

(19) **DANMARK**

(10) **DK 2017 70338 A1**



(12) **PATENTANSØGNING**

Patent- og
Varemærkestyrelsen

-
- (51) Int.Cl.: **G 06 F 17/30 (2006.01)** **G 06 F 17/20 (2006.01)** **G 10 L 15/22 (2006.01)**
- (21) Ansøgningsnummer: **PA 2017 70338**
- (22) Indleveringsdato: **2017-05-15**
- (24) Løbedag: **2017-05-15**
- (41) Alm. tilgængelig: **2017-12-09**
- (30) Prioritet: **2016-06-08 US 62/347,480** **2016-09-15 US 15/266,956**
2016-09-23 US PCT/US2016/053445
- (71) Ansøger: **Apple Inc., 1 Infinite Loop, Cupertino, 95014 California, USA**
- (72) Opfinder: **Ryan Orr, c/o Apple Inc., 1 Infinite Loop, Cupertino, 95014 CA California, USA**
Matthew P. Bernardo, 1105 Abelia Irvine, 92606 CA California, USA
Daniel J. Mandel, c/o Apple Inc, 1 Infinite Loop, Cupertino, 95014 CA California, USA
- (74) Fuldmægtig: **Zacco Denmark A/S, Arne Jacobsens Allé 15, 2300 København S, Danmark**
- (54) Benævnelse: **Intelligent automated assistant for media exploration**
- (56) Fremdragne publikationer:
US 8239206 B1
US 2014/0310001 A1
US 2016/0125048 A1
US 2014/0324884 A1
CA 2631796 A1
US 2015/0382047 A1
WO 2012/019637 A1
- (57) Sammendrag:
Systems and processes for operating an intelligent automated assistant to explore media items are provided. In one example process, a speech input representing a request for one or more media items is received from a user. The process determines whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items. In response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items, at least one media item is obtained from a user-specific corpus of media items. The user-specific corpus of media items is generate based on data associated with the user. The at least one media item is provided.

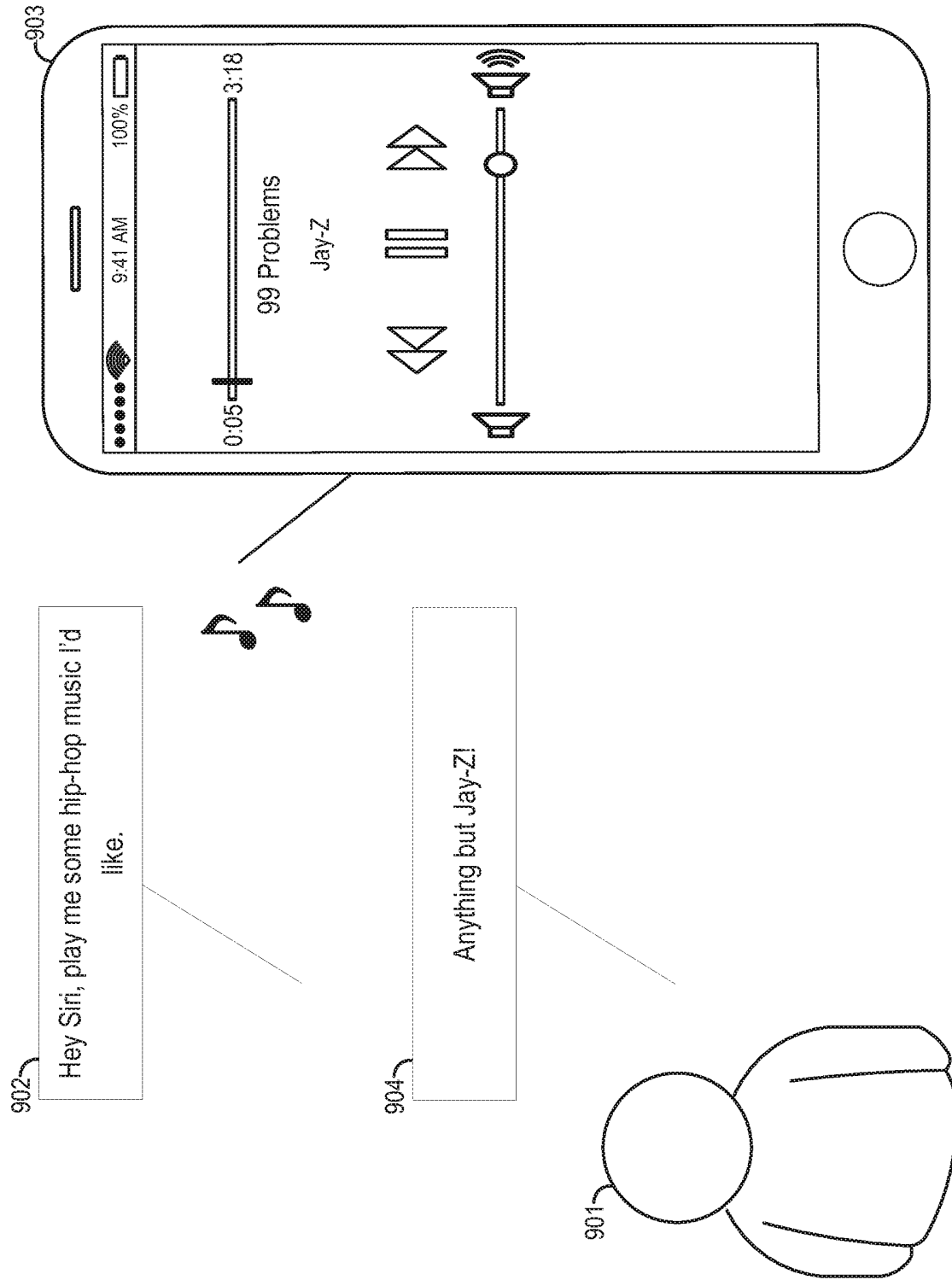


FIG. 9A

INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION

Cross-Reference to Related Applications

[0001] This application claims priority to U.S. Provisional Serial No. 62/347,480, filed on June 8, 2016, entitled INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION; U.S. Non-Provisional Serial No. 15/266,956, filed on September 15, 2016, entitled INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION; and PCT Application Serial No. PCT/US2016/053445, filed on September 23, 2016, entitled INTELLIGENT AUTOMATED ASSISTANT FOR MEDIA EXPLORATION, all of which are hereby incorporated by reference in their entirety for all purposes.

Field

[0002] This relates generally to intelligent automated assistants and, more specifically, to intelligent automated assistants 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] When managing music or other media, a digital assistant can be helpful in searching for or playing back specific media, particularly in a hands-free environment. In particular, a digital assistant can respond effectively to a request to play a specific media item, such as an album or a song identified specifically by title or by artist. However, digital assistants can struggle with discovering relevant media items based on vague open-ended natural language requests, such as, for example, a request to recommend a song or album.

Summary

[0005] Systems and processes for operating an intelligent automated assistant to explore media items are provided. In one example process, a speech input representing a request for one or more media items is received from a user. The process determines whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items. In response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items, at least one media item is obtained from a user-specific corpus of media items. The user-specific corpus of media items is generated based on data associated with the user. The at least one media item is provided.

Brief Description of the Drawings

[0006] FIG. 1 is a block diagram illustrating a system and environment for implementing a digital assistant, according to various examples.

[0007] FIG. 2A is a block diagram illustrating a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0008] FIG. 2B is a block diagram illustrating exemplary components for event handling, according to various examples.

[0009] FIG. 3 illustrates a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0010] FIG. 4 is a block diagram of an exemplary multifunction device with a display and a touch-sensitive surface, according to various examples.

[0011] FIG. 5A illustrates an exemplary user interface for a menu of applications on a portable multifunction device, according to various examples.

[0012] 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.

[0013] FIG. 6A illustrates a personal electronic device, according to various examples.

[0014] FIG. 6B is a block diagram illustrating a personal electronic device, according to various examples.

[0015] FIG. 7A is a block diagram illustrating a digital assistant system or a server portion thereof, according to various examples.

[0016] FIG. 7B illustrates the functions of the digital assistant shown in FIG. 7A, according to various examples.

[0017] FIG. 7C illustrates a portion of an ontology, according to various examples.

[0018] FIGs. 8A-C illustrate a process for operating a digital assistant for media exploration, according to various examples.

[0019] FIGs. 9A-B illustrate a user operating a digital assistant for media exploration, according to various examples.

[0020] FIG. 10 illustrates a user operating a digital assistant for media exploration, according to various examples.

[0021] FIG. 11 illustrates a user operating a digital assistant for media exploration, according to various examples.

[0022] FIG. 12 illustrates a functional block diagram of an electronic device, according to various examples.

Detailed Description

[0023] In the following description of examples, reference is made to the accompanying drawings in which it is 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.

[0024] Conventional techniques for exploring media content using digital assistants can be generally cumbersome and inefficient. In particular, media-related requests that are in natural language form are, for example, overly broad or ambiguous, and thus it is difficult to accurately infer the user's intent that corresponds to the request. For example, the media-related request "Play me something good" is vague and open-ended, and thus with existing techniques, digital assistants may retrieve media items that are incompatible with the user's preferences, may overwhelm the user with too many media items, or may return nothing at

all. This can result in a large number of follow-up interactions between the user and the digital assistant to clarify the user's intent. This can negatively impact user experience. Additionally, a large number of follow-up interactions is inefficient with respect to energy consumption of the device. This consideration is particularly important for battery-operated devices.

[0025] In accordance with some systems, computer-readable media, and processes described herein, media exploration is performed by a digital assistant in a more efficient and accurate manner. In one example process, a speech input representing a request for one or more media items is received from a user. The process determines whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items. In response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items, at least one media item is obtained from a user-specific corpus of media items. The at least one media item is obtained using a user-specific media ranking model. The user-specific corpus of media items and the user-specific media ranking model are generated based on data associated with the user. The at least one media item is then provided to the user. By using a user-specific corpus of media items and a user-specific media ranking model to obtain the at least one media item, the likelihood that the at least one media item matches the preferences of the user increases. As a result, media items that are more relevant to the user are recommended, which increases the efficiency and usefulness of the digital assistant.

[0026] 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.

[0027] 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.

[0028] 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

[0029] 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.

[0030] 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.

[0031] 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.

[0032] 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 (e.g., media service(s) 120-1, navigation service(s) 120-2, messaging service(s) 120-3, information service(s) 120-4, calendar service 120-5, telephony service 120-6, etc.) through network(s) 110 for task completion or information acquisition. I/O interface to external services 118 facilitates such communications.

[0033] In particular, DA server 106 communicates to media service(s) to perform tasks that include searching for and obtaining media items. Media service(s) 120-1 is implemented, for example, on one or more remote media servers and is configured to provide media items, such as songs, albums, playlists, videos, or the like. For example, media service(s) includes media streaming services, such as Apple Music or iTunes Radio™

(services of Apple, Inc. of Cupertino, California). Media service(s) 120-1 is configured to receive a media search query (e.g., from DA server 106) and in response, provide one or more media items that satisfy the media search queries. Specifically, in accordance with the media search query, one or more corpuses of media items are searched to identify one or more media items and the identified one or more media items are provided. Further, media service(s) are configured to provide media information associated with media items, such as the names of artists associated with specific media items, the release dates of specific media items, or the lyrics of specific media items.

[0034] Media service(s) 120-1 include various corpuses of media items. The corpuses of media items include a plurality of user-specific corpuses of media items. Each user-specific corpus of media items is generated based on media-related data associated with a respective user. The media-related data includes, for example, user input indicating media items that were previously viewed, selected, requested, acquired, or rejected by the user. Additionally, the media-related data includes the media items found in the personal library of media items associated with the user. Thus, the media items contained in each user-specific corpus of media items reflect the media preferences of the respective user. In some examples, each user-specific corpus of media items is identified and accessed based on user information, such as user log-in information and/or user password information. In some examples, the corpuses of media items in media service(s) 120-1 further include one or more second corpuses of media items generated based on the release dates of the media items. For example, the one or more second corpuses of media items only contain media items having release dates that are within a predetermined time range from the current date.

[0035] In some examples, each media item in the corpuses of media items includes metadata that indicates one or more media parameters. The media parameters include, for example, {title}, {artist}, {subgenre}, {release date}, {mood}, {occasion}, {editorial list}, {political preference}, {technical proficiency}, or the like. Media items in the corpuses of media items are thus searched and retrieved based on the media parameters indicated in the metadata of the media items. Additional description regarding media parameters associated with media items is provided below with reference to FIGs. 8A-C.

[0036] 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.

[0037] 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.

[0038] 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.

[0039] 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.

[0040] 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.

[0041] 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

[0042] 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.

[0043] 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).

[0044] 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.

[0045] 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.

[0046] 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.

[0047] 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.

[0048] 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.

[0049] 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.

[0050] 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).

[0051] 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).

[0052] 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, which is hereby incorporated by reference in its entirety. 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.

[0053] 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.

[0054] 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.

[0055] 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.

[0056] 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, each of which is hereby incorporated by reference in its entirety. However, touch screen 212 displays visual output from device 200, whereas touch-sensitive touchpads do not provide visual output.

[0057] 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. All of these applications are incorporated by reference herein in their entirety.

[0058] 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.

[0059] 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.

[0060] 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.

[0061] 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.

[0062] 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.

[0063] 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," which are hereby incorporated by reference in their entirety. 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).

[0064] 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.

[0065] 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," both of which are incorporated by reference herein in their entirety. 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.

[0066] 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.

[0067] 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.

[0068] 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.

[0069] 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.

[0070] 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).

[0071] 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.

[0072] 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.

[0073] 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.

[0074] 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.

[0075] 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).

[0076] 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).

[0077] 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.

[0078] 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.

[0079] 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.

[0080] 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 include 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.

[0081] 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.

[0082] 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.

[0083] 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.

[0084] 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.

[0085] 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.

[0086] 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.

[0087] 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.

[0088] 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.

[0089] 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).

[0090] 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.

[0091] 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.

[0092] 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.

[0093] 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.

[0094] 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.

[0095] 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).

[0096] 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).

[0097] 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.

[0098] 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.).

[0099] 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.

[0100] 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.

[0101] 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, the contents of which are hereby incorporated by reference in their entirety.

[0102] 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.

[0103] 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.

[0104] 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.

[0105] 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).

[0106] 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.

[0107] 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.

[0108] 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.

[0109] 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).

[0110] In some embodiments, event sorter 270 also includes a hit view determination module 272 and/or an active event recognizer determination module 273.

[0111] 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.

[0112] 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.

[0113] 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.

[0114] 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.

[0115] 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.

[0116] 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.

[0117] 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.

[0118] 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).

[0119] 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.

[0120] 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.

[0121] 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.

[0122] 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.

[0123] 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.

[0124] 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.

[0125] 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.

[0126] 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.

[0127] 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.

[0128] 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.

[0129] 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.

[0130] 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.

[0131] 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.

[0132] 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.

[0133] 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.

[0134] 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.

[0135] Attention is now directed towards embodiments of user interfaces that can be implemented on, for example, portable multifunction device 200.

[0136] 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:

[0137] 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.

[0138] 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.

[0139] 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.

[0140] 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.

[0141] 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.

[0142] 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.

[0143] 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, each of which is hereby incorporated by reference in their entirety.

[0144] 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.

[0145] 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.

[0146] 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.

[0147] 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.

[0148] 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, and/or 600 (FIGS. 2, 4, and 6). For example, an image (e.g., icon), a button, and text (e.g., hyperlink) each constitutes an affordance.

[0149] 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).

[0150] 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.

[0151] 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.

[0152] 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.

[0153] 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.

[0154] 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).

[0155] 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).

[0156] 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

[0157] 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, or 600) 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.

[0158] 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.

[0159] 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).

[0160] 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, or 600 in FIGs. 2A, 4, 6A-B, 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, or 600).

[0161] 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.

[0162] 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.

[0163] 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.

[0164] 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.

[0165] 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.).

[0166] 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.

[0167] 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.

[0168] 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, phrases, intentions, etc.); determining the task flow for fulfilling the inferred intent; and executing the task flow to fulfill the inferred intent.

[0169] 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.

[0170] 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.

[0171] 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, the entire disclosure of which is incorporated herein by reference.

[0172] 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'meɪrou/ and /tə'matou/. 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.

[0173] In some examples, the candidate pronunciations are ranked based on the commonness of the candidate pronunciation. For example, the candidate pronunciation /tə'meɪrou/ is ranked higher than /tə'matou/, 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'meɪrou/ is associated with the United States, whereas the candidate pronunciation /tə'matou/ 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'meɪrou/ (associated with the United States) is ranked higher than the candidate pronunciation /tə'matou/ (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).

[0174] 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'meɪrou/ 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.”

[0175] 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 can determine that the sequence of phonemes /tə'meɪrou/ corresponds to the word “tomato,” even if that particular sequence of phonemes is not one of the candidate sequence of phonemes for that word.

[0176] Natural language processing module 732 (“natural language processor”) of the digital assistant can take the sequence of words or tokens (“token sequence”) generated by STT processing module 730, and attempt to associate the token sequence with one or more “actionable intents” recognized by the digital assistant. An “actionable 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 depends on the assistant’s ability to infer the correct “actionable intent(s)” from the user request expressed in natural language.

[0177] 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 token sequence 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.

[0178] 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.

[0179] 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).

[0180] 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.

[0181] 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.

[0182] 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.”

[0183] In some examples, the “find media items” domain includes a super domain containing many actionable intent nodes associated with finding or obtaining media items. For example, the “find media items” domain includes the actionable intent nodes, such as “obtain media items having recent release data,” “obtain personalized recommendations for media items,” or “obtain information associated with media items.”

[0184] 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.

[0185] 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.”

[0186] 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.

[0187] Natural language processing module 732 receives the token sequence (e.g., a text string) from STT processing module 730, and determines what nodes are implicated by the words in the token sequence. In some examples, if a word or phrase in the token sequence 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.

[0188] 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.

[0189] 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, the entire disclosure of which is incorporated herein by reference.

[0190] 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.

[0191] In some examples, natural language processing module 732 passes the generated structured query (including any completed parameters) to task flow processing module 736 (“task flow processor”). Task flow processing module 736 is configured to receive the structured query 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.

[0192] 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.

[0193] 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.

[0194] 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.

[0195] 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.

[0196] 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.

[0197] 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.

[0198] 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.

[0199] 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, the entire disclosures of which are incorporated herein by reference.

4. Processes for Operating Digital Assistants for Media Exploration

[0200] FIGs. 8A-C illustrate process 800 for operating a digital assistant for media exploration, according to various examples. FIGs. 9A-B, 10, and 11 illustrate interactions user 901 operating a digital assistant on user device 903 for media exploration, according to various examples. Process 800 is performed, for example, using one or more electronic devices implementing a digital assistant. In some examples, the process is performed at a client-server system (e.g., system 100) implementing a digital assistant. In some examples, the process is performed at a user device (e.g., device 104, 200, 400, or 600). In process 800, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. Further, it should be recognized that in some examples, only a subset of the features described below with reference to FIGs. 8A-C are performed in process 800.

[0201] At block 802, a speech input is received (e.g., at I/O processing module 728 and via microphone 213) from a user. The speech input represents a request for one or more media items. For example, with reference to FIG. 9A, the speech input is “Hey Siri, play me some hip-hop music I’d like.” In another example shown in FIG. 10, the speech input is “Hey Siri, play me some music for barbecuing.” In yet another example shown in FIG. 11, the speech input is “Hey Siri, play me some pop music that just came out.” Other examples of speech input that represent a request for one or more media items include, “What should I listen to,” “Recommend some music,” “What’s for me today,” “Hey Siri, be my DJ,” “Spin me some tasty beats,” “Find a recommended playlist,” “Any good album I should play,” “Play me something I’d like,” “Any recommended workout music,” “Find recently released music,” “Hot new rock tracks please,” or the like.

[0202] At block 804, a determination is made (e.g., using natural language processing module 732) as to whether the speech input of block 802 corresponds to the user intent of obtaining personalized recommendations for media items. In particular, the determination includes determining the user intent (e.g., actionable intent) corresponding to the speech input. The user intent is determined in a similar manner as discussed above with reference to FIGs. 7A-C. In particular, the words or phrases in the speech input are parsed and compared to the words or phrases of a vocabulary index (e.g., vocabulary index 744). The words or phrases of the vocabulary index are associated with the various nodes (e.g., actionable intent nodes or domains) of an ontology (e.g., ontology 760) and thus, based on the comparison,

nodes corresponding to the words or phrases in the speech input are “triggered” or “activated.” The node having the highest confidence value among the activated nodes is selected. The determined user intent corresponding to the speech input of block 802 is thus the actionable intent corresponding to the selected node.

[0203] Determining whether the speech input corresponds to the user intent of obtaining personalized recommendations for media items is based on the selected actionable intent node. If the selected node has a corresponding actionable intent of obtaining personalized recommendations for media items, then the speech input is determined to correspond to the user intent of obtaining personalized recommendations for media items. Conversely, if the selected node has a corresponding actionable intent other than obtaining personalized recommendations for media items, then the speech input is determined to not correspond to the user intent of obtaining personalized recommendations for media items.

[0204] In some examples, determining whether the speech input corresponds to the user intent of obtaining personalized recommendations for media items includes determining whether the speech input includes one or more of a plurality of predetermined phrases. In particular, the vocabulary index includes a plurality of predetermined phrases that correspond to the actionable intent node of obtaining personalized recommendations for media items. The plurality of predetermined phrases include, for example, “Recommend me... [music],” “Be my DJ,” “Spin some tunes/beats,” “What should I play,” “Play me some [music] I’d like,” “Find some good [music] for...,” or the like. Based on the speech input containing one or more of these phrases, the speech input is mapped to the actionable intent of obtaining personalized recommendations for media items, and the speech input is determined to correspond to the user intent of obtaining personalized recommendations for media items. For example, in FIG. 9A, speech input 902 contains the phrase “play me some [music] I’d like,” which is one of the plurality of predetermined phrases corresponding to the actionable intent node of obtaining personalized recommendations for media items. Thus, in this example, speech input 902 is determined to correspond to the user intent of obtaining personalized recommendations for media items.

[0205] In some examples, determining whether the speech input corresponds to the user intent of obtaining personalized recommendations for media items includes determining whether a number of parameters defined in the speech input is less than a predetermined threshold number. In particular, the speech input is determined to correspond to the user

intent of obtaining personalized recommendations for media items if the number of parameters (e.g., media parameters) defined in the speech input is less than a predetermined threshold number. For example, the speech input “What should I play?” is a request related to playing music. However, the request is broad and vague as it does not define any media parameters, such as the desired artist, album, genre, or release date. In this example, the speech input is determined to correspond to the user intent of obtaining personalized recommendations for media items, because the number of media parameters defined in the speech input is less than a predetermined threshold number (e.g., one).

[0206] In some examples, determining whether the speech input corresponds to the user intent of obtaining personalized recommendations for media includes determining whether the speech input refers to the user. Specifically, the speech input is parsed to determine whether it contains words or phrases referring to the user (e.g., “me,” “for me,” “I,” “my,” etc.). For example, the following phrases are determined to contain words referring to the user: “What’s for me,” “Surprise me,” “Anything for me today?” In some examples, the determination is based on determining whether the speech input contains words or phrases referring to the user in combination with words or phrases related to media (e.g., “listen,” “music,” “play,” “tunes,” “DJ,” etc.). For example, the following phrases are determined to contain words referring to the user in combination with words or phrases related to media: “Recommend me some hip-hop,” “Be my DJ,” “What should I listen to,” “What do songs do you have for me,” or “Play me some tunes.” Thus, the speech input is determined to correspond to the user intent of obtaining personalized recommendations for media based on the speech inputs containing words or phrases that refer to the user.

[0207] In response to determining that the speech input corresponds to the user intent of obtaining personalized recommendations for media items, block 806 is performed. At block 806, at least one media item is obtained from a user-specific corpus of media items (e.g., using natural language processing module 732, task flow processing module 736, and/or service processing module 738). In some examples, the at least one media item includes a song, album, video, movie, or playlist. The user-specific corpus of media items is a personalized corpus of media items that is specific to the user. In particular, the user-specific corpus of media items is generated based on data associated with the user. A more detailed description of the user-specific corpus of media items is provided below with reference to block 810. Obtaining the at least one media item from the user-specific corpus of media

items at block 806 includes performing one or more of blocks 808-816, described below. Blocks 808-816 are performed, for example, using one or more of natural language processing module 732, task flow processing module 736, and/or service processing module 738.

[0208] At block 808, media parameters defined in the speech input are determined (e.g., using natural language processing module 732). A structured query corresponding to the actionable intent of obtaining personalized recommendations for media items is then generated with the defined media parameters. In particular, the vocabulary index (e.g., vocabulary index 744) includes words or phrases corresponding to each of a plurality of media parameters. Media parameters defined in the speech input are thus determined by comparing the words or phrases of the speech input with the words or phrases in the vocabulary index. For example, the vocabulary index includes words or phrases associated with the media parameter {genre}. The words or phrases include, for example, “hip-hop,” “R&B,” “jazz,” “punk,” “rock,” “pop,” “classical,” “bluegrass,” or the like. In the example of FIG. 9A, speech input 902 is determined to define the media parameter {genre} = “hip-hop” based on detecting the phrase “hip-hop” in speech input 902. Thus, in this example, the structured query corresponding to the actionable intent of obtaining personalized recommendations for media items are generated to include the media parameter {genre} = “hip-hop.”

[0209] Another media parameter that can be determined from the speech input is {release date}. The media parameter {release date} refers to the release dates of the media items the user is interested in. The release date is, for example, a specific date or a range of dates. Words or phrase associated with the media parameter {release date} include, for example, “seventies,” “eighties,” “nineties,” “the last ten years,” “2008,” “after March 2016,” or the like. In one example, based on the word “eighties” in the speech input “Play some eighties tunes for me,” the speech input is determined to define the media parameter {release date} = “1980-1989.” Thus, in this example, the structured query corresponding to the actionable intent of obtaining personalized recommendations for media items is generated to include the media parameter {release date} = “1980-1989.”

[0210] In some examples, based on the context of the speech input, dates or time periods in the speech input are interpreted as defining a subgenre rather than defining a release date. For example, based on the phrase “’70s” in the speech input “Play me some ’70s punk,” the

speech input is determined to define the time period “1970-1979.” In response to determining that the speech input defines this time period, a determination is made as to whether the speech input defines a genre in association with this time period. In the present example, the speech input includes the phrase “punk,” which corresponds to the media parameter {genre}. Since the time period “’70s” modifies the genre “punk” in the speech input, it is determined that the speech input defines the genre “punk” in association with the time period “’70s.” In response to determining that the speech input defines a genre in association with the defined time period, a subgenre is determined based on the defined time period and the defined genre. For example, based on the defined time period “’70s” and the defined genre “punk,” the subgenre “’70s punk” is determined. Thus, in this example, the structured query generated for the actionable intent of obtaining personalized recommendations for media items includes the media parameter {subgenre} = “’70s punk.” Notably, rather than interpreting the defined time period as the media parameter {release date}, the defined time period is interpreted more accurately as part of the media parameter {subgenre}. In this way, the speech input is interpreted to more accurately reflect the user’s actual intent and thus, more relevant media items are provided to the user. For example, the at least one media item is obtained at block 806 based on the determined subgenre “’70s punk,” which includes media items with release dates outside of the time period 1970-1979. Specifically, each media item in the at least one media item includes metadata that indicates the subgenre of “’70s punk.”

[0211] Other media parameters that are determined as being defined in the speech input includes, for example, {activity}, {mood}, {occasion}, {editorial list}, {political preference}, or {technical proficiency}. Each of these media parameters is described below in turn. For example, the media parameter {activity} refers to an activity performed by the user and is associated with words or phrases, such as “working out,” “exercising,” “barbecuing,” “sleeping,” “driving,” “studying,” “painting,” or the like. In one example, based on the word “studying” in the speech input “Find some music that’s good for studying,” the speech input is determined to define the media parameter {activity} = “studying.” In another example shown in FIG. 10, the phrase “barbecuing” is detected in speech input 1002 as corresponding to the media parameter {activity}. Thus, in this example, speech input 1002 is determined as defining the media parameter {activity} = “barbecuing.”

[0212] The media parameter {mood} refers to a feeling or state of mind of the user and is associated with words or phrases, such as “happy,” “sad,” “angry,” “relaxing,” “powerful,” “excited,” “romantic,” or the like. In one example, based on the word “happy” in the speech input “Recommend me some happy music,” the speech input is determined to define the media parameter {mood} = “happy.”

[0213] The media parameter {occasion} refers to an occasion associated with a specific time period and is associated with words or phrases, such as “Christmas,” “birthday,” “summer,” “winter,” “Halloween,” “New Years,” “Easter,” or the like. In one example, based on the word “Christmas” in the speech input “Play some Christmas music,” the speech input is determined to define the media parameter {occasion} = “Christmas.”

[0214] The media parameter {editorial list} refers to a predetermined list of media items compiled by a media establishment, such as Rolling Stones magazine, Billboard magazine, Shazam, or the like. Exemplary editorial lists include, for example, the Billboard Hot 100, Billboard Hit Parade, Billboard 200, American top 40, Rolling Stone’s 500 Greatest Songs of All Time, Rolling Stone’s 500 Greatest Albums of All Time, Rolling Stones 100 Greatest Artists, or the like. The media parameter {editorial list} is associated with words or phrases that correspond to these lists. For example, based on the phrase “Billboard Hot 100” in the speech input “Play me songs from the Billboard Hot 100,” the speech input is determined to define the media parameter {editorial list} = “Billboard Hot 100.”

[0215] The media parameter {political preference} refers to a political preference of the user and is associated with words or phrases, such as “conservative,” “liberal,” “right wing,” “right leaning,” “left wing,” “left leaning,” or the like. In one example, based on the phrase “conservative” in the speech input “Find me conservative news,” the speech input is determined to define the media parameter {political preference} = “conservative.” In this example, the candidate media items determined at block 812 are more likely associated with conservative media sources (e.g., Fox news, drudge report, etc.) than liberal media sources (e.g., Huffington Post, New York Times, etc.).

[0216] The media parameter {technical proficiency} refers to how proficient the user is with technical subject matter. This is relevant when making requests for documentaries that discuss technical subject matter. In particular, the media parameter {technical proficiency} is associated with words and phrases, such as “very technical,” “layman,” “scientific,” “easily

understandable,” “simple,” “advanced,” or the like. In one example, based on the phrase “very technical” in the speech input “Find me some very technical documentaries on space crafts,” the speech input is determined to define the media parameter {technical proficiency} = “high.” In some examples, the media parameter {technical proficiency} is inferred based on a user’s familiarity with the requested subject matter. In particular, if the user frequently requests documentaries on space crafts (e.g., based on user request logs) or if the user has a large collection of documentaries on space crafts in his/her personal media library, then it can be determined that the user is very familiar with the subject matter of space crafts and thus, in this example, the media parameter {technical proficiency} is inferred to be “high.” ??

[0217] At block 810, a user-specific corpus of media items is determined. Determining the user-specific corpus of media items includes obtaining user identification information associated with the user. The user identification information contains, for example, user account log-in information or user password information for accessing a corresponding user-specific corpus of media items. The user identification information is then used to identify and access the appropriate user-specific corpus of media items from among a plurality of user-specific corpuses of media items to obtain the at least one media item.

[0218] In some examples, the user device at which the speech input of block 802 was received is associated with a unique user profile (e.g., stored in user data 748) that contains the user identification information. The user identification information is thus retrieved at block 810 based on the user profile associated with the user device. The corresponding user-specific corpus of media items is thus identified based on the retrieved user identification information.

[0219] In some examples, the user identification information is retrieved upon verifying the identity of the user. In particular, the identity of the user is verified by performing speaker identification using the speech input of block 802. Speaker identification is performed, for example, by comparing a voice print generated from the speech input of block 802 with a reference voice print associated with a specific user. The identity of the user is verified if the voice print generated from the speech input of block 802 is determined to match the reference voice print beyond a threshold confidence value. It should be recognized that other methods of identity verification can be implemented, such as fingerprinting authentication, passcode verification, or the like. Upon successfully verifying the identity of the user, the user identification information corresponding to the verified identity of the user

is retrieved (e.g., from the user profile). The user identification information is then used to identify and access the corresponding user-specific corpus of media items. Based on the determined identity of the user, the respective user-specific corpus of media items is determined from the plurality of user-specific corpora of media items.

[0220] In some examples, the user-specific corpus of media items is stored on a remote server separate from the user device. For example, the user-specific corpus of media items is stored as part of a media service (e.g., media service(s) 120-1) that provides media items. The user identification information is required to access the user-specific corpus of media items. In some examples, an encrypted token containing the user identification information is generated at the user device and sent to the media service. The media service then decrypts the token and utilizes the user identification information from the decrypted token to access the corresponding user-specific corpus of media items to obtain the at least one media item.

[0221] In some examples, the user-specific corpus of media items is customized to the media preferences of a specific user. For example, previous media-related data associated with the user is utilized to generate the user-specific corpus of media items. In particular, the user-specific corpus of media items is generated based on the media items previously selected, requested, or rejected by the user. For example, if the user is determined to frequently request, browse, select, or play media items having certain media parameters (e.g., {genre} = “pop” or {artist} = “Katy Perry”), then the user-specific corpus of media items is generated to favor media items having those parameters. Similarly, if the user is determined to consistently reject recommended media items having certain other parameters (e.g., {mood} = sad), then the user-specific corpus of media items is generated to disfavor media items having those other parameters.

[0222] In some examples, the user-specific corpus of media items is generated based on information from the user’s profile. The user profile contains information that characterizes the user, such as the country associated with the user, the user’s spoken languages, the user’s age, or the activities frequently engaged by the user. Based on this information, the user-specific corpus of media items is generated to favor media items having media parameters that complement this information. For example, if the user profile indicates that the user primarily speaks English and is twenty years old, then the user-specific corpus of media items is generated to favor media items that are spoken or sung in English and have a recent (e.g., last 5 years) release date.

[0223] Further, in some examples, the user-specific corpus of media items is generated based on a personal library of media items associated with the user. The personal library of media items include media items (e.g., songs, movies, etc.) that were acquired by the user. The personal library of media items is stored on the user device and/or stored on a remote server in association with the user's account. The user-specific corpus of media items is generated to favor media items having media parameters that are similar to those in the user's personal library of media items. For example, if the user's personal library of media items contains many albums by the artist Katy Perry, then the user-specific corpus is generated to favor media items associated with the artist Katy Perry and/or artists similar to Katy Perry, such as Avril Lavigne.

[0224] In some examples, the user-specific corpus of media items is generated such that each media item in the user-specific corpus of media items includes metadata that indicates a plurality of media parameters corresponding to the respective media item. In particular, the metadata for each media item defines any of the media parameters discussed above, such as {artist}, {genre}, {subgenre}, {release date}, {activity}, {mood}, {occasion}, {editorial list}, {political preference}, or {technical proficiency}. The metadata is used to recommend suitable media items to the user based on the media parameters defined in the user's speech input. For example, the user-specific corpus of media items include the instrumental song "Chariots of Fire" having metadata that indicates the following media parameters: {title} = "Chariots of Fire", {genre} = "soundtrack; instrumental," {composer} = Vangelis, {release date} = "March 1981," {activity} = "running," and {mood} = "inspirational." Thus, if the speech input received at block 802 were "Play me some inspirational instrumental music good for running," then based on the media parameters defined in the speech input (i.e., {genre} = "instrumental," {activity} = "running," and {mood} = "inspirational") the song "Chariots of Fire" is a candidate media item that is identified in the user-specific corpus of media items and recommended to the user.

[0225] In some example, the metadata of the media items in the user-specific corpus of media items is intelligently generated based on analyzing specific characteristics associated with the media items. In particular, the music tempo (e.g., beats per minute) of each media item is determined by analyzing the audio data of the media items. Based on the determined music tempo, the {activity} media parameter is determined for media items in the user-specific corpus of media items. For example, media items having a faster music tempo are

associated with more lively activities, such as working out, hiking, or the like. Conversely, media items having a slower music tempo are associated with more passive activities, such as sleeping, meditating, or the like. The relevant {activity} media parameter determined based on music tempo is thus included in the metadata of the respective media items.

[0226] Additionally, in some examples, the {mood} media parameter for media items in the user-specific corpus of media items is determined based on the music key of each media item. For example, the audio data of each media item is analyzed to determine the music key (e.g., C major, G major, A minor, etc.) associated with the audio data. Media items having a major music key are associated with more positive and happy moods, such as “happy,” “upbeat,” “cheerful,” “excited,” or the like, whereas media items having a minor music key are associated with more somber moods, such as “sad,” “mournful,” or the like.

[0227] At block 812, a plurality of candidate media items are determined from the user-specific corpus of media items based on the determined media parameters of block 808. For example, a search is performed using the determined media parameters of block 808 to identify candidate media items in the user-specific corpus of media items having metadata that include the determined media parameters of block 808. For example, returning to FIG. 9A, the media parameter {genre} = “hip-hop” is determined at block 808 to be defined in speech input 902. In response to determining that the speech input defines the media parameter {genre} = “hip-hop,” the user-specific corpus of media items can be searched to identify media items having metadata that include the media parameter {genre} = “hip-hop.” For example, media items such as “Topsy” by J-Kwon, “99 Problems” by Jay-Z, and “Over” by Drake each have metadata that include the media parameter {genre} = “hip-hop.” Thus, in this example, the plurality of candidate media items determined from the user-specific corpus of media items include these media items.

[0228] In another example shown in FIG. 10, user 901 provides speech input 1002 “Hey Siri, play me some music for barbecuing.” In this example, the speech input is determined at block 808 to define the media parameter {activity} = “barbecuing.” In response to determining that the speech input defines the media parameter {activity} = “barbecuing,” the user-specific corpus of media items is searched to identify media items having metadata that include the media parameter {activity} = “barbecuing.” For example, media items such as “She Moves in Her Own Way” by The Kooks, “Hot n Cold” by Katy Perry, and “Fun Fun Fun” by The Beach Boys each has metadata that include the media parameter {activity} =

“barbecuing.” Thus, in this example, the plurality of candidate media items determined from the user-specific corpus of media items include these media items.

[0229] Although the examples of FIGs. 9A-B and 10 are described with respect to specific media parameters, it should be recognized that the plurality of candidate media items are determined from the user-specific corpus of media items based on any media parameter(s) defined in the speech input of block 902. For example, in addition to the media parameters {genre} and {barbecue} described in the examples of FIGs. 9A-B and 10, the media parameters include, {artist}, {media type}, {subgenre}, {release date}, {mood}, {occasion}, {editorial list}, {political preference}, {technical proficiency}, or the like.

[0230] At block 814, the plurality of candidate media items of block 812 are ranked using a user-specific media ranking model. User-specific media ranking model is stored, for example, in user data 748 or data & models 116. Using the user-specific media ranking model, a user-specific ranking score is generated for each of the plurality of candidate media items. The plurality of candidate media items are thus ranked based on the user-specific ranking scores. The user-specific ranking score represents a likelihood that the user will accept the candidate media item given the media parameters associated with the candidate media item. The user-specific media ranking model is a statistical machine-learned model (e.g., neural network model, Bayesian model, etc.) that is trained using user-specific data, such as information from the user profile, previous media-related inputs from the user, or media items associated with the user. Further, the user-specific media ranking model is continuously updated based on subsequently received user-specific data. For example, the user-specific media ranking model is updated based on the speech input of block 802 or any speech contained in the audio input of block 824, described below.

[0231] Information from the user profile includes the user’s age, ethnicity, location, profession, or the like. This information is used to generate the user-specific media ranking model. For example, if information from the user profile indicates that the user is a scientist living in the conservative state of Idaho, the user-specific media ranking model is trained to generate more favorable scores for media items that are associated with a higher technical proficiency or a more conservative political orientation.

[0232] Previous media-related inputs from the user are used to generate the user-specific media ranking model. In particular, previous media-related inputs from the user include

media-related requests, selections, or rejections received prior to receiving the speech input of block 802. For example, if previous media-related requests from the user indicate that the user typically requests for pop music and rejects rap music, then based on this previous media-related input, the user-specific media ranking model is trained to generate more favorable ranking scores for pop music and less favorable ranking scores for rap music. In another example, previous media-related input indicates that, when the user browses through online music stores, the user frequently views music items with release dates in the 1970s. Based on this determination, the user-specific media ranking model is trained to generate more favorable scores for media items having release dates that are in the 1970s.

[0233] Media items associated with the user include media items found in the user's personal media library. In some examples, the media items in the user's personal media library are used to generate the user-specific media ranking model. In particular, the user-specific media ranking model is trained to favor the media items having media parameters similar to those of the media items in the user's personal media library. For example, based on the user's personal media library having many albums by Jay-Z, the user-specific media ranking model is trained to generate more favorable scores for media items related to the artist Jay-Z or artists similar to Jay-Z.

[0234] In some examples, the ranking of block 814 is performed (additionally or alternatively) using a general media ranking model. In particular, a general ranking score is generated for each of the plurality of candidate media items using the general media ranking model. The plurality of candidate media items are thus ranked based on the general ranking scores. The general media ranking model is similar to the user-specific media ranking model except that the general media ranking model is trained using media-related data from a large population of users rather than from one specific user. The general ranking scores represent general popularity ratings of the media items. In particular, the general media ranking model generates more favorable ranking scores for media items that are most frequently requested, viewed, or selected by the large population of users.

[0235] It should be recognized that the ranking of block 814 is, in some examples, performed based on a combination of the user-specific ranking score from the user-specific media ranking model and the general ranking score from the general media ranking model. For example, the scores are interpolated to generate a combined ranking score for each candidate media item. The plurality of candidate media items are then ranked based on the

combined ranking scores. Further, it should be recognized that in some examples, the general media ranking model is integrated with the user-specific ranking model. For example, the user-specific ranking model is generated using media-related data from a large population of users, but adjusted to favor user preferences indicated in user-specific data.

[0236] At block 816, the at least one media item is selected from the plurality of candidate media items based on the ranking of block 814. For example, the at least one media item includes the highest ranked candidate media item or the highest N ranked candidate media items among the plurality of candidate media items, wherein N is an integer greater than zero. The at least one media item obtained at block 806 is the selected at least one media item of block 816. The selected at least one media item is retrieved from the user-specific corpus of media items (e.g., at media service(s) 120-1) and provided to the user at block 818.

[0237] In some examples, the at least one media item is selected based on an inferred degree of familiarity of the user with one or more media parameters associated with the at least one media item. For example, the speech input received at block 802 is “Play me some Michael Jackson songs.” In this example, the speech input is determined at block 808 to define the media parameter {artist} = “Michael Jackson.” Based on this determination, a plurality of candidate Michael Jackson songs are identified at block 812 from the user-specific corpus of media items. The plurality of candidate Michael Jackson songs are ranked at block 814 based on general popularity (e.g., according to the general media ranking model) and/or based on the user-specific media ranking model. A degree of familiarity of the user with the artist “Michael Jackson” is determined. The determination is made based on user-specific data associated with the artist “Michael Jackson.” For example, a high degree of familiarity of the user with the artist “Michael Jackson” is determined based on previous media-related input from the user indicating that the user frequently browses, purchases, listens to, and/or requests songs by Michael Jackson or based on the user’s personal media library containing a large number of songs by Michael Jackson. Conversely, a low degree of familiarity of the user with the artist “Michael Jackson” is determined based on media-related input from the user indicating that the user infrequently browses, purchases, listens to, and/or requests songs by Michael Jackson or based on the user’s personal media library containing very few songs by Michael Jackson. The songs selected at block 816 from the plurality of candidate Michael Jackson songs are based on the determined degree of familiarity. For

example, if the degree of familiarity of the user with the artist “Michael Jackson” is determined to be low, then the most popular or the highest ranked candidate Michael Jackson songs are selected at block 814. In particular, the highest ranked N candidate Michael Jackson songs from the plurality of candidate Michael Jackson songs are selected to play as a playlist. In contrast, if the degree of familiarity of the user with the artist “Michael Jackson” is determined to be high, then a combination of popular (e.g., higher ranked) and less popular (e.g., lower ranked) candidate Michael Jackson songs are selected to play as a playlist. In particular, a greater proportion of less popular candidate Michael Jackson songs is selected based on a higher degree of familiarity of the user with the artist “Michael Jackson.” This is desirable because a user who is very familiar with the artist Michael Jackson would likely already be familiar with the most popular Michael Jackson songs. Such a user would enjoy listening to a combination of Michael Jackson songs that includes popular, highly commercialized songs and less popular, less commercialized (e.g., “deep cuts”) songs. Thus, in this example, the average popularity rating of the Michael Jackson songs selected at block 816 is based on the determined degree of familiarity of the user with the artist “Michael Jackson.”

[0238] It should be recognized that in some examples, the degree of familiarity of the user with the one or more media parameters associated with the at least one media item is directly factored into the user-specific media ranking model. For example, based on a determination that the user is very familiar with the artist “Michael Jackson,” the user-specific media ranking model is configured to generate higher ranking scores for some less popular Michael Jackson songs. In this way, the highest ranked N candidate Michael Jackson songs include a mix of highly commercialized popular Michael Jackson songs and lesser known Michael Jackson songs. In these examples, the selected at least one media item includes the highest ranked N candidate Michael Jackson songs.

[0239] Although block 806 is described above as being performed using a user-specific corpus of media items, it should be recognized and in other examples, other corpuses of media items can be used in lieu of the user-specific corpus of media items. For instance, in some examples, the at least one media item is obtained from a general (user-independent) corpus of media items or a corpus of media items generated based on one or more specific media parameters.

[0240] At block 818, the at least one media item is provided. In particular, the at least one media item is provided at the user device. In some examples, the at least one media item is played at the user device (e.g., using speaker 211). In other examples, the at least one media item is displayed on the user device (e.g., on touchscreen 212) for the user to view and/or select. In yet other examples, the at least one media item is provided in a spoken response to the user (e.g., using speaker 211).

[0241] With reference back to the example shown in FIG. 9A, the candidate hip-hop media items determined at block 812 are ranked at block 814 using the user-specific media ranking model. In particular, the candidate hip-hop media items are ranked such that the candidate media item “99 Problems” by Jay-Z is highest ranked among the candidate hip-hop media items determined at block 812. Thus, in this example, the at least one media item selected at block 816 includes the media item “99 Problems” by Jay-Z and the media item is played to user 901 on user device 903.

[0242] With reference now to the example shown in FIG. 10, the “barbecue” candidate media items determined at block 812 is ranked at block 814 using the user-specific media ranking model. In this example, the candidate media item “She Moves in Her Own Way” by The Kooks is the highest ranked among the candidate media items determined at block 812. Thus, the at least one media item selected at block 816 includes the media item “She Moves in Her Own Way” by The Kooks and the media item is obtained and played on to user 901 on user device 903. It should be recognized that the selected at least one media item can include other media items. For example, the candidate media items “Hot n Cold” by Katy Perry and “Fun Fun Fun” by The Beach Boys are the second and third highest ranked media items among the candidate media items determined at block 812. The at least one media item selected at block 816 includes these media items. Thus, in these examples, the media items “Hot n Cold” by Katy Perry and “Fun Fun Fun” by The Beach Boys are played on user device after the media item “She Moves in Her Own Way” by The Kooks.

[0243] In some examples, process 800 enables the user to provide a follow-up request upon providing the at least one media item at block 818. For example, the user rejects the at least one media item provided at block 818 or request additional information related to the at least one media item. Blocks 820-826 describe aspects where a follow-up spoken request from the user is received and a response to the follow-up spoken request is provided.

[0244] At block 820, a determination is made as to whether a domain corresponding to the speech input is one of a plurality of predetermined domains. In particular, only certain predetermined domains are likely to elicit a follow-up request from the user. Thus, to enhance efficiency, the capability to receive follow-up spoken requests from the user is implemented only for certain predetermined domains. For example, the plurality of predetermined domains include domains associated with items having a large amount of metadata, such as the “find media items” domain or the “find restaurant” domain. Items having a large amount of metadata, such as media items and restaurant items frequently elicit follow-up requests from users. In response to determining that a domain corresponding to the speech input is one of a plurality of predetermined domains, audio input is received at block 820 (e.g., microphone 213 is activated). Conversely, in response to determining that a domain corresponding to the speech input is not one of a plurality of predetermined domains, process 800 forgoes receiving audio input at block 822 (e.g., microphone 213 is not activated).

[0245] At block 824, audio input is received. Specifically, the audio input is received upon providing the at least one media item at block 818. For example, with reference to FIG. 9A, user device 903 begins receiving audio input via a microphone of user device 903 once the media item “99 Problems” by Jay-Z begins playing on user device 903.

[0246] At block 826, a determination is made as to whether the audio input contains speech. The determination is made while receiving the audio input. In particular, as audio input is received, the audio input is analyzed to determine whether it contains acoustic features that correspond to those of speech. In particular, time domain (e.g., zero crossing rates, short-time energy, spectral energy, or spectral flatness) and/or frequency domain features (e.g., mel-frequency cepstral coefficients, linear predictive cepstral coefficients, or mel-frequency discrete wavelet coefficients) are extracted from the received audio input and compared to a human speech model to determine the likelihood that the audio input contains speech. The audio input is determined to contain speech if the likelihood is determined to exceed a predetermined value. Conversely, the audio input is determined not to contain speech if the likelihood is less than the predetermined value. In response to determining that the audio input does not contain speech, process 800 ceases to receive audio input at block 828 after a predetermined amount of time. For example, with reference to FIG. 9A, user

device 903 ceases to receive audio input after receiving a predetermined duration of audio input that is determined not to contain any speech.

[0247] In some examples, the predetermined amount of time is based on the degree of ambient noise detected in the audio input. In particular, block 826 includes determining the amount of ambient noise (e.g., background noise) in the audio input. The predetermined amount of time in which audio input not containing any speech is received at block 824 decreases based on detecting a high degree of ambient noise in the audio input. For example, if the amplitude of ambient noise in the audio input is determined not to exceed a predetermined threshold value, process 800 ceases to receive audio input at block 828 after a predetermined amount of time (e.g., 7 seconds). However, if the amplitude of ambient noise in the audio input is determined to exceed a predetermined threshold value, process 800 ceases to receive audio input at block 828 after a second predetermined amount of time (e.g., 4 seconds) that is less than the predetermined amount of time.

[0248] In response to determining that the audio input contains speech, block 830 is performed. At block 830, a determination is made as to whether the speech of the audio input corresponds to a same domain as the speech input. The determination includes determining the user intent corresponding to the speech of the audio input. The user intent is determined in a similar manner as described above in block 804. Determining the user intent corresponding to the speech of the audio input includes determine a domain corresponding to the speech of the audio input. A determination is then made as to whether the domain corresponding to the speech of the audio input is the same as the domain corresponding to the speech input of block 802. In response to determining that the speech of the audio input does not correspond to a same domain as the speech input, process 800 forgoes providing a response to the audio input at block 832. This is desirable to filter out babble noise. For example, with reference to FIG. 9A, speech input 902 corresponds to the “find media items” domain. If audio input is received while playing the song “99 Problems” by Jay-Z and the audio input contains babble noise that is unrelated to finding media items, then the babble noise is determined to be irrelevant to speech input 902 and no follow-up response would be provided to the user (block 832).

[0249] In response to determining that the speech of the audio input corresponds to a same domain as the speech input, block 834 is performed. At block 834, a response is provided in accordance with the user intent corresponding to the speech of the audio input.

The response is provided in a similar manner as described above with respect to FIGs. 7A-C. In particular, a structured query is generated based on the determined user intent corresponding to the speech of the audio input. One or more tasks corresponding to the user intent are then performed in accordance with the generated structured query. A response is provided based on the one or more performed tasks.

[0250] Blocks 820-834 are further described with reference to the examples of FIGs. 9A-B and 10. In FIG. 9A, while user device 903 is playing the obtained at least one media item “99 Problems” by Jay-Z (block 818), audio input containing second speech input 904 “Anything but Jay-Z!” is received from user 901 (block 824). A user intent corresponding to second speech input 904 is determined (block 830). Based on the phrase “Jay-Z” in second speech input 904 and the context of user device 903 playing the media item “99 Problems” by Jay-Z, second speech input 904 is determined to correspond to the same domain as speech input 902. Specifically, the domain corresponding to speech input 904 is determined to be the “find media items” domain. Further, a determination is made as to whether second speech input 904 corresponds to a rejection of the media item “99 Problems.” The determination is made based on the user intent corresponding to second speech input 904. In this example, based on interpreting the phrase “Anything but” in the context of the media item “99 Problems” being played, it is determined that second speech input 904 corresponds to the user intent of rejecting the media item “99 Problems” and obtaining alternative recommendations for media items. One or more tasks corresponding to the user intent are then performed (block 834). In particular, in response to determining that the second speech input corresponds to a rejection of the at least one media item, the candidate hip-hop media items previously determined and ranked (e.g., at blocks 812 and 814) are re-ranked based on the rejection. The re-ranking is similar to the ranking of block 814, except that unfavorable ranking scores are generated for media items having media parameter {artist} = Jay-Z. For example, the candidate hip-hop media items are re-ranked such that the candidate media item “Topsy” by J-Kwon is the highest ranked media item among the candidate hip-hop media items whereas the candidate media item “99 Problems” by Jay-Z is one of the lowest ranked media items among the candidate hip-hop media items. Based on this re-ranking, as shown in FIG. 9B, the media item “Topsy” by J-Kwon is obtained and played on user device 903. Furthermore, as discussed above, the user-specific media ranking model is continuously updated based on any subsequent speech input received from the user. Thus, in response to determining that the second speech input corresponds to a rejection of the at least one media

item, the user-specific media ranking model is updated in accordance with the rejection. For example, the user-specific media ranking model is updated to subsequently generate less favorable ranking score for candidate media items having media parameter {artist} = Jay-Z. Thus, digital assistant would be less likely to recommend media items associated with the artist Jay-Z when the user subsequently requests to obtain recommendations for media items.

[0251] Referring now to the example of FIG. 10, user 901 provides second speech input 1004 “When was this released?” Specifically, while user device 903 plays the media item “She Moves in Her Own Way” by The Kooks, second speech input 1004 is received from user 901 (block 824). A user intent corresponding to second speech input 1004 is determined (block 830). Based on the word “this” in second speech input 1004 and the context of user device 903 playing the media item “She Moves in Her Own Way,” second speech input 1004 is determined to correspond to the same domain as speech input 902. Specifically, the domain corresponding to speech input 904 is determined to be the “find media items” domain. Further, in this example, based on interpreting the words “this” and “released” in the context of the media item “She Moves in Her Own Way,” being played, it is determined (block 830) that second speech input 904 corresponds to the user intent of obtaining the release date associated with the media item “She Moves in Her Own Way.” In response to this determination, one or more tasks corresponding to the user intent is performed (block 834). In particular, the release date of the song “She Moves in Her Own Way” is retrieved (e.g., from media service(s) 120-1) and provided to the user (block 834). For example, as shown in FIG. 10, spoken response 1006 is provided at user device 903 to user 901 in accordance with the determined user intent. Specifically, spoken response 1006 indicates that the release date of the song “She Moves in Her Own Way” is “June 2006.” In some examples, the release data is additionally or alternatively displayed on user device 903 in response to second speech input 1004.

[0252] Returning back to block 804, in response to determining that the speech input of block 802 does not correspond to the user intent of obtaining personalized recommendations for media items, block 836 of FIG. 8C is performed. At block 836, a determination is made as to whether the speech input of block 802 corresponds to the user intent of obtaining media items having a recent release date. As discussed above, the user intent corresponding to the speech input is determined at block 804. The determination of block 836 is based on the actionable intent node selected in the ontology (e.g., ontology 760). If the selected node has a

corresponding actionable intent of obtaining media items having a recent release date, then the speech input is determined to correspond to the user intent of obtaining media items having a recent release date. Conversely, if the node has a corresponding actionable intent other than obtaining media items having a recent release date, then the speech input is determined to not correspond to the user intent of obtaining media items having a recent release date.

[0253] In some examples, determining whether the speech input corresponds to the user intent of obtaining media items having a recent release date includes determining whether the speech input includes one or more of a second plurality of predetermined phrases. In particular, the actionable intent node corresponding to the user intent of obtaining media items having a recent release date is associated with the second plurality of predetermined phrases. The second plurality of predetermined phrases are stored in a vocabulary index (vocabulary index 744) in association with the actionable intent node that corresponds to the user intent of obtaining media items. The second plurality of predetermined phrases include phrases, such as, “new music,” “recently released,” “latest releases,” “just came out,” or the like. Based on the speech input of block 802 containing one or more of the second plurality of predetermined phrases, the speech input is mapped to the actionable intent node that corresponds to the user intent of obtaining media items having a recent release date. Thus, the speech input of block 802 is determined to correspond to the user intent of obtaining media items having a recent release date. For example, with reference to FIG. 11, speech input 1102 “Hey Siri, play me some pop music that just came out” is received from user 901. Based on speech input 1102 containing the phrase “just came out,” the actionable intent node corresponding to the user intent of obtaining media items having a recent release date is selected. Thus, speech input 1102 is determined to correspond to the user intent of obtaining media items having a recent release date.

[0254] In response to determining that the speech input corresponds to the user intent of obtaining media items having a recent release date, block 838 is performed. Conversely, in response to determining that the speech input does not correspond to the user intent of obtaining media items having a recent release date, process 800 forgoes performing block 838. For example, as shown in FIG. 8C, process 800 ends in response to determining that the speech input does not correspond to the user intent of obtaining media items having a recent release date.

[0255] At block 838, at least one second media item from a second corpus of media items is obtained. Block 838 is similar to block 806, except that block 838 is performed using the second corpus of media items rather than the user-specific corpus of media items. Further, block 838 includes blocks similar to blocks 808-816, except again that the blocks are performed with respect to the second corpus of media items rather than the user-specific corpus of media items. The second corpus of media items is, for example, a general corpus of media items that is generated based on the release date of media items. In particular, each media item in the second corpus of media items has a release date that is within a predetermined time range of a current date. For example, the second corpus of media items includes only media items having a release date that is within three months of the current date. In some examples, the second corpus of media items is generated based on other factors, such as the popularity of each media item.

[0256] At block 840, the at least one second media item is provided. Block 840 is similar to block 818. In particular, the at least one second media item is provided at the user device. In some examples, the at least one media item is played at the user device. In other examples, the at least one media item is displayed on the user device (e.g., on touchscreen 212) for the user to view and/or select. In yet other examples, the at least one media item is provided in a spoken response to the user.

[0257] Blocks 838-840 are further described with reference to FIG. 11. For example, in response to determining that speech input 1102 corresponds to the user intent of obtaining media items having a recent release date, the digital assistant implemented on user device 903 obtains at least one second media item from a second corpus of media items. The second corpus of media items contain only media items having a release date that is within three months of the current date. In the present example, if the current date is June 1, 2016, then each media item in the second corpus of media items has a release date that is not earlier than March 1, 2016. Thus, the at least one second media item obtained from the second corpus of media items also has a release data that is not earlier than March 1, 2016. In this example, the at least one second media item includes the song “Dangerous Woman” by Ariana Grande, which has a release date of March 11, 2016. As shown, the song “Dangerous Woman” is obtained (e.g., from media service(s) 120-1) and played on user device 903 in response to speech input 1102.

5. Other Electronic Devices

[0258] FIG. 12 shows a functional block diagram of electronic device 1200 configured in accordance with the principles of the various described examples. The functional blocks of the device are optionally implemented by hardware, software, or a combination of hardware and software to carry out the principles of the various described examples. It is understood by persons of skill in the art that the functional blocks described in FIG. 12 are optionally combined or separated into sub-blocks to implement the principles of the various described examples. Therefore, the description herein optionally supports any possible combination, separation, or further definition of the functional blocks described herein.

[0259] As shown in FIG. 12, electronic device 1200 includes touch screen display unit 1202 configured to display a graphical user interface and to receive touch input from the user, audio input unit 1204 configured to receive audio input (e.g., speech input), speaker unit 1205 configured to output audio (e.g., speech and/or media content), and communication unit 1206 configured to transmit and receive information. Electronic device 1200 further includes processing unit 1208 coupled to touch screen display unit 1202, audio input unit 1204, and communication unit 1206. In some examples, processing unit 1208 includes receiving unit 1210, determining unit 1212, obtaining unit 1214, providing unit 1216, ranking unit 1218, updating unit 1220, ceasing unit 1222, forgoing unit 1224, and selecting unit 1226.

[0260] In accordance with some embodiments, processing unit 1208 is configured to receive (e.g., with receiving unit 1210 and via audio input unit 1204) from a user, speech input (e.g., speech input of block 802) representing a request for one or more media items. Processing unit 1208 is further configured to determine (e.g. with determining unit 1212) whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items, obtain (e.g., with obtaining unit 1214) at least one media item (e.g., at least one media item of block 806) from a user-specific corpus of media items. The user-specific corpus of media items (e.g., user-specific corpus of media items of block 806) is generated based on data associated with the user. Processing

unit 1208 is further configured to provide (e.g., with providing unit and using touch screen display unit 1202 and/or speaker unit 1205) the at least one media item (e.g., block 818).

[0261] In some examples, determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether a number of parameters defined in the speech input is less than a threshold number (e.g., block 804).

[0262] In some examples, determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether the speech input includes one of a plurality of phrases corresponding to the user intent of obtaining personalized recommendations for media (e.g., block 804).

[0263] In some examples, determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media comprises determining whether the speech input refers to the user (e.g., block 804).

[0264] In some examples, the user-specific corpus of media items (e.g., user-specific corpus of media items of block 806) is generated based on media items previously selected or requested by the user.

[0265] In some examples, the user-specific corpus of media items is generated based on media items previously rejected by the user (e.g., block 806).

[0266] In some examples, the user-specific corpus of media items is generated based on a personal library of media items associated with the user (e.g., block 806).

[0267] In some examples, processing unit 1208 is further configured to rank (e.g., with ranking unit 1218) a plurality of candidate media items from the user-specific corpus of media items using a user-specific media ranking model (e.g., block 814). The user-specific media ranking model is generated based on a plurality of previous media-related requests from the user. Obtaining the at least one media item includes selecting the at least one media item from the plurality of candidate media items based on the ranking (e.g., block 816).

[0268] In some examples, processing unit 1208 is further configured to, receive (e.g., with receiving unit 1210 and via audio input unit 1204) from the user, a second speech input. Processing unit 1208 is further configured to determine (e.g., with determining unit 1212)

whether the second speech input corresponds to a rejection of the at least one media item. Processing unit 1208 is further configured to, in response to determining that the second speech input corresponds to a rejection of the at least one media item, update (e.g., with updating unit 1220) the user-specific media ranking model in accordance with the rejection.

[0269] In some examples, processing unit 1208 is further configured to re-rank (e.g., with ranking unit 1218) the plurality of candidate media items from the user-specific corpus of media items based on the rejection of the at least one media item. Processing unit 1208 is further configured to select (e.g., with selecting unit 1226) at least one second media item from the plurality of candidate media items based on the re-ranking.

[0270] In some examples, the plurality of candidate media items are ranked based on a popularity rating of each media item of the plurality of candidate media items (e.g., block 814).

[0271] In some examples, each media item in the user-specific corpus of media items includes metadata that indicates an activity associated with the media item. The activity is associated with the media item based on a music tempo of the media item.

[0272] In some examples, each media item in the user-specific corpus of media items includes metadata that indicates a mood associated with the media item. The mood is associated with the media item based on a music key of the media item.

[0273] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the speech input defines an occasion associated with a time period (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input defines an occasion associated with a time period, obtain (e.g., with obtaining unit 1214) the at least one media item based the occasion, wherein the at least one media item includes metadata indicating the occasion (e.g., block 806).

[0274] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the speech input defines an editorial list associated with a media establishment (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input defines an editorial list associated with a media establishment, obtain (e.g., with obtaining unit) the at least one media item based on the editorial list associated with the media establishment (e.g., block 806). The at least one

media item includes metadata indicating the editorial list associated with the media establishment.

[0275] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the speech input defines a mood (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input defines a mood, obtain (e.g., with obtaining unit 1214) the at least one media item based on the mood, where the at least one media item includes metadata indicating the mood (e.g., block 806).

[0276] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the speech input defines an activity (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input defines an activity, obtain (e.g., with obtaining unit 1214) the at least one media item based on the activity, wherein the at least one media item includes metadata indicating the activity (e.g., block 806).

[0277] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the speech input defines a time period (e.g., block 804). Processing unit 1208 is further configured to, in response to determining that the speech input defines a time period, determine (e.g., with determining unit 1212) whether the speech input defines a genre in association with the time period. Processing unit 1208 is further configured to, in response to determining that the speech input defines a genre in association with the time period, determine (e.g., with determining unit 1212) a subgenre based on the time period and the genre. The at least one media item is obtained based on the subgenre, and the at least one media item includes metadata indicating the subgenre (e.g., block 806).

[0278] In some examples, the speech input defines a category of media items and obtaining the at least one media item includes obtaining a plurality of media items associated with the category of media items. Processing unit 1208 is further configured to determine (e.g., with determining unit 1212) a degree of familiarity of the user with the category of media items (e.g., degree of familiarity of block 816). An average popularity rating of the plurality of media items is based on the degree of familiarity of the user with the category of media items.

[0279] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) an identity of the user by performing speaker identification using the speech input. Processing unit 1208 is further configured to, based on the determined identity of the user, determine (e.g., with determining unit 1212) the user-specific corpus of media items from a plurality of user-specific corpuses of media items.

[0280] In some examples, obtaining the at least one media item comprises sending an encrypted token to a remote server. The encrypted token contains user identification information. The encrypted token is required to access the user-specific corpus of media items via the remote server.

[0281] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether a domain (e.g., domain of block 820) corresponding to the speech input is one of a plurality of predetermined domains. Processing unit 1208 is further configured to, in response to determining that a domain corresponding to the speech input is one of a plurality of predetermined domains, receive (e.g., with receiving unit 1210 and via audio input unit 1204) audio input (e.g., audio input of block 824) upon providing the at least one media item. Processing unit 1208 is further configured to determine (e.g., with determining unit 1212) whether the audio input contains speech (e.g., block 826). Processing unit 1208 is further configured to, in response to determining that the audio input does not contain speech, cease (e.g., with ceasing unit 1222) to receive audio input after a predetermined amount of time (e.g., block 828).

[0282] In some examples, processing unit 1208 is further configured to, in response to determining that the audio input contains speech, determine (e.g., with determining unit 1212) whether the speech of the audio input corresponds to a same domain as the speech input (e.g., block 830). Processing unit 1208 is further configured to, in response to determining that the speech of the audio input corresponds to a same domain as the speech input, determine (e.g., with determining unit 1212) a user intent (e.g., user intent of block 820) corresponding to the speech of the audio input. Processing unit 1208 is further configured to provide (with providing unit 1216) a response (e.g., response of block 834) to the audio input in accordance with the user intent corresponding to the speech of the audio input.

[0283] In some examples, processing unit 1208 is further configured to, in response to determining that the speech of the audio input does not correspond to a same domain as the speech input, forgo (e.g., with forgoing unit 1224) providing a response to the audio input (e.g., block 832).

[0284] In some examples, the predetermined amount of time is based on a degree of ambient noise detected in the audio input.

[0285] In some examples, providing the at least one media item comprises playing a media item. Processing unit 1208 is further configured to, while playing the media item, receive (e.g., with receiving unit 1210 and via audio input unit 1204) a third speech input (e.g., speech input in audio input of block 824). Processing unit 1208 is further configured to, based on the playing media item and the third speech input, determine (e.g., with determining unit 1212) a user intent (e.g., user intent of block 820) that corresponds to the third speech input. Processing unit 1208 is further configured to, provide (e.g., with providing unit 1216) a response (e.g., response of block 834) in accordance with the user intent that corresponds to the third speech input.

[0286] In some examples, processing unit 1208 is further configured to, in response to determining that the speech input does not correspond to a user intent of obtaining personalized recommendations for media items, determine (e.g., with determining unit 1212) whether the speech input corresponds to a user intent of obtaining media items having a recent release date (e.g., block 836). Processing unit 1208 is further configured to, in response to determining that the speech input corresponds to a user intent of obtaining media items having a recent release date, obtain (e.g., with obtaining unit 1214) at least one second media item (e.g., at least one second media item of block 838) from a second corpus of media items. Each media item in the second corpus of media items has a release date that is within a predetermined time range of a current date. Processing unit 1208 is further configured to provide (e.g., with providing unit 1216) the at least one second media item (e.g., block 840).

[0287] In some examples, determining whether the speech input corresponds to a user intent of obtaining media items having a recent release date comprises determining whether the speech input includes one of a second plurality of phrases corresponding to the user intent of obtaining media items having a recent release date (e.g., block 836).

[0288] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) a political preference associated with the user (e.g., block 814). The determining is based on previous media items requested or consumed by the user. The at least one media item is obtained based on the determined political preference.

[0289] In some examples, processing unit 1208 is further configured to determine (e.g., with determining unit 1212) a technical proficiency associated with the user (e.g., block 814). The determining is based on previous media items requested or consumed by the user. The at least one media item is obtained based on the determined technical proficiency.

[0290] The operations described above with reference to FIGs. 8A-C are optionally implemented by components depicted in FIGs. 1-4, 6A-B, and 7A-C. For example, the operations of process 800 may be implemented by one or more of operating system 718, applications module 724, I/O processing module 728, STT processing module 730, natural language processing module 732, vocabulary index 744, task flow processing module 736, service processing module 738, media service(s) 120-1, or processor(s) 220, 410, 704. 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.

[0291] 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.

[0292] 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.

[0293] 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.

[0294] 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.

[0295] Exemplary methods, non-transitory computer-readable storage media, systems, and electronic devices are set out in the following items:

1. A method for operating a digital assistant to explore media items, the method comprising:
 - at an electronic device with one or more processors and memory:
 - receiving, from a user, speech input representing a request for one or more media items;
 - determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items; and
 - in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items:
 - obtaining at least one media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with the user; and
 - providing the at least one media item.
2. The method of item 1, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether a number of parameters defined in the speech input is less than a threshold number.
3. The method of any of items 1-2, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether the speech input includes one of a plurality of phrases corresponding to the user intent of obtaining personalized recommendations for media items.
4. The method of any of items 1-3, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether the speech input refers to the user.

5. The method of any of items 1-4, wherein the user-specific corpus of media items is generated based on media items previously selected or requested by the user.
6. The method of any of items 1-5, wherein the user-specific corpus of media items is generated based on media items previously rejected by the user.
7. The method of any of items 1-6, wherein the user-specific corpus of media items is generated based on a personal library of media items associated with the user.
8. The method of any of items 1-7, further comprising:
ranking a plurality of candidate media items from the user-specific corpus of media items using a user-specific media ranking model, the user-specific media ranking model generated based on a plurality of previous media-related requests from the user, wherein obtaining the at least one media item includes selecting the at least one media item from the plurality of candidate media items based on the ranking.
9. The method of item 8, wherein the plurality of candidate media items are ranked based on a popularity rating of each media item of the plurality of candidate media items.
10. The method of any of items 8-9, further comprising:
receiving, from the user, a second speech input;
determining whether the second speech input corresponds to a rejection of the at least one media item; and
in response to determining that the second speech input corresponds to a rejection of the at least one media item:
updating the user-specific media ranking model in accordance with the rejection.
11. The method of item 10, further comprising:
re-ranking the plurality of candidate media items from the user-specific corpus of media items based on the rejection of the at least one media item; and

selecting at least one second media item from the plurality of candidate media items based on the re-ranking.

12. The method of any of items 1-11, wherein each media item in the user-specific corpus of media items includes metadata that indicates an activity associated with the media item, and wherein the activity is associated with the media item based on a music tempo of the media item.

13. The method of any of items 1-12, wherein each media item in the user-specific corpus of media items includes metadata that indicates a mood associated with the media item, and wherein the mood is associated with the media item based on a music key of the media item.

14. The method of any of items 1-13, further comprising:
determining whether the speech input defines an occasion associated with a time period; and
in response to determining that the speech input defines an occasion associated with a time period, obtaining the at least one media item based the occasion, wherein the at least one media item includes metadata indicating the occasion.

15. The method of any of items 1-14, further comprising:
determining whether the speech input defines an editorial list associated with a media establishment; and
in response to determining that the speech input defines an editorial list associated with a media establishment, obtaining the at least one media item based on the editorial list associated with the media establishment, wherein the at least one media item includes metadata indicating the editorial list associated with the media establishment.

16. The method of any of items 1-15, further comprising:
determining whether the speech input defines a mood; and
in response to determining that the speech input defines a mood, obtaining the at least one media item based on the mood, wherein the at least one media item includes metadata indicating the mood.

17. The method of any of items 1-16, further comprising:
determining whether the speech input defines an activity; and
in response to determining that the speech input defines an activity, obtaining the at least one media item based on the activity, wherein the at least one media item includes metadata indicating the activity.
18. The method of any of items 1-17, further comprising:
determining whether the speech input defines a time period;
in response to determining that the speech input defines a time period, determining whether the speech input defines a genre in association with the time period; and
in response to determining that the speech input defines a genre in association with the time period, determining a subgenre based on the time period and the genre, wherein the at least one media item is obtained based on the subgenre, and wherein the at least one media item includes metadata indicating the subgenre.
19. The method of any of items 1-18, wherein the speech input defines a category of media items, wherein obtaining the at least one media item includes obtaining a plurality of media items associated with the category of media items, and further comprising:
determining a degree of familiarity of the user with the category of media items, wherein an average popularity rating of the plurality of media items is based on the degree of familiarity of the user with the category of media items.
20. The method of any of items 1-19, further comprising:
determining an identity of the user by performing speaker identification using the speech input; and
based on the determined identity of the user, determining the user-specific corpus of media items from a plurality of user-specific corpuses of media items.
21. The method of any of items 1-20, wherein obtaining the at least one media item comprises sending an encrypted token to a remote server, the encrypted token

containing user identification information, and wherein the encrypted token is required to access the user-specific corpus of media items via the remote server.

22. The method of any of items 1-21, further comprising:
 - determining whether a domain corresponding to the speech input is one of a plurality of predetermined domains;
 - in response to determining that a domain corresponding to the speech input is one of a plurality of predetermined domains:
 - upon providing the at least one media item, receiving audio input;
 - determining whether the audio input contains speech; and
 - in response to determining that the audio input does not contain speech, ceasing to receive audio input after a predetermined amount of time.
23. The method of item 22, further comprising:
 - in response to determining that the audio input contains speech:
 - determining whether the speech of the audio input corresponds to a same domain as the speech input;
 - in response to determining that the speech of the audio input corresponds to a same domain as the speech input:
 - determining a user intent corresponding to the speech of the audio input; and
 - providing a response to the audio input in accordance with the user intent corresponding to the speech of the audio input.
24. The method of item 23, further comprising:
 - in response to determining that the speech of the audio input does not correspond to a same domain as the speech input:
 - forgo providing a response to the audio input.
25. The method of any of items 22-24, wherein the predetermined amount of time is based on a degree of ambient noise detected in the audio input.

26. The method of any of items 1-25, wherein providing the at least one media item comprises playing a media item of the at least one media item, and further comprising:
- while playing the media item, receiving a third speech input;
 - determining, based on the playing media item and the third speech input, a user intent corresponding to the third speech input; and
 - providing a response in accordance with the user intent that corresponds to the third speech input.
27. The method of any of items 1-26, further comprising:
- in response to determining that the speech input does not correspond to a user intent of obtaining personalized recommendations for media items:
 - determining whether the speech input corresponds to a user intent of obtaining media items having a recent release date; and
 - in response to determining that the speech input corresponds to a user intent of obtaining media items having a recent release date:
 - obtaining at least one second media item from a second corpus of media items, wherein each media item in the second corpus of media items has a release date that is within a predetermined time range of a current date; and
 - providing the at least one second media item.
28. The method of any of items 1-27, wherein determining whether the speech input corresponds to a user intent of obtaining media items having a recent release date comprises determining whether the speech input includes one of a second plurality of phrases corresponding to the user intent of obtaining media items having a recent release date.
29. The method of any of items 1-28, further comprising:
- determining a political preference associated with the user, the determining based on previous media items requested or consumed by the user, wherein the at least one media item is obtained based on the determined political preference.
30. The method of any of items 1-29, further comprising:

determining a technical proficiency associated with the user, the determining based on previous media items requested or consumed by the user, wherein the at least one media item is obtained based on the determined technical proficiency.

31. A non-transitory computer-readable medium having instructions stored thereon, the instructions, when executed by one or more processors, cause the one or more processors to perform any of the methods of items 1-30.

32. A system comprising:
one or more processors; and
memory having instructions stored thereon, the instructions, when executed by the one or more processors, cause the one or more processors to perform any of the methods of items 1-30.

33. An apparatus comprising means for performing the method of any of items 1-30.

34. A non-transitory computer-readable medium having instructions stored thereon, the instructions, when executed by one or more processors, cause the one or more processors to:
receive, from a user, speech input representing a request for one or more media items;
determine whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items; and
in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items:
obtain at least one media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with the user; and
provide the at least one media item.

35. A system comprising:
one or more processors; and

memory having instructions stored thereon, the instructions, when executed by the one or more processors, cause the one or more processors to:

receive, from a user, speech input representing a request for one or more media items;

determine whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items; and

in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items:

obtain at least one media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with the user; and
provide the at least one media item.

36. An electronic device comprising:

an audio input unit;

a speaker unit; and

a processing unit coupled to the audio input unit and the speaker unit, the processing unit configured to:

receive, from a user, speech input representing a request for one or more media items;

determine whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items; and

in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items:

obtain at least one media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with the user; and
provide the at least one media item.

[0296] 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.

[0297] 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.

[0298] 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.

[0299] 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.

[0300] 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.

[0301] 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.

[0302] 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.

WHAT IS CLAIMED IS:

1. A method for operating a digital assistant to explore media items, the method comprising:
 - at an electronic device with one or more processors and memory:
 - receiving, from a user, speech input representing a request for one or more media items;
 - determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items; and
 - in response to determining that the speech input corresponds to a user intent of obtaining personalized recommendations for media items:
 - obtaining at least one media item from a user-specific corpus of media items, the user-specific corpus of media items generated based on data associated with the user; and
 - providing the at least one media item.
2. The method of claim 1, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether a number of parameters defined in the speech input is less than a threshold number.
3. The method of any of claims 1-2, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether the speech input includes one of a plurality of phrases corresponding to the user intent of obtaining personalized recommendations for media items.
4. The method of any of claims 1-3, wherein determining whether the speech input corresponds to a user intent of obtaining personalized recommendations for media items comprises determining whether the speech input refers to the user.
5. The method of any of claims 1-4, wherein the user-specific corpus of media items is generated based on media items previously selected or requested by the user.
6. The method of any of claims 1-5, wherein the user-specific corpus of media items is generated based on media items previously rejected by the user.

7. The method of any of claims 1-6, wherein the user-specific corpus of media items is generated based on a personal library of media items associated with the user.
8. The method of any of claims 1-7, further comprising:
 - ranking a plurality of candidate media items from the user-specific corpus of media items using a user-specific media ranking model, the user-specific media ranking model generated based on a plurality of previous media-related requests from the user, wherein obtaining the at least one media item includes selecting the at least one media item from the plurality of candidate media items based on the ranking.
9. The method of claim 8, wherein the plurality of candidate media items are ranked based on a popularity rating of each media item of the plurality of candidate media items.
10. The method of any of claims 8-9, further comprising:
 - receiving, from the user, a second speech input;
 - determining whether the second speech input corresponds to a rejection of the at least one media item; and
 - in response to determining that the second speech input corresponds to a rejection of the at least one media item:
 - updating the user-specific media ranking model in accordance with the rejection.
11. The method of claim 10, further comprising:
 - re-ranking the plurality of candidate media items from the user-specific corpus of media items based on the rejection of the at least one media item; and
 - selecting at least one second media item from the plurality of candidate media items based on the re-ranking.
12. The method of any of claims 1-11, wherein each media item in the user-specific corpus of media items includes metadata that indicates an activity associated with the media item, and wherein the activity is associated with the media item based on a music tempo of the media item.

13. The method of any of claims 1-12, wherein each media item in the user-specific corpus of media items includes metadata that indicates a mood associated with the media item, and wherein the mood is associated with the media item based on a music key of the media item.
14. The method of any of claims 1-13, further comprising:
 - determining whether the speech input defines an occasion associated with a time period; and
 - in response to determining that the speech input defines an occasion associated with a time period, obtaining the at least one media item based the occasion, wherein the at least one media item includes metadata indicating the occasion.
15. The method of any of claims 1-14, further comprising:
 - determining whether the speech input defines an editorial list associated with a media establishment; and
 - in response to determining that the speech input defines an editorial list associated with a media establishment, obtaining the at least one media item based on the editorial list associated with the media establishment, wherein the at least one media item includes metadata indicating the editorial list associated with the media establishment.
16. The method of any of claims 1-15, further comprising:
 - determining whether the speech input defines a mood; and
 - in response to determining that the speech input defines a mood, obtaining the at least one media item based on the mood, wherein the at least one media item includes metadata indicating the mood.
17. The method of any of claims 1-16, further comprising:
 - determining whether the speech input defines an activity; and
 - in response to determining that the speech input defines an activity, obtaining the at least one media item based on the activity, wherein the at least one media item includes metadata indicating the activity.
18. The method of any of claims 1-17, further comprising:
 - determining whether the speech input defines a time period;

in response to determining that the speech input defines a time period, determining whether the speech input defines a genre in association with the time period; and

in response to determining that the speech input defines a genre in association with the time period, determining a subgenre based on the time period and the genre, wherein the at least one media item is obtained based on the subgenre, and wherein the at least one media item includes metadata indicating the subgenre.

19. The method of any of claims 1-18, wherein the speech input defines a category of media items, wherein obtaining the at least one media item includes obtaining a plurality of media items associated with the category of media items, and further comprising:

determining a degree of familiarity of the user with the category of media items, wherein an average popularity rating of the plurality of media items is based on the degree of familiarity of the user with the category of media items.

20. The method of any of claims 1-19, further comprising:

determining an identity of the user by performing speaker identification using the speech input; and

based on the determined identity of the user, determining the user-specific corpus of media items from a plurality of user-specific corpuses of media items.

21. The method of any of claims 1-20, wherein obtaining the at least one media item comprises sending an encrypted token to a remote server, the encrypted token containing user identification information, and wherein the encrypted token is required to access the user-specific corpus of media items via the remote server.

22. The method of any of claims 1-21, further comprising:

determining whether a domain corresponding to the speech input is one of a plurality of predetermined domains;

in response to determining that a domain corresponding to the speech input is one of a plurality of predetermined domains:

upon providing the at least one media item, receiving audio input;

determining whether the audio input contains speech; and

in response to determining that the audio input does not contain speech, ceasing to receive audio input after a predetermined amount of time.

23. The method of claim 22, further comprising:
in response to determining that the audio input contains speech:
determining whether the speech of the audio input corresponds to a same domain as the speech input;
in response to determining that the speech of the audio input corresponds to a same domain as the speech input:
determining a user intent corresponding to the speech of the audio input; and
providing a response to the audio input in accordance with the user intent corresponding to the speech of the audio input.
24. The method of claim 23, further comprising:
in response to determining that the speech of the audio input does not correspond to a same domain as the speech input:
forgo providing a response to the audio input.
25. The method of any of claims 22-24, wherein the predetermined amount of time is based on a degree of ambient noise detected in the audio input.
26. The method of any of claims 1-25, wherein providing the at least one media item comprises playing a media item of the at least one media item, and further comprising:
while playing the media item, receiving a third speech input;
determining, based on the playing media item and the third speech input, a user intent corresponding to the third speech input; and
providing a response in accordance with the user intent that corresponds to the third speech input.
27. The method of any of claims 1-26, further comprising:
in response to determining that the speech input does not correspond to a user intent of obtaining personalized recommendations for media items:
determining whether the speech input corresponds to a user intent of obtaining media items having a recent release date; and

in response to determining that the speech input corresponds to a user intent of obtaining media items having a recent release date:

obtaining at least one second media item from a second corpus of media items, wherein each media item in the second corpus of media items has a release date that is within a predetermined time range of a current date; and
providing the at least one second media item.

28. The method of any of claims 1-27, wherein determining whether the speech input corresponds to a user intent of obtaining media items having a recent release date comprises determining whether the speech input includes one of a second plurality of phrases corresponding to the user intent of obtaining media items having a recent release date.

29. The method of any of claims 1-28, further comprising:

determining a political preference associated with the user, the determining based on previous media items requested or consumed by the user, wherein the at least one media item is obtained based on the determined political preference.

30. The method of any of claims 1-29, further comprising:

determining a technical proficiency associated with the user, the determining based on previous media items requested or consumed by the user, wherein the at least one media item is obtained based on the determined technical proficiency.

31. A computer-readable medium having instructions stored thereon, the instructions, when executed by one or more processors, cause the one or more processors to perform any of the methods of claims 1-30.

32. An electronic device comprising:

one or more processors; and

memory having instructions stored thereon, the instructions, when executed by the one or more processors, cause the one or more processors to perform any of the methods of claims 1-30.

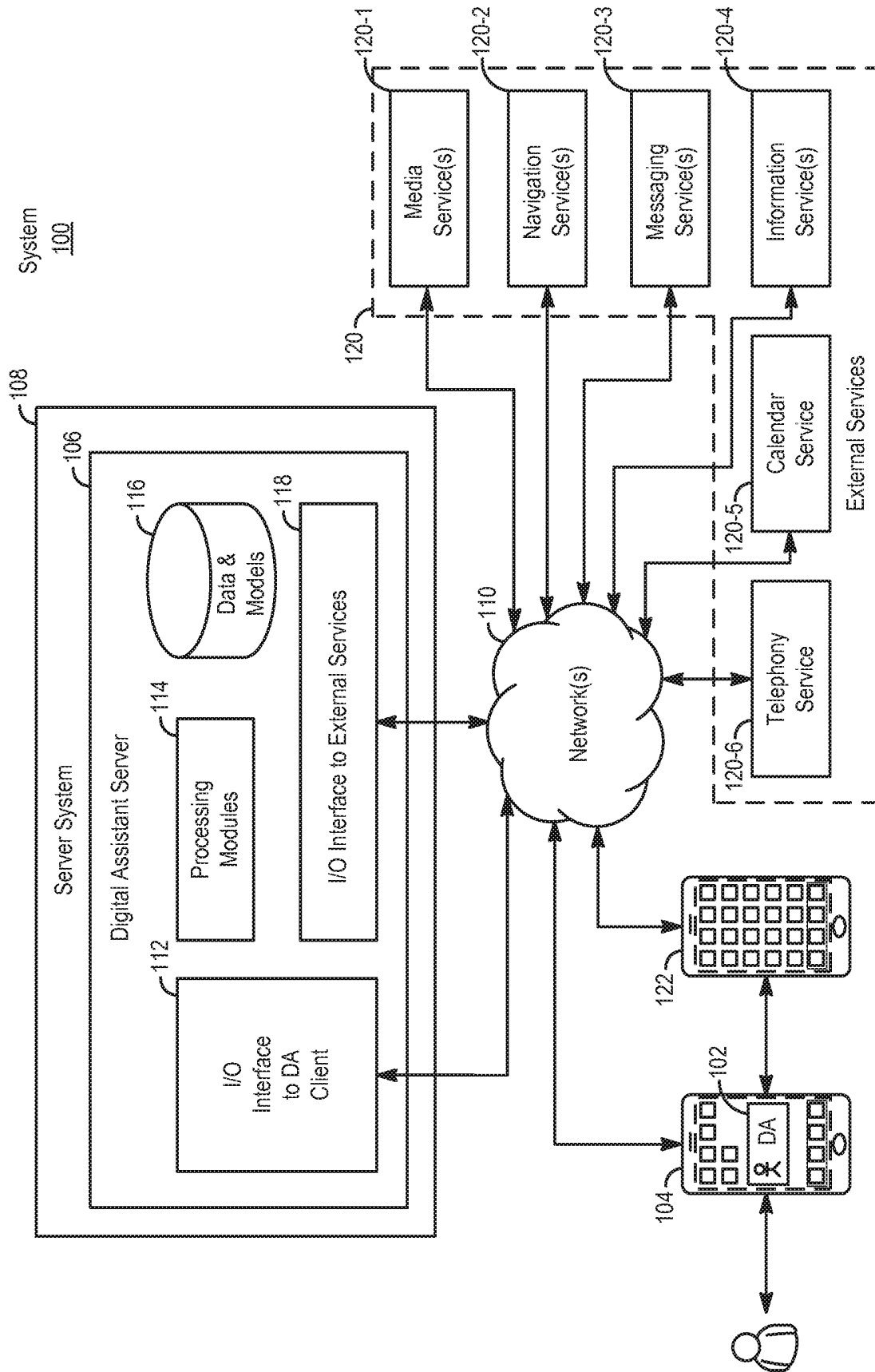


FIG. 1

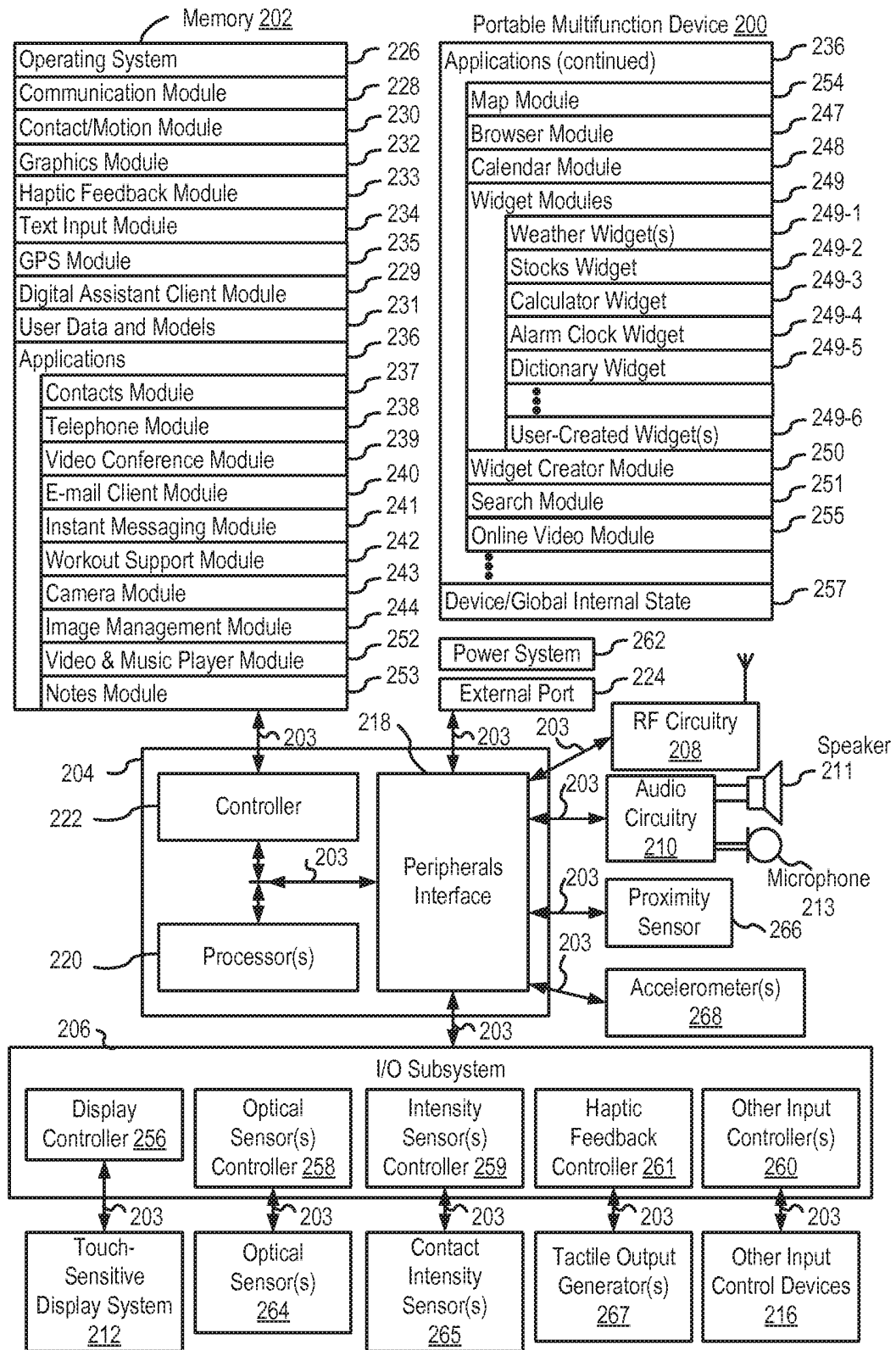


FIG. 2A

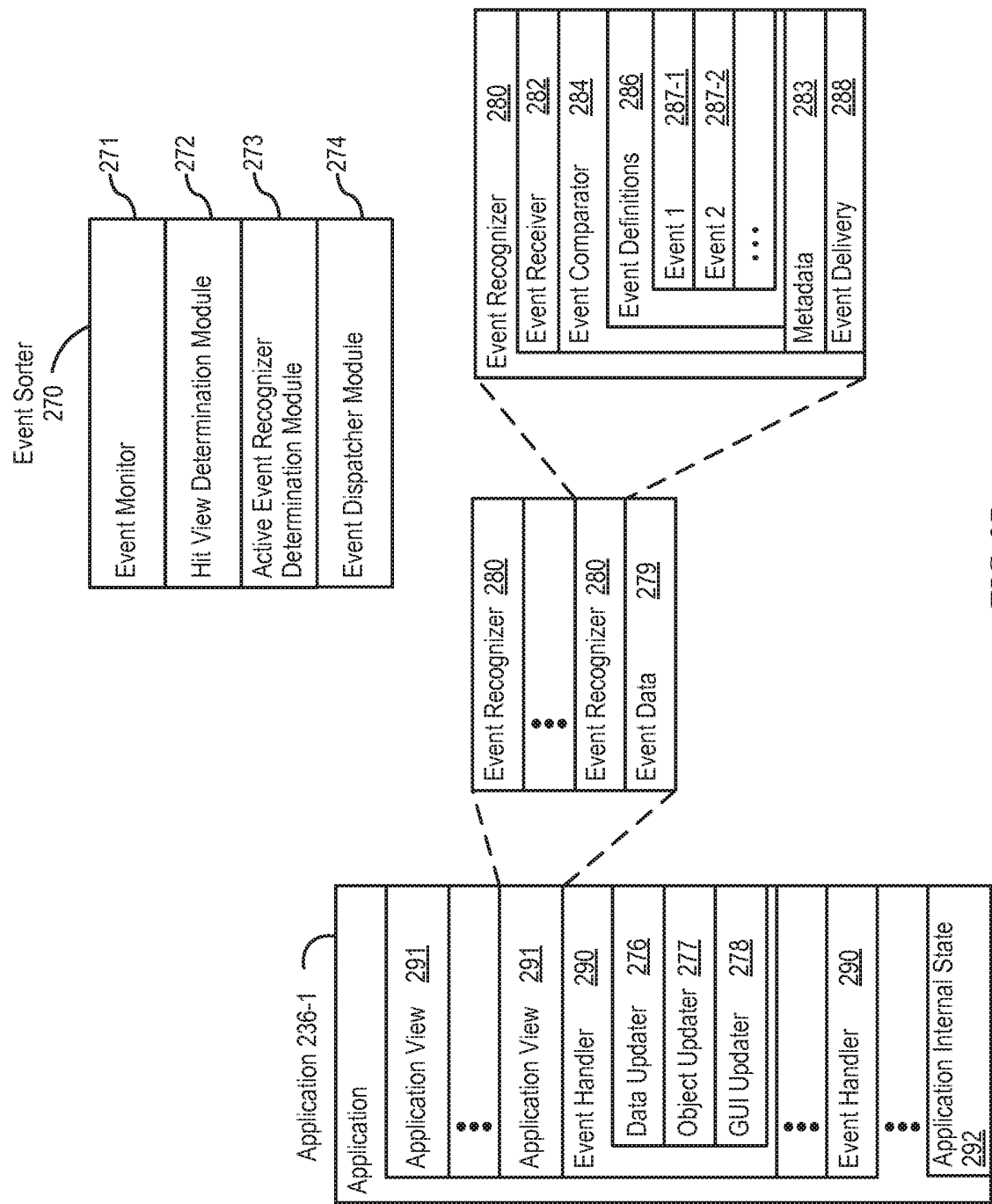


FIG. 2B

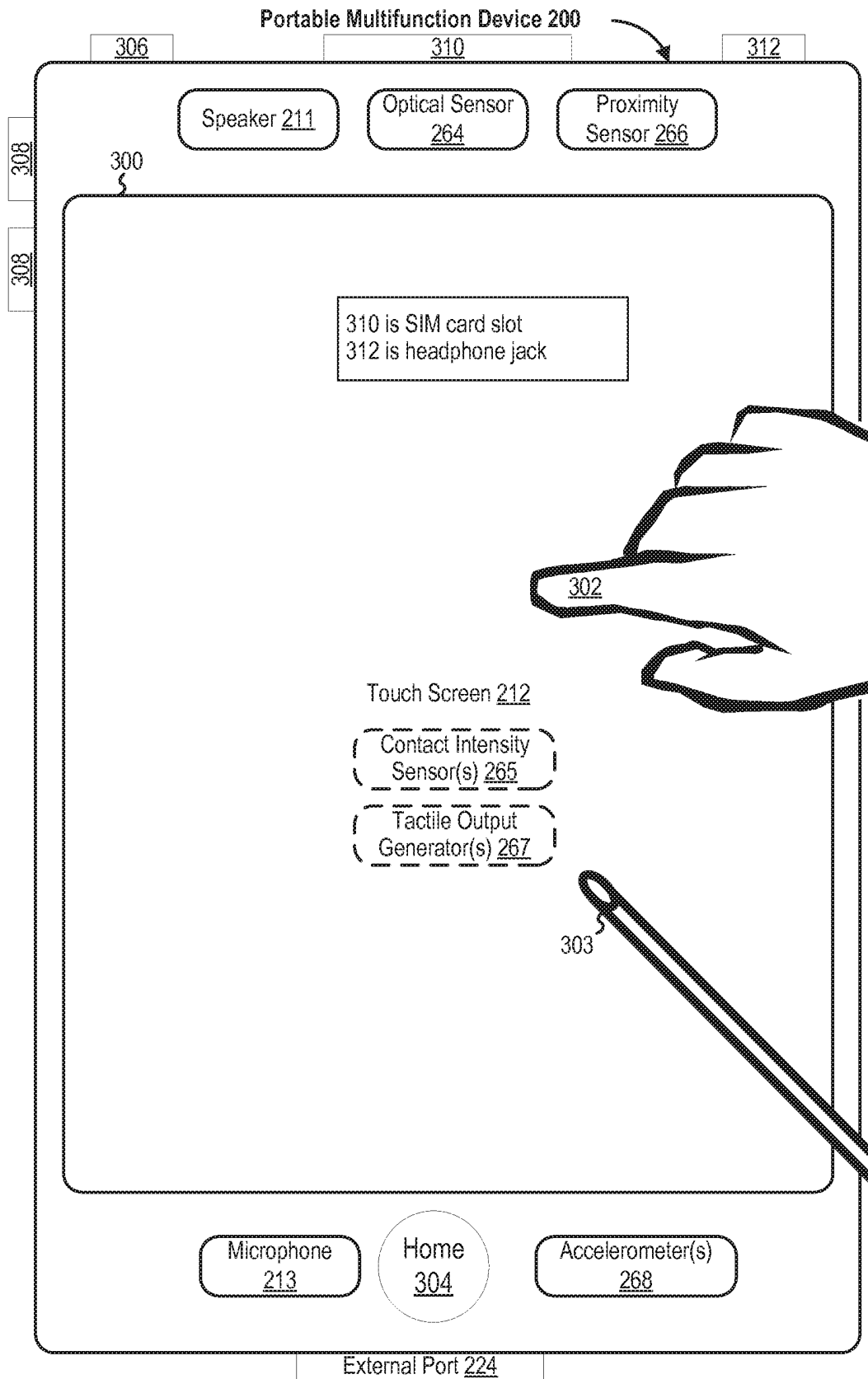


FIG. 3

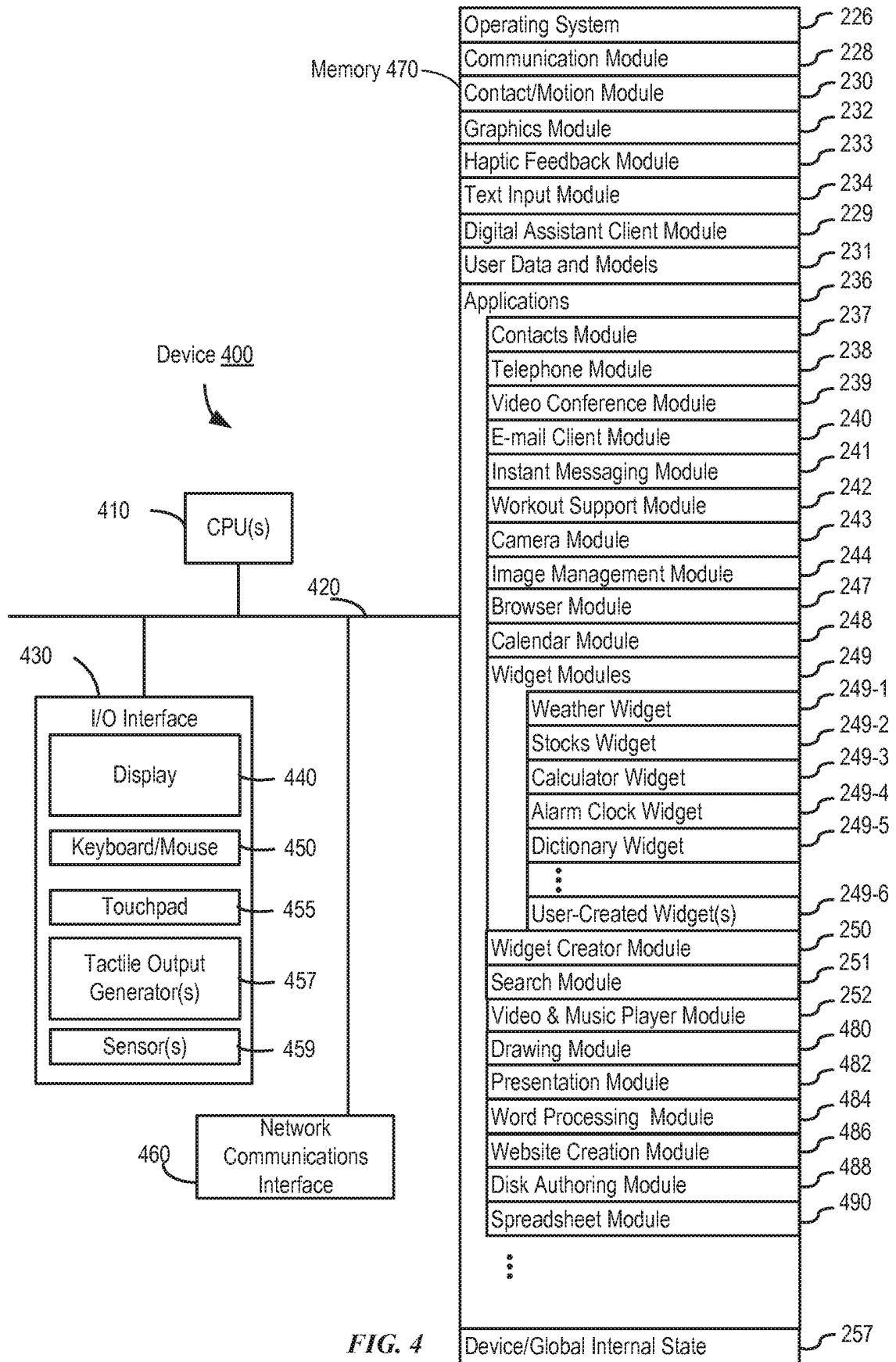


FIG. 4

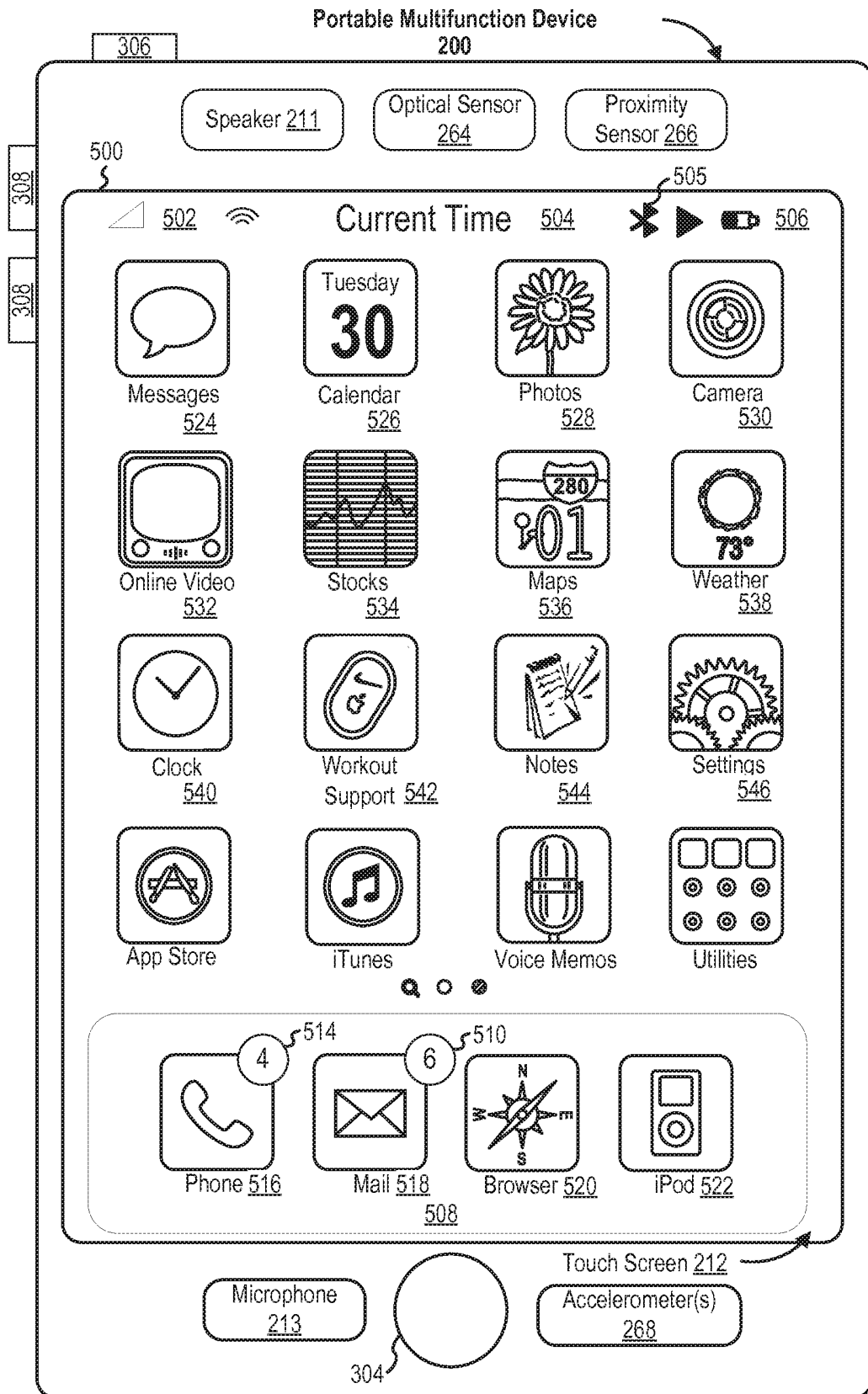


FIG. 5A

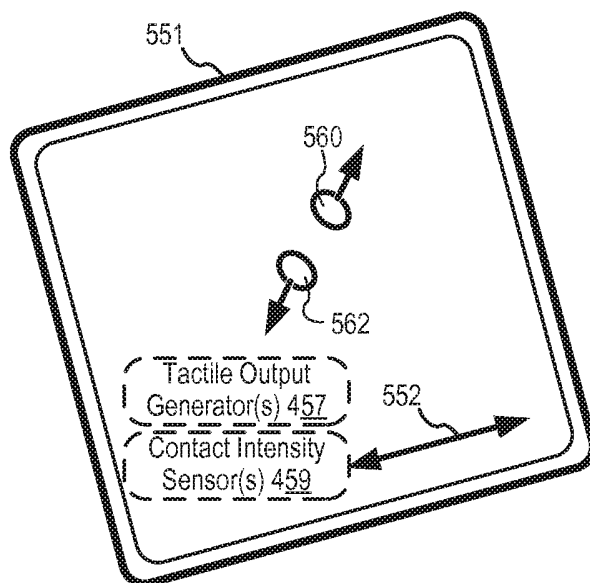
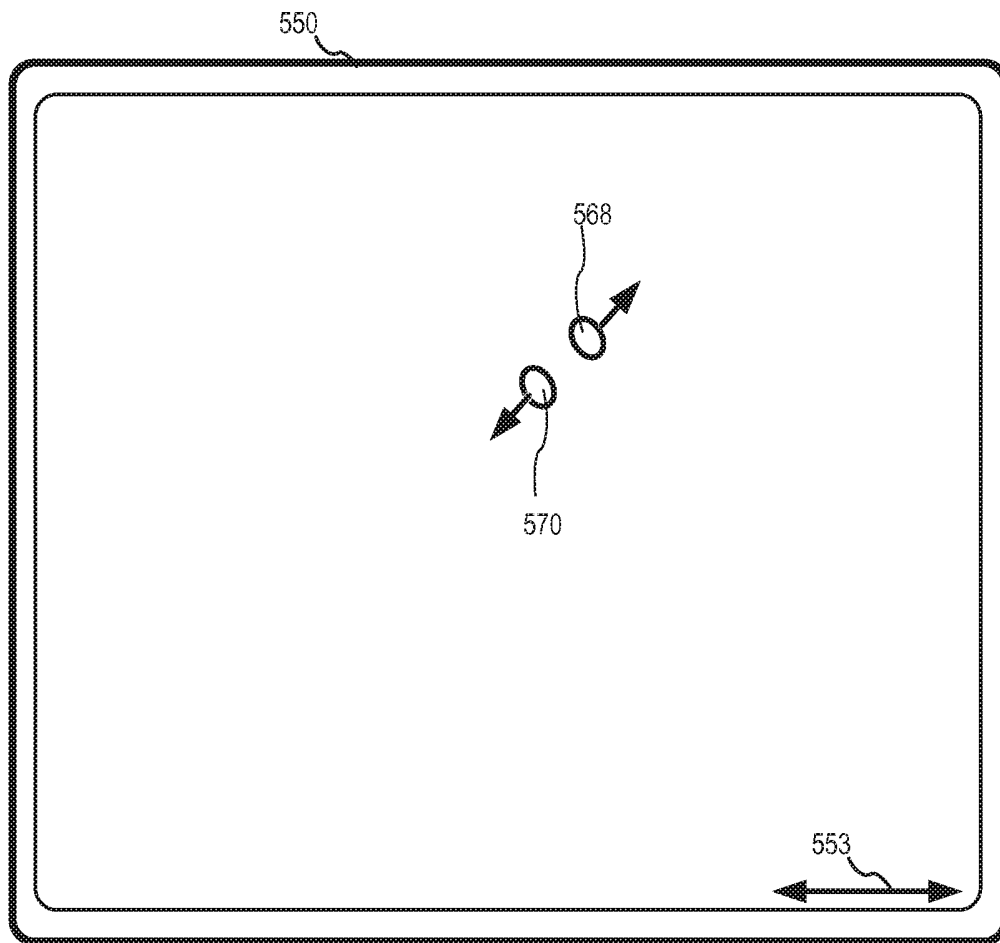
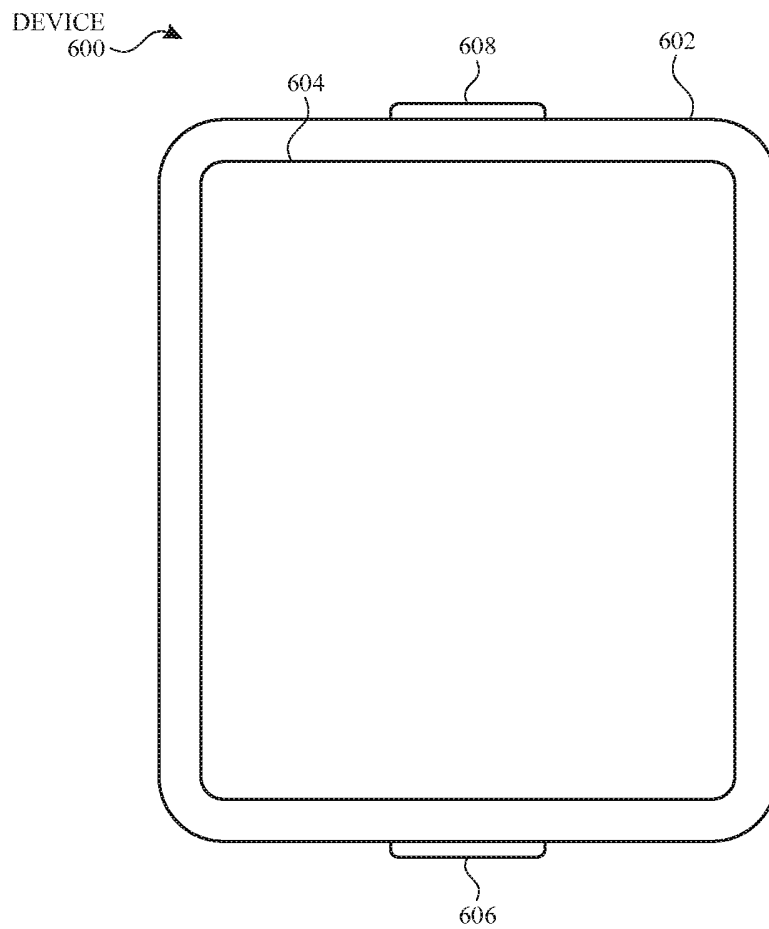


FIG. 5B

**FIG. 6A**

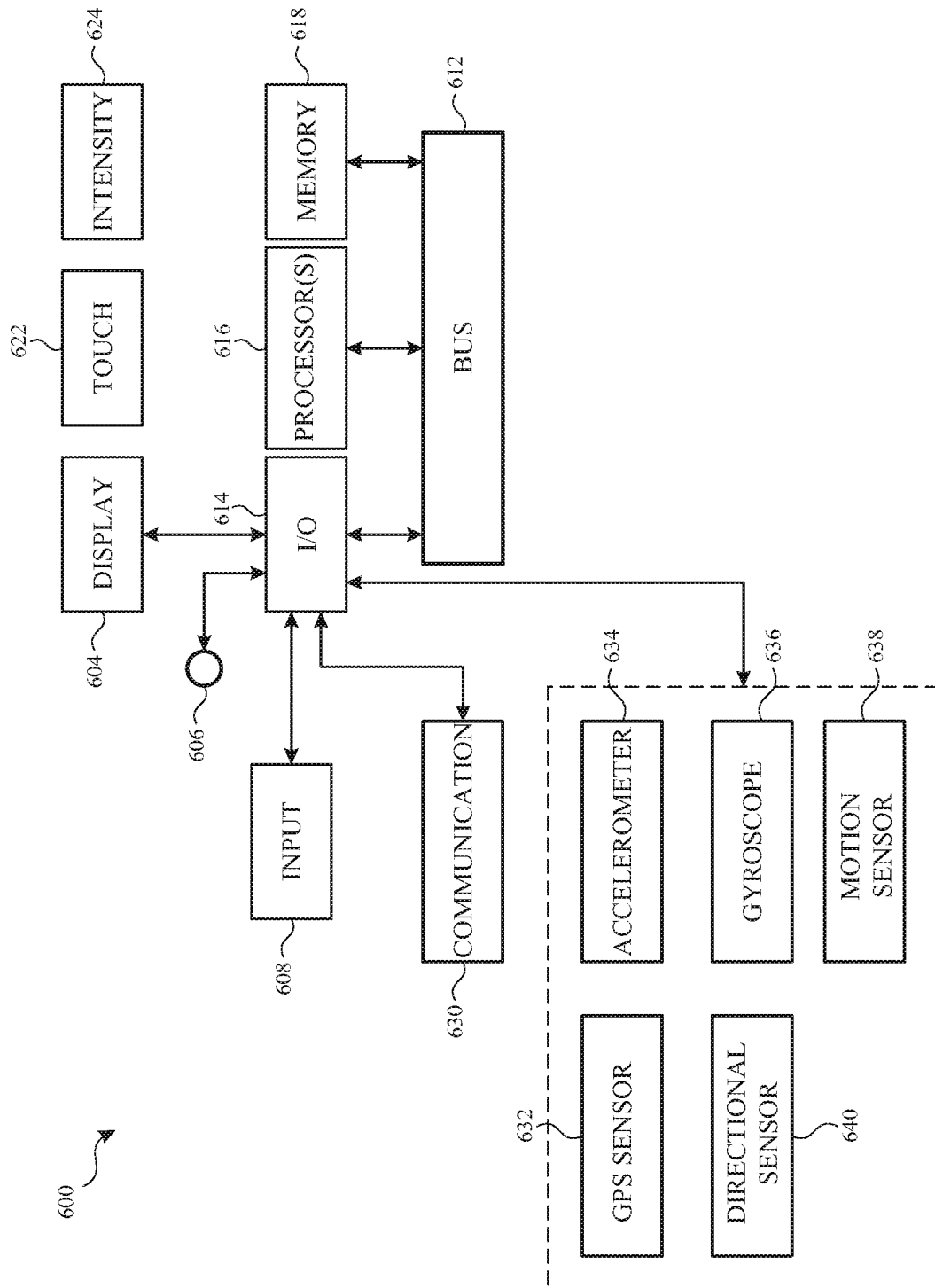


FIG. 6B

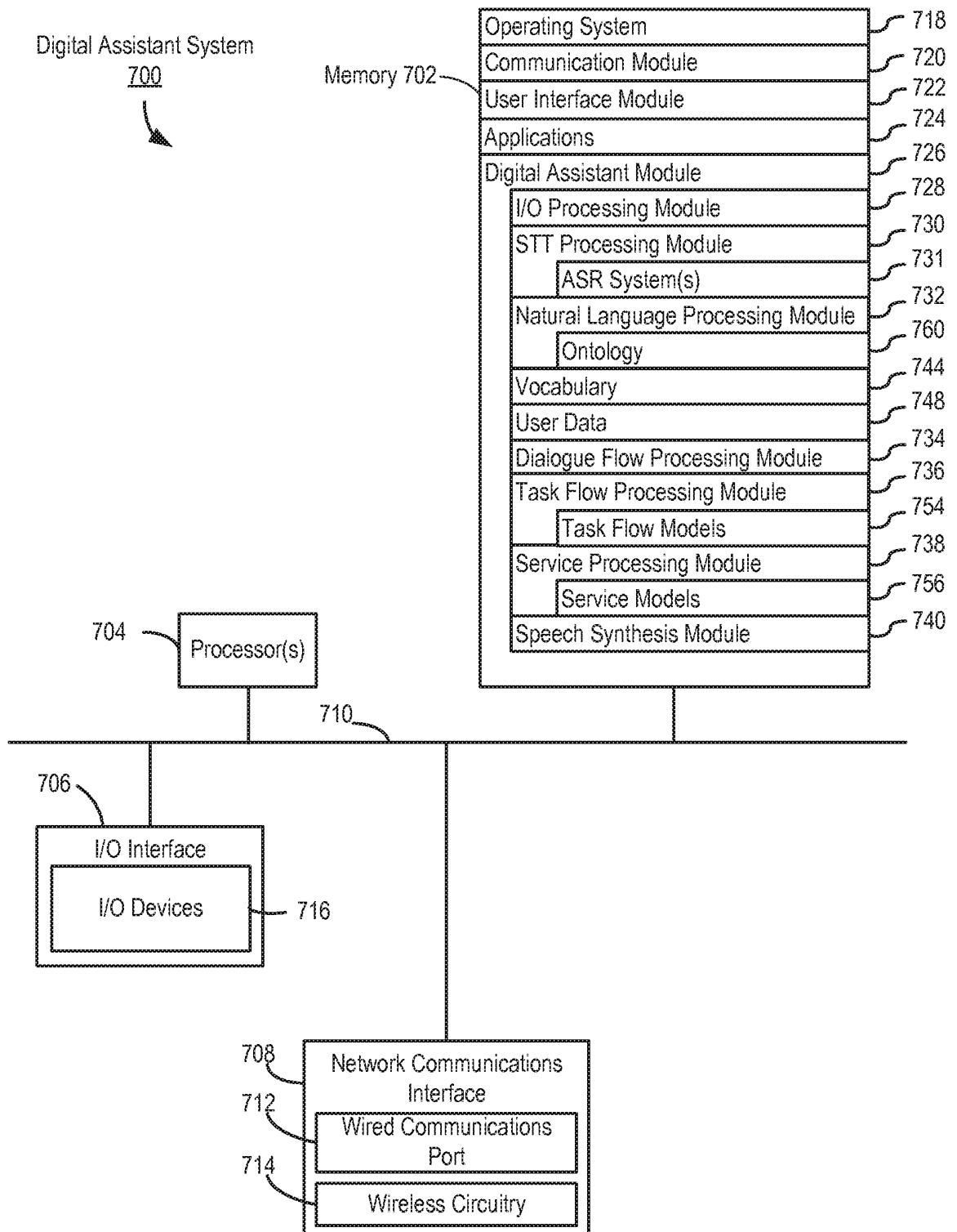


FIG. 7A

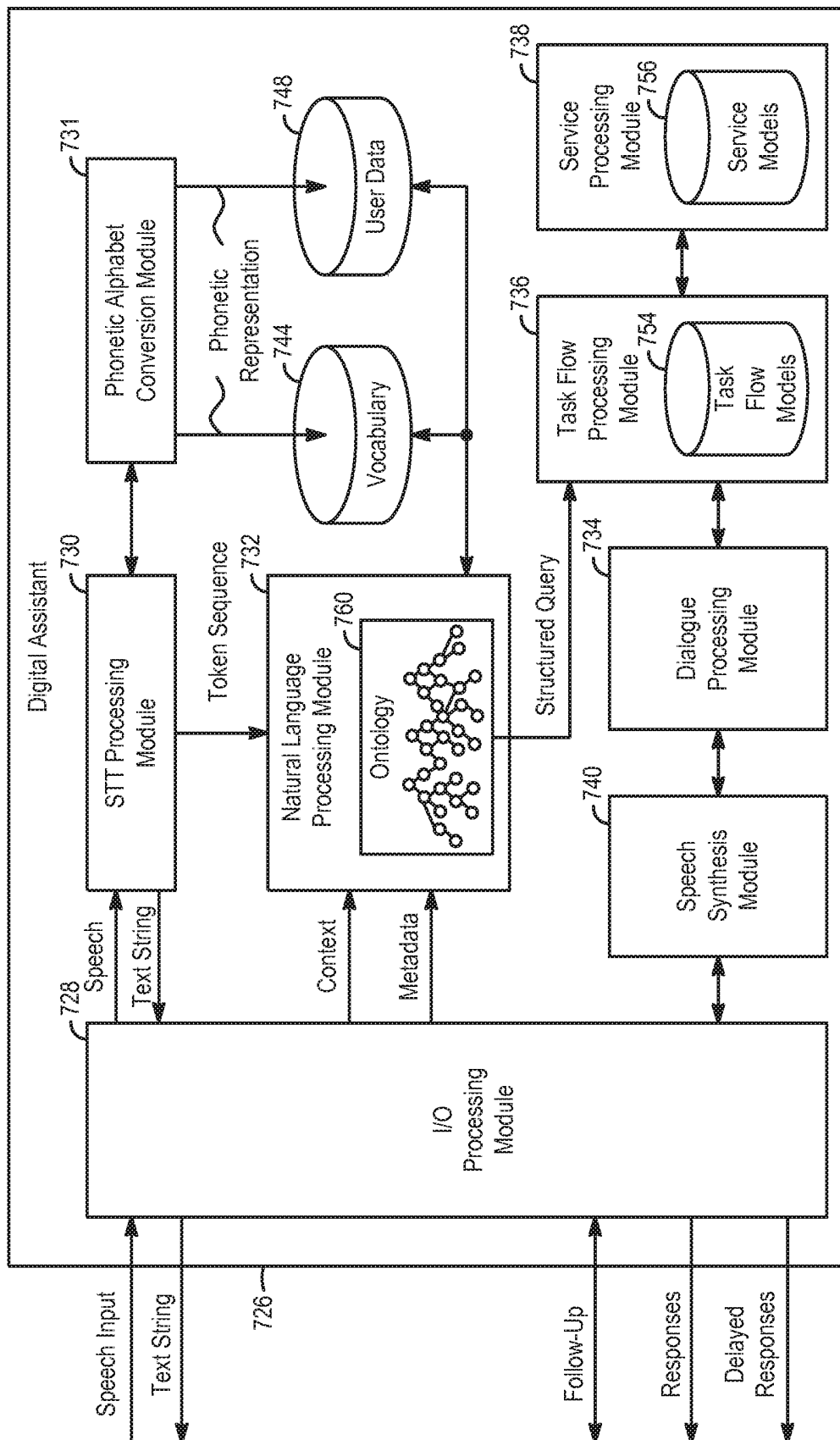


FIG. 7B

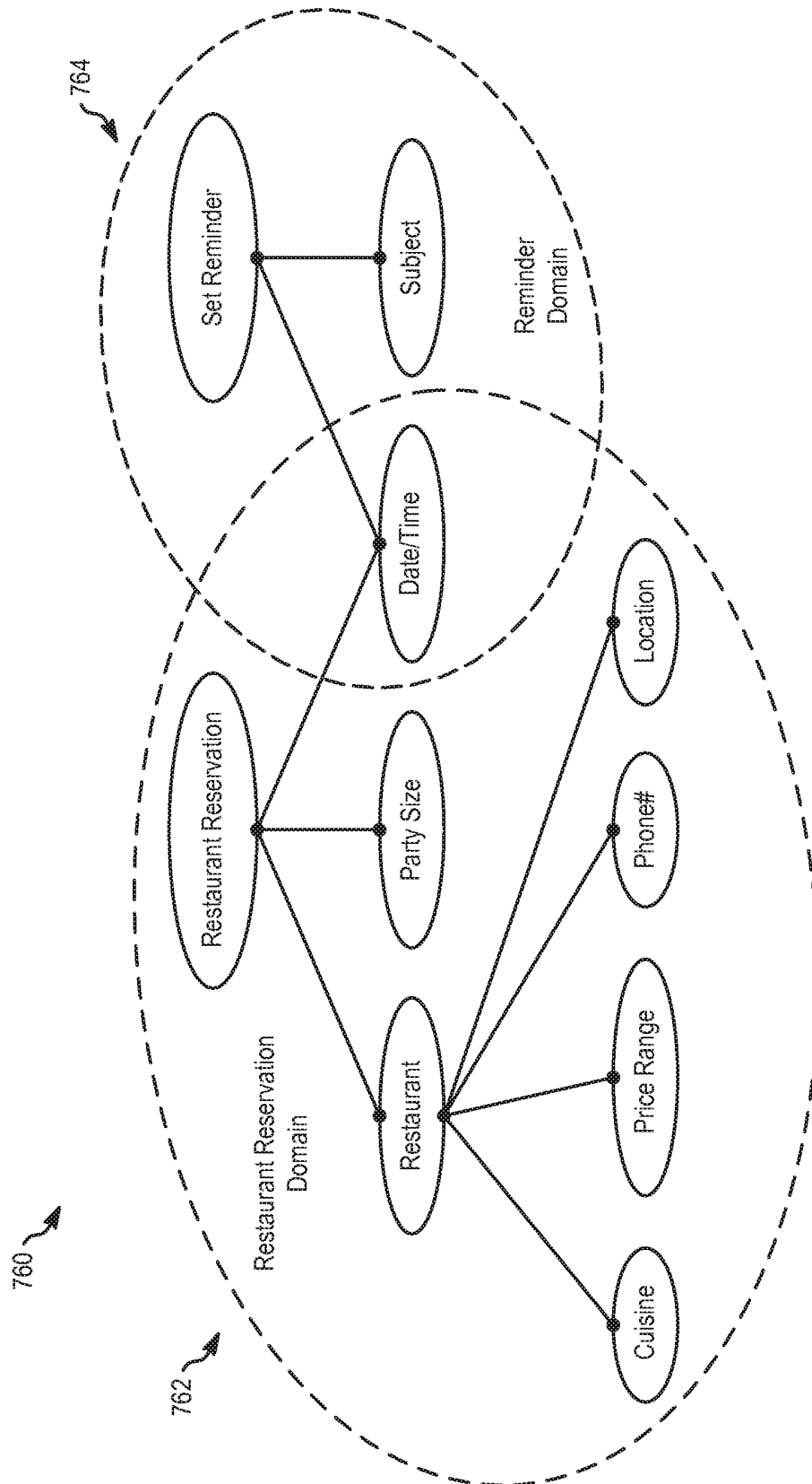
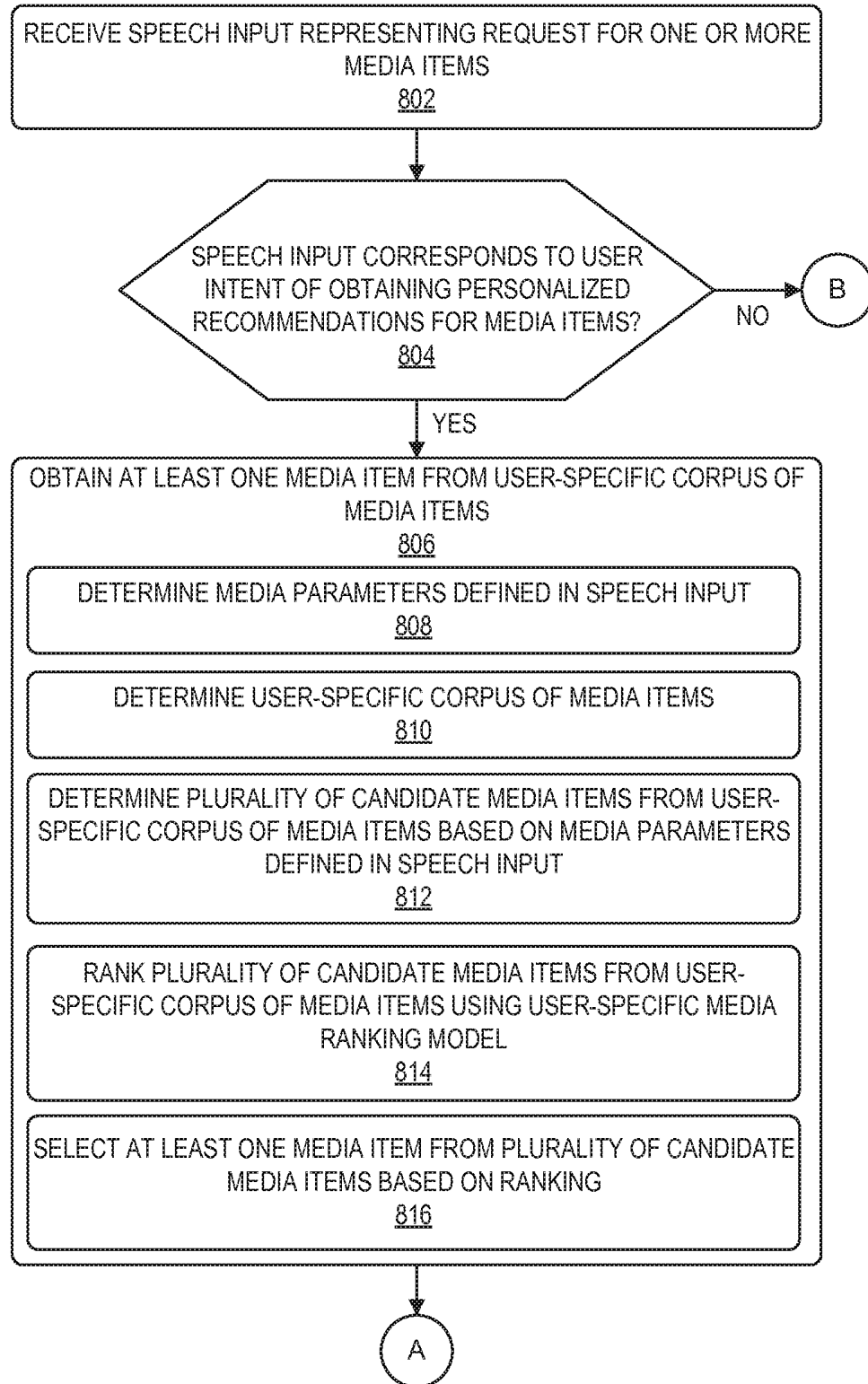


FIG. 7C

Process 800**FIG. 8A**

Process 800
(cont.)

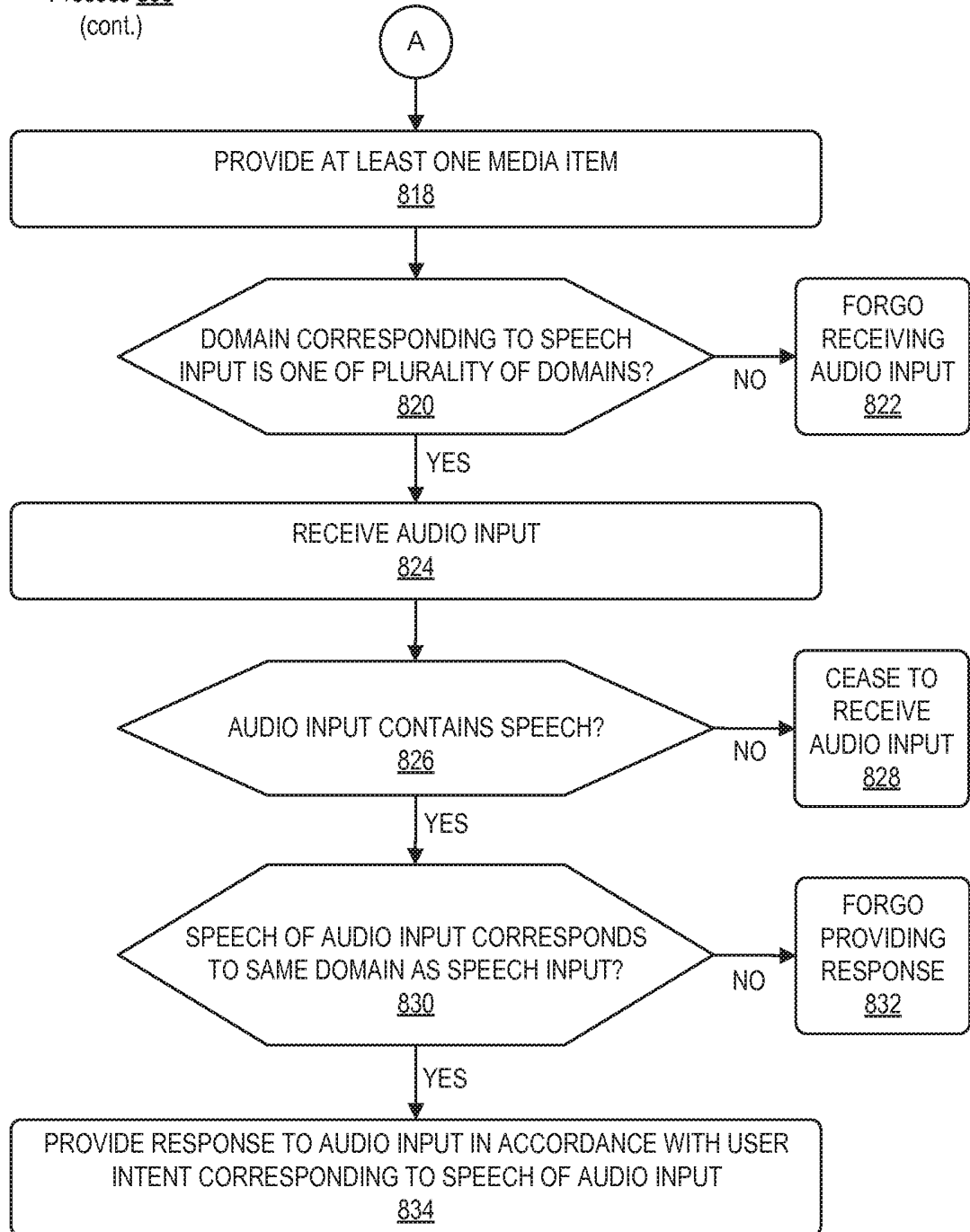
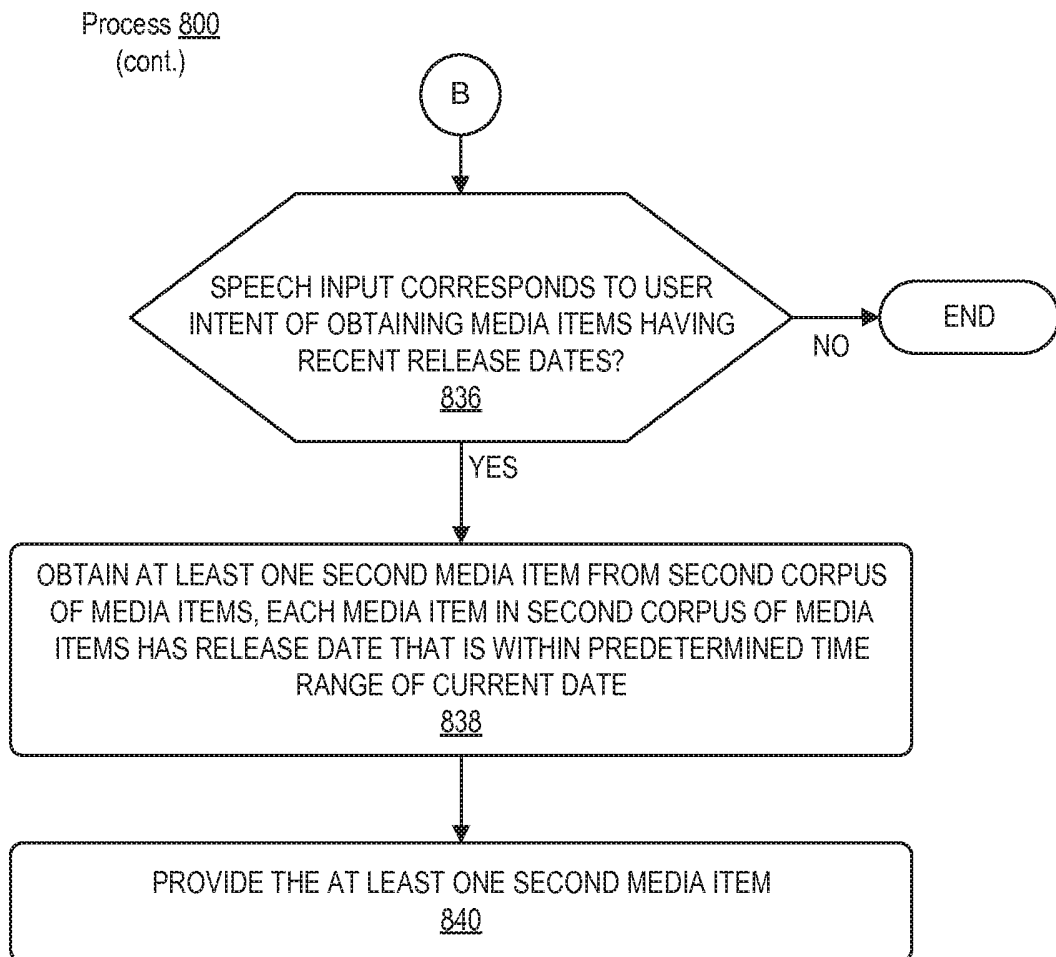


FIG. 8B

**FIG. 8C**

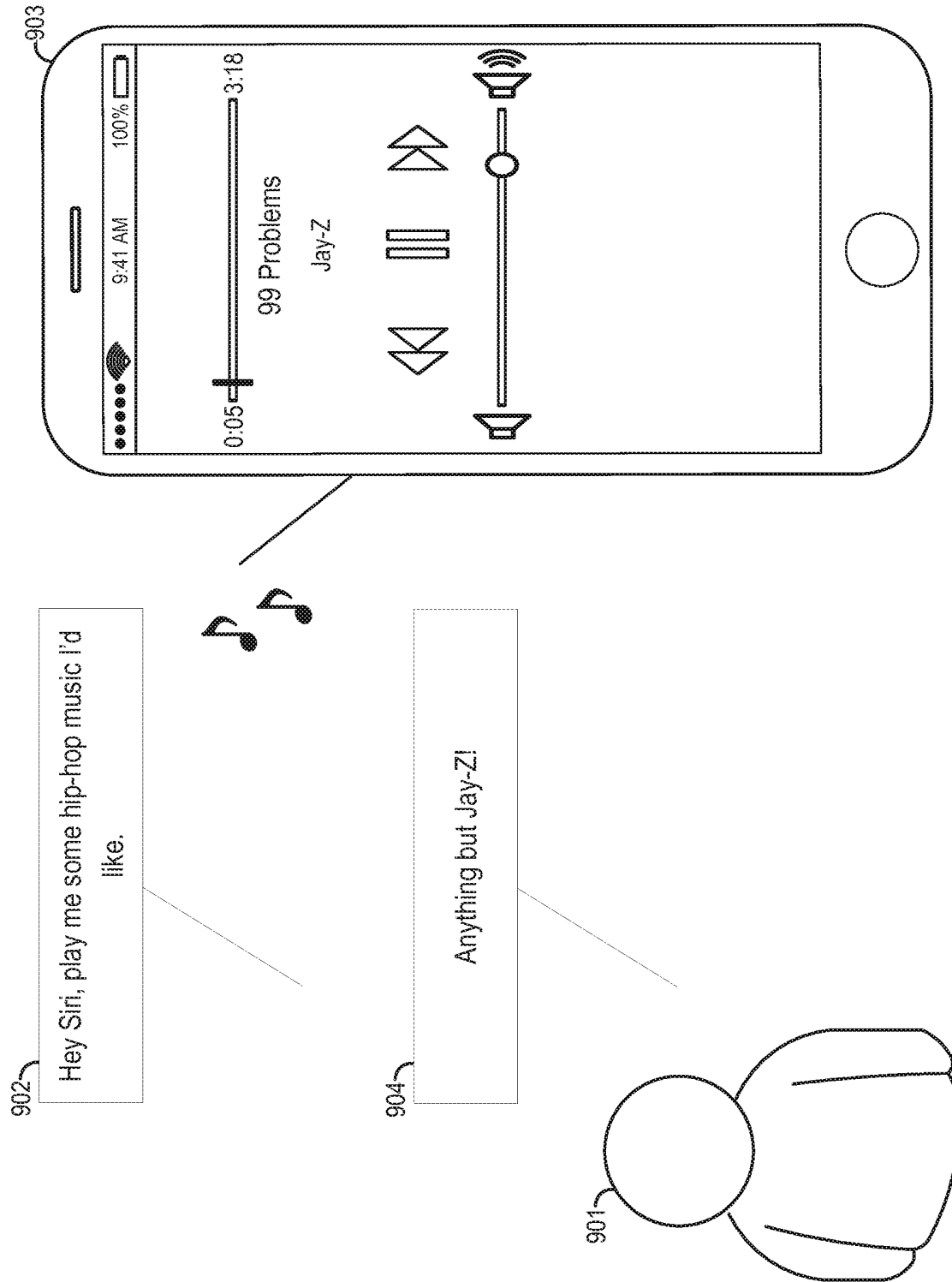


FIG. 9A

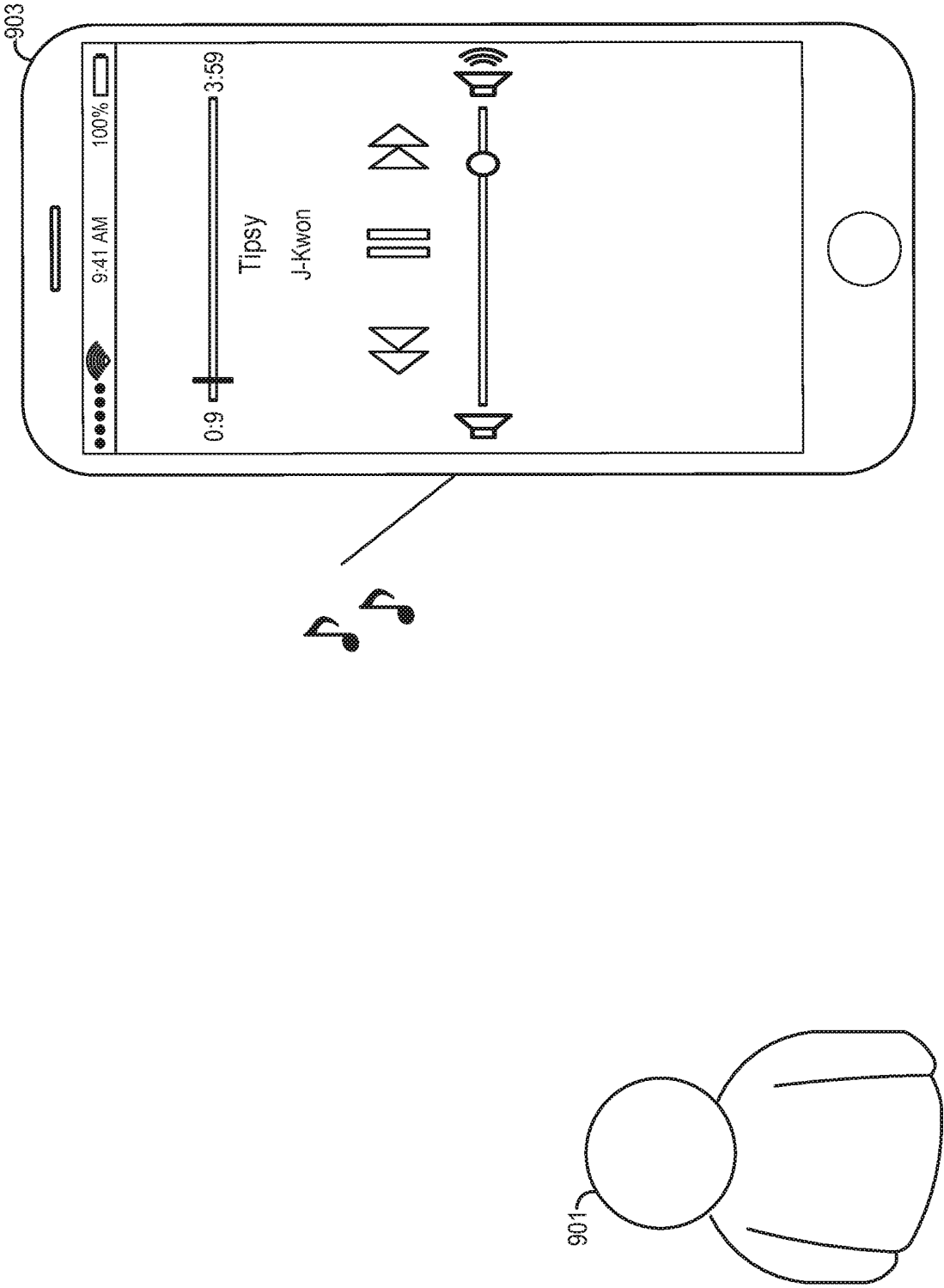


FIG. 9B

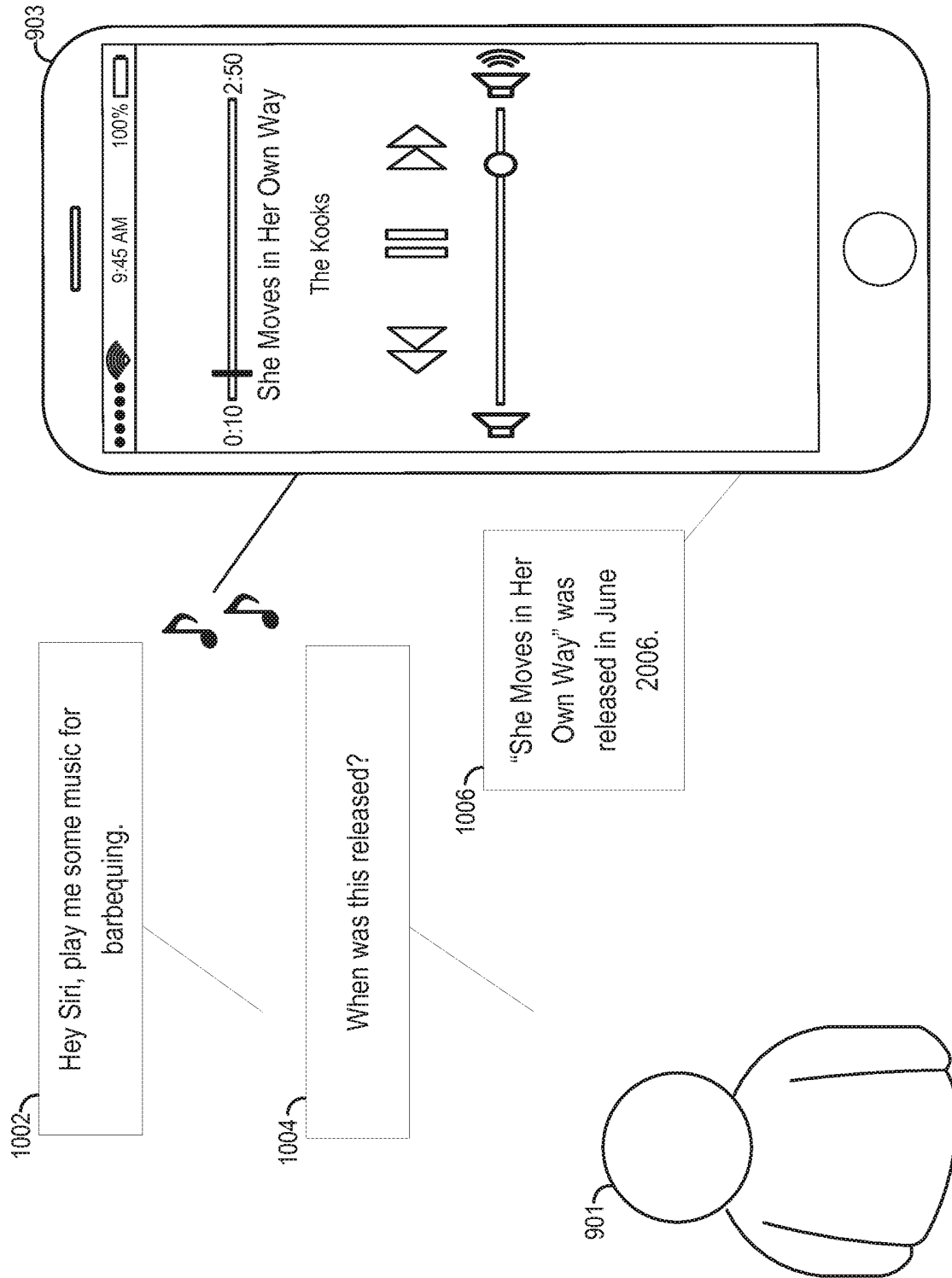


FIG. 10

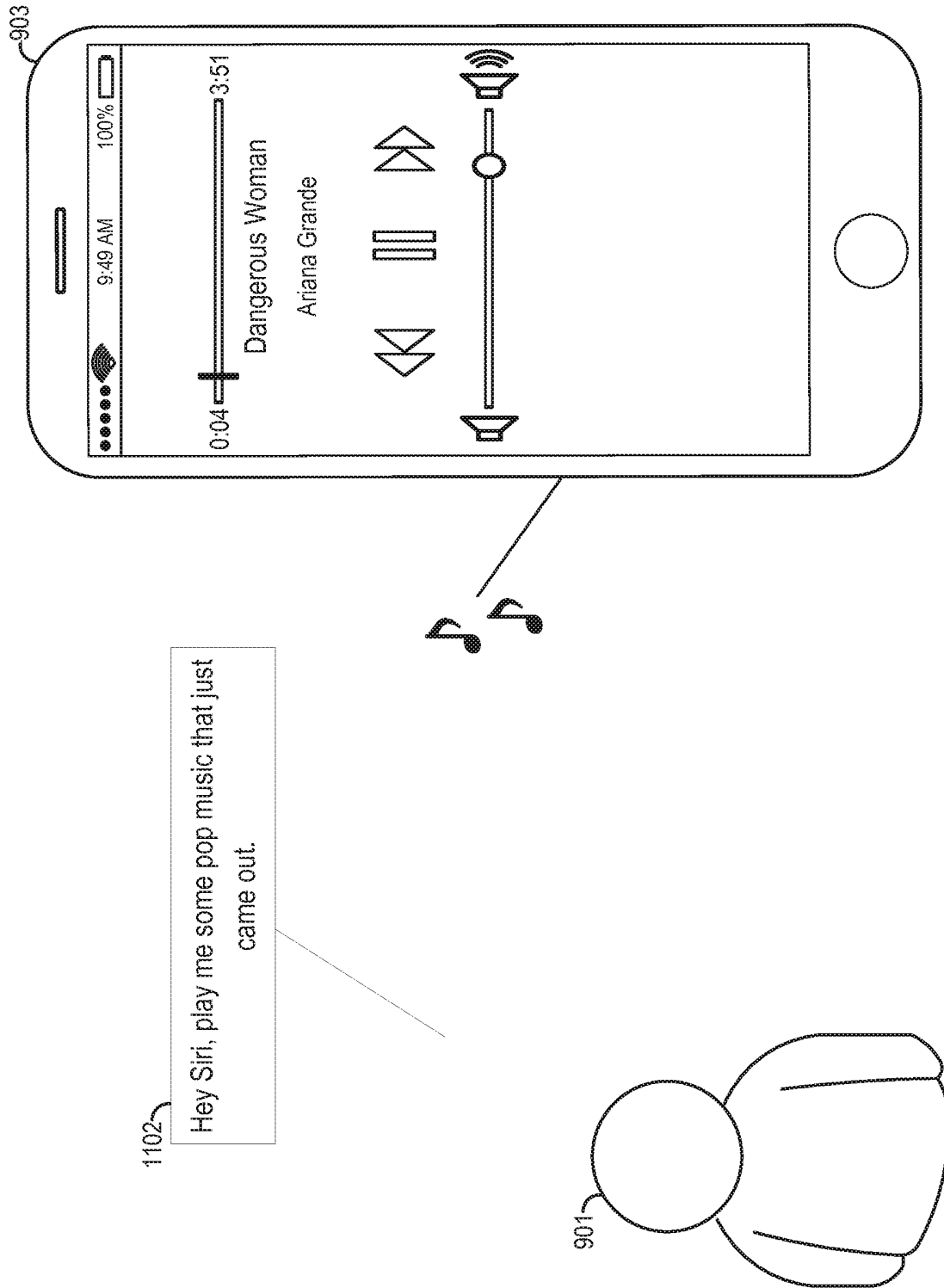
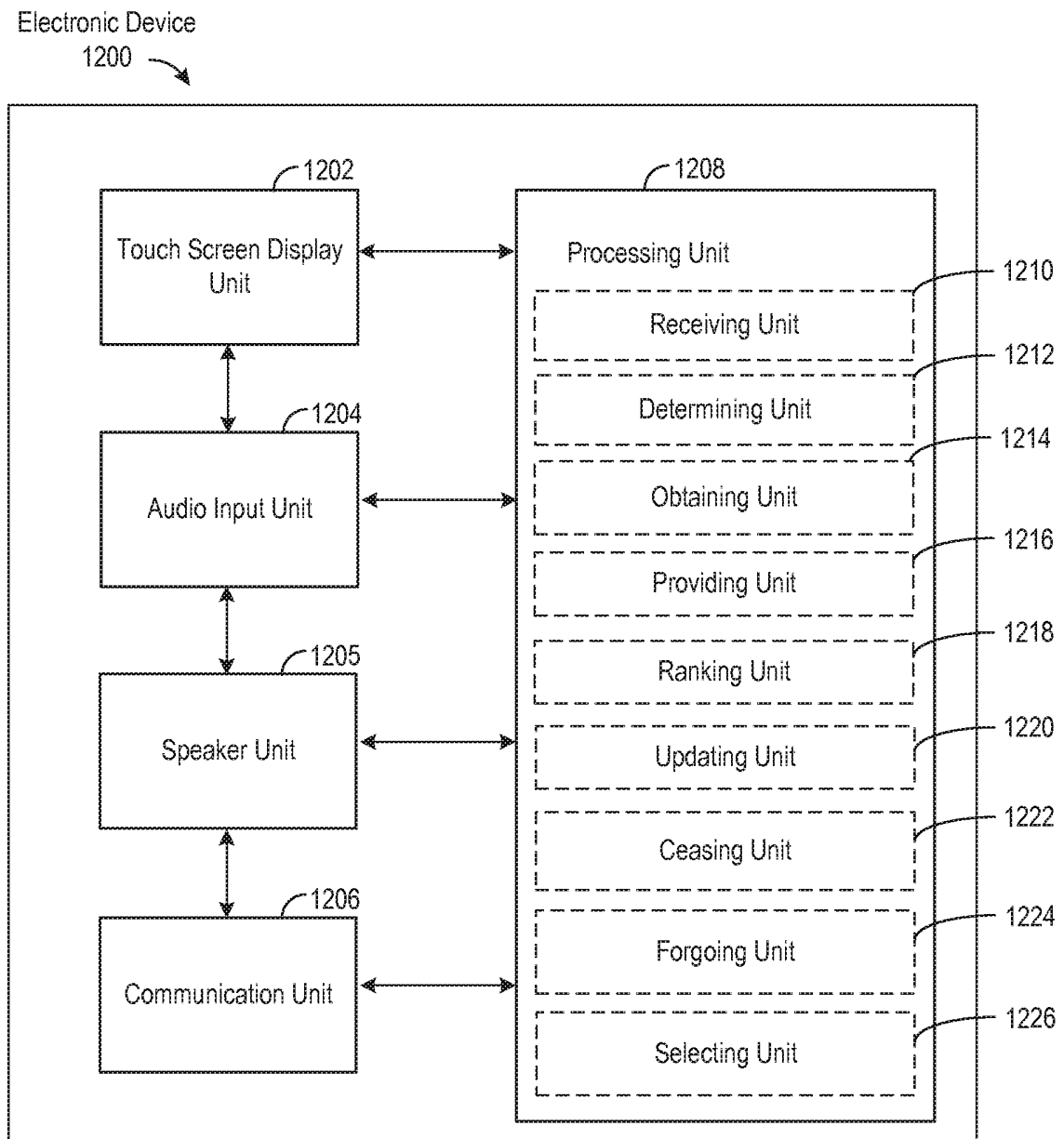


FIG. 11

**FIG. 12**

SEARCH REPORT - PATENT		Application No. PA 2017 70338
1. ☐ Certain claims were found unsearchable (See Box No. I). 2. ☐ Unity of invention is lacking prior to search (See Box No. II).		
A. CLASSIFICATION OF SUBJECT MATTER G 06 F 17/30 (2006.01); G 06 F 17/20 (2006.01); G 10 L 15/22 (2006.01) According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED PCT-minimum documentation searched (classification system followed by classification symbols) IPC/CPC: G06F, G10L		
Documentation searched other than PCT-minimum documentation DK, NO, SE, FI: IPC-classes as above.		
Electronic database consulted during the search (name of database and, where practicable, search terms used) Patent databases: EPODOC, WPI, FULL TEXT: ENGLISH; Non-patent literature: Inspec, IEEE		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant for claim No.
X	US 8239206 B1 (LEBEAU et al.) 2012-08-07; column 6, line 62-column 7, line 48; column 17, line 46	1-32
A	US 2014/0310001 A1 (KALNS et al.) 2014-10-16; para. 0056, 0068-0069	1-32
A	US 2016/0125048 A1 (HAMADA) 2016-05-05, claim 1	1-32
A	US 2014/0324884 A1 (LINDAHL et al.) 2014-10-30, para. 0052, 0059	1-32
A	WO 2012/019637 A1 (DAHULE et al.) 2012-02-16; page 2, lines 2-5; page 3, lines 1-4	1-32
☒ Further documents are listed in the continuation of Box C.		
* Special categories of cited documents: "A" Document defining the general state of the art which is not considered to be of particular relevance. "D" Document cited in the application. "E" Earlier application or patent but published on or after the filing date. "L" Document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified). "O" Document referring to an oral disclosure, use, exhibition or other means.	"P" Document published prior to the filing date but later than the priority date claimed. "T" Document not in conflict with the application but cited to understand the principle or theory underlying the invention. "X" Document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone. "Y" Document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. "&" Document member of the same patent family.	
Danish Patent and Trademark Office Helgeshøj Allé 81 DK-2630 Taastrup Denmark Telephone No. +45 4350 8000 Facsimile No. +45 4350 8001	Date of completion of the search report 18 July 2017 Authorized officer Jesper Bergstrand Telephone No. +45 4350 8364	

SEARCH REPORT - PATENT		Application No. PA 2017 70338
C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant for claim No.
A	CA 2631796 A1 (UNITED VIDEO PROPERTIES INC.) 2007-07-12; claims 1, 3, 8	1-32
A	US 2015/0382047 A1 (VAN OS et al.) 2015-12-31; para. 0048, 0138; claims 1, 5	1-32

SEARCH REPORT - PATENT	Application No. PA 2017 70338
SUPPLEMENTAL BOX	
Continuation of Box [.]	