September 06, 2016

SYSTEM AND METHOD FOR SPEECH RECOGNITION

Dimitri Kanevsky

Tara Sainath

SYSTEM AND METHOD FOR SPEECH RECOGNITION

BACKGROUND

In typical speech recognition systems, a computing device, such as a mobile computing device (smart phone, tablet, etc.), will receive a speech input from a user. The speech input can be received via a microphone of the computing device and converted to an audio signal for processing. The audio signal is parsed and otherwise processed to recognize speech utterances, which can be output as text and/or utilized by the computing device to perform a function (for example, voice search, email or text message dictation, or other command). A user may not, however, utilize the full speech capabilities of her/his computing device in public spaces. For example only, a user may choose to avoid providing speech inputs in public places due to noisy conditions, which may degrade the performance of the speech recognition system, and/or to avoid providing a relatively loud speech input due to privacy concerns, whether actual or perceived.

It would be desirable to provide an improved speech recognition system and method that provides better performance in noisy conditions while also providing increased privacy for the user.

SUMMARY

The present disclosure is directed to a system and method for speech recognition. A user will provide a speech input to a computing device, which will be recognized and processed. The user can also provide authorization to the computing device to sense and process the speech input via one or more sensor(s) that are utilized to detect the position, movement, etc. of the user's lips, and in some aspects

tongue. If explicitly authorized by the user, the computing device can receive the speech input via the one or more sensors that are utilized to detect the position, movement, etc. of the user's lips, and in some aspects tongue. The one or more sensors can, for example, be contactless (proximity detection) sensors and/or touch sensors. The speech input can (but need not) be audible speech of the user that is also received by a microphone of the computing device.

Further areas of applicability of the present disclosure will become apparent from the detailed description provided hereinafter. It should be understood that the detailed description and specific examples are intended for purposes of illustration only and are not intended to limit the scope of the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

The present disclosure will become more fully understood from the detailed description and the accompanying drawings, wherein:

FIG. 1 is a diagram of a computing system including an example server computing device and user computing device according to some implementations of the present disclosure;

FIG. 2 is a functional block diagram of the example user computing device of FIG. 1;

FIG. 3 is a functional block diagram of the example server computing device of FIG. 1; and

FIG. 4 is a flow diagram of a technique for speech recognition according to some implementations of the present disclosure.

DETAILED DESCRIPTION

As mentioned above, there is a need for an improved speech recognition system and method that provides better performance in noisy conditions while also providing increased privacy for the user. In order to address this need, the present disclosure is directed to an improved system and method that utilizes, only after receiving user authorization to do so, one or more sensors configured to detect the position, movement, etc. of a user's lips while providing a speech input to a computing device in order to perform speech recognition. While a user may provide an audible speech input to the computing device (e.g., via a microphone), in some implementations the sensor(s) can be utilized to detect a speech input that is not actually spoken out loud by the user. In this manner, a user can provide a soundless speech input to her/his computing device, which may increase the utility of speech recognition, e.g., by permitting a user to provide a speech input in a public place without the worry/annoyance of being overheard. Furthermore, for a user who may stutter while speaking, the present techniques may eliminate the need for audible speech and therefore permit the user to utilize speech recognition in a manner not previously possible.

Referring now to FIG. 1, a diagram of an example computing system 100 for performing the disclosed techniques is illustrated. The computing system 100 can include a user computing device 104 that is operated by/associated with a user 108. Examples of the user computing device 104 include, but are not limited to, a tablet computer and a mobile phone. The computing system 100 can further include a server computing device 150 that is configured to communicate with the user computing device 104 via a network 112. As used herein, the term "server computing device" can refer to

any suitable hardware computer server, as well as both a single server and multiple servers operating in a parallel or distributed architecture. The network 112 can include a local area network (LAN), a wide area network (WAN), e.g., the Internet, or a combination thereof. In some implementations, the user computing device 104 includes peripheral components, such as a touchscreen display 116, a camera 120 to capture photographs and/or video, a microphone 124, and a speaker 128. In some implementations, the user computing device 104 can also one or more other input/output devices (not shown), such as a button, a switch, or one or more ports (e.g., a headphone port, a charging port, and the like).

Referring now to FIG. 2, a functional block diagram of an example user computing device 104 is illustrated. The user computing device 104 can include a communication device 200, a processor 204, and a memory 208. The user computing device 104 can also include the touchscreen display 116, the camera 120, the microphone 124, and the speaker 128, as well as one or more sensors 212 (referred to herein individually and collectively as “user interface device(s) 216”). The user interface devices 216 are configured for interaction with the user 108.

The communication device 200 is configured for communication between the processor 204 and other devices, e.g., the server computing device 150, via the network 112. The communication device 200 can include any suitable communication components, such as a transceiver. The memory 208 can be configured to store information at the user computing device 104, such as one or more photographs of individuals. The memory 208 can be any suitable storage medium (flash, hard disk, etc.).

The processor 204 can be configured to control operation of the user computing device 104. It should be appreciated that the term “processor” as used herein can refer to both a single processor and two or more processors operating in a parallel or distributed architecture. The processor 204 can be configured to perform general functions including, but not limited to, loading/executing an operating system of the user computing device 104, controlling communication via the communication device 200, and controlling read/write operations at the memory 208. The processor 204 can also be configured to perform specific functions relating to at least a portion of the present disclosure, which are described in greater detail below.

Referring now to FIG. 3, a functional block diagram of the server computing device 150 is illustrated. It should be appreciated that the server computing device 150 can have the same or similar structure to the user computing devices 104 described above. The server computing device 150 can include a communication device 152, a processor 154, and a memory 158. As described above, the term “processor” as used herein can refer to both a single processor and multiple processors operating in a parallel or distributed architecture. The communication device 152 can include any suitable communication components (e.g., a transceiver) for communication via the network 112. The memory 158 can be any suitable storage medium (flash, hard disk, etc.) for storing information at the server computing device 150. The processor 154 can control operation of the server computing device 150 and can implement at least a portion of the techniques of the present disclosure, which are described in greater detail below.

Various implementations of the techniques of the present disclosure will be described in the context of the disclosed computing system 100 that includes the example user computing device 104 and server computing device 150. It should be appreciated that, while various aspects of the techniques will be described as being performed by one of the user computing devices 104 or the server computing device 150, any aspect of the techniques may be performed by one or more user computing devices 104, the server computing device 150, or a combination thereof.

As briefly mentioned above, the user 108 can utilize his/her user computing device 104 to provide a speech input for speech recognition. In some implementations, the user 108 can provide a speech input by moving the user computing device 104 close to or in contact with the mouth of the user 108. Due to the proximity of the microphone 124 to the mouth of the user 108, the user computing device 104 may be able to obtain relatively low volume speech inputs (e.g., whispers). Further, the placement of the user computing device 104 over the mouth of the user 108 may inhibit other persons in the general geographic area from hearing the speech input and/or performing lip-reading or other non-hearing eavesdropping techniques.

In addition to the microphone 124, the present techniques contemplate the use of one or more sensors 212 (only after receiving user authorization do so) configured to receive the speech input of the user 108. The sensors 212, which are more fully described below, can be configured to sense the position, movement (speed and/or direction), and force of the user's lips and tongue to assist the speech recognition techniques. The sensors 212 may include touch sensors and/or contactless position sensors to detect the lips/tongue/etc. of the user 108.

With respect to contactless sensors, the sensors 212 can comprise proximity sensors, e.g., infrared (IR) sensor(s). In some aspects, an IR light source (e.g., an IR light emitting diode) can be combined with IR receiver to provide a proximity sensing functionality. For example only, the IR light source and/or IR receiver can be placed under a plastic/glass surface (e.g., the outer surface of the touchscreen display 116). The IR light source can emit IR signals. When the mouth (lips, tongue, etc.) of the user 108 is near the surface, the IR receiver can detect the reflection of the IR signals and utilize the detected reflections to determine a distance to the portions of the mouth of the user 108 (e.g., the intensity of the reflection may be greater as the distance to the lips of the user 108 decreases). The IR receiver may continuously or periodically scan the input received from the IR receiver to derive features related to the mouth of the user 108, as described more fully below. In a further example, a capacitive proximity detection sensor that detects variations in the capacitive coupling between the sensor and the lips/mouth of the user 108 could be utilized.

In some aspects, the user computing device 104 may derive features from the reflected IR light received by the IR receiver. For example only, the pair-wise time delay between the sensor data of two channels (e.g., the IR light source output and the IR receiver input) can be measured and correlated with movement/position/etc. of the mouth (or portions thereof) of the user 108. In another example, the local slope of the signal segment within a frame can be estimated and correlated with how quickly the lips of the user 108 are moving towards or away from the proximity sensors. The slope can, e.g., by calculated by first-order linear regression, and then summed with the slopes of one or more previous frames to better capture the trend of slopes rather than sudden

changes. In yet another example, the mean and variance of the raw data in the current frame in combination with previous frames can be calculated and utilized.

The sensors 212 could also or alternatively comprise touch sensors to detect the lips/tongue/etc. of the user 108. In implementations where touch sensors are utilized, in addition to the position of the lips or other portions of the user's 108 mouth, the sensors 212 may detect the amplitude and directionality of the force of the user's 108 touch. In some implementations, and only after receiving user authorization to do so, the camera 120 could also be utilized to image and thereby detect the position/movement of the mouth/lips of the user 108. This image data could be combined with the other aspects of the speech input to more accurately perform speech recognition. It should be appreciated, however, that the camera 120 may not be useful to capture the mouth of the user 108 when the user 108 places the user computing device 104 directly in front of his/her lips.

As mentioned above, the speech input, e.g., comprising an audible sound detected by the microphone 124, the lip proximity and/or touch data captured by the sensor(s) 212 (only user authorization to do so has been received), the image data captured by the camera 120, or a combination thereof, can be received by the user computing device 104. The user computing device 104 can extract features from the speech input (e.g., the data from the various sensors) and a decision-tree classifier or other mechanism can be utilized to classify the speech input as corresponding to a particular recognized speech representation (character, set of characters, word, etc.). In some embodiments, the various types of data of the speech input (sound, lip movement/position, image data) can be separately analyzed and classified by

independent models. In this manner, a statistical module could then combine the separate classifications from the independent models to derive a recognition result. Other forms of classification are contemplated.

Referring now to FIG. 4, a flow diagram of an example method 400 for speech recognition utilizing lip sensing is illustrated. For ease of description, the method 400 will be described in the context of the computing system 100 and the components thereof. It should be appreciated that although the method 400 may be described as being performed at the user computing device 104, the method 400 or portions thereof may be performed by the server computing device 150, alone or in combination with the user computing device 104.

At 410, the computing device 104 determines if the user 108 has provided user authorization to utilize one or more sensors configured to detect the position, movement, etc. of a user's lips while providing a speech input in order to perform speech recognition. If not, the method 400 returns to 410 or ends. If, however, the user 108 has provided such use authorization, the method proceeds to 420. At 420, the user computing device 104 can detect the lip position information of the user 108 via the one or more sensors 212. Optionally, at 430, the user computing device 104 could also detect audible speech information via the microphone 124. As mentioned above, the techniques of the present disclosure may not require audible speech to perform speech recognition. The various types of received speech data can be processed at 440. Processing may include, e.g., any form of averaging, filtering, or other processing useful to more accurately provide a quality speech recognition result. At 450, one or more models (a lip model, an audible speech signal recognition model, etc.) can be utilized to

classify the speech data and obtain a speech recognition result, which can be output at 460.

The above techniques provide for an improved speech recognition system and method that – upon receiving user authorization – utilizes lip position information as the speech input or portion thereof. In this manner, a user of the speech recognition system may more accurately and/or privately provide a speech input and have a superior user experience.

One or more systems and methods discussed herein do not require collection or usage of user personal information. In situations in which certain implementations discussed herein may collect or use personal information about users (e.g., user data, information about a user's social network, user's location and time, user's biometric information, user's activities and demographic information), users are provided with one or more opportunities to control whether the personal information is collected, whether the personal information is stored, whether the personal information is used, and how the information is collected about the user, stored and used. That is, the systems and methods discussed herein collect, store and/or use user personal information only upon receiving explicit authorization from the relevant users to do so. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. As one example, a user's identity may be treated so that no personally identifiable information can be determined. As another example, a user's geographic location may be generalized to a larger region so that the user's particular location cannot be determined.

Example embodiments are provided so that this disclosure will be thorough, and will fully convey the scope to those who are skilled in the art. Numerous specific details are set forth such as examples of specific components, devices, and methods, to provide a thorough understanding of embodiments of the present disclosure. It will be apparent to those skilled in the art that specific details need not be employed, that example embodiments may be embodied in many different forms and that neither should be construed to limit the scope of the disclosure. In some example embodiments, well-known procedures, well-known device structures, and well-known technologies are not described in detail.

The terminology used herein is for the purpose of describing particular example embodiments only and is not intended to be limiting. As used herein, the singular forms "a," "an," and "the" may be intended to include the plural forms as well, unless the context clearly indicates otherwise. The term "and/or" includes any and all combinations of one or more of the associated listed items. The terms "comprises," "comprising," "including," and "having," are inclusive and therefore 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. The method steps, processes, and operations described herein are not to be construed as necessarily requiring their performance in the particular order discussed or illustrated, unless specifically identified as an order of performance. It is also to be understood that additional or alternative steps may be employed.

Although the terms first, second, third, etc. may be used herein to describe various elements, components, regions, layers and/or sections, these elements, components, regions, layers and/or sections should not be limited by these terms. These terms may be only used to distinguish one element, component, region, layer or section from another region, layer or section. Terms such as "first," "second," and other numerical terms when used herein do not imply a sequence or order unless clearly indicated by the context. Thus, a first element, component, region, layer or section discussed below could be termed a second element, component, region, layer or section without departing from the teachings of the example embodiments.

The techniques described herein may be implemented by one or more computer programs executed by one or more processors. The computer programs include processor-executable instructions that are stored on a non-transitory tangible computer readable medium. The computer programs may also include stored data. Non-limiting examples of the non-transitory tangible computer readable medium are nonvolatile memory, magnetic storage, and optical storage.

Some portions of the above description present the techniques described herein in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. These operations, while described functionally or logically, are understood to be implemented by computer programs. Furthermore, it has also proven convenient at times to refer to these arrangements of operations as modules or by functional names, without loss of generality.

Unless specifically stated otherwise as apparent from the above discussion, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system memories or registers or other such information storage, transmission or display devices.

Certain aspects of the described techniques include process steps and instructions described herein in the form of an algorithm. It should be noted that the described process steps and instructions could be embodied in software, firmware or hardware, and when embodied in software, could be downloaded to reside on and be operated from different platforms used by real-time network operating systems.

The present disclosure also relates to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general-purpose computer selectively activated or reconfigured by a computer program stored on a computer readable medium that can be accessed by the computer. Such a computer program may be stored in a tangible computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, application specific integrated circuits (ASICs), or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus. Furthermore, the computers

referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

The algorithms and operations presented herein are not inherently related to any particular computer or other apparatus. Various general-purpose systems may also be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatuses to perform the required method steps. The required structure for a variety of these systems will be apparent to those of skill in the art, along with equivalent variations. In addition, the present disclosure is not described with reference to any particular programming language. It is appreciated that a variety of programming languages may be used to implement the teachings of the present disclosure as described herein, and any references to specific languages are provided for disclosure of enablement and best mode of the present invention.

The present disclosure is well suited to a wide variety of computer network systems over numerous topologies. Within this field, the configuration and management of large networks comprise storage devices and computers that are communicatively coupled to dissimilar computers and storage devices over a network, such as the Internet.

The foregoing description of the embodiments has been provided for purposes of illustration and description. It is not intended to be exhaustive or to limit the disclosure. Individual elements or features of a particular embodiment are generally not limited to that particular embodiment, but, where applicable, are interchangeable and can be used in a selected embodiment, even if not specifically shown or described. The same may also be varied in many ways. Such variations are not to be regarded as a departure

from the disclosure, and all such modifications are intended to be included within the scope of the disclosure.

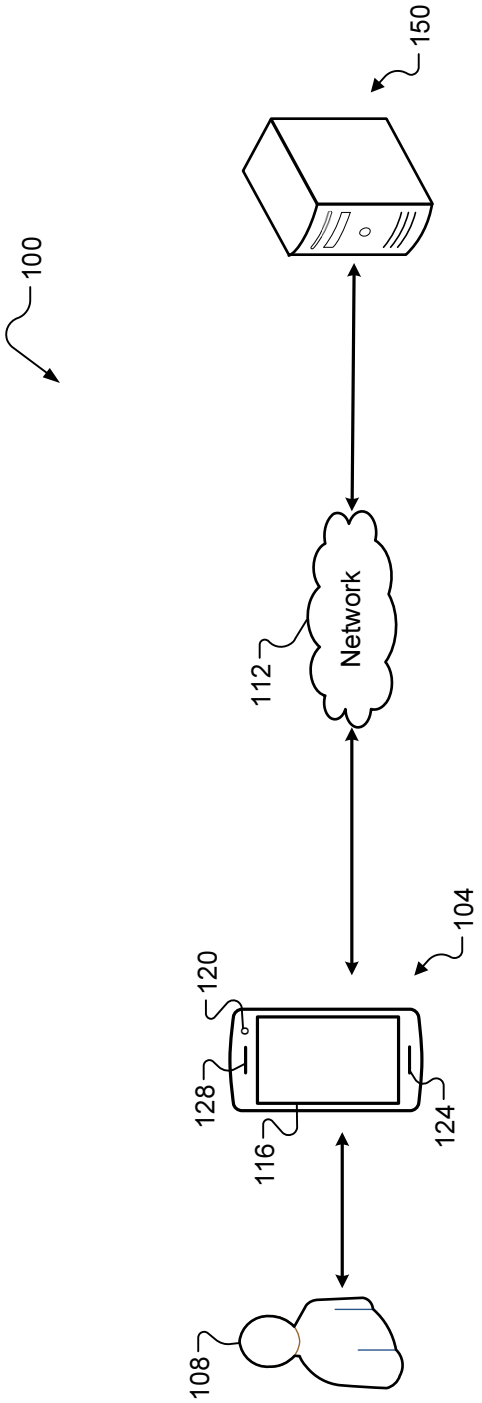

FIG. 1

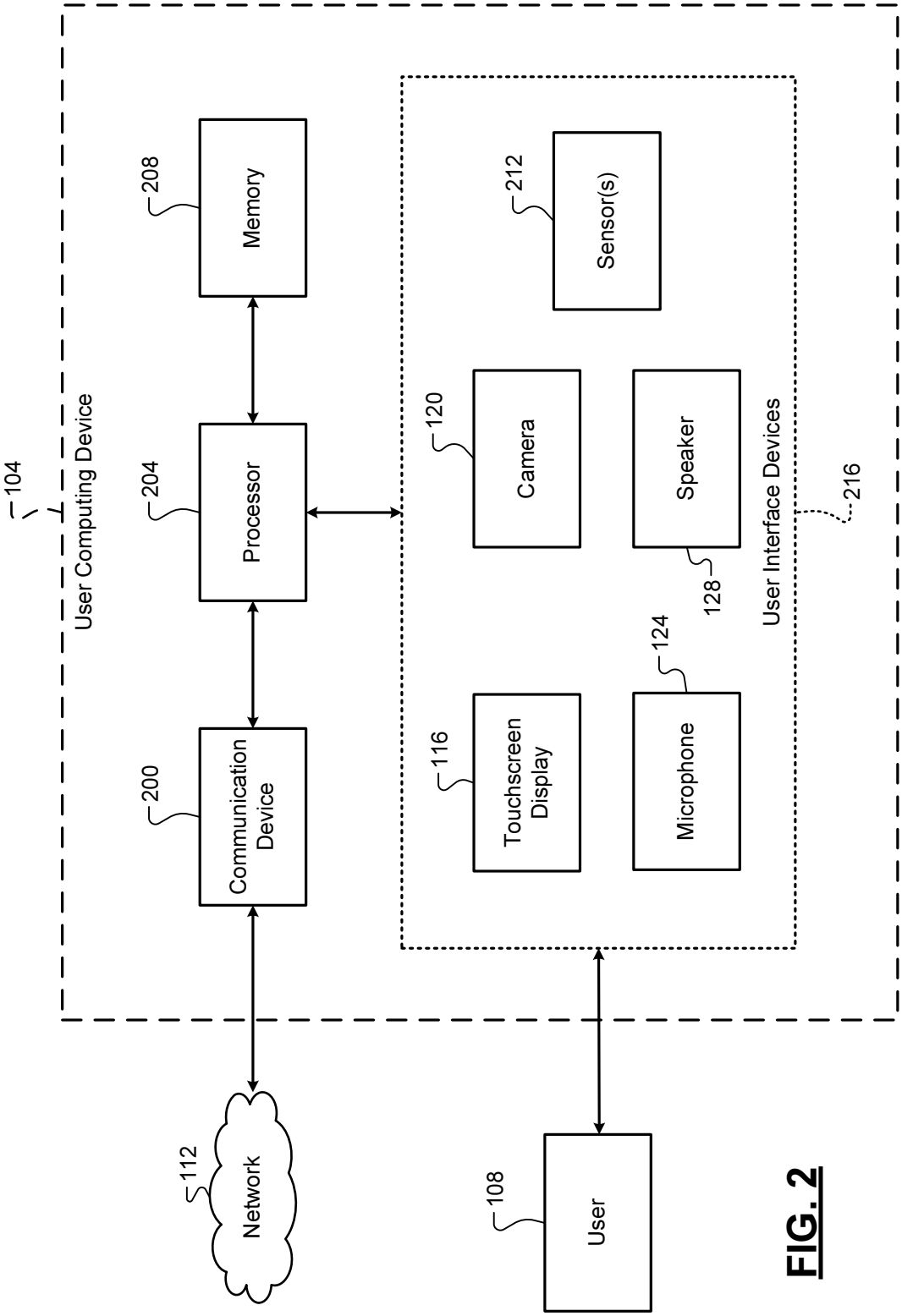
104
User Computing Device
208
Memory
204
Processor
200
Communication Device
212
Sensor(s)
120
Camera
Speaker
128
User Interface Devices
216
124
116
Touchscreen Display
Microphone
112
Network
108
User

FIG. 2

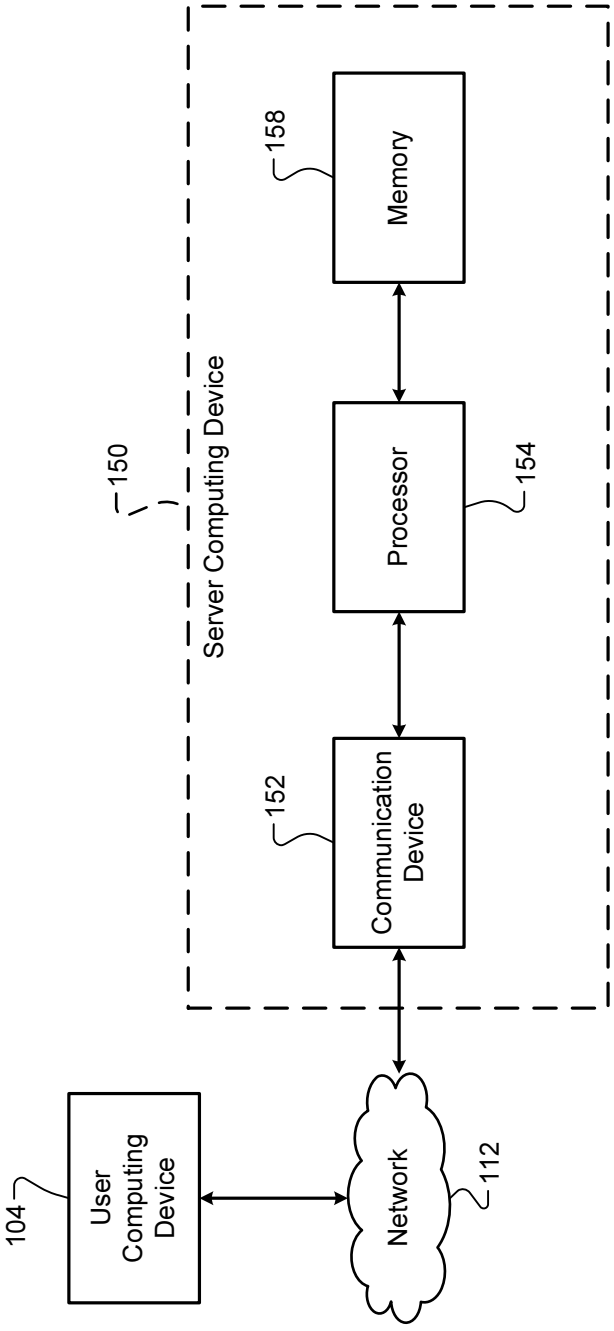
Server Computing Device
150
Memory
158
Processor
154
Communication Device
152
User Computing Device
104
Network
112

FIG. 3

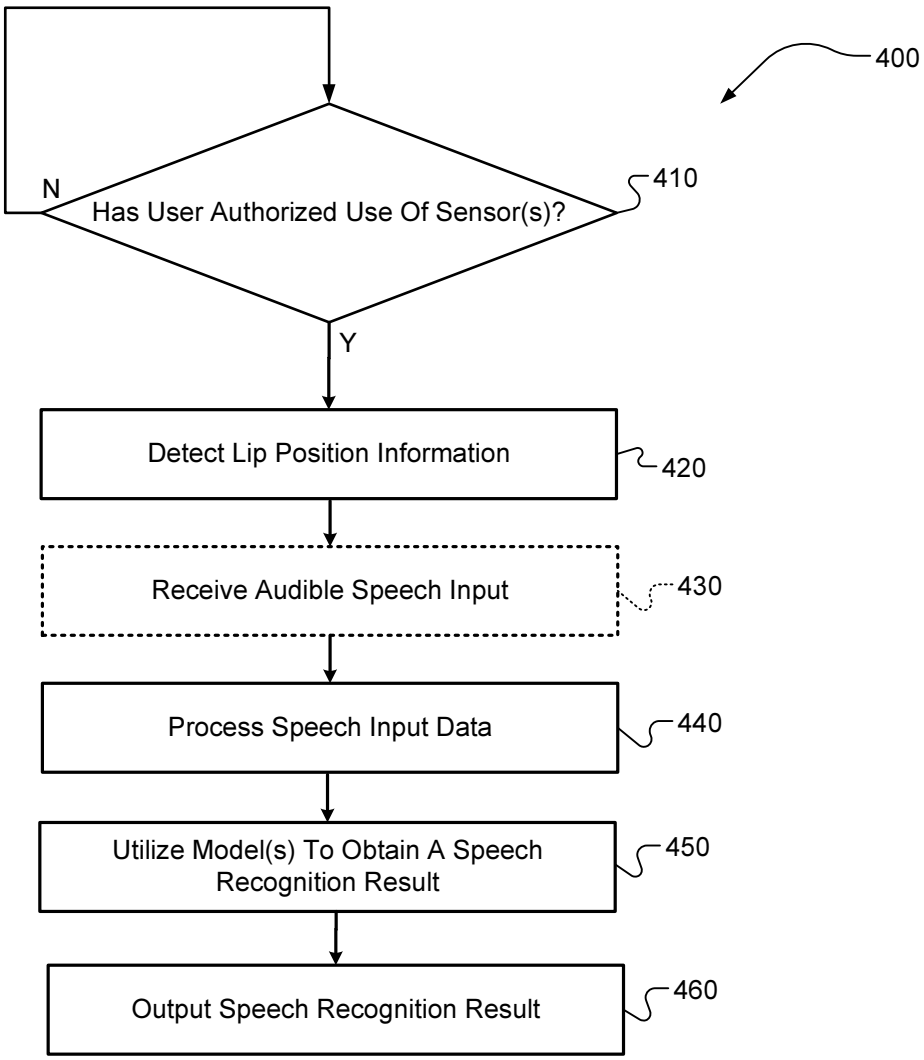
400
Has User Authorized Use Of Sensor(s)?
410
N
Y
Detect Lip Position Information
420
Receive Audible Speech Input
430
Process Speech Input Data
440
Utilize Model(s) To Obtain A Speech Recognition Result
450
Output Speech Recognition Result
460

FIG. 4