VoiceGuard: Secure and Private Speech Processing

Ferdinand Brasser¹, Tommaso Frassetto¹, Korbinian Riedhammer², Ahmad-Reza Sadeghi¹, Thomas Schneider¹, Christian Weinert¹

¹TU Darmstadt, Germany ²University of Applied Sciences Rosenheim, Germany
{ferdinand.brasser, tommaso.frassetto, ahmad.sadeghi}@trust.tu-darmstadt.de, korbinian@ieee.org, tschneider, weinert1@encrypto.cs.tu-darmstadt.de

1. Motivation & Contribution

Current Situation: Devices Providing Voice-based Interfaces are Omnipresent
- > 2B smartphone users (Amazon Alexa, Apple Siri, Google Assistant, Microsoft Cortana)
- Increasing number of smart-home devices (Amazon Echo, Apple Home-Pod, Google Home)

Risks: Voice Data Contains Sensitive Biometric Information as well as Spoken Words
- Impersonation attacks, fake recordings, extracting intimate / secret content, ...

Problem Statement: Naive Solution of Performing Speech Processing on Client-side fails
- Low-end devices are incapable of performing computationally demanding tasks
- Shipping the required machine learning model parameters to the clients contradicts the business interests of vendors

Contribution: Secure and Private Speech Processing Architecture dubbed “VoiceGuard”
- Efficiently protects speech processing tasks using Intel Software Guard Extensions (SGX)
- Supports user-specific models and generalizes to on-promises solutions

2. Related Work

Privacy-Preserving Machine Learning
- Via Secure Multi-Party Computation
  - Multiple parties jointly compute a publicly known function without revealing private inputs to each other by executing an interactive cryptographic protocol
  - Orders of magnitude higher computation time and communication cost
  - Impractical for on-the-fly processing due to repeated initialization costs
- Via Homomorphic Encryption
  - Operations are performed on encrypted data such as the decryption of the computation result equals the outcome when performing the same operations on plaintext data
  - Currently far from suitable for speech recognition in real time due to high overhead

Privacy-Preserving Speech Processing (Pathak et al., IEEE Signal Processing Magazine’13)
- Speech recognition and speaker verification via homomorphic encryption
- > 3 hours to encrypt 1s of audio & to recognize a word out of a 10 word vocabulary
- Speaker verification via secure string matching
  - Approach cannot be transferred to some processing tasks like speech recognition

Privacy-Preserving Encrypted Phonetic Search of Speech Data (Glackin et al., ICASSP’17)
- Requires the vendor to give the acoustic model to the user in the clear

3. Background on Intel SGX

Intel Software Guard Extensions (SGX)
- Enables processing of confidential data on untrusted systems via so-called enclaves
- Enclave: program that is executed in isolation from all other software, including privileged software (e.g., OS or a hypervisor)
- Confidential data (e.g., user input) is provisioned to an enclave over a secure channel

Remote Attestation (RA)
- Allows an external party to verify whether an enclave was created correctly
- Cryptographic hash of the initial memory state (memory measurement 𝑀) of the enclave is digitally signed by the platform signing key 𝛽, which is built into the CPU

Sealing
- Encrypt confidential data using an enclave-specific key and write to external storage
- Allows an enclave to use confidential data (e.g., acoustic model 𝑀, language model 𝐿𝑀, user-specific adaptation data 𝐴𝑀) across multiple instantiations

Pros & Cons
- Widely available in recent Intel CPUs (≥ 6th Core-i generation)
- Almost native execution speed
- Enclave code must incorporate defense mechanism to protect against side-channel attacks

4. VoiceGuard Architecture

5. Evaluation

Evaluation of Prototype Implementation (based on the kaldi toolkit)
- DARPA Resource Management (RM)
  - Training on 4000 utterances
  - 3MB DNN (750k parameters); 0.5MB / 2MB (uni / bigram) decoding graphs
  - Test on joint set of six test runs with ≥ 1500 utterances
- Wall Street Journal (WSJ)
  - Training on the 46B S2SJ4 set & training of m12u2_online system with i-vectors
  - 14MB DNN (3.6M parameters), 641MB pruned trigram decoding graph

Baseline kaldi vs. VoiceGuard (Run-Time in s) on Core i7-7700 @ 3.60GHz

<table>
<thead>
<tr>
<th></th>
<th>Baseline kaldi</th>
<th>VoiceGuard</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM-bigram</td>
<td>351</td>
<td>522</td>
</tr>
<tr>
<td>RM-unigram</td>
<td>358</td>
<td>835</td>
</tr>
<tr>
<td>WSJ (dev93)</td>
<td>1427</td>
<td>2854</td>
</tr>
<tr>
<td>WSJ (eval92)</td>
<td>876</td>
<td>1736</td>
</tr>
<tr>
<td>WSJ (eval93)</td>
<td>518</td>
<td>1058</td>
</tr>
</tbody>
</table>

6. Conclusion & Future Work

VoiceGuard: Novel Architecture for Privacy-Preserving and Efficient Speech Processing
- Protects the user’s sensitive voice data & the vendor’s IP (i.e., model parameters)
- Supports user-specific models, such as feature transformations (e.g., fMLLR), i-vectors, or model transformations (e.g., custom output layers)
- Deployment either in the cloud or on-promises
- Prototype implementation demonstrates applicability for speech recognition in real time
- Generic ⇒ also works for related tasks (speaker verification or voice biometrics, including emotion recognition and medical speech processing)

Open Problems & Future Work
- SGX enclave memory is limited to 96MB memory
  - Requires the vendor to give the acoustic model to the user in the clear
- Typically high-accuracy ASR systems use larger models than evaluated here
- Possible solution: distributing the processing across multiple SGX-enabled nodes
- Prototype code does not employ protection mechanisms against side-channel attacks
- Curious service providers could exploit micro-architectural effects to extract secret data
- Possible solution: use enclave hardening frameworks & measure performance impact