

A METHOD AND SYSTEM FOR VOICE INTERACTION

FIELD OF INVENTION

5 The present invention relates to a method and system for voice interaction and specifically relates to a voice interaction method, voice interaction 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 interaction method, characterized in that the voice interaction method is applied to a voice interaction system, wherein the voice interaction system comprises a client and a first server, the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit, the first server
30 comprises or connects to the language model, the language model comprises one or more language sub-model, the language sub-model correspond to one or more languages, and the method comprises:

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

the terminal processing unit obtaining the text information corresponding to the voice information based on the voice information, and sending the text information to the first server;

the first server inputting the text information to the target language sub-model in the language model, which the target language sub-model corresponds to the language category of the text information; obtaining the text response information corresponding to the text information returned by the target language sub-model, which the text response information has the same language category as the text information; sending the text response information to the terminal processing unit;

the terminal processing unit obtaining the voice response information corresponding to the text response information based on the text response information, and controlling the voice playback unit to play the voice response information in the language category of the text information.

In another aspect, there is provided a voice interaction method, characterized in that the voice interaction method is applied to a client, which the client and the first server can be communicatively connected; the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; the method comprises:

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

the terminal processing unit obtaining the text information corresponding to the voice information based on the voice information, sending the text information to the first server, so that the first server inputting the text information to the target language sub-model in the language model, and obtaining the text response information corresponding to the text information returned by the target language sub-model; the target language sub-model corresponds to the language category of the text information, which the text response information has the same language category as the text information;

the terminal processing unit obtaining the voice response information corresponding to the text response information based on the text response information, and

controlling the voice playback unit to play the voice response information in the language category of the text information.

5 In another aspect, there is provided a voice interaction method characterized in that the voice interaction method is applied to a first server, which the client and the first server can be communicatively connected; the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; the
10 method comprises:
the first server receiving text information sent by the terminal processing unit; the terminal processing unit is used for obtaining text information corresponding to the voice information based on the voice information collected by the voice receiving unit; the first server inputting the text information to the target language sub-model in the
15 language model, which the target language sub-model corresponds to the language category of the text information; obtaining the text response information corresponding to the text information returned by the target language sub-model, which the text response information has the same language category as the text information;
the first server sending the text response information to the terminal processing unit,
20 so that the terminal processing unit obtaining the voice response information corresponding to the text response information based on the text response information, and controlling the voice playback unit to play the voice response information.

25 In another aspect, there is provided a client, characterized in that the client and the first server can be communicatively connected, wherein the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; wherein:
30 the voice receiving unit is configured to collect voice information and send the voice information to the terminal processing unit;
the terminal processing unit is configured to obtain text information corresponding to the voice information based on the voice information, send the text information to the first server, so that the first server inputs the text information to the target language

sub-model in the language model, and obtain the text response information corresponding to the text information returned by the target language sub-model; the target language sub-model corresponds to the language category of the text information, which the text response information has the same language category as the text information;

the terminal processing unit is further configured to obtain the voice response information corresponding to the text response information based on the text response information, and control the voice playback unit to play the voice response information in the language category of the text information.

In a final aspect, there is provided a first server, characterized in that the client and the first server can be communicatively connected, wherein the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; wherein:

the first server is configured to receive text information sent by the terminal processing unit; the terminal processing unit is used to obtain text information corresponding to the voice information based on the voice information collected by the voice receiving unit;

the first server is configured to input the text information to a target language sub-model in the language model, which the target language sub-model corresponds to the language category of the text information; obtain the text response information corresponding to the text information returned by the target language sub-model, which the text response information has the same language category as the text information;

the first server is configured to send the text response information to the terminal processing unit, so that the terminal processing unit obtains the voice response information corresponding to the text response information based on the text response information, and control the voice playback unit to play the voice response information.

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,
5 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:

10

FIG 1 is an architecture diagram of a voice interaction system provided in the
embodiment of the application;

FIG 2 is a flowchart of a voice interaction method provided in the embodiment of the
application;

15 FIG 3 is a schematic diagram of a voice interaction method provided in the
embodiment of the application;

FIG 4 is a flowchart of a voice interaction method provided in the embodiment of the
application;

20 FIG 5 is a flowchart of another voice interaction method provided in the embodiment
of the application;

FIG 6 is a signaling diagram of a voice interaction method provided in the embodiment
of the application;

FIG 7 is a flowchart of a voice interaction method provided in the embodiment of the
application;

25 FIG 8 is a flowchart of another voice interaction method provided in the embodiment
of the application;

FIG 9 is a schematic diagram of an illustrative server of an embodiment of the
application.

30 DETAILED DESCRIPTION

Exemplary embodiments of the present application will be described in more detail
below with reference to all accompanying drawings. Although the exemplary
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 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 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 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.

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 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 solution for the voice interaction system to understand and respond to users is: the client receives the voice content input by the user, the server recognizes and understands the voice content, generates the

corresponding voice reply, and then the client plays the voice reply. The applicant believes that the processing and response speed of such solution is slow, the interaction efficiency is low, and affects the user's interaction experience. To improve the response and processing speed of the voice interaction system, certain
5 embodiments of the application provide a voice interaction method, which is applied to the voice interaction system.

FIG 1 shows an architecture diagram of a voice interaction system 98 related to the embodiment of the application.

10

The voice interaction system 98 provided in the embodiment comprises a client 100 and a first server 200. Wherein, the client 100 comprises a terminal processing unit 101, a voice receiving unit 102, and a voice playback unit 103. The terminal processing unit 101 is configured to process the data collected by the voice receiving unit 102 and
15 control the playback of the voice playback unit 103. The terminal processing unit 101 is also configured to communicate with the first server 200, and use the first server 200 to process the data collected by the voice receiving unit 102. In some examples, the voice receiving unit 102 is, for example, a receiver or pickup, such as a microphone; the voice playback unit 103 is for example, a speaker, such as a broadcaster.

20

The client 100 can optionally comprise a display unit 104, and the terminal processing unit 101 is also configured to control the display of the display unit 104. In some examples, 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-
25 emitting diode display screen (LED), plasma display screen (PDP), electronic paper display screen (E-Ink), touch screen display screen, curved display screen, flexible display screen, 3D display screen, transparent display screen, etc.

30

In other embodiments, the voice interaction system 98 also optionally comprises a camera device 105, which may optionally be integrated into the client 100. In some examples, the camera device 105 is a camera. In the voice interaction system 98 shown in FIG 1, the terminal processing unit 101 is configured to process the data collected by the voice receiving unit 102 and the camera device 105, control the playback of the voice playback unit 103 and the display of the display unit 104, and

also the first server 200 is configured to process the data collected by the voice receiving unit 102 and the camera device 105.

The first server 200 comprises a text processing module 201 and a language response module 202. The text processing module 201 and the language response module 202 can both be configured to classify the text information sent by the client 100. The text processing module 201 can also be configured with a communication protocol with the client 100, and the first server 200 uses the text processing module 201 to communicate with the client 100. The language response module 202 comprises or connects to the language models, which comprise one or more language sub-model corresponding to one or more language categories. In one or more language sub-model comprised in the language model, comprising a target language sub-model corresponding to the language category of the text information, the language response module 202 can be configured to use the target language sub-model corresponding to the language identifier in the language model to semantically understand the text information, and produce text response information through the target language sub-model.

It should be noted that the language response module 202 and the text processing module 201 in the application can be located on the first server 200, or on separate servers. The embodiments of the application are not limited here.

The above content introduces the architecture diagram of the voice interaction system provided in the embodiment of the application, which is applied to the voice interaction system 98. Following is an introduction to the voice interaction method applied to the voice interaction system in the embodiment of the application.

As shown in FIG 2, the application provides a voice interaction method 2000 comprising at least steps 2001-2004.

At step 2001, the voice receiving unit 102 collects voice information and sends it to the terminal processing unit 101.

In the embodiment, the object of outputting voice is the voice object, and in some embodiments, the voice object is a user or device. Wherein, the device has voice playback function, such as smartphones, robots, etc. In some examples of the application, the user is used as an example of the voice object to illustrate.

5

In some examples, the user outputs voice information, and the voice receiving unit 102 collects the voice information and sends it to the terminal processing unit 101.

At step 2002, the terminal processing unit 101 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 200.

10

In the embodiment, the process of obtaining the text information corresponding to the voice information is invoked decoding the voice information. In some examples, decoding the voice information is completed by the client 100 to obtain the text information corresponding to the voice information. In another example, the decoding of voice information is completed by the first server 200. In the specific implementation, the terminal processing unit 101 is configured to upload the voice information to a second server (not shown), and the second server is configured to decode the voice information. For example, the voice to text module of the second server decodes the voice information to obtain the corresponding text information, and the second server sends the text information returned by the backend server to the terminal processing unit. The embodiment does not restrict the execution subject of decoding voice information.

15

20

25

In a practical operation, decoding voice information can obtain the corresponding text content and language identifier of the voice information, which is the text information in the application comprises text content and language identifier, wherein the language identifier identifies the language category of the voice information. Assuming that the client 100 of the application supports language categories comprising Japanese, English, and Arabic, the language identifier comprises language identifier A corresponding to English, language identifier B corresponding to Japanese, and language identifier C corresponding to Arabic. In one example, when the language

30

corresponding to the voice message is Arabic, the text information may comprise the language identifier C corresponding to Arabic.

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 for further processing. In practical operation, the client 100 comprises a camera unit 105. The camera unit 105 is configured to collect image information of the voice object, and the image information of the voice object is configured to assist the terminal processing unit 101 in processing the voice information. In one example, the terminal 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 terminal 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 terminal processing unit 101 is configured to extract key voice information from the voice information collected by the voice receiving unit 102, without the need to decode all the collected voice information, reducing the computational power required for decoding voice information and saving hardware resources.

At step 2003, the first server 200 is configured to input text information to the target language sub-model in the language model, obtain the text response information corresponding to the text information returned by the target language sub-model, and send the text response information to the terminal processing unit 101.

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, currently large language models that support multiple languages at the same time typically have parameter magnitudes of billions, the parameter volume is very large. In order to improve the response speed of the language model as much as possible, the embodiment divides the language model into one or more relatively independent language sub-model based on the language category. 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 language

large model that supports multiple languages simultaneously, the language sub-model that supports one or a small number of languages in the embodiment has a smaller volume and a faster response speed to text information.

- 5 In certain embodiments of the application, the language model comprises one or more language sub-model, each corresponding to one or more language categories, i.e. each language sub-model is used to process input text information for the corresponding language category and return text response information for the corresponding language category. For example, the language model consists of three
- 10 language sub-model, namely language sub-model 1, language sub-model 2, and language sub-model 3. Wherein, language sub-model 1 corresponds to Japanese, language sub-model 2 corresponds to English and French, and language sub-model 3 corresponds to Thai. Therefore, language sub-model 1 is used to process the input Japanese information and return the corresponding Japanese response information;
- 15 language sub-model 2 is used to process input English information, return corresponding English response information, as well as to process input French information and return corresponding French response information; similarly, language sub-model 3 is used to process input Thai information and return corresponding Thai response information. It should be understood that the
- 20 embodiments of the application do not limit the types of languages, and the specific types of languages can be set according to the actual situation.

In one example, the text information comprises language identifier. The first server inputs the text information to the target language sub-model corresponding to the

25 language identifier in the language model, and obtains the text response information corresponding to the language identifier. For example, the language model in the application comprises language sub-model A, language sub-model B, and language sub-model C. Language sub-model A supports English, language sub-model B supports Japanese, and language sub-model C supports Arabic. The language

30 identifier in the application comprises 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. The user outputs the language "hello!", and the terminal processing unit 101 obtains the language identifier A in the text information based on the voice information collected by the voice receiving

unit 102. After receiving the text information containing the language identifier A, the language model inputs the text information into the language sub-model A to obtain the text response information. The language identifier A can also be comprised in the text response information.

5

The solution provided in the embodiment can select a matching target language sub-model from one or more language sub-models in the language model based on the language category of the voice information, input the text information corresponding to the voice information to the target language sub-model, and return the text response information corresponding to the text information through the target language sub-model. The text response information returned by the target language sub-model is the same as the language category of the text information and voice information. It can be seen that the embodiment automatically recognizes the language category and text information of the input voice information, processes the text information of the language category through the target language sub-model corresponding to the language category, and returns text response information of the same language category. The target language sub-model, due to its smaller parameter size compared to using a complete language model in related technologies, has a faster processing speed for text information and a faster generation speed for text response information. It can quickly return text response information of the same language category, which can accelerate the response speed of the voice interaction system to the voice object while maintaining the quality of the voice interaction method in the application, and improve the response speed and efficiency of the entire interaction process.

25 After receiving the text response information returned by the target language sub-model, the first server 200 is configured to send the text response information to the terminal processing unit 101.

At step 2004, the terminal processing unit 101 is configured to obtain the voice response information corresponding to the text response information based on the text response information, and control the voice playback unit 103 to play the voice response information in the language category of the text information.

30

The terminal processing unit 101 is configured to receive text response information, generate voice response information corresponding to the text response information, and control the voice playback unit 103 to play the voice response information. In one example, the terminal processing unit 101 is configured to recognize the language category of the text response information and generate voice response information of the same language category. In another example, the text response information comprises language identifier, and the terminal processing unit 101 is configured to generate corresponding language category voice response information based on the language identifier. In the specific implementation, the terminal processing unit 101 is configured to invoke the TTS (Text to Voice) function to generate the corresponding voice response information for the text response information. The process of converting text information into voice information in the embodiments of the application can also be completed by the server, and the embodiments of the application are not limited here. As a possible implementation, the terminal processing unit 101 is configured to upload text information to a second server, which is different from the first server 200 in step 2003. A text to voice module of the second server processes the voice information, and the terminal processing unit 101 is configured to obtain the voice information returned by the second server.

In some practical operation, the client 100 in the application may comprise a display unit 104, and the terminal processing unit 101 is configured to control the display unit 104 to play display content based on text response information. The display unit 104 is configured to display interactive objects that interact with voice objects, such as virtual human objects, intelligent robot objects, etc. The terminal processing unit 101 is configured to control the display unit 104 to display the display content corresponding to the text response information based on the text response information. As an example, the terminal processing unit 101 is configured to control the display unit 104 to display the text corresponding to the text response information, and control the interaction object on the display unit 104 to interact with the voice object based on the text response information, as well as control the interaction object displayed on the display unit 104 to perform the interaction action corresponding to the text response information, such as matching the mouth shape or gesture of the interaction object with the audio played by the voice playback unit 103, thereby facilitating the voice

object to understand the reply content from multiple perspectives such as, for example, visual and auditory, and improving the user experience of the voice object.

As an example, the text response information may comprise a language identifier, and the terminal processing unit 101 may also be configured to display the language category corresponding to the language identifier on the display unit 104 through elements such as images or text based on the language identifier, to remind the voice object of the language category used in the interaction process. For example, the terminal processing unit 101 is configured to control the display unit 104 to display language identifier for text response information.

The solution provided in the embodiment automatically recognizes the language category and text information of the input voice information, processes the text information of the language category through the target language sub-model corresponding to the language category, and returns the text response information of the same language category. Due to the fact that the language category of the text response information is the same as that of the voice response information, the terminal processing unit can generate voice response information that is the same as the input language category, and play the same language category of voice response information through the voice playback unit. Through the implementation example of the application, it is possible to achieve a full voice communication interaction process of "input certain language of voice information → play same language of voice response information", such as the voice object inputting English voice information "Hello!", and the client playing English voice response information "Hello, Dear guest!".

The existing technology sends 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 the embodiment is completed by the client instead of the first server, saving resources on the first server and improving the processing speed of the first server. Moreover, the embodiment sends 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. The target language sub-model, due to its small parameter size, can quickly return text

response information of the same language category, thereby improving the response speed and efficiency of the entire voice interaction process.

As shown in FIG 3, the first server 200 in the embodiment may comprise a language response module 202 and a text processing module 201. The language response module 202 comprises or connects to the language model, and the text processing module 201 comprises a communication protocol with the client. As a possible implementation, the first server 200 in the embodiment can directly understand the semantic information of the text through a language sub-model and provide feedback on response instruction or text response information. As shown in FIG 4, the following provides a detailed introduction to an implementation method of the aforementioned step 2003 broken down to steps 401-406.

At step 401, the text processing module 201 is configured to receive the text information sent by the terminal processing unit 101 and send the text information to the language response module 202.

As a possible implementation, the text processing module 201 comprises a communication protocol with the client 100. After receiving the text information sent by the client 100, the text information can be formatted and sent to the language response module 202.

Due to the slow speed of semantic understanding of text information directly by the language sub-model, in order to improve the response speed of the voice interaction system to customers, as another possible implementation, the text processing module 201 is configured to perform preliminary classification on the text information. If the text information is instruction type information, the text processing module 201 is configured to feedback to the terminal processing unit 101. If the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response model 202, which processes the text information and generate a text response information. As shown in FIG 5, step 401 can comprise steps 501-504.

At step 501, the text processing module 201 is configured to receive the text information sent by the terminal processing unit 101, classify the text information, and obtain the classification result.

- 5 In the embodiment, the text processing module 210 is configured to obtain classification result based on the text content of the text information. In the embodiment, the text content of the text information may comprise Q&A type information or instruction type information. When the text content of the text information comprises instruction type information, the client 100 needs to perform the
- 10 instruction corresponding to the instruction type information. For example, if the text content of the text information comprises instruction type information "sing a song", the client needs to perform the instruction "sing" corresponding to "sing a song". When the text content of the text information comprises Q&A type information, the client needs to answer the user's question. For example, if the text content of the text
- 15 information comprises Q&A type information such as "how to get to the boarding gate", the client 100 needs to answer the question "how to get to the boarding gate". In practical operation, the text processing module 201 can be configured to classify the text content of text information through keyword recognition, relatively simple classification models, or other methods. The embodiment is not limited here. The voice
- 20 interaction method provided in the application can accelerate the response speed of the voice interaction system to customers while maintaining the quality of answers by classifying text information relatively simply in advance. If it is instruction type information, it will no longer be recognized through subsequent language models.
- 25 In some embodiments, the text processing module 201 is configured to store a series of instruction keywords. When the text content of the text information hits the series of keywords, it is determined that the text information is instruction type information, that is, the category of the text information is instruction. The text processing module 201 is configured to perform the following step 502 of determining the response instruction
- 30 that matches the instruction, or any combination of the response instruction and the following content: text response information corresponding to the language identifier, image response information, and returning the above response information to the client 100 (such as response instruction, or any combination of response instruction and text response information/image response information). When the text content of

the text information does not match the keywords of the series, it is determined that the text information is Q&A type, that is, the text information category is Q&A. The text processing module 201 is also configured to perform the following step 504 of inputting the text information to the language model to obtain the text response information corresponding to the language identifier.

In other embodiments, the text processing module 201 is configured to store keywords for various languages which corresponding to the language identifier, and correspondingly set or store response instruction. When the text content of a text information hits a keyword identified by a language identifier, it is determined that the text information is instruction type information, then perform the following step 502 of determining the response instruction corresponding to the matching keyword, and return the response instruction to the client 100 to perform the action indicated by the response instruction. For example, the text processing module 201 is configured to store keywords such as "dance" in Japanese and "sing a song" in English, where "dance" corresponds to the response instruction of "dance" and "sing a song" corresponds to the response instruction of "sing". When the corresponding text information "Please sing a song" matches "sing a song" based on the English voice information "Please sing a song", it is determined that the text information "Please sing a song" is instruction type information, and the client needs to perform the response instruction "sing". The text processing module 201 is configured to return the response instruction "sing" to the terminal processing unit 101 of the client 100, and the voice processing unit 101 is configured to drive the voice playback unit 103 to play the song.

When the text content of a text information does not match the keyword of any language identifier, it is determined that the text information is Q&A type information. The text processing module 201 is configured to perform the following step 504 of sending the text information to the language response model 202, and the language response model 202 obtains the text response information corresponding to the language identifier through the language model.

At step 502, if the classification result indicates that the text information is instruction type information, the text processing module 201 sends the response instruction corresponding to the text information to the terminal processing unit 101.

It should be noted that if the classification result indicates that the text information is instruction type information, the text processing module 201 of the first server 200 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101.

Due to the fact that the response information corresponding to the generated instruction type information 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 text response information corresponding to the instruction type information 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. The response information corresponding to instruction type information comprises response instruction, or any combination of response instruction with the following content: text response information, image response information.

At step 503, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or control the display unit 104 to play content based on the response instruction.

In the embodiment, the terminal processing unit 101 is configured to play the audio corresponding to the response instruction through the voice playback unit 103 based on the response instruction. The terminal processing unit 101 is also configured to invoke the display content corresponding to the response instruction and play the display content through the display unit 104. It should be noted that the terminal processing unit 101 is configured to store a mapping relationship between response instruction and audio, as well as a mapping relationship between response instruction and display content. When the terminal processing unit 101 is configured to receive the response instruction, it invokes the audio and display content corresponding to the response instruction for playback based on the response instruction. For example, when the terminal processing unit 101 receives the response instruction A, it invokes the audio A and display content A corresponding to the response instruction A for playback.

The terminal processing unit 101 is configured to generate text information based on the voice information, which can comprise language identifier. Language identifier identifies the language category of the voice information, so that the response instruction generated based on the text information can also comprise language identifier. The terminal processing unit 101 is configured to display content or audio corresponding to the language category based on language identifier for display or playback, thereby enabling voice objects in different languages to enjoy the highest interaction efficiency experience. For example, when the language identifier identifies the language as Arabic, the display unit 104 is configured to display interaction objects that match Arabic for communication with voice objects. The terminal processing unit 101 is also configured to display the language category corresponding to the language identifier on the display unit through elements such as images or text, to prompt the language category used by the voice object in the interaction process, so that the voice object can effectively receive and understand response information, improve the effectiveness and efficiency of voice interaction, and enhance the interaction experience.

At step 504, if the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202.

If the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202. The language response module 202 is configured to input text information into the language model to obtain text response information. Afterwards, the language response module 202 is configured to return the text response information to the text processing module 201, which sends the text response information to the terminal processing unit 101. The language response module 202 in the embodiment comprises or is connected to the language model. In one example, the language model is stored in the language response module. In another example, the language model is set separately, which is connected to the language response module 202 and can interact with each other.

In some practical operation, text processing modules may recognize text information that is partially instruction type information as Q&A type information while ensuring speed. 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, the embodiment will perform a secondary classification processing on the received text information through the language model after being processed by the text processing module 201, thereby increasing the accuracy of text information classification while maintaining response speed, as shown in the steps 402-404.

At step 402, the language response module 202 is configured to use the target language sub-model corresponding to language identifier to semantically understand text information and obtain understanding results.

In one example, the language response module 202 is configured to input text information into the target language sub-model corresponding to the language identifier, and the target language sub-model will perform semantic understanding of the text information to obtain the understanding result. Understanding results can indicate the classification result of text information, comprising but not limited to Q&A type information and instruction type information. In the embodiment of the application, the classification of text information can be determined based on the text content of the text information, for example, the text content may comprise Q&A type information or instruction type information. When the text content of the text information comprises instruction type information, the client 100 is configured to perform the instruction corresponding to the instruction type information. For example, if the text content of the text information comprises instruction type information "sing a song", the client 100 is configured to perform the instruction "sing" corresponding to "sing a song". When the text content of the text information comprises Q&A type information, the client 100 is configured to answer the user's question. For example, if the text content of the text information comprises Q&A type information such as "how to get to the boarding gate", the client 100 is configured to answer the question "how to get to the boarding gate".

At step 403, if the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to send the

response instruction corresponding to the text information to the terminal processing unit 101 through the text processing module 201.

5 If the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to generate response instruction corresponding to the text information through the language model. After inputting text information into the target language sub-model corresponding to the language identifier, the target language sub-model can output response instruction corresponding to the text information. For example, after inputting the text information
10 as "have a dance" into the target language sub-model, the target language sub-model outputs the response instruction "command: dance". Then, the language response module 202 is configured to send the response instruction "command: dance" to the text processing module 201. The text processing module 201 is configured to process the response instruction based on the communication protocol with the client (such as
15 message encapsulation), and send the processed response instruction to the terminal processing unit 101.

At step 404, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or control the display unit to play content based
20 on the response instruction.

After receiving the response instruction sent by the language response module 202 through the text processing module 201, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or display unit 104 to play
25 display content based on the response instruction as per step 503 above. The embodiments of the application will not be further elaborated here.

At step 405, if the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to generate text
30 response information corresponding to the text information by identifying the language category represented by the language identifier through the target language sub-model.

If the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to generate text response information corresponding to the text information through the target language sub-model. After inputting the text information into the target language sub-model corresponding to the language identifier, the target language sub-model can output the text response information corresponding to the text information. For example, after inputting the text information "Where is the restroom?" into the corresponding target language sub-model in English, the target language sub-model outputs the text response information (in English) "The restroom is 500 meters ahead of you."

At step 406, the language response module 202 is configured to send text response information to the terminal processing unit 101 through the text processing module 201.

After the language response module 202 is configured to obtain the text response information corresponding to the text information, the language response module 202 can send the text response information to the terminal processing unit 101 through the text processing module 201. For example, the language response module 202 is configured to send the text response information "The restroom is 500 meters ahead of you" corresponding to the text information "Where is the restroom?" to the text processing module 201. The text processing module 201 is configured to process the text response information (such as message encapsulation) based on the communication protocol with the client 100, and send the processed text response information to the terminal processing unit 101.

In order to understand the voice interaction method provided in the embodiment better, the embodiment also provides the embodiment to facilitate understanding of the voice interaction method in the application.

As shown in FIG 6, after the voice object outputs voice, the voice receiving unit 102 is configured to collect the voice information output by the voice object and sends the voice information to the terminal processing unit 101. Correspondingly, the camera unit 105 is also configured to collect image information of the voice object and send the image information of the voice object to the terminal processing unit 101. The

terminal processing unit 101 intercepts the 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. As an example, when the voice object is a user, the terminal processing unit 101 is configured to determine whether the user has a need or intention to interact with the client 100 when the image information of the voice object indicates that the user's eyes are fixed on the client 100, or when the user's face is facing the client 100, intercept the voice information output by the user at the moment, and obtain key voice information. When the user is not looking at the client 100 or not facing the client 100, it is determined that the user has no need or intention to interact with the client 100. At the moment, the voice information output by the user is non critical voice information, and subsequent recognition of this non critical voice information can be omitted. By recognizing key voice information, it is possible to obtain and recognize the voice information required for user interaction, without the need to decode all input voice information at all times, greatly saving the resources of terminal processing unit 101 and improving the recognition rate of key voice information during user interaction; at the same time, it also avoids the first server 200 from processing the text information of voice information input at all times, improving the efficiency of the first server 200.

The text information obtained by the terminal processing unit 101 comprises the text content corresponding to the voice information and the language identifier corresponding to the voice information. The terminal processing unit 101 is configured to send the obtained text information to the text processing module 201, which classifies the text information and obtains the classification result.

If the classification result indicates that the text information is instruction type information, the text processing module 201 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101. The response instruction also comprises language identifier. The terminal processing unit 101 is configured to control the voice playback unit 103 to play the audio corresponding to the response instruction, and control the display unit 104 to play the display content corresponding to the response instruction. Wherein, the language used for the audio corresponding to the response instruction can match the language identifier. The image of the interaction object in the displayed content can match the

language identifier, and the displayed content can also comprise the language category corresponding to the language identifier directly displayed through images or text. In one example, the display unit 104 is configured to play the display content corresponding to the response instruction comprises controlling the virtual human interaction object in the display unit 104 to perform the action corresponding to the response instruction.

If the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202. The language response module 202 is configured to select the target language sub-model corresponding to the language identifier from several language sub-models in the language model based on the language identifier in the text information, and input the text information into the target language sub-model for semantic understanding and obtains the understanding result. If the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to obtain the corresponding response instruction of the text information through the target language sub-model. The language response module 202 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101 through the text processing module 201. The terminal processing unit 101 is configured to control the voice playback unit 103 to play the audio corresponding to the response instruction, and control the display unit 104 to play the display content corresponding to the response instruction. Wherein, the language used for the audio corresponding to the response instruction can match the language identifier. The image of the interaction object in the displayed content can match the language identifier, and the displayed content can also comprise the language identifier corresponding to the language identifier directly displayed through images or text.

If the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to obtain the text response information through the target language sub-model. Language identifier can still be comprised in the text response information. The language response module 202 is configured to send the text response information to the terminal processing unit 101 through the text processing module 201. The terminal processing unit 101 is

configured to control the voice playback unit 103 to play the audio corresponding to the text response information based on the text response information. The terminal processing unit 101 is configured to control the display unit 104 to display the text content corresponding to the text response information, and control the display of the interaction object of the display unit 104 to display the image corresponding to the language identifier, as well as control the matching of the gesture and mouth shape of the interaction object with the audio played by the voice playback unit 103. The terminal processing unit 101 is also configured to display the language category corresponding to the language identifier on the display unit 104 through images or text based on the language identifier, to remind the user of the language category currently used by the interaction object. For example, multiple language identifier elements are displayed on the display unit, comprising language identifier "日本語", language identifier "English", and language identifier "بيبي", which are used to represent Japanese, English, and Arabic, respectively. As another example, multiple language identifiers may comprise language identifier "中文", language identifier "English", language identifier "بيبي", and language identifier "日本語", respectively, used to represent Chinese, English, Arabic, and Japanese. When the display unit 104 is configured to display the text information "Hello!" corresponding to the language identifier "English", and the voice playback unit 103 is configured to play the audio of the text information "Hello!" in the language identifier "English", the element corresponding to the language identifier "English" is highlighted. In some examples, the highlighted way is that the element corresponding to the language identifier "English" is framed by the focus box; in another example, highlighting is done by controlling the element corresponding to the language identifier "English" to blink.

In summary, the voice interaction method provided in the embodiment of the application converts voice information into text information, and the server 200 selects the target language sub-model corresponding to the language category of the text information from one or more language sub-model to process the text information and obtain the text response information. Then the terminal processing unit 101 is configured to respond to the user based on the text response information. Due to the fact that the language model in the application is divided into one or more language sub-model, which only supports one or a few languages, the language sub-model in

the application has a smaller parameter level compared to the use of complete large language models in related technologies, faster processing speed for text information, and faster generation of text response information. Therefore, while maintaining the quality of the voice interaction method in the application, the response speed of the voice interaction system to customers can be accelerated.

According to the aforementioned embodiments, another voice interaction method is also provided in the application. As shown in FIG 7, the voice interaction method provided in the embodiment is applied to a client 100. In one example, combined with FIG 1, the client 100 comprises a terminal processing unit 101, a voice receiving unit 102, and a voice playback unit 103. The client 100 communicates and connects with the first server 200, which comprises or is configured to connect to the language model. The language model comprises one or more language sub-model, which correspond to one or more languages. The method comprises at least steps 701-703.

At step 701, the voice receiving unit 102 is configured to collect voice information and send it to the terminal processing unit 101.

In the embodiment, the object of outputting voice is a voice object, and in some embodiments, the voice object is a user or device. Wherein, the device has voice playback function, such as smartphones, robots, etc. In some examples of the application, taking the user as an example of the voice object, the user outputs voice information, and the voice receiving unit 102 is configured to collect the voice information and send it to the terminal processing unit 101.

At the step 702, the terminal processing unit 101 is configured to obtain the text information corresponding to the voice information based on the voice information, send the text information to the first server 200, so that the first server 200 is configured to input the text information to the target language sub-model in the language model, and obtain the text response information corresponding to the text information returned by the target language sub-model. The target language sub-model corresponds to the language category of the text information, and the language category of the text information and the text response information is the same.

In the embodiment, the process of obtaining the text information corresponding to the voice information is called decoding the voice information. In some examples, decoding the voice information is completed by the client 100 to obtain the text information corresponding to the voice information. In another example, the decoding
5 of voice information is completed by the first server 200. In the specific implementation, the terminal processing unit 101 is configured to upload the voice information to a second server (not shown in the figure), and the second server is configured to decode the voice information. For example, a voice to text module of the second server is configured to decode the voice information to obtain the corresponding text information,
10 and the second server is configured to send the text information to the terminal processing unit 101. The embodiment does not restrict the execution subject of decoding voice information.

In a practical operation, decoding voice information can obtain the corresponding text
15 content and language identifier of the voice information. That is, the text information in the application comprises text content and language identifier, where the language identifier identifies the language category of the voice information. Assuming that the client 100 is configured to support language categories comprising Japanese, English, and Arabic, the language identifier comprises language identifier A corresponding to
20 English, language identifier B corresponding to Japanese, and language identifier C corresponding to Arabic. In one example, when the language corresponding to the voice message is Arabic, the text information may comprise the language identifier C corresponding to Arabic.

25 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 for further processing. In practical operation, the client 100 can optionally comprise a camera unit 105. The camera unit 105 is configured to collect image information of the voice object, and the image information of the voice object is
30 configured to assist the terminal processing unit 101 in processing the voice information. In one example, the terminal processing unit 101 can 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 terminal 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 terminal processing unit 101 is configured to extract key voice information from the voice information collected by the voice receiving unit 102, without the need to decode all the collected voice information, reducing the computational power required for decoding voice information and saving hardware resources.

In some possible implementations, the text information comprises language identifier, which identifies the language category of voice information. Language identifier is used to input text information to the target language sub-model corresponding to the language identifier in the language model by the first server 200, and obtain the text response information corresponding to the language identifier.

At step 703, the terminal processing unit 101 is configured to obtain the corresponding voice response information based on the text response information, and control the voice playback unit 103 to play the voice response information in the language category of the text information.

The terminal processing unit 101 is configured to receive text response information, generate voice response information corresponding to the text response information, and control the voice playback unit 103 to play the voice response information. In one example, the terminal processing unit 101 is configured to recognize the language category of the text response information and generate voice response information of the same language category. In another example, the text response information comprises language identifier, and the terminal processing unit 101 is configured to generate corresponding language category voice response information based on the language identifier. In the specific implementation, the terminal processing unit 101 is configured to invoke the TTS (Text To voice) function to generate the corresponding voice response information for the text response information. The process of converting text information into voice information in the embodiments of the application can also be completed by the server 200, and the embodiments of the application are not limited here. As a possible implementation, the terminal processing unit 101 is configured to upload text information to a second server, which is different from the first server 200. The text to voice module of the second server processes the voice

information, and the terminal processing unit 101 is configured to obtain the voice information returned by the second server.

In some practical operation, the client 100 may comprise a display unit 104, and the terminal processing unit 101 can be configured to control the display unit 104 to play display content based on text response information. The display unit 104 is configured to display interactive objects that interact with voice objects, such as virtual human objects, intelligent robot objects, etc. The terminal processing unit 101 is configured to control the display unit 104 to display the display content corresponding to the text response information based on the text response information. As an example, the terminal processing unit 101 in the application is configured to control the display unit 104 to display the text corresponding to the text response information, and control the interaction object on the display unit 104 to interact with the voice object based on the text response information, control the interaction object displayed on the display unit 104 to perform the interaction action corresponding to the text response information, such as matching the mouth shape or gesture of the interaction object with the audio played by the voice playback unit 103, thereby facilitating the voice object to understand the reply content from multiple perspectives such as visual and auditory, and improving the user experience of the voice object.

As an example, the text response information may comprise the language identifier, and the terminal processing unit 101 may also be configured to display the language category corresponding to the language identifier on the display unit 104 through elements such as images or text based on the language identifier, to remind the voice object of the language category used in the interaction process. For example, the terminal processing unit 101 is configured to control the display unit 104 to display language identifier for text response information.

The solution provided in the embodiment automatically recognizes the language category and text information of the input voice information, processes the text information of the language category through the target language sub-model corresponding to the language category, and returns text response information of the same language category. Due to the fact that the language category of text response information is the same as that of voice response information, the terminal processing

unit 101 is configured to generate voice response information that is the same as the input language category, and play the same language category of voice response information through the voice playback unit. Through the implementation example of the application, it is possible to achieve a full voice communication interaction process of "input certain language of voice information → play same language of voice response information", such as the voice object inputting English voice information "Hello!", and the client 100 being configured to play English voice response information "Hello, Dear guest!".

In some possible embodiments, the client 100 may also comprise a camera unit, which is configured to capture image information of the voice object. In some examples, after the voice object outputs voice, the voice receiving unit 102 is configured to collect the voice information output by the voice object and send the voice information to the terminal processing unit 101. Correspondingly, the camera unit 105 is also configured to collect image information of the voice object and send the image information of the voice object to the terminal processing unit 101. The terminal processing unit 101 is configured to intercept the voice information based on the image information of the voice object to obtain key voice information, and decode the key voice information to obtain the corresponding text information of the key voice information. As an example, when the voice object is a user, the terminal processing unit 101 is configured to determine whether the user has a need or intention to interact with the client 100 when the image information of the voice object indicates that the user's eyes are fixed on the client 100, or when the user's face is facing the client 100, intercept the voice information output by the user at the moment, and obtain key voice information. When the user is not looking at the client 100 or not facing the client 100, it is determined that the user has no need or intention to interact with the client 100. At the moment, the voice information output by the user is non critical voice information, and subsequent recognition of the non critical voice information can be omitted. By recognizing key voice information, it is possible to obtain and recognize the voice information required for user interaction, without the need to decode all input voice information at all times, greatly saving the resources of terminal processing unit 101 and improving the recognition rate of key voice information during user interaction; at the same time, it also avoids the first server 200 from processing the text information of voice information input at all times, improving the efficiency of the first server 200.

According to the voice interaction method provided in the above embodiments, the application also provides a voice interaction method, combined with FIG 1, as shown in FIG 8, the voice interaction method provided in the embodiment is applied to the first server 200, which is configured to communicate and connect with the client 100. The client 100 comprises at least a terminal processing unit 101, a voice receiving unit 102, and a voice playback unit 103. The specific description of the client 100 can be found in the aforementioned embodiments and will not be repeated here. The first server 200 comprises or connects the language model, which comprises one or more language sub-model corresponding to one or more languages. The voice interaction method shown in FIG 8 comprises at least steps 801-803.

At step 801, the first server 200 is configured to receive the text information sent by the terminal processing unit 101.

The terminal processing unit 101 is configured to obtain text information corresponding to the voice information based on the voice information collected by the voice receiving unit 102. The process of obtaining text information by the terminal processing unit 101 can be found in the aforementioned step 2002, and will not be repeated here.

At step 802, the first server 200 is configured to input text information to the target language sub-model in the language model to obtain the text response information corresponding to the text information. The target language sub-model corresponds to the language category of the text information, and the language category of the text information and the text response information are the same.

In most cases, large language models which support multiple languages at the same time have a larger volume and slower understanding and response speed to textual information. For example, large language models which support multiple languages at the same time typically have parameter magnitudes of billions, which is a very large parameter volume. In order to improve the response speed of the language model as much as possible, the embodiment divides the language model into one or more relatively independent language sub-model based on the language category. 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 large language model that supports multiple languages simultaneously, the language sub-model that supports one or a small number of languages in the embodiment has a smaller volume and a faster response speed to text information.

In certain embodiments of the application, the language model comprises one or more language sub-model, each corresponding to one or more language categories, i.e. each language sub-model is used to process input text information for the corresponding language category and return text response information for the corresponding language category. For example, the language model consists of three language sub-model, namely language sub-model 1, language sub-model 2, and language sub-model 3. Wherein, language sub-model 1 corresponds to Japanese, language sub-model 2 corresponds to English and French, and language sub-model 3 corresponds to Thai. Therefore, language sub-model 1 is used to process the input Japanese information and return the corresponding Japanese response information; language sub-model 2 is used to process input English information, return corresponding English response information, as well as to process input French information and return corresponding French response information; similarly, language sub-model 3 is used to process input Thai information and return corresponding Thai response information. It should be understood that the embodiments of the application do not limit the types of languages, and the specific language can be set according to the actual situation.

In one example, the text information comprises a language identifier. The first server 200 is configured to input the text information to the target language sub-model corresponding to the language identifier in the language model, and obtain the text response information corresponding to the language identifier. For example, the language model in the application comprises language sub-model A, language sub-model B, and language sub-model C. Language sub-model A supports English, language sub-model B supports Japanese, and language sub-model C supports Arabic. The language identifier in the embodiment comprises 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. The user outputs the language "hello!", and the terminal processing unit 101 is configured to obtain the language identifier A in the text information based on the voice information collected by the voice receiving unit 102. After receiving the text information containing the language identifier A, the language model inputs the text information into the language sub-model A to obtain the text response information. The language identifier A can also be comprised in the text response information.

The solution provided in the embodiment can select a matching target language sub-model from one or more language sub-models in the language model based on the language category of the voice information, input the text information corresponding to the voice information to the target language sub-model, and return the text response information corresponding to the text information through the target language sub-model. The text response information returned by the target language sub-model has the same language category as the text information and voice information. It can be seen that the embodiment automatically recognizes the language category and text information of the input voice information, processes the text information of the language category through the target language sub-model corresponding to the language category, and returns text response information of the same language category. The target language sub-model, due to its smaller parameter size compared to using a complete large language model in related technologies, has a faster processing speed for text information and a faster generation speed for text response information, allow to quickly return text response information of the same language category, which can accelerate the response speed of the voice interaction system to the voice object while maintaining the quality of the voice interaction method in the application, and improve the response speed and efficiency of the entire interaction process.

At step 803, the first server 200 is configured to send text response information to the terminal processing unit 101, so that the terminal processing unit 101 is configured to obtain the voice response information corresponding to the text response information, and control the voice playback unit to play the voice response information.

After receiving the text response information returned by the target language sub-model, the first server 200 is configured to send the text response information to the terminal processing unit 101.

- 5 The terminal processing unit 101 is configured to obtain the corresponding voice response information based on the text response information, and control the voice playback unit 103 to play the voice response information in the language category of the text information. Refer to the step 2004 for the process, which will not be repeated here.

10

In some possible embodiments, the first server 200 comprises a language response module 202, which comprises or connects to the language model. The first server 200 is configured to input text information to the target language sub-model corresponding to the language identifier in the language model to obtain the text response information corresponding to the language identifier, comprising:

15

the language response module 202 configured to input text information to the target language sub-model corresponding to the language identifier in the language model, and obtain the text response information corresponding to the language identifier.

20

In some possible embodiments, the language response module 202 is configured to input text information to the target language sub-model corresponding to the language identifier in the language model, comprising:

- 25 the language response module 202 is configured to use the target language sub-model corresponding to the language identifier in the language model to semantically understand the text information and obtain the understanding result;

if the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to generate text response information corresponding to the text information by identifying the language category represented by the language identifier through the target language sub-model.

30

In some possible embodiments, the first server 200 further comprises a text processing module 201, and the method further comprises:

5 if the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to send the corresponding response instruction of the text information to the terminal processing unit 101 through the text processing module 201.

In some possible embodiments, the method further comprises:

10

the text processing module 201 is configured to receive the text information sent by the terminal processing unit 101, classify the text information, and obtain the classification result;

15 if the classification result indicates that the text information is instruction type information, the text processing module 201 is configured to send the corresponding response instruction of the text information to the terminal processing unit;

20 if the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202.

In order to provide a clearer explanation of the technical solution for voice interaction method based on the first server 200, combined with FIGs 3, 4 and 5, the first server 25 200 in the embodiment may optionally comprise a language response module 202 and a text processing module 201. The language response module 202 comprises or is configured to connect to the language model, and the text processing module 201 stores a communication protocol with the client. As a possible implementation, the first server 200 in the embodiment can directly understand the semantic information of the 30 text through the language sub-model and provide feedback on response instruction or text response information. As shown in FIG 4, the following provides a detailed introduction to an implementation method of the aforementioned step 2003 through steps 401-405.

At step 401, the text processing module 201 is configured to receive the text information sent by the terminal processing unit 101 and send the text information to the language response module 202.

- 5 As a possible implementation, the text processing module 201 in the application is configured to store a communication protocol with the client 100. After receiving the text information sent by the client 100, the text information is processed and sent to the language response module 202.
- 10 Due to the slow speed of semantic understanding of text information directly by the language sub-model, in order to improve the response speed of the voice interaction system to customers, as another possible implementation, the text processing module 201 is configured to classify the text information preliminary. If the text information is instruction type information, the text processing module 201 is configured to feedback
- 15 to the terminal processing unit 101. If the text information is Q&A type information, the text processing module 201 is configured to sends the text information to the language response model 202, which processes the text information and generates a text response information. As shown in FIG 5, it specifically comprises the steps 501-504.
- 20 At step 501, The text processing module 201 is configured to receive the text information sent by the terminal processing unit 101, classify the text information, and obtain the classification result.

In the embodiment, the text processing module 201 is configured to obtain

25 classification result based on the text content of the text information. In the embodiment, the text content of the text information may comprise Q&A type information or instruction type information. When the text content of the text information comprises instruction type information, the client needs to perform the instruction corresponding to the instruction type information. For example, if the text

30 content of the text information comprises instruction type information "sing a song", the client needs to perform the instruction "sing" corresponding to "sing a song". When the text content of the text information comprises Q&A type information, the client needs to answer the user's question. For example, if the text content of the text information comprises Q&A type information such as "how to get to the boarding gate",

the client needs to answer the question "how to get to the boarding gate". In practical operation, the text processing module 201 is configured to classify the text content of text information through keyword recognition, relatively simple classification models, or other methods. The embodiment is not limited here. The voice interaction method
5 provided in the application can accelerate the response speed of the voice interaction system to customers while maintaining the quality of answers by classifying text information preliminary in advance. If it is instruction type information, it will no longer be recognized through language models.

10 In some embodiments, the text processing module 201 is configured to store a series of instruction keywords. When the text content of the text information hits the series of keywords, it is determined that the text information is instruction type information, that is, the category of the text information is instruction. The text processing module 201 is configured to perform the following step 502 of determining the response instruction
15 that matches the instruction, or any combination of the response instruction and the following content: text response information corresponding to the language identifier, image response information, and returning the above response information to the client 100 (such as response instruction, or any combination of response instruction and text response information/image response information). When the text content of
20 the text information does not match the keywords, it is determined that the text information is Q&A type, that is, the text information category is Q&A. The text processing module 201 is configured to perform the step 504 of inputting the text information to the language model to obtain the text response information corresponding to the language identifier.

25 In other embodiments, the text processing module 201 is configured to store keywords for various languages that correspond to language identifier and corresponding response instruction are set or stored. When the text content of a text information hits any keyword identified by the language identifier, it is determined that the text
30 information is instruction type information, perform the step 502 to determine the response instruction corresponding to the matching keyword, and return the response instruction to the client 100 to perform the action indicated by the response instruction. For example, the text processing module 201 is configured to store keywords such as "dance" in Japanese and "sing a song" in English, where "dance" corresponds to the

response instruction of "dance", and "sing a song" corresponds to the response instruction of "sing". When the corresponding text information "Please sing a song" matches "sing a song" based on the English voice information "Please sing a song", it is determined that the text information "Please sing a song" is instruction type information, and the client needs to perform the response instruction "sing". The text processing module 201 is configured to feedback the response instruction "sing" to the terminal processing unit 101 of the client 100, and the voice processing unit 101 is configured to drive the voice playback unit 103 to play the song.

When the text content of a text information does not match the keyword of any language identifier, it is determined that the text information is Q&A type information. The text processing module 201 is configured to performs the step 504 of sending the text information to the language response model 202, and the language response model 202 being configured to obtain the text response information corresponding to the language identifier through the language model.

At step 502, if the classification result indicates that the text information is instruction type information, the text processing module 201 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101.

It should be noted that if the classification result indicates that the text information is instruction type information, the text processing module 201 of the first server 200 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101.

Due to the fact that the response information corresponding to the generated instruction type information 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 text response information corresponding to the instruction type information 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. The response information corresponding to instruction type information comprises response instruction, or any combination of

response instruction with the following content: text response information, image response information.

At step 503, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or control the display unit to play display content based on the response instruction.

In the embodiment, the terminal processing unit 101 is configured to play the audio corresponding to the response instruction through the voice playback unit 103 based on the response instruction. The terminal processing unit 101 is also configured to invoke the display content corresponding to the response instruction and play the display content through the display unit 104. It should be noted that the terminal processing unit 101 is configured to store a mapping relationship between response instruction and audio, as well as a mapping relationship between response instruction and display content. When the terminal processing unit 101 is configured to receive the response instruction, it invokes the audio and display content corresponding to the response instruction for playback based on the response instruction. For example, when the terminal processing unit 101 is configured to receive the response instruction A, it invokes the audio A and display content A corresponding to the response instruction A for playback.

The terminal processing unit 101 is configured to generate text information based on the voice information, which can comprise language identifier. Language identifier identifies the language category of the voice information, so that the response instruction generated based on the text information can also comprise language identifier. The terminal processing unit 101 is configured to select display content or audio corresponding to the language category based on language identifier for display or playback, thereby enabling voice objects in different languages to enjoy the highest interaction efficiency experience. For example, when the language identifier identifies the language category as Arabic, the display unit 104 can display interaction objects that match Arabic for communication with voice objects. The terminal processing unit 101 is configured to also display the language category corresponding to the language identifier on the display unit 104 through elements such as images or text, to prompt the language category used by the voice object in the interaction process, so that the

voice object can effectively receive and understand response information, improve the effectiveness and efficiency of voice interaction, and enhance the interaction experience.

- 5 At step 504, if the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202.

10 If the classification result indicates that the text information is Q&A type information, the text processing module 201 is configured to send the text information to the language response module 202. The language response module 202 is configured to input text information into the language model to obtain text response information. Afterwards, the language response module 202 is configured to return the text response information to the text processing module 201, which sends the text
15 response information to the terminal processing unit 101. The language response module 202 is configured to store or is connected to the language model. In one example, the language model is stored in the language response module. In another example, the language model is set separately, which is connected to the language response module 202 and can interact with each other.

20

In some practical operation, text processing modules may recognize text information that is partially instruction type information as Q&A type information while ensuring speed. Although the language model in the application has a slower response speed compared to the text processing module, it has better comprehension and judgment
25 abilities for text information. Therefore, in some possible embodiments, there is a secondary classification processing on the received text information through the language model after being processed by the text processing module 201, thereby increasing the accuracy of text information classification while maintaining response speed, as shown in the steps 402-404.

30

At step 402, the language response module 202 is configured to use the target language sub-model corresponding to language identifier to semantically understand text information and obtain understanding results.

In one example, the language response module 202 is configured to input text information into the target language sub-model corresponding to the language identifier, and the target language sub-model will perform semantic understanding of the text information to obtain the understanding result. Understanding results can indicate the classification result of text information, comprising but not limited to Q&A type information and instruction type information. In the embodiment of the application, the classification of text information can be determined based on the text content of the text information, for example, the text content may comprise Q&A type information or instruction type information. When the text content of the text information comprises instruction type information, the client needs to perform the instruction corresponding to the instruction type information. For example, if the text content of the text information comprises instruction type information "sing a song", the client needs to perform the instruction "sing" corresponding to "sing a song". When the text content of the text information comprises Q&A type information, the client needs to answer the user's question. For example, if the text content of the text information comprises Q&A type information such as "how to get to the boarding gate", the client needs to answer the question "how to get to the boarding gate".

At step 403, if the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to send the response instruction corresponding to the text information to the terminal processing unit 101 through the text processing module 201.

If the understanding result indicates that the text information is instruction type information, the language response module 202 is configured to generate response instruction corresponding to the text information through the language model. After inputting text information into the target language sub-model corresponding to the language identifier, the target language sub-model can output response instruction corresponding to the text information. For example, after inputting the text information as "have a dance" into the target language sub-model, the target language sub-model outputs the response instruction "command: dance". Then, the language response module 202 is configured to send the response instruction "command: dance" to the text processing module 201. The text processing module 201 is configured to process the response instruction based on the communication protocol with the client (such as

message encapsulation), and send the processed response instruction to the terminal processing unit 101.

5 At step 404, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or display content based on the response instruction.

10 After receiving the response instruction sent by the language response module through the text processing module 201, the terminal processing unit 101 is configured to control the voice playback unit 103 to play audio and/or display unit 104 to play display content based on the response instruction. Refer to step 503 for details, the embodiments of the application will not be further elaborated here.

15 At step 405, if the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to generate text response information corresponding to the text information by identifying the language category represented by the language identifier through the target language sub-model.

20 If the understanding result indicates that the text information is Q&A type information, the language response module 202 is configured to generate text response information corresponding to the text information through the target language sub-model. After inputting the text information into the target language sub-model corresponding to the language identifier, the target language sub-model can output
25 the text response information corresponding to the text information. For example, after inputting the text information "Where is the restroom?" into the corresponding target language sub-model in English, the language sub-model outputs the text response information (in English) "The restroom is 500 meters ahead of you."

30 At step 406, the language response module 202 is configured to send text response information to the terminal processing unit 101 through the text processing module 201.

After the language response module 202 is configured to obtain the text response information corresponding to the text information, the language response module 202 is configured to send the text response information to the terminal processing unit 101 through the text processing module 201. For example, the language response module 5 202 is configured to send the text response information "The restroom is 500 meters ahead of you" corresponding to the text information "Where is the restroom?" to the text processing module 201. The text processing module 201 is configured to process the text response information (such as message encapsulation) based on the communication protocol with the client 100, and send the processed text response 10 information to the terminal processing unit 101.

It should be noted that the voice interaction method applied in the first server 200 of the embodiment can achieve the various processes of the aforementioned implementation of the voice interaction method and achieve the same effect and 15 function, which will not be repeated here.

In the description of the specification, the reference to terms such as "some possible embodiments", "some embodiments", "examples", "specific examples", or "some examples" means that the specific features, structures, materials, or features 20 described in combination with the embodiments or examples are included in at least one embodiment or example of the application, and the above terms may not necessarily represent the same embodiment or example. Moreover, the specific features, structures, materials, or features described can be combined in an appropriate manner in any one or more embodiments or examples. In addition, those 25 skilled in the art may combine and mix different embodiments or examples described in this specification, as well as features of different embodiments or examples, without contradicting each other.

The method flowchart of the application describes certain operations as different steps 30 executed in a certain order. The flowchart is explanatory rather than restrictive. Some steps described in the description can be grouped together and executed in a single operation, or some steps can be divided into multiple sub steps and executed in a different order from that shown in the description. The various steps shown in the flowchart can be implemented in any manner by any circuit structure and/or tangible

mechanism, such as software running on a computer device, hardware (e.g., logic functions implemented by a processor or chip), etc., and/or any combination thereof.

Those skilled in the art can understand that in the specific implementation methods described, the order of each step shown does not imply a strict execution order. The specific execution order of each step should be determined by its function and possible internal logic.

According to the voice interaction method provided in the above embodiments, the application also provides a voice interaction system 98, as shown in FIG 3. The voice interaction system 98 comprises a client 100 and a first server 200, wherein the client 100 comprises a voice receiving unit 102, a voice playback unit 103, and a terminal processing unit 101. The first server 200 comprises or is configured to connect the language model, which comprises one or more language sub-model corresponding to one or more languages; wherein:

the voice receiving unit 102 is configured to collect voice information and send voice information to the terminal processing unit 101;

the terminal processing unit 101 is configured to obtain text information corresponding to the voice information based on the voice information, and send the text information to the first server 200;

the first server 200 is configured to input text information into the target language sub-model in the language model, which corresponds to the language category of the text information; obtain the text response information corresponding to the text information returned by the target language sub-model, with the same language category as the text response information; send text response information to the terminal processing unit;

the terminal processing unit 101 is configured to obtain voice response information corresponding to the text response information based on the text response information, and controls the voice playback unit 103 to play the voice response information in the language category of the text information.

In some possible embodiments, text information comprises language identifier, which identifies the language category of voice information;

- 5 the first server 200 is configured to input text information to the target language sub-model corresponding to the language identifier in the language model, and obtain the text response information corresponding to the language identifier.

10 In some possible embodiments, the first server 200 comprises a language response module, which comprises or connects to the language model;

the language response module 202 is configured to input text information to the target language sub-model corresponding to the language identifier in the language model, and obtain the text response information corresponding to the language identifier.

15

In some possible embodiments, the language response module 202 is configured as:

the language response module 202 uses the target language sub-model corresponding to the language identifier in the language model to semantically understand the text information and obtain the understanding result;

20

if the understanding result indicates that the text information is Q&A type information, the language response module 202 generates text response information corresponding to the text information in the language category characterized by the language identifier by using the target language sub-model.

25

In some possible embodiments, the first server 200 further comprises a text processing module 201 and a language response module 202, which are configured as:

30

if the understanding result indicates that the text information is instruction type information, the language response module 202 being configured to send the response instruction corresponding to the text information to the terminal processing unit through the text processing module 201.

In some possible embodiments, the client 100 may also comprise a display unit 104,

the terminal processing unit 101 being configured to control the display unit to display
5 the text corresponding to the text response information in the language category
identified by the language identifier, and control the interaction object displayed on the
display unit to perform the interaction action corresponding to the text response
information.

10 In some possible embodiments, the text response information comprises language
identifier;

the terminal processing unit 101 being configured to control the display unit 104 to
display elements corresponding to the language identifier based on the text response
15 information.

In some possible embodiments, the text processing module 201 is configured as:

the text processing module 201 is configured to receive the text information sent by
20 the terminal processing unit 101 and classify the text information to obtain
classification result;

if the classification result indicates that the text information is instruction type
information, the text processing module 201 is configured to send the response
25 instruction corresponding to the text information to the terminal processing unit 101;

if the classification result indicates that the text information is Q&A type information,
the text processing module 201 is configured to send the text information to the
language response module 202.

30 In some possible embodiments, the terminal processing unit 101 is configured to:

decode the voice information to obtain the text information corresponding to the voice
information;

alternatively, upload the voice information to a second server for decoding and obtain the text information returned by the second server.

5 In some possible embodiments, the client 100 may also comprise a camera unit 105;

the camera unit 105 is configured to capture image information of voice objects;

the terminal processing unit 101 is configured to:

10

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;

obtain text information corresponding to the key voice information based on the key
15 voice information.

It should be noted that the voice interaction system 98 in the embodiment can achieve the same effect and function by implementing the various processes of the
aforementioned voice interaction method, and will not be repeated here.

20

It should be noted that the first server in the embodiment can implement the various processes of the voice interaction method applied to the first server as described above, and achieve the same effect and function, and will not be repeated here.

25

According to some embodiments of the application, the application provides a non-volatile computer storage medium for the client, on which computer executable instructions are stored. These computer executable instructions are configured to execute, when run by a processor, the actions performed by the client in the voice
30 interaction method described in the above embodiments. According to some embodiments of the application, the application also provides a non-volatile computer storage medium for a first server, on which computer executable instructions are stored. These computer executable instructions are configured to execute, when run

by a processor, the actions performed by the first server in the voice interaction method described in the above embodiments.

5 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 embodiments. According to some embodiments of the application, the application also provides a non-volatile computer storage medium on a server, which stores computer
10 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.

15 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
20 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.
25 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
30 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 9000

Referring to FIG 9, a server 9000 is typically administered by an entity operating the aforementioned methods, typically whenever cloud server services are not utilised. The use of the server 9000 is typically necessary in a situation where the entity operating the aforementioned methods desires full control of infrastructure enabling the aforementioned methods. The first server 200 can be illustrated by the server 9000.

The server 9000 typically carries out processes to enable the carrying out of the aforementioned methods while receiving and transmitting relevant data. While the server 9000 is illustrated as a single computing system in FIG 9, it should be appreciated that the server 9000 can be a distributed set-up utilising a plurality of computing systems.

The components of the server 9000 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 9, the server 9000 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 9000 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 9000 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:
- 5 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
 - 10 c. a display adapter 330c, which is connected to a display device 334 such as a liquid-crystal display (LCD) panel device.

The server 9000 includes a plurality of standard software modules, including:

- 15 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
- 20 an SQL database 316.

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

25 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 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

30 344, including markup language (e.g., HTML, XML, Java) scripts, and the like.

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 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 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.

Respective steps of processes of the server 9000 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 server 9000 to perform the functions of the module.

The server 9000 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 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.

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.

5 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 interaction method, characterized in that the voice interaction method is
5 applied to a voice interaction system, wherein the voice interaction system comprises
a client and a first server, the client comprises a voice receiving unit, a voice playback
unit, and a terminal processing unit, the first server comprises or connects to the
language model, the language model comprises one or more language sub-model,
the language sub-model correspond to one or more languages, and the method
10 comprises:
the voice receiving unit collecting voice information and sending the voice information
to the terminal processing unit;
the terminal processing unit obtaining the text information corresponding to the voice
information based on the voice information, and sending the text information to the
15 first server;
the first server inputting the text information to the target language sub-model in the
language model, which the target language sub-model corresponds to the language
category of the text information; obtaining the text response information corresponding
to the text information returned by the target language sub-model, which the text
20 response information has the same language category as the text information; sending
the text response information to the terminal processing unit;
the terminal processing unit obtaining the voice response information corresponding
to the text response information based on the text response information, and controlling
the voice playback unit to play the voice response information in the language category
25 of the text information.
2. The method according to claim 1, characterized in that the text information
comprises a language identifier, which identifies the language category of the voice
information;
30 the first server inputting the text information to the target language sub-model in the
language model; obtaining the text response information corresponding to the text
information returned by the target language sub-model, comprising:

the first server inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and obtaining the text response information corresponding to the language identifier.

5 3. The method according to claim 2, characterized in that the first server comprises a language response module, wherein the language response module comprises or connects to the language model, the first server inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and obtaining the text response information corresponding to the language
10 identifier, comprising:

the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and obtaining the text response information corresponding to the language identifier.

15 4. The method according to claim 3, characterized in that the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and obtaining the text response information corresponding to the language identifier, comprising:

the language response module using the target language sub-model corresponding to
20 the language identifier in the language model to semantically understand the text information and obtaining the understanding result;

if the understanding result indicates that the text information is Q&A type information, the language response module generating text response information corresponding to the text information in the language category characterized by the language identifier
25 by using the target language sub-model.

5. The method according to claim 4, characterized in that the first server further comprises a text processing module, and the method further comprises:

if the understanding result indicates that the text information is instruction type
30 information, the language response module sending the response instruction corresponding to the text information to the terminal processing unit through the text processing module.

6. The method according to claim 2, characterized in that the client further comprises a display unit, and the method further comprises:

the terminal processing unit controlling the display unit to display the text corresponding to the text response information in the language category identified by the language identifier, and controlling the interaction object displayed on the display unit to perform the interaction action corresponding to the text response information.

7. The method according to claim 6, characterized in that the text response information comprises language identifier, and the method further comprises:

the terminal processing unit controlling the display unit to display elements corresponding to the language identifier based on the text response information.

8. The method according to claim 5, characterized in that before the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, the method further comprises:

the text processing module receiving the text information sent by the terminal processing unit and classifying the text information to obtain classification result;

if the classification result indicates that the text information is instruction type information, the text processing module sending the response instruction corresponding to the text information to the terminal processing unit;

if the classification result indicates that the text information is Q&A type information, the text processing module sending the text information to the language response module.

9. The method according to claim 1, characterized in that the terminal processing unit obtains text information corresponding to the voice information based on the voice information, comprising:

the terminal processing unit decoding the voice information to obtain the text information corresponding to the voice information;

alternatively, the terminal processing unit uploading the voice information to the second server for decoding and obtaining the text information returned by the second server.

10. The method according to claim 1, characterized in that the client further comprises a camera unit;

the camera unit is configured for capturing image information of voice objects;

the terminal processing unit obtaining text information corresponding to the voice information based on the voice information collected by the voice receiving unit, comprising:

the terminal processing unit intercepting the voice information collected by the voice receiving unit based on the image information of the voice object to obtain key voice information;

the terminal processing unit obtaining text information corresponding to the key voice information based on the key voice information.

11. A voice interaction method, characterized in that the voice interaction method is applied to a client, which the client and the first server can be communicatively connected; the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; the method comprises:

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

the terminal processing unit obtaining the text information corresponding to the voice information based on the voice information, sending the text information to the first server, so that the first server inputting the text information to the target language sub-model in the language model, and obtaining the text response information corresponding to the text information returned by the target language sub-model; the target language sub-model corresponds to the language category of the text information, which the text response information has the same language category as the text information;

the terminal processing unit obtaining the voice response information corresponding to the text response information based on the text response information, and controlling the voice playback unit to play the voice response information in the language category of the text information.

12. The method according to claim 11, characterized in that the client further comprises a display unit, and the method further comprises:

the terminal processing unit controlling the display unit to display the text corresponding to the text response information in the language category identified by the language identifier, and controlling the interaction object displayed on the display unit to perform the interaction action corresponding to the text response information.

13. The method according to claim 12, characterized in that the text response information comprises language identifier, and the method further comprises:

the terminal processing unit controlling the display unit to display elements corresponding to the language identifier based on the text response information.

14. The method according to claim 11, characterized in that the terminal processing unit obtaining text information corresponding to the voice information based on the voice information, comprising:

the terminal processing unit decoding the voice information to obtain the text information corresponding to the voice information;

alternatively, the terminal processing unit uploading the voice information to the second server for decoding and obtaining the text information returned by the second server.

15. The method according to claim 11, characterized in that the client further comprises a camera unit;

the camera unit is configured for capturing image information of voice objects;

the terminal processing unit obtaining text information corresponding to the voice information based on the voice information collected by the voice receiving unit, comprising:

the terminal processing unit intercepting the voice information collected by the voice receiving unit based on the image information of the voice object to obtain key voice information;

the terminal processing unit obtaining text information corresponding to the key voice information based on the key voice information.

16. A voice interaction method characterized in that the voice interaction method is applied to a first server, which the client and the first server can be communicatively connected; the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; the method comprises:
- the first server receiving text information sent by the terminal processing unit; the terminal processing unit is used for obtaining text information corresponding to the voice information based on the voice information collected by the voice receiving unit;
- the first server inputting the text information to the target language sub-model in the language model, which the target language sub-model corresponds to the language category of the text information; obtaining the text response information corresponding to the text information returned by the target language sub-model, which the text response information has the same language category as the text information;
- the first server sending the text response information to the terminal processing unit, so that the terminal processing unit obtaining the voice response information corresponding to the text response information based on the text response information, and controlling the voice playback unit to play the voice response information.
17. The method according to claim 16, characterized in that the text information comprises a language identifier, which identifies the language category of the voice information;
- the first server inputting the text information to the target language sub-model in the language model; obtaining the text response information corresponding to the text information returned by the target language sub-model, comprising:
- the first server inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and obtaining the text response information corresponding to the language identifier.
18. The method according to claim 17, characterized in that the first server comprises a language response module, wherein the language response module comprises or connecting to the language model, the first server inputting the text information to the target language sub-model corresponding to the language identifier in the language

model, and obtaining the text response information corresponding to the language identifier, comprising:

the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, and

5 obtaining the text response information corresponding to the language identifier.

19. The method according to claim 18, characterized in that the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, comprising:

10 the language response module using the target language sub-model corresponding to the language identifier in the language model to semantically understand the text information and obtaining the understanding result;

if the understanding result indicates that the text information is Q&A type information, the language response module generating text response information corresponding to

15 the text information in the language category characterized by the language identifier by using the target language sub-model.

20. The method according to claim 19, characterized in that the first server further comprises a text processing module, and the method further comprises:

20 if the understanding result indicates that the text information is instruction type information, the language response module sending the response instruction corresponding to the text information to the terminal processing unit through the text processing module.

25 21. The method according to claim 20, characterized in that before the language response module inputting the text information to the target language sub-model corresponding to the language identifier in the language model, the method further comprises:

the text processing module receiving the text information sent by the terminal processing unit and classifying the text information to obtain classification result;

30 if the classification result indicates that the text information is instruction type information, the text processing module sending the response instruction corresponding to the text information to the terminal processing unit;

if the classification result indicates that the text information is Q&A type information, the text processing module sending the text information to the language response module.

- 5 22. A client, characterized in that the client and the first server can be communicatively connected, wherein the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages; wherein:
- 10 the voice receiving unit is configured to collect voice information and send the voice information to the terminal processing unit;
- the terminal processing unit is configured to obtain text information corresponding to the voice information based on the voice information, send the text information to the first server, so that the first server inputs the text information to the target language
- 15 sub-model in the language model, and obtain the text response information corresponding to the text information returned by the target language sub-model; the target language sub-model corresponds to the language category of the text information, which the text response information has the same language category as the text information;
- 20 the terminal processing unit is further configured to obtain the voice response information corresponding to the text response information based on the text response information, and control the voice playback unit to play the voice response information in the language category of the text information.
- 25 23. A first server, characterized in that the client and the first server can be communicatively connected, wherein the client comprises a voice receiving unit, a voice playback unit, and a terminal processing unit; the first server comprises or connects the language model, the language model comprises one or more language sub-model, and the language sub-model correspond to one or more languages;
- 30 wherein:
- the first server is configured to receive text information sent by the terminal processing unit; the terminal processing unit is used to obtain text information corresponding to the voice information based on the voice information collected by the voice receiving unit;

the first server is configured to input the text information to a target language sub-model in the language model, which the target language sub-model corresponds to the language category of the text information; obtain the text response information corresponding to the text information returned by the target language sub-model,
5 which the text response information has the same language category as the text information;

the first server is configured to send the text response information to the terminal processing unit, so that the terminal processing unit obtains the voice response information corresponding to the text response information based on the text response
10 information, and control the voice playback unit to play the voice response information.