Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems

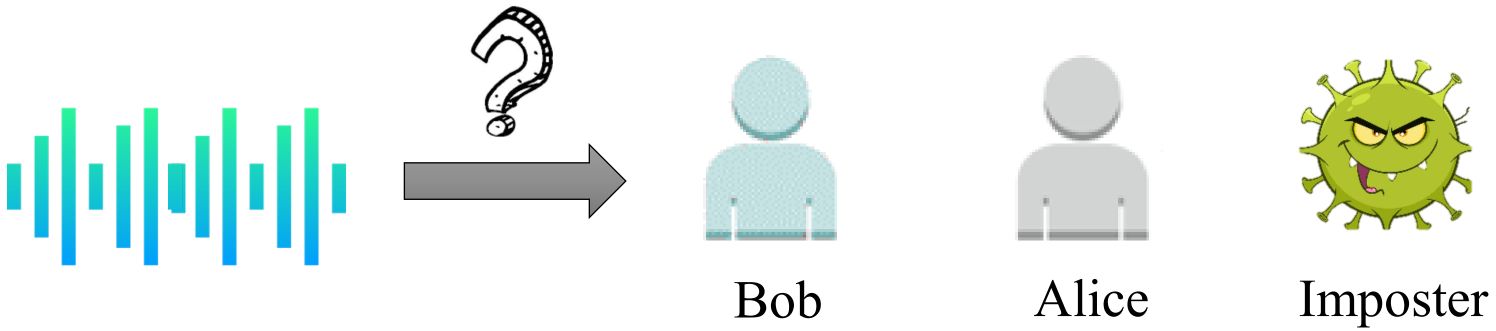
Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du,
Zhe Zhao, Fu Song, Yang Liu

上海科技大学 ShanghaiTech University
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
Shanghai Institute of Microsystem and Information Technology
UCAS 中国科学院大学 University of Chinese Academy of Sciences
天津大学 Tianjin University

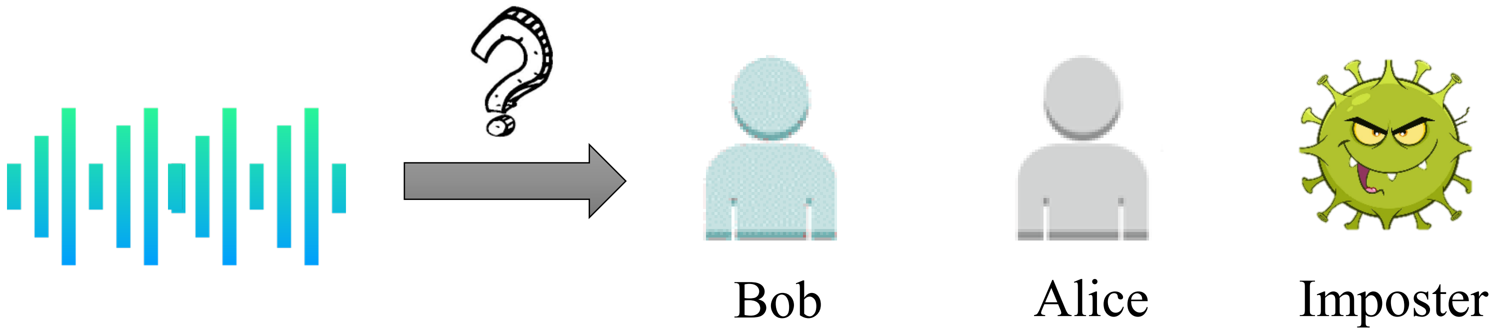
Guangke Chen (GuangkeChen@outlook.com)
✉ Fu Song (songfu@shanghaitech.edu.cn)

Speaker Recognition Systems (SRSs)   a.k.a, Voiceprint Recognition Systems

Speaker Recognition Systems (SRSs)

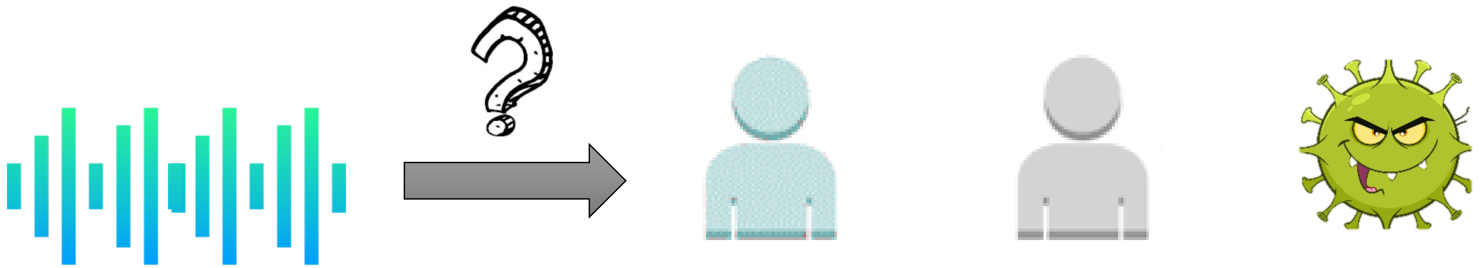


Speaker Recognition Systems (SRSs)



Ubiquitous Application

Speaker Recognition Systems (SRSs)

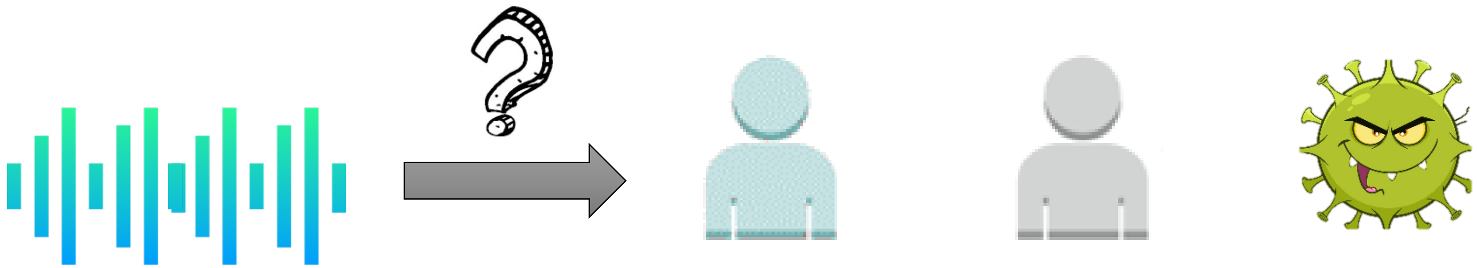


Bob Alice Imposter

Ubiquitous Application

Voice assistant wake up

Speaker Recognition Systems (SRSs)



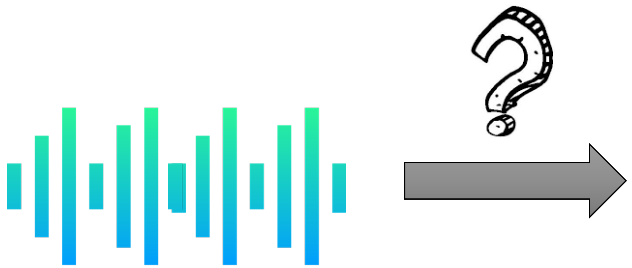
Bob Alice Imposter

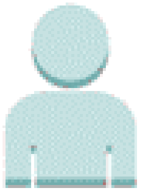
Ubiquitous Application

Voice assistant wake up

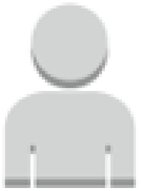
Personalized service
on smart home

Speaker Recognition Systems (SRSs)

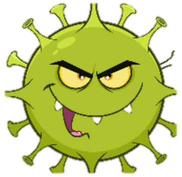




Bob



Alice



Imposter

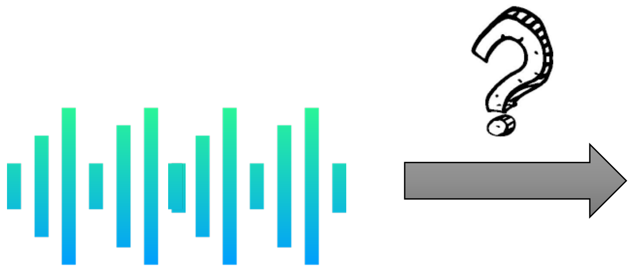
Ubiquitous Application

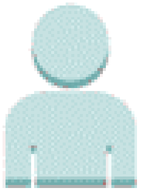
Voice assistant wake up

Personalized service
on smart home

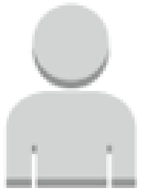
Financial
transaction

Speaker Recognition Systems (SRSs)

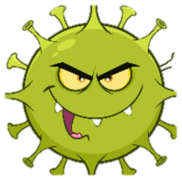




Bob



Alice



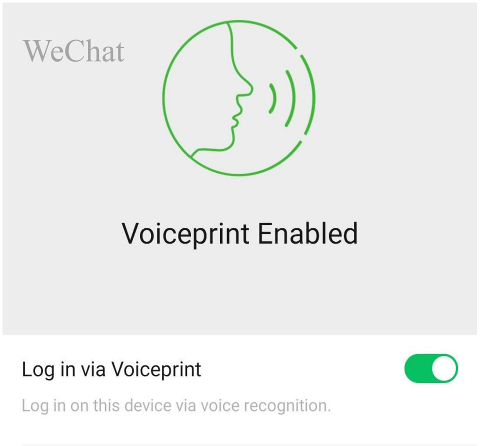
Imposter

Ubiquitous Application

Voice assistant wake up

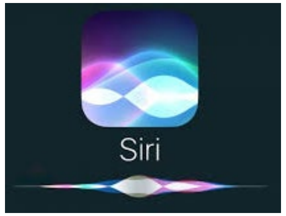
Personalized service on smart home

Financial transaction



App log in

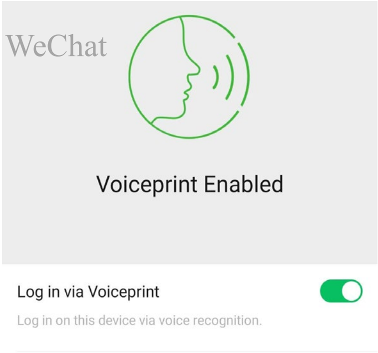
Ubiquitous Application



Voice assistant wake up

Personalized service on smart home

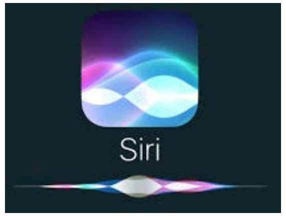
Financial transaction



App log in

Safety-critical scenario

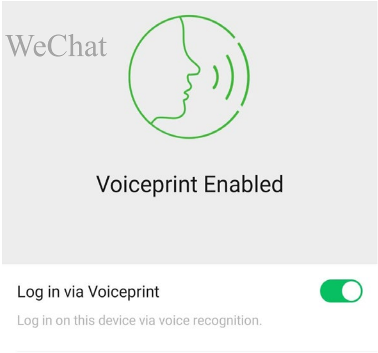
Ubiquitous Application



Voice assistant wake up

Personalized service on smart home

Financial transaction

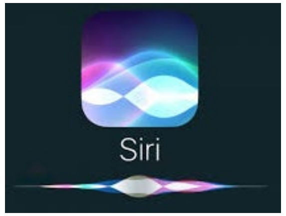


App log in

Safety-critical scenario

Once broken

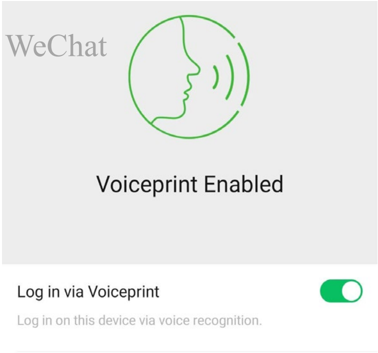
Ubiquitous Application



Voice assistant wake up

Personalized service
on smart home

Financial
transaction



App log in

Safety-critical scenario

Once broken

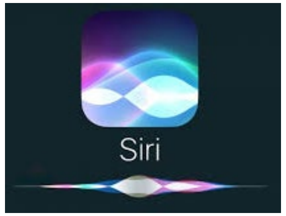
property damage

reputation degrade

sensitive information leak

…

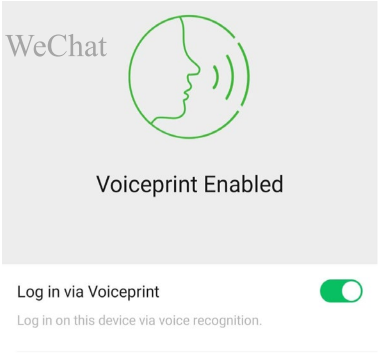
Ubiquitous Application



Voice assistant wake up

Personalized service on smart home

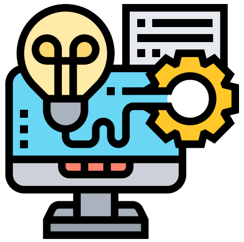
Financial transaction



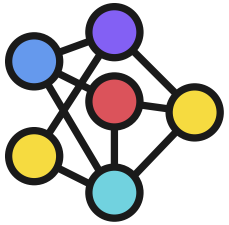
App log in

Security of SRSs!!!

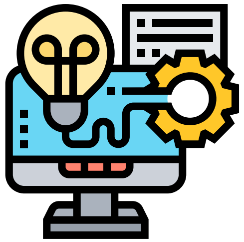
Mainstream implementation of SRSs



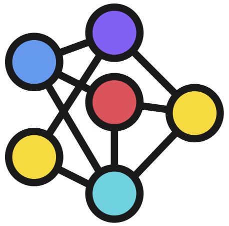
Machine Learning
(ML)



Mainstream implementation of SRSs

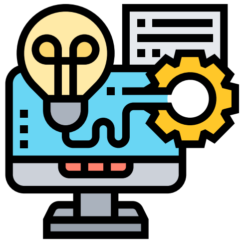


Machine Learning
(ML)

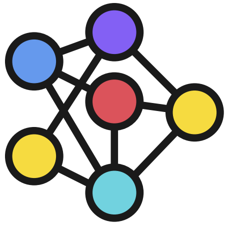


However, ML is vulnerable to adversarial examples

Mainstream implementation of SRSs



Machine Learning
(ML)



However, ML is vulnerable to adversarial examples

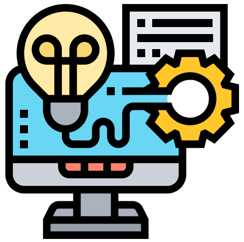
Benign example
Result: Panda
Confidence: 57.7%

Perturbation

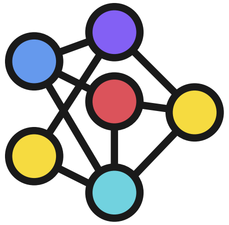
Adversarial example
Result: Gibbon
Confidence: 99.3%

Ian Goodfellow et al.

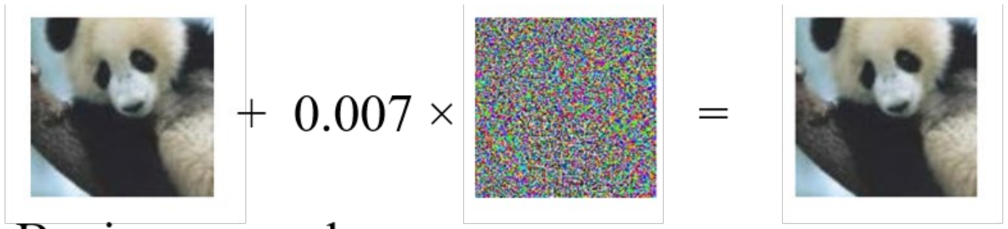
Mainstream implementation of SRSs



Machine Learning
(ML)



However, ML is vulnerable to adversarial examples

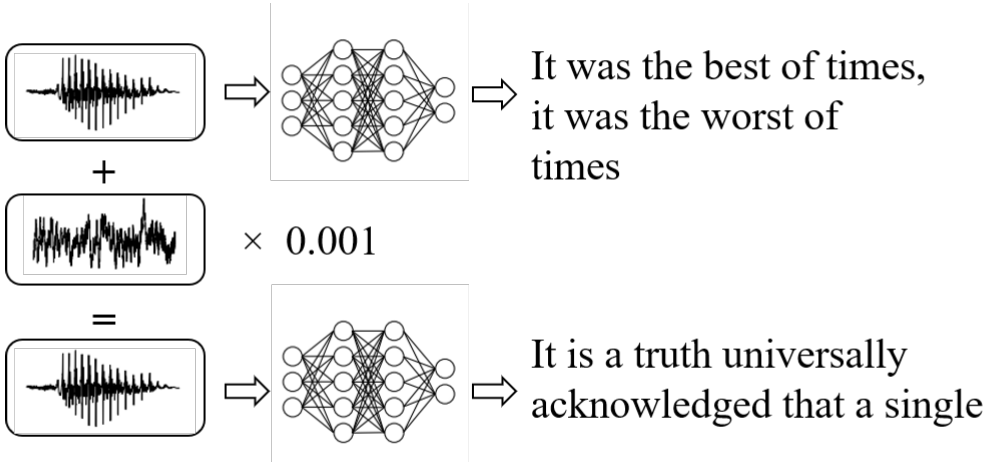


Benign example
Result: Panda
Confidence: 57.7%

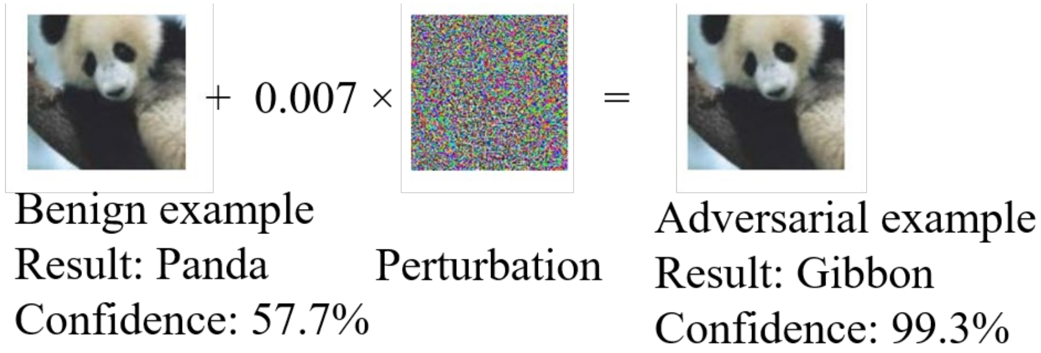
Perturbation

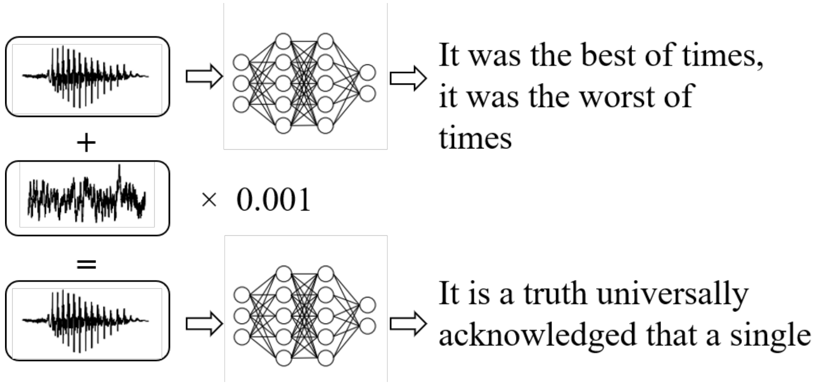
Adversarial example
Result: Gibbon
Confidence: 99.3%

Ian Goodfellow et al.

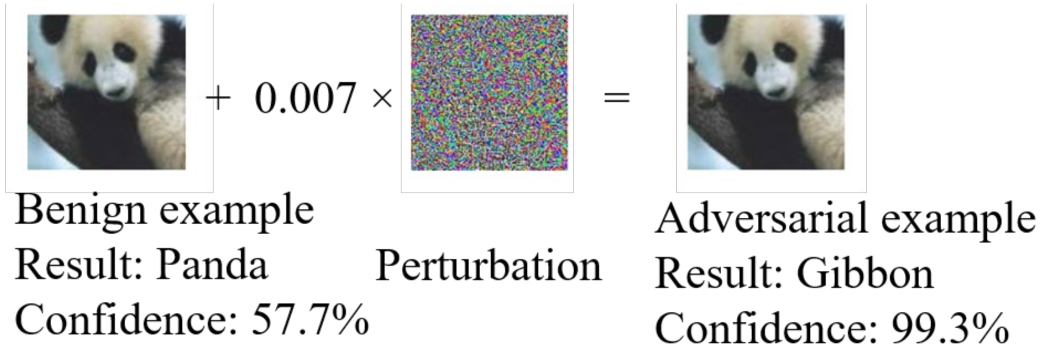


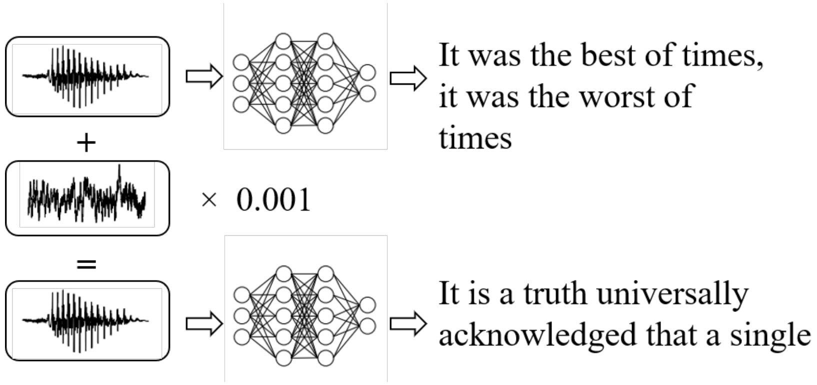
Nicholas Carlini et al.





Is adversarial attack practical on SRSs ?





Is adversarial attack practical on SRSs ?

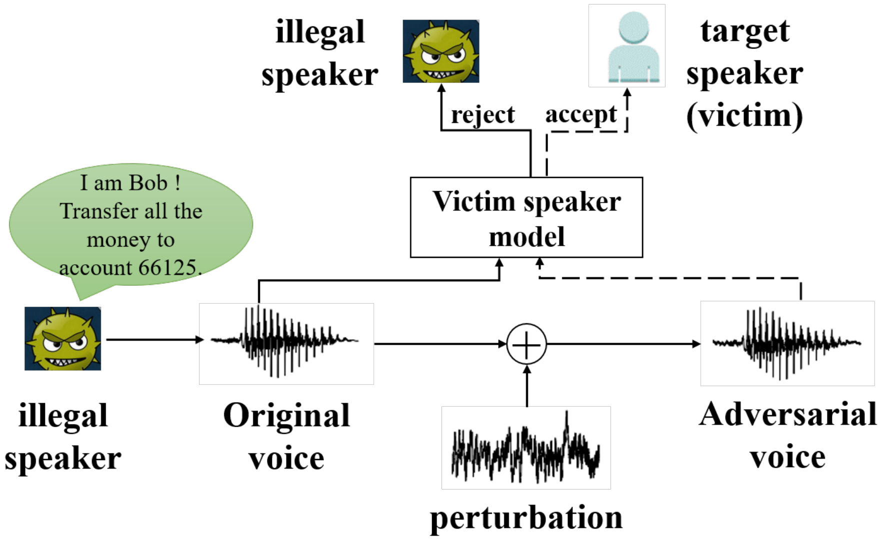
FAKEBOB

- Black-box
- Appliable to general SRS task
- Effective on commercial SRSs
- Effective in over-the-air attack

Threat model

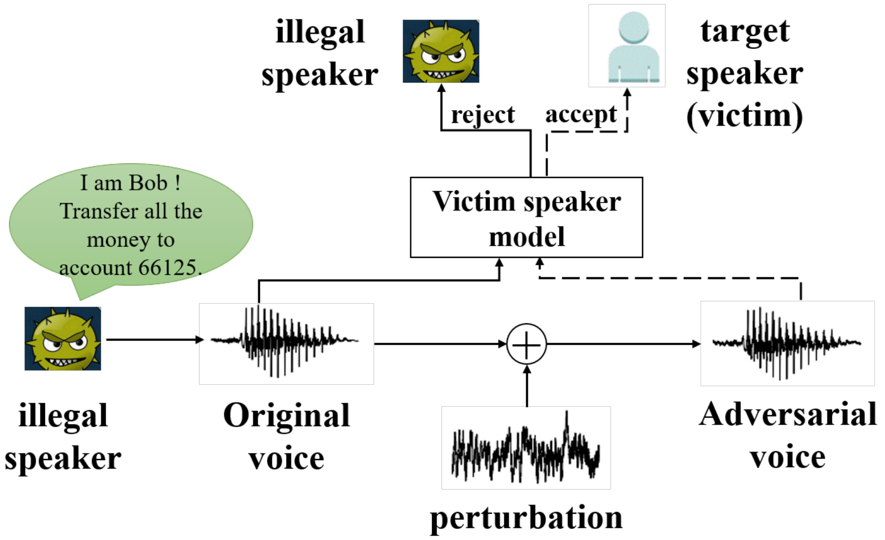
Threat model

- Attacker Goal: pass voice authentication; gain access to privilege



Threat model

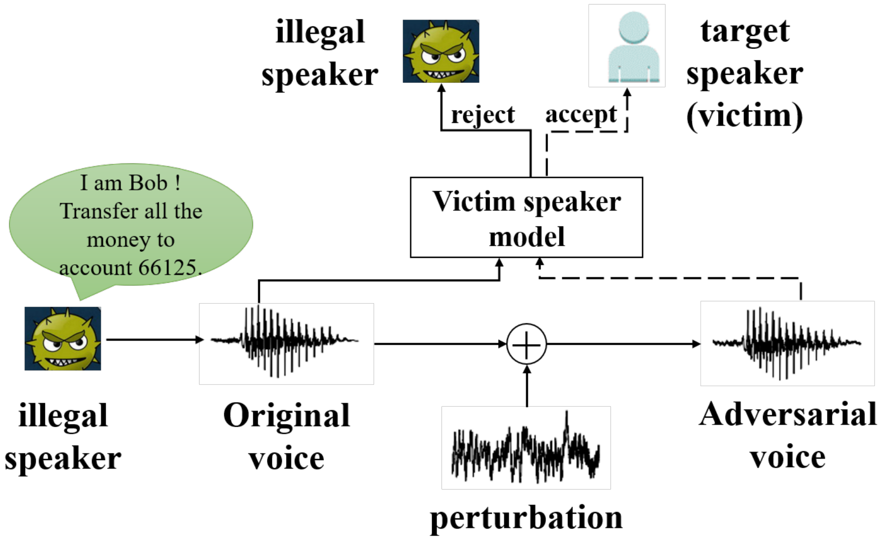
- Attacker Goal: pass voice authentication; gain access to privilege



- Attacker Capability: no information about model structure / parameter;

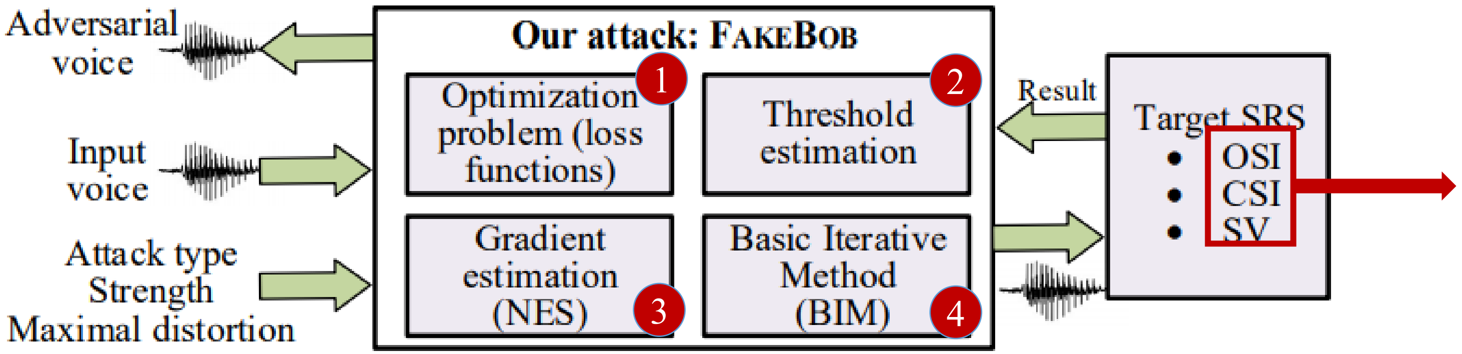
Threat model

- Attacker Goal: pass voice authentication; gain access to privilege



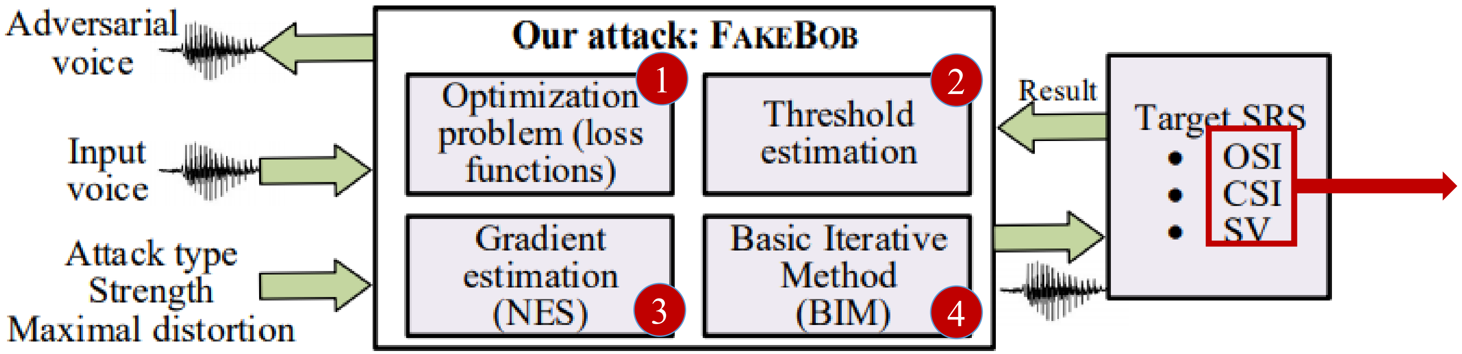
- Attacker Capability: no information about model structure / parameter; limited to query the speak model of the victims

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

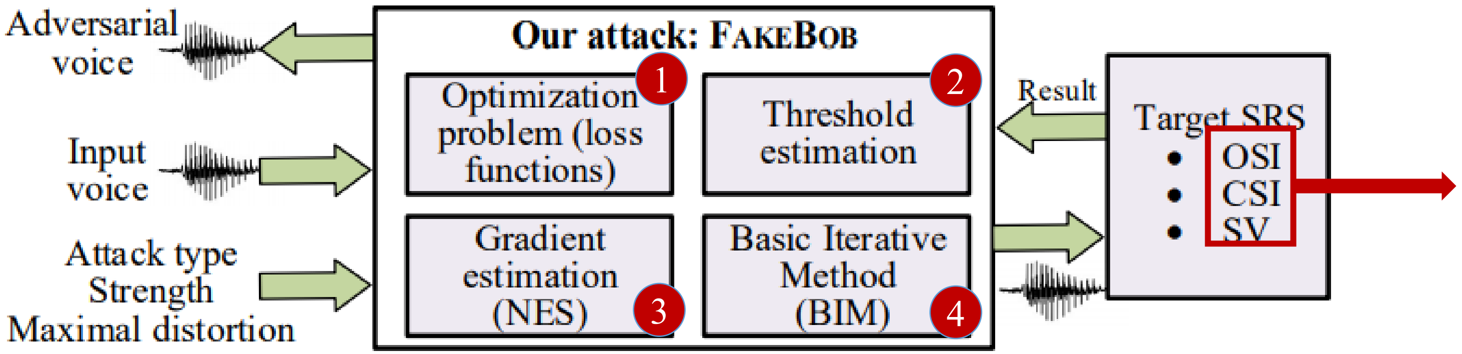
Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1 Effective **loss function** design.

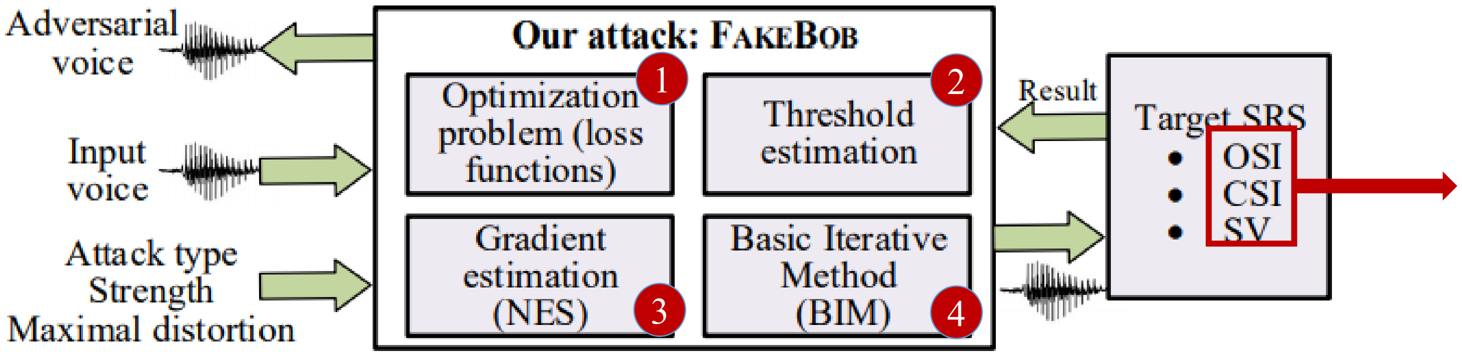
Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1 Effective **loss function** design. Goal: $f(x) \le 0 \leftrightarrow$ attack succeeds

Overview of FAKEBOB

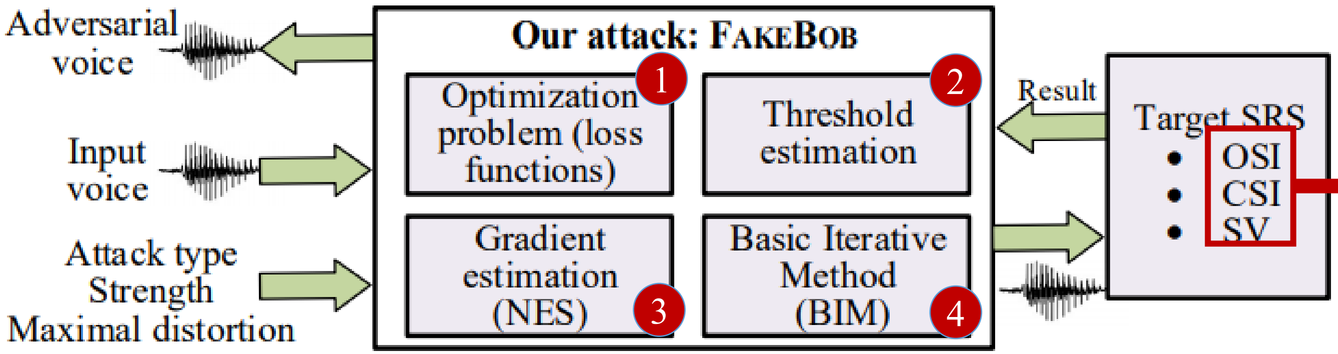


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1. Effective **loss function** design. Goal: $f(x) \leq 0 \leftrightarrow$ attack succeeds

   Based on **scoring** and **decision-making** mechanism

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1 Effective **loss function** design. Goal: $f(x) \leq 0 \leftrightarrow$ attack succeeds

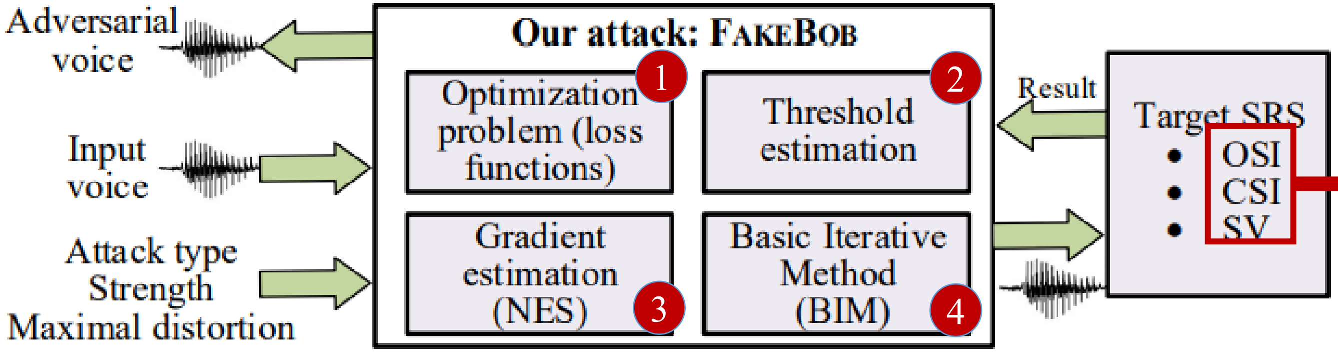
Based on **scoring** and **decision-making** mechanism

e.g., for OSI

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}} \, [S(x)]_i, & \text{if } \underset{i \in G}{\max} \, [S(x)]_i \geq \boxed{\theta}; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

$S(x)$: scores
$D(x)$: decision
θ: threshold

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1 Effective **loss function** design. Goal: $f(x) \leq 0 \leftrightarrow$ attack succeeds

Based on **scoring** and **decision-making** mechanism

e.g., for OSI

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}}\ [S(x)]_i, & \text{if } \underset{i \in G}{\max}\ [S(x)]_i \geq \boxed{\theta}; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

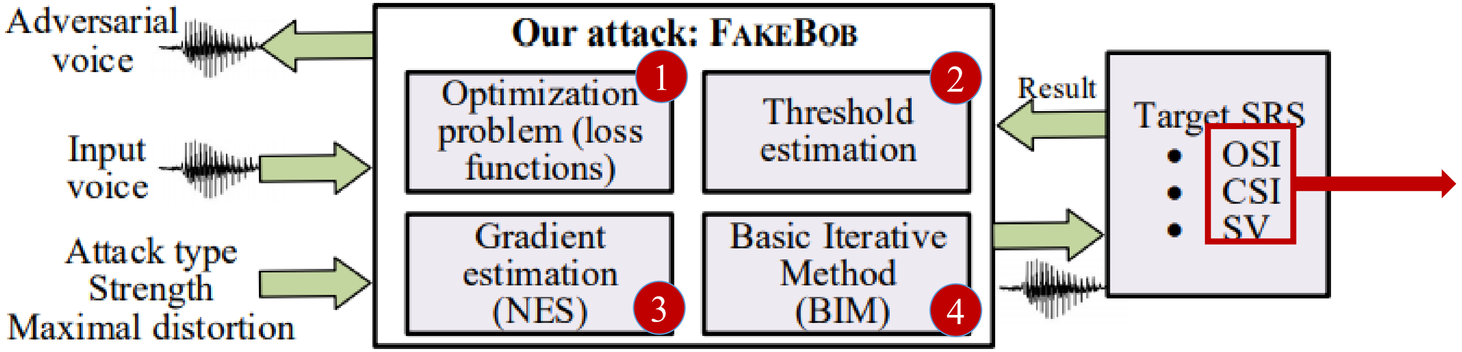
$S(x)$: scores

$D(x)$: decision

θ: threshold

$$f(x) = \max\{\boxed{\theta}, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

1 Effective **loss function** design. Goal: $f(x) \leq 0 \leftrightarrow$ attack succeeds

Based on **scoring** and **decision-making** mechanism

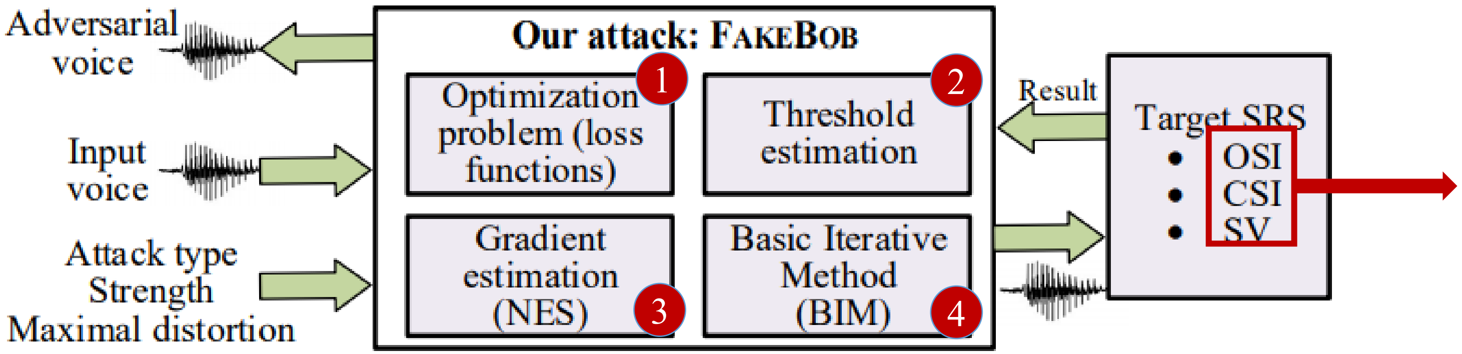
OSI: $$f(x) = \max\{\theta, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Tailored for different SRSs **tasks**: CSI, SV, OSI

CSI: $$f(x) = \max_{i \neq t}[S(x)]_i + \kappa - [S(x)]_t$$

SV: $$f(x) = \theta + \kappa - S(x)$$

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs

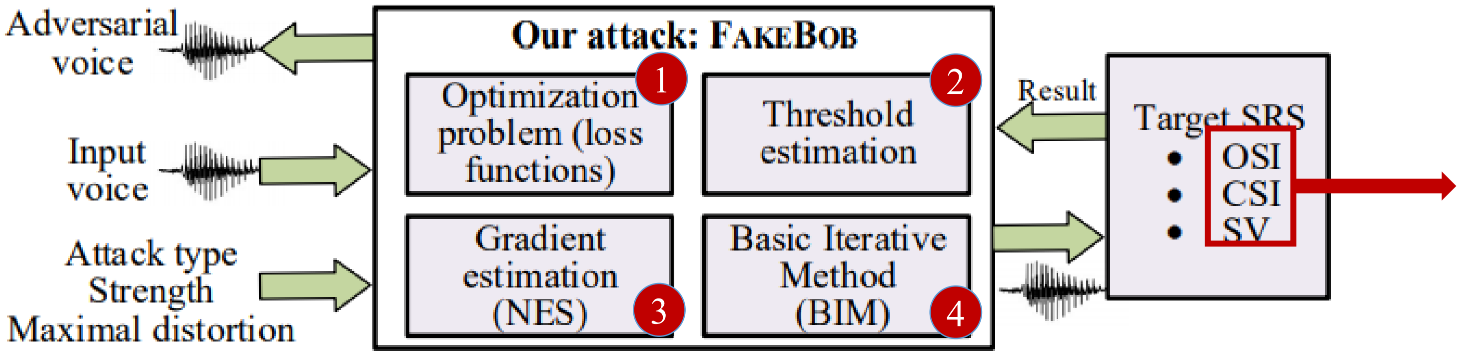
OSI:

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}} \, [S(x)]_i, & \text{if } \max_{i \in G} [S(x)]_i \geq \boxed{\theta}; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

SV:

$$D(x) = \begin{cases} \texttt{accept} & \text{if } S(x) \geq \boxed{\theta}; \\ \texttt{reject} & \text{otherwise.} \end{cases}$$

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs

OSI:

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}} \, [S(x)]_i, & \text{if } \max_{i \in G} [S(x)]_i \geq \boxed{\theta}; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

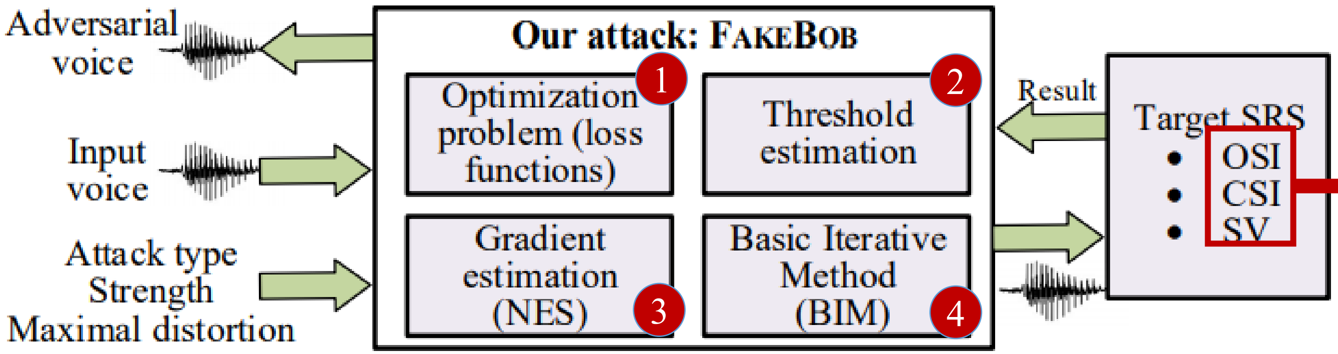
SV:

$$D(x) = \begin{cases} \texttt{accept} & \text{if } S(x) \geq \boxed{\theta}; \\ \texttt{reject} & \text{otherwise.} \end{cases}$$

$\geq \theta$: attack succeeds

$< \theta$: attack fails

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs; unknown to attacker

OSI:

$$D(x) = \begin{cases} \underset{i \in G}{\operatorname{argmax}} \, [S(x)]_i, & \text{if } \max_{i \in G} [S(x)]_i \geq \boxed{\theta}; \\ \texttt{reject}, & \text{otherwise.} \end{cases}$$

?

SV:

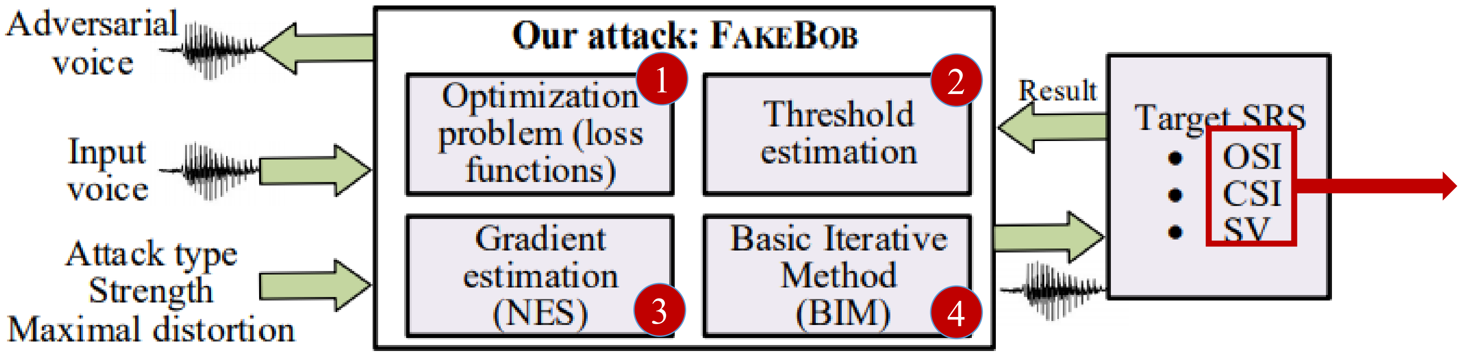
$$D(x) = \begin{cases} \texttt{accept} & \text{if } S(x) \geq \boxed{\theta}; \\ \texttt{reject} & \text{otherwise.} \end{cases}$$

?

$\geq \theta$: attack succeeds

$< \theta$: attack fails

Overview of FAKEBOB

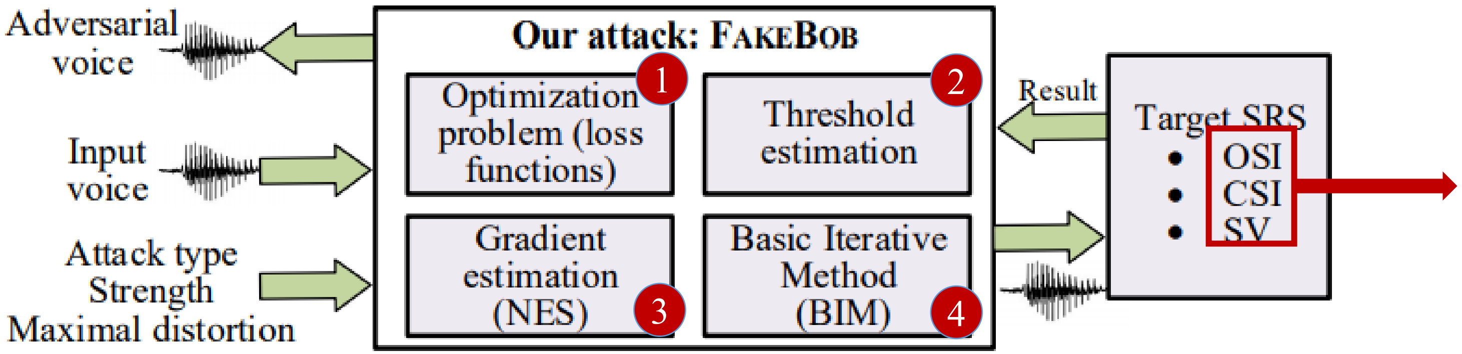


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs; unknown to attacker

Novel **threshold estimation** algorithm

Overview of FAKEBOB



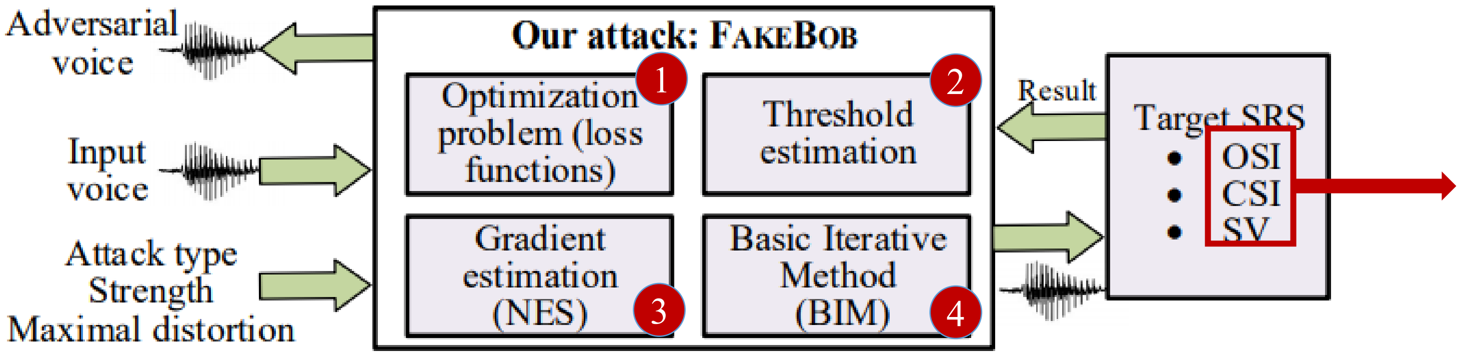
different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs; unknown to attacker

Novel **threshold estimation** algorithm

$$f(x) = \max\{\theta, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Overview of FAKEBOB



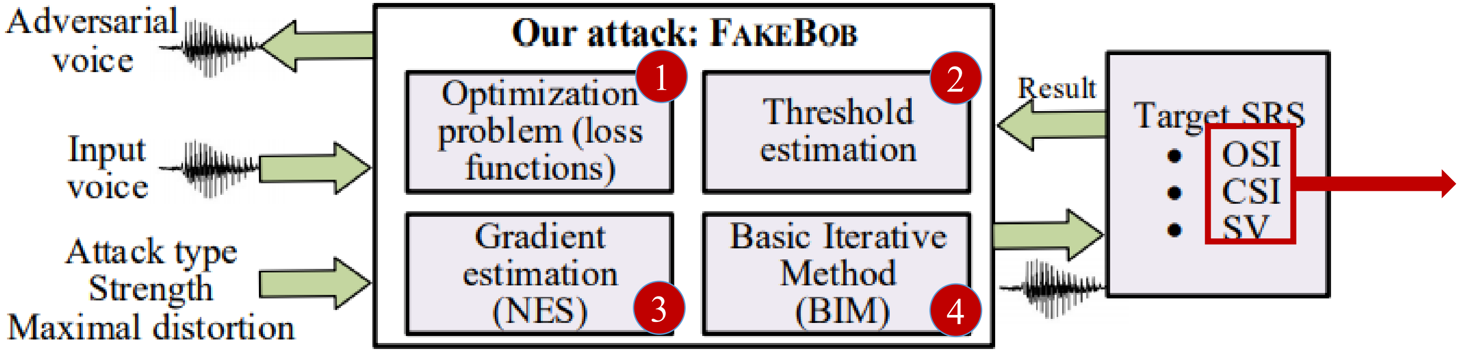
different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs; unknown to attacker

Novel **threshold estimation** algorithm

$$f(x) = \max\{\theta, \max_{i\neq t}[S(x)]_i\} + \kappa - [S(x)]_t \xrightarrow[\hat{\theta} \approx \theta]{\hat{\theta} > \theta \ \&\&} f(x) = \max\{\boldsymbol{\hat{\theta}}, \max_{i\neq t}[S(x)]_i\} + \boldsymbol{\kappa} - [S(x)]_t$$

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

2 **Threshold**: specialness of SRSs; unknown to attacker

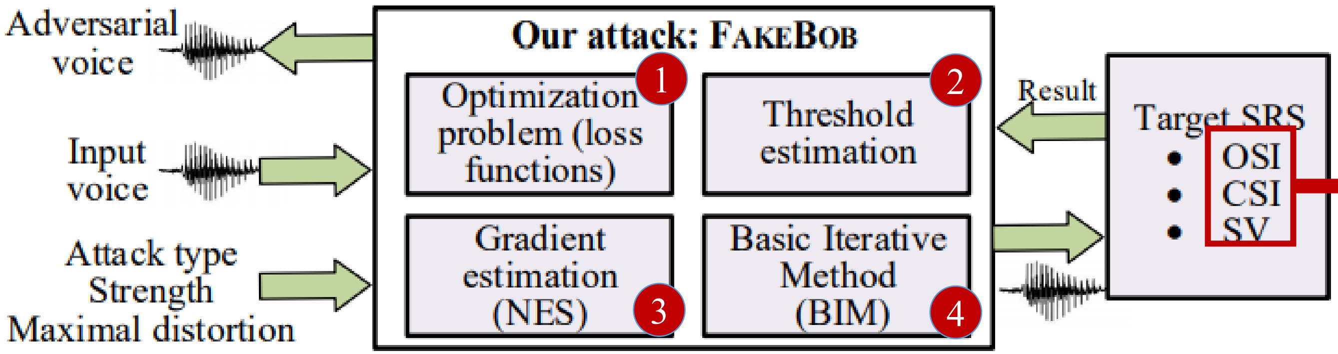
Novel **threshold estimation** algorithm

$$f(x) = \max\{\theta, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t \xrightarrow[\hat{\theta} \approx \theta]{\hat{\theta} > \theta \ \&\&} f(x) = \max\{\hat{\theta}, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

$\hat{\theta} > \theta$: make sure attack succeeds

$\hat{\theta} \approx \theta$: attack not too expensive

Overview of FAKEBOB

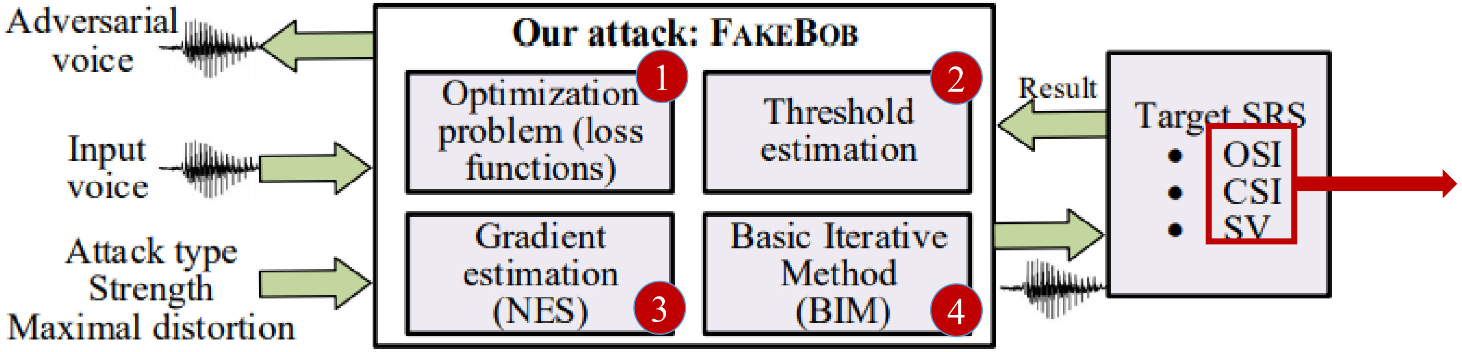


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

3 NES-based gradient estimation

white-box: backpropagation → exact gradient

Overview of FAKEBOB

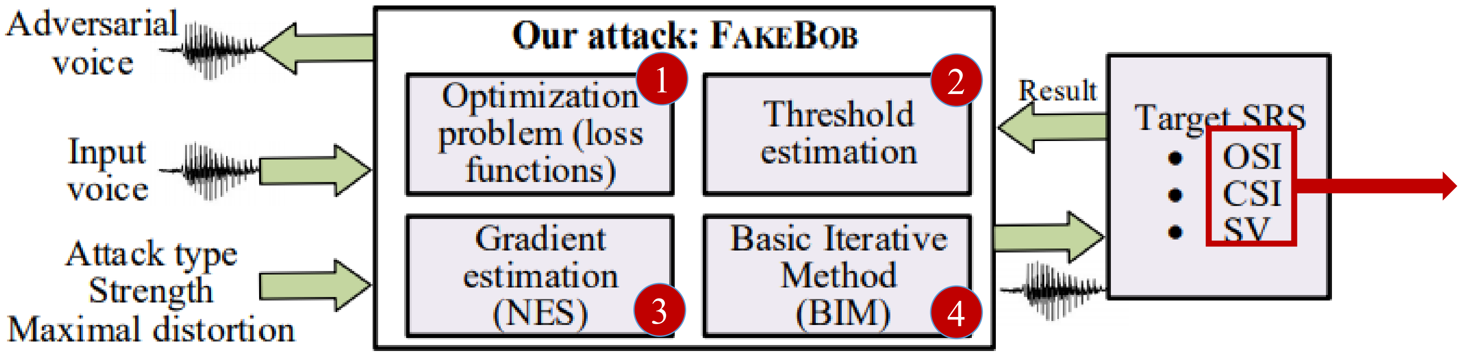


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

3 NES-based gradient estimation

white-box: backpropagation → exact gradient

Overview of FAKEBOB



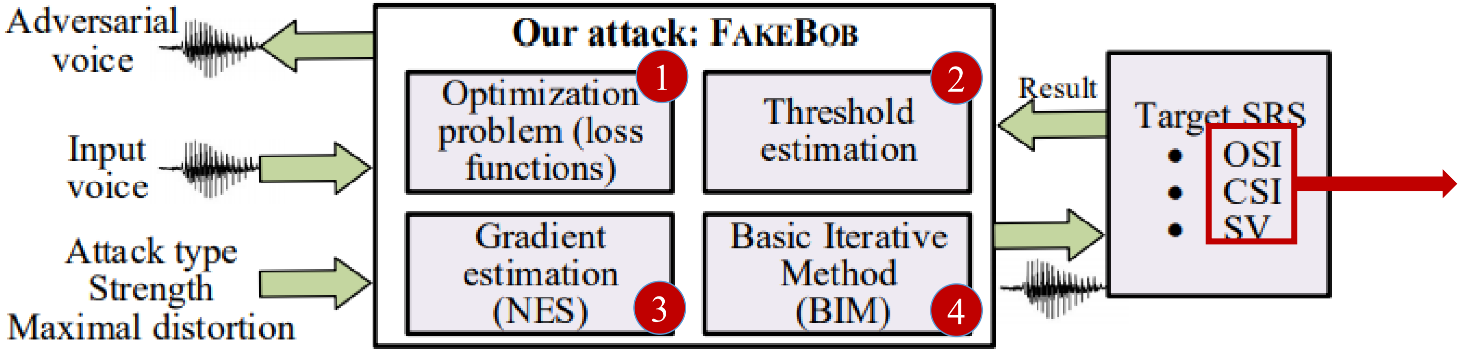
different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

3 NES-based gradient estimation

white-box: backpropagation → exact gradient

black-box: NES-based method → estimated gradient

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

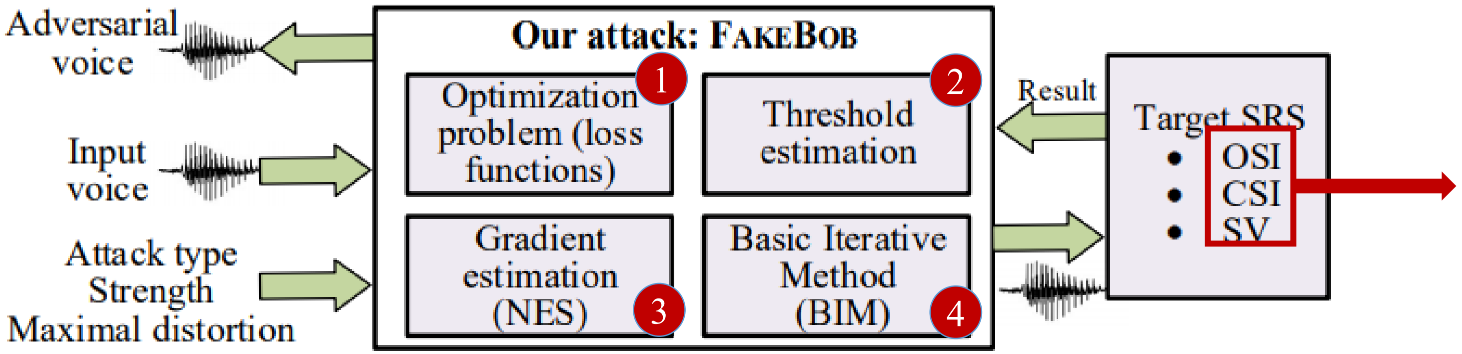
3 NES-based gradient estimation

white-box: backpropagation → exact gradient

black-box: NES-based method → estimated gradient

only rely on scores and decisions returned by victim speaker model

Overview of FAKEBOB



different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

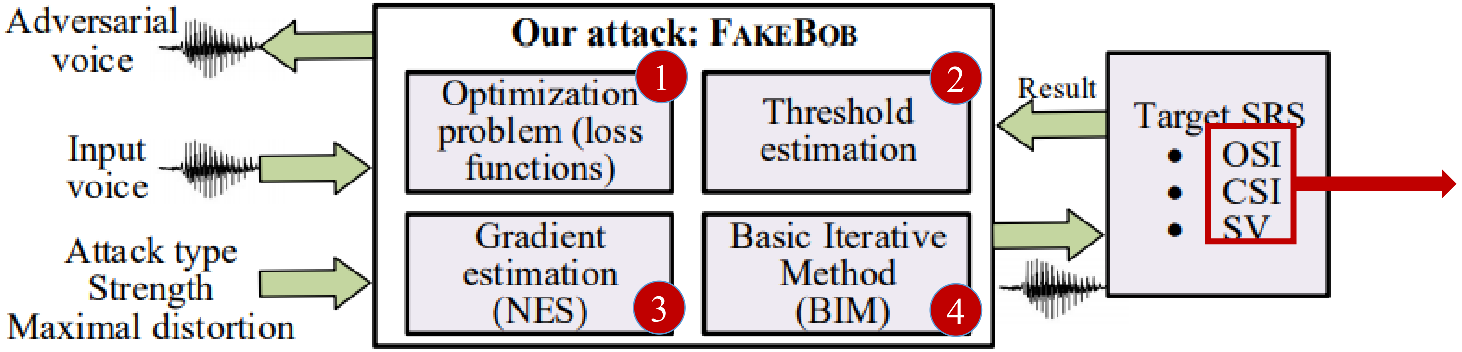
3 NES-based gradient estimation

white-box: backpropagation → exact gradient

black-box: NES-based method → estimated gradient

only rely on scores and decisions returned by victim speaker model → Black-box

Overview of FAKEBOB



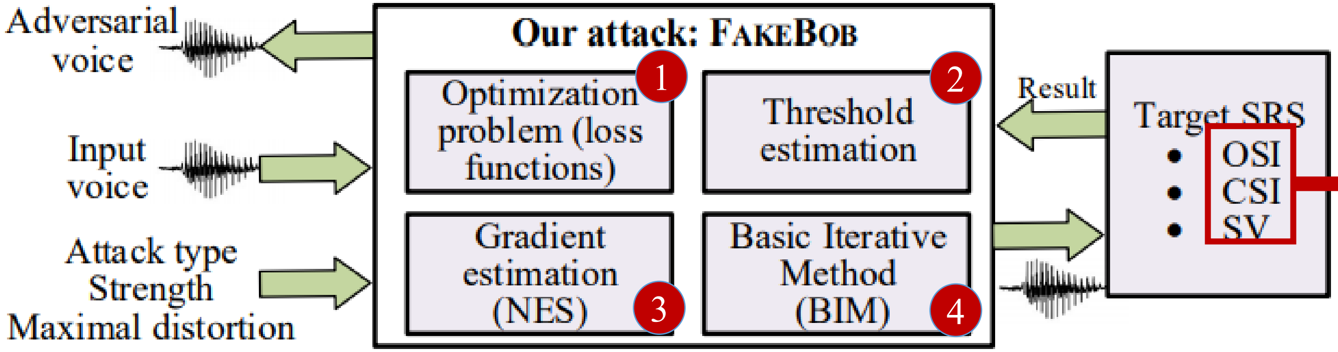
different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

3 NES-based gradient estimation

estimated gradient information

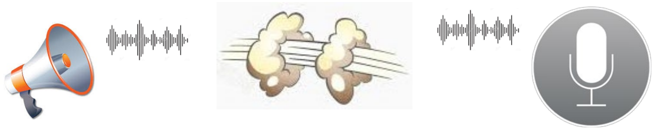
4 Solve the optimization problem by gradient descent

Overview of FAKEBOB

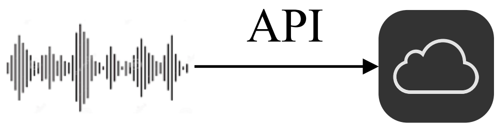


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

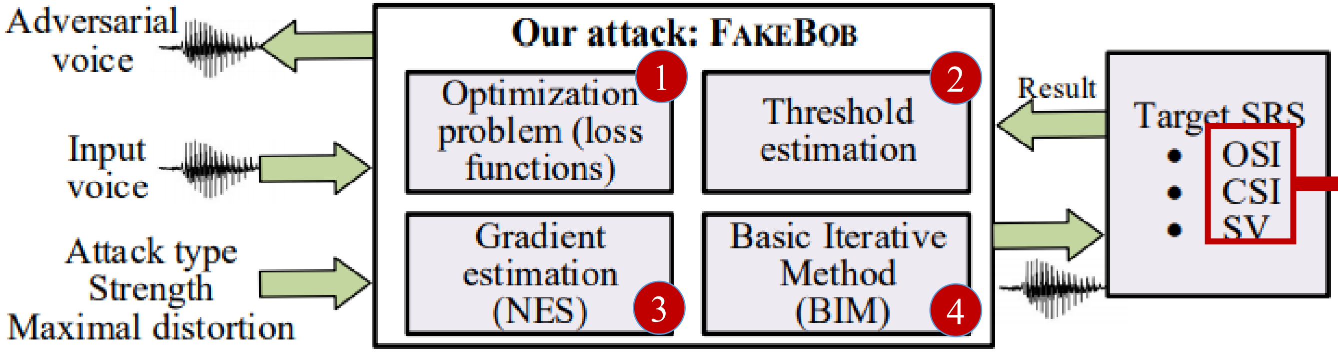
5 Over-the-air attack



API Attack

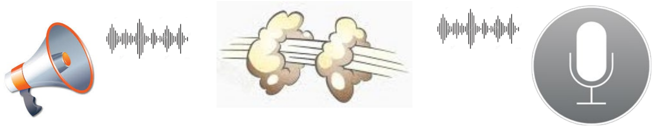


Overview of FAKEBOB

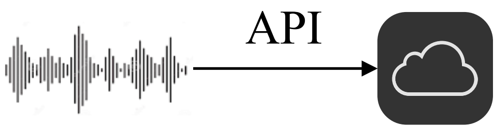


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

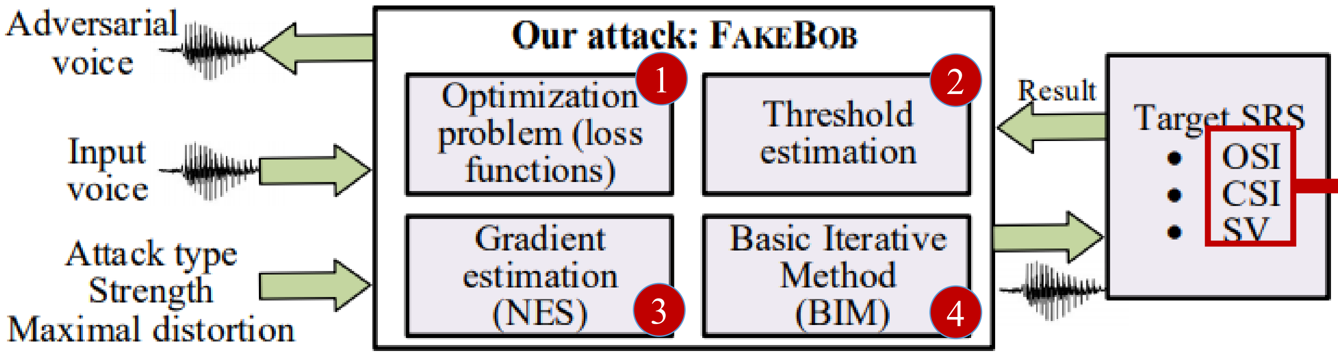
5 Over-the-air attack



API Attack

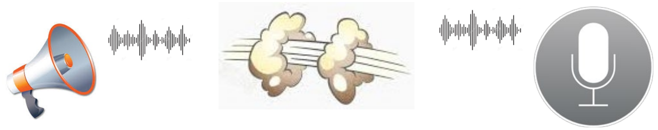


Overview of FAKEBOB

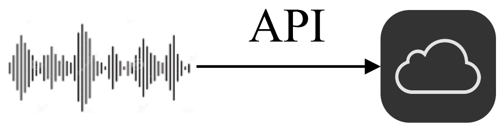


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

5 Over-the-air attack

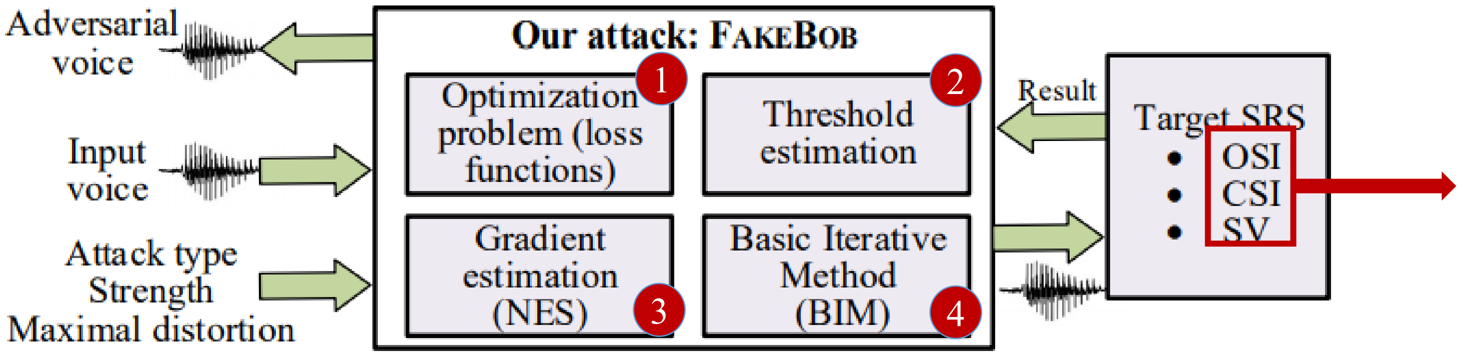


API Attack



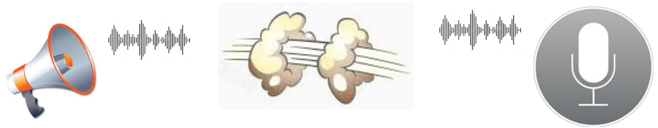
Challenge: noise in air channel makes attack ineffective

Overview of FAKEBOB

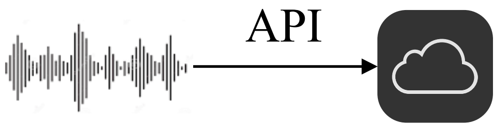


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

5 Over-the-air attack



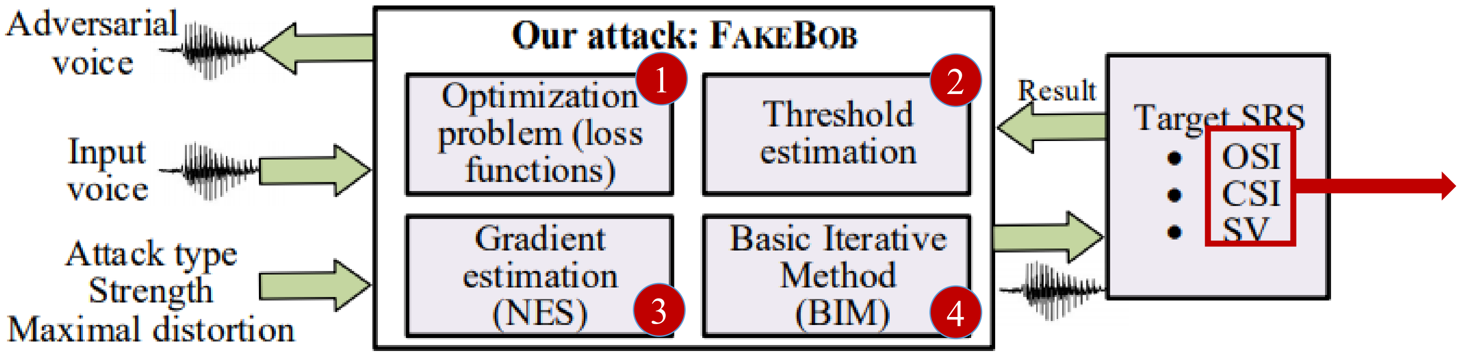
API Attack



Challenge: noise in air channel makes attack ineffective

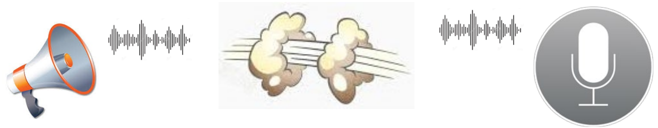
previous work: model the noise during generation

Overview of FAKEBOB

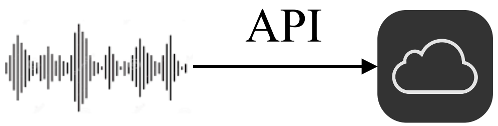


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

5 Over-the-air attack



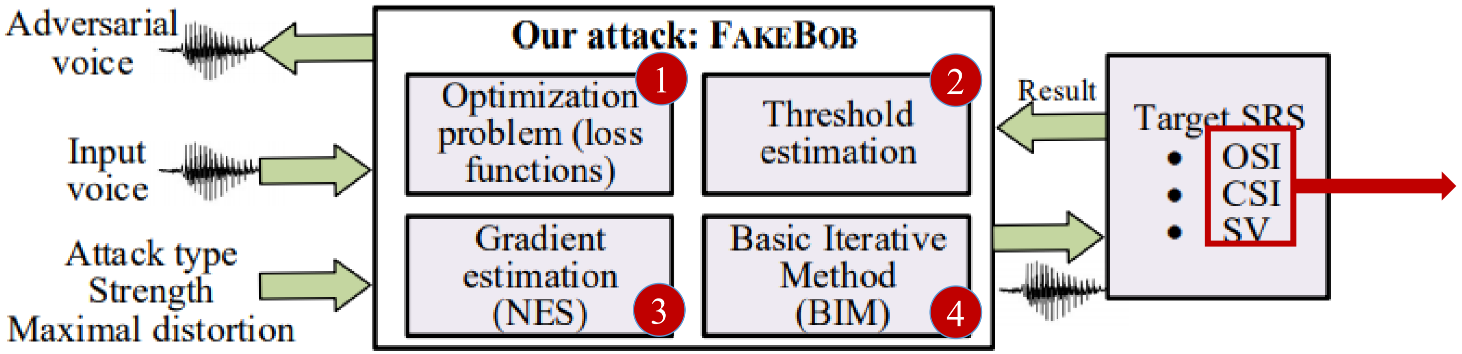
API Attack



Challenge: noise in air channel makes attack ineffective

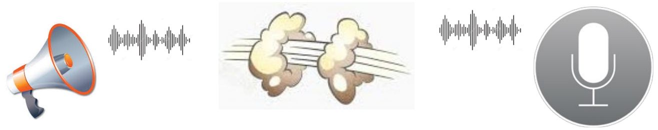
previous work: model the noise during generation → somehow environment- and device- dependent

Overview of FAKEBOB

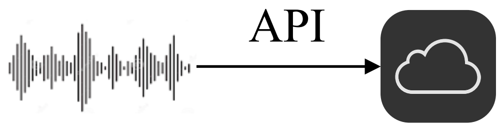


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

5 Over-the-air attack



API Attack



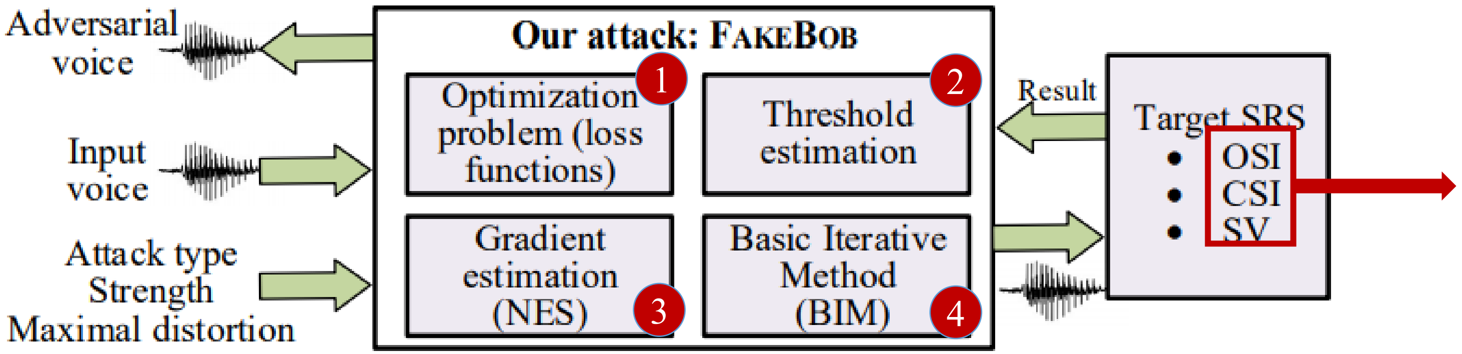
Challenge: noise in air channel makes attack ineffective

previous work: model the noise during generation → somehow environment- and device- dependent

ours: improve confidence κ

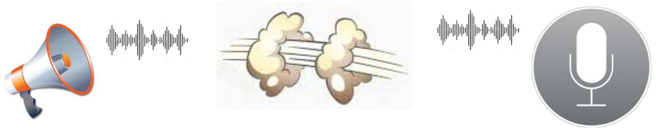
$$f(x) = \max\{\hat{\theta}, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Overview of FAKEBOB

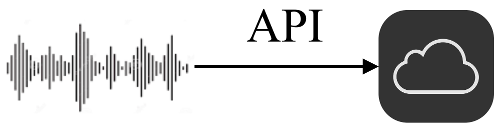


different tasks:
OSI: open-set identification
CSI: close-set identification
SV: speaker verification

5 Over-the-air attack



API Attack



Challenge: noise in air channel makes attack ineffective

previous work: model the noise during generation → somehow environment- and device- dependent

ours: improve confidence

Simple but Effective (will shown later)

Experimental result

Experimental result

- Attack **Open-source** KALDI

Experimental result

- Attack **Open-source** KALDI

✔ ≈ 100% attack success rate (ASR)

Experimental result

- Attack **Open-source** KALDI
  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**

Experimental result

- Attack **Open-source** KALDI
  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - **Talentedsoft** return scores and decisions

Experimental result

- Attack **Open-source** KALDI

✔ ≈ 100% attack success rate (ASR)

- Attack **Commercial**

✔ TalentedSoft **Talentedsoft** return scores and decisions → direct attack by query

Experimental result

- Attack **Open-source** KALDI
  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - TalentedSoft **Talentedsoft** return scores and decisions → direct attack by query

    100% ASR; 2500 query on average

Experimental result

- Attack **Open-source** KALDI
  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - TalentedSoft **Talentedsoft** return scores and decisions → direct attack by query
    100% ASR; 2500 query on average
  - Microsoft Azure **Microsoft Azure** return only decisions

Experimental result

- Attack **Open-source** KALDI
  - ✔ ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - ✔ TalentedSoft **Talentedsoft** return scores and decisions → direct attack by query
    100% ASR; 2500 query on average
  - ✔ Microsoft Azure **Microsoft Azure** return only decisions → transfer attack

Experimental result

- Attack **Open-source** KALDI
  - ✓ ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - ✓ TalentedSoft **Talentedsoft** return scores and decisions → direct attack by query
    100% ASR; 2500 query on average
  - ✓ Microsoft Azure **Microsoft Azure** return only decisions → transfer attack
    26% ASR

Experimental result

- Attack **Open-source**

  - ≈ 100% attack success rate (ASR)

- Attack **Commercial**

  - **Talentedsoft** return scores and decisions → direct attack by query

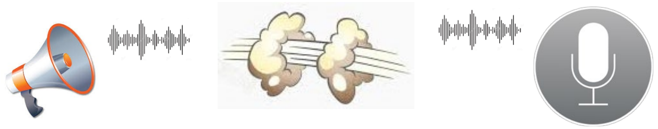
    100% ASR; 2500 query on average

  - **Microsoft Azure** return only decisions → transfer attack

    26% ASR

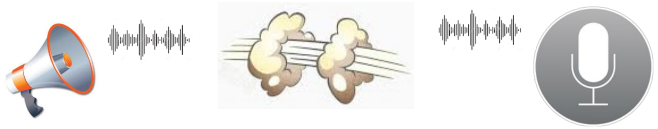
Experimental result

- Over the air Attack



Experimental result

- Over the air Attack

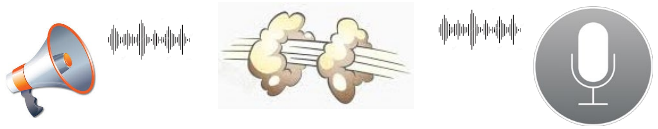


  - different distance between loundspeaker and microphone

Distance (meter)	**0.25**	**0.5**	**1**	**2**	**4**	**8**
ASR (%)	100	100	100	70	40	10

Experimental result

- Over the air Attack



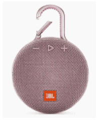
- ✔ different distance between loundspeaker and microphone

Distance (meter)	**0.25**	**0.5**	**1**	**2**	**4**	**8**
ASR (%)	100	100	100	70	40	10

- ✔ Different devices (at least 70% ASR)

Loundspeaker:

Laptop

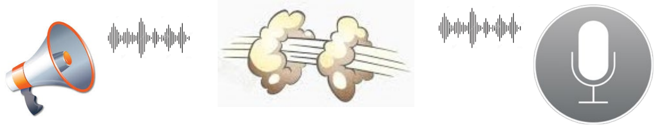


JBL portable speaker

Shinco broadcast equipment

Experimental result

- Over the air Attack



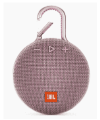
- different distance between loundspeaker and microphone

Distance (meter)	0.25	0.5	1	2	4	8
ASR (%)	100	100	100	70	40	10

- Different devices (at least 70% ASR)

Loundspeaker:

Laptop



JBL portable speaker

Shinco broadcast equipment

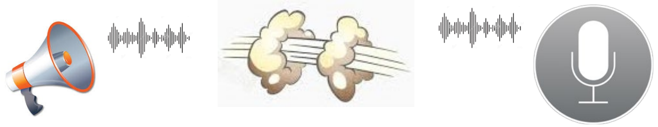
Microphone:

Apple iPhone

OPPO

Experimental result

- Over the air Attack



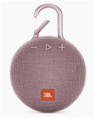
- different distance between loundspeaker and microphone

Distance (meter)	**0.25**	**0.5**	**1**	**2**	**4**	**8**
ASR (%)	100	100	100	70	40	10

- Different devices (at least 70% ASR)

Loundspeaker:

Laptop



JBL portable speaker

Shinco broadcast equipment

Microphone:

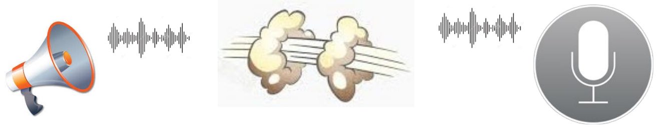
Apple iPhone

OPPO

Device independent

Experimental result

- **Over the air Attack**



- different distance between loundspeaker and microphone

Distance (meter)	0.25	0.5	1	2	4	8
ASR (%)	100	100	100	70	40	10

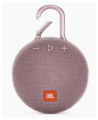
- different acoustic environments
  White / Bus / Restaurant / Music noise
  at least **48%** ASR when noise < 60 dB

Environment independent

- Different devices (at least 70% ASR)

Loundspeaker:

DELL Laptop



JBL portable speaker

Shinco broadcast equipment

Microphone:

Apple iPhone

OPPO

Device independent

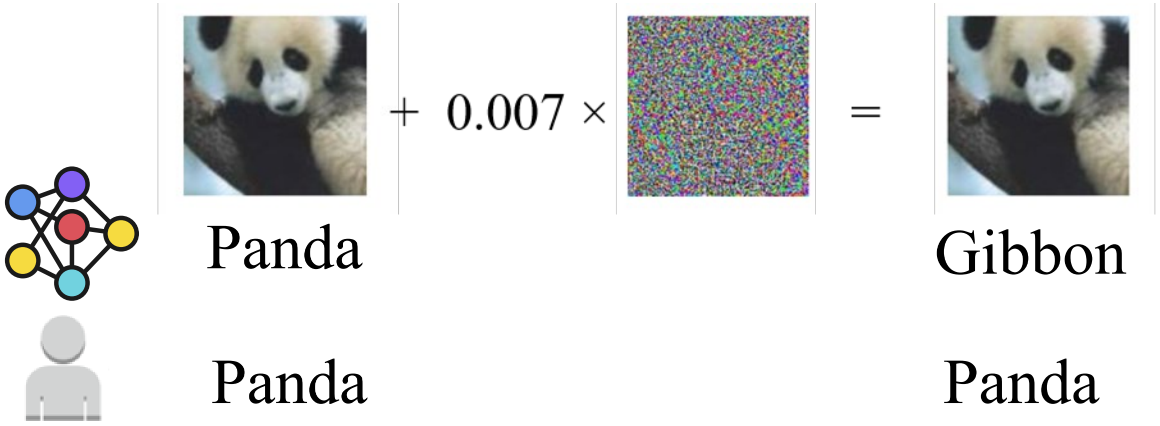
Imperceptibility

Imperceptibility

Imperceptibility has different meaning in different domains

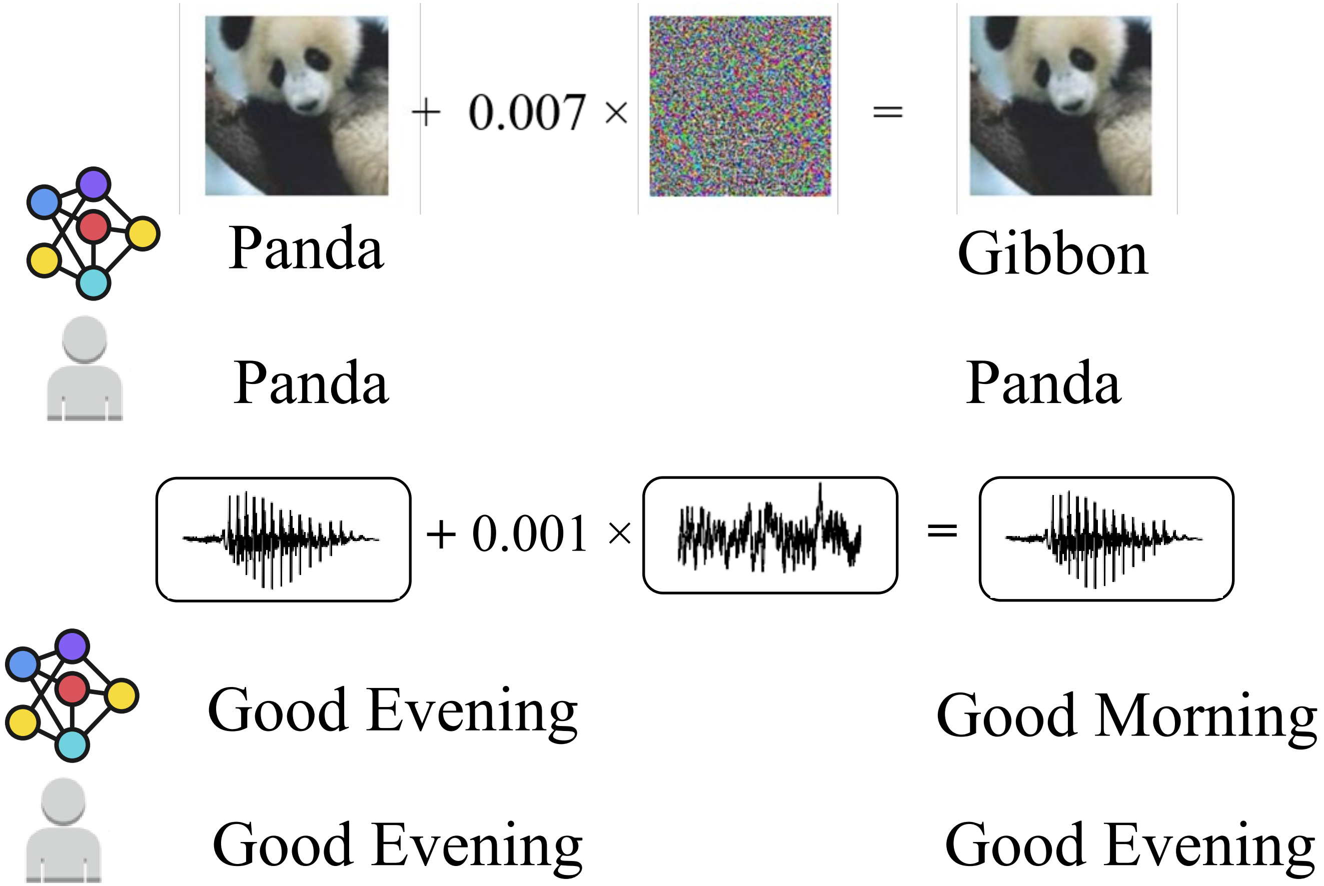
Imperceptibility

Imperceptibility has different meaning in different domains

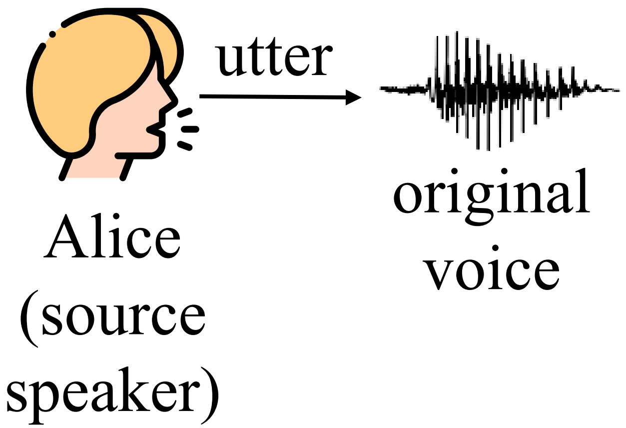


Imperceptibility

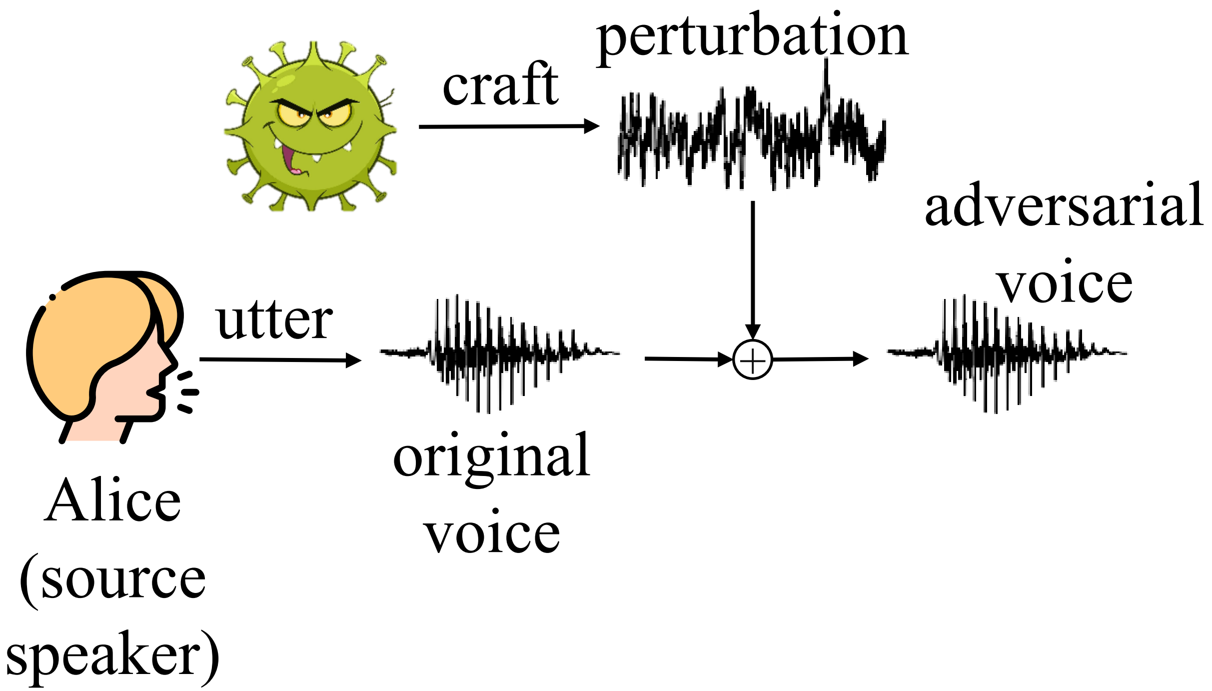
Imperceptibility has different meaning in different domains



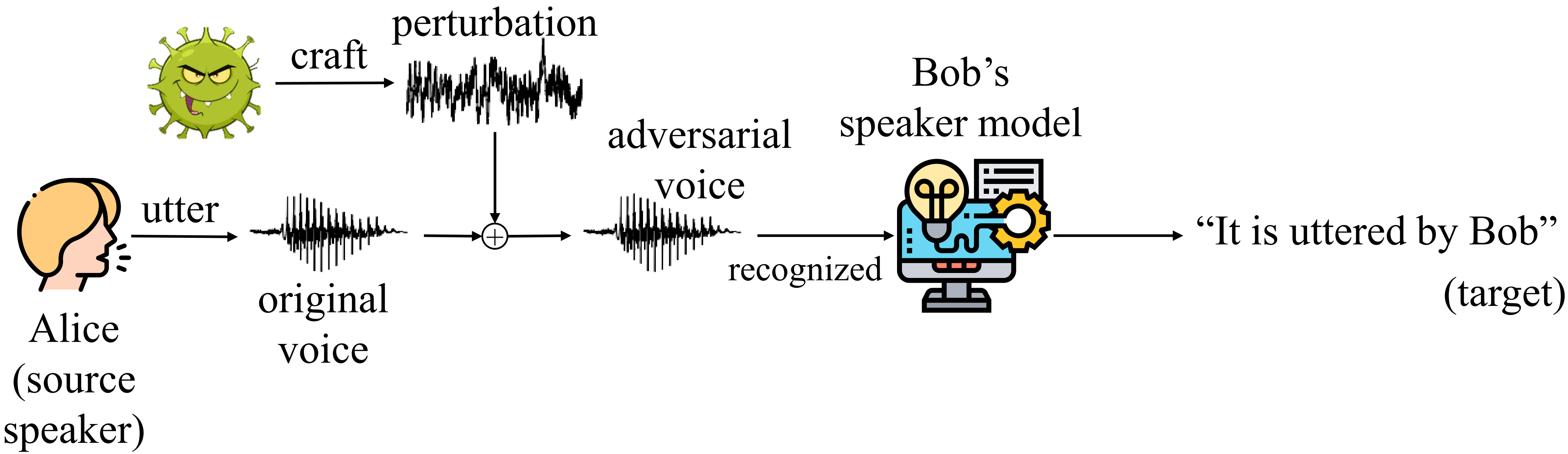
Imperceptibility



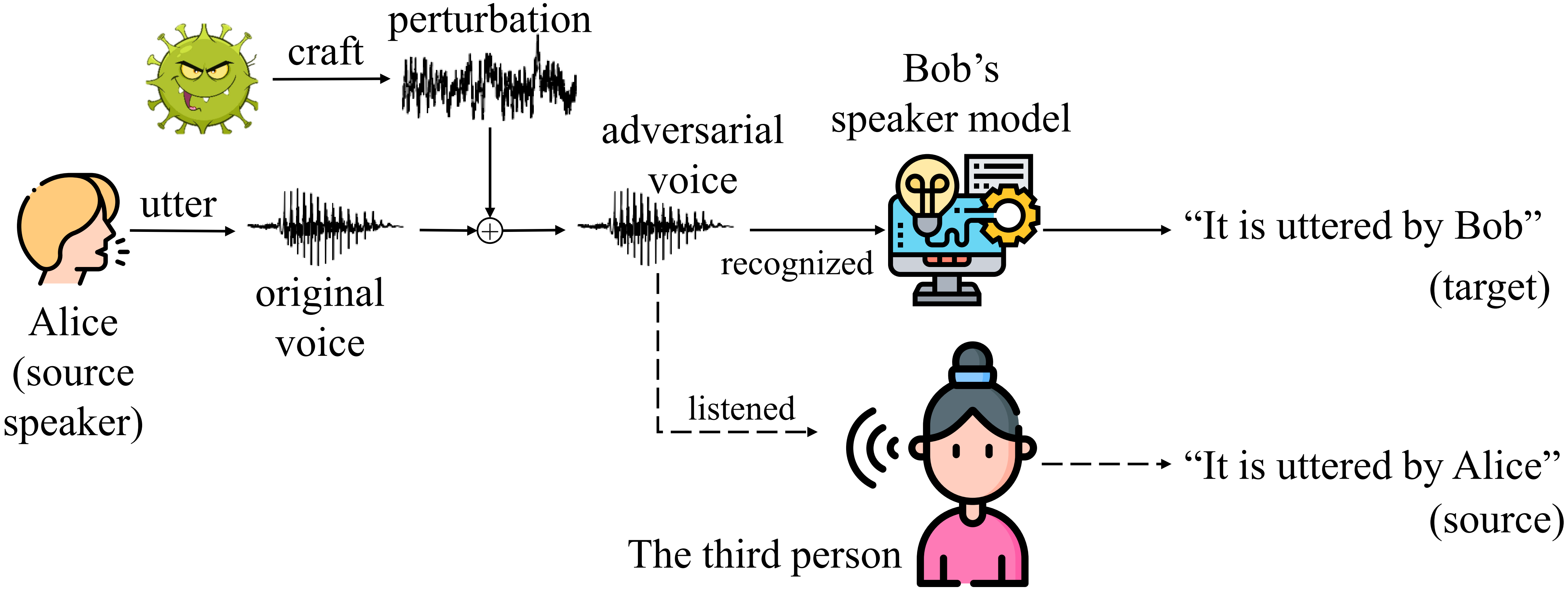
Imperceptibility



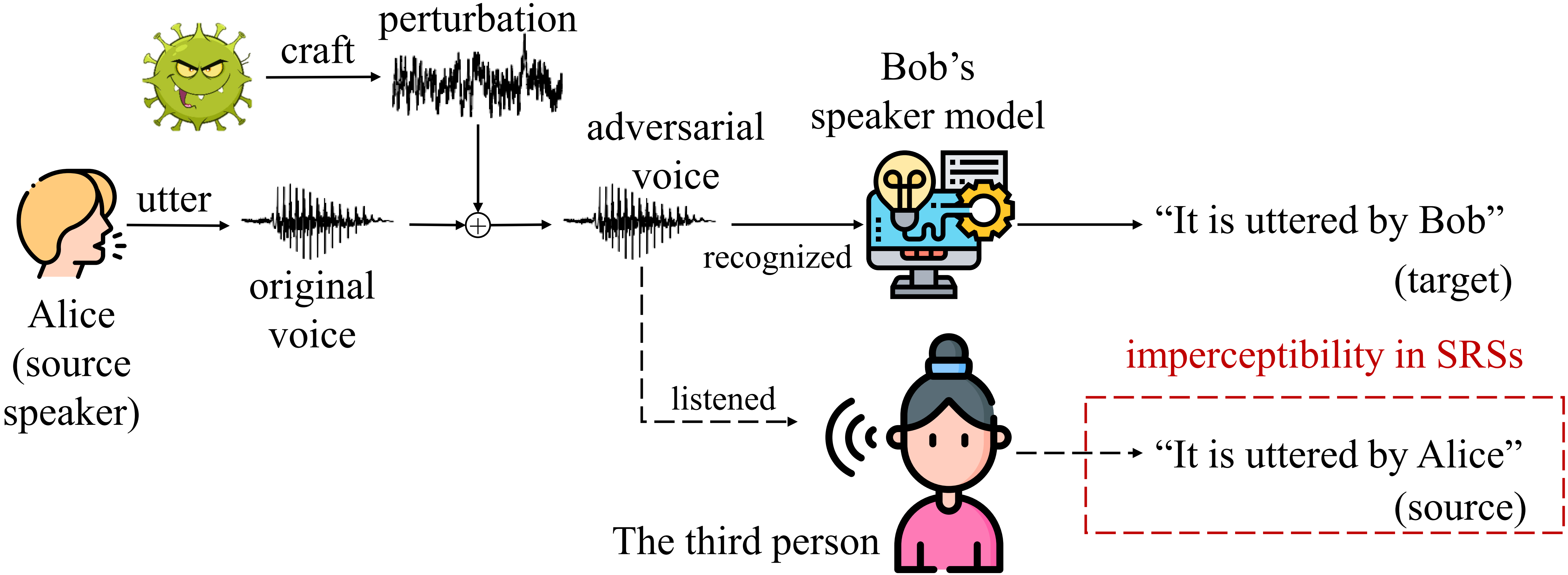
Imperceptibility



Imperceptibility



Imperceptibility



Imperceptibility

- quantitative analysis of imperceptibility

Imperceptibility

- quantitative analysis of imperceptibility

  Q: How many people think adversarial and original voices are uttered by the same speaker ?

Imperceptibility

- quantitative analysis of imperceptibility

  Q: How many people think adversarial and original voices are uttered by the same speaker ?

  A: Human Study on Amazon MTurk

Imperceptibility

- quantitative analysis of imperceptibility

  Q: How many people think adversarial and original voices are uttered by the same speaker ?

  A: Human Study on Amazon MTurk

  - API attack: 64.9% same

Imperceptibility

- quantitative analysis of imperceptibility

  Q: How many people think adversarial and original voices are uttered by the same speaker ?

  A: Human Study on Amazon MTurk

  - API attack: 64.9% same
  - Over-the-air attack: 34.0% same

Imperceptibility

- quantitative analysis of imperceptibility

  Q: How many people think adversarial and original voices are uttered by the same speaker ?

  A: Human Study on Amazon MTurk

  - API attack: 64.9% same
  - Over-the-air attack: 34.0% same

Take away:

1. Black-box and practical adversarial attack against speaker recognition systems
2. Effective to commercial speaker recognition services
3. Effective in over-the-air attack
4. Imperceptible to human hearing

S3L Lab
WeChat QR Code

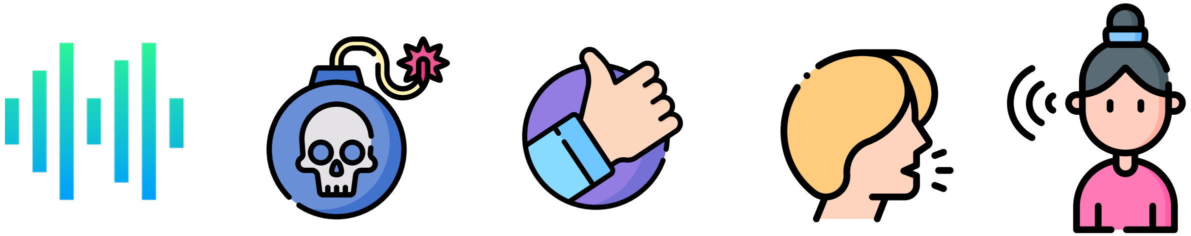
fakebob

FAKEBOB Website:
https://sites.google.com/view/fakebob/home

FAKEBOB Code:
https://github.com/FAKEBOB-adversarial-attack/FAKEBOB

S3Lab
系统与软件安全实验室
System and Software Security Laboratory

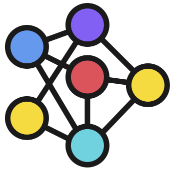
System and Software Security Lab (S3L), ShanghaiTech University:
http://s3l.shanghaitech.edu.cn/



Icon made by Freepik from www.flaticon.com

Icon made by Eucalyp from www.flaticon.com

Icon made by xnimrodx from www.flaticon.com



Icon made by Becris from www.flaticon.com

Take away:

1. Black-box and practical adversarial attack against speaker recognition systems
2. Effective to commercial speaker recognition services
3. Effective in over-the-air attack
4. Imperceptible to human hearing

S3L Lab
WeChat QR Code

FAKEBOB Website:
https://sites.google.com/view/fakebob/home

FAKEBOB Code:
https://github.com/FAKEBOB-adversarial-attack/FAKEBOB

System and Software Security Lab (S3L), ShanghaiTech University:
http://s3l.shanghaitech.edu.cn/