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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 January 2014 (30.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/017876 A1

(51) International Patent Classification:
G10L 15/00 (2013.01) *G06K 9/20* (2006.01)

(21) International Application Number:
PCT/KR2013/006749

(22) International Filing Date:
26 July 2013 (26.07.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10-2012-0081837 26 July 2012 (26.07.2012) KR

(71) Applicant: **SAMSUNG ELECTRONICS CO., LTD.** [KR/KR]; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 443-742 (KR).

(72) Inventors: **LEE, Dongyeol**; No. 804-1504, Heungdeok Maeul 8 Danji Hankuk Adelium Apt., Yeongdeok-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-908 (KR). **SUH, Sangbum**; No. 102-1009, Sinbanpo Chonggu Apt., Jamwon-dong, Seocho-gu, Seoul 137-030 (KR).

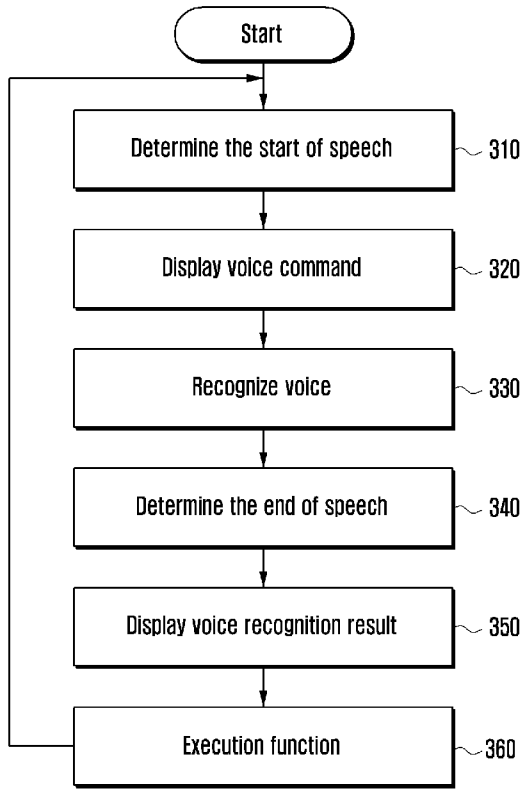
(74) Agent: **YOON, Dong Yol**; Yoon & Lee International Patent & Law Firm, 3rd Fl, Ace Highend Tower-5, 226, Gasan Digital 1-ro, Geumcheon-gu, Seoul 153-803 (KR).

(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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KZ, LA, LC, LK, LR, LS, LT, 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, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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, 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,

[Continued on next page]

(54) Title: VOICE RECOGNIZING METHOD AND APPARATUS USING VIDEO RECOGNITION



(57) Abstract: Provided are a method and an apparatus for performing exact start and end recognition of voice based on video recognition. The method includes determining whether a speech starts based on at least one of first video and audio data before conversion into a voice recognition mode, converting into the voice recognition mode and generating second audio data including a voice command, when it is determined that speech starts, and determining whether the speech is terminated based on at least one of second video and audio data after conversion into the voice recognition mode.

WO 2014/017876 A1

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))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

Description

Title of Invention: VOICE RECOGNIZING METHOD AND APPARATUS USING VIDEO RECOGNITION

Technical Field

[1] The present invention relates generally to a voice recognizing method and apparatus by using video recognition in a terminal, and more particularly, to a method and apparatus for detecting a speech start time and a speech end time through video recognition by a camera without a separate user gesture, and increasing accuracy of voice recognition.

Background Art

[2] Voice recognition technology, when substituted for physical input, aims to enable a user to conveniently use electronic devices without movement. For example, the voice recognition technology may be implemented in various electronic devices such as a smart phone, a television, and a vehicle navigation device.

[3] FIG. 1 illustrates a display screen of a terminal for recognizing voice according to the related art. The voice recognition technology in FIG. 1 requires a user to record start, speak, record end, and perform result computation by operating a specific program. The related art shown in FIG. 1 is implemented by a pre-defined key word or a structure for general free voice recognition, rather than a technology of processing a command according to a current state of a device.

Disclosure of Invention

Technical Problem

[4] The voice recognition technology statistically analyzes and classifies input voice. It is important to minimize noise or silent section of recorded data for exact voice recognition. However, when considering various situations of a user recording voice, a noise other than a speaker's voice is likely to be included in voice recording data, and it is difficult to exactly recognize a voice speaking state.

[5] The user needs a separate operation to start voice recording. For example, when the user is driving a vehicle or carries a burden may be considered. Voice recognition is a very valuable function in such instances, since a terminal function may be executed by using only a user's voice without separate key or gesture input. Accordingly, there is a need in the art for a voice recognition technology that does not require separate gesture user input from the start of voice recording.

Solution to Problem

[6] The present invention has been made in view of the above problems, and provides a method capable of executing a voice recognition terminal function without a finger

motion by the user, and an apparatus thereof.

[7] The present invention further provides a method of recognizing a speech start time and a speech end time through video recognition and exactly and rapidly recognizing voice there-through, and an apparatus thereof.

Advantageous Effects of Invention

[8] In accordance with an aspect of the present invention, a method of recognizing a voice command by an electronic device includes determining whether a speech starts based on at least one of first video data and first audio data before conversion into a voice recognition mode, converting the electronic device into the voice recognition mode and generating second audio data including a voice command of a user, when it is determined that speech starts, and determining whether the speech is terminated based on at least one of second video data and second audio data after conversion into the voice recognition mode.

[9] In accordance with another aspect of the present invention, an electronic device for recognizing a voice command includes an audio processor which collects and records a voice input, a camera unit which collects and records a video input, and a controller which controls to generate first video data or first audio data before conversion into a voice recognition mode, determine whether a speech starts based on at least one of the first video data and the first audio data, convert into the voice recognition mode when it is determined that the speech starts, control to generate second audio data or second video data including a voice command of a user, and determine whether the speech is terminated based on at least one of the second video data and the second audio data.

[10] In accordance with another aspect of the present invention, a platform for recognizing a voice command includes a multi-media framework which collects and records voice input or video input to generate first video data or first audio data before conversion into a voice recognition mode, and generate second video data or second audio data after conversion into the voice recognition mode, and a voice framework which determines whether a speech starts based on at least one of the first video data or the first audio data, and determines whether the speech is terminated based on at least one of the second video data or the second audio data.

Brief Description of Drawings

[11] The objects, features and advantages of the present invention will be more apparent from the following detailed description in conjunction with the accompanying drawings, in which:

[12] FIG. 1 illustrates a display screen of a terminal for recognizing voice according to the related art;

[13] FIG. 2 illustrates a configuration of a terminal according to an embodiment of the

present invention;

[14] FIG. 3 illustrates a method of recognizing voice according to an embodiment of the present invention;

[15] FIG. 4 illustrates an example of using voice recognition through a TV according to an embodiment of the present invention;

[16] FIG. 5 illustrates a configuration of an apparatus for recognizing voice according to an embodiment of the present invention;

[17] FIG. 6 illustrates a detail method of using voice recognition using a speech video according to an embodiment of the present invention;

[18] FIG. 7 illustrates a detail method of using voice according to an embodiment of the present invention when the terminal is controlled according to a voice command;

[19] FIG. 8 illustrates an example of a concrete graphic interface displaying the voice command during the process of voice recognition according to an embodiment of the present invention;

[20] FIG. 9 illustrates an example of a widget for the voice command of an application and a source code capable of registering the voice command in the widget;

[21] FIG. 10 illustrates a sequence configuring a screen displaying the voice command by determining a mouth shape of a user;

[22] FIG. 11 illustrates an operation of a platform which checks a speech time point of the user; and

[23] FIG. 12 illustrates an operation of a platform which checks a speech end time point of the user.

Mode for the Invention

[24] Embodiments of the present invention are described with reference to the accompanying drawings in detail. The same reference numerals are used throughout the drawings to refer to the same or like parts. Detailed descriptions of well-known functions and structures incorporated herein may be omitted to avoid obscuring the subject matter of the present invention.

[25] The terminal 200 of the present invention includes all electronic devices supporting a voice recognition function.

[26] For example, the terminal 200 includes a portable phone, a Portable Multimedia Player (PMP), a digital broadcasting player, a vehicle navigation device, a Personal Digital Assistant (PDA), a music player (e.g., MP3 player, a portable game terminal, a tablet Personal Computer (PC), and a smart phone. The terminal 200 includes various devices supporting a voice recognition function including home appliances such as a television, a refrigerator, and a washing machine installed at a fixed location.

[27] FIG. 2 illustrates a configuration of a terminal 200 according to an embodiment of

the present invention.

[28] Referring to FIG. 2, the terminal 200 of the present invention includes a Radio Frequency (RF) communication unit 210, a key input unit 220, a camera unit 230, an audio processor 240, a touch screen 250, a memory 260, and a controller 270.

[29] The RF communication unit 210 forms a communication channel for each of voice call, video call, and transmitting data such as a video or a message under control of the controller 270 (hereinafter, data communication channel).

[30] The key input unit 220 includes a plurality of input keys and function keys for receiving numeric or character information, and setting various functions. The function keys include arrow keys, side keys, and hot keys set for executing a specific function. The key input unit 220 generates a key signal associated with user setting and function control of the terminal 200, and sends the generated key signal to the controller 270. When the touch screen 250 of the terminal 200 is supported in the form of a full touch screen, the key input unit 220 includes side keys, home keys, and other function keys that are provided at a side of a case of the terminal 100.

[31] Particularly, the key input unit 220 of the present invention includes a voice recognition function key set for executing a voice recognition function. The key input unit 220 sends a voice recognition function key event generated from the voice recognition function key to the controller 270, which determines start and end of a voice recognition mode according to a request signal of the voice recognition function key.

[32] The camera unit 230 provides collected video by shooting a subject, and may be activated according to a signal generated from the touch screen 250 or the key input unit 220 to collect the videos. The camera unit 230 includes a camera sensor which converts an optical signal into an electric signal, an image signal processor which converts an analog video signal into a digital video signal, and a Digital Signal Processor (DSP) which processes a video signal (scaling, noise removal, Red, Green, Blue (RGB) signal conversion) output from the image signal processor in order to display the video signal on the touch screen 250. The camera sensor includes a Charge-coupled Device (CCD) sensor or a Complementary Metal-Oxide Semiconductor (CMOS) sensor. The DSP configuration may be omitted and may be implemented by a DSP chip.

[33] The camera unit 230 drives a camera in an idle mode of the terminal 200 to record a user image under control of the controller 270, sends the recoded video to the controller 270 to provide image recognition data capable of recognizing a user's face, and may be activated in the idle mode of the terminal 200 according to user setting, or according to separate user input.

[34] The audio processor 240 includes a speaker (SPK) for playing transceived audio data

during a call, audio data included in a received message, and audio data according to playback of an audio file stored in the memory 260, and a microphone (MIC) for collecting a user's voice or other audio signals during the call.

[35] Particularly, the audio processor 240 drives the MIC in a voice recognition mode to record the user's voice collected through the MIC under control of the controller 270. The audio processor 240 sends the recorded voice to the controller 270 so that the recognition for the recorded voice may be performed. If the voice recognition mode starts or is terminated, the audio processor 240 may or may not output a corresponding effect sound, depending on a user setting.

[36] The touch screen 250 includes a touch panel 253 and a display unit 256, and the touch panel 253 may be disposed at a front surface of the display unit 256. The size of the touch screen 250 may be determined by the size of the touch panel 253. The touch screen 250 may display a screen according to execution of a user function and detect a touch event related with control of the user function.

[37] The touch panel 253 is disposed in at least one of upper and lower portions of the display unit 256, and a sensor constituting the touch panel 253 is arranged in a matrix pattern. Accordingly, the touch panel 253 may generate a touch event according to a contact or approach distance of a touched object on the touch panel 253, and send the generated touch event to the controller 270. The touch event includes a touched type and location information.

[38] The display unit 256 displays information input by the user or information provided to the user as well as various menus. That is, the display unit 256 may provide en execution screen of various user function according to use of the terminal 200. The display unit 256 may be configured as a Liquid Crystal Display (LCD) or an Organic Light Emitted Diode (OLED). The display unit 256 may be disposed at an upper or lower portion of the touch panel 253.

[39] When the controller 270 determines that a current state of a user is speech start, the display unit 256 displays a voice command in the pop-up form under control of the controller 270. When the controller 270 succeeds in the voice recognition, the display unit 256 displays the recognized voice command in the popup form.

[40] The memory 260 stores at least one application necessary for a function operation, user data generated by the user, a message transceived with a network, and data according to execution of the application. The memory 160 includes, for example, a program area and a data area (not shown). The program area stores an Operating System (OS) for booting the terminal 200 and for operating the foregoing constituent elements, and downloaded and installed applications. Particularly, the program area of the present invention further stores a video recognition-operating program and a voice recognition operation program.

[41] The video recognition-operating program analyzes collected video data to determine a speech start time point and a speech end time point of the user. The video recognition-operating program is driven in an idle state of the terminal 200, and according to separate user input.

[42] The voice recognition-operating program supports the terminal 200 such that a function of the terminal 200 is executed using voice recognition. In particular, the voice recognition-operating program according to the present invention is used to determine the speech start time point and the speech end time point of the user in co-operation with the video recognition-operating program. The voice recognition-operating program includes a routine of executing a corresponding function based on the voice recognition result when it is determined that one of preset voice commands is input.

[43] The data area stores data generated according to use of the terminal 200.

[44] The data area stores data to be used or generated during execution of the data video recognition-operating program and the voice recognition-operating program. In addition, the data area may store various statistic models and a voice recognition result for voice recognition in connection with the voice recognition-operating program.

[45] The controller 270 controls an overall operation of the terminal 200. Particularly, the controller 270 of the present invention determines the speech start time point and the speech end time point, determines a voice command input by the user, and controls a series of procedures for executing a function of the terminal 200 connected to the voice command.

[46] FIG. 3 illustrates a method of recognizing voice according to an embodiment of the present invention.

[47] The method of FIG. 3 starts from an idle mode of a terminal in which only the camera unit 230 is activated. The method of recognizing voice according to the embodiment of the present invention may also start from an active mode of the terminal. In this instance, the display unit 256 displays a user-function execution screen that is configured with widgets including at least one image, such as an icon, a thumbnail or a character. Each of the image constituent elements is associated with a specific function. Accordingly, if a voice command is recognized, the controller 270 may control such that a function associated with the specific image constituent element is executed.

[48] Steps 310 and 320 are for determining the start of speech to determine whether to convert to a voice recognition mode.

[49] A camera unit 230 collects video data, and an audio processor 240 collects audio data through recording (310). The controller 270 simultaneously analyzs the video data and the audio data to determine the start of speech. The video data may be collected by

driving only the camera unit 230, without driving the audio processor 240. When it is determined that a mouth shape of a user recorded in the video data is open, the controller 270 determines that the open mouth shape of the user is the speech start for starting a voice recognition function.

[50] In detail, a method of checking a speech start time point according to the embodiment of the present invention includes a first method of determining only recorded video data. That is, the first method analyzes a recorded video to find a mouth shape of the user, and analyzes the mouth shape of the user to find a speech time point by the user.

[51] A second method is simultaneously analyzing a recorded video and a recorded voice to find a speech time point.

[52] According to the second method, instances when a human user does not speak and just moves his/her mouth or when a sound is barely audible may be considered. The second method is advantageous in that it recognizes a mouth shape and magnitude of the recorded voice and determines whether the sound is a human voice to exactly recognize the speech time point.

[53] In the second method, when it is determined that both of results of analyzing the recorded video and the recorded voice are speech start, the controller 270 determines that the speech starts.

[54] However, when it is determined that only the result of analyzing the video is the speech start, the controller 270 attenuates a voice recording data determination threshold and re-determines speech start. When the result of the re-determination is valid, the controller 270 determines that a current state of the user is the speech start. When the result of the re-determination is invalid, the controller 270 determines that the current state of the user is not the speech start.

[55] When it is determined that only the result of analyzing the voice is the speech start, for example, when a user's face does not appear in video recognition, the controller 270 determines that a current state of the user is not the speech start. However, if the face is recognized, the controller 270 attenuates a determination threshold of motion of the mouth shape and re-determines. When the result of the re-determination is valid, the controller 270 determines that the current state of the user is the speech start. When the result of the re-determination is invalid, the controller 270 determines that the current state of the user is not the speech start.

[56] When the controller 270 determines that the user starts speech for voice recognition, the controller 270 controls the display unit 256 to display a voice command connected to currently displayed widgets in step 320. When the audio processor 240 is not driven in step 310, the controller 270 drives the audio processor 240 to convert a mode to a voice recognition mode in step 330.

[57] According to the present invention, when the terminal 200 converts the mode to a voice recognition mode, the controller 170 may display voice commands associated with a current execution screen on the display unit 256, since the user may view a command to be executed on a screen when the command is a voice command for controlling a device instead of a general dictation function. For example, in order to execute a function touching an icon of a screen through voice, the controller 170 displays an executable command to the user and the user may speak a command to be executed.

[58] For example, when the user opens his mouth, the controller 270 determines that the speech starts in step 310, a voice command is displayed on a popup window, and recording begins through a recorder in step 320. However, if the user shuts his mouth before conversion to the voice recognition mode, the command is not displayed and recording by the recorder ends.

[59] In the foregoing example, the controller 270 determines whether a voice included in data recorded by the recorder is valid. When the voice is valid, voice recognition for the recorded data starts. If speech terminal is checked in steps 340 and 350 of checking a speech end time point, the recording is stopped, and a silent section of actually recorded data is analyzed to integrally determine a voice recognition result.

[60] The controller 270 may control the display unit 256 to display a voice command of 'calendar' within a set distance in a zone in which a specific icon, such as a calendar icon, is displayed in the home screen in step 320. The controller 270 may output a voice command of 'Help' to any part of a blank of an execution screen. This function is for minimizing an amount of processing data for voice recognition and provides rapid and exact voice recognition service by limiting a voice recognition function to specific voice commands.

[61] Particularly, when determining a display location of a voice command, if there is an image constituent element associated with a function connected to the voice command, the controller 270 selects a display location of the voice command as a periphery of the image constituent element, and determines whether the image constituent element exists with reference to a map of an execution screen. In addition, the controller 270 selects the display location of the voice command with reference to the map of the execution screen.

[62] For example, the controller 270 may display the voice command at a location in which an icon connected to a video play function is displayed, in a video execution screen. That is, the controller 270 may display a pause voice command around a pause icon.

[63] In addition, the controller 270 may output the voice command in a preset display scheme. For example, the controller 270 may display the voice command to overlay an

execution screen in the form of tool-tip or speech balloon. When there is a plurality of voice commands such that a plurality of functions is connected with one image constituent element, the controller 270 may display voice command in list form. The controller 270 may control the display unit 256 to display integral voice commands on one pop-up window or a separate screen. A detailed user interface displaying the voice commands will be described with reference to FIG. 8.

[64] The controller 270 recognizes a voice through the audio processor 240 in step 330.

[65] A user's face may be deviated from a video recording area during the voice recognition. When the user is deviated from the video recording area during recording, the controller 270 may check a speech end section by only voice analysis, and report this instance to the user through the display unit 256 or a speaker of the audio processor 240.

[66] The user may be deviated from a recording area before speech start. Since it may be assumed that the user does not use the device, the controller 270 does not start voice recognition until the user's face enters the recording area. The controller 270 may report this instance to the user through the display unit 256 or a speaker of the audio processor 240.

[67] Typically, the controller 170 classifies phonemes from recorded voice data and recognizes a word (or word sequence) configured by the phonemes. Particularly, the controller 270 may perform voice recognition based on voice commands previously displayed on an execution screen according to step 320.

[68] For example, the controller 270 may recognize the voice for each phoneme or word, and compare the recognized voice with phonemes or words of the voice commands displayed on the display unit 256 to shorten a voice recognition time/ In this case, the terminal 100 performs the voice recognition instead of the voice recognition being consigned to an external server.

[69] However, the present invention is not limited thereto, and may be implemented according to the related art supporting intelligent voice recognition. For example, when receiving a voice command to check a current day schedule from the user, the controller 270 may analyze a natural language of the user to determine a corresponding command in step 330.

[70] If it is determined that audio data including the voice command is deviated from noise, error, or recognition range, the controller 270 reports voice recognition end to the user, and may return to step 310. This includes a case of previously displaying the voice command on the execution screen.

[71] In steps 340 and 350, the controller 270 checks the end of speech to determine whether to terminate the voice recognition mode.

[72] The camera unit 230 may collect video data, and the audio processor 240 may collect

audio data through recording to determine the end of speech in step 340. The controller 270 determines the end of speech by simultaneously analyzing the video data and the audio data. Alternatively, the audio processor 240 may only be driven to collect audio data without driving the camera unit 230. When the user's voice is not input for at least a preset time in the audio data, the controller 270 determines a current state of the user as speech end for terminating the voice recognition function.

[73] A method of checking a speech end time point according to the embodiment of the present invention may be depend on the method of checking the speech end time point using both of the recorded video and the recorded voice. When both of the results of analyzing the recorded video and the recorded voice are the speech end, the controller 270 determines this situation as the speech end.

[74] However, when only the result of analyzing the video is determined to be the speech end, the controller 270 attenuates a voice recording data determination threshold and re-determines speech end. When it is determined that the result of re-determination the voice is valid, the controller 270 determines this situation as the speech end. When it is determined that the result of re-determination the voice is invalid, the controller 270 does not determine this situation as the speech end.

[75] When only the result of analyzing the voice is determined to be the speech end, for example, when a user's face does not appear in video recognition, the controller 270 determines this situation as the speech end. However, if the face is recognized, the controller 270 attenuates a determination threshold of motion of the mouth shape and re-determines. When it is determined that the result of analyzing the voice is valid, the controller 270 determines that speech is terminated. When it is determined that the result of analyzing the voice is not valid, the controller 270 determines that the speech is not terminated.

[76] When the controller 270 determines that the user terminates speech for voice recognition in step 340, the controller 270 controls the display unit 256 to display a voice recognition result in step 350.

[77] For example, when receiving a voice command to search a neighboring famous restaurant in step 330, the controller 270 may display a voice recognition result such as <searching a neighboring famous restaurant> through the display unit 256. Thereafter, the controller 270 executes a function for a recognized voice command and terminates the voice recognition function instep 360. The controller 270 may return to step 310 and perform another voice recognition procedure.

[78] In order to improve speed and memory, after determining the speech start in step 320, the controller 270 may analyze the end of the speech by only the voice without performing analysis through the video. When returning to step 310 after step 360, analysis by the video can begin.

[79] FIG. 4 illustrates an example of using voice recognition through a TV according to an embodiment of the present invention.

[80] When a device is controlled through a voice, it is difficult to set a range of voice recognition. However, in the present invention, the user may set the terminal to recognize one ACTION at a time, and easily and continuously apply a voice command.

[81] For example, as shown in FIG. 4, when using a smart TV, the user may command using a voice instead of button input of a general remote controller to execute a specific function of the smart TV. The user may fetch an application operating in a background to a foreground or easily adjust a setting menu option.

[82] FIG. 5 illustrates a configuration of an apparatus for recognizing voice according to an embodiment of the present invention. As shown in FIG. 5, a portable terminal including a camera and a microphone may implement a voice recognition function using video recognition. As shown in FIG. 5, the apparatus for recognizing a voice includes the camera unit 230 and a microphone included in the audio processor 240. The controller 270 includes a constituent element for processing and analyzing a video recorded by a camera and a voice recorded by the microphone.

[83] FIG. 6 illustrates a detail method of using voice recognition using a speech video according to an embodiment of the present invention. Since the implementation method of the present invention was described in reference to FIG. 3, FIG. 6 will illustrate a representative embodiment of the present invention

[84] The controller 270 analyzes a video from the camera 230 in step 610 and, in detail, analyzes a mouth shape of a user to find a speech time point in step 620. If it is determined that the user speaks, the controller 270 operates a recorder 240 to start recording in step 630, and performs voice recognition using recorded data in step 640.

[85] The controller 270 may analyze a video from a camera and recorded voice data to determine a speech end time point in step 650. If it is determined that speech is terminated, the controller 270 stops the recording and performs voice recognition in step 660.

[86] When the controller 270 succeeds in the detection of the voice recognition result in step 670, the controller 270 performs an operation corresponding to the voice recognition result in step 680. If the detection of the voice recognition result is unsuccessful, or succeeds in the detection of the voice recognition result and a corresponding operation is completed, the controller 270 returns to a first step the result of analyzing the voice and places the speech start of the user on standby.

[87] FIG. 7 illustrates a detailed method of using voice according to an embodiment of the present invention when the terminal is controlled through a voice command. Since the implementation method of the present invention was described in reference to FIG. 3, FIG. 7 will illustrate a representative embodiment of the present invention.

[88] The controller 270 analyzes a video from the camera 230 in step 710 and, in detail, analyzes a mouth shape of a user to find a speech time point in step 720. When it is determined that the user opens a mouth before speech start, the controller 270 may display a voice command on a display unit 256. Concurrently, the controller 270 operates a recorder 240 to start recording in step 730, and performs voice recognition using recorded data in step 740.

[89] The controller 270 may analyze recorded voice data and video data from camera to determine a speech end time point in step 750. If it is determined that speech is terminated, the controller 270 stops recording and terminates display of the voice command by the display unit 256, and performs voice recognition in step 760.

[90] When the controller 270 succeeds in the detection of the voice recognition result in step 770, the controller 270 performs an operation corresponding to the voice recognition result in step 780. When the detection of the voice recognition result is unsuccessful, or succeeds in the detection of the voice recognition result and a corresponding operation is completed, the controller 270 returns to a first step the result of analyzing the voice and places speech start of the user on standby.

[91] FIG. 8 illustrates an example of a concrete graphic interface displaying the voice command during the process of voice recognition according to an embodiment of the present invention.

[92] As shown in FIG. 8, voice recognition may be performed when a general screen displaying an icon for various applications is displayed. When the controller 270 determines the state of a user as speech start, the controller 270, as shown in FIG. 8, may display a voice command for executing an application on the display unit 256 together with an icon.

[93] FIG. 9 illustrates an example of a widget for the voice command of an application and a source code capable of registering the voice command in the widget.

[94] As shown in FIG. 9, a plurality of widgets displayed on the display unit 256 is respectively connected to an application. A voice such as <CREATE> may be registered as a voice command for executing an application connected to a <CREATE NEW MESSAGE> widget.

[95] FIG. 10 illustrates a sequence configuring a screen displaying the voice command by determining a mouth shape of a user.

[96] A record module starts recording through a camera 230 at numeral 1. At numerals 2 and 3, the camera unit 230 sends a recorded video to a face recognition engine, and the face recognition engine analyzes the video. At numeral 4, if the face recognition engine determines that the mouth shape of the user is open, the face recognition engine requests a voice command configuration to a user interface framework.

[97] At numeral 5, the user interface framework may collect and configure commands of

widgets of an application. Then, the user interface framework sends a command collected from numeral 6 to a voice framework and uses the command as a candidate group of voice recognition.

[98] At numeral 7, the voice framework sends corresponding contents to a voice recognition engine to prepare the start of the voice recognition. At numeral 8, if the voice framework checks that a configured command having no problem, the user interface framework displays the configured command through a display unit so that the user can have recognition performed.

[99] FIG. 11 illustrates an operation of a platform that checks a speech time point of the user.

[100] A record module starts recording of video and voice through a camera or a microphone at numeral 1. However, when a speech time point is determined by only a recording video, a recording function is not driven.

[101] At numeral 2, the recorder module sends recording data of video and voice to a voice framework, and, at numeral 3, the recording data of video is sent to a face recognition engine and the recording data of voice is sent to a voice recognition engine so that the start of speech is determined. That is, at numeral 4, the controller 270 determines the start of speech based on the determination of each engine.

[102] When it is determined that the speech starts, at numeral 5, the controller 270 reports voice recognition start to a voice recognition engine and continuously sends recorded data to the voice recognition engine. Concurrently, at numeral 6, the controller 270 may request the user interface framework to remove a command displayed on a current screen, and at numeral 7, the user interface framework removes a command displayed on a screen.

[103] FIG. 12 illustrates an operation of a platform which checks a speech end time point of the user.

[104] In numeral 1, a speech end checking time point is when a recorder module performs recording of video and voice through the camera and the microphone. However, when a speech end time point is determined by only recorded voice data, a recording function through the camera may not be driven.

[105] At numeral 2, the recorded data of video and voice are sent to the voice framework, and, at numeral 3, the recorded data of video and voice are sent to the face recognition engine and the voice recognition engine so that the recorded data is used to determine the end of speech.

[106] At numeral 4, the controller 270 determines the end of speech based on determination of each engine. If it is determined that the speech ends at numeral 5, the controller 270 controls the voice recognition engine to stop the recording of voice, and determines a final voice recognition result from the recorded data of voice. That is, at

numeral 6, the controller 270 receives the final recognition result through the voice recognition engine.

[107] At numeral 7, the controller 270 detects ACTION corresponding to the final voice recognition result to request execution of the ACTION to a corresponding module.

[108] The present invention may be implemented when there is more than one user. For example, the present invention includes a display module such as a camera module, a recorder module, a projector and a monitor, and a video conferencing system having a voice recognition module and a video recognition module.

[109] The video conferencing system enables a plurality of users to remotely participate in a video conference and simultaneously displays faces of a plurality of users on a monitor. When at least one or more speakers exist among a plurality of users during the video conference, speech start and end time points of a plurality of participators in the video conference may be clearly determined.

[110] For example, if user A opens his mouth while users A, B, C, and D progress a video conference, the present invention determines that the user A starts to speak. The present invention may simultaneously analyze a recorded video and a recorded voice to determine that speech of the user A is terminated and speech of the user B starts.

[111] When considering the speciality of the video conference, according to another embodiment of the present invention, if the user C lifts his hand, the recorded video is analyzed to determine that speech of the user C starts. If the user C lowers his hand, it may be determined that the speech of the user C is terminated.

[112] Based on such a determination, in the video conferencing system, a video conference user interface may be designed in such a manner that the face of the speaker is fully displayed on a monitor. Further, only the speaker may use the voice recognition function for controlling the video conferencing system.

[113] The user can drive the voice recognition function without using hands, and can continuously perform voice recognition.

[114] A speech time point of the user can be exactly detected to minimize an amount of analyzed recording data of voice so that accuracy and speed of the voice recognition can be improved.

[115] It is clear that the present invention can be realized by hardware, software (i.e., a program), or a combination thereof. This program can be stored in a volatile or non-volatile recording medium readable by a machine such as a computer. This medium can be a storage device such as a Read-Only Memory (ROM), a memory such as a Random-Access Memory (RAM), a memory chip, or an integrated circuit, or an optical or magnetic recording medium such as a Compact Disk (CD), a Digital Versatile Disk (DVD), a magnetic disk, or a magnetic tape.

[116] Although embodiments of the present invention have been described in detail

hereinabove, it should be clearly understood that many variations and modifications of the basic inventive concepts herein taught which may appear to those skilled in the present art will still fall within the spirit and scope of the present invention, as defined in the appended claims.

Claims

[Claim 1] A method of recognizing a voice command by an electronic device, the method comprising:
determining whether a speech starts based on at least one of first video data and first audio data before conversion into a voice recognition mode;
converting the electronic device into the voice recognition mode and generating second audio data including a voice command of a user, when it is determined that speech starts; and
determining whether the speech is terminated based on at least one of second video data and second audio data after conversion into the voice recognition mode.

[Claim 2] The method of claim 1, further comprising displaying a voice command for executing an icon displayed on a screen after determining whether the speech starts, and removing display of the voice command after determining whether the speech is terminated.

[Claim 3] The method of claim 2, wherein determining whether the speech starts comprises identifying a lip of the user from the first video data and determining whether the speech starts based on motion of the lip.

[Claim 4] The method of claim 3, further comprising attenuating a threshold for determining the speech start to re-determine whether the speech starts when at least one of result values of analyzing the first video data and the first audio data is less than or equal to the threshold, after determining whether the speech starts.

[Claim 5] The method of claim 4, wherein determining whether the speech starts comprises determining that the speech does not start when a face of the user is not included in the first video data.

[Claim 6] The method of claim 5, wherein determining whether the speech is terminated comprises determining that the speech is terminated when a voice of the user is not included in the second audio data for at least a preset time.

[Claim 7] The method of claim 6, further comprising attenuating a threshold for determining the speech end to re-determine whether the speech is terminated when at least one of result values of analyzing the second video data and the second audio data is less than or equal to the threshold, after determining whether the speech is terminated.

[Claim 8] The method of claim 4, wherein determining whether the speech is

terminated comprises determining that the speech is terminated when a face of the user is not included in the second video data.

[Claim 9] An electronic device which recognizes a voice command, the electronic device comprising:
an audio processor which collects and records a voice input;
a camera unit which collects and records a video input; and
a controller which controls to generate first video data or first audio data before conversion into a voice recognition mode, determine whether a speech starts based on at least one of the first video data and the first audio data, convert into the voice recognition mode when it is determined that the speech starts, control to generate second audio data or second video data including a voice command of a user, and determine whether the speech is terminated based on at least one of the second video data and the second audio data.

[Claim 10] The electronic device of claim 9, further comprising a display unit which displays a screen,
wherein the controller controls the display unit to display a voice command for executing an icon displayed on the screen when the speech start, and controls the display unit to remove display of the voice command when the speech is terminated.

[Claim 11] The electronic device of claim 10, wherein the controller identifies a lip of a user from the first video data, and determines whether the speech starts based on motion of the lip.

[Claim 12] The electronic device of claim 11, wherein the controller attenuates a threshold for determining the speech start to re-determine whether the speech starts when at least one of result values of analyzing the first video data and the first audio data is less than or equal to the threshold.

[Claim 13] The electronic device of claim 12, wherein the controller determines that the speech does not start when a face of the user is not included in the first video data.

[Claim 14] The electronic device of claim 13, wherein the controller determines that the speech is terminated when a voice of the user is not included in the second audio data for at least a preset time.

[Claim 15] The electronic device of claim 14, wherein the controller attenuates a threshold for determining the speech end to re-determine whether the speech is terminated when at least one of result values of analyzing the second video data and the second audio data is less than or equal to the threshold.

[Claim 16] The electronic device of claim 15, wherein the controller determines that the speech is terminated when a face of the user is not included in the second video data.

[Claim 17] A platform for recognizing a voice command, the platform comprising:
a multi-media framework which collects and records voice input or video input to generate first video data or first audio data before conversion into a voice recognition mode, and generates second video data or second audio data after conversion into the voice recognition mode; and
a voice framework which determines whether a speech starts based on at least one of the first video data or the first audio data, and determines whether the speech is terminated based on at least one of the second video data or the second audio data.

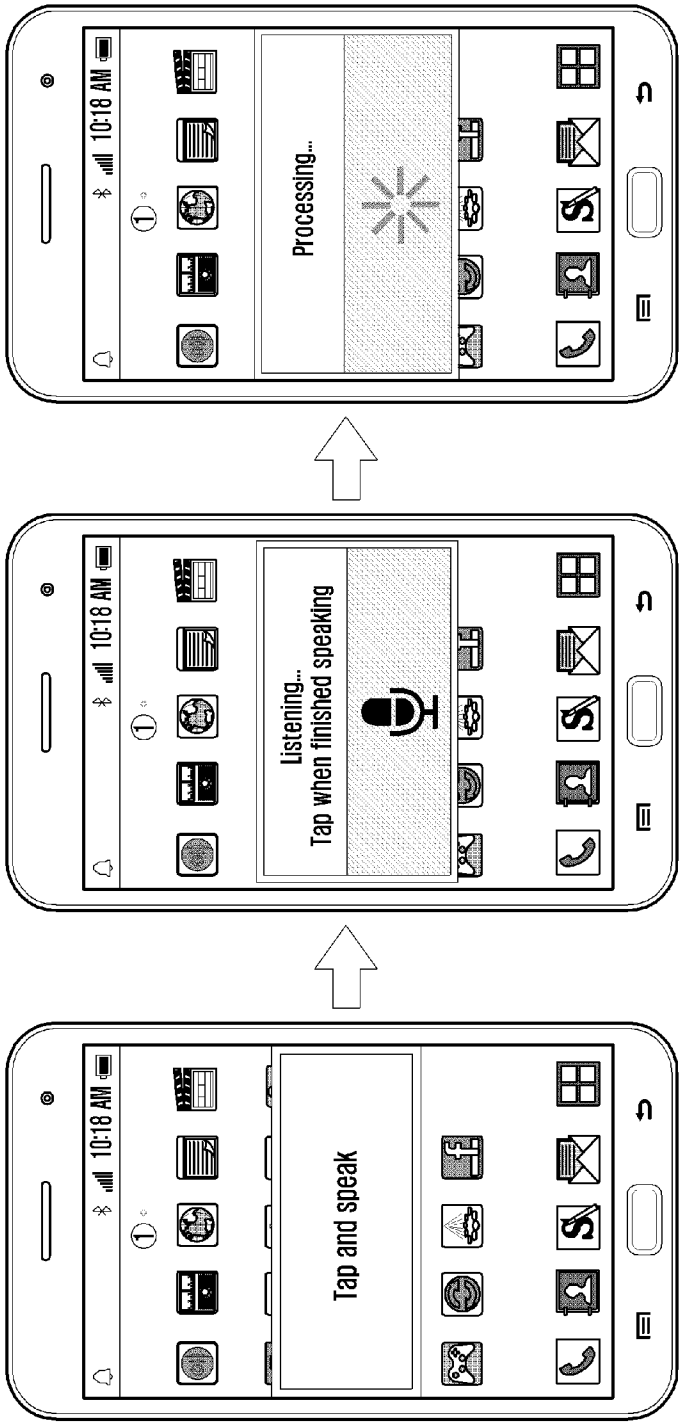
[Claim 18] The platform of claim 17, further comprising a user interface framework which displays a screen,
wherein the user interface framework displays a voice command for executing an icon displayed on the screen when the voice framework determines that the speech starts, and removes display of the voice command when the voice framework determines that the speech is terminated.

[Claim 19] The platform of claim 18, wherein the voice framework attenuates a threshold for determining the speech start to re-determine whether the speech starts when at least one of result values of analyzing the first video data and the first audio data is less than or equal to the threshold, and attenuates a preset threshold for determining the speech end to re-determine whether the speech is terminated when at least one of result values of analyzing the second video data and the second audio data is less than or equal to the threshold.

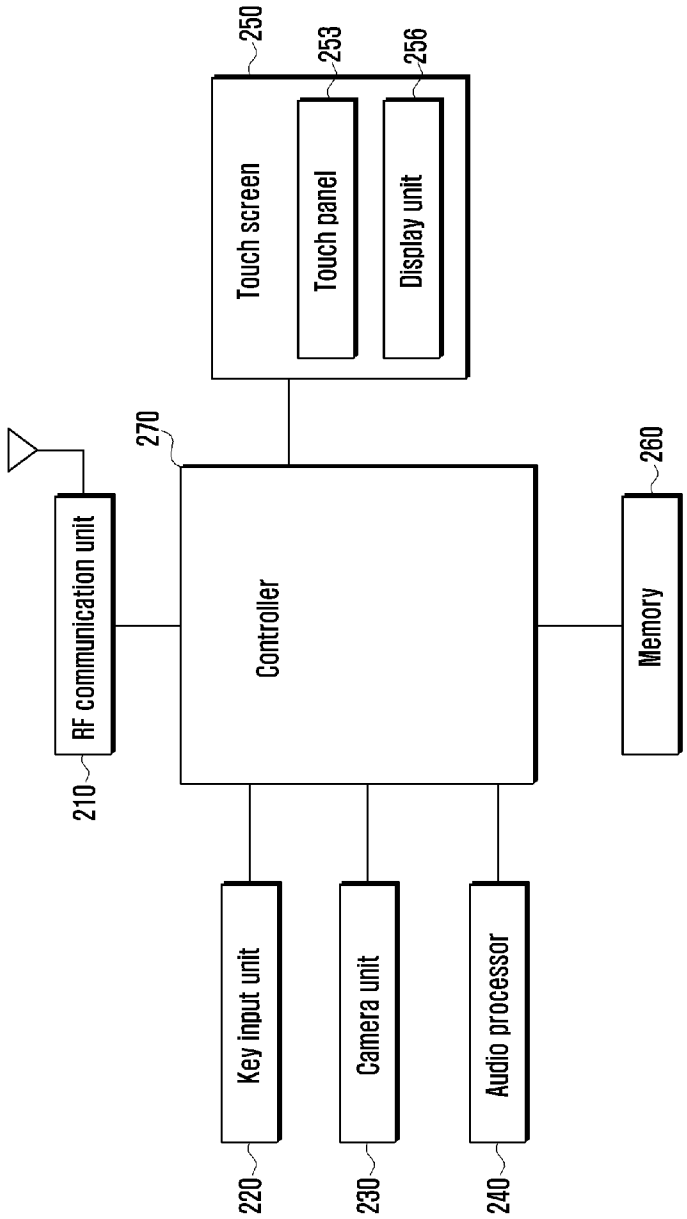
[Claim 20] The platform of claim 19, wherein the voice framework identifies a lip of a user from the first video data to determine whether the speech starts based on motion of the lip, and determines that the speech is terminated when a voice of the user is not included in the second audio data for at least a preset time.

[Claim 21] The platform of claim 20, wherein the voice framework determines that the speech is not started when a face of the user is not included in the first video data, and determines that the speech is terminated when a face of the user is not included in the second video data.

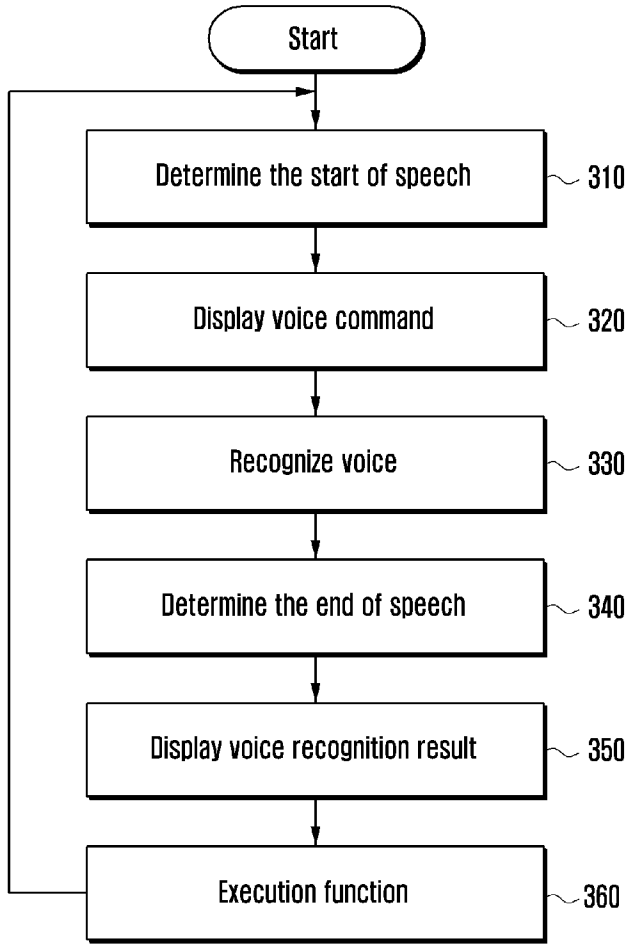
[Fig. 1]


10:18 AM
Tap and speak
10:18 AM
Listening...
Tap when finished speaking
10:18 AM
Processing...

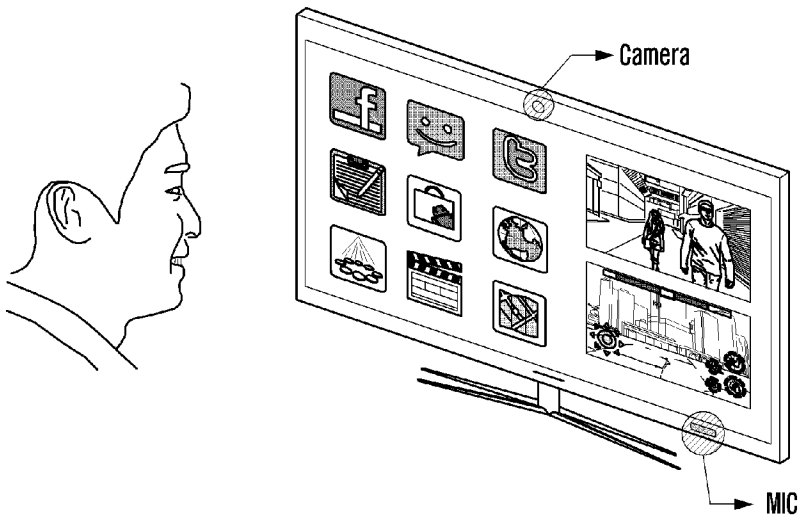
[Fig. 2]


250
253
256
Touch screen
Touch panel
Display unit
270
260
RF communication unit
Controller
Memory
210
Key input unit
Camera unit
Audio processor
220
230
240

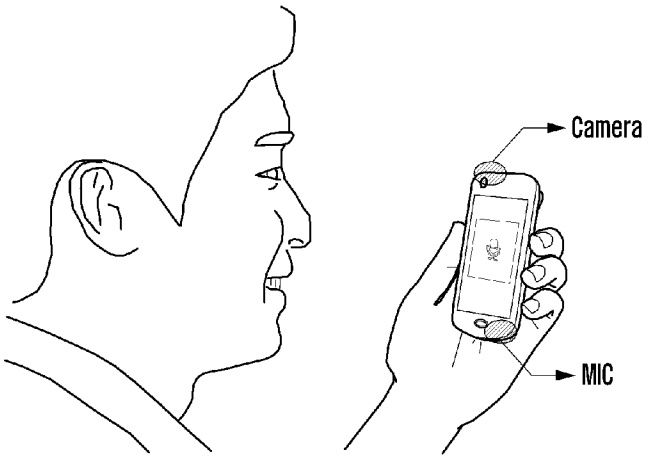
[Fig. 3]

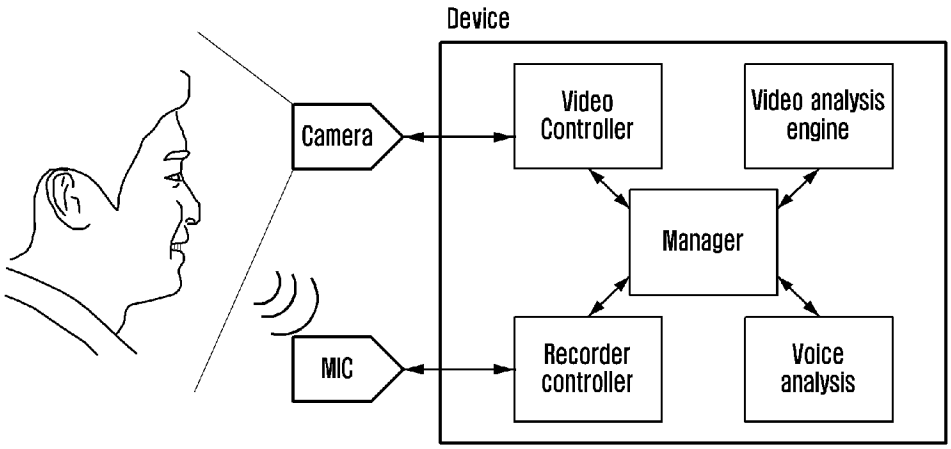

Start
Determine the start of speech
310
Display voice command
320
Recognize voice
330
Determine the end of speech
340
Display voice recognition result
350
Execution function
360

[Fig. 4]

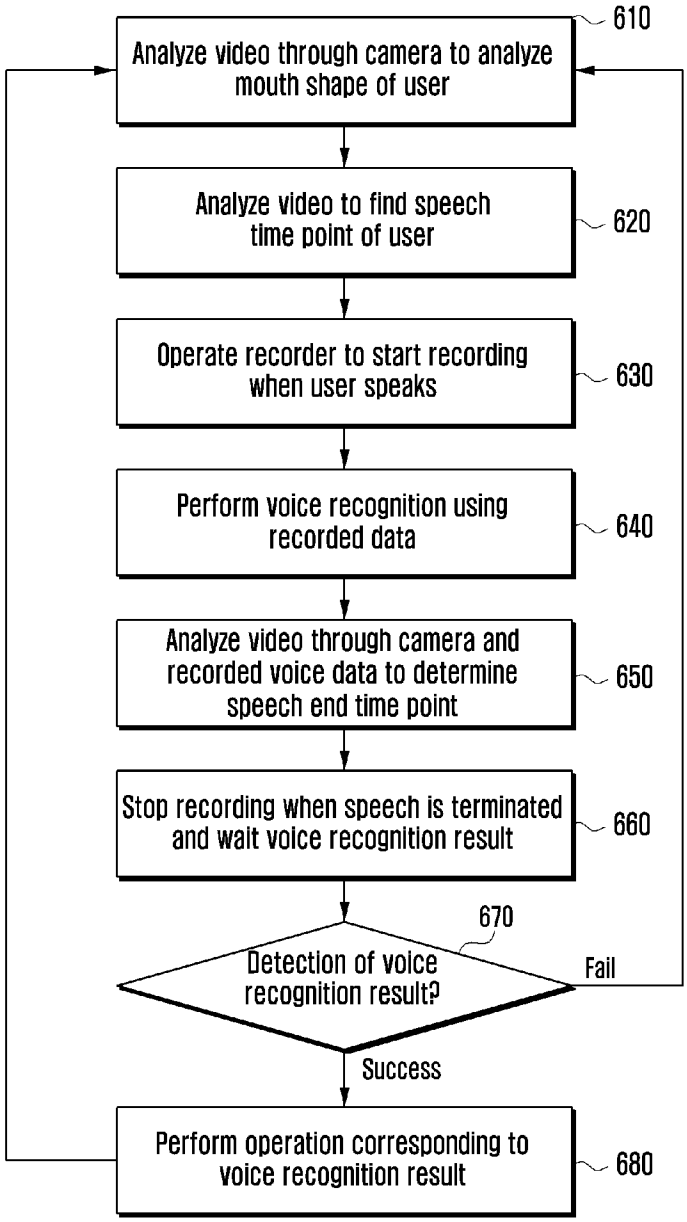

Camera
MIC

[Fig. 5]

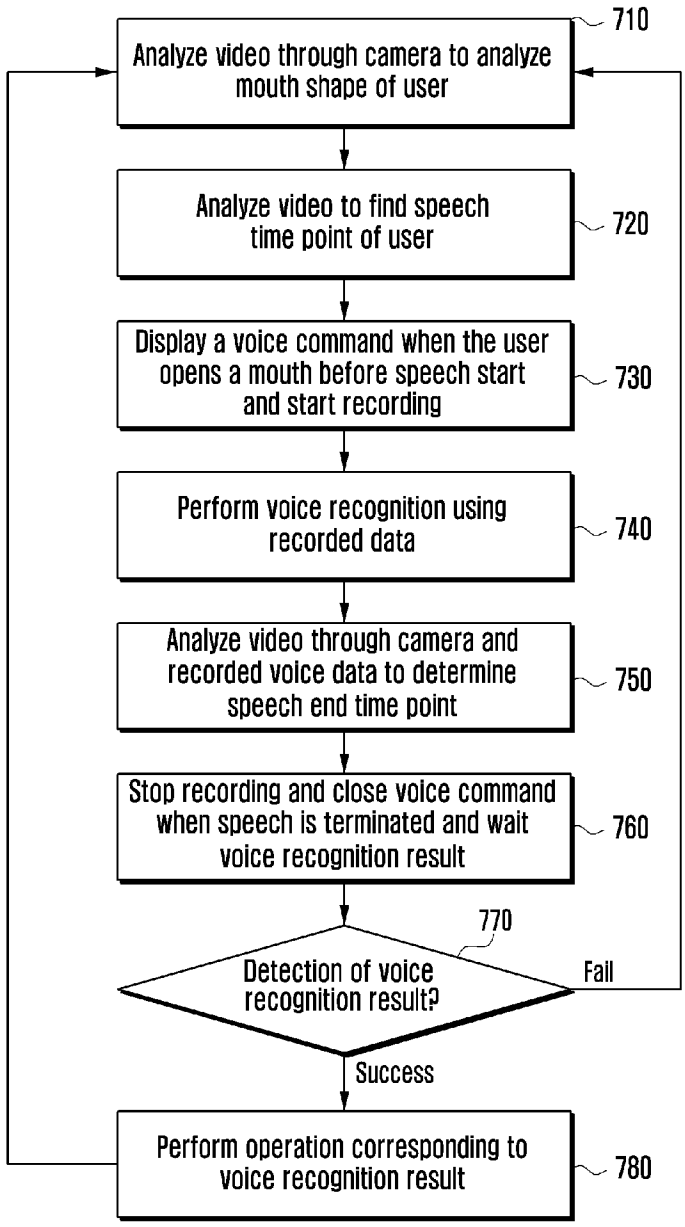

Camera
MIC


Device
Camera
MIC
Video Controller
Video analysis engine
Manager
Recorder controller
Voice analysis

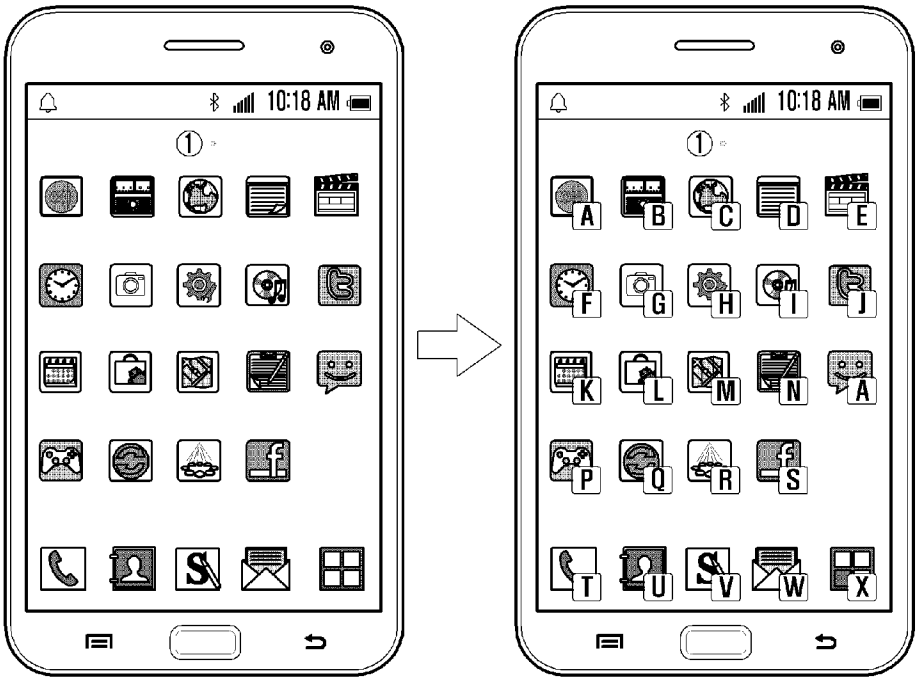
[Fig. 6]


610
Analyze video through camera to analyze mouth shape of user
Analyze video to find speech time point of user
620
Operate recorder to start recording when user speaks
630
Perform voice recognition using recorded data
640
Analyze video through camera and recorded voice data to determine speech end time point
650
Stop recording when speech is terminated and wait voice recognition result
660
670
Detection of voice recognition result?
Fail
Success
Perform operation corresponding to voice recognition result
680

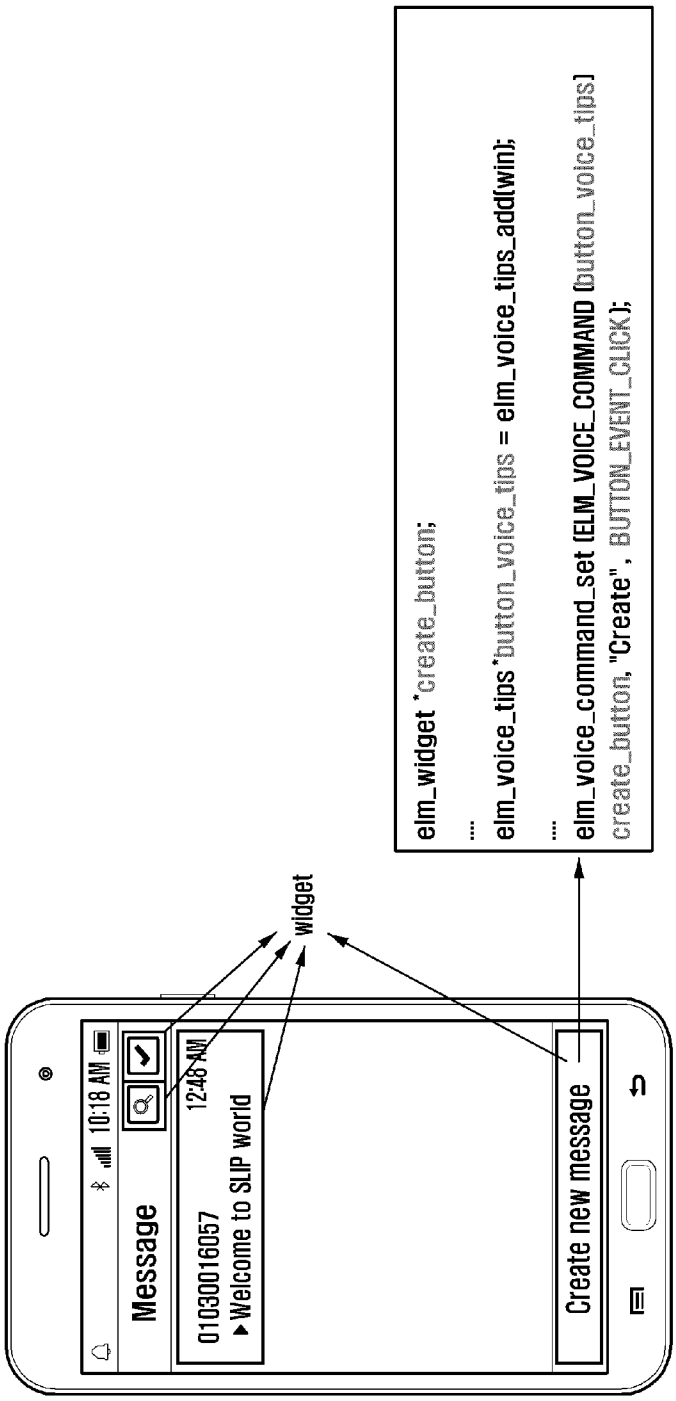
[Fig. 7]


Analyze video through camera to analyze mouth shape of user
710
Analyze video to find speech time point of user
720
Display a voice command when the user opens a mouth before speech start and start recording
730
Perform voice recognition using recorded data
740
Analyze video through camera and recorded voice data to determine speech end time point
750
Stop recording and close voice command when speech is terminated and wait voice recognition result
760
770
Detection of voice recognition result?
Fail
Success
Perform operation corresponding to voice recognition result
780

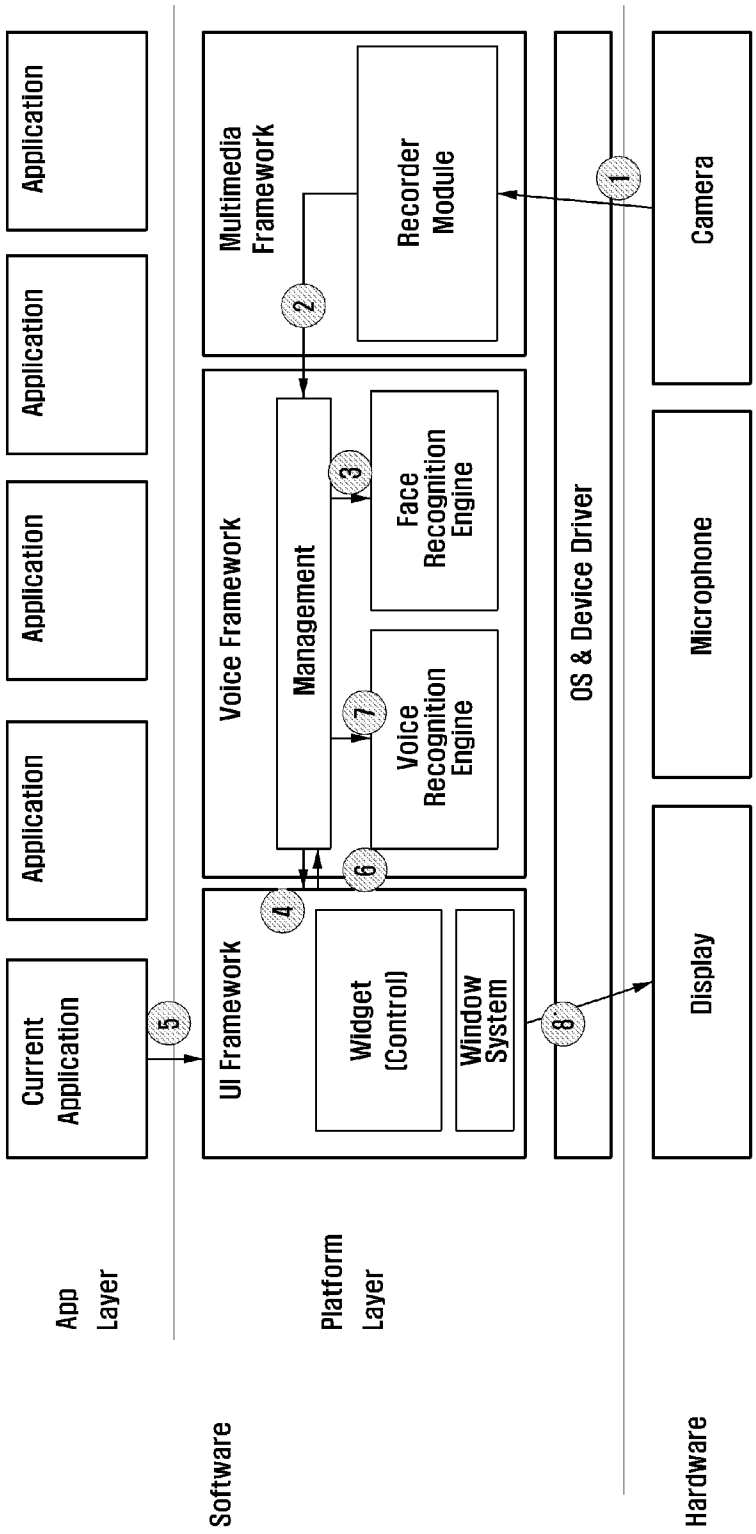
[Fig. 8]



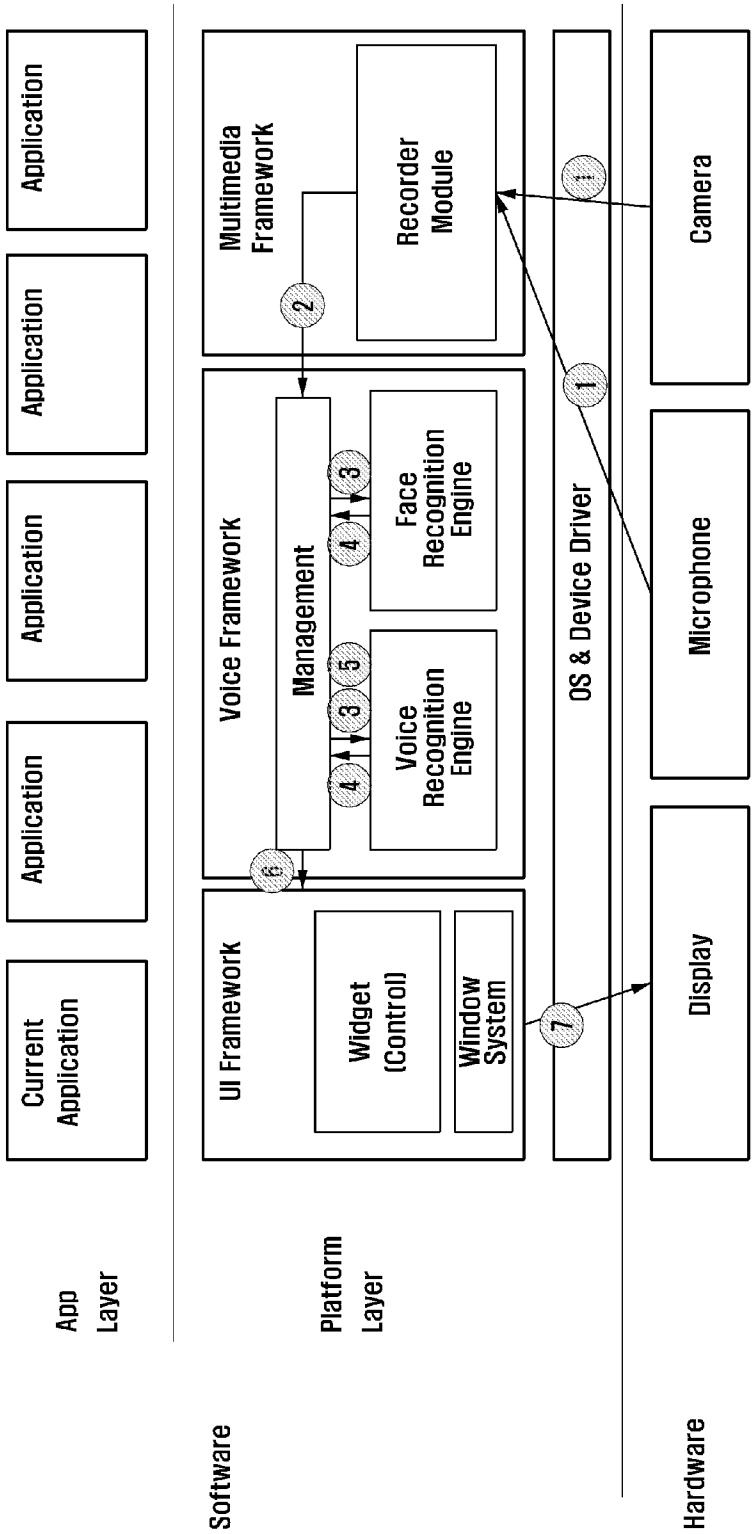
[Fig. 9]


elm_widget *create_button;
....
elm_voice_tips *button_voice_tips = elm_voice_tips_add(win);
....
elm_voice_command_set (ELM_VOICE_COMMAND (button_voice_tips)
create_button, "Create", BUTTON_EVENT_CLICK);
widget
10:18 AM
Message
12:48 AM
01030016057
▶Welcome to SLIP world
Create new message

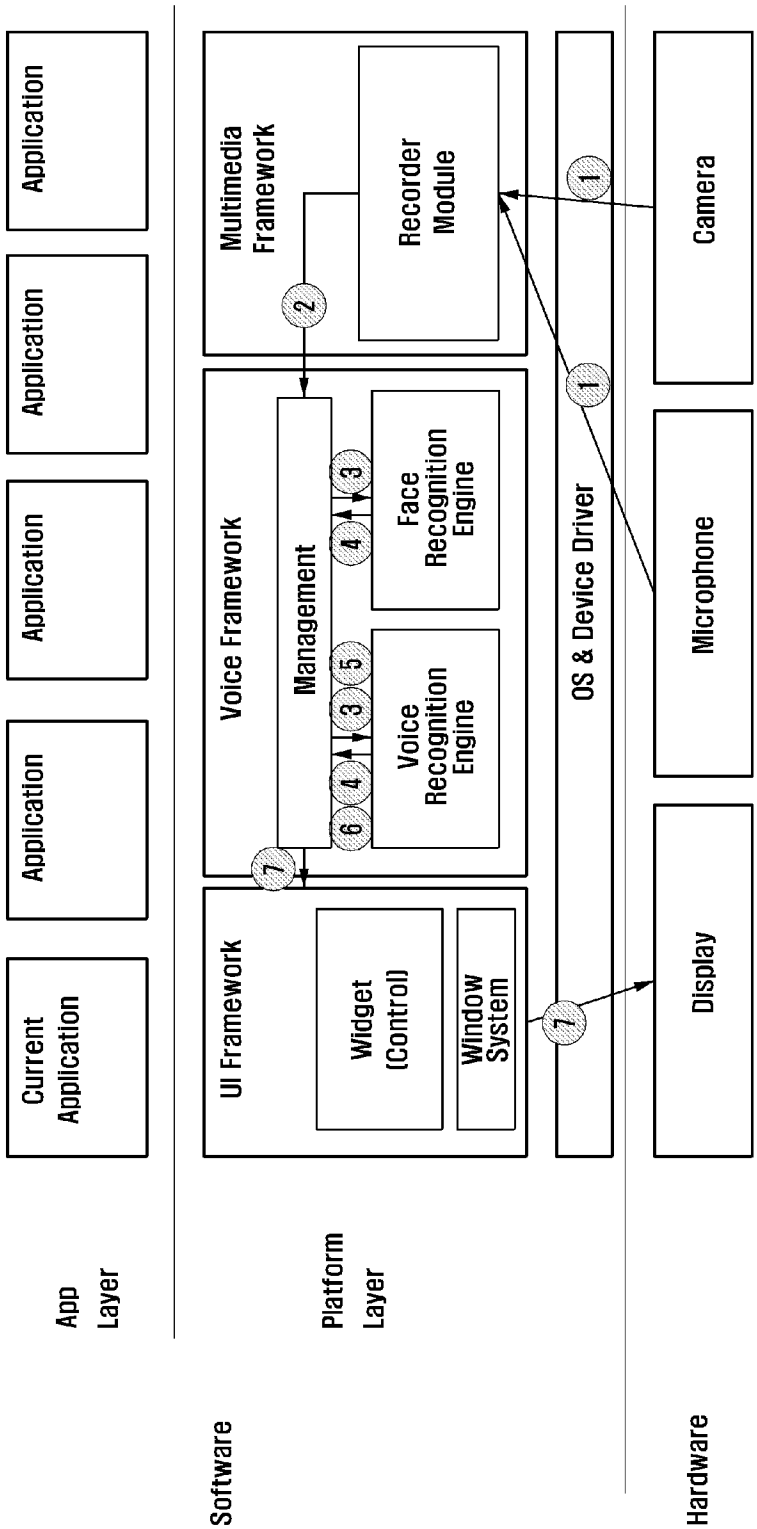
[Fig. 10]


Application
Application
Application
Application
Current Application
Multimedia Framework
Recorder Module
Voice Framework
Management
Face Recognition Engine
Voice Recognition Engine
UI Framework
Widget (Control)
Window System
OS & Device Driver
Camera
Microphone
Display
1
2
3
4
5
6
7
8
App Layer
Platform Layer
Software
Hardware

[Fig. 11]


Application
Application
Application
Application
Current Application
App Layer
Multimedia Framework
Recorder Module
Voice Framework
Management
Face Recognition Engine
Voice Recognition Engine
UI Framework
Widget (Control)
Window System
Platform Layer
OS & Device Driver
Camera
Microphone
Display
Software
Hardware
1
2
3
4
5
6
7

[Fig. 12]


App Layer
Platform Layer
Software
Hardware
Current Application
Application
Application
Application
Application
UI Framework
Widget (Control)
Window System
Voice Framework
Management
Voice Recognition Engine
Face Recognition Engine
Multimedia Framework
Recorder Module
OS & Device Driver
Display
Microphone
Camera
1
2
3
4
5
6
7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2013/006749

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/00(2006.01)i, G06K 9/20(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 15/00; G09C 1/00; G10L 11/00; G10L 19/00; G06K 9/00; G10L 21/00; G06K 9/20

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: voice, recognition, video, start

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2002-0035475 A1 (SHOUTAROU YODA) 21 March 2002 See paragraphs [0025], [0030], [0033], [0041]-[0051]; and figures 1, 2, 4, 7.	1-3,9-11,17,18
A		4-8,12-16,19-21
A	US 6243683 B1 (GEOFFREY W. PETERS) 05 June 2001 See column 2, line 39 - column 4, line 6; and figures 1-3.	1-21
A	US 6690815 B2 (ISAO MIHARA et al.) 10 February 2004 See column 10, line 10 - column 11, line 61; and figures 15-17.	1-21
A	US 2009-0326944 A1 (TAKEHIDE YANO et al.) 31 December 2009 See paragraphs [0021]-[0044]; and figures 1, 2.	1-21
A	US 2005-0165604 A1 (TOSHIYUKI HANAZAWA) 28 July 2005 See paragraphs [0032]-[0046]; and figures 1, 2.	1-21

☐ 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
"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 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
22 November 2013 (22.11.2013)	**22 November 2013 (22.11.2013)**

Name and mailing address of the ISA/KR
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701, Republic of Korea
Facsimile No. +82-42-472-7140

Authorized officer
KIM, Do Weon
Telephone No. +82-42-481-5560

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/KR2013/006749

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2002-0035475 A1	21/03/2002	DE 60124471 D1	28/12/2006
		DE 60124471 T2	06/09/2007
		EP 1187099 A2	13/03/2002
		EP 1187099 A3	04/09/2002
		EP 1187099 B1	15/11/2006
		JP 2002-091466 A	27/03/2002
		US 6604073 B2	05/08/2003
US 6243683 B1	05/06/2001	None	
US 6690815 B2	10/02/2004	JP 03688879 B2	31/08/2005
		JP 11-219421 A	10/08/1999
		US 2002-0126879 A1	12/09/2002
		US 2003-0048930 A1	13/03/2003
		US 6504944 B2	07/01/2003
US 2009-0326944 A1	31/12/2009	JP 2010-008854 A	14/01/2010
		US 8364484 B2	29/01/2013
US 2005-0165604 A1	28/07/2005	CN 1628337 A	15/06/2005
		EP 1513135 A1	09/03/2005
		WO 03-107326 A1	24/12/2003