August 2023

Contextual Remapping of Audio Input and Output for Fluid Virtual Assistant Interaction

D Shin

Contextual Remapping of Audio Input and Output for Fluid Virtual Assistant Interaction

ABSTRACT

When a user moves in physical space while engaging in conversational hands-free interaction with a voice-based virtual assistant on a device, the interaction is interrupted if the user steps too far from the device. Currently, devices that support such interaction do not provide dynamic audio switching or mechanisms to switch an ongoing conversational interaction to a different device. This disclosure describes techniques for seamless dynamic switching of audio input and output from one device to another based on presence detection using data from device sensors. The appropriate devices for the audio input and the output, as well as the device that acts as the host of the virtual assistant can be determined by following any suitable arbitration procedure, guided by prespecified or inferred user preferences for virtual assistant interaction. Automated dynamic remapping of audio input and out and/or virtual assistant host device can enhance the user experience (UX) by enabling users to engage in seamless and fluid conversation interactions with a virtual assistant while moving around.

KEYWORDS

- Virtual assistant
- Voice interaction
- Conversational interaction
- User presence
- Presence detection
- Field of View (FOV)
- Computer vision
- Smart display

BACKGROUND

Users often employ a conversational approach when interacting with voice-based virtual assistants provided via various devices, such as smartphones, smart speakers, smartwatches, earbuds, etc.). When a user interacts with a virtual assistant using voice, the input is received via a device microphone while the spoken response from the virtual assistant is delivered as audio output. The microphone and speaker for input and output can be built into a device (e.g., smartphone, smart speaker, etc.) or connected separately as in the case of external microphones, speakers or earbuds coupled to the device via a wired or wireless connection.

Users often move around in physical space while being engaged in a conversational hands-free interaction with a voice-based virtual assistant on a given stationary device, such as a computer or a smart speaker. If the movement results in a user moving too far from the fixed device, the device microphone may not be able to capture user speech, thus breaking the flow of the voice interaction until the user moves close enough to be audible via the input of the fixed device. For instance, when a user at home starts a conversational interaction with a virtual assistant via a stationary laptop while wearing wireless earbuds for listening, the user's speech may be captured via the laptop microphone while the virtual assistant response is delivered to the earbuds. Having moved without the laptop to another room in the home a bit later, the user is too far away from the laptop for the laptop microphone to capture the user's voice. As a result, the user is unable to continue the interaction until moving back to a location closer to the laptop.

Many users own multiple devices with virtual assistant capabilities. When a user moves too far to be audible at a device, it is desirable to switch an ongoing conversational interaction with a voice-based virtual assistant to a different input source near the user (e.g., on a wearable device worn by the user) instead of interrupting the flow. For instance, when a user moves too

far to be audible to the microphone of a laptop, the user may wish to switch the input source from the built-in laptop microphone to the microphone in the user's earbuds connected wirelessly to the laptop. Yet, current devices do not enable dynamic switching of input sources in the middle of a conversational interaction with a virtual assistant. Similarly, current devices do not include a mechanism that enables a user to switch an ongoing conversational interaction with a virtual assistant from one device to another. For instance, a user cannot currently begin an interaction with a virtual assistant on a laptop and continue it by switching to the virtual assistant on the smartphone when appropriate.

DESCRIPTION

This disclosure describes techniques for seamless dynamic switching of audio input and output from one device to another based on relevant contextual information obtained with user permission from one or more device sensors, such as camera, microphone, inertial measurement unit (IMU), etc. The contextual information can be used to determine user presence at or near various devices and to automatically remap the audio input and/or output to the most suitable option. The automated remapping can enable users to engage in uninterrupted conversational interactions with a virtual assistant while moving around and switching physical locations.

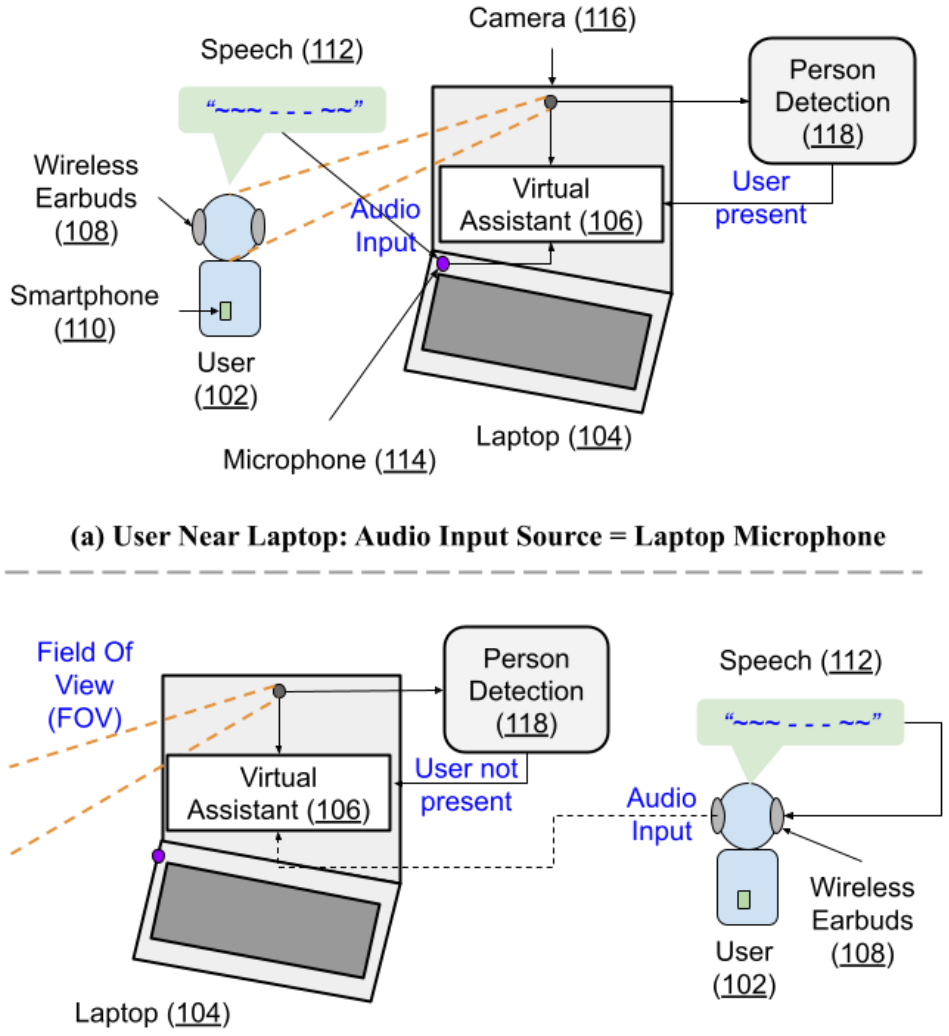

Fig. 1: Automatically switching source of audio input to virtual assistant when user moves

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. As shown in Fig. 1(a), a user (102) interacts via speech (112) with a virtual assistant (106) provided via a laptop (114) while wearing wireless earbuds (108) (e.g., that may be paired to the laptop) and carrying a smartphone (110). With user permission, the field of view (FOV) of the laptop camera (116) is scanned to check if the user is present at or near the laptop. Since the user is detected to be present at the laptop, the laptop microphone (114) is used as the input

source for obtaining the audio of the user's speech which is passed to the virtual assistant. While Fig. 1(a) illustrates the use of the laptop camera for presence detection, any suitable available sensor such as infrared camera, radar, keyboard, mouse, other input device, ambient light sensor, proximity sensor, etc. can be used for presence detection. Presence detection may utilize data from the sensors and activity recognition techniques that can determine the current user activity, e.g., the user is holding a smartphone in their hand, the user is walking while holding a smartphone, the user has taken earbuds off, etc.

As shown in Fig. 1(b), when the user moves away from the laptop, the field of view of the laptop camera indicates that the user is no longer present at the laptop. Since the user can still maintain an audio connection to the laptop via the paired wireless earbuds, the input audio source is automatically switched to the wireless earbuds. The audio of the user's speech is detected by the microphones in the wireless earbuds and relayed to the laptop for use by the virtual assistant, thereby seamlessly continuing the conversational interaction the user initiated earlier, as shown in Fig. 1(a). The audio output can also be switched as necessary in the same manner.

When a switch in audio input and/or output mechanisms is necessary, the audio input and/or output can be switched to any device with audio input and/or output capability that is available to the user at the new location. For instance, such a device can be one that the user is wearing or carrying (e.g., smartphone, smartwatch, earbuds) or one present in the user's vicinity (e.g., smart speaker, smart appliance).

When only a single device is available for the switch the input and/or output are automatically switched to that device. For instance, if the user moves around while wearing wireless earbuds but not carrying a smartphone, the audio input and output can both be switched to the wireless earbuds when required. In case multiple devices with audio input and output

capabilities are available, the appropriate devices for the audio input and the output can be determined by following any suitable arbitration procedure, guided by prespecified or inferred user preferences for virtual assistant interactions. For instance, if specified or inferred user preferences indicate that a user carrying a smartphone while wearing earbuds prefers to listen via earbuds but talk into the smartphone, the audio input can be switched to the smartphone and the audio output to the earbuds. Arbitration rules can specify detection rules as well as device priority for input and output.

Apart from switching audio input and output, the above approach can also be employed to switch the primary device for user presence detection during an ongoing conversational interaction with a virtual assistant. For instance, consider when a user initiates the interaction while in front of a computer in one room and later moves to another room that has a smart display with a camera. In such a case, the user can be detected to be within the camera FOV of the smart display, which can be made the new host device for the ongoing virtual assistant interaction.

If users permit, presence detection can employ suitable computer vision models (e.g., presence detection models) that serve as binary classifiers for the presence of a person in the camera FOV. However, in spaces in which multiple people are present, detecting simply whether the camera FOV contains a person can be ambiguous because the person can be any one of the multiple people. Moreover, if different individuals in different locations within the space are in the FOV of the cameras of different devices at those locations, the presence of a person in the camera FOV can be registered at multiple devices simultaneously. When appropriate, such ambiguities can be resolved by recognizing specific users (with appropriate user permissions)

based on applying suitable user-permitted authentication mechanisms, such as face or body calibration data.

With user permission, the devices that are on or near a user can be detected based on obtaining and analyzing relevant contextual information from one or more sensors within the devices. For example:

- data from smartphone sensors can indicate whether the user is holding the phone or walking with the phone;
- sensors, such as ambient light sensors (ALS) in wearable devices (e.g., smartwatch, earbuds) can be employed to detect whether the user is wearing the device; and
- devices with cameras can be used to determine the presence of the user near the device as described above for a laptop in Fig. 1;

The techniques described in this disclosure can be implemented with user permission to support seamless continuous voice interaction with a virtual assistant provided via any device. Automated dynamic remapping of audio input and output, and/or virtual assistant host device can enhance the user experience (UX) by enabling users to engage in seamless and fluid conversation interactions with a virtual assistant while moving around in physical space.

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 devices, sensor data from devices, a user's authentication information for a virtual assistant, a user's preferences including preferred devices, 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 describes techniques for seamless dynamic switching of audio input and output from one device to another based on presence detection using data from device sensors. The appropriate devices for the audio input and the output, as well as the device that acts as the host of the virtual assistant can be determined by following any suitable arbitration procedure, guided by prespecified or inferred user preferences for virtual assistant interaction. Automated dynamic remapping of audio input and out and/or virtual assistant host device can enhance the user experience (UX) by enabling users to engage in seamless and fluid conversation interactions with a virtual assistant while moving around.