Practical adversarial attack against speech recognition platforms

Shengzhi Zhang
Department of Computer Science
Metropolitan College
Boston University

BOSTON UNIVERSITY

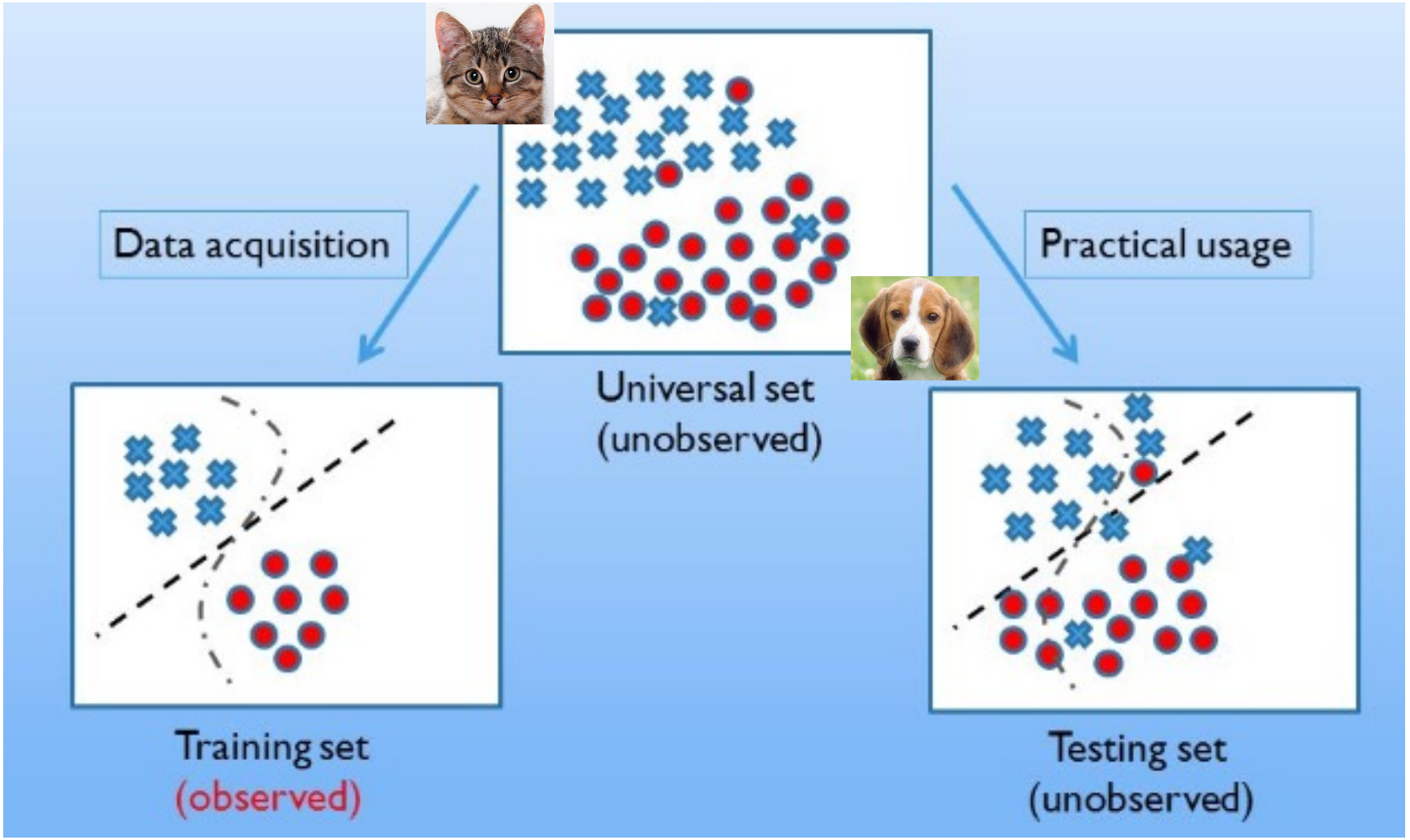
About me

- Shengzhi Zhang
  - Ph.D., Penn State University, 2012
  - Bachelor, Tongji University, 2006
- Assistant professor, CS, Boston University MET College (2018.07 –)
- Experience
  - Assistant professor, CS, Florida Tech (2014.01 – 2018.07)
  - IBM research lab, Honeywell aerospace lab, Cisco R&D
- Cybersecurity research:
  - Machine learning security
  - Vehicle security
  - Operating system security
  - Zeroization verification
  - Cloud computing
  - …

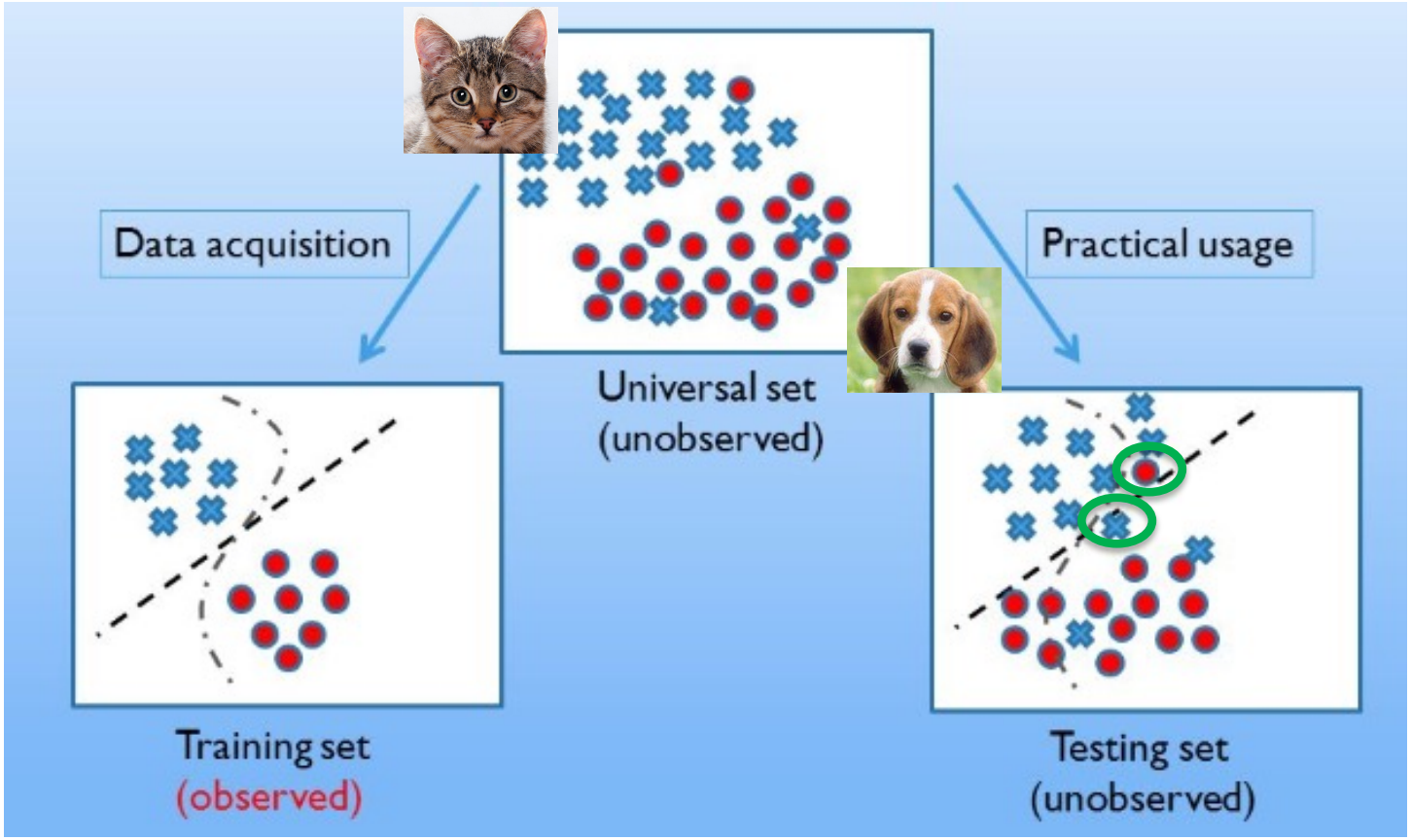
Outline

- **What is adversarial attack?**
- Our research
- Conclusion

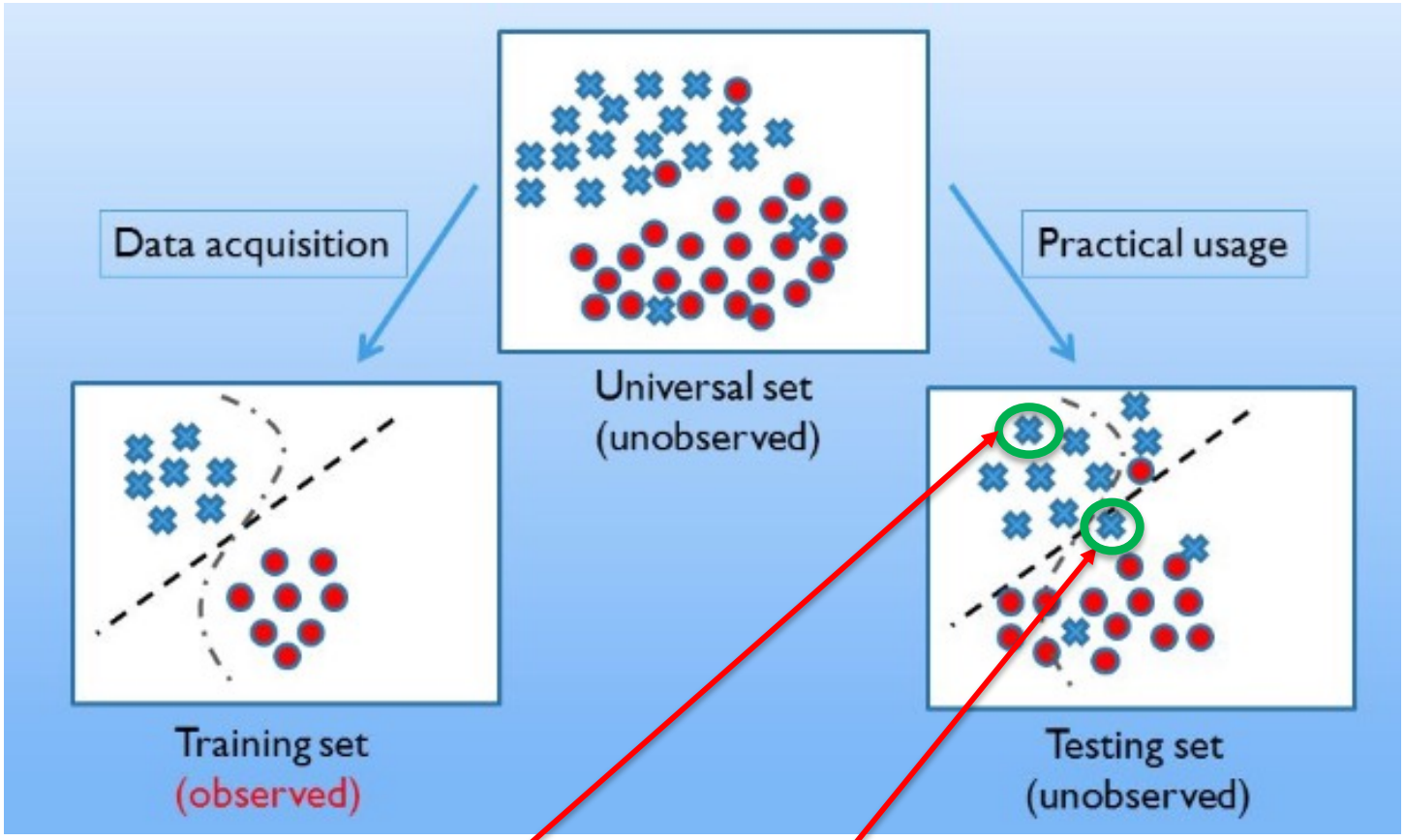
What is machine learning?



Adversarial Attack



Adversarial Attack



Confusion for self-driving vehicles

Original Inputs

Modified Inputs

Wrong ML Detection

Existing Research

- **Image (lots of works)**
  - White box, black box
  - Digital, physical
- **Video (leverage findings on image)**
  - White box, limited black box and blind
  - Digital, limited physical

- **Audio (few works)**
  - Mostly white box
  - Mostly digital
  -

Existing Research

- **Image (lots of works)**
  - White box, black box
  - Digital, physical
- **Video (leverage findings on image)**
  - White box, limited black box and blind
  - Digital, limited physical
  - Our work: both the black box and physical (ACM CCS 2019)
- **Audio (few works)**
  - Mostly white box
  - Mostly digital
  - Our work: both black box and physical (USENIX Security 2018, 2020)

Outline

- What is adversarial attack?
- **Our research**
- Conclusion

Speech Recognition

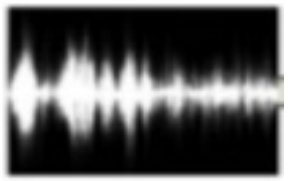
Echo, unlock the front door

Echo, unlock the front door

Adversarial attacks against speech recognition?

Echo, unlock the front door

Soft music with perturbation added



Hmm, quality of the soft music is not so good.

Step 1: White Box Attack

- Methodology:
  - Impact many users in an automated fashion
    - Revise song/music
  - Generate impossible or difficult to be noticed adversarial samples
    - Revise enough
    - Revise little
  - Attack in the physical world (practical attack)
    - Modeling random noise to accommodate background noise, electronic noise from speakers

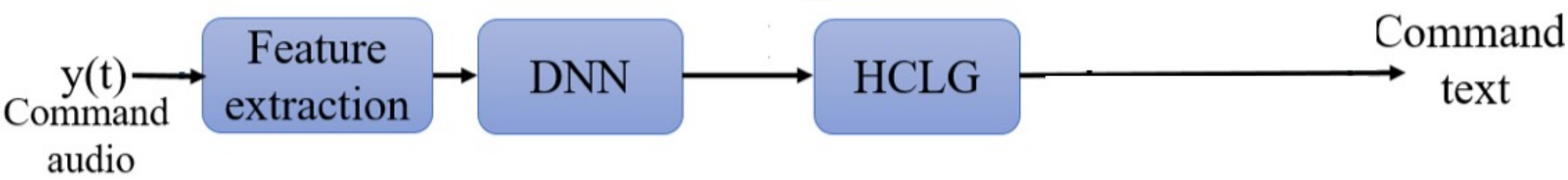
CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition , USENIX Security '18

Step 1: White Box Attack

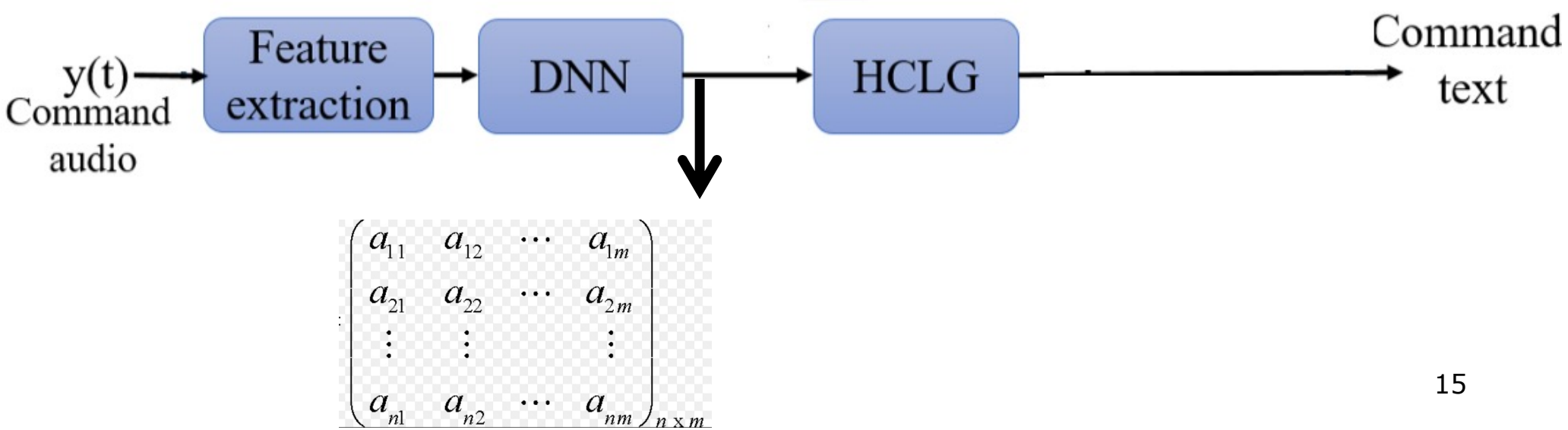
- Methodology:
  - Impact many users in an automated fashion
    - Revise song/music
  - Generate impossible or difficult to be noticed adversarial samples
    - Revise enough
    - Revise little
  - Attack in the physical world (practical attack)
    - Modeling random noise to accommodate background noise, electronic noise from speakers

okay google call one one zero one one nine one two zero

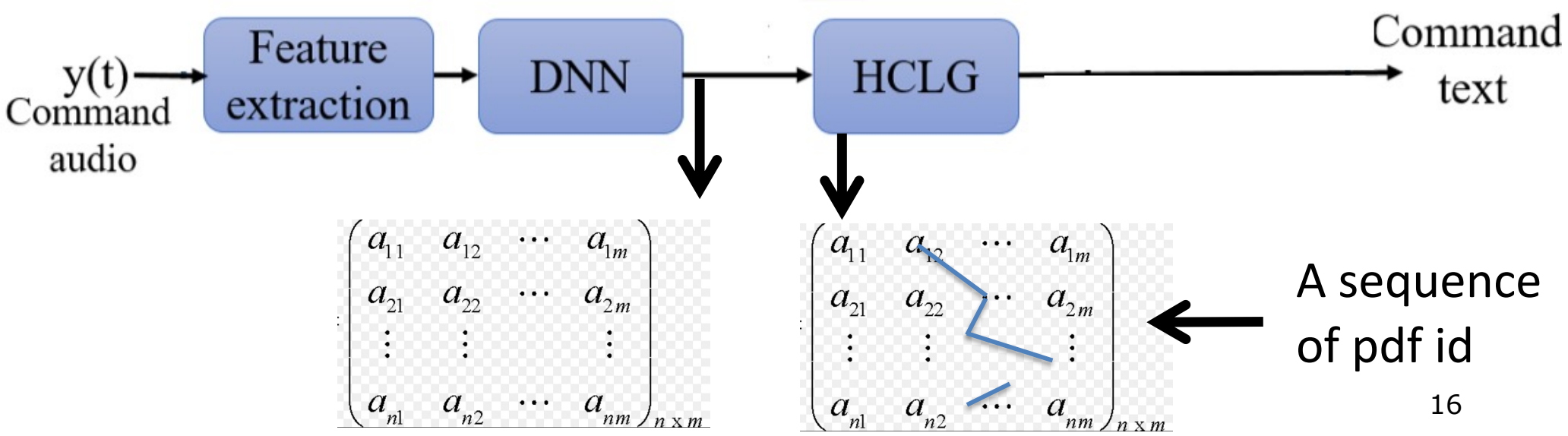
Typical Speech Recognition: Kaldi



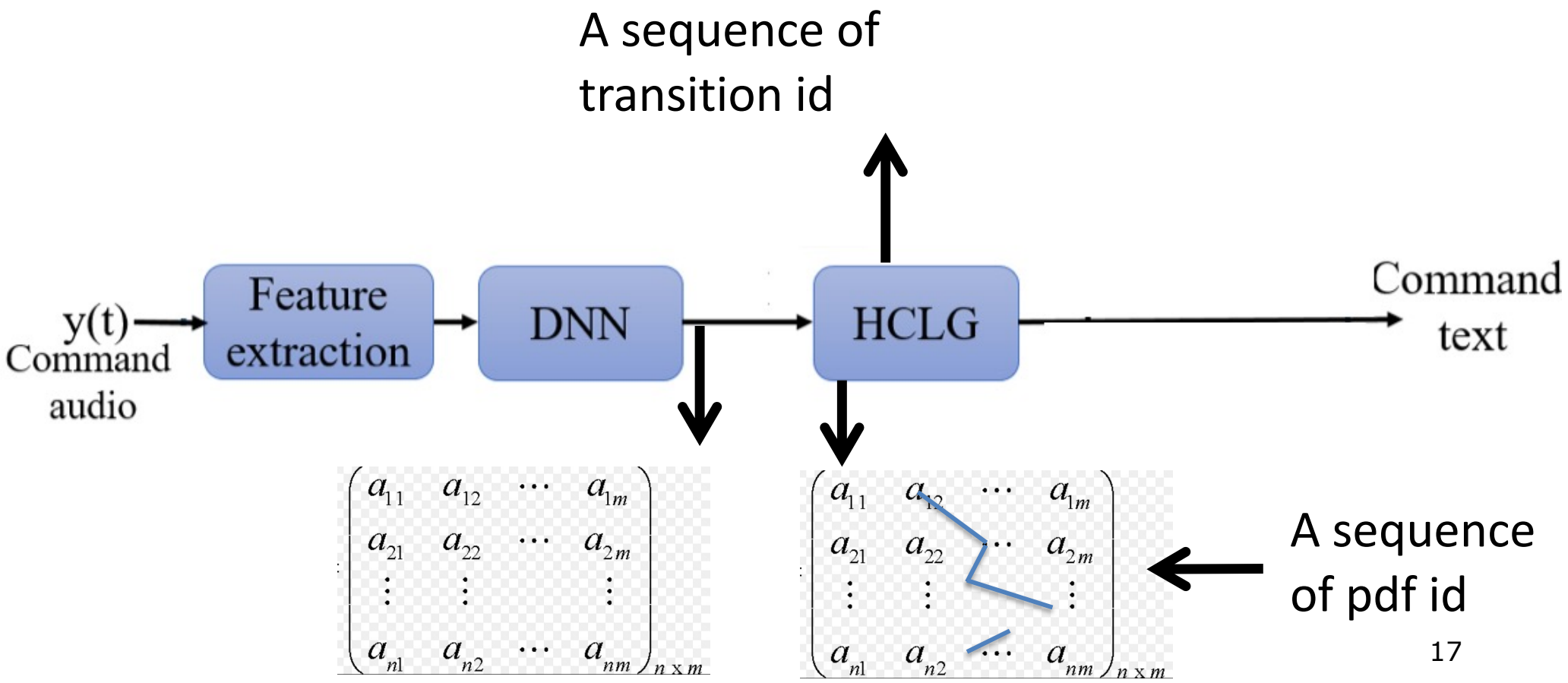
Our Approach



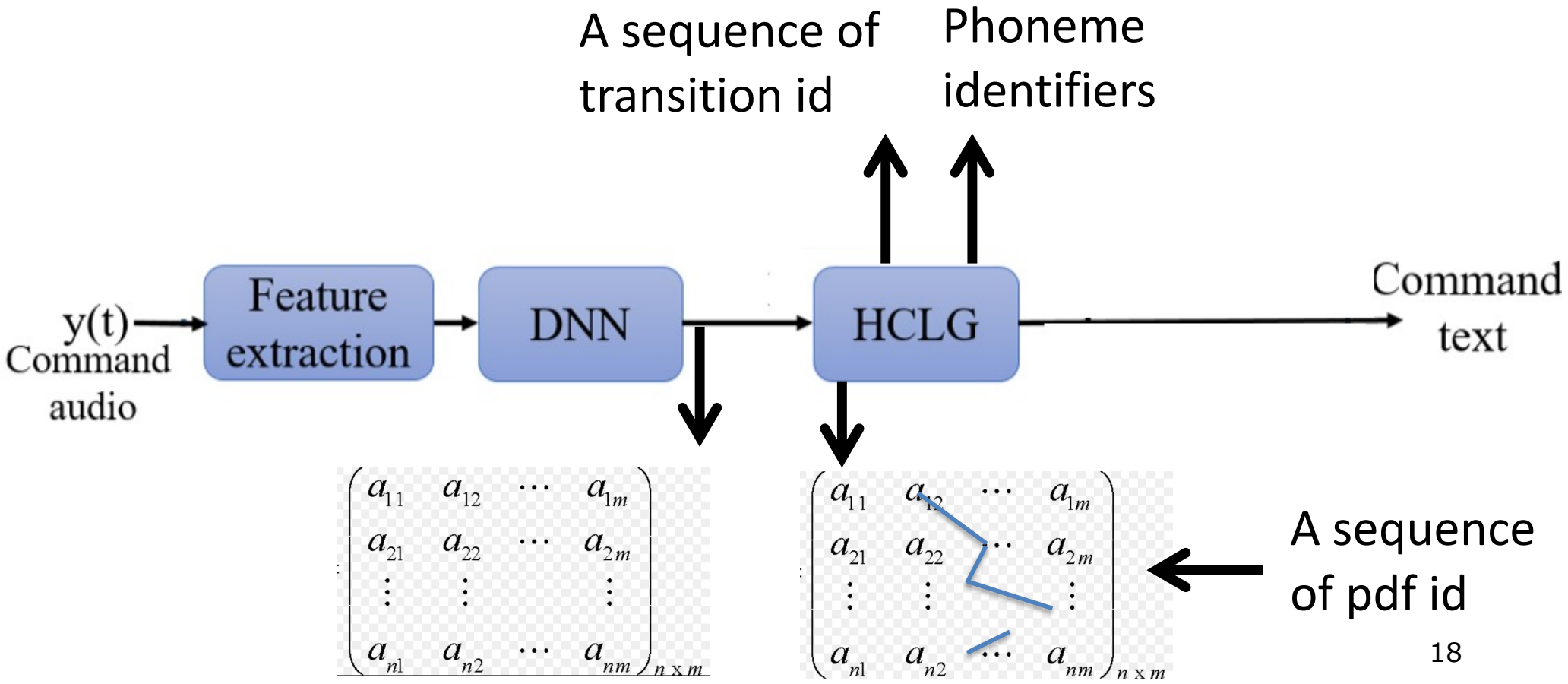
Our Approach



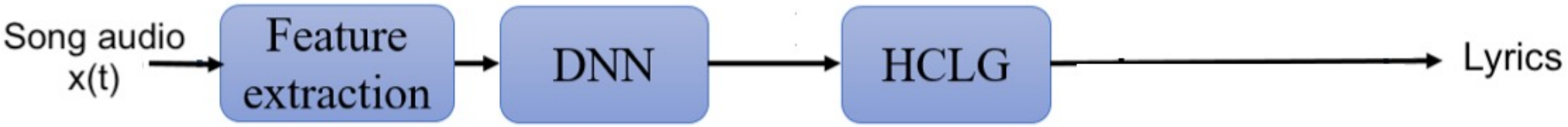
Our Approach

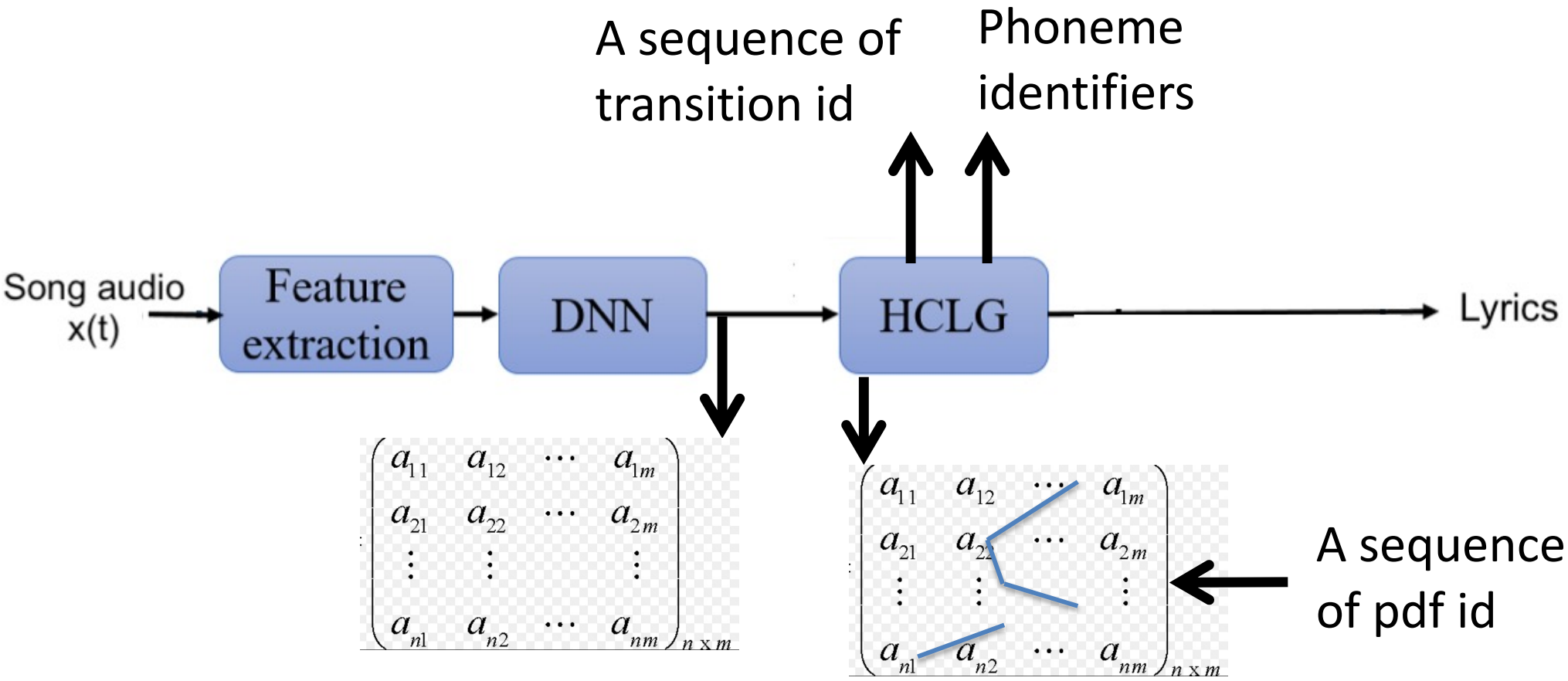


Our Approach

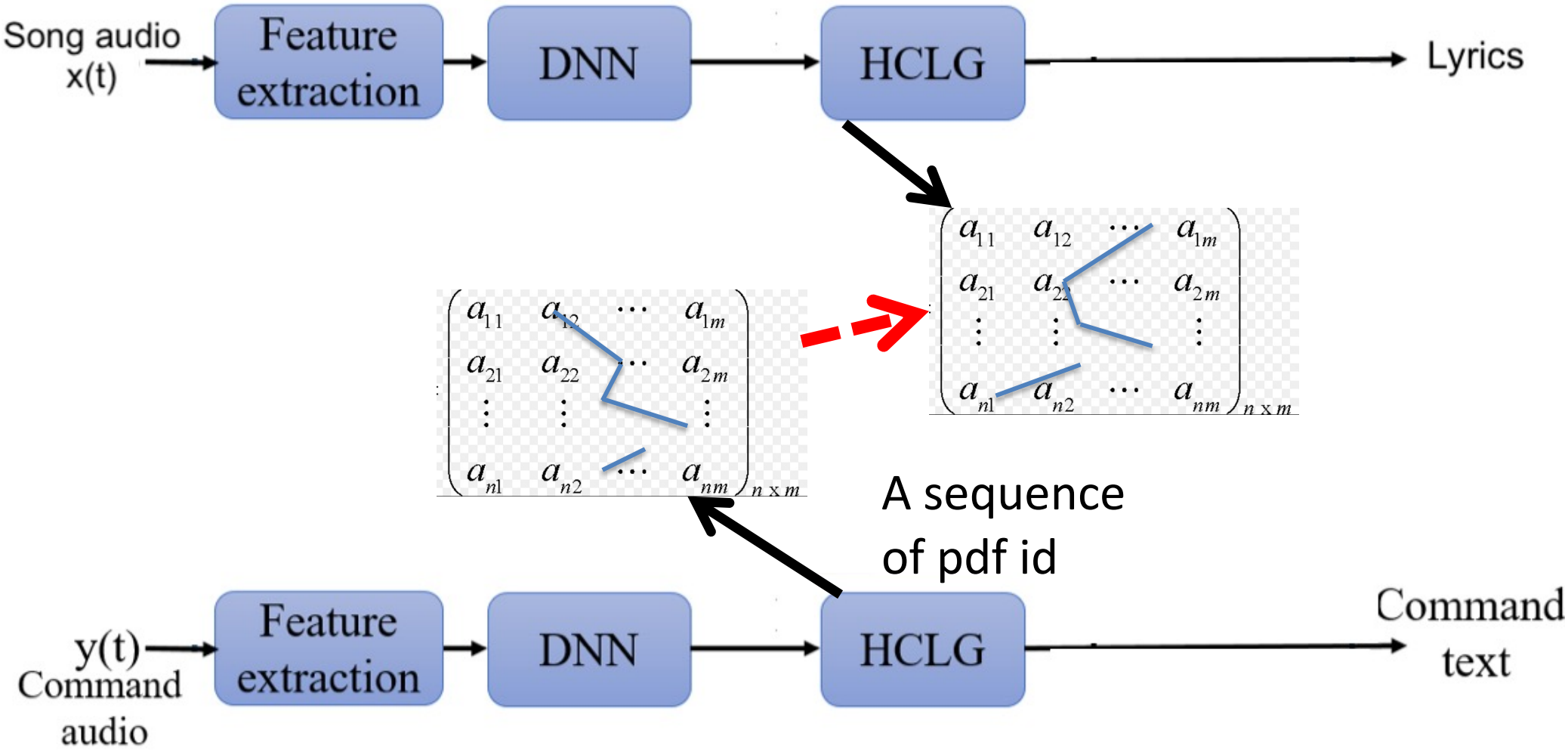


Our Approach

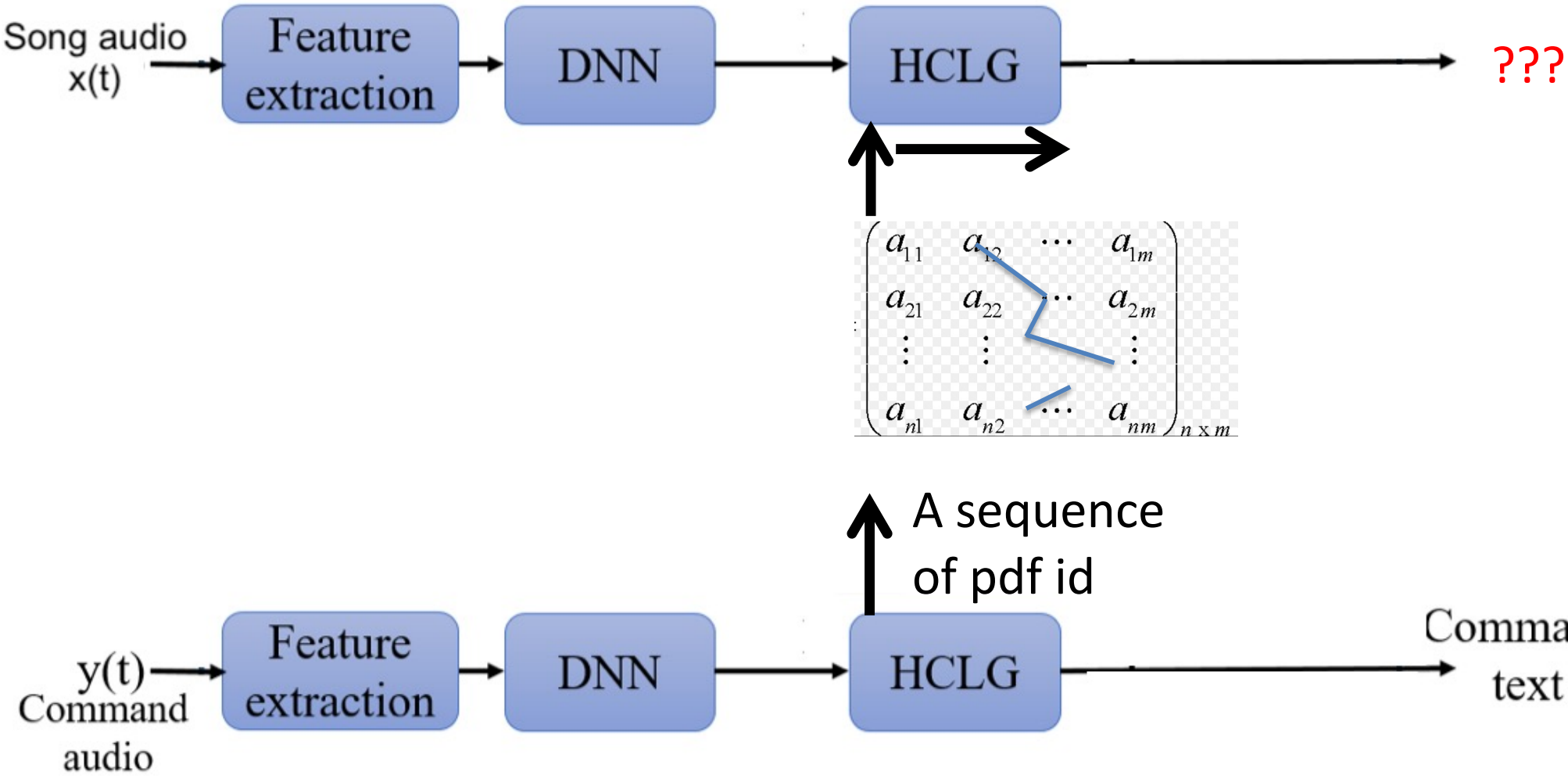



A sequence of transition id
Phoneme identifiers
Song audio x(t)
Feature extraction
DNN
HCLG
Lyrics
A sequence of pdf id

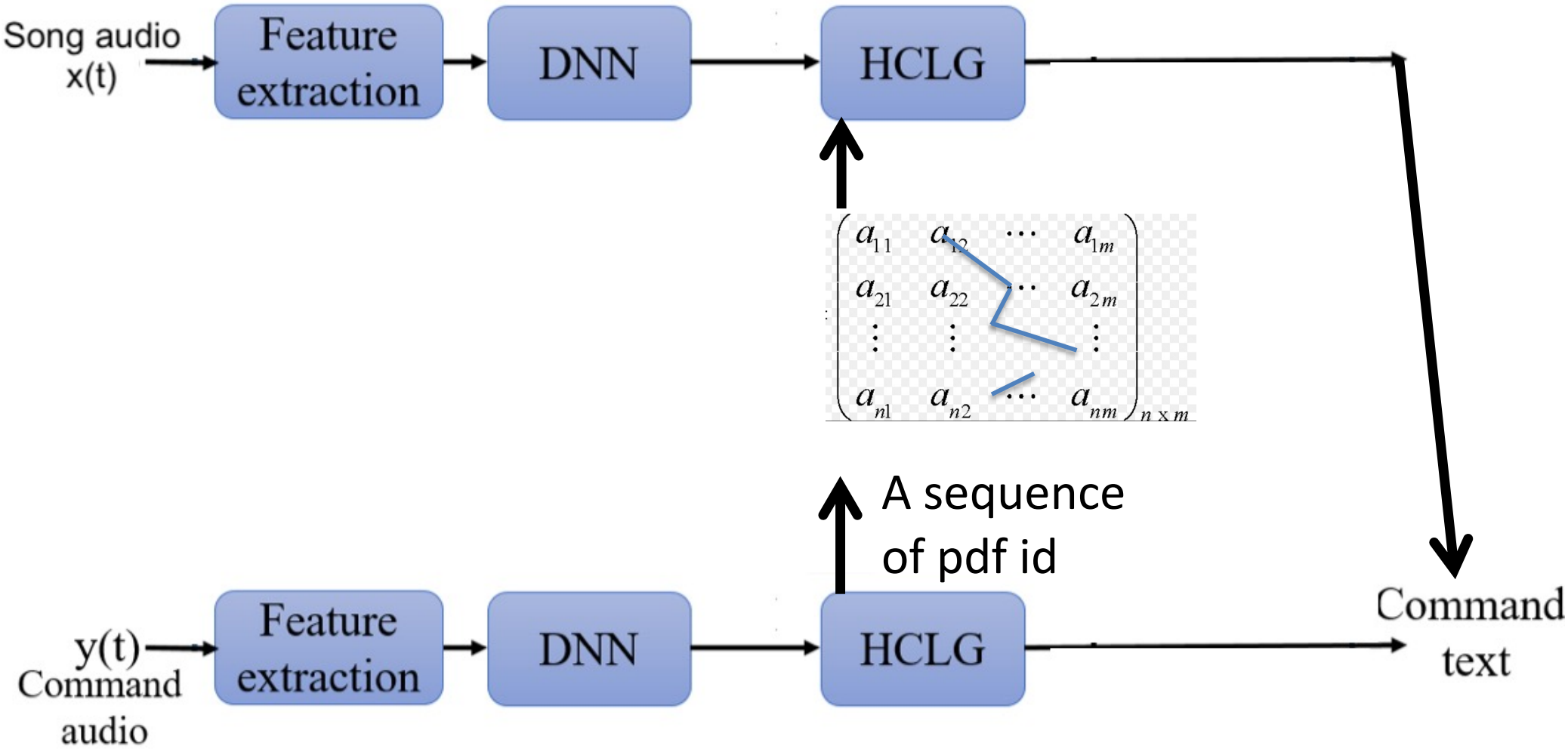
Our Approach



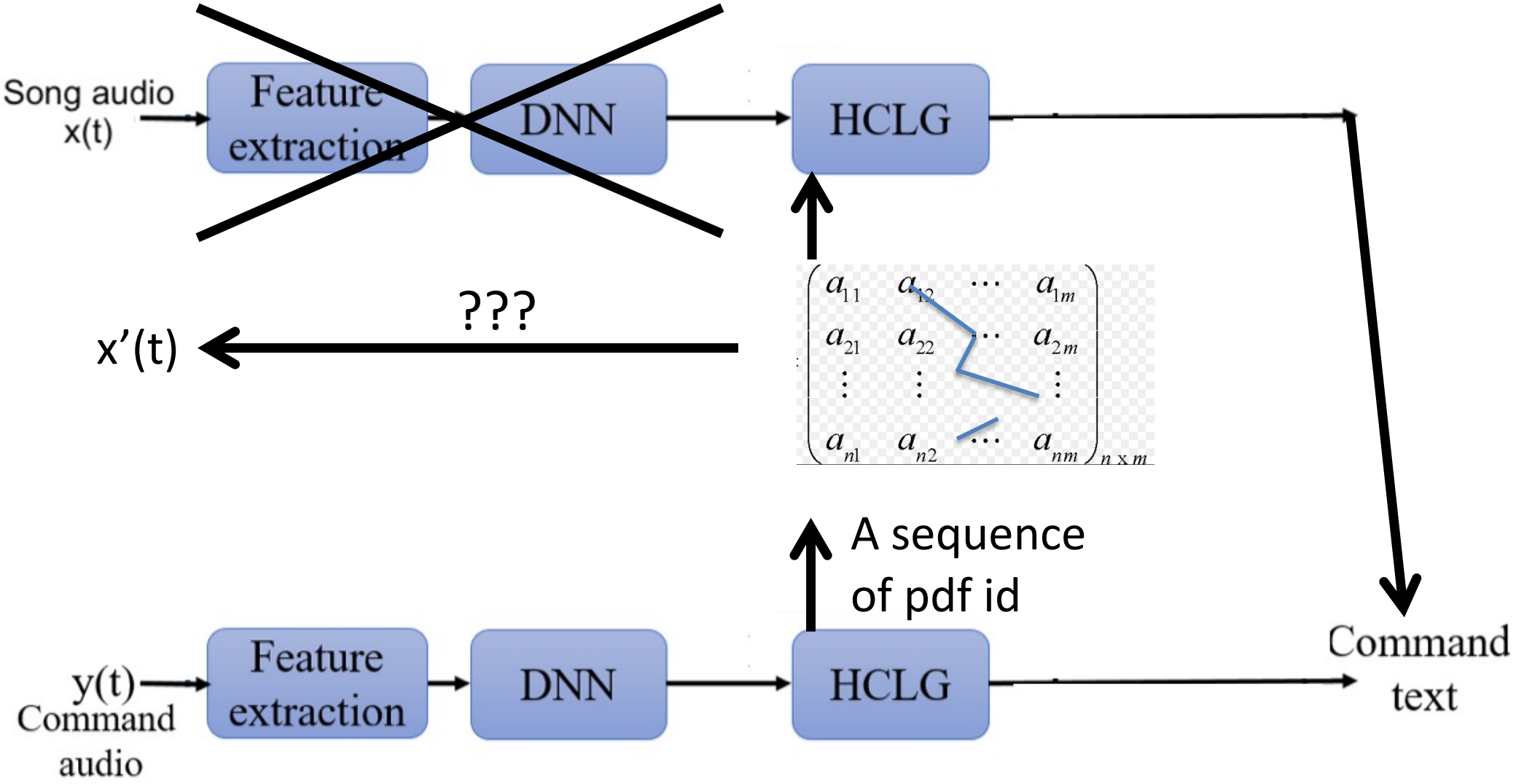
Our Approach



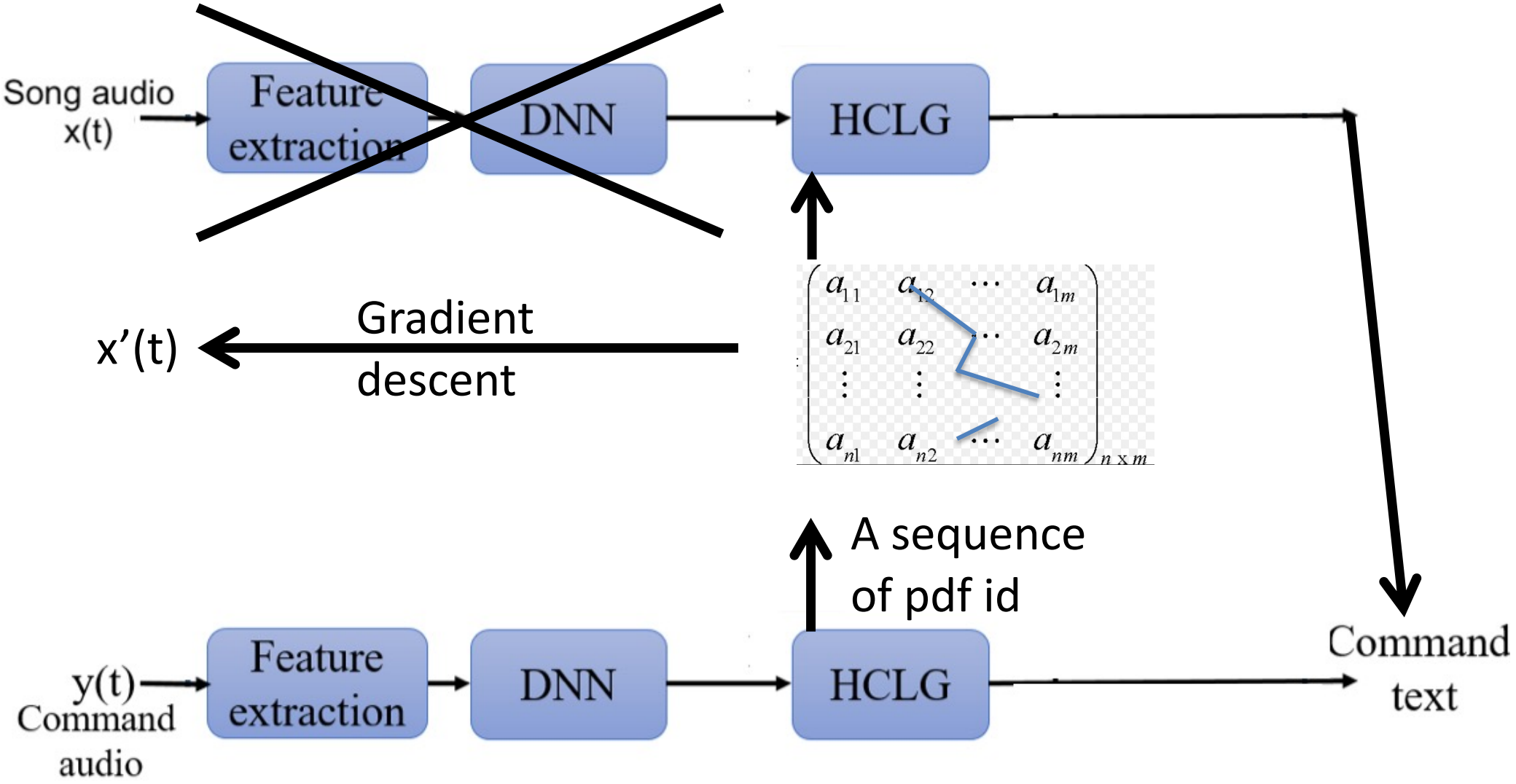
Our Approach



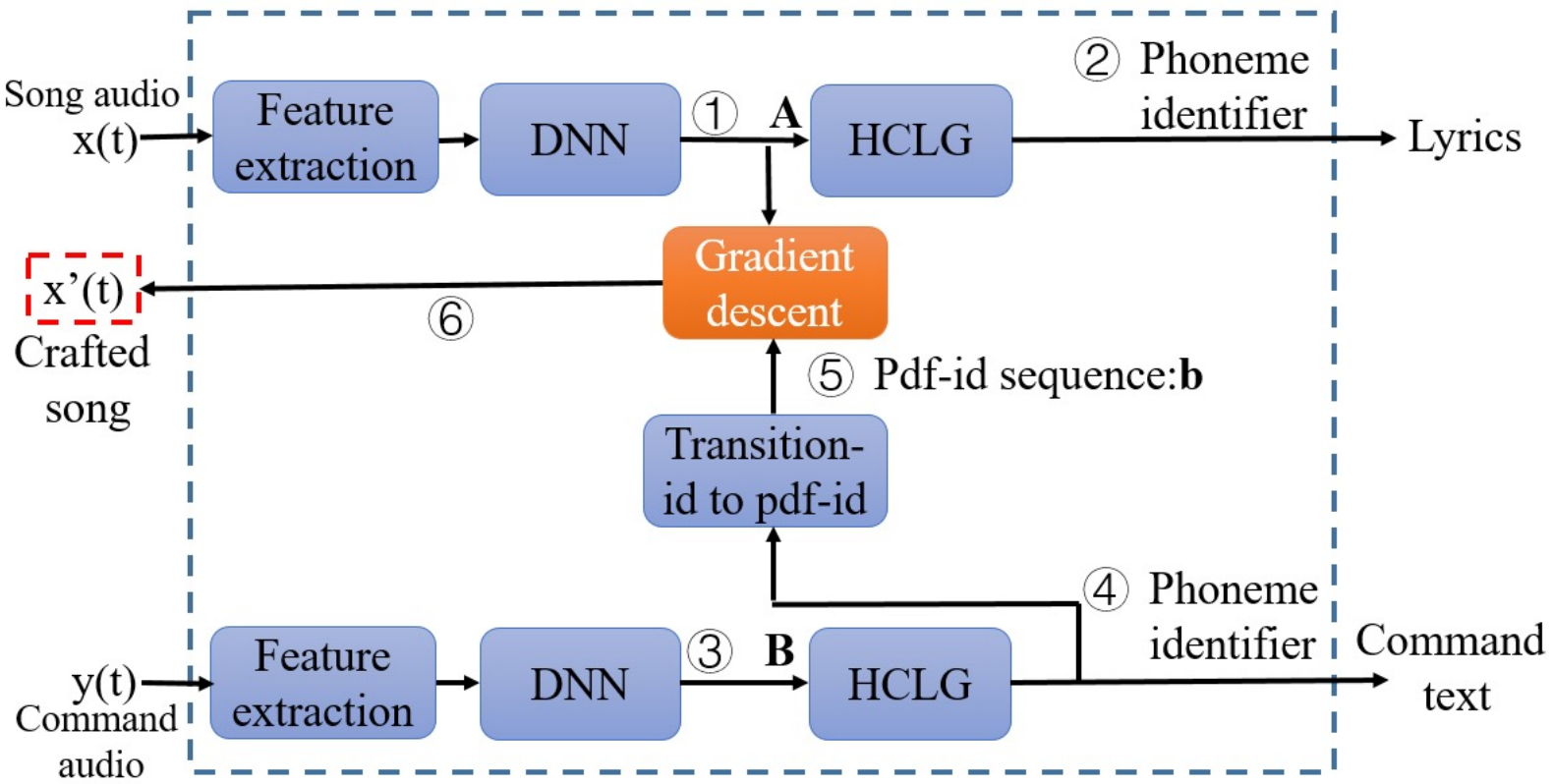
Our Approach



Our Approach



Dealing with Noise



$$\arg\min_{\mu(t)} \quad \|g\left(x(t)+\mu(t)+n(t)\right)-\mathbf{b}\|_1, \qquad (2)$$

Step 2: Black-Box Attack

- Attack what?
  - Google Home, Amazon Echo
  - Google Assistant, Microsoft Cortana, Apple Siri running on smartphone
- Challenges (even you know how to attack physically)
  - You know nothing about the model
  - It fails to respond even if you directly talk to it!

Devil's Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices, USENIX Security 20

Step 2: Black-Box Attack

- Methodology:
  - Target at wakeup word + those frequently-used commands
    - Trained to be quite sensitive to the above phrases

Step 2: Black-Box Attack

- Methodology:
  - Target at wakeup word + those frequently-used commands
    - Trained to be quite sensitive to the above phrases
  - Train a local substitute model approximating the target on the above phrases
    - Build special dataset with twisted speech
    - Query the corresponding speech to text API service

Step 2: Black-Box Attack

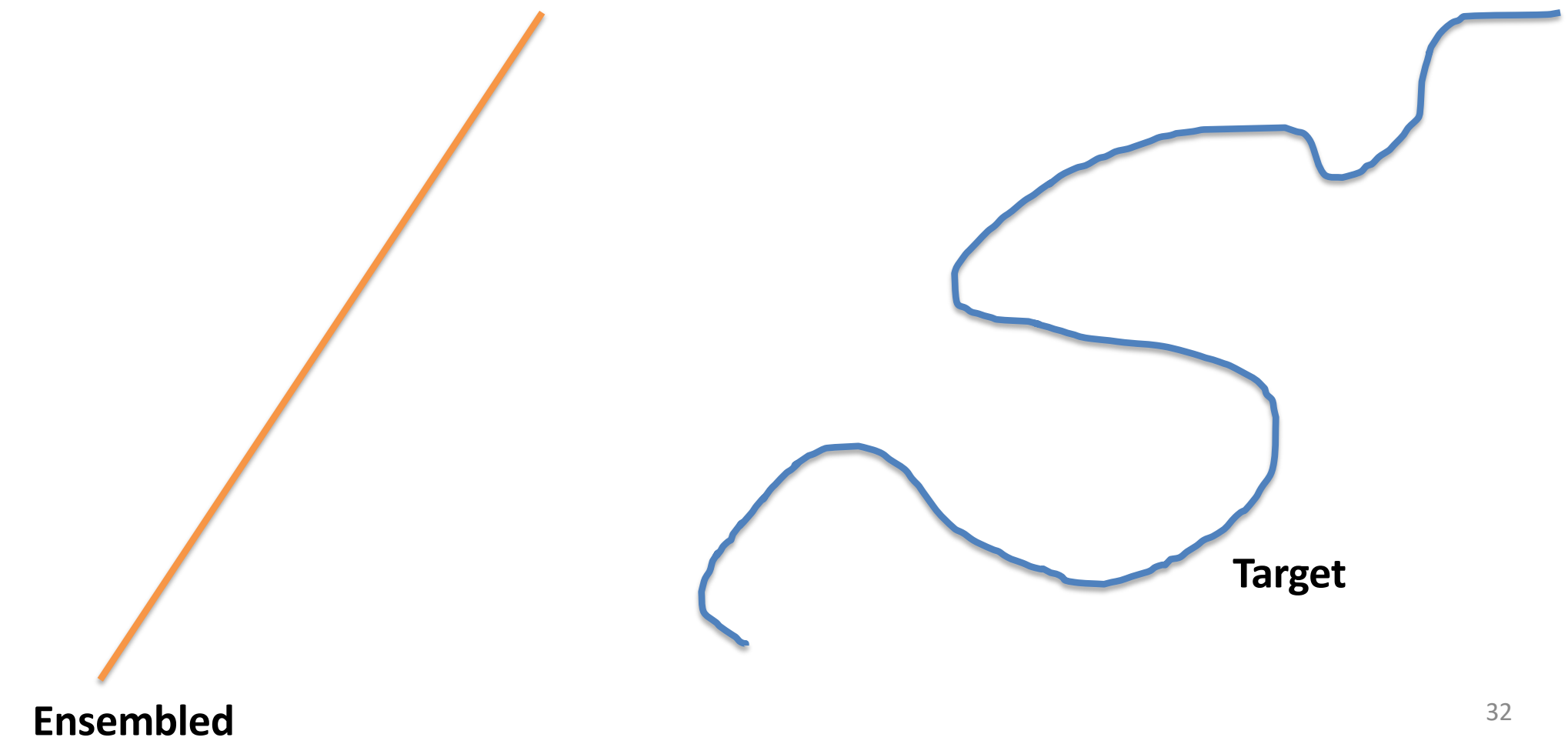
- Methodology:
  - Target at wakeup word + those frequently-used commands
    - Trained to be quite sensitive to the above phrases
  - Train a local substitute model approximating the target on the above phrases
    - Build special dataset with twisted speech
    - Query the corresponding speech to text API service
  - Enhance the substitute using a complete base model
    - Limited capacity of local substitute model

Step 2: Black-Box Attack

- Methodology:
  - Target at wakeup word + those frequently-used commands
    - Trained to be quite sensitive to the above phrases
  - Train a local substitute model approximating the target on the above phrases
    - Build special dataset with twisted speech
    - Query the corresponding speech to text API service
  - Enhance the substitute using a complete base model
    - Limited capacity of local substitute model
  - Finally
    - Ensemble a local substitute model and a complete model
    - Both are white box

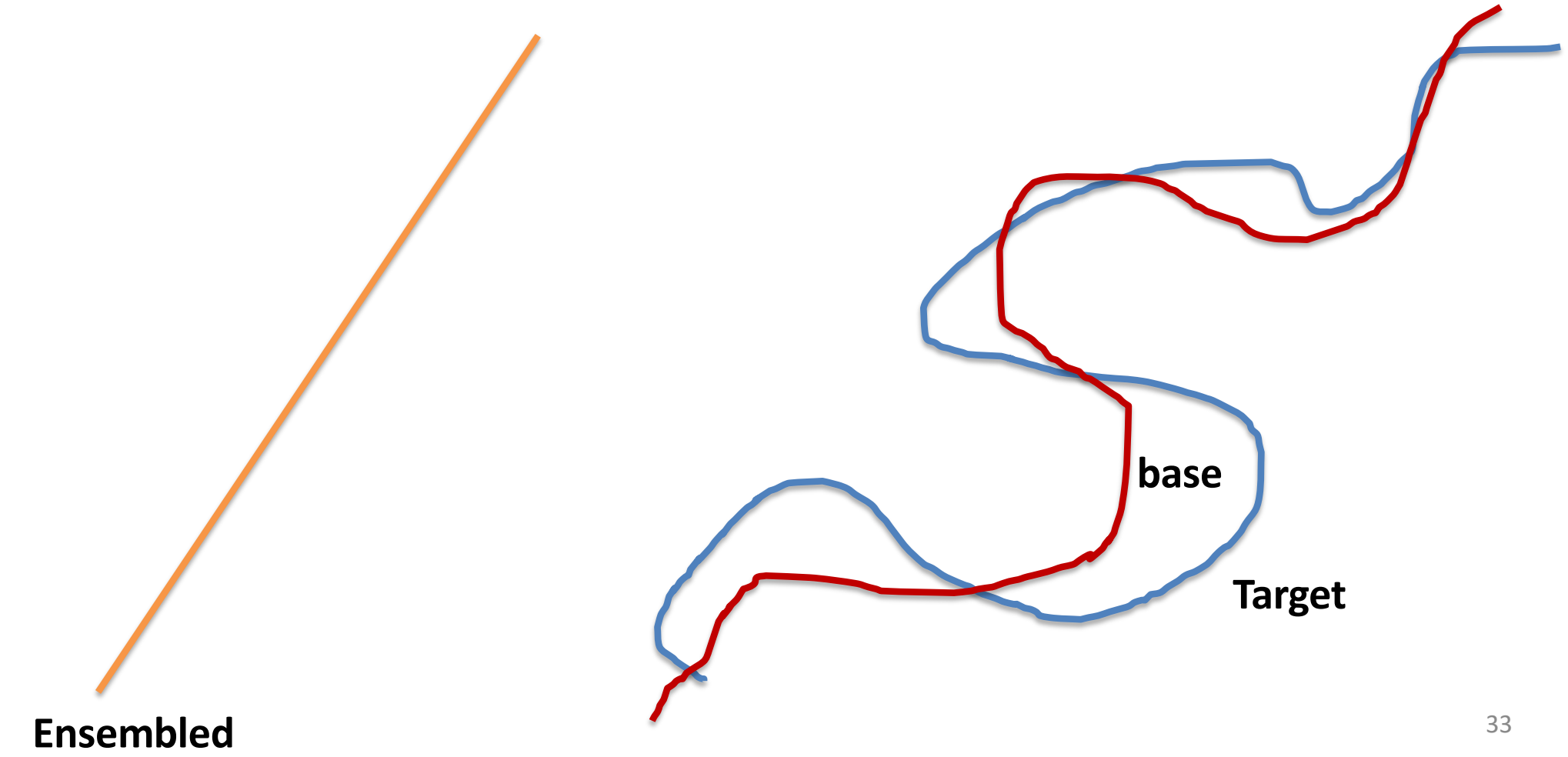
Step 2: Black-Box Attack

- Interpretability
  - Decision boundary in high dimensional space



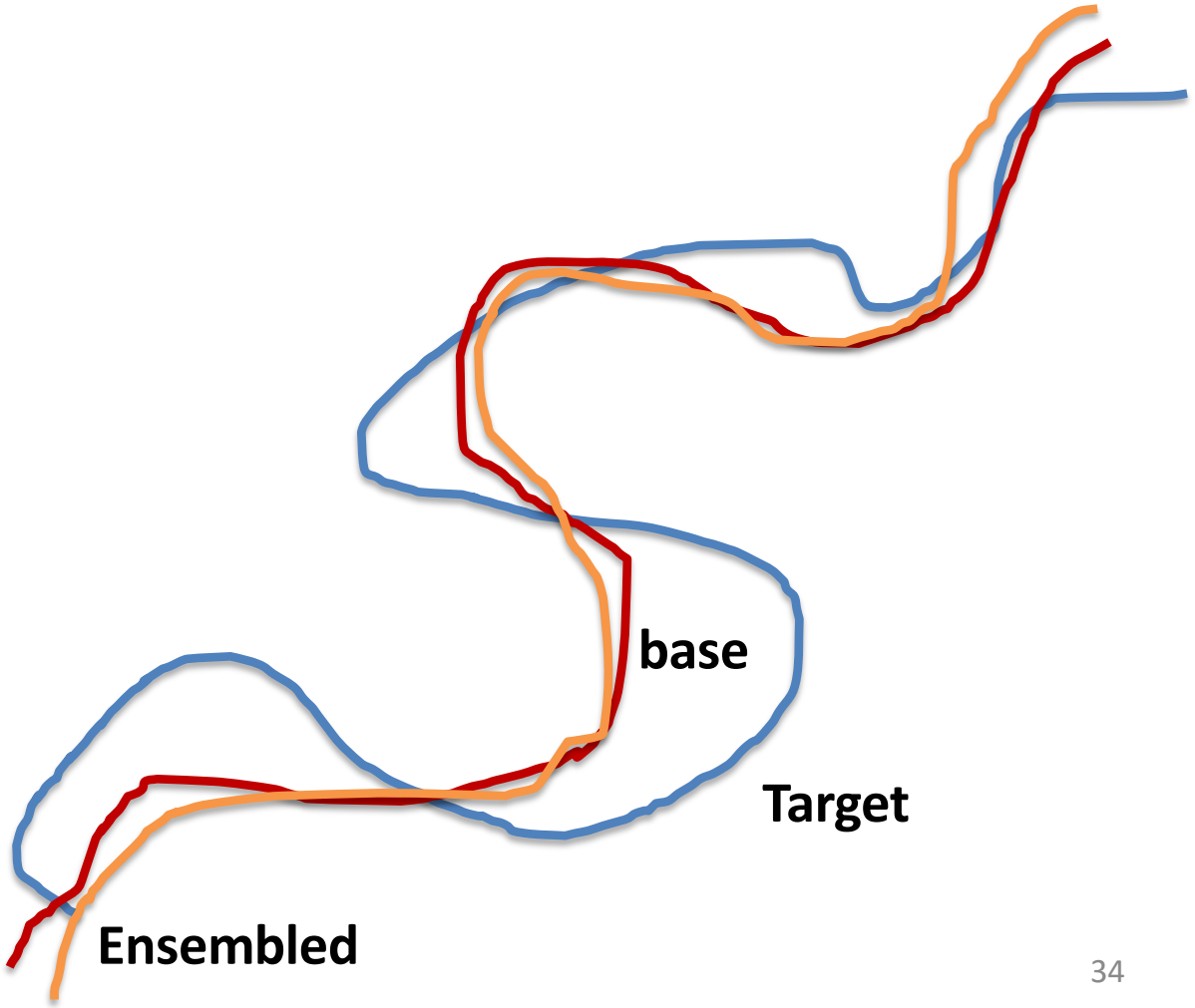
Step 2: Black-Box Attack

- Interpretability
  - Decision boundary in high dimensional space



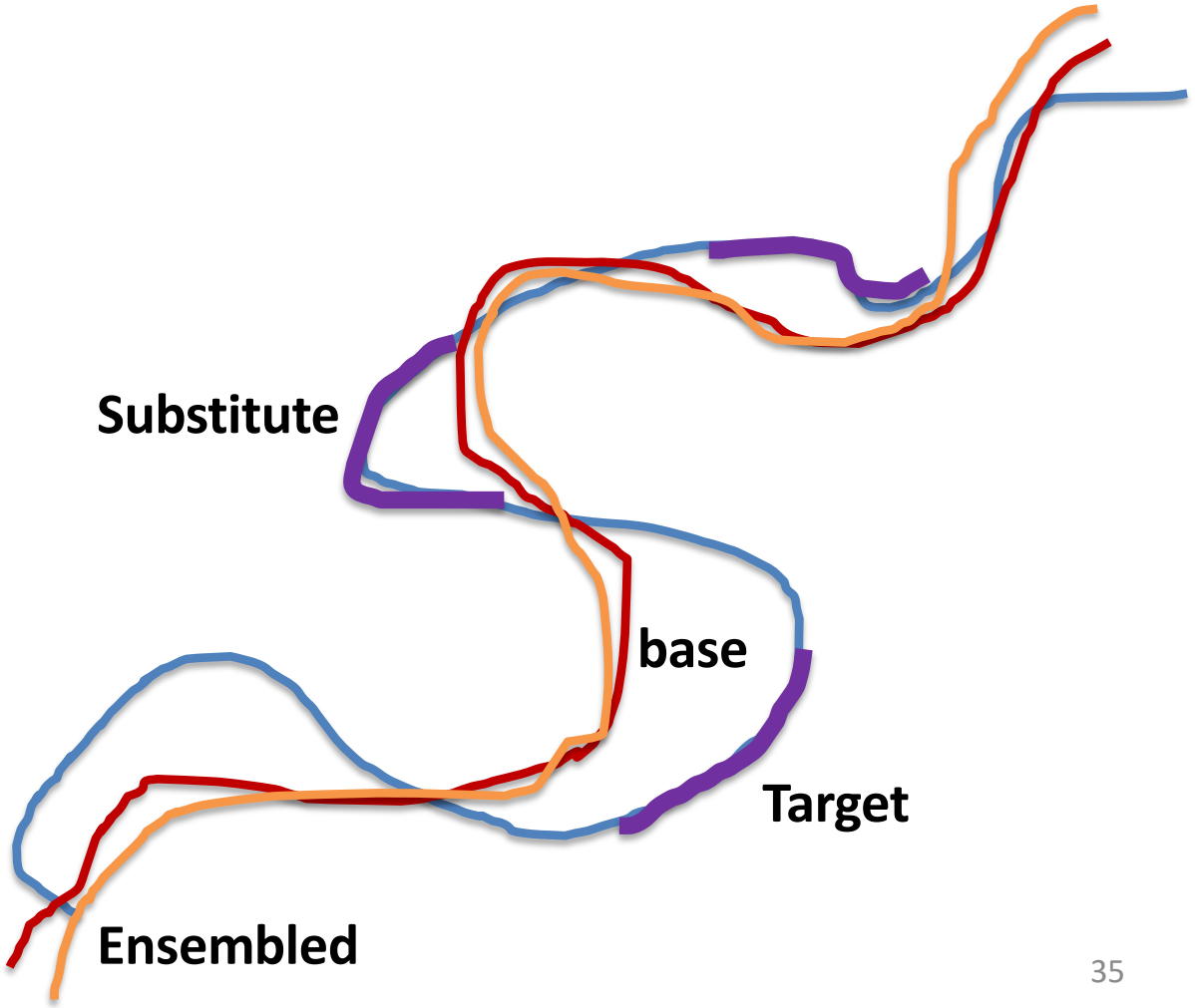
Step 2: Black-Box Attack

- Interpretability
  - Decision boundary in high dimensional space



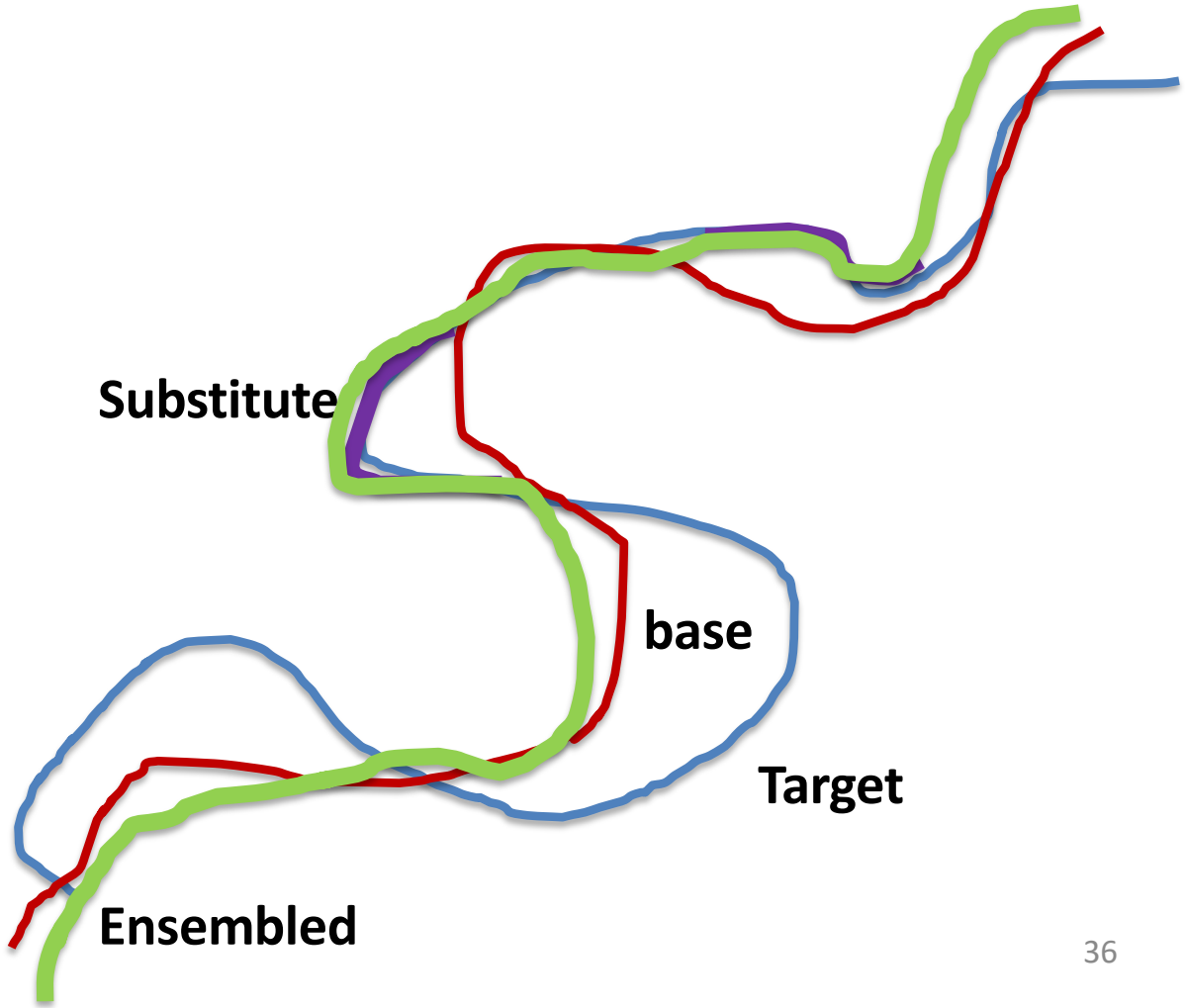
Step 2: Black-Box Attack

- Interpretability
  - Decision boundary in high dimensional space



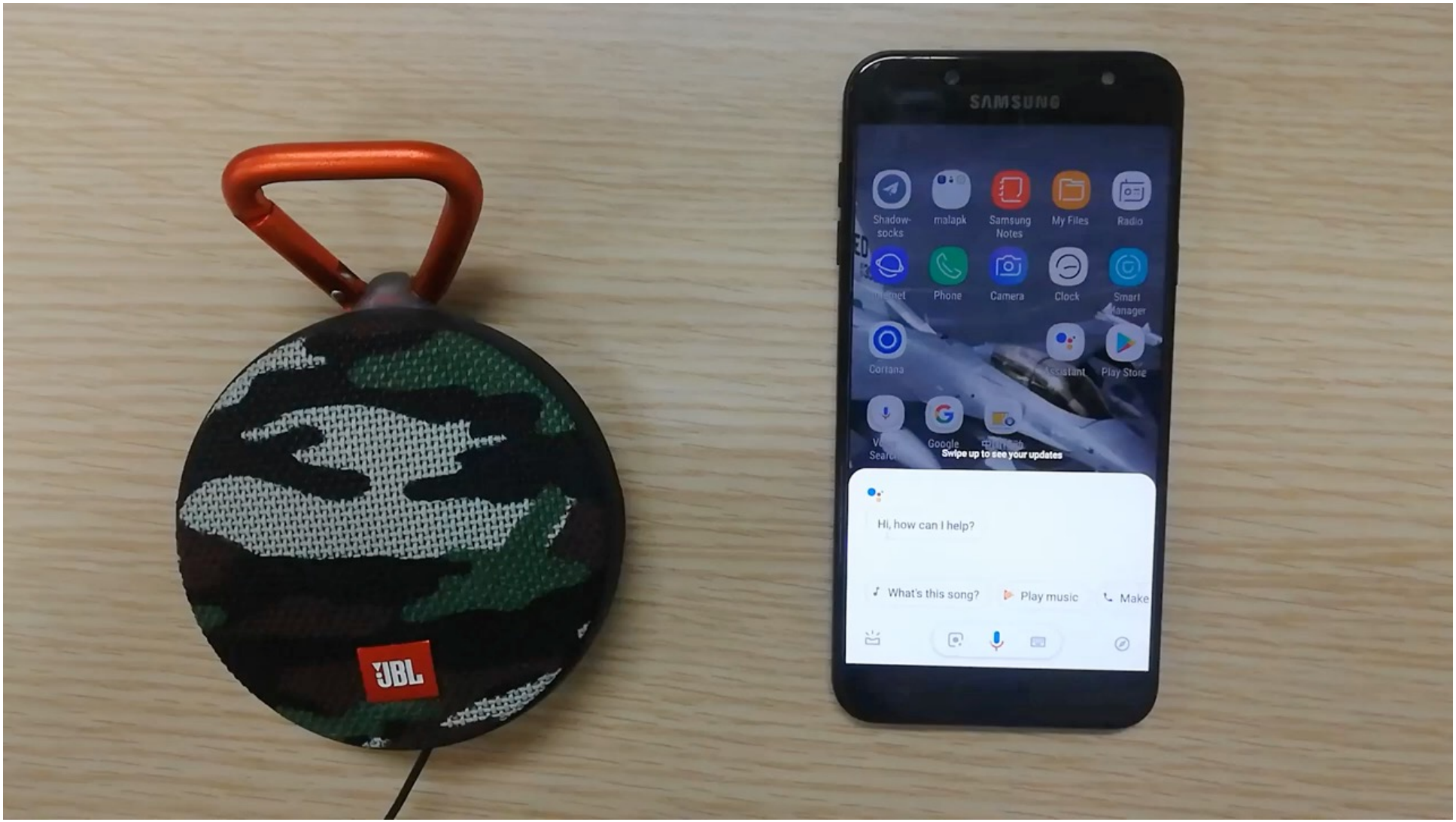
Step 2: Black-Box Attack

- Interpretability
  - Decision boundary in high dimensional space

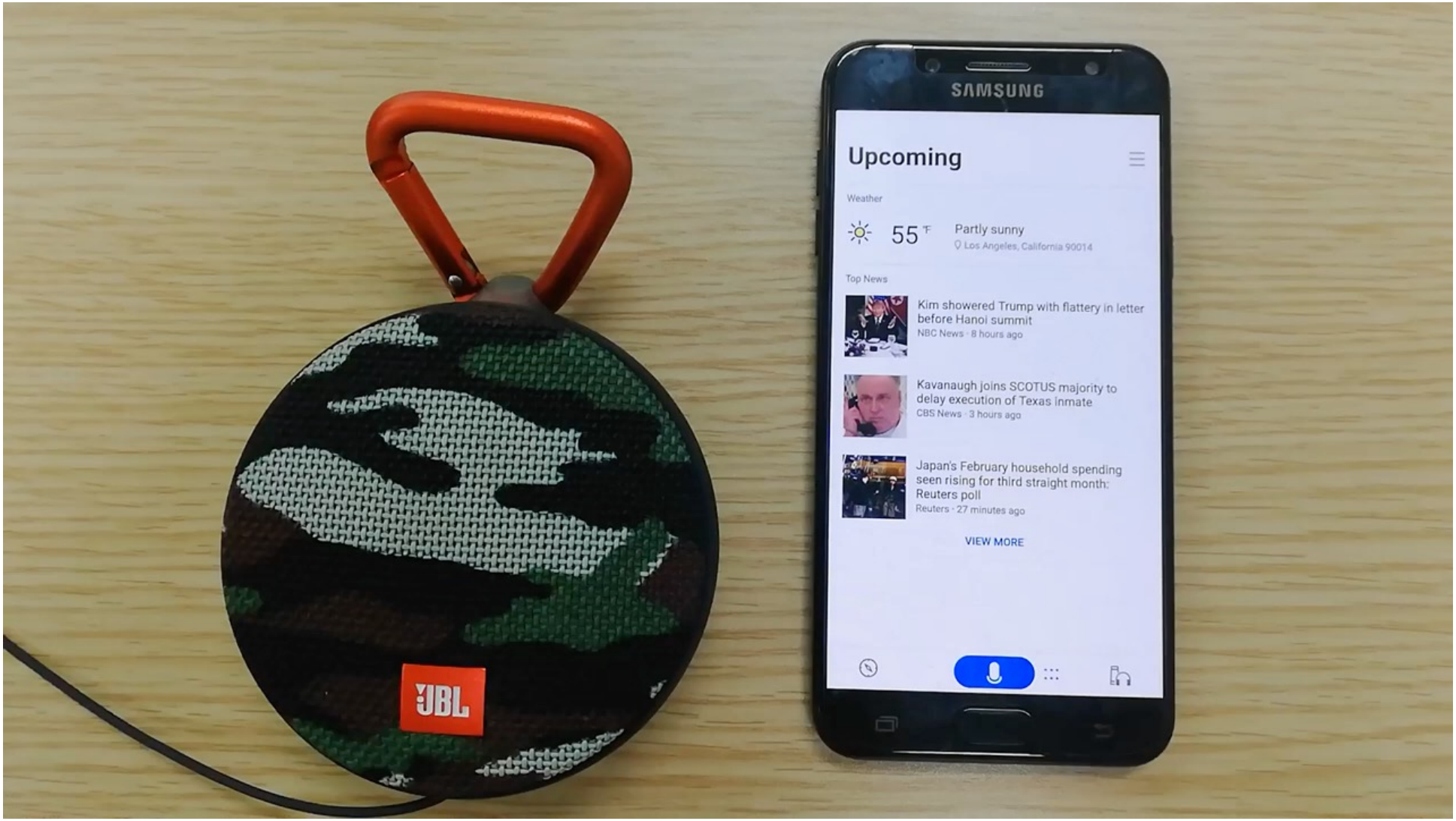


Step 2: Black-Box Attack Demo

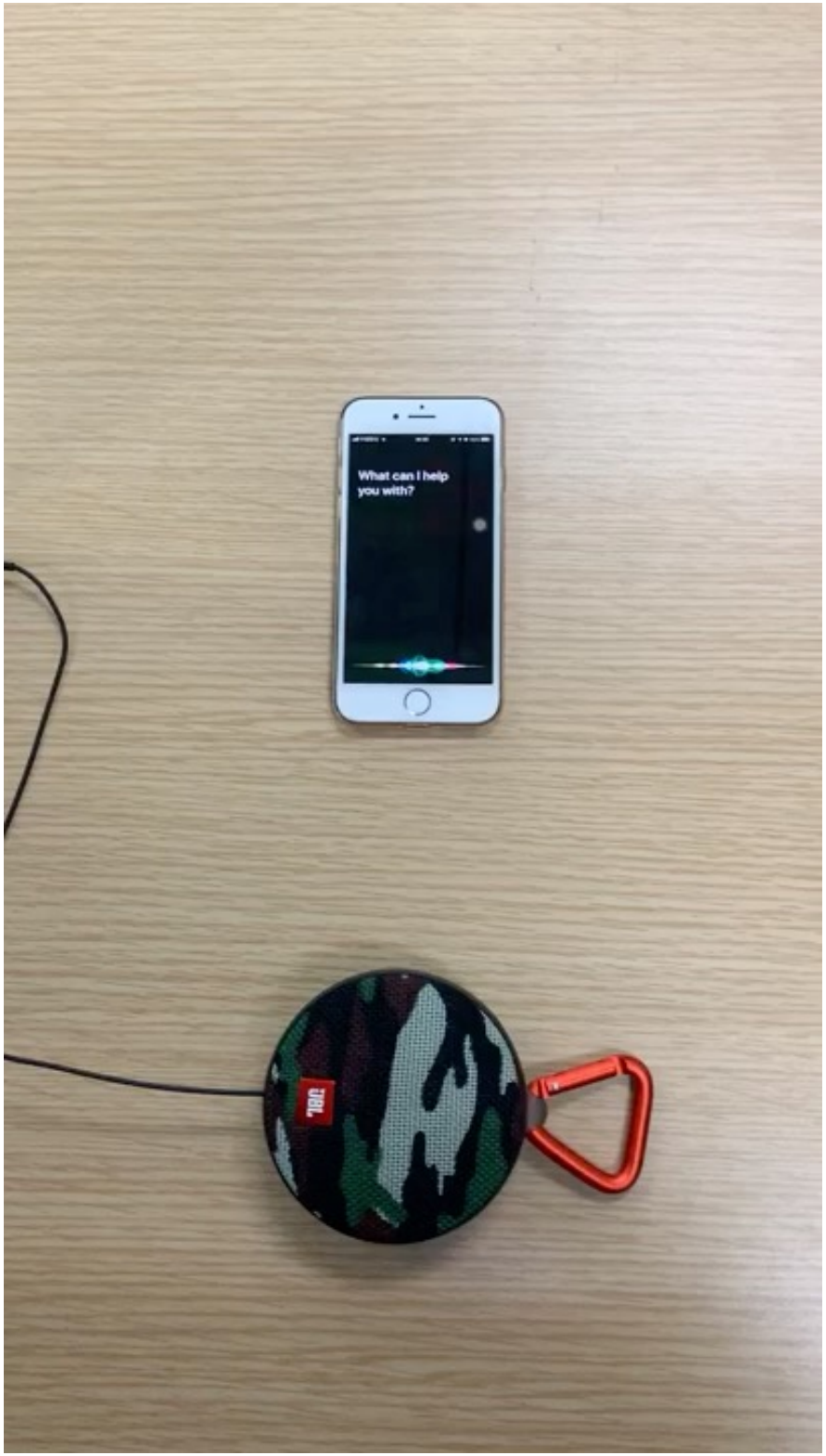
Step 2: Black-Box Attack Demo



Step 2: Black-Box Attack Demo



Step 2: Black-Box Attack Demo



Outline

- What is Adversarial Attack?
- Our research
- **Conclusion**

Conclusion

- Attacks are feasible
  - Image, audio, video
  - Digital, physical
  - White box, black box
- **Defense**
  - Quite limited and specific
    - Examine the distribution of training dataset
    - Smoothen the gradient
  - General approach to defend machine learning is demanded


Thank You!