CommanderSong: A Systematic Approach For Practical Adversarial Voice Recognition

Xuejing Yuan[1,2], Yuxuan Chen[3], Yue Zhao[1,2], Yunhui Long[4], Xiaokang Liu[1,2], Kai Chen[1,2], Shengzhi Zhang[3, 5], Heqing Huang, XiaoFeng Wang[6], and Carl A. Gunter[4]

[1]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Department of Computer Sciences, Florida Institute of Technology, USA
[4]Department of Computer Science, University of Illinois at Urbana-Champaign, USA
[5]Department of Computer Science, Metropolitan College, Boston University, USA
[6]School of Informatics and Computing, Indiana University Bloomington, USA

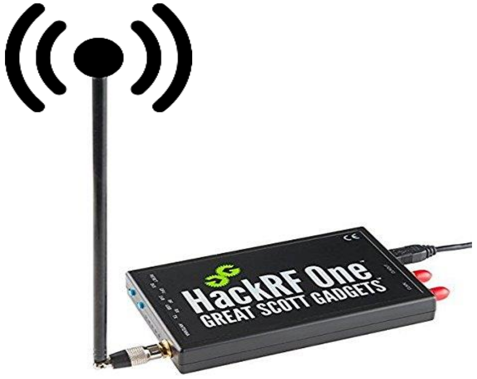
Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion

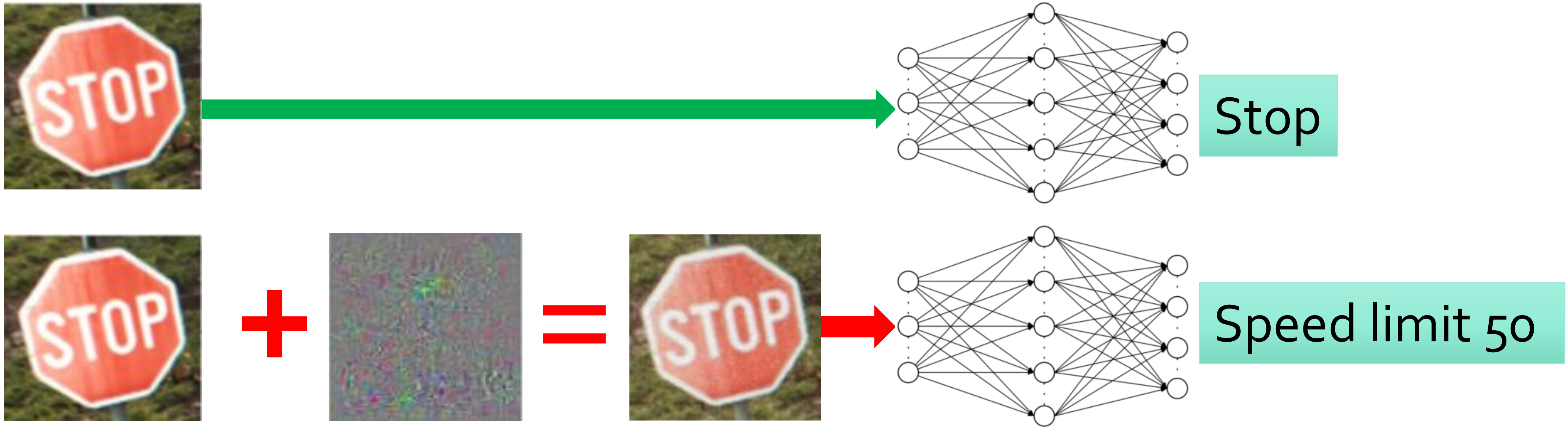
Background

Automatic speech recognition (ASR)

- **Traditional attack**



- **Adversarial sample**



Stop

Speed limit 50

Outline

- Background
- **Motivation**
- Approach
- Evaluation
- Conclusion

Motivation

Hidden voice command attack:
noise-like voice command is abnormal

Dolphin attack:
need a proper transmitter

Recent adversarial audio sample:
is not effective in the physical world

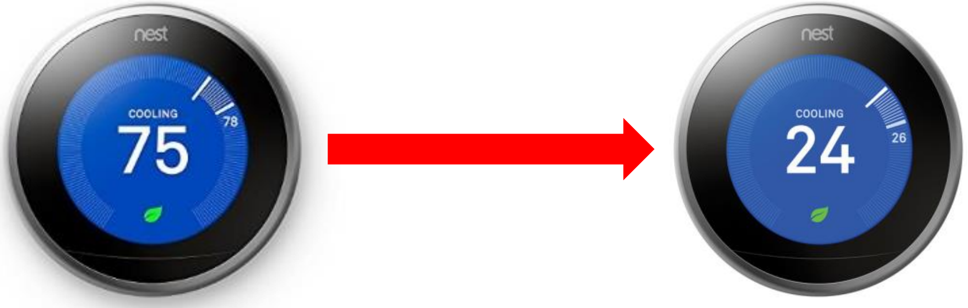
So can we design an approach that is:
using normal sound to make a physical world attack?

- ✓Automatical
- ✓Practical
- ✓Surreptitious
- ✓Spread
- ✓Transferable

CommanderSong Attack

CommanderSong Attack

I think the command in the sound wants me to set the temprature to 24 degree.
----IVC device



Challenges Of The Attack

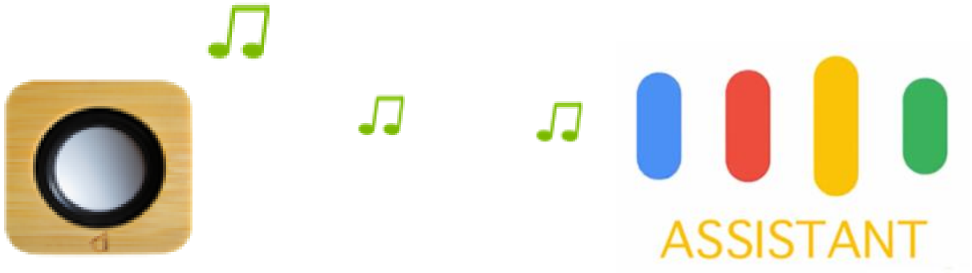
- Human realization
- Influence of the speakers and environment
- Transfer

Outline

- Background
- Motivation
- **Approach**
- Evaluation
- Conclusion

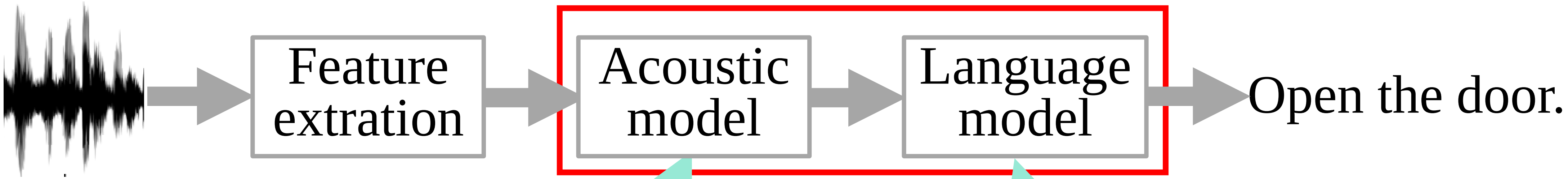
Approach

- step1: WTA (WAV-To-API) attack
- step2: WAA (WAV-Air-API) attack



ASR system: Kaldi (open source platform)

Decoding Principle Of Kaldi



Feature extration

Acoustic model

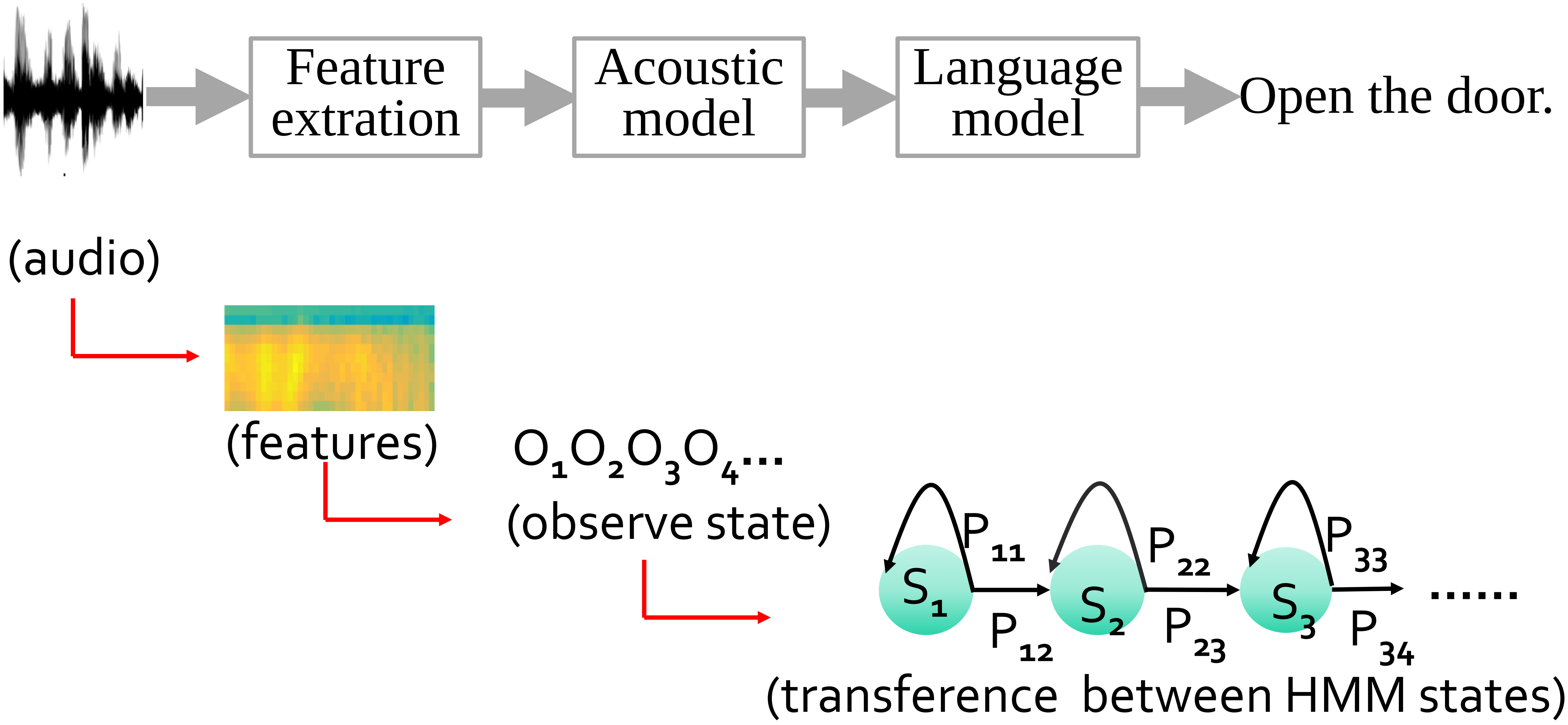
Language model

Open the door.

Deep neural network (DNN): represents the probability between features and phonemes.
phoneme: the smallest unit composing a word.

Weighted Finite State Transducers (WFST): probability distribution over sequence of words.

Decoding Principle Of Kaldi



Feature extration → Acoustic model → Language model → Open the door.

(audio)

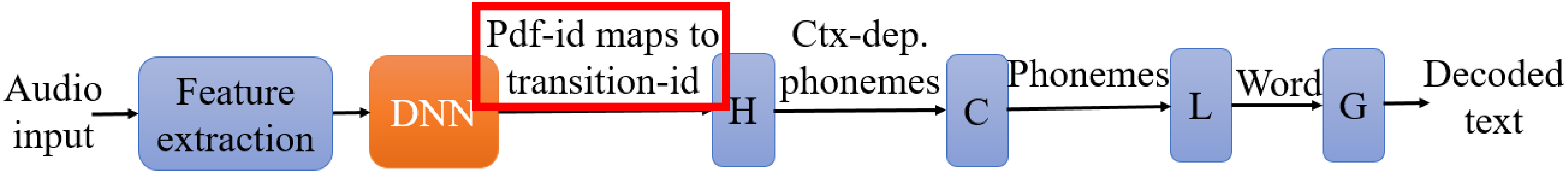
(features)

$O_1O_2O_3O_4$...

(observe state)

(transference between HMM states)

Decoding Principle Of Kaldi



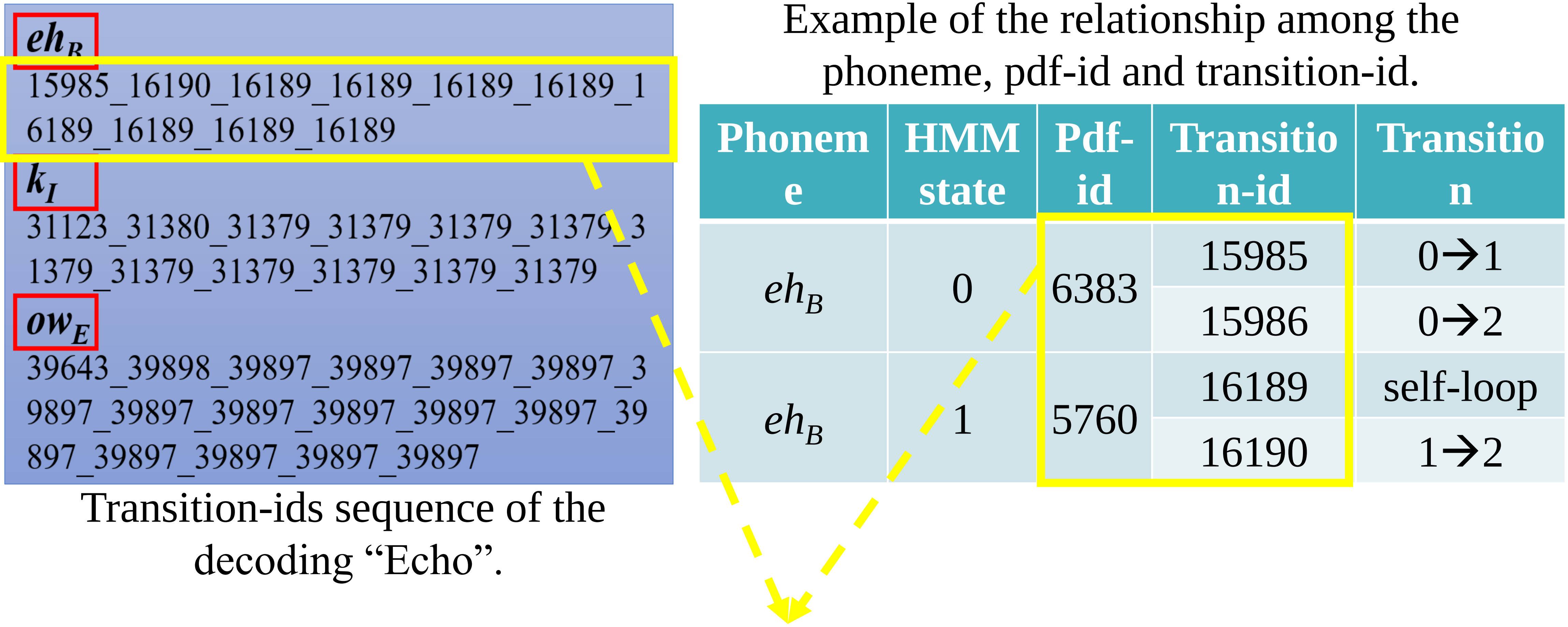
- pdf-id: indicates the probability of every phoneme

  (column number of the DNN output matrix)

- transition-id: uniquely identifies the HMM state transition

  (a sequence of transition-ids can identify a phoneme)

Example Of Kaldi Decoding Results



eh_B

15985_16190_16189_16189_16189_16189_16189_16189_16189_16189

k_I

31123_31380_31379_31379_31379_31379_31379_31379_31379_31379_31379_31379

ow_E

39643_39898_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897_39897

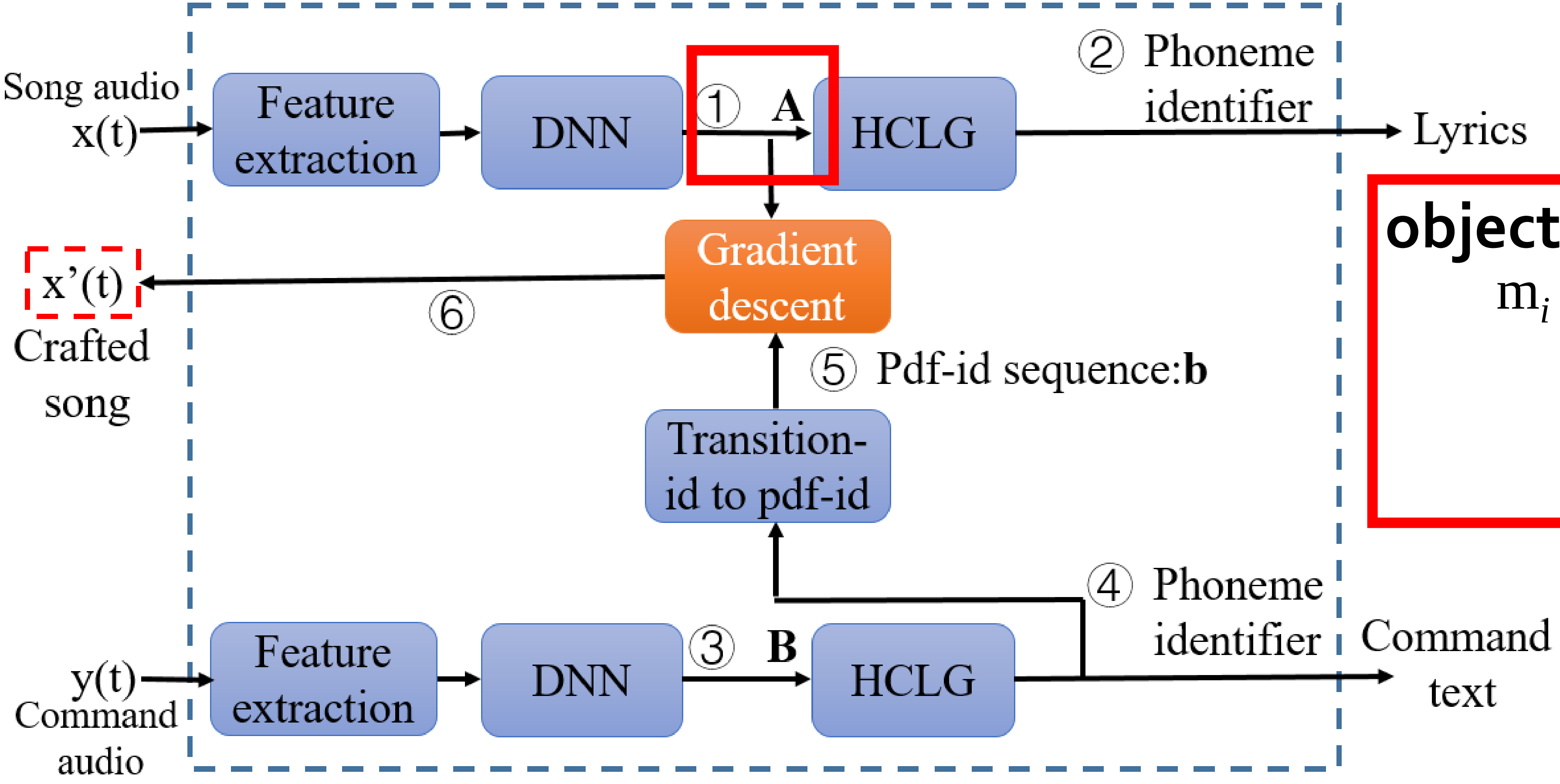
Transition-ids sequence of the decoding “Echo”.

Example of the relationship among the phoneme, pdf-id and transition-id.

Phoneme	HMM state	Pdf-id	Transition-id	Transition
eh_B	0	6383	15985	0→1
			15986	0→2
eh_B	1	5760	16189	self-loop
			16190	1→2

pdf-ids sequence： 6383, 5760,5760, 5760, 5760, 5760, 5760, 5760, 5760, 5760 ……

WTA Attack Approach



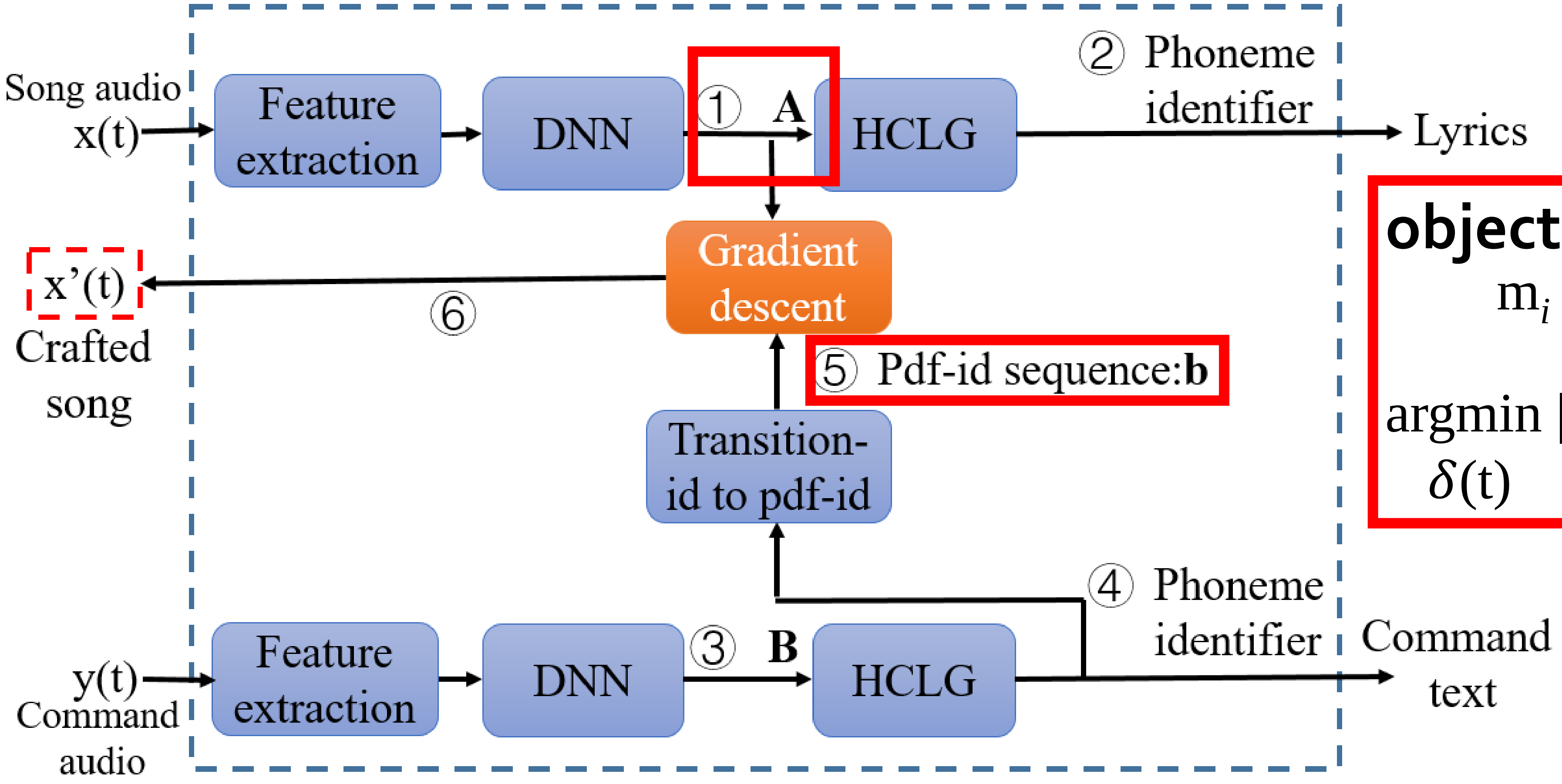
objective function:

$$m_i = \arg\max a_{i,j}\ ,$$

$$g(x(t)) = \mathbf{m}.$$

Pdf-id sequence matching method

WTA Attack Approach



objective function:

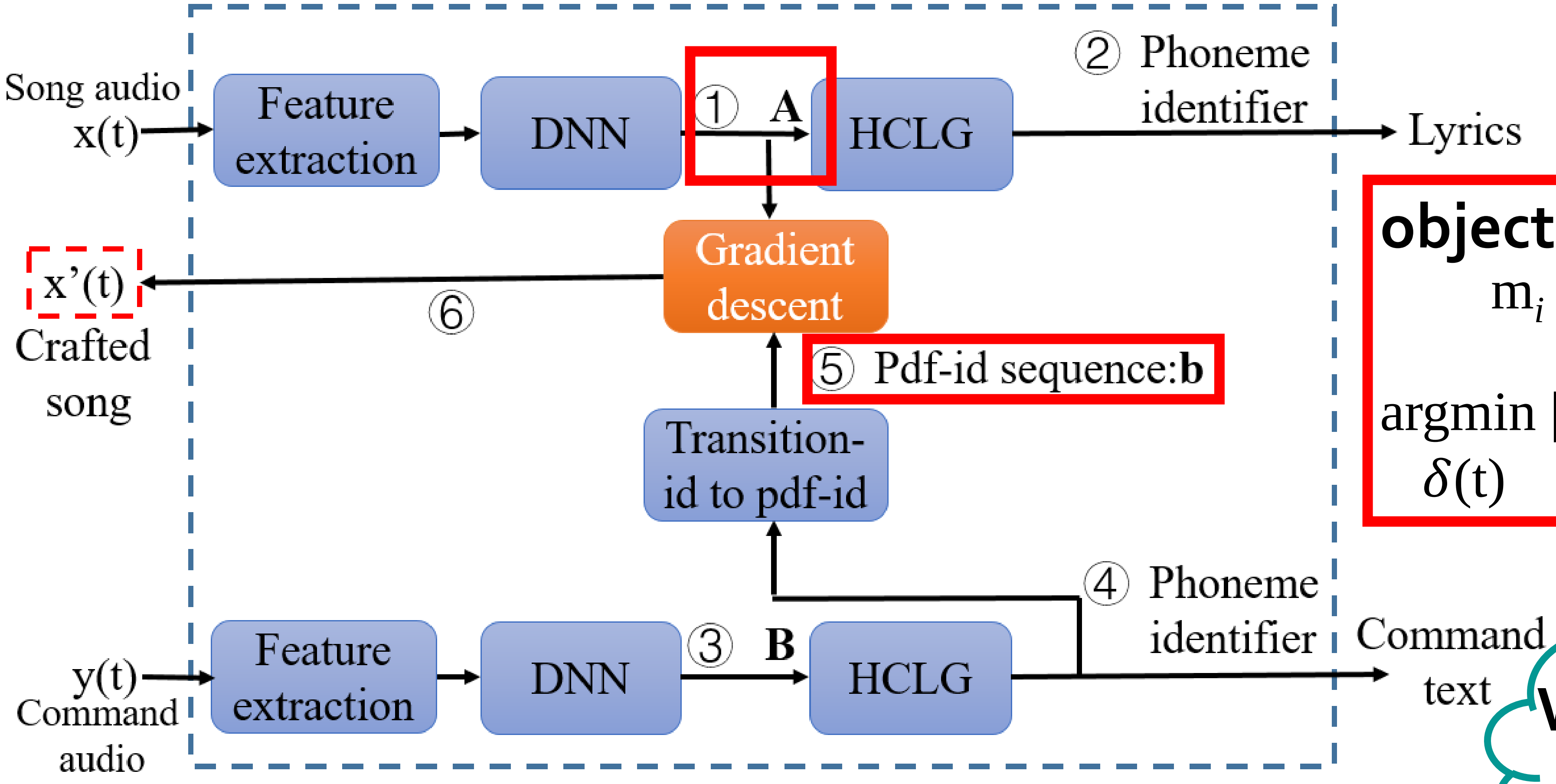
$$m_i = \arg\max a_{i,j} \ ,$$

$$g(x(t))=\mathbf{m}.$$

$$\underset{\delta(t)}{\arg\min} \ \|g(x(t)+\delta(t))-\mathbf{b}\|_1.$$

Pdf-id sequence matching method

WTA Attack Approach



objective function:

$$m_i = \arg\max a_{i,j}\ ,$$

$$g(x(t))=\mathbf{m}.$$

$$\underset{\delta(t)}{\operatorname{argmin}}\ \|g(x(t)+\delta(t))-\mathbf{b}\|_1.$$

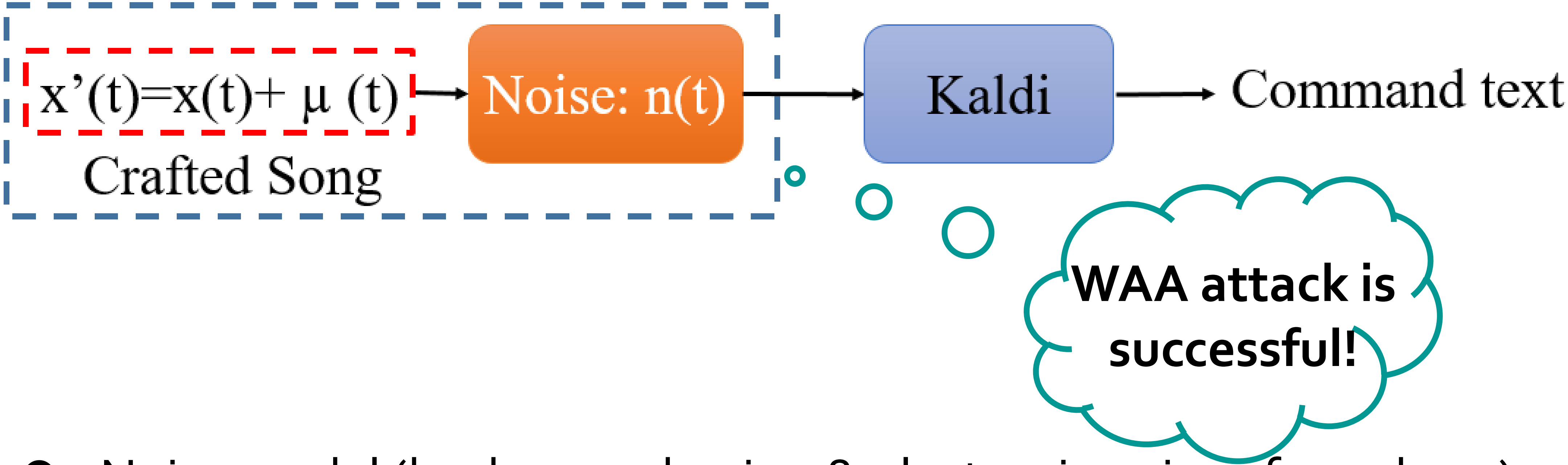
WTA attack is successful!

Pdf-id sequence matching method

WTA Attack samples for the real world attack?

❌

WAA Attack Approach



- Noise model (background noise & electronic noise of speakers)
  (needs to accsess to the speaker and receiver)
- random noise model
  (easily generate and universally applicable)

Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion

Evaluation

WTA attack results

Command	Success rate (%)
Okay google restart phone now.	100
Okay google flashlight on.	100
Okay google read mail.	100
Okay google clear notification.	100
Okay google airplane mode on.	100
Okay google turn on wireless hot spot.	100
Okay google read last sms from boss.	100
Echo open the front door.	100
Echo turn off the light.	100

Evaluation

WAA attack results

Command	Speaker	Success rate (%)
Echo ask capital one to make a credit card payment.	JBL speaker	90
	ASUS Laptop	82
	SENMATE Broadcast	72
Okay google call one one zero one one nine one two zero.	JBL speaker	96
	ASUS Laptop	60
	SENMATE Broadcast	70

Evaluation

Human comprehension (a survey on Amazon Mechanical Turk)

- Have you ever heard this original song before?
- Do you think the song is abnormal?
- Where do you think the noise in the abnormal song comes from?
- How many times have you listened before you can recognize the words.

Evaluation

Human comprehension of the WTA attack samples

Music classification	Listened (%)	Abnormal (%)	Recognize Command (%)
Soft music	13	15	0
Rock	33	28	0
Popular	32	26	0
Rap	41	23	0

Evaluation

Human comprehension of the WAA attack samples

Song name	Listened (%)	Abnormal (%)	Noise-speaker (%)	Noise-song (%)
Did You Need It	15	67	42	1
Outlaw of Love	11	63	36	2
The Saltwater Room	27	67	39	3
Sleepwalker	13	67	41	0
Under neath	13	68	45	3
Feeling Good	38	59	36	4
Average	19.5	65.2	40	2.2

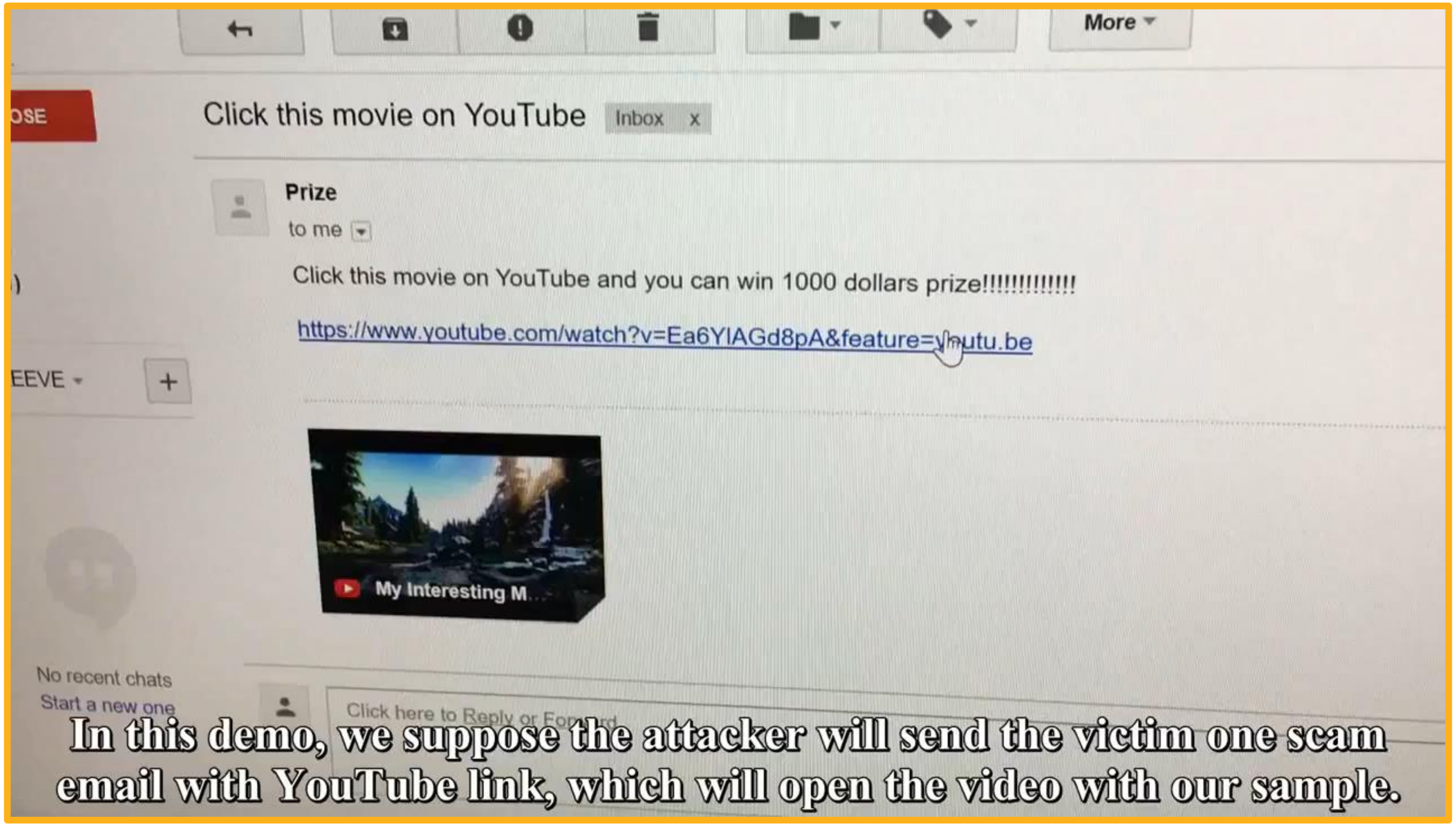
Evaluation

Transferability from Kaldi to iFLYTEK

Command	iFLYREC (%)	iFLYTEK Input (%)
Airplane mode on.	66	0
Open the door.	100	100
Good night.	100	100

Evaluation

Spread and attack iFlytek



Click this movie on YouTube

Prize

to me

Click this movie on YouTube and you can win 1000 dollars prize!!!!!!!!!!!!!

https://www.youtube.com/watch?v=Ea6YIAGd8pA&feature=youtu.be

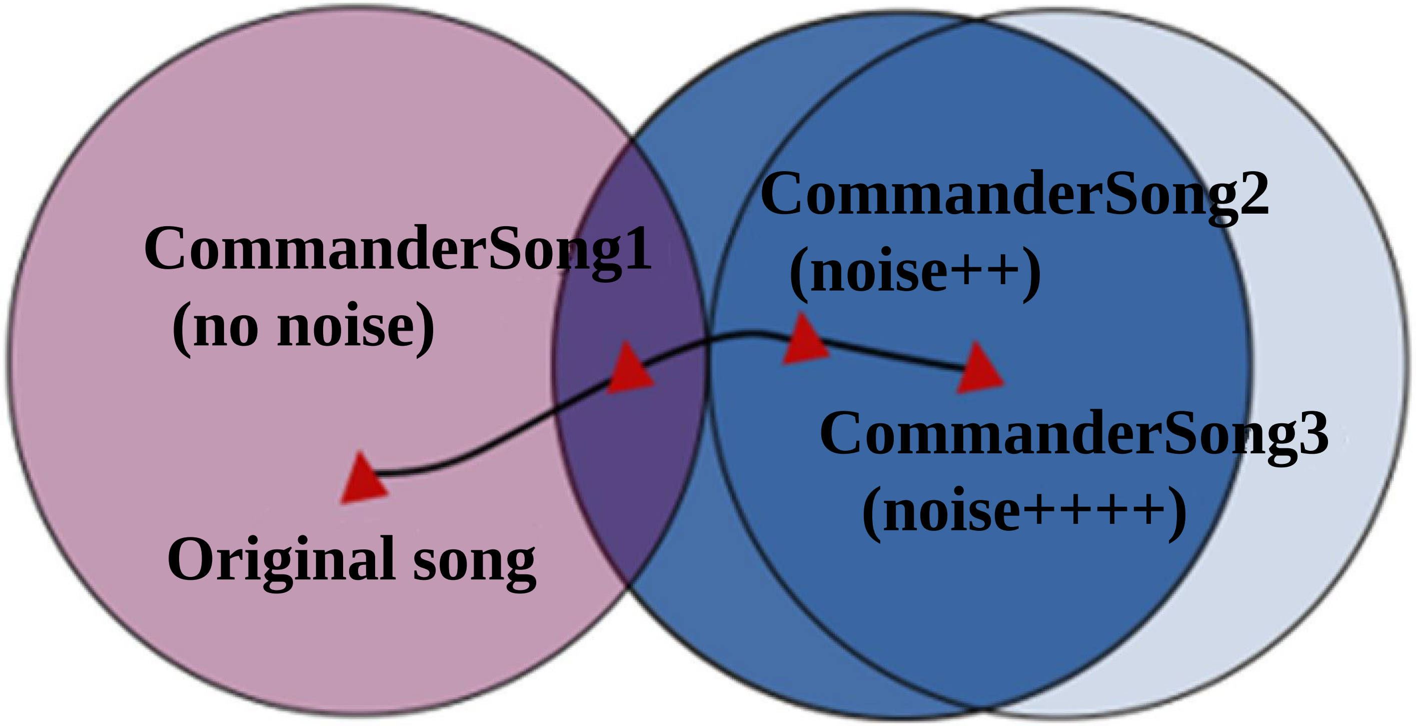
In this demo, we suppose the attacker will send the victim one scam email with YouTube link, which will open the video with our sample.

Understanding Of The Attacks

● Kaldi recognize as command

● Human recognize as command

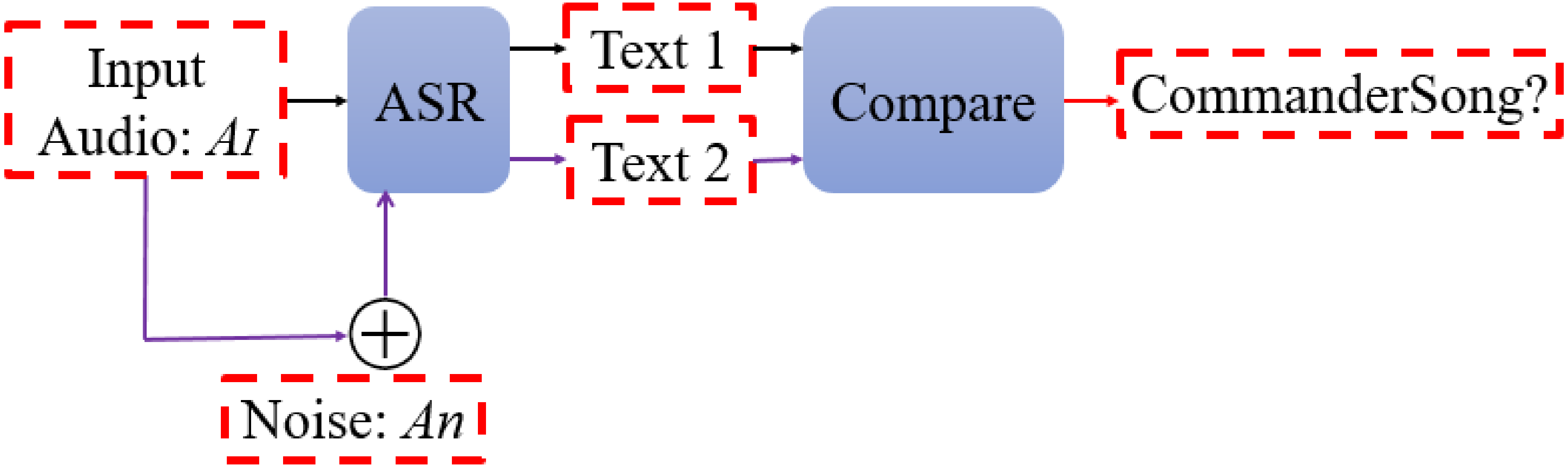
● Human recognize as song



Explaination of Kaldi and human recognize of the audios.

Defense

- Audio turbulence defense



- Audio squeezing defense

Outline

- Background
- Motivation
- Approach
- Evaluation
- Conclusion

Conclusion

- Practical adversarial attack automatic speech recognition
- Can be transferred to iFlytek
- Can be spread through the Internet and radio
- Surreptitious to human