VentricLock:
Exploring voice-based authentication systems

Chaouki Kasmi & José Lopes Esteves
ANSSI, France
WHO WE ARE

Chaouki Kasmi and José Lopes Esteves

- ANSSI-FNISA / Wireless Security Lab
- Electromagnetic threats on information systems
- RF communications security
- Embedded systems
- Signal processing
AGENDA

- Context: Voice command interpreters
- Voice as biometrics
- From brain to computer’s model
- Testing voice authentication engines
- Conclusion and future work
Voice Command Interpreters

Definitions and security analysis
VOICE COMMAND INTERPRETERS

Where?

Who?

What?

APIs
Google’s Super Bowl ad accidentally set off a lot of Google Homes

Burger King triggers Google Home devices with TV ad

Amazon’s Alexa started ordering people dollhouses after hearing its name on TV

Siri opens “smart” lock to let neighbor walk into a locked house
THREAT OF UNAUTHORIZED USE

- Silent voice command injection with a radio signal by front-door coupling on headphones cables [5]
THREAT OF UNAUTHORIZED USE

- Silent voice command injection with a radio signal by back-door coupling [6]
THREAT OF UNAUTHORIZED USE

- Malicious application playing voice commands through the phone’s speaker [1]
- Mangled commands understandable by the system but not the user [3]
- Same technique, embedded in multimedia files [2,4]
SECURITY IMPACTS

- Tracking
- Eavesdropping
- Cost abuse
- Reputation / Phishing
- Malicious app trigger/payload delivery
- Advanced compromising
- Unauthorized use of applications / services / smart devices…
SECURITY MEASURES

- Personalize keyword
- Carefully choose available commands (esp. Pre-auth)
- Limit critical commands
- Provide finer-grain settings to user
- Enable feedbacks (sound, vibration…)
- Voice recognition
Voice as biometrics

Using voice for authentication
"automated recognition of individuals based on their biological and behavioural characteristics"

"biological and behavioural characteristic of an individual from which distinguishing, repeatable biometric features can be extracted for the purpose of biometric recognition"

Biometrics

Physical
- Head
  - Face
  - Iris
  - Ear
  - Etc.
- Hand
  - Fingerprint
  - Palmprint
  - Hand geometry
  - Vein pattern
  - Etc.
- Others
  - DNA
  - Etc.

Behavioral
- Voice
- Others
  - Writing
  - Typing
  - Gait
  - Etc.
ENROLLMENT

Application

Acquisition → Signal processing → Feature extraction → Template / Model

Acquisition → Signal processing → Feature extraction → Comparison / Decision
VOICE BIOMETRICS

Applications:
- Speaker verification/authentication,
- Speaker identification...

Two main cases:
- Text independent
- Text dependent

http://www.busim.ee.boun.edu.tr
VOICE BIOMETRICS

Applications:

- Speaker verification/authentication,
- Speaker identification...

Two main cases:

- Text independent
- Text dependent

http://www.busim.ee.boun.edu.tr
VOICE BIOMETRICS

Enrollment
- 3 to 5 repetitions of the keyword

Model derivation
- The more samples, the more reliable

Speaker verification
- A comparison metrics and a threshold
VOICE BIOMETRICS

- **Pros:**
  - Acquisition device (microphone) widespread and low cost
  - Remote operation possible and natively supported

- **Cons:**
  - Voice changes over time (accuracy vs. usability)
  - Malicious acquisition very easy
  - Generation, modification tools available
  - Submission of test vectors affordable (speaker)
  - Liveness detection not trivial
VOICE BIOOMETRICS

- Reliability issues:
  - “At the present time, there is no scientific process that enables one to uniquely characterize a person’s voice” (2003) [10]
  - “Especially when:
    - The speaker does not cooperate
    - There is no control over recording equipment
    - Recording conditions are not known
    - One does not know if the voice was disguised
    - The linguistic content is not controlled”
VOICE BIOMETRICS

Reliability issues:

<table>
<thead>
<tr>
<th>BIOMETRIC</th>
<th>FINGERPRINT</th>
<th>FACE</th>
<th>HAND GEOMETRY</th>
<th>IRIS</th>
<th>VOICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barriers to universality</td>
<td>Worn ridges; hand or finger impairment</td>
<td>None</td>
<td>Hand impairment</td>
<td>Visual impairment</td>
<td>Speech impairment</td>
</tr>
<tr>
<td>Distinctiveness</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Permanence</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Collectibility</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Performance</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Acceptability</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Potential for circumvention</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Extract from [12]
From brain to computer’s model

Feature extraction techniques
FROM BRAIN TO COMPUTER’S MODEL

- Voice characteristics
- What we hear?

Dan Jurafsky

“Lecture 6: Feature Extraction and Acoustic Modeling”
Voice characteristics – Specificities

- Signal processing of non-stationnary signals
- Characteristics function of the time
Voice characteristics – Specificities

- Sensitivity of human hearing not linear
- Less sensitive at higher frequencies > 1 kHz

Dan Jurafsky
“Lecture 6: Feature Extraction and Acoustic Modeling”
FROM BRAIN TO COMPUTER’S MODEL

- **Linear prediction cepstral coefficient (LPCC)**
  - Energy values of *linearly* arranged filter banks
  - Mimic the human speech production

- **Discrete Wavelet Transform (DWT)**
  - Decomposition *separates* the *lower* frequency contents and *higher* frequency contents.
  - Only the low pass signal is further split

- **Wavelet Packet Decomposition (WPD)**
  - *Low* and *High pass signals* are further split
FROM BRAIN TO COMPUTER’S MODEL

- Mel-frequency cepstral coefficients (MFCC)
  - Frequency bands are placed logarithmically
  - Model the human system closely
  - Easier to implement
  - Voice to text and voice recognition engines
  - Widely used for feature extraction (many papers published by voice recognition editors ex. Google)
Mel-frequency cepstral coefficients (MFCC)

- Preprocessing before feature extraction;
- Framing the signal are splits in time domain, then on each individual frame then windowing them;
- Converting each frame TD to FD with DFT;
- Filter bank is created by calculating number of picks spaced on Mel-scale and again transforming back to the normal frequency scale;
- Converting back the mel spectrum coefficient to TD coefficient to the time domain with Discrete Cosine Transform
Testing voice authentication engines

Testing in a black-box context existing solutions
TESTING APPROACH

- We consider the verification system as a black box
- We use publicly available toolsets
- We set up test scenarios based on the attack’s prerequisites
  - Knows target language?
  - Knows target’s keyword?
  - Possesses target’s voice samples?
EXPERIMENTAL SETUP

- Target 1 (Siri)
- Target 2 (S-voice)
- Target 3 (Google now)
The attacker hears the target saying the keyword
He tries to impersonate the target’s voice

We are not professional impersonators
But we succeeded on all tested targets
  - Within less than 15 attempts
TESTS: REPLAY

- The attacker has a recording of the target saying the keyword
- Our demo last year at Hack In Paris [6]
The attacker has a recording of the target saying the keyword

Our demo last year at Hack In Paris [6]

Additionnal tests

Looking to boundaries with legit sample modifications (Filtering, Pitch, Time-Scale, SNR)

Target 1 (Siri) is shifting pre-auth. ????
TESTS: MODEL SHIFTING

- The attacker knows the keyword
- If the model is updated for each submitted sample
- It can shift so as to accept any voice sample
- By submitting the same sample repeatedly until it passes the authentication
## Tests: Model Shifting

- **Results related to target 1**
  - Try 1 : 10 use by legit user
  - Try 2 : 50 use by legit user
  - Number of try required to trigger target 1
  - Legit user still able to trigger target 1 (+ OK, - NOK)

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1, +</td>
<td>1, +</td>
<td>16, +</td>
<td>25, +</td>
<td>101, -</td>
<td>21, +</td>
<td>34, -</td>
<td>70, +</td>
<td>385, -</td>
</tr>
<tr>
<td>1 bis</td>
<td>1, +</td>
<td>4, +</td>
<td>30, +</td>
<td>48, +</td>
<td>98, -</td>
<td>33, -</td>
<td>24, +</td>
<td>54, -</td>
<td>402, -</td>
</tr>
</tbody>
</table>
TESTS: TD RECONSTRUCTION

- The attacker knows the keyword
- The attacker has a recording of the target
- Contains all the phonemes of the keyword
- He reconstructs the keyword by concatenating the phonemes in time domain

Video 1
The attacker knows the MFCC features extracted from the target pronouncing the keyword.

He can modify the MFCC and reconstruct several time domain samples from the features [3,4].
TESTS: FD RECONSTRUCTION

- MFCC and MFCC inverse
  - MFCC inverse of legit user
  - MFCC inverse of a composition of samples

- Targets 2 and 3: the MFCC seems to contain enough of the information required to authenticate

Video 2
- The attacker knows the keyword
- He has access to several other voice samples saying the keyword
- He generates test vectors by superimposing several voice samples
TESTS: KEYWORD COMPOSITION

Enrolled voice

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 try

2-100 try

> 100 try – NOK

Video 3
## TESTS: KEYWORD COMPOSITION

### Superimposed voices

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Enrolled voice

- **1 try**: Green
- **2-100 try**: Yellow
- **> 100 try – NOK**: Red

### Removed

- **Removed**
## TEST SUMMARY

<table>
<thead>
<tr>
<th>Test</th>
<th>Target 1 (Siri)</th>
<th>Target 2 (S-voice)</th>
<th>Target 3 (Google now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impersonation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Replay</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TD reconstruction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FD reconstruction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Keyword composition</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Model shifting</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Conclusion and future work
CONCLUSION

- Voice interface is getting widespread
  - Devices without other UI
  - Hands free commodity
- Available commands and actions
  - Getting richer and more critical
- Voice authentication seemed as a viable countermeasure
- But it is still inefficient and immature
- Gives a false security feeling
TEST RESULTS LIMITATIONS

- Results cannot be generalized, depend on
  - Language
  - Keyword
  - Legitimate user’s voice during enrollment
  - Model and decision metrics

- And we don’t know the model
  - It can be updated
  - There can be several models/approaches
  - It can shift

- That’s why we don’t provide result statistics
COUNTERMEASURES

- Prevent unlimited successive failed authentication attempts (as Google does)
- Prove the command originates from the user:
  - Use the phone’s sensors [7, 9]
- Add entropy and interaction
  - N-staged process, with challenge-response
- Enhance the user’s voice model
  - Qualcomm patent: continuous voice authentication [8]
The attacker knows the keyword
He has access to several other voice samples saying the keyword
He extracts features for all samples and generates test vectors from statistical characteristics of the features distribution
Trying to preserve the keyword recognition
<table>
<thead>
<tr>
<th>Test</th>
<th>Target 1 (Siri)</th>
<th>Target 2 (S-voice)</th>
<th>Target 3 (Google now)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impersonation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Replay</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TD reconstruction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FD reconstruction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Keyword composition</td>
<td>✓</td>
<td>✓</td>
<td>?</td>
</tr>
<tr>
<td>Model shifting</td>
<td>✓</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Features bruteforce</td>
<td>WIP</td>
<td>WIP</td>
<td>WIP</td>
</tr>
</tbody>
</table>
OPEN QUESTIONS

- Is it possible, for a given language and keyword:
  - To generate a « masterkey »?
  - To derive a verified sample by bruteforce? At which complexity?

- Is it possible, knowing the model and features:
  - To estimate the probability and/or the number of masterkeys?
  - To estimate the robustness of the authentication system against impersonation?

- Can voice authentication vendors tell us:
  - How easily can it be circumvented according to my language, keyword and voice characteristics?
  - And how confident could we be about the answer?
Voice command usability vs. security

Apple response to our disclosure:

« Voice recognition in Siri is not a security feature »
...the federal CFAA (Computer Fraud and Abuse Act) broadly prohibits anyone from accessing a computer without authorization. There’s no doubt that Google Home and its associated Google-based systems are computers, and I know that I didn't give Burger King permission to access and use my Google Home or my associated Google account. Nor did millions of other users. And it’s obvious that Google didn’t give that permission either.

- By using unsecure settings, does the user give permission to access the system?
Thank You

We thank the manufacturers and the editors for their interesting feedbacks.
REFERENCES


QUESTIONS ?

- José Lopes Esteves, jose.lopes-esteves@ssi.gouv.fr
- Chaouki Kasmi, chaouki.kasmi@ssi.gouv.fr

---

Your voice is your password.
Introducing Voice ID.

Shawn McMahon @syberghost · Jun 12
Hey HSBC, I accidentally used my voice in public and now it’s compromised. How do I change it?

HSBC US @HSBC_US · Jun 12
Hi! Did you mean, you were calling at a public place and they may have heard you say your voice id? 🤔 RC

Shawn McMahon @syberghost

Replying to @HSBC_US
No, I just used my voice in public, and you guys are using those as passwords now, so I need to know how to change my voice.