Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity

Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen
Illinois Institute of Technology

Taeho Jung
University of Notre Dame

Xiang-Yang Li
University of Science and Technology of China
Introduction

Building Blocks:
• Voice conversion
• Application scenarios

Approach:
• Attack basic voice conversion
• Compound warping functions
• Add differential privacy
• Security analysis

Evaluation
Voice is becoming a major way of human-computer interaction for mobile devices
Application: keyboard apps

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Application: AI assistants

Hi, I’m Cortana.

"Ok Google"

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Our voice data is being stored...

Apple stores your voice data for two years

The iPhone and iPad maker holds on to the data from Siri and Dictation for two years, so long as it abides by its own privacy policy — which, as you might expect, is fairly vague.

... and shared to third parties

The Hacker News

Apple Admits Siri Voice Data is Being shared with Third Parties

March 11, 2015  Wang Wei

Samsung Admits Private Conversations Being sent to Internet Server

Smart TVs prompt privacy backlash

Paul Joseph Watson - FEBRUARY 9, 2015  0 Comments
Existing voice input scenario

User
Mic

Speech recognition

Voice input cloud

Voice input app

Voice input service provider

Apps waiting for input

Speech
Text

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Security risks

Voiceprint (voice fingerprint) can be extracted

Fake speech can be generated for spoofing, fraud, framing, blackmail, etc.

Voiceprint

Voice cloning

Speech synthesis

Once lost, always lost
Privacy risks

• Voice can be used for linkage attacks
  – When the cloud shares unlabeled speech recordings to 3rd parties
    • User identity can be recognized by speaker recognition
    • Sensitive speech content can be used for profiling, e.g. SMS, search history
Goal

• Main goal
  – To protect the voiceprints of voice input users from being disclosed while maintaining the user experience

• One stone, two birds
  – It strengthens identity privacy as well
    • Prevents 3rd parties from de-anonymize unlabeled speech data via speaker recognition
Challenges

• Existing privacy definitions do not apply
  – Speech = voice + text (speech content)
  – Privacy in text is already hard to define

• Voiceprints are hard to remove
  – Tightly coupled with speech content
  – Voice conversion can be reversed

• Need to preserve the user experience of voice input
  – Speech recognition accuracy, latency, power consumption
Introduction

Building Blocks:
• Voice conversion
• Application scenarios

Approach:
• Attack basic voice conversion
• Compound warping functions
• Add differential privacy
• Security analysis

Evaluation
Voice conversion

Pitch marking
• Detect signal periodicity

Segmentation
• Split signal into frames

FFT
• Convert to frequency domain

VTLN
• Warp the frequency axis

IFFT
• Convert to time domain

PSOLA
• Improve speech quality

Use a warping function
Warping function

Bilinear function \( f(\omega, \alpha) = \left| -i \ln \frac{e^{i\omega - \alpha}}{1 - \alpha e^{i\omega}} \right| \),
\( \omega \) is original frequency, \( \alpha \) is warping factor

Other warping functions: power, quadratic, symmetric, etc.
Rebuild the voice input scenario

- We design VoiceMask to anonymize the voice data before sending it to the cloud
Rebuild the voice input scenario (cont’d)

Comparison

Into the OS

More **practical**
Compatible with existing voice input services

As an app

More **secure**
Harder for the cloud to collect PII

Both protect voiceprints

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Introduction

Building Blocks:

• Voice conversion
• Application scenarios

Approach:

• Attack basic voice conversion
• Compound warping functions
  • Add differential privacy
  • Security analysis

Evaluation
Hide voiceprint with voice conversion

Adjust the warping factor $\alpha$

Demo

$\alpha$

-0.15
Low-pitched
(deep voice)

0

0.15
High-pitched
(sharp voice)
Find the proper range of $\alpha$:

- **Metrics**
  - Security: speaker recognition accuracy
  - Service: speech recognition accuracy

- **Balance security and voice input service**
  - Larger $|\alpha|$ → more distortion
  - → speaker/speech recognition accuracy both decreases
  - → better security $\smile$, worse service $\frown$

- **Luckily, speaker recognition accuracy decreases faster!**

- **Proper range:** $0.08 \leq |\alpha| \leq 0.10$
  - Set $\alpha < 0$ to deepen female voices
  - Set $\alpha > 0$ to sharpen male voice
Reversing attack

Problem:
- Warping function is invertible
  - \( f(\omega, \alpha, -\alpha) = \omega \)
- So voice conversion can be reversed!

Solution:
- Randomly select \( \alpha \) from the proper range for each speech recording
  - Every speech recording from the same speaker sounds different
Reducing attack

• Problem
  – Warping function is reducible
    • \( f(f(\omega, \alpha_1), \alpha_2) = f(\omega, \alpha_1 + \alpha_2) \)
  – Suppose \( \alpha \) is randomly chosen from \([0.08, 0.10]\)
  – The attacker can partially recover the voice by applying voice conversion with \( \alpha = -0.09 \)

• Solution
  – Compound 2 warping functions (see next slide)
Compound warping functions

• We introduce another warping function

  – Quadratic function \( g(\omega, \beta) = \omega + \beta \left( \frac{\omega}{\pi} - \left( \frac{\omega}{\pi} \right)^2 \right) \),

  \( \beta \) is the warping factor

  \[ \begin{aligned}
  &-1 & \text{Low-pitched (deep voice)} \\
  &0 & \text{0} \\
  &1 & \text{High-pitched (sharp voice)}
  \end{aligned} \]

• Compound bilinear and quadratic functions:

  \( h(\omega, \alpha, \beta) = g(f(\omega, \alpha), \beta) \)

  Two warping factors make reversing/reducing attacks harder
How to set $\alpha, \beta$ to properly distort the voice?

• First, define *distortion strength*!

• Intuition
  – The closer a warping function’s curve is to the identity function ($f(\omega) = \omega$), the less distortion it produces

• The *distortion strength* of $h$ is the area between the curves of $h$ and the identity function

$$\text{dist}_h = \int_0^\pi |h(\omega) - \omega|$$
How to set $\alpha, \beta$ to properly distort the voice?

- Recall for the bilinear function $f(\omega, \alpha)$, the proper range is $0.08 \leq |\alpha| \leq 0.10$
- By simple calculation, $dist_f(\omega, 0.08) = 0.32$, $dist_f(\omega, 0.10) = 0.40$
- So the proper range of distortion strength is $0.32 \leq dist \leq 0.40$

Choose $\alpha, \beta$ from this ring to produce a proper distortion.
Security analysis

• Resistance to the reversing attack
  – Impossible to guess $\alpha, \beta$
    • $\alpha, \beta$ is randomly chosen
  – Impossible to reverse $h(\alpha, \beta)$ with a distinct pair of $\alpha', \beta'$.
    • Even if they produce the same distortion strength, the distortion orientations are different
Security analysis

• Resistance to the reducing attack
  – Impossible to reduce distortion overall
    • $\alpha, \beta$ is randomly chosen from the ring for every speech recording. $\mathbb{E}(\alpha) = \mathbb{E}(\beta) = 0$
  – Unlikely to reduce distortion individually
    • The success rate of random guess is reduced to 35% from 100% (basic voice conversion)
Add differential privacy

- Add noise to the signal after IFFT
Add differential privacy

• Time-domain samples of the signal

\[ Y_j = \text{real} \left( \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{\frac{i2\pi kj}{N}} \right), \quad 0 \leq j < N. \]

• Add a Laplace noise to each sample

\[ \hat{Y}_j = Y_j + \text{Lap} \left( \frac{2\Delta m}{\epsilon} \right) \]

Now it’s even harder to recover the original voice!
Introduction

Building Blocks:
- Voice conversion
- Application scenarios

Approach:
- Attack basic voice conversion
- Compound warping functions
  - Add differential privacy
  - Security analysis

Evaluation
Datasets

- **PDA**
  - 16 speakers each speaking over 50 short sentences
- **LibriSpeech**
  - Audios of 251 native speakers reading an English book
- **Volunteers**
  - Collected by us from 14 students with various accents

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Speakers</th>
<th>#Speeches</th>
<th>Hours</th>
<th>English accents</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDA</td>
<td>16</td>
<td>836</td>
<td>1.8h</td>
<td>Mostly native</td>
</tr>
<tr>
<td>LibriSpeech</td>
<td>251</td>
<td>27.7k</td>
<td>100h</td>
<td>All native</td>
</tr>
<tr>
<td>Volunteers</td>
<td>14</td>
<td>240</td>
<td>0.7h</td>
<td>Various accents</td>
</tr>
</tbody>
</table>

For each speaker, use 5-10 speeches as the training set
Tools

• Microsoft Azure Cognitive Services
  – Speaker Recognition API
    • for speaker identification
  – Bing Speech API
    • for speech recognition
Steps of evaluation

1. **Build model**
   - Use the training set to train a voice model for every speaker

2. **Anonymize voice**
   - Use VoiceMask to distort the speeches in the test set

3. **Identify voice**
   - Try identifying the speaker of the distorted speeches

4. **Recognize speech**
   - Try recognize the speech and measure the accuracy

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Metrics

• Security
  – Speaker recognition accuracy
    • The fraction of correctly identified utterances. (We identify a speaker from a pool of 50 candidates)

• Service / User experience
  – Speech recognition accuracy
    • Word accuracy = 1 – word error rate
  – Real-time coefficient (Latency)
    • The ratio between processing time and audio’s duration
Performance

- Speaker recognition accuracy
  - Decreases by 84%
- Speech recognition accuracy
  - Decreases by 19%

![Accuracy Graph]

**Demo**

Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity
Performance (cont’d)

- On a laptop

- On Android phones

<table>
<thead>
<tr>
<th>Device</th>
<th>Memory</th>
<th>OS</th>
<th>Real-time coef.</th>
<th>Power cons.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Nexus 5</td>
<td>16 GB</td>
<td>Android 5.1.1</td>
<td>3.93 ± 0.45</td>
<td>0.70 W</td>
</tr>
<tr>
<td>Google Nexus 6</td>
<td>32 GB</td>
<td>Android 5.1.1</td>
<td>4.90 ± 0.15</td>
<td>0.78 W</td>
</tr>
<tr>
<td>Google Nexus 7</td>
<td>16 GB</td>
<td>Android 5.1.1</td>
<td>4.11 ± 0.03</td>
<td>0.83 W</td>
</tr>
<tr>
<td>MEIZU Pro 6</td>
<td>32 GB</td>
<td>Android 6.0</td>
<td>2.42 ± 0.05</td>
<td>0.49 W</td>
</tr>
</tbody>
</table>
Analysis of external factors

• Ambient noise
  – Home, office, car, outdoors
• Speaker’s motion
  – Walking, running, driving
• Speaker’s gender, accent
• Device brand

Please refer to the paper for more details
Q&A

Thank you for Listening!

Jianwei Qian
https://sites.google.com/view/jqian