Technical Disclosure Commons

Defensive Publications Series

April 05, 2018

Voice input for authentication

Nick Felker

Shuyang Chen

Sachit Mishra

Voice input for authentication

ABSTRACT

Recent years have seen a proliferation of systems designed to respond to voice, e.g., smart speakers, smartphones, and other voice-activated devices. Such systems are convenient and intuitive to use. However, they also open the possibility for malicious actors to gain control of sensitive information through fraudulent use of speech commands. For example, a malicious actor can spoof the voice recognition system through a recording of the user's voice or by using a computer to synthesize the speech mimicking the user. The techniques of this disclosure thwart such attacks by providing a random phrase for the user to repeat and verifying the resulting audio and transcription for authenticity. Spoofed audio sounds choppy and exhibits sharp transitions, while authentic audio is smooth and lacks unnatural artifacts. A trained machine learning model is used to tell the difference.

KEYWORDS

- Voice UI
- Authentication
- Voice spoofing
- Voice fingerprint
- Speech recognition
- Smart speaker
- Challenge response
- Voice assistant

BACKGROUND

Voice recognition systems are increasingly integrated into consumer devices, with the result that systems designed to respond to voice, e.g., smartphones, smart speakers, smart assistants, internet-of-things (IoT) devices, etc., are now nearly ubiquitous. Such voice recognition and response systems match users based on a given voice fingerprint. Although convenient and intuitive to use, use of voice recognition technologies open up the possibility of malicious actors fraudulently using the voice user interface to perform unauthorized actions. For example, a malicious actor may compile a unique query in a user's voice by splicing together unrelated voice recordings of the user. Alternatively, a malicious actor could use a computer to synthesize text that approximates the user's speech. For sensitive applications of voice-enabled systems, e.g., financial transactions, such spoofing can enable a malicious actor to perform undesirable actions.

DESCRIPTION

The techniques of this disclosure enable users to protect sensitive actions from being performed over a voice user interface without authorization. Prior to responding to or completing a user request, the voice user interface asks the user to utter a randomly generated phrase. The audio of the user response is checked using trained machine learning models to verify validity of the transcription and smoothness of the audio. With a spoofed audio track, e.g., where several audio recordings are spliced together, the audio sounds choppy or exhibits sharp changes at transition points. These types of artifacts are identified using the trained machine learning model, and if detected, are used to reject the user's request. The techniques described herein, to authenticate users over a voice interface, are referred to as audio

CAPTCHAs (“Completely Automated Public Turing test to tell Computers and Humans Apart”).

Voice input from the user is obtained and utilized for the purpose of responding to user commands. Recording of user uttered speech, if necessary, is performed only if permitted by the user. Such recordings are utilized for the purpose of responding to user requests, and are retained for a limited period of time, if permitted by the user.

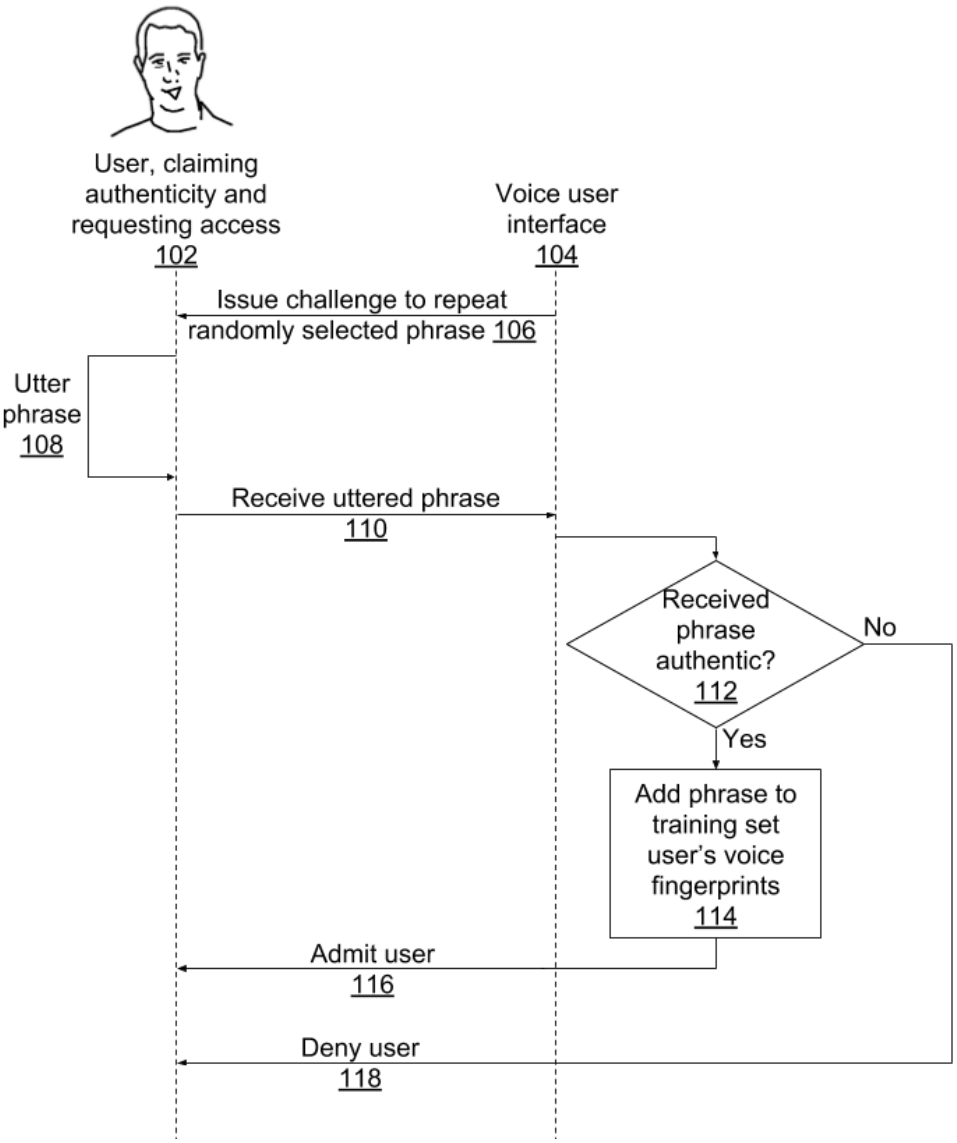

Fig. 1: Authentication over a voice user interface

Fig. 1 illustrates authentication over a voice user interface, per techniques of this disclosure. A user (102), claiming authenticity, queries or requests access to a device using a voice user interface (104). The voice user interface responds by issuing a challenge (106) to prove authenticity, e.g., requesting the user to repeat a randomly generated phrase. The phrase is designed to last only for a few seconds and includes words that typically don't occur together in normal usage. These attributes of the challenge phrase make it difficult for a malicious actor to generate a matching audio clip that sounds smooth.

The user utters the phrase (108). The user has a small time window, e.g., a few seconds, to complete utterance of the challenge phrase. A small time window for response increases the difficulty for a malicious actor to splice together or synthesize mimicked speech. The voice user interface receives the uttered phrase (110) and tests it for authenticity (112).

Authenticity is tested by checking for smoothness within the captured audio, and checking for unnatural artifacts such as sharp transitions within the captured audio. Natural speech by a true user is smooth and free of sharp transitions. A trained machine learning model determines if the received phrase originates from previously recorded clips by identifying whether the audio came from a person or from an electronic speaker. Metrics such volume, frequency, noise levels, etc. are used to make the determination relating to authenticity.

If the captured audio is determined to have originated from a true user, then the user's access request is admitted (116), and the correct audio response to the challenge added to the training set of the user's voice fingerprint model (114), if permitted by the user. Addition of the correct audio response to the training set continually improves the accuracy of the user's voice fingerprint. If authentication fails, the user is denied access (118).

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 provides techniques that protect devices with voice user interfaces against unauthorized access. When a user requests access or issues a command to a device over a voice interface, the techniques request the user to repeat a phrase comprising randomly selected words. The phrase is designed such that it is difficult to artificially synthesize without introducing artifacts. A true user uttering the phrase is identified as such by a trained machine learning model that tests for attributes such as smoothness of captured audio, presence or absence of artifacts, etc. Upon successful verification, the user's request is admitted.