Carnegie Mellon University
Software Engineering Institute

FEASIBILITY OF VOICE RECOGNITION TECHNOLOGY IN CONTROLLED DEFENSE INFORMATION ENVIRONMENTS

Eliezer Kanal, Linda Parker Gates, Jeff Davenport

January 2018

Background

The Amazon Echo is a voice-activated device that processes commands using the cloud-based Alexa artificial intelligence engine. The DIA received an Echo device from Amazon and are interested in exploring how it could be integrated into the Discovery Centers' operations.

Problem Definition

Voice recognition technology has made incredible strides in the past decade. With many devices and in many contexts, users can use free-form spoken word to give commands and not only have their command understood but acted upon within seconds. While the benefits of such technology are immediately visible, companies working with controlled defense information (CDI) and other sensitive information may want to consider more carefully how such technologies work and what information is exfiltrated before adopting these modern devices for their own use.

Voice Recognition Technology

The phrase "voice recognition" can refer to a number of capabilities, with widely varying levels of both sophistication and processing requirements.

Voice-to-Text

The most basic level of voice recognition is simple voice-to-text transcription. This technology converts audio input to written text. The process behind this seemingly simple process is highly complex; a good overview of the process can be found on the wiki of one of the open-source speech recognition systems.

These systems typically work for a single language, with many offerings allowing the user to extend the dictionary with user-specified word lists (e.g., company-specific jargon, names, etc.). Note that this is a fundamental capability for virtually all other technologies, but has long been offered as a standalone capability for transcription needs.

Pre-Defined Dictionary

One of the most basic voice recognition capabilities includes recognition of individual keywords, with specific actions mapped to each phrase. For example, early cell phone-based voice recognition technology had the user speak specific phrases, such as "call <user>" or "text <user>". This combines keywords from a list of commands ("call", "text", "open", etc.) with the words present in a user-defined dictionary (e.g., contact list entries). This technique can yield a fairly powerful system, as the action mapped to each keyword can be user-defined. An interested party could create a system with keywords corresponding to specific business interests.

Unstructured Text Interpretation

The capability offered by modern voice recognition systems offers unstructured text interpretation. The main differentiation is the ability of these systems to interpret free-form sentences rather than listening for specific keywords. From a functional standpoint, this extended capability doesn't offer significantly more functionality, as the capabilities of the system are still limited by the actions provided by the developers. However, this capability allows users to use natural language when interacting with the system. This capability significantly reduces training required to use the system as well as enabling much more freedom with the nature of the input. For example, users can state "send Bob a meeting invite for tomorrow at 11". The backend to this type of voice recognition allows the system to make a likely guess as to which "Bob" the user is referring to, whether "Bob" is actually "Robert", infer that "tomorrow" means "the day after today" and convert that to an actual date, understand that people often leave off the AM and PM designators when specifying time in spoken language, and that most users don't want to meet at 11 PM. None of the preceding relate to the actual words being spoken but instead relate to the semantic interpretation of those words in context.

Always-On Systems vs. User-Initiated Interactions

Historically, voice recognition systems were turned on when needed and turned off at all other times. For example, consider a transcription program using voice-to-text technology to automatically process physician notes to a patient's medical record, or a cell phone command recognition program which can call users through voice commands. These systems would be turned on through user interaction, voice recognition would be enabled, and when the commands were completed the system would be turned off.

Contrast these types of user-initiated systems with modern systems such as Amazon Go, Google Home, or Apple's "hey Siri" feature. Each of these devices maintains an active microphone at all time listening for a keyword signaling user interaction. While the audio may not initiate any activity, it is still being recorded and possibly transmitted to remote servers for processing. Indeed, analysis of network activity of seemingly inactive devices have revealed that, even while in “standby” mode, such devices are likely transmitting audio signal to remote servers.

Client- vs. Server-Based Processing

It should be noted that the computational power required for the different options listed earlier vary widely. Both text transcription and single word recognition are able to be performed on local machines or handheld devices. Unstructured text recognition as described above, due to its reliance on computationally expensive machine learning algorithms, is most commonly performed in the server environment, with the client machine acting as a microphone. Some commercial vendors have introduced an “offline mode” enabling use of these technologies even while unconnected to a network. While the machines can be used in this capacity, as soon as network connectivity is restored data transmission resumes. The vendors have made clear in marketing material for customers that these features are available for convenience; the devices are intended to be used while connected to the remote servers.

Information Leakage

In environments where CDI is likely to be discussed, remote-connected devices may pose an issue. While the data conveyed back to the servers is often encrypted in transit, it is still being decrypted and processed by the vendor. Such information leakage poses a number of issues, from policy violations to practical limitations on types of information that can be discussed in the presence of the device to concerns about vulnerabilities present in the vendor hardware and software stack.

In the context of the Amazon Echo, Alexa AI, and other consumer devices with always-on microphones, it is difficult to tell exactly what information is being captured and exfiltrated, as the data is almost always encrypted in transfer. That said, simply based on device functionality, we infer a significant amount. As the most basic, recorded spoken audio signals are definitely being transmitted to external servers to aid in improving voice recognition systems. As a subset of the above, audio signals from different individuals are likely identified and separated, so as to differentiate different users of the systems. To that extent it is likely that digital fingerprints of the individual users are kept as well. Combining this with timestamps of each audio clip we, it’s likely that leaked information includes the timing of a discussion, the content of a discussion, and the individuals present during the discussion. Given that the device can determine approximate location through GeoIP lookup, location is often leaked as well. Information about discussion topic is easily extractable from the spoken text, and modern sentiment analysis systems are able to determine with relatively high accuracy (80%) whether a speaker is happy, displeased, angry, etc., providing more context about individuals in the discussion.

While the preceding relates to information directly extractable from data sent to servers, side-channel information can be used to infer a great deal more information. It has long been standard practice in advertising to use as much historical information as possible. For example, many search engines keep historical search query information for the purpose of improving future search results… the more I know about the user the better the search engine can disambiguate homonyms. In the case of voice recognition IoT devices, old search information is likely kept. In the context of CDI, discussion often revolves around what does and does not work, and even without directly mentioning a phrase the nature of the topic can be inferred. On a different note, given that some individuals may have purchased such a device for personal use and linked it to their personal company account, the parent company may be able to cross-identify a given user simply based on voice pattern signatures. For example, if Alice purchases a Google Home device and links it to her personal account, Google will have a sample of Alice's voice directly linked with her identity. When Alice uses a Google Home device in a location where CDI may be discussed, Google may be able to identify Alice based on her voice patterns. Extending this further, if Alice uses her Google account to make purchases (or even her Google-enabled phone payment technology), Google may be able to infer when Alice is on travel, and linking all this information together may allow Google to determine both the destination, the reason, and the outcome of Alice's CDI-related travel.

Alternative Techniques, Devices & Recommendations

Given all the preceding concerns, the best recommendation regarding voice recognition technology is to only use either client-based solutions or server-based solutions where the server is controlled by government IT. Client-based solutions retain all data on the local computer without transmitting it to a remote location. To that extent, any safeguards used to protect the machine protect information contained therein. Server-based solutions controlled by government entities are also acceptable, provided that information is secured both on the local machine during capture, while in transit, and on the remote server.

Additionally, an always-on device is likely not a desired end state. Ideally, such a device would be turned on only when needed, and would have a clear indicator—visual and audible—that audio is being recorded. In an environment where operational security is a requirement necessary safeguards should be taken.

Depending on the needs of the client there are many alternatives to using COTS voice-recognition software. A variety of software packages exist that enable developers with no voice recognition expertise to add voice recognition capability to their products. By taking advantage of these products users can gain the capability without the same risk. Additionally, COTS offerings do exists which follow the guidelines above, either by keeping data on the client system or by allowing the government to control the servers. Future work could include a full feature comparison as determined by the needs of the DIA.

Further Reading

Home devices transmitting data while in standby mode:

- https://www.wired.com/2016/12/alexa-and-google-record-your-voice/
- https://www.techworld.com/security/does-amazon-alexa-listen-to-my-conversations-3661967/

CMUSphinx tutorial webpage: https://cmusphinx.github.io/wiki/tutorialconcepts/

Open-source voice recognition toolkits: https://blog.neospeech.com/top-5-open-source-speech-recognition-toolkits/

Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

Phone: 412/268.5800 | 888.201.4479
Web: www.sei.cmu.edu | www.cert.org
Email: info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

This report was prepared for the SEI Administrative Agent AFLCMC/AZS 5 Eglin Street Hanscom AFB, MA 01731-2100

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM18-0112