Technical Disclosure Commons

Defensive Publications Series

April 10, 2018

Secure audio processing

Thad Hughes

Ignacio Lopez Moreno

Aleksandar Kracun

Pedro Moreno

Secure audio processing

ABSTRACT

Automatic speech recognizers (ASR) are now nearly ubiquitous, finding application in smart assistants, smartphones, smart speakers, and other devices. An attack on an ASR that triggers such a device into carrying out false instructions can lead to severe consequences. Typically, speech recognition is performed using machine learning models, e.g., neural networks, whose intermediate outputs are not always fully concealed. Exposing such intermediate outputs makes the crafting of malicious input audio easier. This disclosure describes techniques that thwart attacks on speech recognition systems by moving model inference processing to a secure computing enclave. The memory of the secure enclave and signals are inaccessible to the user and untrusted processes, and therefore, resistant to attacks.

KEYWORDS

- Secure audio processing
- Secure enclave
- Speech recognition
- Secure computing

BACKGROUND

Recent years have seen a proliferation of smart assistants in the marketplace. Such devices, e.g., smart speakers, smartphones, etc., rely on automatic speech recognition (ASR) systems to decipher instructions. ASR systems use machine learning models, e.g., based on multi-layer neural networks to recognize speech. An adversary can use intermediate parameters of the machine learning model, e.g., intermediate weights of a multi-layer neural network, to

compromise speech recognition. It is easy to imagine the severe consequences that would result if a speech recognition system is compromised.

DESCRIPTION

Speech processing and recognition is typically performed on digital signal processors (DSP) or other embedded speech-processing engines. The implementation of ASR on such processors makes intermediate parameters such as neural network weights or intermediate outputs visible. Such visibility, alongside the visibility of the output target labels, make the crafting of malicious input audio simpler.

The techniques of this disclosure relocate inference processing by the machine learning model from a DSP or embedded processor to a separate integrated circuit (IC) or hardware block. The memory and side-channel signals of this IC or hardware block are inaccessible to the user or untrusted processes. Such a separate and inaccessible hardware block is known as a secure enclave.

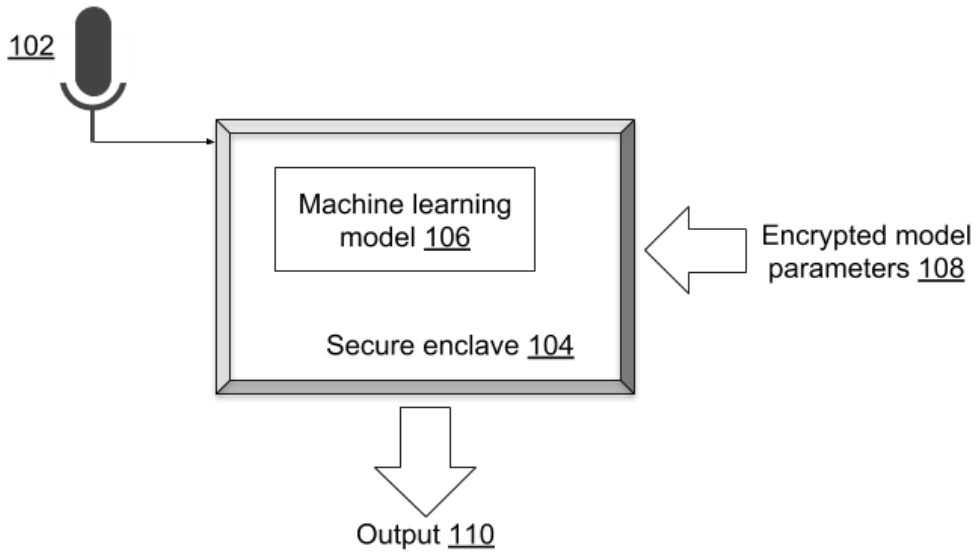

Fig. 1: Secure audio processing

Such secure audio processing is illustrated in Fig. 1. With user consent and permission, signals from a microphone (102) are directly routed to secure enclave (104) that securely encloses a machine learning model (106) for speech recognition. The parameters of the machine learning model are encrypted and uploaded (108) to the enclave.

Inference processing by the machine learning model in the enclave does not expose intermediate output. Only the ASR output (110), e.g., text transcripts or binary decisions, are exposed. Confidence scores or full output weights either do not leave the enclave, or are encrypted before exiting the enclave. The secure enclave multiplexes other hardware blocks that access audio, e.g. the telephony stack, additional DSPs, other trusted downstream enclaves, etc.

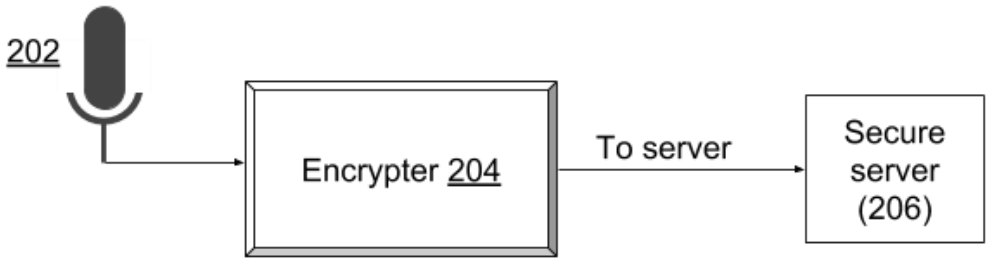

Fig. 2: Secure audio processing using server-side processing

Fig. 2 illustrates an alternate framework for secure audio that leverages server-side processing. In this framework, with user permission and consent, signals from microphone (202) are encrypted (204) immediately after analog processing. The encrypted audio is not accessible to the local device, and is sent to a secured server (206) for processing, e.g., for speech recognition. In this manner, an attacker that targets the local device is thwarted.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein

may enable collection of user information (e.g., information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes a speech processing or recognition system that is resistant to adversarial attacks. With user consent and permission, microphone signals are directly routed to a secure enclave. The memory of the secure enclave and the signals are inaccessible by the user or by untrusted processes. Inference processing by a machine learning model is performed within the secure enclave. Intermediate outputs of the model are not exposed outside the enclave. Model parameters are encrypted prior to uploading to the secure enclave, and only ASR outputs, such as text transcripts or binary decisions, are made available as return values from the secure enclave. Alternately, microphone output is securely encrypted and sent to a server for processing, such that the encrypted audio is inaccessible to the local device. In this manner, audio recognition and processing is made resistant to attacks.