

A METHOD AND SYSTEM FOR VOICE PROCESSING

FIELD OF INVENTION

5 The present invention relates to a method and system for voice processing and specifically relates to a voice processing method, voice processing system, client, and server.

BACKGROUND

10

Service robots are being deployed in an increasing number of applications/locations, such as, for example, airports, banks, museums, schools, hotels, and so forth.

15

Typically, the service robots use voice processing systems to facilitate the provision of services such as, for example, consultation, Q&A, games, and so forth. At present, the service robots use voice processing systems to understand user input content and to provide consequential output. However, the processing time and response speed of the service robots are often slow, and consequently, interaction with the service robots is an undesirable experience.

20

Therefore, improving real-time response durations of voice processing system(s)/method(s) is highly desirable.

SUMMARY

25

In a first aspect, there is provided a voice processing method, characterized in that applied to a voice processing system comprising a client and a first server, wherein the client and the first server can be communicatively connected, and the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit. It is preferable that the method comprises:
30 the voice receiving unit collecting voice information and sending the voice information to the voice processing unit;

the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server;

the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit;

the voice processing unit receiving the first response information , and obtaining the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.

In a second aspect, there is provided a voice processing method characterized in that applied to a client, wherein the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, wherein the client and the first server can be communicatively connected. Preferably, the method comprises:

the voice receiving unit collecting voice information and sending the voice information to the voice processing unit;

the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server; wherein, the first server is used to determine the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit;

the voice processing unit receiving the first response information , and obtaining the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.

In a third aspect, there is provided a voice processing method characterized in that applied to a first server, wherein the first server and the client can be communicatively connected, the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit. It is preferable that the method comprises:

the first server receiving text information sent by the voice processing unit, and the text information is generated based on the voice information sent by the voice receiving unit;

the first server determining the category corresponding to the text information, executes processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit to generate the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.

In a fourth aspect, there is provided a voice processing system, characterized in that the voice processing system comprises a client and a first server, wherein the client and the first server can be communicatively connected, the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit. Preferably,

the voice receiving unit is configured to collect voice information and send the voice information to the voice processing unit;

the voice processing unit is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to the first server;

the first server is configured to determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit, the first response information at least includes the text response information corresponding to the text information;

the voice processing unit is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

In a fifth aspect, there is provided a client, characterized in that the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, the client and the first server can be communicatively connected. Preferably, the voice receiving unit is configured to collect voice information and send the voice information to the voice processing unit;

the voice processing unit is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to the first server;

the voice processing unit is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

In a final aspect, there is provided a first server, characterized in that the first server and the client can be communicatively connected, the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit. It is preferable that the first server is configured as:

receives text information sent by the voice processing unit, and the text information is generated based on the voice information sent by the voice receiving unit;

determines the category corresponding to the text information, executes processing logic corresponding to the category, obtains the first response information of the text information corresponding to the category, and sends the first response information to the voice processing unit to generate the corresponding second response information

based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

It will be appreciated that the broad forms of the invention and their respective features can be used in conjunction, interchangeably and/or independently, and reference to separate broad forms is not intended to be limiting.

DESCRIPTION OF FIGURES

By reading the detailed description of exemplary embodiments in the following text, those skilled in the art will understand the advantages and benefits described herein, as well as other advantages and benefits. The accompanying drawings are only for the purpose of demonstrating exemplary embodiments and are not considered a limitation on the application. And throughout all drawings, the same components are represented by the same numbers. In the attached drawings:

FIG 1 is an embodiment of a framework diagram of a voice processing system of the application;

FIG 2 is a flowchart of an embodiment of a voice processing method of the application;

FIG 3 is a flowchart of another embodiment of a voice processing method of the application;

FIG 4 is a schematic diagram of an embodiment of a voice processing system of the application;

FIG 5 is a signaling diagram of an embodiment of a voice processing method of the application;

FIG 6 is a flowchart of an embodiment of another voice processing method of the application;

FIG 7 is a flowchart of an embodiment of another voice processing method of the application;

FIG 8 is a schematic diagram of an embodiment of another voice processing system of the application;

FIG 9 is a schematic diagram of an embodiment of a client provided of the application;
and

FIG 10 is a schematic diagram of an embodiment of an illustrative server of the
application.

5

DETAILED DESCRIPTION

Exemplary embodiments of the present application will be described in more detail
below with reference to all accompanying drawings. Although the exemplary
10 embodiments of the present application are shown in the accompanying drawings, it
should be understood that the present application can be implemented in various
forms and should not be limited by the embodiments described here. On the contrary,
these embodiments are provided to enable a more thorough understanding of the
present application and to fully convey the scope of the present application to those
15 skilled in the art.

In the description of the embodiments of the application, it should be understood that
terms such as "including" or "having" are intended to indicate the presence of
disclosed features, numbers, steps, actions, components, parts, or combinations
20 thereof in this specification, and do not exclude the possibility of the existence of one
or more other features, numbers, steps, actions, components, parts, or combinations
thereof.

Unless otherwise specified, "/" represents the meaning of "or". For example, A/B can
25 represent A or B; "and/or" in this document is just a way to describe the relationship
between associated objects, indicating that there can be three types of relationships.
For example, A and/or B can represent three situations: the existence of A alone, the
simultaneous existence of A and B, and the existence of B alone.

30 Terms such as "first", "second", and the like are used to distinguish between identical
or similar technical features for descriptive convenience only and should not be
interpreted as indicating or implying the relative importance or quantity of these
technical features. Thus, features defined by "first", "second", etc., can explicitly or

implicitly include one or more of these features. In the description of the embodiments of the application, unless otherwise specified, the term "multiple" means two or more.

It should also be noted that, without conflict, the embodiments and features in the
5 embodiments in the application can be combined with each other. The following will
refer to the accompanying drawings and combine embodiments to illustrate the
application in detail.

The applicant found that the current service robots can use the voice processing
10 system to understand the user's input content and provide corresponding feedback.
The solution is: the client receives the user's input voice content, the server recognizes
and understands the voice content, generates the corresponding voice reply, and then
the client plays the voice reply. The applicant considers that, currently, the processing
and response speed is slow, and the interaction efficiency is low, which affects the
15 user's interactive experience.

To improve the response and processing methods of the voice processing system,
certain embodiments of a voice processing method are applied to the voice processing
system. FIG 1 is an architecture diagram of a voice processing system related to an
20 embodiment of the application.

As shown in FIG 1, a voice processing system 98 provided in an embodiment of the
application includes a client 100 and a first server 200. Wherein, the client 100 includes
a voice processing unit 101, a voice receiving unit 102, a voice playback unit 103, and
25 a display unit 104. The voice processing unit 101 is configured to process data
collected by the voice receiving unit 102, and control, respectively the playback and
display of the voice playback unit 103 and the display unit 104. The voice processing
unit 101 can also communicate with the first server 200, and use the first server 200
to process the data collected by the voice receiving unit 102.

30 In other embodiments, the voice processing system also includes a camera (image
capturing) device 105 (not shown in FIG 1, shown in FIG 4), and the voice processing
unit 101 is configured to process the data collected by the voice receiving unit 102 and
the camera device 105, respectively control the playback and display of the voice

playback unit 103 and the display unit 104, and also use the first server 200 to process the data processed by the voice processing unit 101.

In some examples, the voice receiving unit 102 is a receiver or pickup, such as, for example, a microphone; the voice playback unit 103 is a speaker, such as, for example, a broadcaster; the display unit 104 is a display screen, such as, for example, a liquid crystal display screen (LCD), organic light-emitting diode display screen (OLED), light-emitting diode display screen (LED), plasma display screen (PDP), electronic paper display screen (E-Ink), touchable display screen, curved display screen, flexible display screen, 3D display screen, transparent display screen, etc; the camera device 105 is a camera.

The first server 200 is configured to classify the text information sent by the client 100, determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain a first response information of the text information corresponding to the category, and send the first response information to the client 100. In an embodiment, the first server 200 is configured to classify the text information sent by the voice processing unit 101, determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit 101. In some embodiments shown in FIG 1, the first server 200 includes a text processing module 201 and a language reply module 202. The text processing module 201 is configured to classify the text information sent by the client 100 and execute corresponding processing logic based on the classified categories. The text processing module 201 can also be configured with a communication protocol with the client 100, and the first server 200 can use the text processing module 201 to communicate with the client 100. The language reply module 202 stores or connects language models. The language response module 202 is configured to understand the semantic meaning of text information through a language model and produce text response information. In some examples, the text processing module 201 and the language reply module 202 are located on the same server as the first server 200; in some other examples, the text processing module 201 and the language reply module

202 are located on or distributed on different servers from the first server 200, and the embodiment of the application are not limited.

The above content introduces the architecture diagram of a voice processing system provided in some embodiments of the application, which is applied to the voice processing system of the aforementioned embodiments. The following is an introduction to the voice processing method of the embodiments of the application.

Referring to FIG 2, there is shown an embodiment of a voice processing method 2000, comprising at least steps 2001-2004. The method 2000 can be implemented with the system 98 but can also be implemented using other set-ups using physical/virtual hardware.

At step 2001, the voice receiving unit 102 is configured to collect voice information and sends it to the voice processing unit 101.

In some examples, the user or device outputs voice information, and the voice receiving unit 102 collects the voice information and sends it to the voice processing unit 101. Wherein, the device has voice playback function, such as smartphones, robots, etc.

At step 2002, the voice processing unit 101 is configured to receive voice information, obtains the text information corresponding to the voice information, and sends the text information to the first server 200.

It should be noted that the text information corresponding to the obtained voice information in an embodiment of the application can be completed by the client 100 or the server 200, and there is no limitation in the embodiment. In an embodiment of the application, the voice processing unit 101 is configured to analyze the voice information to obtain the corresponding text information. It should be noted that the process of analyzing the voice information in the embodiment of the application is the process of decoding the voice information. As another possible implementation, the voice processing unit 101 can also be configured to upload voice information to a second server (not shown in the figure), which is used to analyze the voice information

and obtain the corresponding text content and language identifier. In an example, the voice to text module of the second server is configured to decode the voice information, obtain the text content and language identifier corresponding to the voice information, and return the text content and language identifier to the voice processing unit 101.

- 5 The voice processing unit 101 is configured to obtain the text content and language identifier returned by the second server, which is the text information.

In a practical operation, the voice processing unit 101 is configured to obtain the text content corresponding to the voice information and a language identifier when
10 decoding the voice information, which the language identifier indicates the language type of the voice information. In other words, the text information in the application also includes language identifier. In an example, the client 100 of the application can be configured to recognize languages including Chinese, English, and Arabic. Therefore, the language identifier recognized by the voice processing unit 101
15 includes language identifier A corresponding to English, language identifier B corresponding to Chinese, and language identifier C corresponding to Arabic. For example, when the language corresponding to the collected voice information is English, the text information obtained by the voice processing unit 101 includes the language identifier A corresponding to English.

20

As shown in FIG 3, after obtaining the text information corresponding to the voice information, the client 100 is configured to send the text information to the first server 200, so that the first server 200 can proceed with the next step of processing the text information. In practical operation, the client 100 may also include a camera device
25 105. The camera device 105 is configured to collect image information of the voice object, and the image information of the voice object is used to assist the voice processing unit 101 in processing the voice information. In an embodiment of the application, the voice object refers to the object that emits voice information, such as users, devices, etc.

30

In an example, the voice processing unit 101 can process the voice information collected by the voice receiving unit 102 based on the image information of the voice object to obtain key voice information; the voice processing unit 101 being configured to obtain text information corresponding to the key voice information based on the key

voice information. In this way, the voice processing unit 101 can extract the key voice information from the voice information collected by the voice receiving unit 102 and obtain the text information of the key voice information, without decoding all the collected voice information, correspondingly reducing computing power consumed by decoding the originally collected voice information and improving decoding efficiency.

At step 2003, the first server 200 is configured to determine the category corresponding to the text information, and executes processing logic corresponding to the category, then obtaining the first response information of the text information corresponding to the category, and sends the first response information to the voice processing unit 101.

In an embodiment of the application, the first server 200 may include a text processing module 201. The text processing module 201 is configured to obtain classification results based on the text content attributes of the text information, which is the category corresponding to the text information. In an embodiment of the application, the attributes of the text content may include Q&A attributes or instruction attributes, and the categories corresponding to the text information include Q&A or instruction. When receiving the text information of the instruction attribute, the client 100 then is configured to execute the action command corresponding to the text information. For example, if the text content of the text information is "dance", then the text content attribute is the instruction attribute, and the category corresponding to the text information is the instruction. The client 100 is configured to execute the action command "dance". When receiving text information with Q&A attributes, the client 100 needs to answer the user's question. For example, if the text content of the text message is "Where is the restroom?", then the text content attribute is Q&A, and the corresponding category of the text message is Q&A, the client is configured to answer the question "Where is the restroom?". In practical operation, the text processing module 201 can filter the text information or classify the attributes of text information through keyword recognition, relatively simple language models, or other methods. The embodiments of the application are not limited herein.

In some embodiments, the text processing module 201 is configured to screen the command keywords from the text information through keyword recognition, a relatively

simple language model, or other methods. When a command keyword is matched, the text processing module 201 returns a matching action response command to the client, or any combination of the action response command and the following content: text response information and image response information corresponding to the language identifier. When no command keyword is matched, the text information is sent to the language sub model corresponding to the language type of the text information in the language model. The language sub model understands the text information and, when determining that the text information is command information, returns a matching action response command, or any combination of the action response command and the following content: text response information and image response information corresponding to the language identifier, and returns the above response information (such as the action response command, or any combination of the action response command and text response information/image response information) to the client 100. When the language sub model determines that the text information is not command information, such as Q&A or query information, it returns text response information of the same language type as the text information. It is possible to use the text processing module 201 to quickly identify the command keywords in the text information which are easily recognizable to quickly respond with matching action response commands, text response information, and image response information. When the command keywords in the text information are relatively vague and difficult to recognize, such as when the command information indicated by the text information needs to be understood in conjunction with context or scenario information, i.e., there are no command keywords in the text processing module 201 that match the text information, but the content of the text information contains relevant command information, the text processing module 201 is configured to send the text information to the language sub model corresponding to the language type of the text information in the language model. The language sub model is then configured to recognize whether the text information contains command information based on the content of the text information, thereby achieving accurate classification of the text information. For example, the text processing module 201 stores the command keywords: Chinese "have a dance" and English "sing a song", where the action response command corresponding to "have a dance" is "dance", and the action response command corresponding to "sing a song" is "sing". When the text content of the text information is "let's have a dance", it is determined that the Chinese command keyword "have a dance" is matched, and the

matching action response command is "dance". The text processing module 201 is configured to feedback the action response command "dance" to the voice processing unit 101 of the client 100, and the voice processing unit 101 is then configured to drive the virtual object displayed on the display screen to perform the action of dancing.

5 When the text content of the text information is "let's dance to the song", there is the command information "dance" in this text information, but the Chinese command keyword "have a dance" cannot be matched. At this time, the text processing module 201 is configured to send the text information to the Chinese sub model in the language model. The Chinese sub model is configured to recognize that there is the
10 command information "dance" in the text information, so the Chinese sub model is configured to feedback the action response command "dance" to the voice processing unit 101 of the client 100 through the text processing module 201, or the Chinese sub model is configured to feedback the action response command "dance" to the voice processing unit 101 of the client 100. The voice processing unit 101 drives the virtual
15 object displayed on the display screen to perform the action associated with "let's dance to the song". Another example, when the text content of the text information is "sing a song", it is determined that the English command "sing" is matched, and the matching action response command is "sing". The text processing module 201 is configured to feedback the action response command "sing" to the voice processing
20 unit 101 of the client 100, and the voice processing unit 101 is configured to drive the voice playback unit 103 to play a song.

In some other embodiments, the text processing module 201 is configured to store a series of instruction keywords. When the text information hits this series of instruction
25 keywords, it is determined that the text content attribute of the text information is an instruction attribute, that is, the category of the text information is an instruction. The text processing module 201 is configured to determine the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language
30 identifier, image response information, and returns the above response information to the client 100 (such as action response instruction, or any combination of action response instruction and text response information/image response information). When the text information misses the instruction keywords of the series, it is configured to determined that the text content attribute of the text information is a Q&A

attribute, that is, the text information category is Q&A. The text processing module 201 is configured to input the text information to the language model to obtain the text response information corresponding to the language identifier.

5 In an example, the text processing module 201 is configured to store instruction keywords for various languages, which correspond to language identifiers and set or store action response instructions accordingly. When the text information hits an instruction keyword identified by a language, it is determined that the text information is an instruction attribute. Then, the action response instruction corresponding to the
10 matching instruction keyword is determined, and the action response instruction is returned to the client 100 to execute the action response instruction. For example, the text processing module 201 is configured to store instruction keywords: Chinese "have a dance" and English "sing a song". Wherein, the action response instruction corresponding to "have a dance" is "dance", and the action response instruction
15 corresponding to "sing a song" is "sing". When the text content of the text information is "have a dance", it is determined that the Chinese instruction keyword "have a dance" has been hit. At this point, the matched action response instruction is "dance". The text processing module 201 is configured to feedback the action response instruction "dance" to the voice processing unit 101 of the client 100, which drives the virtual
20 object displayed on the display screen to perform the dance action. For example, when the text content of the text information is "sing a song", it is determined that the English instruction "sing" has been hit, and the matched action response instruction is "sing". The text processing module 201 is configured to feedback the action response instruction "sing" to the voice processing unit 101 of the client 100, which drives the
25 voice playback unit 103 to play the song.

In another example, the text processing module 201 is configured to store instruction keywords for various languages, which correspond to language identifiers and set or store action response instructions accordingly. When the text information fails to hit
30 the instruction keyword of any language identifier, the text processing module 201 is configured to input the text information to the language model to obtain the text response information corresponding to the language identifier. In most cases, large language models that support multiple languages at the same time have a larger volume and slower understanding and response speed to textual information. For

example, the parameter scale of large language models that currently support multiple languages is usually in the billions, with a very large volume. In order to improve the response speed of the language model as much as possible, an embodiment of the application divides the language model into one or more relatively independent language sub models based on the language type. The language sub model can support one language or a small number of languages, and the parameter quantity of the language sub model is smaller than that of the large language model, which can be in the billions. Compared with the single large model that supports multiple languages simultaneously, the language sub model that supports one or a small number of languages in an embodiment of the application has a smaller volume and a faster response speed to text information. In some specific implementations, the first server 200 includes or connects a language model, which includes one or more language sub models, each corresponding to one or more language types. Each language sub model is used to understand the input text content and output the corresponding text response content in a language with the same text content. When the text information fails to hit the instruction keyword of any language identifier, the text processing module 201 is configured to input the text information to the target language sub model corresponding to the language identifier of the text information to obtain the text response information returned by the target language sub model. The language identifier of the text response information is the language identifier of the text information. For example, the text processing module 201 is configured to store instruction keywords such as "have a dance" in Chinese and "sing a song" in English. When the text content of the text message is "where is the restroom", the instruction keywords by each language identifier are not hit. The text processing module 201 is configured to send text information to the target language sub model corresponding to the language identifier "Chinese" (specifically, the text content of the text message "where is the restroom"), and the target language sub model understands and generates a Chinese text response message "the restroom is on your left", which is sent to the text processing module 201.

In an embodiment of the application, the first server 200 is configured to classify and process text information of different categories based on the classification results, execute different processing logic, and obtain the first response information corresponding to different categories. The text information of different categories in an

embodiment of the application includes text information of instruction attributes and text information of Q&A attributes. The first server 200 is configured to perform different categories of logical processing on the text information of instruction attributes and Q&A attributes based on the text content attributes of the text information, and obtain the first response information corresponding to different categories of text information. By classifying text information before processing, the first server can better recognize and process text information in a targeted manner. For different categories of text information, it can selectively generate the first response information corresponding to the category by itself or the language model, in order to obtain the second response information corresponding to different categories. Specifically, when the category of text information is instruction, an embodiment of the application matches and returns action response instructions based on its own stored information. When the category of text information is Q&A, the text response information is returned through a language model. Due to the fact that the first response information corresponding to the generated instruction category is based on self stored information matching, without the necessary for semantic analysis, semantic understanding, text generation, and other operations, the speed of generating the first response information corresponding to the instruction category is fast, which can accelerate the response speed of the voice processing system to customers while maintaining the quality of the voice processing method in the application. In addition, generating the first response information corresponding to the Q&A category does not require the text processing module of the first server, but is processed by a language model with stronger voice processing capabilities, which further accelerates the response speed of the voice processing system to customers.

At step 2004, the voice processing unit 101 is configured to receive the first response information and is configured to obtain the second response information corresponding to the first response information. The second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

After receiving the first response information, the voice processing unit 101 is configured to obtain the voice response information corresponding to the first response information and controls the voice playback unit 103 to play the voice response information. In some embodiments, the first response information includes any combination of text response information, image response information, and action response instructions. In an embodiment, the first response information includes at least text response information, and the voice processing unit 101 is configured to generate the corresponding voice response information for the text response information. In other embodiment, the voice processing unit 101 may also be configured to control the display unit 104 to display the first response information, such as displaying text response information. In another embodiment, the voice processing unit 101 is configured to control the voice playback unit 103 to play the voice response information after generating the voice response information, and is configured to control the display unit 104 to display the text response information of the first response information. In another embodiment, the voice processing unit 101 is configured to control the display unit 104 to display the first response information after receiving it.

In some embodiments, the first response information of the voice processing unit 101 may contain different content and obtain different second response information. When the first response information includes text response information, the voice processing unit 101 is configured to generate a voice response information corresponding to the text response information, control the voice playback unit 103 to play the voice response information, and control the display unit 104 to display the text response information. When the first response information also includes action response instructions and image response information, the voice processing unit 101 is configured to also control the client to execute the action indicated by the action response instruction and control the display screen to display the image response information.

In an example, the display screen displays a virtual object, and the first response information includes text response information, image response information, and action response instructions. The voice processing unit 101 is configured to control the display unit 104 to display text response information in the text display area, image

response information in the image display area, and control the virtual object to execute the action corresponding to the action response instruction. In specific implementation, the voice processing unit 101 is configured to also invoke the display content corresponding to the action response instruction, and play the display content through the display unit 104. It should be noted that the voice processing unit 101 is configured to store the mapping relationship between action response instructions and audio/video/image/function, as well as the mapping relationship between action response instructions and display content. After receiving the action response instruction, the voice processing unit 101 is configured to call the audio/video/image/function and display content corresponding to the action response instruction for playback. For example, the action response instruction A corresponds to audio A, voice playback unit, video A, and video player. When the voice processing unit 101 receives the action response instruction A, it calls the voice playback unit to play the corresponding audio A, and calls the video player to play the corresponding video A.

The existing technology transmits voice information from the client to the server for decoding. The applicant believes that compared with the existing technology, the text information corresponding to the voice information obtained in embodiment of the application is completed by the client or a non first server part, saving resources on the first server and improving the processing speed of the first server. Moreover, an embodiment of the application transmits text information between the client and the first server instead of voice information with a large amount of data, saving network resources and improving the transmission speed and overall voice processing efficiency of the information.

In an embodiment of the application, the first server 200 is configured to classify and process text information of different categories based on the classification results, execute different processing logic, and obtain the first response information corresponding to different categories. The text information of different categories in an embodiment of the application includes text information of instruction attributes and text information of Q&A attributes. The first server can perform different categories of logical processing on the text information of instruction attributes and Q&A attributes based on the text content attributes of the text information, and obtain the first

response information corresponding to different categories of text information. By classifying the text information before processing, the first server can better recognize and process the text information in a targeted manner, in order to obtain the first response information corresponding to different categories. Compared with the fact that all response information is generated by the server in related technologies, the second response information of different categories in the embodiments of the application is generated by the client, reducing the process of the first server 200 and thus accelerating the response speed of the voice processing system to customers while maintaining the quality of the voice processing method in the application.

In practical operation, an embodiment of the application matches and returns action response instructions based on its own stored information when the category of text information is instruction. When the category of text information is Q&A, it returns text response information through a language model. Due to the fact that the first response information corresponding to the generated instruction category is based on self stored information matching, without the need for semantic analysis, semantic understanding, text generation, and other operations, the speed of generating the first response information corresponding to the instruction category is fast, which can accelerate the response speed of the voice processing system to customers while maintaining the quality of the voice processing method in the application. In addition, generating the first response information corresponding to the Q&A category does not require the text processing module of the first server, but is processed by a language model with stronger voice processing capabilities, which further accelerates the response speed of the voice processing system to customers.

It should be noted that in order to improve the response speed of the voice processing method in the application, the text processing module usually classifies text information through keyword matching, relatively simple language classification, or other faster classification methods. Therefore, while ensuring speed, the text processing module may recognize some text information that belongs to instruction categories as problem categories. Although the language model in the application has a slower response speed compared to the text processing module, it has better comprehension and judgment abilities for text information. Therefore, in some possible embodiments, after being classified by the text processing module first, the received

text information will be further classified by a language model, thereby increasing the accuracy of text information classification while maintaining response speed. Specifically, the language response module 202 is configured to utilize a language model to semantically understand text information and obtain comprehension results.

5 If the comprehension result indicates that the category of the text information is instruction, the language response module 202 is configured to send the action response instruction corresponding to the text information to the voice processing unit 101 through the text processing module 201. If the comprehension result indicates that the category of text information is Q&A, the language model generates text
10 response information based on the text information. The language response module 202 is configured to send the text response information to the voice processing unit 101 through the text processing module 201.

In order to provide a more comprehensive understanding of the scheme provided in
15 the embodiments of the application, reference is made to FIGs 4 and 5. FIG 4 shows the architecture diagram of the voice processing system 4000 provided in some embodiments, and FIG 5 shows an interactive schematic diagram based on the voice processing method provided in FIG 4.

20 As shown in FIG 4, the voice processing system 4000 includes a second client 400 and a second server 500. Wherein, the second client 400 includes a voice processing unit 401, a voice receiving unit 402, a voice playback unit 403, a display unit 404, and a camera device 405. Wherein, the voice processing unit 401 is equipped with a virtual human interaction device, which displays a virtual human object through the display
25 unit 404. Users can interact with the virtual object through Q&A, interaction, and other interactive operations. It should be understood that the virtual human interaction device can also be installed in other units, for example, the client is also provided with a processing unit to install the virtual interaction device, and the voice processing unit 401 is configured to send control information to the processing unit to control or drive
30 the virtual object to execute actions, and display the process of the virtual object executing actions on the display unit 404. In the architecture diagram shown in FIG 4, the voice receiving unit 402 takes the pickup as an example, the voice playback unit 403 takes the speaker as an example, the display unit 404 takes the liquid crystal display screen (LCD) as an example, and the camera device 405 takes the camera as

an example. The second server 500 includes a text processing module 501 and a language response module 502. The language response module 502 is connected to a language model, which includes three language sub models: language sub model A, language sub model B, and language sub model C. Wherein, language sub model A is used for text information in English, language sub model B is used for text information in Japanese, and language sub model C is used for text information in Arabic. Correspondingly, the language identifier in an embodiment of the application includes language identifier A, language identifier B, and language identifier C, where language identifier A indicates English, language identifier B indicates Japanese, and language identifier C indicates Arabic.

Referring to FIG 5, a voice object (such as the user) outputs voice in English, and the voice receiving unit 402 collects the voice information output by the voice object (such as the user) and sends the voice information to the voice processing unit 401. Camera 405 can also capture image information of voice objects (such as users) and send the image information of voice objects to voice processing unit 401. For the convenience of understanding, the following will take the voice object as an example for explanation.

The voice processing unit 401 is configured to intercept the user's output voice information based on the image information of the voice object, obtains key voice information, and decodes the key voice information to obtain the text information corresponding to the key voice information. The text information includes the text content corresponding to the voice information and the language identifier A, which represents the language type (i.e. English) used by the user to output the voice information. As an example, the voice processing unit 401 can determine that there is an interaction need between the user and the client when the image information of the voice object indicates that the user's eyes are fixed on the client, or when the user's face is facing the client, intercept the voice information output by the user At this point, and obtain key voice information. When the user is not looking at the client or their face is not facing the client, it is determined that there is no need for interaction between the user and the client. At this point, the voice information output by the user is non critical and can be ignored for subsequent recognition. By recognizing key voice information, it is possible to obtain and recognize the voice information required for user interaction, without the need to decode the input voice information at all times,

greatly saving the resources of the voice processing unit 401 and improving the recognition rate of key voice information during user interaction; and it also avoids processing the text information of voice information input at all times by the second server 500, improving the efficiency of the second server's 500 use.

5

The voice processing unit 401 is configured to send text information to the text processing module 501.

10

The text processing module 501 is configured to classify the text information and determines the category corresponding to the text information, that is, determines the category to which the text information belongs. The text processing module 501 is configured to execute processing logic corresponding to the category, obtains the first response information of the text information corresponding to the category, and sends the first response information to the voice processing unit 401. In an example, the first response information includes at least the text response information corresponding to the text information. In another example, the first response information includes any combination of text response information, image response information, and action response instructions.

15

20

The process of classifying text information in text processing module 501 is as follows: there is a series of instruction keywords stored in text processing module 501. When the text information hits this series of instruction keywords, the text content attribute of the text information is determined as the instruction attribute, that is, the category of the text information is the instruction; when the text information fails to hit the instruction keyword identified by any language, it is determined that the text content attribute of the text information is a Q&A attribute, that is, the category of the text information is a Q&A attribute. In other examples, while ensuring speed, the text processing module 501 may recognize some text information that belongs to the instruction category as a problem category. Therefore, in some possible embodiments, after the text processing module 501 is configured to perform a classification process, the received text information will be further classified through a language model, thereby increasing the accuracy of text information classification while maintaining response speed. Specifically, the language response module 502 is configured to utilize a language model to semantically understand text information and obtain

25

30

comprehension results. If the comprehension result indicates that the category of the text information is instruction, the language response module 502 is configured to send the action response instruction corresponding to the text information to the voice processing unit 401 through the text processing module 501. If the comprehension result indicates that the category of text information is Q&A, the language model generates text response information based on the text information. The language response module 502 is configured to send the text response information to the voice processing unit 401 through the text processing module 501.

10 The process of executing the processing logic corresponding to the category in the text processing module 501 is as follows (1) - (2).

(1) If the classification result indicates that the category corresponding to the text information is instruction, the text processing module 501 is configured to send the action response instruction corresponding to the text information to the voice processing unit 401, or the text processing module 501 sends the action response instruction corresponding to the text information and any combination of the following content to the voice processing unit 401: text response information, image response information. In this example, the action response instruction also includes a language identifier A to display content in the language (i.e. English) represented by the language identifier when the display unit 404 is configured to display content. For example, the voice processing unit 401 is configured to control the voice playback unit 403 to play the audio corresponding to the action response instruction, and controls the display unit 404 to display the display content corresponding to the action response instruction represented by the language identifier. Wherein, the language used for the audio corresponding to the action response instruction is English, which is represented by the language identifier. The image of the virtual object in the displayed content can match the language identifier, for example, the appearance of the virtual object represents the appearance of the user object represented by the language identifier, and the clothing of the virtual object represents the clothing of the user object represented by the language identifier. In this example, the display content can also include a language element that identifies the language type, that is, the language element is a display element that identifies "English" as the language type. For

example, a language element that displays the language type in a preset area of the display unit is an image or text "English".

(2) If the classification result indicates that the category corresponding to the text information is Q&A, the text processing module 501 is configured to send the text information to the language response module 502. Language response module 502 selects the target language sub model corresponding to language identifier A from one or more language sub models in the language model based on language identifier A in the text information, and inputs the text content from the text information into the target language sub model. The target language sub model is used to understand and analyze text content in English, generate English text response information, and return the text response information to the language response module 502. In other examples, the target language sub model will also return the language identifier A to the language response module 502 to determine whether the language of the text response information is accurate.

The language response module 502 is configured to send the text response information and language identifier A to the voice processing unit 401 through the text processing module 501. The language response module 502 is configured to send the language identifier A to the voice processing unit 401, so that the voice processing unit 401 is configured to quickly generate corresponding language voice response information based on the language identifier. It should be understood that in other examples, language identifier A may not be sent to voice processing unit 401, which can generate corresponding language voice response information based on the textual language of the text response information. The following is an example of using language response module 502 being configured to send language identifier.

The voice processing unit 401 is configured to generate a voice response information corresponding to the text response information, based on the text response information and language identifier A, representing the language represented by language identifier A, i.e. audio. The voice response information is played in the language "English" represented by language identifier A.

The voice processing unit 401 is configured to control the voice playback unit 403 to play voice response information in English, controls the display unit 404 to display the image of the virtual object as the image corresponding to language identifier A, such as a British gentleman, and controls the gestures and mouth shapes of the virtual object to match the audio played by the voice playback unit. The voice processing unit 401 is configured to also control the display unit 404 to display language elements corresponding to language identifier A, to remind the user of the language type currently used by the interaction object. For example, the display unit 404 is configured to display display elements corresponding to multiple language identifiers, including language identifier A, language identifier B, and language identifier C. Wherein, language identifier A indicates English, language identifier B indicates Japanese, and language identifier C indicates Arabic. When the display unit 404 is configured to display text information 1 corresponding to language identifier A, and the voice playback unit plays audio of text information 1 corresponding to language identifier A in same language, the display element corresponding to language identifier A is highlighted. In an example, the highlighting method is flashing; in another example, the highlighting method is that the element corresponding to language identifier A is surrounded by a focus box.

In summary, the voice processing method provided in the application converts voice information into text information, and the server classifies the text information. If the category of the text information is instruction, the first response information corresponding to the text information is sent to the client. If the category of the text information is Q&A, the text information is input into the language model to obtain the first response information, and the client then displays or plays it based on the first response information. Through the voice processing method provided in the embodiments of the application, if the category of text information is instruction, it will no longer be recognized through subsequent language models. This can accelerate the response speed of the voice response system to customers while maintaining answer quality.

Referring to FIG 6, an embodiment of the application also provides a voice processing method 6000 applied to a client, which the client includes a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, wherein the client and

the first server can be communicatively connected. Referring to FIG 1, the voice processing method shown in FIG 6 at least includes steps 601-603.

At step 601, the voice receiving unit 102 is configured to collect voice information and
5 send it to the voice processing unit.

In some examples, the user or device outputs voice information, and the voice receiving unit 102 is configured to collect the voice information and send it to the voice processing unit 101. Wherein, the device has voice playback function, such as
10 smartphones, robots, etc.

At step 602, the voice processing unit 101 is configured to receive voice information, obtain the corresponding text information of the voice information, and sends the text information to the first server 200; wherein, the first server 200 is configured to
15 determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit. In an example, the first response information includes at least the text response information corresponding to the text information. In some
20 embodiments, the first response information includes any combination of text response information, image response information, and action response instructions.

It should be noted that the text information corresponding to the obtained voice information in an embodiment of the application can be completed by the client or the
25 server, and there is no limitation in the embodiment of the application. In an embodiment of the application, the voice processing unit 101 is configured to analyze the voice information to obtain the corresponding text information. It should be noted that the process of analyzing the voice information in the embodiment of the application is the process of decoding the voice information. As another possible
30 implementation, the voice processing unit 101 is configured to also upload voice information to a second server (not shown in the figure), which is used to analyze the voice information and obtain the corresponding text content and language identifier. In an example, the voice to text module of the second server decodes the voice information, obtains the text content and language identifier corresponding to the voice

information, and returns the text content and language identifier to the voice processing unit 101. The voice processing unit 101 is configured to obtain the text content and language identifier returned by the second server, which is the text information. The existing technology transmits voice information from the client to the server for decoding. The applicant believes that compared with the existing technology, the text information corresponding to the voice information obtained in an embodiment of the application is completed by the client or a non first server part, saving resources on the first server 200 and improving the processing speed of the first server 200. Moreover, the embodiment of the application transmits text information between the client 100 and the first server 200 instead of voice information with a large amount of data, saving network resources and improving the transmission speed and overall voice processing efficiency of the information.

In practical operation, the voice processing unit 101 is configured to obtain the text content corresponding to the voice information and a language identifier when decoding the voice information, which indicates the language type of the voice information. In other words, the text information in the application also includes language identifier. In an example, the client of the application can recognize languages including Chinese, English, and Arabic. Therefore, the language identifier recognized by the voice processing unit 101 includes language identifier A corresponding to English, language identifier B corresponding to Chinese, and language identifier C corresponding to Arabic. For example, when the language corresponding to the collected voice information is English, the text information obtained by the voice processing unit 101 includes the language identifier A corresponding to English.

As shown in FIG 3 above, after obtaining the text information corresponding to the voice information, the client 100 is configured to send the text information to the first server 200, so that the first server 200 can proceed with the next step of processing the text information. In practical operation, the client 100 may also include a camera 105. The camera 105 is configured to collect image information of the voice object, and the image information of the voice object is used to assist the voice processing unit 101 in processing the voice information. In an embodiment of the application, the

voice object refers to the object that emits voice information, such as users, devices, etc.

5 In an example, the voice processing unit 101 is configured to intercept the voice information collected by the voice receiving unit 102 based on the image information of the voice object to obtain key voice information; the voice processing unit 101 is configured to obtain text information corresponding to the key voice information based on the key voice information. In this way, the voice processing unit 101 is configured to extract key voice information from the voice information collected by the voice
10 receiving unit 102 and obtain the text information of the key voice information, without the need to decode all the collected voice information, reducing the computational cost of decoding the original collected voice information and improving decoding efficiency.

At step 603, the voice processing unit 101 is configured to receive the first response
15 information and obtain the second response information corresponding to the first response information. The second response information includes any combination of the following: the first response information and the voice response information corresponding to the text response information; control the voice playback unit to play voice response information and/or display the first response information on the display
20 unit.

After receiving the first response information, the voice processing unit 101 is configured to obtain the voice response information corresponding to the text response information and control the voice playback unit 103 to play the voice response
25 information. In other embodiments, the display unit 104 is configured to also be controlled to display the first response information, such as displaying text response information. In another embodiment, the voice processing unit 101 is configured to control the voice playback unit 103 to play the voice response information after generating the voice response information, and control the display unit 104 to display
30 the text response information of the first response information. In another embodiment, the voice processing unit 101 is configured to control the display unit 104 to display the first response information after receiving it.

In some embodiments, the first response information of the voice processing unit 101 may contain different content and obtain different second response information. When the first response information includes text response information, the voice processing unit 101 is configured to generate a voice response information corresponding to the text response information, control the voice playback unit 103 to play the voice response information, and control the display unit 104 to display the text response information. When the first response information also includes action response instructions and image response information, the voice processing unit 101 is configured to also control the client to execute the action indicated by the action response instruction and control the display screen to display the image response information.

In an example, the display screen displays a virtual object, and the first response information includes text response information, image response information, and action response instructions. The voice processing unit 101 is configured to control the display unit 104 to display text response information in the text display area, image response information in the image display area, and controls the virtual object to execute the action corresponding to the action response instruction. In specific implementation, the voice processing unit 101 is configured to also invoke the display content corresponding to the action response instruction, and play the display content through the display unit 104. It should be noted that the voice processing unit 101 is configured to store the mapping relationship between action response instructions and audio/video/image/function, as well as the mapping relationship between action response instructions and display content. After receiving the action response instruction, the voice processing unit 101 is configured to call the audio/video/image/function and display content corresponding to the action response instruction for playback. For example, the action response instruction A corresponds to audio A, voice playback unit, video A, and video player. When the voice processing unit 101 is configured to receive the action response instruction A, it invokes the voice playback unit to play the corresponding audio A, and invokes the video player to play the corresponding video A.

Referring to FIG 7, an embodiment of the application also provides a voice processing method 7000 applied to a first server 200, wherein the first server 200 and the client

100 can be communicatively connected, which the client includes a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit. Referring to FIG 1, the voice processing method 7000 shown in FIG 7 at least includes steps 701-702.

5 At step 701, the first server 200 is configured to receive text information sent by the voice processing unit, and the text information is generated based on the voice information of the voice receiving unit. The process of generating text information can be found in the aforementioned step 2002 of method 2000, and will not be further elaborated here.

10

At step 702, the first server 200 is configured to determine the category corresponding to the text information, execute processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit to generate the
15 corresponding second response information based on the first response information. The second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play voice response information and/or display the first response information on the display unit. In an
20 embodiment, the first response information includes at least text response information, and the voice processing unit 101 is configured to generate the corresponding voice response information for the text response information.

In some possible embodiments, the first response information may also comprise any
25 combination of the following: text response information, image response information, action response instruction; the first server 200 is configured to determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, comprises:

30

in the case where the category is instruction, determine the action response instruction that matches the text information, or any combination of the action response instruction and the following content: text response information, image response information;

in the case where the category is Q&A, input the text information into the language model to obtain the text response information corresponding to the text information.

- 5 In an embodiment of the application, the first server 200 may include a text processing module 201. The text processing module 201 is configured to obtain classification results based on the text content attributes of the text information, which is the category corresponding to the text information. In an embodiment of the application, the attributes of the text content may include Q&A attributes or instruction attributes, and the categories corresponding to the text information include Q&A or instruction. When receiving the text information of the instruction attribute, the client 100 is configured to execute the action command corresponding to the text information. For example, if the text content of the text information is "have a dance", then the text content attribute is the instruction attribute, and the category corresponding to the text information is the instruction. The client 100 is configured to execute the action command "dance". When receiving text information with Q&A attributes, the client 100 is configured to answer the user's question. For example, if the text content of the text message is "Where is the restroom?", then the text content attribute is Q&A, and the corresponding category of the text message is Q&A, the client 100 is configured to answer the question "Where is the restroom?". In practical operation, the text processing module 201 is configured to classify the attributes of text information through keyword recognition, relatively simple language models, or other methods. The embodiments of the application are not limited here.
- 25 In some embodiments, the text processing module 201 is configured to store a series of instruction keywords. When the text information hits this series of instruction keywords, it is determined that the text content attribute of the text information is an instruction attribute, that is, the category of the text information is an instruction. The text processing module 201 is configured to determine the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information, and returns the above response information to the client 100 (such as action response instruction, or any combination of action response instruction and text response information/image response information).

When the text information misses the instruction keywords of the series, it is determined that the text content attribute of the text information is a Q&A attribute, that is, the text information category is Q&A. The text processing module 201 is configured to input the text information to the language model to obtain the text response information corresponding to the language identifier.

In some possible embodiments, the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the first server 200 is configured to determine the category corresponding to the text information, executes the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, comprises:

when the first server 200 is configured to determine that the text information matches the instruction corresponding to the language identifier, determine the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

when the first server 200 is configured to determine that the text information does not match the instruction corresponding to the language identifier, input the text information to the language model to obtain the text response information corresponding to the language identifier.

In an example, the text processing module 201 is configured to store instruction keywords for various languages, which correspond to language identifiers and set or store action response instructions accordingly. When the text information hits an instruction keyword identified by a language, it is determined that the text information is an instruction attribute. Then, the action response instruction corresponding to the matching instruction keyword is determined, and the action response instruction is returned to the client 100 being configured to execute the action response instruction. For example, the text processing module 201 stores instruction keywords: Chinese "have a dance" and English "sing a song". Wherein, the action response instruction corresponding to "have a dance" is "dance", and the action response instruction

corresponding to "sing a song" is "sing". When the text content of the text information is "have a dance", it is determined that the Chinese instruction keyword "have a dance" has been hit. At this point, the matched action response instruction is "dance". The text processing module 201 is configured to feedback the action response instruction "dance" to the voice processing unit 101 of the client 100, which drives the virtual object displayed on the display screen to perform the dance action. For example, when the text content of the text information is "sing a song", it is determined that the English instruction "sing" has been hit, and the matched action response instruction is "sing". The text processing module 201 is configured to feedback the action response instruction "sing" to the voice processing unit 101 of the client 100, which is configured to drive the voice playback unit 103 to play the song.

In another example, the text processing module 201 is configured to store instruction keywords for various languages, which correspond to language identifiers and set or store action response instructions accordingly. When the text information fails to hit the instruction keyword of any language identifier, the text processing module 201 is configured to input the text information to the language model to obtain the text response information corresponding to the language identifier. In most cases, large language models that support multiple languages at the same time have a larger volume and slower understanding and response speed to textual information. For example, the parameter scale of large language models that currently support multiple languages is usually in the billions, with a very large volume. In order to improve the response speed of the language model as much as possible, an embodiment of the application divides the language model into one or more relatively independent language sub models based on the language type. The language sub model can support one language or a small number of languages, and the parameter quantity of the language sub model is smaller than that of the large language model, which can be in the billions. Compared with the single large model that supports multiple languages simultaneously, the language sub model that supports one or a small number of languages in an embodiment of the application has a smaller volume and a faster response speed to text information. In some specific implementations, the first server includes or connects a language model, which includes one or more language sub models, each corresponding to one or more language types. Each language sub model is used to understand the input text content and output the corresponding text

response content in a language with the same text content. When the text information fails to hit the instruction keyword of any language identifier, the text processing module 201 is configured to input the text information to the target language sub model corresponding to the language identifier of the text information to obtain the text response information returned by the target language sub model. The language identifier of the text response information is the language identifier of the text information. For example, the text processing module 201 is configured to store instruction keywords such as "have a dance" in Chinese and "sing a song" in English. When the text content of the text message is "where is the restroom", the instruction keywords by each language identifier are not hit. The text processing module 201 is configured to send text information to the target language sub model corresponding to the language identifier "Chinese" (specifically, the text content of the text message "where is the restroom"), and the target language sub model understands and generates a Chinese text response message "the restroom is on your left", which is sent to the text processing module 201.

In an embodiment of the application, the first server 200 is configured to classify and process text information of different categories based on the classification results, execute different processing logic, and obtain the first response information corresponding to different categories. The text information of different categories in an embodiment of the application includes text information of instruction attributes and text information of Q&A attributes. The first server 200 is configured to perform different categories of logical processing on the text information of instruction attributes and Q&A attributes based on the text content attributes of the text information, and obtain the first response information corresponding to different categories of text information. By classifying text information before processing, the first server 200 is configured to better recognize and process text information in a targeted manner. For different categories of text information, it can selectively generate the first response information corresponding to the category by itself or the language model, in order to obtain the second response information corresponding to different categories. Specifically, when the category of text information is instruction, an embodiment of the application matches and returns action response instructions based on its own stored information. When the category of text information is Q&A, the text response information is returned through a language model. Due to the fact that the first response information

corresponding to the generated instruction category is based on self stored information matching, without the necessary for semantic analysis, semantic understanding, text generation, and other operations, the speed of generating the first response information corresponding to the instruction category is fast, which can accelerate the response speed of the voice processing system to customers while maintaining the quality of the voice processing method in the application. In addition, generating the first response information corresponding to the Q&A category does not require the text processing module of the first server, but is processed by a language model with stronger voice processing capabilities, which further accelerates the response speed of the voice processing system to customers.

In practical operation, an embodiment of the application matches and returns action response instructions based on its own stored information. When the category of text information is Q&A, the text response information is returned through a language model. Due to the fact that the first response information corresponding to the generated instruction category is based on self stored information matching, without the necessary for semantic analysis, semantic understanding, text generation, and other operations, the speed of generating the first response information corresponding to the instruction category is fast, which can accelerate the response speed of the voice processing system to customers while maintaining the quality of the voice processing method in the application. In addition, generating the first response information corresponding to the Q&A category does not require the text processing module of the first server, but is processed by a language model with stronger voice processing capabilities, which further accelerates the response speed of the voice processing system to customers.

It should be noted that in order to improve the response speed of the voice processing method in the application, the text processing module usually classifies text information through keyword matching, relatively simple language classification, or other faster classification methods. Therefore, while ensuring speed, the text processing module may recognize some text information that belongs to instruction categories as problem categories. Although the language model in the application has a slower response speed compared to the text processing module, it has better comprehension and judgment abilities for text information. Therefore, in some possible embodiments, after being classified by the text processing module first, the received

text information will be further classified by a language model, thereby increasing the accuracy of text information classification while maintaining response speed. Specifically, the language response module 202 is configured to utilize a language model to semantically understand text information and obtain comprehension results.

- 5 If the comprehension result indicates that the category of the text information is instruction, the language response module 202 is configured to send the action response instruction corresponding to the text information to the voice processing unit 101 through the text processing module 201. If the comprehension result indicates that the category of text information is Q&A, the language model generates text
10 response information based on the text information. The language response module 202 is configured to send the text response information to the voice processing unit 101 through the text processing module 201.

Refer to FIG 8, an embodiment of the application provides a voice processing system
15 8000 comprising a third client 800 and a third server 900, wherein the third client 800 and the third server 900 can be communicatively connected. The third client 800 comprises a voice processing unit 801, a voice receiving unit 802, a voice playback unit 803, and a display unit 804, wherein:

- 20 the voice receiving unit 802 is configured to collect voice information and send the voice information to the voice processing unit;
the voice processing unit 801 is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to the first server;
25 the third server 900 is configured to determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit, the first response information at least includes the text response information corresponding to the text
30 information;
the voice processing unit 801 is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response

information corresponding to the first response information; control the voice playback unit 803 to play the voice response information and/or display the first response information on the display unit 804.

- 5 In some possible embodiments, the first response information may also comprise any combination of the following: text response information, image response information, action response instruction; the third server 900 is configured as:

10 in the case where the category is instruction, determine the action response instruction that matches the text information, or any combination of the action response instruction and the following content: text response information, image response information;

in the case where the category is Q&A, input the text information into the language model to obtain the text response information corresponding to the text information.

15

In some possible embodiments, the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the third server 900 is further configured as:

20 when the third server 900 determines that the text information matches the instruction corresponding to the language identifier, determines the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

25 when the third server 900 determines that the text information does not match the instruction corresponding to the language identifier, inputs the text information to the language model to obtain the text response information corresponding to the language identifier.

30 In some possible embodiments, the third server 900 comprises or connects a language model, wherein the language model comprises one or more language sub models, each of which corresponds to one or more language types; the first server is configured as:

input the text information to the target language sub model corresponding to the language identifier, which the target language sub model is used to understand the text information and output the text response information corresponding to the language type represented by the language identifier;

5 receive the text response information.

In some possible embodiments, the voice processing unit 801 is configured to:

10 analyzes the voice information, obtains the text content and language identifier corresponding to the voice information, and sends the text information including the text content and language identifier to the third server 900; alternatively,
sends the voice information to another server, which the other server is used to analyze the voice information and obtain the text content and language identifier corresponding to the voice information; receive the text content and language identifier
15 returned by the other server, and send the text information including the text content and language identifier to the first server.

In some possible embodiments, the third client 800 may also include a camera device for capturing images of voice objects; the voice processing unit 801 is configured as
20 follows:

receives the voice information and the image information of the voice object transmitted by the camera device;
intercepts key voice information of the voice information based on the image information of the voice object;
25 recognizes the key voice information and obtains the text information corresponding to the key voice information.

In some possible embodiments, the first response information includes text response information, image response information, and action response instructions;

30

the display unit 804 is configured to display virtual objects;
the voice processing unit 801 is configured to control the display unit to display text response information in the text display area, display image response information in

the image display area, and control the virtual object to execute actions corresponding to instructions.

Referring to FIG 9, an embodiment of the application provides a fourth client 950, which includes a voice receiving unit 902, a voice processing unit 901, a voice playback unit 903, and a display unit 904, the fourth client 950 and the third server 900 can be communicatively connected; wherein:

the voice receiving unit 902 is configured to collect voice information and send the voice information to the voice processing unit;

the voice processing unit 901 is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to the first server;

the voice processing unit 901 is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

In some possible embodiments, the first response information includes at least the text response information corresponding to the text information, and the voice response information is the voice information corresponding to the text response information.

In other embodiments, the first response information includes any combination of text response information, image response information, and action response instructions.

In some possible embodiments, the voice processing unit 901 is configured to:

analyzes the voice information, obtains the text content and language identifier corresponding to the voice information, and sends the text information including the text content and language identifier to the third server 900; alternatively, sends the voice information to another server, which the other server is used to analyze the voice information and obtain the text content and language identifier

corresponding to the voice information; receive the text content and language identifier returned by the other server, and send the text information including the text content and language identifier to the third server 900.

- 5 In some possible embodiments, the fourth client 950 may also include a camera device for capturing images of voice objects; the voice processing unit 901 is configured as follows:

receives the voice information and the image information of the voice object
10 transmitted by the camera device;
intercepts key voice information of the voice information based on the image information of the voice object;
recognizes the key voice information and obtains the text information corresponding to the key voice information.

15

In some possible embodiments, the first response information includes text response information, image response information, and action response instructions;
the display unit 904 is configured to display virtual objects;
the voice processing unit 901 is configured to control the display unit to display text
20 response information in the text display area, display image response information in the image display area, and control the virtual object to execute actions corresponding to instructions.

It should be noted that the client in the embodiment of the application can implement
25 various functions of the aforementioned implementation example of the voice processing system and achieve the same effect, which will not be repeated here.

In certain embodiments, the application also provides a third server 900 that the third server 900 and the fourth client 950 can be communicatively connected, the client
30 comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, wherein the first server is configured as:

receives text information sent by the voice processing unit, and the text information is generated based on the voice information sent by the voice receiving unit;

determines the category corresponding to the text information, executes processing logic corresponding to the category, obtains the first response information of the text information corresponding to the category, and sends the first response information to the voice processing unit to generate the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

10

In some possible embodiments, the first response information includes at least the text response information corresponding to the text information, and the voice response information is the voice information corresponding to the text response information.

15 In other embodiments, the first response information further includes any combination of the following: image response information, action response instructions; the third server 900 is configured as:

20 in the case where the category is instruction, determine the action response instruction that matches the text information, or any combination of the action response instruction and the following content: text response information, image response information;

in the case where the category is Q&A, input the text information into the language model to obtain the text response information corresponding to the text information.

25

In some possible embodiments, the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the third server 900 is further configured as:

30 when the third server 900 determines that the text information matches the instruction corresponding to the language identifier, determines the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

when the third server 900 determines that the text information does not match the instruction corresponding to the language identifier, inputs the text information to the language model to obtain the text response information corresponding to the language identifier.

5

In some possible embodiments, the third server 900 comprises or connects a language model, wherein the language model comprises one or more language sub models, each of which corresponds to one or more language types; the third server 900 is configured as:

10

input the text information to the target language sub model corresponding to the language identifier, which the target language sub model is used to understand the text information and output the text response information corresponding to the language type represented by the language identifier;

15 receive the text response information.

It should be noted that the third server 900 in the embodiment of the application can implement various functions of the aforementioned implementation example of the voice processing system and achieve the same effect, which will not be repeated here.

20

According to some embodiments of the application, the application provides a non-volatile computer storage medium for a client, on which computer executable instructions are stored, which are set to be executed when run by a processor: the actions executed by the client in the voice processing method described in the above
25 embodiments. According to some embodiments of the application, the application also provides a non-volatile computer storage medium on a server, which stores computer executable instructions set to be executed when run by a processor: the action executed by the first server in the voice processing method described in the above embodiments.

30

According to some embodiments of the application, the application provides a computer-readable storage medium, wherein the computer-readable storage medium stores a program that, when executed by a multi-core processor, causes the multi-core processor to execute the voice processing method applied to the client as

described above. the application also provides a computer-readable storage medium, wherein the computer-readable storage medium stores a program that, when executed by a multi-core processor, causes the multi-core processor to execute the voice processing method applied to the first server. Computer readable media include permanent and non permanent, movable and non movable media, and can be implemented for information storage by any method or technology. Information can be computer-readable instructions, data structures, program modules, or other data. Examples of computer-readable storage media include but are not limited to phase change memory (PRAM), static random access memory (SRAM), dynamic random access memory (DRAM), other types of random access memory, read-only memory, electrically erasable programmable read-only memory (EEPROM), flash memory or other memory technologies CD-ROM、 Digital multifunctional optical discs (DVDs) or other optical storage, magnetic cassette tapes, magnetic tape disk storage, or other magnetic storage devices or any other non transmission media can be used to store information that can be accessed by computing devices. Furthermore, although the operations of the present method are described in a specific order in the accompanying drawings, this does not require or imply that these operations must be performed in that specific order, or that all shown operations must be performed to achieve the desired results. Additionally, certain steps can be omitted, and multiple steps can be merged into one step for execution, and/or a step can be decomposed into multiple sub steps for execution.

SERVER 10000

Referring to FIG 10, a server 10000 is typically administered by an entity operating the aforementioned methods, typically whenever cloud server services are not utilised. The use of the server 10000 is typically necessary in a situation where the entity operating the aforementioned methods desires full control of infrastructure enabling the aforementioned methods.

The server 10000 typically carries out processes to enable the carrying out of the aforementioned methods while receiving and transmitting relevant data. While the server 10000 is illustrated as a single computing system in FIG 10, it should be

appreciated that the server 10000 can be a distributed set-up utilising a plurality of computing systems.

The components of the server 10000 can be configured in a variety of ways. The components can be implemented entirely by software to be executed on standard computer server hardware, which may comprise one hardware unit or different computer hardware units distributed over various locations, some of which may require the communications network 150 for communication.

In the example shown in FIG 10, the server 10000 is a commercially available server computer system based on a 32 bit or a 64 bit Intel architecture, and the processes and/or methods executed or performed by the server 10000 are implemented in the form of programming instructions of one or more software components or modules 322 stored on non-volatile (e.g. hard disk) computer-readable storage 324.

The server 10000 includes at least one or more of the following standard, commercially available, computer components, all interconnected by a BUS 335:

1. random access memory (RAM) 326;

2. at least one computer processor 328, and

3. external computer interfaces 330:

- a. universal serial bus (USB) interfaces 330a (at least one of which is connected to one or more user-interface devices, such as a keyboard, a pointing device (e.g., a mouse 332 or touchpad),

- b. a network interface connector (NIC) 330b which connects the central server 140 to the data communications network 150; and

- c. a display adapter 330c, which is connected to a display device 334 such as a liquid-crystal display (LCD) panel device.

The server 10000 includes a plurality of standard software modules, including:

1. an operating system (OS) 336 (e.g., Linux or Microsoft Windows);

2. web server software 338 (e.g., Apache, available at <http://www.apache.org>);

3. Javascript or Python modules 340; and

4. structured query language (SQL) modules 342 (e.g., MySQL, available from <http://www.mysql.com>), which allow data to be stored in and retrieved/accessed from an SQL database 316.

5 Together, the web server 338, Javascript module 340, and SQL modules 342 provide the server 10000 with the general ability to allow users with client computing devices equipped with standard web browser software to access the server 10000 and in particular to provide data to and receive data from the database 316. It will be understood by those skilled in the art that the specific functionality provided by the
10 server 10000 to such users is provided by scripts accessible by the web server 338, including the one or more software modules 322 implementing the processes performed by the central server 140, and also any other scripts and supporting data 344, including markup language (e.g., HTML, XML, Java) scripts, and the like.

15 The boundaries between the modules and components in the software modules 322 are exemplary, and alternative embodiments may merge modules or impose an alternative decomposition of functionality of modules. For example, the modules discussed herein may be decomposed into submodules to be executed as multiple computer processes, and, optionally, on multiple computers. Moreover, alternative
20 embodiments may combine multiple instances of a particular module or submodule. Furthermore, the operations may be combined or the functionality of the operations may be distributed in additional operations in accordance with the invention. Alternatively, such actions may be embodied in the structure of circuitry that implements such functionality, such as the micro-code of a complex instruction set
25 computer (CISC), firmware programmed into programmable or erasable/programmable devices, the configuration of a field- programmable gate array (FPGA), the design of a gate array or full-custom application- specific integrated circuit (ASIC), or the like.

30 Respective steps of processes of the server 10000 may be executed by a module (of software modules 322) or a portion of a module. The processes may be embodied in a non-transient machine -readable and/or computer-readable medium for configuring a computer system to execute the method. The software modules may be stored within

and/or transmitted to a computer system memory to configure the central server 140 to perform the functions of the module.

5 The server 10000 normally processes information according to a program (a list of internally stored instructions such as a particular application program and/or an operating system) and produces resultant output information via input/output (I/O) devices 330. A computer process typically includes an executing (running) program or portion of a program, current program values and state information, and the resources used by the operating system to manage the execution of the process. A parent
10 process may spawn other, child processes to help perform the overall functionality of the parent process. Because the parent process specifically spawns the child processes to perform a portion of the overall functionality of the parent process, the functions performed by child processes (and grandchild processes, etc.) may sometimes be described as being performed by the parent process.

15 Throughout this specification and claims which follow, unless the context requires otherwise, the word “comprise”, and variations such as “comprises” or “comprising”, will be understood to imply the inclusion of a stated integer or group of integers or steps but not the exclusion of any other integer or group of integers.

20 Persons skilled in the art will appreciate that numerous variations and modifications will become apparent. All such variations and modifications which become apparent to persons skilled in the art, should be considered to fall within the spirit and scope that the invention broadly appearing before described.

CLAIMS

1. A voice processing method, characterized in that applied to a voice processing system comprising a client and a first server, wherein the client and the first server can be communicatively connected, and the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit; the method comprises:
the voice receiving unit collecting voice information and sending the voice information to the voice processing unit;
the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server;
the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit;
the voice processing unit receiving the first response information , and obtaining the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following:
the first response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.
2. The method according to claim 1, characterized in that the first response information comprises any combination of the following: text response information, image response information, action response instruction; the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, comprises:
in the case where the category is instruction, determining the action response instruction that matches the text information, or any combination of the action response instruction and the following content: text response information, image response information;

in the case where the category is Q&A, inputting the text information into the language model to obtain the text response information corresponding to the text information.

3. The method according to claim 1, characterized in that the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, comprises:

when the first server determining that the text information matches the instruction corresponding to the language identifier, determining the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

when the first server determining that the text information does not match the instruction corresponding to the language identifier, inputs the text information to the language model to obtain the text response information corresponding to the language identifier.

4. The method according to claim 3, characterized in that the first server comprises or connects a language model, wherein the language model comprises one or more language sub models, each of which corresponds to one or more language types; inputs the text information to the language model to obtain the text response information corresponding to the language identifier, comprises:

inputting the text information to the target language sub model corresponding to the language identifier, which the target language sub model is used to understand the text information and output the text response information corresponding to the language type represented by the language identifier; receiving the text response information.

5. The method according to claim 1, characterized in that the voice processing unit receives the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server, comprises:

the voice processing unit analyzing the voice information, obtaining the text content and language identifier corresponding to the voice information, and sending the text information including the text content and language identifier to the first server; alternatively,

- 5 the voice processing unit sending the voice information to a second server, which the second server is used to analyze the voice information and obtain the text content and language identifier corresponding to the voice information; receiving the text content and language identifier returned by the second server, and sending the text information including the text content and language identifier to the first server.

10

6. The method according to claim 1, characterized in that the client further comprises camera device for capturing images of voice objects; the voice processing unit receives the voice information, obtains the text information corresponding to the voice information, comprises:

- 15 the voice processing unit receiving the voice information and the image information of the voice object transmitted by the camera device;
the voice processing unit intercepting key voice information of the voice information based on the image information of the voice object;
the voice processing unit recognizing the key voice information and obtaining the text
20 information corresponding to the key voice information.

7. The method according to claim 1, characterized in that the first response information includes text response information, image response information, and action response instructions, and the display unit displaying a virtual object; displaying the first
25 response information on the display unit, comprises:

the voice processing unit controlling the display unit to display the text response information in the text display area, displaying the image response information in the image display area, and controlling the virtual object to execute the action corresponding to the action response instructions.

30

8. A voice processing method characterized in that applied to a client, wherein the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, wherein the client and the first server can be communicatively connected; the method comprises:

the voice receiving unit collecting voice information and sending the voice information to the voice processing unit;

the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server; wherein, the first server is used to determine the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit;

the voice processing unit receiving the first response information , and obtaining the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.

9. The method according to claim 8, characterized in that the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, and sending the text information to the first server, comprises:

the voice processing unit analyzing the voice information, obtaining the text content and language identifier corresponding to the voice information, and sending the text information including the text content and language identifier to the first server; alternatively,

the voice processing unit sending the voice information to a second server, which the second server is used to analyze the voice information and obtain the text content and language identifier corresponding to the voice information; receiving the text content and language identifier returned by the second server, and sending the text information including the text content and language identifier to the first server.

10. The method according to claim 8, characterized in that the client further comprises camera device for capturing images of voice objects; the voice processing unit receiving the voice information, obtaining the text information corresponding to the voice information, comprises:

the voice processing unit receiving the voice information and the image information of the voice object transmitted by the camera device;

the voice processing unit intercepting key voice information of the voice information based on the image information of the voice object;

- 5 the voice processing unit recognizing the key voice information and obtains the text information corresponding to the key voice information.

11. The method according to claim 8, characterized in that the first response information includes text response information, image response information, and
10 action response instructions, and the display unit displaying a virtual object; displaying the first response information on the display unit, comprises:

the voice processing unit controlling the display unit to display the text response information in the text display area, displaying the image response information in the image display area, and controlling the virtual object to execute the action
15 corresponding to the action response instructions.

12. A voice processing method characterized in that applied to a first server, wherein the first server and the client can be communicatively connected, the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display
20 unit; the method comprises:

the first server receiving text information sent by the voice processing unit, and the text information is generated based on the voice information sent by the voice receiving unit;

the first server determining the category corresponding to the text information, executes processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, and sending the first response information to the voice processing unit to generate the corresponding second response information based on the first response information, wherein the second response information includes any combination of the following: the first
25 response information, the voice response information corresponding to the first response information; controlling the voice playback unit to play the voice response information and/or displaying the first response information on the display unit.
30

13. The method according to claim 12, characterized in that the first response information comprises any combination of the following: text response information, image response information, action response instruction; the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtains the first response information of the text information corresponding to the category, comprises:

in the case where the category is instruction, determining the action response instruction that matches the text information, or any combination of the action response instruction and the following content: text response information, image response information;

in the case where the category is Q&A, inputting the text information into the language model to obtain the text response information corresponding to the text information.

14. The method according to claim 12, characterized in that the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the first server determining the category corresponding to the text information, executing the processing logic corresponding to the category, obtaining the first response information of the text information corresponding to the category, comprises:

when the first server determining that the text information matches the instruction corresponding to the language identifier, determining the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

when the first server determining that the text information does not match the instruction corresponding to the language identifier, inputting the text information to the language model to obtain the text response information corresponding to the language identifier.

15. The method according to claim 14, characterized in that the first server comprises or connects a language model, wherein the language model comprises one or more language sub models, each of which corresponds to one or more language types; inputs the text information to the language model to obtain the text response information corresponding to the language identifier, comprises:

inputting the text information to the target language sub model corresponding to the language identifier, which the target language sub model is used to understand the text information and output the text response information corresponding to the language type represented by the language identifier;

5 receive the text response information.

16. A voice processing system, characterized in that the voice processing system comprises a client and a first server, wherein the client and the first server can be communicatively connected, the client comprises a voice receiving unit, a voice
10 processing unit, a voice playback unit, and a display unit, wherein:

the voice receiving unit is configured to collect voice information and send the voice information to the voice processing unit;

the voice processing unit is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to
15 the first server;

the first server is configured to determine the category corresponding to the text information, execute the processing logic corresponding to the category, obtain the first response information of the text information corresponding to the category, and send the first response information to the voice processing unit, the first response
20 information at least includes the text response information corresponding to the text information;

the voice processing unit is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any
25 combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

30 17. The system according to claim 16, characterized in that the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the first server is further configured as:
when the first server determines that the text information matches the instruction corresponding to the language identifier, determines the action response instruction

that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;

5 when the first server determines that the text information does not match the instruction corresponding to the language identifier, inputs the text information to the language model to obtain the text response information corresponding to the language identifier.

10 18. A client, characterized in that the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, the client and the first server can be communicatively connected; wherein:

the voice receiving unit is configured to collect voice information and send the voice information to the voice processing unit;

15 the voice processing unit is configured to receive the voice information, obtain the text information corresponding to the voice information, and send the text information to the first server;

the voice processing unit is further configured to receive the first response information, and obtain the corresponding second response information based on the first response information, wherein the second response information includes any
20 combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

25 19. A first server, characterized in that the first server and the client can be communicatively connected, the client comprises a voice receiving unit, a voice processing unit, a voice playback unit, and a display unit, wherein the first server is configured as:

30 receives text information sent by the voice processing unit, and the text information is generated based on the voice information sent by the voice receiving unit;

determines the category corresponding to the text information, executes processing logic corresponding to the category, obtains the first response information of the text information corresponding to the category, and sends the first response information to the voice processing unit to generate the corresponding second response information

based on the first response information, wherein the second response information includes any combination of the following: the first response information, the voice response information corresponding to the first response information; control the voice playback unit to play the voice response information and/or display the first response information on the display unit.

20. The server according to claim 19, characterized in that the text information comprises a language identifier, wherein the language identifier represents the language type of the voice information; the first server is further configured as:

- 10 when the first server determines that the text information matches the instruction corresponding to the language identifier, determines the action response instruction that matches the instruction, or any combination of the action response instruction and the following content: text response information corresponding to the language identifier, image response information;
- 15 when the first server determines that the text information does not match the instruction corresponding to the language identifier, inputs the text information to the language model to obtain the text response information corresponding to the language identifier.