(19) DANMARK (10) DK 179559 B1

(12) PATENTSKRIFT

Patent- og
Varemærkestyrelsen

(51) Int.Cl.: *G06F 17/30 (2006.01)*

(21) Ansøgningsnummer: **PA 2017 70425**

(22) Indleveringsdato: **2017-06-01**

(24) Løbedag: **2017-06-01**

(41) Alm. tilgængelig: **2018-11-17**

(45) Patentets meddelelse bkg. og publiceret den: **2019-02-15**

(30) Prioritet:
2017-05-16 US 62/506,981

(73) Patenthaver:
Apple Inc., 1 Infinite Loop Cupertino 95014 CA California, USA

(72) Opfinder:
David Chance Graham, c/o Apple Inc. 1, Infinite Loop Cupertino 95014 CA California, USA
Cyrus Daniel Irani, c/o Apple Inc. 1 Infinite Loop Cupertino 95014 CA California, USA
Thomas Alsina, c/o Apple Inc. 1 Infinite Loop Cupertino 95014 CA California, USA
Aimee PIERCY, c/o Apple Inc. 1 Infinite Loop Cupertino 95014 CA California, USA
Garrett L. WEINBERG, c/o Apple Inc. 1 Infinite Loop Cupertino 95014 CA California, USA

(74) Fuldmægtig:
COPA COPENHAGEN PATENTS K/S, Rosenørns Allé 1, 2. sal, 1970 Frederiksberg C, Danmark

(54) Titel: **INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION**

(56) Fremdragne publikationer:
US 2016/0173960 A1
US 2017/0068423 A1
US 2014/0081633 A1
US 2016/0378747 A1
US 2017/0068670 A1

(57) Sammendrag:
Systems and processes for operating an intelligent automated assistant are provided. In accordance with one example, a method includes, at an electronic device with one or more processors and memory, receiving a first natural-language speech input indicative of a request for media, where the first natural-language speech input comprises a first search parameter; providing, by a digital assistant, a first media item identified based on the first search parameter. The method further includes, while providing the first media item, receiving a second natural language speech input and determining whether the second input corresponds to a user intent of refining the request for media. The method further includes, in accordance with a determination that the second speech input corresponds to a user intent of refining the request for media: identifying, based on the first parameter and the second speech input, a second media item and providing the second media item.

Fortsættes...

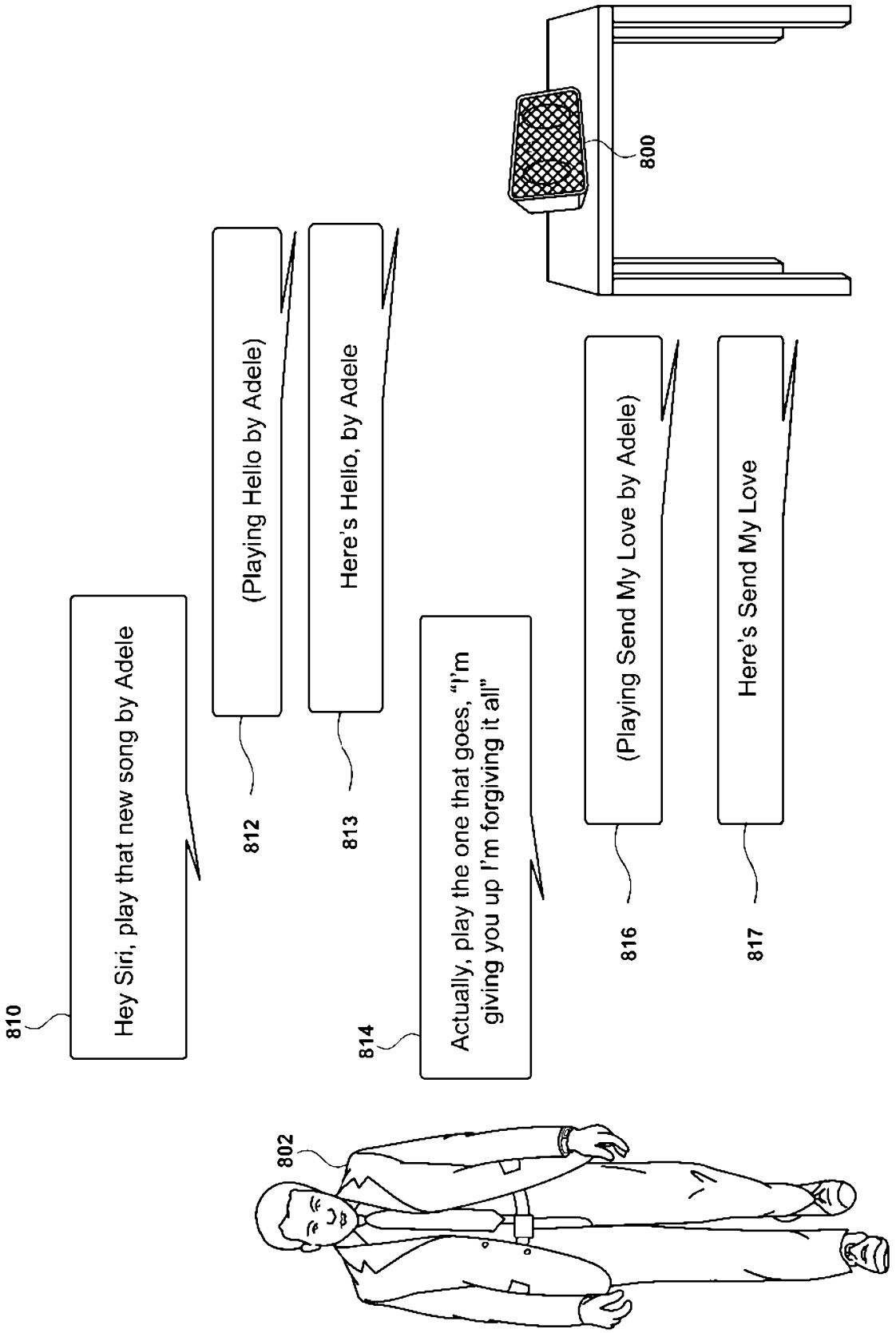

800
802
810
Hey Siri, play that new song by Adele
812
(Playing Hello by Adele)
813
Here's Hello, by Adele
814
Actually, play the one that goes, "I'm giving you up I'm forgiving it all"
816
(Playing Send My Love by Adele)
817
Here's Send My Love

FIG. 8A

INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application Serial No. 62/506,981, "INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION," filed on May 16, 2017.

Field

[0002] This relates generally to intelligent automated assistants and, more specifically, to providing an auditory-based interface of a digital assistant for media exploration.

Background

[0003] Intelligent automated assistants (or digital assistants) can provide a beneficial interface between human users and electronic devices. Such assistants can allow users to interact with devices or systems using natural language in spoken and/or text forms. For example, a user can provide a speech input containing a user request to a digital assistant operating on an electronic device. The digital assistant can interpret the user's intent from the speech input and operationalize the user's intent into tasks. The tasks can then be performed by executing one or more services of the electronic device, and a relevant output responsive to the user request can be returned to the user.

[0004] In most cases, users rely, at least in part, on conventional, graphical user interfaces to interact with an electronic device. In some instances, however, a digital assistant may be implemented on an electronic device with limited or no display capabilities. US 2017/0068670 A1 discloses systems and processes for operating a digital assistant in a media environment. An audio input containing a media-related request can be received. A primary user intent corresponding to the media-related request can be determined. In accordance with a determination that the primary user intent comprises a user intent to narrow the primary media

search query, a second primary media search query corresponding to the primary user intent can be generated.

Summary

[0005] The scope of the invention is defined by the claims. Additional non-claimed embodiments are described below and are examples of related techniques to help understand the claimed invention. Example methods are disclosed herein. The invention provides a method that includes, at an electronic device having one or more processors and memory, receiving a first natural-language speech input indicative of a request for media, where the first natural-language speech input comprises a first search parameter; providing, by the digital assistant, a first media item, where the first media item is identified based on the first search parameter; while providing the first media item, receiving a second natural-language speech input; determining whether the second natural-language speech input corresponds to a user intent of refining the request for media. The method further includes, in accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media: identifying, based on the first parameter and the second natural-language speech input, a second media item different from the first media item; and providing, by the digital assistant, the second media item. The scope of the method is defined in claim 1.

[0006] An example method includes, at an electronic device having one or more processors and memory, receiving a natural-language speech input; identifying, by the digital assistant, a task based on the natural-language speech input; providing, by the digital assistant, a speech output indicative of a verbal response associated with the identified task; and while providing the speech output indicative of a verbal response: providing, by the digital assistant, playback of a media item corresponding to the verbal response.

[0007] An example method includes, at an electronic device having one or more processors and memory, receiving a speech input indicative of a request for media; in response to receiving the speech input, providing, by the digital assistant, an audio output indicative of a suggestion of a first media item; determining, by the digital assistant, whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold. The method further includes, in accordance with a determination that the number of consecutive non-

affirmative responses does not satisfy the threshold: providing, by the digital assistant, an audio output indicative of a suggestion of a second media item different from the first media item. The method further includes, in accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold: foregoing providing an audio output indicative of a suggestion of a second media item; and providing, by the digital assistant, an audio output indicative of a request for user input.

[0008] An example method includes, at an electronic device having one or more processors and memory, receiving a speech input indicative of a request for media; detecting, by the digital assistant, physical presence of a plurality of users to the electronic device; in response to detecting the physical presence of the plurality of users, obtaining a plurality of preference profiles corresponding to the plurality of users; providing, by the digital assistant, a merged preference profile based on the plurality of preference profiles; identifying, by the digital assistant, a media item based on the merged preference profile; and providing, by the digital assistant, an audio output including the identified media item.

[0009] Example non-transitory computer-readable media are disclosed herein. The invention provides a non-transitory computer-readable storage medium that stores one or more programs. The one or more programs comprise instructions, which when executed by one or more processors of an electronic device, cause the electronic device to receive a first natural-language speech input indicative of a request for media, where the first natural-language speech input comprises a first search parameter; provide, by a digital assistant, a first media item, where the first media item is identified based on the first search parameter; while providing the first media item, receive a second natural-language speech input; determine whether the second natural-language speech input corresponds to a user intent of refining the request for media. The instructions can further cause the electronic device to, in accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media: identify, based on the first parameter and the second natural-language speech input, a second media item different from the first media item; and provide, by the digital assistant, the second media item. The scope of the non-transitory computer-readable media is defined in claim 41.

[0010] An example non-transitory computer-readable storage medium stores one or more programs. The one or more programs comprise instructions, which when executed by one or more processors of an electronic device, cause the electronic device to receive a natural-language speech input; identify, by a digital assistant, a task based on the natural-language speech input; provide, by the digital assistant, a speech output indicative of a verbal response associated with the identified task; while providing the speech output indicative of a verbal response: provide, by the digital assistant, playback of a media item corresponding to the verbal response.

[0011] An example non-transitory computer-readable storage medium stores one or more programs. The one or more programs comprise instructions, which when executed by one or more processors of an electronic device, cause the electronic device to receive a speech input indicative of a request for media; in response to receiving the speech input, provide, by a digital assistant, an audio output indicative of a suggestion of a first media item; determine, by the digital assistant, whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold. The instructions can further cause the electronic device to, in accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold: provide, by the digital assistant, an audio output indicative of a suggestion of a second media item different from the first media item. The instructions can further cause the electronic device to, in accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold: forego providing an audio output indicative of a suggestion of a second media item; and provide, by the digital assistant, an audio output indicative of a request for user input.

[0012] An example non-transitory computer-readable storage medium stores one or more programs. The one or more programs comprise instructions, which when executed by one or more processors of an electronic device, cause the electronic device to receive a speech input indicative of a request for media; detect, by a digital assistant, physical presence of a plurality of users to the electronic device; in response to detecting the physical presence of the plurality of users, obtain a plurality of preference profiles corresponding to the plurality of users; provide, by the digital assistant, a merged preference profile based on the plurality of preference profiles; identify, by the digital assistant, a media item based on the merged preference profile; and provide, by the digital assistant, an audio output including the identified media item.

[0013] Example electronic devices are disclosed herein. The invention provides an electronic device that comprises one or more processors; a memory; and one or more programs, where the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for receiving a first natural-language speech input indicative of a request for media, where the first natural-language speech input comprises a first search parameter; providing, by a digital assistant, a first media item, where the first media item is identified based on the first search parameter; while providing the first media item, receiving a second natural-language speech input; determining whether the second natural-language speech input corresponds to a user intent of refining the request for media. The one or more programs further include instructions for, in accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media: identifying, based on the first parameter and the second natural-language speech input, a second media item different from the first media item; and providing, by the digital assistant, the second media item. The scope of electronic device is defined in claim 40.

[0014] An example electronic device comprises one or more processors; a memory; and one or more programs, where the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for receiving a natural-language speech input; identifying, by a digital assistant, a task based on the natural-language speech input; providing, by the digital assistant, a speech output indicative of a verbal response associated with the identified task; while providing the speech output indicative of a verbal response: providing, by the digital assistant, playback of a media item corresponding to the verbal response.

[0015] An example electronic device comprises one or more processors; a memory; and one or more programs, where the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for receiving a speech input indicative of a request for media; in response to receiving the speech input, providing, by a digital assistant, an audio output indicative of a suggestion of a first media item; determining, by the digital assistant, whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold. The one or more

programs further include instructions for, in accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold: providing, by the digital assistant, an audio output indicative of a suggestion of a second media item different from the first media item. The one or more programs further include instructions for, in accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold: foregoing providing an audio output indicative of a suggestion of a second media item; and providing, by the digital assistant, an audio output indicative of a request for user input.

[0016] An example electronic device comprises one or more processors; a memory; and one or more programs, where the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for receiving a speech input indicative of a request for media; detecting, by a digital assistant, physical presence of a plurality of users to the electronic device; in response to detecting the physical presence of the plurality of users, obtaining a plurality of preference profiles corresponding to the plurality of users; providing, by the digital assistant, a merged preference profile based on the plurality of preference profiles; identifying, by the digital assistant, a media item based on the merged preference profile; and providing, by the digital assistant, an audio output including the identified media item

[0017] An example electronic device comprises means for receiving a first natural-language speech input indicative of a request for media, where the first natural-language speech input comprises a first search parameter; means for providing, by a digital assistant, a first media item, where the first media item is identified based on the first search parameter; means for, while providing the first media item, receiving a second natural-language speech input; means for determining whether the second natural-language speech input corresponds to a user intent of refining the request for media; means for, in accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media: identifying, based on the first parameter and the second natural-language speech input, a second media item different from the first media item; and providing, by the digital assistant, the second media item.

[0018] An example electronic device comprises means for receiving a natural-language

speech input; means for identifying, by a digital assistant, a task based on the natural-language speech input; means for providing, by the digital assistant, a speech output indicative of a verbal response associated with the identified task; means for, while providing the speech output indicative of a verbal response: providing, by the digital assistant, playback of a media item corresponding to the verbal response.

[0019] An example electronic device comprises means for receiving a speech input indicative of a request for media; means for, in response to receiving the speech input, providing, by a digital assistant, an audio output indicative of a suggestion of a first media item; means for determining, by the digital assistant, whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold; means for, in accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold: providing, by the digital assistant, an audio output indicative of a suggestion of a second media item different from the first media item; means for, in accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold: foregoing providing an audio output indicative of a suggestion of a second media item; and providing, by the digital assistant, an audio output indicative of a request for user input.

[0020] An example electronic device comprises means for receiving a speech input indicative of a request for media; means for detecting, by a digital assistant, physical presence of a plurality of users to the electronic device; means for, in response to detecting the physical presence of the plurality of users, obtaining a plurality of preference profiles corresponding to the plurality of users; means for providing, by the digital assistant, a merged preference profile based on the plurality of preference profiles; means for identifying, by the digital assistant, a media item based on the merged preference profile; and means for providing, by the digital assistant, an audio output including the identified media item.

[0021] Receiving a natural-language speech input while providing a media item allows the user to easily steer a media search to obtain desirable content. The digital assistant allows the user to refine a media request at any time, without having to stop a current playback or having to wait for a prompt by the digital assistant. As such, the digital assistant provides the user with full and flexible control over the media search process. Further, receiving a natural-language speech

input while providing a media item provides natural, intuitive, and human-like interactions between the digital assistant and the user, as the digital assistant allows the user to barge into a conversation and steer the conversation at any given time. Providing flexible and intuitive control of the media search process enhances the operability of the device and makes the interactions with the digital assistant more efficient (e.g., by understanding the user intent and giving user full control) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

[0022] Determining whether the natural-language speech input corresponds to a user intent of refining a media request and identifying a media item accordingly allows the user to quickly obtain desirable content with a relatively small number of inputs. The technique reduces the number of user inputs because, for instance, the user does not need to repeatedly provide previously specified parameters when refining the media request. The technique also provides natural and intuitive interactions between the digital assistant and the user because, for example, the user is able to receive recommendations that are increasingly tailored and, via a series of decisions, narrowed down to desirable content. Reducing the number of user inputs and providing an intuitive user interface enhance the operability of the device and make the user-device interface more efficient (e.g., by helping the user to provide proper inputs and reducing user mistakes when operating/interacting with the device) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

[0023] Providing a speech output indicative of a verbal response to a user request while also providing playback of a related media item provides a rich and intuitive auditory interface of the digital assistant. The playback of the media item (e.g., related sound effects, representative samples of content) helps the user quickly understand the content being presented and make more informed decisions without prolonging the duration of the audio output. Further, enabling the user to make more informed decisions reduces the number of user inputs. Providing a rich and intuitive auditory interface enhances the operability of the device and makes the user-device interface more efficient (e.g., by helping the user to provide proper inputs and reducing user mistakes when operating/interacting with the device) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and

efficiently.

[0024] Determining whether a number of consecutive non-affirmative responses to recommendations satisfies a threshold and, if not, providing another recommendation allow the digital assistant to quickly and intuitively present options to a user. The technique reduces the number of user inputs, as the user does not need to repeatedly request new recommendations. Reducing the number of user inputs to obtain recommendations enhances the operability of the device and makes the user-device interface more efficient (e.g., by helping the user to provide proper inputs and reducing user mistakes when operating/interacting with the device) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

[0025] Determining whether a number of consecutive non-affirmative responses satisfies a threshold and, if so, requesting user inputs allow the digital assistant to quickly identify desirable content for a user. The technique reduces the number of user inputs, as the user does not need to repeatedly reject undesirable recommendations, stop the digital assistant from providing undesirable recommendations, and/or start a new search. This technique also provides natural and intuitive interactions between the digital assistant and the user, as the digital assistant automatically prompts for information when appropriate without a user command. Reducing the number of user inputs and providing a natural user interface enhance the operability of the device and make the user-device interface more efficient (e.g., by helping the user to provide proper inputs and reducing user mistakes when operating/interacting with the device) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

[0026] Detecting the physical presence of multiple users and providing a merged preference profile based on the preference profiles of these users allow for quick identification of desirable content for the multiple users. This technique reduces cognitive burden on the multiple users to identify common preferences among them, and reduces the number of inputs needed to specify the common preferences to the digital assistant and/or reject undesirable recommendations. Reducing the number of user inputs and cognitive burden enhance the operability of the device and make the user-device interface more efficient (e.g., by helping the user to provide proper

inputs and reducing user mistakes when operating/interacting with the device) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

Brief Description of the Drawings

[0027] FIG. 1 is a block diagram illustrating a system and environment for implementing a digital assistant, according to various examples.

[0028] FIG. 2A is a block diagram illustrating a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0029] FIG. 2B is a block diagram illustrating exemplary components for event handling, according to various examples.

[0030] FIG. 3 illustrates a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0031] FIG. 4 is a block diagram of an exemplary multifunction device with a display and a touch-sensitive surface, according to various examples.

[0032] FIG. 5A illustrates an exemplary user interface for a menu of applications on a portable multifunction device, according to various examples.

[0033] FIG. 5B illustrates an exemplary user interface for a multifunction device with a touch-sensitive surface that is separate from the display, according to various examples.

[0034] FIG. 6A illustrates a personal electronic device, according to various examples.

[0035] FIG. 6B is a block diagram illustrating a personal electronic device, according to various examples.

[0036] FIG. 7A is a block diagram illustrating a digital assistant system or a server portion thereof, according to various examples.

[0037] FIG. 7B illustrates the functions of the digital assistant shown in FIG. 7A, according

to various examples.

[0038] FIG. 7C illustrates a portion of an ontology, according to various examples.

[0039] FIGS. 8A-B illustrate exemplary user interfaces of an electronic device in accordance with some embodiments.

[0040] FIGS. 9A-B illustrate exemplary user interfaces of an electronic device in accordance with some embodiments.

[0041] FIGS. 10A-B illustrate exemplary user interfaces of an electronic device in accordance with some embodiments.

[0042] FIG. 11 illustrates exemplary user interfaces of an electronic device in accordance with some embodiments.

[0043] FIG. 12 illustrates a process for providing an auditory-based interface of a digital assistant, according to various examples.

[0044] FIG. 13 illustrates a process for providing an auditory-based interface of a digital assistant, according to various examples.

[0045] FIG. 14 illustrates a process for providing an auditory-based interface of a digital assistant, according to various examples.

[0046] FIG. 15 illustrates a process for providing an auditory-based interface of a digital assistant, according to various examples.

Detailed Description

[0047] In the following description of examples, reference is made to the accompanying drawings in which are shown by way of illustration specific examples that can be practiced. It is to be understood that other examples can be used and structural changes can be made without departing from the scope of the various examples.

[0048] As discussed above, a digital assistant may be implemented on an electronic device

with limited or no display capabilities. Thus, there is a need for an electronic device (or digital assistant on the electronic device) that provides primarily auditory-based interfaces. Further, there is a need for such primarily auditory-based interfaces to provide natural, intuitive, and rich interactions between the digital assistant and the user. Such techniques can enhance the operability of the device and make the interactions with the digital assistant more efficient (e.g., by helping the user to quickly obtain and understand the content being presented) which, additionally, reduces power usage and improves battery life of the device by enabling the user to use the device more quickly and efficiently.

[0049] Although the following description uses terms "first," "second," etc. to describe various elements, these elements should not be limited by the terms. These terms are only used to distinguish one element from another. For example, a first input could be termed a second input, and, similarly, a second input could be termed a first input, without departing from the scope of the various described examples. The first input and the second input are both inputs and, in some cases, are separate and different inputs.

[0050] The terminology used in the description of the various described examples herein is for the purpose of describing particular examples only and is not intended to be limiting. As used in the description of the various described examples and the appended claims, the singular forms "a," "an," and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term "and/or" as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms "includes," "including," "comprises," and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0051] The term "if" may be construed to mean "when" or "upon" or "in response to determining" or "in response to detecting," depending on the context. Similarly, the phrase "if it is determined" or "if [a stated condition or event] is detected" may be construed to mean "upon determining" or "in response to determining" or "upon detecting [the stated condition or event]"

or "in response to detecting [the stated condition or event]," depending on the context.

1. System and Environment

[0052] FIG. 1 illustrates a block diagram of system 100 according to various examples. In some examples, system 100 implements a digital assistant. The terms "digital assistant," "virtual assistant," "intelligent automated assistant," or "automatic digital assistant" refer to any information processing system that interprets natural language input in spoken and/or textual form to infer user intent, and performs actions based on the inferred user intent. For example, to act on an inferred user intent, the system performs one or more of the following: identifying a task flow with steps and parameters designed to accomplish the inferred user intent, inputting specific requirements from the inferred user intent into the task flow; executing the task flow by invoking programs, methods, services, APIs, or the like; and generating output responses to the user in an audible (e.g., speech) and/or visual form.

[0053] Specifically, a digital assistant is capable of accepting a user request at least partially in the form of a natural language command, request, statement, narrative, and/or inquiry. Typically, the user request seeks either an informational answer or performance of a task by the digital assistant. A satisfactory response to the user request includes a provision of the requested informational answer, a performance of the requested task, or a combination of the two. For example, a user asks the digital assistant a question, such as "Where am I right now?" Based on the user's current location, the digital assistant answers, "You are in Central Park near the west gate." The user also requests the performance of a task, for example, "Please invite my friends to my girlfriend's birthday party next week." In response, the digital assistant can acknowledge the request by saying "Yes, right away," and then send a suitable calendar invite on behalf of the user to each of the user's friends listed in the user's electronic address book. During performance of a requested task, the digital assistant sometimes interacts with the user in a continuous dialogue involving multiple exchanges of information over an extended period of time. There are numerous other ways of interacting with a digital assistant to request information or performance of various tasks. In addition to providing verbal responses and taking programmed actions, the digital assistant also provides responses in other visual or audio forms, e.g., as text, alerts, music, videos, animations, etc.

[0054] As shown in FIG. 1, in some examples, a digital assistant is implemented according to a client-server model. The digital assistant includes client-side portion 102 (hereafter "DA client 102") executed on user device 104 and server-side portion 106 (hereafter "DA server 106") executed on server system 108. DA client 102 communicates with DA server 106 through one or more networks 110. DA client 102 provides client-side functionalities such as user-facing input and output processing and communication with DA server 106. DA server 106 provides server-side functionalities for any number of DA clients 102 each residing on a respective user device 104.

[0055] In some examples, DA server 106 includes client-facing I/O interface 112, one or more processing modules 114, data and models 116, and I/O interface to external services 118. The client-facing I/O interface 112 facilitates the client-facing input and output processing for DA server 106. One or more processing modules 114 utilize data and models 116 to process speech input and determine the user's intent based on natural language input. Further, one or more processing modules 114 perform task execution based on inferred user intent. In some examples, DA server 106 communicates with external services 120 through network(s) 110 for task completion or information acquisition. I/O interface to external services 118 facilitates such communications.

[0056] User device 104 can be any suitable electronic device. In some examples, user device is a portable multifunctional device (e.g., device 200, described below with reference to FIG. 2A), a multifunctional device (e.g., device 400, described below with reference to FIG. 4), or a personal electronic device (e.g., device 600, described below with reference to FIG. 6A-B.) A portable multifunctional device is, for example, a mobile telephone that also contains other functions, such as PDA and/or music player functions. Specific examples of portable multifunction devices include the iPhone®, iPod Touch®, and iPad® devices from Apple Inc. of Cupertino, California. Other examples of portable multifunction devices include, without limitation, laptop or tablet computers. Further, in some examples, user device 104 is a non-portable multifunctional device. In particular, user device 104 is a desktop computer, a game console, a television, or a television set-top box. In some examples, user device 104 includes a touch-sensitive surface (e.g., touch screen displays and/or touchpads). Further, user device 104 optionally includes one or more other physical user-interface devices, such as a physical

keyboard, a mouse, and/or a joystick. Various examples of electronic devices, such as multifunctional devices, are described below in greater detail.

[0057] Examples of communication network(s) 110 include local area networks (LAN) and wide area networks (WAN), e.g., the Internet. Communication network(s) 110 is implemented using any known network protocol, including various wired or wireless protocols, such as, for example, Ethernet, Universal Serial Bus (USB), FIREWIRE, Global System for Mobile Communications (GSM), Enhanced Data GSM Environment (EDGE), code division multiple access (CDMA), time division multiple access (TDMA), Bluetooth, Wi-Fi, voice over Internet Protocol (VoIP), Wi-MAX, or any other suitable communication protocol.

[0058] Server system 108 is implemented on one or more standalone data processing apparatus or a distributed network of computers. In some examples, server system 108 also employs various virtual devices and/or services of third-party service providers (e.g., third-party cloud service providers) to provide the underlying computing resources and/or infrastructure resources of server system 108.

[0059] In some examples, user device 104 communicates with DA server 106 via second user device 122. Second user device 122 is similar or identical to user device 104. For example, second user device 122 is similar to devices 200, 400, or 600 described below with reference to FIGs. 2A, 4, and 6A-B. User device 104 is configured to communicatively couple to second user device 122 via a direct communication connection, such as Bluetooth, NFC, BTLE, or the like, or via a wired or wireless network, such as a local Wi-Fi network. In some examples, second user device 122 is configured to act as a proxy between user device 104 and DA server 106. For example, DA client 102 of user device 104 is configured to transmit information (e.g., a user request received at user device 104) to DA server 106 via second user device 122. DA server 106 processes the information and return relevant data (e.g., data content responsive to the user request) to user device 104 via second user device 122.

[0060] In some examples, user device 104 is configured to communicate abbreviated requests for data to second user device 122 to reduce the amount of information transmitted from user device 104. Second user device 122 is configured to determine supplemental information to add to the abbreviated request to generate a complete request to transmit to DA server 106. This

system architecture can advantageously allow user device 104 having limited communication capabilities and/or limited battery power (e.g., a watch or a similar compact electronic device) to access services provided by DA server 106 by using second user device 122, having greater communication capabilities and/or battery power (e.g., a mobile phone, laptop computer, tablet computer, or the like), as a proxy to DA server 106. While only two user devices 104 and 122 are shown in FIG. 1, it should be appreciated that system 100, in some examples, includes any number and type of user devices configured in this proxy configuration to communicate with DA server system 106.

[0061] Although the digital assistant shown in FIG. 1 includes both a client-side portion (e.g., DA client 102) and a server-side portion (e.g., DA server 106), in some examples, the functions of a digital assistant are implemented as a standalone application installed on a user device. In addition, the divisions of functionalities between the client and server portions of the digital assistant can vary in different implementations. For instance, in some examples, the DA client is a thin-client that provides only user-facing input and output processing functions, and delegates all other functionalities of the digital assistant to a backend server.

2. Electronic Devices

[0062] Attention is now directed toward embodiments of electronic devices for implementing the client-side portion of a digital assistant. FIG. 2A is a block diagram illustrating portable multifunction device 200 with touch-sensitive display system 212 in accordance with some embodiments. Touch-sensitive display 212 is sometimes called a "touch screen" for convenience and is sometimes known as or called a "touch-sensitive display system." Device 200 includes memory 202 (which optionally includes one or more computer-readable storage mediums), memory controller 222, one or more processing units (CPUs) 220, peripherals interface 218, RF circuitry 208, audio circuitry 210, speaker 211, microphone 213, input/output (I/O) subsystem 206, other input control devices 216, and external port 224. Device 200 optionally includes one or more optical sensors 264. Device 200 optionally includes one or more contact intensity sensors 265 for detecting intensity of contacts on device 200 (e.g., a touch-sensitive surface such as touch-sensitive display system 212 of device 200). Device 200 optionally includes one or more tactile output generators 267 for generating tactile outputs on

device 200 (e.g., generating tactile outputs on a touch-sensitive surface such as touch-sensitive display system 212 of device 200 or touchpad 455 of device 400). These components optionally communicate over one or more communication buses or signal lines 203.

[0063] As used in the specification and claims, the term "intensity" of a contact on a touch-sensitive surface refers to the force or pressure (force per unit area) of a contact (e.g., a finger contact) on the touch-sensitive surface, or to a substitute (proxy) for the force or pressure of a contact on the touch-sensitive surface. The intensity of a contact has a range of values that includes at least four distinct values and more typically includes hundreds of distinct values (e.g., at least 256). Intensity of a contact is, optionally, determined (or measured) using various approaches and various sensors or combinations of sensors. For example, one or more force sensors underneath or adjacent to the touch-sensitive surface are, optionally, used to measure force at various points on the touch-sensitive surface. In some implementations, force measurements from multiple force sensors are combined (e.g., a weighted average) to determine an estimated force of a contact. Similarly, a pressure-sensitive tip of a stylus is, optionally, used to determine a pressure of the stylus on the touch-sensitive surface. Alternatively, the size of the contact area detected on the touch-sensitive surface and/or changes thereto, the capacitance of the touch-sensitive surface proximate to the contact and/or changes thereto, and/or the resistance of the touch-sensitive surface proximate to the contact and/or changes thereto are, optionally, used as a substitute for the force or pressure of the contact on the touch-sensitive surface. In some implementations, the substitute measurements for contact force or pressure are used directly to determine whether an intensity threshold has been exceeded (e.g., the intensity threshold is described in units corresponding to the substitute measurements). In some implementations, the substitute measurements for contact force or pressure are converted to an estimated force or pressure, and the estimated force or pressure is used to determine whether an intensity threshold has been exceeded (e.g., the intensity threshold is a pressure threshold measured in units of pressure). Using the intensity of a contact as an attribute of a user input allows for user access to additional device functionality that may otherwise not be accessible by the user on a reduced-size device with limited real estate for displaying affordances (e.g., on a touch-sensitive display) and/or receiving user input (e.g., via a touch-sensitive display, a touch-sensitive surface, or a physical/mechanical control such as a knob or a button).

[0064] As used in the specification and claims, the term "tactile output" refers to physical displacement of a device relative to a previous position of the device, physical displacement of a component (e.g., a touch-sensitive surface) of a device relative to another component (e.g., housing) of the device, or displacement of the component relative to a center of mass of the device that will be detected by a user with the user's sense of touch. For example, in situations where the device or the component of the device is in contact with a surface of a user that is sensitive to touch (e.g., a finger, palm, or other part of a user's hand), the tactile output generated by the physical displacement will be interpreted by the user as a tactile sensation corresponding to a perceived change in physical characteristics of the device or the component of the device. For example, movement of a touch-sensitive surface (e.g., a touch-sensitive display or trackpad) is, optionally, interpreted by the user as a "down click" or "up click" of a physical actuator button. In some cases, a user will feel a tactile sensation such as an "down click" or "up click" even when there is no movement of a physical actuator button associated with the touch-sensitive surface that is physically pressed (e.g., displaced) by the user's movements. As another example, movement of the touch-sensitive surface is, optionally, interpreted or sensed by the user as "roughness" of the touch-sensitive surface, even when there is no change in smoothness of the touch-sensitive surface. While such interpretations of touch by a user will be subject to the individualized sensory perceptions of the user, there are many sensory perceptions of touch that are common to a large majority of users. Thus, when a tactile output is described as corresponding to a particular sensory perception of a user (e.g., an "up click," a "down click," "roughness"), unless otherwise stated, the generated tactile output corresponds to physical displacement of the device or a component thereof that will generate the described sensory perception for a typical (or average) user.

[0065] It should be appreciated that device 200 is only one example of a portable multifunction device, and that device 200 optionally has more or fewer components than shown, optionally combines two or more components, or optionally has a different configuration or arrangement of the components. The various components shown in FIG. 2A are implemented in hardware, software, or a combination of both hardware and software, including one or more signal processing and/or application-specific integrated circuits.

[0066] Memory 202 includes one or more computer-readable storage mediums. The computer-readable storage mediums are, for example, tangible and non-transitory. Memory 202 includes high-speed random access memory and also includes non-volatile memory, such as one or more magnetic disk storage devices, flash memory devices, or other non-volatile solid-state memory devices. Memory controller 222 controls access to memory 202 by other components of device 200.

[0067] In some examples, a non-transitory computer-readable storage medium of memory 202 is used to store instructions (e.g., for performing aspects of processes described below) for use by or in connection with an instruction execution system, apparatus, or device, such as a computer-based system, processor-containing system, or other system that can fetch the instructions from the instruction execution system, apparatus, or device and execute the instructions. In other examples, the instructions (e.g., for performing aspects of the processes described below) are stored on a non-transitory computer-readable storage medium (not shown) of the server system 108 or are divided between the non-transitory computer-readable storage medium of memory 202 and the non-transitory computer-readable storage medium of server system 108.

[0068] Peripherals interface 218 is used to couple input and output peripherals of the device to CPU 220 and memory 202. The one or more processors 220 run or execute various software programs and/or sets of instructions stored in memory 202 to perform various functions for device 200 and to process data. In some embodiments, peripherals interface 218, CPU 220, and memory controller 222 are implemented on a single chip, such as chip 204. In some other embodiments, they are implemented on separate chips.

[0069] RF (radio frequency) circuitry 208 receives and sends RF signals, also called electromagnetic signals. RF circuitry 208 converts electrical signals to/from electromagnetic signals and communicates with communications networks and other communications devices via the electromagnetic signals. RF circuitry 208 optionally includes well-known circuitry for performing these functions, including but not limited to an antenna system, an RF transceiver, one or more amplifiers, a tuner, one or more oscillators, a digital signal processor, a CODEC chipset, a subscriber identity module (SIM) card, memory, and so forth. RF circuitry 208

optionally communicates with networks, such as the Internet, also referred to as the World Wide Web (WWW), an intranet and/or a wireless network, such as a cellular telephone network, a wireless local area network (LAN) and/or a metropolitan area network (MAN), and other devices by wireless communication. The RF circuitry 208 optionally includes well-known circuitry for detecting near field communication (NFC) fields, such as by a short-range communication radio. The wireless communication optionally uses any of a plurality of communications standards, protocols, and technologies, including but not limited to Global System for Mobile Communications (GSM), Enhanced Data GSM Environment (EDGE), high-speed downlink packet access (HSDPA), high-speed uplink packet access (HSUPA), Evolution, Data-Only (EV-DO), HSPA, HSPA+, Dual-Cell HSPA (DC-HSPDA), long term evolution (LTE), near field communication (NFC), wideband code division multiple access (W-CDMA), code division multiple access (CDMA), time division multiple access (TDMA), Bluetooth, Bluetooth Low Energy (BTLE), Wireless Fidelity (Wi-Fi) (e.g., IEEE 802.11a, IEEE 802.11b, IEEE 802.11g, IEEE 802.11n, and/or IEEE 802.11ac), voice over Internet Protocol (VoIP), Wi-MAX, a protocol for e mail (e.g., Internet message access protocol (IMAP) and/or post office protocol (POP)), instant messaging (e.g., extensible messaging and presence protocol (XMPP), Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions (SIMPLE), Instant Messaging and Presence Service (IMPS)), and/or Short Message Service (SMS), or any other suitable communication protocol, including communication protocols not yet developed as of the filing date of this document.

[0070] Audio circuitry 210, speaker 211, and microphone 213 provide an audio interface between a user and device 200. Audio circuitry 210 receives audio data from peripherals interface 218, converts the audio data to an electrical signal, and transmits the electrical signal to speaker 211. Speaker 211 converts the electrical signal to human-audible sound waves. Audio circuitry 210 also receives electrical signals converted by microphone 213 from sound waves. Audio circuitry 210 converts the electrical signal to audio data and transmits the audio data to peripherals interface 218 for processing. Audio data are retrieved from and/or transmitted to memory 202 and/or RF circuitry 208 by peripherals interface 218. In some embodiments, audio circuitry 210 also includes a headset jack (e.g., 312, FIG. 3). The headset jack provides an interface between audio circuitry 210 and removable audio input/output peripherals, such as

output-only headphones or a headset with both output (e.g., a headphone for one or both ears) and input (e.g., a microphone).

[0071] I/O subsystem 206 couples input/output peripherals on device 200, such as touch screen 212 and other input control devices 216, to peripherals interface 218. I/O subsystem 206 optionally includes display controller 256, optical sensor controller 258, intensity sensor controller 259, haptic feedback controller 261, and one or more input controllers 260 for other input or control devices. The one or more input controllers 260 receive/send electrical signals from/to other input control devices 216. The other input control devices 216 optionally include physical buttons (e.g., push buttons, rocker buttons, etc.), dials, slider switches, joysticks, click wheels, and so forth. In some alternate embodiments, input controller(s) 260 are, optionally, coupled to any (or none) of the following: a keyboard, an infrared port, a USB port, and a pointer device such as a mouse. The one or more buttons (e.g., 308, FIG. 3) optionally include an up/down button for volume control of speaker 211 and/or microphone 213. The one or more buttons optionally include a push button (e.g., 306, FIG. 3).

[0072] A quick press of the push button disengages a lock of touch screen 212 or begin a process that uses gestures on the touch screen to unlock the device, as described in U.S. Patent Application 11/322,549, "Unlocking a Device by Performing Gestures on an Unlock Image," filed December 23, 2005, U.S. Pat. No. 7,657,849. A longer press of the push button (e.g., 306) turns power to device 200 on or off. The user is able to customize a functionality of one or more of the buttons. Touch screen 212 is used to implement virtual or soft buttons and one or more soft keyboards.

[0073] Touch-sensitive display 212 provides an input interface and an output interface between the device and a user. Display controller 256 receives and/or sends electrical signals from/to touch screen 212. Touch screen 212 displays visual output to the user. The visual output includes graphics, text, icons, video, and any combination thereof (collectively termed "graphics"). In some embodiments, some or all of the visual output correspond to user-interface objects.

[0074] Touch screen 212 has a touch-sensitive surface, sensor, or set of sensors that accepts input from the user based on haptic and/or tactile contact. Touch screen 212 and display

controller 256 (along with any associated modules and/or sets of instructions in memory 202) detect contact (and any movement or breaking of the contact) on touch screen 212 and convert the detected contact into interaction with user-interface objects (e.g., one or more soft keys, icons, web pages, or images) that are displayed on touch screen 212. In an exemplary embodiment, a point of contact between touch screen 212 and the user corresponds to a finger of the user.

[0075] Touch screen 212 uses LCD (liquid crystal display) technology, LPD (light emitting polymer display) technology, or LED (light emitting diode) technology, although other display technologies may be used in other embodiments. Touch screen 212 and display controller 256 detect contact and any movement or breaking thereof using any of a plurality of touch sensing technologies now known or later developed, including but not limited to capacitive, resistive, infrared, and surface acoustic wave technologies, as well as other proximity sensor arrays or other elements for determining one or more points of contact with touch screen 212. In an exemplary embodiment, projected mutual capacitance sensing technology is used, such as that found in the iPhone® and iPod Touch® from Apple Inc. of Cupertino, California.

[0076] A touch-sensitive display in some embodiments of touch screen 212 is analogous to the multi-touch sensitive touchpads described in the following U.S. Patents: 6,323,846 (Westerman et al.), 6,570,557 (Westerman et al.), and/or 6,677,932 (Westerman), and/or U.S. Patent Publication 2002/0015024A1. However, touch screen 212 displays visual output from device 200, whereas touch-sensitive touchpads do not provide visual output.

[0077] A touch-sensitive display in some embodiments of touch screen 212 is as described in the following applications: (1) U.S. Patent Application No. 11/381,313, "Multipoint Touch Surface Controller," filed May 2, 2006; (2) U.S. Patent Application No. 10/840,862, "Multipoint Touchscreen," filed May 6, 2004; (3) U.S. Patent Application No. 10/903,964, "Gestures For Touch Sensitive Input Devices," filed July 30, 2004; (4) U.S. Patent Application No. 11/048,264, "Gestures For Touch Sensitive Input Devices," filed January 31, 2005; (5) U.S. Patent Application No. 11/038,590, "Mode-Based Graphical User Interfaces For Touch Sensitive Input Devices," filed January 18, 2005; (6) U.S. Patent Application No. 11/228,758, "Virtual Input Device Placement On A Touch Screen User Interface," filed September 16, 2005; (7) U.S. Patent

Application No. 11/228,700, "Operation Of A Computer With A Touch Screen Interface," filed September 16, 2005; (8) U.S. Patent Application No. 11/228,737, "Activating Virtual Keys Of A Touch-Screen Virtual Keyboard," filed September 16, 2005; and (9) U.S. Patent Application No. 11/367,749, "Multi-Functional Hand-Held Device," filed March 3, 2006.

[0078] Touch screen 212 has, for example, a video resolution in excess of 100 dpi. In some embodiments, the touch screen has a video resolution of approximately 160 dpi. The user makes contact with touch screen 212 using any suitable object or appendage, such as a stylus, a finger, and so forth. In some embodiments, the user interface is designed to work primarily with finger-based contacts and gestures, which can be less precise than stylus-based input due to the larger area of contact of a finger on the touch screen. In some embodiments, the device translates the rough finger-based input into a precise pointer/cursor position or command for performing the actions desired by the user.

[0079] In some embodiments, in addition to the touch screen, device 200 includes a touchpad (not shown) for activating or deactivating particular functions. In some embodiments, the touchpad is a touch-sensitive area of the device that, unlike the touch screen, does not display visual output. The touchpad is a touch-sensitive surface that is separate from touch screen 212 or an extension of the touch-sensitive surface formed by the touch screen.

[0080] Device 200 also includes power system 262 for powering the various components. Power system 262 includes a power management system, one or more power sources (e.g., battery, alternating current (AC)), a recharging system, a power failure detection circuit, a power converter or inverter, a power status indicator (e.g., a light-emitting diode (LED)) and any other components associated with the generation, management and distribution of power in portable devices.

[0081] Device 200 also includes one or more optical sensors 264. FIG. 2A shows an optical sensor coupled to optical sensor controller 258 in I/O subsystem 206. Optical sensor 264 includes charge-coupled device (CCD) or complementary metal-oxide semiconductor (CMOS) phototransistors. Optical sensor 264 receives light from the environment, projected through one or more lenses, and converts the light to data representing an image. In conjunction with imaging module 243 (also called a camera module), optical sensor 264 captures still images or

video. In some embodiments, an optical sensor is located on the back of device 200, opposite touch screen display 212 on the front of the device so that the touch screen display is used as a viewfinder for still and/or video image acquisition. In some embodiments, an optical sensor is located on the front of the device so that the user's image is obtained for video conferencing while the user views the other video conference participants on the touch screen display. In some embodiments, the position of optical sensor 264 can be changed by the user (e.g., by rotating the lens and the sensor in the device housing) so that a single optical sensor 264 is used along with the touch screen display for both video conferencing and still and/or video image acquisition.

[0082] Device 200 optionally also includes one or more contact intensity sensors 265. FIG. 2A shows a contact intensity sensor coupled to intensity sensor controller 259 in I/O subsystem 206. Contact intensity sensor 265 optionally includes one or more piezoresistive strain gauges, capacitive force sensors, electric force sensors, piezoelectric force sensors, optical force sensors, capacitive touch-sensitive surfaces, or other intensity sensors (e.g., sensors used to measure the force (or pressure) of a contact on a touch-sensitive surface). Contact intensity sensor 265 receives contact intensity information (e.g., pressure information or a proxy for pressure information) from the environment. In some embodiments, at least one contact intensity sensor is collocated with, or proximate to, a touch-sensitive surface (e.g., touch-sensitive display system 212). In some embodiments, at least one contact intensity sensor is located on the back of device 200, opposite touch screen display 212, which is located on the front of device 200.

[0083] Device 200 also includes one or more proximity sensors 266. FIG. 2A shows proximity sensor 266 coupled to peripherals interface 218. Alternately, proximity sensor 266 is coupled to input controller 260 in I/O subsystem 206. Proximity sensor 266 is performed as described in U.S. Patent Application Nos. 11/241,839, "Proximity Detector In Handheld Device"; 11/240,788, "Proximity Detector In Handheld Device"; 11/620,702, "Using Ambient Light Sensor To Augment Proximity Sensor Output"; 11/586,862, "Automated Response To And Sensing Of User Activity In Portable Devices"; and 11/638,251, "Methods And Systems For Automatic Configuration Of Peripherals." In some embodiments, the proximity sensor turns off and disables touch screen 212 when the multifunction device is placed near the user's ear (e.g., when the user is making a phone call).

[0084] Device 200 optionally also includes one or more tactile output generators 267. FIG. 2A shows a tactile output generator coupled to haptic feedback controller 261 in I/O subsystem 206. Tactile output generator 267 optionally includes one or more electroacoustic devices such as speakers or other audio components and/or electromechanical devices that convert energy into linear motion such as a motor, solenoid, electroactive polymer, piezoelectric actuator, electrostatic actuator, or other tactile output generating component (e.g., a component that converts electrical signals into tactile outputs on the device). Contact intensity sensor 265 receives tactile feedback generation instructions from haptic feedback module 233 and generates tactile outputs on device 200 that are capable of being sensed by a user of device 200. In some embodiments, at least one tactile output generator is collocated with, or proximate to, a touch-sensitive surface (e.g., touch-sensitive display system 212) and, optionally, generates a tactile output by moving the touch-sensitive surface vertically (e.g., in/out of a surface of device 200) or laterally (e.g., back and forth in the same plane as a surface of device 200). In some embodiments, at least one tactile output generator sensor is located on the back of device 200, opposite touch screen display 212, which is located on the front of device 200.

[0085] Device 200 also includes one or more accelerometers 268. FIG. 2A shows accelerometer 268 coupled to peripherals interface 218. Alternately, accelerometer 268 is coupled to an input controller 260 in I/O subsystem 206. Accelerometer 268 performs, for example, as described in U.S. Patent Publication No. 20050190059, "Acceleration-based Theft Detection System for Portable Electronic Devices," and U.S. Patent Publication No. 20060017692, "Methods And Apparatuses For Operating A Portable Device Based On An Accelerometer." In some embodiments, information is displayed on the touch screen display in a portrait view or a landscape view based on an analysis of data received from the one or more accelerometers. Device 200 optionally includes, in addition to accelerometer(s) 268, a magnetometer (not shown) and a GPS (or GLONASS or other global navigation system) receiver (not shown) for obtaining information concerning the location and orientation (e.g., portrait or landscape) of device 200.

[0086] In some embodiments, the software components stored in memory 202 include operating system 226, communication module (or set of instructions) 228, contact/motion module (or set of instructions) 230, graphics module (or set of instructions) 232, text input

module (or set of instructions) 234, Global Positioning System (GPS) module (or set of instructions) 235, Digital Assistant Client Module 229, and applications (or sets of instructions) 236. Further, memory 202 stores data and models, such as user data and models 231. Furthermore, in some embodiments, memory 202 (FIG. 2A) or 470 (FIG. 4) stores device/global internal state 257, as shown in FIGS. 2A and 4. Device/global internal state 257 includes one or more of: active application state, indicating which applications, if any, are currently active; display state, indicating what applications, views or other information occupy various regions of touch screen display 212; sensor state, including information obtained from the device's various sensors and input control devices 216; and location information concerning the device's location and/or attitude.

[0087] Operating system 226 (e.g., Darwin, RTXC, LINUX, UNIX, OS X, iOS, WINDOWS, or an embedded operating system such as VxWorks) includes various software components and/or drivers for controlling and managing general system tasks (e.g., memory management, storage device control, power management, etc.) and facilitates communication between various hardware and software components.

[0088] Communication module 228 facilitates communication with other devices over one or more external ports 224 and also includes various software components for handling data received by RF circuitry 208 and/or external port 224. External port 224 (e.g., Universal Serial Bus (USB), FIREWIRE, etc.) is adapted for coupling directly to other devices or indirectly over a network (e.g., the Internet, wireless LAN, etc.). In some embodiments, the external port is a multi-pin (e.g., 30-pin) connector that is the same as, or similar to and/or compatible with, the 30-pin connector used on iPod® (trademark of Apple Inc.) devices.

[0089] Contact/motion module 230 optionally detects contact with touch screen 212 (in conjunction with display controller 256) and other touch-sensitive devices (e.g., a touchpad or physical click wheel). Contact/motion module 230 includes various software components for performing various operations related to detection of contact, such as determining if contact has occurred (e.g., detecting a finger-down event), determining an intensity of the contact (e.g., the force or pressure of the contact or a substitute for the force or pressure of the contact), determining if there is movement of the contact and tracking the movement across the touch-

sensitive surface (e.g., detecting one or more finger-dragging events), and determining if the contact has ceased (e.g., detecting a finger-up event or a break in contact). Contact/motion module 230 receives contact data from the touch-sensitive surface. Determining movement of the point of contact, which is represented by a series of contact data, optionally includes determining speed (magnitude), velocity (magnitude and direction), and/or an acceleration (a change in magnitude and/or direction) of the point of contact. These operations are, optionally, applied to single contacts (e.g., one finger contacts) or to multiple simultaneous contacts (e.g., "multitouch"/multiple finger contacts). In some embodiments, contact/motion module 230 and display controller 256 detect contact on a touchpad.

[0090] In some embodiments, contact/motion module 230 uses a set of one or more intensity thresholds to determine whether an operation has been performed by a user (e.g., to determine whether a user has "clicked" on an icon). In some embodiments, at least a subset of the intensity thresholds are determined in accordance with software parameters (e.g., the intensity thresholds are not determined by the activation thresholds of particular physical actuators and can be adjusted without changing the physical hardware of device 200). For example, a mouse "click" threshold of a trackpad or touch screen display can be set to any of a large range of predefined threshold values without changing the trackpad or touch screen display hardware. Additionally, in some implementations, a user of the device is provided with software settings for adjusting one or more of the set of intensity thresholds (e.g., by adjusting individual intensity thresholds and/or by adjusting a plurality of intensity thresholds at once with a system-level click "intensity" parameter).

[0091] Contact/motion module 230 optionally detects a gesture input by a user. Different gestures on the touch-sensitive surface have different contact patterns (e.g., different motions, timings, and/or intensities of detected contacts). Thus, a gesture is, optionally, detected by detecting a particular contact pattern. For example, detecting a finger tap gesture includes detecting a finger-down event followed by detecting a finger-up (liftoff) event at the same position (or substantially the same position) as the finger-down event (e.g., at the position of an icon). As another example, detecting a finger swipe gesture on the touch-sensitive surface includes detecting a finger-down event followed by detecting one or more finger-dragging events, and subsequently followed by detecting a finger-up (liftoff) event.

[0092] Graphics module 232 includes various known software components for rendering and displaying graphics on touch screen 212 or other display, including components for changing the visual impact (e.g., brightness, transparency, saturation, contrast, or other visual property) of graphics that are displayed. As used herein, the term "graphics" includes any object that can be displayed to a user, including ,without limitation, text, web pages, icons (such as user-interface objects including soft keys), digital images, videos, animations, and the like.

[0093] In some embodiments, graphics module 232 stores data representing graphics to be used. Each graphic is, optionally, assigned a corresponding code. Graphics module 232 receives, from applications etc., one or more codes specifying graphics to be displayed along with, if necessary, coordinate data and other graphic property data, and then generates screen image data to output to display controller 256.

[0094] Haptic feedback module 233 includes various software components for generating instructions used by tactile output generator(s) 267 to produce tactile outputs at one or more locations on device 200 in response to user interactions with device 200.

[0095] Text input module 234, which is, in some examples, a component of graphics module 232, provides soft keyboards for entering text in various applications (e.g., contacts 237, email 240, IM 241, browser 247, and any other application that needs text input).

[0096] GPS module 235 determines the location of the device and provides this information for use in various applications (e.g., to telephone 238 for use in location-based dialing; to camera 243 as picture/video metadata; and to applications that provide location-based services such as weather widgets, local yellow page widgets, and map/navigation widgets).

[0097] Digital assistant client module 229 includes various client-side digital assistant instructions to provide the client-side functionalities of the digital assistant. For example, digital assistant client module 229 is capable of accepting voice input (e.g., speech input), text input, touch input, and/or gestural input through various user interfaces (e.g., microphone 213, accelerometer(s) 268, touch-sensitive display system 212, optical sensor(s) 229, other input control devices 216, etc.) of portable multifunction device 200. Digital assistant client module 229 is also capable of providing output in audio (e.g., speech output), visual, and/or tactile forms

through various output interfaces (e.g., speaker 211, touch-sensitive display system 212, tactile output generator(s) 267, etc.) of portable multifunction device 200. For example, output is provided as voice, sound, alerts, text messages, menus, graphics, videos, animations, vibrations, and/or combinations of two or more of the above. During operation, digital assistant client module 229 communicates with DA server 106 using RF circuitry 208.

[0098] User data and models 231 include various data associated with the user (e.g., user-specific vocabulary data, user preference data, user-specified name pronunciations, data from the user's electronic address book, to-do lists, shopping lists, etc.) to provide the client-side functionalities of the digital assistant. Further, user data and models 231 include various models (e.g., speech recognition models, statistical language models, natural language processing models, ontology, task flow models, service models, etc.) for processing user input and determining user intent.

[0099] In some examples, digital assistant client module 229 utilizes the various sensors, subsystems, and peripheral devices of portable multifunction device 200 to gather additional information from the surrounding environment of the portable multifunction device 200 to establish a context associated with a user, the current user interaction, and/or the current user input. In some examples, digital assistant client module 229 provides the contextual information or a subset thereof with the user input to DA server 106 to help infer the user's intent. In some examples, the digital assistant also uses the contextual information to determine how to prepare and deliver outputs to the user. Contextual information is referred to as context data.

[0100] In some examples, the contextual information that accompanies the user input includes sensor information, e.g., lighting, ambient noise, ambient temperature, images or videos of the surrounding environment, etc. In some examples, the contextual information can also includes the physical state of the device, e.g., device orientation, device location, device temperature, power level, speed, acceleration, motion patterns, cellular signals strength, etc. In some examples, information related to the software state of DA server 106, e.g., running processes, installed programs, past and present network activities, background services, error logs, resources usage, etc., and of portable multifunction device 200 is provided to DA server 106 as contextual information associated with a user input.

[0101] In some examples, the digital assistant client module 229 selectively provides information (e.g., user data 231) stored on the portable multifunction device 200 in response to requests from DA server 106. In some examples, digital assistant client module 229 also elicits additional input from the user via a natural language dialogue or other user interfaces upon request by DA server 106. Digital assistant client module 229 passes the additional input to DA server 106 to help DA server 106 in intent deduction and/or fulfillment of the user's intent expressed in the user request.

[0102] A more detailed description of a digital assistant is described below with reference to FIGs. 7A-C. It should be recognized that digital assistant client module 229 can include any number of the sub-modules of digital assistant module 726 described below.

[0103] Applications 236 include the following modules (or sets of instructions), or a subset or superset thereof:

- Contacts module 237 (sometimes called an address book or contact list);
- Telephone module 238;
- Video conference module 239;
- E-mail client module 240;
- Instant messaging (IM) module 241;
- Workout support module 242;
- Camera module 243 for still and/or video images;
- Image management module 244;
- Video player module;
- Music player module;
- Browser module 247;

- Calendar module 248;
- Widget modules 249, which includes, in some examples, one or more of: weather widget 249-1, stocks widget 249-2, calculator widget 249-3, alarm clock widget 249-4, dictionary widget 249-5, and other widgets obtained by the user, as well as user-created widgets 249-6;
- Widget creator module 250 for making user-created widgets 249-6;
- Search module 251;
- Video and music player module 252, which merges video player module and music player module;
- Notes module 253;
- Map module 254; and/or
- Online video module 255.

[0104] Examples of other applications 236 that are stored in memory 202 include other word processing applications, other image editing applications, drawing applications, presentation applications, JAVA-enabled applications, encryption, digital rights management, voice recognition, and voice replication.

[0105] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, contacts module 237 are used to manage an address book or contact list (e.g., stored in application internal state 292 of contacts module 237 in memory 202 or memory 470), including: adding name(s) to the address book; deleting name(s) from the address book; associating telephone number(s), e-mail address(es), physical address(es) or other information with a name; associating an image with a name; categorizing and sorting names; providing telephone numbers or e-mail addresses to initiate and/or facilitate communications by telephone 238, video conference module 239, e-mail 240, or IM 241; and so forth.

[0106] In conjunction with RF circuitry 208, audio circuitry 210, speaker 211, microphone 213, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, telephone module 238 are used to enter a sequence of characters corresponding to a telephone number, access one or more telephone numbers in contacts module 237, modify a telephone number that has been entered, dial a respective telephone number, conduct a conversation, and disconnect or hang up when the conversation is completed. As noted above, the wireless communication uses any of a plurality of communications standards, protocols, and technologies.

[0107] In conjunction with RF circuitry 208, audio circuitry 210, speaker 211, microphone 213, touch screen 212, display controller 256, optical sensor 264, optical sensor controller 258, contact/motion module 230, graphics module 232, text input module 234, contacts module 237, and telephone module 238, video conference module 239 includes executable instructions to initiate, conduct, and terminate a video conference between a user and one or more other participants in accordance with user instructions.

[0108] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, e-mail client module 240 includes executable instructions to create, send, receive, and manage e-mail in response to user instructions. In conjunction with image management module 244, e-mail client module 240 makes it very easy to create and send e-mails with still or video images taken with camera module 243.

[0109] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, the instant messaging module 241 includes executable instructions to enter a sequence of characters corresponding to an instant message, to modify previously entered characters, to transmit a respective instant message (for example, using a Short Message Service (SMS) or Multimedia Message Service (MMS) protocol for telephony-based instant messages or using XMPP, SIMPLE, or IMPS for Internet-based instant messages), to receive instant messages, and to view received instant messages. In some embodiments, transmitted and/or received instant messages include graphics, photos, audio files, video files and/or other attachments as are supported in an

MMS and/or an Enhanced Messaging Service (EMS). As used herein, "instant messaging" refers to both telephony-based messages (e.g., messages sent using SMS or MMS) and Internet-based messages (e.g., messages sent using XMPP, SIMPLE, or IMPS).

[0110] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, GPS module 235, map module 254, and music player module, workout support module 242 includes executable instructions to create workouts (e.g., with time, distance, and/or calorie burning goals); communicate with workout sensors (sports devices); receive workout sensor data; calibrate sensors used to monitor a workout; select and play music for a workout; and display, store, and transmit workout data.

[0111] In conjunction with touch screen 212, display controller 256, optical sensor(s) 264, optical sensor controller 258, contact/motion module 230, graphics module 232, and image management module 244, camera module 243 includes executable instructions to capture still images or video (including a video stream) and store them into memory 202, modify characteristics of a still image or video, or delete a still image or video from memory 202.

[0112] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, and camera module 243, image management module 244 includes executable instructions to arrange, modify (e.g., edit), or otherwise manipulate, label, delete, present (e.g., in a digital slide show or album), and store still and/or video images.

[0113] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, browser module 247 includes executable instructions to browse the Internet in accordance with user instructions, including searching, linking to, receiving, and displaying web pages or portions thereof, as well as attachments and other files linked to web pages.

[0114] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, e-mail client module 240, and browser module 247, calendar module 248 includes executable instructions to create,

display, modify, and store calendars and data associated with calendars (e.g., calendar entries, to-do lists, etc.) in accordance with user instructions.

[0115] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, and browser module 247, widget modules 249 are mini-applications that can be downloaded and used by a user (e.g., weather widget 249-1, stocks widget 249-2, calculator widget 249-3, alarm clock widget 249-4, and dictionary widget 249-5) or created by the user (e.g., user-created widget 249-6). In some embodiments, a widget includes an HTML (Hypertext Markup Language) file, a CSS (Cascading Style Sheets) file, and a JavaScript file. In some embodiments, a widget includes an XML (Extensible Markup Language) file and a JavaScript file (e.g., Yahoo! Widgets).

[0116] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, and browser module 247, the widget creator module 250 are used by a user to create widgets (e.g., turning a user-specified portion of a web page into a widget).

[0117] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, search module 251 includes executable instructions to search for text, music, sound, image, video, and/or other files in memory 202 that match one or more search criteria (e.g., one or more user-specified search terms) in accordance with user instructions.

[0118] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, audio circuitry 210, speaker 211, RF circuitry 208, and browser module 247, video and music player module 252 includes executable instructions that allow the user to download and play back recorded music and other sound files stored in one or more file formats, such as MP3 or AAC files, and executable instructions to display, present, or otherwise play back videos (e.g., on touch screen 212 or on an external, connected display via external port 224). In some embodiments, device 200 optionally includes the functionality of an MP3 player, such as an iPod (trademark of Apple Inc.).

[0119] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, and text input module 234, notes module 253 includes executable instructions to create and manage notes, to-do lists, and the like in accordance with user instructions.

[0120] In conjunction with RF circuitry 208, touch screen 212, display controller 256, contact/motion module 230, graphics module 232, text input module 234, GPS module 235, and browser module 247, map module 254 are used to receive, display, modify, and store maps and data associated with maps (e.g., driving directions, data on stores and other points of interest at or near a particular location, and other location-based data) in accordance with user instructions.

[0121] In conjunction with touch screen 212, display controller 256, contact/motion module 230, graphics module 232, audio circuitry 210, speaker 211, RF circuitry 208, text input module 234, e-mail client module 240, and browser module 247, online video module 255 includes instructions that allow the user to access, browse, receive (e.g., by streaming and/or download), play back (e.g., on the touch screen or on an external, connected display via external port 224), send an e-mail with a link to a particular online video, and otherwise manage online videos in one or more file formats, such as H.264. In some embodiments, instant messaging module 241, rather than e-mail client module 240, is used to send a link to a particular online video. Additional description of the online video application can be found in U.S. Provisional Patent Application No. 60/936,562, "Portable Multifunction Device, Method, and Graphical User Interface for Playing Online Videos," filed June 20, 2007, and U.S. Patent Application No. 11/968,067, "Portable Multifunction Device, Method, and Graphical User Interface for Playing Online Videos," filed December 31, 2007.

[0122] Each of the above-identified modules and applications corresponds to a set of executable instructions for performing one or more functions described above and the methods described in this application (e.g., the computer-implemented methods and other information processing methods described herein). These modules (e.g., sets of instructions) need not be implemented as separate software programs, procedures, or modules, and thus various subsets of these modules can be combined or otherwise rearranged in various embodiments. For example, video player module can be combined with music player module into a single module (e.g.,

video and music player module 252, FIG. 2A). In some embodiments, memory 202 stores a subset of the modules and data structures identified above. Furthermore, memory 202 stores additional modules and data structures not described above.

[0123] In some embodiments, device 200 is a device where operation of a predefined set of functions on the device is performed exclusively through a touch screen and/or a touchpad. By using a touch screen and/or a touchpad as the primary input control device for operation of device 200, the number of physical input control devices (such as push buttons, dials, and the like) on device 200 is reduced.

[0124] The predefined set of functions that are performed exclusively through a touch screen and/or a touchpad optionally include navigation between user interfaces. In some embodiments, the touchpad, when touched by the user, navigates device 200 to a main, home, or root menu from any user interface that is displayed on device 200. In such embodiments, a "menu button" is implemented using a touchpad. In some other embodiments, the menu button is a physical push button or other physical input control device instead of a touchpad.

[0125] FIG. 2B is a block diagram illustrating exemplary components for event handling in accordance with some embodiments. In some embodiments, memory 202 (FIG. 2A) or 470 (FIG. 4) includes event sorter 270 (e.g., in operating system 226) and a respective application 236-1 (e.g., any of the aforementioned applications 237-251, 255, 480-490).

[0126] Event sorter 270 receives event information and determines the application 236-1 and application view 291 of application 236-1 to which to deliver the event information. Event sorter 270 includes event monitor 271 and event dispatcher module 274. In some embodiments, application 236-1 includes application internal state 292, which indicates the current application view(s) displayed on touch-sensitive display 212 when the application is active or executing. In some embodiments, device/global internal state 257 is used by event sorter 270 to determine which application(s) is (are) currently active, and application internal state 292 is used by event sorter 270 to determine application views 291 to which to deliver event information.

[0127] In some embodiments, application internal state 292 includes additional information, such as one or more of: resume information to be used when application 236-1 resumes

execution, user interface state information that indicates information being displayed or that is ready for display by application 236-1, a state queue for enabling the user to go back to a prior state or view of application 236-1, and a redo/undo queue of previous actions taken by the user.

[0128] Event monitor 271 receives event information from peripherals interface 218. Event information includes information about a sub-event (e.g., a user touch on touch-sensitive display 212, as part of a multi-touch gesture). Peripherals interface 218 transmits information it receives from I/O subsystem 206 or a sensor, such as proximity sensor 266, accelerometer(s) 268, and/or microphone 213 (through audio circuitry 210). Information that peripherals interface 218 receives from I/O subsystem 206 includes information from touch-sensitive display 212 or a touch-sensitive surface.

[0129] In some embodiments, event monitor 271 sends requests to the peripherals interface 218 at predetermined intervals. In response, peripherals interface 218 transmits event information. In other embodiments, peripherals interface 218 transmits event information only when there is a significant event (e.g., receiving an input above a predetermined noise threshold and/or for more than a predetermined duration).

[0130] In some embodiments, event sorter 270 also includes a hit view determination module 272 and/or an active event recognizer determination module 273.

[0131] Hit view determination module 272 provides software procedures for determining where a sub-event has taken place within one or more views when touch-sensitive display 212 displays more than one view. Views are made up of controls and other elements that a user can see on the display.

[0132] Another aspect of the user interface associated with an application is a set of views, sometimes herein called application views or user interface windows, in which information is displayed and touch-based gestures occur. The application views (of a respective application) in which a touch is detected correspond to programmatic levels within a programmatic or view hierarchy of the application. For example, the lowest level view in which a touch is detected is called the hit view, and the set of events that are recognized as proper inputs is determined based, at least in part, on the hit view of the initial touch that begins a touch-based gesture.

[0133] Hit view determination module 272 receives information related to sub events of a touch-based gesture. When an application has multiple views organized in a hierarchy, hit view determination module 272 identifies a hit view as the lowest view in the hierarchy which should handle the sub-event. In most circumstances, the hit view is the lowest level view in which an initiating sub-event occurs (e.g., the first sub-event in the sequence of sub-events that form an event or potential event). Once the hit view is identified by the hit view determination module 272, the hit view typically receives all sub-events related to the same touch or input source for which it was identified as the hit view.

[0134] Active event recognizer determination module 273 determines which view or views within a view hierarchy should receive a particular sequence of sub-events. In some embodiments, active event recognizer determination module 273 determines that only the hit view should receive a particular sequence of sub-events. In other embodiments, active event recognizer determination module 273 determines that all views that include the physical location of a sub-event are actively involved views, and therefore determines that all actively involved views should receive a particular sequence of sub-events. In other embodiments, even if touch sub-events were entirely confined to the area associated with one particular view, views higher in the hierarchy would still remain as actively involved views.

[0135] Event dispatcher module 274 dispatches the event information to an event recognizer (e.g., event recognizer 280). In embodiments including active event recognizer determination module 273, event dispatcher module 274 delivers the event information to an event recognizer determined by active event recognizer determination module 273. In some embodiments, event dispatcher module 274 stores in an event queue the event information, which is retrieved by a respective event receiver 282.

[0136] In some embodiments, operating system 226 includes event sorter 270. Alternatively, application 236-1 includes event sorter 270. In yet other embodiments, event sorter 270 is a stand-alone module, or a part of another module stored in memory 202, such as contact/motion module 230.

[0137] In some embodiments, application 236-1 includes a plurality of event handlers 290 and one or more application views 291, each of which includes instructions for handling touch

events that occur within a respective view of the application's user interface. Each application view 291 of the application 236-1 includes one or more event recognizers 280. Typically, a respective application view 291 includes a plurality of event recognizers 280. In other embodiments, one or more of event recognizers 280 are part of a separate module, such as a user interface kit (not shown) or a higher level object from which application 236-1 inherits methods and other properties. In some embodiments, a respective event handler 290 includes one or more of: data updater 276, object updater 277, GUI updater 278, and/or event data 279 received from event sorter 270. Event handler 290 utilizes or calls data updater 276, object updater 277, or GUI updater 278 to update the application internal state 292. Alternatively, one or more of the application views 291 include one or more respective event handlers 290. Also, in some embodiments, one or more of data updater 276, object updater 277, and GUI updater 278 are included in a respective application view 291.

[0138] A respective event recognizer 280 receives event information (e.g., event data 279) from event sorter 270 and identifies an event from the event information. Event recognizer 280 includes event receiver 282 and event comparator 284. In some embodiments, event recognizer 280 also includes at least a subset of: metadata 283, and event delivery instructions 288 (which include sub-event delivery instructions).

[0139] Event receiver 282 receives event information from event sorter 270. The event information includes information about a sub-event, for example, a touch or a touch movement. Depending on the sub-event, the event information also includes additional information, such as location of the sub-event. When the sub-event concerns motion of a touch, the event information also includes speed and direction of the sub-event. In some embodiments, events include rotation of the device from one orientation to another (e.g., from a portrait orientation to a landscape orientation, or vice versa), and the event information includes corresponding information about the current orientation (also called device attitude) of the device.

[0140] Event comparator 284 compares the event information to predefined event or sub-event definitions and, based on the comparison, determines an event or sub event, or determines or updates the state of an event or sub-event. In some embodiments, event comparator 284 includes event definitions 286. Event definitions 286 contain definitions of events (e.g.,

predefined sequences of sub-events), for example, event 1 (287-1), event 2 (287-2), and others. In some embodiments, sub-events in an event (287) include, for example, touch begin, touch end, touch movement, touch cancellation, and multiple touching. In one example, the definition for event 1 (287-1) is a double tap on a displayed object. The double tap, for example, comprises a first touch (touch begin) on the displayed object for a predetermined phase, a first liftoff (touch end) for a predetermined phase, a second touch (touch begin) on the displayed object for a predetermined phase, and a second liftoff (touch end) for a predetermined phase. In another example, the definition for event 2 (287-2) is a dragging on a displayed object. The dragging, for example, comprises a touch (or contact) on the displayed object for a predetermined phase, a movement of the touch across touch-sensitive display 212, and liftoff of the touch (touch end). In some embodiments, the event also includes information for one or more associated event handlers 290.

[0141] In some embodiments, event definition 287 includes a definition of an event for a respective user-interface object. In some embodiments, event comparator 284 performs a hit test to determine which user-interface object is associated with a sub-event. For example, in an application view in which three user-interface objects are displayed on touch-sensitive display 212, when a touch is detected on touch-sensitive display 212, event comparator 284 performs a hit test to determine which of the three user-interface objects is associated with the touch (sub-event). If each displayed object is associated with a respective event handler 290, the event comparator uses the result of the hit test to determine which event handler 290 should be activated. For example, event comparator 284 selects an event handler associated with the sub-event and the object triggering the hit test.

[0142] In some embodiments, the definition for a respective event (287) also includes delayed actions that delay delivery of the event information until after it has been determined whether the sequence of sub-events does or does not correspond to the event recognizer's event type.

[0143] When a respective event recognizer 280 determines that the series of sub-events do not match any of the events in event definitions 286, the respective event recognizer 280 enters an event impossible, event failed, or event ended state, after which it disregards subsequent sub-

events of the touch-based gesture. In this situation, other event recognizers, if any, that remain active for the hit view continue to track and process sub-events of an ongoing touch-based gesture.

[0144] In some embodiments, a respective event recognizer 280 includes metadata 283 with configurable properties, flags, and/or lists that indicate how the event delivery system should perform sub-event delivery to actively involved event recognizers. In some embodiments, metadata 283 includes configurable properties, flags, and/or lists that indicate how event recognizers interact, or are enabled to interact, with one another. In some embodiments, metadata 283 includes configurable properties, flags, and/or lists that indicate whether sub-events are delivered to varying levels in the view or programmatic hierarchy.

[0145] In some embodiments, a respective event recognizer 280 activates event handler 290 associated with an event when one or more particular sub-events of an event are recognized. In some embodiments, a respective event recognizer 280 delivers event information associated with the event to event handler 290. Activating an event handler 290 is distinct from sending (and deferred sending) sub-events to a respective hit view. In some embodiments, event recognizer 280 throws a flag associated with the recognized event, and event handler 290 associated with the flag catches the flag and performs a predefined process.

[0146] In some embodiments, event delivery instructions 288 include sub-event delivery instructions that deliver event information about a sub-event without activating an event handler. Instead, the sub-event delivery instructions deliver event information to event handlers associated with the series of sub-events or to actively involved views. Event handlers associated with the series of sub-events or with actively involved views receive the event information and perform a predetermined process.

[0147] In some embodiments, data updater 276 creates and updates data used in application 236-1. For example, data updater 276 updates the telephone number used in contacts module 237, or stores a video file used in video player module. In some embodiments, object updater 277 creates and updates objects used in application 236-1. For example, object updater 277 creates a new user-interface object or updates the position of a user-interface object. GUI

updater 278 updates the GUI. For example, GUI updater 278 prepares display information and sends it to graphics module 232 for display on a touch-sensitive display.

[0148] In some embodiments, event handler(s) 290 includes or has access to data updater 276, object updater 277, and GUI updater 278. In some embodiments, data updater 276, object updater 277, and GUI updater 278 are included in a single module of a respective application 236-1 or application view 291. In other embodiments, they are included in two or more software modules.

[0149] It shall be understood that the foregoing discussion regarding event handling of user touches on touch-sensitive displays also applies to other forms of user inputs to operate multifunction devices 200 with input devices, not all of which are initiated on touch screens. For example, mouse movement and mouse button presses, optionally coordinated with single or multiple keyboard presses or holds; contact movements such as taps, drags, scrolls, etc. on touchpads; pen stylus inputs; movement of the device; oral instructions; detected eye movements; biometric inputs; and/or any combination thereof are optionally utilized as inputs corresponding to sub-events which define an event to be recognized.

[0150] FIG. 3 illustrates a portable multifunction device 200 having a touch screen 212 in accordance with some embodiments. The touch screen optionally displays one or more graphics within user interface (UI) 300. In this embodiment, as well as others described below, a user is enabled to select one or more of the graphics by making a gesture on the graphics, for example, with one or more fingers 302 (not drawn to scale in the figure) or one or more styluses 303 (not drawn to scale in the figure). In some embodiments, selection of one or more graphics occurs when the user breaks contact with the one or more graphics. In some embodiments, the gesture optionally includes one or more taps, one or more swipes (from left to right, right to left, upward and/or downward), and/or a rolling of a finger (from right to left, left to right, upward and/or downward) that has made contact with device 200. In some implementations or circumstances, inadvertent contact with a graphic does not select the graphic. For example, a swipe gesture that sweeps over an application icon optionally does not select the corresponding application when the gesture corresponding to selection is a tap.

[0151] Device 200 also includes one or more physical buttons, such as "home" or menu button 304. As described previously, menu button 304 is used to navigate to any application 236 in a set of applications that is executed on device 200. Alternatively, in some embodiments, the menu button is implemented as a soft key in a GUI displayed on touch screen 212.

[0152] In one embodiment, device 200 includes touch screen 212, menu button 304, push button 306 for powering the device on/off and locking the device, volume adjustment button(s) 308, subscriber identity module (SIM) card slot 310, headset jack 312, and docking/charging external port 224. Push button 306 is, optionally, used to turn the power on/off on the device by depressing the button and holding the button in the depressed state for a predefined time interval; to lock the device by depressing the button and releasing the button before the predefined time interval has elapsed; and/or to unlock the device or initiate an unlock process. In an alternative embodiment, device 200 also accepts verbal input for activation or deactivation of some functions through microphone 213. Device 200 also, optionally, includes one or more contact intensity sensors 265 for detecting intensity of contacts on touch screen 212 and/or one or more tactile output generators 267 for generating tactile outputs for a user of device 200.

[0153] FIG. 4 is a block diagram of an exemplary multifunction device with a display and a touch-sensitive surface in accordance with some embodiments. Device 400 need not be portable. In some embodiments, device 400 is a laptop computer, a desktop computer, a tablet computer, a multimedia player device, a navigation device, an educational device (such as a child's learning toy), a gaming system, or a control device (e.g., a home or industrial controller). Device 400 typically includes one or more processing units (CPUs) 410, one or more network or other communications interfaces 460, memory 470, and one or more communication buses 420 for interconnecting these components. Communication buses 420 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. Device 400 includes input/output (I/O) interface 430 comprising display 440, which is typically a touch screen display. I/O interface 430 also optionally includes a keyboard and/or mouse (or other pointing device) 450 and touchpad 455, tactile output generator 457 for generating tactile outputs on device 400 (e.g., similar to tactile output generator(s) 267 described above with reference to FIG. 2A), sensors 459 (e.g., optical, acceleration, proximity, touch-sensitive, and/or contact intensity sensors similar to contact intensity sensor(s) 265 described

above with reference to FIG. 2A). Memory 470 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM, or other random access solid state memory devices; and optionally includes non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. Memory 470 optionally includes one or more storage devices remotely located from CPU(s) 410. In some embodiments, memory 470 stores programs, modules, and data structures analogous to the programs, modules, and data structures stored in memory 202 of portable multifunction device 200 (FIG. 2A), or a subset thereof. Furthermore, memory 470 optionally stores additional programs, modules, and data structures not present in memory 202 of portable multifunction device 200. For example, memory 470 of device 400 optionally stores drawing module 480, presentation module 482, word processing module 484, website creation module 486, disk authoring module 488, and/or spreadsheet module 490, while memory 202 of portable multifunction device 200 (FIG. 2A) optionally does not store these modules.

[0154] Each of the above-identified elements in FIG. 4 is, in some examples, stored in one or more of the previously mentioned memory devices. Each of the above-identified modules corresponds to a set of instructions for performing a function described above. The above-identified modules or programs (e.g., sets of instructions) need not be implemented as separate software programs, procedures, or modules, and thus various subsets of these modules are combined or otherwise rearranged in various embodiments. In some embodiments, memory 470 stores a subset of the modules and data structures identified above. Furthermore, memory 470 stores additional modules and data structures not described above.

[0155] Attention is now directed towards embodiments of user interfaces that can be implemented on, for example, portable multifunction device 200.

[0156] FIG. 5A illustrates an exemplary user interface for a menu of applications on portable multifunction device 200 in accordance with some embodiments. Similar user interfaces are implemented on device 400. In some embodiments, user interface 500 includes the following elements, or a subset or superset thereof:

[0157] Signal strength indicator(s) 502 for wireless communication(s), such as cellular and Wi-Fi signals;

- Time 504;
- Bluetooth indicator 505;
- Battery status indicator 506;
- Tray 508 with icons for frequently used applications, such as:
    - Icon 516 for telephone module 238, labeled "Phone," which optionally includes an indicator 514 of the number of missed calls or voicemail messages;
    - Icon 518 for e-mail client module 240, labeled "Mail," which optionally includes an indicator 510 of the number of unread e-mails;
    - Icon 520 for browser module 247, labeled "Browser;" and
    - Icon 522 for video and music player module 252, also referred to as iPod (trademark of Apple Inc.) module 252, labeled "iPod;" and
- Icons for other applications, such as:
    - Icon 524 for IM module 241, labeled "Messages;"
    - Icon 526 for calendar module 248, labeled "Calendar;"
    - Icon 528 for image management module 244, labeled "Photos;"
    - Icon 530 for camera module 243, labeled "Camera;"
    - Icon 532 for online video module 255, labeled "Online Video;"
    - Icon 534 for stocks widget 249-2, labeled "Stocks;"
    - Icon 536 for map module 254, labeled "Maps;"
    - Icon 538 for weather widget 249-1, labeled "Weather;"
    - Icon 540 for alarm clock widget 249-4, labeled "Clock;"

- Icon 542 for workout support module 242, labeled "Workout Support;"
- Icon 544 for notes module 253, labeled "Notes;" and
- Icon 546 for a settings application or module, labeled "Settings," which provides access to settings for device 200 and its various applications 236.

[0158] It should be noted that the icon labels illustrated in FIG. 5A are merely exemplary. For example, icon 522 for video and music player module 252 is optionally labeled "Music" or "Music Player." Other labels are, optionally, used for various application icons. In some embodiments, a label for a respective application icon includes a name of an application corresponding to the respective application icon. In some embodiments, a label for a particular application icon is distinct from a name of an application corresponding to the particular application icon.

[0159] FIG. 5B illustrates an exemplary user interface on a device (e.g., device 400, FIG. 4) with a touch-sensitive surface 551 (e.g., a tablet or touchpad 455, FIG. 4) that is separate from the display 550 (e.g., touch screen display 212). Device 400 also, optionally, includes one or more contact intensity sensors (e.g., one or more of sensors 457) for detecting intensity of contacts on touch-sensitive surface 551 and/or one or more tactile output generators 459 for generating tactile outputs for a user of device 400.

[0160] Although some of the examples which follow will be given with reference to inputs on touch screen display 212 (where the touch-sensitive surface and the display are combined), in some embodiments, the device detects inputs on a touch-sensitive surface that is separate from the display, as shown in FIG. 5B. In some embodiments, the touch-sensitive surface (e.g., 551 in FIG. 5B) has a primary axis (e.g., 552 in FIG. 5B) that corresponds to a primary axis (e.g., 553 in FIG. 5B) on the display (e.g., 550). In accordance with these embodiments, the device detects contacts (e.g., 560 and 562 in FIG. 5B) with the touch-sensitive surface 551 at locations that correspond to respective locations on the display (e.g., in FIG. 5B, 560 corresponds to 568 and 562 corresponds to 570). In this way, user inputs (e.g., contacts 560 and 562, and movements thereof) detected by the device on the touch-sensitive surface (e.g., 551 in FIG. 5B) are used by the device to manipulate the user interface on the display (e.g., 550 in FIG. 5B) of the

multifunction device when the touch-sensitive surface is separate from the display. It should be understood that similar methods are, optionally, used for other user interfaces described herein.

[0161] Additionally, while the following examples are given primarily with reference to finger inputs (e.g., finger contacts, finger tap gestures, finger swipe gestures), it should be understood that, in some embodiments, one or more of the finger inputs are replaced with input from another input device (e.g., a mouse-based input or stylus input). For example, a swipe gesture is, optionally, replaced with a mouse click (e.g., instead of a contact) followed by movement of the cursor along the path of the swipe (e.g., instead of movement of the contact). As another example, a tap gesture is, optionally, replaced with a mouse click while the cursor is located over the location of the tap gesture (e.g., instead of detection of the contact followed by ceasing to detect the contact). Similarly, when multiple user inputs are simultaneously detected, it should be understood that multiple computer mice are, optionally, used simultaneously, or a mouse and finger contacts are, optionally, used simultaneously.

[0162] FIG. 6A illustrates exemplary personal electronic device 600. Device 600 includes body 602. In some embodiments, device 600 includes some or all of the features described with respect to devices 200 and 400 (e.g., FIGS. 2A-4B). In some embodiments, device 600 has touch-sensitive display screen 604, hereafter touch screen 604. Alternatively, or in addition to touch screen 604, device 600 has a display and a touch-sensitive surface. As with devices 200 and 400, in some embodiments, touch screen 604 (or the touch-sensitive surface) has one or more intensity sensors for detecting intensity of contacts (e.g., touches) being applied. The one or more intensity sensors of touch screen 604 (or the touch-sensitive surface) provide output data that represents the intensity of touches. The user interface of device 600 responds to touches based on their intensity, meaning that touches of different intensities can invoke different user interface operations on device 600.

[0163] Techniques for detecting and processing touch intensity are found, for example, in related applications: International Patent Application Serial No. PCT/US2013/040061, titled "Device, Method, and Graphical User Interface for Displaying User Interface Objects Corresponding to an Application," filed May 8, 2013, and International Patent Application Serial

No. PCT/US2013/069483, titled "Device, Method, and Graphical User Interface for Transitioning Between Touch Input to Display Output Relationships," filed November 11, 2013.

[0164] In some embodiments, device 600 has one or more input mechanisms 606 and 608. Input mechanisms 606 and 608, if included, are physical. Examples of physical input mechanisms include push buttons and rotatable mechanisms. In some embodiments, device 600 has one or more attachment mechanisms. Such attachment mechanisms, if included, can permit attachment of device 600 with, for example, hats, eyewear, earrings, necklaces, shirts, jackets, bracelets, watch straps, chains, trousers, belts, shoes, purses, backpacks, and so forth. These attachment mechanisms permit device 600 to be worn by a user.

[0165] FIG. 6B depicts exemplary personal electronic device 600. In some embodiments, device 600 includes some or all of the components described with respect to FIGS. 2A, 2B, and 4. Device 600 has bus 612 that operatively couples I/O section 614 with one or more computer processors 616 and memory 618. I/O section 614 is connected to display 604, which can have touch-sensitive component 622 and, optionally, touch-intensity sensitive component 624. In addition, I/O section 614 is connected with communication unit 630 for receiving application and operating system data, using Wi-Fi, Bluetooth, near field communication (NFC), cellular, and/or other wireless communication techniques. Device 600 includes input mechanisms 606 and/or 608. Input mechanism 606 is a rotatable input device or a depressible and rotatable input device, for example. Input mechanism 608 is a button, in some examples.

[0166] Input mechanism 608 is a microphone, in some examples. Personal electronic device 600 includes, for example, various sensors, such as GPS sensor 632, accelerometer 634, directional sensor 640 (e.g., compass), gyroscope 636, motion sensor 638, and/or a combination thereof, all of which are operatively connected to I/O section 614.

[0167] Memory 618 of personal electronic device 600 is a non-transitory computer-readable storage medium, for storing computer-executable instructions, which, when executed by one or more computer processors 616, for example, cause the computer processors to perform the techniques and processes described below. The computer-executable instructions, for example, are also stored and/or transported within any non-transitory computer-readable storage medium for use by or in connection with an instruction execution system, apparatus, or device, such as a

computer-based system, processor-containing system, or other system that can fetch the instructions from the instruction execution system, apparatus, or device and execute the instructions. Personal electronic device 600 is not limited to the components and configuration of FIG. 6B, but can include other or additional components in multiple configurations.

[0168] As used here, the term "affordance" refers to a user-interactive graphical user interface object that is, for example, displayed on the display screen of devices 200, 400, 600, 800, 900, 1000, and/or 1100 (FIGS. 2, 4, 6, 8A-B, 9A-B, 10A-B, and 11). For example, an image (e.g., icon), a button, and text (e.g., hyperlink) each constitutes an affordance.

[0169] As used herein, the term "focus selector" refers to an input element that indicates a current part of a user interface with which a user is interacting. In some implementations that include a cursor or other location marker, the cursor acts as a "focus selector" so that when an input (e.g., a press input) is detected on a touch-sensitive surface (e.g., touchpad 455 in FIG. 4 or touch-sensitive surface 551 in FIG. 5B) while the cursor is over a particular user interface element (e.g., a button, window, slider or other user interface element), the particular user interface element is adjusted in accordance with the detected input. In some implementations that include a touch screen display (e.g., touch-sensitive display system 212 in FIG. 2A or touch screen 212 in FIG. 5A) that enables direct interaction with user interface elements on the touch screen display, a detected contact on the touch screen acts as a "focus selector" so that when an input (e.g., a press input by the contact) is detected on the touch screen display at a location of a particular user interface element (e.g., a button, window, slider, or other user interface element), the particular user interface element is adjusted in accordance with the detected input. In some implementations, focus is moved from one region of a user interface to another region of the user interface without corresponding movement of a cursor or movement of a contact on a touch screen display (e.g., by using a tab key or arrow keys to move focus from one button to another button); in these implementations, the focus selector moves in accordance with movement of focus between different regions of the user interface. Without regard to the specific form taken by the focus selector, the focus selector is generally the user interface element (or contact on a touch screen display) that is controlled by the user so as to communicate the user's intended interaction with the user interface (e.g., by indicating, to the device, the element of the user interface with which the user is intending to interact). For example, the location of a focus

selector (e.g., a cursor, a contact, or a selection box) over a respective button while a press input is detected on the touch-sensitive surface (e.g., a touchpad or touch screen) will indicate that the user is intending to activate the respective button (as opposed to other user interface elements shown on a display of the device).

[0170] As used in the specification and claims, the term "characteristic intensity" of a contact refers to a characteristic of the contact based on one or more intensities of the contact. In some embodiments, the characteristic intensity is based on multiple intensity samples. The characteristic intensity is, optionally, based on a predefined number of intensity samples, or a set of intensity samples collected during a predetermined time period (e.g., 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10 seconds) relative to a predefined event (e.g., after detecting the contact, prior to detecting liftoff of the contact, before or after detecting a start of movement of the contact, prior to detecting an end of the contact, before or after detecting an increase in intensity of the contact, and/or before or after detecting a decrease in intensity of the contact). A characteristic intensity of a contact is, optionally based on one or more of: a maximum value of the intensities of the contact, a mean value of the intensities of the contact, an average value of the intensities of the contact, a top 10 percentile value of the intensities of the contact, a value at the half maximum of the intensities of the contact, a value at the 90 percent maximum of the intensities of the contact, or the like. In some embodiments, the duration of the contact is used in determining the characteristic intensity (e.g., when the characteristic intensity is an average of the intensity of the contact over time). In some embodiments, the characteristic intensity is compared to a set of one or more intensity thresholds to determine whether an operation has been performed by a user. For example, the set of one or more intensity thresholds includes a first intensity threshold and a second intensity threshold. In this example, a contact with a characteristic intensity that does not exceed the first threshold results in a first operation, a contact with a characteristic intensity that exceeds the first intensity threshold and does not exceed the second intensity threshold results in a second operation, and a contact with a characteristic intensity that exceeds the second threshold results in a third operation. In some embodiments, a comparison between the characteristic intensity and one or more thresholds is used to determine whether or not to perform one or more operations (e.g., whether to perform a respective operation or forgo performing the respective operation) rather than being used to determine whether to perform a first operation or a second operation.

[0171] In some embodiments, a portion of a gesture is identified for purposes of determining a characteristic intensity. For example, a touch-sensitive surface receives a continuous swipe contact transitioning from a start location and reaching an end location, at which point the intensity of the contact increases. In this example, the characteristic intensity of the contact at the end location is based on only a portion of the continuous swipe contact, and not the entire swipe contact (e.g., only the portion of the swipe contact at the end location). In some embodiments, a smoothing algorithm is applied to the intensities of the swipe contact prior to determining the characteristic intensity of the contact. For example, the smoothing algorithm optionally includes one or more of: an unweighted sliding-average smoothing algorithm, a triangular smoothing algorithm, a median filter smoothing algorithm, and/or an exponential smoothing algorithm. In some circumstances, these smoothing algorithms eliminate narrow spikes or dips in the intensities of the swipe contact for purposes of determining a characteristic intensity.

[0172] The intensity of a contact on the touch-sensitive surface is characterized relative to one or more intensity thresholds, such as a contact-detection intensity threshold, a light press intensity threshold, a deep press intensity threshold, and/or one or more other intensity thresholds. In some embodiments, the light press intensity threshold corresponds to an intensity at which the device will perform operations typically associated with clicking a button of a physical mouse or a trackpad. In some embodiments, the deep press intensity threshold corresponds to an intensity at which the device will perform operations that are different from operations typically associated with clicking a button of a physical mouse or a trackpad. In some embodiments, when a contact is detected with a characteristic intensity below the light press intensity threshold (e.g., and above a nominal contact-detection intensity threshold below which the contact is no longer detected), the device will move a focus selector in accordance with movement of the contact on the touch-sensitive surface without performing an operation associated with the light press intensity threshold or the deep press intensity threshold. Generally, unless otherwise stated, these intensity thresholds are consistent between different sets of user interface figures.

[0173] An increase of characteristic intensity of the contact from an intensity below the light press intensity threshold to an intensity between the light press intensity threshold and the deep

press intensity threshold is sometimes referred to as a "light press" input. An increase of characteristic intensity of the contact from an intensity below the deep press intensity threshold to an intensity above the deep press intensity threshold is sometimes referred to as a "deep press" input. An increase of characteristic intensity of the contact from an intensity below the contact-detection intensity threshold to an intensity between the contact-detection intensity threshold and the light press intensity threshold is sometimes referred to as detecting the contact on the touch-surface. A decrease of characteristic intensity of the contact from an intensity above the contact-detection intensity threshold to an intensity below the contact-detection intensity threshold is sometimes referred to as detecting liftoff of the contact from the touch-surface. In some embodiments, the contact-detection intensity threshold is zero. In some embodiments, the contact-detection intensity threshold is greater than zero.

[0174] In some embodiments described herein, one or more operations are performed in response to detecting a gesture that includes a respective press input or in response to detecting the respective press input performed with a respective contact (or a plurality of contacts), where the respective press input is detected based at least in part on detecting an increase in intensity of the contact (or plurality of contacts) above a press-input intensity threshold. In some embodiments, the respective operation is performed in response to detecting the increase in intensity of the respective contact above the press-input intensity threshold (e.g., a "down stroke" of the respective press input). In some embodiments, the press input includes an increase in intensity of the respective contact above the press-input intensity threshold and a subsequent decrease in intensity of the contact below the press-input intensity threshold, and the respective operation is performed in response to detecting the subsequent decrease in intensity of the respective contact below the press-input threshold (e.g., an "up stroke" of the respective press input).

[0175] In some embodiments, the device employs intensity hysteresis to avoid accidental inputs sometimes termed "jitter," where the device defines or selects a hysteresis intensity threshold with a predefined relationship to the press-input intensity threshold (e.g., the hysteresis intensity threshold is X intensity units lower than the press-input intensity threshold or the hysteresis intensity threshold is 75%, 90%, or some reasonable proportion of the press-input intensity threshold). Thus, in some embodiments, the press input includes an increase in

intensity of the respective contact above the press-input intensity threshold and a subsequent decrease in intensity of the contact below the hysteresis intensity threshold that corresponds to the press-input intensity threshold, and the respective operation is performed in response to detecting the subsequent decrease in intensity of the respective contact below the hysteresis intensity threshold (e.g., an "up stroke" of the respective press input). Similarly, in some embodiments, the press input is detected only when the device detects an increase in intensity of the contact from an intensity at or below the hysteresis intensity threshold to an intensity at or above the press-input intensity threshold and, optionally, a subsequent decrease in intensity of the contact to an intensity at or below the hysteresis intensity, and the respective operation is performed in response to detecting the press input (e.g., the increase in intensity of the contact or the decrease in intensity of the contact, depending on the circumstances).

[0176] For ease of explanation, the descriptions of operations performed in response to a press input associated with a press-input intensity threshold or in response to a gesture including the press input are, optionally, triggered in response to detecting either: an increase in intensity of a contact above the press-input intensity threshold, an increase in intensity of a contact from an intensity below the hysteresis intensity threshold to an intensity above the press-input intensity threshold, a decrease in intensity of the contact below the press-input intensity threshold, and/or a decrease in intensity of the contact below the hysteresis intensity threshold corresponding to the press-input intensity threshold. Additionally, in examples where an operation is described as being performed in response to detecting a decrease in intensity of a contact below the press-input intensity threshold, the operation is, optionally, performed in response to detecting a decrease in intensity of the contact below a hysteresis intensity threshold corresponding to, and lower than, the press-input intensity threshold.

3. Digital Assistant System

[0177] FIG. 7A illustrates a block diagram of digital assistant system 700 in accordance with various examples. In some examples, digital assistant system 700 is implemented on a standalone computer system. In some examples, digital assistant system 700 is distributed across multiple computers. In some examples, some of the modules and functions of the digital assistant are divided into a server portion and a client portion, where the client portion resides on

one or more user devices (e.g., devices 104, 122, 200, 400, 600, 800, 900, 1000, or 1100) and communicates with the server portion (e.g., server system 108) through one or more networks, e.g., as shown in FIG. 1. In some examples, digital assistant system 700 is an implementation of server system 108 (and/or DA server 106) shown in FIG. 1. It should be noted that digital assistant system 700 is only one example of a digital assistant system, and that digital assistant system 700 can have more or fewer components than shown, can combine two or more components, or can have a different configuration or arrangement of the components. The various components shown in FIG. 7A are implemented in hardware, software instructions for execution by one or more processors, firmware, including one or more signal processing and/or application specific integrated circuits, or a combination thereof.

[0178] Digital assistant system 700 includes memory 702, one or more processors 704, input/output (I/O) interface 706, and network communications interface 708. These components can communicate with one another over one or more communication buses or signal lines 710.

[0179] In some examples, memory 702 includes a non-transitory computer-readable medium, such as high-speed random access memory and/or a non-volatile computer-readable storage medium (e.g., one or more magnetic disk storage devices, flash memory devices, or other non-volatile solid-state memory devices).

[0180] In some examples, I/O interface 706 couples input/output devices 716 of digital assistant system 700, such as displays, keyboards, touch screens, and microphones, to user interface module 722. I/O interface 706, in conjunction with user interface module 722, receives user inputs (e.g., voice input, keyboard inputs, touch inputs, etc.) and processes them accordingly. In some examples, e.g., when the digital assistant is implemented on a standalone user device, digital assistant system 700 includes any of the components and I/O communication interfaces described with respect to devices 200, 400, 600, 800, 900, 1000, or 1100 in FIGs. 2A, 4, 6A-B, 8A-B, 9A-B, 10A-B, and 11, respectively. In some examples, digital assistant system 700 represents the server portion of a digital assistant implementation, and can interact with the user through a client-side portion residing on a user device (e.g., devices 104, 200, 400, 600, 800, 900, 1000, or 1100).

[0181] In some examples, the network communications interface 708 includes wired communication port(s) 712 and/or wireless transmission and reception circuitry 714. The wired communication port(s) receives and send communication signals via one or more wired interfaces, e.g., Ethernet, Universal Serial Bus (USB), FIREWIRE, etc. The wireless circuitry 714 receives and sends RF signals and/or optical signals from/to communications networks and other communications devices. The wireless communications use any of a plurality of communications standards, protocols, and technologies, such as GSM, EDGE, CDMA, TDMA, Bluetooth, Wi-Fi, VoIP, Wi-MAX, or any other suitable communication protocol. Network communications interface 708 enables communication between digital assistant system 700 with networks, such as the Internet, an intranet, and/or a wireless network, such as a cellular telephone network, a wireless local area network (LAN), and/or a metropolitan area network (MAN), and other devices.

[0182] In some examples, memory 702, or the computer-readable storage media of memory 702, stores programs, modules, instructions, and data structures including all or a subset of: operating system 718, communications module 720, user interface module 722, one or more applications 724, and digital assistant module 726. In particular, memory 702, or the computer-readable storage media of memory 702, stores instructions for performing the processes described below. One or more processors 704 execute these programs, modules, and instructions, and reads/writes from/to the data structures.

[0183] Operating system 718 (e.g., Darwin, RTXC, LINUX, UNIX, iOS, OS X, WINDOWS, or an embedded operating system such as VxWorks) includes various software components and/or drivers for controlling and managing general system tasks (e.g., memory management, storage device control, power management, etc.) and facilitates communications between various hardware, firmware, and software components.

[0184] Communications module 720 facilitates communications between digital assistant system 700 with other devices over network communications interface 708. For example, communications module 720 communicates with RF circuitry 208 of electronic devices such as devices 200, 400, and 600 shown in FIG. 2A, 4, 6A-B, respectively. Communications module

720 also includes various components for handling data received by wireless circuitry 714 and/or wired communications port 712.

[0185] User interface module 722 receives commands and/or inputs from a user via I/O interface 706 (e.g., from a keyboard, touch screen, pointing device, controller, and/or microphone), and generate user interface objects on a display. User interface module 722 also prepares and delivers outputs (e.g., speech, sound, animation, text, icons, vibrations, haptic feedback, light, etc.) to the user via the I/O interface 706 (e.g., through displays, audio channels, speakers, touch-pads, etc.).

[0186] Applications 724 include programs and/or modules that are configured to be executed by one or more processors 704. For example, if the digital assistant system is implemented on a standalone user device, applications 724 include user applications, such as games, a calendar application, a navigation application, or an email application. If digital assistant system 700 is implemented on a server, applications 724 include resource management applications, diagnostic applications, or scheduling applications, for example.

[0187] Memory 702 also stores digital assistant module 726 (or the server portion of a digital assistant). In some examples, digital assistant module 726 includes the following sub-modules, or a subset or superset thereof: input/output processing module 728, speech-to-text (STT) processing module 730, natural language processing module 732, dialogue flow processing module 734, task flow processing module 736, service processing module 738, and speech synthesis module 740. Each of these modules has access to one or more of the following systems or data and models of the digital assistant module 726, or a subset or superset thereof: ontology 760, vocabulary index 744, user data 748, task flow models 754, service models 756, and ASR systems.

[0188] In some examples, using the processing modules, data, and models implemented in digital assistant module 726, the digital assistant can perform at least some of the following: converting speech input into text; identifying a user's intent expressed in a natural language input received from the user; actively eliciting and obtaining information needed to fully infer the user's intent (e.g., by disambiguating words, games, intentions, etc.); determining the task flow for fulfilling the inferred intent; and executing the task flow to fulfill the inferred intent.

[0189] In some examples, as shown in FIG. 7B, I/O processing module 728 interacts with the user through I/O devices 716 in FIG. 7A or with a user device (e.g., devices 104, 200, 400, or 600) through network communications interface 708 in FIG. 7A to obtain user input (e.g., a speech input) and to provide responses (e.g., as speech outputs) to the user input. I/O processing module 728 optionally obtains contextual information associated with the user input from the user device, along with or shortly after the receipt of the user input. The contextual information includes user-specific data, vocabulary, and/or preferences relevant to the user input. In some examples, the contextual information also includes software and hardware states of the user device at the time the user request is received, and/or information related to the surrounding environment of the user at the time that the user request was received. In some examples, I/O processing module 728 also sends follow-up questions to, and receive answers from, the user regarding the user request. When a user request is received by I/O processing module 728 and the user request includes speech input, I/O processing module 728 forwards the speech input to STT processing module 730 (or speech recognizer) for speech-to-text conversions.

[0190] STT processing module 730 includes one or more ASR systems. The one or more ASR systems can process the speech input that is received through I/O processing module 728 to produce a recognition result. Each ASR system includes a front-end speech pre-processor. The front-end speech pre-processor extracts representative features from the speech input. For example, the front-end speech pre-processor performs a Fourier transform on the speech input to extract spectral features that characterize the speech input as a sequence of representative multi-dimensional vectors. Further, each ASR system includes one or more speech recognition models (e.g., acoustic models and/or language models) and implements one or more speech recognition engines. Examples of speech recognition models include Hidden Markov Models, Gaussian-Mixture Models, Deep Neural Network Models, n-gram language models, and other statistical models. Examples of speech recognition engines include the dynamic time warping based engines and weighted finite-state transducers (WFST) based engines. The one or more speech recognition models and the one or more speech recognition engines are used to process the extracted representative features of the front-end speech pre-processor to produce intermediate recognitions results (e.g., phonemes, phonemic strings, and sub-words), and ultimately, text recognition results (e.g., words, word strings, or sequence of tokens). In some examples, the speech input is processed at least partially by a third-party service or on the user's device (e.g.,

device 104, 200, 400, or 600) to produce the recognition result. Once STT processing module 730 produces recognition results containing a text string (e.g., words, or sequence of words, or sequence of tokens), the recognition result is passed to natural language processing module 732 for intent deduction. In some examples, STT processing module 730 produces multiple candidate text representations of the speech input. Each candidate text representation is a sequence of words or tokens corresponding to the speech input. In some examples, each candidate text representation is associated with a speech recognition confidence score. Based on the speech recognition confidence scores, STT processing module 730 ranks the candidate text representations and provides the n-best (e.g., n highest ranked) candidate text representation(s) to natural language processing module 732 for intent deduction, where n is a predetermined integer greater than zero. For example, in one example, only the highest ranked (n=1) candidate text representation is passed to natural language processing module 732 for intent deduction. In another example, the five highest ranked (n=5) candidate text representations are passed to natural language processing module 732 for intent deduction.

[0191] More details on the speech-to-text processing are described in U.S. Utility Application Serial No. 13/236,942 for "Consolidating Speech Recognition Results," filed on September 20, 2011.

[0192] In some examples, STT processing module 730 includes and/or accesses a vocabulary of recognizable words via phonetic alphabet conversion module 731. Each vocabulary word is associated with one or more candidate pronunciations of the word represented in a speech recognition phonetic alphabet. In particular, the vocabulary of recognizable words includes a word that is associated with a plurality of candidate pronunciations. For example, the vocabulary includes the word "tomato" that is associated with the candidate pronunciations of /tə'meɪroʊ/ and /tə'mɑtoʊ/. Further, vocabulary words are associated with custom candidate pronunciations that are based on previous speech inputs from the user. Such custom candidate pronunciations are stored in STT processing module 730 and are associated with a particular user via the user's profile on the device. In some examples, the candidate pronunciations for words are determined based on the spelling of the word and one or more linguistic and/or phonetic rules. In some examples, the candidate pronunciations are manually generated, e.g., based on known canonical pronunciations.

[0193] In some examples, the candidate pronunciations are ranked based on the commonness of the candidate pronunciation. For example, the candidate pronunciation /tə'meɪɾoʊ/ is ranked higher than /tə'mɑtoʊ/, because the former is a more commonly used pronunciation (e.g., among all users, for users in a particular geographical region, or for any other appropriate subset of users). In some examples, candidate pronunciations are ranked based on whether the candidate pronunciation is a custom candidate pronunciation associated with the user. For example, custom candidate pronunciations are ranked higher than canonical candidate pronunciations. This can be useful for recognizing proper nouns having a unique pronunciation that deviates from canonical pronunciation. In some examples, candidate pronunciations are associated with one or more speech characteristics, such as geographic origin, nationality, or ethnicity. For example, the candidate pronunciation /tə'meɪɾoʊ/ is associated with the United States, whereas the candidate pronunciation /tə'mɑtoʊ/ is associated with Great Britain. Further, the rank of the candidate pronunciation is based on one or more characteristics (e.g., geographic origin, nationality, ethnicity, etc.) of the user stored in the user's profile on the device. For example, it can be determined from the user's profile that the user is associated with the United States. Based on the user being associated with the United States, the candidate pronunciation /tə'meɪɾoʊ/ (associated with the United States) is ranked higher than the candidate pronunciation /tə'mɑtoʊ/ (associated with Great Britain). In some examples, one of the ranked candidate pronunciations is selected as a predicted pronunciation (e.g., the most likely pronunciation).

[0194] When a speech input is received, STT processing module 730 is used to determine the phonemes corresponding to the speech input (e.g., using an acoustic model), and then attempt to determine words that match the phonemes (e.g., using a language model). For example, if STT processing module 730 first identifies the sequence of phonemes /tə'meɪɾoʊ/ corresponding to a portion of the speech input, it can then determine, based on vocabulary index 744, that this sequence corresponds to the word "tomato."

[0195] In some examples, STT processing module 730 uses approximate matching techniques to determine words in an utterance. Thus, for example, the STT processing module 730 determines that the sequence of phonemes /tə'meɪɾoʊ/ corresponds to the word "tomato," even if that particular sequence of phonemes is not one of the candidate sequence of phonemes for that word.

[0196] Natural language processing module 732 ("natural language processor") of the digital assistant takes the n-best candidate text representation(s) ("word sequence(s)" or "token sequence(s)") generated by STT processing module 730, and attempts to associate each of the candidate text representations with one or more "actionable intents" recognized by the digital assistant. An "actionable intent" (or "user intent") represents a task that can be performed by the digital assistant, and can have an associated task flow implemented in task flow models 754. The associated task flow is a series of programmed actions and steps that the digital assistant takes in order to perform the task. The scope of a digital assistant's capabilities is dependent on the number and variety of task flows that have been implemented and stored in task flow models 754, or in other words, on the number and variety of "actionable intents" that the digital assistant recognizes. The effectiveness of the digital assistant, however, also dependents on the assistant's ability to infer the correct "actionable intent(s)" from the user request expressed in natural language.

[0197] In some examples, in addition to the sequence of words or tokens obtained from STT processing module 730, natural language processing module 732 also receives contextual information associated with the user request, e.g., from I/O processing module 728. The natural language processing module 732 optionally uses the contextual information to clarify, supplement, and/or further define the information contained in the candidate text representations received from STT processing module 730. The contextual information includes, for example, user preferences, hardware, and/or software states of the user device, sensor information collected before, during, or shortly after the user request, prior interactions (e.g., dialogue) between the digital assistant and the user, and the like. As described herein, contextual information is, in some examples, dynamic, and changes with time, location, content of the dialogue, and other factors.

[0198] In some examples, the natural language processing is based on, e.g., ontology 760. Ontology 760 is a hierarchical structure containing many nodes, each node representing either an "actionable intent" or a "property" relevant to one or more of the "actionable intents" or other "properties." As noted above, an "actionable intent" represents a task that the digital assistant is capable of performing, i.e., it is "actionable" or can be acted on. A "property" represents a parameter associated with an actionable intent or a sub-aspect of another property. A linkage

between an actionable intent node and a property node in ontology 760 defines how a parameter represented by the property node pertains to the task represented by the actionable intent node.

[0199] In some examples, ontology 760 is made up of actionable intent nodes and property nodes. Within ontology 760, each actionable intent node is linked to one or more property nodes either directly or through one or more intermediate property nodes. Similarly, each property node is linked to one or more actionable intent nodes either directly or through one or more intermediate property nodes. For example, as shown in FIG. 7C, ontology 760 includes a "restaurant reservation" node (i.e., an actionable intent node). Property nodes "restaurant," "date/time" (for the reservation), and "party size" are each directly linked to the actionable intent node (i.e., the "restaurant reservation" node).

[0200] In addition, property nodes "cuisine," "price range," "phone number," and "location" are sub-nodes of the property node "restaurant," and are each linked to the "restaurant reservation" node (i.e., the actionable intent node) through the intermediate property node "restaurant." For another example, as shown in FIG. 7C, ontology 760 also includes a "set reminder" node (i.e., another actionable intent node). Property nodes "date/time" (for setting the reminder) and "subject" (for the reminder) are each linked to the "set reminder" node. Since the property "date/time" is relevant to both the task of making a restaurant reservation and the task of setting a reminder, the property node "date/time" is linked to both the "restaurant reservation" node and the "set reminder" node in ontology 760.

[0201] An actionable intent node, along with its linked concept nodes, is described as a "domain." In the present discussion, each domain is associated with a respective actionable intent, and refers to the group of nodes (and the relationships there between) associated with the particular actionable intent. For example, ontology 760 shown in FIG. 7C includes an example of restaurant reservation domain 762 and an example of reminder domain 764 within ontology 760. The restaurant reservation domain includes the actionable intent node "restaurant reservation," property nodes "restaurant," "date/time," and "party size," and sub-property nodes "cuisine," "price range," "phone number," and "location." Reminder domain 764 includes the actionable intent node "set reminder," and property nodes "subject" and "date/time." In some examples, ontology 760 is made up of many domains. Each domain shares one or more property

nodes with one or more other domains. For example, the "date/time" property node is associated with many different domains (e.g., a scheduling domain, a travel reservation domain, a movie ticket domain, etc.), in addition to restaurant reservation domain 762 and reminder domain 764.

[0202] While FIG. 7C illustrates two example domains within ontology 760, other domains include, for example, "find a movie," "initiate a phone call," "find directions," "schedule a meeting," "send a message," and "provide an answer to a question," "read a list," "providing navigation instructions," "provide instructions for a task" and so on. A "send a message" domain is associated with a "send a message" actionable intent node, and further includes property nodes such as "recipient(s)," "message type," and "message body." The property node "recipient" is further defined, for example, by the sub-property nodes such as "recipient name" and "message address."

[0203] In some examples, ontology 760 includes all the domains (and hence actionable intents) that the digital assistant is capable of understanding and acting upon. In some examples, ontology 760 is modified, such as by adding or removing entire domains or nodes, or by modifying relationships between the nodes within the ontology 760.

[0204] In some examples, nodes associated with multiple related actionable intents are clustered under a "super domain" in ontology 760. For example, a "travel" super-domain includes a cluster of property nodes and actionable intent nodes related to travel. The actionable intent nodes related to travel includes "airline reservation," "hotel reservation," "car rental," "get directions," "find points of interest," and so on. The actionable intent nodes under the same super domain (e.g., the "travel" super domain) have many property nodes in common. For example, the actionable intent nodes for "airline reservation," "hotel reservation," "car rental," "get directions," and "find points of interest" share one or more of the property nodes "start location," "destination," "departure date/time," "arrival date/time," and "party size."

[0205] In some examples, each node in ontology 760 is associated with a set of words and/or phrases that are relevant to the property or actionable intent represented by the node. The respective set of words and/or phrases associated with each node are the so-called "vocabulary" associated with the node. The respective set of words and/or phrases associated with each node are stored in vocabulary index 744 in association with the property or actionable intent

represented by the node. For example, returning to FIG. 7B, the vocabulary associated with the node for the property of "restaurant" includes words such as "food," "drinks," "cuisine," "hungry," "eat," "pizza," "fast food," "meal," and so on. For another example, the vocabulary associated with the node for the actionable intent of "initiate a phone call" includes words and phrases such as "call," "phone," "dial," "ring," "call this number," "make a call to," and so on. The vocabulary index 744 optionally includes words and phrases in different languages.

[0206] Natural language processing module 732 receives the candidate text representations (e.g., text string(s) or token sequence(s)) from STT processing module 730, and for each candidate representation, determines what nodes are implicated by the words in the candidate text representation. In some examples, if a word or phrase in the candidate text representation is found to be associated with one or more nodes in ontology 760 (via vocabulary index 744), the word or phrase "triggers" or "activates" those nodes. Based on the quantity and/or relative importance of the activated nodes, natural language processing module 732 selects one of the actionable intents as the task that the user intended the digital assistant to perform. In some examples, the domain that has the most "triggered" nodes is selected. In some examples, the domain having the highest confidence value (e.g., based on the relative importance of its various triggered nodes) is selected. In some examples, the domain is selected based on a combination of the number and the importance of the triggered nodes. In some examples, additional factors are considered in selecting the node as well, such as whether the digital assistant has previously correctly interpreted a similar request from a user.

[0207] User data 748 includes user-specific information, such as user-specific vocabulary, user preferences, user address, user's default and secondary languages, user's contact list, and other short-term or long-term information for each user. In some examples, natural language processing module 732 uses the user-specific information to supplement the information contained in the user input to further define the user intent. For example, for a user request "invite my friends to my birthday party," natural language processing module 732 is able to access user data 748 to determine who the "friends" are and when and where the "birthday party" would be held, rather than requiring the user to provide such information explicitly in his/her request.

[0208] It should be recognized that in some examples, natural language processing module 732 is implemented using one or more machine learning mechanisms (e.g., neural networks). In particular, the one or more machine learning mechanisms are configured to receive a candidate text representation and contextual information associated with the candidate text representation. Based on the candidate text representation and the associated contextual information, the one or more machine learning mechanism are configured to determine intent confidence scores over a set of candidate actionable intents. Natural language processing module 732 can select one or more candidate actionable intents from the set of candidate actionable intents based on the determined intent confidence scores. In some examples, an ontology (e.g., ontology 760) is also used to select the one or more candidate actionable intents from the set of candidate actionable intents.

[0209] Other details of searching an ontology based on a token string is described in U.S. Utility Application Serial No. 12/341,743 for "Method and Apparatus for Searching Using An Active Ontology," filed December 22, 2008.

[0210] In some examples, once natural language processing module 732 identifies an actionable intent (or domain) based on the user request, natural language processing module 732 generates a structured query to represent the identified actionable intent. In some examples, the structured query includes parameters for one or more nodes within the domain for the actionable intent, and at least some of the parameters are populated with the specific information and requirements specified in the user request. For example, the user says "Make me a dinner reservation at a sushi place at 7." In this case, natural language processing module 732 is able to correctly identify the actionable intent to be "restaurant reservation" based on the user input. According to the ontology, a structured query for a "restaurant reservation" domain includes parameters such as {Cuisine}, {Time}, {Date}, {Party Size}, and the like. In some examples, based on the speech input and the text derived from the speech input using STT processing module 730, natural language processing module 732 generates a partial structured query for the restaurant reservation domain, where the partial structured query includes the parameters {Cuisine = "Sushi"} and {Time = "7pm"}. However, in this example, the user's utterance contains insufficient information to complete the structured query associated with the domain. Therefore, other necessary parameters such as {Party Size} and {Date} is not specified in the

structured query based on the information currently available. In some examples, natural language processing module 732 populates some parameters of the structured query with received contextual information. For example, in some examples, if the user requested a sushi restaurant "near me," natural language processing module 732 populates a {location} parameter in the structured query with GPS coordinates from the user device.

[0211] In some examples, natural language processing module 732 identifies multiple candidate actionable intents for each candidate text representation received from STT processing module 730. Further, in some examples, a respective structured query (partial or complete) is generated for each identified candidate actionable intent. Natural language processing module 732 determines an intent confidence score for each candidate actionable intent and ranks the candidate actionable intents based on the intent confidence scores. In some examples, natural language processing module 732 passes the generated structured query (or queries), including any completed parameters, to task flow processing module 736 ("task flow processor"). In some examples, the structured query (or queries) for the m-best (e.g., m highest ranked) candidate actionable intents are provided to task flow processing module 736, where m is a predetermined integer greater than zero. In some examples, the structured query (or queries) for the m-best candidate actionable intents are provided to task flow processing module 736 with the corresponding candidate text representation(s).

[0212] Other details of inferring a user intent based on multiple candidate actionable intents determined from multiple candidate text representations of a speech input are described in U.S. Utility Application Serial No. 14/298,725 for "System and Method for Inferring User Intent From Speech Inputs," filed June 6, 2014.

[0213] Task flow processing module 736 is configured to receive the structured query (or queries) from natural language processing module 732, complete the structured query, if necessary, and perform the actions required to "complete" the user's ultimate request. In some examples, the various procedures necessary to complete these tasks are provided in task flow models 754. In some examples, task flow models 754 include procedures for obtaining additional information from the user and task flows for performing actions associated with the actionable intent.

[0214] As described above, in order to complete a structured query, task flow processing module 736 needs to initiate additional dialogue with the user in order to obtain additional information, and/or disambiguate potentially ambiguous utterances. When such interactions are necessary, task flow processing module 736 invokes dialogue flow processing module 734 to engage in a dialogue with the user. In some examples, dialogue flow processing module 734 determines how (and/or when) to ask the user for the additional information and receives and processes the user responses. The questions are provided to and answers are received from the users through I/O processing module 728. In some examples, dialogue flow processing module 734 presents dialogue output to the user via audio and/or visual output, and receives input from the user via spoken or physical (e.g., clicking) responses. Continuing with the example above, when task flow processing module 736 invokes dialogue flow processing module 734 to determine the "party size" and "date" information for the structured query associated with the domain "restaurant reservation," dialogue flow processing module 734 generates questions such as "For how many people?" and "On which day?" to pass to the user. Once answers are received from the user, dialogue flow processing module 734 then populates the structured query with the missing information, or pass the information to task flow processing module 736 to complete the missing information from the structured query.

[0215] Once task flow processing module 736 has completed the structured query for an actionable intent, task flow processing module 736 proceeds to perform the ultimate task associated with the actionable intent. Accordingly, task flow processing module 736 executes the steps and instructions in the task flow model according to the specific parameters contained in the structured query. For example, the task flow model for the actionable intent of "restaurant reservation" includes steps and instructions for contacting a restaurant and actually requesting a reservation for a particular party size at a particular time. For example, using a structured query such as: {restaurant reservation, restaurant = ABC Café, date = 3/12/2012, time = 7pm, party size = 5}, task flow processing module 736 performs the steps of: (1) logging onto a server of the ABC Café or a restaurant reservation system such as OPENTABLE®, (2) entering the date, time, and party size information in a form on the website, (3) submitting the form, and (4) making a calendar entry for the reservation in the user's calendar.

[0216] In some examples, task flow processing module 736 employs the assistance of service processing module 738 ("service processing module") to complete a task requested in the user input or to provide an informational answer requested in the user input. For example, service processing module 738 acts on behalf of task flow processing module 736 to make a phone call, set a calendar entry, invoke a map search, invoke or interact with other user applications installed on the user device, and invoke or interact with third-party services (e.g., a restaurant reservation portal, a social networking website, a banking portal, etc.). In some examples, the protocols and application programming interfaces (API) required by each service are specified by a respective service model among service models 756. Service processing module 738 accesses the appropriate service model for a service and generate requests for the service in accordance with the protocols and APIs required by the service according to the service model.

[0217] For example, if a restaurant has enabled an online reservation service, the restaurant submits a service model specifying the necessary parameters for making a reservation and the APIs for communicating the values of the necessary parameter to the online reservation service. When requested by task flow processing module 736, service processing module 738 establishes a network connection with the online reservation service using the web address stored in the service model, and send the necessary parameters of the reservation (e.g., time, date, party size) to the online reservation interface in a format according to the API of the online reservation service.

[0218] In some examples, natural language processing module 732, dialogue flow processing module 734, and task flow processing module 736 are used collectively and iteratively to infer and define the user's intent, obtain information to further clarify and refine the user intent, and finally generate a response (i.e., an output to the user, or the completion of a task) to fulfill the user's intent. The generated response is a dialogue response to the speech input that at least partially fulfills the user's intent. Further, in some examples, the generated response is output as a speech output. In these examples, the generated response is sent to speech synthesis module 740 (e.g., speech synthesizer) where it can be processed to synthesize the dialogue response in speech form. In yet other examples, the generated response is data content relevant to satisfying a user request in the speech input.

[0219] In examples where task flow processing module 736 receives multiple structured queries from natural language processing module 732, task flow processing module 736 initially processes the first structured query of the received structured queries to attempt to complete the first structured query and/or execute one or more tasks or actions represented by the first structured query. In some examples, the first structured query corresponds to the highest ranked actionable intent. In other examples, the first structured query is selected from the received structured queries based on a combination of the corresponding speech recognition confidence scores and the corresponding intent confidence scores. In some examples, if task flow processing module 736 encounters an error during processing of the first structured query (e.g., due to an inability to determine a necessary parameter), the task flow processing module 736 can proceed to select and process a second structured query of the received structured queries that corresponds to a lower ranked actionable intent. The second structured query is selected, for example, based on the speech recognition confidence score of the corresponding candidate text representation, the intent confidence score of the corresponding candidate actionable intent, a missing necessary parameter in the first structured query, or any combination thereof.

[0220] Speech synthesis module 740 is configured to synthesize speech outputs for presentation to the user. Speech synthesis module 740 synthesizes speech outputs based on text provided by the digital assistant. For example, the generated dialogue response is in the form of a text string. Speech synthesis module 740 converts the text string to an audible speech output. Speech synthesis module 740 uses any appropriate speech synthesis technique in order to generate speech outputs from text, including, but not limited, to concatenative synthesis, unit selection synthesis, diphone synthesis, domain-specific synthesis, formant synthesis, articulatory synthesis, hidden Markov model (HMM) based synthesis, and sinewave synthesis. In some examples, speech synthesis module 740 is configured to synthesize individual words based on phonemic strings corresponding to the words. For example, a phonemic string is associated with a word in the generated dialogue response. The phonemic string is stored in metadata associated with the word. Speech synthesis model 740 is configured to directly process the phonemic string in the metadata to synthesize the word in speech form.

[0221] In some examples, instead of (or in addition to) using speech synthesis module 740, speech synthesis is performed on a remote device (e.g., the server system 108), and the

synthesized speech is sent to the user device for output to the user. For example, this can occur in some implementations where outputs for a digital assistant are generated at a server system. And because server systems generally have more processing power or resources than a user device, it is possible to obtain higher quality speech outputs than would be practical with client-side synthesis.

[0222] Additional details on digital assistants can be found in the U.S. Utility Application No. 12/987,982, entitled "Intelligent Automated Assistant," filed January 10, 2011, and U.S. Utility Application No. 13/251,088, entitled "Generating and Processing Task Items That Represent Tasks to Perform," filed September 30, 2011.

4. Exemplary Techniques for Providing an Auditory-based Interface of a Digital Assistant for Media Exploration

[0223] FIGS. 8A-B, 9A-B, 10A-B, and 11 illustrate exemplary techniques including exemplary user interfaces ("UI") for providing a digital assistant in accordance with some embodiments. These figures are also used to illustrate the processes described below, including the processes 1200, 1300, 1400, and 1500 of FIGS. 12-15, respectively.

[0224] FIG. 8A shows electronic device 800. Electronic device 800 may be any of devices 200, 400, and 600 (FIGS. 2, 4, and 6A-B) in some embodiments. In the illustrated example, the electronic device 800 is an electronic device with one or more speakers, though it will be appreciated that the electronic device may be a device of any type, such as a phone, laptop computer, desktop computer, tablet, wearable device (e.g., smart watch), set-top box, television, speaker, or any combination or subcombination thereof.

[0225] In operation, the electronic device 800 provides for the exchange of natural language speech between a user and an intelligent automated assistant (or digital assistant). In some examples, the exchange is purely auditory. In some examples, the exchange is additionally or alternatively visual (e.g., by way of graphical user interface and/or one or more light indicators) and/or haptic.

[0226] In FIG. 8A, the electronic device 800 receives (e.g., via a microphone) a natural-language speech input 810 indicative of a request to the digital assistant of the electronic device

800. The natural-language speech input 810 can include any request that can be directed to the digital assistant. In some examples, the natural-language speech input includes a predetermined trigger phrase (e.g., "Hey Siri"). In some examples, the natural-language speech input includes a request for media items (e.g., "play music Rich Rubin produced", "play a rap song", "play something from the 80s", "play something upbeat"). With reference to FIG. 8A, user 802 provides the natural-language speech input 810 that includes a trigger phrase and a request for media items of a particular artist: "Hey Siri, play that new song by Adele."

[0227] In some examples, the electronic device 800 processes the natural-language speech input 810 to perform one or more tasks. In some examples, processing the natural-language speech input 810 in this manner includes providing one or more candidate text representations (e.g., text strings) of the natural-language speech input, for instance, using the STT processing module 730. As described, each of the candidate text representations may be associated with a speech recognition confidence score, and the candidate text representations may be ranked accordingly. In other examples, the natural-language input is a textual input (e.g., inputted via a touchpad of the electronic device 800) and is provided as a candidate text representation, where n=1. Textual inputs provided as candidate text representations in this manner may be assigned a maximum speech recognition confidence score, or any other speech recognition confidence score. With reference to FIG. 8A, the digital assistant provides one or more candidate text representations including a candidate text representation "hey Siri, play that new song by Adele".

[0228] In some examples, the electronic device 800 provides one or more candidate intents based on the n-best (e.g., highest ranked) candidate text representations, for instance, using the natural language processing module 732. Each of the candidate intents may be associated with an intent confidence score, and the candidate intents may be ranked accordingly. In some examples, multiple candidate intents are identified for each candidate text representation. Further, in some examples, a structured query (partial or complete) with one or more parameters is generated for each candidate intent. With reference to FIG. 8A, the digital assistant of electronic device 800 provides one or more candidate intents including a candidate intent of "obtaining recommendations for media items", which is based on the candidate text representation "hey Siri, play that new song by Adele". Further, the digital assistant of the

electronic device 800 determines a structured query with multiple parameters: {obtaining recommendations for media items, artist = Adele, media type = song, time period = new}.

[0229] Thereafter, candidate tasks are determined based on the m-best (e.g., highest ranked) candidate intents, for instance, using the task flow processing module 736. In some examples, the candidate tasks are identified based on the structured query for each of the m-best (e.g., highest ranked) candidate intents. By way of example, as described, the structured queries may be implemented according to one or more task flows, such as one or more task flows 754.

[0230] In some examples, the electronic device 800 performs a candidate task based on the identified parameters to obtain one or more results. For example, based on the structured query, a task flow processing module of the electronic device 800 (e.g., task flow processing module 736) invokes programs, methods, services, APIs, or the like, to obtain one or more results. Results may include, for example, information related to one or more media items including but not limited to a song, an audio book, a podcast, a station, a playlist, or any combination thereof. With reference to FIG. 8A, based on the structured query {obtaining recommendations for media items, artist = Adele, media type = song, time period = new}, the digital assistant performs a media search using search parameters "Adele", "song," and "new", and identifies a song titled "Hello" (hereinafter "first media item").

[0231] Thereafter, the electronic device 800 provides the first media item. With reference to FIG. 8A, the digital assistant of the electronic device 800 provides a playback 812 of the song "Hello". As depicted, the digital assistant also provides a natural-language speech output 813 including a description (e.g., verbal description) of the first media item ("Here's Hello, by Adele") while providing the playback of the first media item. The playback of the first media item and the description of the first media item may be provided concurrently in some examples.

[0232] In some examples, the electronic device 800 provides a playback of a portion of the first media item in response to the natural-language speech input 810. The portion of the first media item can be a representative sample of the media item (e.g., the chorus, the first verse). In some examples, the digital assistant provides the playback of the portion (e.g., the chorus) while providing a speech output indicative of a description associated with the first media item (e.g., "Here's Hello, by Adele"). If the user provides an affirmative response (e.g., natural-language

response) to the speech output (e.g., "Ok play this"), the electronic device 800 provides a playback of the first media item in its entirety (e.g., from the beginning). More details of the mechanism for providing layered audio outputs are provided herein.

[0233] In some examples, the electronic device 800 provides a summary and/or a listing of multiple media items identified based on the natural-language speech input 810 (e.g., "You've got a lot from this artist. Here are the first 3 out of 10 songs: Hello, ..."). In some examples, the electronic device 800 provides one or more suggestions (e.g., "Let me know if you hear something you like or if you would like to hear the next five") before providing the listing. In some examples, the electronic device foregoes providing a suggestion after presenting the suggestion for a predetermined number of times. For example, the electronic device can forego providing the suggestion "let me know if you hear something you like" after having providing it three times with respect to the same media request. Additional description of providing media recommendations is provided in U.S. Patent Application 62/399,232, "INTELLIGENT AUTOMATED ASSISTANT," filed September 23, 2016 (Attorney Docket No. 770003001300(P30584USP1)).

[0234] The electronic device 800 can receive a natural-language speech input while providing the playback of a media item. With reference to FIG. 8A, while providing the playback of the song "Hello", the electronic device 800 receives a natural-language speech input 814 ("Actually, play the one that goes, 'I'm giving you up I'm forgiving it all'"). In some examples, in response to receiving the natural-language speech input 814, the electronic device adjusts the manner in which the current playback of the first media item is provided (e.g., providing the playback at a lower volume or rate). Additional exemplary description of adjusting audio output is provided in U.S. Patent Application 62/399,232, "INTELLIGENT AUTOMATED ASSISTANT," filed September 23, 2016 (Attorney Docket No. 770003001300(P30584USP1)).

[0235] The electronic device 800 processes the natural-language input 814 in a manner consistent with what is described above with respect to the natural-language input 812. Specifically, based on the natural-language input 814, the electronic device 800 provides one or more candidate text representations, one or more candidate intents, and performs a task

associated with a highest ranked candidate intent. In the depicted example, one candidate intent corresponding to the natural language speech input 814 ("Actually, play the one that goes, 'I'm giving you up I'm forgiving it all'") is "refining a request for media". In some examples, the candidate intent is one of the m-best candidate intents.

[0236] In some examples, the electronic device 800 derives the user intent of refining a request for media based on one or more predefined phrases and natural-language equivalents of the one or more phrases. Exemplary predefined phrases include, but are not limited to: "yes, but", "what about", "how about", "only", "else", "other", "more", "less", "something more", "something less", "the new one", "the old one", "the one that goes", "the one that sounds like" "actually", "wait", "play", "no", "different", "skip", and "next". As such, the electronic device 800 can process exemplary inputs such as "Nah, what else do you have?", "Play something more upbeat", "Play something else", "Only stuff he produced in the 80s". With reference to FIG. 8A, the electronic device 800 can derive the user intent of refining a media request based on the predefined phrases in the natural-language speech input 814 (e.g., "actually", "play", "the one that goes"). Exemplary techniques for processing the natural-language input 814 are described above with respect to the natural language processing module 732. For instance, the electronic device 800 can receive a candidate text representation of the natural-language input 814 (a text string "Actually play the one that goes I'm giving you up I'm forgiving it all") (e.g., from STT processing module 730) and determine what nodes in an ontology of the digital assistant (e.g., ontology 760) are implicated by the words in the candidate text representation. Based on the quantity and/or relative importance of the activated nodes, the electronic device (more specifically, the natural language processing module) can select one of the actionable intents as the task that the user intended the digital assistant to perform.

[0237] In some examples, the electronic device 800 derives the user intent of refining a request for media based on context information. Context information includes one or more previous user interactions (e.g., user sessions) with the electronic device. For example, if the user's previous request is associated with a user intent of obtaining media recommendations (e.g., speech input 812) and/or if the user's current input corresponds to one or more properties (e.g., property nodes) in the media recommendation domain, the electronic device 800 can derive a user intent of refining the previous media request (e.g., using the one or more specified

properties). The properties in the media recommendation domain can correspond to artist, genre, lyrics, release date, or any of the search parameters described below. With reference to FIG. 8A, the electronic device 800 can derive the user intent of refining a media request based on the user's previous speech input 912 and/or a property specified in the current speech input (lyrics "I'm giving you up I'm forgiving it all").

[0238] In some examples, the electronic device 800 identifies one or more candidate tasks and corresponding parameters based on the natural-language speech input 814. With reference to FIG. 8A, the electronic device 800 identifies a candidate task of "refining a previous media request" and a parameter of "I'm giving you up I'm forgiving it all" for refining the previous media request.

[0239] Parameters identified based on the natural-language speech input can be used to refine a media request. Exemplary parameters are provided herein. In some examples, the parameters correspond to: lyrical content of a media item (e.g., "Hey Jude"), a genre (e.g., "hip hop"), a song or album title (e.g., "Hotel California"), an occasion or a time period (e.g., a season, a holiday, time of the day, a decade), an activity (e.g., working out, driving, sleeping), a location (e.g., the beach, work, home, Hawaii), a mood (e.g., upbeat), an artist (e.g., singer, producer), or any combination thereof.

[0240] In some examples, the parameters correspond to a date (e.g., release date) within a predetermined time frame. For example, the electronic device 800 stores correlations between phrases (and the natural-language equivalents of these phrases) and time frames. For example, the electronic device 800 correlates "new" with a time frame of 1 month, "recent" with a time frame of 3 months, "just came out" and "latest" with a time frame of 1 week.

[0241] In some examples, the parameters correspond to one or more people (e.g., intended audience). For example, the natural-language speech input 814 can include reference to people associated with the user, such as "What are my friends listening to?", "What is Jason playing?", "Play more music from Amy", "Play something that my friends like". The electronic device 800 processes the natural-language speech input to identify words or phrases referring to the user (e.g., "me", "for me", "I", "my"), people other than the user (e.g., "Amy"), or any combination thereof (e.g., "our", "my friends and me"). Based on these words or phrases, the electronic

device 800 obtains identification information from one or more sources (e.g., contact list, software services such as social media services and media services). In some examples, the electronic device 800 obtains the identification information by prompting the user to disambiguate between candidate interpretations (e.g., "Did you mean John Smith or John Doe?"). In some other examples, the electronic device 800 obtains the identification information based on context information such as physical presence of one or more people near the electronic device. Techniques for detecting physical presence of one or more people are discussed in more detail below.

[0242] In some examples, the parameters correspond to a source of media item. For example, the natural-language speech input 814 can include reference to a collection of media items (e.g., "what's in my library?", "play something from my weekend jam list"). As another example, the natural-language speech input 814 can include reference to an owner of media items (e.g., "play something from Jason's collection"). In response, the electronic device 800 obtains identification information and further identifies one or more media items with the appropriate permission settings, as discussed in more detail below.

[0243] In some examples, the electronic device 800 identifies the parameters for refining a media request based at least in part on context information. As discussed above, context information (or contextual information) can include information associated with an environment of the electronic device 800, e.g., lighting, ambient noise, ambient temperature, images or videos of the surrounding environment, etc. In some examples, context information includes the physical state of the electronic device 800, e.g., device orientation, device location, device temperature, power level, speed, acceleration, motion patterns, cellular signals strength, etc. Device location can be absolute (e.g., based on GPS coordinates) or relative (e.g., the device is in the user's living room, garage, bedroom). In some examples, context information includes a current time at the electronic device. In some examples, context information includes information related to a state of the digital assistant server (e.g., DA server 106), e.g., running processes, installed programs, past and present network activities, background services, error logs, resources usage, etc., and of the electronic device 800.

[0244] In some examples, context information comprises identities of people in physical proximity to the electronic device 800. In some examples, electronic device 800 can detect physical presence and/or identities of one or more users by obtaining information from one or more sources and comparing the information with known information about the one or more users to make one or more identifications. For example, the electronic device 800 can detect the physical presence of a person based on information related to an electronic device associated with the person, such as connectivity information of the person's electronic device (e.g., on the same Wi-Fi network, within Bluetooth range, within NFC range). As another example, the electronic device 800 can detect the physical presence of a person based on facial characteristics and/or voice characteristics of the person (captured via, for instance, cameras and microphones). As another example, the electronic device 800 can detect the physical presence of a person based on information available locally, such as contacts listed in a calendar invite (set for the current time) or an email message. As still another example, the electronic device 800 can detect the physical presence of a person based on credentials provided by the person (e.g., user name and password). In some examples, the electronic device 800 prompts for disambiguation input (e.g., "is that Jason or John that I'm hearing?") and/or confirmation (e.g., "Did John just join the party?") after detecting the physical presence of a person.

[0245] In some examples, context information comprises information related to the people in physical proximity with the electronic device 800. For example, context information can include the preferences, media collections, history of each of the people detected to be in physical proximity the electronic device. For example, if the electronic device 800 determines that the user's friend Amy has uttered "Play something I'd also like", the electronic device identifies Amy's preferences (favorite genre, explicit language settings) from one or more sources and uses the preferences as search parameters for refining the media request. Additional information regarding providing a merged preference profile for multiple people is provided below.

[0246] In some examples, context information comprises information related to the first media item. For example, if the user utters "play something more recent than this" in response to the recommendation of the first media item, the electronic device 800 derives a time parameter based on the release date of the first media item.

[0247] It should be appreciated that the above-described parameters for refining a media request are merely exemplary. It should be further appreciated that the electronic device can receive a user request for refining a media request at any time when the electronic device is processing the original media request and/or when the electronic device is providing (e.g. providing information related to or playback of) one or more media items based on the original media request. It should be further appreciated that the use of a natural-language speech input (e.g., the natural-language speech input 814) to refine a media request is merely exemplary. In some examples, the electronic device 800 can initiate a process for refining a media request and/or providing additional media items in response to receiving an input via one or more sensors of the electronic device (e.g., a tactile input, a gesture input, a button press). Additional exemplary descriptions of performing media searches are provided in U.S. Patent Application 62/347,480, "INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION", filed June 8, 2016 (Attorney Docket No. 770003000600(P30491USP1)). Additional exemplary descriptions of obtaining context information is provided in U.S. Patent Application 62/507,056, "PROVIDING AN AUDITORY-BASED INTERFACE OF A DIGITAL ASSISTANT", filed May 16, 2017 (Attorney Docket No. 770003015700(P34183USP1)).

[0248] After determining that the natural-language speech input 814 corresponds to the user intent of refining the previous media request (e.g., one of the m-best candidate intents), the electronic device identifies a second media item different from the first media item. The second media item can be a song, an audio book, a podcast, a station, a playlist, or any combination thereof. With reference to FIG. 8A, the electronic device 800 identifies the second media item (e.g., "Send My Love") based on the parameters in the speech input 812 ("Adele", "new", "song") and the parameters in the speech input 814 ("I've giving you up I'm forgiving it all").

[0249] In some examples, based on the natural-language speech input 812 ("Hey Siri, play that new song by Adele"), the electronic device 800 identifies a first set of media items (e.g., a set of songs by Adele released in the past three months). From the first set of media items, the electronic device selects the song "Hello" (e.g., based on a popularity ranking) to provide to the user. Thereafter, based on the subsequent natural-language speech input 814, the electronic device 800 identifies a subset of the first set of media items based on the specified parameter derived from the natural-language speech input 814 (e.g., only songs including the lyrics "I'm

giving you up I'm forgiving it all" from the first set of songs). In some examples, identifying the subset of the first set of media items comprises determining whether a media item of the first set is associated with content (e.g., lyrics, script) or metadata (genre, release date) that matches the specified parameter in the natural-language input 814. If so, the electronic device 800 then selects the second media item (the song "Send My Love") from the subset to provide to the user. If not, the electronic device 800 foregoes selecting the second media item to provide to the user.

[0250] In some examples, the electronic device identifies the first media item and/or the second media item from a user-specific corpus of media items. In some examples, the electronic device 800 identifies the user-specific corpus based on acoustic information associated with a user input (e.g., natural-language speech input 814). The user-specific corpus is generated based on data associated with the user (e.g., preferences, settings, previous requests, previous selections, previous rejections, previous user purchases, user-specific playlists). In some examples, at least part of the user-specific corpus is generated based on a software service (e.g., a media service or social media service). For example, the user-specific corpus associates media items previously rejected or disliked by the user (e.g., on a software service) with low rankings, or does not include these media items. As another example, the user-specific corpus includes data corresponding to media items owned/purchased by the user on the software service. As yet another example, the user-specific corpus includes data corresponding to media items created by the user on the software service (e.g., a playlist). As discussed above, the electronic device can identify a media item by determining whether a media item in the user-specific corpus is associated with metadata or content that matches the specified search parameters. In some examples, at least one media item in the user-specified corpus includes metadata indicative of: an activity (e.g., working out, sleeping); a mood (e.g., upbeat, calming, sad); an occasion (e.g., birthday); a time period (e.g., 80s), a location; a curator (e.g., Rolling stones lists); a collection (e.g., summer playlist); one or more previous user inputs (previous rejections by user, previous likes by user); or any combination thereof. Additional exemplary descriptions of a user-specific corpus is provided in U.S. Patent Application 62/347,480, "INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION," filed June 8, 2016 (Attorney Docket No. 770003000600(P30491USP1)).

[0251] In some examples, at least one media item in the user-specific corpus includes metadata that is based on information from a person different from the user that has provided the media request. For example, a media item can be associated with the "beach" location based on the frequency at which it is played by all users of a software application (e.g., a media service such as iTunes) at locations corresponding to beaches. As another example, a media item can be associated with an activity (e.g., partying) based on the number of times it has been played by the user's friends (i.e., associated with the user on a social media service) and/or by people from a similar demographic segment. In some examples, the metadata is generated on a remote device different from the electronic device 800. In some examples, at least one media item in the user-specific corpus is a media item that the user is not authorized to access (e.g., has not purchased), but another person in physical proximity to the electronic device 800 is authorized to, as discussed in more detail.

[0252] Thereafter, the electronic device 800 provides the second media item. In some examples, the second media item is provided in a manner consistent with what is described above with respect to the first media item. With reference to FIG. 8A, the digital assistant of the electronic device 800 provides a playback 816 of the song "Send My Love". As depicted, the digital assistant also provides a natural-language speech output 817 including a description of the second media item ("Here's Send My Love"), for instance, while providing the playback of the second media item. In some examples (not depicted), the electronic device 800 provides a playback of a representative sample of the second media item in response to the natural-language speech input 814 and requires a user confirmation before providing the second media item in its entirety. In some examples, the electronic device 800 provides a summary and/or a listing of multiple media items identified based on the natural-language speech input 814 (e.g., "I've found two songs with those lyrics: Send My Love, Send My Love Acoustic Version...").

[0253] In some examples, with reference to FIG. 8B, the electronic device 800 receives a third natural-language speech input 818 ("Hey Siri, add this to my Saturday Morning playlist"). Based on the third natural-language speech input 818, the electronic device determines a user intent of associating a media item with a collection of media items. The electronic device can determine the user intent based on context information (e.g., currently/previously played media items). In the depicted example, the electronic device 800 associates the currently played song

"Send My Love" to a playlist named "Saturday Morning" and provides a speech output 820 indicative of the association ("Done"). In another example (not depicted), the electronic device can receive a natural-language speech input "Add the last 10 songs to a new playlist called New Favs". In response, the electronic device can create a new collection of media items named "New Favs" and associate the previously played 10 songs with the new collection.

[0254] In some examples, the electronic device 800 receives a fourth natural-language speech input 822 ("Is Adele on tour?") while providing the playback of "Send My Love". Based on the fourth natural-language speech input 822, the electronic device determines a user intent of obtaining information (e.g., the artist, the release date, related interviews, back story, meaning of the lyrics, touring information, which of the user's friends have listened to the media item) related to a particular media item. In some examples, the particular media item is identified based on context information (the song being played, the song previously played). In the depicted example, the electronic device 800 determines a user intent of obtaining touring information related to Adele, the singer of the currently played song, and provides a speech output 824 indicative of the information ("Yes, Adele will be in your city next month. Do you want tickets?"). In the depicted example, the user provides a negative response 826 ("Not now"). In another example (not depicted), the user can provide an affirmative response, and the electronic device 800 can initiate a process for purchasing concert tickets. In some examples, the electronic device 800 provides the information automatically without the fourth natural-language speech input 822.

[0255] In some examples, while playing a media item, the electronic device 800 may provide a speech output indicative of another media item. By way of example, while providing the second media item ("Send My Love"), the electronic device 800 provides a speech output 828 indicative of a third media item to be played ("Next up is Someone Like You by Adele"). After providing the second media item, the electronic device 800 provides the third media item. In some examples, while playing a media item, the electronic device 800 may receive a natural-language speech input indicative of a location (e.g., "Play this in the garage"). In response, the electronic device 800 identifies another electronic device (e.g., a speaker associated with the user's garage, a phone that is physically located in the user's garage) based on the specified location and causes the identified electronic device to provide the playback of the media item. In

some examples, the electronic device 800 can send information related to the playback (e.g., identification information of the media item, progress of the playback, playback settings such as volume) to the identified electronic device (e.g., directly or via a remote device). Additional descriptions for processing a natural-language speech input indicative of a location and processing a media request accordingly can be found, for example, in U.S. Utility Application No. 14/503,105, entitled "INTELLIGENT ASSISTANT FOR HOME AUTOMATION", filed September 30, 2014 (Attorney Docket No. 106842108200(P23013US1)), U.S. Provisional Patent Application Serial No. 62/348,015, entitled "INTELLIGENT AUTOMATED ASSISTANT IN A HOME ENVIRONMENT," filed June 9, 2016 (Attorney Docket No. 770003000100(P30331USP1)), and U.S. Provisional Patent Application Serial No. 62/348,896, entitled "INTELLIGENT DEVICE ARBITRATION AND CONTROL," filed June 11, 2016 (Attorney Docket No. 770003001400(P30585USP1)).

[0256] FIGS. 9A-B show electronic device 900. Electronic device 900 may be any of devices 200, 400, 600, and 800 (FIGS. 2, 4, 6A-B and 8A-B) in some embodiments. In the illustrated example, the electronic device 900 is an electronic device with one or more speakers, though it will be appreciated that the electronic device may be a device of any type, such as a phone, laptop computer, desktop computer, tablet, wearable device (e.g., smart watch), set-top box, television, speaker, or any combination or subcombination thereof.

[0257] With reference to FIG. 9A, the electronic device 900 receives (e.g., via a microphone) a natural-language speech input 910 indicative of a request to the digital assistant of the electronic device 900. The natural-language speech input 910 can include any request that can be directed to the digital assistant. In some examples, the natural-language speech input includes a predetermined trigger phrase (e.g., "Hey Siri"). In the depicted example in FIG. 9A, the user 902 provides the natural-language speech input 910 that includes a trigger phrase and a request for media items: "Hey Siri, what music do you have for me today?"

[0258] The electronic device 900 processes natural-language speech inputs in a manner consistent with what is discussed above with respect to electronic device 800. For example, the electronic device 900 processes the natural-language speech input 910 to provide one or more candidate text representations (e.g., a text representation "hey Siri, what music do you have for

me today") and one or more candidate intents (e.g., a user intent of "obtaining media recommendations").

[0259] The electronic device 900 identifies a task based on the natural-language speech input 910. In some examples, the electronic device 900 identifies one or more candidate tasks based on the one or more candidate intents (which in turn are identified based on one or more candidate text representations), as discussed above. Further, the electronic device performs a candidate task to obtain one or more results. The one or more results can include information related to a media item such as a song, an audio book, a podcast, a station, a playlist, or a combination thereof. In the depicted example in FIG. 9A, the electronic device 900 identifies a candidate task of "providing a media item" with parameters "for me" and "music" based on the natural-language speech input 910. Further, the electronic device 900 performs the identified task to obtain information related to a playlist "Transgressive New Releases".

[0260] With reference to FIG. 9A, the electronic device 900 provides a speech output 914 indicative of a verbal response associated with the identified task. Specifically, the electronic device 900 provides a verbal description of the identified playlist ("I've got the playlist Transgressive New Releases"). In some examples, the verbal description provided to the user includes information (e.g., metadata) corresponding to the identified media item, parameters identified from the user request, or a combination thereof. For example, in response to a user request "Play something from my favorite artists", the electronic device 900 can provide a speech output "Here's a song by Adele, one of your favorite singers, released last week". The electronic device 900 can determine the user's favorite singers based on metadata of a user-specific corpus. As discussed with reference to FIG. 8, in some examples, at least one media item in the user-specified corpus includes metadata indicative of one or more previous user inputs (previously rejected by user, previously liked by user, previously searched by user). Alternatively, in some other examples, the electronic device 900 can determine the user's favorite singers based on user preference data (e.g., user data and models 231).

[0261] In some examples, the electronic device 900 provides the speech output 914 in accordance with one or more text-to-speech modes. For example, the speech output 914 can be provided in a voice of the digital assistant, a voice (e.g., artist, DJ) associated with the media

item, or a combination thereof. Additional exemplary descriptions of using different text-to-speech modes is provided in U.S. Patent Application 62/507,056, "PROVIDING AN AUDITORY-BASED INTERFACE OF A DIGITAL ASSISTANT", filed May 16, 2017 (Attorney Docket No. 770003015700(P34183USP1)).

[0262] While providing the speech output indicative of a verbal response (e.g., speech output 914), the electronic device 900 simultaneously provides an audio output 912, which is a playback of a media item corresponding to the verbal response. In some examples, the media item being played back is a portion (e.g., a representative sample) of the identified media item. For example, if the identified media item is a single song, the playback can include the chorus or first verse of the song. As another example, if the identified media item is a playlist, the playback can include a 5-second segment for each of the songs in the playlist.

[0263] In some examples, the electronic device 900 provides the playback at a different volume (e.g., lower) than the speech output. In some examples, the electronic device 900 provides the playback at a different fidelity (e.g., lower) than the speech output. In some examples, the electronic device 900 begins providing the audio output 912 prior to providing the speech output 914. In other examples, the electronic device 900 begins providing the audio output 912 and the speech output 914 simultaneously. In yet other examples, the electronic device 900 begins providing the speech output 914 prior to providing the audio output 912. Additional description of adjusting audio during playback is provided in U.S. Patent Application 62/399,232, "INTELLIGENT AUTOMATED ASSISTANT," filed September 23, 2016 (Attorney Docket No. 770003001300(P30584USP1)).

[0264] In some examples, while providing playback 912, the electronic device receives a natural language speech input 916 ("Play it!"). In response to receiving the natural-language speech input 916, the electronic device 900 provides an audio output 918, which is a playback of the identified media item in its entirety (e.g., from the beginning). In some examples, the playback 918 is provided at a different volume and/or fidelity than the playback 912.

[0265] In some examples, while providing playback 918, the electronic device 800 provides a speech output (not depicted) indicative of information related to a media item. The information can correspond to, for instance, trivia of a song ("this was released last week"), touring

information of an artist ("This artist is coming to California next month. Want tickets?"), or news ("This artist just got engaged. Let me know if you want to know more about that."). The media item can be the media item being played back, previously played, or to be played by the electronic device 900.

[0266] In some examples, instead of simultaneously providing two layers of audio (e.g., a verbal description and a representative sample of the identified media item), the electronic device 900 provides a description of the identified media item without providing a playback of a representative sample. In some examples, the user 902 can provide a follow-up request to hear the representative sample (e.g., "What does it sound like?", "What kind of songs is in the playlist?"). In response, the electronic device 900 provides a playback of the representative sample (e.g., "Let's take a listen. <30-second summary>") and, in some instances, prompts user for additional input ("Would you like me to play it?").

[0267] It should be appreciated that the above-described techniques for providing multiple layers of audio are merely exemplary. Generally, the electronic device 900 can provide layered and/or coordinated audio information as part of any interaction between the digital assistant and the user. For example, with reference to FIG. 9B, the electronic device 900 receives a natural-language speech input 918 ("Hey Siri, how did my team do?"). Based on the input, the electronic device 900 identifies a task (e.g., a candidate task of "obtaining scores of a sports event") and one or more parameters (e.g., "Giants"), and performs the task to obtain one or more results (e.g., scores). In some other examples, the electronic device 900 can determine the one or more parameters based on user preference data (e.g., user data and models 231). The electronic device 900 provides a speech output 922 indicative of a verbal response associated with the identified task. In the depicted example, the speech output 922 is indicative of a verbal description of the obtained results ("The Giants won yesterday, the score was . . .").

[0268] While providing the speech output 922, the electronic device 900 also provides a playback 920 of a media item corresponding to the verbal response. In the depicted example, the media item is a sound effect corresponding to a winning score (e.g., crowd cheering). In some examples, the sound effect is a pre-recorded audio (e.g., generic sound effect, audio recorded at the related sports event) or a live stream (e.g., sound of rain at the current location of the

electronic device). In some examples, the speech output 922 is provided at a different volume (e.g., higher) and/or fidelity (e.g., higher) than the playback 920.

[0269] It should be appreciated that the digital assistant of the electronic device 900 can interact with a user (e.g., provide information) in a variety of text-to-speech modes, voices, and sequences. Generally, the digital assistant can coordinate between the various layers (e.g., background audio, foreground audio) and various types (e.g., sound effects, speech, music) of audio outputs to provide an intuitive, rich, and natural user interface. For example, the electronic device can adjust the timing, volume, fidelity, and content of the background audio based on the timing, volume, fidelity, and content of the foreground audio.

[0270] FIGS. 10A-B show electronic device 1000. Electronic device 1000 may be any of devices 200, 400, 600, 800, and 900 (FIGS. 2, 4, 6A-B, 8A-B, and 9A-B) in some embodiments. In the illustrated example, the electronic device 1000 is an electronic device with one or more speakers, though it will be appreciated that the electronic device may be a device of any type, such as a phone, laptop computer, desktop computer, tablet, wearable device (e.g., smart watch), set-top box, television, speaker, or any combination or subcombination thereof.

[0271] With reference to FIG. 10A, the electronic device 1000 receives (e.g., via a microphone) a natural-language speech input 1010 indicative of a request for media to the digital assistant of the electronic device 1000 ("Hey Siri, what should I listen to?"). The electronic device 1000 processes natural-language speech inputs in a manner consistent with what is described above with respect to electronic devices 800 and 900. For example, the electronic device 1000 processes the natural-language speech input 1010 to provide one or more candidate text representations (e.g., a text representation "hey Siri, what should I listen to?") and one or more candidate intents (e.g., a user intent of "obtaining media recommendations").

[0272] The electronic device 1000 identifies a task based on the natural-language speech input 1010. In some examples, the electronic device 1000 identifies one or more candidate tasks based on the one or more candidate intents, as discussed above, and performs the highest ranking candidate task to obtain one or more results. In some examples, the one or more results include information related to: a song, an audio book, a podcast, a station, a playlist, or a combination thereof. In the depicted example in FIG. 10A, the electronic device 1000 identifies a candidate

task of "providing a media item" with a parameter "for me" and performs the task to obtain a first media item ("The Altar" by Banks).

[0273] In response to receiving the speech input 1010, the electronic device 1000 provides an audio output 1012 indicative of a suggestion of the first media item ("If you are feeling Alternative, I've got 'The Altar' by Banks"). In some examples, the suggestion of the first media item includes additional information to contextualize the media recommendation, such as metadata of the media item (e.g., genre, artist) and reason(s) as to why the media item is recommended (e.g., "If you are feeling Alternative . . . "). In some examples, the electronic device 1000 simultaneously provides a playback of a portion of the recommended media item.

[0274] In some examples, after providing the audio output 1012, the electronic device 1000 receives a speech input 1014 ("Nah"). The electronic device determines whether the speech input 1014 is indicative of a non-affirmative response corresponding to the request for media (e.g., "no", "next", "don't like it", "hate it", "a couple more" and natural-language equivalents of any phrases indicative of a rejection). In accordance with a determination that the speech input 1014 is indicative of a non-affirmative response, the electronic device updates the number of consecutive non-affirmative responses corresponding to the request. On the other hand, in accordance with a determination that the speech input 1014 is not indicative of a non-affirmative response, the electronic device foregoes updating the number. In the depicted example, the electronic device updates the number from 0 to 1 based on the speech input 1014.

[0275] With reference to FIG. 10A, in some examples, the electronic device 1000 provides another audio output 1016 indicative of a suggestion of the second media item ("How about the playlist, When Hip-Hop Goes Left?"). In some examples, the electronic device 900 identifies the first media item and the second media item as part of a single search, and the first media item has a higher confidence score than the second media item and is thus suggested first by the electronic device. In some examples, the electronic device performs two separate searches to identify the first media item and the second media item, respectively, and the second search is performed after the user provides a non-affirmative response to the suggestion of the first media item (e.g., audio input 1014).

[0276] After providing a suggestion of the second media item, the electronic device 1000 receives a speech input 1018 ("Next"). The electronic device determines whether the speech input 1018 is indicative of a non-affirmative response corresponding to the request for media in a manner consistent with what is described with respect to the speech input 1014. In the depicted example, the electronic device determines that the speech input 1018 is indicative of a non-affirmative response and updates the number from 1 to 2.

[0277] With reference to FIG. 10A, in some examples, the electronic device 1000 provides yet another audio output 1020 indicative of a suggestion of the third media item ("I've also got the playlist, 'If You Like Alabama Shakes.'"). The electronic device 900 can identify the first media item, the second media item, and the third media item in a single search or in different searches (e.g., using different search parameters and/or context information).

[0278] After providing a suggestion of the third media item, the electronic device 1000 samples (e.g., via a microphone) for inputs from the user. In some examples, the electronic device determines that a response is not received within a predetermined period of time (e.g., 5 seconds). In accordance with the determination, the electronic device updates the number of consecutive non-affirmative responses corresponding to the request. In the depicted example, the electronic device determines that no response is received within a predetermined period of time (e.g., silence) and updates the number from 2 to 3.

[0279] The electronic device 1000 determines whether the number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold. In some examples, the electronic device makes the determination after providing each of the audio outputs 1012, 1016, and 1020. In accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold, the electronic device provides an audio output indicative of a suggestion of another media item. For example, after receiving audio output 1018 ("Next"), the electronic device determines that the number of consecutive non-affirmative responses (2) is not equal to a predetermined threshold (e.g., 3). Accordingly, the electronic device provides speech output 1020 to suggest another media item different from what has been suggested.

[0280] In accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold, the electronic device foregoes providing a suggestion of another media item and instead provides an audio output indicative of a request for user input. For example, after receiving a non-affirmative response 1022 (e.g., silence for a predetermined amount of time), the electronic device determines that the number of consecutive non-affirmative responses (3) is equal to the predetermined threshold (e.g., 3). Accordingly, with reference to FIG. 10B, the electronic device provides speech output 1024 ("Ok. Can you name an artist you've been enjoying lately?").

[0281] In some examples, the speech output 1024 is indicative of a prompt for one or more parameters for the request for media. In the depicted example, the electronic device 1000 prompts the user for an artist parameter ("Can you name an artist you've been enjoying lately?") and receives a speech input 1026 ("Um... Flume"). The speech input 1026 is indicative of a parameter for the request for media (artist = Flume). Based on the received parameter, the electronic device 1000 identifies another media item different from the previously recommended media items. Accordingly, the electronic device 1000 provides the identified media item via speech output 1028 ("Great, here's the playlist, 'If You Like Flume.'"). In some examples, based on the received parameter, the electronic device updates the user preference data (e.g., user data and models 231) and/or the user-specific corpus accordingly.

[0282] In some examples, alternatively or additionally to providing the speech output 1024, the electronic device 1000 provides a speech output indicative of a prompt for a user selection among a plurality of media items previously suggested (e.g., "Do any of these sound good?"). In some examples, the electronic device receives a speech input indicative of a user selection (e.g., "Yeah, the second one", "the hip hop one", "the one by Adele") and interprets the speech input based on context information. The context information can include the plurality of media items previously suggested by the electronic device.

[0283] FIG. 11 shows electronic device 1100. Electronic device 1100 may be any of devices 200, 400, 600, 800, 900, and 1000 (FIGS. 2, 4, 6A-B, 8A-B, 9A-B, and 10) in some embodiments. In the illustrated example, the electronic device 1100 is an electronic device with one or more speakers, though it will be appreciated that the electronic device may be a device of

any type, such as a phone, laptop computer, desktop computer, tablet, wearable device (e.g., smart watch), set-top box, television, speaker, or any combination or subcombination thereof.

[0284] In operation, the electronic device 1100 receives (e.g., via a microphone) a natural-language speech input 1110 indicative of a request for media to the digital assistant of the electronic device 1100 ("Hey Siri, play something"), uttered by user 1102. In the depicted example, the electronic device 1100 is associated with the user 1102. The electronic device 1100 processes natural-language speech inputs in a manner consistent with what is described above with respect to electronic devices 800, 900, and 1000. For example, the electronic device 1100 processes the natural-language speech input 1110 to provide one or more candidate text representations (e.g., a text representation "hey Siri, play something") and one or more candidate intents (e.g., a user intent of "obtaining media recommendations").

[0285] The electronic device 1100 detects the physical presence of a plurality of users in proximity to the electronic device. In some examples, the electronic device 1100 can detect the physical presence of a person based on information related to an electronic device associated with the person, such as connectivity status of the person's electronic device (e.g., on the same Wi-Fi network, within Bluetooth range, within NFC range), information on the person's electronic device, etc. For instance, if the user's sister is also in physical proximity with the electronic device 1100 and has her phone on her, the electronic device 1100 can receive information corresponding to the sister's phone. For example, the electronic device 1100 can receive identification information (e.g., phone number, user name) from the sister's device (e.g., via a Bluetooth connection). As another example, the electronic device 1100 can receive identification information from a routing device (e.g., a wireless router that both electronic device 1100 and the sister's device are connected to).

[0286] In some examples, the electronic device 1100 can detect the physical presence of a person based on facial characteristics and/or voice characteristics of the person (captured via, for instance, cameras and microphones). In other examples, the electronic device 800 can detect the physical presence of a person based on information available locally, such as contacts listed in a calendar invite or an email message. In still some other examples, the electronic device 800 can detect the physical presence of a person based on credentials provided by the person (e.g., user

name and password). In some examples, the electronic device 1100 prompts for disambiguation input (e.g., "is that Jason or John that I'm hearing?") and/or confirmation input (e.g., "Did John just join the party?") after detecting the physical presence of a person.

[0287] In response to detecting the physical presence of the plurality of users (e.g., family members, visitors), the electronic device 1100 obtains a plurality of preference profiles corresponding to the plurality of users. In some examples, the electronic device 1100 receives a preference profile corresponding to a person other than the user 1102 (e.g., the user's sister) from a remote device (e.g., a server device). In some examples, the electronic device 1100 receives a preference profile corresponding to a person other than the user 1102 (e.g., the user 1102's sister) directly from the person's electronic device (e.g., the sister's phone). In some examples, the electronic device 1100 stores the preference profile of a person other than the user 1102 locally. For instance, user 1102 may have previously asked the digital assistant to store the preference locally (e.g., "Hey Siri, remember that my sister likes The Beatles").

[0288] Based on the plurality of preference profiles, the electronic device 1100 provides a merged preference profile. In some examples, providing a merged preference profile comprises identifying one or more preferences shared by each of the plurality of preference profiles. In the depicted example, the electronic device 1100 provides a merged preference profile based on the user's preference profile and the sister's preference profile. Because both the user and the sister have a preference for The Beatles, the merged preference profile includes a preference for The Beatles. On the other hand, because only the user, but not the sister, has a preference for Banks, the merged preference profile may not include a preference for Banks, in some examples.

[0289] Based on the merged preference profile, the electronic device 1100 identifies a media item. The identified media item can be a song, an audio book, a podcast, a station, a playlist, or any combination thereof. For example, the electronic device 1100 identifies a song "Hey Jude" because the metadata of the song (e.g., artist) matches one or more preferences of the merged profile (e.g., The Beatles). Accordingly, the electronic device 1100 provides an audio output 1112 ("Here's something you both may like, from the Beatles"). The audio output 1112 includes a description of the identified media item ("from The Beatles") and makes a reference to the

merged profile (e.g., "something you both may like"). The electronic device 1100 also provides an audio output 1113 including the identified media (a playback of the song "Hey Jude").

[0290] In some examples, identifying the media item based on the merged preference profile includes identifying the media item from a plurality of media items. In some examples, the plurality of media items includes a first set of media items associated with the first user (e.g., that the first user is authorized to access) and a second set of media items associated with the second user (e.g., that the second user is authorized to access). In the depicted example, the identified media item is not part of the first set of media items but is part of the second set of media items (i.e., the user does not have access to the song "Hey Jude", but the user's sister does).

[0291] In some examples, after detecting physical presence of multiple users including a second user (e.g., the sister), the electronic device 1100 detects a lack of presence of the second user. The electronic device 1100 can detect the lack of presence of the second user using techniques similar to those for detecting the presence of the second user. For example, the electronic device 1100 can detect the lack of presence by obtaining information (e.g., whether the sister's device is still connected to the wireless network) related to the electronic device of the second user. The information can be obtained from the second user's device directly or from a network router. After detecting the lack of presence of the second user, the electronic device 1100 updates the merged preference profile and/or the plurality of media items to search from. For example, if the electronic device 1100 detects a lack of presence of the user's sister, the electronic device 1100 removes media items that only the sister has access to (e.g., the sister's The Beatles collection) from the plurality of media items to search from.

[0292] In some examples, the electronic device 1100 receives a natural-language speech input 1114 indicative of a request for media based on preferences and/or activities of a person other than the user. In the depicted example in FIG. 11, the user 1102 provides an audio output 1114 ("What are my friends listening to?"). In response, the electronic device 1100 identifies one or more people (e.g., via contact list, software services such as social media services and media services, and other user-specific data). The electronic device furthermore obtains information related to the preferences (e.g., preferred genre, preferred artist) and/or activities (recently played songs) of the one or more people from one or more sources (e.g., software

services such as social media services and media services). For example, the electronic device 1100 can identify one or more people that are associated with the user on a software service and identifies a media item that some or all of these people have played using the software service. Alternatively, the electronic device 1100 identifies the media item by searching for media items with the appropriate metadata (e.g., a friend tag) in a database (e.g., the user-specific corpus discussed above). In the depicted example, the electronic device 1100 provides an audio output 1116 ("Here's Hello by Adele") to provide the identified media item.

[0293] In some examples, the electronic device receives a natural-language speech input 1118 indicative of a request for information ("Who's listening to this?"). In response, the electronic device 1100 provides identification information of the one or more people associated with the media item (e.g., using the user-specific corpus, using related software services). The identification information can be obtained locally and/or from one or more remote devices. In the depicted example, the electronic device 1100 provides audio output 1120 ("Your friends John and Jane") to provide the identification information.

4. Processes for Providing an Auditory-based Interface of a Digital Assistant for Media Exploration

[0294] FIG. 12 illustrates process 1200 for providing an auditory-based interface of a digital assistant, according to various examples. Process 1200 is performed, for example, using one or more electronic devices implementing a digital assistant. In some examples, process 1200 is performed using a client-server system (e.g., system 100), and the blocks of process 1200 are divided up in any manner between the server (e.g., DA server 106) and a client device. In other examples, the blocks of process 1200 are divided up between the server and multiple client devices (e.g., a mobile phone and a smart watch). Thus, while portions of process 1200 are described herein as being performed by particular devices of a client-server system, it will be appreciated that process 1200 is not so limited. In other examples, process 1200 is performed using only a client device (e.g., user device 104) or only multiple client devices. In process 1200, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 1200.

[0295] At block 1202, the electronic device receives a first natural-language speech input indicative of a request for media. The first natural-language speech input comprises a first search parameter. In some examples, the electronic device obtains, based on the first natural-language speech input, a text string. Further, the electronic device determines, based on the text string, a representation of user intent of obtaining recommendations for media items. Further, the electronic device determines, based on the representation of user intent, a task and one or more parameters for performing the task, which include the first search parameter.

[0296] At block 1204, the electronic device (or the digital assistant of the electronic device) provides a first media item. The first media item is identified based on the first search parameter. In some examples, the first media item is a song, an audio book, a podcast, a station, a playlist, or any combination thereof.

[0297] In some examples, providing the first media item comprises: providing, by the digital assistant, a speech output indicative of a verbal response associated with the first media item. Providing the first media item further comprises: while providing the speech output indicative of the verbal response, providing, by the digital assistant, playback of a portion of the first media item. In some other examples, providing the first media item comprises providing, by the digital assistant, playback of the first media item. In some other examples, providing the first media item comprises providing, by the digital assistant, a plurality of media items, which includes the first media item.

[0298] At block 1206, while providing the first media item, the electronic device receives a second natural-language speech input. In some examples, in response to receiving the second natural-language speech input, the electronic device adjusts the manner in which the first media item is provided.

[0299] At block 1208, the electronic device determines whether the second natural-language speech input corresponds to a user intent of refining the request for media. In some examples, determining whether the second natural-language speech input corresponds to a user intent of refining the request for media comprises deriving a representation of user intent of refining the request for media based on one or more predefined phrases and natural-language equivalents of the one or more phrases. In some examples, determining whether the second natural-language

speech input corresponds to a user intent of refining the request for media comprises deriving a representation of user intent of refining the request for media based on context information.

[0300] In some examples, the electronic device obtains based on the second natural-language speech input, one or more parameters for refining the request for media. In some examples, a parameter of the one or more parameters corresponds to: lyrical content of a media item, an occasion or a time period, an activity, a location, a mood, a release date within a predetermined time frame, an intended audience, a collection of media items, or any combination thereof. In some examples, the second natural-language speech input is associated with a first user, and a parameter of the one or more parameters corresponds to a second user different from the first user.

[0301] In some examples, obtaining the one or more parameters for refining the request for media comprises determining the one or more parameters based on context information. In some examples, the context information comprises information related to the first media item.

[0302] In some examples, the electronic device detects physical presence of one or more users, and the context information comprises information related to the one or more users. In some examples, the context information comprises a setting associated with one or more users of the electronic device.

[0303] At block 1210, in accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media, the electronic device (or the digital assistant) identifies, based on the first parameter and the second natural-language speech input, a second media item different from the first media item, and provides the second media item. The second media item can be a song, an audio book, a podcast, a station, a playlist, or any combination thereof.

[0304] In some examples, the electronic device obtains, based on the first natural-language speech input, a first set of media items and selects the first media item from the first set of media items. Further, the electronic device obtains, based on the second natural-language speech input, a second set of media items, which is a subset of the first set of media items, and selects the second media item from the set second set of media items. In some examples, obtaining the

second set of media items comprises selecting, from the first set of media items, one or more media items based on the one or more parameters for refining the request for media.

[0305] In some examples, identifying the second media item comprises determining whether content associated with the second media item matches at least one of the one or more parameters. In some other examples, identifying the second media item comprises determining whether metadata associated with the second media item matches at least one of the one or more parameters.

[0306] In some examples, the electronic device obtains the second media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with a user. In some examples, the electronic device identifies the user-specific corpus of media items based on acoustic information associated with the second natural-language speech input. In some examples, a media item in the user-specific corpus of media items includes metadata indicative of: an activity; a mood; an occasion; a location; a time; a curator; a playlist; one or more previous user inputs; or any combination thereof. In some examples, at least a portion of the metadata is based on information from a second user different from the first user.

[0307] In some examples, providing the second media item comprises providing, by the digital assistant, a speech output indicative of a verbal response associated with the second media item. Further, providing the second media item comprises, while providing the speech output indicative of the verbal response, providing, by the digital assistant, playback of a portion of the second media item.

[0308] In some examples, providing the second media item comprises providing, by the digital assistant, playback of the second media item. In some other examples, providing the second media item comprises providing, by the digital assistant, a plurality of media items, which includes the second media item.

[0309] In some examples, the electronic device receives a third natural-language speech input and determines, based on the third natural-language speech input, a representation of user intent of associating the second media item with a collection of media items. Further, the

electronic device associates the second media item with the collection of media items and provides, by the digital assistant, an audio output indicative of the association.

[0310] In some examples, while providing the second media item, the electronic device receives a fourth natural-language speech input. Further, the electronic device determines, based on the fourth natural-language speech input, a representation of user intent of obtaining information related to a particular media item, and provides, by the digital assistant, the information related to the particular media item. In some examples, the electronic device selects the particular media item based on context information.

[0311] In some examples, while providing the second media item, the electronic device (or the digital assistant of the electronic device) provides a speech output indicative of a third media item and, after providing the second media item, provides the third media item.

[0312] In some examples, the electronic device is a computer, a set-top box, a speaker, a smart watch, a phone, or a combination thereof.

[0313] The operations described above with reference to FIG. 12 are optionally implemented by components depicted in FIGS. 1-4, 6A-B, and 7A-C. For example, the operations of process 1200 may be implemented by any device, or component thereof, described herein, including but not limited to, devices 104, 200, 400, 600, 800, 900, 1000, and 1100. It would be clear to a person having ordinary skill in the art how other processes are implemented based on the components depicted in FIGS. 1-4, 6A-B, and 7A-C.

[0314] FIG. 13 illustrates process 1300 for providing an auditory-based interface of a digital assistant, according to various examples. Process 1300 is performed, for example, using one or more electronic devices implementing a digital assistant. In some examples, process 1300 is performed using a client-server system (e.g., system 100), and the blocks of process 1300 are divided up in any manner between the server (e.g., DA server 106) and a client device. In other examples, the blocks of process 1300 are divided up between the server and multiple client devices (e.g., a mobile phone and a smart watch). Thus, while portions of process 1300 are described herein as being performed by particular devices of a client-server system, it will be appreciated that process 1300 is not so limited. In other examples, process 1300 is performed

using only a client device (e.g., user device 104) or only multiple client devices. In process 1300, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 1300.

[0315] At block 1302, the electronic device receives a natural-language speech input. In some examples, the natural-language speech input is indicative of a request for one or more media items.

[0316] At block 1304, the electronic device (or the digital assistant of the electronic device) identifies a task based on the natural-language speech input. In some examples, identifying the task comprises: obtaining a text string based on the natural-language speech input; interpreting the text string to obtain a representation of user intent; and determining the task based on the representation of user intent.

[0317] In some examples, identifying the task based on the natural-language speech input comprises identifying a task of providing one or more media items. In some examples, the electronic device identifies a media item (hereinafter "the second media item") based on the speech input and obtains information corresponding to the media item (e.g., by performing the identified task). In some examples, the second media item comprises: a song, an audio book, a podcast, a station, a playlist, or a combination thereof. In some other examples, the electronic device performs the task to obtain one or more results (e.g., search results).

[0318] At block 1306, the electronic device (or the digital assistant of the electronic device) provides a speech output indicative of a verbal response associated with the identified task. In some examples, providing the speech output comprises providing a verbal description of the second media item. In some examples, the speech output is provided in a voice of the digital assistant, a voice associated with the second media item, or a combination thereof. In some examples, providing a speech output indicative of a verbal response associated with the identified task comprises providing a speech output indicative of a verbal description of a result of the one or more results (e.g., search results).

[0319] At block 1308, while providing the speech output indicative of a verbal response, the electronic device (or the digital assistant of the electronic device) provides playback of a media item (hereinafter "the first media item") corresponding to the verbal response. In some examples, the played back media item corresponds to a portion of the second media item. For example, the played back media item is a representative sample of the second media item.

[0320] In some examples, while providing playback of the first media item, the electronic device receives a second natural-language speech input. In response to receiving the second natural-language speech input, the electronic device provides the playback of the second media item. In some examples, the playback of the second media item is provided at a different volume from the playback of the first media item.

[0321] In some examples, the speech output indicative of a verbal response associated with the identified task is a first speech output. While providing the playback of the second media item, the electronic device provides a second speech output.

[0322] In some examples, providing playback of a media item corresponding to the verbal response comprises providing playback of a sound effect corresponding to the result. In some examples, the speech output indicative of a verbal response associated with the identified task is provided at a first volume, and the playback of a media item (e.g., sound effect) is provided at a second volume different from the first volume.

[0323] In some examples, the electronic device is a computer, a set-top box, a speaker, a smart watch, a phone, or a combination thereof.

[0324] The operations described above with reference to FIG. 13 are optionally implemented by components depicted in FIGS. 1-4, 6A-B, and 7A-C. For example, the operations of process 1300 may be implemented by any device, or component thereof, described herein, including but not limited to, devices 104, 200, 400, 600, 800, 900, 1000, and 1100. It would be clear to a person having ordinary skill in the art how other processes are implemented based on the components depicted in FIGS. 1-4, 6A-B, and 7A-C.

[0325] FIG. 14 illustrates process 1400 for providing an auditory-based interface of a digital assistant, according to various examples. Process 1400 is performed, for example, using one or

more electronic devices implementing a digital assistant. In some examples, process 1400 is performed using a client-server system (e.g., system 100), and the blocks of process 1400 are divided up in any manner between the server (e.g., DA server 106) and a client device. In other examples, the blocks of process 1400 are divided up between the server and multiple client devices (e.g., a mobile phone and a smart watch). Thus, while portions of process 1400 are described herein as being performed by particular devices of a client-server system, it will be appreciated that process 1400 is not so limited. In other examples, process 1400 is performed using only a client device (e.g., user device 104) or only multiple client devices. In process 1400, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 1400.

[0326] At block 1402, the electronic device receives a speech input indicative of a request for media. In some examples, the electronic device obtains, based on the speech input indicative of the request for media, a text string. In some examples, the electronic device further determines, based on the obtained text string, a representation of user intent and obtains, based on the representation of user intent, information related to one or more media items.

[0327] At block 1404, in response to receiving the speech input, the electronic device (or the digital assistant of the electronic device) provides an audio output indicative of a suggestion of a first media item. In some examples, the first media item is part of the one or more media items. In some examples, the first media item is a song, an audio book, a podcast, a station, a playlist, or any combination thereof.

[0328] At block 1406, the electronic device (or the digital assistant of the electronic device) determines whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold.

[0329] In some examples, the speech input indicative of the request for media is a first speech input. Further, determining whether the number of consecutive non-affirmative responses corresponding to the request for media satisfies the threshold comprises: after providing the audio output indicative of the suggestion of the first media item, receiving a second speech input; and determining whether the second speech input is indicative of a non-affirmative

response corresponding to the request for media. Further, in accordance with a determination that the second speech input is indicative of a non-affirmative response, the electronic device updates the number of consecutive non-affirmative responses corresponding to the request. In accordance with a determination that the second speech input is not indicative of a non-affirmative response, the electronic device foregoes updating the number of consecutive non-affirmative responses corresponding to the request.

[0330] In some examples, determining whether the second speech input is indicative of a non-affirmative response to the request for media comprises: determining whether the second speech input is indicative of a rejection.

[0331] In some examples, determining whether the number of consecutive non-affirmative responses corresponding to the request for media satisfies the threshold comprises: after providing the audio output indicative of the suggestion of the first media item, determining that a response corresponding to the request is not received within a predefined period of time. Determining whether the number of consecutive non-affirmative responses corresponding to the request for media satisfies the threshold further comprises: updating the number of consecutive non-affirmative responses corresponding to the request.

[0332] At block 1408, in accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold, the electronic device (or the digital assistant of the electronic device) provides an audio output indicative of a suggestion of a second media item different from the first media item. In some examples, the second media item is part of the one or more media items. In some examples, the second media item is a song, an audio book, a podcast, a station, a playlist, or any combination thereof.

[0333] At block 1410, in accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold, the electronic device (or the digital assistant of the electronic device) foregoes providing an audio output indicative of a suggestion of a second media item and provides an audio output indicative of a request for user input.

[0334] In some examples, providing an audio output indicative of a request for user input comprises: providing, by the digital assistant, a speech output indicative of a prompt for a user

selection among a plurality of media items previously suggested by the digital assistant. In some examples, after providing the audio output indicative of a request for user input, the electronic device receives a speech input indicative of a user selection and interprets, based on context information, the speech input indicative of the user selection. In some examples, the context information includes the plurality of media items previously suggested by the digital assistant.

[0335] In some examples, providing an audio output indicative of a request for user input comprises: providing, by the digital assistant, a speech output indicative of a prompt for one or more parameters for the request for media. In some examples, after providing the audio output indicative of a request for user input, the electronic device receives a speech input indicative of one or more parameters for the request for media. In some examples, the electronic device obtains a third media item based on the one or more parameters. The third media item is different from the first media item and the second media item.

[0336] In some examples, the electronic device is a computer, a set-top box, a speaker, a smart watch, a phone, or a combination thereof.

[0337] The operations described above with reference to FIG. 14 are optionally implemented by components depicted in FIGS. 1-4, 6A-B, and 7A-C. For example, the operations of process 1400 may be implemented by any device, or component thereof, described herein, including but not limited to, devices 104, 200, 400, 600, 800, 900, 1000, and 1100. It would be clear to a person having ordinary skill in the art how other processes are implemented based on the components depicted in FIGS. 1-4, 6A-B, and 7A-C.

[0338] FIG. 15 illustrates process 1500 for providing an auditory-based interface of a digital assistant, according to various examples. Process 1500 is performed, for example, using one or more electronic devices implementing a digital assistant. In some examples, process 1500 is performed using a client-server system (e.g., system 100), and the blocks of process 1500 are divided up in any manner between the server (e.g., DA server 106) and a client device. In other examples, the blocks of process 1500 are divided up between the server and multiple client devices (e.g., a mobile phone and a smart watch). Thus, while portions of process 1500 are described herein as being performed by particular devices of a client-server system, it will be appreciated that process 1500 is not so limited. In other examples, process 1500 is performed

using only a client device (e.g., user device 104) or only multiple client devices. In process 1500, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 1500.

[0339] At block 1502, the electronic device receives a speech input indicative of a request for media. At block 1504, the electronic device (or the digital assistant of the electronic device) detects physical presence of a plurality of users to the electronic device. In some examples, the electronic device is a first electronic device associated with a first user of the plurality of users. Further, detecting physical presence of the plurality of users to the electronic device comprises: receiving information corresponding to a second electronic device associated with a second user of the plurality of users.

[0340] In some examples, receiving information corresponding to the second electronic device comprises: receiving, at the first electronic device, identification information from the second electronic device. In some examples, receiving information corresponding to the second electronic device comprises: receiving identification information from a routing device, which is connected to the first electronic device and the second electronic device.

[0341] At block 1506, in response to detecting the physical presence of the plurality of users, the electronic device obtains a plurality of preference profiles corresponding to the plurality of users. In some examples, the electronic device receives a preference profile corresponding to the second user from a remote device. In other examples, the preference profile corresponding to the second user is stored on the electronic device.

[0342] At block 1508, the electronic device (or the digital assistant of the electronic device) provides a merged preference profile based on the plurality of preference profiles. In some examples, providing the merged preference profile comprises: identifying one or more preferences shared by each of the plurality of preference profiles.

[0343] At block 1510, the electronic device (or the digital assistant of the electronic device) identifies a media item based on the merged preference profile. In some examples, the identified media item is associated with metadata matching the one or more preferences. The identified

media item can be a song, an audio book, a podcast, a station, a playlist, or any combination thereof.

[0344] In some examples, identifying a media item based on the merged preference profile comprises: identifying the media item from a plurality of media items. The plurality of media items comprises a first set of media items associated with the first user and a second set of media items associated with the second user. In some examples, the identified media item is not part of the first set of media items, but is part of the second set of media items.

[0345] At block 1512, the electronic device (or the digital assistant of the electronic device) provides an audio output including the identified media item. In some examples, the audio output includes a verbal description of the identified media item. In some examples, the audio output includes a speech output indicative of the merged profile.

[0346] In some examples, after detecting physical presence of the plurality of users, the electronic device (or the digital assistant of the electronic device) detects a lack of presence of the second user. After detecting the lack of presence of the second user, the electronic device updates the plurality of media items and updates the merged preference profile. In some examples, updating the plurality of media items comprises removing the media item from the plurality of media items.

[0347] In some examples, the electronic device is a computer, a set-top box, a speaker, a smart watch, a phone, or a combination thereof.

[0348] The operations described above with reference to FIG. 15 are optionally implemented by components depicted in FIGS. 1-4, 6A-B, and 7A-C. For example, the operations of process 1500 may be implemented by any device, or component thereof, described herein, including but not limited to, devices 104, 200, 400, 600, 800, 900, 1000, and 1100. It would be clear to a person having ordinary skill in the art how other processes are implemented based on the components depicted in FIGS. 1-4, 6A-B, and 7A-C.

[0349] In accordance with some implementations, a computer-readable storage medium (e.g., a non-transitory computer readable storage medium) is provided, the computer-readable storage medium storing one or more programs for execution by one or more processors of an

electronic device, the one or more programs including instructions for performing any of the methods or processes described herein.

[0350] In accordance with some implementations, an electronic device (e.g., a portable electronic device) is provided that comprises means for performing any of the methods or processes described herein.

[0351] In accordance with some implementations, an electronic device (e.g., a portable electronic device) is provided that comprises a processing unit configured to perform any of the methods or processes described herein.

[0352] In accordance with some implementations, an electronic device (e.g., a portable electronic device) is provided that comprises one or more processors and memory storing one or more programs for execution by the one or more processors, the one or more programs including instructions for performing any of the methods or processes described herein.

[0353] The foregoing description, for purpose of explanation, has been described with reference to specific embodiments. However, the illustrative discussions above are not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The embodiments were chosen and described in order to best explain the principles of the techniques and their practical applications. Others skilled in the art are thereby enabled to best utilize the techniques and various embodiments with various modifications as are suited to the particular use contemplated.

[0354] Although the disclosure and examples have been fully described with reference to the accompanying drawings, it is to be noted that various changes and modifications will become apparent to those skilled in the art. Such changes and modifications are to be understood as being included within the scope of the disclosure and examples as defined by the claims.

[0355] As described above, one aspect of the present technology is the gathering and use of data available from various sources to improve the delivery to users of invitational content or any other content that may be of interest to them. The present disclosure contemplates that in some instances, this gathered data may include personal information data that uniquely identifies or can be used to contact or locate a specific person. Such personal information data can include

demographic data, location-based data, telephone numbers, email addresses, home addresses, or any other identifying information.

[0356] The present disclosure recognizes that the use of such personal information data, in the present technology, can be used to the benefit of users. For example, the personal information data can be used to deliver targeted content that is of greater interest to the user. Accordingly, use of such personal information data enables calculated control of the delivered content. Further, other uses for personal information data that benefit the user are also contemplated by the present disclosure.

[0357] The present disclosure further contemplates that the entities responsible for the collection, analysis, disclosure, transfer, storage, or other use of such personal information data will comply with well-established privacy policies and/or privacy practices. In particular, such entities should implement and consistently use privacy policies and practices that are generally recognized as meeting or exceeding industry or governmental requirements for maintaining personal information data private and secure. For example, personal information from users should be collected for legitimate and reasonable uses of the entity and not shared or sold outside of those legitimate uses. Further, such collection should occur only after receiving the informed consent of the users. Additionally, such entities would take any needed steps for safeguarding and securing access to such personal information data and ensuring that others with access to the personal information data adhere to their privacy policies and procedures. Further, such entities can subject themselves to evaluation by third parties to certify their adherence to widely accepted privacy policies and practices.

[0358] Despite the foregoing, the present disclosure also contemplates embodiments in which users selectively block the use of, or access to, personal information data. That is, the present disclosure contemplates that hardware and/or software elements can be provided to prevent or block access to such personal information data. For example, in the case of advertisement delivery services, the present technology can be configured to allow users to select to "opt in" or "opt out" of participation in the collection of personal information data during registration for services. In another example, users can select not to provide location information

for targeted content delivery services. In yet another example, users can select to not provide precise location information, but permit the transfer of location zone information.

[0359] Therefore, although the present disclosure broadly covers use of personal information data to implement one or more various disclosed embodiments, the present disclosure also contemplates that the various embodiments can also be implemented without the need for accessing such personal information data. That is, the various embodiments of the present technology are not rendered inoperable due to the lack of all or a portion of such personal information data. For example, content can be selected and delivered to users by inferring preferences based on non-personal information data or a bare minimum amount of personal information, such as the content being requested by the device associated with a user, other non-personal information available to the content delivery services, or publically available information. The scope of the invention is defined by the following claims.

PATENTKRAV

1. Fremgangsmåde til betjening af en digital assistent, hvilken fremgangsmåde omfatter:

ved en elektronisk indretning med én eller flere processorer og hukommelse:

modtagelse (1202) af et første naturligt taleinput der indikerer en anmodning om medie, hvor det første naturlige taleinput omfatter en første søgeparameter;

tilvejebringelse (1204), med den digitale assistent, af afspilning af et første medieelement, hvor det første medieelement er identificeret baseret på den første søgeparameter;

modtagelse (1206), samtidig med tilvejebringelse af afspilning af det første medieelement, af et andet naturligt taleinput;

bestemmelse (1208) af om det andet naturlige taleinput svarer til en brugerhensigt om forfinelse af anmodningen om medie;

i overensstemmelse med en bestemmelse af at det andet naturlige taleinput svarer til en brugerhensigt om forfinelse af anmodningen om medie (1210):

identificering, baseret på den første søgeparameter og det andet naturlige taleinput, af et andet medieelement der er forskelligt fra det første medieelement; og

tilvejebringelse, med den digitale assistent, af det andet medieelement.

2. Fremgangsmåde ifølge krav 1, der endvidere omfatter:

tilvejebringelse, baseret på det første naturlige taleinput, af en tekststreng;

bestemmelse, baseret på tekststrengen, af en repræsentation af brugerhensigt om tilvejebringelse af anbefalinger af medieelementer; og

bestemmelse, baseret på repræsentationen af brugerhensigt, af en opgave og én eller flere parametre til udførelse af opgaven, hvor den ene eller flere parametre omfatter den første søgeparameter.

3. Fremgangsmåde ifølge krav 1-2, hvor tilvejebringelse af afspilning af det første medieelement omfatter:

tilvejebringelse, med den digitale assistent, af et taleoutput der indikerer et verbalt svar associeret med det første medieelement; og

tilvejebringelse, med den digitale assistent, samtidig med tilvejebringelse af taleoutputtet der indikerer det verbale svar, af afspilning af en del af det første medieelement.

4. Fremgangsmåde ifølge krav 1-2, hvor tilvejebringelse af afspilning af det første medieelement omfatter:

tilvejebringelse, med den digitale assistent, af en flerhed af medieelementer, hvor flerheden af medieelementer omfatter det første medieelement.

5. Fremgangsmåde ifølge et hvilket som helst af kravene 1-4, der endvidere omfatter:

justering, som reaktion på modtagelse af det andet naturlige taleinput, af måden hvorpå afspilning af det første medieelement tilvejebringes.

6. Fremgangsmåde ifølge et hvilket som helst af kravene 1-5, hvor bestemmelse af om det andet naturlige taleinput svarer til en brugerhensigt om forfinelse af anmodningen om medie omfatter: udledning af en repræsentation af brugerhensigt om forfinelse af anmodningen om medie baseret på én eller flere foruddefinerede fraser og naturlige sprogækvivalenter af den ene eller flere fraser.

7. Fremgangsmåde ifølge et hvilket som helst af kravene 1-6, der endvidere omfatter:

tilvejebringelse, baseret på det andet naturlige taleinput, af én eller flere parametre til forfinelse af anmodningen om medie.

8. Fremgangsmåde ifølge krav 7, hvor bestemmelse af om det andet naturlige taleinput svarer til en brugerhensigt om forfinelse af anmodningen om medie omfatter: udledning af en repræsentation af brugerhensigt om forfinelse af anmodningen om medie baseret på kontekstinformation.

9. Fremgangsmåde ifølge et hvilket som helst af kravene 7-8, hvor en parameter af den ene eller flere parametre svarer til lyrisk indhold for et medieelement.

10. Fremgangsmåde ifølge et hvilket som helst af kravene 7-9, hvor en parameter af den ene eller flere parametre svarer til en anledning eller en tidsperiode.

11. Fremgangsmåde ifølge et hvilket som helst af kravene 7-10, hvor en parameter af den ene eller flere parametre svarer til en aktivitet.

12. Fremgangsmåde ifølge et hvilket som helst af kravene 7-11, hvor en parameter af den ene eller flere parametre svarer til en lokation.

13. Fremgangsmåde ifølge et hvilket som helst af kravene 7-12, hvor en parameter af den ene eller flere parametre svarer til en stemning.

14. Fremgangsmåde ifølge et hvilket som helst af kravene 7-13, hvor en parameter af den ene eller flere parametre svarer til en udgivelsesdato inden for en forudbestemt tidsramme.

15. Fremgangsmåde ifølge et hvilket som helst af kravene 7-14, hvor en parameter af den ene eller flere parametre svarer til et tiltænkt publikum.

16. Fremgangsmåde ifølge et hvilket som helst af kravene 7-15, hvor en parameter af den ene eller flere parametre svarer til en samling af medieelementer.

17. Fremgangsmåde ifølge et hvilket som helst af kravene 7-16,
hvor det andet naturlige taleinput er associeret med en første bruger; og
hvor en parameter af den ene eller flere parametre svarer til en anden bruger der er forskellig fra den første bruger.

18. Fremgangsmåde ifølge et hvilket som helst af kravene 7-17, hvor tilvejebringelse af den ene eller flere parametre til forfinelse af anmodningen om medie omfatter: bestemmelse af den ene eller flere parametre baseret på kontekstinformation.

19. Fremgangsmåde ifølge krav 18, hvor kontekstinformationen omfatter information relateret til det første medieelement.

20. Fremgangsmåde ifølge krav 18, der endvidere omfatter:

detektering af fysisk tilstedeværelse af én eller flere brugere,

hvor kontekstinformationen omfatter information relateret til den ene eller flere brugere.

21. Fremgangsmåde ifølge krav 18, hvor kontekstinformationen omfatter en indstilling associeret med én eller flere brugere af den elektroniske indretning.

22. Fremgangsmåde ifølge et hvilket som helst af kravene 7-21, der endvidere omfatter:

tilvejebringelse, baseret på det første naturlige taleinput, af et første sæt af medieelementer;

valg af det første medieelement fra det første sæt af medieelementer;

tilvejebringelse, baseret på det andet naturlige taleinput, af et andet sæt af medieelementer, hvor det andet sæt af medieelementer er den delmængde af det første sæt af medieelementer; og

valg af det andet medieelement fra det andet sæt af medieelementer.

23. Fremgangsmåde ifølge krav 22, hvor tilvejebringelse af det andet sæt af medieelementer omfatter:

valg, fra det første sæt af medieelementer, af ét eller flere medieelementer baseret på den ene eller flere parametre til forfinelse af anmodningen om medie.

24. Fremgangsmåde ifølge et hvilket som helst af kravene 7-23, hvor identificering af det andet medieelement omfatter:

bestemmelse af om indhold associeret med det andet medieelement matcher mindst én af den ene eller flere parametre.

25. Fremgangsmåde ifølge et hvilket som helst af kravene 7-24, hvor identificering af det andet medieelement omfatter:

bestemmelse af om metadata associeret med det andet medieelement matcher mindst én af den ene eller flere parametre.

26. Fremgangsmåde ifølge et hvilket som helst af kravene 1-25, der endvidere omfatter:

tilvejebringelse af det andet medieelement fra et brugerspecifikt korpus af medieelementer, idet det brugerspecifikke korpus af medieelementer er genereret baseret på data associeret med en bruger.

27. Fremgangsmåde ifølge krav 26, der endvidere omfatter:

identificering af det brugerspecifikke korpus af medieelementer baseret på akustisk information associeret med det andet naturlige taleinput.

28. Fremgangsmåde ifølge et hvilket som helst af kravene 26-27, hvor et medieelement i det brugerspecifikke korpus af medieelementer omfatter metadata der indikerer: en aktivitet; en stemning; en anledning; en lokation; et tidspunkt; en kurator; en afspilningsliste; ét eller flere tidligere brugerinput; eller hvilken som helst kombination deraf.

29. Fremgangsmåde ifølge krav 28, hvor i det mindste en del af metadataene er baseret på information fra en anden bruger der er forskellig fra den første bruger.

30. Fremgangsmåde ifølge et hvilket som helst af kravene 1-29, der endvidere omfatter:

modtagelse af et tredje naturligt taleinput;

bestemmelse, baseret på det tredje naturlige taleinput, af en repræsentation af brugerhensigt om associering af det andet medieelement med en samling af medieelementer;

associering af det andet medieelement med samlingen af medieelementer; og

tilvejebringelse, med den digitale assistent, af et lydoutput der indikerer associeringen.

31. Fremgangsmåde ifølge et hvilket som helst af kravene 1-30, der endvidere omfatter:

modtagelse, samtidig med tilvejebringelse af det andet medieelement, af et fjerde naturligt taleinput;

bestemmelse, baseret på det fjerde naturlige taleinput, af en repræsentation af brugerhensigt om tilvejebringelse af information relateret til et bestemt medieelement;

tilvejebringelse, med den digitale assistent, af informationen relateret til det bestemte medieelement.

32. Fremgangsmåde ifølge krav 31, der endvidere omfatter: valg af det bestemte medieelement baseret på kontekstinformation.

33. Fremgangsmåde ifølge et hvilket som helst af kravene 1-32, der endvidere omfatter:

tilvejebringelse, med den digitale assistent, samtidig med tilvejebringelse af det andet medieelement, af et taleoutput der indikerer et tredje medieelement;

tilvejebringelse, efter tilvejebringelse af det andet medieelement, af afspilning af det tredje medieelement.

34. Fremgangsmåde ifølge et hvilket som helst af kravene 1-33, hvor tilvejebringelse af det andet medieelement omfatter:

tilvejebringelse, med den digitale assistent, af et taleoutput der indikerer et verbalt svar associeret med det andet medieelement; og

tilvejebringelse, med den digitale assistent, samtidig med tilvejebringelse af taleoutputtet der indikerer det verbale svar, af afspilning af en del af det andet medieelement.

35. Fremgangsmåde ifølge et hvilket som helst af kravene 1-33, hvor tilvejebringelse af det andet medieelement omfatter:

tilvejebringelse, med den digitale assistent, af afspilning af det andet medieelement.

36. Fremgangsmåde ifølge et hvilket som helst af kravene 1-33, hvor tilvejebringelse af det andet medieelement omfatter:

tilvejebringelse, med den digitale assistent, af et taleoutput der indikerer en anmodning om et brugervalg blandt en flerhed af medieelementer, hvor flerheden af medieelementer omfatter det andet medieelement.

37. Fremgangsmåde ifølge et hvilket som helst af kravene 1-36, hvor det første medieelement er en sang, en lydbog, en podcast, en station, en afspilningsliste, eller hvilken som helst kombination deraf.

38. Fremgangsmåde ifølge et hvilket som helst af kravene 1-36, hvor det andet medieelement er en sang, en lydbog, en podcast, en station, en afspilningsliste, eller hvilken som helst kombination deraf.

39. Fremgangsmåde ifølge et hvilket som helst af kravene 1-38, hvor den elektroniske indretning er en computer, en tv-boks, en højtaler, et smartwatch, en telefon, eller en kombination deraf.

40. Elektronisk indretning, der omfatter:

én eller flere processorer;

en hukommelse; og

ét eller flere programmer, hvor det ene eller flere programmer er lagret i hukommelsen og er konfigureret til at blive udført på den ene eller flere processorer, idet det ene eller flere programmer omfatter instruktioner til udførelse af fremgangsmåderne ifølge et hvilket som helst af kravene 1-39.

41. Ikke-flygtigt computerlæsbart lagermedium der lagrer ét eller flere programmer, idet det ene eller flere programmer omfatter instruktioner, som ved udførelse på én eller flere processorer i en elektronisk indretning får den elektroniske indretning til at udføre fremgangsmåderne ifølge et hvilken som helst af kravene 1-39.

DK 179559 B1

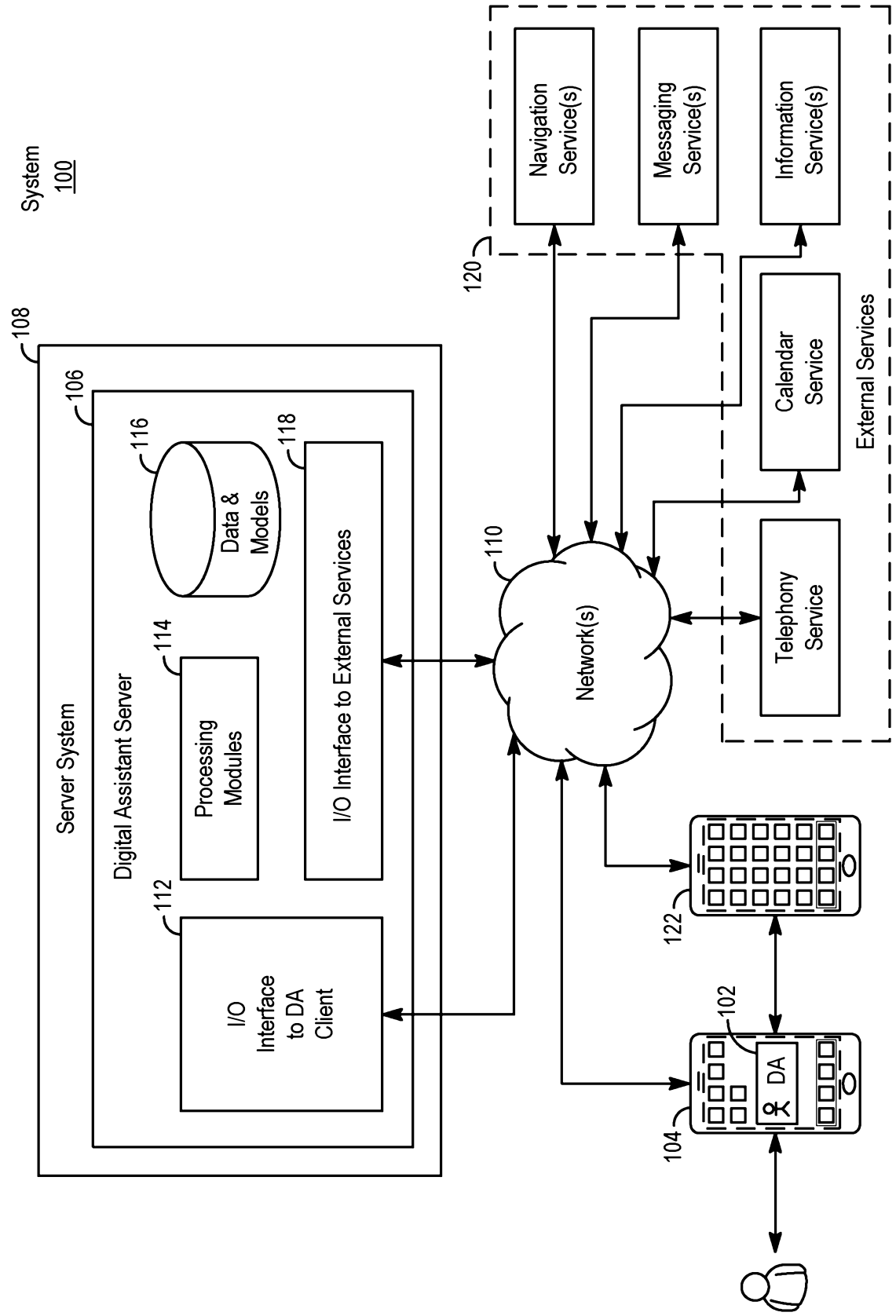


FIG. 1

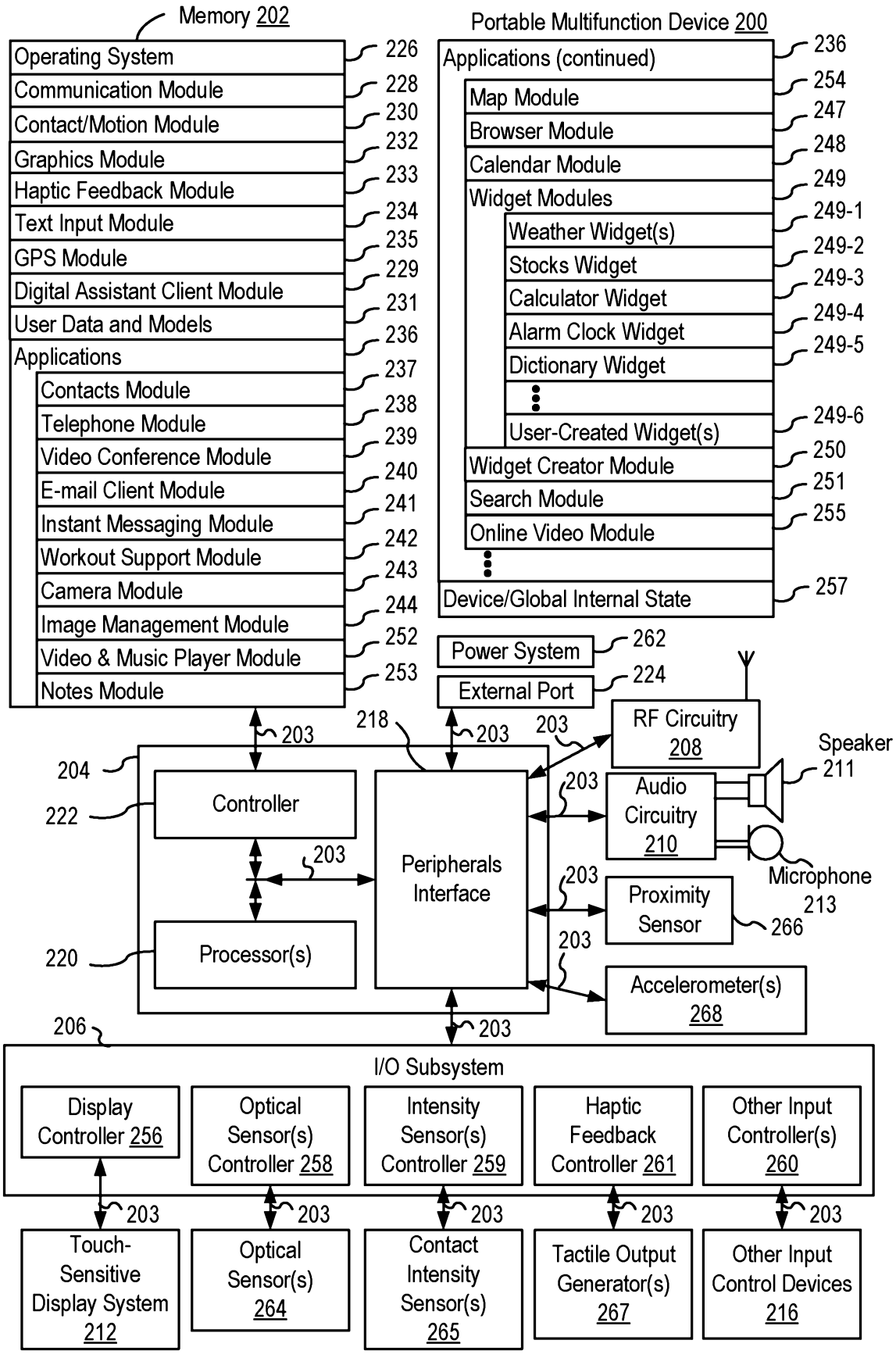

Memory 202
Operating System
Communication Module
Contact/Motion Module
Graphics Module
Haptic Feedback Module
Text Input Module
GPS Module
Digital Assistant Client Module
User Data and Models
Applications
Contacts Module
Telephone Module
Video Conference Module
E-mail Client Module
Instant Messaging Module
Workout Support Module
Camera Module
Image Management Module
Video & Music Player Module
Notes Module
226
228
230
232
233
234
235
229
231
236
237
238
239
240
241
242
243
244
252
253
Portable Multifunction Device 200
Applications (continued)
Map Module
Browser Module
Calendar Module
Widget Modules
Weather Widget(s)
Stocks Widget
Calculator Widget
Alarm Clock Widget
Dictionary Widget
User-Created Widget(s)
Widget Creator Module
Search Module
Online Video Module
Device/Global Internal State
236
254
247
248
249
249-1
249-2
249-3
249-4
249-5
249-6
250
251
255
257
Power System
External Port
262
224
RF Circuitry 208
Speaker 211
Audio Circuitry 210
Microphone 213
Proximity Sensor 266
Accelerometer(s) 268
Controller
Peripherals Interface
Processor(s)
204
222
220
218
203
206
I/O Subsystem
Display Controller 256
Optical Sensor(s) Controller 258
Intensity Sensor(s) Controller 259
Haptic Feedback Controller 261
Other Input Controller(s) 260
Touch-Sensitive Display System 212
Optical Sensor(s) 264
Contact Intensity Sensor(s) 265
Tactile Output Generator(s) 267
Other Input Control Devices 216

FIG. 2A

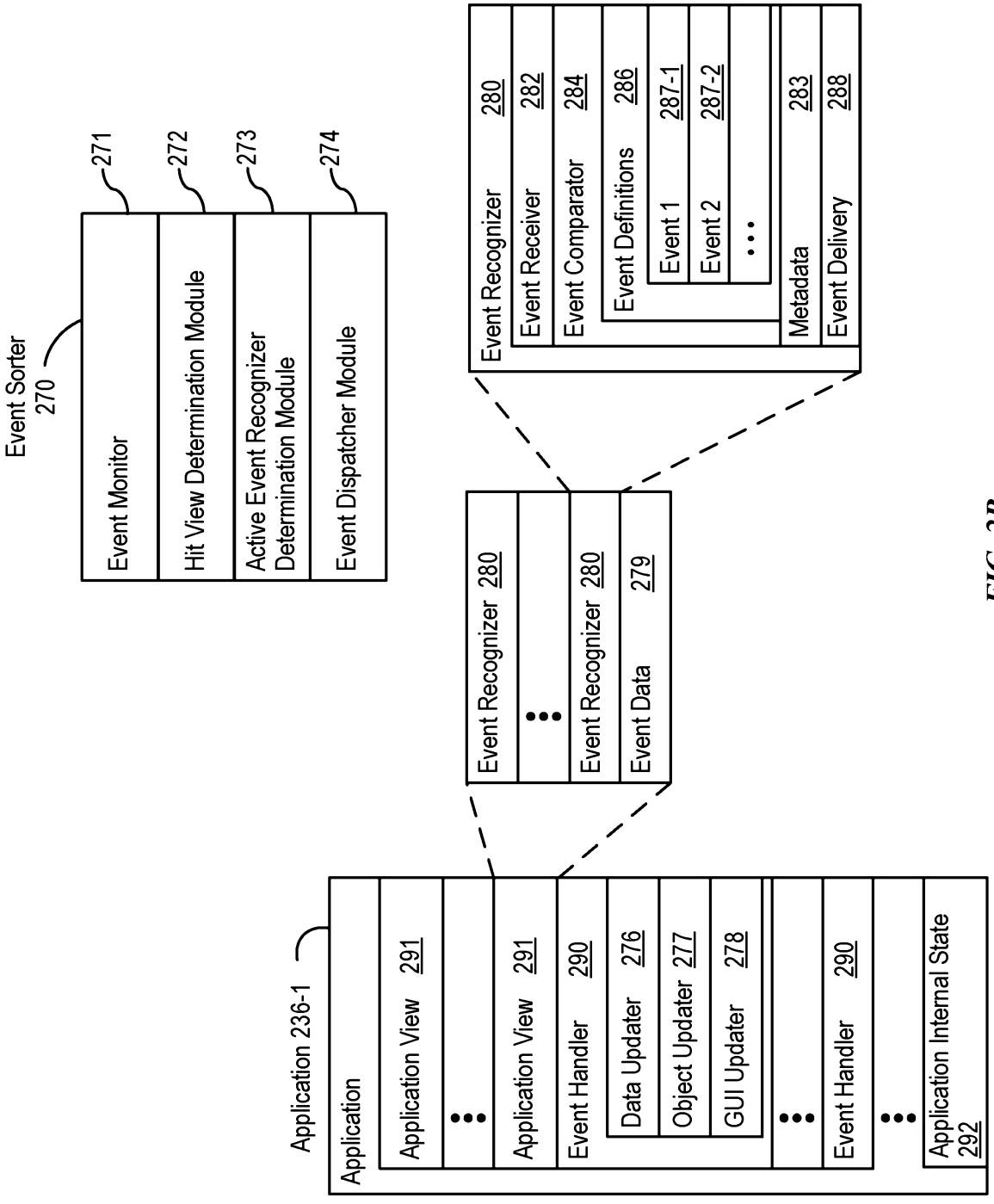

Event Sorter
270
271
Event Monitor
272
Hit View Determination Module
273
Active Event Recognizer Determination Module
274
Event Dispatcher Module
Application 236-1
Application
Application View 291
Application View 291
Event Handler 290
Data Updater 276
Object Updater 277
GUI Updater 278
Event Handler 290
Application Internal State 292
Event Recognizer 280
Event Recognizer 280
Event Data 279
Event Recognizer 280
Event Receiver 282
Event Comparator 284
Event Definitions 286
Event 1 287-1
Event 2 287-2
Metadata 283
Event Delivery 288

FIG. 2B

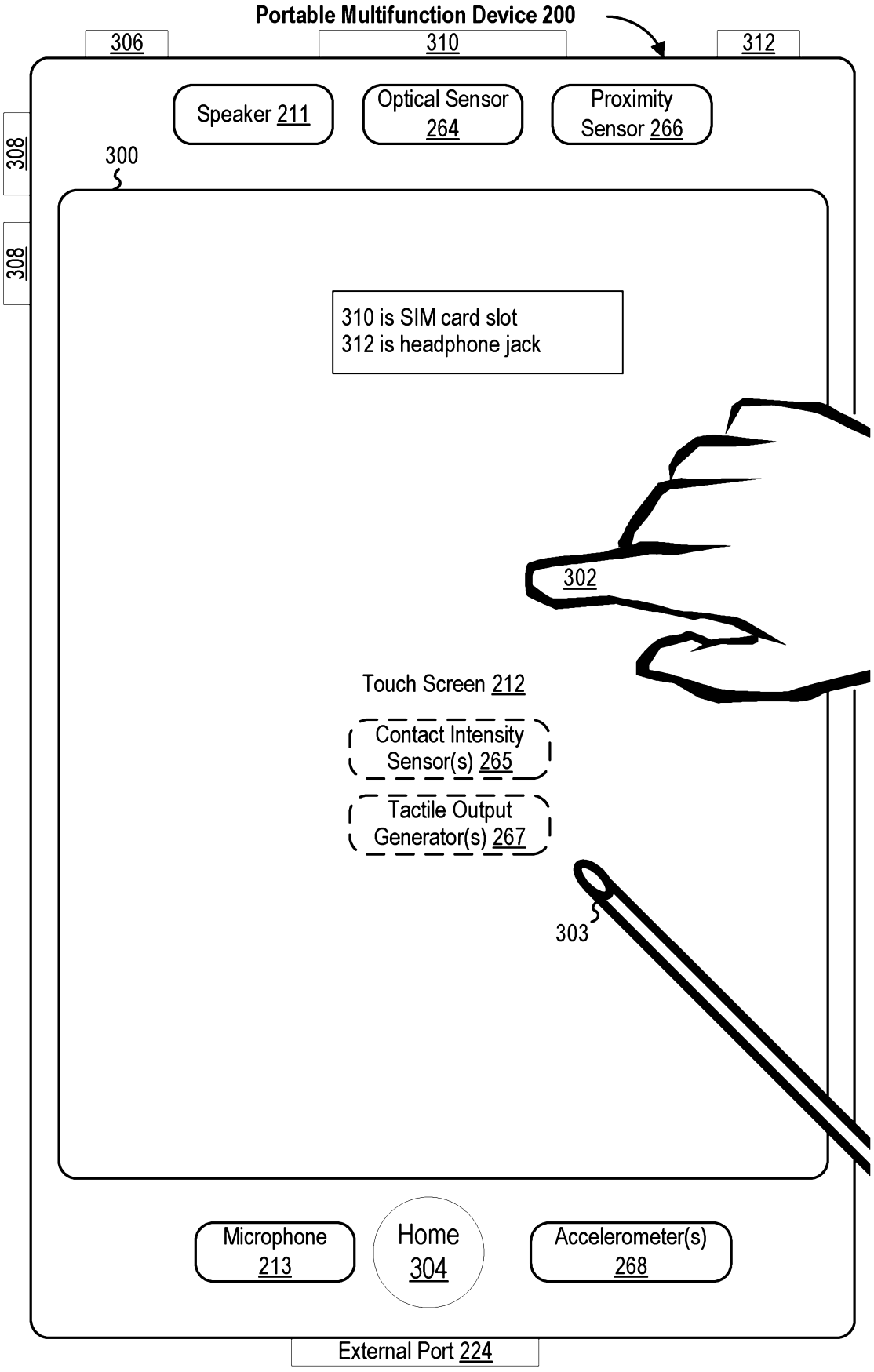

Portable Multifunction Device 200
306
310
312
Speaker 211
Optical Sensor 264
Proximity Sensor 266
308
300
308
310 is SIM card slot
312 is headphone jack
302
Touch Screen 212
Contact Intensity Sensor(s) 265
Tactile Output Generator(s) 267
303
Microphone 213
Home 304
Accelerometer(s) 268
External Port 224

FIG. 3

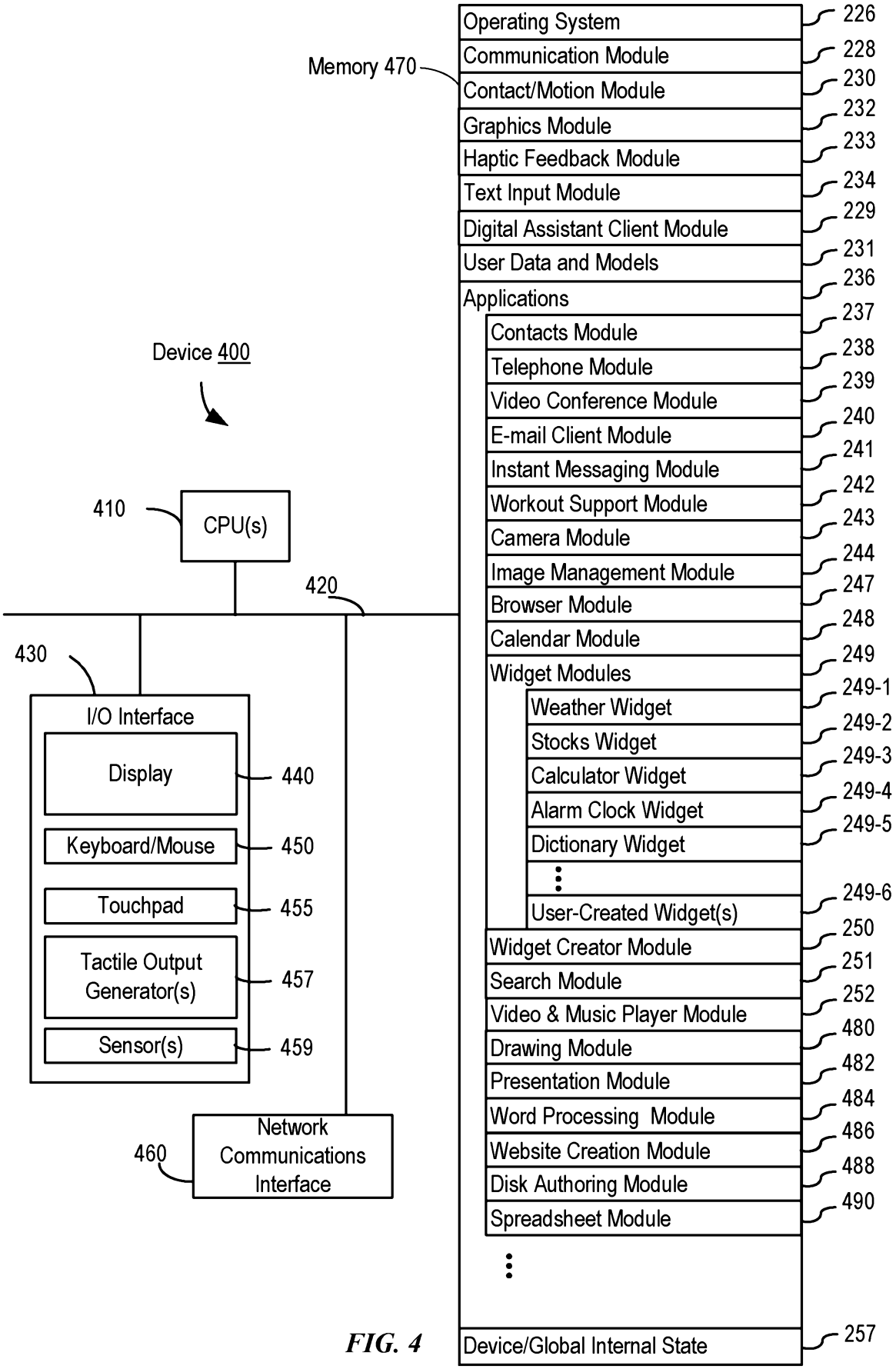

Memory 470
Operating System
226
Communication Module
228
Contact/Motion Module
230
Graphics Module
232
Haptic Feedback Module
233
Text Input Module
234
Digital Assistant Client Module
229
User Data and Models
231
Applications
236
Contacts Module
237
Telephone Module
238
Video Conference Module
239
E-mail Client Module
240
Instant Messaging Module
241
Workout Support Module
242
Camera Module
243
Image Management Module
244
Browser Module
247
Calendar Module
248
Widget Modules
249
Weather Widget
249-1
Stocks Widget
249-2
Calculator Widget
249-3
Alarm Clock Widget
249-4
Dictionary Widget
249-5
User-Created Widget(s)
249-6
Widget Creator Module
250
Search Module
251
Video & Music Player Module
252
Drawing Module
480
Presentation Module
482
Word Processing Module
484
Website Creation Module
486
Disk Authoring Module
488
Spreadsheet Module
490
Device/Global Internal State
257
Device 400
410
CPU(s)
420
430
I/O Interface
Display
440
Keyboard/Mouse
450
Touchpad
455
Tactile Output Generator(s)
457
Sensor(s)
459
460
Network Communications Interface

FIG. 4

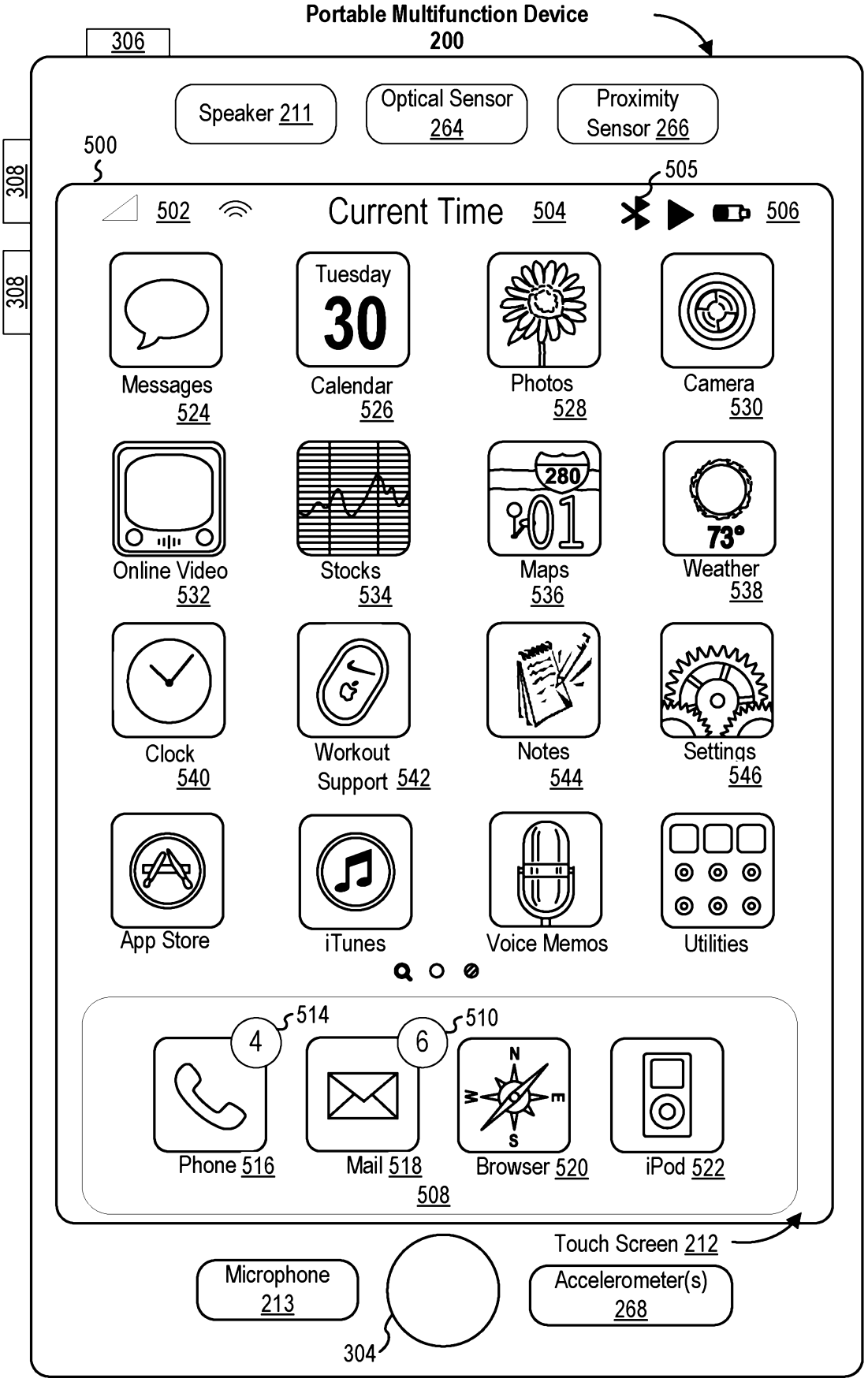
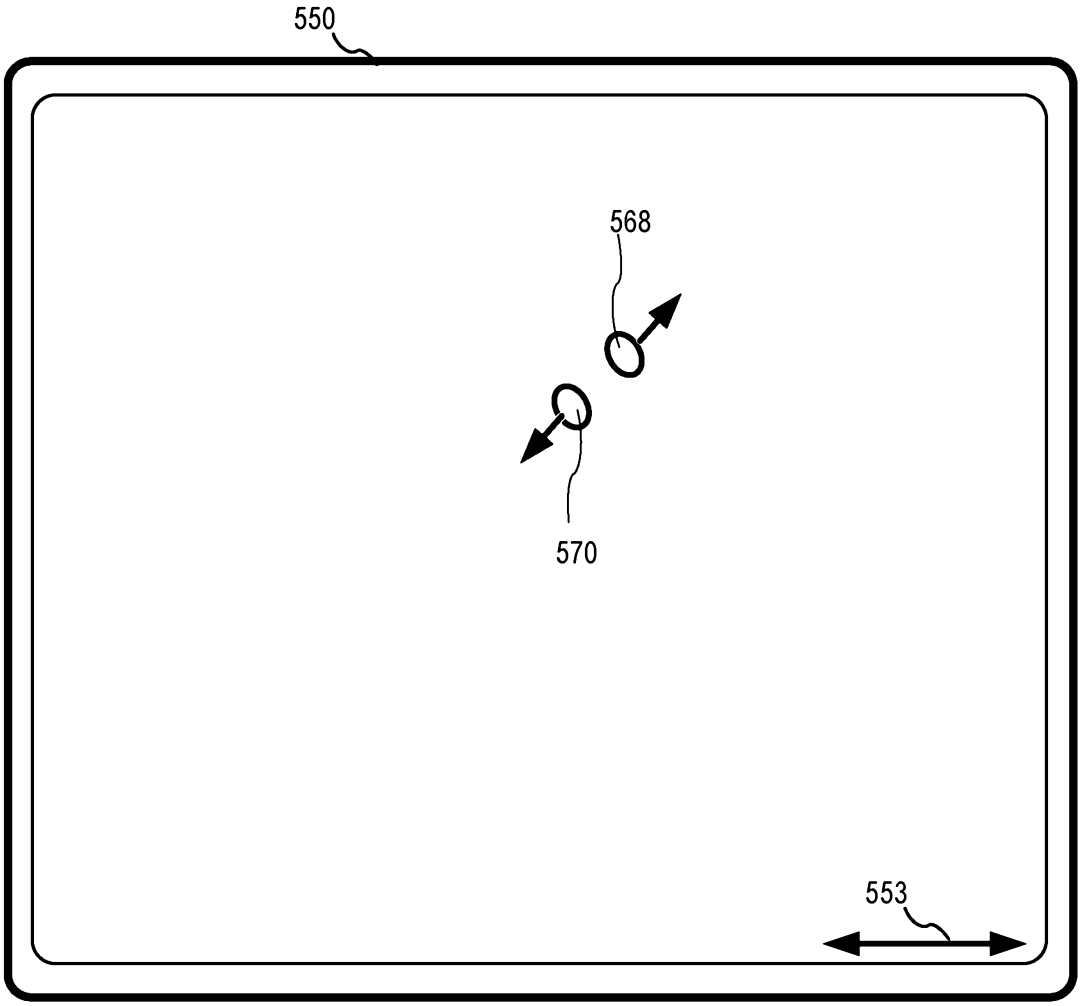

Portable Multifunction Device
200
306
Speaker 211
Optical Sensor 264
Proximity Sensor 266
500
308
308
505
502
Current Time 504
506
Messages 524
Tuesday
30
Calendar 526
Photos 528
Camera 530
Online Video 532
Stocks 534
280
Maps 536
73°
Weather 538
Clock 540
Workout Support 542
Notes 544
Settings 546
App Store
iTunes
Voice Memos
Utilities
514
510
4
6
N
W
E
S
Phone 516
Mail 518
Browser 520
iPod 522
508
Touch Screen 212
Microphone 213
Accelerometer(s) 268
304

FIG. 5A


550
568
570
553

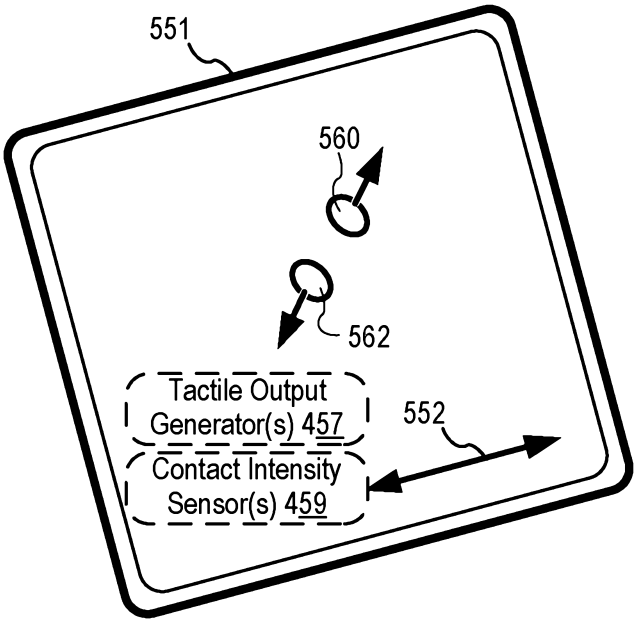

551
560
562
Tactile Output Generator(s) 457
Contact Intensity Sensor(s) 459
552

FIG. 5B

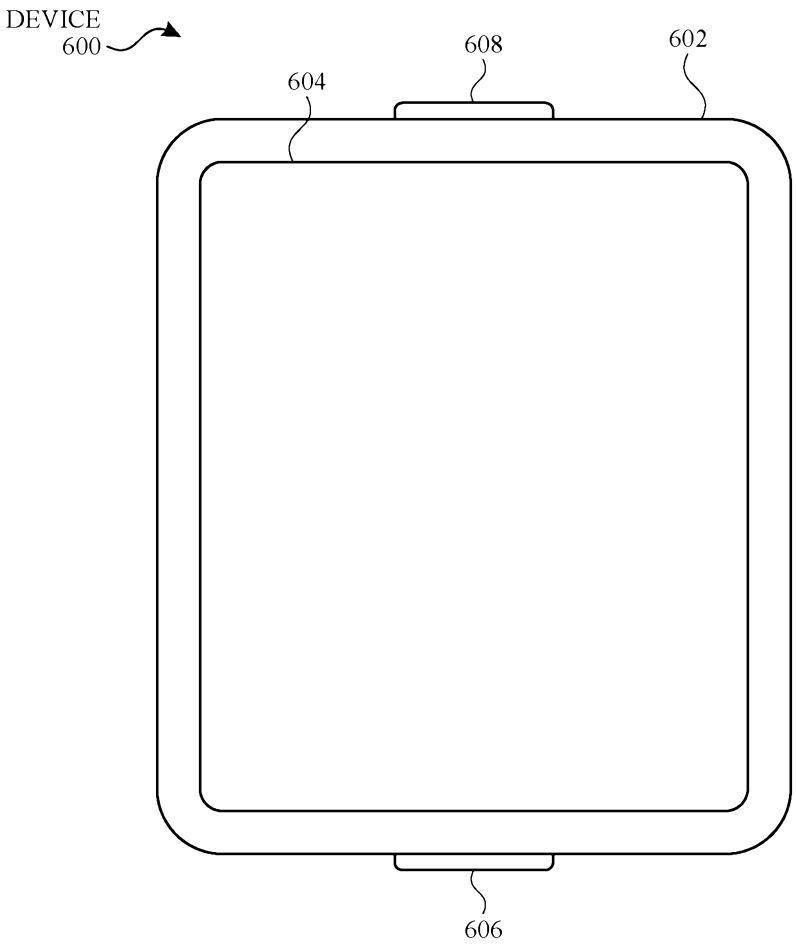


FIG. 6A

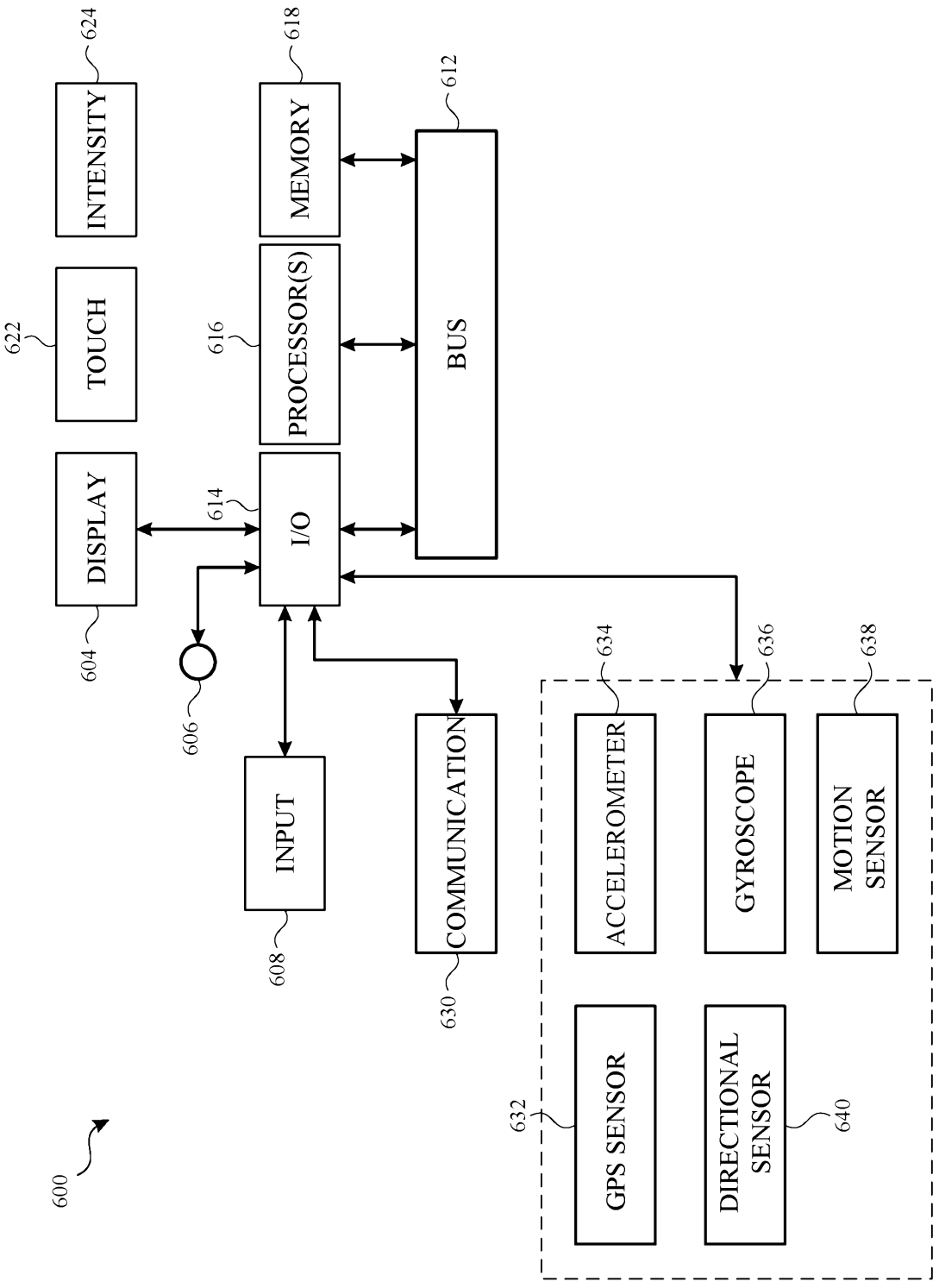

600
604
DISPLAY
622
TOUCH
624
INTENSITY
606
608
INPUT
614
I/O
616
PROCESSOR(S)
618
MEMORY
612
BUS
630
COMMUNICATION
632
GPS SENSOR
634
ACCELEROMETER
636
GYROSCOPE
638
MOTION SENSOR
640
DIRECTIONAL SENSOR

FIG. 6B

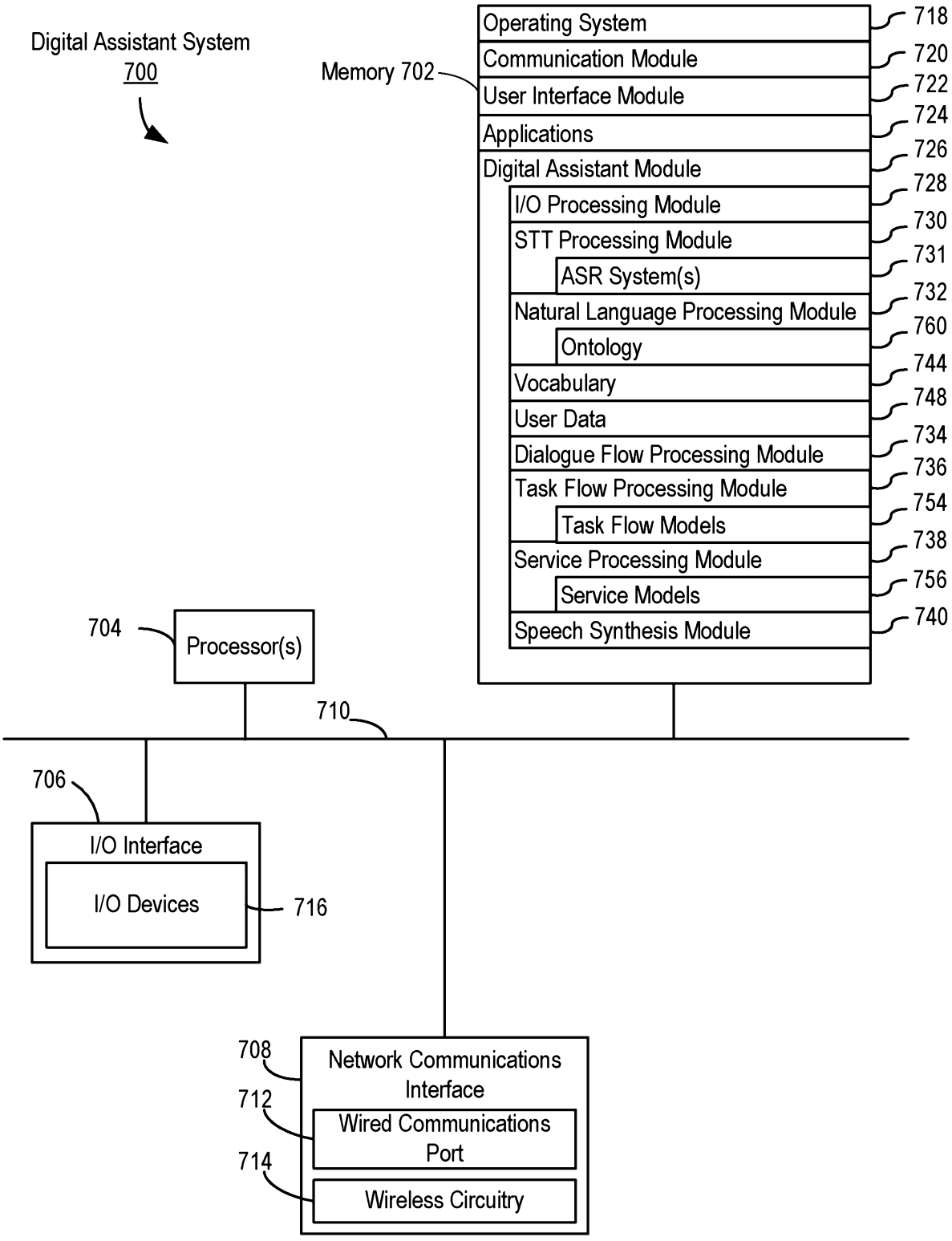

Digital Assistant System
700
Memory 702
Operating System
718
Communication Module
720
User Interface Module
722
Applications
724
Digital Assistant Module
726
I/O Processing Module
728
STT Processing Module
730
ASR System(s)
731
Natural Language Processing Module
732
Ontology
760
Vocabulary
744
User Data
748
Dialogue Flow Processing Module
734
Task Flow Processing Module
736
Task Flow Models
754
Service Processing Module
738
Service Models
756
Speech Synthesis Module
740
704
Processor(s)
710
706
I/O Interface
I/O Devices
716
708
Network Communications Interface
712
Wired Communications Port
714
Wireless Circuitry

FIG. 7A

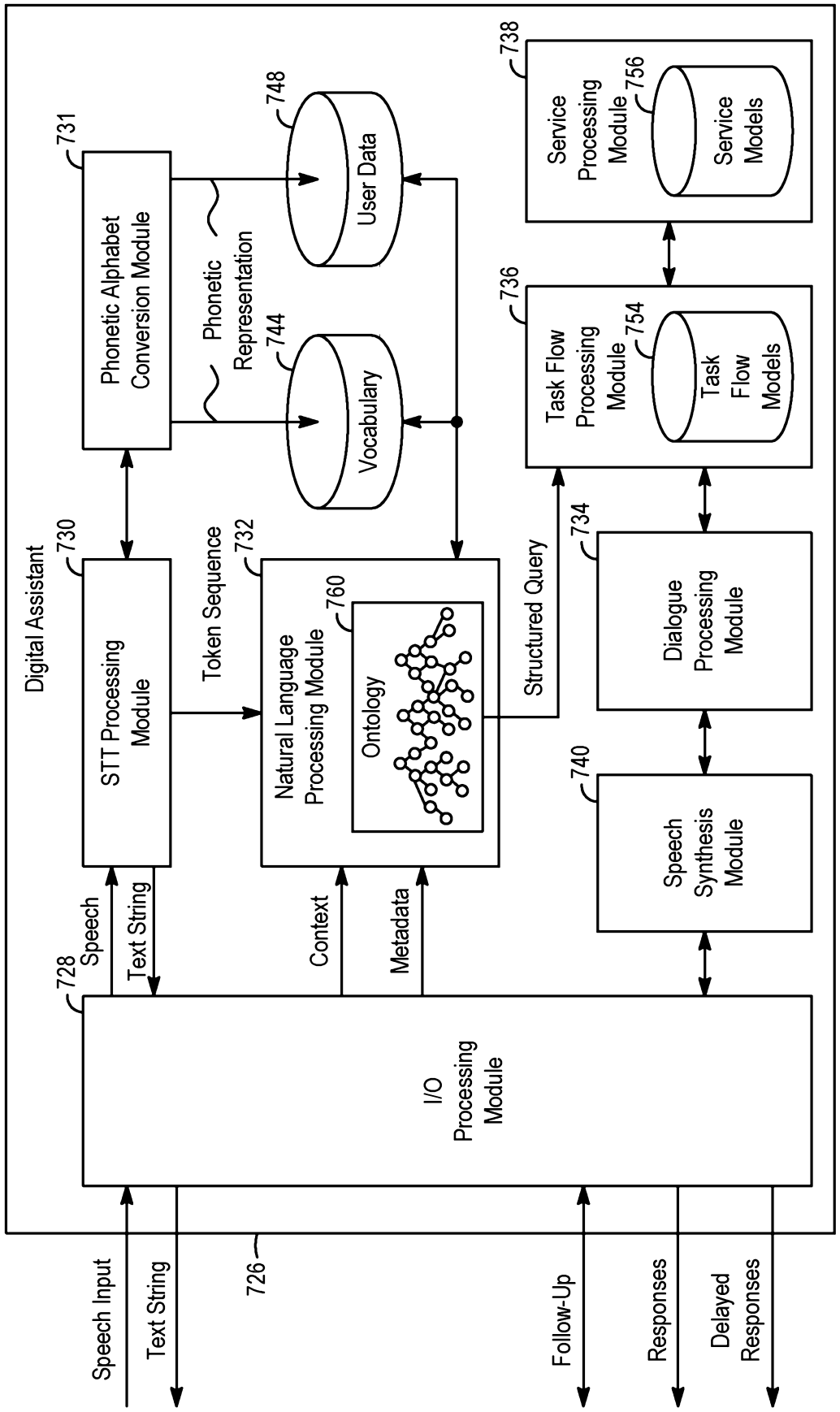

Digital Assistant
731
Phonetic Alphabet Conversion Module
748
User Data
Phonetic Representation
744
Vocabulary
738
Service Processing Module
756
Service Models
736
Task Flow Processing Module
754
Task Flow Models
730
STT Processing Module
Token Sequence
732
Natural Language Processing Module
760
Ontology
Structured Query
734
Dialogue Processing Module
740
Speech Synthesis Module
728
Speech
Text String
Context
Metadata
I/O Processing Module
726
Speech Input
Text String
Follow-Up
Responses
Delayed Responses

FIG. 7B

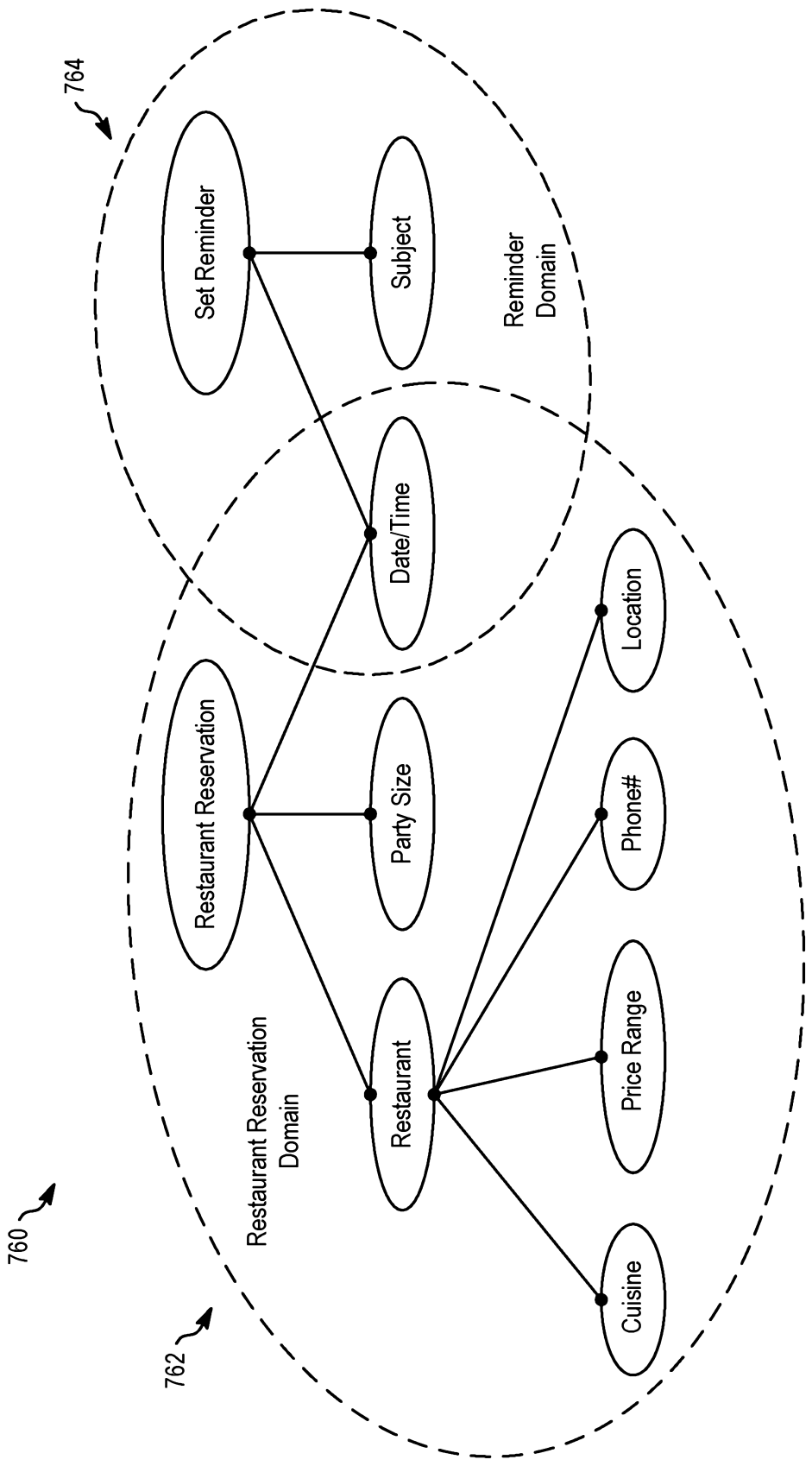

760
762
764
Restaurant Reservation Domain
Reminder Domain
Restaurant Reservation
Party Size
Date/Time
Restaurant
Location
Phone#
Price Range
Cuisine
Set Reminder
Subject

FIG. 7C

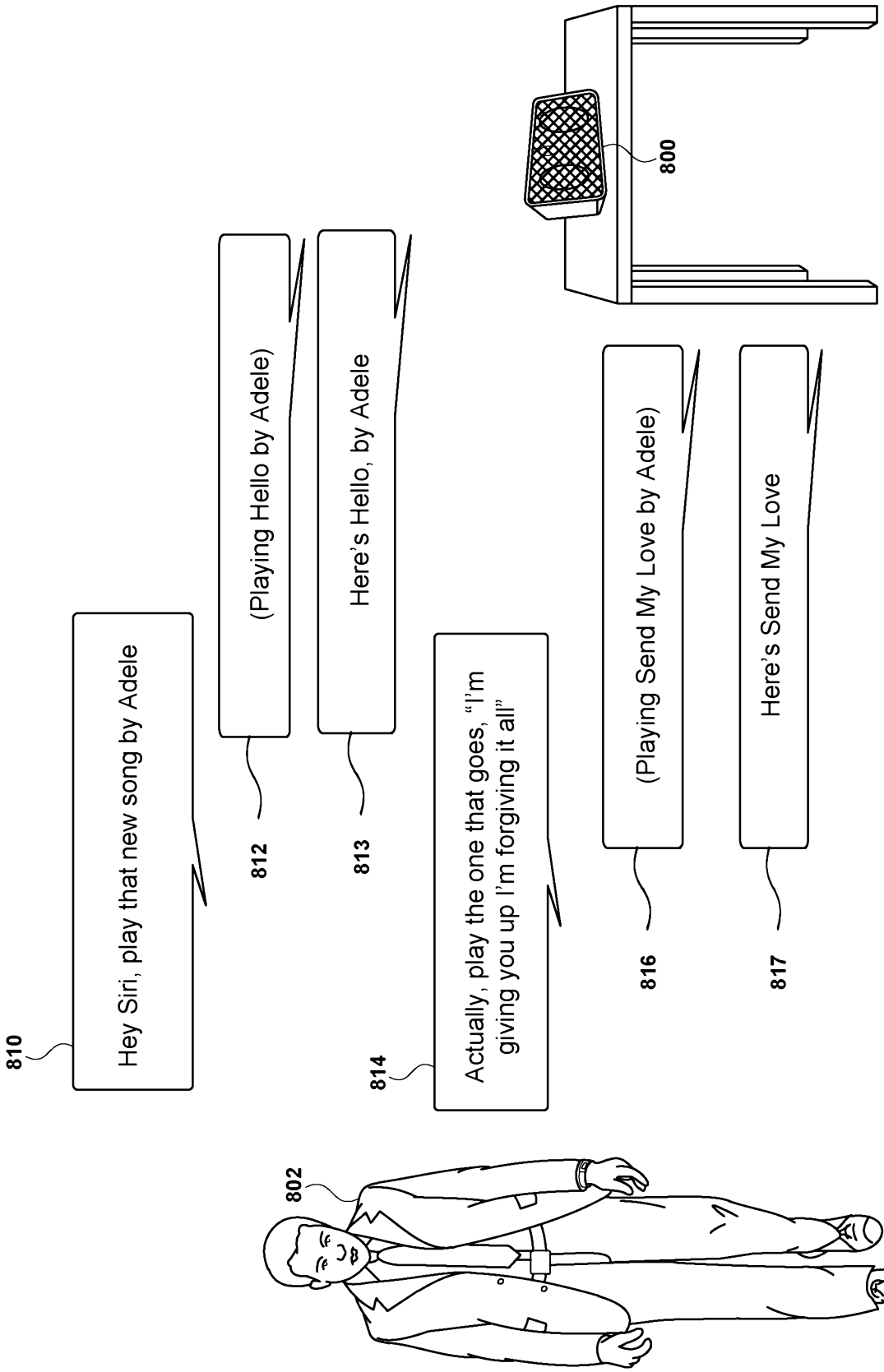

800
802
810
Hey Siri, play that new song by Adele
812
(Playing Hello by Adele)
813
Here's Hello, by Adele
814
Actually, play the one that goes, "I'm giving you up I'm forgiving it all"
816
(Playing Send My Love by Adele)
817
Here's Send My Love

FIG. 8A

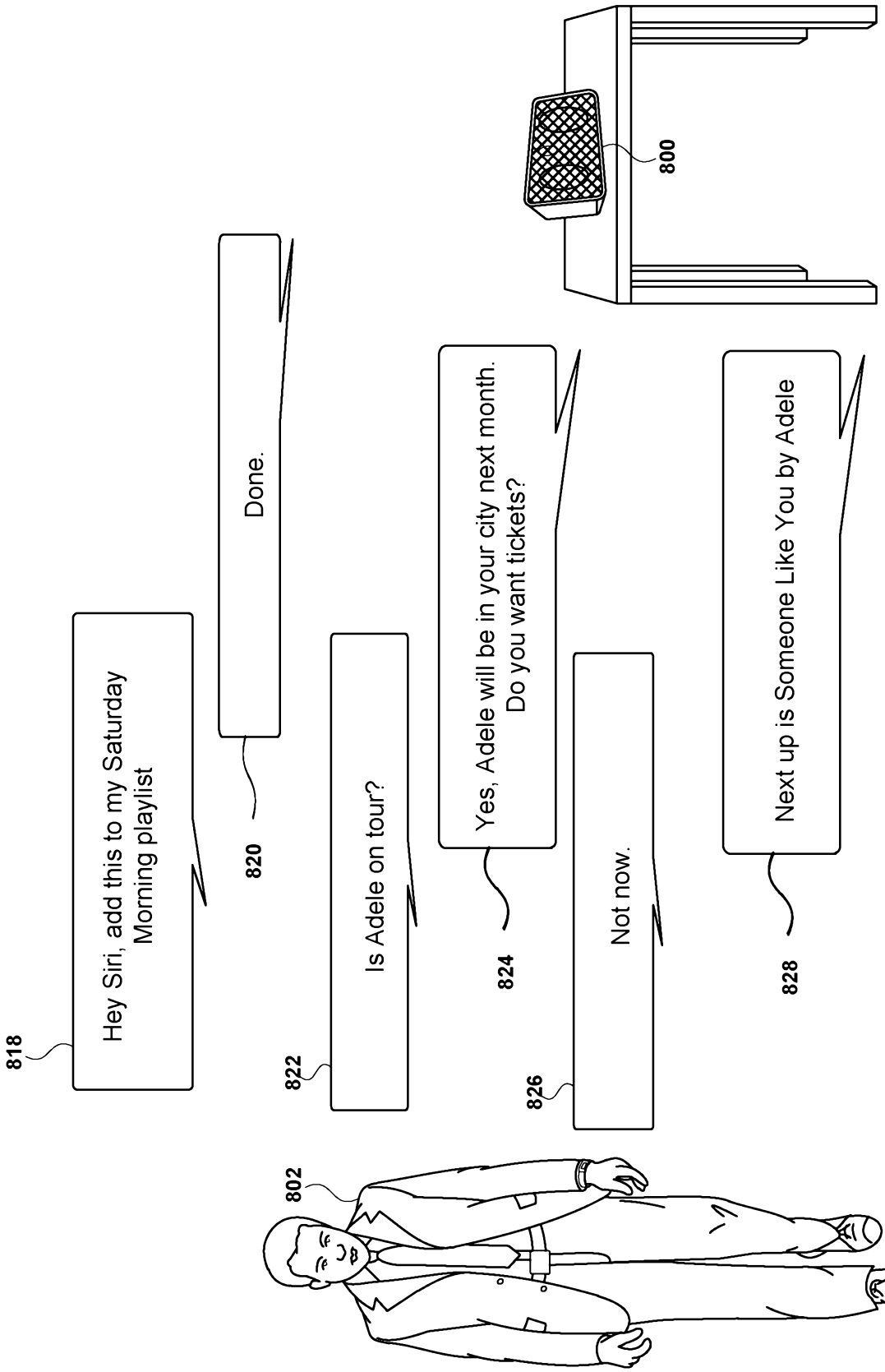

800
802
818
Hey Siri, add this to my Saturday Morning playlist
820
Done.
822
Is Adele on tour?
824
Yes, Adele will be in your city next month. Do you want tickets?
826
Not now.
828
Next up is Someone Like You by Adele

FIG. 8B

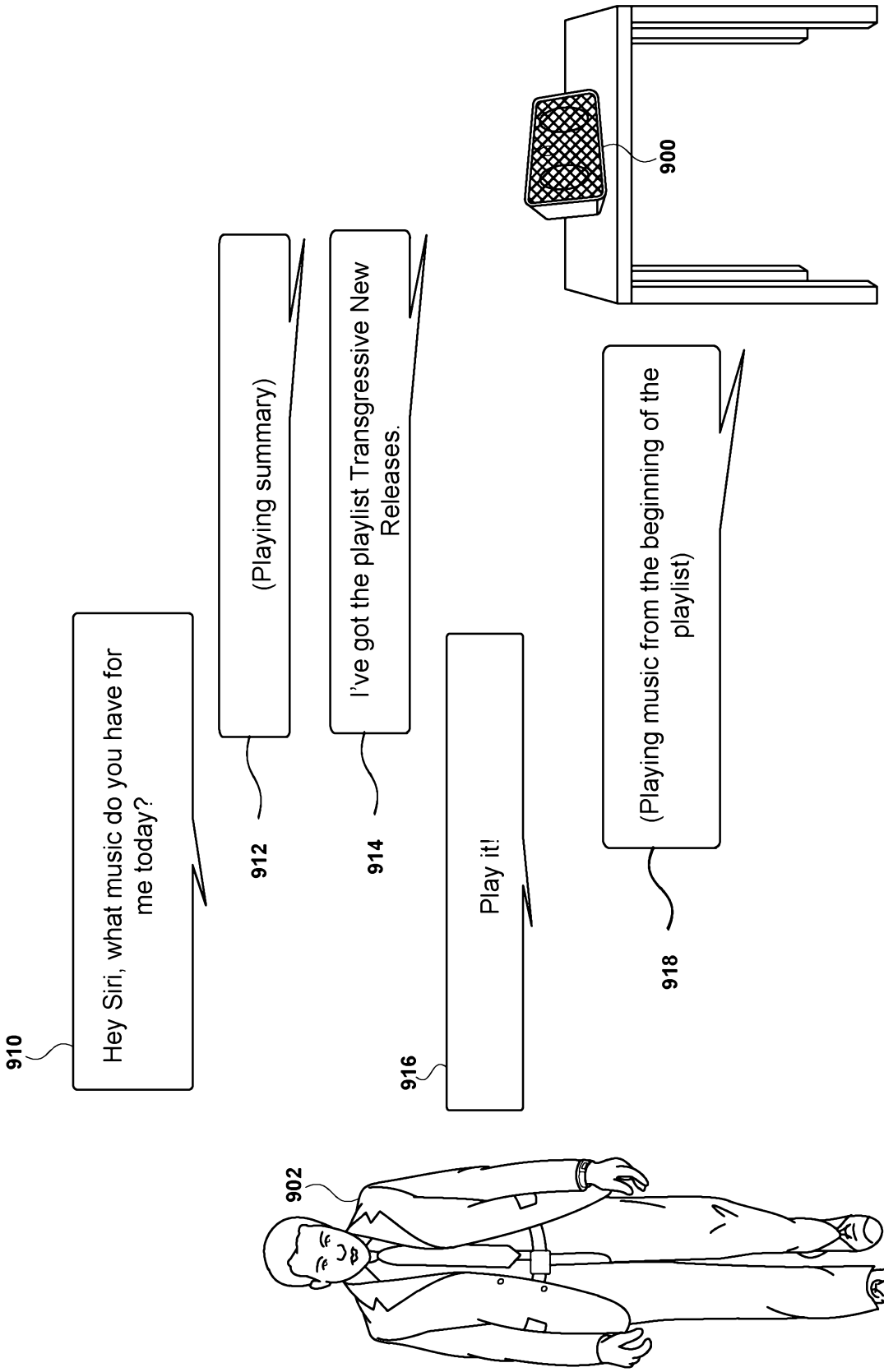

900
902
910
Hey Siri, what music do you have for me today?
912
(Playing summary)
914
I've got the playlist Transgressive New Releases.
916
Play it!
918
(Playing music from the beginning of the playlist)

FIG. 9A

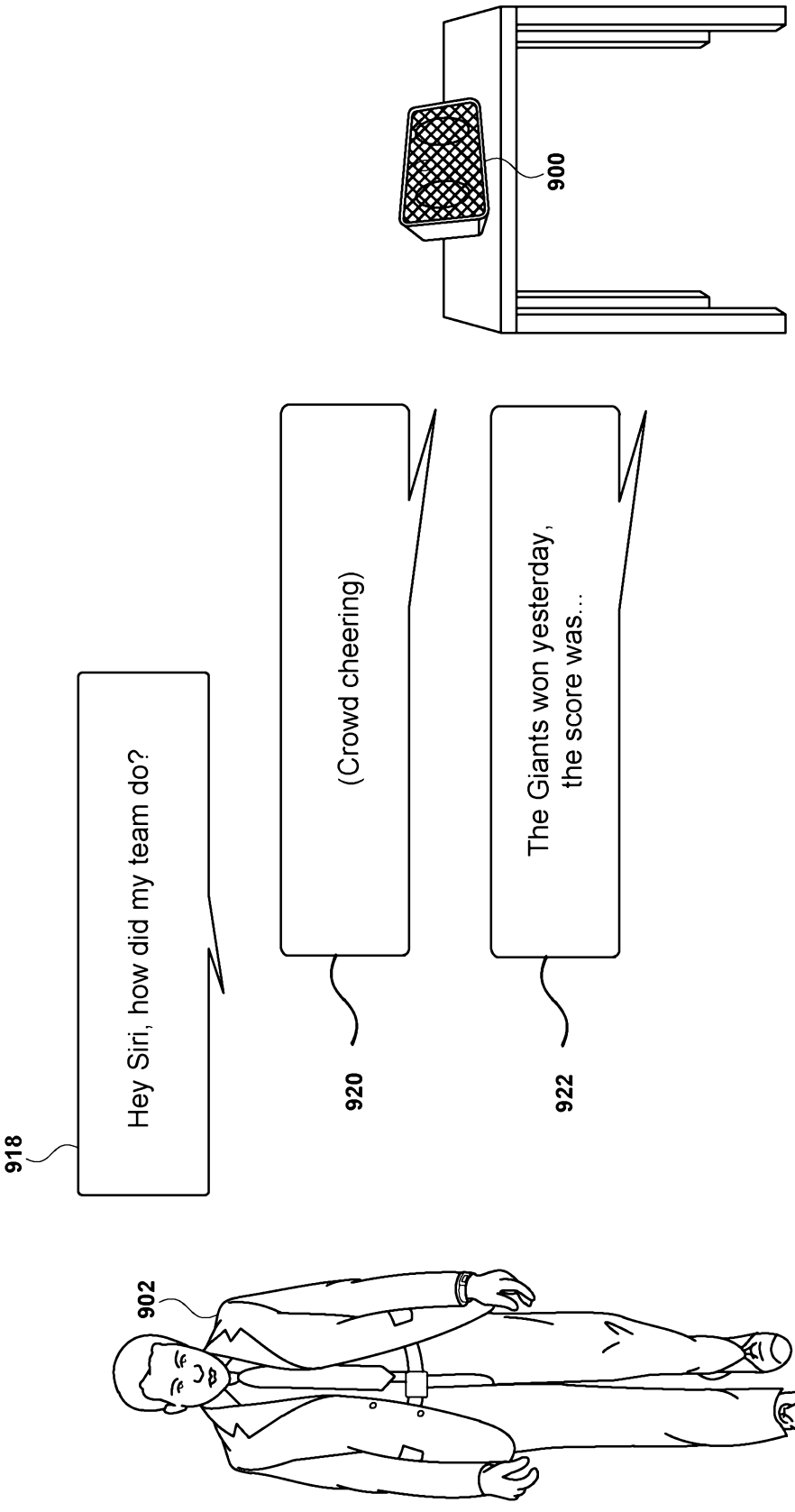

900
(Crowd cheering)
920
The Giants won yesterday, the score was...
922
Hey Siri, how did my team do?
918
902

FIG. 9B

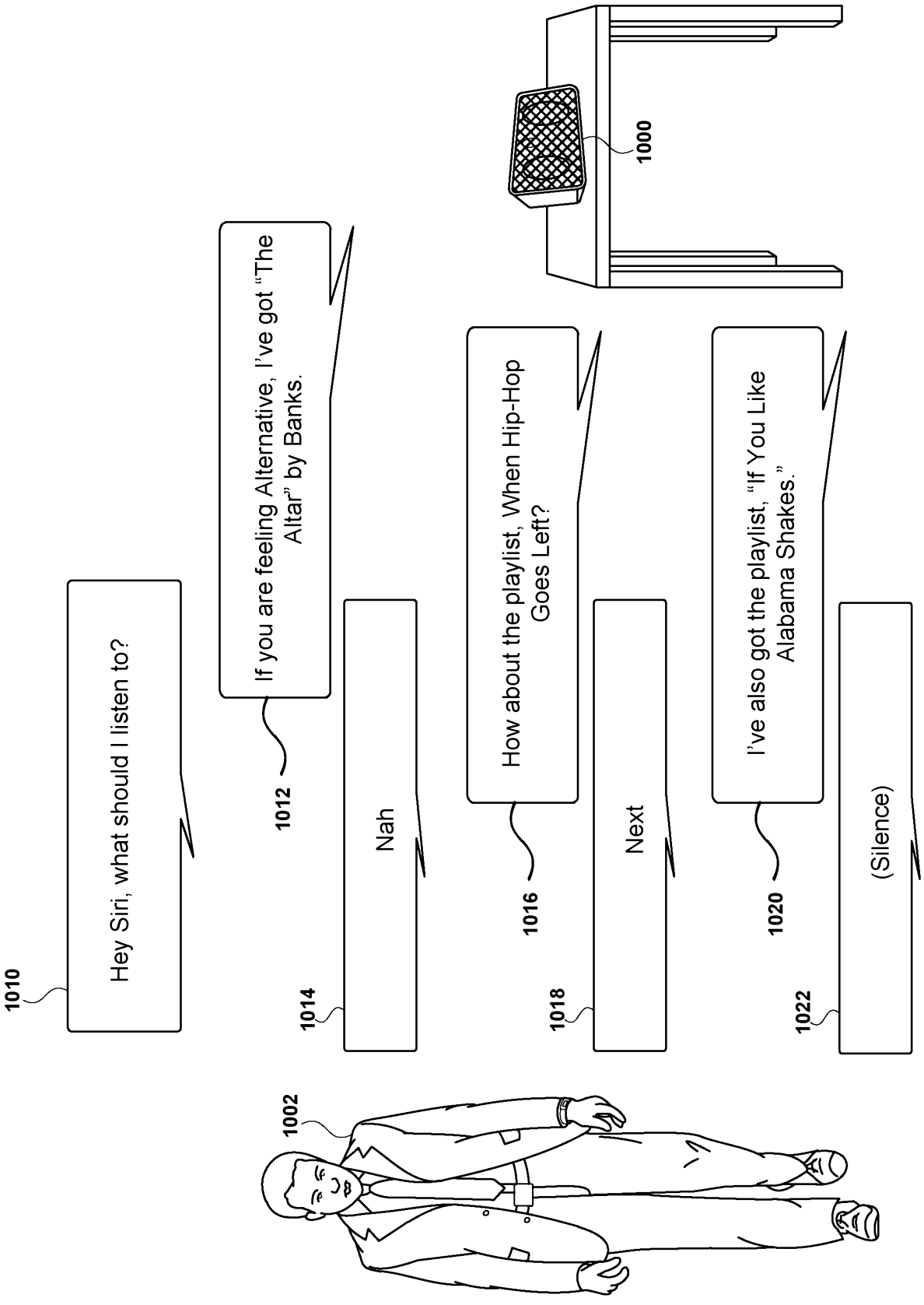

1000
1002
1010
Hey Siri, what should I listen to?
1012
If you are feeling Alternative, I've got "The Altar" by Banks.
1014
Nah
1016
How about the playlist, When Hip-Hop Goes Left?
1018
Next
1020
I've also got the playlist, "If You Like Alabama Shakes."
1022
(Silence)

FIG. 10A

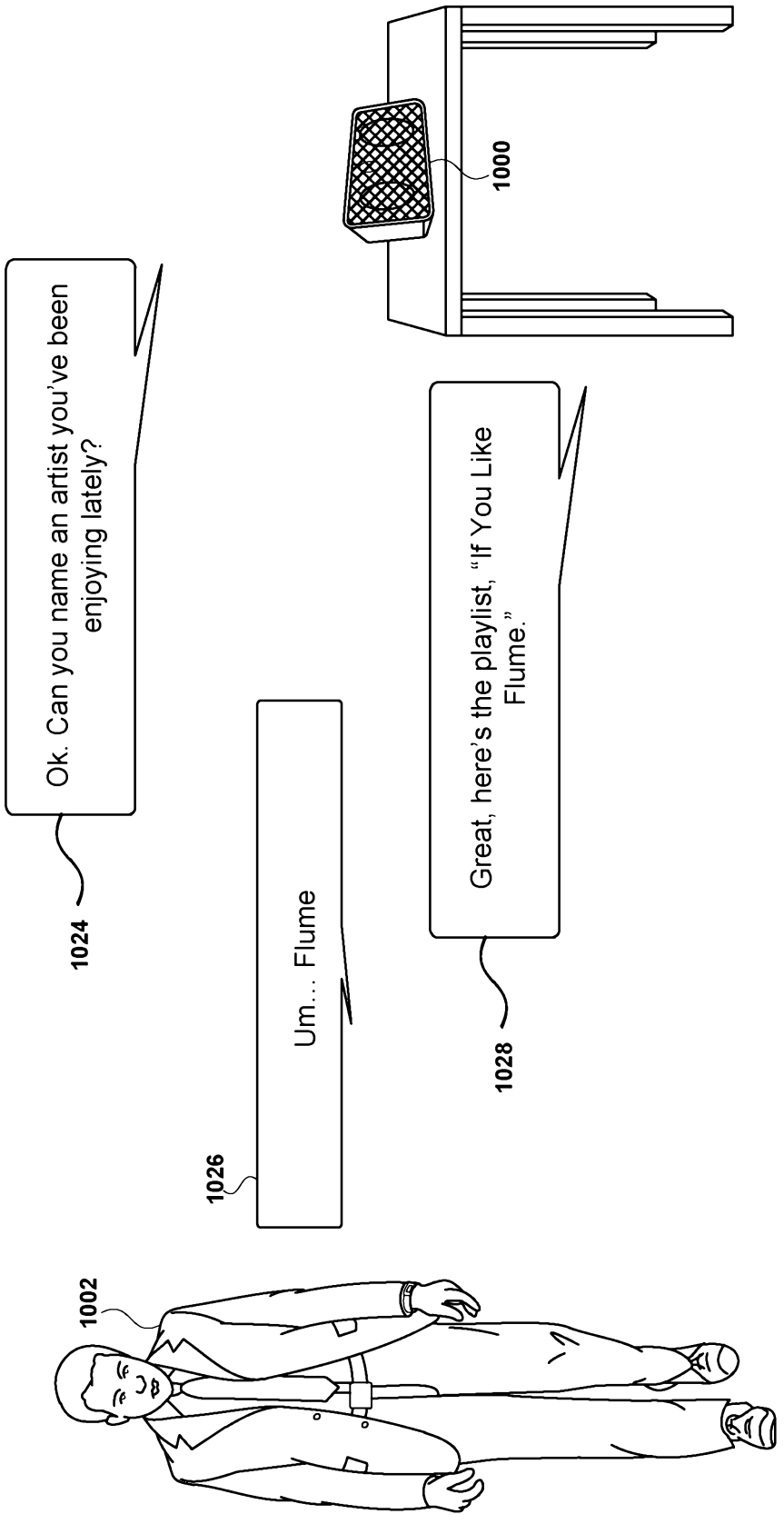

1000
Ok. Can you name an artist you've been enjoying lately?
1024
Um... Flume
1026
Great, here's the playlist, "If You Like Flume."
1028
1002

FIG. 10B

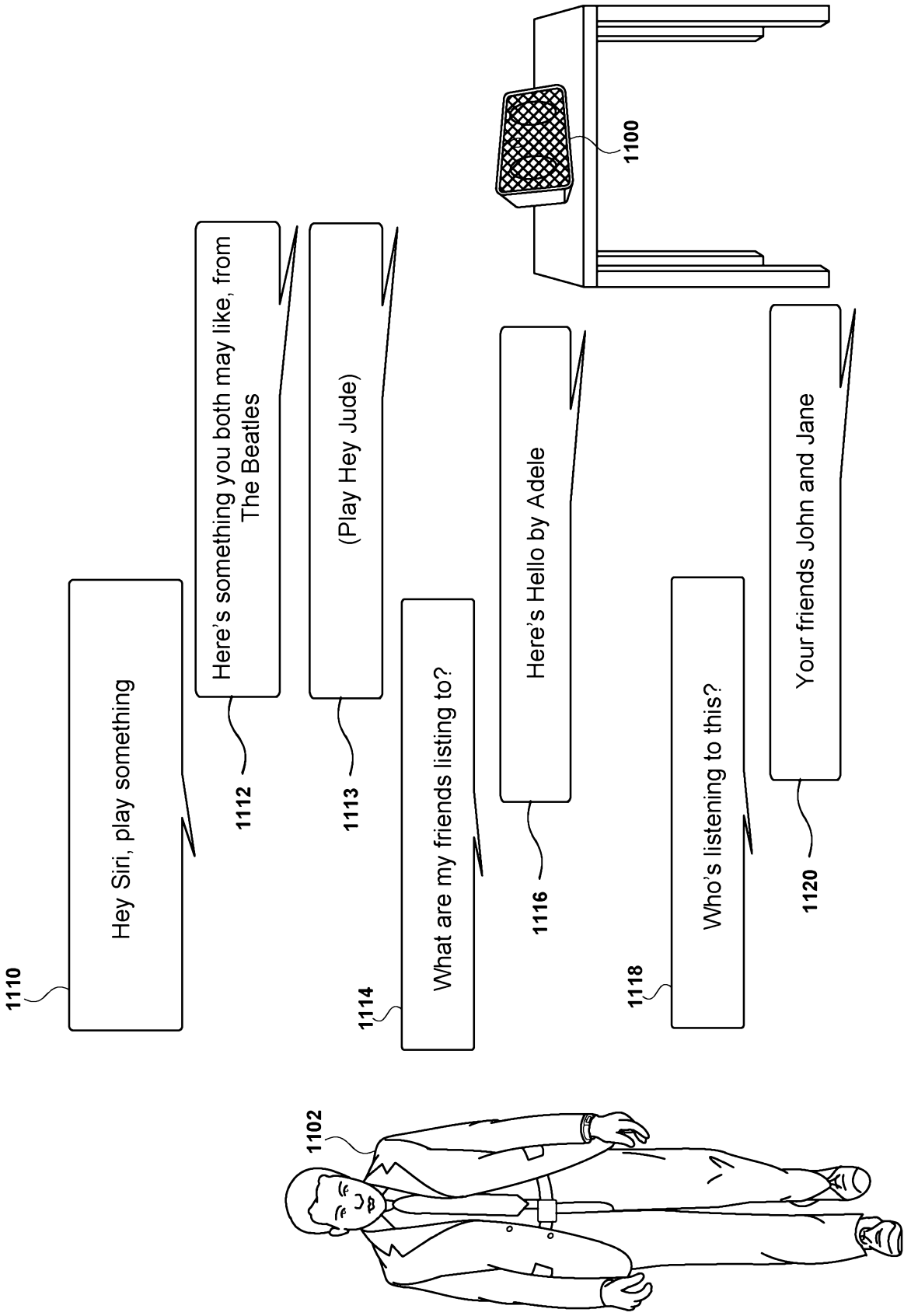

1100
1102
1110
Hey Siri, play something
1112
Here's something you both may like, from The Beatles
1113
(Play Hey Jude)
1114
What are my friends listing to?
1116
Here's Hello by Adele
1118
Who's listening to this?
1120
Your friends John and Jane

FIG. 11

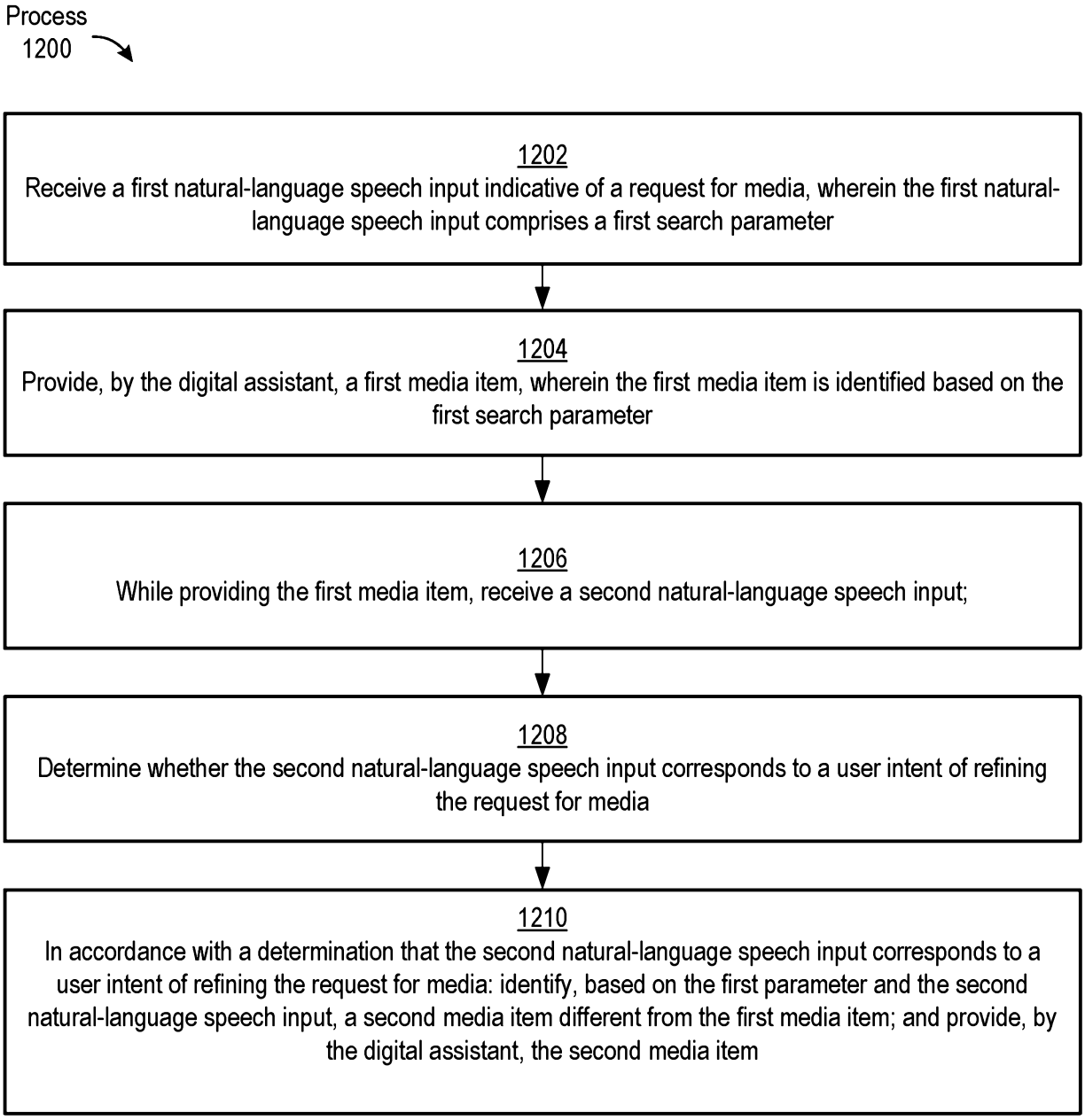

Process
1200
1202
Receive a first natural-language speech input indicative of a request for media, wherein the first natural-language speech input comprises a first search parameter
1204
Provide, by the digital assistant, a first media item, wherein the first media item is identified based on the first search parameter
1206
While providing the first media item, receive a second natural-language speech input;
1208
Determine whether the second natural-language speech input corresponds to a user intent of refining the request for media
1210
In accordance with a determination that the second natural-language speech input corresponds to a user intent of refining the request for media: identify, based on the first parameter and the second natural-language speech input, a second media item different from the first media item; and provide, by the digital assistant, the second media item

FIG. 12

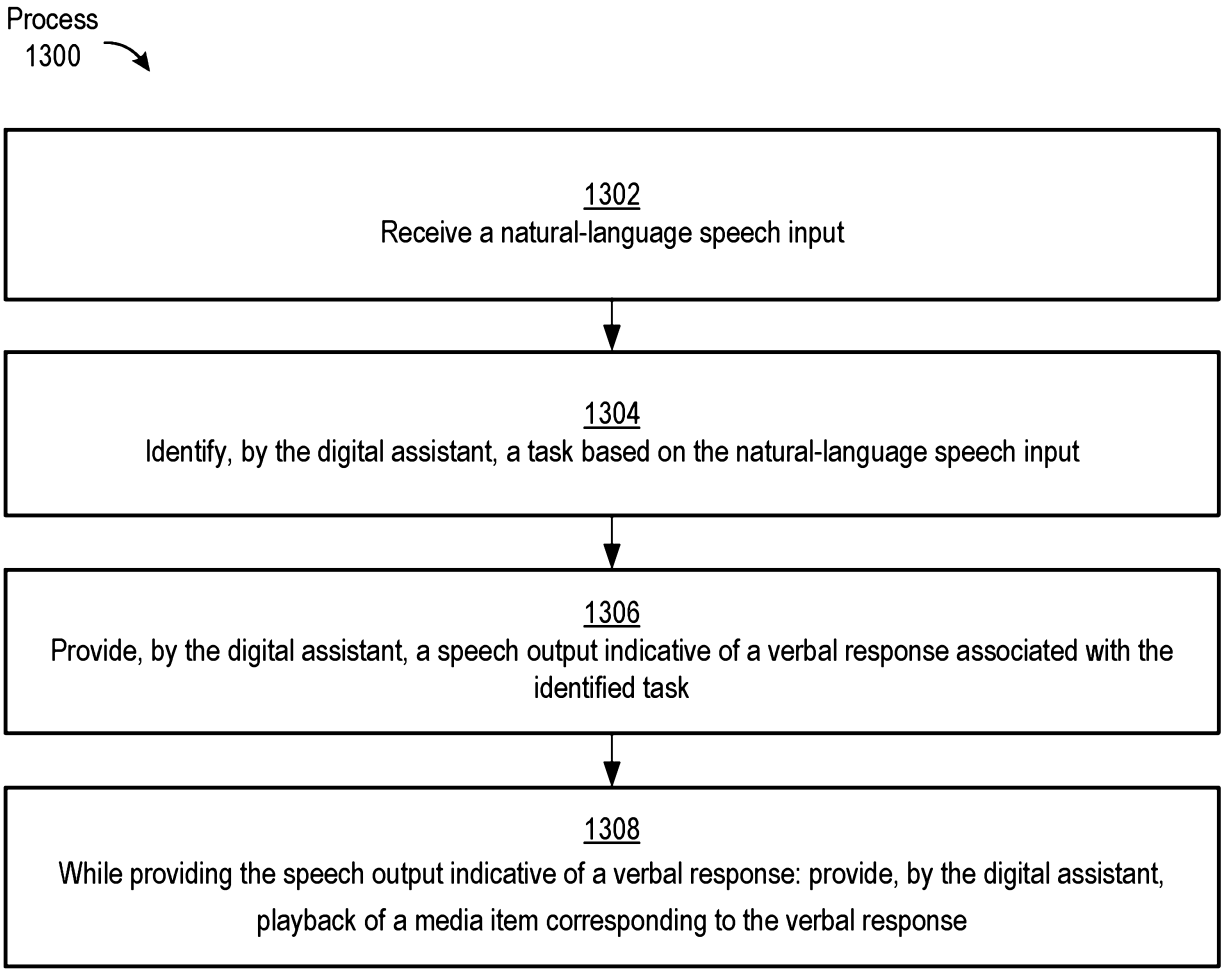

Process
1300
1302
Receive a natural-language speech input
1304
Identify, by the digital assistant, a task based on the natural-language speech input
1306
Provide, by the digital assistant, a speech output indicative of a verbal response associated with the identified task
1308
While providing the speech output indicative of a verbal response: provide, by the digital assistant, playback of a media item corresponding to the verbal response

FIG. 13

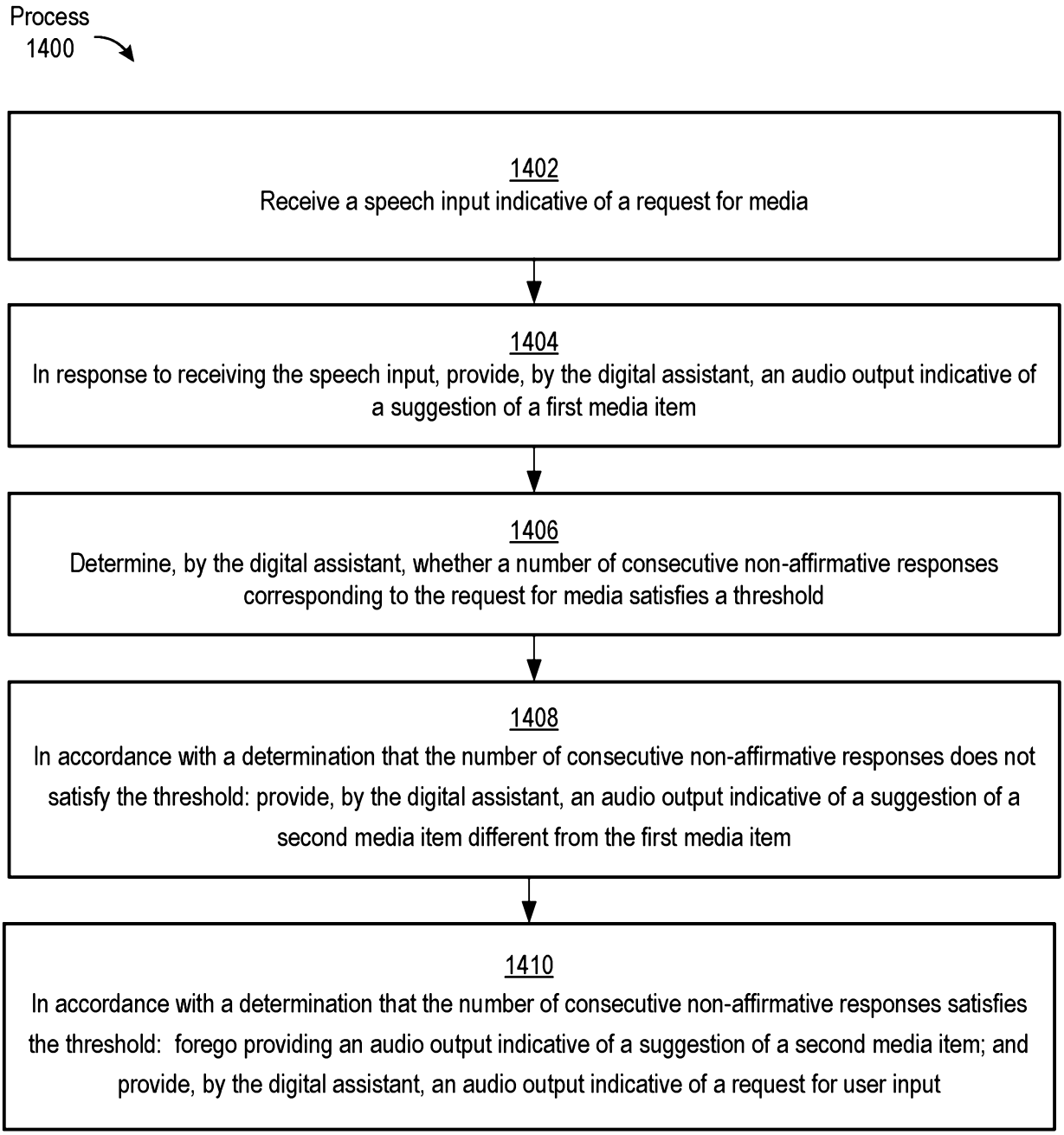

Process
1400
1402
Receive a speech input indicative of a request for media
1404
In response to receiving the speech input, provide, by the digital assistant, an audio output indicative of a suggestion of a first media item
1406
Determine, by the digital assistant, whether a number of consecutive non-affirmative responses corresponding to the request for media satisfies a threshold
1408
In accordance with a determination that the number of consecutive non-affirmative responses does not satisfy the threshold: provide, by the digital assistant, an audio output indicative of a suggestion of a second media item different from the first media item
1410
In accordance with a determination that the number of consecutive non-affirmative responses satisfies the threshold: forego providing an audio output indicative of a suggestion of a second media item; and provide, by the digital assistant, an audio output indicative of a request for user input

FIG. 14

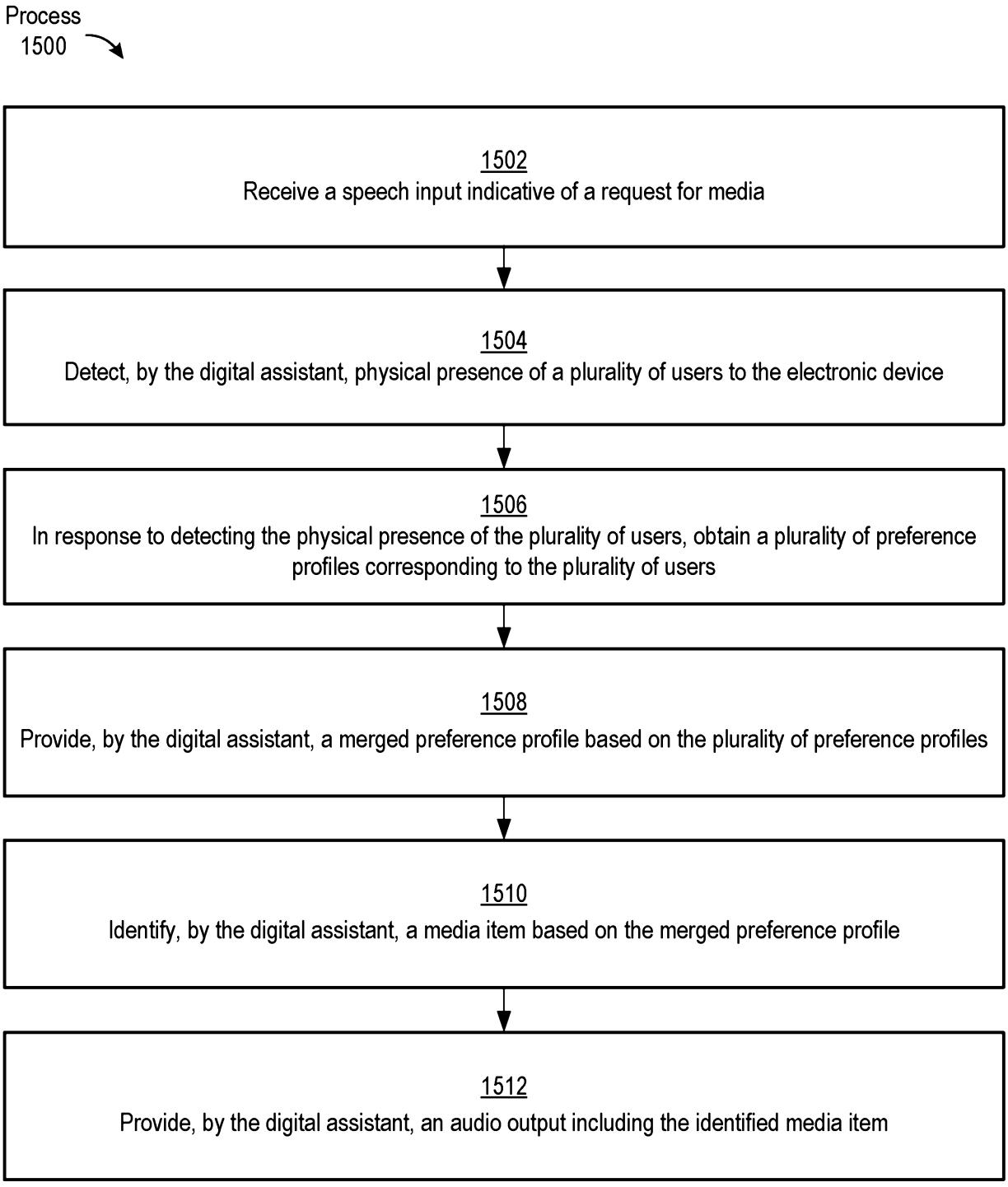

Process
1500
1502
Receive a speech input indicative of a request for media
1504
Detect, by the digital assistant, physical presence of a plurality of users to the electronic device
1506
In response to detecting the physical presence of the plurality of users, obtain a plurality of preference profiles corresponding to the plurality of users
1508
Provide, by the digital assistant, a merged preference profile based on the plurality of preference profiles
1510
Identify, by the digital assistant, a media item based on the merged preference profile
1512
Provide, by the digital assistant, an audio output including the identified media item

FIG. 15