SRVoice: A Robust Sparse Representation-based Liveness Detection System

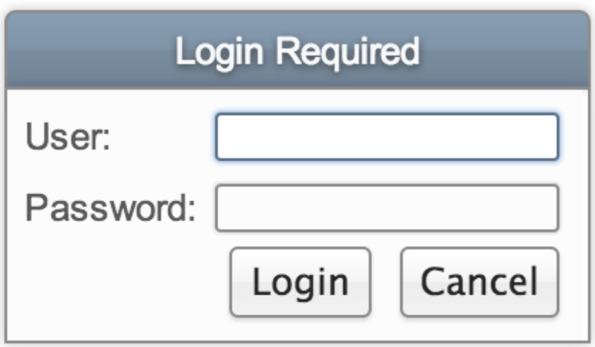
Jiacheng Shang, Si Chen, and Jie Wu

Center for Networked Computing

Temple University

Biometrics: Voiceprint

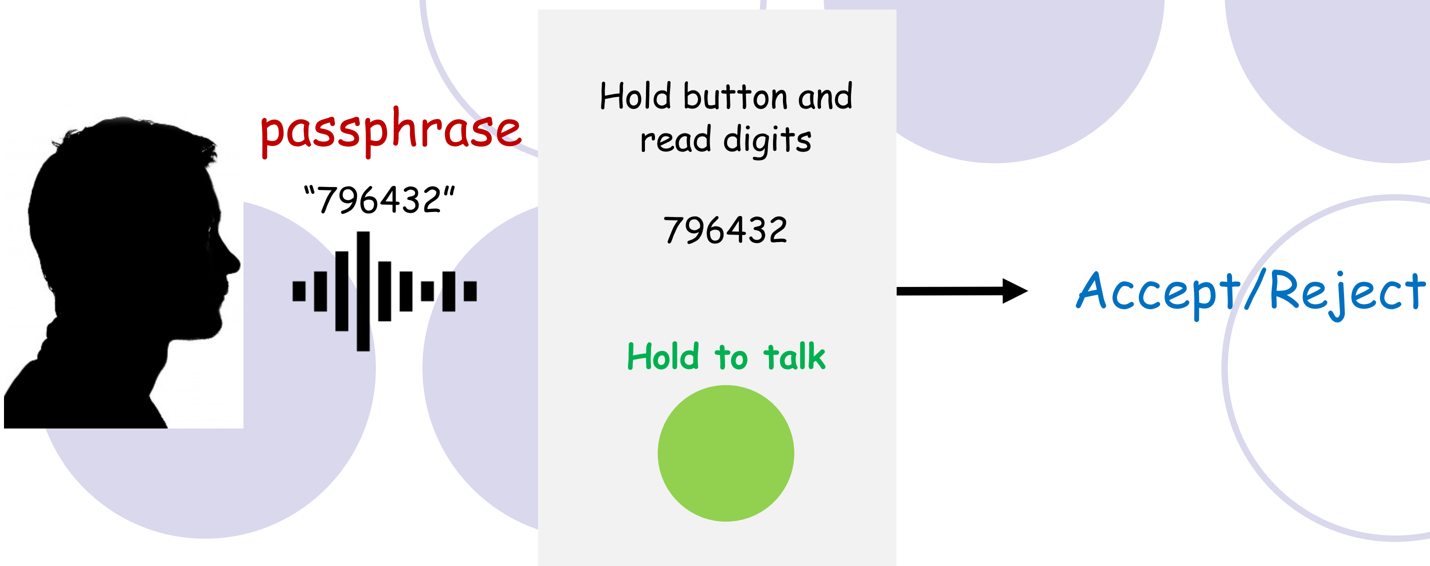
- Voiceprint
  - Promising alternative to password
  - Primary way of communication
  - Better user experience
  - Integration with existing techniques for multi-factor authentication



Applications

Biometrics: Voiceprint

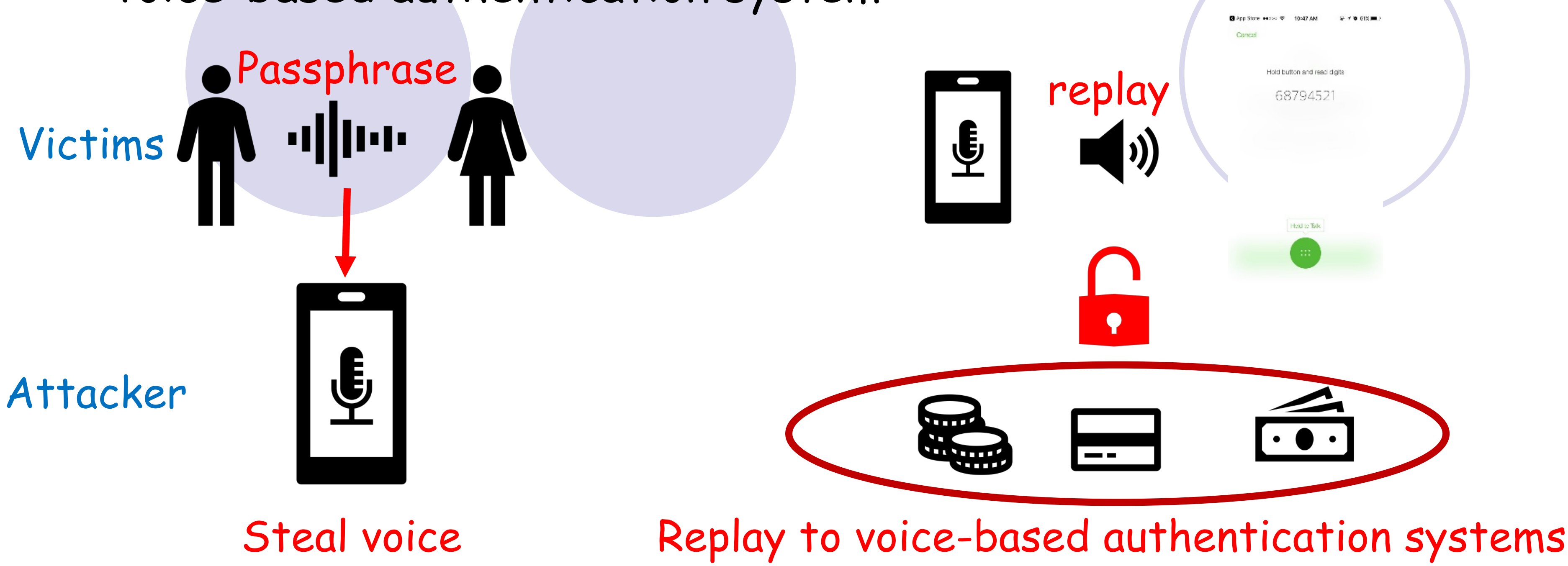
- Voiceprint example



Voiceprint-based authentication

Threats

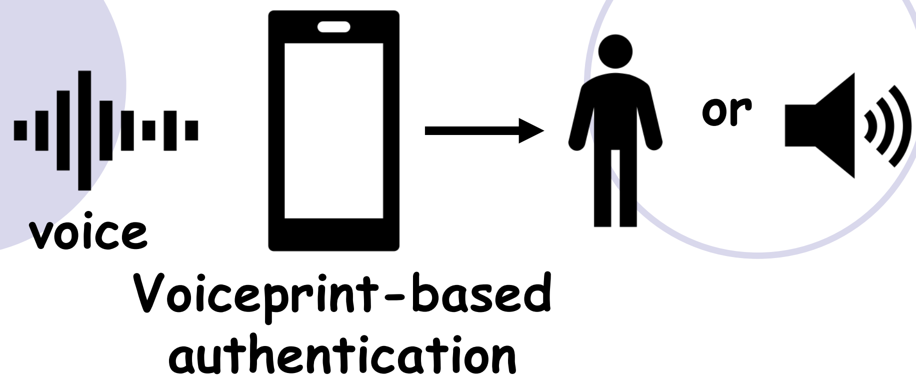
- Human voice is often exposed to the public
- Attackers can "steal" victim's voice with recorders
- Security issues
  - E.g. Adversary could impersonate the victim to spoof the voice-based authentication system



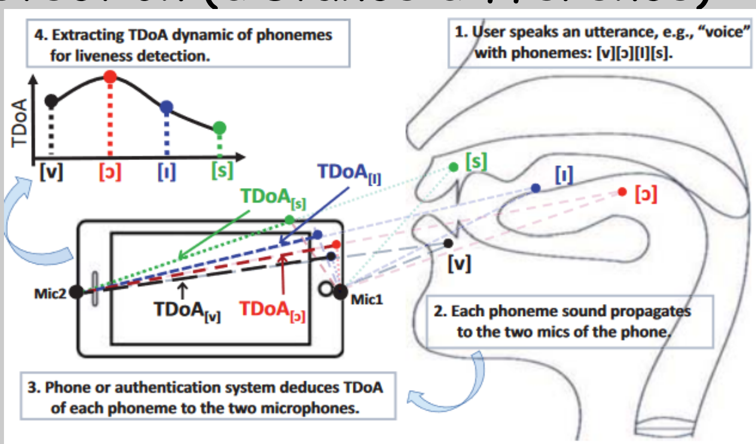
Reverse Turing Test

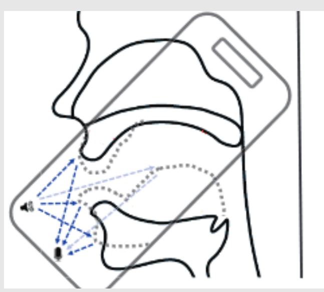
CAPTCHA

Completely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans



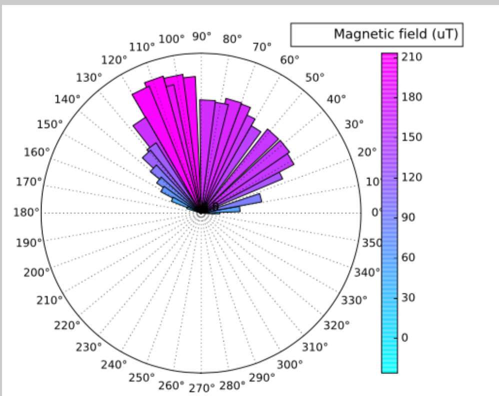
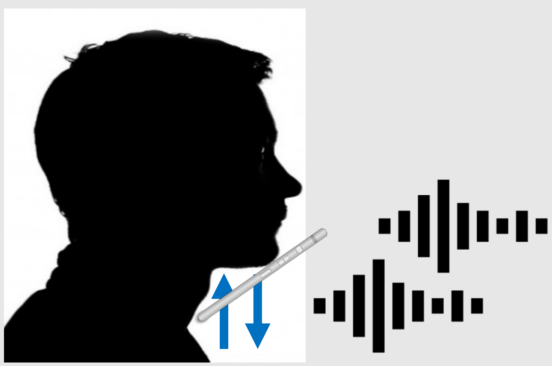
Previous work

<table>
<tr><th>Systems</th><th>Limitations</th></tr>
<tr><td>Phoneme location-based liveness detection (distance difference)

</td><td>• Low true acceptance rate (TAR): the smartphone needs to be static relative to the mouth

VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones (L. Zhang et al. CCS 2016)</td></tr>
<tr><td>Lip motion-based liveness detection (Doppler shift)
</td><td>• Low true acceptance rate (TAR): the smartphone needs to be static relative to the mouth

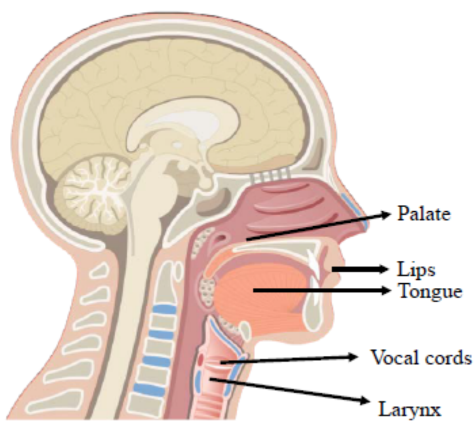
Hearing Your Voice Is Not Enough: An Articulatory Gesture Based Mobile Voice Authentication (L. Zhang et al. CCS 2017)</td></tr>
</table>

Previous work

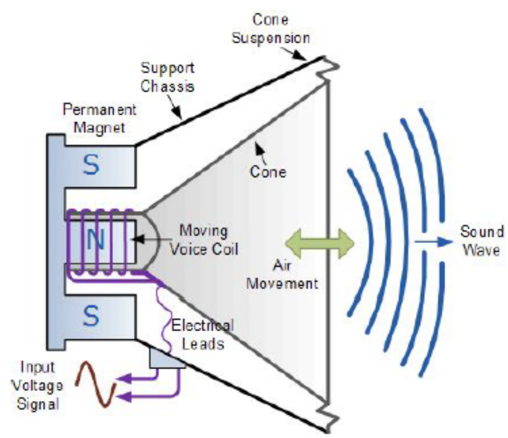
Systems	Limitations
Leveraging the magnetic fields of loudspeakers 	• Low TAR: cannot work if magnetic noise exists • Low true rejection rate (TRR): cannot work if the attacker uses non-conventional loudspeaker You Can Hear But You Cannot Steal: Defending against Voice Impersonation Attacks on Smartphones (S. Chen et al. ICDCS 2017)
Audio and throat motion-based 	• Low TRR: Cannot work if users are performing other activities Defending Against Voice Spoofing: A Robust Software-based Liveness Detection System (J. Shang et al. MASS 2018)

Basic ideas

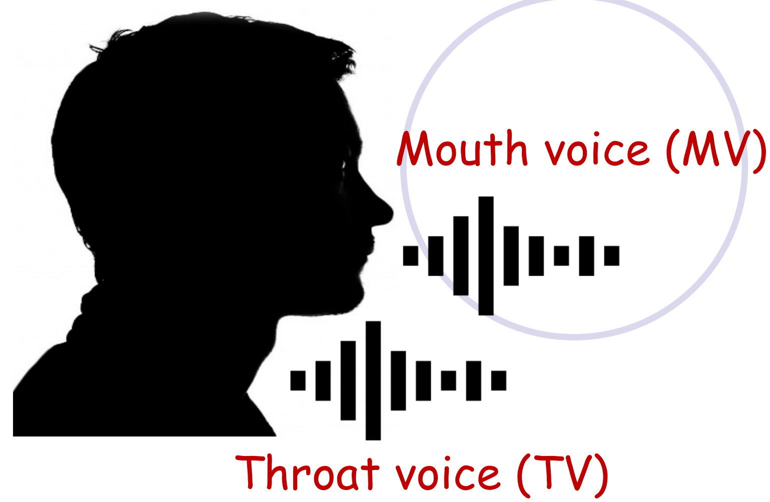
- Leveraging the structural differences between the vocal systems of human and loudspeakers



(a) Human vocal structure

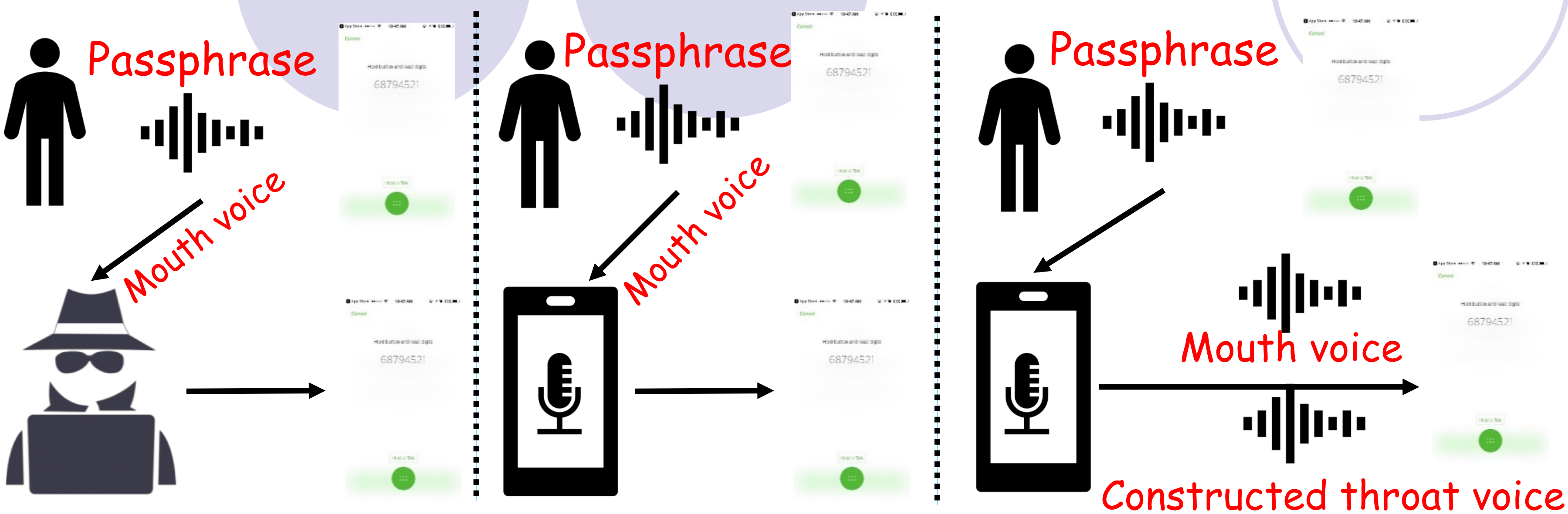


(b) Speaker's structure



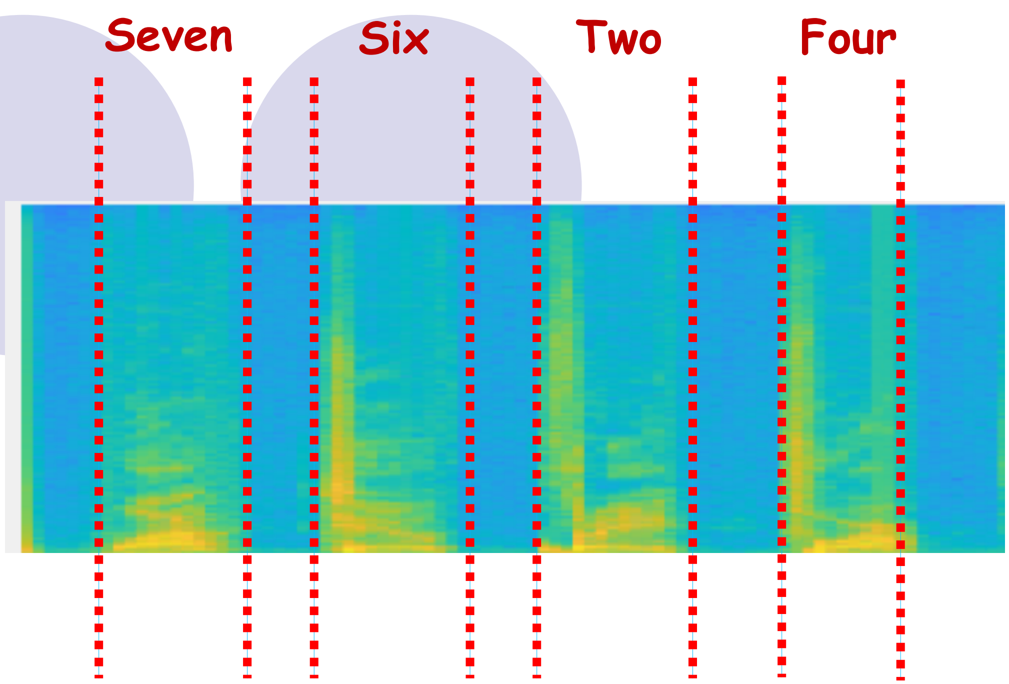
Attack model

- Mimicry attack
  - Attackers imitate victim's voice without extra device
- Replay attack
  - Attackers steal victim's voice at the mouth with recorder
- Reconstruction attack
  - Attackers reconstruct victim's throat voice using low-pass filter

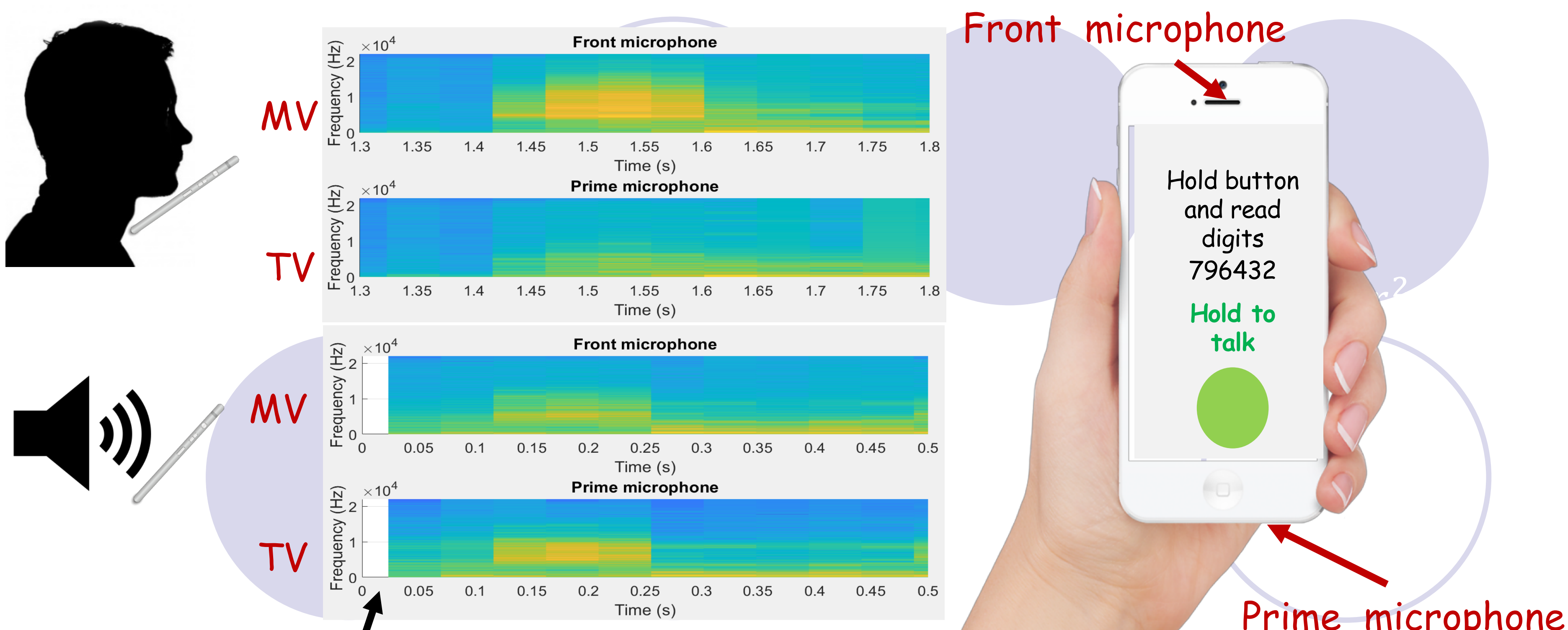


Word Segmentation

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



Feature Extraction



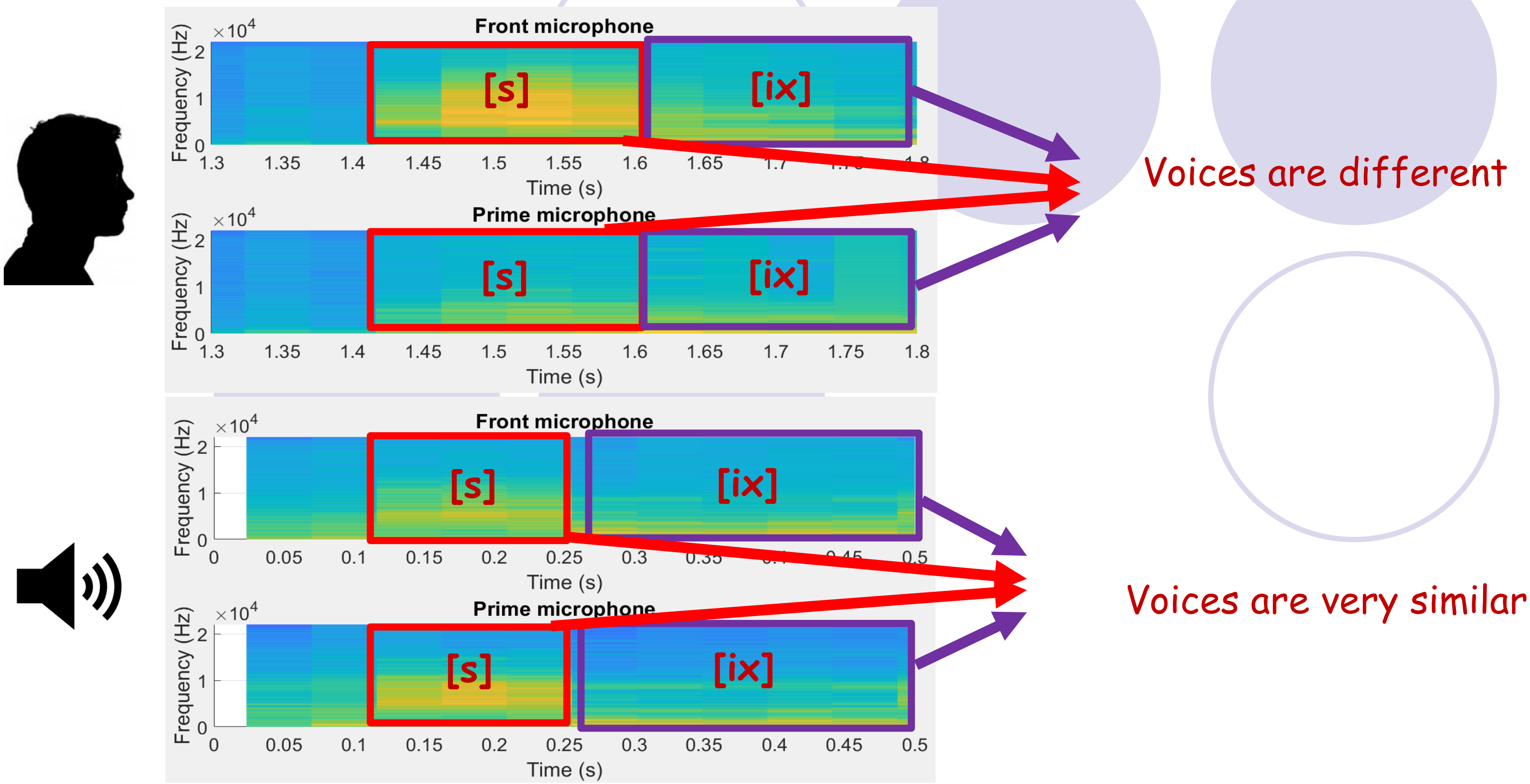
Compute the spectra using Short-time Fourier transform

Time domain to Frequency and time domains

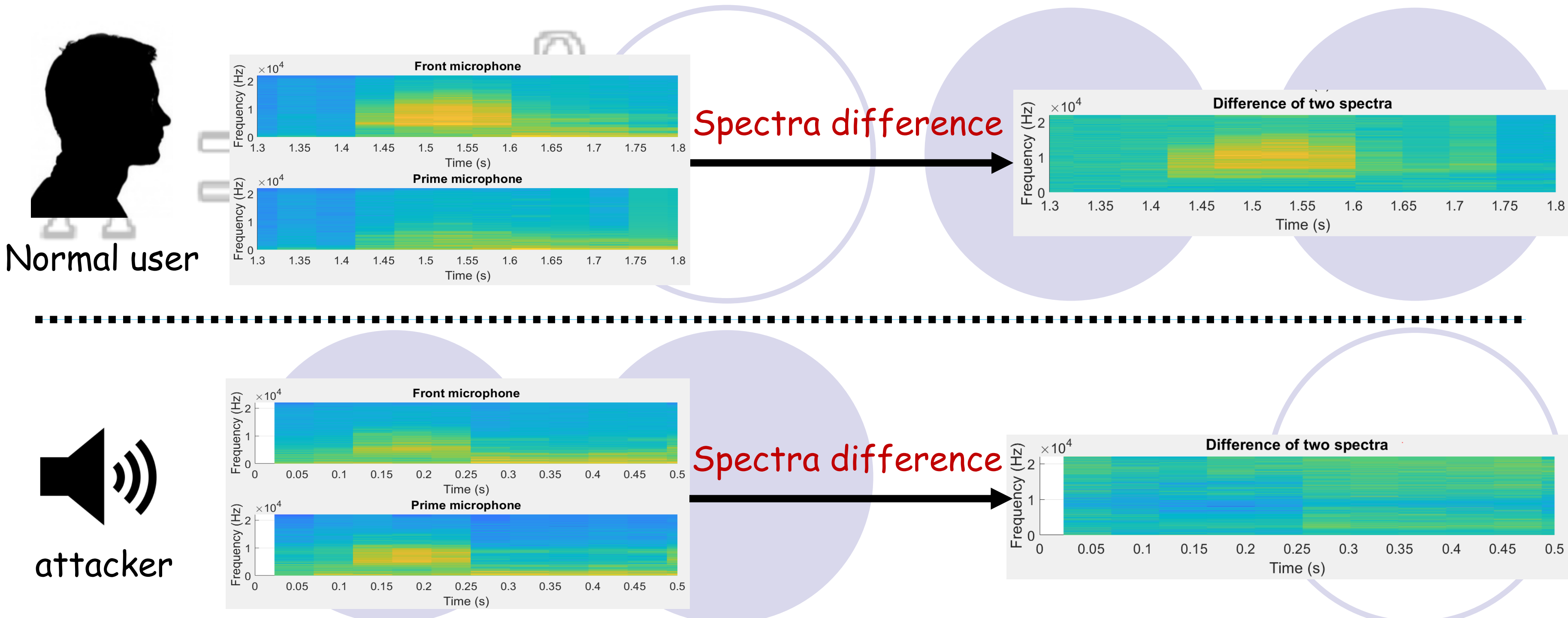
$$spectrogram\{x[t]\}(m, \omega) = \left| \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \right|^2$$ Convolution

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

Feature Extraction



Feature Extraction

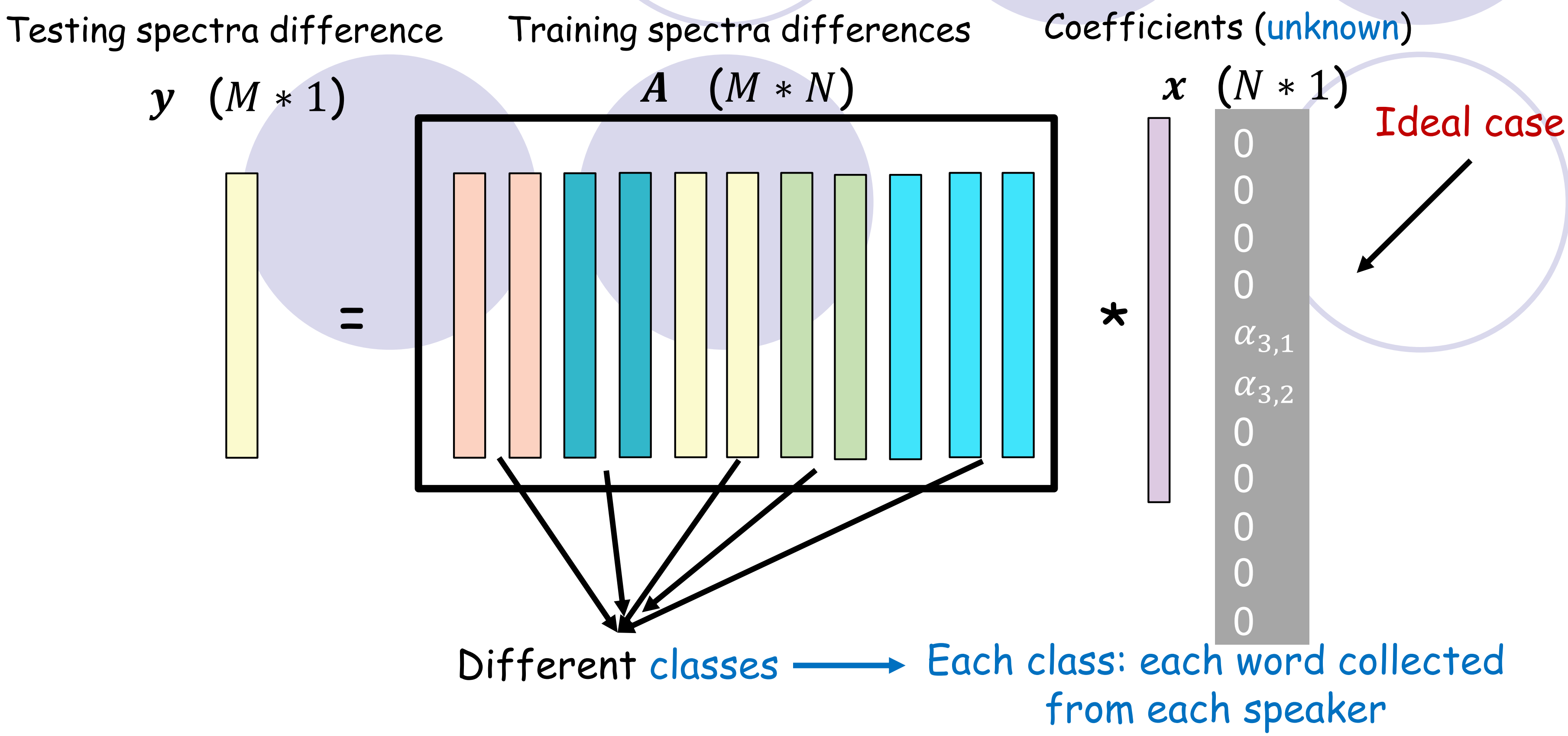


We further convert each spectra difference (matrix) to a vector

E.g. $\begin{matrix} m_1 & m_4 & m_7 \\ m_2 & m_5 & m_8 \\ m_3 & m_6 & m_9 \end{matrix} \longrightarrow [m_1, m_2, m_3, m_4, \ldots, m_9]$

Liveness detection for a single word

- Feature selection among spectra difference is critical
- Sparse representation-based classification
  - Assumption: Samples from a single class do lie on a subspace



Liveness detection for a single word

- If we do not know the label of y
  - We can reversely compute x based on a sparse representation formulation

$$\widehat{x_1} = \arg\min ||x||_1 \ \ subject\ to\ \ y = Ax$$

$$||x||_1 = sum(|x|)$$

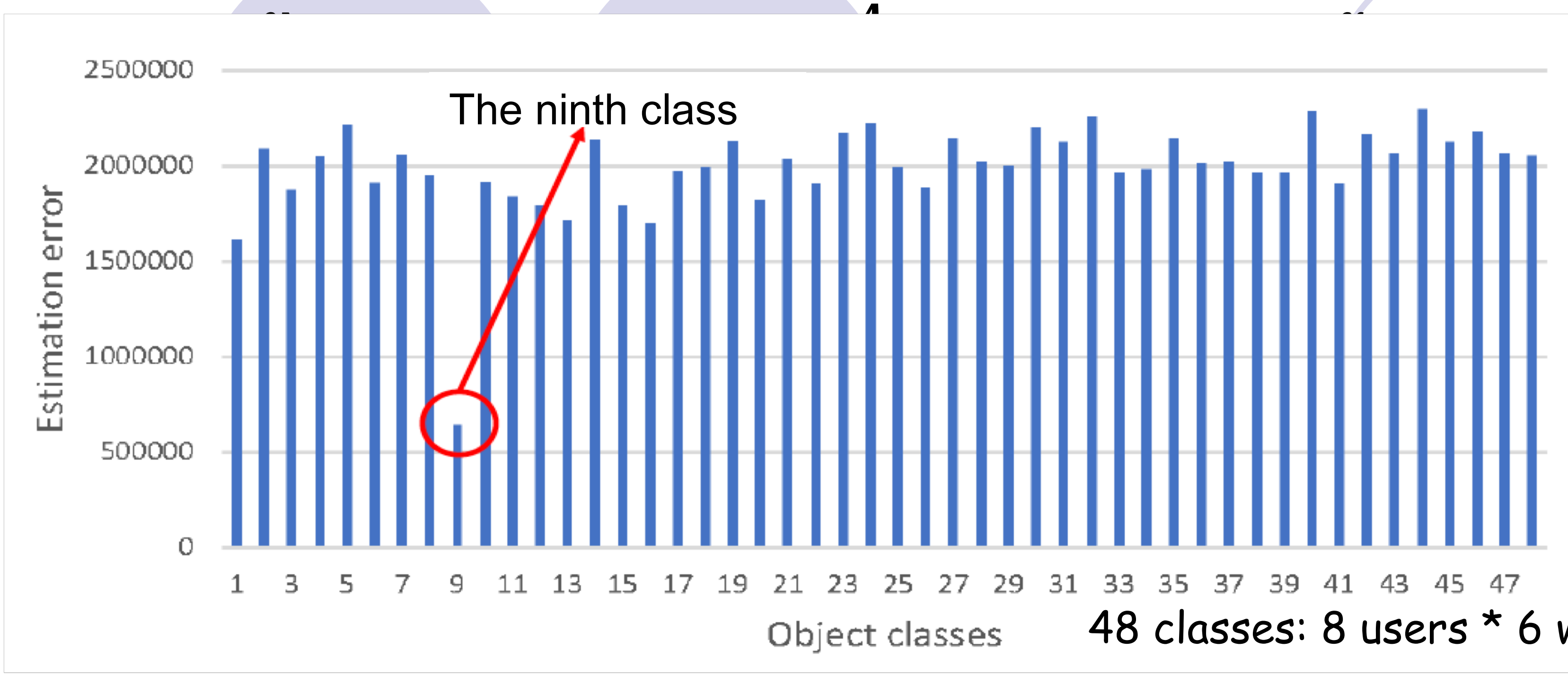
  - If number of object classes is reasonably large, the x should be sparse enough, and this problem can be solved in polynomial time by standard linear programming method

Simple idea: assigning y to the object class with the single largest entry in $\widehat{x_1}$

--> does not harness linear structure of all training samples in the same class

Liveness detection for a single word

- We use estimation error $E(y)$ for each possible class
$$E(y) = mean(||y - A\Delta_i\widehat{x_1}||_1)$$
  - $\Delta_i(\widehat{x_1})$ is the coefficient vector that only contains coefficients associated with the i^{th} class
  - y is labeled as the class whose $E(y)$ is minimal



48 classes: 8 users * 6 words

Liveness detection for a passphrase

- Improving performance by combining results of multiple words in a passphrase (weighted voting)
  - Each player is a tuple (user, word, weight)
  - Weight:
    - If the detected word ≠ the argued word, weight is 0
    - Otherwise, $Weight(w) = 1 + log^{(1+N_{unvoiced}(w))}$

classification results

w: a word

$N_{unvoiced}(w)$: the # of unvoiced phonemes in word w

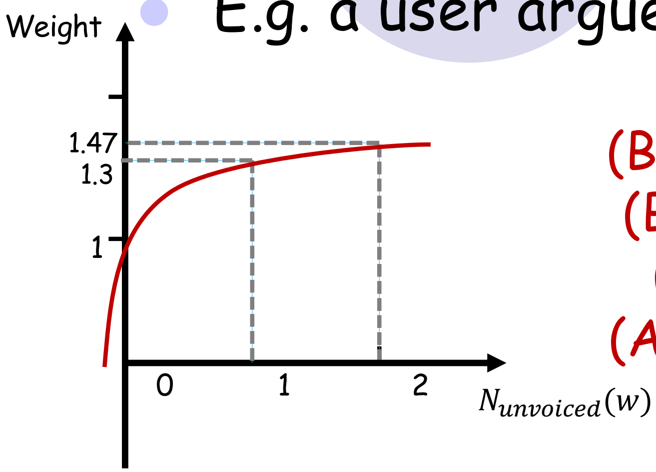
Digital words	Weight
“One”, “Nine”	1
“Two”, “Three”, “Four”, “Five”, “Seven”, “Eight”, “Ten”	1.3
“Six”	1.47

Liveness detection for a passphrase

- Improving performance by combining results of multiple words in a passphrase (weighted voting)
  - Each player is a tuple (user, word, weight)
    - classification results
  - Weight:
    - If the detected word ≠ the argued word, weight is 0
    - Otherwise, $Weight(w) = 1 + log^{(1+N_{unvoiced}(w))}$

      w: a word

      $N_{unvoiced}(w)$: the # of unvoiced phonemes in word w
- E.g. a user argues he/she is Bob (passphrase 7614)



(Bob, "Seven", 1.3)
(Bob, "Six", 1.47)
(Bob, "One", 1)
(Alice, "Four", 1.3)

→ Bob: 3.77>2
Alice: 1.3

Cut-off threshold

→ Bob: matched

Evaluation

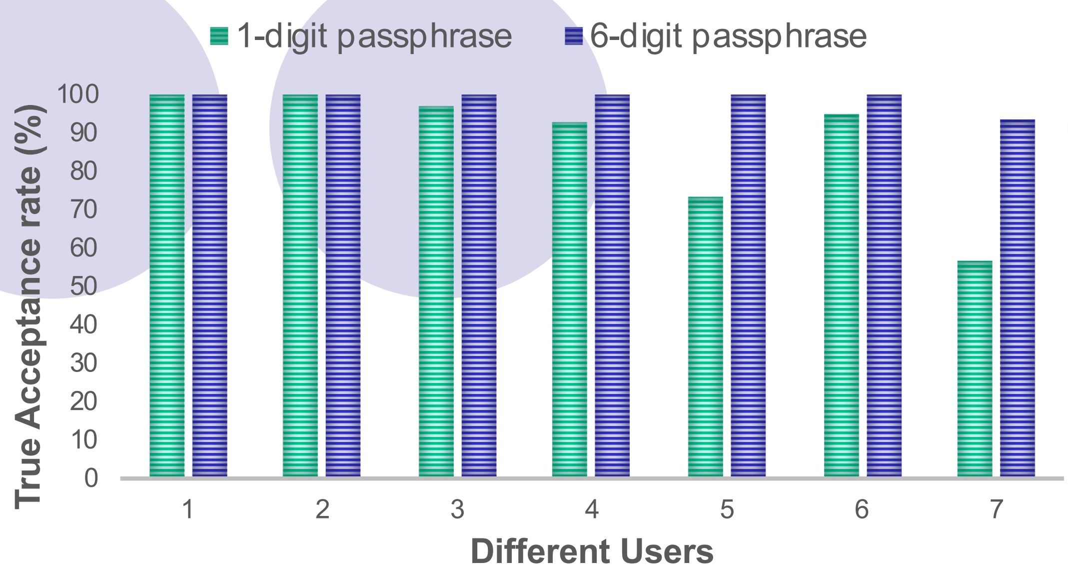
- Methodology
  - Implement our system on real smartphones (nexus 4 and 5)
  - Use two loudspeakers, 50% each, to perform replay attack

Maker	Model	Number of trumpets
Willnorn	SoundPlus	2
Amazon	Echo	2

- Performance metrics
  - The standard automatic speaker verification metrics
    - True Acceptance Rate (TAR)
    - True Rejection Rate (TRR)

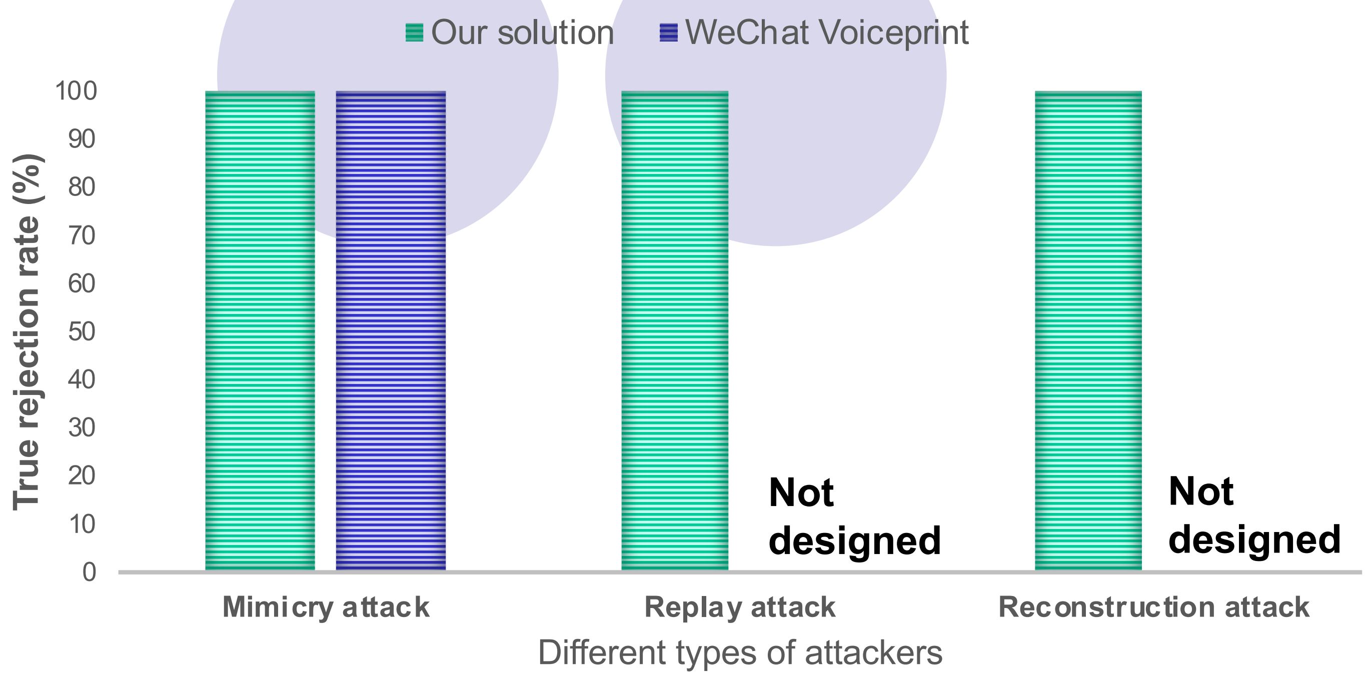
Evaluation

- Performance for normal users
  - Average true acceptance rate for a single word: 87.83%
  - Tolerating mistake by voting: combining detection results of 6 words, average TAR is improved to 99.04%



Evaluation

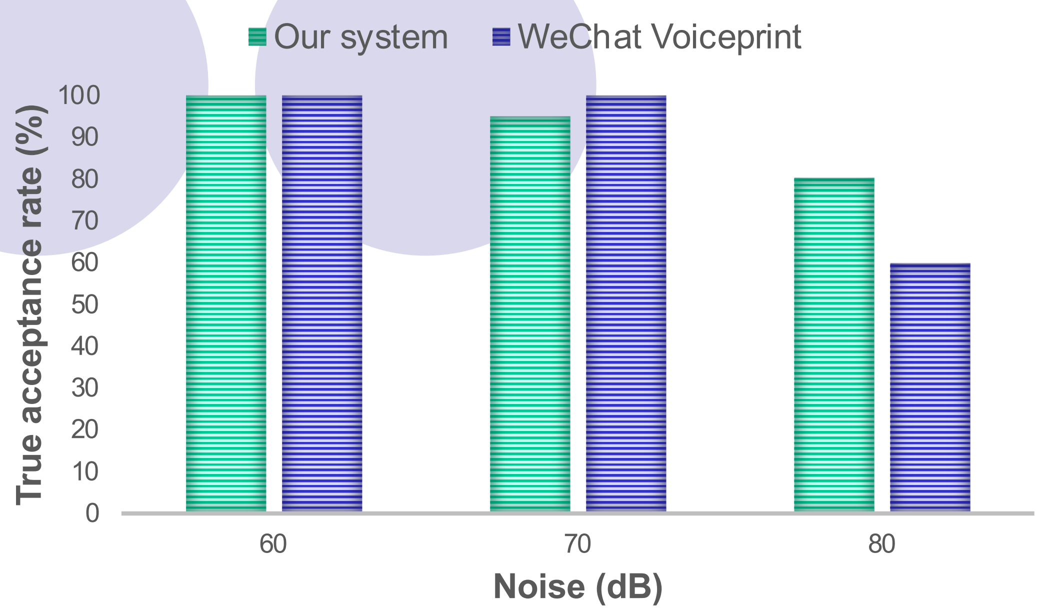
- Performance against attackers
  - Mimicry attack
  - Replay attack
  - Reconstruction attack
    - Attackers reconstruct victim's throat voice using low-pass filter



High true rejection rate of 100%

Evaluation

- Performance under different acoustic environments
  - When noise is under 70 dB, both systems can ensure at least 95% TAR for normal users
  - When the environment is pretty noisy, our system can provide 20% higher TAR than WeChat Voiceprint



Conclusion

- Smartphone-based liveness detection system
  - Leveraging microphones and motion sensors in smartphone – without additional hardware
  - Easy to integrate with off-the-shelf mobile phones (software-based approach)
- Good performance against strong attackers
  - Can detect a live speaker with mean accuracy of 99.04% and reject an attacker with an accuracy of 100%.

Q&A