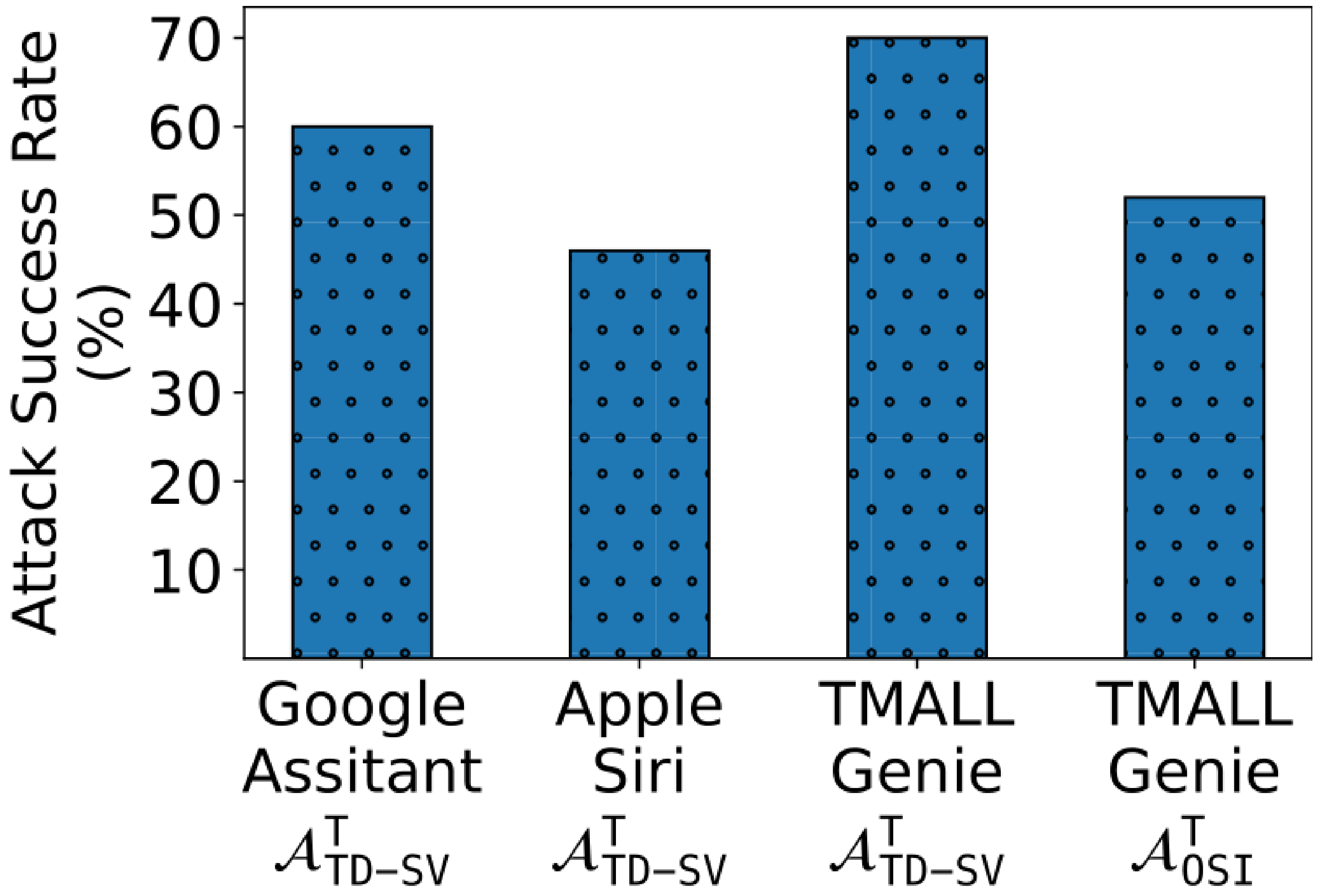
	Microsoft Azure			Jingdong		
	differ-enroll ASR_t	SNR (dB)	PESQ	differ-enroll ASR_t	SNR (dB)	PESQ
SirenAttack	0.49	8.97	1.15	0	10.15	1.18
Kenansville	0	**20.64**	**2.11**	0	**20.64**	**2.11**
Voice Cloning	10	-	-	40	-	-
FakeBob	0.52	13.16	1.28	8	13.32	1.28
FakeBob + 1	0.52	13.16	1.28	8	13.32	1.28
FakeBob + 1 2	16.67	13.14	1.28	11	13.14	1.28
FakeBob + 1 2 3	0.1	13.45	1.30	3	13.45	1.30
BIM	13.01	12.40	1.24	12	12.21	1.23
BIM + 1	13.01	12.40	1.24	12	12.21	1.23
BIM + 1 2	27.78	12.21	1.23	23.5	12.21	1.23
BIM + 1 2 3 (QFA2SR)	**61.86** ↑ 48.85	11.84	1.24	**66** ↑ 26	11.84	1.24

↑: 10%-70% transfer improvement over the most effective baseline

Azure: ≈ 90% targeted
≈ 100% untargetd

QFA2SR: experiments on voice assistants

Voice assistants: Google Assistant, Apple Siri, and TMall Genie

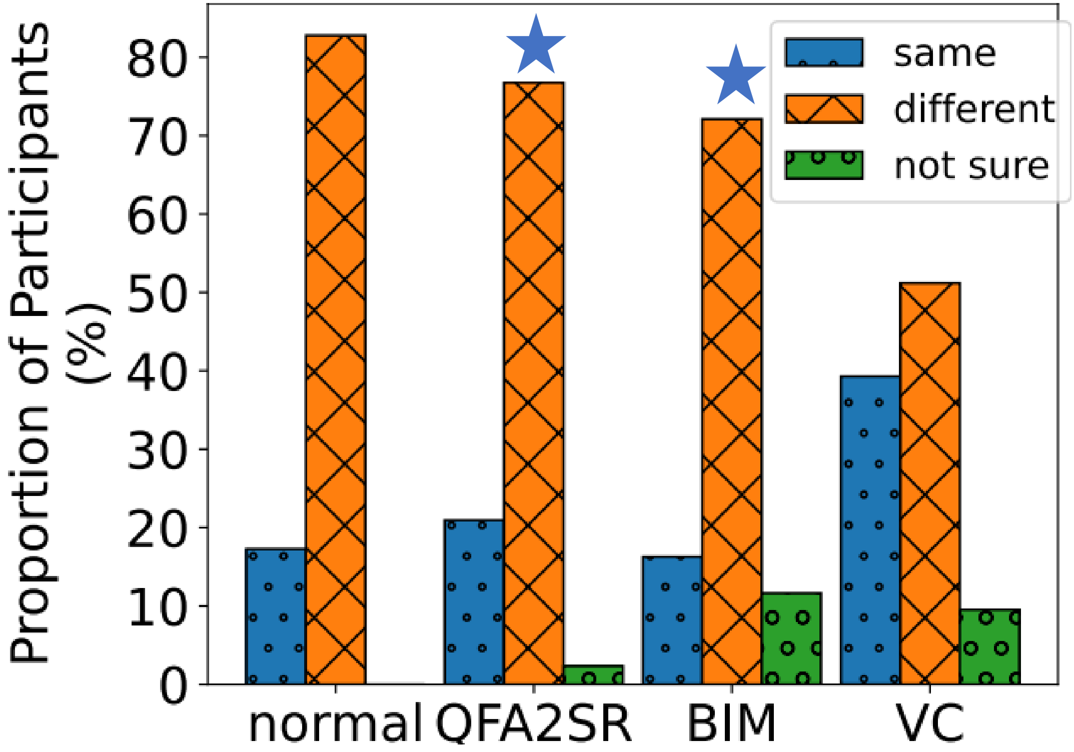


QFA2SR: human study

presented with a pair of voices
tell if they are uttered by the same speaker

126 participants from Amazon Mechanical Turk Platform

- Normal: 2 clean voices from distinct speakers
- QFA2SR: 1 clean voice from the target speaker
1 QFA2SR adversarial voice from imposter
- BIM: 1 clean voice from the target speaker
1 BIM adversarial voice from imposter
- VC: 1 clean voice from the target speaker
1 voice generated by voice cloning



QFA2SR does not worsen imperceptibility

Take away

- Query-free black-box speech adversarial examples against voiceprint recognition
- Leverage transferability
- Equipped with three approaches to boost transferability
- Highly effective against commercial APIs and voice assistants
- Negligible effect on imperceptibility
- Vulnerability disclosure receives acknowledgment or bounty award from vendors

Website (attack audios & videos): https://sites.google.com/view/qfa2sr
Paper: https://arxiv.org/abs/2305.14097

Any Question?
Thanks!

Guangke Chen: https://guangkechen.site
Fu Song: songfu1983@gmail.com