Defending Against Voice Spoofing: A Robust Software-based Liveness Detection System

Jiacheng Shang, Si Chen, and Jie Wu

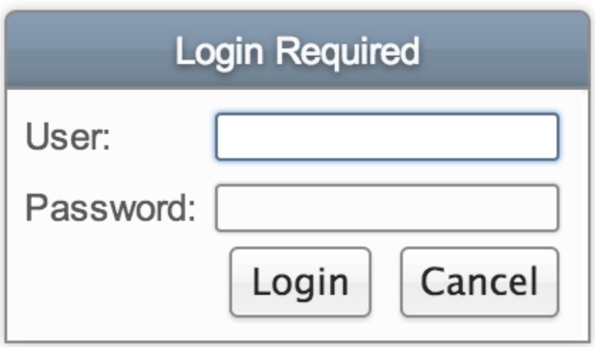
Center for Networked Computing

Dept. of Computer and Info. Sciences

Temple University

Biometrics: Voiceprint

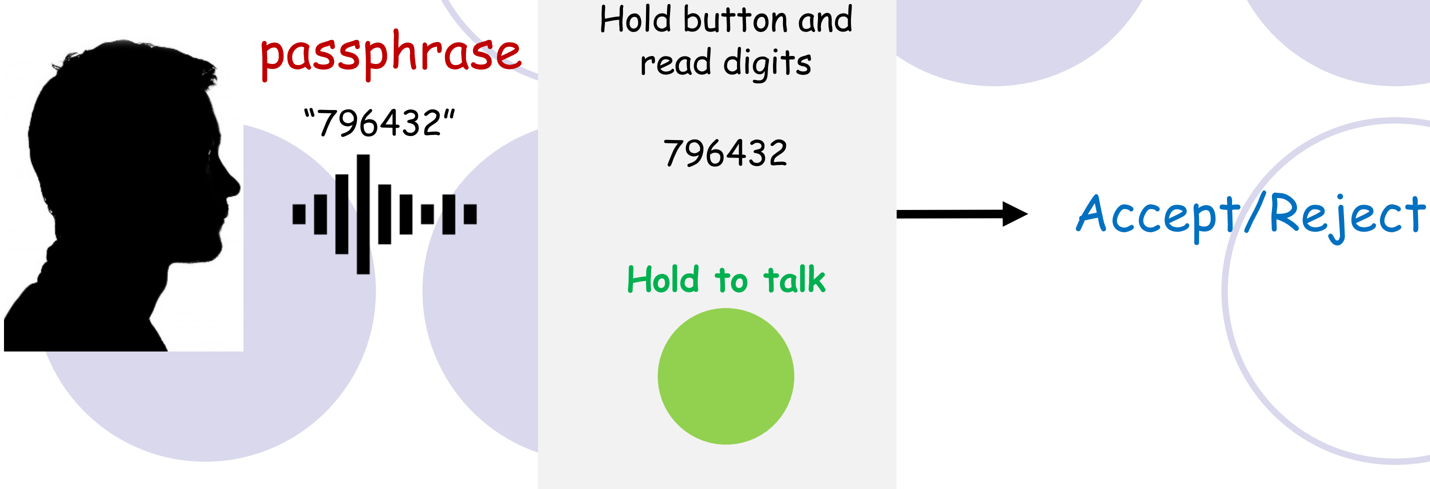
- Voiceprint
  - Promising alternative to password
  - Primary way of communication
  - Better user experience
  - Integration with existing techniques for multi-factor authentication



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

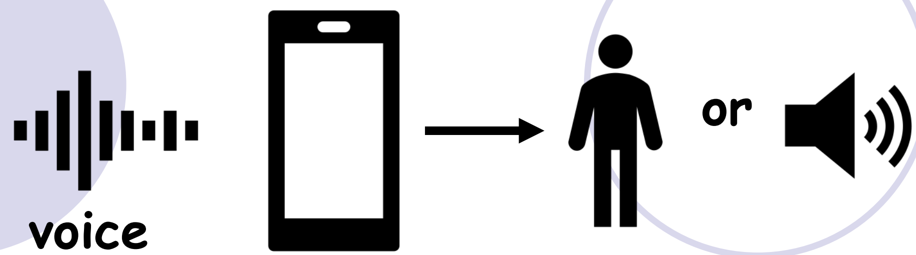
Steal voice

Replay to voice-based authentication systems

Reverse Turing Test

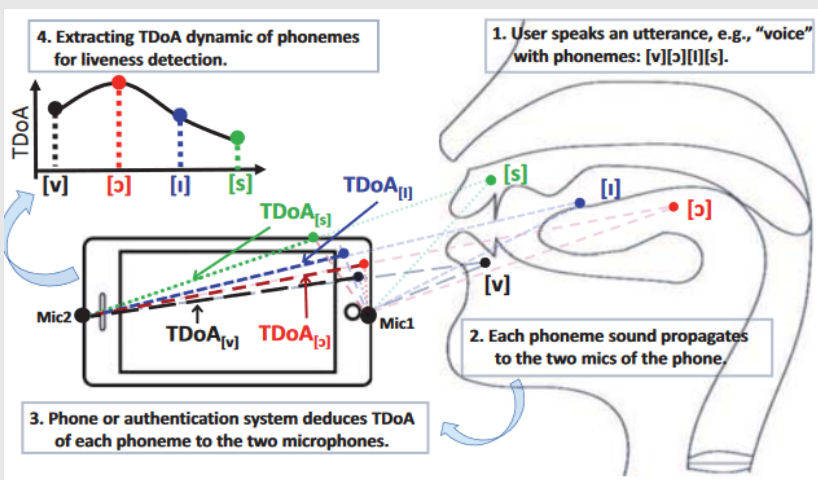
CAPTCHA

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

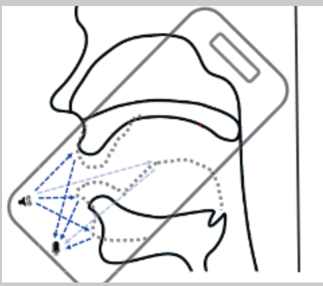
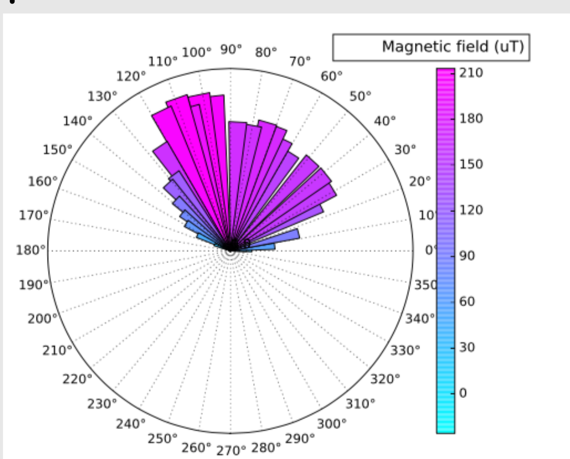


Voiceprint-based authentication

Previous work

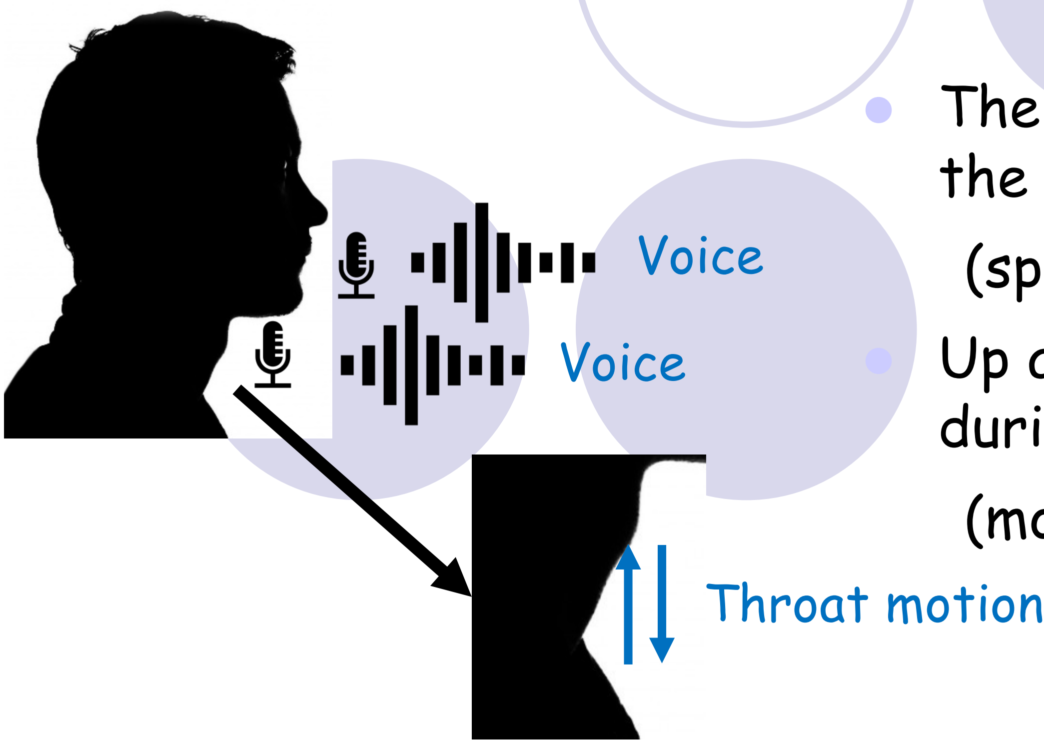
Systems	Limitations
Automatic speaker verification	• Verifying the speaker's identity (Bob or Alice) • Cannot defend against replay attack
Phoneme localization-based liveness detection (distance) 	• Low true acceptance rate (TAR): the smartphone needs to be static relative to the mouth VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones (L. Zhang et al. CCS 2016)

Previous work

Systems	Limitations
Articulatory gesture-based liveness detection (e.g. lip motion)  (Doppler effect)	• 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)
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)

Basic idea

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



- The voices at the mouth and the throat are different

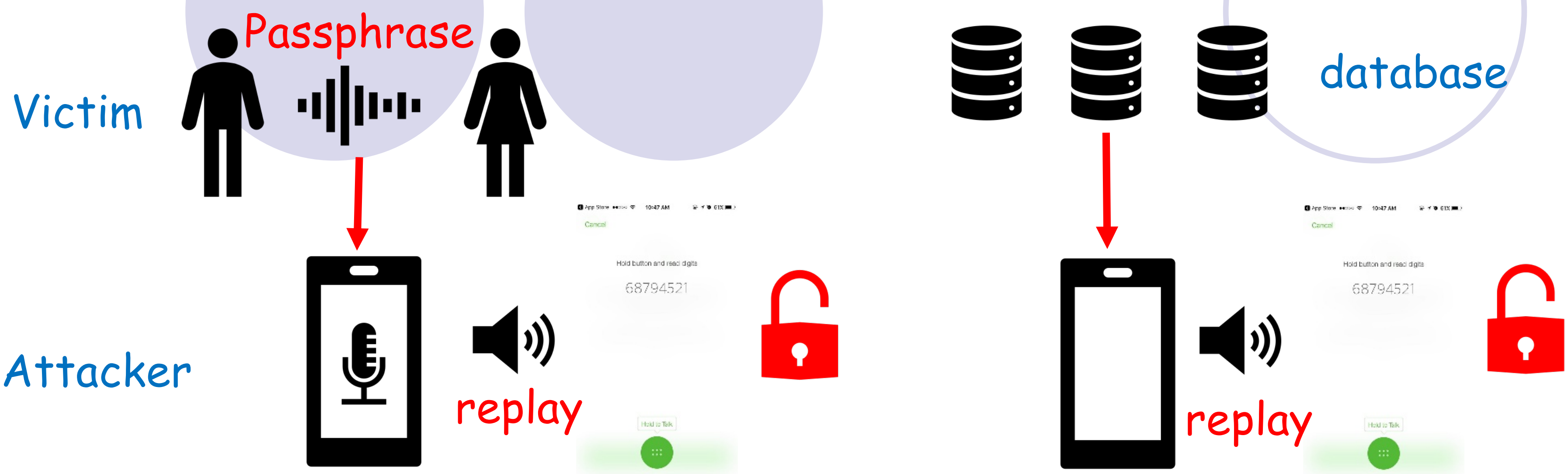
  (spectrum-based approach)

- Up and down motions exist during speaking

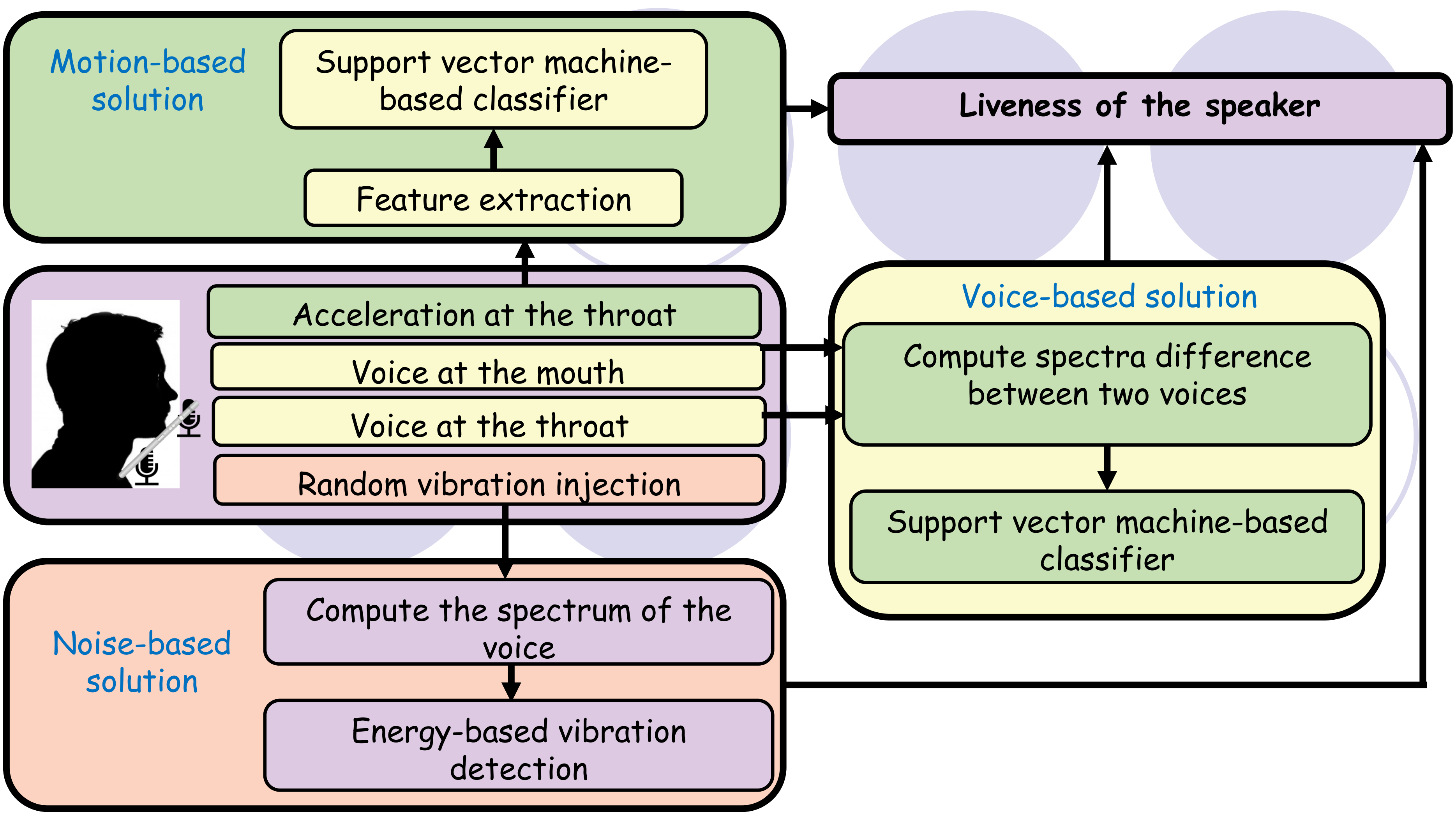
  (motion-based approach)

Attack model

- Attack model:
  - A simple replay attack: only stealing victim's voice at the mouth and replaying it
  - A strong replay attack: stealing victim's throat motions and voices at both mouth and throat from the database and replaying it

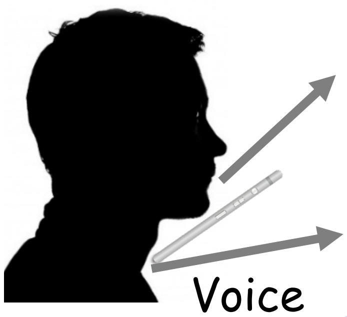


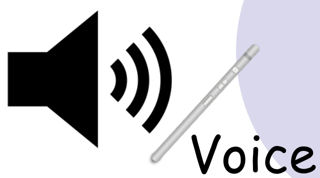
System Architecture

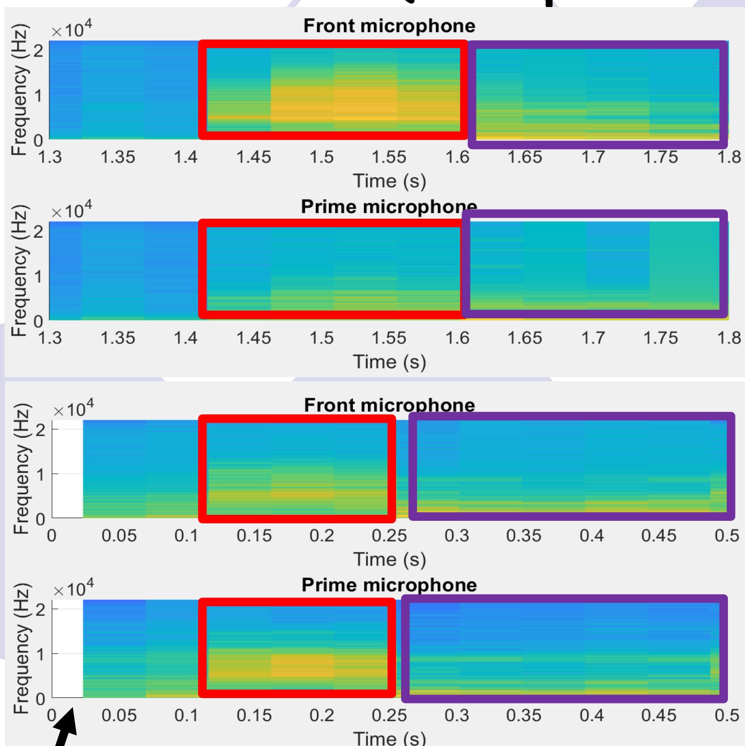


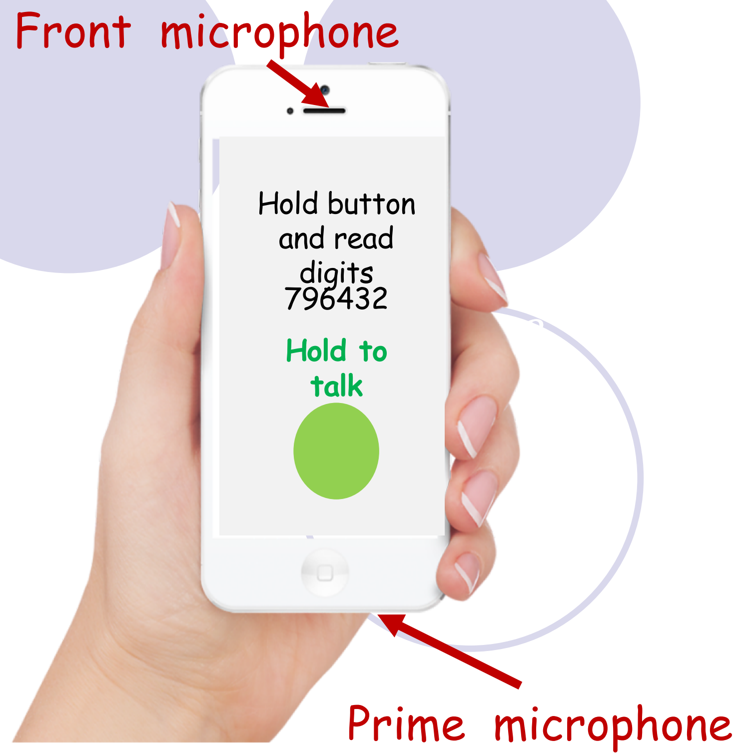
Proposed solutions

- Voice-based solution (Simple attack model)









Computing the spectra using Short-time Fourier transform (STFT)

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

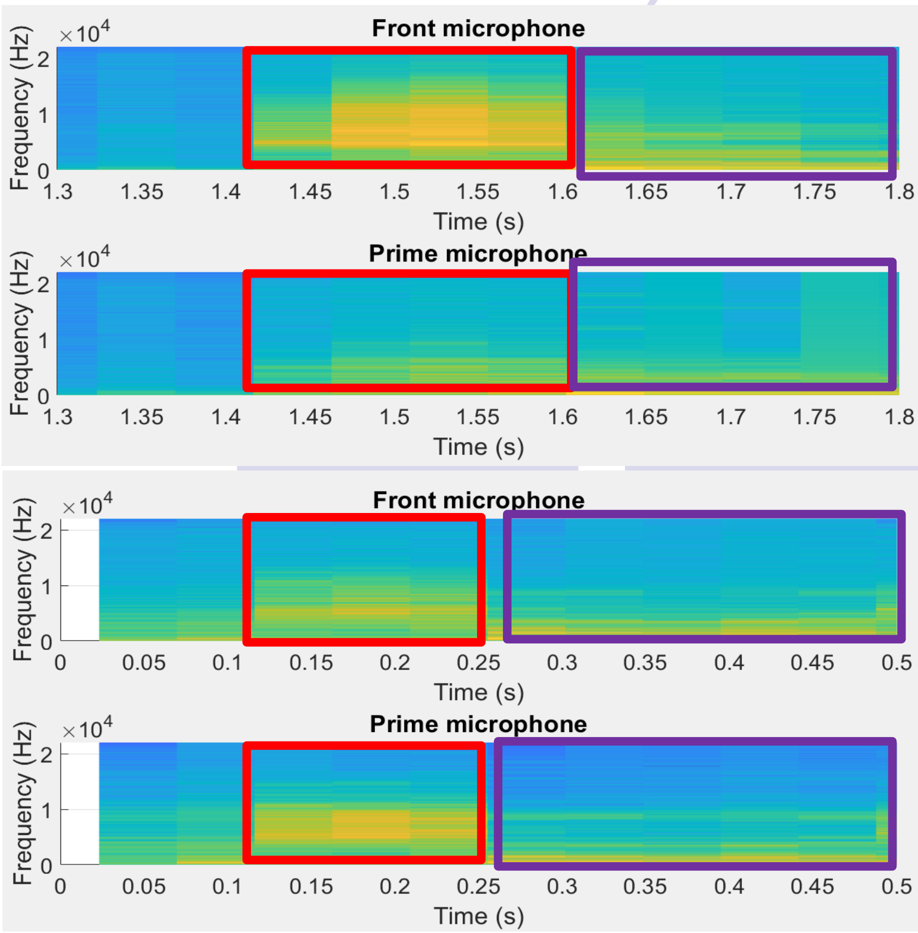
Convolution

Time domain to frequency domain

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

Proposed solutions

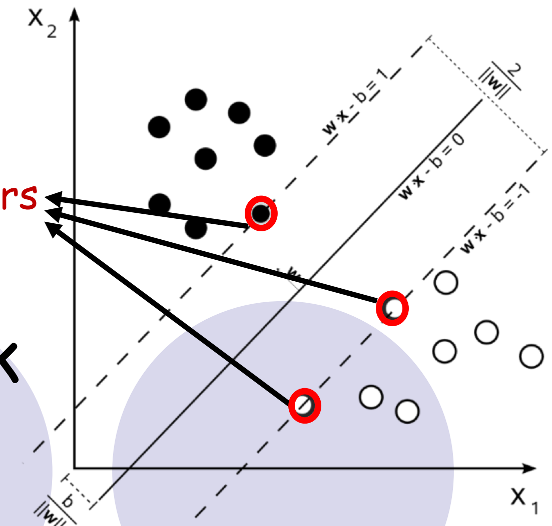
- Voice-based solution for simple attack

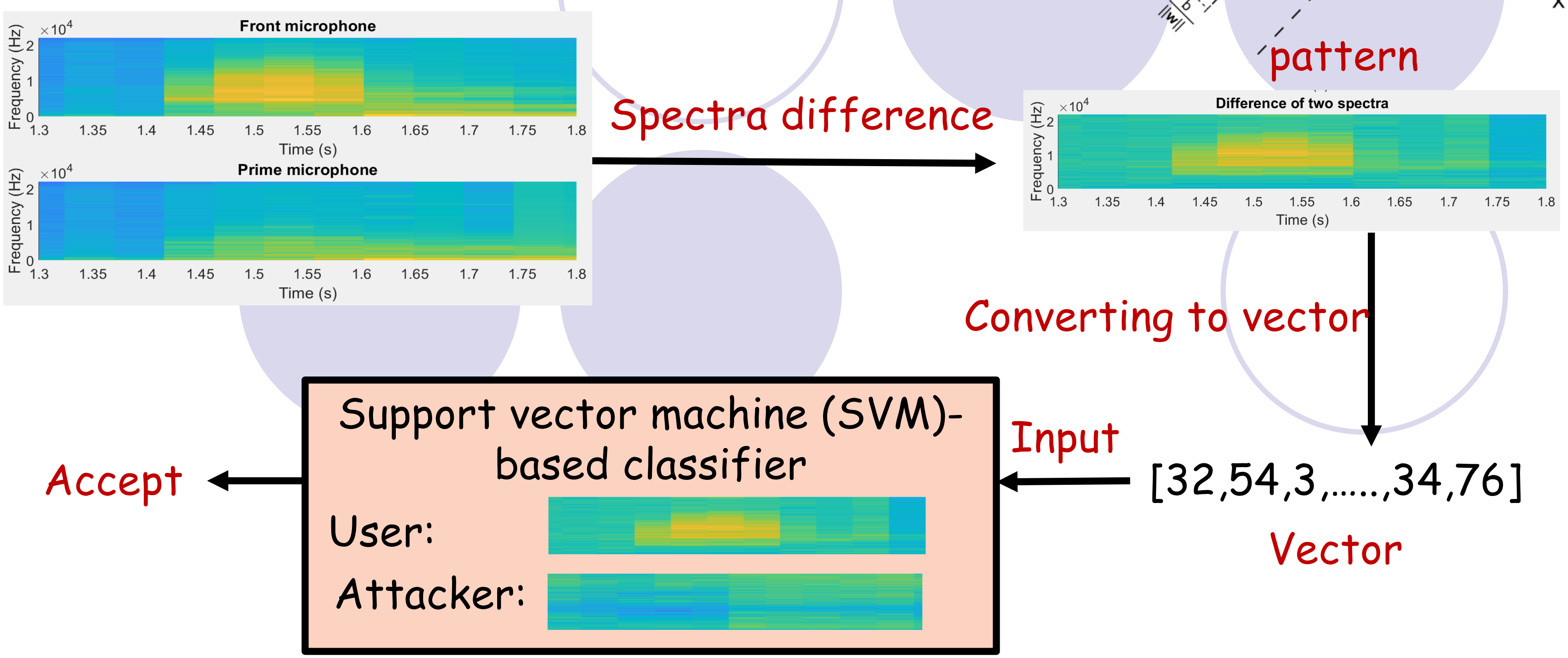


- Normal user: two voices are different
  - The voice (prime microphone) does not contain information of the unvoiced part.
  - The voice (prime microphone) contains low-frequency information of the voiced part.
- Attacker: two voices are similar
  - The voice (prime microphone) contains information of the unvoiced part.
  - The voice (prime microphone) contains most information of the voiced part.

Proposed solutions

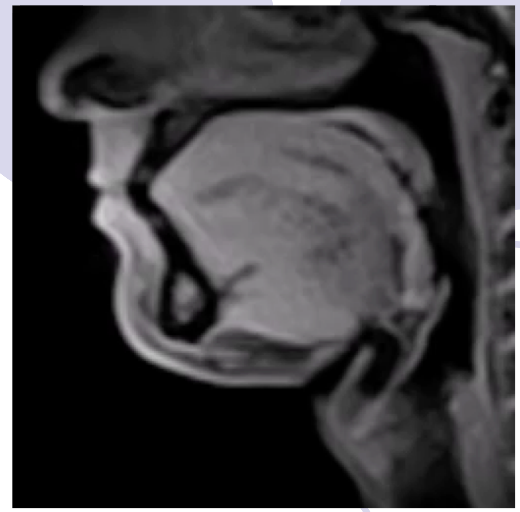
- Voice-based solution for simple attack
  - For normal users



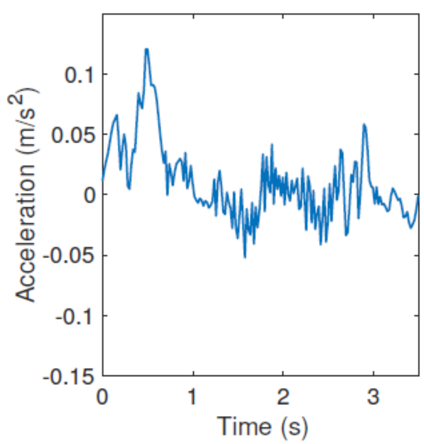


Proposed solutions

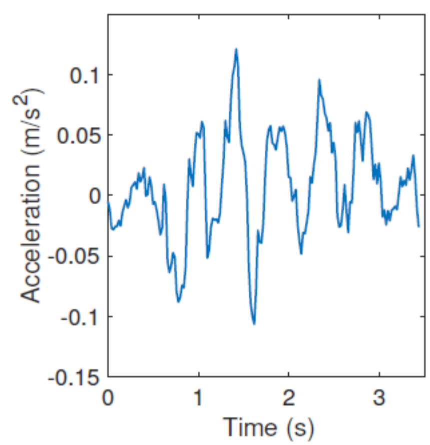
- Motion-based solution for simple attack
  - Using accelerator to capture throat motions
  - 7 features: Variance, minimum, maximum, mean, skewness, kurtosis, standard deviation
  - SVM-based classification model for decision



Normal users

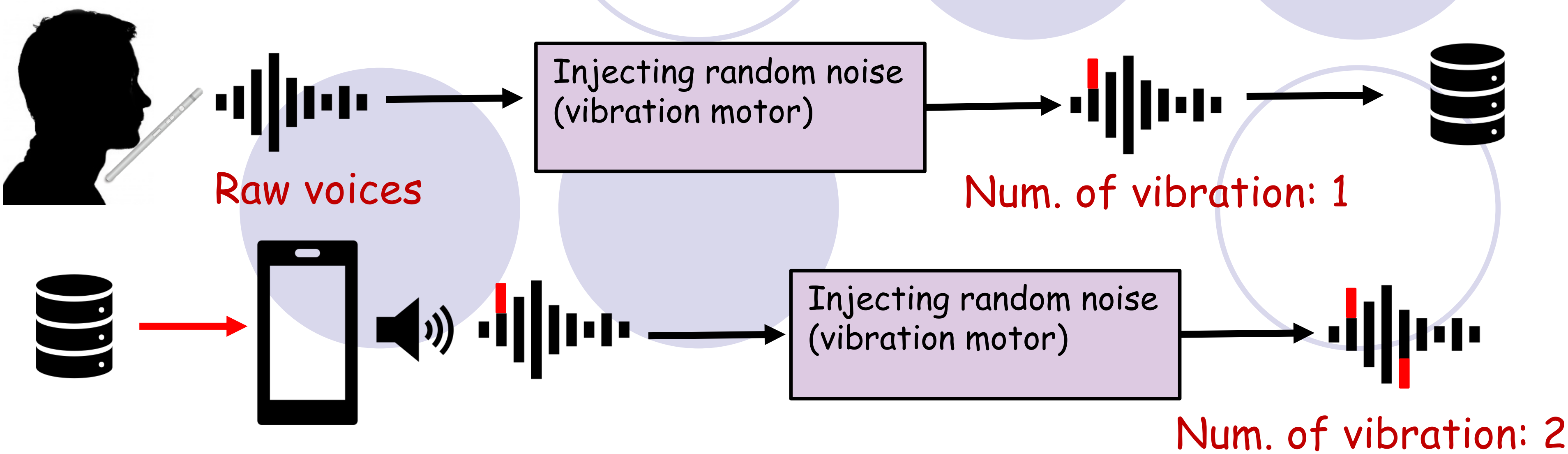


Attacker



Proposed solutions

- Random noise-based solution for strong attack
  - Attackers who can steal victim's voices and throat motions from the database and use multiple loudspeakers to imitate the victim

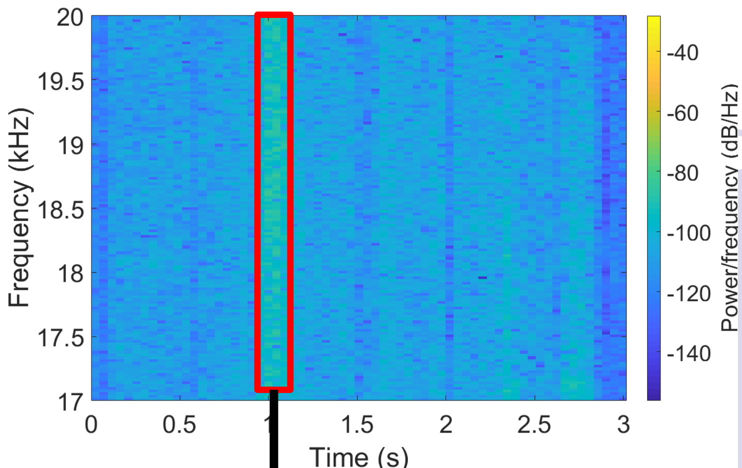


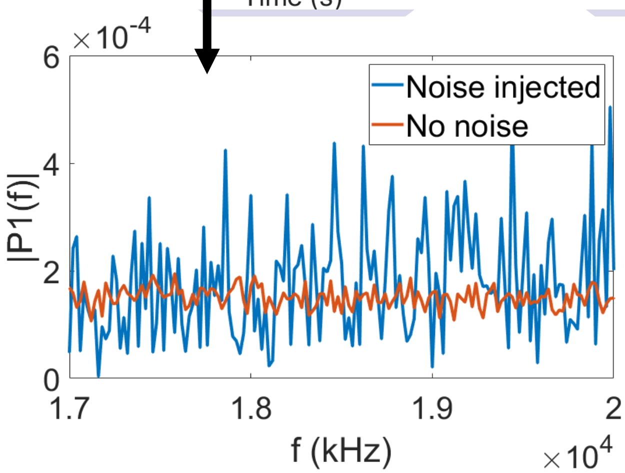
- Our solution:
  - Injecting a random vibration while the user is speaking
  - Checking the number of vibration in the voices

Proposed solutions

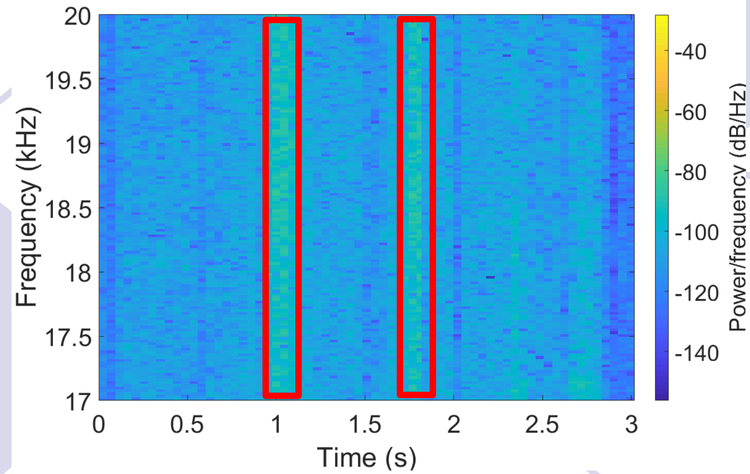
Computed by STFT

- Random noise-based solution
  - For normal users





For the attacker



  - The vibration introduces high energy to the high-frequency band.
  - A vibration is detected if the energy of a moving window exceeds a threshold.

Evaluation

- Methodology
  - Implementing our system on real smartphones
  - Using two loudspeakers 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)

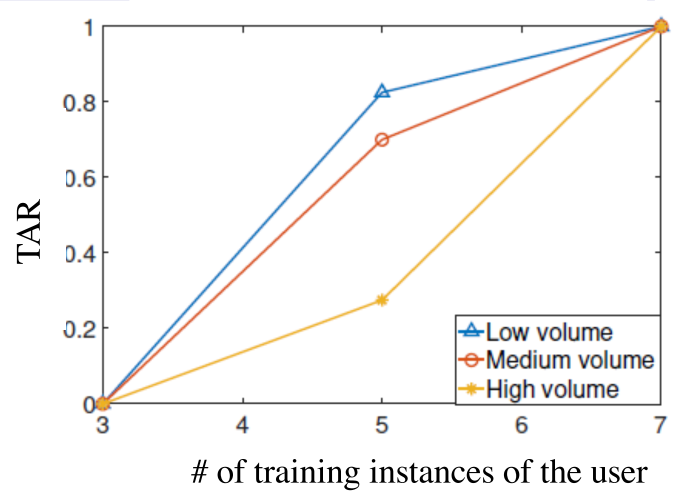
Evaluation

- Influence of locations on random noise-based approach

Locations	TAR	TRR
1	100%	100%
2	100%	100%
3	100%	100%
4	97.5%	100%

- Influence of acoustic noise on spectrum-based approach

  7 training instances from the user are sufficient



Evaluation

- Overall performance
  - Simple replay attack

Solutions	TAR	TRR	Computation cost
Voice-based	100%	100%	Medium (SVM+STFT)
Motion-based	93.3%	88.93%	Low (SVM)

  - Strong replay attack

Solutions	TAR	TRR	Computation cost
Voice-based & random noise	97.5%	100%	High (SVM+2*STFT)
Motion-based & random noise	91.0%	100%	Medium (SVM+STFT)

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

Q&A