Enabling Secure Voice Input on Augmented Reality Headsets using Internal Body Voice

Jiacheng Shang and Jie Wu

Center for Networked Computing

Dept. of Computer and Info. Sciences

Temple University

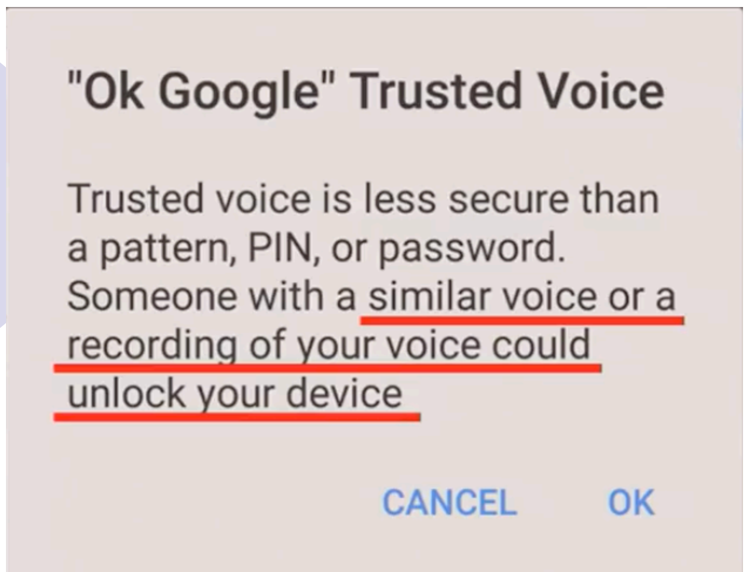
Power of Voice on AR headsets

- Voice on AR headsets
  - Primary way of communication
  - Better user experience
  - Integration with existing techniques
- Applications
  - Voice-based interaction (no identity verification)
  - Voice-based authentication (identity verification)

Applications

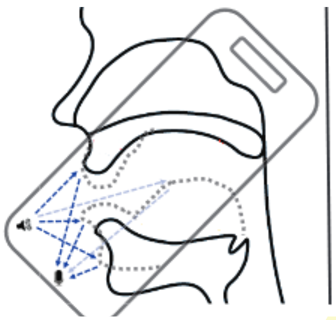
Threats of Voice

- Threats of voice
  - Human voice is often exposed to the public
  - Attackers can "steal" or even generate victim's voice
  - Security issue → replay attack



Goal: Protect the voice input for AR headsets

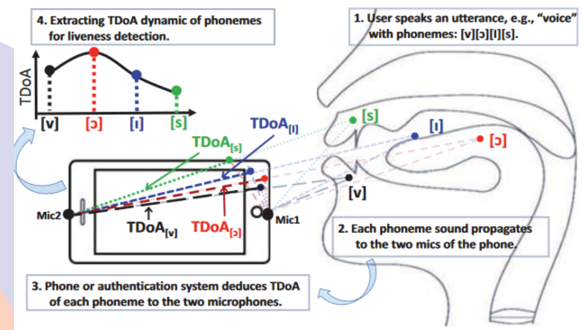
Previous work



CCS 17'

Lip motion based

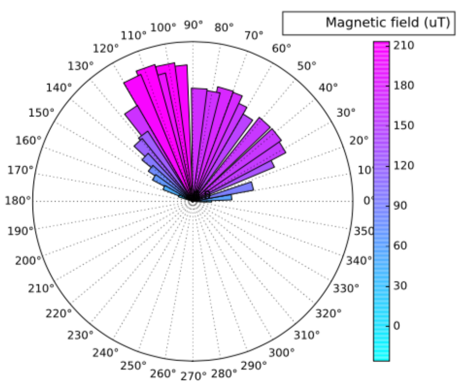
Phoneme location based



CCS 16'

Magnetic fields of loudspeakers

Throat voice based

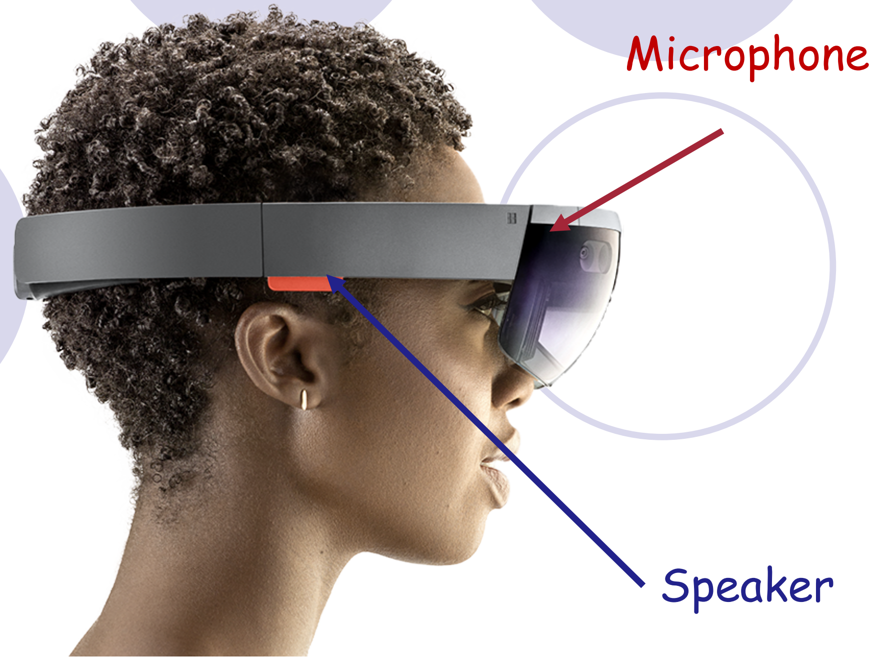


ICDCS 17'

MASS 18'

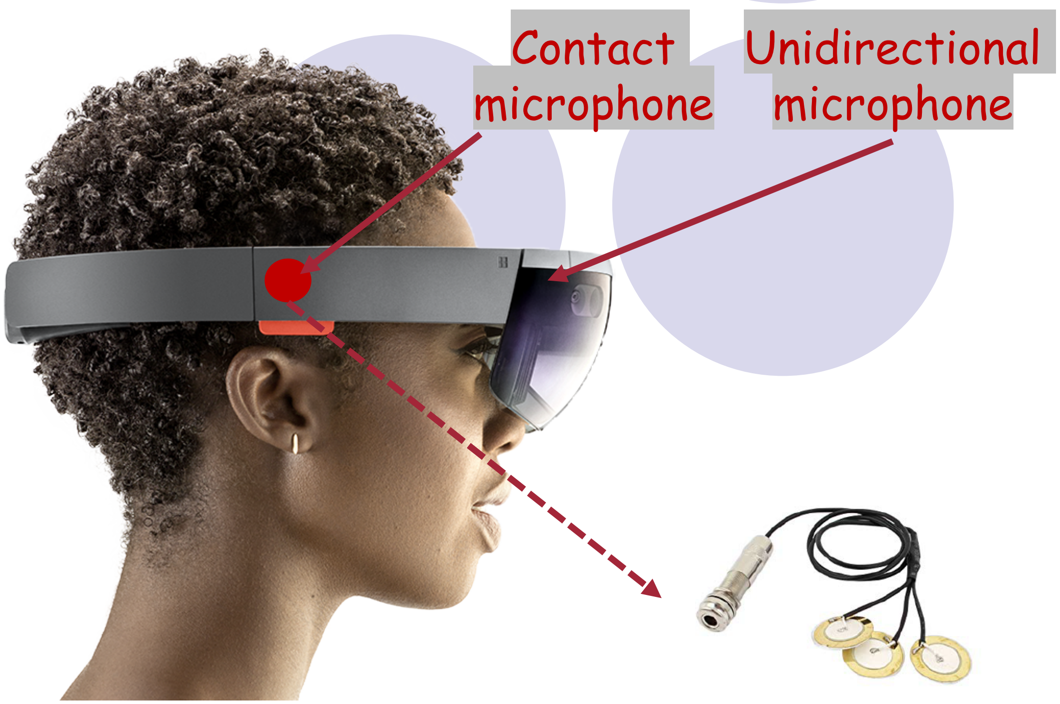
Voice Liveness detection

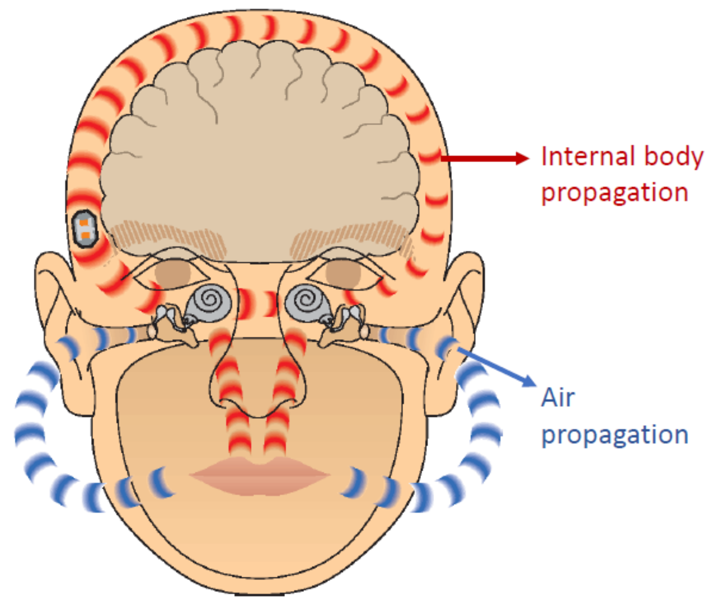
- Limitation of existing works
  - Existing solutions cannot work on AR headset due to special hardware locations
  - Only for replay attack



Voice Liveness detection

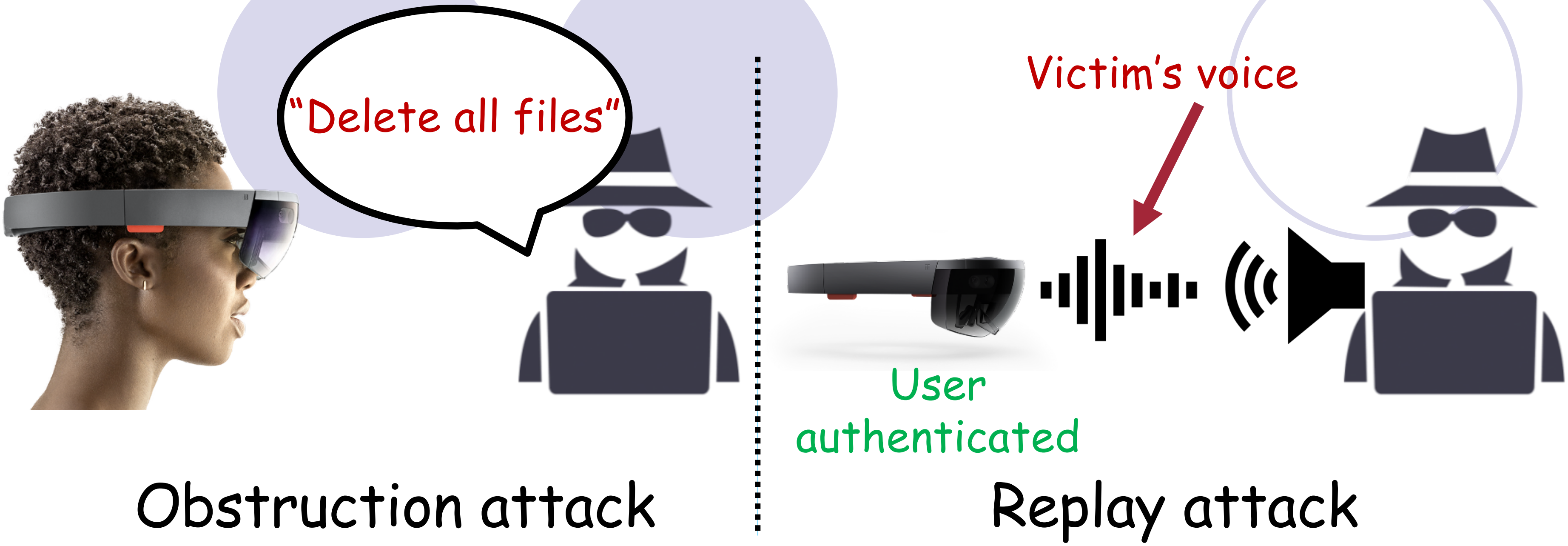
- Our work
  - Solution: voice liveness detection using internal body voice
  - Insight: voice propagates through both air and internal body
  - Collect internal body voice using a contact microphone



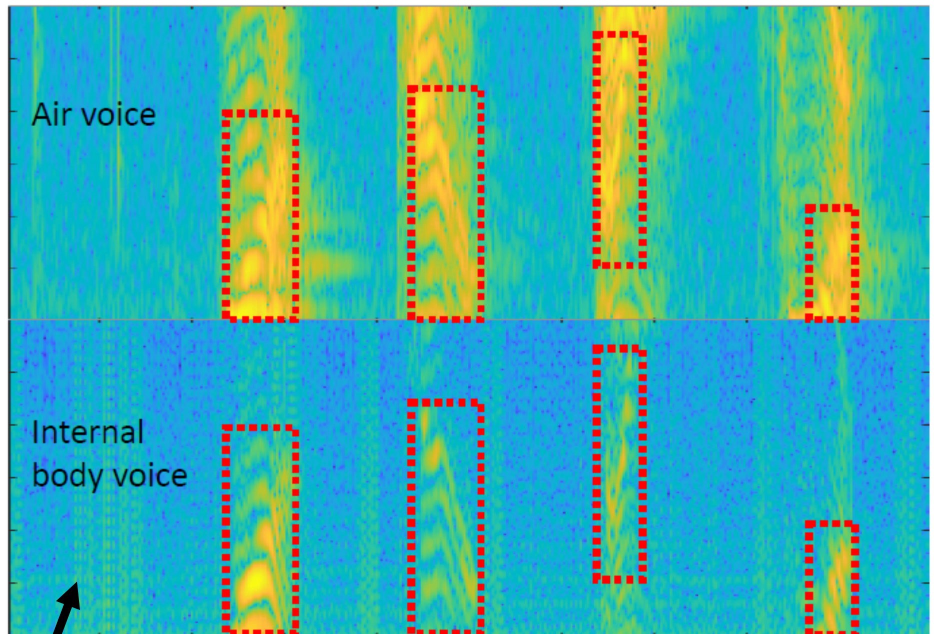


Attack model

- Obstruction attack for voice-based interaction
  - Attacker nearby issues a malicious command (e.g. “delete all files”)
- Replay attack for voice-based authentication
  - Attacker steals victim’s voice at the mouth with recorder and replays it to AR headset



Spectrogram generation



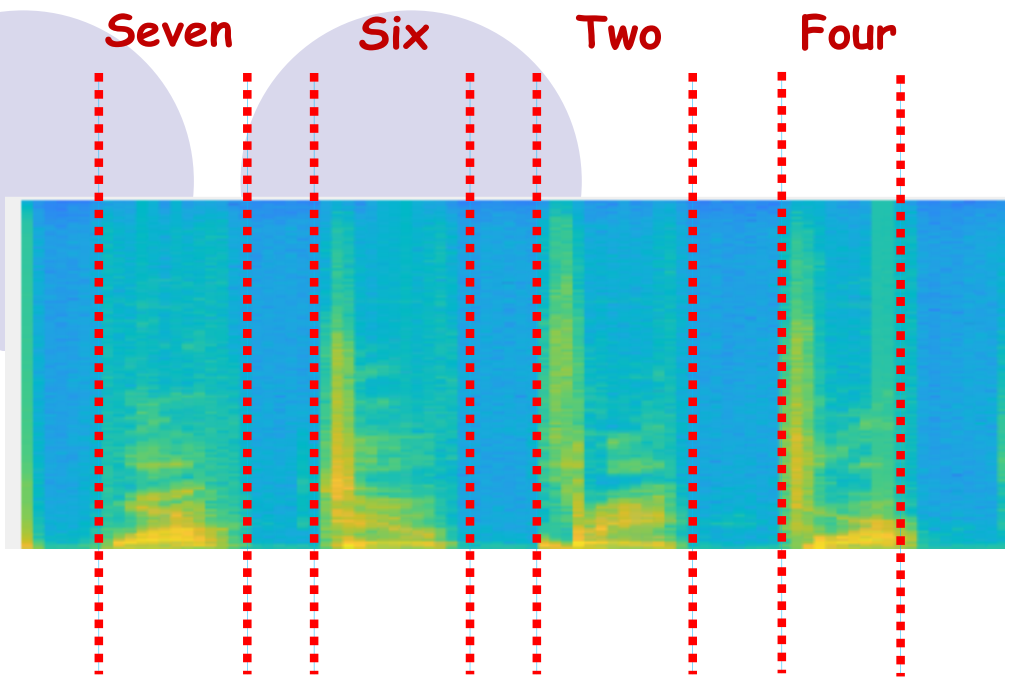
Compute the spectra using Short-time Fourier transform

$$spectrogram\{x[t]\}(m,\omega) = |\sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}|^2$$

$x[n]$: voice in time domain $w[n]$: window ω: angular frequency

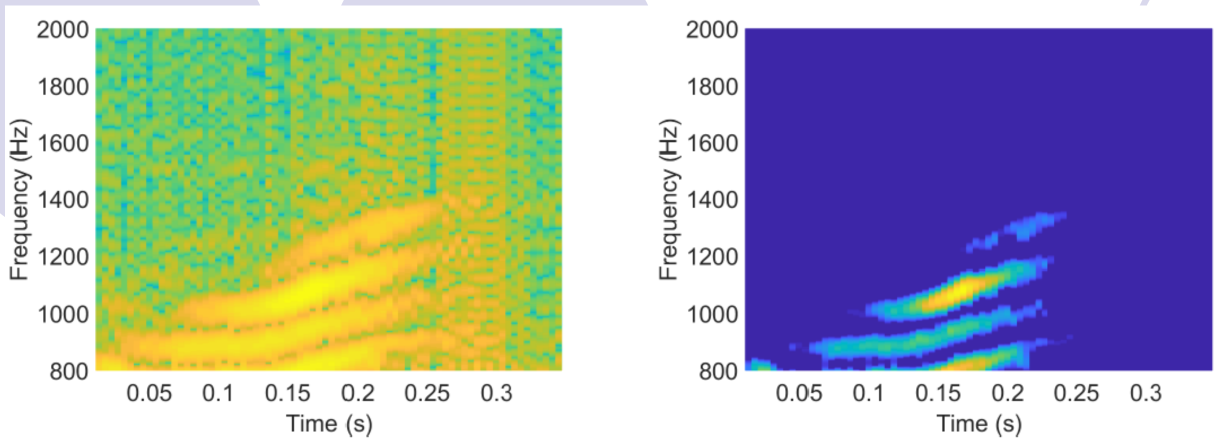
Word Segmentation

- Recorded voice: the sequence of words and noise
- Segmenting each word:
  - Using Hidden Markov Model-based techniques



Spectrum enhancement

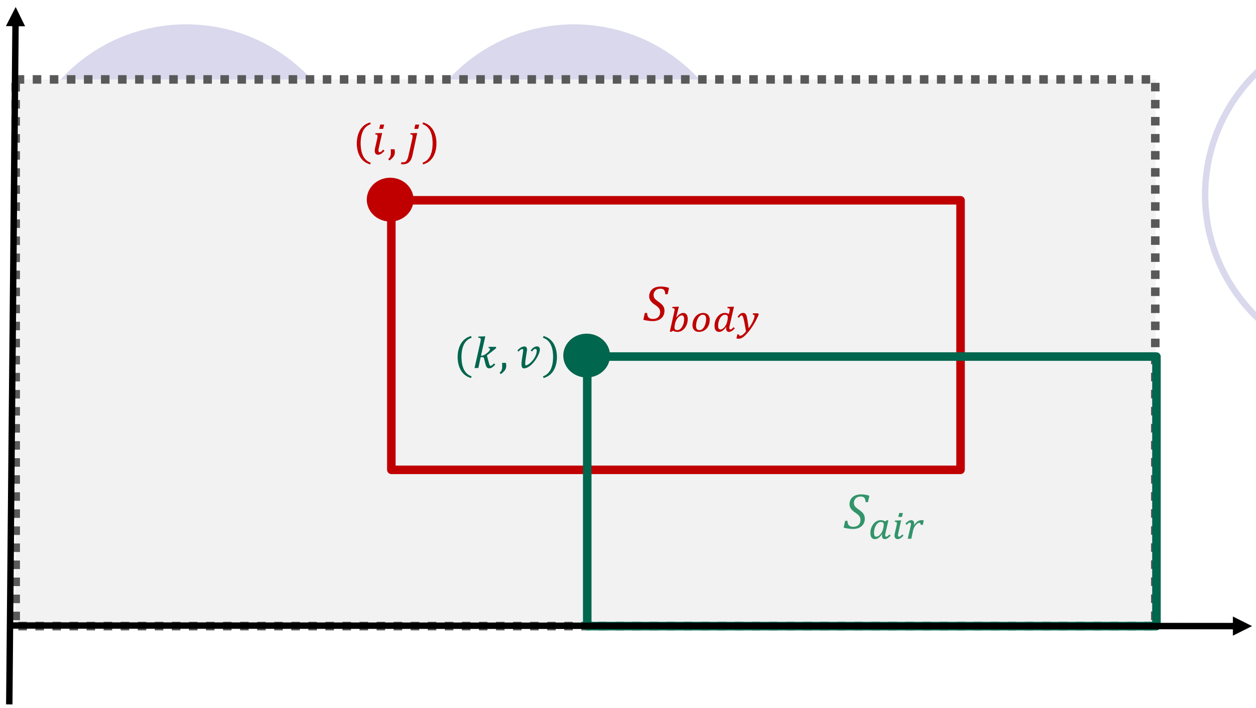
- Spectrogram enhancement: further remove background noise
  - Voice dominates the spectrogram
  - Noise floor: 80% highest power in the spectrogram of each word



(a) Raw internal body voice. (b) Enhenced spectrogram.

Liveness detection for a single word

- Liveness detection for AR headset
  - Observation 1: the energy distributions in two spectrograms $S_{body}(M * N)$ and $S_{air}(M * N)$ are highly correlated
  - If we find a best match, they should be perfectly overlapped

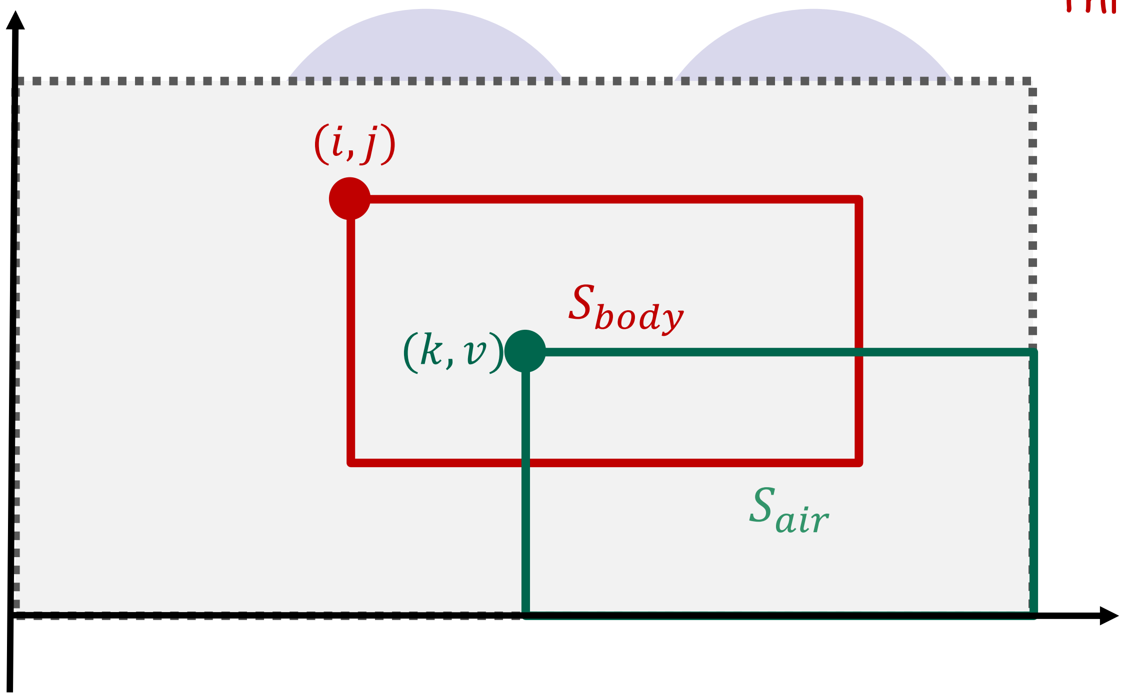


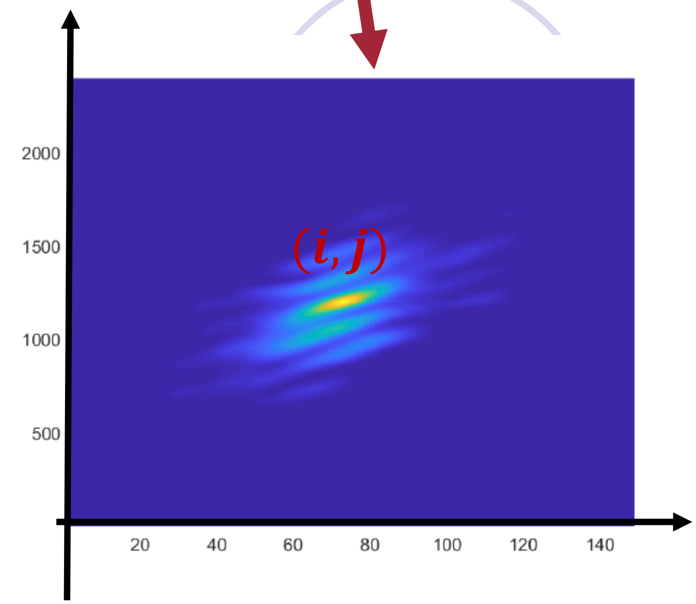
Liveness detection for a single word

- Liveness detection for AR headset
  - (i,j) can be solved by finding the maximum in the correlation matrix

$$\frac{|i-N|}{2N} < \gamma \ \&\& \ \frac{|j-M|}{2M} < \delta$$

Threshold: 0.1

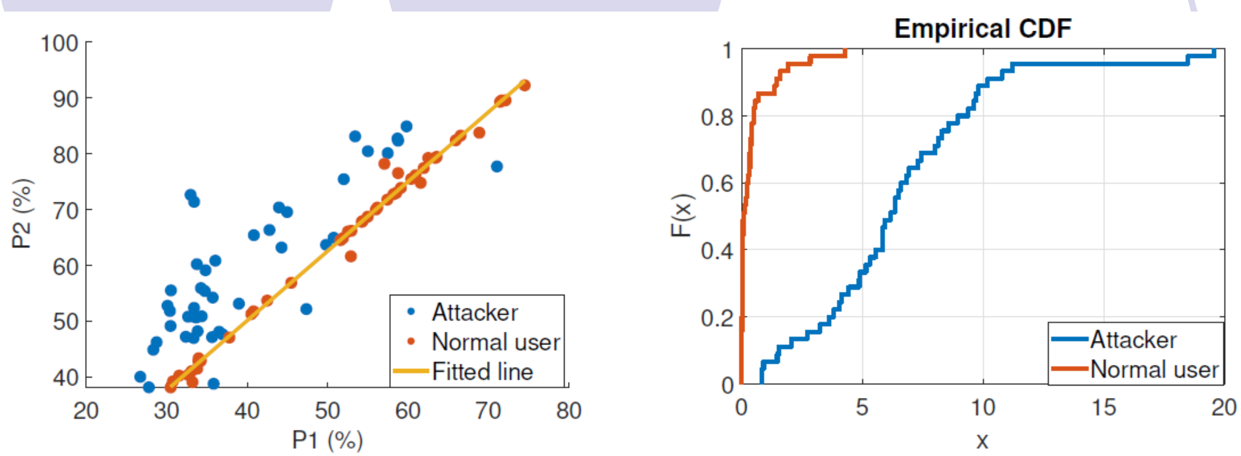




Liveness detection for a single word

$$P_1 = \frac{Sizeof(\{(i,j) \mid S_1[i,j] > 0 \ \& \ S_2[i,j] > 0\})}{Sizeof(\{(i,j) | S_1[i,j] > 0\})}$$

- Observation 2:
  - two spectrograms $S_{body}(M * N)$ and $S_{air}(M * N)$ have much shared information (non-zero entries)
- Two metrics:
  - Shared information: non-zero entries in both spectrograms
  - P_1: the proportion of the shared information that is in S_{body}
  - P_2 : the proportion of the shared information that is in S_{air}



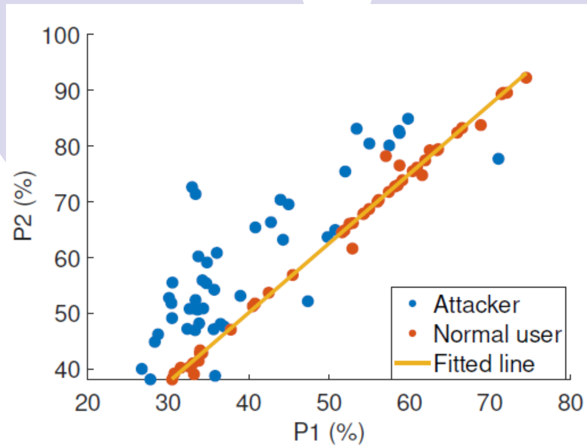
(a) Feature distribution. (b) Distance distribution.

Liveness detection for a single word

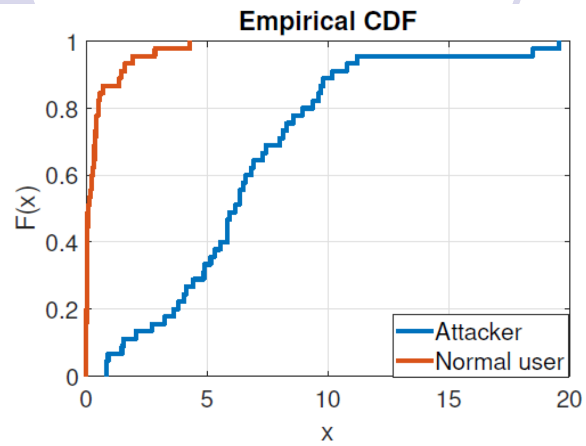
- Fitting a line using normal user's training data: y = ax + b
- If a point is away from the line, it is considered from the attacker

$$\frac{|aP_1 + bP_2 + c|}{\sqrt{a^2 + b^2}} < \gamma$$

Threshold:
95% largest distance of normal user's training data



(a) Feature distribution.

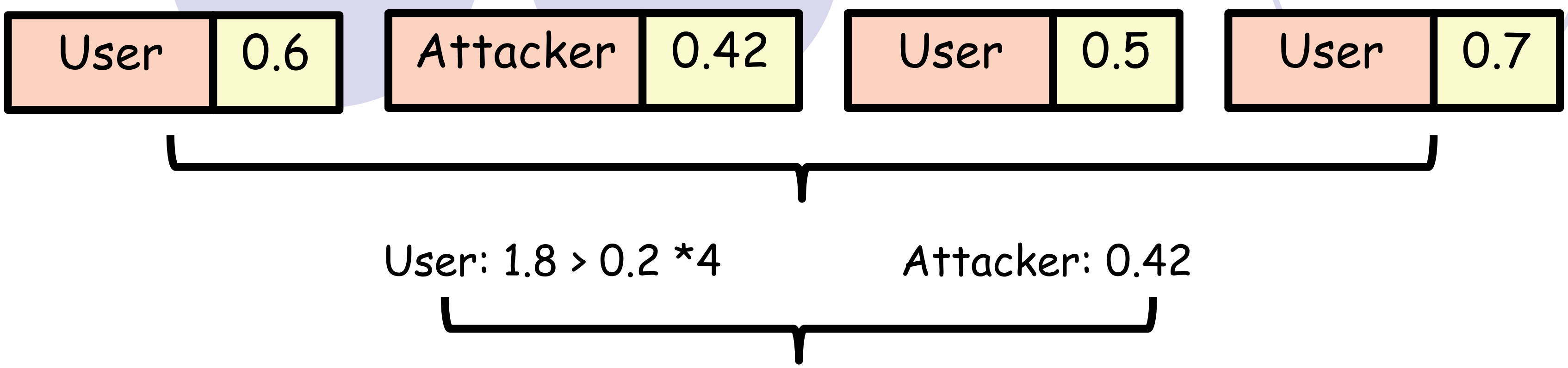


(b) Distance distribution.

Liveness detection for a sentence

- Combining the classification results from multiple words
  - Weighted majority Voting
  - Player: each word
  - Weight: the smaller value of P_1 and P_2
  - Decision threshold: $c * n$

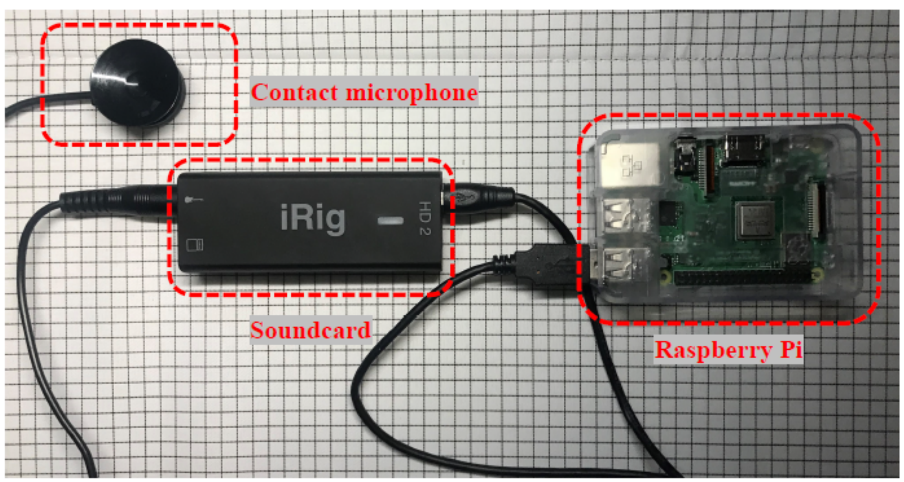
Set to 0.2 by default — The number of words in the sentence



The voice is from the normal user

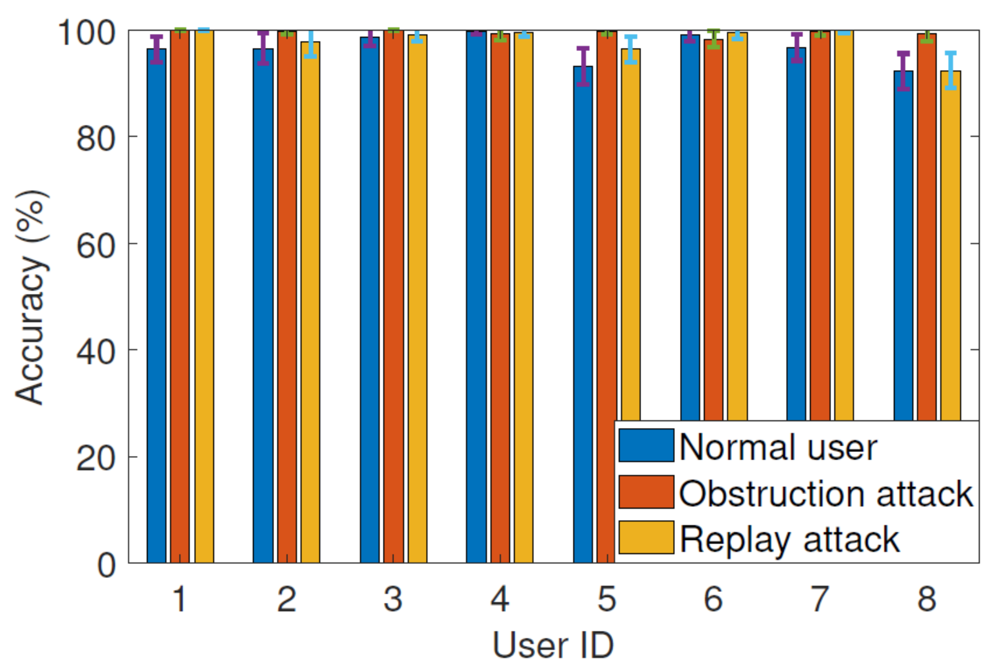
Evaluation

- Body voice: Contact microphone via Raspberry Pi 3 b+ board
- Air voice: A smartphone is used to record and replay mouth voices
- 8 volunteers (5 males and 3 females)



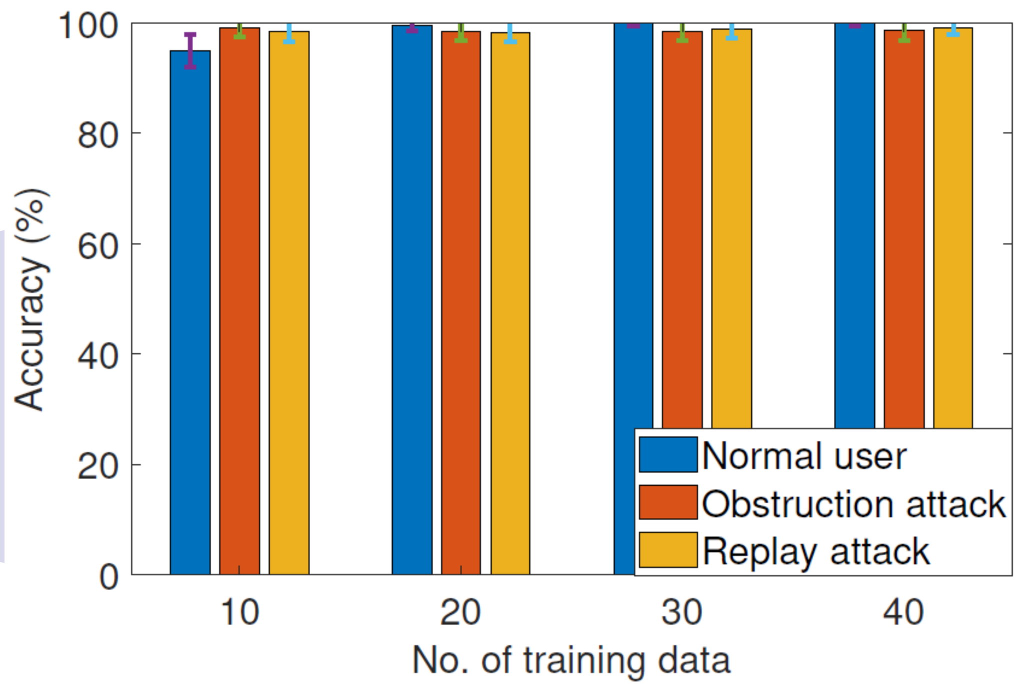
Evaluation

- Overall performance
  - Average authentication accuracy: 92.3%
  - Average true rejection rate of random attack: 99.2%
  - Average true rejection rate of mimicry attack: 98%



Evaluation

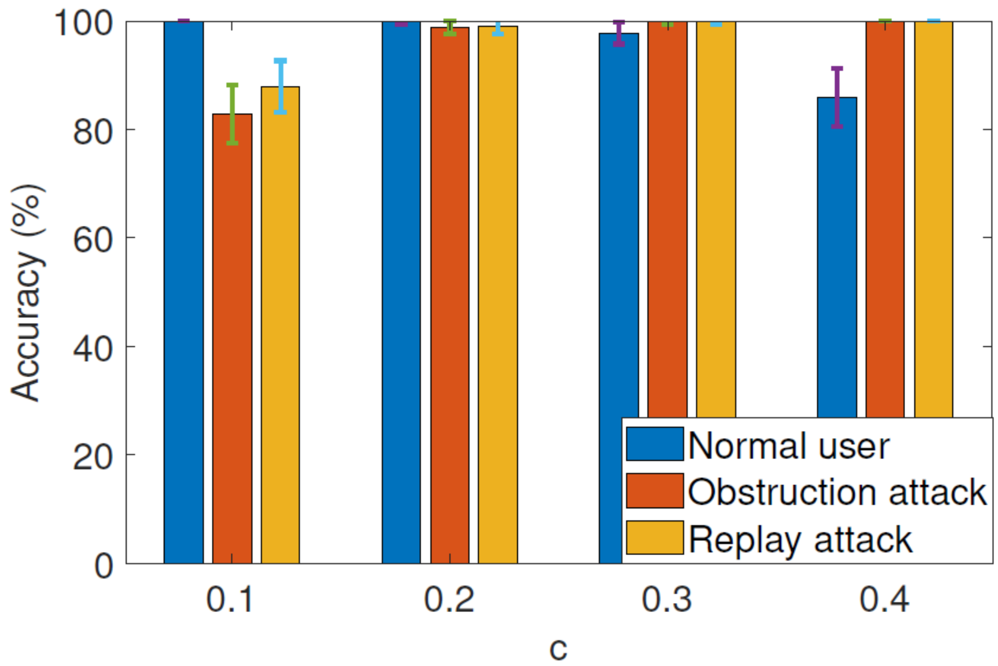
- Impact of training set size



20 words are enough to ensure good performance

Evaluation

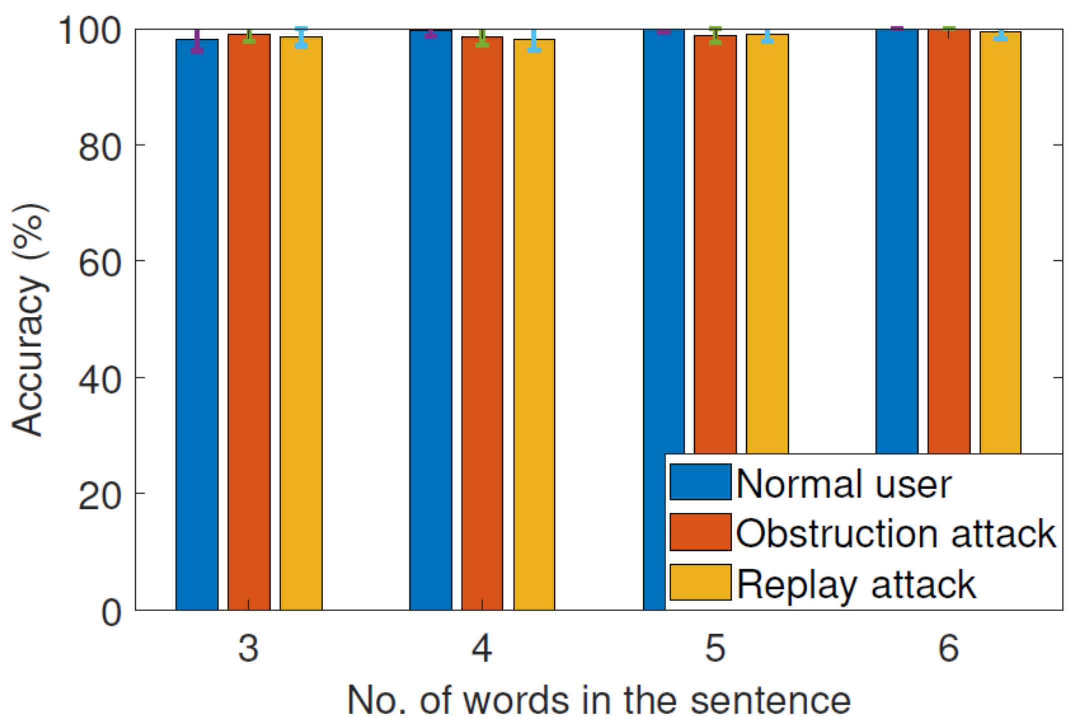
- Impact of voting threshold



The voting threshold should be between 0.2 and 0.3

Evaluation

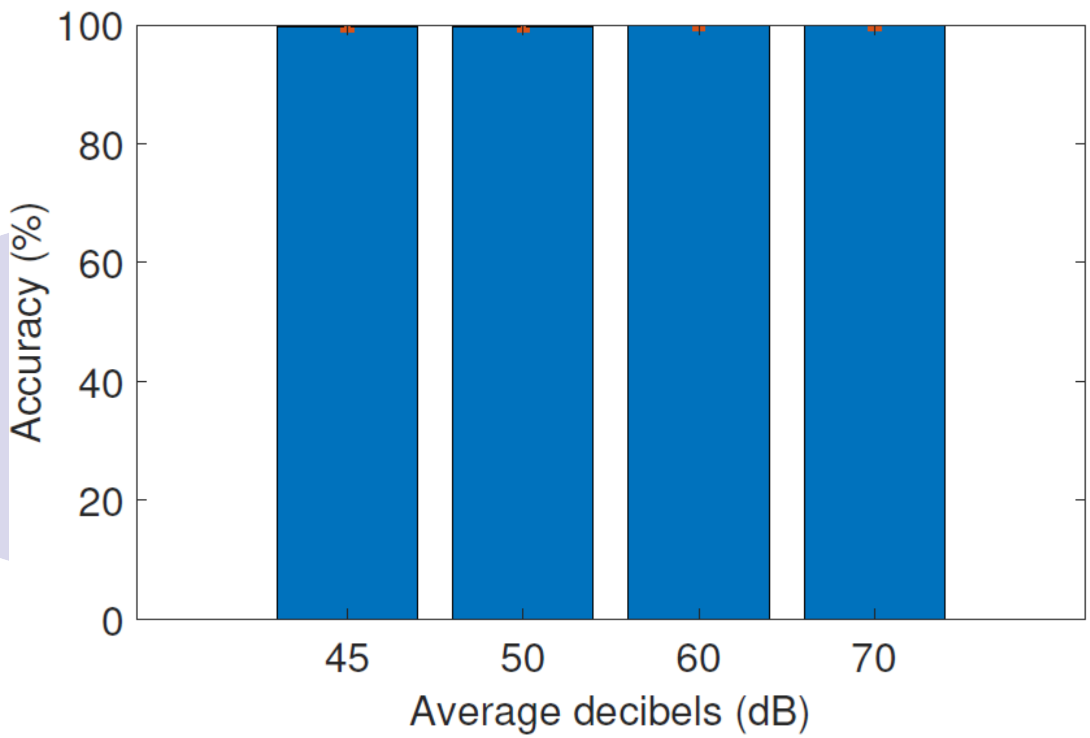
- Impact of number of words in a sentence



Our system can work for most voice commands

Evaluation

- Impact of background noise



Our system is robust to background noise in daily life

Conclusion

- We show that the internal body voice can be used to secure the voice input for AR headsets
- We develop a prototype and conduct comprehensive evaluations.
- Experimental results show that our system can successfully defend against obstruction and replay attacks with an accuracy of at least 98%.

Thanks!
Q&A