

USING TEXT FOR AVATAR ANIMATION

FIELD

[0001] The present disclosure relates generally to avatar animation and, more specifically, to techniques of animating avatars using textual input.

BACKGROUND

[0002] Avatars are used to represent the users of electronic devices. The avatars can represent the appearance of a user or can represent an idealized or completely fictional representation of the user. Avatars can then be associated with a user so that the appearance of the avatar to others triggers an association or link with the user. Avatars can be created and edited for such use, including use in multimedia communications.

[0003] In some circumstances, an avatar may be created based on a representation of the user received by an electronic device. However, such creation may be limited to considering only particular representations or inputs, such as video input including the user. Additionally, creating these avatars may require specific types of video inputs or video inputs that include very specific features. Thus, the creation of avatars in this manner may only function when specific circumstances are met, resulting in an incomplete avatar, an avatar that does not accurately represent the user, or no avatar being created. Further, in some instances an animated avatar may be desired when these types of inputs are not available, such as when a user may only supply a textual input. Additionally, current avatars may not adequately provide a desired emotion because the range of motions considered and used to animate the avatar do not include motions used by humans to convey certain emotions.

SUMMARY

[0004] Example methods are disclosed herein. An example method includes, at an electronic device having one or more processors and memory, receiving text, determining an emotional state, and generating, using a neural network, a speech data set representing the received text and

a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

[0005] Example non-transitory computer-readable media are disclosed herein. An example non-transitory computer-readable storage medium stores one or more programs. The one or more programs comprise instructions, which when executed by one or more processors of an electronic device, cause the electronic device to receive text, determine an emotional state, and generate, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

[0006] Example electronic devices are disclosed herein. An example electronic device comprises one or more processors; a memory; and one or more programs, where the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for receiving text, determining an emotional state, and generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

[0007] An example electronic device comprises means for receiving text, determining an emotional state, and generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

[0008] Generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state allows for more accurate and flexible creation of an avatar representing a user. For example, in some circumstances a user may not wish or may not be able to provide a video input to create an avatar, as is required in conventional avatar creation methods, but may be able to provide a text input (e.g., while in a meeting). By generating the speech data set and the set of parameters representing one or more movements of the avatar based on the received text, the current system may generate an avatar in these types of circumstances, providing the user with greater flexibility when generating avatars. Additionally,

by considering a determined emotional state during the generation of the speech data set and the set of parameters representing one or more movements of the avatar, the avatar may appear more accurate and realistic to a recipient. For example, if the determined emotional state is sad the generated avatar may include movements that convey that the user is sad, resulting in a more realistic avatar that more accurately conveys the desired message to the recipient. Accordingly, generating the speech data set and the set of parameters representing one or more movements of the avatar in this manner provides for greater flexibility and accuracy, resulting in a more enjoyable user experience.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] FIG. 1 is a block diagram illustrating a system and environment for implementing a digital assistant, according to various examples.

[0010] FIG. 2A is a block diagram illustrating a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0011] FIG. 2B is a block diagram illustrating exemplary components for event handling, according to various examples.

[0012] FIG. 3 illustrates a portable multifunction device implementing the client-side portion of a digital assistant, according to various examples.

[0013] FIG. 4 is a block diagram of an exemplary multifunction device with a display and a touch-sensitive surface, according to various examples.

[0014] FIG. 5A illustrates an exemplary user interface for a menu of applications on a portable multifunction device, according to various examples.

[0015] 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.

[0016] FIG. 6A illustrates a personal electronic device, according to various examples.

[0017] FIG. 6B is a block diagram illustrating a personal electronic device, according to

various examples.

[0018] FIGS. 6C-6D illustrate exemplary components of a personal electronic device having a touch-sensitive display and intensity sensors in accordance with some embodiments.

[0019] FIGS. 6E-6H illustrate exemplary components and user interfaces of a personal electronic device in accordance with some embodiments.

[0020] FIG. 7A is a block diagram illustrating a digital assistant system or a server portion thereof, according to various examples.

[0021] FIG. 7B illustrates the functions of the digital assistant shown in FIG. 7A, according to various examples.

[0022] FIG. 7C illustrates a portion of an ontology, according to various examples.

[0023] FIG. 8 illustrates a system for generating and animating an avatar in accordance with some examples.

[0024] FIGS. 9A-9C illustrate generated and animated avatars, according to various examples.

[0025] FIGS. 10A-10C illustrate a process for generating and animating an avatar in accordance with some examples.

DETAILED DESCRIPTION

[0026] In the following description of examples, reference is made to the accompanying drawings in which are shown by way of illustration specific examples that can be practiced. It is to be understood that other examples can be used and structural changes can be made without departing from the scope of the various examples.

[0027] In some examples, a text is received and an emotional state is determined. Based, on the received text and the determined emotional state and a speech data set representing the received text and a set of parameters representing one or more movements is generated using a

neural network. Generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state allows for more accurate and flexible creation of an avatar representing a user. In particular, by generating these outputs based on text the user may create an avatar in many different situations, including those that may not allow the user to provide video or audio input. Further, by considering the determined emotional state the generated avatar may appear more realistic and convey the desired message more accurately.

[0028] 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.

[0029] 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.

[0030] 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

[0031] 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.

[0032] 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.

[0033] 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.

[0034] In some examples, DA server 106 includes client-facing I/O interface 112, one or more processing modules 114, data and models 116, and I/O interface to external services 118. The client-facing I/O interface 112 facilitates the client-facing input and output processing for DA server 106. One or more processing modules 114 utilize data and models 116 to process speech input and determine the user’s intent based on natural language input. Further, one or more processing modules 114 perform task execution based on inferred user intent. In some examples, DA server 106 communicates with external services 120 through network(s) 110 for task completion or information acquisition. I/O interface to external services 118 facilitates such communications.

[0035] User device 104 can be any suitable electronic device. In some examples, user device 104 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 Apple Watch®, iPhone®, iPod Touch®, and iPad® devices from Apple Inc. of Cupertino, California. Other examples of portable multifunction devices include, without limitation, earphones/headphones, speakers, and 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 speaker, 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.

[0036] 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.

[0037] 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.

[0038] 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 returns relevant data (e.g., data content responsive to the user request) to user device 104 via second user device 122.

[0039] 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.

[0040] 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

[0041] 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.

[0042] 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).

[0043] 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.

[0044] 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.

[0045] 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.

[0046] 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.

[0047] 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.

[0048] 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.

[0049] 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).

[0050] 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).

[0051] 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.

[0052] 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.

[0053] 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.

[0054] 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.

[0055] 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.

[0056] 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.

[0057] 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.

[0058] 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.

[0059] 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.

[0060] 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.

[0061] 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.

[0062] 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).

[0063] 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.

[0064] 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.

[0065] 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.

[0066] 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.

[0067] 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.

[0068] 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.

[0069] 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).

[0070] 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.

[0071] 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.

[0072] 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.

[0073] 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.

[0074] 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).

[0075] 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).

[0076] 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) 264, 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.

[0077] 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.

[0078] 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.

[0079] 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.

[0080] 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.

[0081] 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.

[0082] 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.

[0083] 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.

[0084] 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.

[0085] 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.

[0086] 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.

[0087] 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.

[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, 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).

[0089] 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.

[0090] 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.

[0091] 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.

[0092] 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.

[0093] 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.

[0094] 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).

[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, 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).

[0096] 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.

[0097] 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.).

[0098] 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.

[0099] 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.

[0100] 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.

[0101] 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.

[0102] 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.

[0103] 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.

[0104] 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).

[0105] 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.

[0106] 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.

[0107] 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.

[0108] 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).

[0109] In some embodiments, event sorter 270 also includes a hit view determination module 272 and/or an active event recognizer determination module 273.

[0110] 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.

[0111] 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.

[0112] 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.

[0113] 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.

[0114] 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.

[0115] 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.

[0116] 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.

[0117] 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).

[0118] 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.

[0119] 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.

[0120] 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.

[0121] 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.

[0122] 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.

[0123] 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.

[0124] 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.

[0125] 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.

[0126] 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.

[0127] 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.

[0128] 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.

[0129] 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.

[0130] 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.

[0131] 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.

[0132] 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.

[0133] 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.

[0134] Attention is now directed towards embodiments of user interfaces that can be implemented on, for example, portable multifunction device 200.

[0135] 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:

[0136] 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.

[0137] 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.

[0138] 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.

[0139] 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.

[0140] 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.

[0141] 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-4). 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.

[0142] 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.

[0143] 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.

[0144] 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.

[0145] 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.

[0146] 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.

[0147] As used here, the term “affordance” refers to a user-interactive graphical user interface object that is, for example, displayed on the display screen of devices 200, 400, 600, and/or 900 (FIGS. 2A, 4, 6A-B, and 9A-C). For example, an image (e.g., icon), a button, and text (e.g., hyperlink) each constitutes an affordance.

[0148] 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).

[0149] 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.

[0150] 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.

[0151] 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.

[0152] 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.

[0153] 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).

[0154] 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).

[0155] 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.

[0156] FIG. 6C illustrates detecting a plurality of contacts 652A-652E on touch-sensitive display screen 604 with a plurality of intensity sensors 624A-624D. FIG. 6C additionally includes intensity diagrams that show the current intensity measurements of the intensity sensors 624A-624D relative to units of intensity. In this example, the intensity measurements of intensity sensors 624A and 624D are each 9 units of intensity, and the intensity measurements of intensity sensors 624B and 624C are each 7 units of intensity. In some implementations, an aggregate intensity is the sum of the intensity measurements of the plurality of intensity sensors 624A-624D, which in this example is 32 intensity units. In some embodiments, each contact is assigned a respective intensity that is a portion of the aggregate intensity. FIG. 6D illustrates assigning the aggregate intensity to contacts 652A-652E based on their distance from the center of force 654. In this example, each of contacts 652A, 652B, and 652E are assigned an intensity of contact of 8 intensity units of the aggregate intensity, and each of contacts 652C and 652D are assigned an intensity of contact of 4 intensity units of the aggregate intensity. More generally, in some implementations, each contact j is assigned a respective intensity I_j that is a portion of the aggregate intensity, A , in accordance with a predefined mathematical function, $I_j = A \cdot (D_j / \sum D_i)$, where D_j is the distance of the respective contact j to the center of force, and $\sum D_i$ is the sum of the distances of all the respective contacts (e.g., $i=1$ to last) to the center of force. The operations described with reference to FIGS. 6C-6D can be performed using an electronic device similar or identical to device 104, 200, 400, or 600. In some embodiments, a characteristic intensity of a contact is based on one or more intensities of the contact. In some embodiments, the intensity sensors are used to determine a single characteristic intensity (e.g., a single characteristic intensity of a single contact). It should be noted that the intensity diagrams are not part of a displayed user interface, but are included in FIGS. 6C-6D to aid the reader.

[0157] 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.

[0158] The intensity of a contact on the touch-sensitive surface is, optionally, 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.

[0159] 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.

[0160] 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).

[0161] FIGS. 6E-6H illustrate detection of a gesture that includes a press input that corresponds to an increase in intensity of a contact 662 from an intensity below a light press intensity threshold (e.g., “IT_L”) in FIG. 6E, to an intensity above a deep press intensity threshold (e.g., “IT_D”) in FIG. 6H. The gesture performed with contact 662 is detected on touch-sensitive surface 660 while cursor 676 is displayed over application icon 672B corresponding to App 2, on a displayed user interface 670 that includes application icons 672A-672D displayed in predefined region 674. In some embodiments, the gesture is detected on touch-sensitive display 604. The intensity sensors detect the intensity of contacts on touch-sensitive surface 660. The device determines that the intensity of contact 662 peaked above the deep press intensity threshold (e.g., “IT_D”). Contact 662 is maintained on touch-sensitive surface 660. In response to the detection of the gesture, and in accordance with contact 662 having an intensity that goes above the deep

press intensity threshold (e.g., “IT_D”) during the gesture, reduced-scale representations 678A-678C (e.g., thumbnails) of recently opened documents for App 2 are displayed, as shown in FIGS. 6F-6H. In some embodiments, the intensity, which is compared to the one or more intensity thresholds, is the characteristic intensity of a contact. It should be noted that the intensity diagram for contact 662 is not part of a displayed user interface, but is included in FIGS. 6E-6H to aid the reader.

[0162] In some embodiments, the display of representations 678A-678C includes an animation. For example, representation 678A is initially displayed in proximity of application icon 672B, as shown in FIG. 6F. As the animation proceeds, representation 678A moves upward and representation 678B is displayed in proximity of application icon 672B, as shown in FIG. 6G. Then, representations 678A moves upward, 678B moves upward toward representation 678A, and representation 678C is displayed in proximity of application icon 672B, as shown in FIG. 6H. Representations 678A-678C form an array above icon 672B. In some embodiments, the animation progresses in accordance with an intensity of contact 662, as shown in FIGS. 6F-6G, where the representations 678A-678C appear and move upwards as the intensity of contact 662 increases toward the deep press intensity threshold (e.g., “IT_D”). In some embodiments, the intensity, on which the progress of the animation is based, is the characteristic intensity of the contact. The operations described with reference to FIGS. 6E-6H can be performed using an electronic device similar or identical to device 104, 200, 400, 600, or 800.

[0163] 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).

[0164] 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.

[0165] As used herein, an “installed application” refers to a software application that has been downloaded onto an electronic device (e.g., devices 100, 200, 400, 600 and/or 800) and is ready to be launched (e.g., become opened) on the device. In some embodiments, a downloaded application becomes an installed application by way of an installation program that extracts program portions from a downloaded package and integrates the extracted portions with the operating system of the computer system.

[0166] As used herein, the terms “open application” or “executing application” refer to a software application with retained state information (e.g., as part of device/global internal state 157 and/or application internal state 192). An open or executing application is, optionally, any one of the following types of applications:

- an active application, which is currently displayed on a display screen of the device that the application is being used on;

- a background application (or background processes), which is not currently displayed, but one or more processes for the application are being processed by one or more processors; and
- a suspended or hibernated application, which is not running, but has state information that is stored in memory (volatile and non-volatile, respectively) and that can be used to resume execution of the application.

[0167] As used herein, the term “closed application” refers to software applications without retained state information (e.g., state information for closed applications is not stored in a memory of the device). Accordingly, closing an application includes stopping and/or removing application processes for the application and removing state information for the application from the memory of the device. Generally, opening a second application while in a first application does not close the first application. When the second application is displayed and the first application ceases to be displayed, the first application becomes a background application.

3. Digital Assistant System

[0168] FIG. 7A illustrates a block diagram of digital assistant system 700 in accordance with various examples. In some examples, digital assistant system 700 is implemented on a standalone computer system. In some examples, digital assistant system 700 is distributed across multiple computers. In some examples, some of the modules and functions of the digital assistant are divided into a server portion and a client portion, where the client portion resides on one or more user devices (e.g., devices 104, 122, 200, 400, 600, or 900) 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.

[0169] 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.

[0170] 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).

[0171] In some examples, I/O interface 706 couples input/output devices 716 of digital assistant system 700, such as displays, keyboards, touch screens, and microphones, to user interface module 722. I/O interface 706, in conjunction with user interface module 722, receives user inputs (e.g., voice input, keyboard inputs, touch inputs, etc.) and processes them accordingly. In some examples, e.g., when the digital assistant is implemented on a standalone user device, digital assistant system 700 includes any of the components and I/O communication interfaces described with respect to devices 200, 400, 600, or 900 in FIGS. 2A, 4, 6A-B, 9A-C respectively. In some examples, digital assistant system 700 represents the server portion of a digital assistant implementation, and can interact with the user through a client-side portion residing on a user device (e.g., devices 104, 200, 400, 600, or 900).

[0172] 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.

[0173] 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.

[0174] 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.

[0175] 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, 600, and 900 shown in FIGS. 2A, 4, 6A-B, 9A-C, respectively. Communications module 720 also includes various components for handling data received by wireless circuitry 714 and/or wired communications port 712.

[0176] 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.).

[0177] 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.

[0178] 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 processing 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 758.

[0179] 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.

[0180] 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.

[0181] STT processing module 730 includes one or more ASR systems 758. The one or more ASR systems 758 can process the speech input that is received through I/O processing module 728 to produce a recognition result. Each ASR system 758 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 758 includes one or more speech recognition models (e.g., acoustic models and/or language models) and implements one or more speech recognition engines. Examples of speech recognition models include Hidden Markov Models, Gaussian-Mixture Models, Deep Neural Network Models, n-gram language models, and other statistical models. Examples of speech recognition engines include the dynamic time warping based engines and weighted finite-state transducers (WFST) based engines. The one or more speech recognition models and the one or more speech recognition engines are used to process the extracted representative features of the front-end speech pre-processor to produce intermediate recognitions results (e.g., phonemes, phonemic strings, and sub-words), and ultimately, text recognition results (e.g., words, word strings, or sequence of tokens). In some examples, the speech input is processed at least partially by a third-party service or on the user's device (e.g., device 104, 200, 400, or 600) to produce the recognition result. Once STT processing module 730 produces recognition results containing a text string (e.g., words, or sequence of words, or sequence of tokens), the recognition result is passed to natural language processing module 732 for intent deduction. In some examples, STT processing module 730 produces multiple candidate text representations of the speech input. Each candidate text representation is a sequence of words or tokens corresponding to the speech input. In some examples, each candidate text representation is associated with a speech recognition confidence score. Based on the speech recognition confidence scores, STT processing module 730 ranks the candidate text representations and provides the n-best (e.g., n highest ranked) candidate text representation(s) to natural language processing module 732 for intent deduction, where n is a predetermined integer greater than zero. For example, in one example, only the highest ranked (n=1) candidate text representation is passed to natural language processing module 732 for intent deduction. In another example, the five highest ranked (n=5) candidate text representations are passed to natural language processing module 732 for intent deduction.

[0182] 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.

[0183] 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.

[0184] 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'meɪroʊ/ (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).

[0185] 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'meɪroʊ/ 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.”

[0186] In some examples, STT processing module 730 uses approximate matching techniques to determine words in an utterance. Thus, for example, the STT processing module 730 determines that the sequence of phonemes /tə'meɪroʊ/ corresponds to the word “tomato,” even if that particular sequence of phonemes is not one of the candidate sequence of phonemes for that word.

[0187] Natural language processing module 732 (“natural language processor”) of the digital assistant takes the n-best candidate text representation(s) (“word sequence(s)” or “token sequence(s)”) generated by STT processing module 730, and attempts to associate each of the candidate text representations with one or more “actionable intents” recognized by the digital assistant. An “actionable intent” (or “user intent”) represents a task that can be performed by the digital assistant, and can have an associated task flow implemented in task flow models 754. The associated task flow is a series of programmed actions and steps that the digital assistant takes in order to perform the task. The scope of a digital assistant’s capabilities is dependent on the number and variety of task flows that have been implemented and stored in task flow models 754, or in other words, on the number and variety of “actionable intents” that the digital assistant recognizes. The effectiveness of the digital assistant, however, also depends on the assistant’s ability to infer the correct “actionable intent(s)” from the user request expressed in natural language.

[0188] In some examples, in addition to the sequence of words or tokens obtained from STT processing module 730, natural language processing module 732 also receives contextual

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

[0189] 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.

[0190] 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).

[0191] 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.

[0192] An actionable intent node, along with its linked property 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.

[0193] 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.”

[0194] 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.

[0195] 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.”

[0196] 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.

[0197] Natural language processing module 732 receives the candidate text representations (e.g., text string(s) or token sequence(s)) from STT processing module 730, and for each candidate representation, determines what nodes are implicated by the words in the candidate text representation. In some examples, if a word or phrase in the candidate text representation is found to be associated with one or more nodes in ontology 760 (via vocabulary index 744), the word or phrase “triggers” or “activates” those nodes. Based on the quantity and/or relative importance of the activated nodes, natural language processing module 732 selects one of the actionable intents as the task that the user intended the digital assistant to perform. In some examples, the domain that has the most “triggered” nodes is selected. In some examples, the domain having the highest confidence value (e.g., based on the relative importance of its various triggered nodes) is selected. In some examples, the domain is selected based on a combination of

the number and the importance of the triggered nodes. In some examples, additional factors are considered in selecting the node as well, such as whether the digital assistant has previously correctly interpreted a similar request from a user.

[0198] 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.

[0199] It should be recognized that in some examples, natural language processing module 732 is implemented using one or more machine learning mechanisms (e.g., neural networks). In particular, the one or more machine learning mechanisms are configured to receive a candidate text representation and contextual information associated with the candidate text representation. Based on the candidate text representation and the associated contextual information, the one or more machine learning mechanisms are configured to determine intent confidence scores over a set of candidate actionable intents. Natural language processing module 732 can select one or more candidate actionable intents from the set of candidate actionable intents based on the determined intent confidence scores. In some examples, an ontology (e.g., ontology 760) is also used to select the one or more candidate actionable intents from the set of candidate actionable intents.

[0200] Other details of searching an ontology based on a token string are 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.

[0201] 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} are 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.

[0202] In some examples, natural language processing module 732 identifies multiple candidate actionable intents for each candidate text representation received from STT processing module 730. Further, in some examples, a respective structured query (partial or complete) is generated for each identified candidate actionable intent. Natural language processing module 732 determines an intent confidence score for each candidate actionable intent and ranks the candidate actionable intents based on the intent confidence scores. In some examples, natural language processing module 732 passes the generated structured query (or queries), including any completed parameters, to task flow processing module 736 (“task flow processor”). In some examples, the structured query (or queries) for the m-best (e.g., m highest ranked) candidate actionable intents are provided to task flow processing module 736, where m is a predetermined integer greater than zero. In some examples, the structured query (or queries) for the m-best candidate actionable intents are provided to task flow processing module 736 with the corresponding candidate text representation(s).

[0203] Other details of inferring a user intent based on multiple candidate actionable intents determined from multiple candidate text representations of a speech input are described in U.S. Utility Application Serial No. 14/298,725 for “System and Method for Inferring User Intent From Speech Inputs,” filed June 6, 2014, the entire disclosure of which is incorporated herein by reference.

[0204] Task flow processing module 736 is configured to receive the structured query (or queries) from natural language processing module 732, complete the structured query, if necessary, and perform the actions required to “complete” the user’s ultimate request. In some examples, the various procedures necessary to complete these tasks are provided in task flow models 754. In some examples, task flow models 754 include procedures for obtaining additional information from the user and task flows for performing actions associated with the actionable intent.

[0205] 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.

[0206] 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.

[0207] 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 generates requests for the service in accordance with the protocols and APIs required by the service according to the service model.

[0208] 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 sends 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.

[0209] 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 processing 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.

[0210] In examples where task flow processing module 736 receives multiple structured queries from natural language processing module 732, task flow processing module 736 initially processes the first structured query of the received structured queries to attempt to complete the first structured query and/or execute one or more tasks or actions represented by the first structured query. In some examples, the first structured query corresponds to the highest ranked actionable intent. In other examples, the first structured query is selected from the received structured queries based on a combination of the corresponding speech recognition confidence scores and the corresponding intent confidence scores. In some examples, if task flow processing module 736 encounters an error during processing of the first structured query (e.g., due to an inability to determine a necessary parameter), the task flow processing module 736 can proceed to select and process a second structured query of the received structured queries that corresponds to a lower ranked actionable intent. The second structured query is selected, for example, based on the speech recognition confidence score of the corresponding candidate text representation, the intent confidence score of the corresponding candidate actionable intent, a missing necessary parameter in the first structured query, or any combination thereof.

[0211] Speech synthesis processing module 740 is configured to synthesize speech outputs for presentation to the user. Speech synthesis processing 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 processing module 740 converts the text string to an audible speech output. Speech synthesis processing 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 processing 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 processing module 740 is configured to directly process the phonemic string in the metadata to synthesize the word in speech form.

[0212] In some examples, instead of (or in addition to) using speech synthesis processing 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.

[0213] 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. System for generating an avatar

[0214] FIG. 8 illustrates system 800 for generating and animating an avatar in accordance with some examples. In some examples, system 800 is implemented on one or more electronic devices (e.g., devices 104, 122, 200, 400, 600, or 900) and the modules and functions of system 800 may be distributed in any manner between the devices. In some examples, some of the modules and functions of system 800 are divided into a server portion and a client portion, where the client portion resides on one or more user devices (e.g., devices 104, 122, 200, 400, 600, and 900) and communicates with the server portion (e.g., server system 108) through one or more networks, e.g., as shown in FIG. 1. System 800 is implemented using hardware, software, or a combination of hardware and software to carry out the principles discussed herein. In some

examples, the module and functions of system 800 are implemented within a digital assistant module, as discussed above with respect to FIGS. 7A-7C.

[0215] It should be noted that system 800 is exemplary, and thus system 800 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. Further, although the below discussion describes functions being performed at a single component of system 800, it is to be understood that such functions can be performed at other components of system 800 and that such functions can be performed at more than one component of system 800.

[0216] As illustrated in FIG. 8, system 800 includes neural network 810, avatar animator 820, and output processor 830. As discussed in more detail below, system 800 receives text 802 and determines an emotional state.

[0217] In some examples, as discussed further below, system 800 determines an emotional state based on received text 802 and thus, does not require any further indications of emotional state to be provided. In other examples, system 800 may receive indication of emotional state 804, e.g., by prompting a user or from explicit user input. In other examples, system 800 may optionally be provided indication of emotional state 804 with text 802 and determine the emotional state based on both the provided indication of emotional state 804 and the text 802.

[0218] System 800 then generates a speech output (e.g., speech output 832) of text 802 and a set of parameters (e.g., animation parameters 814) representing one or more movements of an avatar (e.g., avatar 906, 914, 920) based on text 802 and the determined emotional state. In this way, more realistic avatars may be generated with greater flexibility. In particular, different movements of the avatar may be informed by the determined emotional state to make the avatar appear more realistic and lifelike. In some examples, the determined emotional state may affect a pose of the avatar. For example, when the determined emotional state is “happy” the avatar may be animated to tilt its head upwards in a pose that indicates this emotional state. Similarly, in some examples, the determined emotional state of may affect movements of various features of the face including movements of the mouth, nose, ears etc. For example, when the determined emotional state is “happy” the avatar may be animated so that the mouth movements resemble a smile, the eyes are opened more and are directed towards a viewer, etc. It will be appreciated that this is merely one example and that the avatar can be influenced by many different emotional

states including happy, sad, angry, neutral, etc. Several different examples are further provided below.

[0219] In some examples, text 802 is received from an input/output component of a user device, server, or a combination of the two (e.g., I/O Subsystem 206, I/O interface 430, and I/O processing module 728). In some examples, text 802 is received from a user. For example, a user may provide text 802 through a user interface of an application (e.g., a messaging application such as a text messaging application, an email application etc.). In some examples, text 802 is received from another electronic device. For example, system 800 may be located on a server and text 802 may be received from a user device located remotely to the server. In some examples, text 802 is received from another application located on a user device. For example, a user may provide text 802 in a messaging application and then indicate they would like to generate an avatar based on text 802 provided to the messaging application.

[0220] In some examples, system 800 processes text 802 before providing text 802 to neural network 810. In some examples, system 800 processes text 802 to determine a representation of text 802. In some examples, the representation of text 802 is a phoneme sequence corresponding to text 802. In some examples, text 802 is determined from a larger text input before being provided to neural network 810. For example, as shown in FIG. 9A and discussed further below, a user provides text input 904 of “send ‘Happy Birthday!’ to my mom,” to a digital assistant of user device 900. The digital assistant may then process the entire text input 904 to determine that text 802 to be provided to neural network 810 should include “Happy Birthday!” and not “send...to my mom.” Thus, neural network 810 receives text 802 to be included in speech output 812 after pre-processing text input 904 by the digital assistant.

[0221] In some examples, text 802 is a textual representation of a digital assistant output. For example, as shown in FIG. 9C and discussed further below, the user may provide input request 916 of “What’s the temperature in Palo Alto?” to the digital assistant of device 900. The digital assistant may then process the user request to determine an appropriate response using one or more modules discussed above with relation to FIGS. 7A-7C including but not limited to input/output processing module 728, speech-to-text (STT) processing module 730, natural language processing module 732, ontology 760, vocabulary index 744, user data 748, task flow models 754, service models 756, and ASR systems 758, at the user device or at another

electronic device such as a server. The digital assistant may then provide the response including “It’s 48°F outside.” as text 802 to system 800 to determine speech output 812 and avatar data 822 for display as avatar 920. Text 802 may then be provided as output 918 along with avatar 920 representing the digital assistant on display 902 of user device 900.

[0222] In some examples, the digital assistant processes the digital assistant output to determine a portion of the output to be provided as text 802. For example, the user may provide an input request at a user device for a digital assistant such as “What restaurants are near me?” In response to the user request “What restaurants are near me?” the digital assistant may determine that the appropriate response is “I’ve found these restaurants near your current location,” followed by a list of nearby restaurants. The digital assistant may determine that “I’ve found these restaurants near your current location,” should be provided as text 802 to system 800 as that is the portion of the output that is desired to be provided as speech output and that the listing of restaurants should be displayed on a screen of the user device. Thus, system 800 will only use the portion of the output provided as text 802 to generate speech output 812 and avatar data 822, and will not use the list of restaurants.

[0223] In some examples, system 800 optionally receives indication of emotional state 804. In some examples, indication of emotional state 804 is determined based on a user input received concurrently with a user input providing text 802. For example, a user may provide a text input of “send ‘Happy Birthday!’ to my mom with a happy avatar,” (e.g., through a user interface of a user device). Based on the input provided by the user, system 800 receives text 802 “Happy Birthday!” and indication of emotional state 804 “happy” based on the text input provided by the user.

[0224] In some examples, indication of emotional state 804 is determined by another neural network separate from neural network 810 before being provided to neural network 810. In some examples, indication of emotional state 804 is determined based on text 802 by another neural network separate from neural network 810. For example, text 802 may be provided to a neural network trained to determine indication of emotional state 804, such as the classifier discussed with regard to training neural network 810 below. This neural network may then evaluate text 802 to determine indication of emotional state 804, using techniques similar to those discussed with the evaluation of text 802 by neural network 810 to determine an emotional state discussed

below. Thus, the separate neural network (e.g., a classifier) may evaluate the individual words, punctuation, or phrases of text 802 to determine if they provide any indication of a particular emotional state (e.g., “happy”). The separate neural network may then provide this determined emotional state as indication of emotional state 804 to neural network 810.

[0225] In some examples, indication of emotional state 804 is received separately from text 802. For example, as shown in FIG. 9B and discussed further below, a user may provide text input 908 of “send ‘I’m not sure I can make it’ to my mom with an avatar,” and then separately indicate that the avatar should have an emotional state of “sad” in another input (e.g., input 912) provided to user device 900. In some examples, indication of emotional state 804 is based on a selection of possible emotional states provided to a user. For example, the user may provide the text input “send ‘Happy Birthday!’ to my mom with an avatar,” and a user device may provide a selection menu of possible emotional states including happy, angry, sad, etc. The user may then select “happy” as the desired emotional state which is provided to system 800 as indication of emotional state 804.

[0226] In some examples, indication of emotional state 804 is based on contextual data of the user or the user device. As discussed above, the contextual data includes user-specific data, vocabulary, and/or preferences relevant to the user input. In some examples, the contextual data includes software and hardware states of the user device at the time the text input is received, and/or information related to the surrounding environment of the user at the time that the text input was received. In some examples, the contextual data includes sensor information collected before, during, or shortly after the text input is received, prior interactions (e.g., dialogue) between a digital assistant of the user device and the user, and the like. In some examples, the contextual data includes one or more default settings of the user or the user device.

[0227] In some examples, system 800 provides text 802 to neural network 810. In some examples, neural network 810 includes a neural network trained to determine speech data set 812, and animation parameters 814 based on text 802. Neural network 810 may include any type of neural network capable of receiving text and determining speech data sets and avatar animation parameters. For example, neural network 810 may include a tacotron network, a modified tacotron network, a long short-term memory (LSTM) network, a recurrent neural network (RNN), any other suitable network, or any combination of these or other suitable

networks. In some examples, neural network 810 includes a machine learning model trained to determine speech data set 812 and animation parameters 814 based on text 802. In some examples, neural network 810 includes one or more of an encoder, a decoder, or any other suitable modules, which may be implemented with a tacotron network, a modified tacotron network, a LSTM network, a RNN network, any other suitable network, or any combination of these or other suitable networks.

[0228] In some examples, neural network 810 determines an emotional state. In some examples, neural network 810 determines the emotional state based on received text 802. In some examples, neural network 810 determines whether received text 802 or a representation of received text 802 includes one or more words that indicate a particular emotional state. For example, when text 802 includes “Happy Birthday!” neural network 810 may determine that text 802 includes the word “happy” and thus, that text 802 indicates an emotional state of happy. As another example, when text 802 includes “I fell and hurt myself” neural network 810 determines that text 802 includes the word “hurt” and thus, indicates an emotional state of sad or worried.

[0229] In some examples, neural network 810 determines whether text 802 or a representation of received text 802 includes punctuation that indicates a particular emotional state. For example, when text 802 includes “Happy Birthday!” as previously discussed, neural network 810 determines that text 802 indicates an emotional state of happy based on the use of an exclamation point. In some examples, neural network 810 determines an emotional state based on words and punctuation included in text 802. For example, neural network 810 may determine that the presence of an exclamation point in “Happy Birthday!” could indicate an emotional state of happy or angry. Thus, by determining that text 802 also includes the word “happy” neural network 810 may determine that text 802 indicates an emotional state of happy and not angry.

[0230] In some examples, when system 800 receives an indication of emotional state 804, system 800 provides indication of emotional state 804 to neural network 810. In some examples, neural network 810 determines an emotional state based on indication of emotional state 804 without consideration of text 802. For example, when indication of emotional state 804 is “happy” neural network 810 determines the emotional state is “happy” regardless of text 802. In some examples, neural network 810 determines the emotional state based on text 802 and

indication of emotional state 804. For example, neural network 810 may determine that the presence of an exclamation point in text 802 of “Happy Birthday!” could indicate an emotional state of happy or angry and thus may reference indication of emotional state 804, which is “happy.” Thus, neural network 810 may determine that the emotional state is “happy.”

[0231] In some examples, neural network 810 determines a vector representation of text 802 and a vector representation of the determined emotional state. In some examples, the vector representation of text 802 and the vector representation of the determined emotional state are determined by an encoder of neural network 810. In some examples, neural network 810 determines a vector representation of both text 802 and the determined emotional state. In some examples, neural network 810 determines a vector representation of both text 802 and the determined emotional state by concatenating the vector representation of text 802 and the vector representation of the determined emotional state. In some examples, the vector representation of both text 802 and the determined emotional state is a first intermediate vector representing text 802 and the determined emotional state. In some examples, the first intermediate vector representing text 802 and the determined emotional state are determined by an encoder of neural network 810.

[0232] In some examples, neural network 810 determines speech data set 812 and animation parameters 814 based on the first intermediate vector representing text 802 and the determined emotional state. Thus, the outputs of neural network 810, including speech data set 812 and animation parameters 814 are based on both text 802 and the determined emotional state, and thus are based on both text 802 and optionally indication of emotional state 804. In some examples, speech data set 812 and animation parameters 814 are determined using a decoder of neural network 810.

[0233] In some examples, neural network 810 determines (e.g., using a decoder) a second intermediate vector from the first intermediate vector representing text 802 and the determined emotional state. For example, neural network 810 may determine a second intermediate vector corresponding to speech data set 812 and animation parameters 814 from the first intermediate vector representing text 802 and the determined emotional state. In some examples, neural network 810 applies a linear projection to the second intermediate vector to determine speech data set 812 and animation parameters 814. In some examples, neural network 810 applies a

different linear projection to the second intermediate vector to determine each respective output. For example, neural network 810 may apply a first linear projection to the second intermediate vector to determine speech data set 812, and a second linear projection to the second intermediate vector to determine animation parameters 814.

[0234] In some examples, neural network 810 determines speech data set 812 based on the vector representation of text 802 or the vector representation of the determined emotional state without considering the other vector representation. For example, speech data set 812 may be generated the same regardless of the determined emotional state and thus may be determined based on the vector representation of text 802 without considering the determined emotional state. Similarly, in some examples, neural network 810 determines animation parameters 814, or a portion of animation parameters 814, based on the vector representation of the determined emotional state or the vector representation of text 802 without considering the other vector representation. Thus, neural network 810 may determine some outputs or a portion of the outputs (e.g., speech data set 812 and animation parameters 814) based on text 802 or indication of emotional state 804 without requiring any other inputs.

[0235] In some examples, animation parameters 814 include one or more sets of parameters. For example, animation parameters 814 may include a first set of parameters for animating movements of the avatar and a second set of parameters for animating a pose of the avatar. Accordingly, neural network 810 may determine each set of parameters included in animation parameters 814 (e.g., the first set of parameters and the second set of parameters) based on the first intermediate vector representing text 802 and the determined emotional state. For example, as discussed above, neural network 810 may determine a second intermediate vector and then apply a respective linear projection to the second intermediate vector to determine each respective output including speech data set 812, the first set of parameters for animating movements of the avatar, and the second set of parameters for animating the pose of the avatar.

[0236] In some examples, each set of parameters included in animation parameters 814 (e.g., the first set of parameters and the second set of parameters) are based on the vector representation of text 802 or the vector representation of the determined emotional state without considering the other vector representation. For example, the first set of parameters for animating movements of the avatar may not change when the determined emotional state changes. Thus,

the first set of parameters for animating movement of the avatar may be determined based only on the vector representation of text 802 and may not consider the determined emotional state. Similarly, the second set of parameters for animating the pose of the avatar may not change when text 802 changes. Thus, the second set of parameters for animating the pose of the avatar may be determined based only on the vector representation of the determined emotional state and may not consider text 802.

[0237] In some examples, prior to receiving text 802 and determining speech data set 812 and animation parameters 814, neural network 810 is trained to determine the speech data set and the set of parameters representing one or more movements of the avatar. To train neural network 810, a set of training text data is provided to neural network 810 along with a reference speech data set and a reference set of parameters representing one or movements of an avatar. Neural network 810 then determines a training speech data set and a set of training parameters representing the movement of the avatar based on the training text data. The training speech data set and the set of training parameters representing the movement of the avatar are then be evaluated to determine if neural network 810 has been successfully trained. In particular, the training speech data set and the set of training parameters representing the movement of the avatar are compared to the reference speech data set and the reference set of parameters representing one or more movements of the avatar respectively. Typically, the training of neural network 810 is accomplished in an iterative manner, by providing training text data to neural network 810 repeatedly so that neural network 810 learns to synthesize a training speech data set and a set of training parameters representing the movement of the avatar that approximates the reference speech data set and the reference set of parameters representing one or more movements of the avatar respectively.

[0238] In some examples, the same training text data may be provided to neural network 810 in an iterative manner. When training the neural network 810 in this manner, the training speech data set and training parameters determined by neural network 810 will become progressively more similar to the reference speech data set and the reference set of parameters representing one or more movements of the avatar respectively. For example, the training text data provided to neural network 810 may be “Happy Birthday.” When this training text is first provided to neural network 810 the determined training speech data set and training parameters may be very different from the reference speech data set representing “Happy Birthday” with reference

parameters that will animate an avatar to make movements as if saying “Happy Birthday.” As the training text is provided to neural network 810 more times during the course of training neural network 810 the training speech data set and training parameters will change until the training speech data set approximates the reference speech data set representing “Happy Birthday” and the training parameters approximate the reference parameters that animate an avatar to make the movements as if saying “Happy Birthday.”

[0239] Similarly, in some examples, training text data including different words or phrases may be provided to neural network 810 during the course of training neural network 810. This training text data including different words or phrases may also be provided to neural network 810 iteratively to effectively train neural network 810 to produce many different speech data sets and parameters as speech data set 812 and animation parameters 814, respectively.

[0240] As discussed above, after each training speech data set and set of training parameters is determined by neural network 810 they are evaluated to determine if neural network 810 has been successfully trained. In some examples, neural network 810 is determined to have been successfully trained when the training speech data set and the set of training parameters sufficiently match a reference speech data set and set of parameters. In some examples, neural network 810 is determined to have been successfully trained when a plurality of training speech data sets and sets of training parameters are correctly determined based on a plurality of sets of training text data. Once it is determined that neural network 810 has been successfully trained then the system may forgo providing training text data to neural network 810.

[0241] In some examples, after training of neural network 810 is complete, neural network 810 is conditioned to recognize and determine an emotional state. In some examples, conditioning neural network 810 to recognize and determine emotional state includes conditioning neural network 810 to determine speech data set 812 and animation parameters 814 based on the emotional state. To condition neural network 810 to recognize and determine an emotional state a predetermined emotional state is provided to neural network 810. In some examples, the predetermined emotional state is provided to neural network 810 as a vector representation of the predetermined emotional state.

[0242] In some examples, the predetermined emotional state is generated by a classifier trained independently to determine an emotional state. The classifier can be provided with

various sets of training text data and a corresponding emotional state of each set of training text data. For example, the classifier may be provided with the training text data “Happy Birthday!” and the corresponding emotional state of “happy.” The classifier may be provided training text data and the corresponding emotional states in a supervised iterative manner until the classifier can correctly identify from the text one of several different emotional state (e.g., happy, sad, angry, neutral, etc.). Once the classifier is trained an intermediate representation of the predetermined emotional state is extracted and then provided to neural network 810. In some examples, an intermediate representation of each different predetermined emotional state (e.g., happy, sad, angry, neutral, etc.) is extracted and provided to neural network 810.

[0243] In some examples, the predetermined emotional state is generated by a classifier trained independently to determine an emotional state based on speech data. The classifier can be provided with various sets of training speech data and a corresponding emotional state of each set of training speech data. For example, the classifier may be provided with the training speech data “Happy Birthday!” and the corresponding emotional state of “happy.” Similarly, the classifier may be provided with the training speech data “I’m sorry” and the corresponding emotional state of “sad.” The classifier may be provided training speech data and the corresponding emotional states in a supervised iterative manner until the classifier can correctly identify from the speech one of several different emotional states (e.g., happy, sad, angry, neutral, etc.). Once the classifier is trained an intermediate representation of the predetermined emotional state is extracted and then provided to neural network 810. In some examples, an intermediate representation of each different predetermined emotional state (e.g., happy, sad, angry, neutral, etc.) is extracted and provided to neural network 810.

[0244] In some examples, neural network 810 is provided with the predetermined emotional state and a set of training text data representative of the predetermined emotional state. For example, when an intermediate representation of the emotional state “happy” is provided to neural network 810 the training text data “Happy Birthday!” is also provided to teach neural network 810 the correlation between similar text data and the emotional state “happy.” Neural network 810 may then generate a training speech data set and a set of training parameters that represent the predetermined emotional state (e.g. happy). As discussed above with respect to training neural network 810, this process of providing a predetermined emotional state and training text data may be repeated iteratively until the training speech data set and set of training

parameters sufficiently matches a set of reference training speech data set and a set of reference parameters respectively. Further, the process may be repeated iteratively with a plurality of sets of training text data and a plurality of predetermined emotional states until neural network 810 is conditioned to recognize many different emotional states and produce speech data sets and sets of animation parameters based on those different emotional states.

[0245] In some examples, the set of training parameters that represent the predetermined emotional state is generated in a similar manner to animation parameters 814. Thus, the set of training parameters is generated by determining a vector representation of the training text data and a vector representation of the predetermined emotional state. In some examples, the vector representation of the predetermined emotional state is the intermediate representation extracted from the trained classifier discussed above. In other examples, the vector representation of the predetermined emotional state is determined by neural network 810 based on another indication of the predetermined emotional state. Once the vector representation of the training text data and the vector representation of the predetermined emotional state are obtained or generated by neural network 810 the vector representations may be concatenated together to create an intermediate vector representing both the training text data and the predetermined emotional state. Neural network 810 may then apply the appropriate linear projection(s) to the intermediate vector representing both the training text data and the predetermined emotional state to determine the training speech data and/or one or more sets of training parameters.

[0246] In some examples, neural network 810 is trained to generate speech outputs and conditioned to recognize emotional states in separate consecutive steps as discussed above. In some examples, neural network 810 is trained to generate speech outputs and conditioned to recognize emotional states at the same time (e.g., simultaneously). Thus, the iterative process of providing training text data and predetermined emotional states to generate training speech data sets and training sets of parameters may overlap and occur at the same time so that neural network 810 is progressively trained to generate both speech data sets and animation parameters based on text data and determined emotional states.

[0247] In some examples, neural network 810 optionally comprises a plurality of neural networks. In some examples, the plurality of neural networks functions similarly to the description of neural network 810 described above. In some examples, the plurality of neural

networks includes a first neural network trained to generate a speech data set representing text 802 based on text 802 and a determined emotional state and a second neural network trained to generate a set of parameters for animating an avatar based on the generated speech data set. In some examples, the plurality of neural networks further includes a neural network trained to determine the emotional state based on text 802 or optionally received indication of emotional state 804. In some examples, the first neural network trained to generate the speech data set is trained to determine the emotional state.

[0248] When neural network 810 comprises the plurality of neural networks as described above, the plurality of neural networks may be trained and conditioned consecutively or simultaneously as described above in relation to neural network 810. Therefore, in some examples, the plurality of neural networks may first be trained to generate the speech data set and the set of parameters for animating an avatar and then conditioned to recognize and generate the outputs based on an emotional state and in some examples, the plurality of neural networks may be trained to generate the outputs and conditioned to recognize the emotional state at the same time.

[0249] In some examples, an audio input may be provided to the plurality of neural networks. In some examples, the audio input is received from a user and includes a request to generate an avatar. In some examples, the audio input may be processed prior to being provided to the plurality of neural networks to determine the user's intent and determine which portions of the audio input are to be provided to the plurality of networks. Additionally, the audio input may be processed to determine a representation of the audio input. In some examples, the representation of audio input is a phoneme sequence corresponding to the audio input.

[0250] In some examples, the audio input or a representation of the audio input is processed to determine whether to provide the audio input to one or more of the plurality of neural networks. In some examples, determining whether to provide the representation of the audio input to one or more of the plurality of neural networks includes determining whether to provide the representation of the audio input to the second neural network. In some examples, determining whether to provide the representation of the audio input to the second neural network includes determining whether the audio input is to be provided with the avatar. In some examples, determining whether the audio input is to be provided with the avatar includes

determining that an audio output based on a text input is not to be generated. In some examples, in accordance with a determination that the audio input is to be provided with the avatar and/or a determination that the representation of the audio input is to be provided to the second neural network, the representation of the audio input is provided to the neural network.

[0251] In some examples, the audio input or a representation of the audio input is also provided to another neural network to determine an emotional state associated with the audio input, such as the classifier trained to recognize the emotional state based on audio data discussed above. In some examples, an emotional state is determined based on another input provided by the user indicating an emotional state. Once an emotional state is determined, the determined emotional state may also be provided to the second neural network.

[0252] Once the audio input or the representation of the audio input and the determined emotional state are provided to the second neural network, the second neural network determines a set of parameters for animating an avatar based on the representation of the audio input and the determined emotional state. In this way an avatar that realistically moves as if providing the user's provided input may be generated and may convey the desired emotional state along with the provided audio input.

[0253] For example, a user may provide the audio input "Send an avatar saying 'Happy Birthday!' to my Mom," to a digital assistant. The digital assistant may then determine based on the audio input that the user wishes to generate and send an avatar to a contact of "Mom" with the audio recording of the user saying "Happy Birthday!" The digital assistant may further provide the portion of the audio input to be sent with the avatar (e.g., "Happy Birthday!") to a neural network (e.g., a classifier) to determine that the emotional state of the audio input is "happy." Alternatively, the user may provide the emotional state in an input by specifying that the emotional state should be "happy." A representation of the selected portion of the audio input can then be provided with the determined emotional state of "happy" to the second neural network which generates a set of parameters for animating the avatar so as to make the avatar appear happy and speak "Happy Birthday!" The parameters for animating the avatar and the audio recording of the user saying "Happy Birthday!" may later be processed (e.g., by output processor 830) to synchronize the avatar with the audio recording so that the avatar can be provided with the audio recording, as discussed further below.

[0254] It will be recognized that this is an example of a multi-network implementation of neural network 810 and that such multi-network implementations may include as many or as few different neural networks as needed. Thus, in some examples, neural network 810 may include three, four, five, six, or more neural networks that all may be trained consecutively or simultaneously to produce the desired outputs.

[0255] As illustrated by FIG. 8, system 800 includes avatar animator 820. In some examples, avatar animator 820 receives animation parameters 814 from neural network 810. Once avatar animator 820 receives animation parameters avatar animator 820 generates avatar data 822 for animating an avatar (e.g., avatar 906, 914, 920) using animation parameters 814. In some examples, avatar animator 820 animates different portions or movements of the avatar with different portions of animation parameters 814. In some examples, avatar animator 820 animates a first feature of the avatar (e.g., head pose) with a first set of parameters of animation parameters 814 (e.g., pose animation parameters) and a second feature of the avatar (e.g., lip movements) with a second set of parameters of animation parameters 814 (e.g., movement animation parameters). For example, when neural network 810 determines that the emotional state is “happy” animation parameters 814 may include a first set of parameters for a “happy” head pose which avatar animator 820 uses to generate avatar data 822 including a head pose of looking up. Similarly, animation parameters 814 may also include a second set of parameters for “happy” movements which avatar animator 820 uses to generate avatar data 822 including lips that are upturned, resembling a smile. In some examples, animation parameters 814 may include a plurality of sets of parameters with each set of parameters corresponding to a different feature of the avatar (e.g., pose, lips, nose, eyebrows, ears, etc.). In some examples, avatar animator 820 uses each set of the plurality of sets of parameters corresponding to a different feature of the avatar to generate a portion of avatar data 822.

[0256] In some examples, avatar animator 820 uses animation parameters 814 to generate avatar data 822 for animating a feature of the avatar’s (e.g., avatar 906, 914, 920) face and/or a state of a feature of the avatar’s face. For example, avatar animator 820 may generate avatar data 822 used to animate a specific position of many different points of the avatar’s lips, how far away the avatar’s top lip is from its nose, the shape of the avatar’s lips, whether the avatar’s cheeks are convex or concave, the distance between the avatar’s eyebrows, the distance between

the avatar's eyebrows and their eyes, whether the avatar's forehead or nose is wrinkled, the position of the avatar's ears, the direction of the avatar's eyes, and if the avatar's eyes are closed.

[0257] In some examples, avatar animator 820 determines a type of avatar to use when generating avatar data 822. In some examples, the types of avatars include different types of animals, human-like avatars, robots, aliens, anthropomorphic inanimate objects, or any other type of avatar. In some examples, avatar animator 820 determines the type of avatar to generate based on a selection by a user. For example, a user may select a desired type of avatar on a user interface of a user device or with a voice input that is then provided to system 800 and provided to avatar animator 820. In some examples, avatar animator 820 determines the type of avatar based on a default setting. In some examples, the avatar animator 820 determines the type of avatar based on a user specific default setting.

[0258] As illustrated by FIG. 8, system 800 also includes output processor 830. In some examples, output processor 830 processes avatar data 822 to produce an avatar (e.g., avatar 906, 914, 920) for display. In some examples, output processor 830 receives speech data set 812 generated by neural network 810 and avatar data 822 generated by avatar animator 820. In some examples, output processor 830 processes speech data set 812 to generate speech output 832. In some examples, output processor 830 includes a neural network (e.g., a WaveRNN) trained to generate speech output 832 based on speech data set 812.

[0259] In some examples, output processor 830 synchronizes at least one movement of avatar data 822 with speech output 832. In some examples, output processor 830 synchronizes the movement of at least one feature of avatar data 822 with speech output 832. For example, for the speech output "Happy Birthday!" output processor 830 may synchronize the movement of the avatar's lips or at least a portion of the movement of the avatar's lips with the provided audio. In some examples, output processor 830 does not synchronize the movement of at least one feature of the avatar with speech output 832. For example, output processor 830 may not synchronize the pose or change in pose of the avatar with speech output 832.

[0260] In some examples, after avatar data 822 and speech output 832 are processed with output processor 830, the avatar (e.g., avatar 906, 914, 920) is displayed on a screen of an electronic device such as screen 902 as shown in FIGS. 9A-9C. In some examples, the avatar is displayed on screen 902 as avatar 906 to provide a preview of avatar 822 that can be edited,

changed, or deleted. In some examples, the avatar may be provided to an application for use by that application. In some examples, the avatar may be included in a message to the user or another user. In some examples, the avatar may be displayed as avatar 920 on screen 902 to provide an interactive digital assistant to the user.

[0261] FIGS. 9A-9C illustrate generated and animated avatars, according to various examples. FIG. 9A illustrates an example in which a user requests a message be sent to a contact “Mom” and an avatar 906 is generated based on default settings. FIG. 9A shows that an input 904 of “Send ‘Happy Birthday! to my Mom with an avatar,” is provided by a user to user device 900 during an interaction with a digital assistant. In some examples, input 904 is provided as a spoken input. In some examples, input 904 is provided as a text input entered with a user interface of user device 900 or the digital assistant which may include, for example, a keyboard or other affordance for providing text.

[0262] As discussed above, input 904 is pre-processed to determine what portion of input 904 should be provided to system 800 as text 802. In this example, user device 900 or the digital assistant may pre-process input 904 to determine that “Happy Birthday!” is to be provided to system 800 as text 802 to determine an audio output and animate an avatar. In particular, user device 900 or the digital assistant may perform natural language understanding as described above to determine that the user’s intent is to send a message to a contact “Mom” in a phone book application and thus that “send...to my Mom with an avatar,” is not part of the intended message.

[0263] After pre-processing the input, text 802 is provided to system 800 and thus to neural network 810. Neural network 810 then determines based on “Happy Birthday!” that the emotional state is “happy” based on the use of the word “Happy” in text 802, as well as the use of an exclamation point. Neural network 810 then determines speech data set 812 representing “Happy Birthday!” and animation parameters 814 as a set of parameters representing one or more movements of avatar 906. Based on the determined emotional state of “happy” animation parameters 814 may include parameters that cause the one or more movements of avatar 906 to indicate a “happy” emotional state. For example, the parameters may cause the head pose of avatar 906 to look up, which is generally associated with a person being happy, or the parameters may cause the lips of avatar 906 to make a movement similar to a smile.

[0264] Animation parameters 814 are then received by avatar animator 820, which generates avatar data 822 for animating avatar 906. In this example, avatar animator 820 determines which avatar to generate as avatar 906 based on a default setting of “monkey.” Avatar animator 820 then generates avatar data 822 for animating avatar 906 based on the set of parameters representing one or more movements of avatar 906 (e.g., animation parameters 814). As discussed above, in this case, avatar animator 820 generates avatar data 822 for animating avatar 906 to look up (exposing the neck of avatar 906) to indicate the emotional state of “happy” in addition to the other movements representing avatar 906 speaking the provided text.

[0265] Avatar data 822 is then provided to output processor 830 along with speech data set 812. Output processor 830 generates speech output 832 based on speech data set 812 and animates avatar data 822 as avatar 906. Output processor 830 then synchronizes some of the movements of avatar 906, in this case movements of the mouth, ears, nose, etc., with speech output 832 so that avatar 906 appears to be speaking speech output 832 when speech output 832 is provided. Output processor 830 then provides avatar 906 on display 902 of user device 900 with speech output 832 so that the user may review the preview of avatar 906. The user may then select to edit avatar 906 or send 906 in the message to “Mom.”

[0266] FIG. 9B illustrates an example in which a user requests a message be sent to a contact “Mom” and further information is requested to generate avatar 914. FIG. 9B shows that an input 908 of “send ‘I’m not sure I can make it’ to my Mom with an avatar” is provided to device 900 and displayed on screen 902. In this examples, input 908 is pre-processed to determine that the portion of input 908 to be provided as text 802 to system 800 is “I’m not sure I can make it.” During pre-processing, the digital assistant may recognize that there is not enough information within input 908 for neural network 810 to determine an emotional state. Thus, the digital assistant provides the output “Okay, what emotion would you like to convey,” and further asks the user “which avatar would you like to generate?” as output 910 on screen 902. The user then provides input 912 specifying the emotional state as “sad” and adding that the desired avatar is the “robot avatar.” In some examples, text 802 may be provided to neural network 810 and after neural network 810 determines it is unable to determine an emotional state based on text 802, the digital assistant provides output 910 requesting more information from the user.

[0267] User device 900 and/or the digital assistant may then process input 912 to determine indication of emotional state 804 to be provided to neural network 810 is “sad.” Neural network 810 then receives text 802 of “I’m not sure I can make it,” and indication of emotional state 804 of “sad.” Neural network 810 then determines speech data set 812 representing “I’m not sure I can make it” and animation parameters 814 as a set of parameters representing one or more movements of avatar 914. Based on the determined emotional state of “sad” animation parameters 814 may include parameters that cause the one or more movements of avatar 914 to indicate a “sad” emotional state. For example, the parameters may cause the head pose of avatar 914 to look down, which is generally associated with a person being sad, or the parameters may cause the lips of avatar 914 to make a movement similar to a frown.

[0268] Animation parameters 814 are then received by avatar animator 820, which generates avatar data 822 for animating avatar 914. In this example, avatar animator 820 determines which avatar to generate as avatar 914 based on the input 912 which specifies that the avatar should be a “robot.” Avatar animator 820 then generates avatar data 822 for animating avatar 914 based on the set of parameters representing one or more movements of avatar 914 (e.g., animation parameters 814). As discussed above, in this case, avatar animator 820 generates avatar data 820 for animating avatar 914 to look down (exposing the top/back of the head of avatar 914) to indicate the emotional state of “sad” in addition to the other movements representing avatar 914 speaking the provided text.

[0269] Avatar data 822 is then provided to output processor 830 along with speech data set 812. Output processor 830 generates speech output 832 based on speech data set 812 and animates avatar data 822 as avatar 914. Output processor 830 then synchronizes some of the movements of avatar 914, in this case movements of the mouth, ears, nose, etc., with speech output 832 so that avatar 914 appears to be speaking speech output 832 when speech output 832 is provided. Output processor 830 then provides avatar 914 on display 902 of user device 900 with speech output 832 so that the user may review the preview of avatar 914. The user may then select to edit avatar 914 or send 914 in the message to “Mom.”

[0270] FIG. 9C illustrates an example in which a digital assistant provides a message to be provided to a user along with avatar 920 representing the digital assistant. FIG. 9C shows that a user has provided input 916 of “What’s the temperature in Palo Alto?” to a digital assistant of

user device 900. As discussed above, in some examples, this user input may be provided as a spoken input and in some examples, the user input may be provided as a text input through a user interface or affordance such as a keyboard.

[0271] After receiving input 916, the digital assistant of user device 900 may determine the user intent is to determine the current temperature in Palo Alto based on natural language understanding techniques as described above. The digital assistant then determines the appropriate response to the user's question is "It's 48°F outside." The digital assistant may further process this response to determine that it should be provided to neural network 810 as text 802.

[0272] Neural network 810 then receives text 802 including "It's 48°F outside." Based on text 802, neural network 810 determines that the emotional state of text 802 is "neutral." Neural network 810 determines this emotional state based on the lack of any words in the text with a strong emotional connotation, as well as the lack of strong punctuation. Neural network 810 may further determine this emotional state because text 802 is a response to an informational questions and thus, is not indicative of any particular emotion. Based on text 802 and the determined emotional state of "neutral," neural network 810 determines speech data set 812 and animation parameters 814 as a set of parameters representing one or more movements of avatar 920. Based on the determined emotional state of "neutral" animation parameters 814 may include parameters that cause the one or more movements of avatar 914 to indicate a "neutral" emotional state. For example, the parameters may cause the head pose of avatar 914 to look straight ahead, which is not generally associated with any particular emotion.

[0273] Animation parameters 814 are then received by avatar animator 820, which generates avatar data 822 for animating avatar 920. In this example, avatar animator 820 determines which avatar to generate as avatar 920 based on a default setting of how the digital assistant should be represented as an avatar. Avatar animator 820 then generates avatar data 822 for animating avatar 920 based on the set of parameters representing one or more movements of avatar 920 (e.g., animation parameters 814). As discussed above, in this case, avatar animator 820 generates avatar data 822 for animating avatar 920 to look straight ahead to indicate the emotional state of "neutral" in addition to the other movements representing avatar 920 speaking the provided text.

[0274] Avatar data 822 is then provided to output processor 830 along with speech data set 812. Output processor 830 generates speech output 832 based on speech data set 812 and animates avatar data 822 as avatar 920. Output processor 830 then synchronizes some of the movements of avatar 920, in this case movements of the mouth, ears, nose, etc., with speech output 832 so that avatar 920 appears to be speaking speech output 832 when speech output 832 is provided. Output processor 830 then provides avatar 920 on display 902 of user device 900 with speech output 832 to communicate the response to the user's question. In the example shown in FIG. 9C, the digital assistant also provides text 802 as output 918 on display 902 of user device 900, to provide the user with different ways to review the digital assistant's response to their question.

[0275] FIGS. 10A-10C illustrates process 1000 for generating and animating an avatar, according to various examples. Process 1000 is performed, for example, using one or more electronic devices implementing system 800. In some examples, process 1000 is performed using a client-server system (e.g., system 100), and the blocks of process 1000 are divided up in any manner between the server (e.g., DA server 106) and a client device. In other examples, the blocks of process 1000 are divided up between the server and multiple client devices (e.g., a mobile phone and a smart watch). Thus, while portions of process 1000 are described herein as being performed by particular devices of a client-server system, it will be appreciated that process 1000 is not so limited. In other examples, process 1000 is performed using only a client device (e.g., user device 104) or only multiple client devices. In process 1000, some blocks are, optionally, combined, the order of some blocks is, optionally, changed, and some blocks are, optionally, omitted. In some examples, additional steps may be performed in combination with the process 1000.

[0276] At block 1002, text (e.g., text 802) is received. At block 1004, a representation of the received text is determined. At block 1006 an emotional state is determined. In some examples, the emotional state is determined based on a user input (e.g., input 912) separate from the received text. In some examples, the user input separate from the received text specifies the emotional state. In some examples, the emotional state is determined based on the received text.

[0277] At block 1008, a speech data set (e.g., speech data set 812) representing the received text (e.g., text 802) and a set of parameters (e.g., animation parameters 814) representing one or

more movements of an avatar (e.g., avatar 906, 914, 920) are generated using a neural network (e.g., neural network 810) based on the received text and the determined emotional state.

[0278] At block 1010, a vector representation of the received text (e.g., text 802) is determined. At block 1012, a vector representation of the determined emotional state is determined. At block 1014, an intermediate vector based on the vector representation of the received text and the vector representation of the determined emotional state is determined. At block 1016, the speech data set (e.g., speech data set 812) representing the received text and the set of parameters (e.g., animation parameters 814) representing one or more movements of the avatar (e.g., avatar 906, 914, 920) is determined based on the intermediate vector.

[0279] In some examples, the neural network (e.g., neural network 810) is a first neural network and further comprises a plurality of neural networks including a second neural network and a third neural network. At block 1018, the speech data set (e.g., speech data set 812) representing the received text (e.g., text 802) is generated using the second neural network based on the received text and the determined emotional state. At block 1020, the set of parameters (e.g., animation parameters 814) representing one or more movements of the avatar (e.g., avatar 906, 914, 920) is generated using the third neural network based on the speech data set representing the received text.

[0280] In some examples, prior to receiving the text, the neural network (e.g., neural network 810) is trained, as shown in block 1022. At block 1024 a set of training text data (e.g., text 802) is provided to the neural network. At block 1026, a first training speech data set (e.g., speech data set 812) and a first training set of parameters (e.g., animation parameters 814) representing a movement of the avatar (e.g., avatar 906, 914, 920) are determined based on the set of training text data. At block 1028, whether the neural network has been trained is determined based on the first training speech data set and the first training set of parameters representing the movement of the avatar. At block 1030, in accordance with a determination that the neural network has not been trained, a second training speech data set (e.g., speech data set 812) and a second training set of parameters (e.g., animation parameters 814) representing the movement of the avatar are determined based on the set of training text data.

[0281] In some examples, prior to receiving the text, the neural network (e.g., neural network 810) is conditioned using a plurality of predetermined emotional states, as shown at block 1032.

At block 1034 a vector representation of the set of training text data (e.g., text 802) and a predetermined emotional state of the plurality of predetermined emotional states is determined. At block 1036, a vector representation of the set of training text data and a vector representation of the predetermined emotional state are concatenated. In some examples, the vector representation of the predetermined emotional state is an intermediate vector extracted from a classifier trained in a supervised manner to recognize an emotional state. At block 1038, a third training set of parameters (e.g., animation parameters 814) representing the movement of the avatar (e.g., avatar 906, 914, 920) is determined based on the vector representing the set of training text data and the predetermined emotional state. At block 1040, whether the neural network has been condition is determined based on the third set of parameters representing the movement of the avatar.

[0282] In some examples, the set of parameters (e.g., animation parameters 814) representing one or more movements of the avatar (e.g., avatar 906, 914, 920) further comprises a first set of parameters for controlling a pose of the avatar. In some examples, the first set of parameters for controlling a pose of the avatar is determined based on the determined emotional state. In some examples, the set of parameters representing one or more movements of the avatar further comprises a second set of parameters for controlling a plurality of features of the avatar.

[0283] In some examples, the avatar (e.g., avatar 906, 914, 920) is generated (e.g., with avatar animator 820) as shown at block 1042. In some examples the avatar to be generated is determined based on a selection (e.g., input 912) by the user. At block 1044, the avatar is animated (e.g., with avatar animator 820 or output processor 830) using the set of parameters (e.g., animation parameters 814) representing one or more movements of the avatar. At block 1046, the animated avatar is displayed on a display (e.g., display 902) of an electronic device (e.g., user device 900). At block 1048, a speech output (e.g., speech output 832) is generated based on the speech data set (e.g., speech data set 812) representing the received text (e.g., text 802).

[0284] In some examples, at least a first movement of the one or more movements of the avatar is synchronized with the audio output. In some examples, at least a second movement of the one or more movements of the avatar is not synchronized with the audio output.

[0285] The operations described above with reference to FIGS. 10A-10C are optionally implemented by components depicted in FIGS. 1-4, 6A-B, 7A-C, 8, and 9A-C. For example, the operations of process 800 may be implemented by system 800. 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, 7A-C, 8, and 9A-C.

[0286] 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.

[0287] 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.

[0288] 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.

[0289] 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.

[0290] 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.

[0291] 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.

[0292] As described above, one aspect of the present technology is the gathering and use of data available from various sources to improve the accuracy and realism of generated avatars. 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, twitter IDs, home addresses, data or records relating to a user's health or level of fitness (e.g., vital signs measurements, medication information, exercise information), date of birth, or any other identifying or personal information.

[0293] 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 generate more believable and realistic avatars. Accordingly, use of such personal information data enables users to calculated control of the delivered content. Further, other uses for personal information data that benefit the user are also contemplated by the present disclosure. For instance, health and fitness data may be used to provide insights into a user's general wellness, or may be used as positive feedback to individuals using technology to pursue wellness goals.

[0294] The present disclosure 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. Such policies should be easily accessible by users, and should be updated as the collection and/or use of data changes. 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/sharing should occur after receiving the informed consent of the users. Additionally, such entities should consider taking 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. In addition, policies and practices should be adapted for the particular types of personal information data being collected and/or accessed and adapted to applicable laws and standards, including jurisdiction-specific considerations. For instance, in the US, collection of or access to certain health data may be governed by federal and/or state laws, such as the Health Insurance Portability and Accountability Act (HIPAA); whereas health data in other countries may be subject to other regulations and policies and should be handled accordingly. Hence different privacy practices should be maintained for different personal data types in each country.

[0295] 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 avatar generation 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 or anytime thereafter. In another example, users can select to use default, non-personalized avatars. In yet another example, users can select to limit the types of data used when generating avatars to specific types. In addition to providing "opt in" and "opt out" options, the present disclosure contemplates providing notifications relating to the access or use of personal information. For instance, a user may be notified upon downloading an app that their personal information data will be accessed and then reminded again just before personal information data is accessed by the app.

[0296] Moreover, it is the intent of the present disclosure that personal information data should be managed and handled in a way to minimize risks of unintentional or unauthorized access or use. Risk can be minimized by limiting the collection of data and deleting data once it is no longer needed. In addition, and when applicable, including in certain health related applications, data de-identification can be used to protect a user's privacy. De-identification may be facilitated, when appropriate, by removing specific identifiers (e.g., date of birth, etc.), controlling the amount or specificity of data stored (e.g., collecting location data at a city level rather than at an address level), controlling how data is stored (e.g., aggregating data across users), and/or other methods.

[0297] 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, avatars can be generated for users by inferring preferences based on non-personal information data or a bare minimum amount of personal information, such as the avatar being requested by the device associated with a user, other non-personal information available, or publicly available information.

[0298] Exemplary methods, non-transitory computer-readable storage media, systems, and electronic devices are set out in the following items:

1. A method, comprising:
at an electronic device with one or more processors and memory:
receiving text;
determining an emotional state; and
generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.
2. The method of item 1, further comprising determining a representation of the received text.
3. The method of any of items 1-2, wherein the emotional state is determined based on a user input separate from the received text.
4. The method of item 3, wherein the user input specifies the emotional state.
5. The method of any of items 1-4, wherein the emotional state is determined based on the received text.

6. The method of any of items 1-5, wherein, generating a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state further comprises:

determining a vector representation of the received text;

determining, a vector representation of the emotional state;

determining an intermediate vector based on the vector representation of the received text and the vector representation of the emotional state; and

determining the speech data set representing the received text and the set of parameters representing one or more movements of the avatar based on the intermediate vector.

7. The method of any of items 1-6 wherein the neural network is a first neural network and further comprises a plurality of neural networks including a second neural network and a third neural network.

8. The method of item 7, wherein, generating a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the indication of the emotional state further comprises:

generating, using an the second neural network, the speech data set representing the received text based on the received text and the determined emotional state; and

generating, using the third neural network, the set of parameters representing one or more movements of the avatar based on the speech data set representing the received text.

9. The method of any of items 1-8, further comprising:

prior to receiving the text:

training the neural network; and

conditioning the neural network using a plurality of predetermined emotional states.

10. The method of item 9, wherein training the neural network further comprises:

providing a set of training text data to the neural network;

determining, based on the set of training text data, a first training speech data set and a first training set of parameters representing a movement of an avatar;

determining, based on the first training speech data set and the first training set of parameters representing the movement of the avatar, whether the neural network has been trained; and

in accordance with a determination that the neural network has not been trained, determining, based on the set of training text data, a second training speech data set and a second training set of parameters representing the movement of the avatar.

11. The method of item 9, wherein conditioning the neural network with a plurality of predetermined emotional states further comprises:

determining a vector representing the set of training text data and a predetermined emotional state of the plurality of predetermined emotional states;

determining, based on the vector representing the set of training text data and the predetermined emotional state, a third training set of parameters representing the movement of the avatar; and

determining, based on the third set of parameters representing the movement of the avatar, whether the neural network has been conditioned.

12. The method of item 11, wherein determining the vector representing the set of training text data and the predetermined emotional state further comprises:

concatenating a vector representation of the set of training text data and a vector representation of the predetermined emotional state.

13. The method of item 12, wherein the vector representation of the predetermined emotional state is an intermediate vector extracted from a classifier trained in a supervised manner to recognize an emotional state.

14. The method of any of items 1-13, wherein the set of parameters representing one or more movements of the avatar further comprises a first set of parameters for controlling a pose of the avatar.

15. The method of item 14 wherein the first set of parameters for controlling the pose of the avatar are determined based on the determined emotional state.
16. The method of any of items 1-15, wherein the set of parameters representing one or more movements of the avatar further comprises a second set of parameters for controlling a plurality of features of the avatar.
17. The method of any of items 1-16, further comprising:
generating the avatar; and
animating the avatar using the set of parameters representing one or more movements of the avatar.
18. The method of item 17, further comprising:
displaying the animated avatar on a display of the electronic device.
19. The method of item 17, wherein the avatar to be generated is determined based on a selection by a user.
20. The method of any of items 1-19, further comprising:
generating a speech output based on the speech data set representing the received text.
21. The method of item 20, wherein at least a first movement of the one or more movements of the avatar is synchronized with the speech output.
22. The method of item 21, wherein at least a second movement of the one or more movements of the avatar is not synchronized with the speech output.
23. An electronic device comprising:
one or more processors;
a memory; and

one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for:

- receiving text;
- determining an emotional state; and
- generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the indication of the emotional state.

24. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to:

- receive text;
- determine an emotional state; and
- generate, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

25. An electronic device, comprising:

- means for receiving text;
- means for determining an emotional state; and
- means for generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.

26. An electronic device, comprising:

- one or more processors;
- a memory; and
- one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for performing the methods of any one of items 1-21.

27. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to perform the methods of any one of items 1-21.
28. An electronic device, comprising:
means for performing the methods of any one of items 1-21.
29. A system comprising means for performing any of the methods of items 1-21.

WHAT IS CLAIMED IS:

1. A method, comprising:
at an electronic device with one or more processors and memory:
receiving text;
determining an emotional state; and
generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.
2. The method of any of claims 1-2, wherein the emotional state is determined based on a user input separate from the received text.
3. The method of claim 2, wherein the user input specifies the emotional state.
4. The method of any of claims 1-3, wherein the emotional state is determined based on the received text.
5. The method of any of claims 1-4, wherein, generating a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state further comprises:
determining a vector representation of the received text;
determining, a vector representation of the emotional state;
determining an intermediate vector based on the vector representation of the received text and the vector representation of the emotional state; and
determining the speech data set representing the received text and the set of parameters representing one or more movements of the avatar based on the intermediate vector.
6. The method of any of claims 1-5 wherein the neural network is a first neural network and further comprises a plurality of neural networks including a second neural network and a third neural network.

7. The method of claim 6, wherein, generating a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the indication of the emotional state further comprises:

generating, using an the second neural network, the speech data set representing the received text based on the received text and the determined emotional state; and

generating, using the third neural network, the set of parameters representing one or more movements of the avatar based on the speech data set representing the received text.

8. The method of any of claims 1-7, further comprising:
prior to receiving the text:

training the neural network; and

conditioning the neural network using a plurality of predetermined emotional states.

9. The method of claim 8, wherein training the neural network further comprises:

providing a set of training text data to the neural network;

determining, based on the set of training text data, a first training speech data set and a first training set of parameters representing a movement of an avatar;

determining, based on the first training speech data set and the first training set of parameters representing the movement of the avatar, whether the neural network has been trained; and

in accordance with a determination that the neural network has not been trained, determining, based on the set of training text data, a second training speech data set and a second training set of parameters representing the movement of the avatar.

10. The method of claim 8, wherein conditioning the neural network with a plurality of predetermined emotional states further comprises:

determining a vector representing the set of training text data and a predetermined emotional state of the plurality of predetermined emotional states;

determining, based on the vector representing the set of training text data and the predetermined emotional state, a third training set of parameters representing the movement of the avatar; and

determining, based on the third set of parameters representing the movement of the avatar, whether the neural network has been conditioned.

11. The method of claim 10, wherein determining the vector representing the set of training text data and the predetermined emotional state further comprises:

concatenating a vector representation of the set of training text data and a vector representation of the predetermined emotional state.

12. The method of claim 11, wherein the vector representation of the predetermined emotional state is an intermediate vector extracted from a classifier trained in a supervised manner to recognize an emotional state.

13. The method of any of claims 1-12, wherein the set of parameters representing one or more movements of the avatar further comprises a first set of parameters for controlling a pose of the avatar.

14. The method of claim 13 wherein the first set of parameters for controlling the pose of the avatar are determined based on the determined emotional state.

15. The method of any of claims 1-14, wherein the set of parameters representing one or more movements of the avatar further comprises a second set of parameters for controlling a plurality of features of the avatar.

16. The method of any of claims 1-15, further comprising:

generating the avatar; and

animating the avatar using the set of parameters representing one or more movements of the avatar.

17. The method of claim 16, further comprising:
displaying the animated avatar on a display of the electronic device.
18. The method of any of claims 1-17, further comprising:
generating a speech output based on the speech data set representing the received text.
19. An electronic device comprising:
one or more processors;
a memory; and
one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for:
receiving text;
determining an emotional state; and
generating, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the indication of the emotional state.
20. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of an electronic device, cause the electronic device to:
receive text;
determine an emotional state; and
generate, using a neural network, a speech data set representing the received text and a set of parameters representing one or more movements of an avatar based on the received text and the determined emotional state.