June 2023

ASSISTANT AUTOMATIC MULTIMODAL SUGGESTIONS

Xin Li

Nitish Murthy

Mukesh Kumar Singh

ASSISTANT AUTOMATIC MULTIMODAL SUGGESTIONS

ABSTRACT

A computing device (e.g., a smartphone, a laptop computer, a tablet computer, a smartwatch, etc.) may provide a user with multimodal suggestions in response to an ambiguous user request. The computing device may evaluate a user request to an assistant and determine that the request is ambiguous as to the user’s intent. The computing device may determine, based on the evaluation, multimodal suggestions to prompt the user to clarify the intent of the user request. The computing device may output the suggestions to the user via various combinations of graphical user interface elements, text displayed via a user interface, and spoken indications. The user may provide an input (e.g., touch, stylus, mouse, keyboard, voice, etc.) to select one of the suggestions. Based on the selected suggestion, the computing device may perform the desired action and may train a user-specific machine learning model to perform the desired action when the user provides the same or similar ambiguous user request in the future.

DETAILED DESCRIPTION

FIG. 1, below, is a conceptual diagram illustrating a computing device 100 that includes an assistant 102, microphone(s) 104, user interface device(s) 106, and speaker(s) 108. In accordance with the various techniques described in this publication, computing device 100 may use assistant 102 to identify multimodal suggested responses to an ambiguous user request.

As shown in FIG. 1, below, computing device 100 may include assistant 102, microphone(s) 104, user interface device(s) 106, and speaker(s) 108. Computing device 100 may be any computing device such as a cellular phone, a smartphone, laptop computer, tablet computer, a portable gaming device, a portable media player, an e-book reader, a wearable device (including computerized watches, rings, glasses, headset, headphones, jackets, etc.), a desktop computer, and/or the like as well as an automotive computing device such as a head unit, infotainment unit, integrated navigation system, and/or the like. Computing device 100 may be a representation of a combination of any one or more of the above computing devices, for example a smartphone and automotive infotainment system connected via wireless communication.

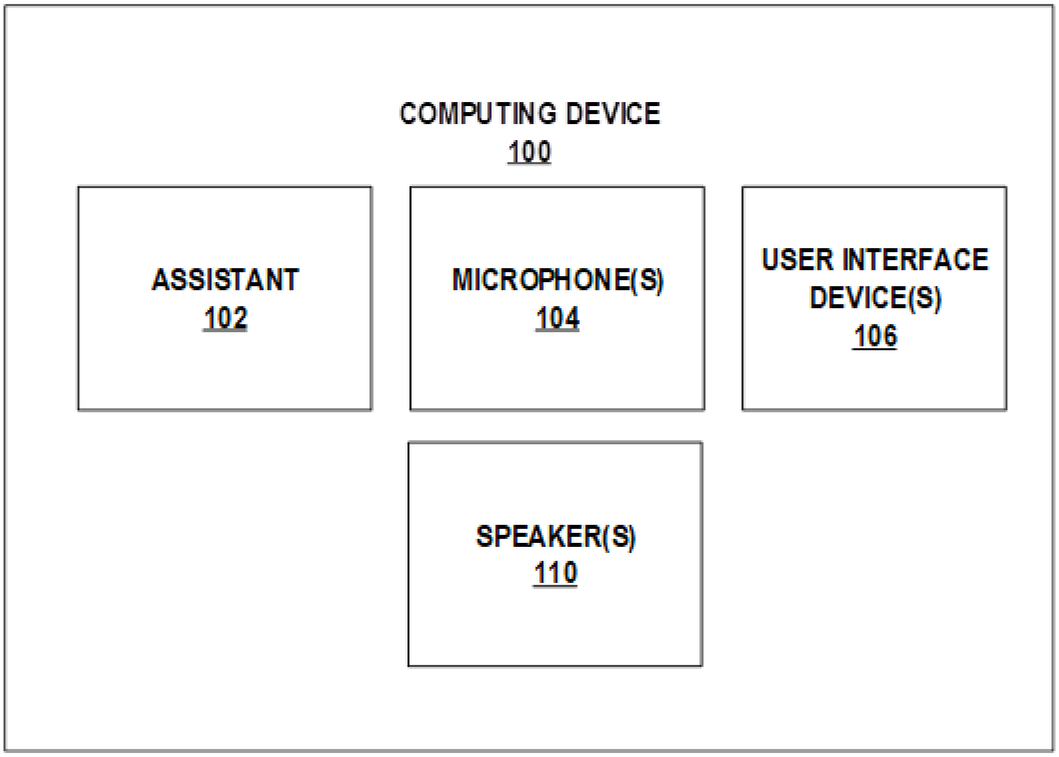

FIG. 1

Assistant 102 may be any software-based assistant executing on computing device 100. Computing device 100 may execute assistant 102 to enable computing device 100 to respond to user requests using the intelligence of assistant 102. Assistant 102 may be invoked in response to user input such as predefined spoken phrases (e.g., “Hi Assistant, please…”). Assistant 102 may instruct components of computing device 100 to perform one or more actions in response to user input (e.g., modify the volume of output by speaker(s) 108, cause speaker(s) 108 to generate output such as spoken words, modify visual elements displayed by user interface device(s) 106, etc.).

Microphone(s) 104 may include one or more components of computing device 100 that can receive audio input from a user. Microphone(s) 104 may additionally include the components necessary for converting received audio signals into digital signals for processing by computing device 100. Microphone(s) 104 may receive input from a user of computing device 100 such as spoken words or other input from a user of computing device 100.

Various interface device hardware may implement user interface device(s) 106. User interface device(s) 106 may function as an input device using a presence-sensitive input component, such as a presence-sensitive screen or touch-sensitive screen that receives tactile input from a user of computing device 100 and/or as physical buttons such as a button on a smartphone. The presence-sensitive input component may determine a contact location (e.g., an (x,y) coordinate) of the presence-sensitive input component at which the object was detected. User interface device(s) may display a user interface via the presence-sensitive or touch-sensitive screen.

Speaker(s) 108 may include one or more components capable of producing audio output. Speaker(s) 108 may include speakers contained internally within computing device 100, as well as speakers physically and/or wirelessly coupled with computing device 100 such as Bluetooth® speakers. Speaker(s) 108 may generate audio output for a user of computing device 100 such as audio indications as well as spoken output (e.g., “Did you want to increase the volume?”).

Various aspects of the techniques described in this publication enable assistant 102 to provide multimodal suggested responses to ambiguous user requests. Assistant 102 may receive an indication of user input from microphone(s) 104. Computing device 100 may cause assistant 102 to process the user input and determine that the request received is ambiguous. Assistant 102 may determine that a user request is ambiguous due to one or more issues with the request. For example, assistant 102 may determine that a user's spoken input that does not define an actionable request (e.g., a request that does not include a command, a request that provides insufficient context for assistant 102 to determine the user's intent, etc.), that some of the words spoken by the user were not comprehensible, or other issues with a user's request that prevents assistant 102 from completing the user's request.

Computing device 100 may utilize assistant 102 to identify one or more potential responses to an ambiguous user request. Assistant 102 may identify the one or more potential responses that include one or more suggested actions based on the content of the ambiguous user request. Assistant 102 may utilize particular words or phrases within the user request to identify potential responses to the request (e.g., if the user states "radio", the potential responses will be associated with different functions of the radio). Assistant 102 may utilize one or more machine learning models to identify potential responses to the user request.

Assistant 102 may utilize machine learning to analyze user requests and determine multimodal responses to user requests. For example, assistant 102 may perform speech-to-text on voice input provided by the user and provide the resulting text to a machine learning model. In some examples, computing device 100 may execute the machine learning model on-device. The machine learning model used by assistant 102 may output one or more possible actions. In some instances, the machine learning model may also output a confidence value associated with each of the one or more possible actions. If all of the confidence values do not satisfy a threshold confidence level, assistant 102 may determine that the voice input was an ambiguous voice input.

Assistant 102 may communicate with a server via one or more components of computing device 100 to process the user requests and responses. Computing device 100 may provide the audio data to a server so that the server can apply a larger machine learning model to the audio data that would not be practical for computing device 100 to execute. The server may process the audio data and determine multimodal responses to the use requests. The server may provide the multimodal responses to assistant 102 for presentation to the user.

Assistant 102 may select a subset of the identified potential responses for display to the user of computing device 100 as it may not be practical to display all of the potential responses to a user and overwhelm them. Assistant 102 may cause computing device 100 to update user interface device(s) 106 to display visual indicators of the subset of potential responses as multimodal suggestions to the user. Additionally, assistant 102 may cause speaker(s) 108 to play spoken indicators of the multimodal suggestions.

In an example case, a user asks their tablet to “radio” while their tablet is playing music from a radio station. The tablet provides the data regarding the request to assistant 102. Assistant 102 processes the data regarding the user request and determines multimodal responses to the user request. Assistant 102 causes the user’s tablet to generate audio output requesting the user to select one of the multimodal responses (“I heard you speak radio, did you mean change the radio station or stop playing radio?”) to resolve the ambiguity of the user request.

Assistant 102 may record data regarding user selections of multimodal responses to optimize processing of future user requests. Assistant 102 may utilize the data regarding previous user selections of multimodal responses to identify patterns of user interaction with multimodal responses. Computing device 100 may cause assistant 102 to utilize the identified patterns to determine a response to a nominally ambiguous user request without requiring further user input.

As an example, a user says “home” to an assistant on their smartphone while sitting in their car and preparing to leave a restaurant. The user’s smartphone receives the user’s input and provides data regarding the input to assistant 102 executing on the smartphone. Assistant 102 processes the user input and compares the user input to identified patterns of user input. Assistant 102 determines that the user has a pattern of stating “home” when they wish to navigate to their home address. Computing device 100 provides navigational instructions to navigate to the user’s home address in response to the determination by assistant 102.

Assistant 102 may utilize aggregated user data from other users of assistants on other computing devices to determine potential responses to an ambiguous user request. Computing device 100 may receive aggregated user data that includes identified responses to ambiguous user requests. Computing device 100 may provide this data to assistant 102 for use in determining responses to user requests.

In an example usage of aggregated user data, a user tells the name of a well-known fast food restaurant to their smartwatch. Assistant 102 may receive data regarding the user input from the user’s smartwatch. Assistant 102 may determine, based on the user data and aggregated data maintained by computing device 100, that the user’s request is similar to the requests of other users who then requested navigation to the well-known fast food restaurant. Based on the determination, assistant 102 may cause the user’s smartwatch to provide navigation instructions to the nearest location of the well-known fast food restaurant.

Assistant 102 may perform automatic completion of user input and cause computing device 100 to perform the user’s request. Computing device 100 may provide data consistent with a user request to assistant 102 for assistant 102 to process and determine multimodal responses to the user request. Assistant 102 may identify multimodal responses that are a completion or correction of the user’s request similar to text autofill and autocorrection.

In a use case, a user asks their vehicle’s infotainment system to navigate them to an address that includes “2000”. Assistant 102, executing on the infotainment system, may receive data regarding the user input and determine whether the user’s input was “2000” or “to 1000” due to the similarly sounding phrases. Based on a determination that “2000” is the user’s intended input, assistant 102 causes the infotainment system to navigate to the appropriate address.

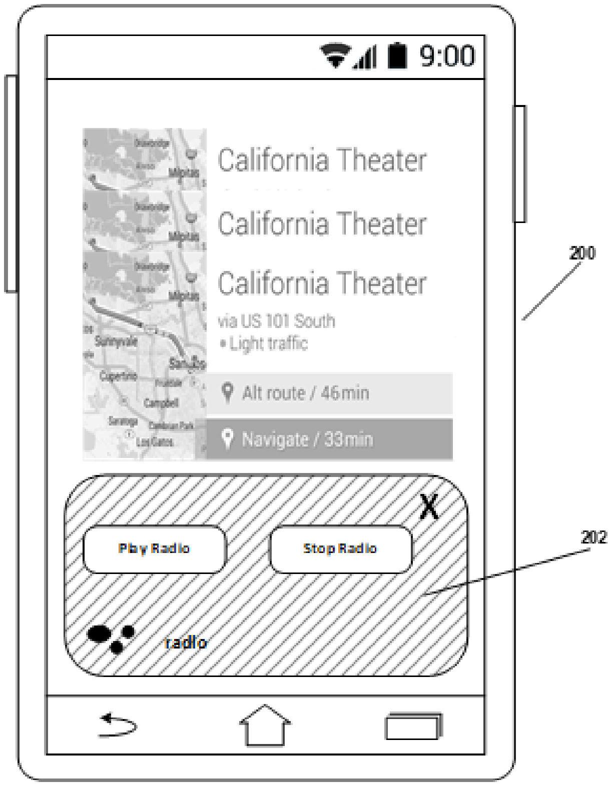

FIG. 2

FIG. 2 illustrates an example computing device 200, similar to computing device 100 as illustrated in FIG. 1, displaying an example user interface with a multimodal response displayed by the assistant. FIG. 2 illustrates computing device 200 displaying a user interface, with an example navigational app in the background of the user interface and an assistant prompt 202 in the foreground of the user interface.

Assistant 102 may cause computing device 200 to display an assistant prompt 202 in the foreground of the user interface of computing device 200 in response to a user request. Computing device 200 may generate an updated user interface with visual elements associated with assistant 102 visually displayed over the rest of the user interface in response to a request to update the user interface from assistant 102.

In an example similar to that illustrated by FIG. 2, a user is riding in a car and using a navigation app on their smartphone to assist the driver. The user tells their smartphone “radio” with no further context. The user’s smartphone receives the user input and provides data regarding the input to assistant 102. Assistant 102 processes the input and determines that the user has either indicated that they wish for the user’s smartphone to begin playing the radio or for the user’s smartphone to cease playing the radio. In response to the determination, assistant 102 causes the user’s smartphone to update the UI of the smartphone to display a visual indicator of assistant 102 and two interactable user interface elements with one labeled “Play Radio” and the other labeled “Stop Radio”.

FIG. 3 illustrates an example computing device 300, similar to computing device 100 as illustrated in FIG. 1, displaying a user interface that includes a user interface element for an assistant. FIG. 3 includes computing device 300 displaying a user interface that includes assistant user interface element 302 (“UI element 302”).

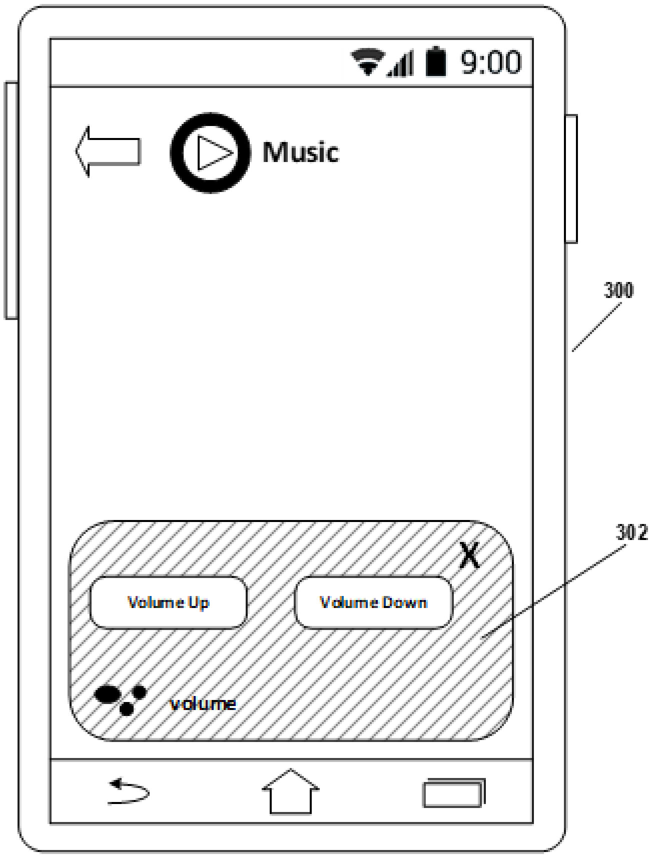

FIG. 3

Assistant 102 may cause computing device 300 to display assistant prompt 302 in the foreground of the user interface of computing device 200 in response to a user request. In the example of FIG. 3, assistant 102 has caused computing device 300 to display UI element 302 visually placed above a music application executing on computing device 300.

In an example case, a user is listening to music via a music app on their smartphone. The user tells their smartphone “volume,” with the rest of the request lost due to a sudden increase in background noise. Assistant 102 determines that the user’s request could be to increase the volume of the music application or to decrease the volume. Assistant 102 causes the user’s smartphone to update the UI of the smartphone with a user element such as UI element 302 visually placed over the music application to request that the user select one of the multimodal responses identified by assistant 102.

It is noted that the techniques of this disclosure may be combined with any other suitable technique or combination of techniques. As one example, the techniques of this disclosure may be combined with the techniques described in *Speech recognition interface design for in-vehicle system* (Zhang Hua et al.). In another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2007/0005206. In yet another example, the techniques of this disclosure may be combined with the techniques described in International Patent Publication No. 2012/174515.