REAL-TIME SPEECH TRANSLATION METHOD, SYSTEM, DEVICE, AND STORAGE MEDIUM

BACKGROUND OF THE INVENTION

1. Field of the invention

The present invention relates to the field of online translation technology, and more particularly, to a real-time speech translation method, system, device, and storage medium.

2. Description of the prior art

With the rise of international communication, the use of different languages for communication is becoming more frequent. During communication, due to language differences, parties rely on their language skills or human interpreters for verbal translation, which causes inconvenience for both parties. Therefore, using translation devices or wearable devices paired with earphones, along with robotic speech, to achieve real-time online translation will become the most convenient method.

During a call, the robotic translation system translates the speaker's content into the listener's language for playback. However, the translation can deviate from the speaker's intent, and there is a delay between the spoken content and the translation. Since the speaker cannot understand the translation, they are unable to quickly determine when specific parts of their message have been understood by the listener or if they were received accurately. This prevents the speaker from assessing the timing and accuracy of their message, forcing them to pause and wait for the translation to finish before assuming the listener has understood. As a result, this does not ensure

effective and accurate communication between both parties.

SUMMARY OF THE INVENTION

To solve the problem of the prior art, the invention provides a real-time speech translation method, system, device, and storage medium that enhances the accuracy of cross-language translation and the communication efficiency between parties.

To achieve the above purpose, the present invention provides a real-time speech translation method, comprising the following steps of:

performing speech recognition of a first speech information input by a first user to obtain first text data;

translating the first speech information to obtain a second synthesized speech corresponding to a second language category;

obtaining a first synthesized speech corresponding to a first language category on the basis of the first text data; and

outputting the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to a second user.

Optionally, the step of translating the first speech information to obtain a second synthesized speech corresponding to a second language category further comprises the following steps of:

translating the first text data obtained from the speech recognition of the first speech information to obtain a second text data corresponding to the second language category; and

obtaining the second synthesized speech corresponding to the second language category on the basis of the second text data.

Optionally, the step of obtaining a first synthesized speech according to

the first text data, corresponding to a first language category further comprises the following steps of:

calculating an estimated duration for completing the playback of the second synthesized speech to obtain a first estimated duration;

obtaining a second timestamp information related to the second synthesized speech on the basis of the first estimated duration; and

performing speech synthesis on the basis of the first text data to obtain the first synthesized speech; and synchronizing a first timestamp corresponding of the first synthesized speech with the second timestamp information.

Optionally, the real-time speech translation method further comprises the following steps of:

continuing collecting a next segment of the first speech information input by the first user when receiving a positive feedback from the first user regarding the first synthesized speech; and

collecting the first speech information re-input by the first user and outputting a notification message of the re-input of the first speech information to the second user when receiving a negative feedback from the first user regarding the first synthesized speech.

Optionally, the real-time speech translation method further comprises the following steps of:

respectively extracting interest tags and the corresponding timestamp information from the first text data and the second text data; and

respectively marking the interest tags on the first text data and the second text data according to the first timestamp information and the second timestamp information.

Optionally, before the step of outputting the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to the second user, further comprising the following steps of:

respectively collecting voiceprint information from the first user and the second user;

respectively identifying the first user and the second user to obtain identity information on the basis of each of the voiceprint information;

determining a relationship type between the first user and the second user on the basis of each of the identity information; and

filtering the interest tags on the basis of the relationship type between the first user and the second user and a preset identity relationship interest database, and designating the remaining interest tags as target interest tags;

Wherein the step of outputting the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to the second user further comprises the following step of:

providing a voice notification for the voice associated with the target interest tags in the first synthesized speech and second synthesized speech.

Optionally, the step of performing speech recognition on a first speech information from a first user to obtain first text data further comprises the following steps of:

obtaining preceding context data of the first speech information;

performing speech recognition of the first speech information to obtain initial text data; and

correcting the initial text data according to the preceding context data to obtain the first text data.

Optionally, before the step of outputting the first synthesized speech to

the first user and simultaneously outputting the second synthesized speech to the second user, further comprising the following step of:

when detecting a pair of the earphones for speech output worn by different users, controlling the earphones to operate in a first state; when detecting two pairs of earphones worn by the first user and the second user respectively, controlling the pairs of earphones to operate in a second state;

wherein, in the first state, each of the earphones is configured as a speech output channel of the user wearing it and outputs different speech from each other; in the second state, both earphones in one pair of earphones are configured as a speech output channel of the user wearing the and output the same speech.

Optionally, the pair of earphones comprises an UWB communication module; before the step of outputting the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to the second user, the real-time speech translation method further comprises the following steps of:

the UWB communication module detecting the distance between the two earphones in the pair of earphones in real-time;

when the distance between the earphones is larger than a first preset threshold, controlling the earphones to operate in the first state; and

when the distance between the earphones is less than the first preset threshold, controlling the earphones to operate in the second state.

Optionally, the real-time speech translation method further comprises the following steps of:

forming an audio adjustment database on the basis of un-played audio frames of the first synthesized speech and audio frames of the collected first speech information that have not been converted into the first synthesized

speech; and

adjusting a playing speed of the first synthesized speech on the basis of the number of audio frames in the audio adjustment database;

wherein, the playing speed of the first synthesized speech increases with an increase in the number of audio frames in the audio adjustment database and decreases with a decrease in the number of audio frames in the audio adjustment database.

The present invention also provides a real-time speech translation system for implementing the real-time speech translation method mentioned above. The real-time speech translation system comprises:

a first text data generation module, configured to performing speech recognition of a first speech information input by a first user to obtain first text data;

a second synthesized speech generation module, configured to translate the first speech information to obtain a second synthesized speech corresponding to a second language category;

a first synthesized speech generation module, configured to obtain a first synthesized speech corresponding to a first language category on the basis of the first text data; and

a synthesized speech playback module, configured to the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to a second user.

The present invention also provides a real-time speech translation device. The real-time speech translation device comprises:

a processor; and

a memory, configured to store programs for the processor to execute;

wherein, the processor is configured to execute the programs to implemented the steps of the real-time speech translation method.

The present invention also provides a storage medium for storing programs, wherein the programs are implemented the steps of the real-time speech translation method by a processor.

Compared with the prior art, the present invention has the following advantages and significant effects:

The present invention provides the real-time speech translation method, system, device, and storage medium that can synthesize the first synthesized speech according to the first text data obtained through speech recognition from the speaker's content. While playing the second synthesized speech to the listener, the first synthesized speech is simultaneously played back to the speaker. This allows the speaker to be notified of any errors in the translation content caused by mistakes in the speech recognition process in a timely manner, thereby improving the accuracy of cross-language translation and the communication efficiency between the parties in cross-language communication. As a result, communication becomes more fluent in cross-language.

BRIEF DESCRIPTION OF THE APPENDED DRAWINGS

The other features, objectives, and advantages of the present invention will become more apparent through a detailed description of non-limiting embodiments with reference to the drawings.

FIG. 1 is a schematic diagram illustrating a real-time speech translation method according to an embodiment of the present invention.

FIG. 2 is a schematic diagram illustrating the application scenario related to the real-time speech translation method according to the embodiment of the present invention.

FIG. 3 is a schematic diagram illustrating another application scenario related to the real-time speech translation method according to the embodiment of the present invention.

FIG. 4 is a schematic diagram illustrating another application scenario related to the real-time speech translation method according to the embodiment of the present invention.

FIG. 5 is a schematic diagram illustrating the real-time speech translation method according to another embodiment of the present invention.

FIG. 6 is a schematic diagram illustrating the synchronous working principle of the real-time speech translation method according to the embodiment of the present invention.

FIG. 7 is a schematic diagram illustrating the real-time speech translation method according to another embodiment of the present invention.

FIG. 8 is a schematic diagram illustrating the real-time speech translation method according to another embodiment of the present invention.

FIG. 9 is a schematic diagram illustrating the real-time speech translation method according to another embodiment of the present invention.

FIG. 10 is a structure schematic diagram illustrating the real-time speech translation system according to an embodiment of the present invention.

FIG. 11 is a structure schematic diagram illustrating the real-time speech translation device according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

The appended drawings provide a more comprehensive description of the exemplary embodiments. However, the exemplary embodiments can be

implemented in various forms and should not be construed as limited to the embodiments described herein. Instead, these embodiments are provided to ensure that the present invention is comprehensive and complete, and to fully convey the concepts of the exemplary embodiments to those skilled in the art. Identical reference numerals in the drawings indicate the same or similar structures, and therefore, repeated descriptions of them will be omitted.

As shown in FIG. 1, an embodiment of the present invention discloses a real-time speech translation method. This embodiment is applied to speech recognize and translate in scenarios where two users communicate with cross-language in real-time. The method can be executed by an online real-time speech translation device, which can be implemented through hardware and/or software, and can be configured in any terminal or network element. Through online real-time speech translation, translation services can be provided to both users in real time, allowing smooth communication between users who speak different languages. The real-time speech translation method includes the following steps:

S110: performing speech recognition of a first speech information input by a first user to obtain first text data.

Please refer to FIG. 2, in the step mentioned above, the first user 21 and the second user 22 communicate with cross-language. The first user 21 can input the first speech information in the first language, and the online real-time speech translation device can perform speech recognition to obtain first text data. The speech from the first user 21 can be captured through the earphones worn by the first user 21.

In this embodiment, the first user 21 is regarded as the speaker at the beginning of the cross-language communication, and the second user 22 is regarded as the listener at the beginning of the communication. Therefore, at the beginning of the communication, the step S110 collects the first speech information input by the first user 21, and then performs recognition on the

first speech information of the first user.

Since the communication is bidirectional, in subsequent ongoing communications, the step also collects the second speech information input by the second user 22 and performs recognition on the second speech information. Wherein, the online real-time speech translation device can be the translation device 23 as shown in FIG. 2, or it can be a server. The invention does not limit the device to these specific forms.

On the other hand, when the online real-time speech translation device is a translation device, the embodiment of the invention can include more than one translation device. For example, as shown in FIG. 3, the first user 21 and the second user 22 can each use a translation device. For example, the first user 21 uses the first translation device 24, and the second user 22 uses the second translation device 25. The first translation device 24 and the second translation device 25 communicate with each other and exchange data. The text translation between different languages can be performed by either the first translation device 24 or the second translation device 25. The first translation device 24 and the second translation device 25 are also used to perform speech synthesis for their respective users.

It should be noted that FIG. 2 and FIG. 3 illustrate a face-to-face real-time translation scenario. The embodiment can also be applied to long-distance online translation scenarios. FIG. 4 illustrates a schematic diagram of a long-distance online translation scenario. As shown in FIG. 4, the first user 21 and the second user 22 perform online translation using their respective notebook computers. In this illustration, first user 21 and the second user 22 perform online translation using their respective notebook computers.

S120: translating the first speech information to obtain a second synthesized speech corresponding to a second language category. Specifically, online speech translation generally involves two stages: The

first stage is speech recognition, where the first language speech input by the first user is recognized as text information; the second stage is to translate the text information according to a translation corpus and then generate speech or text information in the second language, which is provided to the second user. Therefore, in some embodiments, step S120 includes:

S121: translating the first text data to obtain the second text data corresponding to the second language category; and

S122: obtaining the second synthesized speech corresponding to the second language category on the basis of the second text data.

The steps mentioned above indicate that translate the first text data in the first language into the second text data in the second language, translating the first text data in the first language into the second text data in the second language, and then synthesizing speech on the basis of the second text data.

S130: obtaining a first synthesized speech corresponding to a first language category on the basis of the first text data.

Wherein, the first synthesized speech is used to allow the first user, as the speaker at the time, to know the progress of the playback of the translated speech through their own device, and promptly detect any errors occurring during the speech recognition process. This enables the speaker to accurately verify the correctness of the translated content and the timing of its delivery to the other party, thereby allowing communication between parties speaking different languages to correct errors in real-time and enhance interactivity.

In order to keep the playback content of the speaker's device and the listener's device synchronized, the first synthesized speech and the second synthesized speech need to be synchronized with timestamps. Therefore, in some embodiments, as shown in FIG. 5, step S130 includes:

S131: calculating an estimated duration for completing the playback of the second synthesized speech to obtain a first estimated duration;

S132: obtaining a second timestamp information related to the second synthesized speech on the basis of the first estimated duration; and

S133: performing speech synthesis on the basis of the first text data to obtain the first synthesized speech, and synchronizing a first timestamp corresponding of the first synthesized speech with the second timestamp information.

This embodiment utilizes the steps mentioned above to estimate the time required for synthesizing the speech in the second language and synchronizes the timestamps accordingly. In practice, for example, the two synthesized speeches can be divided into multiple audio segments, and the audio segments of the first synthesized speech can be synchronized with the corresponding audio segments of the second synthesized speech to achieve timestamp synchronization.

This way, the invention establishes the synchronization mechanism between the synthesized speech in the first language and the synthesized speech in the second language, ensuring that the playback content of both parties maintains the same progress.

S140: outputting the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to the second user.

This embodiment synchronously plays the two synthesized speeches, which helps ensure that the playback content maintains the same progress. This allows the speaker to get the accuracy of the translation and the timing of its delivery to the listener, thereby enabling communication between parties speaking different languages to correct errors in real-time and enhance interactivity.

Furthermore, during synchronized playback, the speaker can continue speaking or temporarily pause to listen to the first synthesized speech. If the speaker continues speaking, the playback volume of the first synthesized speech can be set to not exceed a fourth preset threshold below the speaker's own volume, so as not to interfere with the speaker's expression while allowing the speaker to hear the content through the earphones.

Since real-time online translation is divided into two scenarios: face-to-face scenario and long-distance online scenario, in practice, in the face-to-face translation scenario, the two users can share a single pair of earphones or each wear their own. When sharing a single pair of earphones, the first synthesized speech can be output through the left channel to the first user, while the second synthesized speech is simultaneously output through the right channel to the second user. In long-distance online scenario, both users wear their own pair of earphones.

FIG. 6 is a schematic diagram illustrating the synchronous working principle of the real-time speech translation method according to the embodiment of the present invention. As shown in FIG. 6, User A is the first user, and A's speech robot refers to the device used to play the first synthesized speech to User A. User B is the second user, and B's speech robot refers to the device used to play the second synthesized speech to user B.

Sx (including S1, S2, and S3) represents the original voice of the first user. S1, S2, and S3 correspond to the first, second, and third segments of speech emitted by User A, respectively. Sx' refers to the synthesized machine voice generated by A's speech robot after recognizing the content of Sx as the first language, i.e., the synthesized speech after recognition. S1', S2', and S3' respectively represent the synthesized speech in the first language played to User A, corresponding to the first segment of speech S1, the second segment of speech S2, and the third segment of speech S3.

TSx’ represents the synthesized machine voice generated by B's speech robot after translating the content identified from Sx into text in the second language, i.e., the translated speech. Therefore, TS1’, TS2’, and TS3’ respectively represent the synthesized speech in the second language played to User B, corresponding to the first segment of speech S1, the second segment of speech S2, and the third segment of speech S3.

In another embodiment of the invention, another real-time speech translation method is disclosed. As shown in FIG. 7, this method further includes the following steps according to the corresponding embodiment of the FIG. 1 mentioned above:

S150: determining whether to receive positive or negative feedback from the first user regarding the first synthesized speech. If so, proceed to step S160. Otherwise, proceed to step S170;

S160: continuing collecting the next segment of the first speech information input by the first user; and

S170: collecting the first speech information re-input by the first user and outputting the notification message of the re-input of the first speech information to the second user.

Specifically, when the first user confirms the translation is accurate, the system continues to collect the next segment of the first speech information input by the first user. When the first user confirms the translation is error, the system notifies the first user to re-input the content and collects again, while notifying the second user that the previously played second synthesized speech contains an error and will be replayed.

Regarding the method for determining whether the translation is accurate, in other embodiments, if the system does not receive a predefined vocal response from the first user within a predetermined time period after the first synthesized speech has finished playing, it is determined that the

translation is accurate. Otherwise, it is determined that there is an error in the translation. This allows communication between parties speaking different languages to promptly correct errors and enhance interactivity.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, step S110 includes:

performing voiceprint detection on the first speech information and conducting speech recognition only on the audio segments corresponding to the speech of the first user, and then obtaining the first text data.

Step S120 includes: performing voiceprint detection on the first speech information and only translating the audio segments spoken by the first user to obtain the second synthesized speech. That is, during the translation process, when audio segments not belonging to the first user are detected, the translation is paused. When audio segments spoken by the first user are detected, the translation continues until completion.

This helps ensure the accuracy of real-time translation during cross-lingual communication.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, before step S140, the method further includes the following steps of:

respectively collecting environmental voice data from the first user and the second user;

performing voiceprint detection on each of the environmental voice data;

sending a first preset alert message to the first user when detecting other people's voiceprints in the environmental voice data; the first preset alert message is used to remind the first user that there are other people nearby, reminding them to be cautious about protecting their privacy;

pausing the translation of first speech information when detecting other people's voiceprints in the environmental voice data from the second user, and sending a second preset alert message to the second user; and

resuming the translation of the first speech information when the other people's voiceprints are no longer detected in the environmental voice data from the second user.

This helps protect the privacy of both the first and second users. The steps mentioned above can be positioned between step S130 and step S140.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, before step S140, the method further includes the following steps of:

respectively extracting interest tags and the corresponding timestamp information from the first text data and the second text data; and

respectively marking the interest tags on the first text data and the second text data according to the first timestamp information and the second timestamp information.

Exemplarily, the steps mentioned above can be positioned between step S130 and step S140.

Step S140 includes: providing a voice notification for the voice associated with the target interest tags when playing the first synthesized speech and the second synthesized speech.

Wherein, each of the interest tags mentioned above corresponds to the timestamp. By using the interest tags, users can easily find out the key content when reviewing the speech. The interest tags can be displayed either in text or audio form.

When certain content is repeatedly mentioned, or key points from speech tone or semantic analysis are identified, or when sensitive privacy is

involved, or if the listener is temporarily unable to focus but wants to review the content later, the invention can proactively activate the function of the interest tags to record timestamps and the corresponding interest tags, and stores them. When the user reviews the interest tags, key points of the translated conversation can be displayed. The user can skip to view the content before and after the key points or replay the conversation history around those key points, enhancing the user experience. Additionally, the interest tags can help both parties in the conversation focus on key points, avoiding misinterpretation or neglect, thus improving the cross-language communication experience.

In some embodiments, according to the embodiment mentioned above, before step S140, the method further includes the following steps of:

respectively collecting voiceprint information from the first user and the second user;

respectively identifying the first user and the second user to obtain identity information on the basis of each of the voiceprint information;

determining a relationship type between the first user and the second user on the basis of each of the identity information; and

filtering the interest tags on the basis of the relationship type between the first user and the second user and a preset identity relationship interest database, and designating the remaining interest tags as target interest tags.

Exemplarily, the steps mentioned above can be positioned between step S130 and step S140.

Step S140 includes: providing the voice notification for the voice associated with the target interest tags in the first synthesized speech and second synthesized speech.

Wherein, the determination of the relationship type between the first

user and the second user can be achieved on the basis of each of a preset identity relationship interest database.

In some embodiments, according to the embodiment mentioned above of FIG. 1 mentioned above, step S110 further includes the following steps of:

S111: obtaining preceding context data of the first speech information;

S112: performing speech recognition of the first speech information to obtain initial text data; and

S113: correcting the initial text data on the basis of the preceding context data to obtain the first text data.

Specifically, obtaining other speech information spoken by the user before the first speech information, and correcting the recognized text on the basis of the other speech information helps improve translation accuracy, thereby enhancing the fluency of cross-language communication. For example, if the preceding context data translates the same speech as a first word, while this step translates the speech as a second word, the second word will then be corrected to the first word.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, as shown in FIG. 8, before step S140, the method further includes the following steps of:

S180: when detecting a pair of the earphones for speech output worn by different users, controlling the earphones to operate in a first state; when detecting two pairs of earphones worn by the first user and the second user respectively, controlling the pairs of earphones to operate in a second state.

wherein, in the first state, each of the earphones is configured as a speech output channel of the user wearing it and outputs different speech from each other; in the second state, both earphones in one pair of earphones are configured as a speech output channel of the user wearing the and output

the same speech.

Exemplarily, the step mentioned above can be positioned between step S130 and step S140.

Step S140 is replaced with step S141: outputting the first synthesized speech to the first user through the earphones worn by the first user, and simultaneously outputting the second synthesized speech to a second through the earphones worn by the second user.

Specifically, for example, when each of the earphones worn by different users, ie, the first user wears of the earphone and the second user wears the other, the earphones can be controlled to operate in face-to-face translation mode. When each user wears their own pair of earphones, the earphones can be controlled to operate in long-distance translation mode. In face-to-face translation mode, the synthesized speech corresponding to each user needs to be encoded as the audio of the channel corresponding to the earphone by that user. For example, if the first user wears the left earphone, the first synthesized speech is encoded as the left channel audio. Correspondingly, the second synthesized speech is encoded as the right channel audio.

The earphones worn by the first user is configured as the first speech output channel, and the earphones worn by the second user is configured as the second speech output channel. The first synthesized speech is output to the first user through the first speech output channel, and the second synthesized speech is output to the second user through the second speech output channel.

In some embodiments, according to the embodiment mentioned above, the pair of earphones includes an UWB (Ultra Wide Band) communication module. Step S180 includes:

the UWB communication module detecting the distance between the

two earphones in the pair of earphones in real-time.

When the distance between the earphones is larger than a first preset threshold, controlling the earphones to operate in the first state; when the distance between the earphones is less than the first preset threshold, controlling the earphones to operate in the second state.

Specifically, when the distance between the earphones is larger than the first preset threshold, it indicates that the two earphones are worn by two different users. When the distance between the earphones is less than the first preset threshold, it indicates that the pair of earphones is worn by the same user.

Exemplarily, the first preset threshold can be 25 cm. In other embodiments, the distance between the earphones can also be detected through an ultrasonic signal module arranged in the earphones, but does not impose limitations of the invention.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, as shown in FIG. 9, the method further includes the following steps of:

S190: forming an audio adjustment database on the basis of un-played audio frames of the first synthesized speech and audio frames of the collected first speech information that have not been converted into the first synthesized speech; and

S200: adjusting a playing speed of the first synthesized speech on the basis of the number of audio frames in the audio adjustment database.

Wherein, the playing speed of the first synthesized speech increases with an increase in the number of audio frames in the audio adjustment database and decreases with a decrease in the number of audio frames in the audio adjustment database. In some embodiments, the playing speed of the

second synthesized speech can also be adjusted simultaneously.

For example, when the user has already spoken up to the S6 audio segment, the earphones are still playing back the S3 audio segment, which would hinder the speaker from promptly noticing errors in the speech recognition process. This helps to avoid situations mentioned above, thereby preventing the parties speaking different languages from being able to correct mistakes in a timely manner.

In some embodiments, according to the corresponding embodiment of FIG. 1 mentioned above, before the step S110, the method further includes the following steps of:

collecting the environmental voice data from the first user's environment and extracting multiple speech information from the environmental voice data when detecting that the first user is wearing only one earphone of the pair and the distance between the earphones is larger than a first preset threshold;

determining the language category corresponding to the multiple speech information mentioned above as the second language category; and

playing the notification message to the external environment through the other earphone of the pair; the notification message is configured to notify the second user to wear the other earphone of the pair.

Step S140 includes:

outputting the second synthesized speech to the second user through the second earphone of the pair when detecting the second user has worn the second earphone.

This feature is useful when the first user is in an environment where a foreign language is spoken, such as when traveling to another country, and needs face-to-face translation with the second user. The invention helps the

first user quickly establish communication with the second user, thereby enhancing the fluidity of cross-linguistic communication.

It should be noted that all the disclosed embodiments of the invention can be freely combined. The technical methods obtained from such combinations are also included within the scope of protection of the invention.

As shown in FIG. 10, an embodiment of the present invention further discloses a real-time speech translation system 10, which includes:

a first text data generation module 101, configured to performing speech recognition of a first speech information input by a first user to obtain first text data;

a second synthesized speech generation module 102, configured to translate the first speech information to obtain a second synthesized speech corresponding to a second language category;

a first synthesized speech generation module 103, configured to obtain a first synthesized speech corresponding to a first language category on the basis of the first text data; and

a synthesized speech playback module 104, configured to the first synthesized speech to the first user and simultaneously outputting the second synthesized speech to a second user.

It should be understood that the real-time speech translation system of the present invention can also include other existing functional modules that support the operation of the real-time speech translation system. The real-time speech translation system shown in FIG. 10 is merely an example and should not impose any limitations on the functions and scope of the embodiments of the invention.

The embodiment of the real-time speech translation system is used to

implement the real-time speech translation method mentioned above. Therefore, the specific implementation steps of the real-time speech translation system can refer to the description of the real-time speech translation method mentioned above, and will not be repeated here.

An embodiment of the present invention also discloses a real-time speech translation device, including a processor and a memory. Wherein, the memory is configured to store programs for the processor to execute; the processor is configured to execute the programs to execute the steps of the real-time speech translation method mentioned above. FIG. 11 is a structure schematic diagram illustrating the real-time speech translation device according to an embodiment of the present invention. Please refer to FIG. 11 to describe the electronic device 600 of the embodiment of the invention. The electronic device 600 shown in FIG. 11 is merely an example and should not impose any limitations on the functions and scope of use of the embodiments of the present invention.

As shown in FIG. 11, the electronic device 600 is represented as a general-purpose computing device. The electronic device 600 includes at least one processing unit 610, at least one storage unit 620, a bus 630 that connects different platform components (including the storage unit 620 and processing unit 610), and a display unit 640, and others.

Wherein, the storage unit stores program code, which can be executed by the processing unit 610, allowing the processing unit 610 to execute the steps of the real-time speech translation method described in the invention, according to various exemplary embodiments of the present invention. For example, the processing unit 610 can execute the steps shown in FIG. 1.

The storage unit 620 can include readable media in the form of volatile memory, such as a random access memory (RAM) 6201 and/or a cache memory 6202, and can also include a read-only memory (ROM) 6203.

The storage unit 620 can further include programs/utilities 6204 with a set of (at least one) program modules 6205. These program modules 6205 can include, but are not limited to: an operating system, one or more applications, other program modules, and program data, each of which or any combination thereof can include implementations of network environments.

The bus 630 can represent one or more types of bus structures, including a memory bus or memory controller, a peripheral bus, a accelerated graphics port, a processor, or a local bus using any of a variety of bus architectures.

The electronic device 600 can communicate with one or more external devices 700 (such as a keyboard, pointing device, Bluetooth devices, etc.), can communicate with one or more devices that allow the user to interact with the electronic device 600, and/or with any devices that allows the electronic device 600 to communicate with one or more other computing devices(such as routers, modems, etc.). This communication can be carried out through the input/output (I/O) interface 650. Moreover, the electronic device 600 can communicate with one or more networks (such as a local area network (LAN), wide area network (WAN), and/or public network, such as the Internet) via a network adapter 660. The network adapter 660 can communicate with other modules of the electronic device 600 through the bus 630. It should be understood that, although not shown in the figure, other hardware and/or software modules can be used in conjunction with the electronic device 600, including but not limited to: microcode, device drivers, redundant processing units, external disk drive arrays, RAID systems, tape drives, and data backup storage platforms.

An embodiment of the present invention also discloses a storage medium for storing programs. Wherein, the programs are implemented the steps of the real-time speech translation method mentioned above by a processor. In some embodiments, various aspects of the invention can also be

implemented as a form of program product, which includes program code. When the program product is executed on a terminal device, the program code enables the terminal device to perform the steps described in the real-time speech translation method section of this specification, according to various exemplary embodiments of the present invention.

As described above, when the program stored on the computer-readable storage medium of this embodiment is executed, it synthesizes the first synthesized speech on the basis of the first text data obtained through speech recognition of the speaker's spoken content. While the second synthetic speech is being played to the listener, the first synthetic speech is simultaneously played back to the speaker, allowing the speaker to be promptly aware of any errors in the translation caused by mistakes in the speech recognition process. This improves the accuracy of cross-language translation and enhances the communication efficiency between both parties in cross-language conversations.

An embodiment of the present invention discloses a storage medium. The storage medium is a program product for implementing the method mentioned above, which can be in the form of a portable compact disc read-only memory (CD-ROM) and includes program code, and it can run on terminal devices, such as personal computers. However, the program product of the invention is not limited hereto. In the invention, the storage medium can be any tangible medium that contains or stores programs that can be used or executed by a system, device, or apparatus that can execute instructions.

The program product can take any combination of one or more readable media. A readable medium can be a readable signal medium or a readable storage medium. A readable storage medium can be, but is not limited to, an electrical, magnetic, optical, electromagnetic, infrared, or semiconductor system, device, or apparatus, or any combination thereof. More specific examples (a non-exhaustive list) of readable storage media include: electrical

connections with one or more wires, USB flash disk, hard disks, random access memory (RAM), read-only memory (ROM), erasable programmable read-only memory (EPROM or flash memory), optical fibers, portable compact disc read-only memory (CD-ROM), optical storage devices, magnetic storage devices, or any suitable combination of the above.

The computer-readable storage medium can also include data signals transmitted on a baseband or as part of a carrier wave, which carry readable program code. These propagated data signals can take various forms, including but not limited to electromagnetic signals, optical signals, or any suitable combination thereof. The readable storage medium can also be any readable medium other than a readable storage medium, which can send, propagate, or transmit programs for use by or in conjunction with an instruction execution system, device, or apparatus. The program code contained on a readable storage medium can be transmitted by any suitable medium, including but not limited to wireless, wired, optical cable, RF, or any suitable combination of the above.

The program code for executing the operations of the invention can be written in any combination of one or more programming languages, including object-oriented programming languages such as Java, C++, and conventional procedural programming languages such as the "C" language or similar programming languages. The program code can be executed entirely on the user’s computing device, partly on the user’s device, as executing software package, partly on the user’s computing device and partly on a remote computing device, or entirely on a remote computing device or server. In the case of a remote computing device, the remote computing device can be connected to the user’s computing device through any type of network, including a local area network (LAN) or a wide area network (WAN), or can be connected to an external computing device (for example, through an internet service provider over the Internet).

The present invention provides the real-time speech translation method, system, device, and storage medium that can synthesize the first synthesized speech according to the first text data obtained through speech recognition from the speaker's content. While playing the second synthesized speech to the listener, the first synthesized speech is simultaneously played back to the speaker. This allows the speaker to be notified of any errors in the translation content caused by mistakes in the speech recognition process in a timely manner, thereby improving the accuracy of cross-language translation and the communication efficiency between the parties in cross-language communication. As a result, communication becomes more fluent in cross-language.

With the examples and explanations mentioned above, the features and spirits of the invention are hopefully well described. More importantly, the present invention is not limited to the embodiment described herein. Those skilled in the art will readily observe that numerous modifications and alterations of the device may be made while retaining the teachings of the invention. Accordingly, the above disclosure should be construed as limited only by the metes and bounds of the appended claims.