CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition

-Yuan et. al

Presented by Deeksha Dixit and Naman Awasthi

Background

Automatic

Speech

Recognition

Intelligent **V**oice **C**ontrol

Hi. I'm Cortana.

Ask me a question!

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

Slide credits: Xuejing Yuan et al.

Effective attack requirements

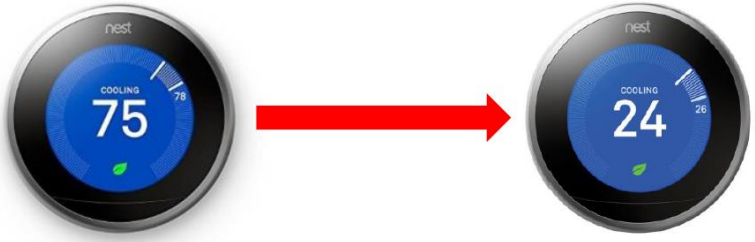
- Effective in real world noisy environment
  - Background sound
  - Electronic noise from speaker
- Stealthy
  - White noise
  - Unnoticeable to ordinary user
- Remotely deliverable at scale
  - Youtube, spotify, TV, etc.
  - Vibrations delivered via other devices

Commander Song

Slide credits: Xuejing Yuan et al.

Commander Song

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



Slide credits: Xuejing Yuan et al.

Approach

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

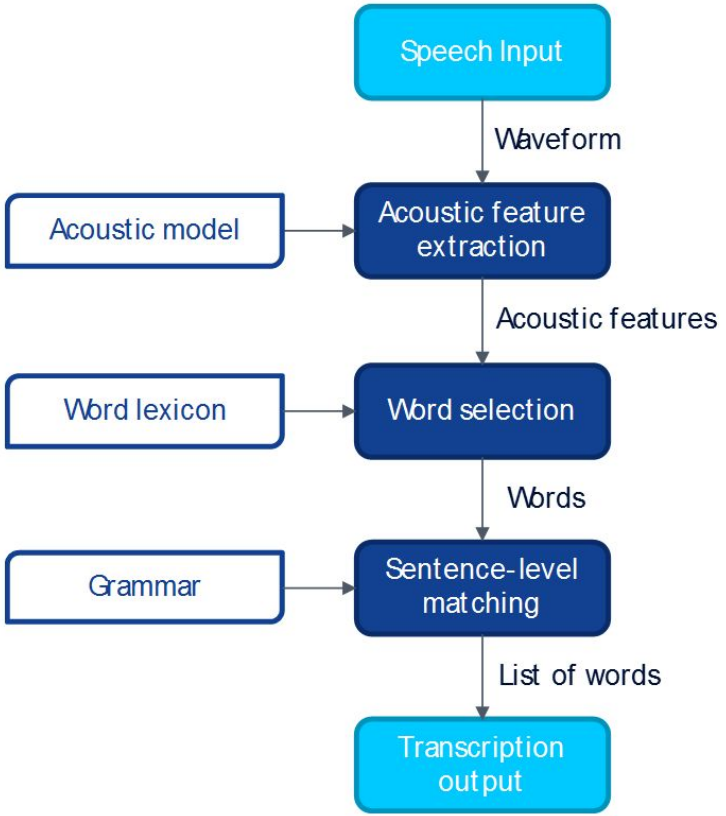
ASR system: Kaldi (open source platform)

Components of this attack: Kaldi

Kaldi is a speech recognition toolkit written in C++

They use the Aspire Chain module from Kaldi for ASR

Aspire's generic block diagram->

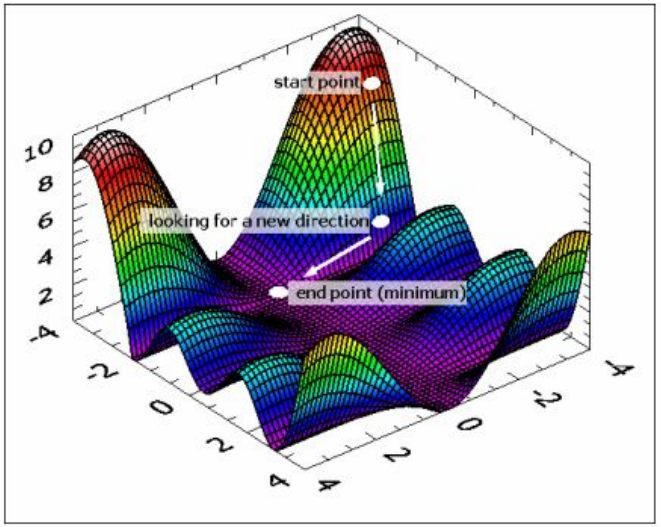


61 eh_B
15985_16190_16189_16189_16189_16189_
16189_16189_16189_16189
99 k_I
31123_31380_31379_31379_31379_31379_
31379_31379_31379_31379_31379_31379
118 ow_E
39643_39898_39897_39897_39897_39897_
39897_39897_39897_39897_39897_39897_
39897_39897_39897_39897_39897

Aspire Documentation

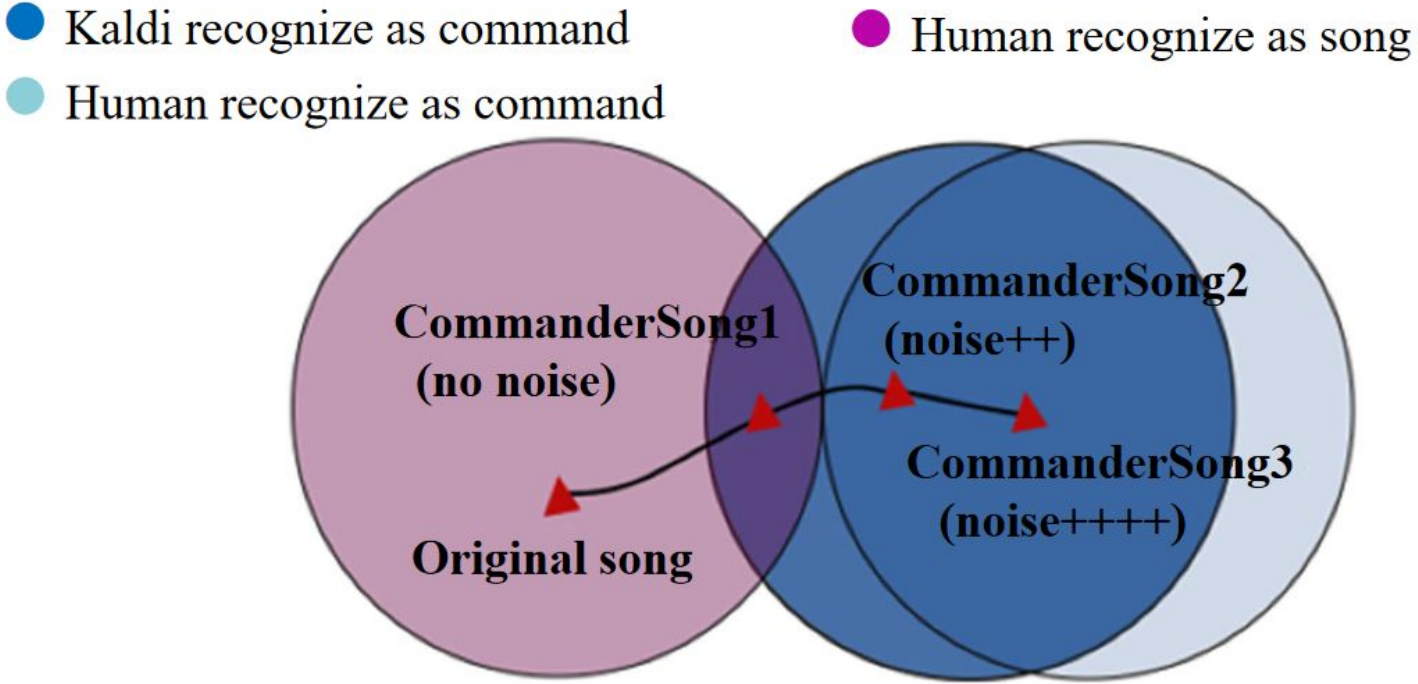
Components of this attack: Gradient Descent

- Idea
  - Start somewhere
  - Take steps based on the gradient vector of the current position till convergence
- Convergence :
  - happens when change between two steps < ε

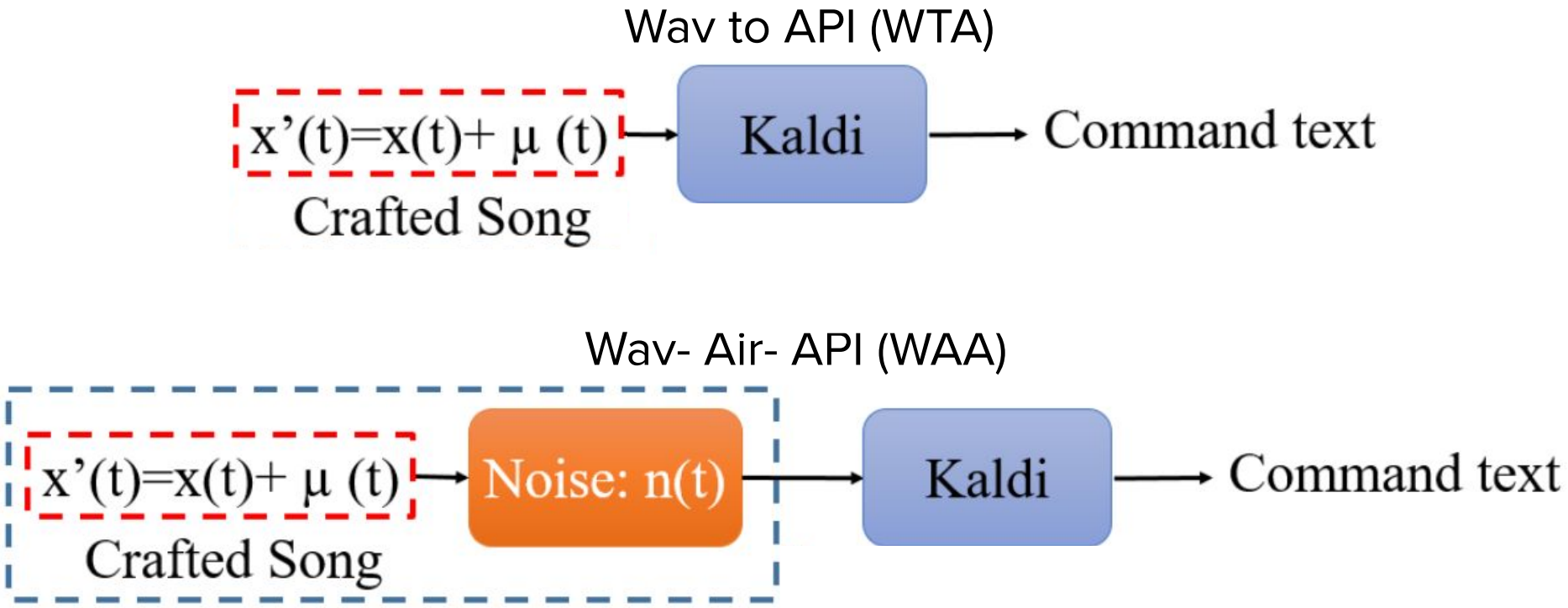


Slide credits: Michail Michailidis & Patrick Maiden

The Three Spaces

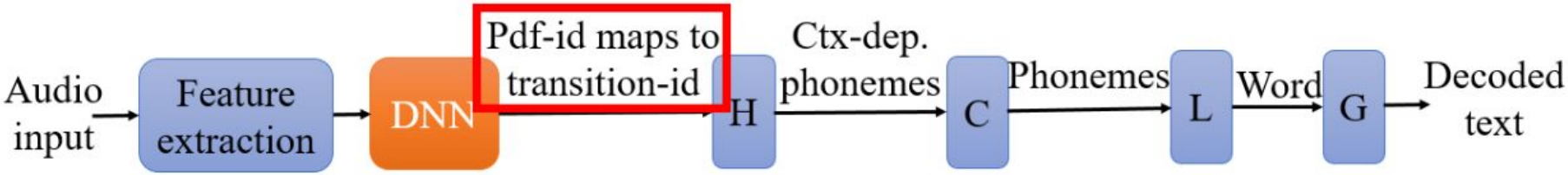


Commander Song: Flow Chart



Slide credits: Xuejing Yuan et al.

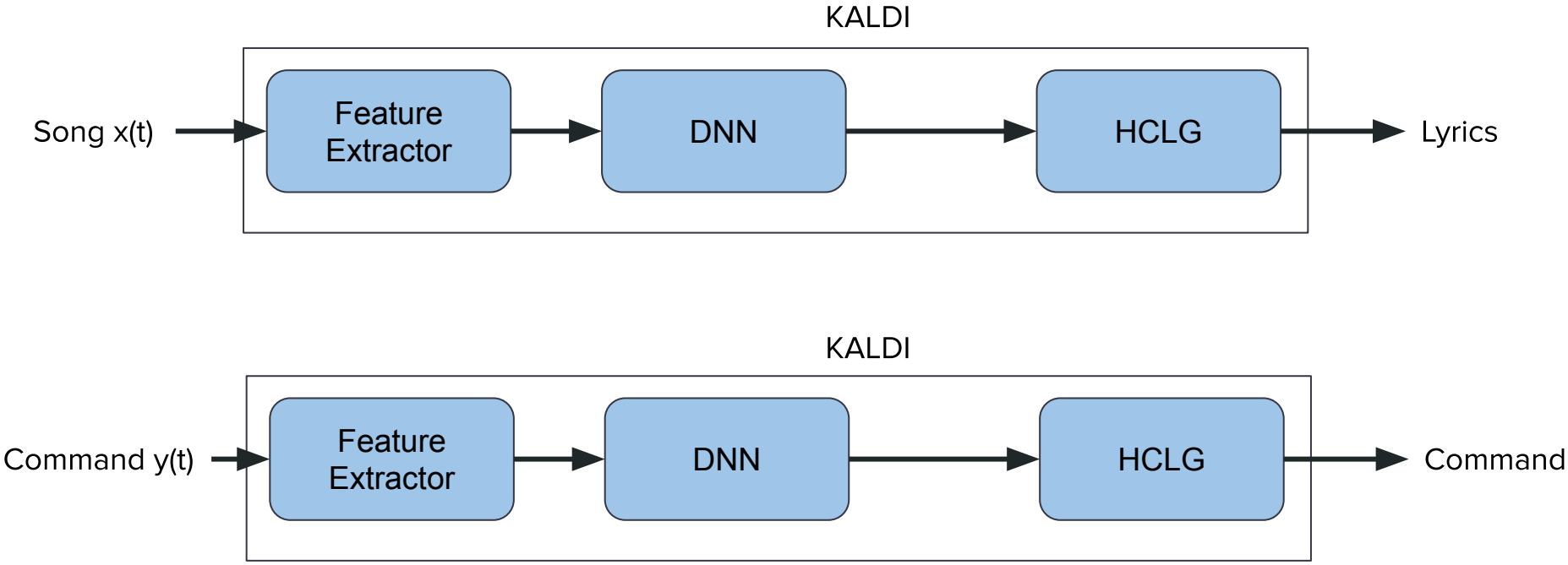
Kaldi's inner working



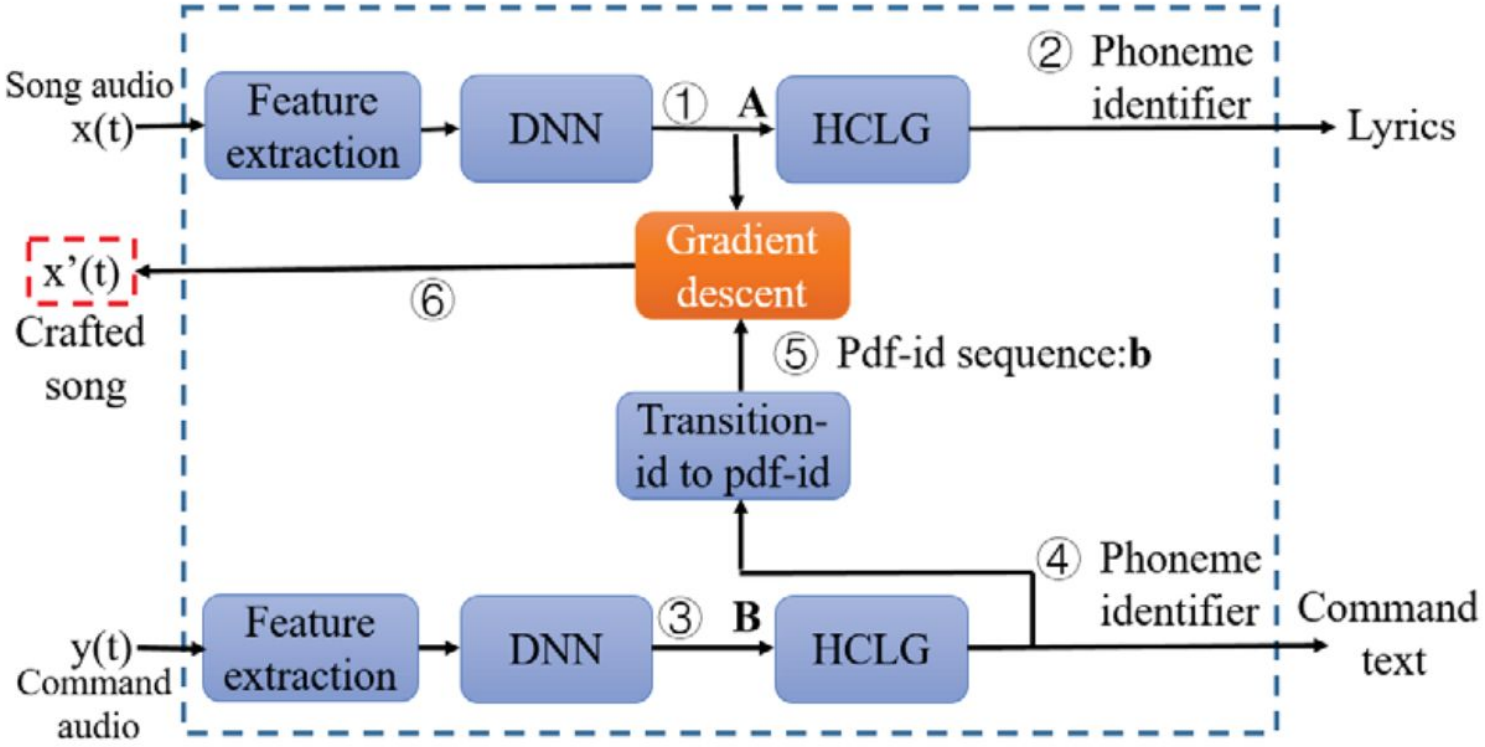
	transducer	input sequence	output sequence
G	word-level grammar	words	words
L	pronunciation lexicon	phones	words
C	context-dependency	CD phones	phones
H	HMM	HMM states	CD phones

Slide credits: Xuejing Yuan et al. and Aspire Documentation

The two signals

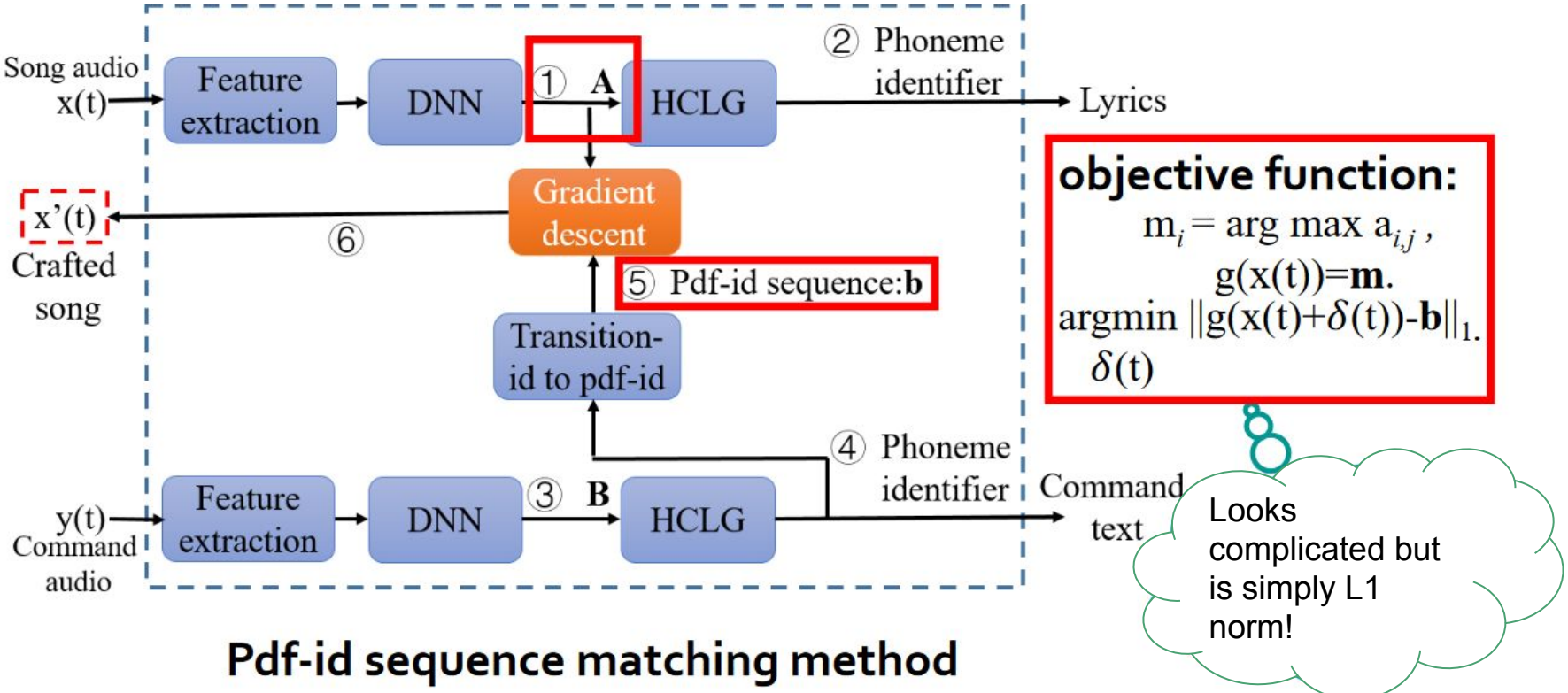


Steps of attack!



Slide credits: Xuejing Yuan et al.

Combining the two



Slide credits: Xuejing Yuan et al.

Explanation of the objective

Let the output of DNN be A (with elements $a_{i,j}$)

A is n X k (n: number of frames, k different phonemes)

m is the sequence of phoneme with highest probability

Thus, we try to introduce a δ(t) which brings the original signal x(t) close to the command's phonemes.

objective function:

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

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

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

Evaluation: Audio directly fed to Kaldi's ASR

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: Audio directly fed to Kaldi's ASR

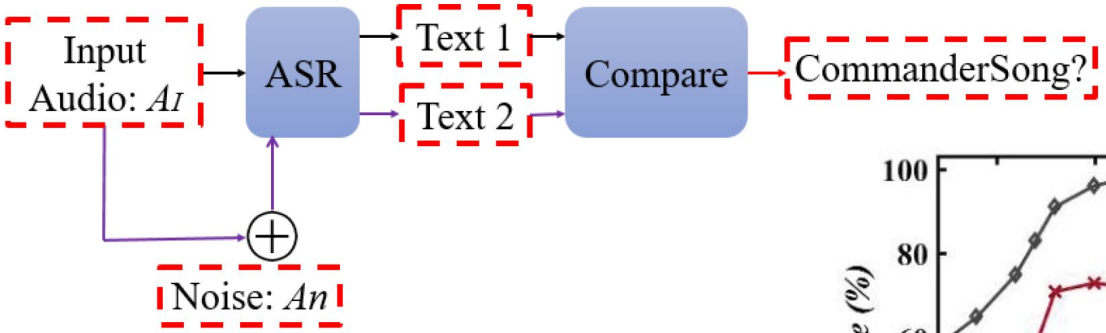
Command	Speaker	Success rate (%)
Echo ask capital one to make a credit card payment.	JBL speaker	90
	ASUS Laptop	82
	SENMATE Broadcast	72
Okay google call one one zero one one nine one two zero.	JBL speaker	96
	ASUS Laptop	60
	SENMATE Broadcast	70

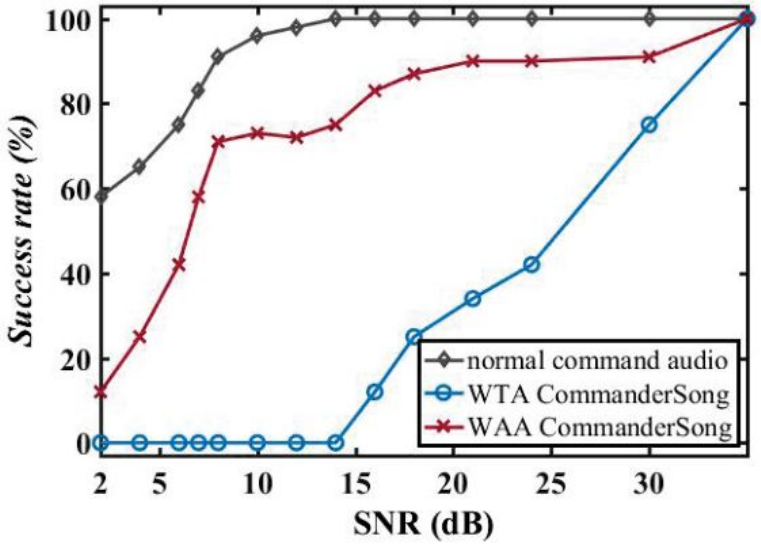
Evaluation: Fooling Humans!

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

Defense

- Audio turbulence defense

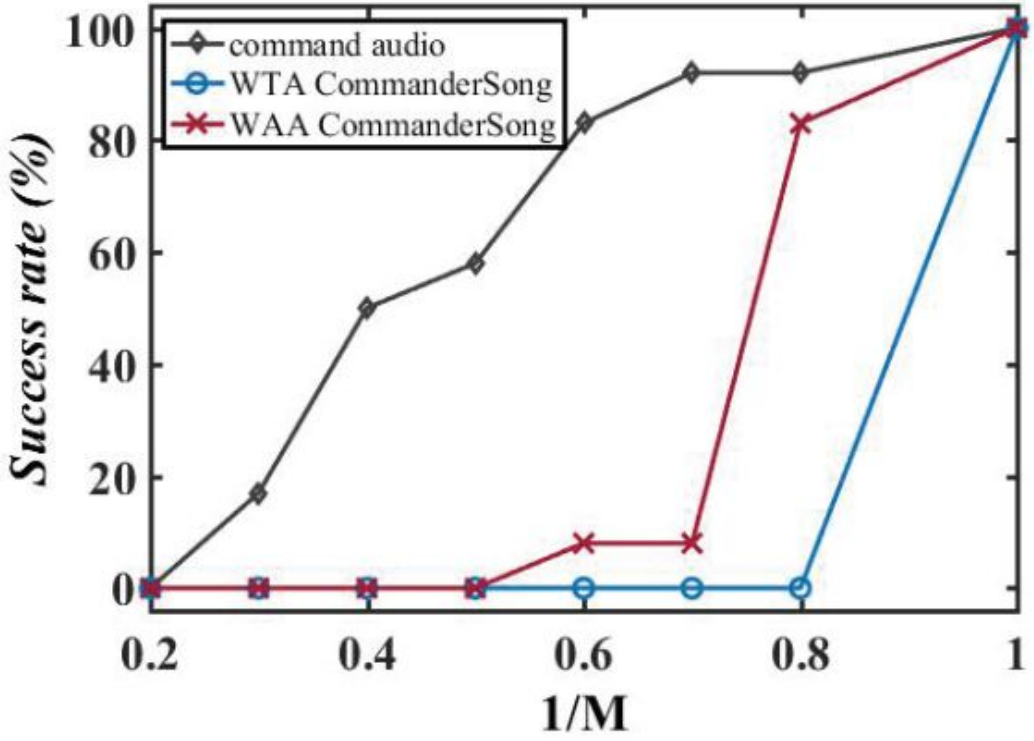




Defense

Audio squeezing

Change Sampling Rate by a factor M



Conclusion

- **Gradient descent based attack** on DNN based ASR systems.
- CommanderSong makes ASR systems **execute the command** while being played **over the air**
- CommanderSong can be **transferred to iFLYTEK**, impacting popular apps such as WeChat, Sina Weibo, and JD with **billions of users**
- CommanderSong can be **spread through YouTube and radio**
- **Audio turbulence and audio squeezing** can be used to **defend** against CommanderSong attacks

Insights

- **First attack** proposed for DNN based ASR
- Bringing **songs** into the mix made the attacks more practical for a wide scale distribution
- One of the defences proposed ”Air turbulence” works by introducing noise to the signal. This just shows that WTA models would not work well in WAA pipeline. So **WTA models are a subset of WAA** models (where environment noise = 0)
- We are working on Adversarial attacks for ASR for our project and CommanderSong has been the **motivation paper for many** of them!

Thank you!

Let us know if there are any questions about the paper