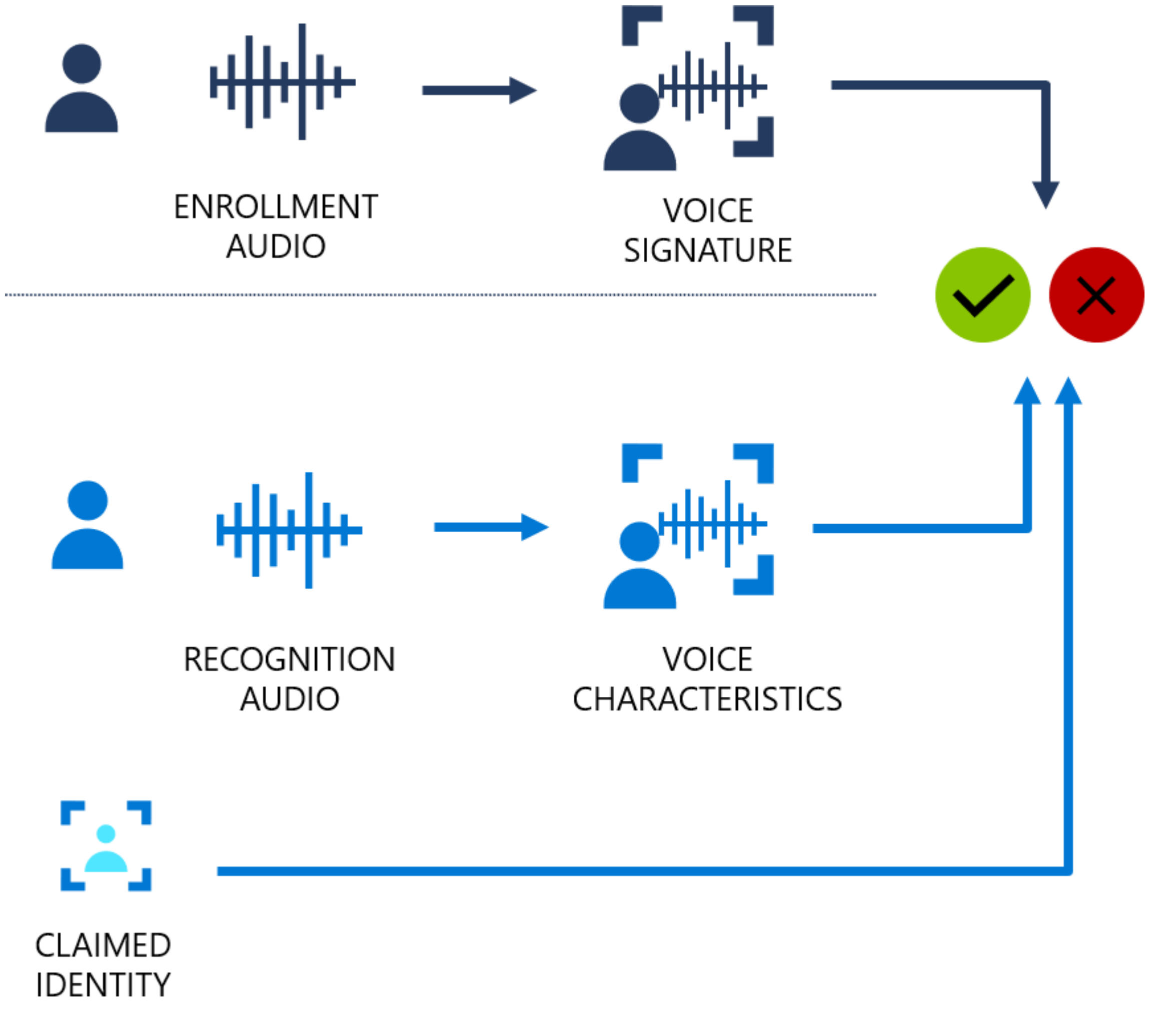
Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image

Jiang et al.
USENIX ‘24

Part 1

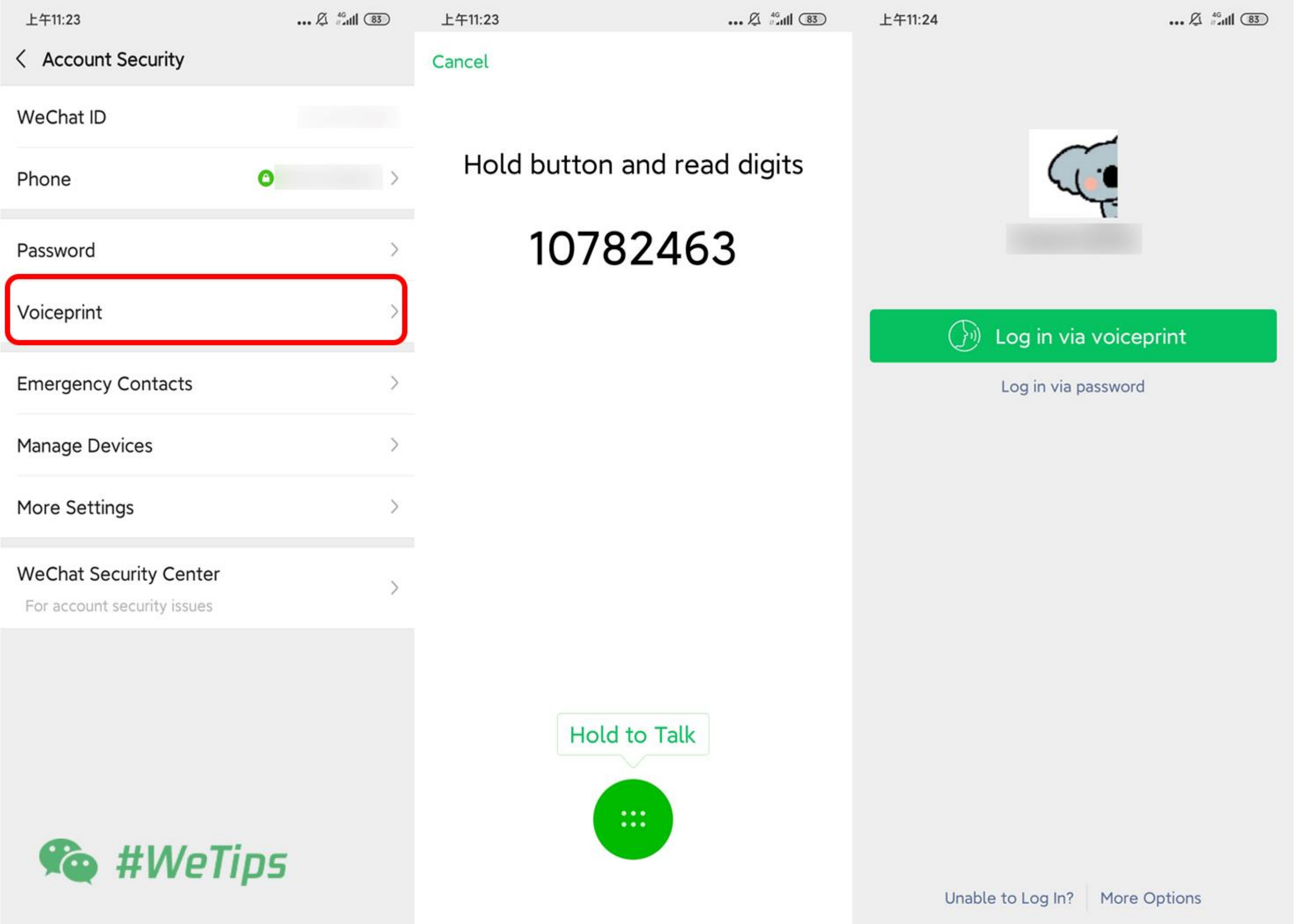
Introduction

Voice Authentication?



https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speaker-recognition-overview

Voice Authentication is widely used

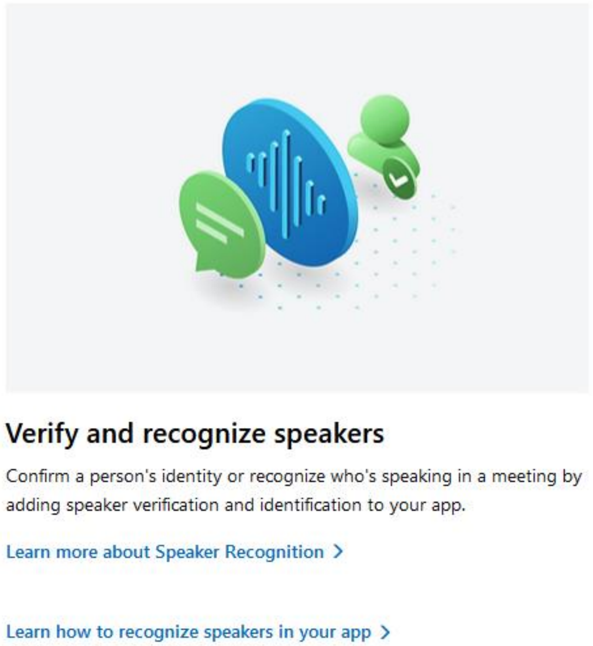


Hold button and read digits

10782463

https://www.facebook.com/wechatapp/posts/have-you-ever-tried-to-use-voiceprint-%EF%B8%8F-%EF%B8%8F-to-log-in-to-your-wechat-account-it-is/2598409853529264/

Voice Authentication is widely used

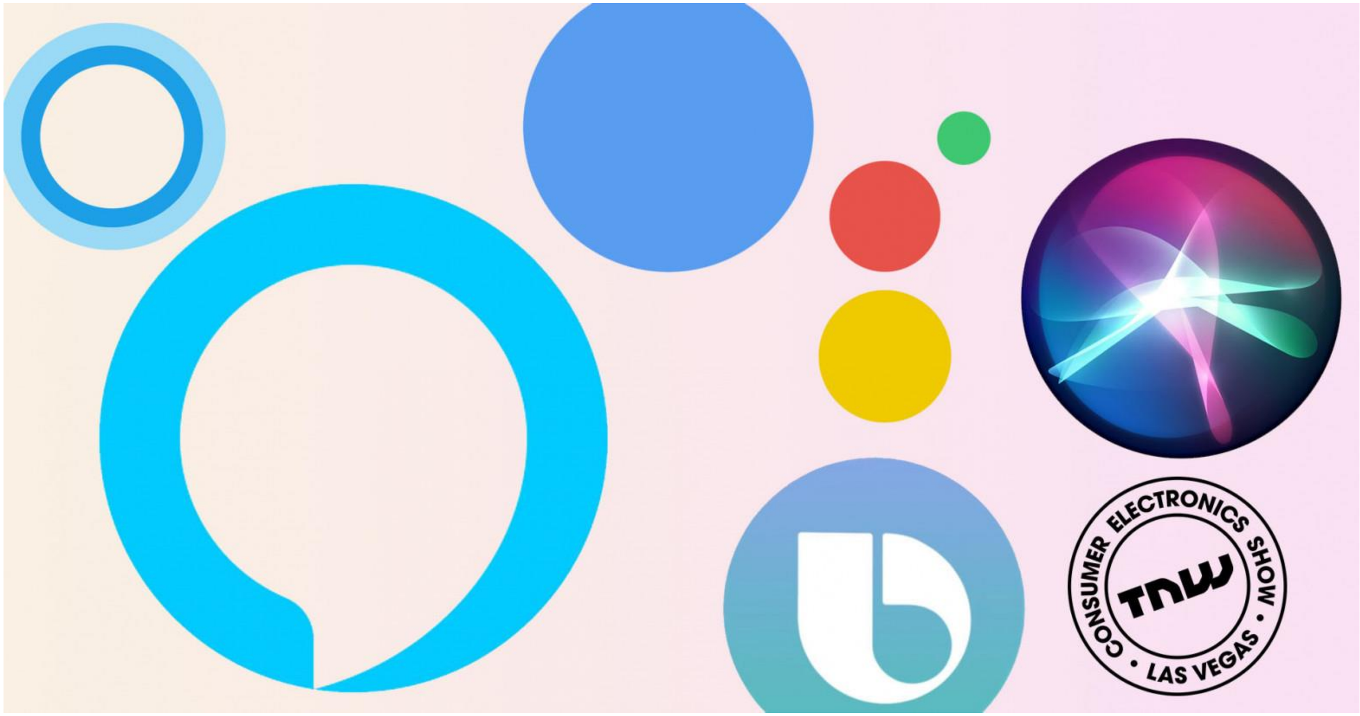


Verify and recognize speakers

Confirm a person's identity or recognize who's speaking in a meeting by adding speaker verification and identification to your app.

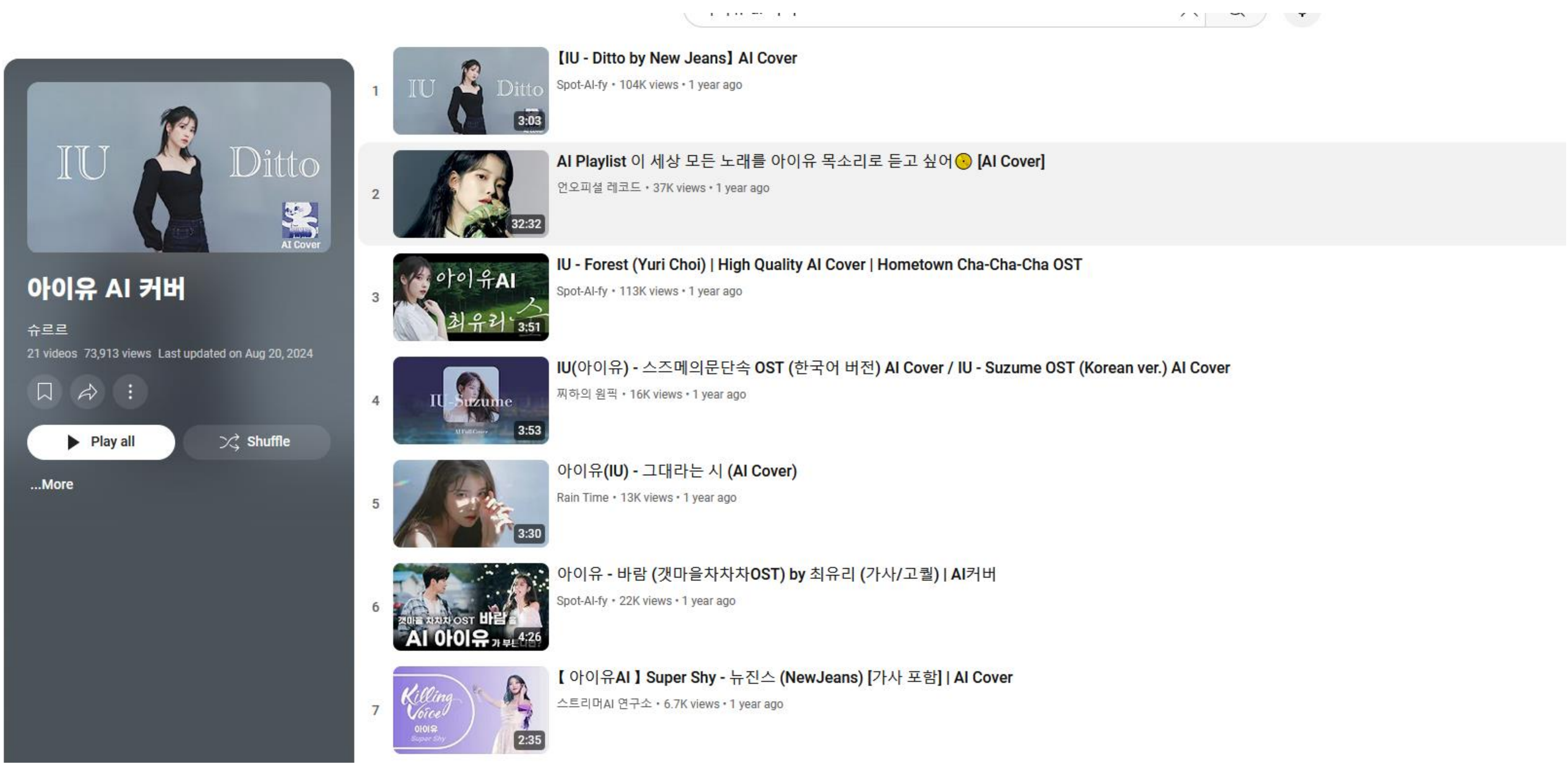
Learn more about Speaker Recognition >

Learn how to recognize speakers in your app >



https://thenextweb.com/news/its-about-time-alexa-and-google-assistant-learn-to-play-nice

Audio Deepfake



아이유 AI 커버

슈르르

21 videos 73,913 views Last updated on Aug 20, 2024

1 【IU - Ditto by New Jeans】 AI Cover
Spot-AI-fy • 104K views • 1 year ago
3:03

2 AI Playlist 이 세상 모든 노래를 아이유 목소리로 듣고 싶어💿 [AI Cover]
언오피셜 레코드 • 37K views • 1 year ago
32:32

3 IU - Forest (Yuri Choi) | High Quality AI Cover | Hometown Cha-Cha-Cha OST
Spot-AI-fy • 113K views • 1 year ago
3:51

4 IU(아이유) - 스즈메의문단속 OST (한국어 버전) AI Cover / IU - Suzume OST (Korean ver.) AI Cover
찌하의 원픽 • 16K views • 1 year ago
3:53

5 아이유(IU) - 그대라는 시 (AI Cover)
Rain Time • 13K views • 1 year ago
3:30

6 아이유 - 바람 (갯마을차차차OST) by 최유리 (가사/고퀄) | AI커버
Spot-AI-fy • 22K views • 1 year ago
4:26

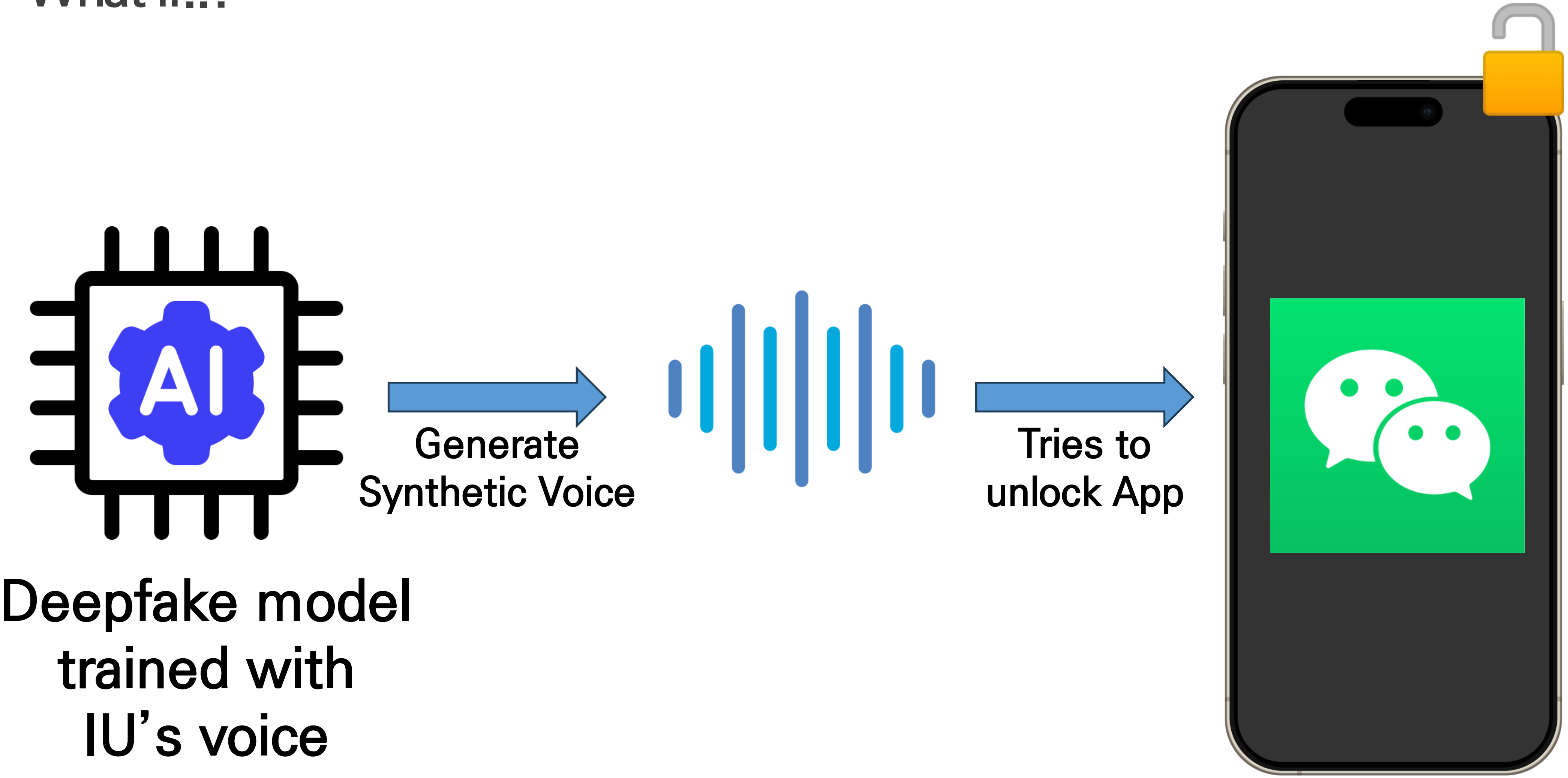
7 【 아이유AI 】 Super Shy - 뉴진스 (NewJeans) [가사 포함] | AI Cover
스트리머AI 연구소 • 6.7K views • 1 year ago
2:35

https://www.youtube.com/playlist?list=PLpBfg7XImIEgHyuStO0g3F7t8xnbnqv8C

Part 1

Introduction

What if..?

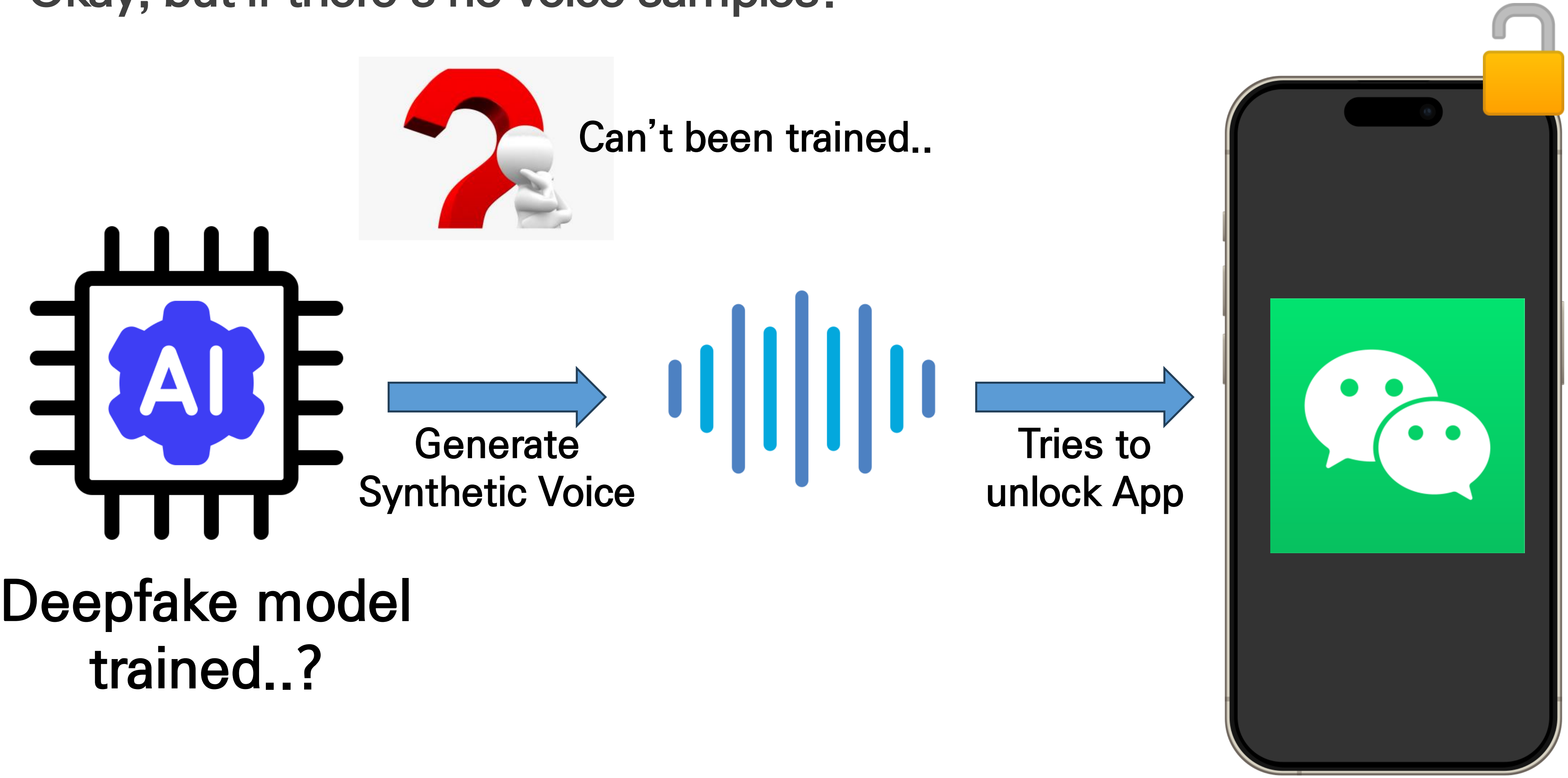


Generate Synthetic Voice

Tries to unlock App

Deepfake model trained with IU’s voice

Okay, but if there's no voice samples?



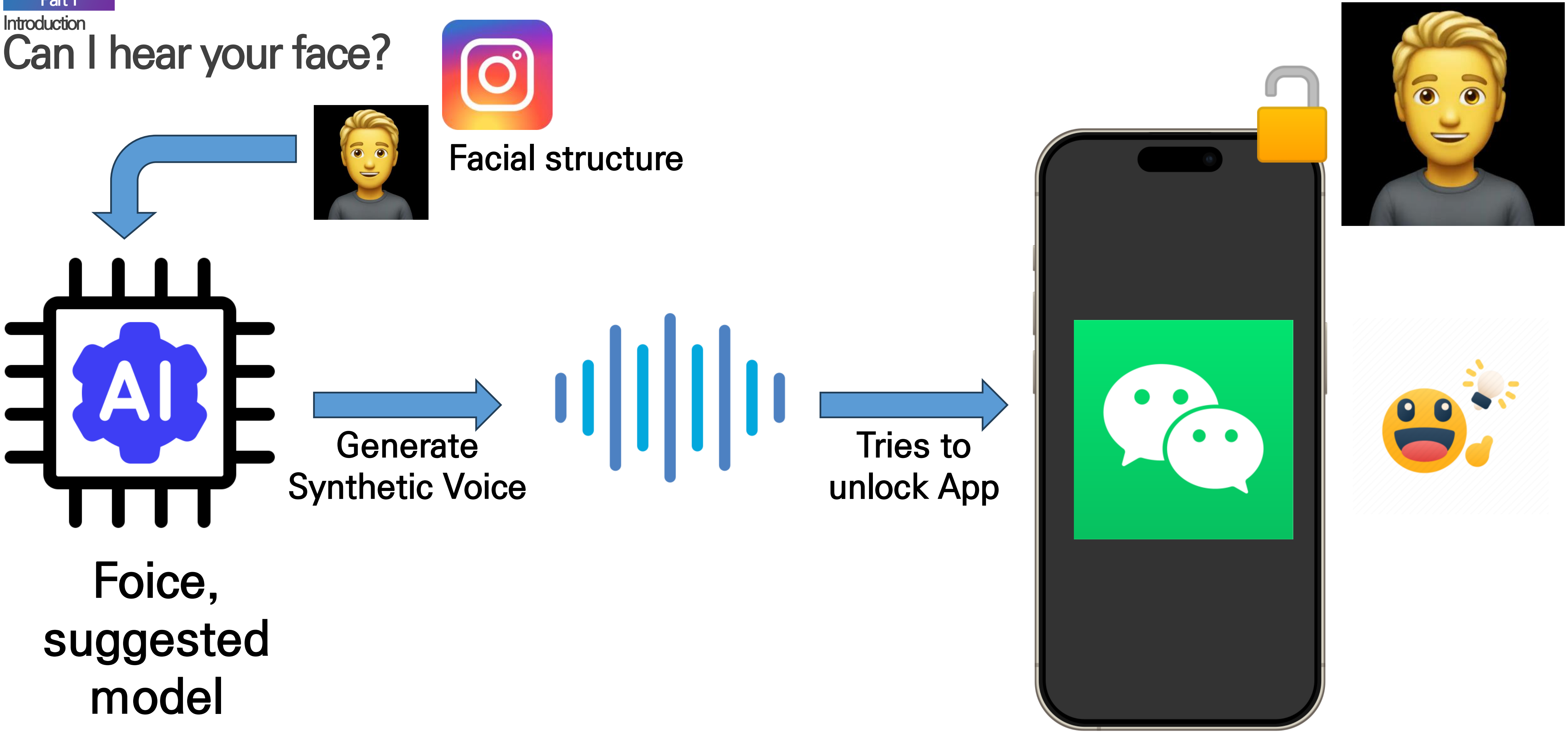
Can't been trained..

Generate Synthetic Voice

Tries to unlock App

Deepfake model trained..?

Can I hear your face?



Facial structure

Foice, suggested model

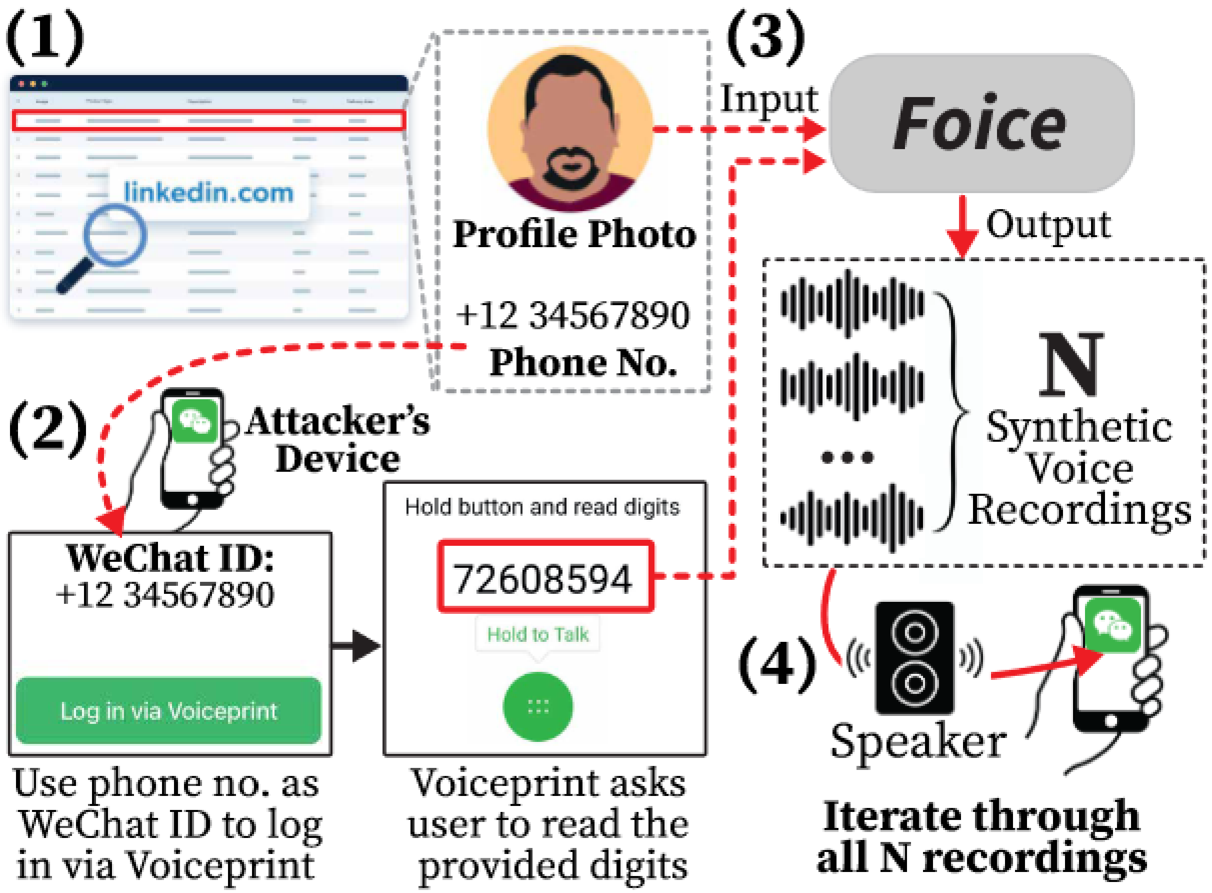
Generate Synthetic Voice

Tries to unlock App

Part 2

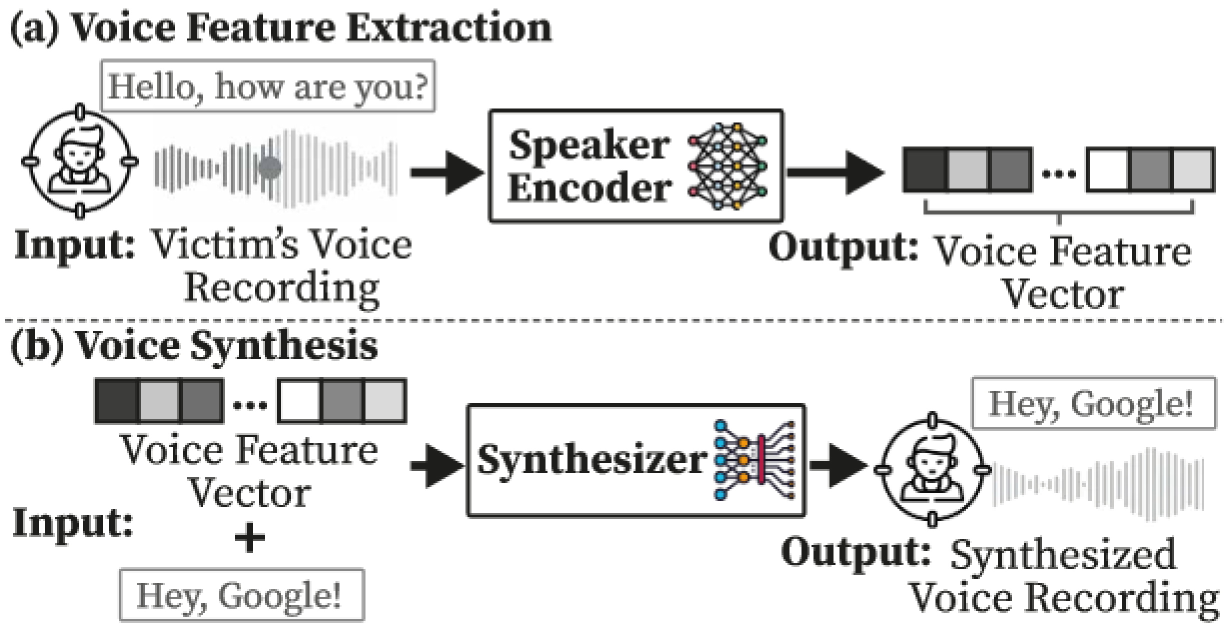
Paper Review

Brief Overview


(1)
linkedin.com
Profile Photo
+12 34567890
Phone No.
(3)
Input
Foice
Output
N
Synthetic Voice Recordings
(2)
Attacker's Device
WeChat ID:
+12 34567890
Log in via Voiceprint
Use phone no. as WeChat ID to log in via Voiceprint
Hold button and read digits
72608594
Hold to Talk
Voiceprint asks user to read the provided digits
(4)
Speaker
Iterate through all N recordings

Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

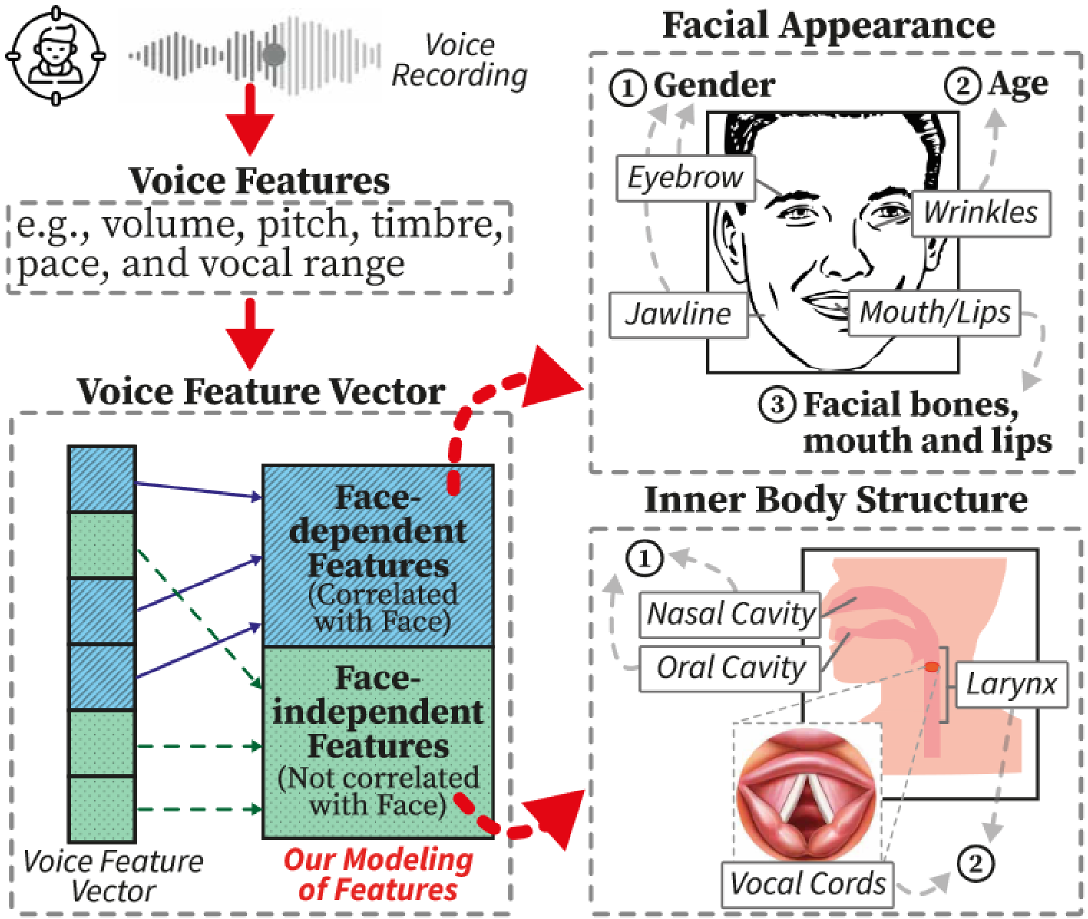
How to generate synthetic voice?


(a) Voice Feature Extraction
Hello, how are you?
Speaker Encoder
Input: Victim's Voice Recording
Output: Voice Feature Vector
(b) Voice Synthesis
Voice Feature Vector
Input:
+
Hey, Google!
Synthesizer
Hey, Google!
Output: Synthesized Voice Recording

Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

How to generate synthetic voice?

English	Korean	Chinese
Nasal cavity	비강	鼻腔
Oral cavity	구강	口腔
Vocal cords	성대	声带
Larynx	후두	喉



Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

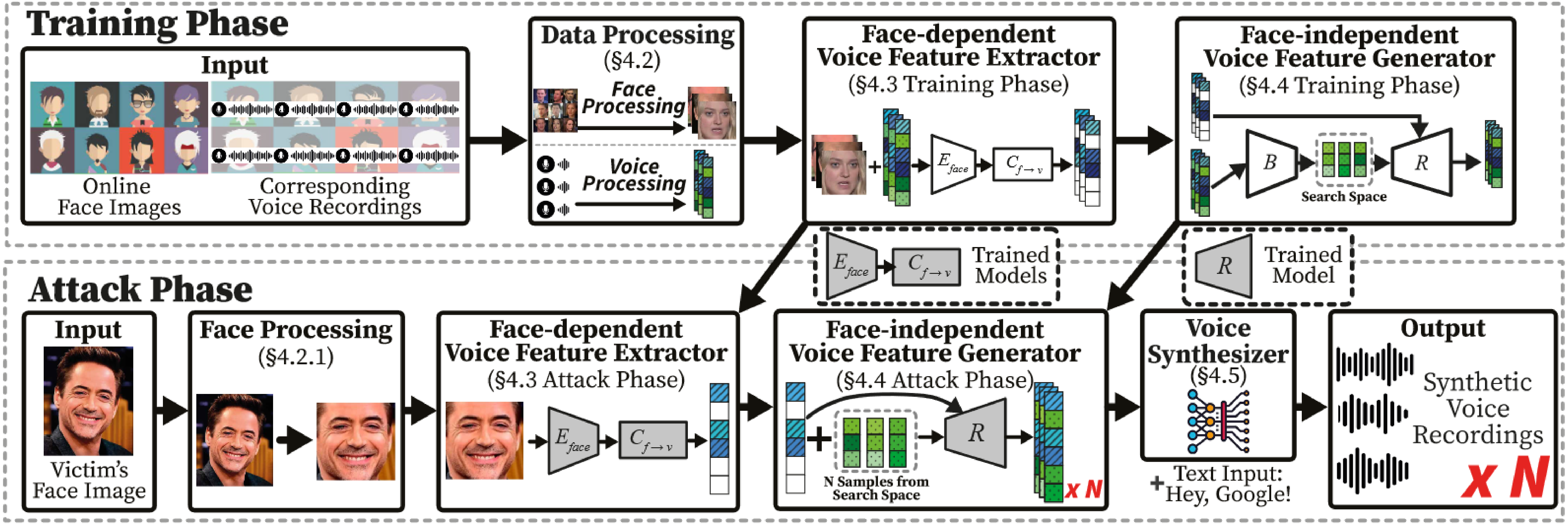
Few things to consider

- Voice Authentication
  – Enrollment Phase(i.e. system displays ###)
  – Authentication Phase(i.e. Speak displayed #)
- System set low threshold to ensure user can authenticate in noisy environments.
- Many voice authentication system do not restrict the number of authentication attempts.

Part 2

Paper Review

System Design

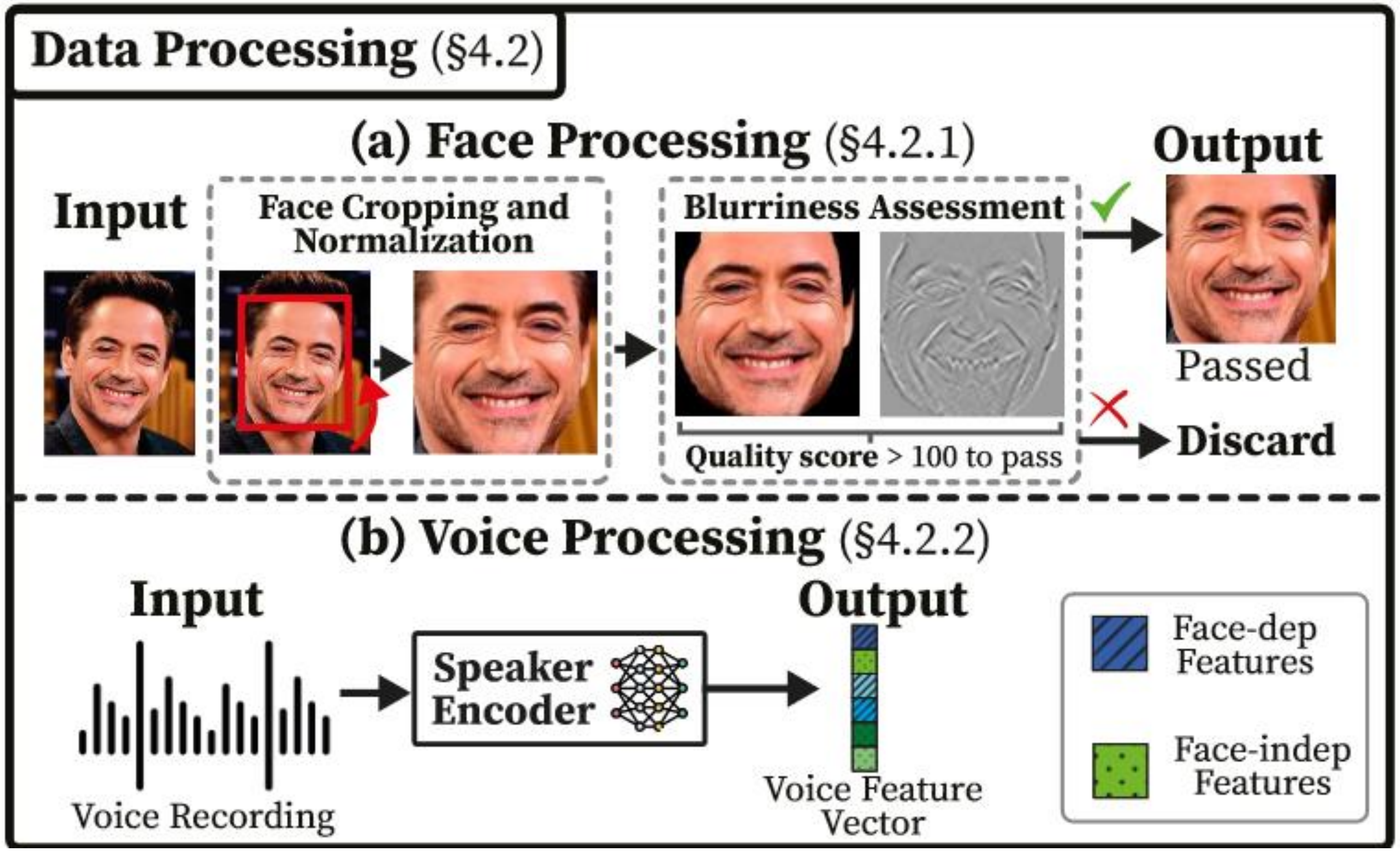

Training Phase
Input
Online Face Images
Corresponding Voice Recordings
Data Processing (§4.2)
Face Processing
Voice Processing
Face-dependent Voice Feature Extractor (§4.3 Training Phase)
E_{face}
$C_{f\rightarrow v}$
Face-independent Voice Feature Generator (§4.4 Training Phase)
B
R
Search Space
Trained Models
Trained Model
Attack Phase
Input
Victim's Face Image
Face Processing (§4.2.1)
Face-dependent Voice Feature Extractor (§4.3 Attack Phase)
Face-independent Voice Feature Generator (§4.4 Attack Phase)
N Samples from Search Space
x N
Voice Synthesizer (§4.5)
+ Text Input: Hey, Google!
Output
Synthetic Voice Recordings
x N

Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

Part 2

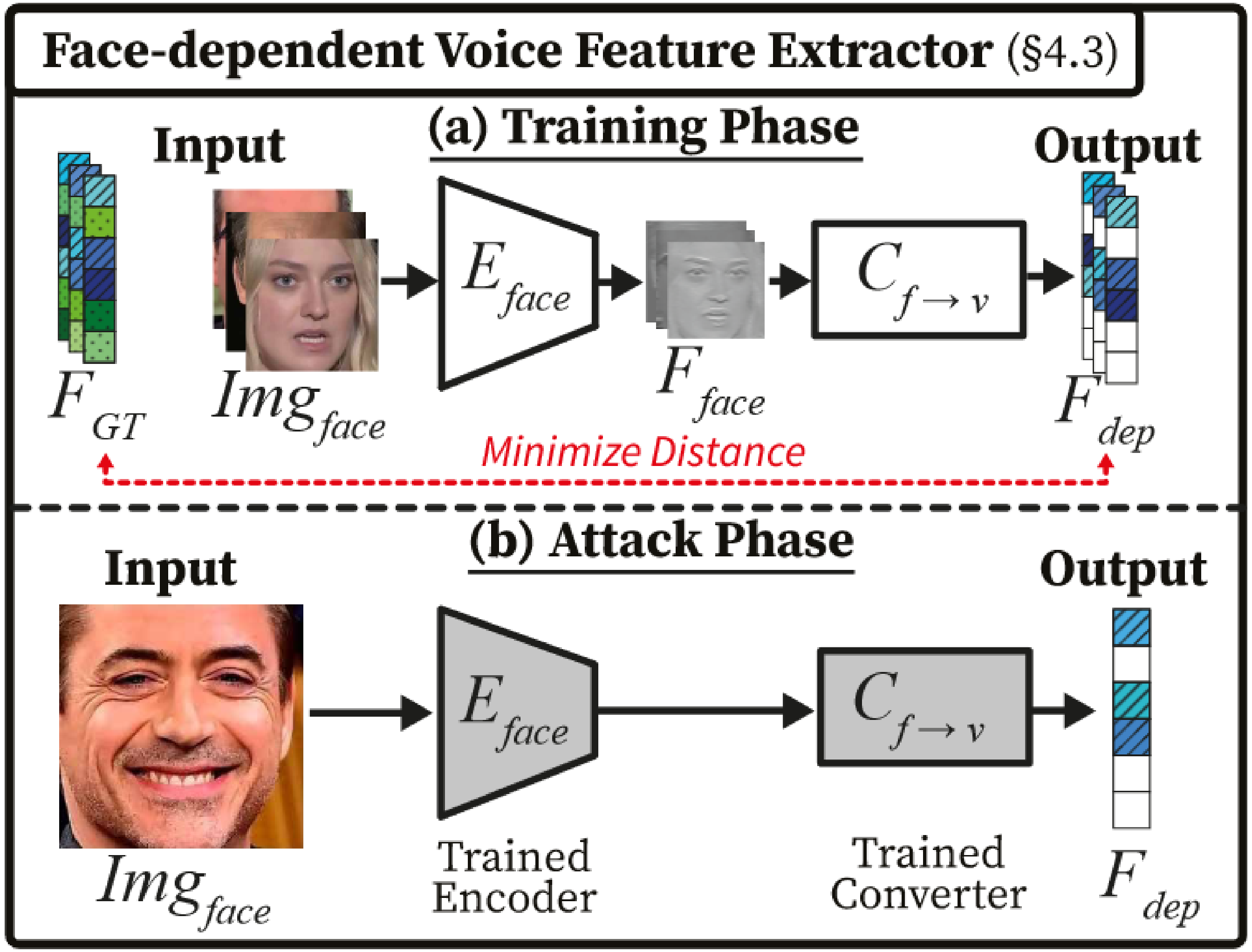
Paper Review

Face & Voice Processing



Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

Face−Dependent Feature Generator



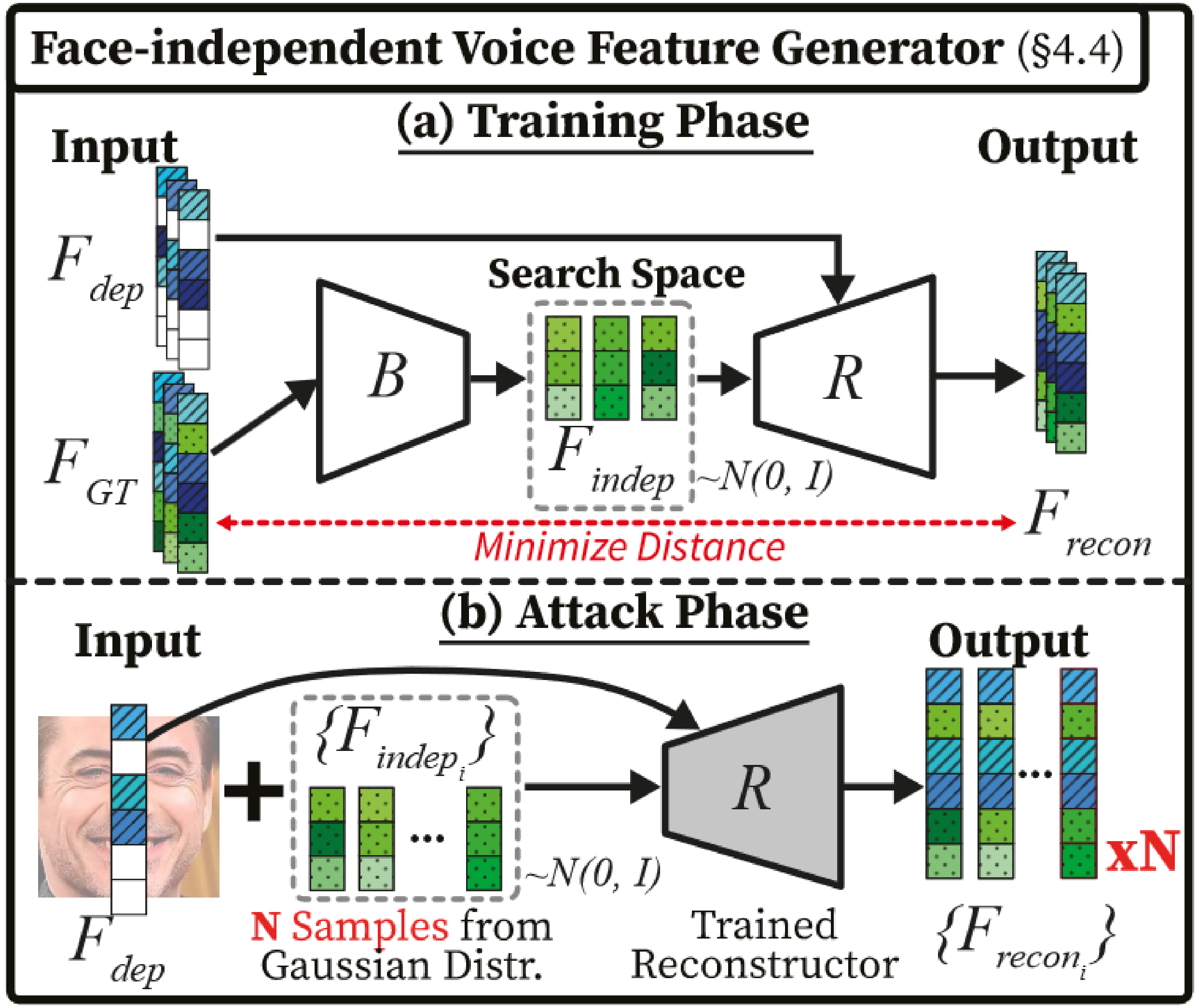
- F: Feature
  - GT: Ground Truth
  - dep: Face−Dependent
- E: Encoder
  - face: Encoder trained with facial image
- C: Converter
  - f−>v: facial feature to F_dep

$$F_{face} = E_{face}(Img_{face}), \qquad F_{dep} = C_{f\rightarrow v}(F_{face}),$$

$$\min_{E_{face}(\cdot), C_{f\rightarrow v}(\cdot)} Err(F_{dep}, F_{GT})$$

Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

Face−Independent Feature Generator



- F: Feature
  − GT: Ground Truth
  − dep: Face−Dependent
  − recon: reconstructed
  − indep: Face−Independent
- B: Bottleneck
  − Project F to smaller indep. search space, tries F to follow gaussian distribution
- R: Reconstructor
  − Search space to reconstructed F

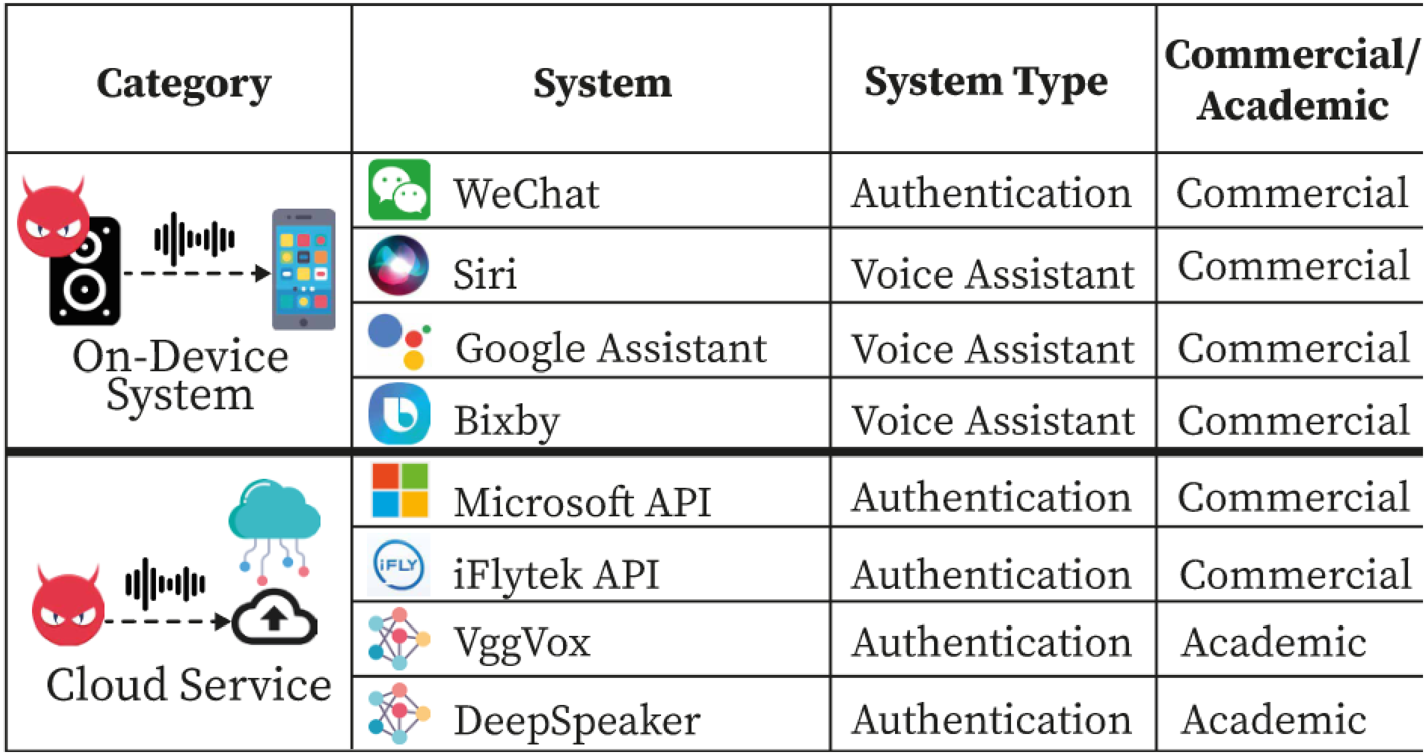
$$F_{indep} = B(F_{GT}), \qquad F_{recon} = R(F_{indep}, F_{dep})$$

$$\min_{B(\cdot),R(\cdot,\cdot)} Err(F_{GT}, F_{recon}) + KL[P_{F_{indep}}(\cdot) \,\|\, \mathcal{N}(0,I)],$$

Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX'24

Experimental Setup

- Voice Authentication Systems
  - On-device: Commercial System installed on smartphones
  - Cloud: Microsoft, iFlytek, VGGVox, DeepSpeaker



Category	System	System Type	Commercial/ Academic
On-Device System	WeChat	Authentication	Commercial
	Siri	Voice Assistant	Commercial
	Google Assistant	Voice Assistant	Commercial
	Bixby	Voice Assistant	Commercial
Cloud Service	Microsoft API	Authentication	Commercial
	iFlytek API	Authentication	Commercial
	VggVox	Authentication	Academic
	DeepSpeaker	Authentication	Academic

Experimental Setup

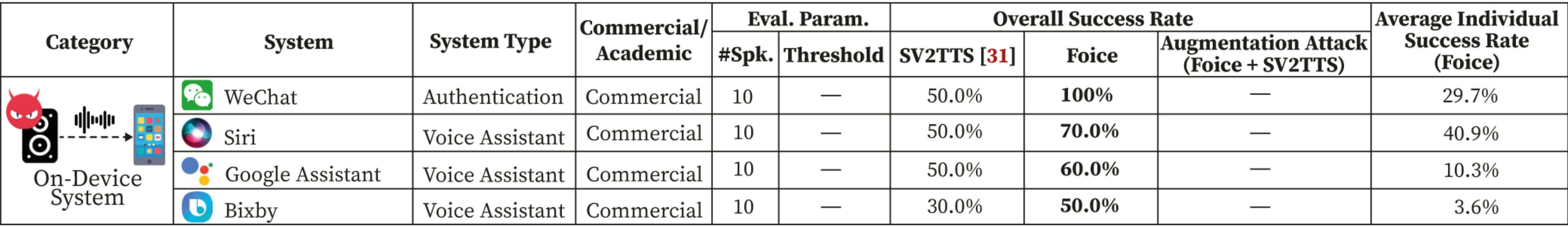
- Benchmark Voice Deepfake System:
  – SV2TTS: SoTA TTS System can produce voice of the un–seen speaker in training phase “naturally”.
- Speaker Dataset:
  – VoxCeleb1(100K Videos, 1251 celebs) – For evaluation
  – VoxCeleb2(1M Videos, 6112 celebs) – For Training
  – 10 Participants recorded in quiet environment.
- Performance Metrics
  – Overall Success Rate: Percentage of speakers with at least one successful voice attack
  – Individual SR: Percentage of successful synthetic voice attacks for specific person
  – Foice Individual SR: Fraction of voice cloned from single face image passes verification

RQs

- Is Foice attack effective against diverse modern implementations of speaker authentication systems and voice assistants?
- Can Foice provide more voice information other than age and gender?
- Can we combine Foice and the existing voice deepfake system to improve the attack's effectiveness?
- How do different experimental conditions affect the effectiveness of Foice?

RQ1: Is Foice attack effective against diverse modern implementations of speaker authentication systems and voice assistants?

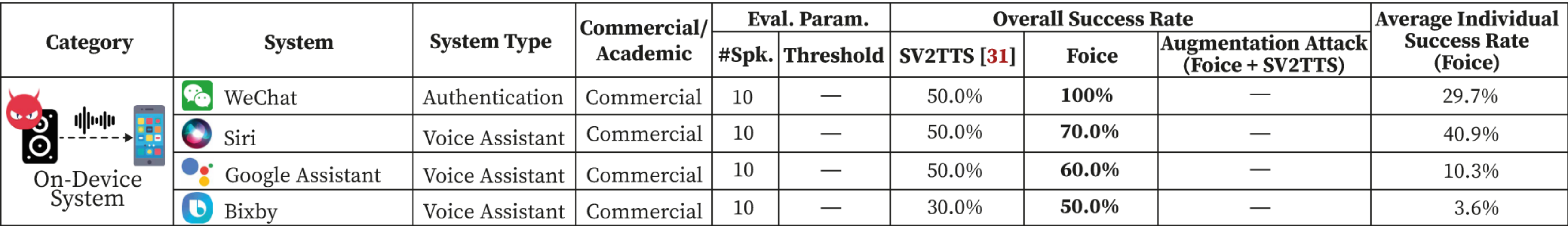
- Method
  - Used custom dataset(10)
  - Foice: 100 synthetic voice recordings per participants
  - SV2TTS: 1 synthetic voice recording per participant
  - Laptop Speaker played synthetic recording to cell phone



Category	System	System Type	Commercial/ Academic	Eval. Param.		Overall Success Rate			Average Individual Success Rate (Foice)
				#Spk.	Threshold	SV2TTS [31]	Foice	Augmentation Attack (Foice + SV2TTS)	
On-Device System	WeChat	Authentication	Commercial	10	—	50.0%	**100%**	—	29.7%
	Siri	Voice Assistant	Commercial	10	—	50.0%	**70.0%**	—	40.9%
	Google Assistant	Voice Assistant	Commercial	10	—	50.0%	**60.0%**	—	10.3%
	Bixby	Voice Assistant	Commercial	10	—	30.0%	**50.0%**	—	3.6%

RQ1: Is Foice attack effective against diverse modern implementations of speaker authentication systems and voice assistants?: On-device

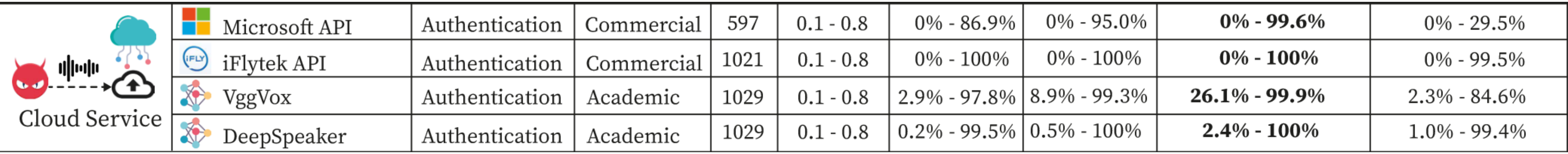
- Method
  - Used custom dataset(10)
  - Foice: 100 synthetic voice recordings per participants w\ img
  - SV2TTS: 1 synthetic voice recording per participant
  - Laptop Speaker played synthetic recording to cell phone
- Analysis
  - SV2TTS struggles, input voice is not noise-free(i.e. echo).



Category	System	System Type	Commercial/ Academic	Eval. Param.		Overall Success Rate			Average Individual Success Rate (Foice)
				#Spk.	Threshold	SV2TTS [31]	Foice	Augmentation Attack (Foice + SV2TTS)	
On-Device System	WeChat	Authentication	Commercial	10	—	50.0%	**100%**	—	29.7%
	Siri	Voice Assistant	Commercial	10	—	50.0%	**70.0%**	—	40.9%
	Google Assistant	Voice Assistant	Commercial	10	—	50.0%	**60.0%**	—	10.3%
	Bixby	Voice Assistant	Commercial	10	—	30.0%	**50.0%**	—	3.6%

RQ1: Is Foice attack effective against diverse modern implementations of speaker authentication systems and voice assistants?: Cloud

- Method
  - Used VoxCeleb1 dataset(1029)
  - Foice: Synthetic voice recordings w\ celeb's img
  - SV2TTS: Synthetic voice recording from new voice recording
- Analysis
  - SV2TTS ~= Foice to commercial APIs
  - SV2TTS 〈 Foice on Academic models

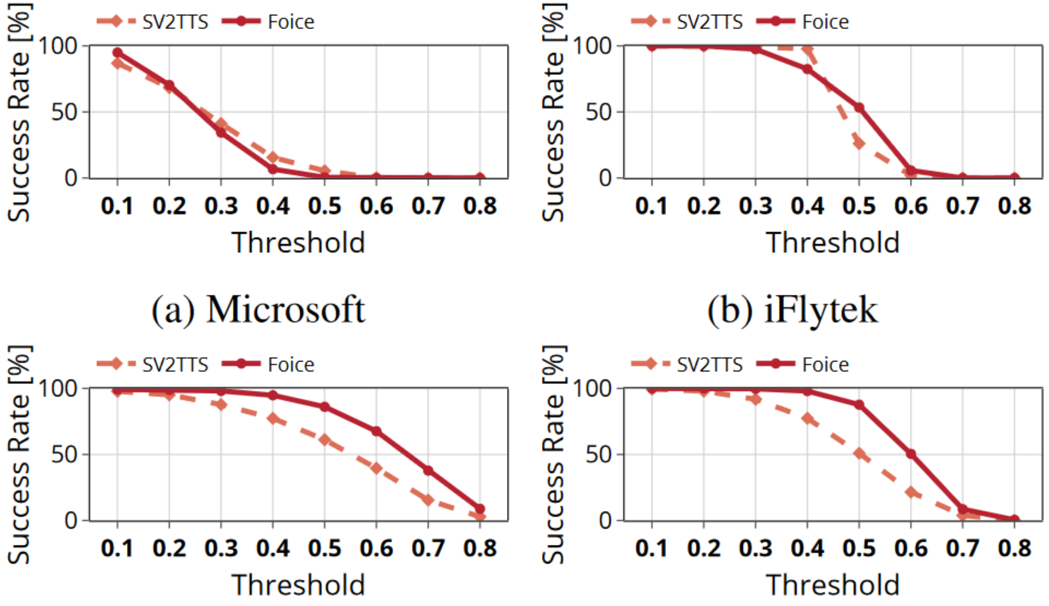


Cloud Service	Microsoft API	Authentication	Commercial	597	0.1 - 0.8	0% - 86.9%	0% - 95.0%	**0% - 99.6%**	0% - 29.5%
	iFlytek API	Authentication	Commercial	1021	0.1 - 0.8	0% - 100%	0% - 100%	**0% - 100%**	0% - 99.5%
	VggVox	Authentication	Academic	1029	0.1 - 0.8	2.9% - 97.8%	8.9% - 99.3%	**26.1% - 99.9%**	2.3% - 84.6%
	DeepSpeaker	Authentication	Academic	1029	0.1 - 0.8	0.2% - 99.5%	0.5% - 100%	**2.4% - 100%**	1.0% - 99.4%

Part 2

Paper Review

RQ1: Is Foice attack effective against diverse modern implementations of speaker authentication systems and voice assistants?: Cloud

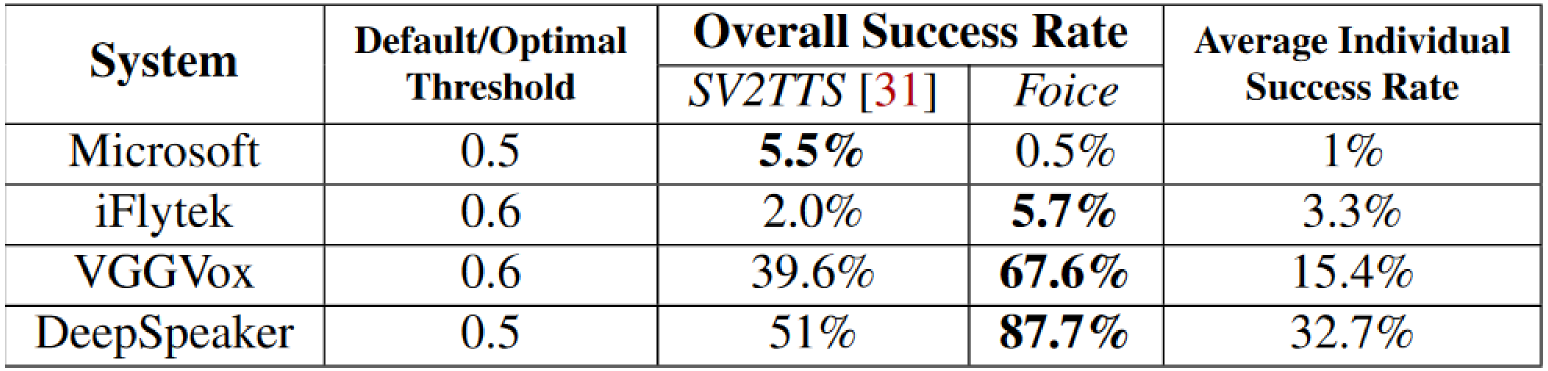


(a) Microsoft

(b) iFlytek

(c) VGGVox

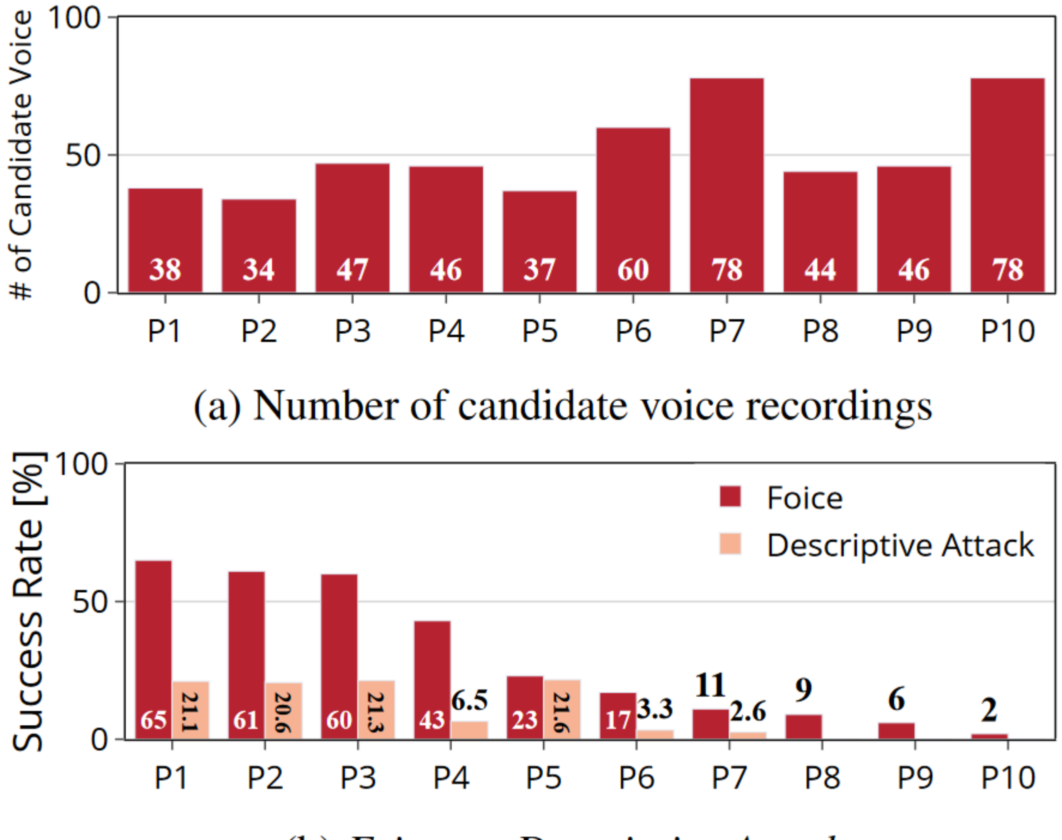
(d) DeepSpeaker



System	Default/Optimal Threshold	Overall Success Rate		Average Individual Success Rate
		SV2TTS [31]	*Foice*	
Microsoft	0.5	**5.5%**	0.5%	1%
iFlytek	0.6	2.0%	**5.7%**	3.3%
VGGVox	0.6	39.6%	**67.6%**	15.4%
DeepSpeaker	0.5	51%	**87.7%**	32.7%

RQ2: Can Foice provide more voice information other than age and gender?

- Method: Target system: WeChat
  – P1..10: Cluster VoxCeleb1(1029) to 10, using age and gender

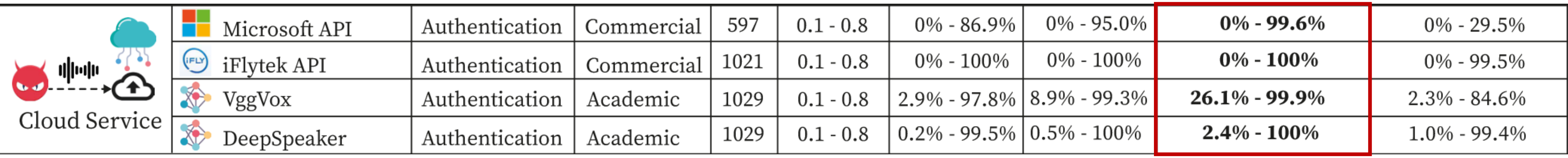


(a) Number of candidate voice recordings

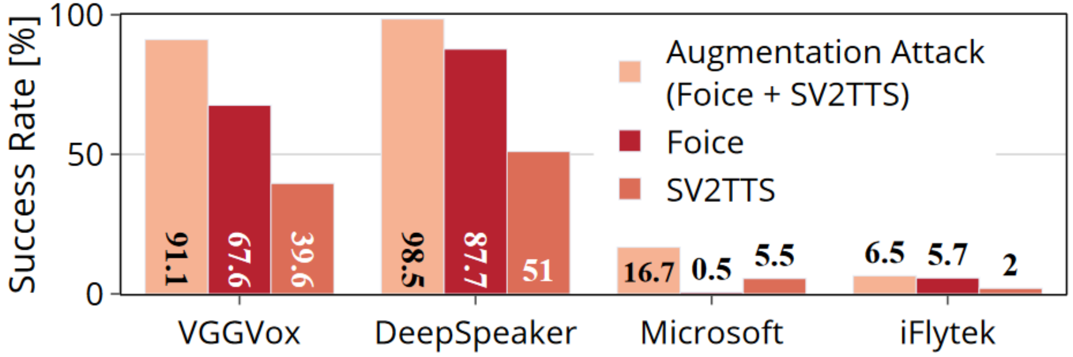
(b) *Foice* vs. *Descriptive Attack*

RQ3: Can we combine Foice and the existing voice deepfake system to improve the attack's effectiveness?

- Method:
  - Average 100 voice feature vector of SV2TTS and Foice
  - Generate 100 synthetic voice from averaged vector.

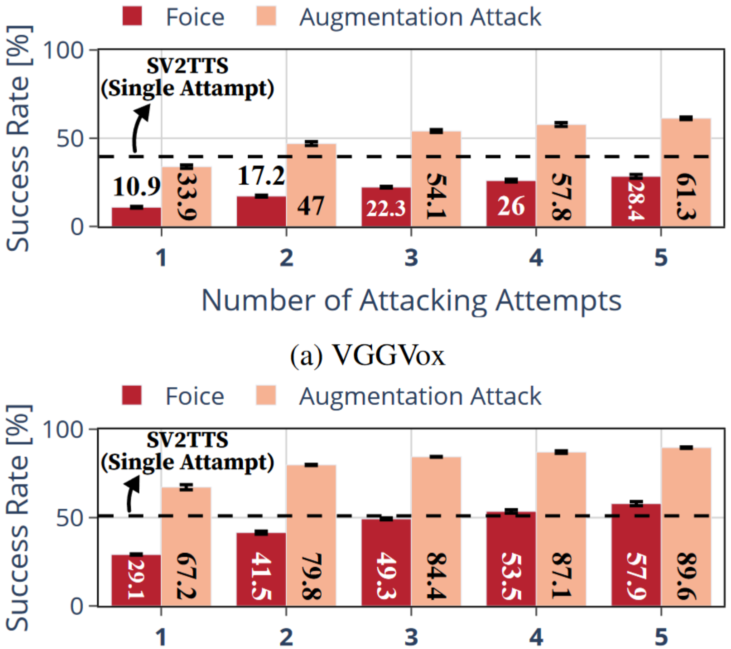


Cloud Service	Microsoft API	Authentication	Commercial	597	0.1 - 0.8	0% - 86.9%	0% - 95.0%	**0% - 99.6%**	0% - 29.5%
	iFlytek API	Authentication	Commercial	1021	0.1 - 0.8	0% - 100%	0% - 100%	**0% - 100%**	0% - 99.5%
	VggVox	Authentication	Academic	1029	0.1 - 0.8	2.9% - 97.8%	8.9% - 99.3%	**26.1% - 99.9%**	2.3% - 84.6%
	DeepSpeaker	Authentication	Academic	1029	0.1 - 0.8	0.2% - 99.5%	0.5% - 100%	**2.4% - 100%**	1.0% - 99.4%



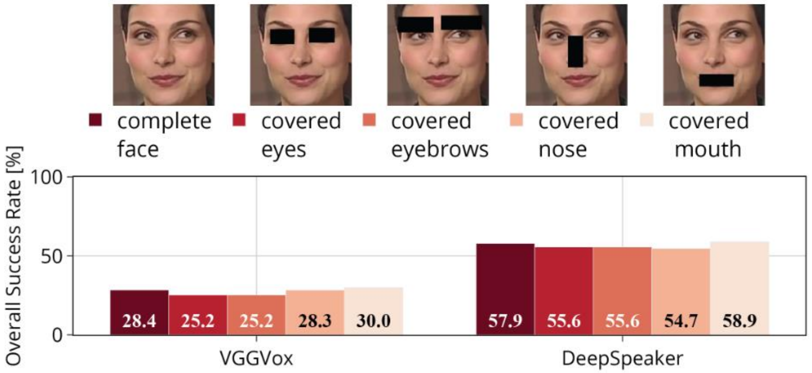
RQ4: How do different experimental conditions affect the effectiveness of Foice?

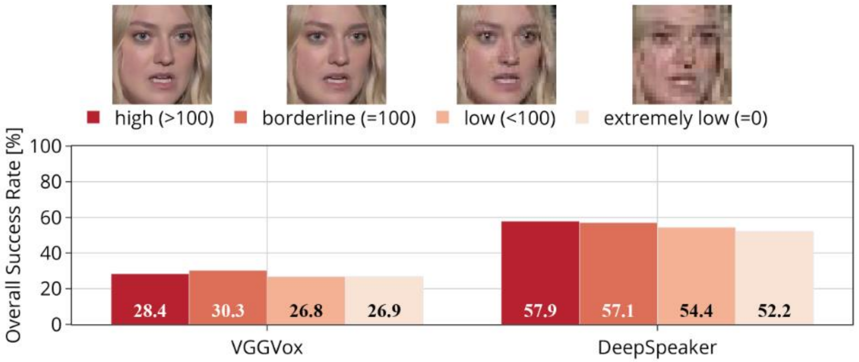
- Varying number of attacking attempts, image occlusion, image resolution



(a) VGGVox

(b) DeepSpeaker





Attacking attempts

Image occlusion

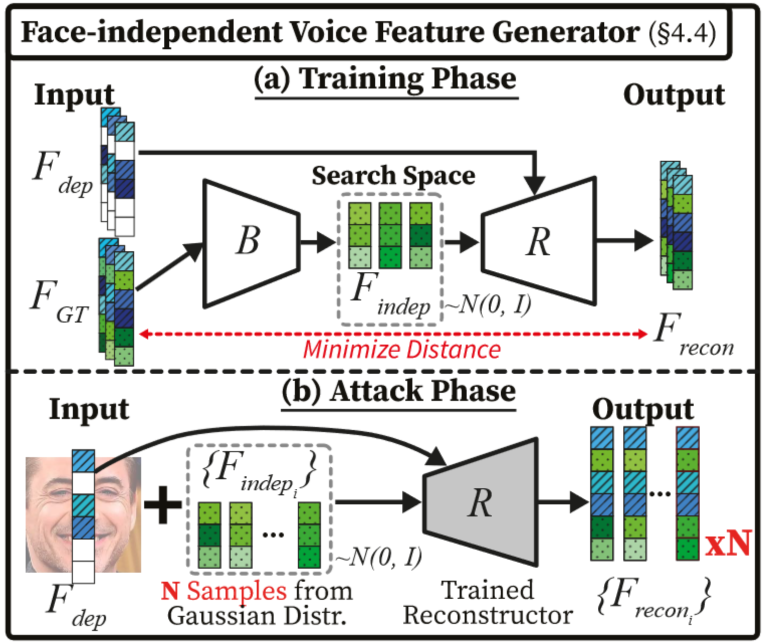
Image resolution

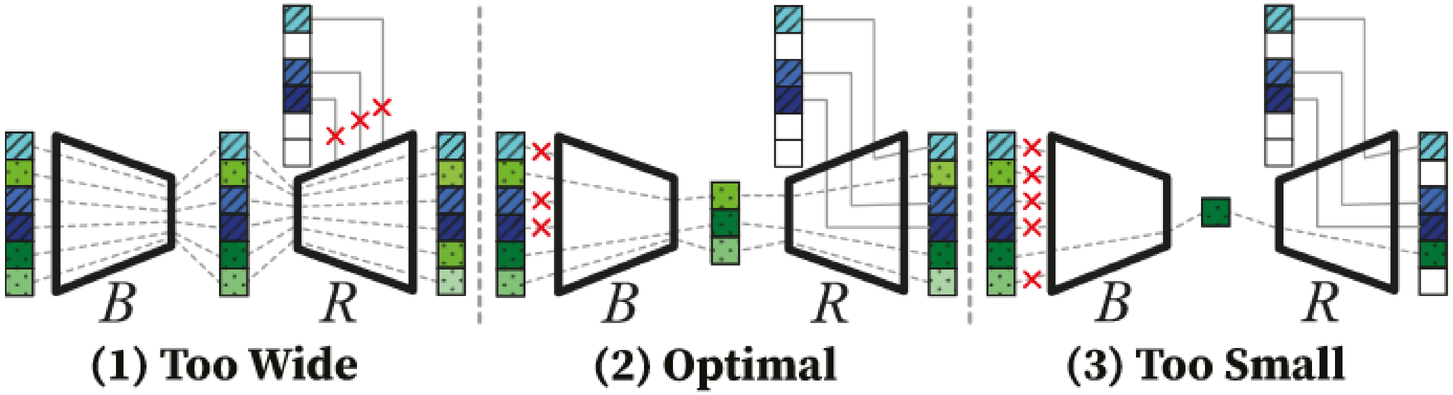
Discussion

- Voice Authentication is trending, performance will increase with 3D Face photos + Voice (i.e. Face ID + Voice)
- Safety
  – Restricting Login Attempts
  – Deepfake detection algorithm
- Foice shows how “vulnerable” current voice authentication systems are; safety enhancement is needed.

Thank You! Any Questions?

Dimension of B – finding optimal is necessary





Jiang et al., Can I Hear Your Face? Pervasive Attack on Voice Authentication Systems with a Single Face Image, USENIX’24