VOICEFOX: LEVERAGING INBUILT TRANSCRIPTION TO ENHANCE THE SECURITY OF MACHINE-HUMAN SPEAKER VERIFICATION AGAINST VOICE SYNTHESIS ATTACKS

Maliheh Shirvanian†
mshirvan@visa.com
Visa Research

Manar Mohammed*
mohamem@MiamiOH.edu
Miami University

Nitesh Saxena
saxena@uab.edu
University of Alabama at Birmingham

S Abhishek Anand†
sanand53@bloomberg.net
Bloomberg LP

† Work done at UAB
* Now at Google
Speaker Verification

- Machine-Human based process of authenticating a claimant by their voice
- Applications: personalized virtual assistant apps and devices, secure authenticated key exchange

Diagram:
- Synthesized or Natural Human Voice
- Similarity Test (Classification)
- Speaker’s Reference Model
- Likelihood Score?
  - Reject
  - Accept
Is Voice Secure?

Voice Replay

Voice Synthesis

Voice Conversion

Mimicking

Inaudible Attacks
Contributions

- We offer Voicefox, a mitigation scheme against voice synthesis attacks, based on speech-to-text transcription
- We ran a study to measure the accuracy of speech-to-text techniques when confronted with synthesized samples
- We propose several post-transcription techniques to reduce False Reject and False Accept Rate
Threat Model

- Attacker has access to high-quality audio recording devices
- Attacker can collect a few minutes of the victim’s voice
- Attacker has access to off-the-shelf speech synthesis tools
Application 1: Machine-Auth
Application 1: Attack on Machine-Auth

Challenge-Response Passphrase
Application 2: Human-E2EE
Application 2: Attack on Human-E2EE
Voicefox Design

Synthesized or Natural Human Voice

Speaker Verification

Accept

Reference Text

Error Calculator

Accept

Reject

Comparator (Error < Threshold)

Yes

No

Accept

Reject
Decision Logic

- System rejects the input if the speaker verification does not verify user’s voice, for accepted samples:
- Voicefox compares output of transcriber against the reference text
- Errors higher than a defined threshold indicate a possible attacks

**Improvements with post transcription processing:**
- To reduce FRR, ignore transcription error of up to 1/2 input size
- To reduce FAR, reject if transcribed to words available in dictionary
Rationale of Our Work

Limitations of Speech Synthesis Attack

- **Quality**, requires large high quality training data, affect transcription accuracy
- **Intelligibility**, requires large training data, affect transcription accuracy
- **Similarity**, affects speaker verification

Speech Transcription

- Speech to text transcribers are already built in application
- Accuracy of the transcribers, to a great extent, depends on the training data (natural voice)

Spectrum Differences (next)
Spectrum Differences: Natural vs. Synthesized Voices

Power spectral density spectrum of the original speech sample arctic_a0101 for speaker "sit"

Power spectral density spectrum of the synthesized speech sample generated using Festvox Transform
Study Design


- **4 Datasets**: Natural Speaker’s Voice, Festvox Voice Conversion, Lyrebird Speech Synthesis, Google Speech Synthesis

- **3 Speech Transcription Tools**: Google Cloud Speech-to-Text, IBM Watson Speech-to-Text, Mozilla DeepSpeech

- **Metrics**: Word Error Rate (WER), Word Recognition Rate (WRR), Sentence Error Rate (SER), False Reject Rate (FRR), False Accept Rate (FAR)
(a) Dataset 1: Natural Voice Samples – Sentences

- IBM Watson Speech-to-Text
- Google Cloud Speech-to-Text
- Mozilla DeepSpeech

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>WRR</th>
<th>FRR (=SER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>8.09%</td>
<td>10.58%</td>
<td>14.51%</td>
</tr>
<tr>
<td>Google</td>
<td>93.27%</td>
<td>90.47%</td>
<td>87.47%</td>
</tr>
<tr>
<td>Mozilla</td>
<td>39.60%</td>
<td>52.40%</td>
<td>64.40%</td>
</tr>
<tr>
<td>Voicefox</td>
<td>1.20%</td>
<td>1.80%</td>
<td>4.40%</td>
</tr>
</tbody>
</table>

(a) Dataset 3: Lyrebird Synthesized Samples – Sentences

- IBM Watson Speech-to-Text
- Google Cloud Speech-to-Text
- Mozilla DeepSpeech
- Voicefox

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>WRR</th>
<th>FAR (=1-SER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>20.82%</td>
<td>30.07%</td>
<td>22.42%</td>
</tr>
<tr>
<td>Google</td>
<td>80.73%</td>
<td>71.76%</td>
<td>79.17%</td>
</tr>
<tr>
<td>Mozilla</td>
<td>34.10%</td>
<td>20.00%</td>
<td>22.30%</td>
</tr>
<tr>
<td>Voicefox</td>
<td>12.10%</td>
<td>11.16%</td>
<td>4.94%</td>
</tr>
</tbody>
</table>
(b) Dataset 1: Natural Voice Samples – PGP Words

IBM Watson Speech-to-Text
Google Cloud Speech-to-Text
Mozilla DeepSpeech

WER
WRR
FRR (=SER)
FRR_{Voicefox}

(b) Dataset 4: Tacotron Synthesized Samples – PGP Words

IBM Watson Speech-to-Text
Google Cloud Speech-to-Text
Mozilla DeepSpeech

WER
WRR
FAR (=1-SER)
FAR_{Voicefox}
Takeaways

- Voicefox relies on transcribers higher performance for natural voices and lower performance for synthesized voice.
- Our extensive evaluation shows that on average the WER for the synthesized voices are 2-3 times than the WER for natural voices.
- Using simple post-transcription processing rules can minimize or even eliminate the errors.