

DATE OF TRANSLATION: 25-Oct-22

ELECTRONIC FILE NAME: JPA 2003223188

SOURCE LANGUAGE: Japanese (JP)

TARGET LANGUAGE: English (US)

TRANSPERFECT JOB ID: US1410674

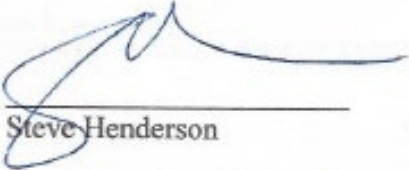
TransPerfect is globally certified under the standards ISO 9001:2015 and ISO 17100:2015. This Translation Certificate confirms the included documents have been completed in conformance with the Quality Management System documented in its ISO process maps and are, to the best knowledge and belief of all TransPerfect employees engaged on the project, full and accurate translations of the source material.

TCert v. 2.0



July 24, 2023

I, Steve Henderson, am a professional translator, and I am fluent in English and Japanese. I translated the enclosed/attached JPA 2003223188 file. All statements made in this declaration are of my own knowledge are true and all statements made on information and belief are believed to be true. I make these statements under the understanding that willful false statements and the like are punishable by fine or imprisonment, both (18 U.S.C. 1001).



Steve Henderson

(12) **Japanese Unexamined
Patent Application Publication**
(A)

(51) Int.Cl. ⁷	Identification codes	F1	Theme code (reference)
G10L 15/28		G06F 3/16	320H 5D015
G06L 3/16	320		340A
	340	G10L 3/00	571A
G10L 15/00			551A
			561H

Request for review Not invoiced Number of claims 18 OL (17 pages in total)

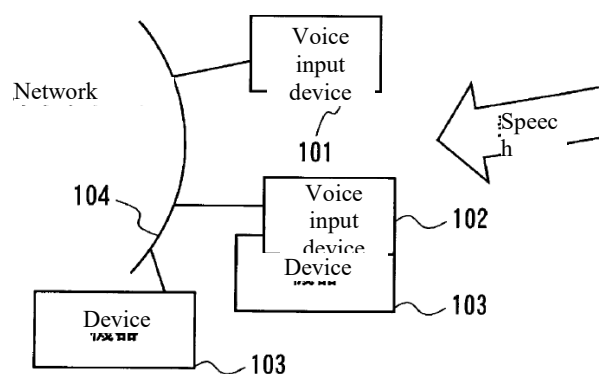
(21) Application number	Patent application 2002 - 19457 (P2002 - 19457)	(71) Applicant	000003078 Toshiba Corporation 1-1-1 Shibaura, Minato-ku, Tokyo
(22) Date of application	January 29, 2002	(72) Inventor	Masahide Ariu Within Toshiba Research and Development Center, Inc., 1-1 Komotoshiba-cho, Yuki-ku, Kawasaki-shi, Kanagawa Prefecture
		(74) Representative	100083161 Attorney Hideaki Togawa
			F-term 5D015 KK02 LL05 LL07 (reference)

(54) [Name of Invention] Voice input system, voice input method, and voice input program

(57) [Summary]

[Issue] A voice input system, voice input method, and voice input program are provided to enable processing of speech without burdening the user when the voice of the user can be used for voice input on multiple voice input devices.

[Solution means] Multiple voice input devices are connected through a network, and when these voice input devices share information, and when voice is input, the determined information is sent to other voice input devices, and each voice input device can process the voice input without burdening the user by making a processing decision regarding the voice based on the processing results of the voice input device and other voice input devices.



[Scope of Claims]

[Claim 1] A voice input system characterized by a plurality of voice input devices connected to a net

work, each voice input device connected to the network detecting voice information input in each voice input device connected to the network, providing and receiving the information regarding the voice detected in each voice input device connected to the network to other voice input devices via the network as determinative information, with each voice input device connected to the network determines processing and execution of the voice information based on the determinative information in the voice input device and other voice input devices.

[Claim 2] The voice input system as claimed in Claim 1, characterized by a plurality of voice input devices connected to said network configured in a rank relationship based on predetermined rules, and each voice input device connected to said network using said rank information in the voice input device itself and other voice input devices as said determination information to determine processing and execution of said voice information.

[Claim 3] The voice input system according to Claim 1, characterized in that a plurality of voice input devices connected to said network are grouped together based on predetermined rules, and comprise a region that stores information about said group, and the storage region related to said group comprises a mechanism for performing work in association with a storage region related to a voice input device connected to a network, and that the processing and execution of said voice information are determined using information of the storage region related to said group as the determinative information.

[Claim 4] The voice input system of Claim 1, characterized in that a plurality of voice input devices connected to the network have common time information, and when the voice information is detected by each voice input device connected to the network, the detection time information is transmitted to other voice input devices, and each voice input device connected to the network uses the detection time information in the applicable voice input device itself and other voice input devices as information to determine processing and execution of the voice information.

[Claim 5] The voice input system of Claim 1, characterized in that a plurality of voice input devices connected to the network comprises a common measure of the volume of the input voice information, and when the voice information is detected by each voice input device connected to the network, the volume information is transmitted to other voice input devices, and each voice input device connected to the network uses the volume information in the applicable voice input device itself and other voice input devices as information to determine processing and execution of the voice information.

[Claim 6] The voice input system according to Claim 1, characterized in that a plurality of voice input devices connected to the network comprise a means for estimating a signal-to-noise ratio of input voice information, and wherein each voice

input device connected to the network comprises a means for transmitting said signal-to-noise ratio information to other voice input systems when said voice information is detected by each voice input device connected to the network, and wherein each voice input device connected to the network uses said signal-to-noise ratio and volume information in the applicable voice input device itself and other voice input devices as said determinative information to determine processing and execution of said voice information.

[Claim 7] The voice input system according to Claim 1, characterized in that a plurality of voice input devices connected to said network comprises a storage area that stores past status history, each voice input device connected to said network communicates said past status history to other voice input devices, and each voice input device connected to said network uses said past status history of the applicable voice input device itself and other voice input devices as said information to determine processing and execution of said voice information.

[Claim 8] The voice input system of Claim 1, characterized in that a plurality of voice input devices connected to said network comprises a means for storing acoustic and verbal information and for matching and speech recognition using said acoustic and verbal information, and each voice input device connected to said network, once said voice information is detected, transmits said voice recognition information based on said voice information to other voice input devices, and each voice input device connected to said network uses the applicable voice input device itself and other voice input devices to determine processing and execution of said voice information as said determinative information.

[Claim 9] The voice input system of Claim 1, wherein a plurality of voice input devices connected to the network comprises a means that can be identified with respect to the voice source of the voice information, with each voice input device connected to the network conveying information about the voice source to other voice input devices when it detects the voice information, and each voice input device connected to the network making a decision on processing and execution of the voice information as determinative information regarding the voice source in the applicable voice input device and other voice input devices.

[Claim 10] The voice input system of Claim 1, wherein a plurality of voice input devices connected to the network comprise a common control instruction system that can control each voice input device and other voice input devices, and wherein the control instructions are transmitted to other voice input devices on the network for the captured voice, and the control instructions are received from other voice input devices, with the content of the instructions is executed.

[Claim 11] The voice input system of Claim 1, wherein multiple voice input devices connected to the network comprise a region that stores information of controllable devices connected to network in addition to the applicable voice input device and other voice input devices, and wherein the voice input information is processed and the information is received from controllable devices by using other controllable device information in relation to the captured voice information.

[Claim 12] The voice input system as claimed in Claim 1,

3

characterized in that each voice input device connected to the network comprises a sensor device and signal processing means other than the voice input means, and uses other sensor device information as information to determine processing and execution of voice information for the captured voice information.

[Claim 13] The voice input system as claimed in Claim 1, characterized in that the criteria for determining the processing of voice information detected based on information of each voice input device and other voice input devices connected to the network can be changed using information of other voice input devices, or user settings.

[Claim 14] The voice input system of Claim 1, wherein a plurality of voice input devices connected to the network comprise an indication of the state of each system, and a function is provided that controls how the captured voice information is to be conveyed to the user, and is characterized in that processing decisions are made and execution and information are displayed.

[Claim 15] The voice input system as claimed in Claim 1, characterized in that some or all means of said plurality of voice input devices connected to said network are shared with a means in a function other than the said voice input device .

[Claim 16] The voice input system as claimed in Claim 1, characterized in that even if some of the functions of the plurality of voice input devices connected to the network are physically separate, it can function as a single system by network technology connection.

[Claim 17] A voice input method characterized in that the voice input device comprises the steps of detecting voice information input in a plurality of voice input devices connected to a network, respectively, receiving the detected voice information from each voice input device connected to the network to other voice input devices via the network as determination information, and determining the processing and execution of the voice information based on the determinative information in the applicable voice input device itself and other voice input devices connected to the network.

[Claim 18] A voice input program characterized in that voice information input in a plurality of voice input devices connected to a network is detected with said voice information detected in the network received from other voice input devices via the network as determinative information, and each voice input device connected to said network realizes a function to make processing decisions and execution decisions for said voice information based on the determinative information in the applicable voice input device itself, and other voice input devices.

[Detailed description of invention]

[0001]

4

[Technical field to which the invention belongs] The present invention relates to devices that handle voice, and in particular, to voice input systems, voice input methods, and voice input programs where the speech of a user may be input into a plurality of voice inputs.

[0002]

[Prior art] Up until now, when controlling the device by voice or entering voice into a device, it was assumed for the most part that the user and voice input device would be supported one-to-one. However, it may be sufficient if the voice of the user is used to enter voice into a plurality of voice input devices, for example if there are multiple such voice input devices in a room. In that case, in the past, it was necessary for the user to specify the target device specifically, and to perform operations such as suppressing voice input for devices other than those intended as the voice input.

[0003]

[Issues to be solved by the invention] The purpose of the present invention is to provide a voice input system, voice input method, and voice input program for inputting the voice of the user into a voice input device without burdening the user.

[0004]

[Methods for solving problems] The voice input system of the present invention is characterized by a plurality of voice input devices being connected to a network, detecting each voice information input in each voice input device connected to the network, sharing this with other voice input devices via the network in the form of determinative information regarding the voice information that is processed as necessary by each voice input device connected to the network, with the voice input devices connected to the network making determinations on processing and execution of said voice information based on the determinative information in the voice input device and other voice input devices.

[0005] Also, the voice input method of the present invention is characterized by detecting voice information input into a plurality of voice input devices connected to a network, respectively, and sharing information regarding said voice detected by each voice input device connected to the network with other voice input devices via the network as determinative information, with each voice input device connected to the network using the determinative information in the applicable voice input device itself, and other voice input devices to determine processing and execution of the voice information.

[0006] In addition, the voice input program of the present invention is characterized by detecting voice information input into a plurality of voice input devices connected to a network, sharing information on said voice detected by each voice input device connected to the network with other voice input devices via the network as determinative information, and realizing the function of each voice input device connected to the network to make processing and execution decisions

regarding said voice information based on the determinative information in the self-voice input device and other voice input devices.

[0007]

[Form of Invention] The voice input system according to the present invention will be described with reference to the drawings below. An o

verview of the present invention in its entirety will be described using Figure 1.

[0008] In the voice input system of the present invention, a device 103 having multiple single voice input devices 101 and 102 connected to network 104, with these voice input devices 101 and 102 mounted on the standalone voice input device 103, measures the voice commands and messages or conversations, etc., spoken by the user, and converts the input voice signals to appropriate signals by means of signal processing. The voice input system can then perform processing on voice input devices 101 and 102 mounted on device 103 using this converted signal.

[0009] In addition, voice input device 102 installed on the standalone voice input devices 101 and 103 can receive and receive information via network 104, and in processing the input voice, can transmit and receive information with other standalone voice input devices on the network and voice input devices installed in the device.

[0010] At this time, it is acceptable to use a relay method in which information is transferred to each voice input device, or a broadcast method in which information is sent from one voice input device to multiple voice input devices at the same time. However, since real-time processing of voice is an important application, this method will be explained with the broadcast method in mind.

[0011] The present invention solves the question of what to do with each voice input device when the voice of the user is input to a plurality of networked voice input devices. It can also be included in the process of the present invention even if the user's voice was only input into a single voice input device.

[0012] Although an example of the present invention mainly describes the human voice of a user, the present invention is not limited to human voice. The sound of machinery and animals, may, depending on the purpose, be acceptable.

[0013] Next, the voice input device comprising the voice input system in the embodiment of the present invention is described using Figure 2. The voice input devices (20-1 through 20-3) are connected to network 21, with voice input device 20-1 mounted to video data recorder 26 (hereinafter referred to as video), voice input device 20-3 is mounted to air conditioner 27 (hereinafter referred to as air conditioner), and voice input device 20-2 is treated as a standalone unit. Operation of video 26 is performed by voice input to voice input device 20-1, and operation of air conditioner 27 is performed by voice input to voice input device 20-3. Furthermore, as described below, each device can process

voice input even if the voice is not going to its own voice input device.

[0014] Each voice input device (20-1 through 20-3) consists of a microphone 201, a signal processing unit 202, a central processing unit 203, a storage unit 204, a network connection unit 205, and an information presentation unit 206, respectively.

[0015] The voice input that the user speaks is input to microphone 201, and the speech of the user is measured with this microphone 201. This is practical for microphones that are generally available. These microphones are made up of one microphone, multiple microphones (microphone arrays), directional and omnidirectional microphones, and other items that can be used as microphones.

[0016] The voice signal acquired from the microphone is processed in the format necessary for processing in the latter stage by signal processing unit 202. For example, this process may constitute compression of the voice signal by MPEG, or it may be converted to a cepstrum feature used in speech recognition. It should be noted that signal-processing unit 202 can be configured to perform appropriate processing in accordance with the application of the voice input device.

[0017] In addition, this signal-processing unit 202 also includes a format-conversion function for converting to a format that is communicated to information presentation unit 206 in response to instructions from central processing unit 203, which is explained below. Furthermore, in this information display portion 206, speech synthesis is performed from the content of the message conveyed to the user from central processing portion 203 to convert it into a synthetic sound signal.

[0018] In addition, it is also possible to configure the processing to be converted to the display content, or to enable processing to be performed according to the application of the voice input device in information display 206.

[0019] However, it does not matter whether the processing of the voice signal from this microphone and the processing of the information sent to information presentation unit 206 are performed by the same processing mechanism. In other words, the mechanism for the above processing is collectively referred to as signal processing unit 202.

[0020] In addition, sensor devices other than microphones are also feasible as a means to input to signal processing unit 202. Examples include motion images from cameras, tactile sensors, switches, etc. Inputs from other sensor devices shall also be configurable to be processed according to the application of the voice input device. This will be discussed later.

[0021] Central processing unit 203 controls the processing of the entire voice input device. Central processing unit 203 controls the state of the voice input device and sends instructions to each processing mechanism as necessary. The content to be controlled can be determined based on information from signal processing unit 202, network connection unit 205 and memory unit 204. Also, control information is sent to other voice input devices. The manner in which voice is processed as a voice input system of the present invention is described below.

[0022] Memory unit 204 is a mechanism that holds the

processing performed by the central processing unit 203, its work area, information from signal processing unit 202, and information from network connection unit 205. It should be noted that memory unit 204 may be on a separate circuit, such as for information storage

e in signal processing unit 202 or for storage of information from the network connection unit.

[0023] In other words, the information retention mechanism in the voice input device is collectively referred to as memory 204. Storage unit 204 may feasibly be used with a mechanism such as semiconductor memory or magnetic disk, and can be configured with any mechanism capable of holding data, even though semiconductor memory is used in this embodiment.

[0024] The method of use of memory unit 204 and the information to be stored will be described later along with the explanation of the processing of central processing unit 203.

[0025] Network connection unit 205 is a mechanism for sending and receiving information between voice input devices through network 21, and can be achieved by inter-device communication technology such as network connection over LAN and wireless technology such as Bluetooth. Network connection over LAN is used here.

[0026] In addition, it is acceptable for each or all of the mechanisms of voice input devices such as the above to share the mechanisms with those of systems with other functions. If voice-visual devices such as video systems include voice input devices, common signal processing circuits may be used to achieve mutual functions, or the same central processing circuit may be used to control the functions of voice input devices and video systems.

[0027] Although there are other examples in which voice input devices and other system functions can be realized through a common mechanism, no such information is presented here.

[0028] In addition, the above shall also include cases in which voice input devices and other systems are not separate circuit mechanisms, but rather can be controlled as separate systems in a programmatic process, even though they are shared circuits.

[0029] Next, Figure 3 shows how central processing unit 203 processes voice based on voice signals from signal processing unit 202, information from network connection unit 205, and information possessed by memory unit 204. In Fig. 3, an example is shown in which voice is input for voice input device 20-1 (hereinafter referred to as voice input device A) installed in video 26 of Fig. 2 and voice input device 20-3 (hereinafter referred to as voice input device B) installed in air conditioner 27. Further, the user is currently interacting with voice input device B, and voice input device A indicates a waiting state.

[0030] First, when a user speaks to voice input device A and voice input device B (step 301), voice is detected from a user captured by microphone 201 and signal is processed in signal

processing unit 202 of each voice input device (step 302).

[0031] In this case, since voice input device B is already interacting with the user, if voice input device B is in the process of interacting with the user and other systems are not in an interactive state, voice input device B makes a choice to process the content spoken by the user. (step 303)Next, central processing unit 202 of voice input device B processes the voice captured according to the function of the voice input device , operates the device according to the content of the voice, and waits again after the interaction ends (step 304).

10 [0032] Conversely, in voice input device A, voice input device B is in an interactive state with the user, so after signal processing is performed (step 302), no further processing is performed (step 305), and it enters a waiting state.

[0033] This makes it possible for users to easily access only voice input devices that are currently targeted for voice input, even if the voice of the user is detected by multiple voice input devices. Also, while the above mentioned examples have been given where a user speaks to multiple voice input devices, the user does not need to intentionally speak to be detected by multiple voice input devices, and this is the same in the following examples.

20 [0034] In addition, conditions such as processing if other voice input devices not in an interactive state may be prescribed by the user at their discretion or as a setting by the voice input device for conditions other than the above.

[0035] In addition, the dialogue here is not limited to one-on-one voice interactions between humans and systems, but rather may include the return of unilateral speech from humans to systems, visual responses from the system side, or responses from the system to any person, and the same applies to the dialogue used in the following explanations.

30 [0036] In addition, the voice input device has a rule-based ordering relationship, and it is possible to determine the processing of the captured voice information based on that ordering relationship. Specific examples of rules include the processing capacity of voice input devices, user settings, usage frequency, set values based on the function of voice input devices, sensor information from sources other than microphones, and combinations of these, etc.

[0037] Next, an example of ranking by the above voice input device function is described using Fig. 4.

40 [0038] There is a wearable computer (hereinafter referred to as "voice input device C") as the device equipped with voice input device and a video system (hereinafter referred to as "voice input device D") as the device equipped with voice input device , and the former is more highly ranked for specific users, and the video system is lower ranked because it can be used by unspecified users.

[0039] Here, the user uses voice input device C and voice

is spoken to input device D (step 401), and each voice input device transmits a rank order of the applicable voice input device itself when voice is detected by a user captured by microphone 201 on signal processing unit 202 (step 402).

[0040] Next, the voice input device C with the higher rank when compared to the rank

of other voice input devices, processes the voice of the user (step 403).

[0041] Voice input device D with the lower rank is not processed (step 404) and remains waiting.

[0042] In the above example, rank information is transmitted, but it is acceptable to have information other than rank in the transmitted information, and it is acceptable to exchange information in advance rather than after detecting voice, or to make a decision on whether to process it using a self-voice input device based on the preset rank information.

[0043] According to the above examples, emergency devices such as fire alarms and emergency alarms are higher in order than any other devices as devices equipped with voice input devices, and it is possible to first prioritize voice input for emergency devices, regardless of how much voice instruction is registered as a voice command with ordinary devices in response to a voice saying "help" for example.

[0044] Also, a mechanism for processing time can be provided in the voice input device, and this can be used as a reference for processing decisions. Give an example in Figure 5.

[0045] Figure 5 shows an example of a voice input for voice input devices installed in video (hereinafter referred to as voice input device E) and voice input devices installed in an air conditioner (hereinafter referred to as voice input device F), where voice input device E is installed closer to the user than voice input device F.

[0046] Here, the user speaks to voice input device E and voice input device F (step 501), and each voice input device sends a voice detection time of the voice input device when voice is detected by a user captured by microphone 201 in signal processing unit 202 (step 502).

[0047] Next, the detection time from other voice input devices that detected voice and the detection time of the applicable voice input device itself are compared, and if the applicable voice input device is the fastest, the sound is processed (step 503) and the decision is made not to process the said sound (step 504) so that the voice input device closest to the user can process the sound even if the user does not specify it.

[0048] In addition, it is possible to assume that the voice input device with the longest voice detection time was able to detect the voice of the user from the beginning to the end, and to use the time information other than the speed of voice detection as the determinative criteria, such as the voice input device processing said voice.

[0049] In addition, the volume of the user's voice can be measured from the voice captured from the microphone to be used as a reference for processing decisions. An example of the present invention utilizing volume information will be described using Figure 6.

[0050] In this case, when there is voice input device E and voice input device F described above, the user speaks to voice input

device E and voice input device F (step 601), and each voice input device transmits volume information when it detects voice from a user captured by microphone 201 at signal processing unit 202 (step 602). That is, the volume of the user's voice is measured from the voice captured from the microphone and communicated to other voice input devices on the network.

[0051] Next, the volume information from other voice input devices that detected the voice is compared with the volume information of the self-voice input device, and if the self-voice input device is the largest, the voice is processed (step 603) and, if not, the decision is made to not otherwise process it (step 604) so that the voice input device closest to the user can process the voice even if the user does not specify it, or can process the original voice with the best recorded voice. This volume information includes sound pressure level, sound power level, or units such as phon and sone.

[0052] In addition, the signal-to-noise ratio of the user's voice to the ambient noise can be calculated from the voice captured from the microphone to be used as a reference for processing decisions. An example of the present invention utilizing a signal-to-noise ratio will be described using Figure 7.

[0053] Figure 7 shows an example of voice input for a voice device installed in video (hereinafter referred to as voice input device G) and voice input device installed in an air conditioner (hereinafter referred to as voice input device H), and there is a noise source, and voice input device G is located farther from the sound source than voice input device H.

[0054] At the beginning, each voice input device constantly captures voice and measures surrounding noise information (step 701).

[0055] Next, the user speaks to voice input device G and voice input device H (step 702), and each voice input device detects voice from the user captured by microphone 201 at signal processing unit 202, calculates the signal-to-noise ratio based on the noise information when the user's voice is captured from the microphone, and communicates this to other voice input devices on the network (step 703).

[0056] Next, the signal-to-noise ratio information from other voice input devices that detected sound is compared with the signal-to-noise ratio information of the voice input device, and if the voice input device is the largest, the voice is processed (step 704). Otherwise, a decision is made that the voice is not to be processed (step 705).

[0057] This allows the voice input device closest to the user to process voice even if the user does not specify it, or to process the original voice with the best recorded voice. In the example given here ambient sounds were constantly captured even during non-vocalization to calculate the noise, but the noise may be estimated based on intervals of silence during voice, for example, after detecting voice.

[0058] In addition, past history of usage conditions can be kept in the memory unit, and can be used to determine processing. An example of the invention utilizing past history will be described using Figure 8.

[0059] Figure 8 shows an example of voice input for a voice

input device installed in video (hereinafter referred to as voice input device I) and voice input device installed on an air conditioner (hereinafter referred to as voice input device J), where voice input device I is used more frequently than voice input device J.

[0060] To start, the user speaks to both voice input devices (step 801), and in response to this vocalization, the time of use and the number of times of use are communicated to other voice input devices over the network (step 802).

[0061] Meanwhile, with voice input device I, comparisons with the usage history of voice input device J are made, and it is decided to perform voice processing if voice input device I is most commonly used (step 803), so that voice input device I, which is often used without the user having to specify it, can be used.

[0062] Moreover, in comparison to the usage history of voice input device I using voice input device J, if voice input device J is not used very much, voice processing is not performed (step 804), it remains in waiting mode.

[0063] Also, it is possible to use the recognition results to determine the processing of the captured voice by means of voice recognition. The information from the signal-processing unit is processed by a voice recognition mechanism and the result is passed to the central processing unit. Speech recognition performed at this time may be handled by the central processing unit.

[0064] Also, the method used for speech recognition can be a method generally utilized, such as HMM and DP matching using mixed normal distribution as models, and the HMM and language model used at this time can be in the memory unit. The vocabulary of voice recognition may be different for each voice input device or common to all such devices. Further, the vocabulary can be matched with control instructions to enable voice commands. An example of the present invention utilizing this voice recognition is described in Figure. 9.

[0065] Figure 9 shows an example of voice input for a voice input device installed in video (hereinafter referred to as video input device K) and a voice input device installed in an air conditioner (hereinafter referred to as video input device L).

[0066] In the first instance, when the user says [play] as associated with voice input device K (step 901, each voice input device detects and recognizes the voice (step 902).

[0067] The central processing unit receives the voice recognition results, and determines whether or not the voice input device was spoken to based on the recognition results (step 903) and communicates the determinative results and recognition results to other voice input devices via the network (step 904).

[0068] On the other hand, by looking at the determinative results and recognition results of other voice input devices, voice input device K makes it possible to use voice input devices that are subject to voice even if the user does not specify them, by performing processing for said voice if Voice

input device is determined to have spoken (step 905).

[0069] On the other hand, because voice input device L does not determine whether voice is directed at it (step 906), it remains in waiting mode.

[0070] Also, there is a means for identifying the sound source, and it is possible to use the identification results to determine the processing of the voice. The types of sound sources, such as humans, machines, and animals, may be considered depending on the purpose of use, but we will explain the case where human voice is used as an example. Speaker identification is performed on the voice information of the user from the signal-processing unit, and the results are communicated to the central processing unit. The method of identifying a speaker may be a method that can identify the characteristics of an individual or a speaker (e.g., gender, age group, etc.), such as one that determines the likelihood of HMM learned or adapted for each speaker, or one that selects the category closest to the model for each gender or age group.

[0071] An example of the invention using this speaker identification is described using Figure 10 below.

[0072] In Figure 10, an example is shown in which voice is input for the voice input device installed in the video (hereinafter referred to as voice input device M) and the voice input device installed in the air conditioner (hereinafter referred to as voice input device N), where the user can only process voice on standalone voice input device M.

[0073] In the first instance, if voice is heard by a user for each voice input device (step 1001), the voice input device that detects the voice of the user performs speaker identification (step 1002), determines whether or not the voice should be processed by the applicable voice input device to which the speech is directed (step 1003), and communicates the determinative result and speaker identification result to other voice input devices via the network (step 1004).

[0074] In addition, by looking at the determinative results and speaker recognition results of the applicable voice input device itself and other voice input devices, if it is possible to speak to the applicable voice input device itself and make a determination (step 1005), with processing performed on said voice, and conversely, if the other voice input device N does not communicate to the voice input device to which the speech is directed (step 1006), processing is not performed, so that if a voice input device is available to a specific user, it is possible to use the voice input device for speech not specified by the user.

[0075] In addition, if the reliability of speaker identification is low or there are multiple speaker candidates, the system side may further ask for a PIN number, a standard phrase, or free speech to obtain further data to increase the identification accuracy, and then perform processing after speaker identification.

[0076] Further, although human speaker recognition is described here, it is possible to perform identification and subsequent processing in accordance with the sound of a disabled person or an animal.

[0077] Also, shared instructions with voice input devices and

other devices are possible on the network, with control of them within the range mutually allowed. This can inhibit the operation of other voice input devices or improve compatibility between voice input devices.

[0078] This example is described in Figure 11.

[0079] For example, when all voice input devices 1101 connected to network 1102 have shared power management instructions such as [Power ON] [Power OFF] and [Power saving], instructions for operating the power supply of any voice input device 1

101, including multiple systems at once, these can be sent over the network from personal computer 1103 connected to network 1102, and each voice input device can execute the instructions.

[0080] In addition, it is possible to execute control commands with simpler and more reliable voice by providing a means to match the input voice with the voice control commands shared with voice input devices and other devices on the network. This example will be described using the flow diagram in Figure. 12.

[0081] In the example shown in Figure 12, where there is a video with a voice input device (voice input device O) and an air conditioner with a voice input device (voice input device P), the user selects [video] after which common commands such as [power supply ON] and [power supply OFF] are provided.

[0082] In this case, if the user says [video] [power ON] (step 1201), voice input device O and voice input device P recognize the device name and device instructions using the matching method used for the aforementioned voice recognition (step 1202), and determine whether they are instructions to their system or can be processed (step 1203).

[0083] The results can be communicated to other voice input devices and controllable devices on the network (step 1204), and these results and results from other voice input devices can be used to determine if the voice input device should process these (step 1205) and then processing corresponding to the control instructions can be performed.

[0084] It is noteworthy that using the results obtained from multiple voice input devices for shared instructions differs from previous voice remote control and voice commanded devices.

[0085] In addition, when there are multiple controlled devices by voice on the network, by providing a mechanism that allows the memory unit to store information about all or some of the control instructions, and a means for matching the input voice with those instructions, it is possible to execute control instructions by voice with greater ease and confidence.

[0086] This example is described using Figure 13 and Figure 14 below. Assuming that there is video (voice input device Q) and air conditioner (voice input device R) that can be controlled by voice input devices on the network, if voice input device Q is [play] [stop] and voice input device R is [temperature increase] [temperature decrease], etc., for voice instructions, each voice input device on the network will be able to associate recognition words with the target device and store them.

[0087] Figure 13 represents the concept of linking this recognition word to the target device and processing content. The linking of a recognized word to the process, such as in this

Figure 13, shall be realized through simple tabulations, object orientation, and high-order knowledge processing, the details of which are omitted here.

[0088] When the user says "video" "playback" as shown in the flow diagram in Figure 14 (step 1401), voice input device Q and voice input device R detect and recognize voice (step 1402).

[0089] Furthermore, using the concept shown in Fig. 13, the content of speech is determined (step 1403), the results are transmitted to other voice input devices on the network (step 1404), and based on these results and the results sent by other voice input devices, it is determined whether the speech sent by another voice input device should have been processed (step 1405), and the processing corresponding to the control instructions is performed.

[0090] In the case of [video] and [playback] as described above, it can be determined that the voice of both voice input devices was an instruction for playback of the video by linking the recognition word to the target device and the processing content as shown in Figure 13. Furthermore, the information transmitted over the network can be used to uniquely interpret the voice, and the voice input device can perform processing corresponding to the recognition results.

[0091] Although we have basically mentioned examples of word recognition in the previous example of speech recognition, even if word spotting and continuous speech recognition technologies are used, and there are differences in the specifications of speech recognition in each voice input device, one may assume that the concept of mapping the recognition results and processing content as shown in Figure 13 is acceptable.

[0092] In addition, with regard to the example shown in Figure 14 above, it shall be possible to process controlled devices connected to networks other than voice input devices. An example is described using Figure 15.

[0093] As shown in Figure 15, an air conditioner 1501 with a voice input device, a single voice input device 1502, and video 1503 are connected to network 1504, where the user uses speech to operate video 1503.

[0094] This voice input device detects and recognizes voice using the flow shown in the flow diagram of Figure 14, and links the recognition results to the processing content using a concept similar to that shown in Figure 13. Then, the recognition results and processing content are determined before sending to other systems on network 1504.

[0095] As a result, video 1503 receives the processing content according to the recognition results, and can perform vocalization. Thus, video 1503 itself can be controlled by voice by information that can be controlled by oneself flowing to a network without the need for a voice input device, and a concept for recognition results and processing content for each voice input device as shown in Figure 13 can be created.

[0096] With regard to the example of the present invention using voice recognition described in Figures 12 to 15, an example is provided in which the results of voice recognition and determination are transmitted to all systems of the network using a broadcast

method, but it is also acceptable to directly convey the results of recognition and determination solely to the target device based on the results of recognition.

[0097] In addition, when there is a sensor other than the microphone voice input in the voice input device, the content of the detected voice can be determined using the sensor information. This example is described using Figure 16.

[0098] As shown in Figure 16, an air conditioner 1601 with a voice input device and a single voice input device 1602 are connected to network 1603. Also, this single voice input device has a camera and can incorporate surrounding image information from the camera. The input of this camera is input to signal processing unit 202 of Figure 2 and image processing is performed.

[0099] In this voice input system, the user speaks to the voice input device of air conditioner 1601. Here, the camera attached to standalone voice input device 1602 estimates which direction the speaker is facing. It should be noted that, with regards to which voice input device that this speaker faces, such techniques as extracting humans from images, estimating facial parts and their orientation, and estimating voice detected from movements of the mouth are available, but no details are presented here.

[0100] If it is determined from the face of the assumed speaker that he or she is facing air conditioner 1601, the target device for voice is determined to be the air conditioner, and each voice input device notifies the other voice input devices on network 1603 of the results and determines the processing as described in the previous example.

[0101] This is an example of utilizing image information when using a camera, but it is possible to consider direct sensor devices such as switches, and microphone arrays for sound source localization, as there is no limit to the kind of measurement technology that may be used.

[0102] Also, as stated in the configuration of the voice input device in Figure 2, microphone 201, information presentation unit 206, signal processing unit 202, central processing unit 203, storage unit 204, and network connection unit 205 are collective terms for each functions in the voice input device, so it is acceptable to consider that each may be physically divided into multiple parts through the network or directly connected. This example is described in Figure 17.

[0103] As shown in Figure 17, even if the voice input device is physically divided into two voice input devices (1701, 1702), it is connected by network 1703, and appropriate information can be exchanged. At this time, the two voice input devices (1701, 1702) can act as one voice input device for the voice of the user.

[0104] In addition, the criteria for determination with regards to voice input devices as described above can be changed depending on the information and user settings of other voice input devices. For example, in addition to information such as

detection and recognition results when voice is detected, the voice input device can also exchange processing status, processing performance, recognizable vocabulary, and processing details of other voice processing systems at regular intervals and store this information in the storage unit of the voice input device.

[0105] Using such information, it is possible for the user to control the above-mentioned determination according to their own preferences, such as when it is currently not possible to process using a given voice input device, or when correcting recognition errors by focusing more on the recognition results of voice input devices with better performance than their own devices.

[0106] In addition, the means for determining input for voice input devices as described conventionally may be combined with those described above. For example, a voice input device with an early detection time may handle voice, but within a certain allowable time, there may be no time difference, and in the same time, it may be determined by volume, or it may handle voice with the highest scored voice input device by weighing the likelihood of speech recognition and the rank of the voice input device.

[0107] Also, it may be possible to make a decision using the information obtained through the combination of the above-mentioned means of determination in a high-level agent system or knowledge processing system.

[0108] In addition, the means for determining voice input device processing as described above will not be required to be the same between voice input devices on the network. For example, if there are two voice input devices, one based only on the detection time of the voice, and the other only on the volume information, the information that is exchanged and received after the voice is detected cannot necessarily be handled, but if the processing in that case is set according to the purpose of the device, it is possible for each voice input device to make a processing decision with each voice input device without causing the processing to fail as a voice input system. In addition, when the means for determination of the voice input device as described above is different for each voice input device, it may be possible to determine processing in an agent system or knowledge processing system of a higher order than the voice input device based on information sent or received through the network.

[0109] In addition, in the determination of input for voice input devices as described above, if it is not possible to uniquely determine what type of voice input the user has made for which device based on information such as voice detection time and volume, and information such as voice recognition results and identification results, one of the voice input devices may conduct interactive processing with the user to make a decision, or use other conditions such as sensor information other than the microphone to make a decision.

[0110] Next, the information presentation unit 206 of Figure 2, which has been described above in the voice input device as described so far, and an example that serves as a supplement to the previous explanation will be described using the following Figure 18.

[0111] As shown in Figure 18, an air conditioner 1801 having

a voice input device , a single voice input device 1802, and a video 1803 having a voice input device are connected to network 1804. These voice input devices also include information presentation unit 206 of Figure 2.

[0112] As mentioned previously in this voice input system, each voice input device

exchanges its own information, i.e., recognition vocabulary, processing content, and media that can be expressed in particular as the presence or absence of information display, and stores this in the memory unit while waiting.

[0113] The information presentation unit of each voice input device in this example is equipped with a speaker, and it is possible to return the voice of any text synthesized by the central processing unit and the signal-processing unit to the user. In addition, it is assumed that some of the control instructions to the information presentation unit are shared by voice input devices. In other words, instead of one voice input device returning a response from its own information presentation unit to the user, it enables a response from the information presentation unit of another voice input device on the network to the user.

[0114] When the user says [video] and [play], the voice input device connected to the air conditioner and the standalone voice input device detect the user's voice. It should be noted that, in this case, the user is positioned closest to the single voice input device .

[0115] According to the procedures described so far, both voice input devices detect and recognize voice, and eventually determine that it is a "playback command" for "video," then communicate this to other voice input devices on the network. The voice input device connected to the video does not detect direct voice, but receives information from another voice input device on the network, interprets it as a command to the applicable voice input device itself, and performs processing when a playback instruction is issued.

[0116] At this time, since the standalone voice input device is closer to the user, when making a determination based on the volume and signal-to-noise ratio information sent on the network, it can be determined that the standalone voice input device is more suitable for voice processing than the video-connected voice input device .

[0117] Therefore, it is possible to determine that a standalone voice input device and a video voice input device are each standalone voice input devices that provide and receive voice to and from a user.

[0118] The voice input device of the video that received the playback instruction sends the video a control command for playback, while in order to inform the user that playback has started, the instruction is generated to return the synthetic voice called "playback started" from the standalone voice input device to the user, and is communicated via a network to the standalone voice input device . At this time, the control instructions sent from the voice input device of the video may be sent directly to a single standalone voice input device , along with previous transmissions of information to the network, or

may be communicated to all voice input devices in broadcast format, including the instructions to the standalone voice input device .

[0119] In this way, a response instruction sent from a video voice input device to a user can be interpreted, and a standalone voice input device can convey to the user a message of "started playing" with a synthetic voice.

[0120] In addition, we have already stated an example in which by setting a flag, a standalone voice input device and a video voice input device are in the process of interacting with the user for a certain period of time, the next utterance of the user can be processed preferentially, and it is not necessary to process it with the air conditioner voice input device .

[0121] Next, an example of a voice input device grouped according to certain criteria in a voice input device such as that described above is described using Figure 19.

[0122] In this example, the groups [Kitchen] 1901, [Wearable] 1902, and [Living] 1903 are all connected by network 1904, based on the location of the voice input device . In addition, there are voice input devices within each group, and each voice input device within each of these groups has information that can be used to identify other groups.

[0123] However, the types of information held by the memory unit in relation to other voice input devices in the group and the information held in relation to other groups do not necessarily have to be the same. Specifically, it is assumed that there is no recognition vocabulary for each voice input device in other groups, or corresponding information or processing content on the target device thereto.

[0124] The user said [Living] [Video] [Play], which was detected by the [Kitchen] and [Wearable] voice input devices. As with the examples described so far, as a result of determining whether the detected voice input device should be recognized and processed using its own voice input device , it is determined that [the voice] is not spoken to its own group but rather to the [living room] group, and the voice information and determination results are communicated to the [living room] voice input device .

[0125] By sending information only to groups that were basically identified at this time, the advantage of grouping is that only the required voice input devices can exchange information when many voice input devices are connected to the network.

[0126] Accordingly, when the voice input device of [living room] group receives information about voice addressed to its own group, it can determine [play] instructions for [video] within its own group and perform the corresponding processing. It should be noted that the present invention is, of course, applicable to voice input programs.

[0127]

[Effects of the Invention] As explained above, the present invention uses information from other voice input devices for the voice of the user. Thus, it can determine the processing for the voice without burdening the user.

[Brief description of drawings]

19

[Figure 1] A diagram showing the configuration of a voice input system according to one embodiment of the present invention.

[Fig. 2] A diagram showing a voice input device comprised of a voice input system according to one embodiment of the present invention.

[Figure 3] A flow diagram showing the operation of a voice input system according to one embodiment of the present invention.

[Figure 4] A flow diagram showing another operation of a voice input system according to one embodiment of the present invention.

[Figure 5] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 6] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 7] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 8] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 9] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

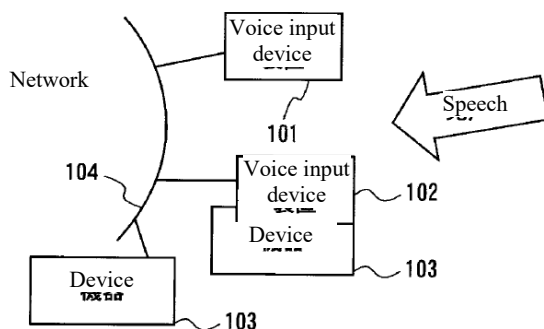
[Figure 10] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 11] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Figure 12] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 13] A diagram showing the concept of linking recognized words, subject devices, and processing content in relation to the voice input system according to one embodiment of the present invention.

[Figure 1]



[Figure 13]

Recognition vocabulary	Target device	Processing details
[Video]	Video	Specify video
[Air conditioner]	Air conditioner	Specify air conditioner
[Power ON]	Shared Instructions	Turn on the target device
[Power OFF]	Shared Instructions	Turn off the target device
[Play]	Video	Play video
[Stop]	Video	Stop playing video
[Turn up temperature]	Air conditioner	Operates by increasing the temperature setting of the air conditioner
[Turn down temperature]	Air conditioner	Operation by lowering the air conditioner temperature setting

20

* [Figure 14] A flow diagram showing another operation of the voice input system according to one embodiment of the present invention.

[Figure 15] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Figure 16] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Figure 17] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Figure 18] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Figure 19] A diagram showing the configuration of another voice input system according to one embodiment of the present invention.

[Explanation of references]

101 . . . Voice input device

102 . . . Voice input device

103 . . . device

104 . . . network

201 . . . microphone

202 . . . signal processing unit

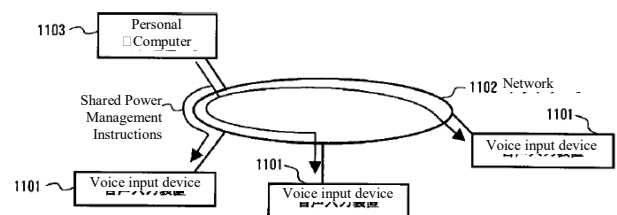
203 . . . central processing unit

204 . . . memory unit

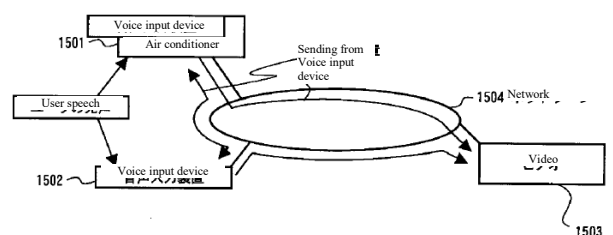
205 . . . network connection unit

206 . . . information presentation unit

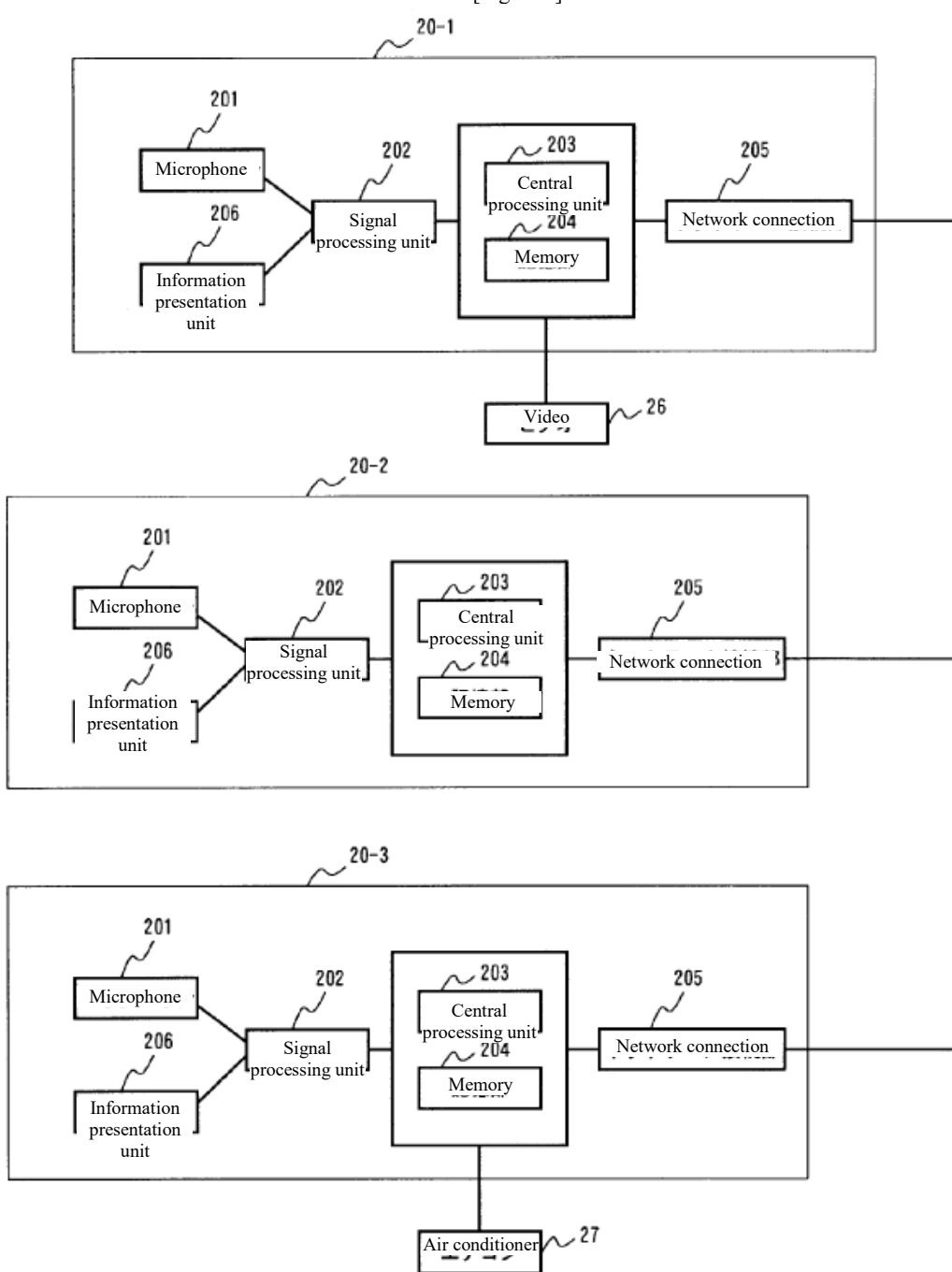
[Figure 11]



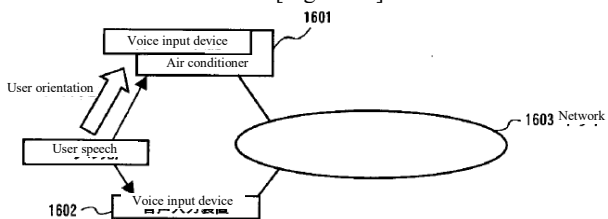
[Figure 15]



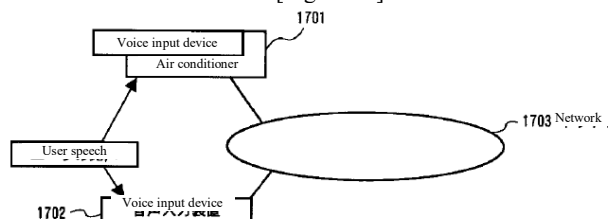
[Figure 2]



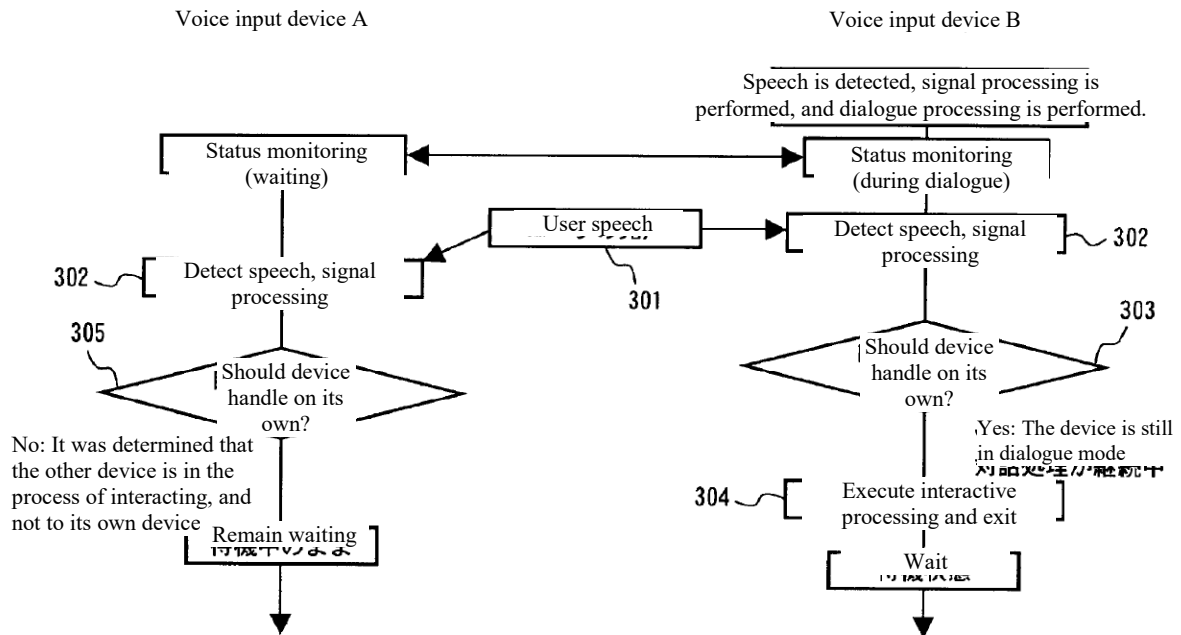
[Figure 16]



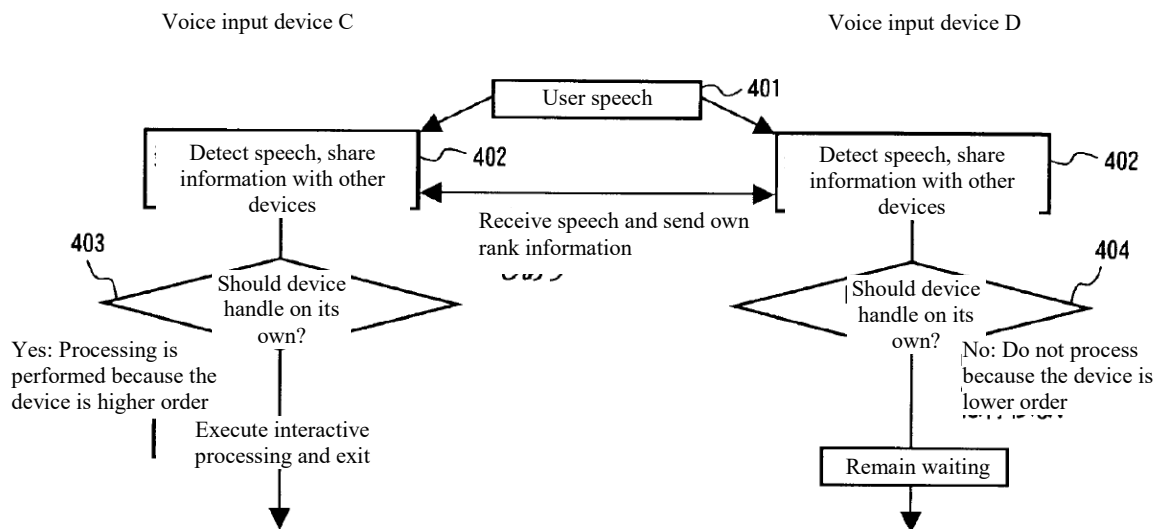
[Figure 17]



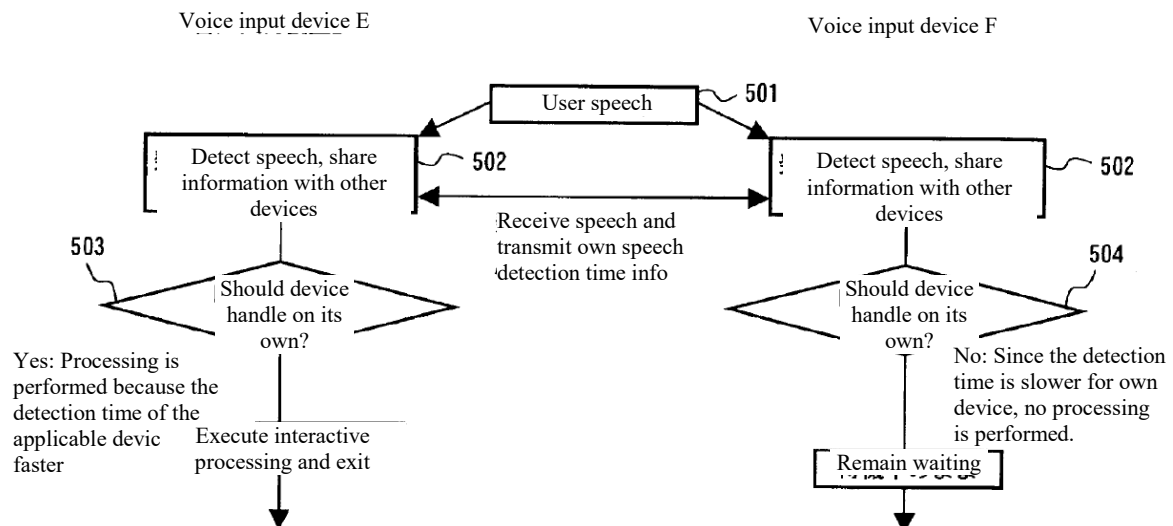
[Figure 3]



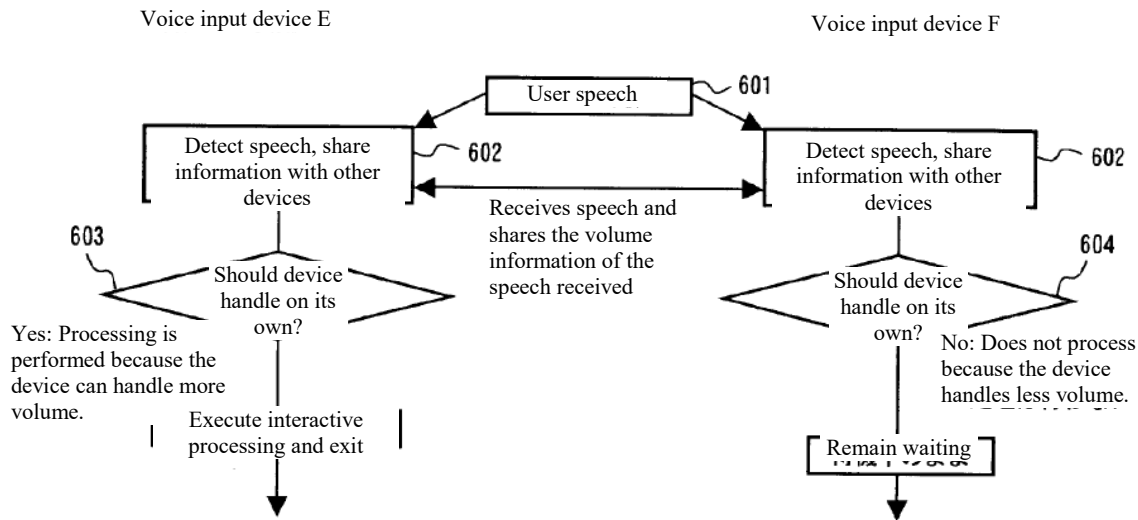
[Figure 4]



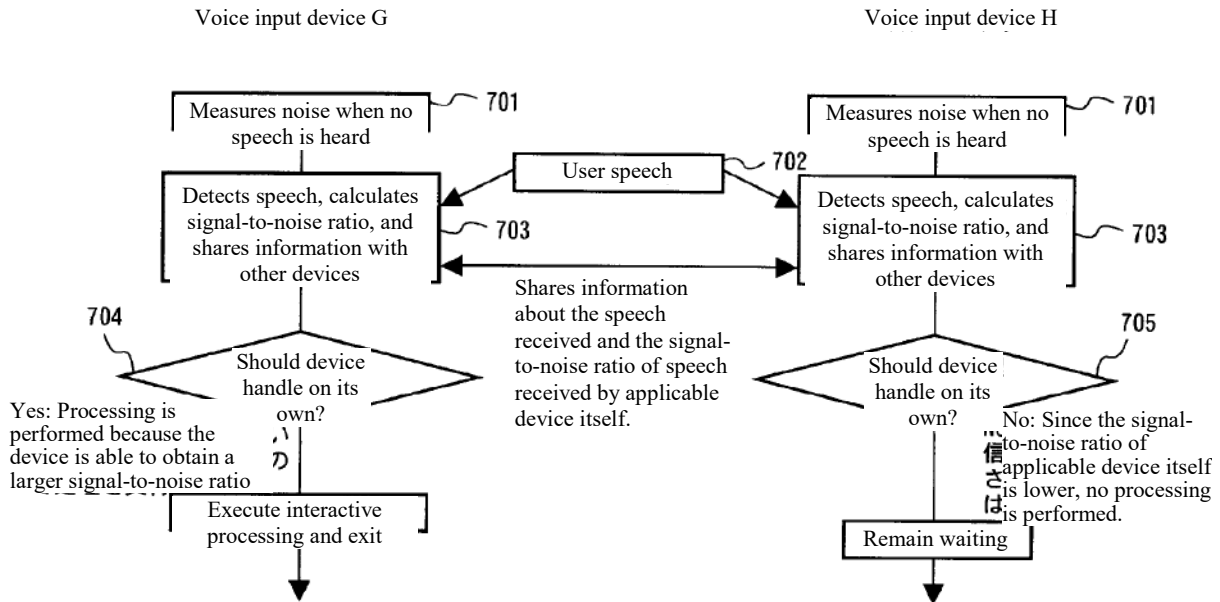
[Figure 5]



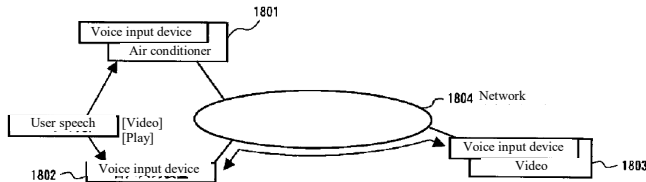
[Figure 6]



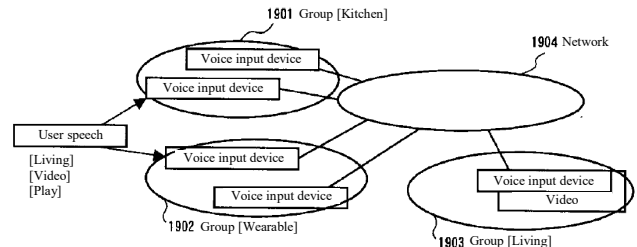
[Figure 7]



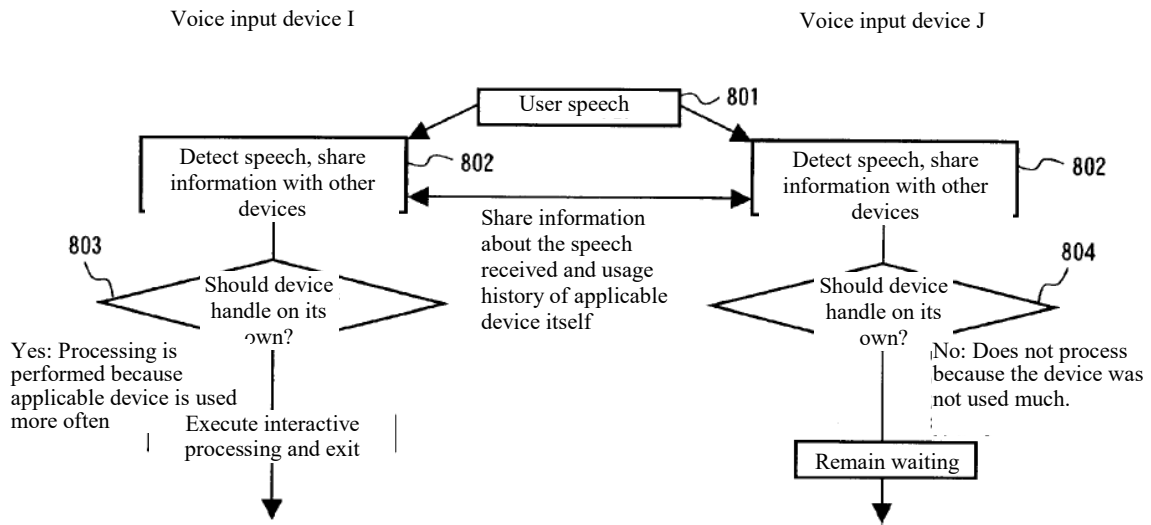
[Figure 18]



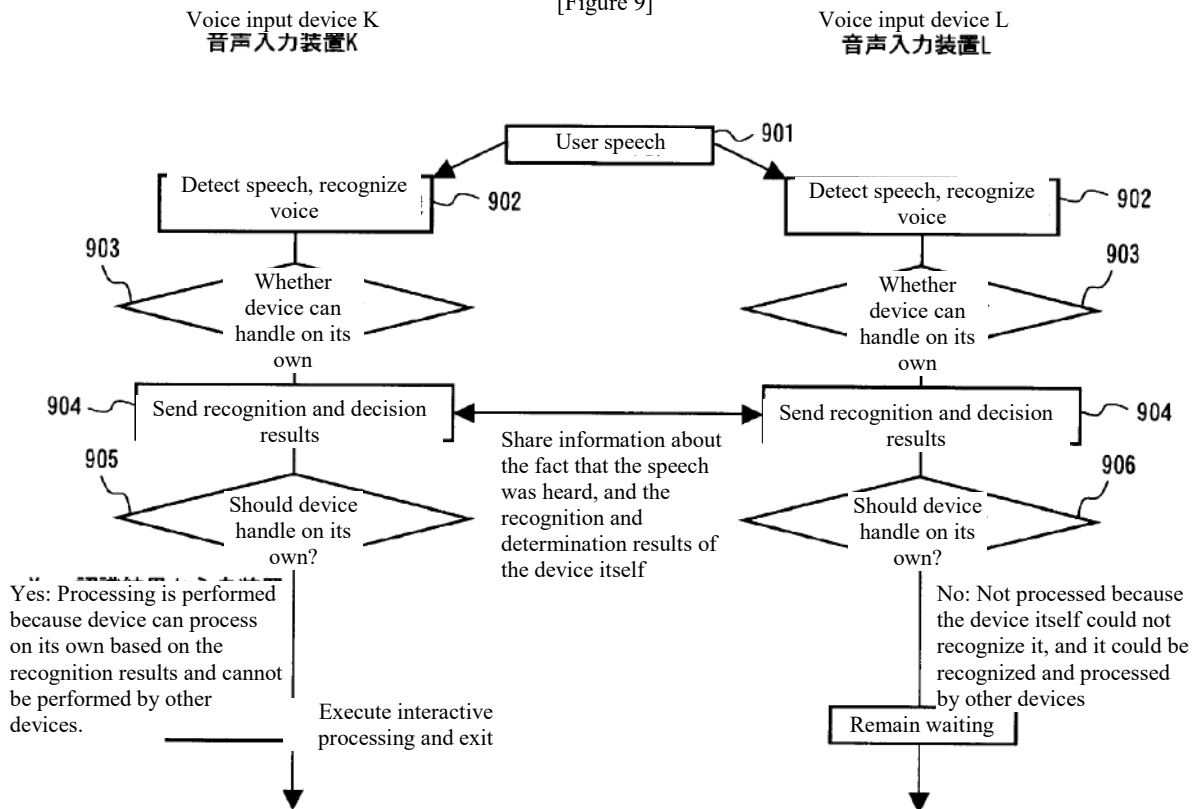
[Figure 19]



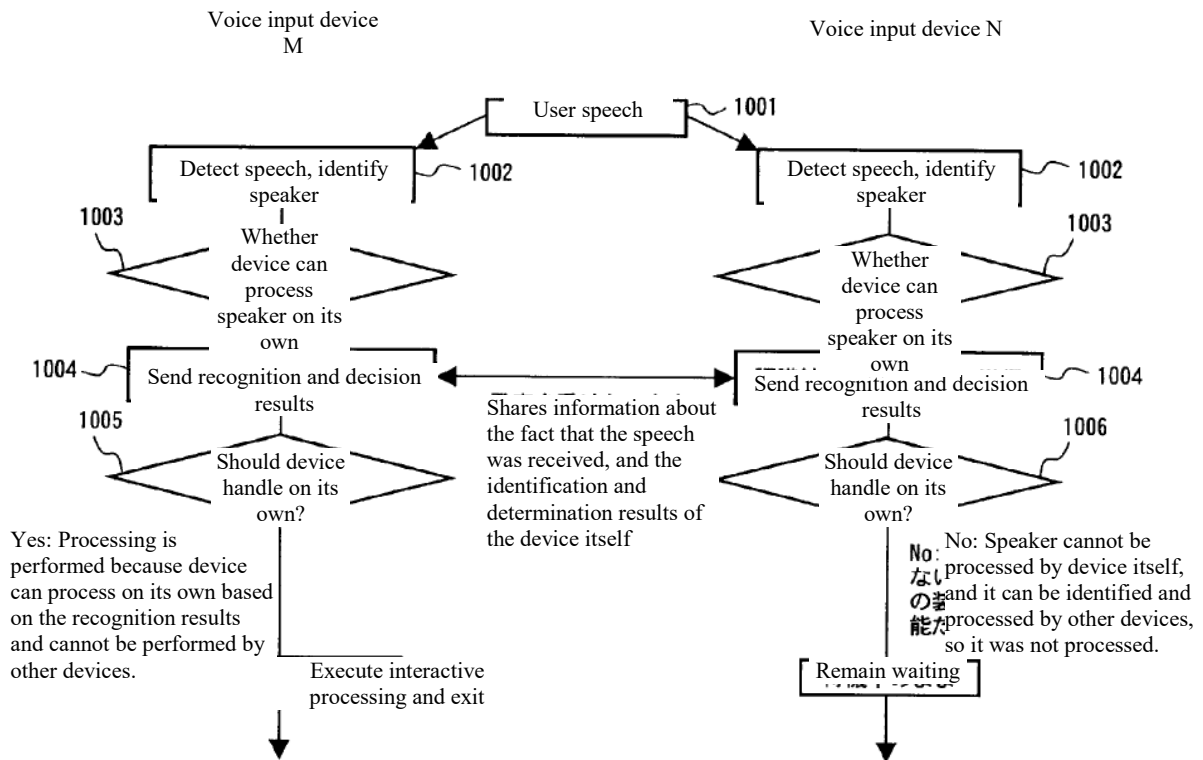
[Figure 8]



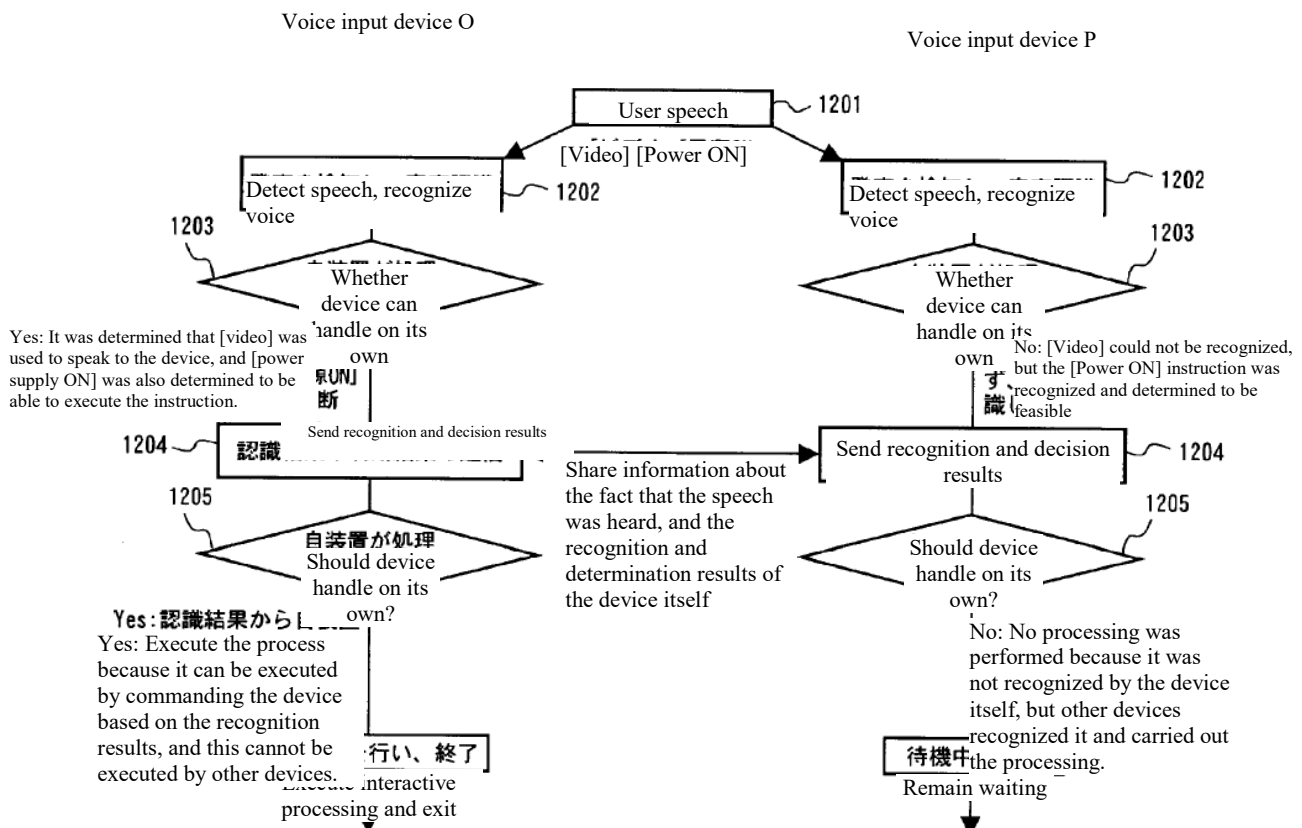
[Figure 9]



[Figure 10]



[Figure 12]



[Figure 14]

