March 27, 2019

Client-side masking for voice queries

Aleksandar Kracun

Matthew Sharifi

Client-side masking for voice queries

ABSTRACT

Many voice-based assistive technologies transmit the voice input received from users to a server for processing. The transmitted audio includes the speaker's voice which can identify the person. Users of such technologies therefore face a tradeoff between convenient voice interfaces with reduced privacy or less convenient non-voice input with higher privacy. Techniques described herein mask a user's voice by locally processing the voice input received by a device. The masked voice cannot personally identify the user while still enabling server-side processing that allows recognition of spoken phrases. Application of the proposed techniques provides the user with greater privacy without diminishing the user experience for voice input in terms of recognition, latency, and other operational characteristics.

KEYWORDS

- Voice assistant
- Voice input
- Voice query
- Spoken input
- Speech recognition
- Voice masking
- Voice morphing
- Privacy
- Smart speaker

BACKGROUND

In many situations, it is more convenient for users to issue a query via voice instead of using another form of input such as tapping, clicking, typing, etc. As a result, voice-based assistive technologies that can handle voice input are increasingly used in a variety of everyday situations. For example, virtual assistant or other voice-activated software accepts voice input in devices such as smart speakers, smartphones, wearable devices, home appliances, etc.

Typically, the voice input received by devices that employ such technologies is transmitted to a server for processing. The transmitted audio includes the speaker's voice which can often identify the person. In contrast, the same query issued via text input does not require the user's device to transmit the user's voice characteristics to the server, which provides better privacy. Therefore, users face a tradeoff between more convenient voice interfaces with reduced privacy or less convenient non-voice input with higher privacy.

DESCRIPTION

Techniques described herein mask a user's voice by locally processing the voice input received by a device. The masked voice cannot personally identify the user while still enabling server-side processing that allows recognition of spoken phrases. Application of the proposed techniques provides the user with greater privacy without diminishing the user experience for voice input in terms of recognition, latency, and other operational characteristics.

With permission from the user, voice input is processed locally, e.g., on a user device such as a smartphone, smart speaker, computer, etc. The local processing includes normalizing the audio by filtering out the unique audio characteristics of the user's voice while preserving the ability of a speech recognition system to recognize the spoken phrases, e.g., for the purpose of converting the phrase from audio to text. The local audio processing can be carried out in one or

more of several ways such as randomizing the pitch to remove the speaker's voice characteristics, using speech-to-speech voice normalization software, applying speaker diarization to segment audio to separate the voice of the main speaker from that of any bystander, transforming the voice via standard voice morphing algorithms, etc.

If the user permits, the locally processed audio is transmitted to the server for converting the audio to text via speech recognition. The user's spoken request is thus converted to a text form that is then passed on as input to the appropriate system that can handle the request. For instance, if the user's voice request has asked for directions to a location, the converted text is provided to a mapping application so that the application can deliver the desired directions to the user. Additionally, if the user permits, the masked audio request can be stored to a log that can be accessed by the user. Since the audio stored to such a historical record of voice requests is the audio received by the server after masking on the user's device, the removal of the user's voice characteristics from the audio is evident to the user by playing the audio records in the audio history.

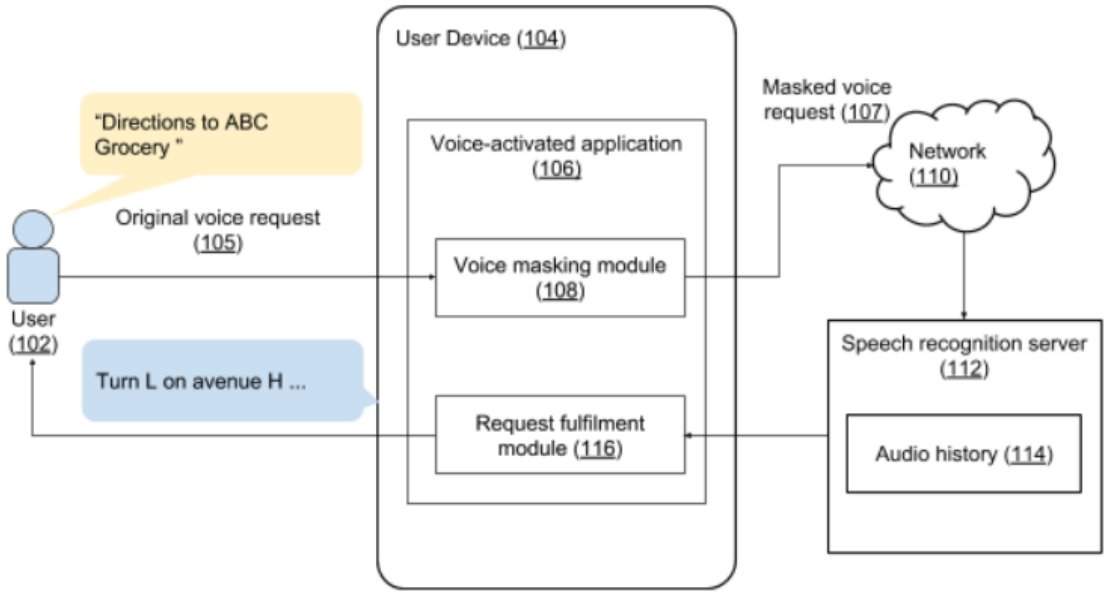

Fig. 1: Masking voice input prior to transmission to a speech recognition server

Fig. 1 shows an example implementation of the techniques described in this disclosure. A user (102) issues a request ("Directions to ABC Grocery") via voice to a device (104) supports voice-based input, e.g., via a voice-activated application (106). The user's original voice request (105) is processed locally on the device by a voice masking module (108). The audio output of the voice masking module removes identifiable voice characteristics of the user while retaining the speech content of the user's voice request.

The masked voice request (107) is sent over a network (110) to a speech recognition server (112) that converts the speech within the masked voice request to corresponding text. The result is the textual form of the user's original request issued via voice. The server optionally stores the received masked voice request in audio history (114), if permitted and enabled by the user. The text is passed to an appropriate application, e.g., a server-based map application. The response from the application is provided to a request fulfillment module (116) on the device, which delivers the result ("Turn L on avenue H…") to the user.

The audio history (114) that serves as a log of masked voice requests received from the user. The user can inspect and play back the audio history to confirm that the audio passed to the server was masked. While Fig. 1 shows the request fulfillment module implemented within the device, the module can be located within the server or another entity separate from the server and the device.

The user is provided with options to enable or disable the voice masking feature. The feature can be made active at all times, at certain predefined times, or can be enabled on-demand to provide a private voice input mode, e.g., similar to the private browsing feature of modern web browsers.

An alternative to employing the proposed techniques is to avoid transmitting voice data to the server altogether by employing speech recognition locally on device that receives the voice input. However, the quality of speech recognition performed locally on the device can be inferior to that achieved via server-side speech processing. Further, locally performed speech recognition can use processing and memory resources and increase the usage of the device battery, e.g., for battery-powered devices such as smartphones, wearables, etc. Moreover, not all devices include the hardware and software resources necessary to perform local speech recognition.

The techniques of this disclosure can be applied to any device or system that incorporates server-based speech processing for processing voice-based input. Examples of such systems include speech-to-text software, smart speaker, digital assistant, smart TVs, etc. The described techniques allow users to benefit from the convenient user experience of voice-based interfaces while eliminating the need for transmitting personally identifiable voice information to a server-side system for processing the speech and fulfilling the user's request. As a result, application of the proposed techniques provides the user with improved privacy without diminishing the user experience for voice input in terms of recognition, latency, and other operational characteristics. Therefore, the techniques described herein improve the privacy of server-based speech recognition - original speech input with the user's identifiable voice does not leave the user's device.

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

Many voice-based assistive technologies transmit the voice input received from users to a server for processing. The transmitted audio includes the speaker's voice which can identify the person. Users of such technologies therefore face a tradeoff between convenient voice interfaces with reduced privacy or less convenient non-voice input with higher privacy. Techniques described herein mask a user's voice by locally processing the voice input received by a device. The masked voice cannot personally identify the user while still enabling server-side processing that allows recognition of spoken phrases. Application of the proposed techniques provides the user with greater privacy without diminishing the user experience for voice input in terms of recognition, latency, and other operational characteristics.