July 2021

VISUAL AUDIO MESSAGES

Akshay Kannan

Adrien Olczak

Shumin Zhai

Xu Liu

VISUAL AUDIO MESSAGES

ABSTRACT

Computing devices (e.g., a cellular phone, a smartphone, a desktop computer, a laptop computer, a tablet computer, a portable gaming device, a watch, etc.). may enable users to exchange electronic communication including both a recorded message, such as an audio recording, a video recording, etc., as well as a transcript of the recorded message. In some examples, a first computing device may record audio from a first user and perform speech-to-text to generate a transcript of the recorded audio. The first computing device may then send the recorded message and the transcript of the recorded message in a single electronic communication to a second computing device (e.g., being used by a second user). Because the electronic communication includes the recorded message and the transcript of the recorded message, the second user can both listen to and read the recorded message, which may improve consumption of the recorded message (e.g., because background noise may make listening to the recorded message difficult, reading a transcript of the recorded message may be faster than listening to the recorded message, etc.).

To facilitate a hands-free user experience, the computing device may include a voice user interface (VUI) by which a user may compose the electronic communication. For example, the user may provide voice commands (e.g., “clear”, “send”, “browse”, etc.) to cause the computing device to perform corresponding functions with respect to the electronic communication. Furthermore, the computing device may provide one or more instructions for using voice commands. In some cases, the instructions may relate to the action currently being taken by the user, a context of the electronic communication, etc.

DESCRIPTION

FIG. 1 below is a conceptual diagram illustrating a first computing device 100 that exchanges electronic communication with a second computing device 120. As shown in FIG. 1, first computing device 100 includes one or more processors 102, a display 104, one or more communication components 106 (“COMM components 106”), one or more speakers 108, one or more microphones 110, and one or more storage devices 112. As further shown in FIG. 1, second computing device 120 includes one or more processors 122, a display 124, one or more communication components 126 (“COMM components 126”), one or more speakers 128, one or more microphones 130, and one or more storage devices 132.

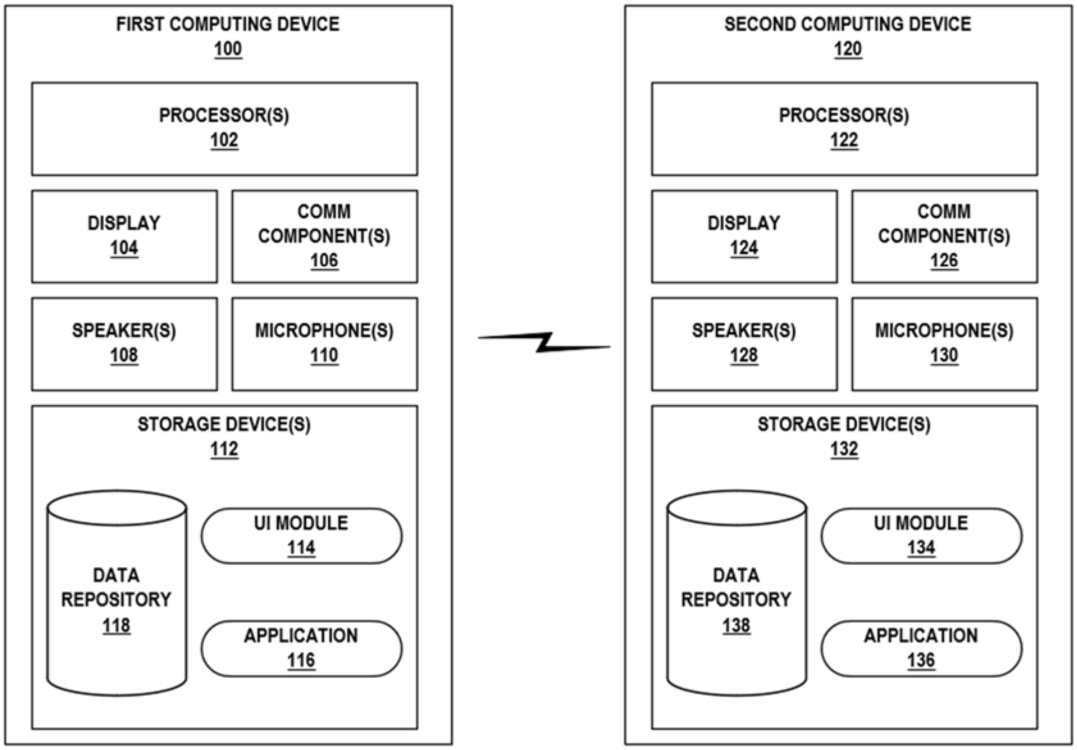


FIG. 1

First computing device 100 (including any component thereof) may be substantially similar to second computing device 120 (including any component thereof). As such, the description of one may apply equally to the other except for any differences described herein.

First computing device 100 may be any mobile or non-mobile computing device, such as a cellular phone, a smartphone, a desktop computer, a laptop computer, a tablet computer, a portable gaming device, a portable media player, an e-book reader, a watch (including a so-called smartwatch), a gaming controller, and/or the like.

Processors 102 may implement functionality and/or execute instructions associated with first computing device 100. Examples of processors 102 may include one or more of an application specific integrated circuit (ASIC), a field programmable gate array (FPGA), an application processor, a display controller, an auxiliary processor, a central processing unit (CPU), a graphics processing unit (GPU), one or more sensor hubs, and any other hardware configure to function as a processor, a processing unit, or a processing device. Processors 102 may retrieve and execute instructions stored by storage devices 112 that cause processors 102 to perform the operations described in this disclosure.

Display 104 of first computing device 100 may be a presence-sensitive display that functions as an input device and as an output device. For example, display 104 may function as an input device using a presence-sensitive input component, such as a resistive touchscreen, a surface acoustic wave touchscreen, a pressure-sensitive screen, an acoustic pulse recognition touchscreen, or another presence-sensitive display technology. Additionally, display 104 may function as an output (e.g., display) device using any of one or more display components, such as a liquid crystal display (LCD), dot matrix display, light emitting diode (LED) display, active-matrix organic light-emitting diode (AMOLED) display, etc.

COMM components 106 of first computing device 100 may include wireless communication devices capable of transmitting and/or receiving communication signals, such as a cellular radio, a 3G radio, a 4G radio, a 5G radio, a Bluetooth® radio (or any other PAN radio), an NFC radio, or a Wi-Fi™ radio (or any other wireless local area network (WLAN) radio). COMM components 106 may be configured to send and receive information via a network (e.g., a local area network (LAN), wide area network (WAN), a global network, such as the Internet, etc.).

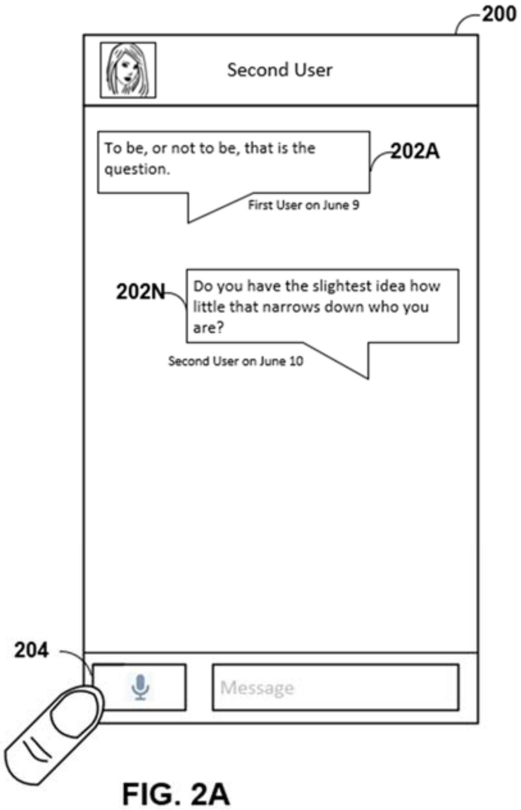
Storage devices 112 of first computing device 100 may include one or more computer-readable storage media. For example, storage devices 112 may be configured for long-term, as well as short-term storage of information, such as instructions, data, or other information used by first computing device 100. In some examples, storage devices 112 may include non-volatile storage elements. Examples of such non-volatile storage elements include magnetic hard discs, optical discs, solid state discs, and/or the like. Additionally or alternatively, storage devices 112 may include one or more so-called "temporary" memory devices, meaning that a primary purpose of these devices may not be long-term data storage. For example, the devices may comprise volatile memory devices, meaning that the devices may not maintain stored contents when the devices are not receiving power. Examples of volatile memory devices include random-access memories (RAM), dynamic random-access memories (DRAM), static random-access memories (SRAM), etc.

As shown in FIG. 1, storage devices 112 may include a user interface module 114 ("UI module 114"). UI module 114 may provide, manage, update, and/or control one or more user interfaces (UIs) (e.g., a graphical user interface (GUI), a voice user interface (VUI), etc.) by which a user may interact with first computing device 100. For example, UI module 114 may generate a GUI that first computing device 100 displays via display 104. The GUI may include one or more graphical elements in one or more layouts. Graphical elements may include, but are not limited to, buttons, icons, pictures, text boxes, menus, thumbnails, scroll bars, hyperlinks, etc. UI module 114 may output and format graphical elements in any one of a variety of layouts and may transition between the various layouts to, for example, show different GUIs. UI module 114 may also reorganize (e.g., rearrange, reformat, resize, replace, remove, etc.) graphical elements in response to user input.

A first user using first computing device 100 may use a communication application 116 ("application 116"), such as a text messaging application, an e-mail application, etc., to exchange electronic communication with a second user using second computing device 120. In general, the electronic communication may include a variety of content, such as texts, emoticons, pictures, audio recordings, videos recordings, data files, etc. However, while some forms of electronic communication may be easy for a sender (e.g., the first user) to create (e.g., audio recordings), those same forms of electronic communication may be difficult for a recipient (e.g., the second user) to consume. For example, the recipient of an electronic communication including a recording (e.g., an audio recording, a video recording, etc.) may have trouble listening to the recording in a noisy environment. In such cases (as well as in others), a transcript of the recording may be desirable, if not necessary, to help the recipient consume the content of the electronic communication.

In accordance with techniques of this disclosure, first computing device 100 and second computing device 120 (and, in some cases, additional computing devices, e.g., a third computing device, a fourth computing device, etc.) may enable users to exchange, via application 116, electronic communication including both a recording, such as an audio recording, a video recording, etc., and a transcript of the recording. As shown in FIG. 2A below, UI module 114 may output, for display by display 104, a GUI of a chat session 200 ("chat session GUI 200"). Chat session GUI 200 may include electronic communications 202A-202N (collectively, "electronic communications 202") exchanged between users of application 116. In some examples, UI module 114 may organize electronic communications 202 to indicate the chronological relationship of electronic communications 202. For example, the newest electronic communications may be displayed below the oldest electronic communications in chat session GUI 200.

A first user (i.e., the user of first computing device 100) may provide user input to first computing device 100 to record a message. For example, the first user may provide a touch input (e.g., a tap) to an audio recording graphical element 204 that causes microphones 110 to record audio from the first user. The visual appearance of audio recording graphical element 204 may change to visually indicate to the first user that microphones 110 are recording audio. For example, the color of audio recording graphical element 204 may change, audio recording



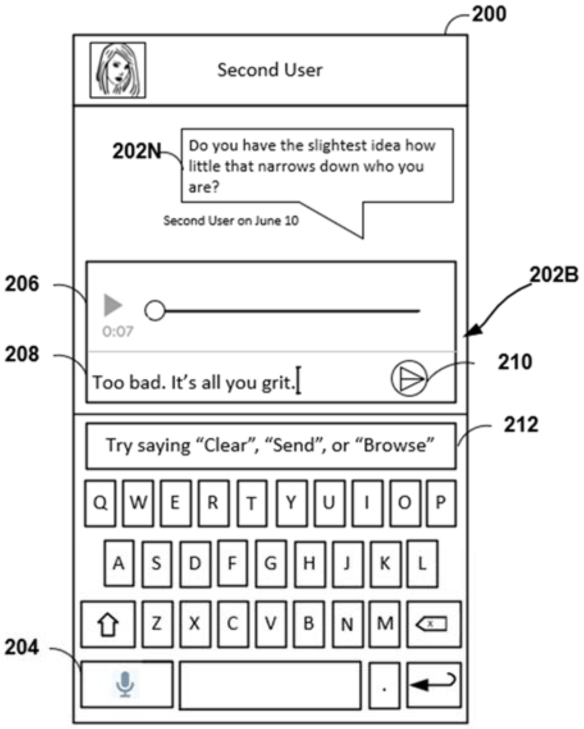
To be, or not to be, that is the question.

First User on June 9

Do you have the slightest idea how little that narrows down who you are?

Second User on June 10

FIG. 2A



Do you have the slightest idea how little that narrows down who you are?

Second User on June 10

Too bad. It's all you grit.

Try saying "Clear", "Send", or "Browse"

FIG. 2B

graphical element 204 may be animated, etc. In some examples, the first user may provide a voice input (e.g., a voice command) to a VUI provided by UI module 114 that causes microphones 110 to record audio. In such an example, microphones 110 may be always active (but not always recording) to be able to detect the voice input.

Responsive to microphones 110 recording the audio from the first user, first computing device 100 may store data associated with the recorded audio ("audio data") in data repository 118. In some examples, first computing device 100 may include a speech-to-text module ("STT module") that analyzes the audio input data stored in data repository 118 to generate a transcript of the recorded audio. In some examples, the STT module may be a machine learning model trained to convert speech to text. In such examples, the STT module may represent one more neural networks, such as one or more recurrent neural networks. In some instances, at least some of the nodes of a recurrent neural network may form a cycle. That is, the STT module may pass or retain information from a previous portion of the input data sequence to a subsequent portion of the input data sequence through the use of recurrent or directed cyclical node connections, where the input data sequence includes words in a sentence for natural language processing, speech detection or processing, etc.

Thus, the STT module may output a transcript of the recorded message, which first computing device 100 may store in data repository 118. Application 116 may obtain the recorded message and the transcript of the recorded message and include both in an electronic communication to be sent (e.g., electronic communication 202B) to second computing device 120. The visual appearance of electronic communication 202B may indicate, to the first user, that electronic communication 202B includes a first component 206 for a recording (e.g., an audio recording, a video recording, etc.) and a second component 208 for other data types (e.g., text, pictures, etc.). In some examples, the first user may delete at least a portion of first component 206 and/or second component 208 by providing a user input (e.g., a touch input, a voice input, etc.), resulting in corresponding changes to electronic communication 202B. For example, the first user may say “clear” to delete electronic communication 202B, thereby removing first component 206 and second component 208 from chat session GUI 200.

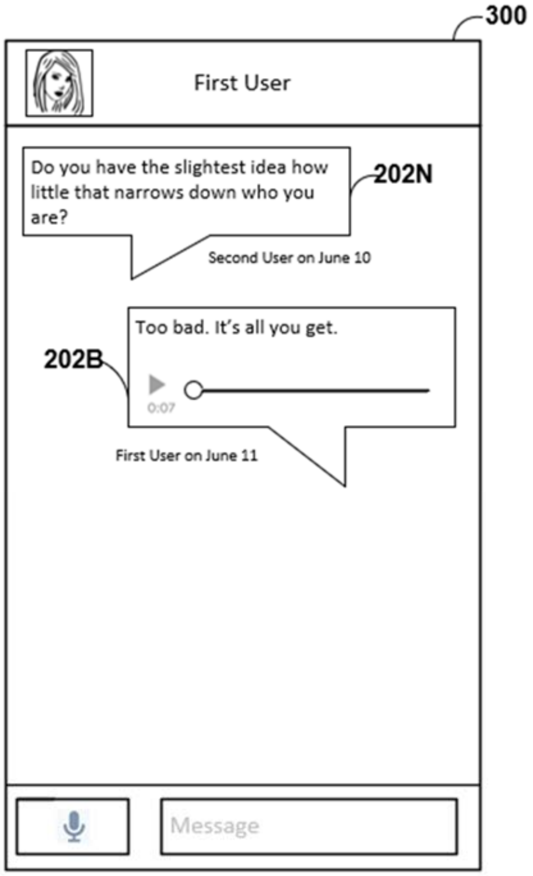
The first user may use a VUI provided by UI module 114 to compose second component 208 of electronic communications 202B. For example, the first user may provide voice commands via microphones 110 to add emoticons, pictures, and other data files to second component 208, in this way enabling a hands-free user experience. In some examples, chat session GUI 200 includes one or more instructions 212 (e.g., usage guides) for using application 116. For example, as shown in FIG. 2B, chat session GUI 200 includes instructions 212 with the text “Try saying “Clear”, “Send”, or “Browse”. Instructions 212 may relate to the action currently being taken by the user, a context of the electronic communication, etc. For example, in FIG. 2B, instructions 212 may include the voice commands “clear”, “send”, and “browse” because the first user is currently composing electronic communication 202B. If the first user provides the voice command “browse”, UI module 114 may generate a GUI of voice commands, which may be in the form of a list, an array, etc.

As the transcript of the recorded message may require editing, the first user may provide user input to correct the transcript of the recorded message. In some examples, UI module 114 may output suggestions (e.g., graphical elements of a selection of words for making corrections). The suggestions may be based on previous electronic communication, user preferences, a dictionary, a spell checker, etc. UI module 114 may generate the suggestions approximately where instructions 212 are located. In other words, UI module 114 may replace instructions 212 with the suggestions.

In some cases, the first user may select the suggestions by providing a voice command. For example, if the first user selects (e.g., via a touch input) the word “grit” in electronic communication 202B, UI module 114 may generate as a suggestion the word “get”. The first user may then say “get”, resulting in the substitution of the word “grit” with the word “get”. Other voice commands that may modify text in electronic communication 202B include bolding, italicizing, highlighting, copying, cutting, deleting, etc.

When the first user is finished composing electronic communication 202B, the first user may provide user input to first computing device 100 to send electronic communication 202B to second computing device 120 (e.g., being used by the second user). For example, the first user may select a send graphical element 210 to cause application 116 to send electronic communication 202B to second computing device 120 (e.g., which is being used by a second user). In another example, the first user may provide a voice command, such as “send”, to cause application 116 to send electronic communication 202B to second computing device 120. In examples where first computing device 100 is a wearable device, a home device, or some device other than a phone of the first user, first computing device 100 may first send electronic communication 202 to the phone of the first user, which in turn may send electronic communication 202 to second computing device 120.

Responsive to second computing device 120 receiving electronic communication 202B, a UI module 134 may update a GUI of a chat session 300 (“chat session 300”), as shown in FIG. 3 below. Because electronic communication 202B includes the recorded message and the transcript of the recorded message, the second user can both listen to and read the message via an application 136.



Do you have the slightest idea how little that narrows down who you are?

Second User on June 10

Too bad. It's all you get.

First User on June 11

FIG. 3

Although primarily described here as being performed at first computing device 100, it should be understood that conversion of the speech in the recorded message into a transcript may be performed elsewhere. For example, second computing device 120 may include a STT module that analyzes audio input data received from first computing device 100 and stored in a data repository 138 via a network to generate the transcript of the recorded audio. In another example, a remote computing system (e.g., the cloud), may include a STT module that analyzes audio input data received from first computing device 100 to generate the transcript of the recorded audio. The remote computing system may then transmit the recorded message and the transcript of the recorded message to second computing device 120.

It should also be understood that although primarily described here as being performed with respect to audio recordings, the techniques of this disclosure may also apply to video recordings, which include audio data. Thus, the techniques may be used to send electronic communications including video recordings and transcripts of the video recordings.

It is noted that the techniques of this disclosure may be combined with any other suitable technique or combination of techniques. As one example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2017/0085506A1. In another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2015/0312175A1. In yet another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2021/0142782A1. In yet another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2020/0272485A1.