(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 April 2021 (15.04.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/071110 A1

(51) International Patent Classification:
G10L 15/22 (2006.01) *G10L 15/26* (2006.01)
G10L 15/14 (2006.01) *G10L 13/08* (2006.01)
G10L 15/183 (2013.01) *G10L 17/24* (2013.01)
G06T 7/11 (2017.01) *G06F 3/16* (2006.01)
G10L 21/0208 (2013.01)

(21) International Application Number:
PCT/KR2020/012198

(22) International Filing Date:
09 September 2020 (09.09.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10-2019-0125172 10 October 2019 (10.10.2019) KR

(71) Applicant: **SAMSUNG ELECTRONICS CO., LTD.** [KR/KR]; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677 (KR).

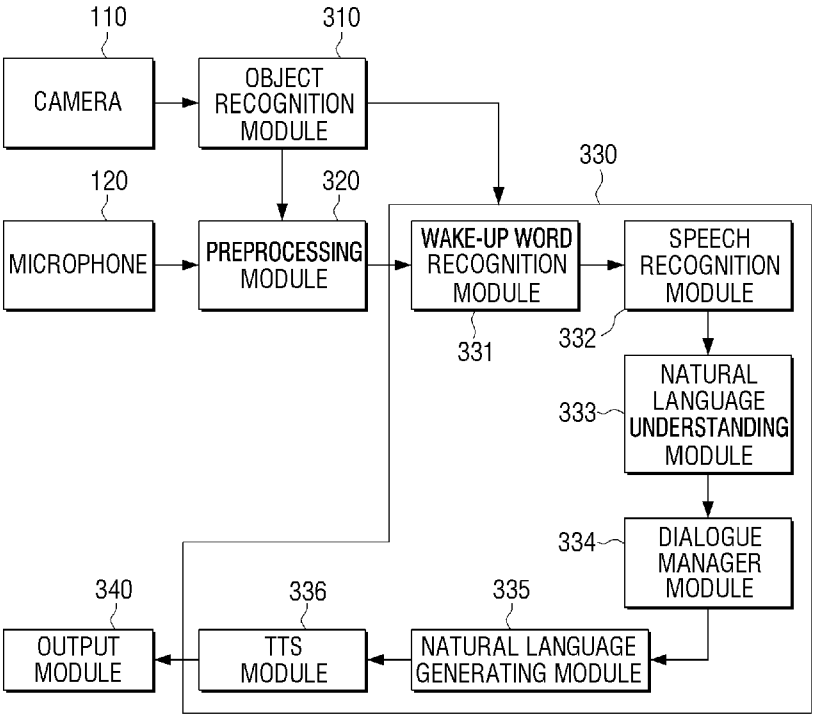
(72) Inventors: **LIM, Hyeontaek**; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677 (KR). **KWAK, Sejin**; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677 (KR). **KIM, Youngjin**; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677 (KR).

(74) Agent: **KIM, Tae-hun** et al.; 9th Floor, Shinduk Bldg., 343, Gangnam-daero, Seocho-gu, Seoul 06626 (KR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

WO 2021/071110 A1

(54) **Title:** ELECTRONIC APPARATUS AND METHOD FOR CONTROLLING ELECTRONIC APPARATUS



(57) **Abstract:** An electronic apparatus and a control method thereof are provided. The electronic apparatus includes a microphone, a camera, a memory storing an instruction, and a processor configured to control the electronic apparatus coupled with the microphone, the camera and the memory, and the processor is configured to, by executing the instruction, obtain a user image by photographing a user through the camera, obtain the user information based on the user image, and based on a user speech being input from the user through the microphone, recognize the user speech by using a speech recognition model corresponding to the user information among a plurality of speech recognition models.

[Continued on next page]

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

Description

Title of Invention: ELECTRONIC APPARATUS AND METHOD FOR CONTROLLING ELECTRONIC APPARATUS

Technical Field

[1] The disclosure relates to an electronic apparatus and a method for controlling the electronic apparatus. More particularly, the disclosure relates to an electronic apparatus capable of providing a response to a user speech by recognizing the user speech and a method for controlling the electronic apparatus.

Background Art

[2] In recent years, artificial intelligence systems have been employed in various fields. Unlike existing rule-based smart systems, the artificial intelligence systems are machines that learn and judge themselves. The more the artificial intelligence system is used, recognition rate improves and user preference is more accurately understood. Thus, existing rule based smart systems are gradually being replaced with deep learning based artificial intelligence systems.

[3] In particular, recently, an artificial intelligence personal assistant system (or dialogue system) capable of providing a response to a user speech or controlling an electronic apparatus based on the user speech is being developed by using the artificial intelligence system.

[4] The artificial intelligence personal assistance system described above includes various neural network models such as a speech recognition module for recognizing user speech, a natural language understanding module for understanding user speech, a text to speech (TTS) for outputting the generated response in speech data format, and the like.

[5] However, because a plurality of neural network models included in the artificial intelligence personal assistant system of the related art is generically trained, there are many constraints to usability. That is, because a response to a user speech is provided by using a speech recognition module, a natural language understanding module, a TTS module and the like which does not consider user characteristics, there has been problems such as providing a response the context of which is not desired by the user or providing a response in a manner not suitable to the user.

[6] The above information is presented as background information only to assist with an understanding of the disclosure. No determination has been made, and no assertion is made, as to whether any of the above might be applicable as prior art with regard to the disclosure.

Disclosure of Invention

Technical Problem

[7] Aspects of the disclosure are to address at least the above-mentioned problems and/or disadvantages and to provide at least the advantages described below. Accordingly, an aspect of the disclosure is to provide an electronic apparatus and a method for controlling the electronic apparatus.

Solution to Problem

[8] Additional aspects will be set forth in part in the description which follows and, in part, will be apparent from the description, or may be learned by practice of the presented embodiments.

[9] In accordance with an aspect of the disclosure, an electronic apparatus is provided. The electronic apparatus includes a microphone, a camera, a memory storing an instruction, and a processor configured to control the electronic apparatus by being coupled to the microphone, the camera and the memory, and the processor is configured to, by executing the instruction, obtain a user image by photographing a user through the camera, obtain the user information based on the user image, and based on a user speech being input from the user through the microphone, recognize the user speech by using a speech recognition model corresponding to the user information among a plurality of speech recognition models.

[10] The processor, by executing the instruction, may be further configured to obtain an environment information based on information on an object other than the user comprised in the user image by analyzing the user image, and perform preprocessing on a speech signal comprising the user speech based on the environment information and the user information.

[11] The processor, by executing the instruction, may be further configured to perform preprocessing by identifying a preprocessing filter for removing noise comprised in the speech signal based on the environment information, and perform preprocessing by identifying a parameter for enhancing the user speech comprised in the speech signal based on the user information.

[12] Each of the plurality of speech recognition models may comprise a language model and an acoustic model, and the processor, by executing the instruction, may be further configured to identify a language model and an acoustic model corresponding to the user information, and obtain a text data on the user speech by using the identified language model and the acoustic model.

[13] The processor, by executing the instruction, may be further configured to identify a natural language understanding model of a plurality of natural language understanding models based on the user information, and perform a natural language understanding on the obtained text data through the identified natural language understanding model.

[14] The processor, by executing the instruction, may be further configured to obtain a response information on the user speech based on a result of the natural language understanding, and output a response speech on the user by inputting the response information to a text to speech (TTS) model corresponding to the user speech of a plurality of TTS models.

[15] The processor, by executing the instruction, may be further configured to identify an output method on the response information based on the user information, and the output method on the response information may comprise a method of outputting the response information through a display and a method of outputting the response information through a speaker.

[16] The processor, by executing the instruction, may be further configured to obtain the user information by inputting the obtained user image to an object recognition model trained to obtain information on an object comprised in the obtained user image.

[17] The memory may store a user image of a registered user matched with a registered user information, and The processor, by executing the instruction, may be further configured to obtain the user information by comparing the obtained user image with the user image of the registered user.

[18] The memory may comprise a plurality of wake-up models, and the processor, by executing the instruction, may be further configured to identify a wake-up word in the user speech based on a wake-up model, of the plurality of wake-up models, corresponding to the user information.

[19] In accordance with another aspect of the disclosure, a control method of an electronic apparatus is provided. The control method includes obtaining a user image by photographing a user through a camera, obtaining information on the user based on the user image, and based on a user speech being input from the user through the microphone, recognizing the user speech by using a speech recognition model corresponding to the user information among a plurality of speech recognition models.

[20] The control method may further comprise obtaining an environment information based on information on an object other than the user comprised in the user image by analyzing the user image and performing preprocessing on a speech signal comprising the user speech based on the environment information and the user information.

[21] The performing of the preprocessing may comprise performing preprocessing by identifying a preprocessing filter for removing noise comprised in the speech signal based on the environment information and performing preprocessing by identifying a parameter for enhancing the user speech comprised in the speech signal based on the user information.

[22] Each of the plurality of the speech recognition models may comprise a language model and an acoustic model, and the recognizing may comprise identifying a

language model and an acoustic model corresponding to the user information and obtaining a text data on the user speech by using the identified language model and the acoustic model.

[23] The control method may further comprise identifying a natural language understanding model of a plurality of natural language understanding models based on the user information and performing natural language understanding on the obtained text data through the identified natural language understanding model.

[24] The control method may further comprise obtaining a response information on the user speech based on a result of the natural language understanding and outputting a response speech on the user by inputting the response information to a text to speech (TTS) model corresponding to the user speech of a plurality of TTS models.

[25] The control method may further comprise identifying an output method on the response information based on the user information, and the output method of the response information may comprise a method of outputting the response information through a display and a method of outputting the response information through a speaker.

[26] The obtaining of the user information may comprise obtaining the user information by inputting the obtained user image to an object recognition model trained to obtain information on an object comprised in the obtained user image.

[27] The electronic apparatus may matche and store a user image of a registered user with a registered user information, and The obtaining the user information may comprise obtaining the user information by comparing the obtained user image with the user image of the registered user.

[28] The electronic apparatus may comprise a plurality of wake-up models, and the method may further comprise identifying a wake-up word in the user speech based on a wake-up model, of the plurality of wake-up models, corresponding to the user information.

[29] Other aspects, advantages, and salient features of the disclosure will become apparent to those skilled in the art from the following detailed description, which, taken in conjunction with the annexed drawings, discloses various embodiments of the disclosure.

Advantageous Effects of Invention

[30] As described above, according to an embodiment, because the electronic apparatus provides a response on the user speech by selecting the speech recognition corresponding to the user information and the neural network model for the natural language understanding, usability of the electronic apparatus may be greatly improved for being able to provide a response suitable to the user characteristic.

Brief Description of Drawings

[31] The above and other aspects, features and advantages of certain embodiments of the disclosure will be more apparent from the following description taken in conjunction with the accompanying drawings, in which:

[32] FIG. 1 is a diagram illustrating an embodiment of providing a response to a user by selecting a neural network model comprised in a dialogue system based on a user information according to an embodiment of the disclosure;

[33] FIG. 2 is a drawing of a block diagram schematically illustrating a configuration of an electronic apparatus according to an embodiment of the disclosure;

[34] FIG. 3 is a drawing illustrating a block diagram comprising a configuration for providing a response to a user speech according to an embodiment of the disclosure;

[35] FIG. 4 is a diagram illustrating a method for providing a user speech by obtaining a user information based on a user image and selecting a neural network model of a dialogue system based on the obtained user information according to an embodiment of the disclosure;

[36] FIGS. 5A and 5B are diagrams illustrating an embodiment of providing a response by different methods based on an age group of a user according to various embodiments of the disclosure;

[37] FIG. 6 is a flowchart illustrating a control method of an electronic apparatus according to an embodiment of the disclosure;

[38] FIG. 7 is a sequence diagram illustrating an electronic apparatus in association with a server providing a response to a user according to an embodiment of the disclosure; and

[39] FIG. 8 is a drawing of a block diagram illustrating in detail a configuration of an electronic apparatus according to an embodiment of the disclosure.

[40] Throughout the drawings, like reference numerals will be understood to refer to like parts, components, and structures.

Best Mode for Carrying out the Invention

[41] The following description with reference to the accompanying drawings is provided to assist in a comprehensive understanding of various embodiments of the disclosure as defined by the claims and their equivalents. It includes various specific details to assist in that understanding but these are to be regarded as merely exemplary. Accordingly, those of ordinary skill in the art will recognize that various changes and modifications of the various embodiments described herein can be made without departing from the scope and spirit of the disclosure. In addition, descriptions of well-known functions and constructions may be omitted for clarity and conciseness.

[42] The terms and words used in the following description and claims are not limited to the bibliographical meanings, but, are merely used by the inventor to enable a clear and

consistent understanding of the disclosure. Accordingly, it should be apparent to those skilled in the art that the following description of various embodiments of the disclosure is provided for illustration purpose only and not for the purpose of limiting the disclosure as defined by the appended claims and their equivalents.

[43] It is to be understood that the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a component surface" includes reference to one or more of such surfaces.

[44] In the disclosure, expressions such as "comprise," "may comprise," "consist of," "may consist of," or the like are used to designate a presence of a corresponding characteristic (e.g., elements such as numerical value, function, operation, or component, etc.), and not to preclude a presence or a possibility of additional characteristics.

[45] In the disclosure, expressions such as "A or B," "at least one of A and/or B," or "one or more of A and/or B" should be understood to include all possible combinations of the items listed together. For example, "A or B," "at least one of A and B," or "at least one of A or B" should be understood to represent all cases including (1) at least one of A, (2) at least one of B, or (3) at least one of A and at least one of B.

[46] The expressions such as "first," "second," "1st," or "2nd" used herein may denote various elements, regardless of order and/or importance, and may be used to distinguish one element from another, without limiting the corresponding elements. For example, a first user device and a second user device may indicate the different user devices, regardless of order and importance. For example, a first element may be denoted as a second element, and similarly a second element may also be denoted as a first element without departing from the scope of the disclosure.

[47] The term such as "module," "unit," "part," and so on is used to refer to an element that performs at least one function or operation, and such element may be implemented as hardware or software, or a combination of hardware and software. Further, except for when each of a plurality of "modules," "units," "parts," and the like need to be implemented in an individual hardware, the components may be integrated in at least one module or chip and implemented in at least one processor.

[48] When a certain element (e.g., first element) is indicated as being "(operatively or communicatively) coupled with/to" or "connected to" another element (e.g., second element), it could be understood as the certain element being directly coupled with/to the other element or as being coupled through another element (e.g., third element). On the other hand, when a certain element (e.g., first element) is indicated as "directly coupled with/to" or "connected to" another element (e.g., second element), it could be understood as another element (e.g., third element) not being present between the certain element and the other element.

[49] The expression "configured to... (or set up to)" used in the disclosure may be used in-

terchangeably with, for example, "suitable for" "having the capacity to..." "designed to..." "adapted to..." "made to,,," or "capable of..." based on circumstance. The term "configured to...(or set up to)" may not necessarily mean "specifically designed to..." in terms of hardware. Rather, in a certain circumstance, the expression "a device configured to" may mean something that the device "may be configured with..." together with another device or components. For example, the phrase "a processor configured to (or set up to) perform A, B, or C" may mean a processor dedicated to perform a corresponding operation (e.g., embedded processor), or a generic-purpose processor (e.g., a central processing unit (CPU) or an application processor (AP)) capable of performing the corresponding operations by executing one or more software programs stored in the memory apparatus.

[50] The terms used herein have merely been used to describe a specific embodiment, and not to limit the scope of another embodiment. A singular expression includes a plural expression, unless otherwise specified. The terms used in the disclosure, including technical or scientific terms, may have the same meaning as the terms generally understood by those of ordinary skill in the related field of art. Of the terms used herein, the terms which are defined in a typical dictionary may be interpreted to meanings identical or similar to the contextual meanings thereof in the related art. Unless clearly defined otherwise, the terms may not be interpreted to ideal or excessively formal meanings. In some cases, even if the term is defined in the disclosure, the terms may not be interpreted to exclude the embodiments of the disclosure.

[51] The disclosure will be described in greater detail below with reference to the accompanying drawings. However, in describing the disclosure, in case it is determined that the detailed description of related known technologies may unnecessarily confuse the gist of the disclosure, the detailed description thereof will be omitted. With respect to the description on the drawings, like reference numerals may be used to indicate like elements.

[52] The disclosure has been devised to solve the above-described problem, and an object of the disclosure lies in providing an electronic apparatus capable of obtaining a user information based on a user image and providing a response to a user speech by using a trained neural network model corresponding to the obtained user information, and a control method thereof.

[53] The disclosure will be described in greater detail below with reference to the accompanying drawings.

[54] FIG. 1 is a diagram illustrating an embodiment of providing a response to a user by selecting a neural network model comprised in a dialogue system based on a user information according to an embodiment of the disclosure.

[55] Referring to FIG. 1, the electronic apparatus according to an embodiment may be a

robot, but this is merely one embodiment, and may be implemented as a portable terminal such as, for example, and without limitation, a smart phone, tablet personal computer (PC), and the like or a home appliance product or the like such as, for example, and without limitation, a television (TV), a refrigerator, a washer, and the like. The electronic apparatus 100 may include a camera for photographing a user and a microphone for receiving input of a user speech, but this is merely one embodiment, and may be coupled with an external camera and an external microphone.

[56] The electronic apparatus 100 may obtain a user image by photographing the user through the camera. If the electronic apparatus 100 is implemented as a robot and if the user is approaching in the direction of the electronic apparatus 100, the electronic apparatus 100 may detect a proximity of the user by using a sensor. When the user is detected as being in close proximity, the electronic apparatus 100 may obtain a user image by photographing the user. The user image may not only include the user, but also an object (e.g., furniture, home appliance product, etc.) included in the environment the electronic apparatus 100 is located.

[57] The electronic apparatus 100 may obtain a user information based on the obtained user image. In an embodiment, the electronic apparatus 100 may obtain user information by inputting a user image to an object recognition model trained to recognize an object. In another embodiment, the electronic apparatus 100 may identify whether a user is a registered user based on a specific body part (e.g., iris, face, etc.) included in the user image. If identified as a registered user, the electronic apparatus 100 may obtain the registered user information. The user information may include information on an age group, gender, disabilities of the user.

[58] In addition, the electronic apparatus 100 may analyze the user image and obtain environment information based on information on an object other than the user included in the user image. The environment information may be information on an environment the user is located, and may include information on a place the user is located, and information on an object included in the place the user is located.

[59] When user information is obtained, the electronic apparatus 100 may execute a wake-up model corresponding to user information among a plurality of wake-up models. A wake-up word may be a word for executing or loading an artificial intelligence program (e.g., dialogue system) to provide a response to a user speech. The wake-up model, as a neural network model capable of recognizing the wake-up words, may be trained according to user characteristics. For example, the wake-up model may include a first wake-up model trained with a wake-up word uttered by children (e.g., ages 4 to 12), a second wake-up model trained with a wake-up word uttered by adults (ages 13 to 59), and a third wake-up model trained with a wake-up word uttered by seniors (e.g., ages 60 and over). That is, if the recognized user is a child, the electronic apparatus

100 may execute a first wake-up model.

[60] In addition, if a user is detected as being in close proximity, the electronic apparatus 100 may output a guide message 10 guiding on a wake-up word to control the electronic apparatus 100. For example, the electronic apparatus 100 may output a guide message 10 indicating "hello young customer, please call me retail-bot and tell me how I may assist you" as illustrated in FIG. 1. In FIG. 1, the guide message 10 has been output in an auditory form through a speaker, but this is merely one embodiment, and the guide message 10 may be output in a visual form through a display, or output concurrently in the visual form and the auditory form.

[61] If a user utters a wake-up word 20, the electronic apparatus 100 may recognize the wake-up word 20 uttered by the user. For example, if the user utters "retail-bot" as illustrated in FIG. 1, the electronic apparatus 100 may recognize the wake-up word 20 uttered by the user. The electronic apparatus 100 may identify whether the wake-up word 20 is included among the user utterances by using the wake-up model corresponding to the user information of the plurality of wake-up models. For example, if the recognized user is a child, the electronic apparatus 100 may identify whether the wake-up word 20 was included in a speech uttered by the user based on the first wake-up model.

[62] When the wake-up word 20 is recognized, the electronic apparatus 100 may execute models corresponding to the user information among a plurality of models included in the dialogue system. For example, the electronic apparatus 100 may execute a speech recognition model corresponding to the user information of a plurality of speech recognition models, execute a natural language understanding model corresponding to the user information of a plurality of natural language understanding models, and execute a TTS model corresponding to the user information of a plurality of TTS models. The electronic apparatus 100 may execute a plurality of neural network models concurrently, but this is merely one embodiment, and may also execute the plurality of neural network models sequentially.

[63] The electronic apparatus 100 may receive input of an audio signal included with a user speech uttered by the user through the microphone. For example, the user speech 30 may be a speech asking, "which are the most popular smart phones currently being sold?" The electronic apparatus 100 may detect a user speech section from the audio signal which includes the user speech 30.

[64] The electronic apparatus 100 may perform a preprocessing on the user speech 30. The electronic apparatus 100 may perform preprocessing on a speech signal included with a user speech based on an environment information and a user information. Specifically, the electronic apparatus 100 may perform preprocessing by identifying a preprocessing filter for removing noise included in the speech signal based on the en-

vironment information. In addition, the electronic apparatus 100 may perform pre-processing by identifying a parameter for enhancing user speech included in the speech signal based on the user information.

[65] When preprocessing is performed, the electronic apparatus may perform a speech recognition on the user speech 30. Specifically, the electronic apparatus 100 may obtain a text corresponding to the user speech 30 by performing speech recognition on the user speech 30 through a speech recognition model corresponding to user information of a plurality of speech recognition models. Specifically, the electronic apparatus 100 may store a plurality of speech recognition models trained for each user characteristic (e.g., per age, per gender, whether or not there is disability, etc.), and perform speech recognition on the user speech 30 by using the speech recognition model corresponding to user information of the plurality of speech recognition models. For example, the electronic apparatus 100 may perform speech recognition on the user speech 30 by using the speech recognition model trained based on the speech of a child.

[66] When the text is obtained through the speech recognition model corresponding to the user information, the electronic apparatus 100 may perform a natural language understanding based on the obtained text. The electronic apparatus 100 may perform a natural language understanding on the obtained text by using a natural language understanding model corresponding to the user information of a plurality of natural language understanding models trained for each user characteristic, and may obtain an intent of the user speech and a slot for performing the intent as a performance result. The electronic apparatus 100 may perform the natural language understanding on the user speech 30 by using the natural language understanding model trained based on the text frequently uttered by children.

[67] When the intent and slot of the user speech 30 is obtained, the electronic apparatus 100 may obtain a response information on the user speech 30 based on the intent and slot of the user speech 30. The electronic apparatus 100 may obtain the response information on the user speech 30 by using a knowledge database (DB) corresponding to information on the user information. For example, the electronic apparatus 100 may obtain the response information on the user speech 30 by using the knowledge DB corresponding to a child.

[68] When the response information on the user speech 30 is obtained, the electronic apparatus 100 may generate a natural language based on the response information. The electronic apparatus 100 may generate a response in a form of a natural language by using a natural language generating model corresponding to user information of the plurality of natural language generating models. For example, the electronic apparatus 100 may generate a response in the form of the natural language of, "The product most

popularly sold among kids-phones is the XX phone. Let me show you."

[69] The electronic apparatus 100 may perform a text to speech (TTS) operation to output a response in the form of a natural language to a speech data. The electronic apparatus 100 may generate a response in the form of a natural language to a speech data by using a TTS model corresponding to the user information of a plurality of TTS models. For example, the electronic apparatus 100 may generate the response in the form of a natural language to speech data by using the TTS model (e.g., a TTS model in a voice of an animation character preferred by a child) corresponding to a child of the plurality of TTS models.

[70] The electronic apparatus 100 may output a response generated in the form of a natural language. Specifically, as illustrated in FIG. 1, the electronic apparatus 100 may, through the speaker, output the response 40 in the form a natural language of, "The product most popularly sold among kids-phones is the XX phone. Let me show you."

[71] In addition, the electronic apparatus 100 may identify an output method on the response information based on the user information. The output method on the response information may include a method of outputting a response information through a display, and a method of outputting the response information through the speaker. For example, if the user is a child, the electronic apparatus 100 may, as illustrated in FIG. 1, provide the response information by using both the method of outputting the response information through the display and the method of outputting the response information through the speaker. However, if the user is an adult, the electronic apparatus 100 may provide the response information by using the method of outputting the response information through the speaker.

[72] The electronic apparatus 100 may store each of a plurality of models included in a dialogue system to provide a response to the user speech, but this is merely one embodiment, and may receive a model corresponding to the user information of the plurality of models from an external server. For example, the electronic apparatus 100 may transmit user information to a server storing the plurality of speech recognition models trained for each user characteristic, and perform speech recognition on the user speech 30 by using a speech recognition model corresponding to user information from the server.

[73] FIG. 2 is a drawing of a block diagram schematically illustrating a configuration of an electronic apparatus according to an embodiment of the disclosure.

[74] Referring to FIG. 2, the electronic apparatus 100 may include a camera 110, a microphone 120, a memory 130, and a processor 140. However, the configuration of FIG. 2 is merely one embodiment, and other configurations may be added or some configurations may be omitted according to the embodiment of the electronic apparatus 100.

[75] The camera 110 may obtain a user image by photographing a user through an image sensor. The user image may include not only the user but also an object (e.g., furniture, home appliance, display item, etc.) included in an environment the user is located. In addition, a plurality of cameras may be included in the main body of the electronic apparatus 100, but this is merely one embodiment, and a plurality of cameras may be electrically coupled with the electronic apparatus 100 by being externally located.

[76] The microphone 120 may receive input of an audio signal including the user speech. The microphone 120 may receive an audio signal including a wake-up word (or trigger word) and an audio signal including the user speech for inquiring about a specific information or controlling an apparatus. The microphone 120 may be provided in plurality on the main body of the electronic apparatus 100, but this is merely one embodiment, and may be electrically coupled with the electronic apparatus 100 by being externally located.

[77] The memory 130 may store at least one instruction or data related to another element of the electronic apparatus 100. Specifically, the memory 130 may include a non-volatile memory and a volatile memory, and may, for example, be implemented as a flash-memory, a hard disk drive (HDD), a solid state drive (SSD), or the like. The memory 130 may be accessed by the processor 140, and reading, writing, modifying, deleting, updating, or the like of the data may be performed by the processor 140. In addition, the memory 130 may store an artificial intelligence agent for operating a dialogue system. Specifically, the electronic apparatus 100 may use the artificial intelligence agent to generate a natural language in response to user utterance or obtain a control instruction. The artificial intelligence agent may be a program dedicated to providing an artificial intelligence based service (e.g., a speech recognition server, a personal assistant service, a translation server, a search service, etc.). Specifically, the artificial intelligence agent may be executed by a generic-purpose processor (e.g., CPU) or a separate AI dedicated processor according to the related art (e.g., graphics processing unit (GPU), neural processing unit (NPU), etc.). In addition, the memory 130 may include a plurality of configurations (or modules) comprising the dialogue system as illustrated in FIG. 3, which will be described in greater detail with reference to FIG. 3.

[78] In addition, the memory 130 may include a plurality of artificial intelligence models (or neural network models) with respect to each of the modules for providing a response to the user speech. For example, the memory 130 may include a plurality of wake-up models, a plurality of speech recognition models, a plurality of natural language understanding models, a plurality of natural language generating models, a plurality of TTS models, and the like.

[79] The processor 140 may be electrically coupled with the memory 130 and control the

overall operation of the electronic apparatus 100. Specifically, the processor 140 may, by executing at least one instruction stored in the memory 130, obtain a user image by photographing the user through the camera 110, obtain a user information based on the user image, and based on a user speech being input from the user through the microphone 120, recognize the user speech by using a speech recognition model corresponding to the user information of the plurality of speech recognition models.

[80] Specifically, the processor 140 may obtain the user information by analyzing the user image. In an embodiment, the processor 140 may obtain user information by inputting the obtained user image to the object recognition model, which is trained to obtain information on an object included in the image, and may obtain the user information by comparing the user image on at least one user pre-registered in the memory 130 and the user image obtained by the camera 110.

[81] The processor 140 may perform preprocessing on the speech signal. Specifically, the processor 140 may obtain not only the user information but also an environment information based on information on an object other than the user included in the user image by analyzing the user image, and may perform preprocessing on the speech signal included with the user speech based on the environment information and the user information. Specifically, the processor 140 may perform preprocessing by identifying a preprocessing filter to remove noise included in the speech signal based on the environment information, and may perform preprocessing by identifying a parameter for enhancing user speech included in the speech signal based on the user information.

[82] The processor 140 may identify whether a wake-up word is included in the user speech based on the wake-up model corresponding to the user information of the plurality of wake-up models. If the wake-up word is included in the user speech, the processor may execute a dialogue system based on the wake-up word.

[83] In addition, the processor 140 may identify a language model and an acoustic model corresponding to the user information, and obtain a text data on the user speech by using the identified language model and the acoustic model. The processor 140 may then identify one of the plurality of natural language understanding models based on the user information, and perform a natural language understanding on the text data obtained through the identified natural language understanding model. The processor 140 may obtain a response information on the user speech based on a result of the natural language understanding, and output a response speech on the user by inputting the response information to the TTS model corresponding to the user speech of the plurality of TTS models. The processor 140 may identify an output method (e.g., an auditory output method or a visual output method) on the response information based on the user information.

[84] Specifically, a function related to the artificial intelligence according to the disclosure may be operated through the processor 140 and the memory 130. The processor 140 may be configured as one or a plurality of processors. The one or plurality of processors may be a generic-purpose processor such as, for example, and without limitation, a CPU, an AP, a digital signal processor (DSP), or the like, a graphics dedicated processor such as, for example, and without limitation, a GPU, a vision processing unit (VPU), or the like, or an AI dedicated processor such as, for example, and without limitation, an NPU. The one or plurality of processors may control to process the input data through a pre-defined operation rule or an artificial intelligence model stored in the memory 130. Alternatively, if the one or plurality of processors is an AI dedicated processor, the AI dedicated processor may be designed to a hardware structure specializing in the processing of a specific artificial intelligence model.

[85] The pre-defined operation rule or artificial intelligence model is characterized by being generated through learning. Here, being generated through learning refers to a pre-defined operation rule or an artificial intelligence model set to perform a desired feature (or, object) being generated as a basic artificial intelligence model is trained by a learning algorithm using a plurality of training data. The learning may be carried out in the device in which the artificial intelligence according to the disclosure is performed on its own, or carried out through a separate server and/or system. Examples of the learning algorithm may include a supervised learning, an unsupervised learning, a semi-supervised learning, or a reinforcement learning, but the embodiment is not limited thereto.

[86] The artificial intelligence model may be comprised of a plurality of neural network layers. Each of the plurality of neural network layers may include a plurality of weighted values, and perform a neural network processing through a processing result of a previous layer and the processing between a plurality of weighted values. The plurality of weight values included in the plurality of network layers may be optimized by the learning result of the artificial intelligence model. For example, the plurality of weight values may be updated so that a loss value or a cost value obtained by the artificial intelligence model during the learning process is reduced or minimized. The artificial neural network may include a Deep Neural Network (DNN), and other examples may include, for example, a Convolutional Neural Network (CNN), a Deep Neural Network (DNN), a Recurrent Neural Network (RNN), a Restricted Boltzmann Machine (RBM), a Deep Belief Network (DBN), a Bidirectional Recurrent Deep Neural Network (BRNDD), a Deep Q-Networks, or the like, but the embodiment is not limited thereto.

[87] The linguistic understanding may, as a technique for recognizing and applying/

processing the language/characters of humans, include natural language processing, machine translation, dialogue system, question answering, speech recognition/ synthesis, and the like.

[88] FIG. 3 is a drawing illustrating a block diagram comprising a configuration for providing a response to a user speech according to an embodiment of the disclosure.

[89] Referring to FIG. 3, the configurations may be a configuration for performing a dialogue with a virtual artificial intelligence agent through a natural language or controlling the electronic apparatus 100, and the electronic apparatus 100 may include an object recognition module 310, a preprocessing module 320, a dialogue system 330, and an output module 340. The dialogue system 330 may be included with a wake-up word recognition module 331, a speech recognition module 332, a natural language understanding module 333, a dialogue manager module 334, a natural language generating module 335, and a TTS module 336. According to an embodiment, the modules included in the dialogue system 330 may be stored in the memory 130 of the electronic apparatus 100, but this is merely one embodiment, and may be implemented as a combined form of a hardware and a software. In addition, the at least one module included in the dialogue system 330 may be included in at least one external server.

[90] The object recognition module 310 may obtain information on the user and the object included in the user image photographed through the camera 110 by using the object recognition model trained to recognize the object. Specifically, the object recognition module 310 may obtain the age group, gender, whether or not there is disability, physical information, and the like of the user included in the user image through the object recognition model. In addition to the above, the object recognition module 310 may obtain information on the object included in the user image through the object recognition model, and obtain the environment information based on the information on the object. For example, the object recognition module 310 may obtain information on a place the user is located based on the information on the object, information on the object that generates noise around the user, and the like.

[91] The preprocessing module 320 may perform preprocessing on an audio signal received through the microphone 120. Specifically, the preprocessing module 320 may receive an audio signal in an analog form including the user speech through the microphone, and may convert the analog signal to a digital signal. The preprocessing module 320 may then extract a user speech section by calculating an energy of the converted digital signal.

[92] Specifically, the preprocessing module 320 may identify whether an energy of the digital signal is greater than or equal to a pre-set value. If the energy of the digital signal is greater than or equal to a pre-set value, the preprocessing module 320 may either remove noise on the input digital signal or enhance user speech by identifying as

the speech section, or if the energy of the digital signal is less than or equal to a pre-set value, the preprocessing module 320 may not perform signal processing on the input digital signal and may wait for another input. Accordingly, unnecessary power consumption may be prevented because the audio processing process as a whole is not activated by a sound that is not the user speech.

[93] The preprocessing module 320 may perform preprocessing which performs filtering on the audio signal including the user speech based on the environment information, and preprocessing for enhancing the audio signal including the user speech based on the user information. Specifically, the preprocessing module 320 may select a filtering parameter corresponding to a place information the user is currently located of a plurality of filtering parameters, and perform a filtering operation based on the selected filtering parameter. The electronic apparatus 100 may perform the filtering operation by using the different filtering parameters according to whether the place the user is currently located is the kitchen, the living-room, the bathroom. In addition, the preprocessing module 320 may select an enhancing parameter corresponding to the current user information, and perform the enhancement preprocessing on the user speech by using the selected enhancement parameter. In an example, if the user is a child, the preprocessing module 320 may perform the enhancement preprocessing on the user speech by using the enhancement parameter which enhances a high-pitch range of the user speech. In another example, if the user is an adult male, the preprocessing module 320 may perform enhancement preprocessing on the user speech by using the enhancement parameter which enhances a low-pitch range of the user speech.

[94] The wake-up word recognition module 331 may identify whether the wake-up word is included in the user speech through the wake-up model. The wake-up word (or trigger word) may be a command (e.g., Bixby, Galaxy) indicating that the user has started the speech recognition. So, the electronic apparatus 100 may execute the dialogue system in response to the wake-up word being input. The wake-up word may be pre-set at the time of manufacture, but this is merely one embodiment, and may be changed according to user setting. The wake-up word recognition module 331 may include a plurality of wake-up models trained for each user characteristic. The wake-up word recognition module 331 may then identify whether the wake-up word is included in the user speech by using a wake-up model corresponding to the user information of the plurality of wake-up models.

[95] The speech recognition module 332 may convert the user speech in the form of an audio data received from the preprocessing module 320. The speech recognition module 221 may include a plurality of speech recognition models trained for each user characteristic, and each of the plurality of speech recognition models may include an acoustic model and a language model. The acoustic model may include information as-

sociated with vocalization and the language model may include information on unit phoneme information and information on a combination of unit phoneme information. The speech recognition module 332 may convert the user speech to text data by using information associated with vocalization and information on unit phoneme information. The information on the acoustic model and the language model may, for example, be stored in an automatic speech recognition database (ASR DB). Specifically, the speech recognition module 332 may include a plurality of language models and acoustic models trained for each of the plurality of user characteristics (e.g., age group, gender, etc.). For example, the speech recognition module 332 may be a first language model trained with a corpus including words frequently used by children, and a second language model trained with a corpus including words frequently used by adults. In another example, the speech recognition module 332 may include a first acoustic model trained with a voice of a child, and a second acoustic model trained with a voice of an adult.

[96] The speech recognition module 332 may perform speech recognition on the user speech by using a speech recognition model corresponding to user information of the plurality of speech recognition models. For example, if the user is identified as a child based on the user information, the speech recognition module 332 may perform speech recognition on the user speech by using a first language model trained with a corpus including words frequently used by the child and a first acoustic model trained through the voice of the child.

[97] The natural language understanding module 333 may identify a domain and a user intent on the user speech by performing a syntactic analysis or a semantic analysis based on the text data on the user speech obtained through speech recognition. The syntactic analysis divides the user input into syntactic units (e.g., word, phrase, morpheme, etc.), and identify which syntactic elements the divided units include. The semantic analysis may be performed by using sematic matching, rule matching, formula matching and the like.

[98] The natural language understanding module 333 may include a plurality of natural language understanding models trained by each user characteristic, and may perform natural language understanding based on the natural language understanding model corresponding to the user information of the plurality of natural language understanding models. For example, if the user is identified as a child based on the user information, the natural language understanding module 333 may obtain a natural language understanding on the user speech by using the natural language understanding model trained based on the text frequently used by the child.

[99] The natural language understanding module 333 may obtain the natural language understanding result, category of user speech, intent of user speech, and a slot (or, entity,

parameter, etc.) for performing the intent of the user speech.

[100] The dialogue manager module 334 may obtain a response information on the user speech based on the user intent and slot obtained in the natural language understanding module 333. The dialogue manager module 334 may provide a response on the user speech based on the knowledge DB. The knowledge DB may be included in the electronic apparatus 100, but this is merely one embodiment, and may be included in an external server. In addition, the dialogue manager module 334 may include a plurality of knowledge DB for each user characteristic, and obtain a response information on the user speech by using the knowledge DB corresponding to the user information of the plurality of knowledge DB. For example, if the user is identified as a child based on the user information, the dialogue manager module 334 may obtain a response information on the user speech by using the knowledge DB corresponding to the child.

[101] In addition, the dialogue manager module 334 may identify whether the user intent identified by the natural language understanding module 333 is clear. For example, the dialogue manager module 334 may identify whether the user intent is clear based on whether there is sufficient information on the slot. In addition, the dialogue manager module 334 may identify whether the slot identified in the natural language understanding module 333 is sufficient for performing a task. According to an embodiment, the dialogue manager module 334 may perform feedback requesting necessary information to the user if the user intent is not clear.

[102] The natural language generating module 335 may change the response information or designated information obtained through the dialogue manager module 334 to a text form. The information changed to the text form may be in the form of a natural language utterance. The designated information may, for example, be information on additional input, information guiding a completion of operation corresponding to the user input, or information (e.g., feedback information on user input) guiding for an additional input by the user. The information changed to the text form may be displayed on a display of the electronic apparatus 100 or changed to a speech form by the TTS module 336.

[103] The natural language generating module 335 may also include a plurality of natural language generating models trained to generate a natural language for each user characteristic or a plurality of natural language templates generated for each user characteristic. The natural language generating module 335 may generate a response on the user speech in a natural language form by using a natural language generating model (or a natural language template corresponding to user information of a plurality of natural language generating models) corresponding to the user information of the plurality of natural language generating models. For example, if the user is identified

as a child based on the user information, the natural language generating module 335 may generate a response on the user speech as a natural language by using the natural language generating model or the natural language template corresponding to the child.

[104] The TTS module 336 may change the information in the text form to an information in a speech form. The TTS module 336 may include a plurality of TTS models for generating a response in various voices, and the TTS module 336 may obtain a response speech in speech form by using the TTS model corresponding to the user information of the plurality of TTS models. For example, if the user is identified as a child based on the user information, the TTS module 336 may obtain a response speech by using the TTS model (e.g., a TTS model for generating a voice of an animation character favored by the child) corresponding to the child.

[105] The output module 340 may output information in a form of a speech data received from the TTS module 336. The output module 340 may output information in the form of speech data through the speaker or a speech output terminal. Alternatively, the output module 340 may output information in the form of text data obtained through the natural language generating module 335 through the display or an image output terminal.

[106] FIG. 4 is a diagram illustrating a method for providing a user speech by obtaining a user information based on a user image and selecting a model of a dialogue system based on the obtained user information according to an embodiment of the disclosure.

[107] Referring to FIG. 4, first, the electronic apparatus 100 may obtain a user image at operation S410. The electronic apparatus 100 may obtain the user image through the camera 110, but this is merely one embodiment, and may receive the user image from an external apparatus including a camera.

[108] The electronic apparatus 100 may obtain the user information and the environment information by analyzing the user image at operation S420. Specifically, the electronic apparatus 100 may obtain the user information based on the user image. In an embodiment, the electronic apparatus 100 may identify whether the user is a registered user by comparing a characteristic area (e.g., iris, face, etc.) of the user included in the user image with a pre-registered image. If identified as the registered user, the electronic apparatus 100 may obtain the user information based on information on the registered user. In another embodiment, the electronic apparatus 100 may obtain the user information included in the user image by inputting the user image to the trained object recognition model. The user information may include the age group, gender, whether or not there is disability, physical information, and the like of the user, but this is merely one embodiment, and may also include other user information. For example, the electronic apparatus 100 may identify that the user is a child by analyzing the user image. In addition, the electronic apparatus 100 may obtain the environment in-

formation based on information on another object included in the user image. The electronic apparatus 100 may obtain information on the object included in the user image by inputting the user image to the trained object recognition model, and obtain the environment information of the environment the user is located based on information on the obtained object. For example, the electronic apparatus 100 may identify that the place the user is located is a shop by analyzing the user image. That is, the electronic apparatus 100 may obtain information on the shop as information on the place the user is located.

[109] The electronic apparatus 100 may receive input of the user speech at operation S430. The user speech may be in a speech data form, and may include at least one of the wake-up word, the text for inquiring information, and the text for controlling the electronic apparatus. For example, the input user speech may be, "Retail-bot, which are the most sold smart phones?"

[110] The electronic apparatus 100 may perform preprocessing on the input user speech at operation S440. The electronic apparatus 100 may perform preprocessing by identifying a parameter which enhances the user speech based on the obtained user information. In addition, the electronic apparatus 100 may perform preprocessing by identifying a parameter for filtering an external noise based on the obtained environment information. For example, if the user is a child, the electronic apparatus 100 may perform preprocessing on the user speech by identifying a parameter for enhancing a high-pitch range. In addition, if the place the user is located is a shop, the electronic apparatus 100 may perform preprocessing on the audio signal which includes the user speech by identifying a filter for removing the noise generated in the shop.

[111] The electronic apparatus 100 may recognize the wake-up word in the obtained speech at operation S450. The wake-up word may be a word for executing the dialogue system, and may be a word such as, for example, "retail-bot." Specifically, the electronic apparatus 100 may recognize the wake-up word based on the wake-up model corresponding to the user information of the plurality of wake-up models capable of recognizing the wake-up word for each user characteristic.

[112] The electronic apparatus 100 may perform speech recognition at operation S460. The electronic apparatus may obtain the text data corresponding to the user speech by performing speech recognition based on the speech recognition model. Specifically, the electronic apparatus 100 may obtain the text data corresponding to the user speech based on the speech recognition model corresponding to the user information of the plurality of speech recognition models trained for each user characteristic.

[113] The electronic apparatus 100 may perform the natural language understanding at operation S470. The electronic apparatus 100 may identify the category, intent and slot

of the user speech by performing a natural language understanding on the text data based on the natural language understanding model. Specifically, the electronic apparatus 100 may obtain the category, intent, slot and the like corresponding to the user speech based on the natural language understanding model corresponding to the user information of the plurality of natural language understanding models trained for each user characteristic. For example, if the text data is "which are the most sold smart phones?" the electronic apparatus 100 may obtain "search" as the user intent, and "mobile phone" and "most sold" as the information on the slot through the natural language understanding model.

[114] The electronic apparatus 100 may obtain the response information at operation S480. The electronic apparatus 100 may obtain the response information based on the intent and slot of the user speech obtained based on the natural language understanding. The electronic apparatus 100 may obtain the response information by using the knowledge DB corresponding to the user information of the plurality of knowledge DB stored in the electronic apparatus 100. Alternatively, the electronic apparatus 100 may transmit the intent and slot of the user speech and the user information to the external server, and receive the response information obtained based on the transmitted intent and slot of the user speech and the user information. For example, if the user is a child, the electronic apparatus 100 may obtain the response information by using the knowledge DB on kids-phones.

[115] The electronic apparatus 100 may generate the natural language by using the response information at operation S490. The electronic apparatus 100 may generate a response in the form of a natural language by inputting the response information to the natural language generating model or the natural language template. The electronic apparatus 100 may generate a response in the form of a natural language by using the natural language generating model corresponding to the user information of the plurality of natural language generating models trained for each user characteristic. Alternatively, the electronic apparatus 100 may generate a response in the form of a natural language by using the natural language template corresponding to the user information of the plurality of natural language templates pre-registered for each user characteristic. For example, if the user is a child, the electronic apparatus 100 may generate a response in the form of a natural language by using the natural language generating model or the natural language template corresponding to the child.

[116] The electronic apparatus 100 may synthesize the speech on the generated natural language at operation S495. The electronic apparatus 100 may process the response in the form of a natural language generated by using the TTS model to a response speech in the form of a speech data. The electronic apparatus 100 may obtain the response speech by using the TTS model corresponding to the user information of the plurality

of TTS models trained for each user characteristic. For example, if the user is a child, the electronic apparatus 100 may obtain the response speech by using the TTS model capable of generating the speech of the animation character favored by the child.

[117] The electronic apparatus 100 may output the response at operation S497. The electronic apparatus 100 may identify an output method on the response information based on the user information. The output method on the response information may include a method of outputting the response information through the display and a method of outputting the response information through the speaker. For example, if the user is a child, the electronic apparatus 100 may output the response information by using both the method of outputting the response information through the display and the method of outputting the response information through the speaker, and if the user is an adult, the electronic apparatus 100 may output the response information by using the method of outputting the response information through the speaker. In addition, the electronic apparatus 100 may output the response information to different output states based on the user information.

[118] FIGS. 5A and 5B are diagrams illustrating an embodiment of providing a response by different methods based on an age group of a user according to various embodiments of the disclosure.

[119] Referring to FIGS. 5A and 5B, for example, if the user is a child, the electronic apparatus 100 may output the response information to a size of a first character, and if the user is a senior, the electronic apparatus 100 may output the response information to a size of a second character, which is larger than the size of the first character.

[120] FIG. 6 is a flowchart illustrating a control method of an electronic apparatus according to an embodiment of the disclosure.

[121] Referring to FIG. 6, the electronic apparatus 100 may obtain the user image by photographing the user at operation S610.

[122] The electronic apparatus 100 may obtain the user information based on the user image at operation S620. The electronic apparatus may obtain the user information by inputting the user image to the trained object recognition model, and obtain the user information by identifying whether the photographed user is a pre-registered user based on the user image.

[123] The electronic apparatus 100 may receive input of the user speech at operation S630.

[124] At operation S640, the electronic apparatus 100 may recognize the user speech by using the speech recognition model corresponding to the user information of the plurality of speech recognition models. That is, the electronic apparatus 100 may further raise the accuracy on speech recognition by recognizing the user speech based on the speech recognition model corresponding to the user information obtained by analyzing the image from the plurality of speech recognition models trained for each

user characteristic.

[125] In the above-described embodiment, the electronic apparatus 100 has been described as comprising all configurations, but this is merely one embodiment, and some configurations of the dialogue system may be included in the external server.

[126] FIG. 7 is a sequence diagram illustrating an electronic apparatus in association with a server providing a response to a user according to an embodiment of the disclosure. Meanwhile, operations S705 to S730 illustrated in FIG. 7 overlap with operations S410 to S460 described in FIG. 4 and thus, a detailed description thereof will be omitted.

[127] Referring to FIG. 7, the electronic apparatus 100 may transmit the obtained text data and the user information to the server 700 at operation S735.

[128] The server 700 may perform the natural language understanding based on the obtained text data and the user information at operation S740. Specifically, the server 700 may obtain the category, intent, slot, and the like corresponding to the user speech based on the natural language understanding model corresponding to the user information of the plurality of natural language understanding models trained for each user characteristic.

[129] The server 700 may obtain the response information at operation S745. Specifically, the server 700 may obtain the response information based on the intent and slot of the user speech obtained based on the natural language understanding. The server 700 may obtain the response information by using the knowledge DB corresponding to the user information of the plurality of knowledge DB. The server 700 may obtain the plurality of response information for each of the plurality of output methods based on the user information. For example, if the user is a child, the server 700 may obtain the response information provided through the display and the response information provided through the speaker.

[130] The server 700 may transmit the obtained response information to the electronic apparatus 100 at operation S750. The server 700 may transmit the information on the output method of the response information together with the response information thereto.

[131] The electronic apparatus 100 may generate the natural language by using the received response information at operation S755. The electronic apparatus 100 may generate a response in the form of a natural language by using the natural language generating model corresponding to the user information of the plurality of natural language generating models trained for each user characteristic. Alternatively, the electronic apparatus 100 may generate a response in the form of a natural language by using the natural language template corresponding to the user information of the plurality of natural language templates pre-registered for each user characteristic.

[132] The electronic apparatus 100 may synthesize the speech on the generated natural

language at operation S760. The electronic apparatus 100 may obtain the response speech by using the TTS model corresponding to the user information of the plurality of TTS models trained for each user characteristic. For example, if the user is a child, the electronic apparatus 100 may obtain the response speech by using the TTS model capable of generating the speech of the animation character favored by the child.

[133] The electronic apparatus 100 may output the response at operation S765. Specifically, the electronic apparatus 100 may output the response information based on the output method of the response information received from the server 700.

[134] In the above-described embodiment, the natural language understanding module and the dialogue manager module of the dialogue system has been described as being included in the server 700, but this is merely one embodiment. Further, other configurations (e.g., speech recognition module and natural language generating module) may be included in the server, and some configurations (e.g., one of the natural language understanding module or the dialogue manager module) may also be included in the electronic apparatus.

[135] FIG. 8 is a drawing of a block diagram illustrating in detail a configuration of an electronic apparatus according to an embodiment of the disclosure.

[136] Referring to FIG. 8, the electronic apparatus 800 may be implemented as a robot capable of movement, and the electronic apparatus 800 may include a camera 810, a microphone 820, a speaker 830, a display 840, a memory 850, a communication interface 860, an input interface 870, a sensor 880, a driving unit 890, and a processor 895. Because the camera 810, the microphone 820, the memory 850, and the processor 895 illustrated in FIG. 8 overlap with the camera 110, the microphone 120, the memory 130, and the processor 140 described in FIG. 2, the redundant descriptions will be omitted. In addition, according to an embodiment of the electronic apparatus 800, some of the configurations of FIG. 8 may be removed or other configurations may be added.

[137] The speaker 830 may be a configuration for the electronic apparatus 800 to audibly provide information. The electronic apparatus 800 may include one or more speakers 830, and may output a response on the input user speech, an inquiry on the user speech, alarm information, and the like as an audio signal through the speaker 830. The configuration for outputting the audio signal may be implemented as the speaker 830, but this is merely one embodiment, and may be implemented as an output terminal.

[138] The display 840 may be a configuration for the electronic apparatus 800 to visually provide information. The electronic apparatus 800 may include one or more displays 840, and display a response on the input user speech, inquiry on the user speech, alarm information, and the like through the display 840. The display 840 may be implemented as a liquid crystal display (LCD), a plasma display panel (PDP), an organic

light emitting diodes (OLED), transparent OLED (TOLED), a micro LED, and the like. In addition, the display 840 may be implemented in the form of a touch screen capable of detecting a touch operation of the user, and may also be implemented as a flexible display which may be folded or bent.

[139] The communication interface 860 may be a configuration which is capable of performing communication with the external apparatus. The communication interface 860 communicatively coupling with the external apparatus may include communicating through a third device (e.g., relay, hub, access point, server, gateway, or the like). The wireless communication may include a cellular communication which uses at least one of, for example, an long term evolution (LTE), an LTE advance (LTE-A), a code division multiple access (CDMA), a wideband CDMA (WCDMA), a universal mobile telecommunications system (UMTS), a wireless broadband (WiBro), or a global system for mobile communications (GSM). According to an embodiment, the wireless communication may include at least one of, for example, wireless fidelity (WiFi), Bluetooth, Bluetooth low energy (BLE), ZigBee, near field communication (NFC), magnetic secure transmission, radio frequency (RF) or body area network (BAN). The wired communication may include at least one of, for example, a universal serial bus (USB), a high definition multimedia interface (HDMI), a recommended standard 232 (RS-232), a power line communication (PLC), or plain old telephone service (POTS). A network in which the wireless communication or the wired communication is performed may include a telecommunication network, for example, at least one of a computer network (e.g., local area network (LAN) or wide area network (WAN)), the Internet, or a telephone network.

[140] The communication interface 860 may provide the dialogue system service by performing communication with an external server. Specifically, the communication interface 860 may transmit the user speech (or the text corresponding to the user speech) to the external server, and receive the response information on the user speech from the external server.

[141] Alternatively, the communication interface 860 may receive the user image or the user speech from the external apparatus which includes the camera, the microphone, and the like, and provide the user with the received user image or the user speech by transmitting the response information on the user speech to the external apparatus which includes the speaker or the display.

[142] The input interface 870 may be a configuration for receiving user input to control the electronic apparatus 800. For example, the input interface 870 may be implemented as a touch panel, a button, a dial, and the like to receive input of a user touch to control the electronic apparatus 800, but this is merely one embodiment, and may be implemented as an input apparatus such as a keyboard and a mouse.

[143] The sensor 880 may be a configuration for detecting information of a surrounding state of the electronic apparatus 800. Specifically, the sensor 880 may include a proximity sensor for detecting the proximity to the user or the object, a gyro sensor for obtaining movement information, an acceleration sensor, and the like, and may include a sensor for obtaining biometric information of a user and a sensor for obtaining information (e.g., temperature, humidity, etc.) on a space the electronic apparatus 800 is located.

[144] The driving unit 890 may be a configuration for moving the electronic apparatus 800. Specifically, the driving unit 890 may include an actuator for driving of the electronic apparatus 800. In addition, the actuator for driving the motion of other physical configurations (e.g., arm, etc.) of the electronic apparatus 800 may be included in addition to the driving unit 890. For example, the electronic apparatus 800 may control the actuator to travel or move in a direction of the user detected through the sensor 880.

[145] One or more embodiments may be implemented with software including instructions stored in a machine-readable storage media (e.g., computer). The machine may call an instruction stored in the storage medium, and as an apparatus capable of operating according to the called instruction, may include an electronic apparatus (e.g., electronic apparatus 100) according to embodiments. Based on the instruction being executed by the processor, the processor may directly or under the control of the processor perform a function corresponding to the instruction using different elements. The instructions may include a code generated by a compiler or executed by an interpreter. The machine-readable storage media may be provided in the form of a non-transitory storage medium. Herein, "non-transitory" merely means that the storage medium is tangible and does not include a signal and does not distinguish that data is permanently or temporarily stored in the storage medium. For example, the 'non-transitory storage medium' may include a buffer in which data is temporarily stored.

[146] According to an embodiment, a method according to one or more embodiments may be provided in a computer program product. The computer program product may be exchanged between a seller and a purchaser as a commodity. The computer program product (e.g., downloadable app) may be distributed in the form of a machine-readable storage medium (e.g., a compact disc read only memory (CD-ROM)) or distributed online through an application store (e.g., PLAYSTORE™). In the case of online distribution, at least a portion of the computer program product may be at least stored temporarily in a storage medium such as a manufacturer's server, a server of an application store, or a memory of a relay server, or temporarily generated.

[147] Each of the elements (e.g., a module or a program) according to various embodiments may be comprised of a single entity or a plurality of entities, and some sub-elements of the abovementioned sub-elements may be omitted, or different sub-

elements may be further included in various embodiments. Alternatively or additionally, some elements (e.g., modules or programs) may be integrated into one entity to perform the same or similar functions performed by each respective element prior to integration. Operations performed by a module, program, or other element, in accordance with various embodiments, may be performed sequentially, in a parallel, repetitive, or heuristically manner, or at least some operations may be performed in a different order, omitted or a different operation may be added.

[148] While the disclosure has been shown described with reference to various embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the disclosure as defined by the appended claims and their equivalents.

Claims

[Claim 1] An electronic apparatus, comprising:
a microphone;
a camera;
a memory storing an instruction; and
a processor configured to control the electronic apparatus coupled with the microphone, the camera and the memory,
wherein the processor, by executing the instruction, is further configured to:
obtain a user image by photographing a user through the camera,
obtain user information based on the user image, and
based on a user speech being input from the user through the microphone, recognize the user speech by using a speech recognition model, among a plurality of speech recognition models, corresponding to the user information.

[Claim 2] The electronic apparatus of claim 1, wherein the processor, by executing the instruction, is further configured to:
obtain an environment information based on information on an object other than the user comprised in the user image by analyzing the user image, and
perform preprocessing on a speech signal comprising the user speech based on the environment information and the user information.

[Claim 3] The electronic apparatus of claim 2, wherein the processor, by executing the instruction, is further configured to:
perform preprocessing by identifying a preprocessing filter for removing noise comprised in the speech signal based on the environment information, and
perform preprocessing by identifying a parameter for enhancing the user speech comprised in the speech signal based on the user information.

[Claim 4] The electronic apparatus of claim 2,
wherein each of the plurality of speech recognition models comprise a language model and an acoustic model, and
wherein the processor, by executing the instruction, is further configured to:
identify a language model and an acoustic model corresponding to the user information, and

obtain a text data on the user speech by using the identified language model and the acoustic model.

[Claim 5] The electronic apparatus of claim 4, wherein the processor, by executing the instruction, is further configured to:
identify a natural language understanding model of a plurality of natural language understanding models based on the user information, and
perform a natural language understanding on the obtained text data through the identified natural language understanding model.

[Claim 6] The electronic apparatus of claim 5, wherein the processor, by executing the instruction, is further configured to:
obtain a response information on the user speech based on a result of the natural language understanding, and
output a response speech on the user by inputting the response information to a text to speech (TTS) model corresponding to the user speech of a plurality of TTS models.

[Claim 7] The electronic apparatus of claim 6,
wherein the processor, by executing the instruction, is further configured to identify an output method on the response information based on the user information, and
wherein the output method on the response information comprises a method of outputting the response information through a display and a method of outputting the response information through a speaker.

[Claim 8] The electronic apparatus of claim 1, wherein the processor, by executing the instruction, is further configured to obtain the user information by inputting the obtained user image to an object recognition model trained to obtain information on an object comprised in the obtained user image.

[Claim 9] The electronic apparatus of claim 1,
wherein the memory stores a user image of a registered user matched with a registered user information, and
wherein the processor, by executing the instruction, is further configured to obtain the user information by comparing the obtained user image with the user image of the registered user.

[Claim 10] The electronic apparatus of claim 1,
wherein the memory comprises a plurality of wake-up models, and
wherein the processor, by executing the instruction, is further configured to identify a wake-up word in the user speech based on a

wake-up model, of the plurality of wake-up models, corresponding to the user information.

[Claim 11] A control method of an electronic apparatus, the method comprising:
obtaining a user image by photographing a user through a camera;
obtaining user information based on the user image; and
based on a user speech being input from the user through a microphone, recognizing the user speech by using a speech recognition model, among a plurality of speech recognition models, corresponding to the user information.

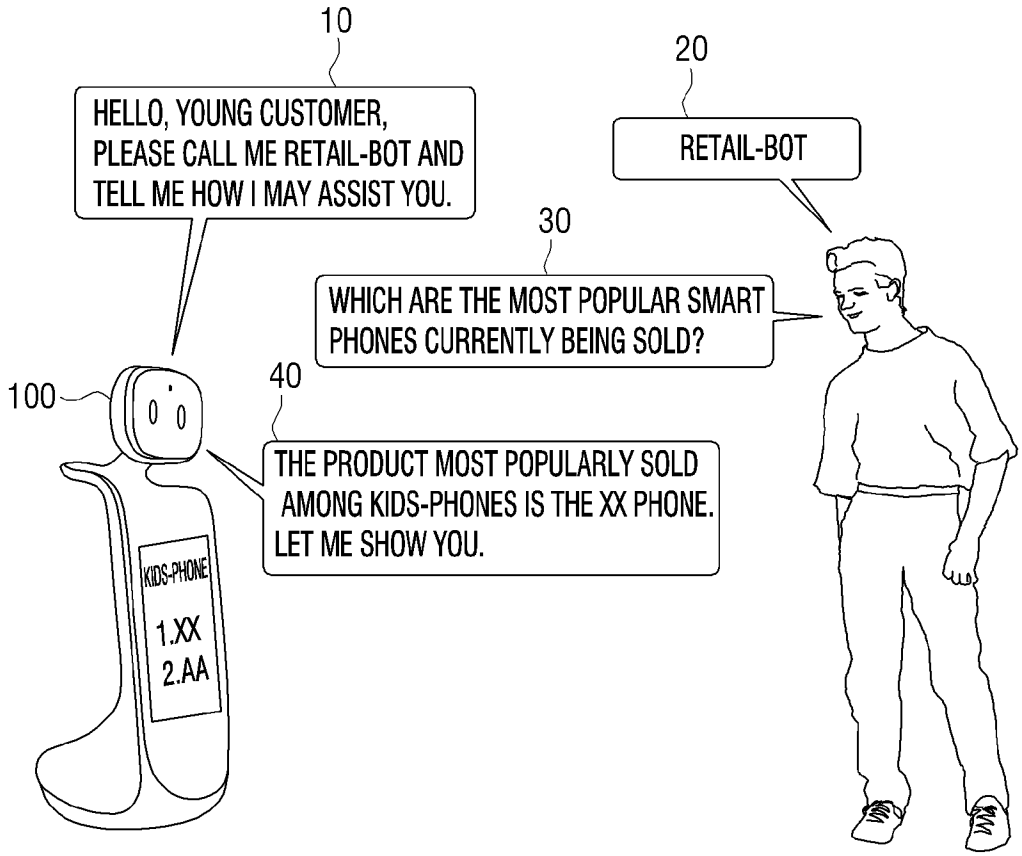
[Claim 12] The control method of claim 11, comprising:
obtaining an environment information based on information on an object other than the user comprised in the user image by analyzing the user image; and
performing preprocessing on a speech signal comprising the user speech based on the environment information and the user information.

[Claim 13] The control method of claim 12, wherein the performing of the pre-processing comprises:
performing preprocessing by identifying a preprocessing filter for removing noise comprised in the speech signal based on the environment information; and
performing preprocessing by identifying a parameter for enhancing the user speech comprised in the speech signal based on the user information.

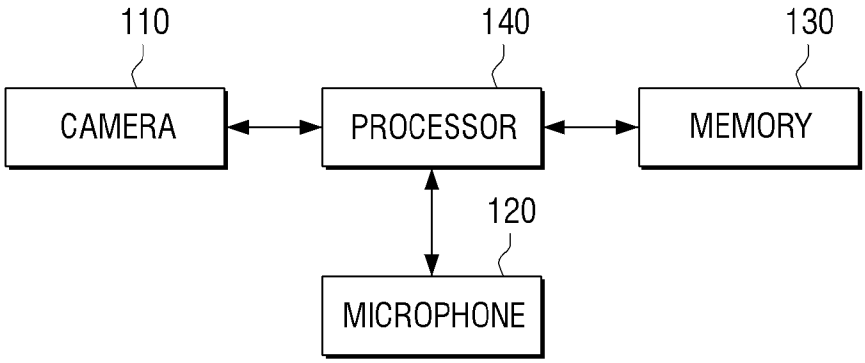
[Claim 14] The control method of claim 12,
wherein each of the plurality of the speech recognition models comprises a language model and an acoustic model, and
wherein the recognizing comprises:
identifying a language model and an acoustic model corresponding to the user information; and
obtaining a text data on the user speech by using the identified language model and the acoustic model.

[Claim 15] The control method of claim 14, comprising:
identifying a natural language understanding model of a plurality of natural language understanding models based on the user information; and
performing natural language understanding on the obtained text data through the identified natural language understanding model.

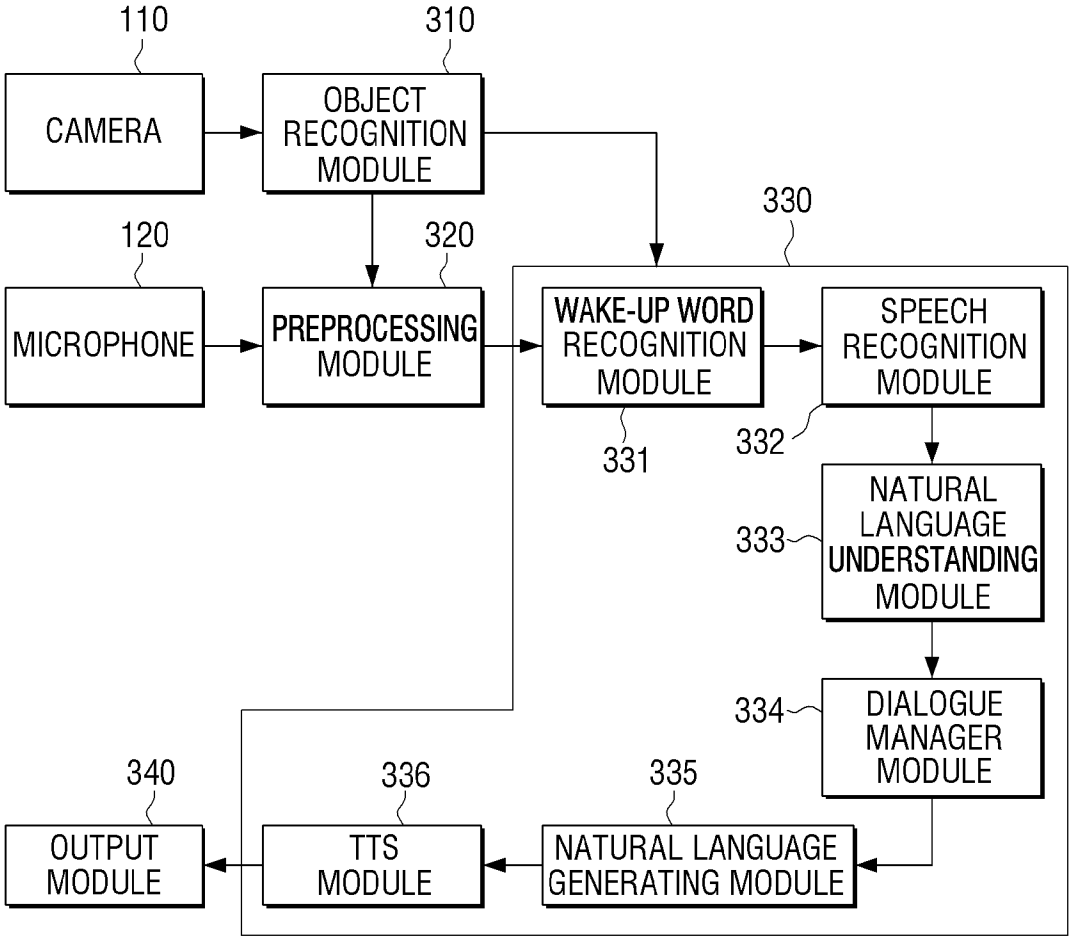
[Fig. 1]


10
HELLO, YOUNG CUSTOMER, PLEASE CALL ME RETAIL-BOT AND TELL ME HOW I MAY ASSIST YOU.
20
RETAIL-BOT
30
WHICH ARE THE MOST POPULAR SMART PHONES CURRENTLY BEING SOLD?
100
40
THE PRODUCT MOST POPULARLY SOLD AMONG KIDS-PHONES IS THE XX PHONE. LET ME SHOW YOU.
KIDS-PHONE
1.XX
2.AA

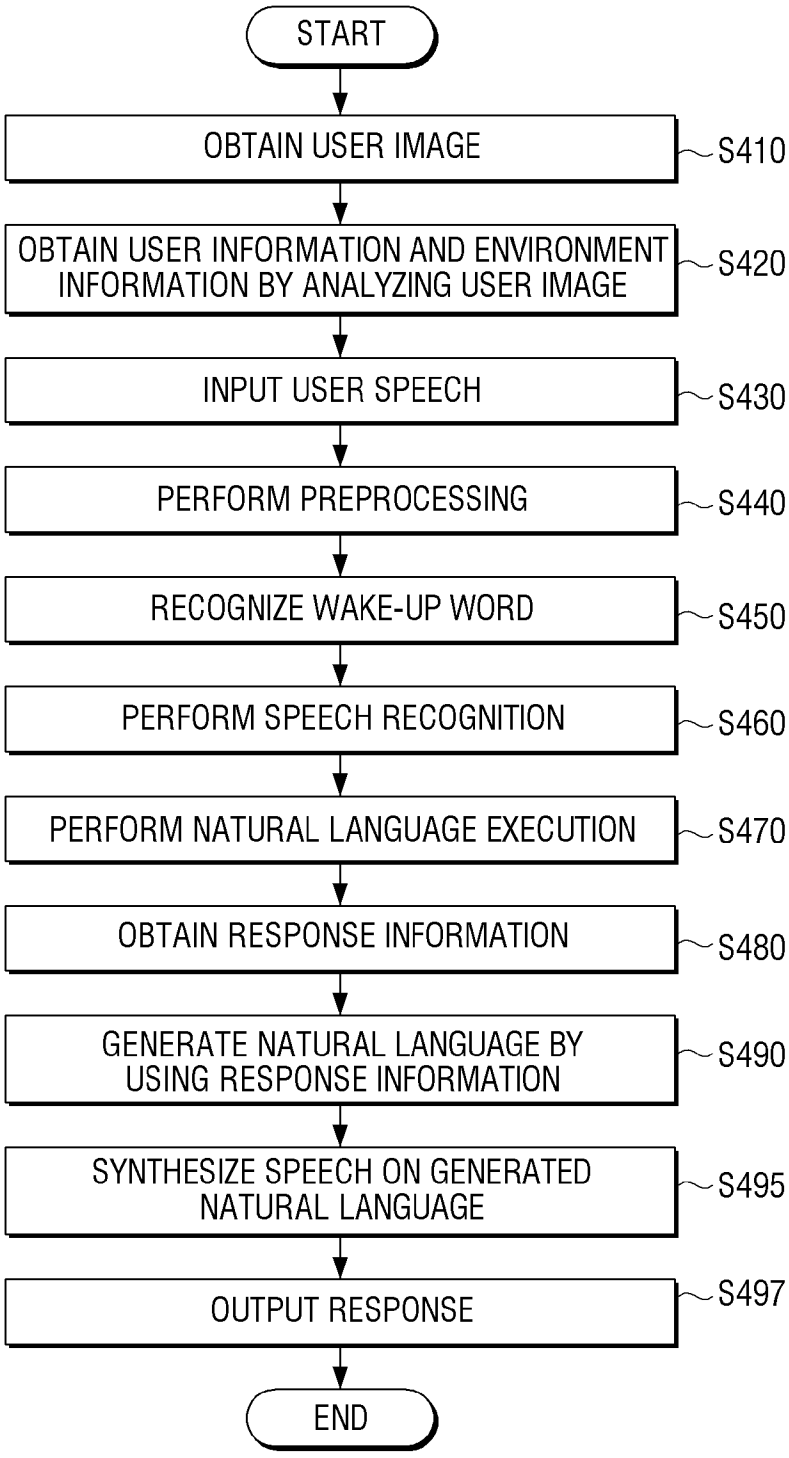
[Fig. 2]


110
CAMERA
140
PROCESSOR
130
MEMORY
120
MICROPHONE

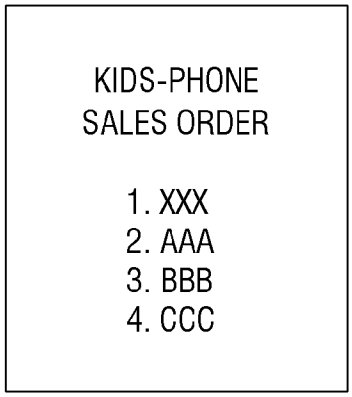
[Fig. 3]


110
CAMERA
310
OBJECT RECOGNITION MODULE
120
MICROPHONE
320
PREPROCESSING MODULE
330
WAKE-UP WORD RECOGNITION MODULE
331
SPEECH RECOGNITION MODULE
332
NATURAL LANGUAGE UNDERSTANDING MODULE
333
DIALOGUE MANAGER MODULE
334
NATURAL LANGUAGE GENERATING MODULE
335
TTS MODULE
336
OUTPUT MODULE
340

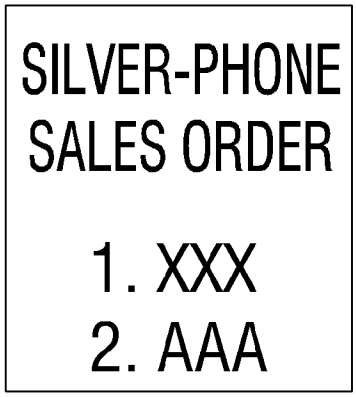
[Fig. 4]


START
OBTAIN USER IMAGE
S410
OBTAIN USER INFORMATION AND ENVIRONMENT INFORMATION BY ANALYZING USER IMAGE
S420
INPUT USER SPEECH
S430
PERFORM PREPROCESSING
S440
RECOGNIZE WAKE-UP WORD
S450
PERFORM SPEECH RECOGNITION
S460
PERFORM NATURAL LANGUAGE EXECUTION
S470
OBTAIN RESPONSE INFORMATION
S480
GENERATE NATURAL LANGUAGE BY USING RESPONSE INFORMATION
S490
SYNTHESIZE SPEECH ON GENERATED NATURAL LANGUAGE
S495
OUTPUT RESPONSE
S497
END

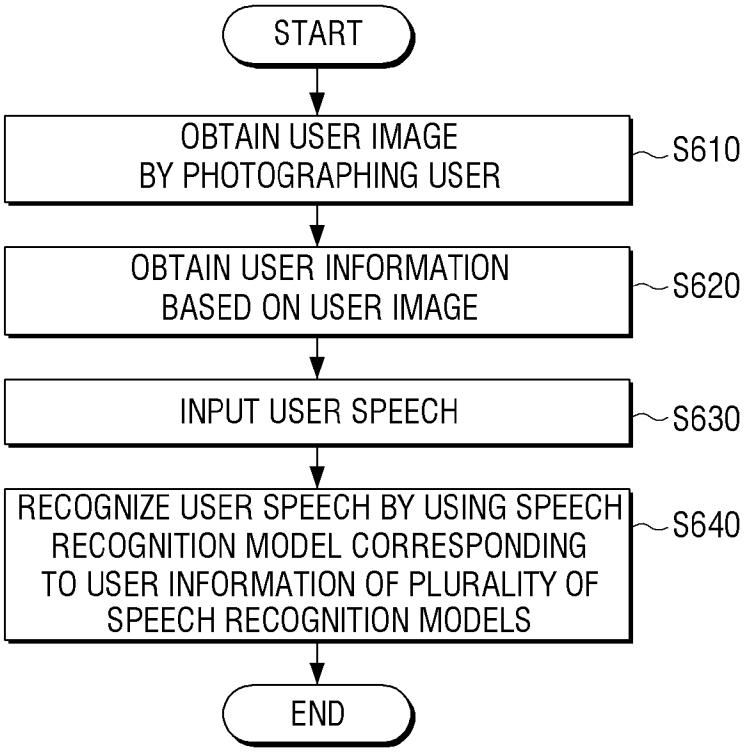
[Fig. 5A]


KIDS-PHONE
SALES ORDER
1. XXX
2. AAA
3. BBB
4. CCC

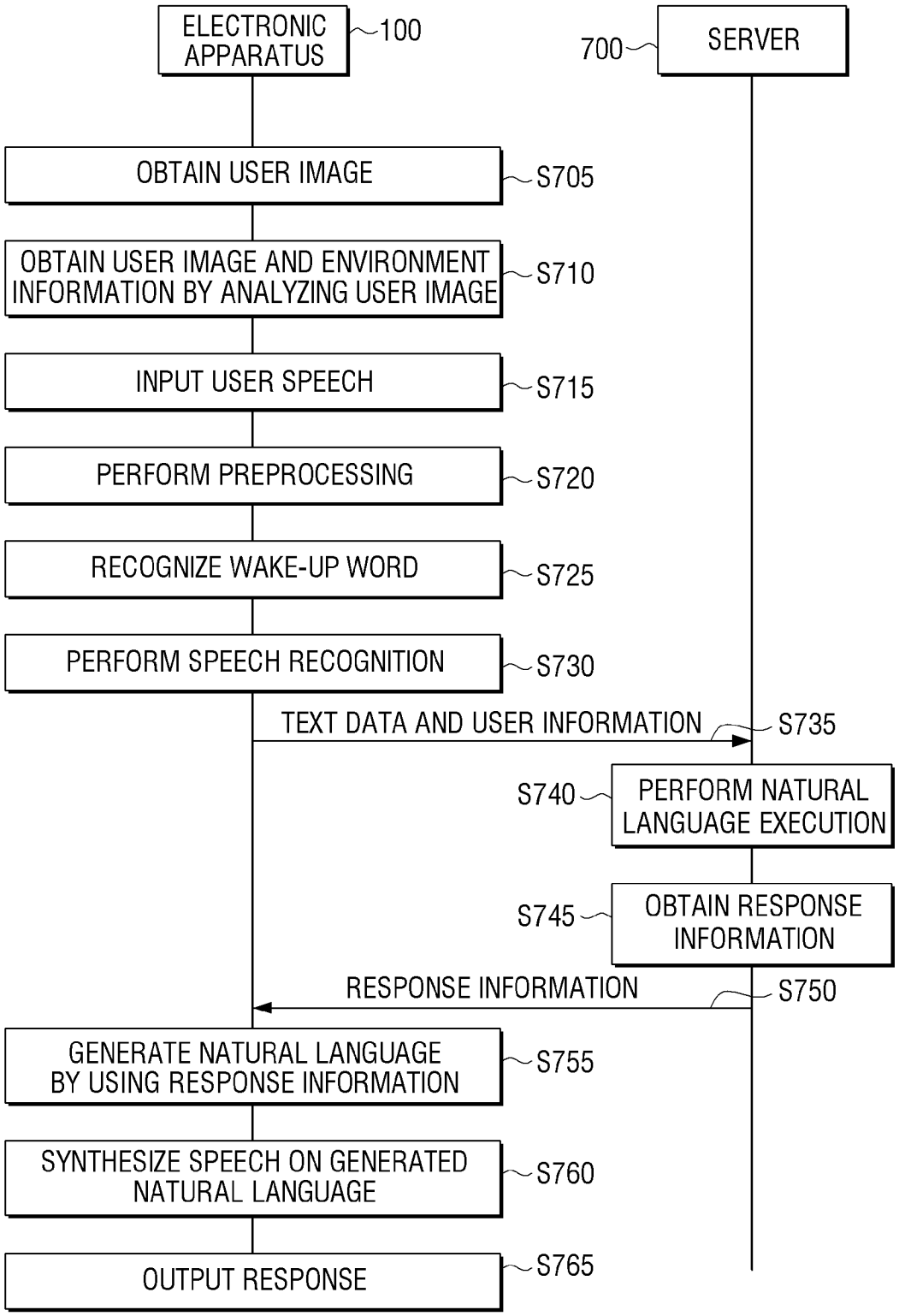
[Fig. 5B]


SILVER-PHONE
SALES ORDER
1. XXX
2. AAA

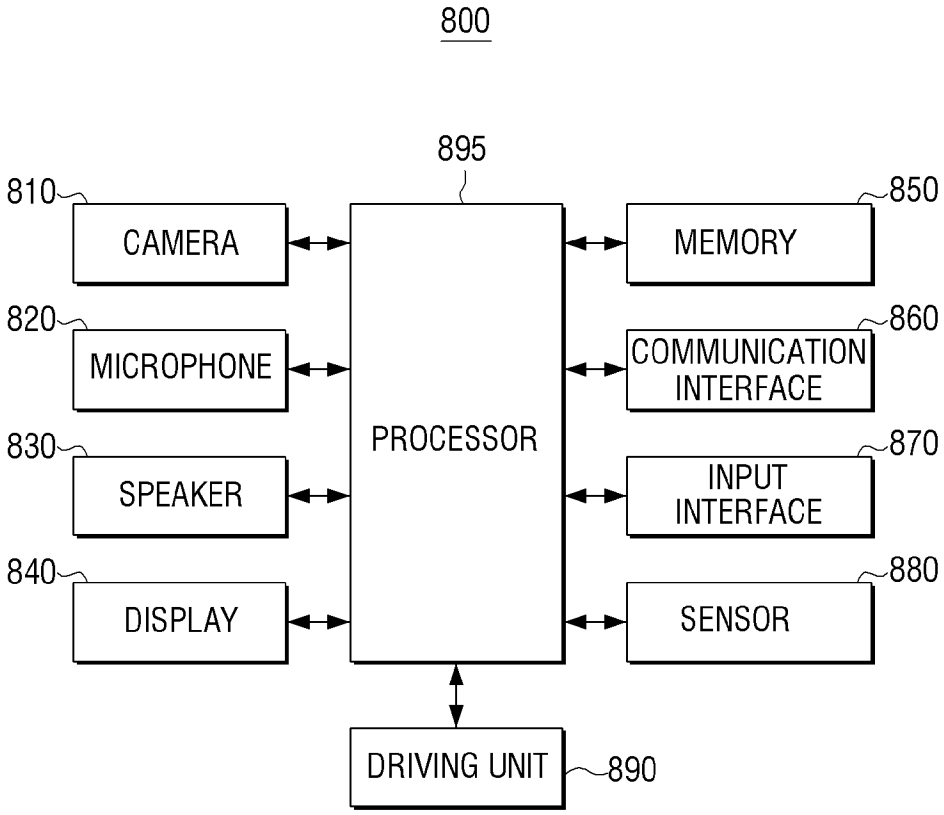
[Fig. 6]


START
OBTAIN USER IMAGE BY PHOTOGRAPHING USER
S610
OBTAIN USER INFORMATION BASED ON USER IMAGE
S620
INPUT USER SPEECH
S630
RECOGNIZE USER SPEECH BY USING SPEECH RECOGNITION MODEL CORRESPONDING TO USER INFORMATION OF PLURALITY OF SPEECH RECOGNITION MODELS
S640
END

[Fig. 7]


ELECTRONIC APPARATUS
100
700
SERVER
OBTAIN USER IMAGE
S705
OBTAIN USER IMAGE AND ENVIRONMENT INFORMATION BY ANALYZING USER IMAGE
S710
INPUT USER SPEECH
S715
PERFORM PREPROCESSING
S720
RECOGNIZE WAKE-UP WORD
S725
PERFORM SPEECH RECOGNITION
S730
TEXT DATA AND USER INFORMATION
S735
S740
PERFORM NATURAL LANGUAGE EXECUTION
S745
OBTAIN RESPONSE INFORMATION
RESPONSE INFORMATION
S750
GENERATE NATURAL LANGUAGE BY USING RESPONSE INFORMATION
S755
SYNTHESIZE SPEECH ON GENERATED NATURAL LANGUAGE
S760
OUTPUT RESPONSE
S765

[Fig. 8]


800
895
810
CAMERA
820
MICROPHONE
830
SPEAKER
840
DISPLAY
PROCESSOR
850
MEMORY
860
COMMUNICATION INTERFACE
870
INPUT INTERFACE
880
SENSOR
DRIVING UNIT
890

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2020/012198

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/22(2006.01)i, G10L 15/14(2006.01)i, G10L 15/183(2013.01)i, G06T 7/11(2017.01)i, G10L 21/0208(2013.01)i, G10L 15/26(2006.01)i, G10L 13/08(2006.01)i, G10L 17/24(2013.01)i, G06F 3/16(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L 15/22; G06F 16/33; G06F 16/332; G06F 17/30; G10L 15/02; G10L 17/14; G10L 17/24; H04L 12/58; G10L 15/14; G10L 15/183; G06T 7/11; G10L 21/0208; G10L 15/26; G10L 13/08; G06F 3/16

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: speech, recognition, image, AI, learning, instruction

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	CN 106503275 A (CAPITAL NORMAL UNIVERSITY) 15 March 2017 paragraphs [0003]-[0004], [0033]-[0039], [0073]-[0085]; and claims 1-4	1-2,4-5,8,11-12 ,14-15 3,6-7,9-10,13
Y	KR 10-2018-0084392 A (SAMSUNG ELECTRONICS CO., LTD.) 25 July 2018 paragraph [0093]; and claim 1	3,10,13
Y	KR 10-2019-0106011 A (KYUNGPOOK NAT. UNIV. IND. ACADEMIC COOP FOUND) 18 September 2019 paragraphs [0040]-[0071]	6-7,9
A	US 2018-0090140 A1 (MUNIR NIKOLAI ALEXANDER GEORGES et al.) 29 March 2018 claims 1-7; and figures 1-2	1-15
A	CN 109376225 A (GUANGZHOU PINGDAO INFORMATION TECHNOLOGY CO., LTD.) 22 February 2019 claims 1-6	1-15

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 December 2020 (14.12.2020)	**14 December 2020 (14.12.2020)**

Name and mailing address of the ISA/KR

International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

KWON, Sungho

Telephone No. +82-42-481-3547

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/KR2020/012198

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CN 106503275 A	15/03/2017	None	
KR 10-2018-0084392 A	25/07/2018	EP 3567584 A1	13/11/2019
		US 2020-0051554 A1	13/02/2020
		WO 2018-135753 A1	26/07/2018
KR 10-2019-0106011 A	18/09/2019	None	
US 2018-0090140 A1	29/03/2018	CN 109661704 A	19/04/2019
		EP 3520103 A1	07/08/2019
		US 10147423 B2	04/12/2018
		US 2019-0348036 A1	14/11/2019
		WO 2018-063619 A1	05/04/2018
CN 109376225 A	22/02/2019	None	