idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Ruhr-Universität Bochum
Dr. Julia Weiler

24.09.2018
http://idw-online.de/de/news702681

Forschungsergebnisse
Informationstechnik
überregional

Geheime Botschaften für Alexa und Co.

Einem Team der Ruhr-Universität Bochum ist es gelungen, geheime Befehle für das Spracherkennungssystem Kaldi – welches vermutlich in Amazons Alexa und vielen anderen Systemen enthalten ist – in Audiodateien einzubauen. Für das menschliche Ohr sind diese nicht hörbar, aber Kaldi reagiert darauf. Die Forscherinnen und Forscher zeigten, dass sie beliebige Befehle in unterschiedlichen Arten von Audiosignalen, etwa in Sprache, Vogelgezwitscher oder Musik, verstecken können und dass Kaldi diese versteht.

Die Ergebnisse veröffentlichte die Gruppe um Lea Schönherr, Prof. Dr. Dorothea Kolossa und Prof. Dr. Thorsten Holz vom Horst-Görtz-Institut für IT-Sicherheit im Internet (https://adversarial-attacks.net/).

„Als eines von vielen Beispielen, wo ein solcher Angriff ausgenutzt werden könnte, kann man sich einen Sprachassistenten vorstellen, der Online-Bestellungen ausführen kann“, sagt Thorsten Holz. „Wir könnten eine Audiodatei, etwa einen Song, der im Radio abgespielt wird, so manipulieren, dass sie den Befehl enthält, ein bestimmtes Produkt einzukaufen.“

Ähnliche Angriffe, in der Fachsprache als Adversarial Examples bezeichnet, wurden vor einigen Jahren bereits für Bilderkennungssoftware beschrieben. Für Sprachsignale sind sie komplizierter umzusetzen, da die Bedeutung eines Audiosignals erst über die Zeit hinweg entsteht und zu einem Satz wird.

MP3-Prinzip genutzt

Um die Befehle in die Audiosignale einzubauen, nutzen die Forscherinnen und Forscher das psychoakustische Modell des Hörens, genauer gesagt den lautstärke- und frequenzabhängigen Effekt der Maskierung. „Wenn das Gehör damit beschäftigt ist, einen lauten Ton einer bestimmten Frequenz zu verarbeiten, können wir für einige Millisekunden auf dieser Frequenz andere, leisere Töne nicht mehr wahrnehmen“, erklärt Dorothea Kolossa.

Diese Tatsache wird auch beim MP3-Format genutzt, das nicht hörbare Bereiche ausspart, um die Dateigröße zu minimieren. Genau in diesen Bereichen versteckten die Forscher die Befehle für den Sprachassistenten. Die hinzugefügten Komponenten klingen für den Menschen wie zufälliges Rauschen, das im Gesamtsignal nicht oder kaum auffällt. Für die Maschine ändert es jedoch den Sinn. Während der Mensch Aussage A hört, versteht die Maschine Aussage B. Beispiele für die manipulierten Dateien und die durch Kaldi erkannten Sätze finden sich auf der Webseite der Forscher (https://adversarial-attacks.net/).

Die Berechnungen, um zehn Sekunden einer Audiodatei mit versteckten Informationen zu versehen, dauern weniger als zwei Minuten und sind somit deutlich schneller als zuvor beschriebene Angriffe auf Spracherkennungssysteme.

Funktioniert noch nicht mit Luftübertragung

Die Angriffe haben die Bochumer Wissenschaftler bisher nicht über den Luftweg ausgeführt; sie haben die manipulierten Audiodateien direkt als Daten in Kaldi eingespielt. In künftigen Studien wollen sie zeigen, dass der Angriff auch funktioniert, wenn das Signal über einen Lautsprecher abgespielt und durch die Luft zum Sprachassistenten transportiert wird. „Durch die Hintergrundgeräusche wird der Angriff nicht mehr ganz so effizient sein", vermutet Lea Schönherr. „Aber wir gehen davon aus, dass es immer noch funktioniert."

Neuronale Netze bislang wenig verstanden

Moderne Sprachassistenten beruhen auf neuronalen Netzen, sogenannten Deep Neural Networks, für die es momentan kaum Ansätze gibt, beweisbar sichere Systeme zu entwickeln. Die Netze bestehen aus mehreren Schichten; der Input, also die Audiodatei, gelangt in die erste Schicht und wird in den tieferliegenden Schichten verarbeitet. Die letzte Schicht generiert den Output, in diesem Fall also den erkannten Satz. „Die Funktion der verdeckten Schichten zwischen Input und Output ist in vielen Anwendungen nicht hinlänglich spezifiziert, was von einem Angreifer ausgenutzt werden kann", sagt Dorothea Kolossa.

Bislang kein wirksamer Schutz

Ziel der Forschung ist es, Sprachassistenten auf Dauer robuster gegen Angriffe zu machen. Für den jetzt vorgestellten Angriff sei es etwa denkbar, dass die Systeme standardmäßig berechnen, welche Anteile eines Audiosignals für Menschen nicht hörbar sind, und diese entfernen. „Allerdings gibt es sicher auch andere Möglichkeiten, um die geheimen Befehle in den Dateien zu verstecken, als das MP3-Prinzip", erklärt Kolossa. Und die würden wieder andere Schutzmechanismen erforderlich machen.

Bezüglich des aktuellen Gefahrenpotenzials gibt Holz jedoch Entwarnung: „Unser Angriff funktioniert derzeit noch nicht über die Luftschnittstelle. Außerdem sind Sprachassistenten derzeit nicht in sicherheitsrelevanten Bereichen im Einsatz, sondern dienen lediglich dem Komfort." Daher seien die Folgen möglicher Angriffe überschaubar. „Da die Systeme aber immer ausgefeilter und beliebter werden, muss weiter an den Schutzmechanismen gearbeitet werden", ergänzt der IT-Sicherheitsexperte.

wissenschaftliche Ansprechpartner:

Prof. Dr. Dorothea Kolossa
Kognitive Signalverarbeitung
Horst-Görtz-Institut für IT-Sicherheit
Ruhr-Universität Bochum
Tel.: 0234 32 28965
E-Mail: dorothea.kolossa@rub.de

Prof. Dr. Thorsten Holz
Systemsicherheit
Horst-Görtz-Institut für IT-Sicherheit
Ruhr-Universität Bochum
Tel.: 0234 32 25199
E-Mail: thorsten.holz@rub.de

Originalpublikation:

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, Dorothea Kolossa: Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding, 2018, Online-Vorabveröffentlichung: https://arxiv.org/abs/1808.05665

Das Bochumer Projektteam: Thorsten Holz, Lea Schönherr, Steffen Zeiler und Dorothea Kolossa (von links)
© RUB, Kramer