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

Passphrase: "796432"

Hold button and read digits

796432

Hold to talk

Accept/Reject

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

Victims

Passphrase

Attacker

Steal voice

Replay to voice-based authentication systems
Reverse Turing Test

**CAPTCHA**

Completely Automated Public Turing test to tell Computers and Humans

Voiceprint-based authentication
## Previous work

<table>
<thead>
<tr>
<th>Systems</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| **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

<table>
<thead>
<tr>
<th>Systems</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Articulatory gesture-based liveness detection (e.g. lip motion)</strong></td>
<td>• Low true acceptance rate (TAR): the smartphone needs to be static relative to the mouth</td>
</tr>
<tr>
<td><img src="image" alt="Doppler effect" /></td>
<td><strong>Hearing Your Voice Is Not Enough: An Articulatory Gesture Based Mobile Voice Authentication</strong> (L. Zhang et al. CCS 2017)</td>
</tr>
</tbody>
</table>
| **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 |
| ![Magnetic field](image) | **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

Motion-based solution
- Support vector machine-based classifier
- Feature extraction

Voice-based solution
- Acceleration at the throat
- Voice at the mouth
- Voice at the throat
- Random vibration injection
- Compute spectra difference between two voices
- Support vector machine-based classifier

Noise-based solution
- Compute the spectrum of the voice
- Energy-based vibration detection

Liveness of the speaker
Proposed solutions

- Voice-based solution (Simple attack model)

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

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

- **Convolution**
- **Time domain to frequency domain**

- **x[n]: voice**
- **w[n]: window**
- **\(\omega\): 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

![Spectra difference](image)

Input → Converting to vector → Support vector machine (SVM)-based classifier

User: [32, 54, 3, ..., 34, 76]

Accept

Supporting vectors
Proposed solutions

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

![Acceleration graphs for normal users and attacker](image-url)
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

- **Random noise-based solution**
  - For normal users
  - The vibration introduces high energy to the high-frequency band.
  - For the attacker
  - A vibration is detected if the energy of a moving window exceeds a threshold.

Computed by STFT
Evaluation

- **Methodology**
  - Implementing our system on real smartphones
  - Using two loudspeakers to perform replay attack

<table>
<thead>
<tr>
<th>Maker</th>
<th>Model</th>
<th>Number of trumpets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willnorn</td>
<td>SoundPlus</td>
<td>2</td>
</tr>
<tr>
<td>Amazon</td>
<td>Echo</td>
<td>2</td>
</tr>
</tbody>
</table>

- **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

<table>
<thead>
<tr>
<th>Locations</th>
<th>TAR</th>
<th>TRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>97.5%</td>
<td>100%</td>
</tr>
</tbody>
</table>

- 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