PHYS.ORG

Secret messages for Alexa and Co

September 24 2018

Project team from Bochum: Thorsten Holz, Lea Schönherr, Steffen Zeiler, and Dorothea Kolossa (from the left). Credit: RUB, Kramer

A team from Ruhr-Universität Bochum has succeeded in integrating secret commands for the Kaldi speech recognition system – which is believed to be contained in Amazon's Alexa and many other systems – into audio files. These are not audible to the human ear, but Kaldi reacts to them. The researchers showed that they could hide any sentence they liked in different types of audio signals, such as speech, birds' twittering, or music, and that Kaldi understood them. The results were published on the Internet by the group involving Lea Schönherr, Professor Dorothea Kolossa, and Professor Thorsten Holz from the Horst Görtz Institute for IT Security (adversarial-attacks.net/).

"A virtual assistant that can carry out online orders is one of many examples where such an attack could be exploited," says Thorsten Holz. "We could manipulate an audio file, such as a song played on the radio, to contain a command to purchase a particular product."

Similar attacks, known as adversarial examples in technical jargon, were already described a few years ago for image recognition software. They are more complicated to implement for speech signals as the meaning of an audio signal only emerges over time and becomes a sentence.

MP3 principle used

In order to incorporate the commands into the audio signals, the researchers use the psychoacoustic model of hearing, or, more precisely, the masking effect, which is dependent on volume and frequency. "When the auditory system is busy processing a loud sound of a certain frequency, we are no longer able to perceive other, quieter sounds at this frequency for a few milliseconds," explains Dorothea Kolossa.

This fact is also used in the MP3 format, which omits inaudible areas to minimise file size. It was in these areas that the researchers hid the commands for the voice assistant. For humans, the added components sound like random noise that is not or hardly noticeable in the overall signal. For the machine, however, it changes the meaning. While the human hears statement A, the machine understands statement B. Examples of the manipulated files and the sentences recognised by Kaldi can be found on the researchers' website (adversarial-attacks.net/).

The calculations for adding hidden information to ten seconds of an audio file take less than two minutes and are thus much faster than previously described attacks on speech recognition systems.

Not yet working with airborne transmission

The researchers from Bochum have not yet carried out the attacks over the air; they have passed the manipulated audio files directly to Kaldi as input data. In future studies, they want to show that the attack also works when the signal is played through a loudspeaker and reaches the voice assistant through the air. "Due to the background noise, the attack will no longer be quite as efficient," Lea Schönherr suspects. "But we assume that it will still work."

Modern speech recognition assistants are based on so-called deep neural networks, for which there are currently few attempts to develop provably secure systems. The networks consist of several layers; the input, i.e. the audio file, reaches the first layer and is processed in the deeper layers. The last layer generates the output, in this case the recognised sentence. "The function of the hidden layers between input and output, which can be exploited by an attacker, is not sufficiently specified in many applications," says Dorothea Kolossa.

No effective protection so far

The aim of the research is to make speech recognition assistants more robust against attacks over the long term. For the attack presented here, it is conceivable that the systems could calculate which parts of an audio signal are inaudible to humans and remove them. "However, there are certainly other ways to hide the secret commands in the files besides the

MP3 principle," explains Kolossa. And these would again require other protection mechanisms.

However, Holz does not believe there is cause for concern regarding the current potential for danger: "Our attack does not yet work via the air interface. In addition, speech recognition assistants are not currently used in safety-relevant areas, but are only for convenience." The consequences of possible attacks are therefore manageable. "Nevertheless, we must continue to work on the protection mechanisms as the systems become more sophisticated and popular," adds the IT security expert.

More information: Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. arxiv.org/abs/1808.05665

Provided by Ruhr-Universitaet-Bochum

Citation: Secret messages for Alexa and Co (2018, September 24) retrieved 25 April 2024 from https://phys.org/news/2018-09-secret-messages-alexa.html