May 2023

Customized Virtual Assistant Invocation Based on Device Orientation and Mode

Tuan Anh Nguyen

Sana Mithani

Sergei Volnov

Yunfan Ye

Liang-yu Chen

See next page for additional authors

Inventor(s)
Tuan Anh Nguyen, Sana Mithani, Sergei Volnov, Yunfan Ye, Liang-yu Chen, Krunal Shah, Alexey Galata, Qiong Huang, Tzu-Chan Chuang, Sai Aditya Chitturu, and Will Truong

Customized Virtual Assistant Invocation Based on Device Orientation and Mode

ABSTRACT

Hotword-less virtual assistant interaction, referred to as look-to-talk, can be triggered when the user is within a certain distance from a device that provides the virtual assistant and looks at the device. With user permission, machine learning (ML) models analyze real-time video, audio, and text to determine if a given user utterance is intended as an instruction to the virtual assistant or an utterance directed to someone else in the room. This disclosure describes look-to-talk functionality that is applicable to not only stationary devices, but also mobile devices such as smartphones or tablets. The techniques detect changes in device orientation or mode to trigger look-to-talk functionality. In recognition of the fact that sensor data captured by a stationary device can be different that that captured by a mobile device, look-to-talk parameters and ML models are adapted to the device orientation and mode.

KEYWORDS

- Virtual assistant
- Virtual assistant invocation
- Virtual assistant triggering
- Hotword detection
- Device orientation
- Device mode
- Engagement model
- Look-to-talk

BACKGROUND

In natural conversations, people do not call out their conversational partner's names every time they speak to each other or at each conversational turn. Rather, they rely on contextual signaling mechanisms to initiate conversations; eye contact is often all it takes. In contrast, virtual assistant applications that take spoken commands and provide audio responses rely on a hotword mechanism to wake up and act on a user's instructions.

A more intuitive user-assistant interaction, known as look-to-talk [1], triggers the virtual assistant into action when the user is within a certain distance from a device via which the virtual assistant is accessed (e.g., a smart speaker, a smart display, or other device) and looks at the device (e.g., the device screen). With appropriate user permissions, machine learning (ML) models analyze real-time video, audio, and text, and use such parameters as proximity, face match, head orientation, gaze direction, active-speaker (lip movement) detection, voice matching, contextual awareness, intent classification, etc., to determine if a given user utterance is intended as an instruction to the virtual assistant as opposed to e.g., a passing glance or an utterance directed to someone else in the room.

Look-to-talk has been available on stationary devices, such as smart displays, smart speakers, etc. Making look-to-talk available on mobile devices such as smartphones, tablets, etc. is difficult due to changing device orientations, camera fields-of-view, etc. Additionally, a user looking at a smartphone or scrolling through displayed content is no confirmation that the user wants to trigger a virtual assistant on the smartphone.

DESCRIPTION

This disclosure describes look-to-talk functionality that can be made available on mobile devices such as smartphones or tablets. The techniques detect changes in device orientation (e.g.,

a device being picked up by the user) or mode (e.g., a tablet being switched from docked mode to carry-around mode) to trigger look-to-talk functionality. In recognition of the fact that sensor data (e.g., image) captured by a stationary device (e.g., body pose) can be different that that captured by a mobile device (e.g., face), look-to-talk parameters and ML models are adapted to the device orientation and mode.

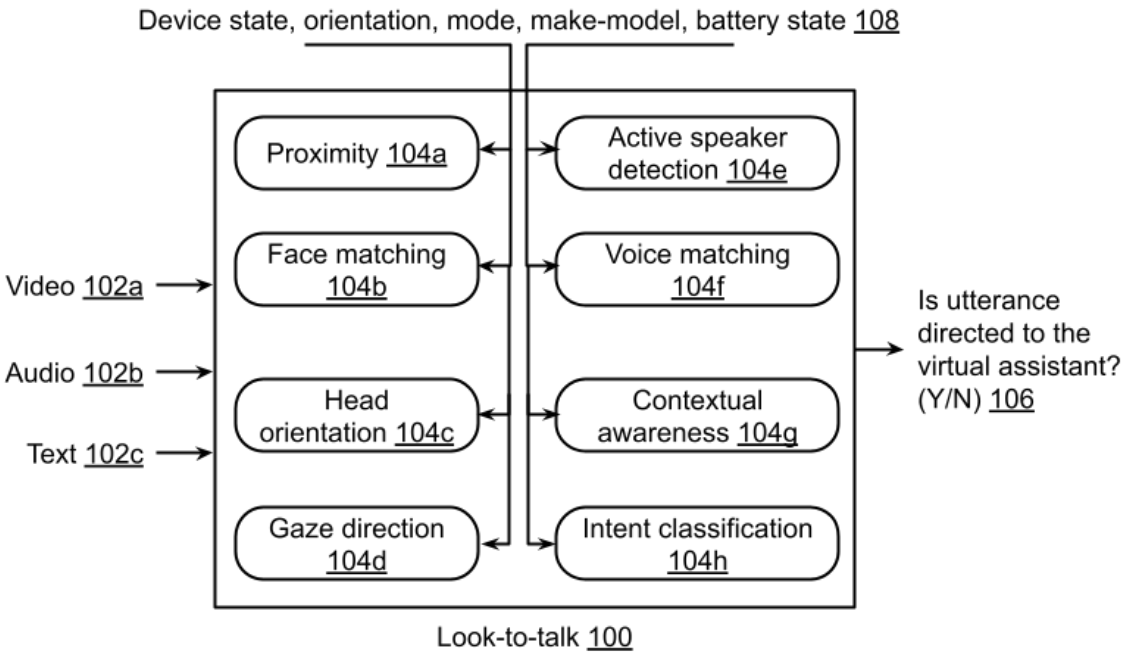

Fig. 1: Changing parameters or engagement model in look-to-talk in response to a change in device orientation or mode

Fig. 1 illustrates changing parameters or engagement models used to detect user intent in look-to-talk in response to a change in device orientation or mode. With user permission, mobile device look-to-talk (100) analyzes real-time video (102a), audio (102b), and/or text (102c) to detect user intent, e.g., if a user's utterance is directed to the virtual assistant (106). Machine learning models rely on parameters such as proximity (104a), face matching (104b), head orientation (104c), gaze direction (104d), active-speaker (lip-movement) detection (104e), voice

matching (104f), contextual awareness (104g), intent classification (104h), etc. to determine if a given user utterance is intended as an instruction to the virtual assistant.

In contrast to traditional look-to-talk implemented on stationary devices, the relative weights given to different parameters can vary based on whether the look-to-talk is performed on a stationary or mobile device. In particular, the ML models are informed by the device state, orientation, mode, make-model, battery state, (108) etc.

Example 1

Look-to-talk on a stationary device can select a relatively large weight for body pose, since stationary cameras typically capture both body and face (this can be referred to as 'far-away' mode). Look-to-talk on a mobile device can select a relatively low weight for body pose, since mobile front-facing cameras typically capture not the body but the face (this can be referred to as 'close-up' mode).

Example 2

Look-to-talk on a stationary device can select a relatively low weight for gaze, since users in a typical scenario (smart display on the kitchen countertop, with the user carrying out chores while interacting with the virtual assistant) may not gaze at a stationary device for extended periods of time. Look-to-talk on a mobile device can select a relatively large weight for gaze, as a user engaging with a mobile device virtual assistant is likely to look at the device screen when providing a spoken command. Similarly, audio input can have differing weights for look-to-talk implementations on stationary versus mobile devices.

Example 3

Stationary devices such as smart displays, smart speakers, etc. are usually plugged into the wall electric supply, and therefore rarely experience power supply problems. However, mobile devices such as smartphones run on battery, and hence can be subject to dwindling energy supply. Look-to-talk in stationary devices accordingly does not account for battery status, while look-to-talk in mobile devices can account for battery status. For example, look-to-talk can be automatically deactivated below certain battery levels. Alternatively, look-to-talk can be made to gracefully degrade with falling battery levels, e.g., it can take longer for look-to-talk to respond when the battery is at a 20% level as compared to an 80% level. Similarly, look-to-talk features and performance can take into account the computational capabilities of the mobile device: a sophisticated, high-end smartphone can have the full feature set of look-to-talk, while an entry-level smartphone can have an essential feature set supported by the computational capabilities of the device.

Example 4

ML models used in look-to-talk can take device orientation, and changes thereto, as an input. Their training data can include device orientation as a parameter. A mobile device can be in several orientations; for example, it can be held in the hand, it can be lying flat on a table, it can be docked in a docking/charging station, etc., whereas a stationary device (such as a smart display) typically has only one orientation. Look-to-talk on mobile devices is therefore triggered (and behaves) differently based on the device orientation: A mobile device detected to be held in the hand can put a relatively large weight on the user's gaze to trigger look-to-talk, while a mobile device that is detected to be lying flat may put a relatively low weight on the user's gaze.

A mobile device docked in a docking/charging station can default to the look-to-talk behavior of stationary devices, since the physical setup after docking is similar to a stationary smart display.

In this manner, even if look-to-talk is turned on for multiple devices used by a person, the actual triggering and behavior of look-to-talk is based on circumstances that differ from device to device. Look-to-talk is thereby made portable across different device types.

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 the collection of user information (e.g., information about a user's gaze, pose, face, voice, 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 describes look-to-talk functionality that is applicable to not only stationary devices, but also mobile devices such as smartphones or tablets. The techniques detect changes in device orientation or mode to trigger look-to-talk functionality. In recognition of the fact that sensor data captured by a stationary device can be different that that captured by a mobile device, look-to-talk parameters and ML models are adapted to the device orientation and mode.

REFERENCES

[1] “Look and Talk: Natural Conversations with Google Assistant,” available online at https://ai.googleblog.com/2022/07/look-and-talk-natural-conversations.html accessed May 6, 2023.