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

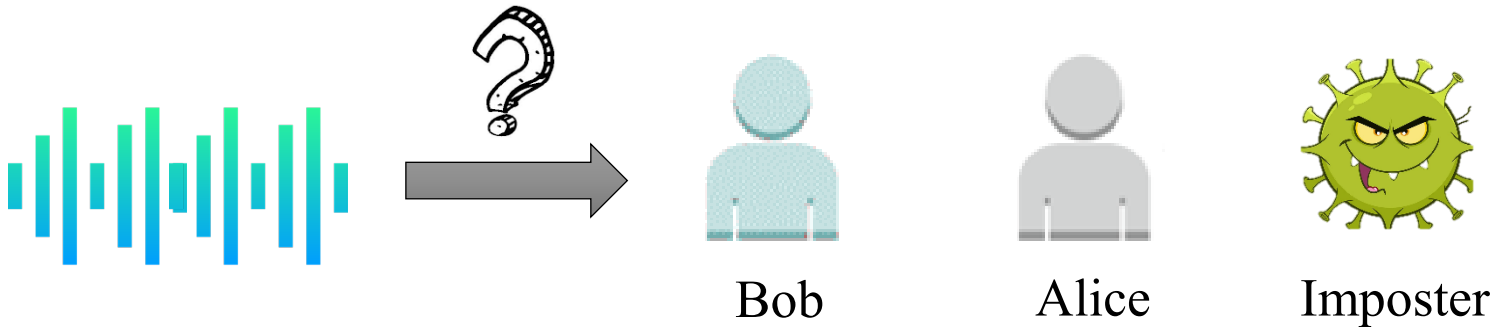
UCAS 中国科学院大学 University of Chinese Academy of Sciences

天津大学 Tianjin University

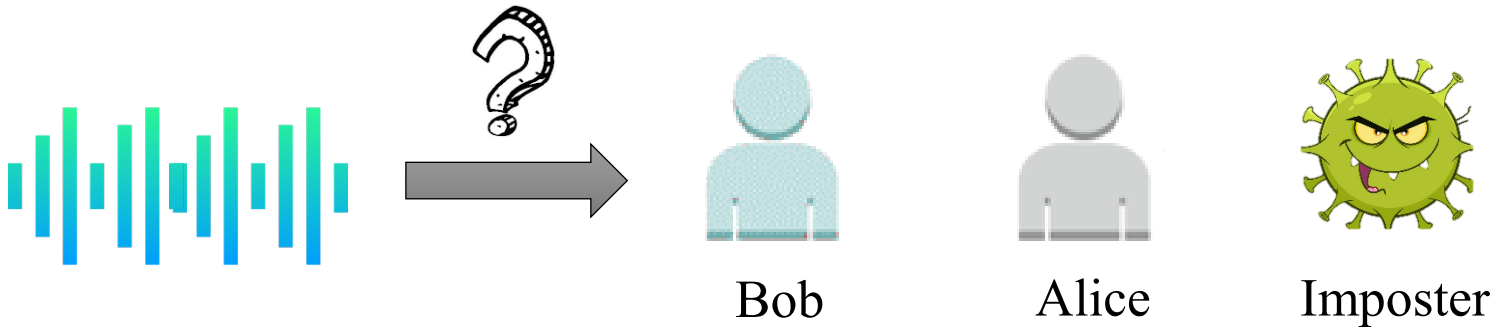
✉ Fu Song (songfu@shanghaitech.edu.cn)

Speaker Recognition Systems (SRSs)

Speaker Recognition Systems (SRSs)

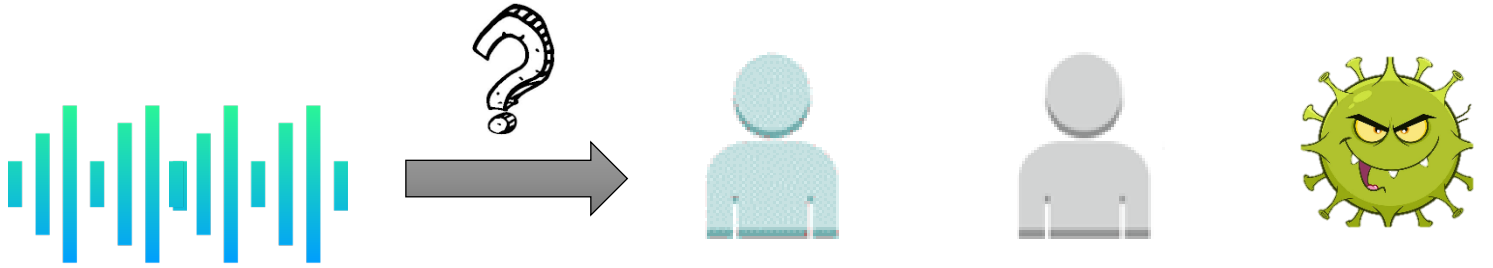


Speaker Recognition Systems (SRSs)



Ubiquitous Application

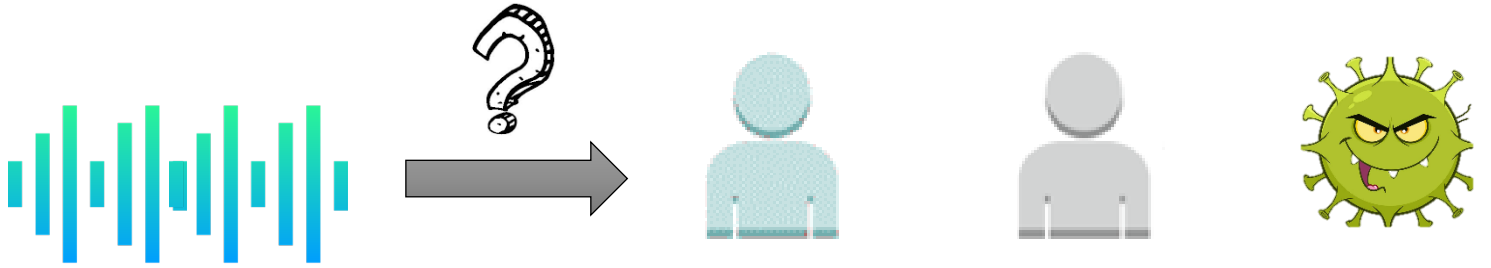
Speaker Recognition Systems (SRSs)



Ubiquitous Application

Voice assistant wake up

Speaker Recognition Systems (SRSs)



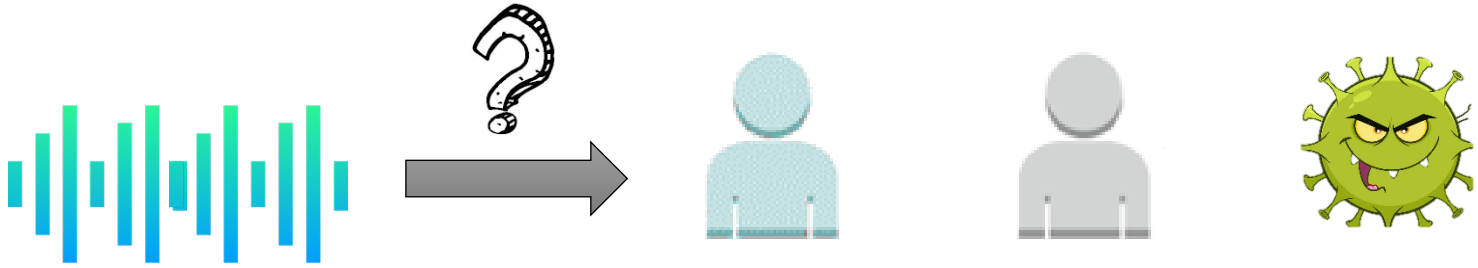
Bob Alice Imposter

Ubiquitous Application

Voice assistant wake up

Personalized service
on smart home

Speaker Recognition Systems (SRSs)



Bob Alice Imposter

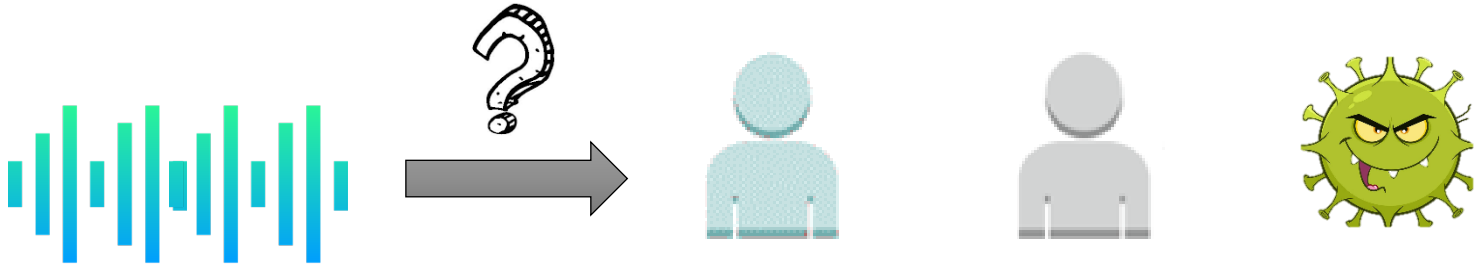
Ubiquitous Application

Voice assistant wake up

Personalized service on smart home

Financial transaction

Speaker Recognition Systems (SRSs)



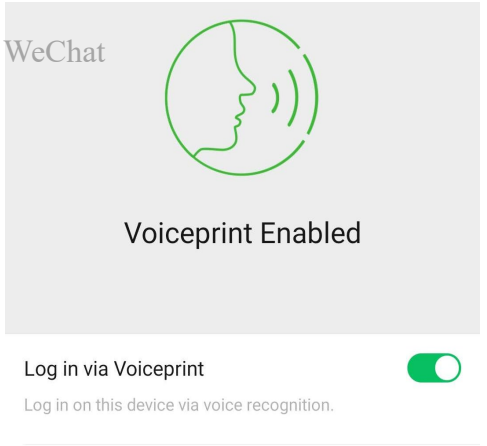
Bob Alice Imposter

Ubiquitous Application

Voice assistant wake up

Personalized service on smart home

Financial transaction



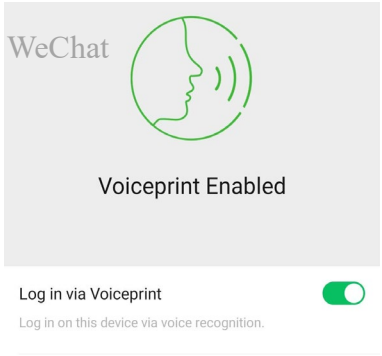
App log in

Ubiquitous Application

Voice assistant wake up

Personalized service on smart home

Financial transaction



App log in

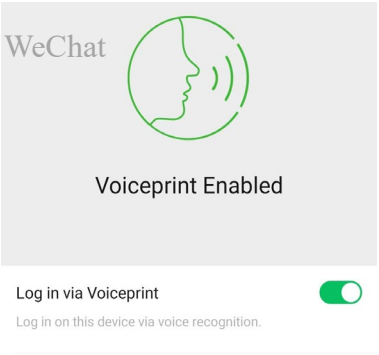
Safety-critical scenario

Ubiquitous Application

Voice assistant wake up

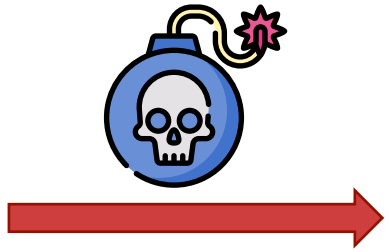
Personalized service on smart home

Financial transaction



App log in

Safety-critical scenario



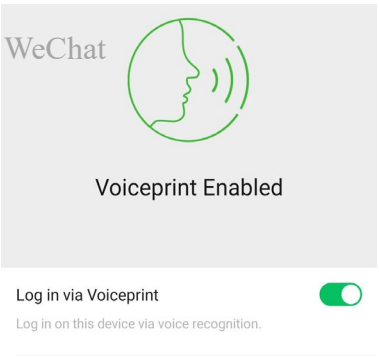
Once broken

Ubiquitous Application

Voice assistant wake up

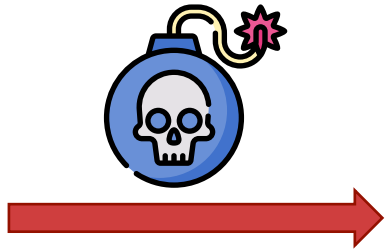
Personalized service on smart home

Financial transaction



App log in

Safety-critical scenario



Once broken

property damage

reputation degrade

sensitive information leak

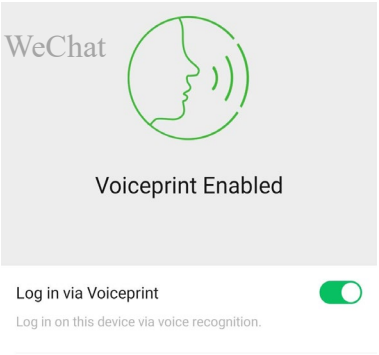
…

Ubiquitous Application

Voice assistant wake up

Personalized service on smart home

Financial transaction

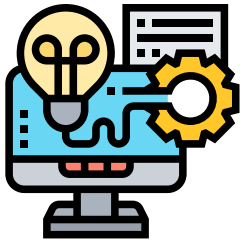


App log in

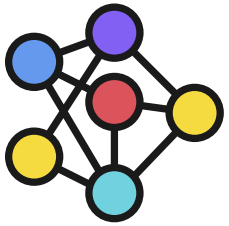
Security of SRSs!!!

Mainstream implementation of SRSs

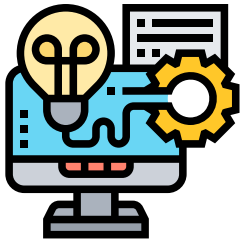
Mainstream implementation of SRSs



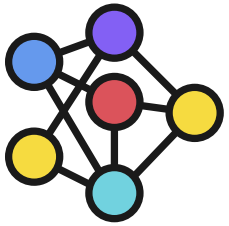
Machine Learning
(ML)



Mainstream implementation of SRSs

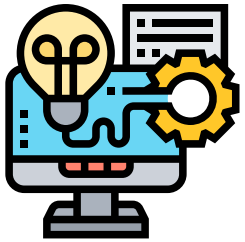


Machine Learning
(ML)

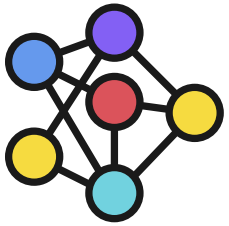


However,

Mainstream implementation of SRSs

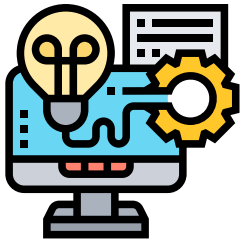


Machine Learning
(ML)

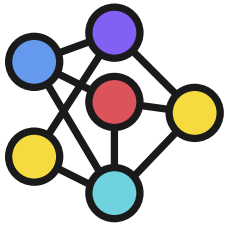


However, ML is vulnerable to adversarial examples

Mainstream implementation of SRSs



Machine Learning
(ML)



However, ML is vulnerable to adversarial examples

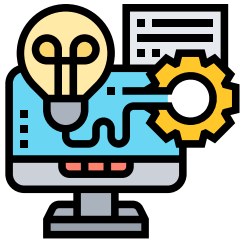
Benign example
Result: Panda
Confidence: 57.7%

Perturbation

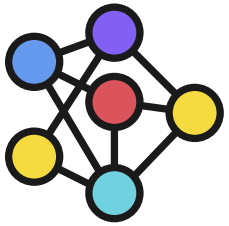
Adversarial example
Result: Gibbon
Confidence: 99.3%

Ian Goodfellow et al.

Mainstream implementation of SRSs



Machine Learning
(ML)



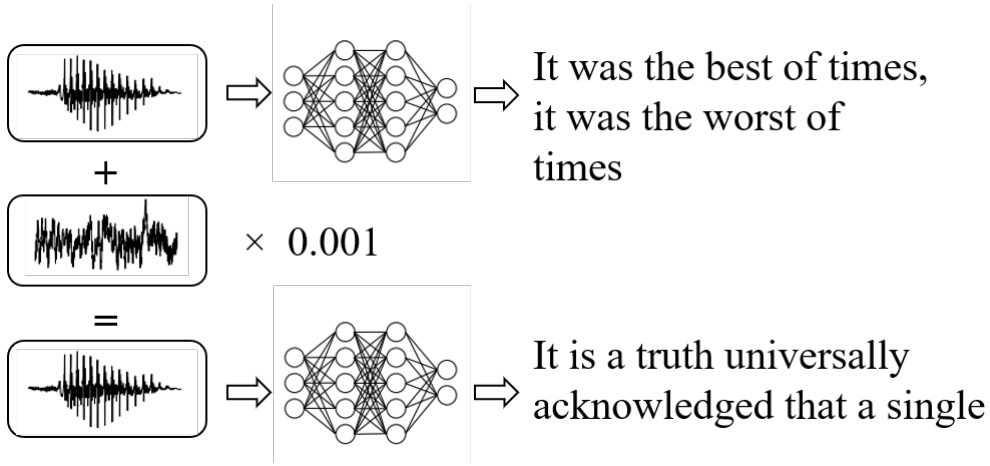
However, ML is vulnerable to adversarial examples

Benign example
Result: Panda
Confidence: 57.7%

Perturbation

Adversarial example
Result: Gibbon
Confidence: 99.3%

Ian Goodfellow et al.

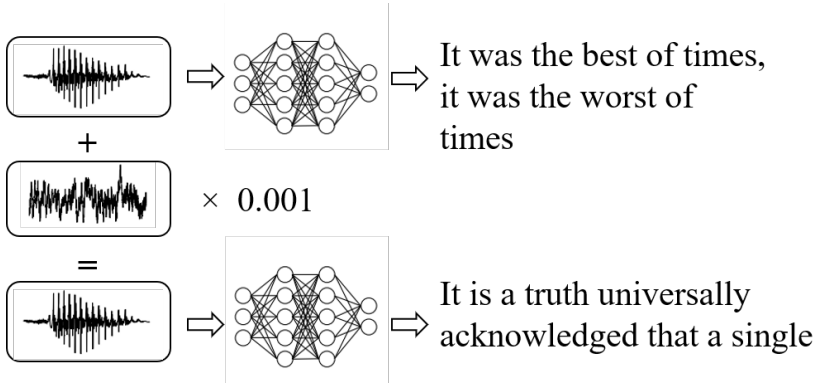


Nicholas Carlini et al.

Benign example
Result: Panda
Confidence: 57.7%

Perturbation

Adversarial example
Result: Gibbon
Confidence: 99.3%

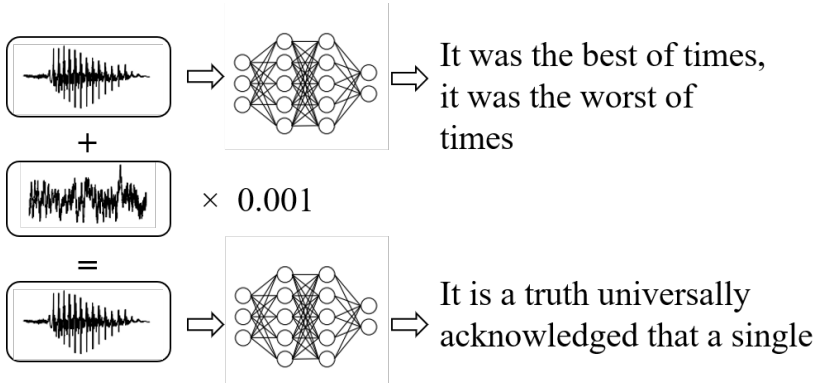


Is adversarial attack practical on SRSs ?

Benign example
Result: Panda
Confidence: 57.7%

Perturbation

Adversarial example
Result: Gibbon
Confidence: 99.3%



Is adversarial attack practical on SRSs ?

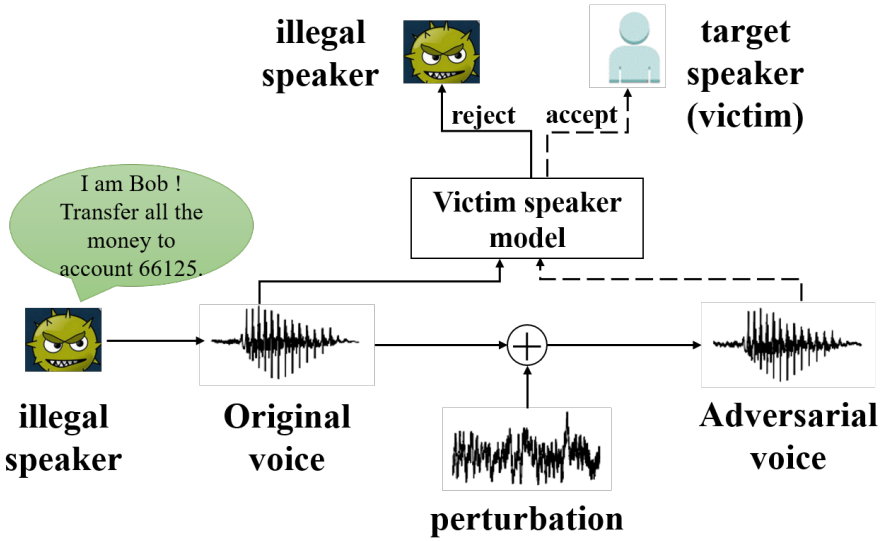
FAKEBOB

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

Threat model

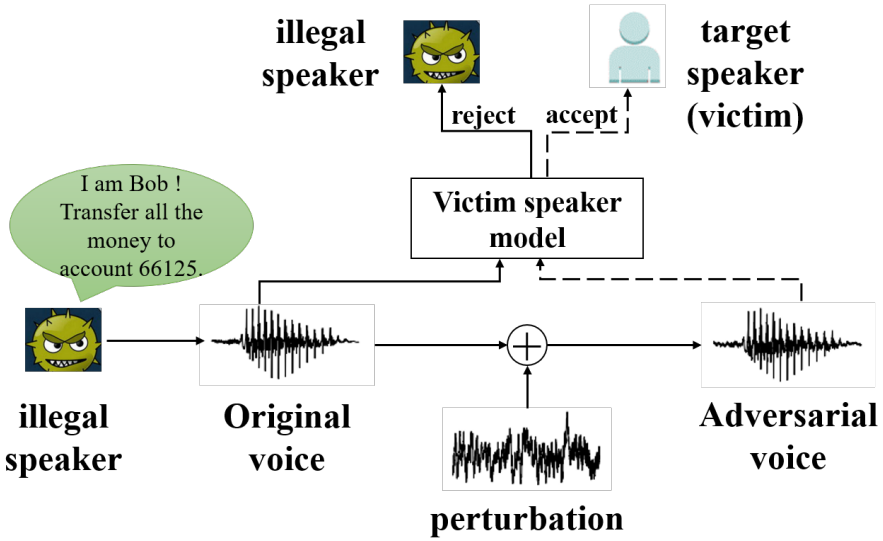
Threat model

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



Threat model

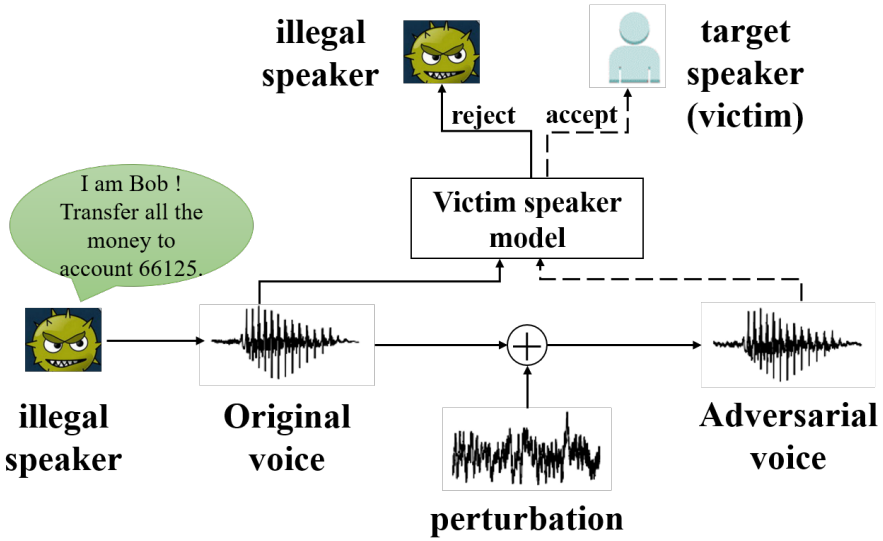
- Attacker Goal: pass voice authentication; gain access to privilege



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

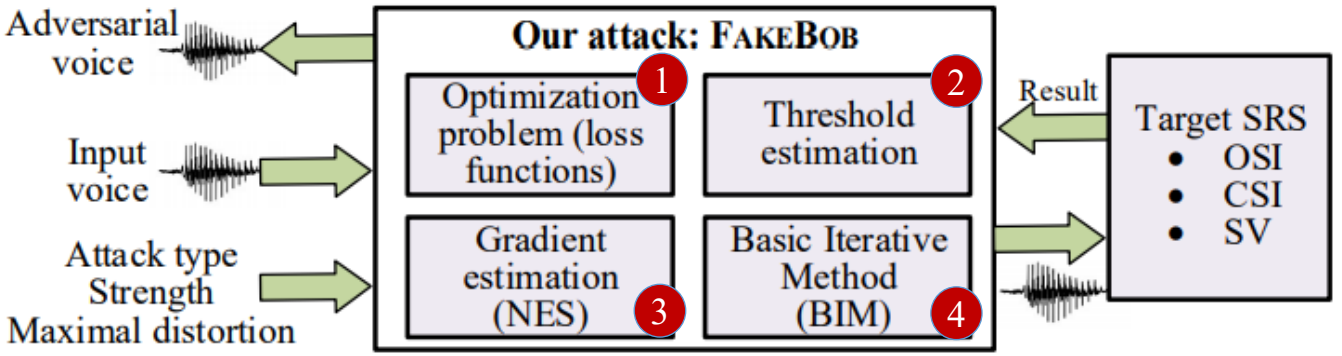
Threat model

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

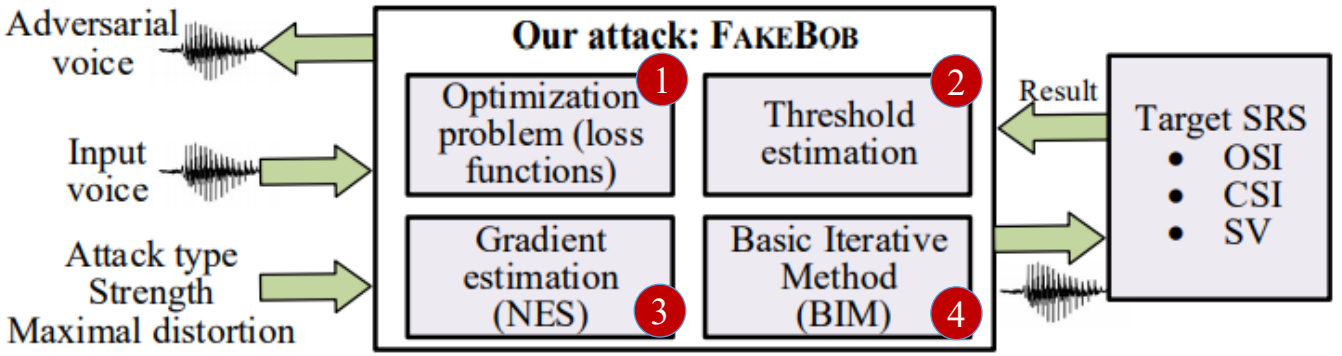


- Attacker Capability: no information about model structure / parameter; limited to query the speak model of the victims

Overview of FAKEBOB

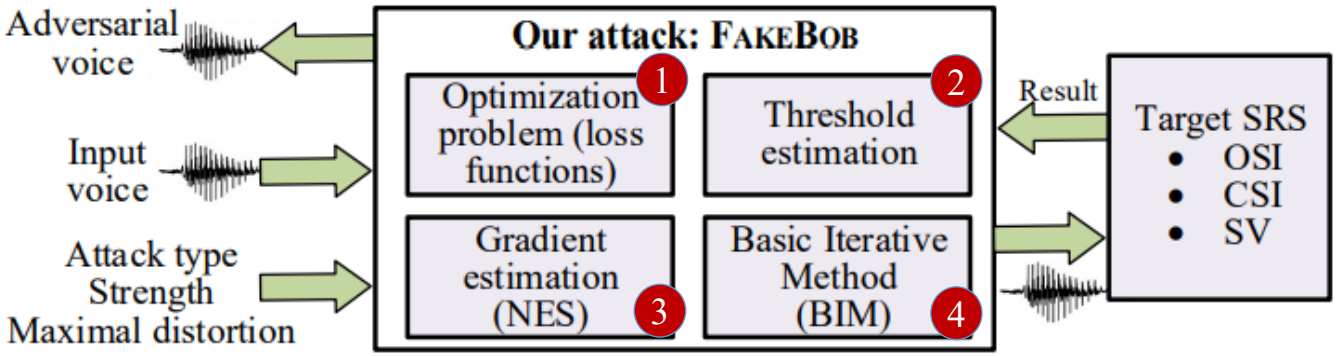


Overview of FAKEBOB



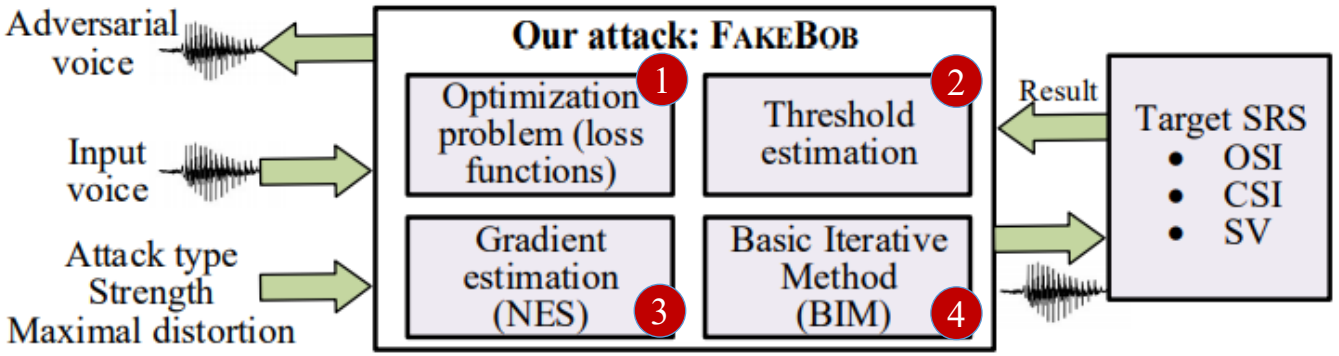
1. Effective **loss function** design.

Overview of FAKEBOB



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

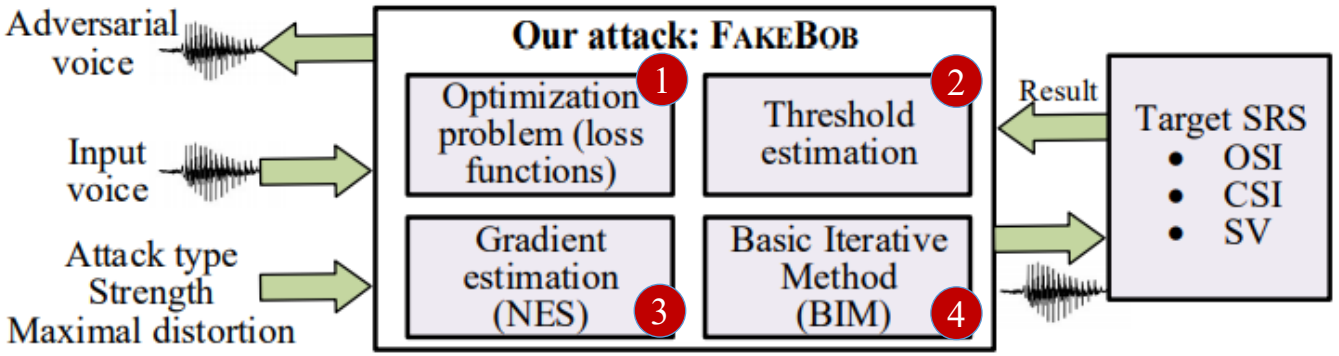
Overview of FAKEBOB



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

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

Overview of FAKEBOB



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

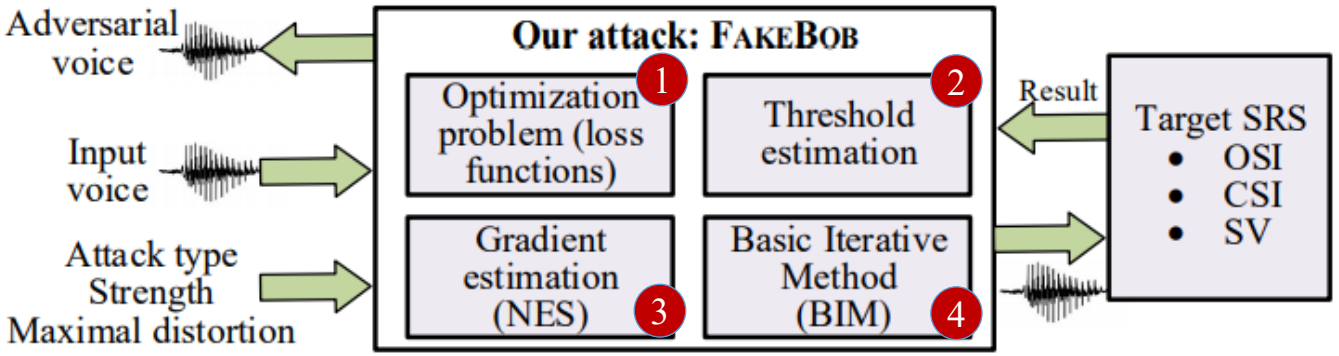
   Based on **scoring** and **decision-making** mechanism

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

Open-set identification (OSI) task

θ: threshold

Overview of FAKEBOB



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

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

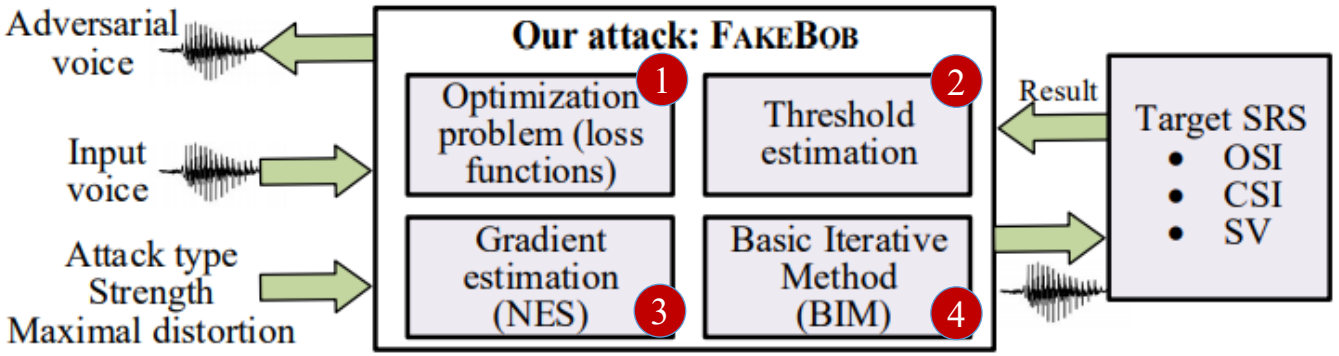
$$f(x) = \max\{\theta, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Open-set identification (OSI) task

θ: threshold

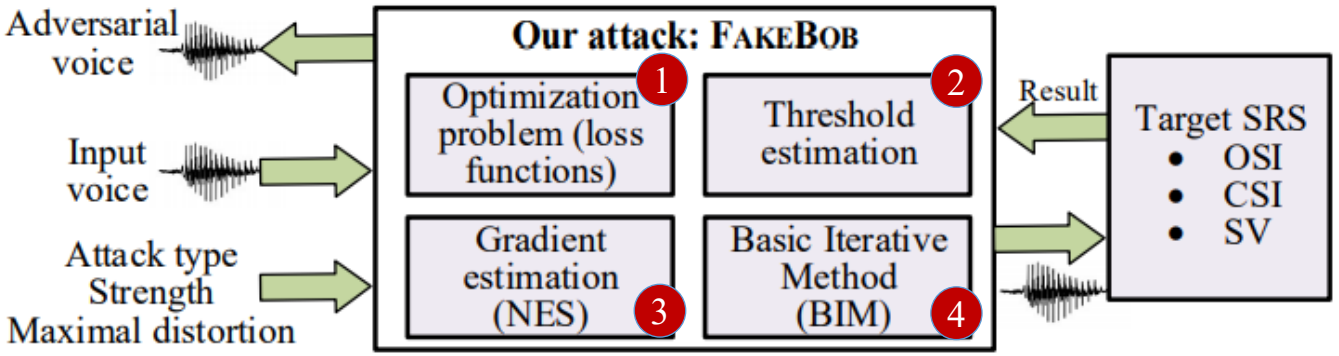
Tailored for different SRSs **tasks**: CSI, SV, OSI

Overview of FAKEBOB



2 **Threshold**: unique in VPR; unknown to attacker

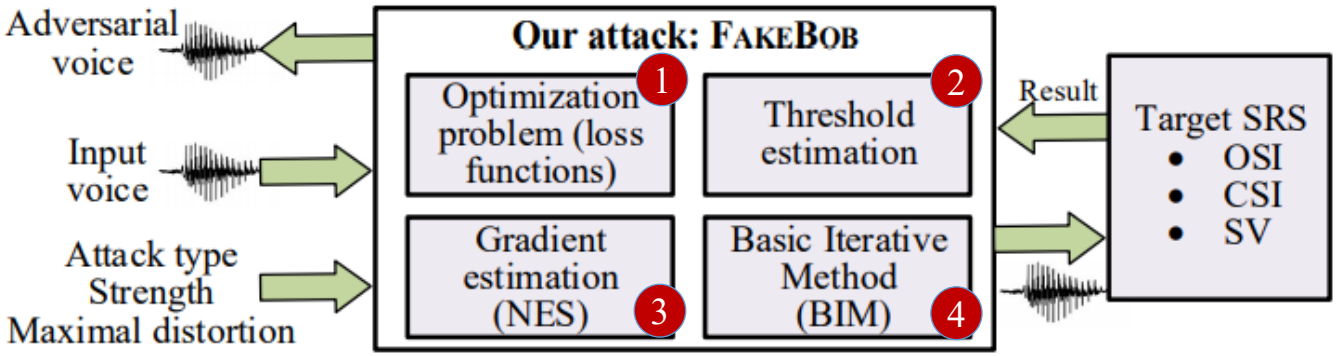
Overview of FAKEBOB



2 **Threshold**: unique in VPR; unknown to attacker

Novel **threshold estimation** algorithm

Overview of FAKEBOB

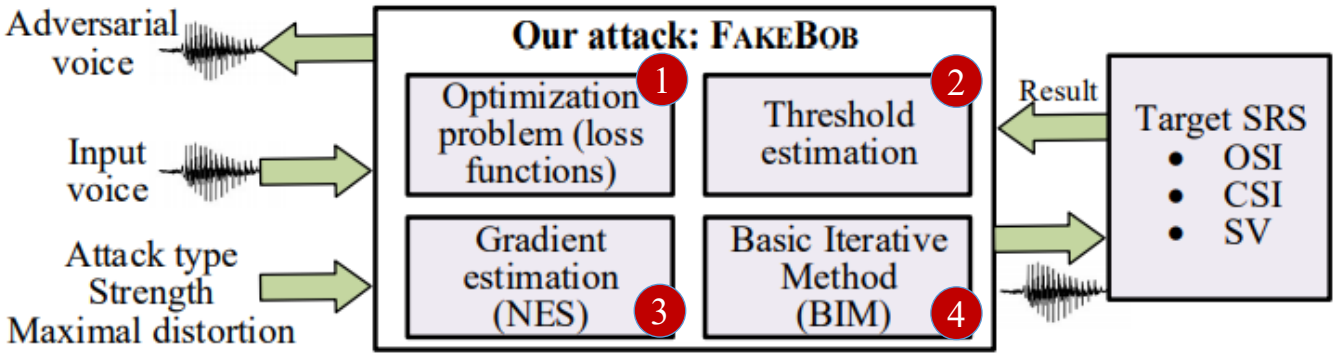


2 **Threshold**: unique in VPR; unknown to attacker

Novel **threshold estimation** algorithm

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

Overview of FAKEBOB

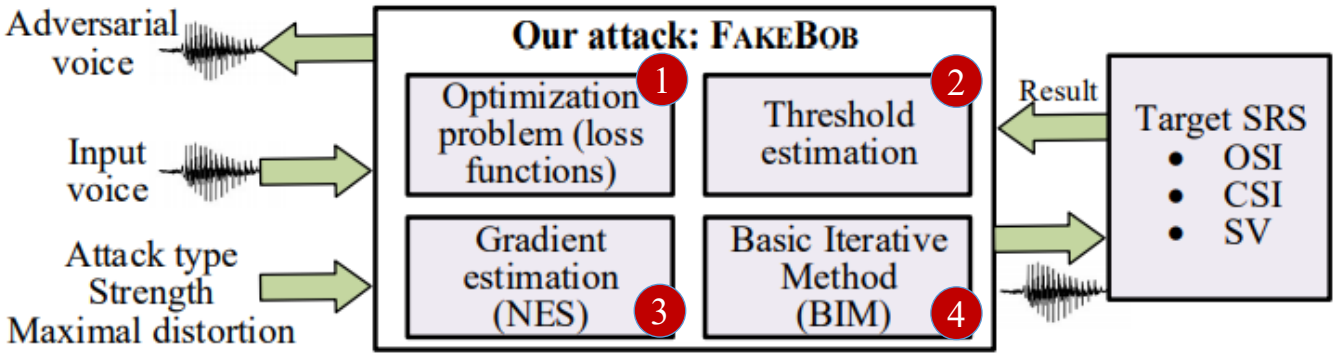


2 **Threshold**: unique in VPR; unknown to attacker

Novel **threshold estimation** algorithm

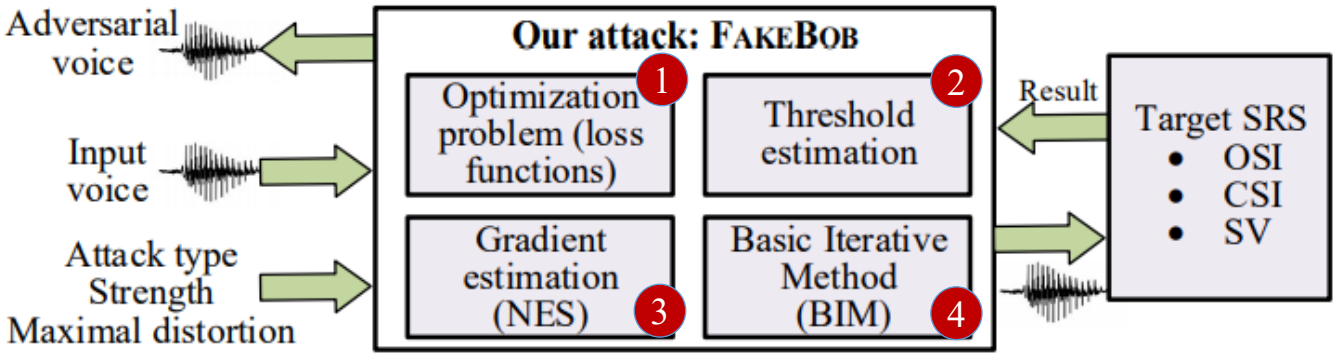
$$f(x) = \max\{\theta, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t \xrightarrow{\hat{\theta} \approx \theta} f(x) = \max\{\hat{\theta}, \max_{i \neq t}[S(x)]_i\} + \kappa - [S(x)]_t$$

Overview of FAKEBOB



3 NES-based gradient estimation

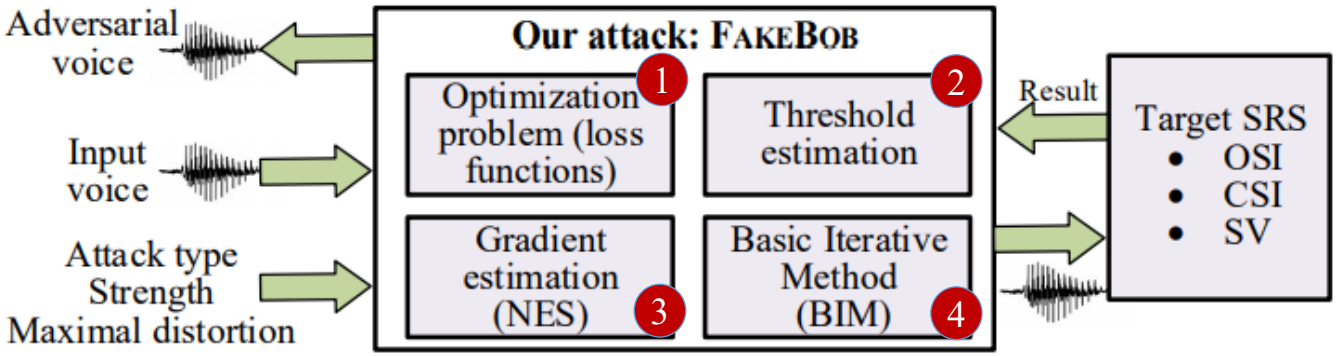
Overview of FAKEBOB



3 NES-based gradient estimation

rely on scores and decisions by querying victim speaker model

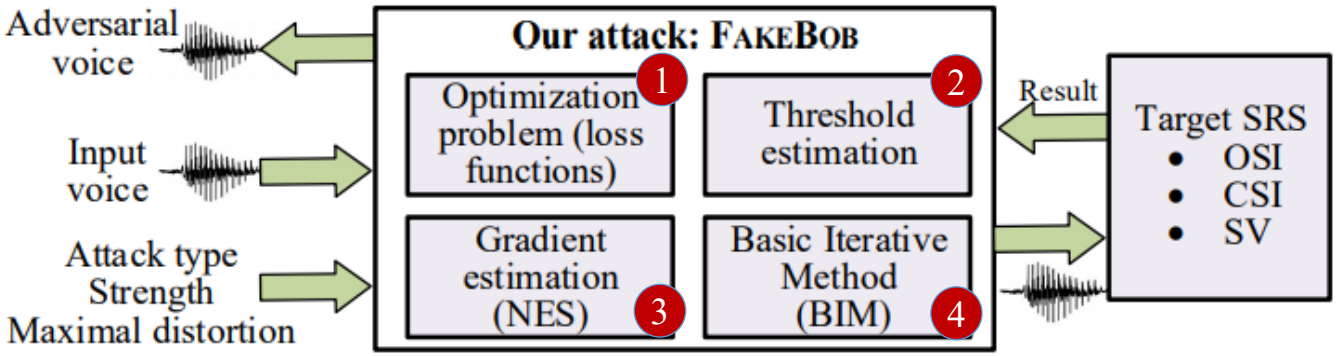
Overview of FAKEBOB



3 NES-based gradient estimation

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

Overview of FAKEBOB

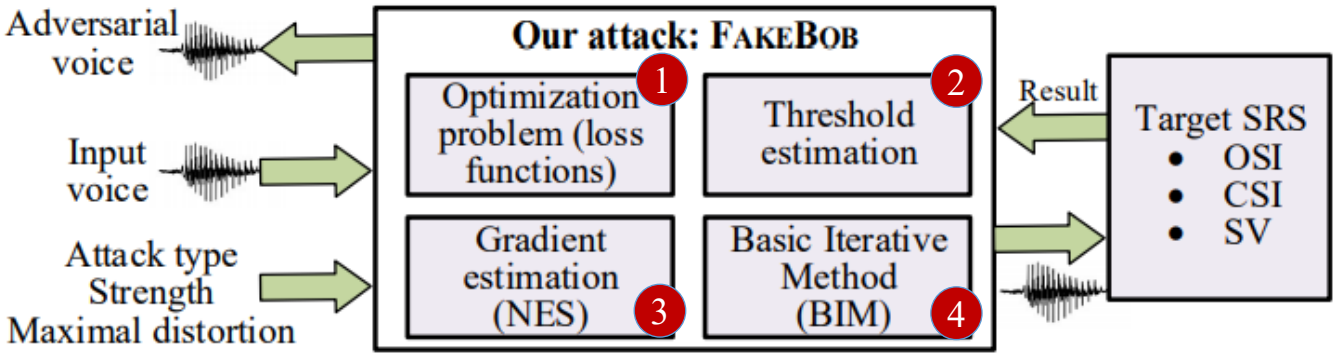


3. NES-based gradient estimation

gradient information

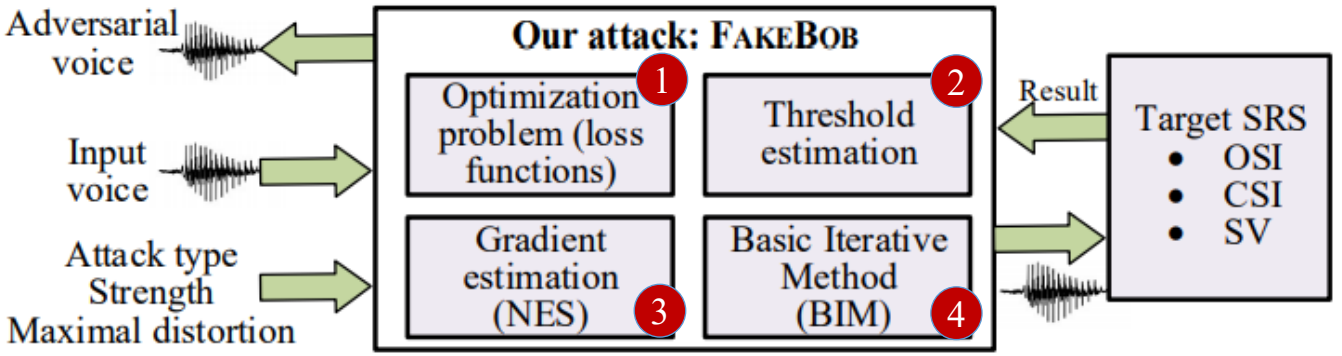
4. Solve the optimization problem by gradient descent

Overview of FAKEBOB



5 Over-the-air attack

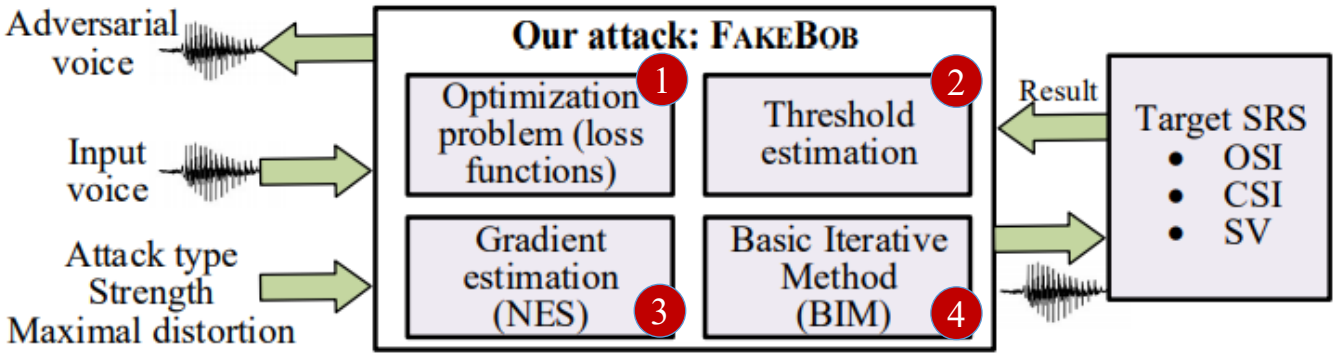
Overview of FAKEBOB



5. Over-the-air attack

noise in air makes attack ineffective

Overview of FAKEBOB

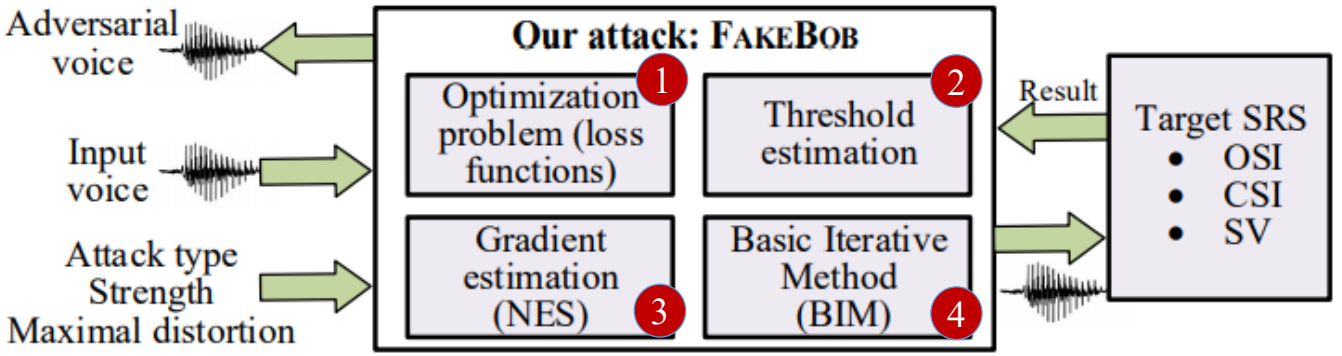


5. Over-the-air attack

noise in air makes attack ineffective

previous work: noise model

Overview of FAKEBOB

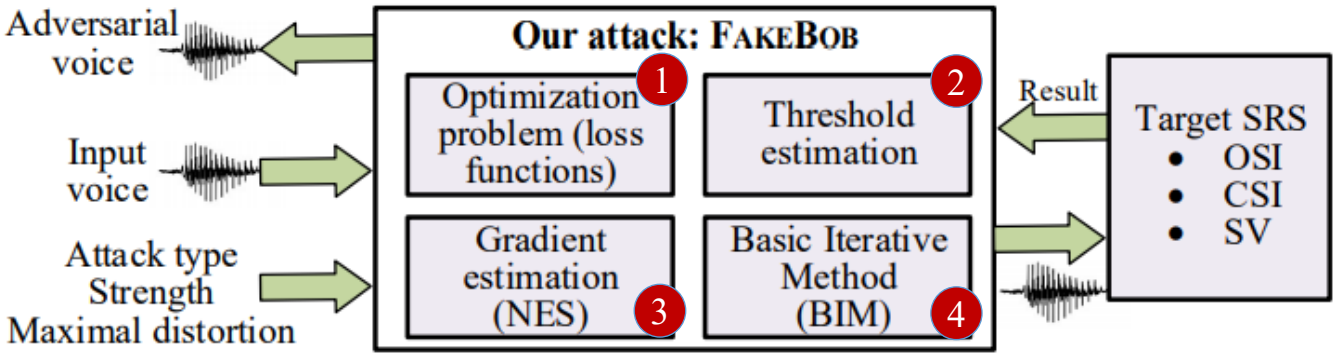


5 Over-the-air attack

noise in air makes attack ineffective

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

Overview of FAKEBOB



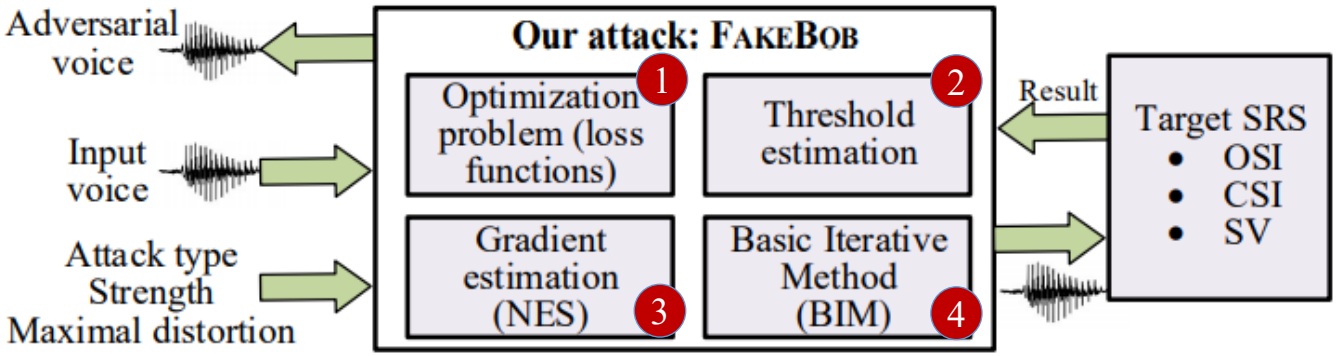
5 Over-the-air attack

noise in air makes attack ineffective

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

ours: improve confidence

Overview of FAKEBOB



5 Over-the-air attack

noise in air makes attack ineffective

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

ours: improve confidence κ

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

Experimental result

Experimental result

- Attack **Open-source** 

Experimental result

- Attack **Open-source** 

  - ≈ 100% attack success rate (ASR)

Experimental result

- Attack **Open-source** 

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

Experimental result

- Attack **Open-source** 

  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - **Talentedsoft:** 100% ASR; 2500 query on average 

Experimental result

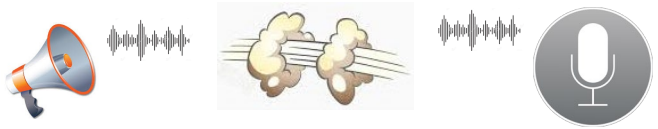
- Attack **Open-source** 

  - ≈ 100% attack success rate (ASR)
- Attack **Commercial**
  - **Talentedsoft:** 100% ASR; 2500 query on average 

  - **Microsoft Azure:** 26% ASR 

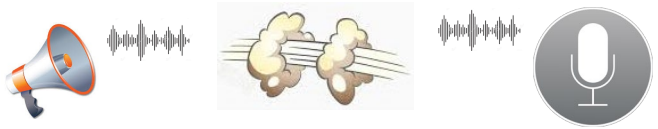
Experimental result

- Over the air Attack



Experimental result

- Over the air Attack

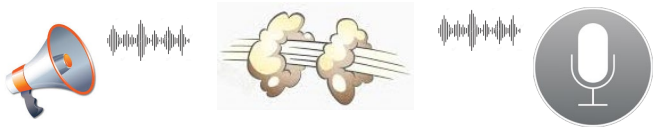


  - different distance between loundspeaker and microphone

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

Experimental result

- Over the air Attack



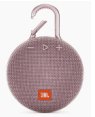
- different distance between loundspeaker and microphone

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

Microphone:

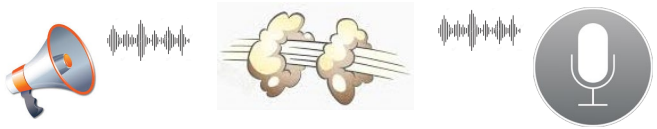
Apple iPhone

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

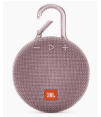
- different acoustic environments
  White / Bus / Restaurant / Music noise
  at least **48%** ASR when noise < 60 dB

Environment independent

- Different devices (at least 70% ASR)

Loundspeaker:

Laptop



JBL portable speaker

Shinco broadcast equipment

Microphone:

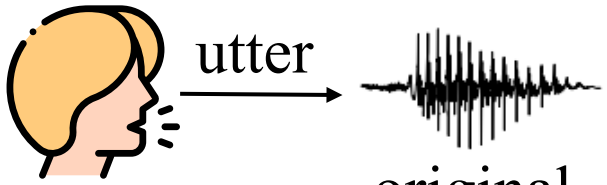
Apple iPhone

OPPO

Device independent

Imperceptibility

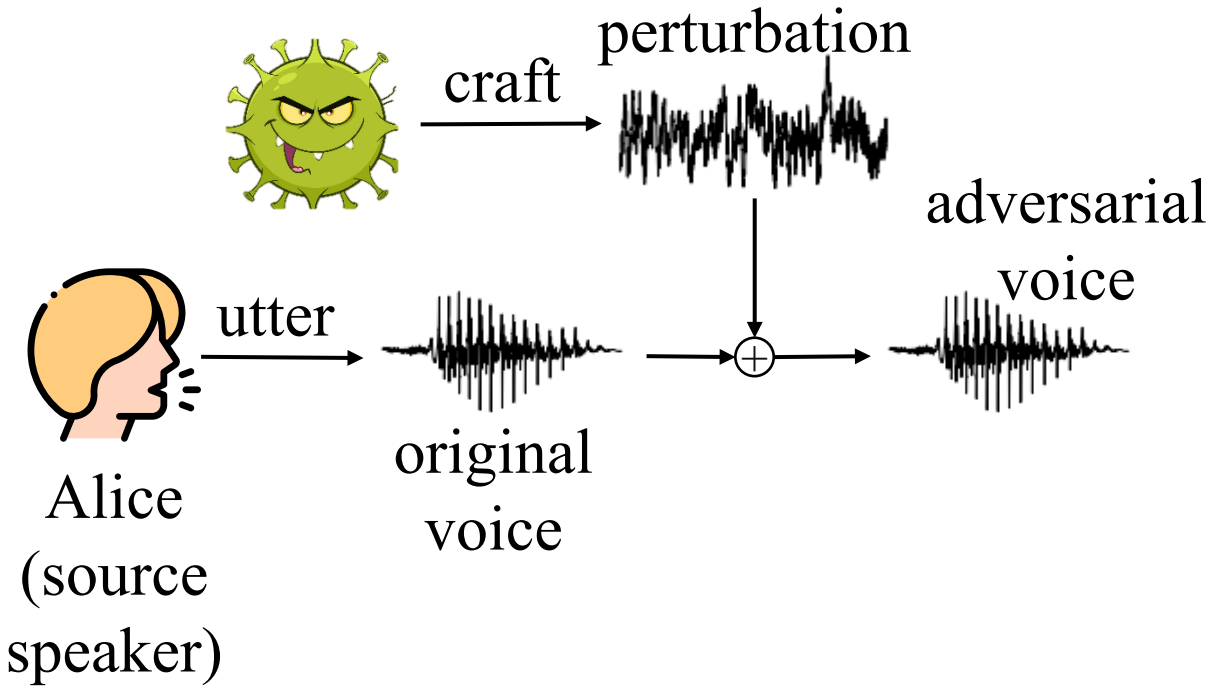
Imperceptibility



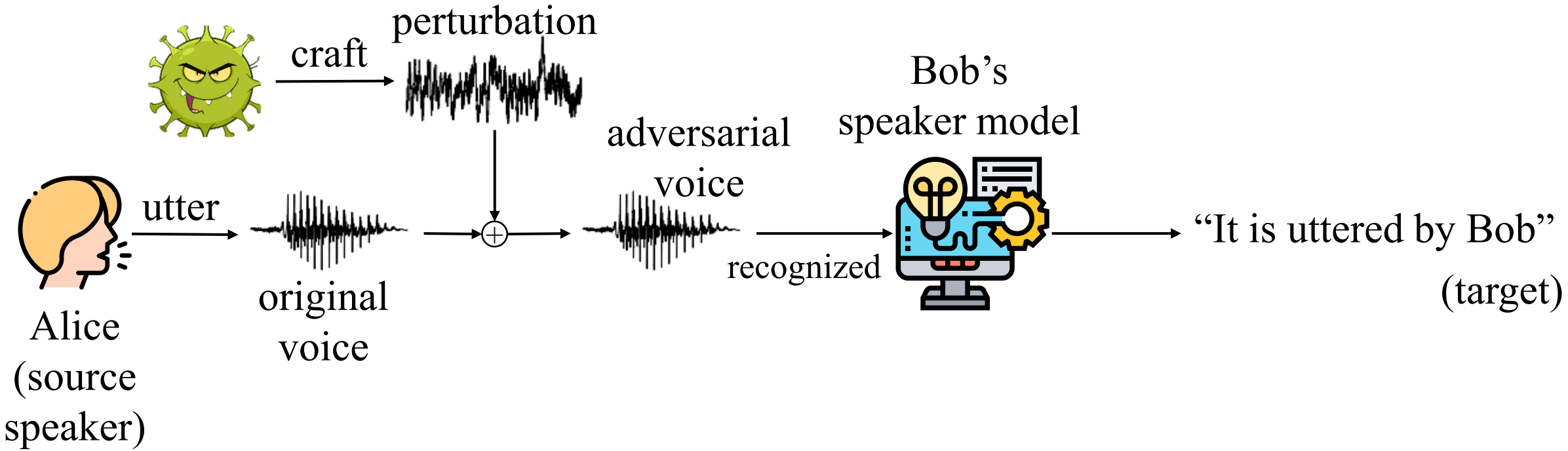
Alice (source speaker)

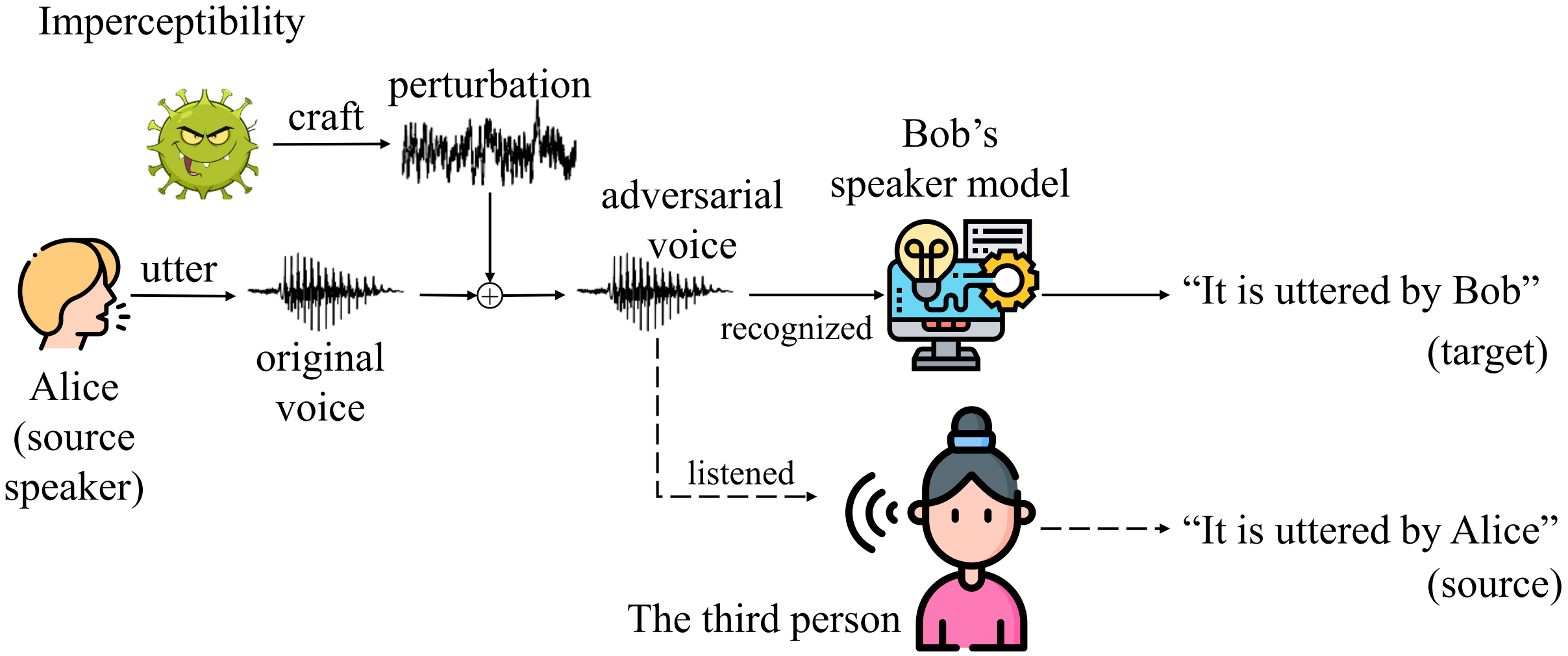
original voice

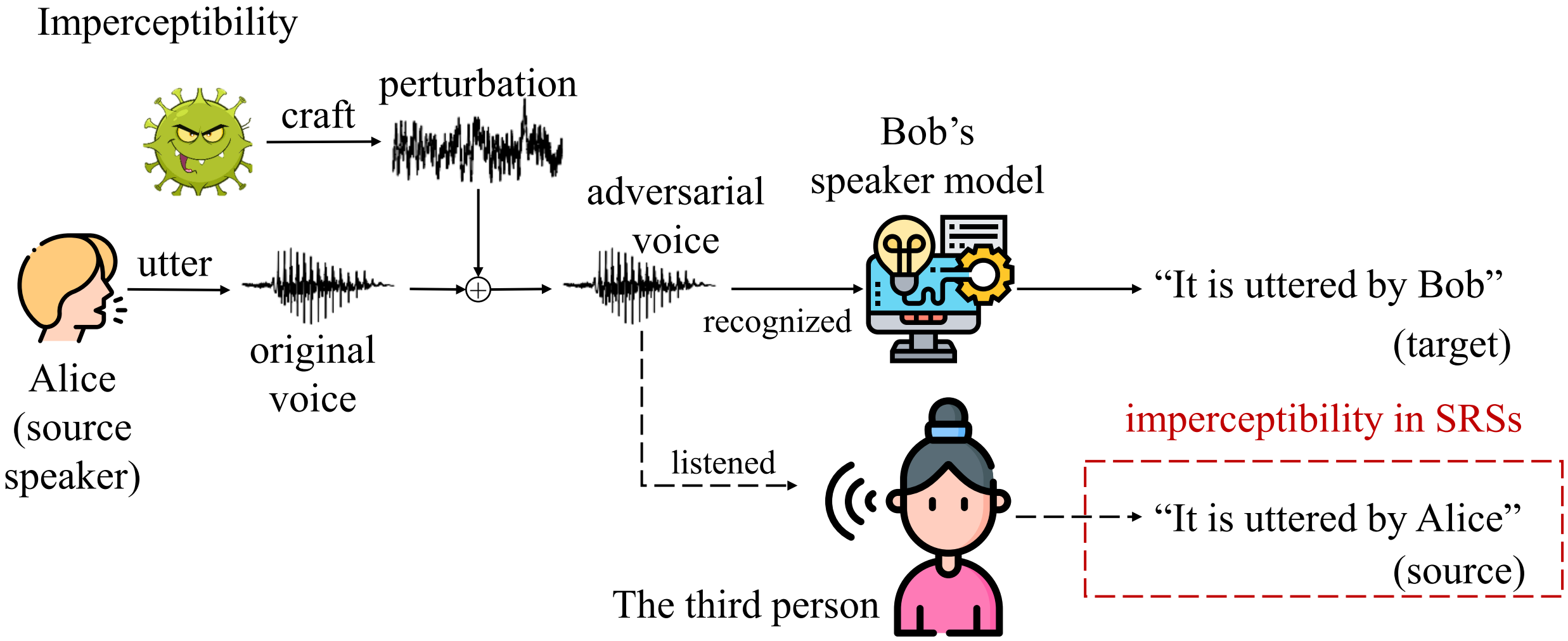
Imperceptibility



Imperceptibility




Imperceptibility
craft
perturbation
Bob's
speaker model
adversarial
voice
utter
original
voice
recognized
"It is uttered by Bob"
(target)
Alice
(source
speaker)
listened
"It is uttered by Alice"
(source)
The third person


Imperceptibility
craft
perturbation
Bob's
speaker model
adversarial
voice
utter
original
voice
Alice
(source
speaker)
recognized
"It is uttered by Bob"
(target)
imperceptibility in SRSs
listened
"It is uttered by Alice"
(source)
The third person

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

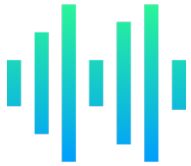
  - API attack: 64.9% same
  - Over-the-air attack: 34.0% same

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

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

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





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

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

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