Can I Hear Your Face?

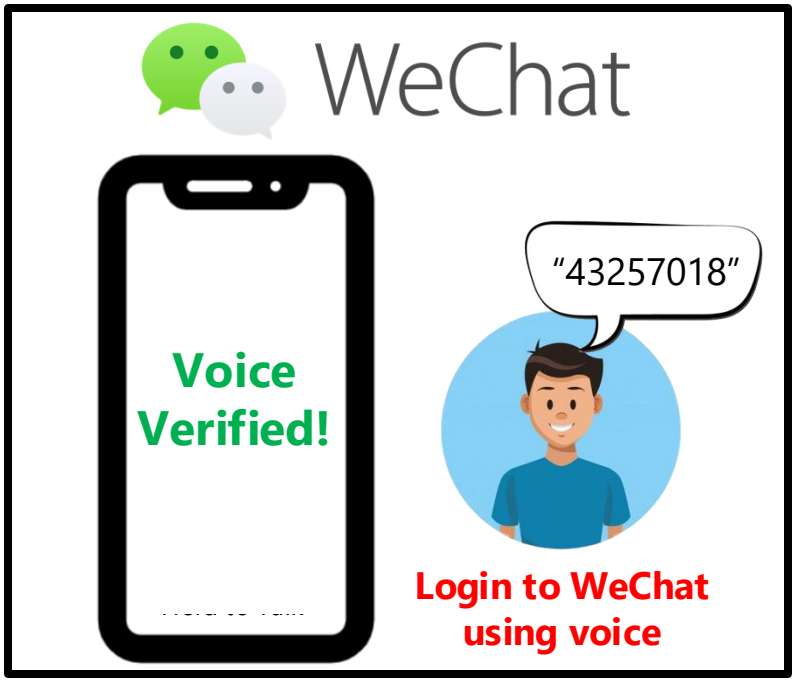
Pervasive Attack on Voice Authentication Systems with a Single Face Image

Nan Jiang, Bangjie Sun, Terence Sim, and Jun Han*

National University of Singapore * KAIST

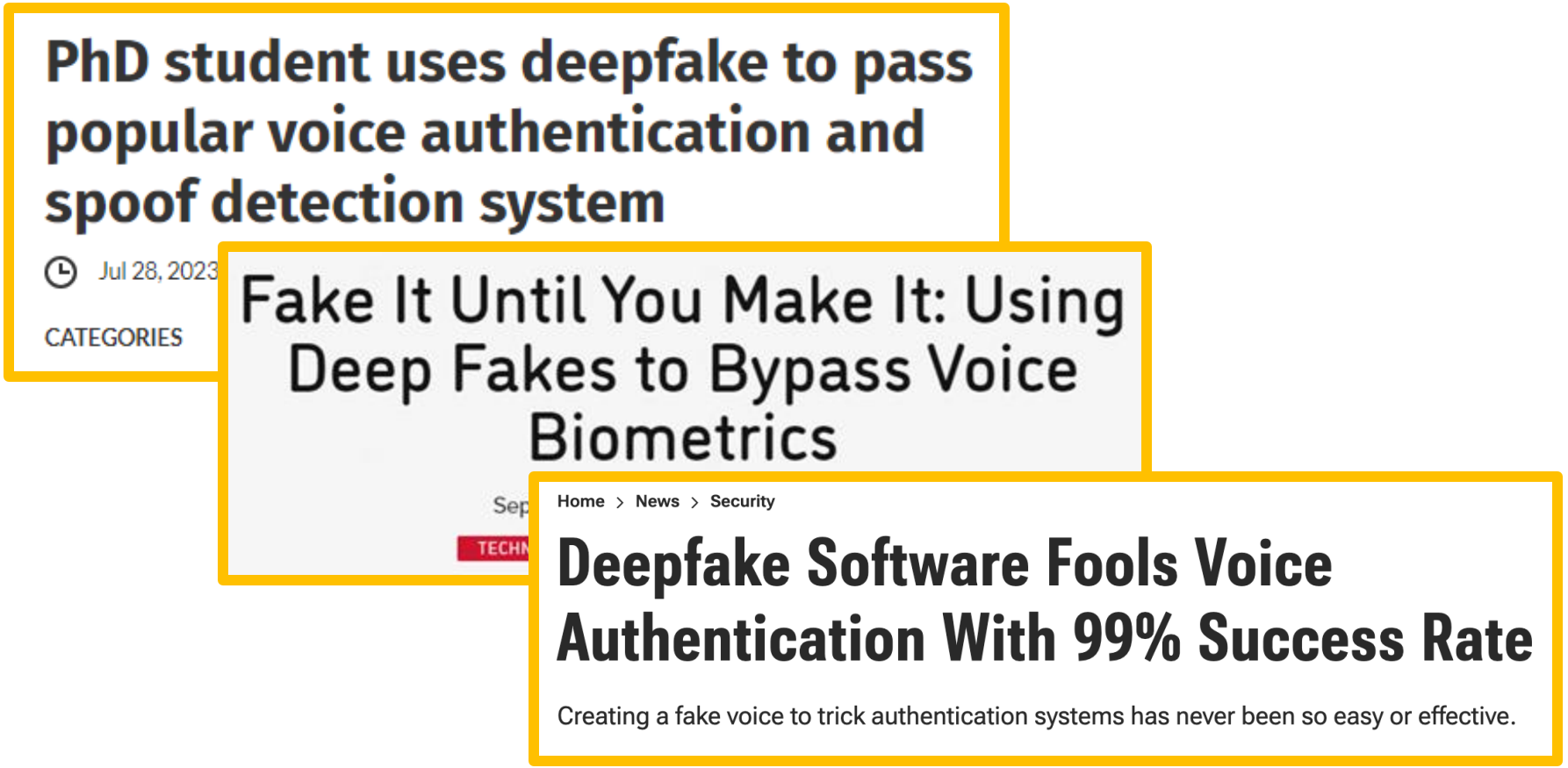
Voice Authentication in Daily Life

Login to Account



Activate Voice Assistant

Voice Deepfake Can Bypass Voice Authentication



Limitation of Voice Deepfake Attack

- Require **high-quality** recording of the victim's voice

(a) Download from the network

Limited range of victims

Limitation of Voice Deepfake Attack

- Require **high-quality** recording of the victim's voice

(a) Download from the network

Limited range of victims

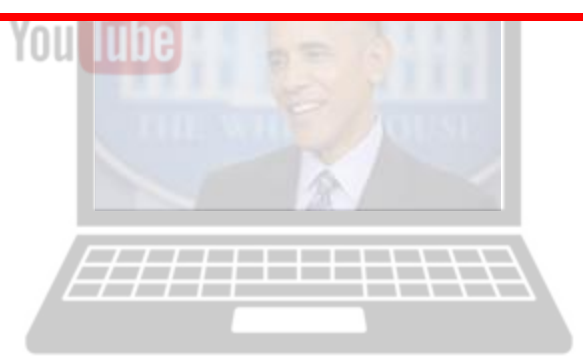
(b) Secretly record the victim's speech

Poor quality voice recordings

Limitation of Voice Deepfake Attack

- Require **high-quality** recording of the victim's voice

(a) Download from the network



Limited range of victims

(b) Secretly record the victim's

Poor quality voice recordings

How to Overcome this Limitation?

How About Replacing Voice with Face Images?

- Easier to obtain face images

(a) Download from the network

Wide availability of face images

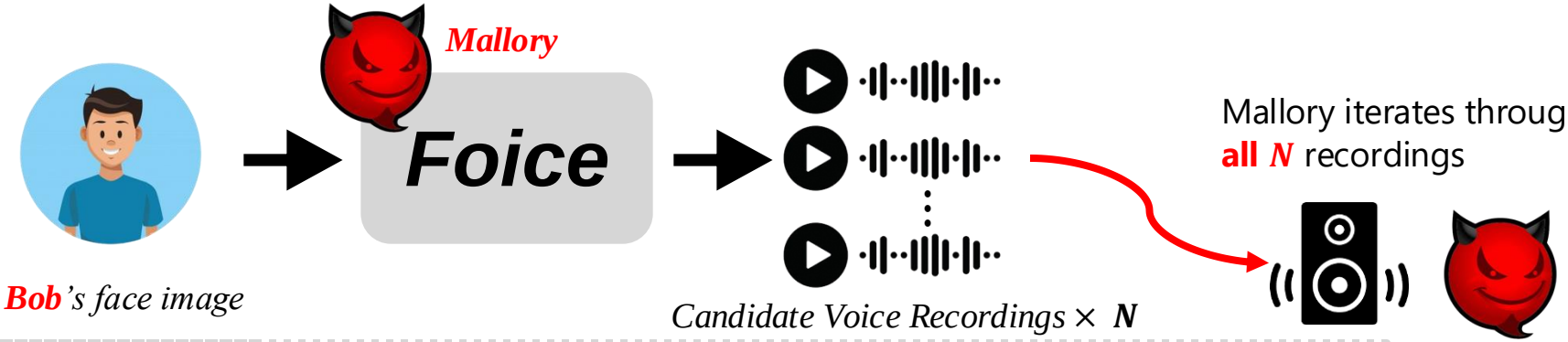
(b) Stealthy photography

Easy capture of clear images

Can we launch a novel attack leveraging only a single image of the victim without requiring voice recordings?

Our Work: *Foice*

- Synthesizes **voice recordings** from a **single face image**



Threat Model

- **Attacker's Goal:** compromise verification
  - **Gain unauthorised access** to the victim's private account
  - **Execute unauthorised commands** on the victim's personal voice assistant

Can You Guess What He Sounds Like?

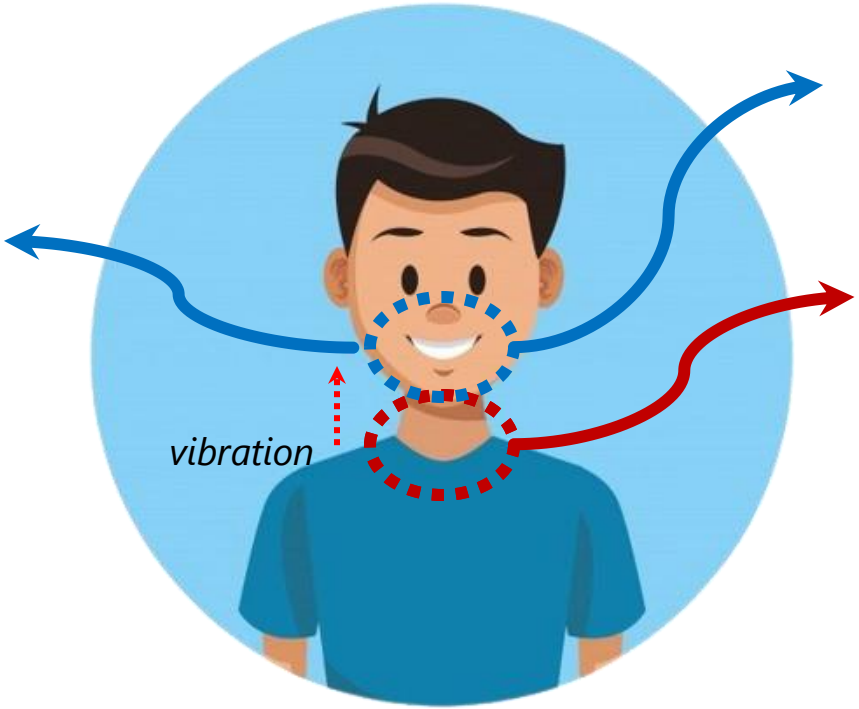
A.

B.

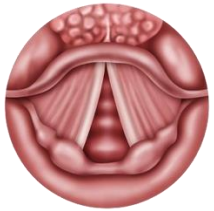
Background: Face and Voice Correlation

- **Facial appearance** affects how human voice sounds like

Vibration here causes **facial** resonance



Narrow space = Bright sound



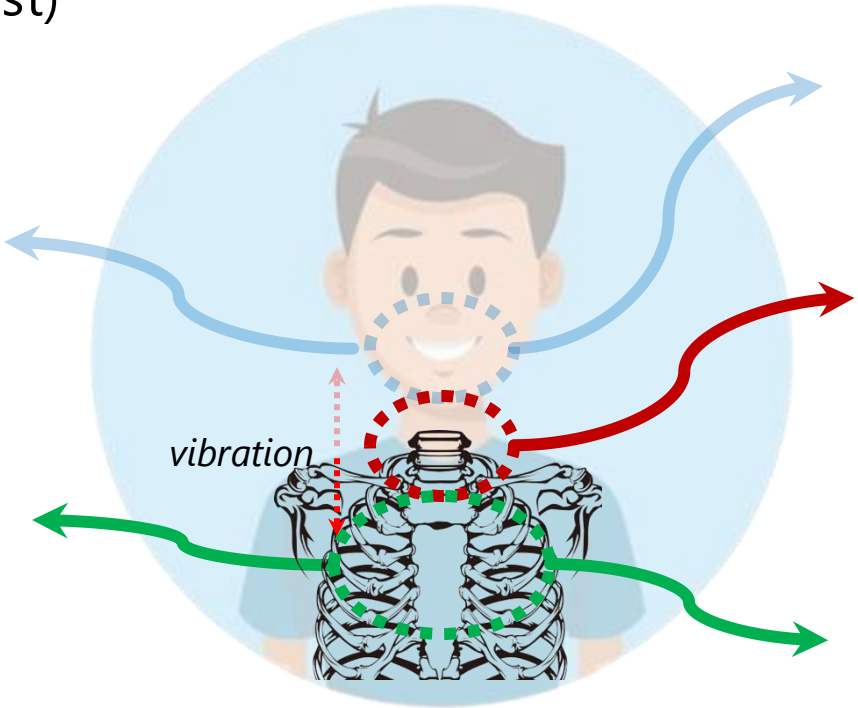
Vocal cords

Background: Face Offers Limited Information

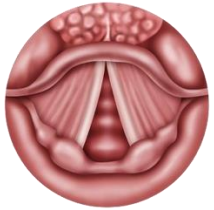
- Human voice is primarily affected by the **inner body structure** (e.g., vocal cords, chest)

Vibration here causes **facial** resonance

Vibration here causes **chest** resonance



Narrow space = Bright sound



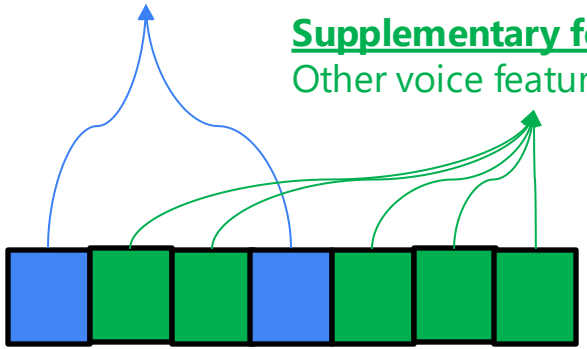
Vocal cords

Wide space = Deep sound

Core Ideas of *Foice*

Face-dependent features:
Voice features related to the face

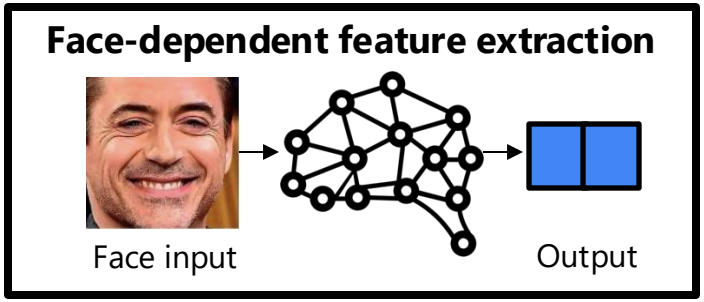
Supplementary features:
Other voice features



Voice features

Our Model

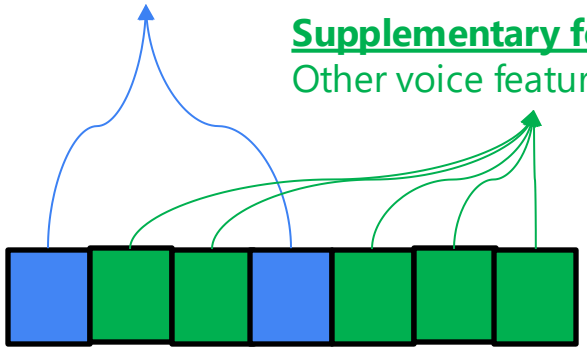
- **Core idea (1)**: **extract** face-dependent features from face image input



Core Ideas of *Foice*

Face-dependent features:
Voice features related to the face

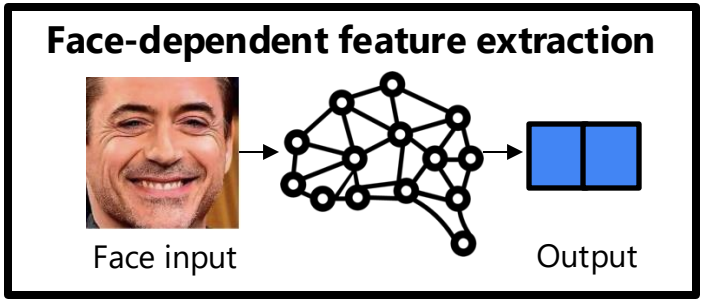
Supplementary features:
Other voice features



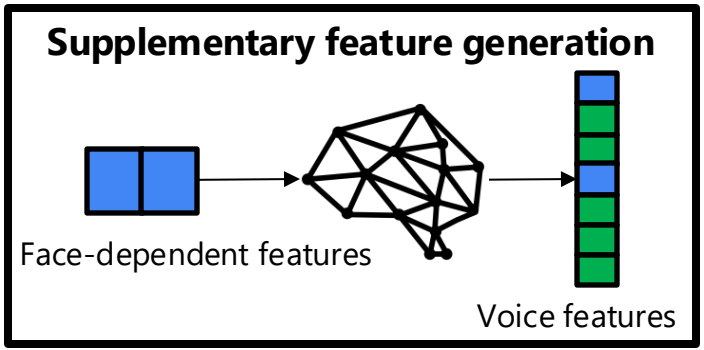
Voice features

Our Model

- **Core idea (1)**: **extract** face-dependent features from face image input



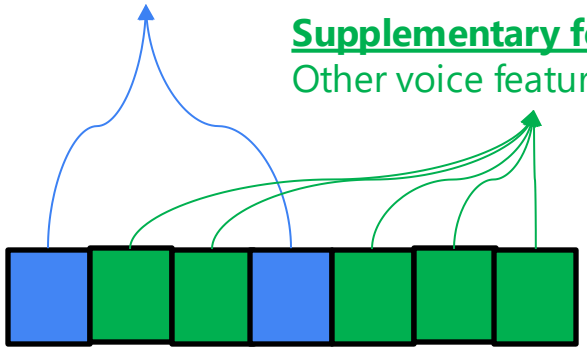
- **Core idea (2): generate** candidate supplementary features



Core Ideas of *Foice*

Face-dependent features:
Voice features related to the face

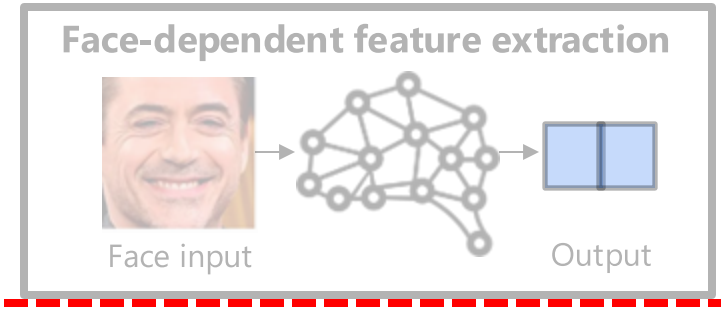
Supplementary features:
Other voice features



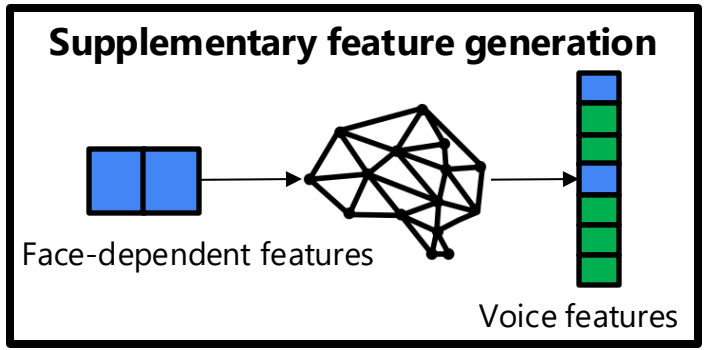
Voice features

Our Model

- **Core idea (1)**: **extract** face-dependent features from face image input

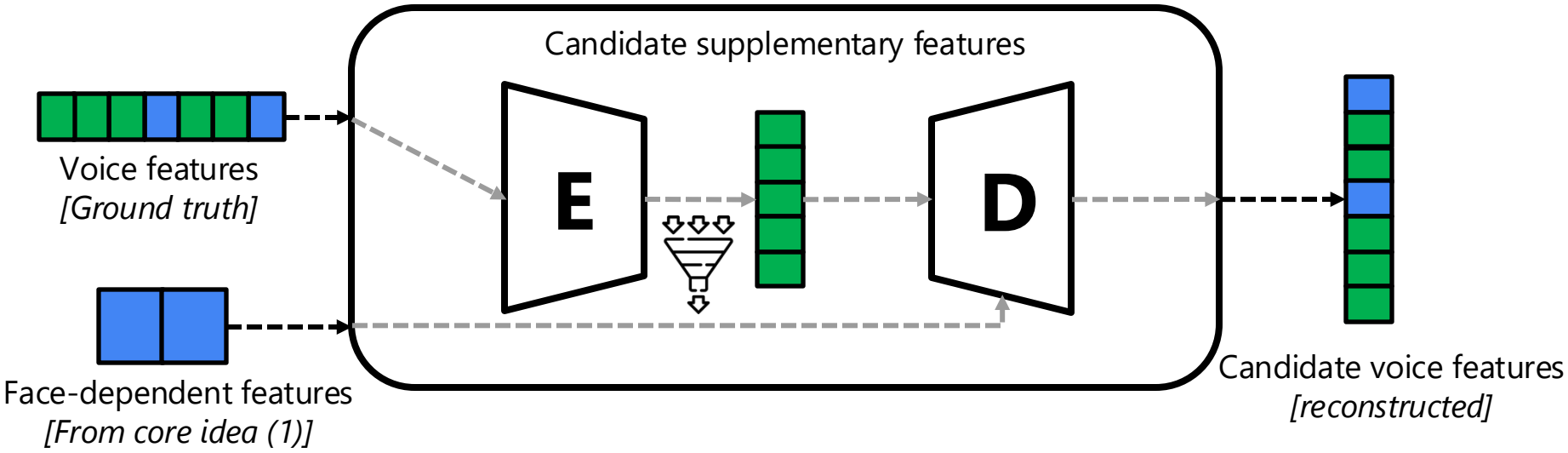


- **Core idea (2): generate** candidate supplementary features



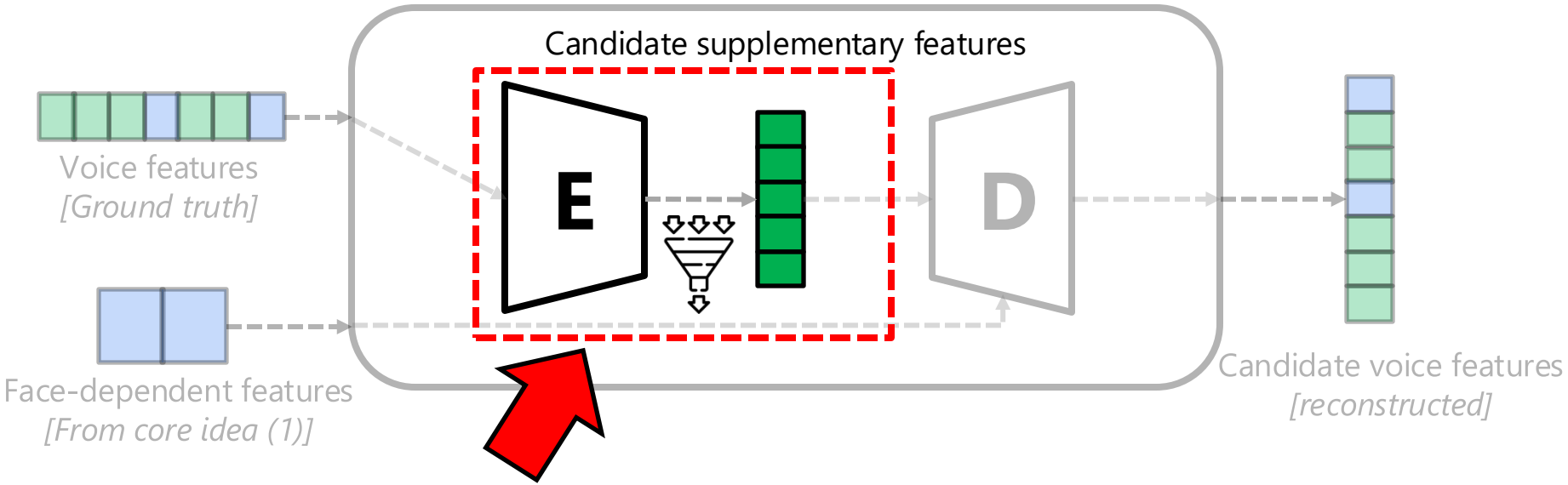
Training: Supplementary Feature Generation

Goal: To generate candidate supplementary features



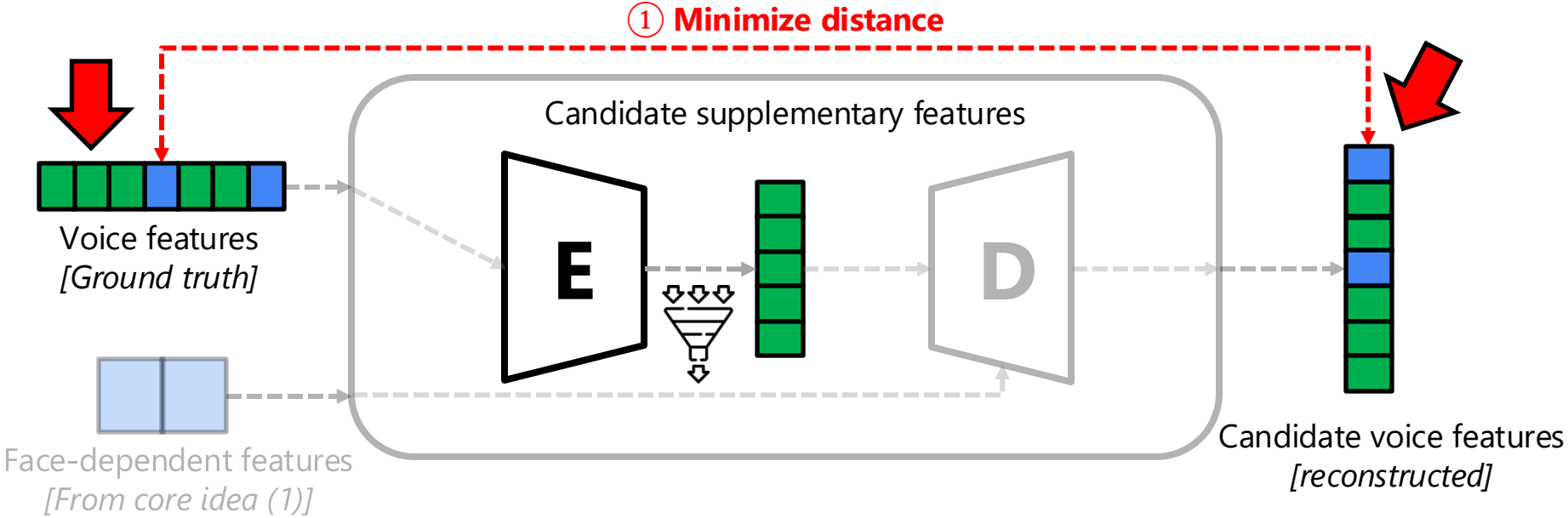
Training: Supplementary Feature Generation

Goal: To generate candidate supplementary features



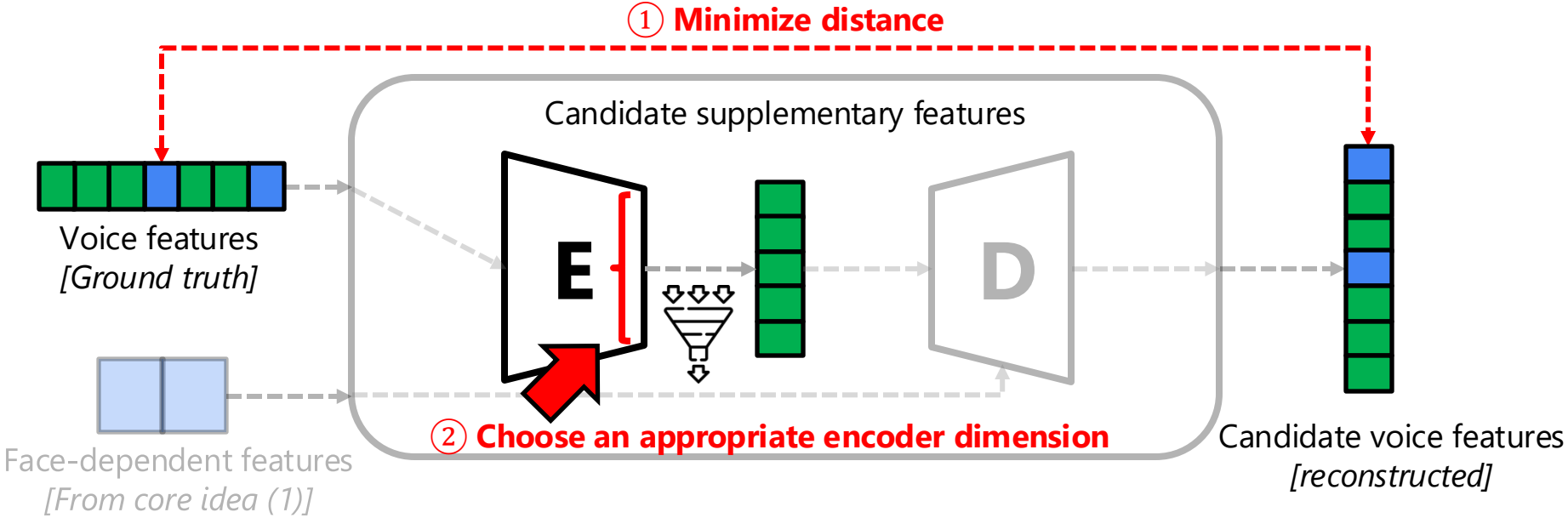
Training: Supplementary Feature Generation

Goal: To generate candidate supplementary features

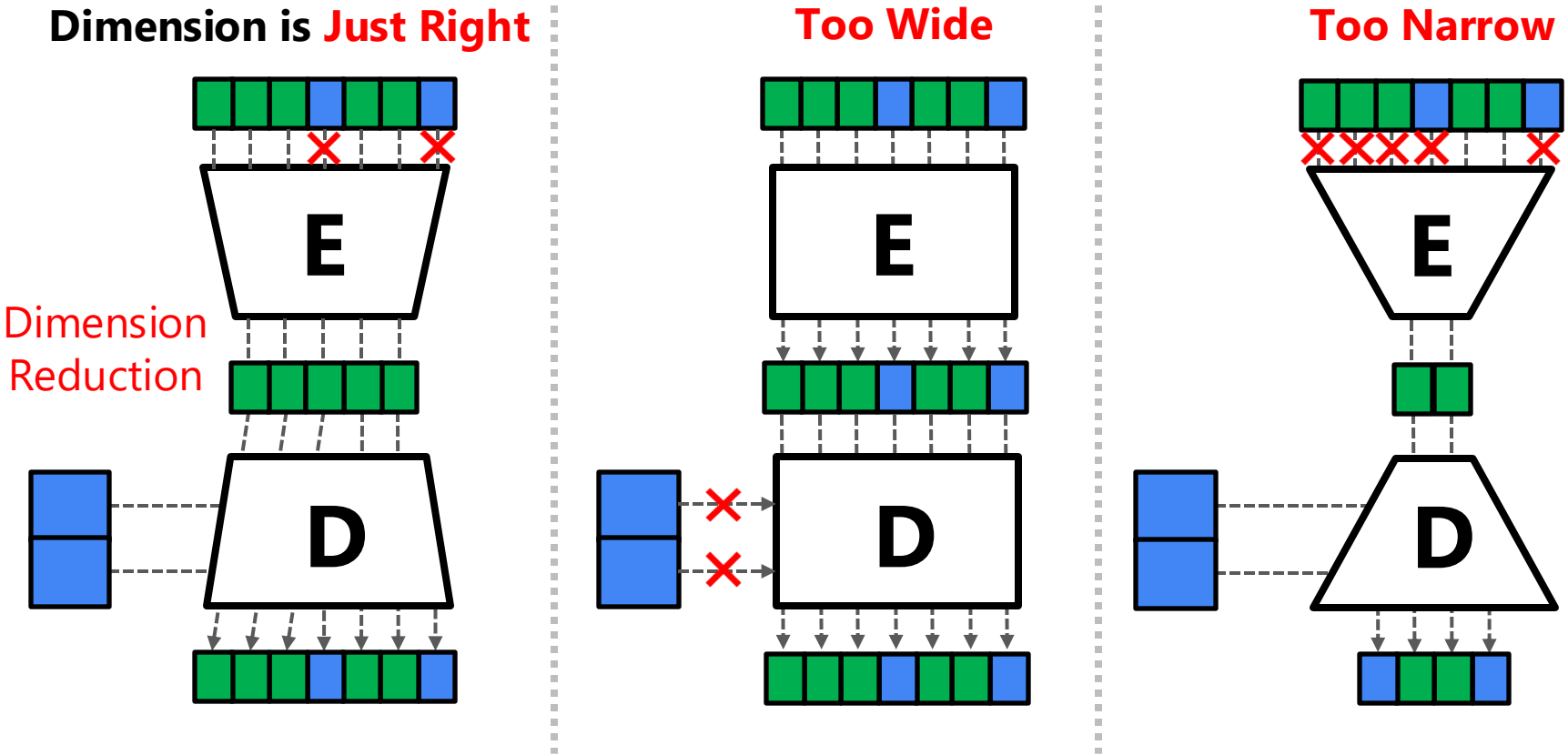


Training: Supplementary Feature Generation

Goal: To generate candidate supplementary features

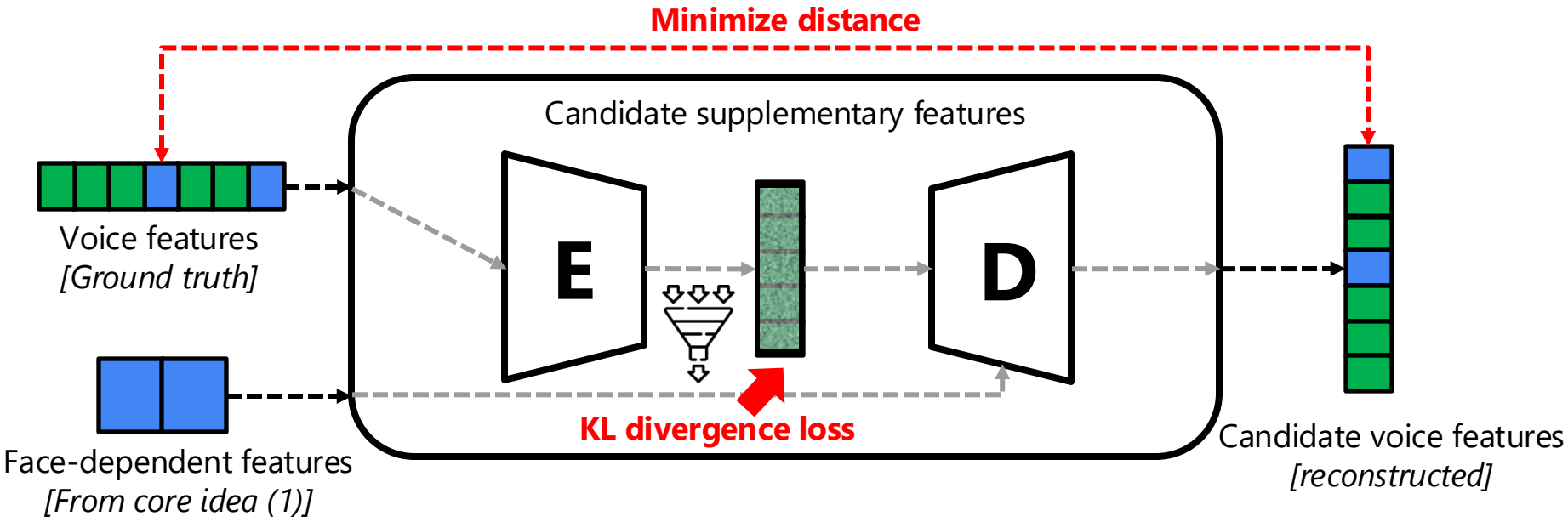


Training: Supplementary Feature Generation



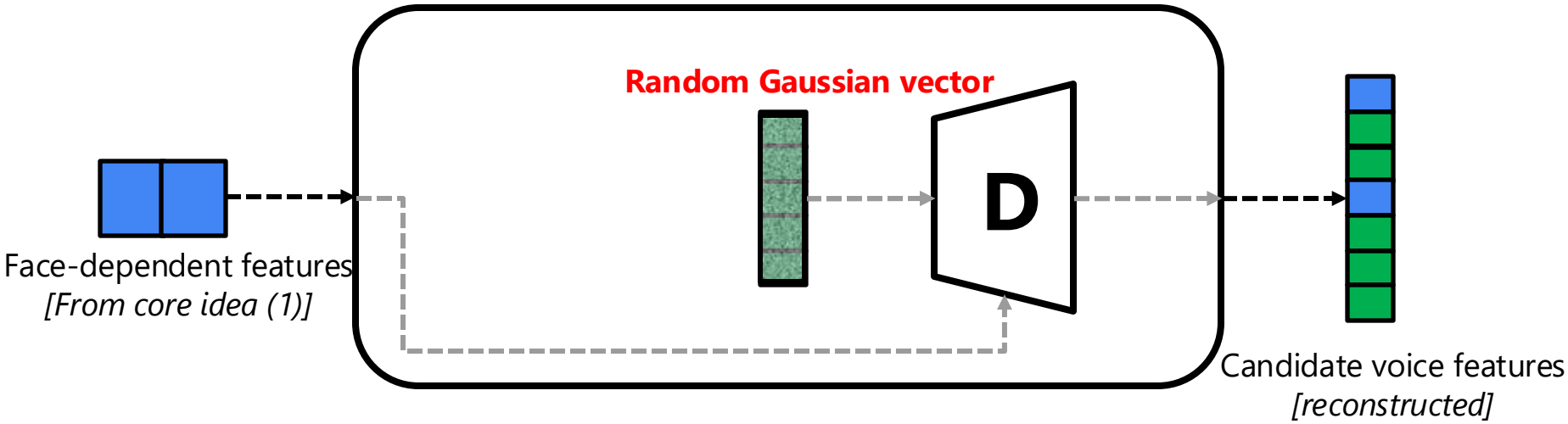
Training: Supplementary Feature Generation

Goal: To generate candidate supplementary features



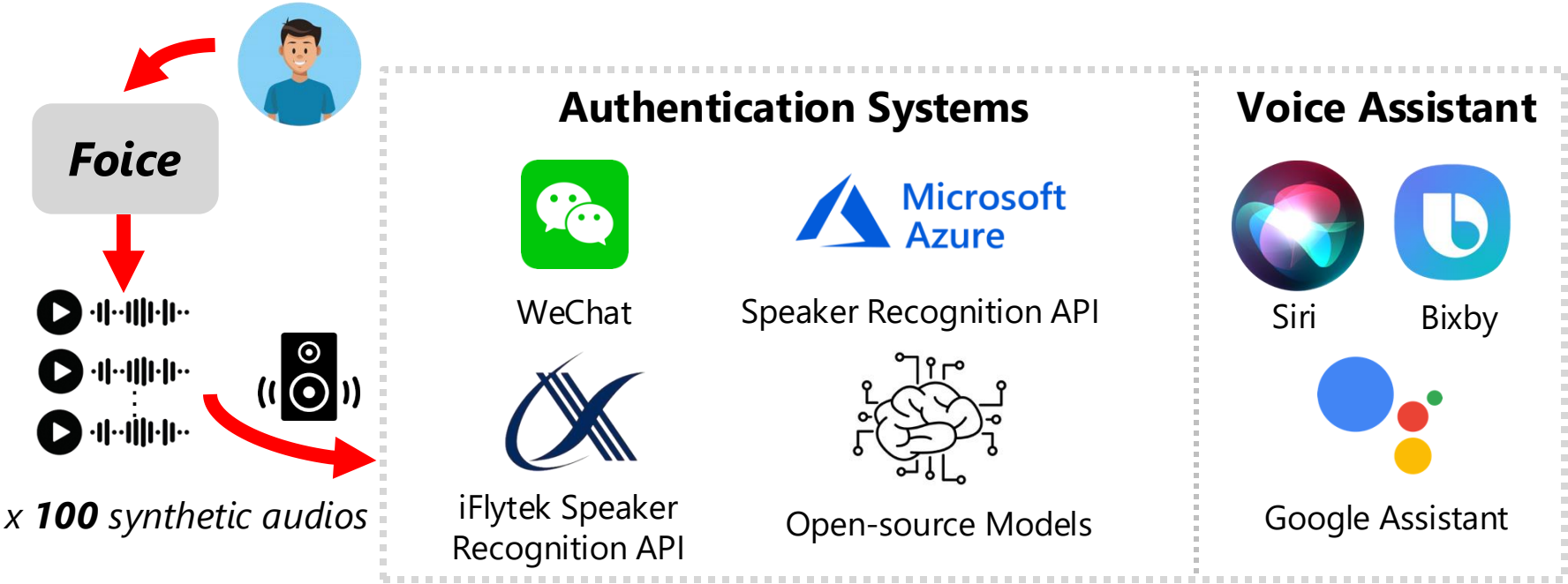
Supplementary Feature Generation

Testing Phase



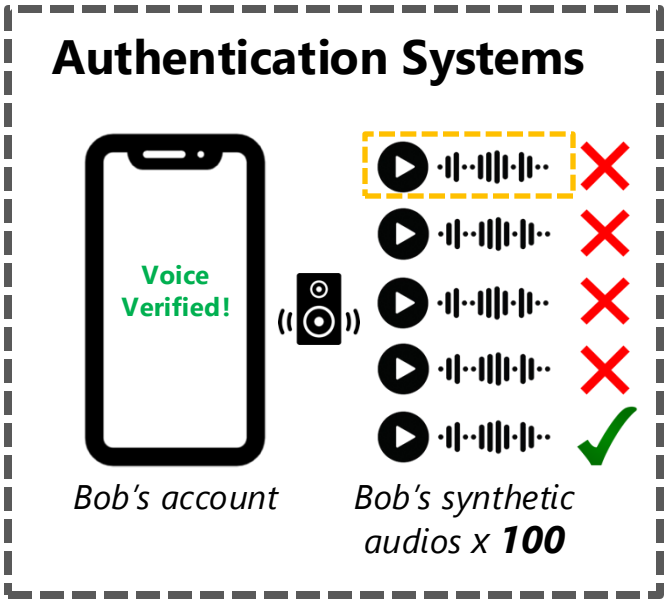
Evaluation Setup

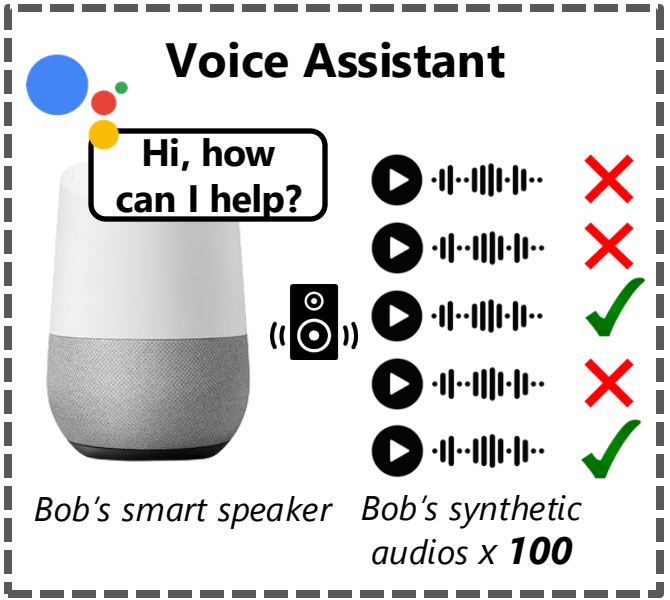
- Evaluate *Foice* with **authentication systems** and **voice assistants**
  - **100** candidate synthetic audios for each subject in the test dataset



Evaluation Setup

- For each subject, a successful attack is defined as:





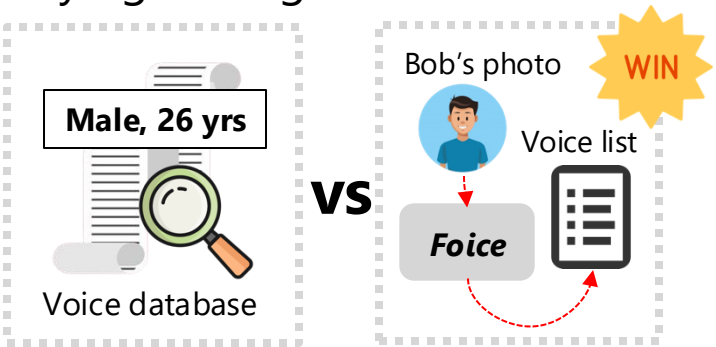
Summary of Results

- **All** tested systems vulnerable to *Foice* attack

- **Comparable** attack performance to voice deepfake attack

- Attack success rate improved by **more than three times** when combing face and voice

- *Foice* outperforming attacks using only age and gender information



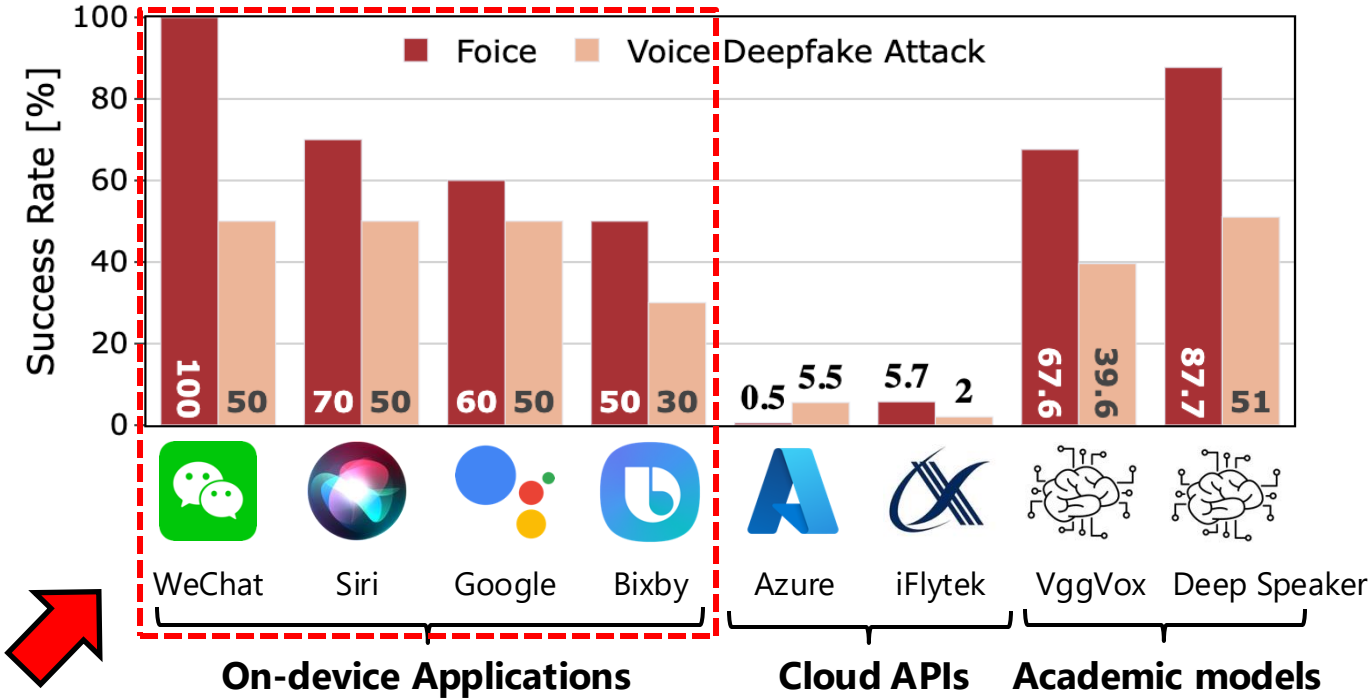
- *Foice* **robust** to various types of image conditions

Occlusion

Resolution

Main Result

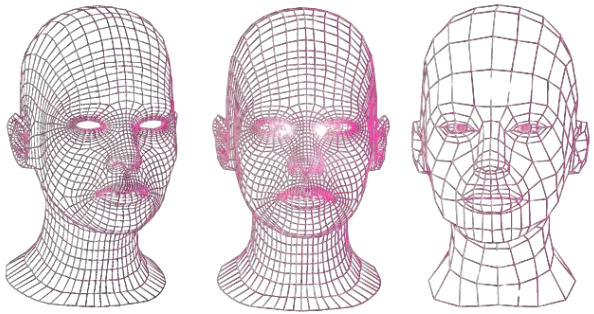
- Achieves **comparable** attack performance to the state-of-the-art voice deepfake attack



Discussion

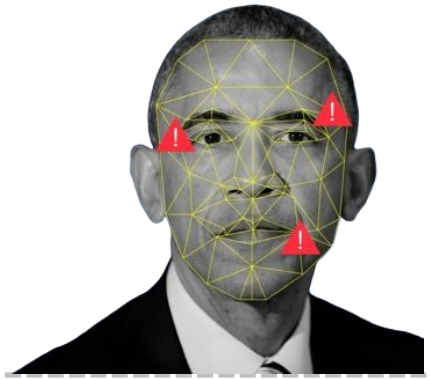
① Improving *Foice*

- **Dynamic** face provides more details on the facial bone structure
- Optimize model structure



② Countermeasures

- Deepfake detection and liveness detection
- Lack of adoption in real-world systems



Conclusion

- Synthesize **voice** recordings using a single **face** image
- Highlight the need to protect voice verification against deepfake attacks