VOICE COMMUNICATION~~CALL~~ CONTROL METHOD AND APPARATUS, COMPUTER-READABLE MEDIUM AND ELECTRONIC DEVICE

[0001] This application claims priority to Chinese Patent Application No. 202110083388.6, entitled "VOICE CALL CONTROL METHOD AND APPARATUS, COMPUTER-READABLE MEDIUM AND ELECTRONIC DEVICE" and filed with the China National Intellectual Property Administration on January 21, 2021, which is incorporated herein by reference in its entirety.

FIELD

[0002] This application relates to the field of computer technologies, and specifically, to a voice communication control method and apparatus, a computer-readable medium, and an electronic device.

BACKGROUND

[0003] In many voice communication scenarios, the quantity of speakers or the timbre of speakers needs to be identified for subsequent voice control. In the related art, a speaker scenario detection system is trained based on a large quantity of labeled voice segments, where a label of each segment includes the quantity of speakers. A voice segment is used during testing, and the system predicts the current quantity of speakers. Such a processing manner causes a relatively long delay for voice detection, especially in a real-time communication scenario, which greatly reduces the efficiency of voice recognition and affect the effect of real-time voice control.

SUMMARY

[0004] A voice communication control method and apparatus, a computer-readable medium, and an electronic device are provided according to embodiments of the present disclosure, which can ensure the accuracy of detecting the quantity of speakers in a voice communication to a certain extent, and improve the efficiency of recognizing the quantity of speakers in a voice communication and the efficiency of controlling the voice communication.

[0005] Other features and advantages of the present disclosure become apparent through the following detailed descriptions, or may be partially learned through the practice of the present disclosure.

CLAIMS

1. A voice communication control method, performed by a server, the method comprising:

obtaining a mixed communication voice, the mixed communication voice comprising at least one branch voice;

performing frequency domain transformation on the communication voice to determine energy information of the communication voice at each frequency point in a frequency domain;

performing separation on the energy information at each frequency point by using a neural network to determine an energy proportion of each branch voice at the frequency point in total energy at the frequency point;

determining a quantity of the at least one branch voice comprised in the communication voice based on the energy proportion of each branch voice at each frequency point; and

controlling the voice communication by setting a communication voice control manner based on the quantity of the branch voices.

2. The method according to claim 1, wherein the performing frequency domain transformation on the communication voice to determine energy information of the communication voice at each frequency point in a frequency domain comprises:

performing framing on the communication voice to obtain at least one audio frame of voice information;

performing frequency domain transformation on each of the at least one audio frame of voice information to obtain a voice energy spectrum in the frequency domain; and

determining, based on the voice energy spectrum, the energy information of the communication voice at each frequency point in the frequency domain.

3. The method according to claim 2, wherein the performing frequency domain transformation on each audio frame of the voice information to obtain a voice energy spectrum in the frequency domain comprises:

performing Fourier transform on each of the at least one audio frame of voice information in

the time domain to obtain a voice energy spectrum of the audio frame of voice information in the frequency domain.

4. The method according to claim 2, wherein the determining the energy information of the communication voice at each frequency point in the frequency domain based on the voice energy spectrum comprises:

performing modulo processing on an amplitude of the voice energy spectrum at each frequency point to obtain an amplitude spectrum corresponding to the voice energy spectrum; and

calculating a square value of the amplitude spectrum, and performing a logarithmic operation on the square value to generate the energy information of the communication voice at each frequency point in the frequency domain.

5. The method according to claim 1, wherein the neural network comprises a long-short-term memory neural network, and

the performing separation on the energy information at each frequency point by using a neural network to determine an energy proportion of each branch voice at each frequency point in total energy at the frequency point comprises:

inputting the energy information into a preset voice separation model, performing convolution processing by using the long short-term memory neural network, to determine the branch voice at each frequency point; and

determining, based on the energy information of each branch voice at each frequency point, the energy proportion of the branch voice at the frequency point.

6. The method according to claim 5, further comprising:

obtaining a first voice sample corresponding to a single voice, and a second voice sample corresponding to a mixed voice comprising the single voice;

extracting a first voice feature from the first voice sample, and extracting a second voice feature from the second voice sample;

inputting the second voice feature into a voice separation model constructed based on the long short-term memory artificial neural network, to determine a predicted voice separated from

the second voice feature and a predicted energy proportion corresponding to the predicted voice in the second voice sample; and

updating parameters of the voice separation model according to a comparison result between an actual energy proportion of the first voice sample in the second voice sample and the predicted energy proportion.

7. The method according to claim 5, wherein the determining a quantity of the at least one branch voice comprised in the communication voice based on the energy proportion of each branch voice at each frequency point comprises:

calculating, based on the energy proportion of each branch voice at each frequency point, an average value of energy proportions of the branch voice; and

determining the quantity of the at least one branch voice comprised in the communication voice based on the average value of each branch voice and a set threshold.

8. The method according to claim 7, wherein the determining the quantity of the at least one branch voice comprised in the communication voice based on the average value of each branch voice and a set threshold comprises:

determining that the quantity of the at least one branch voice is more than one in a case that an absolute value of a difference between the average value of each branch voice and the set threshold is less than a difference threshold; and

determining that the quantity of the at least one branch voices is one in a case that the absolute value of the difference between the average value of each branch voice and the set threshold is greater than or equal to the difference threshold.

9. The method according to any one of claims 1 to 8, wherein the controlling the voice communication by setting a communication voice control manner based on the quantity of the branch voices in response to determining that the quantity of the branch voices is more than one comprises:

extracting a host voice by using a set voice extraction manner.

10. The method according to claim 9, wherein the extracting a host voice based on a set voice extraction manner in response to determining that the quantity of the branch voices is more than one comprises:

recognizing, based on an energy proportion corresponding to each of the more than one branch voice at each frequency point, a branch voice corresponding to a maximum value of the energy proportion as the host voice;

determining frequency information corresponding to the host voice from the energy information; and

extracting the host voice from the communication voice based on the frequency information.

11. The method according to any one of claims 1 to 8, wherein the quantity of the branch voices is one or at least two; and

the controlling the voice communication by setting a communication voice control manner based on the quantity of the branch voices comprises:

in a case that the quantity of the at least one branch voice is one, based on a set single talk echo processing manner, recognizing a return voice of the branch voice, and eliminating a single talk echo of the return voice; and

in a case that the quantity of the at least one branch voice is at least two, based on a set double talk echo processing manner, recognizing a return voice corresponding to each of the at least two branch voices, and eliminating a double talk echo of the return voice.

12. A voice communication control apparatus, comprising:

an obtaining unit, configured to obtain a mixed communication voice, the mixed communication voice comprising at least one branch voice;

a transform unit, configured to perform frequency domain transformation on the communication voice to determine energy information of the communication voice at each frequency point in a frequency domain;

a separation unit, configured to perform separation processing on the energy information at each frequency point by using a neural network to determine an energy proportion of each branch voice at each frequency point in total energy at the frequency point;

a quantity unit, configured to determine a quantity of the at least one branch voice comprised in the communication voice based on the energy proportion of each branch voice at each frequency point; and

a control unit, configured to control the voice communication by setting a communication voice control manner based on the quantity of the branch voices.

~~13. The apparatus according to claim 12, wherein the transform unit comprises:~~

~~a framing unit, configured to perform framing on the communication voice to obtain at least one audio frame of voice information;~~

~~a frequency domain transform unit, configured to perform frequency domain transformation on each of the at least one audio frame of the voice information to obtain a voice energy spectrum in the frequency domain; and~~

~~an energy unit, configured to determine, based on the voice energy spectrum, energy information of the communication voice at each frequency point in the frequency domain.~~

~~14. The apparatus according to claim 13, wherein the frequency domain transform unit is configured to:~~

~~perform Fourier transform on each of the at least one audio frame of the voice information in the frequency domain to obtain a voice energy spectrum of the audio frame of the voice information in the frequency domain.~~

~~15. The apparatus according to claim 13, wherein the energy unit is configured to:~~

~~perform modulo processing on an amplitude of the voice energy spectrum at each frequency point to obtain an amplitude spectrum corresponding to the voice energy spectrum; and~~

~~calculate a square value of the amplitude spectrum, and perform a logarithmic operation on the square value to generate the energy information of the communication voice at each frequency point in the frequency domain.~~

~~16. The apparatus according to claim 12, wherein the neural network comprises a long-short-~~

~~term memory neural network, and~~

~~the separation unit is configured to:~~

~~input the energy information into a preset voice separation model, performing convolution processing by using the long short-term memory neural network, and determine the branch voice at each frequency point; and~~

~~determine an energy proportion of each branch voice at each frequency point at each frequency point based on the energy information of the branch voice at the frequency point.~~

~~17. The apparatus according to claim 16, further comprising an update unit, configured to:~~

~~obtain a first voice sample corresponding to a single voice, and a second voice sample corresponding to a mixed voice comprising the single voice;~~

~~extract a first voice feature from the first voice sample, and extracting a second voice feature from the second voice sample;~~

~~input the second voice feature into a voice separation model constructed based on the long short-term memory artificial neural network, to determine a predicted voice separated from the second voice feature and a predicted energy proportion corresponding to the predicted voice in the second voice sample; and~~

~~update parameters of the voice separation model according to a comparison result between an actual energy proportion of the first voice sample in the second voice sample and the predicted energy proportion.~~

~~18. The apparatus according to claim 16, wherein the quantity unit comprises:~~

~~an average unit, configured to calculate, based on the energy proportion of each branch voice at each frequency point, an average value of energy proportions of the branch voice; and~~

~~a voice quantity unit, configured to determine the quantity of the at least one branch voice comprised in the communication voice based on the average value of each branch voice and a set threshold.~~

~~19. The apparatus according to claim 18, wherein the voice quantity unit comprises:~~

a first quantity unit, configured to determine that the quantity of the at least one branch voice is more than one in a case that an absolute value of a difference between the average value of each branch voice and the set threshold is less than a difference threshold; and

a second quantity unit, configured to determine that the quantity of the at least one branch voice is one in a case that the absolute value of the difference between the average value of each branch voice and the set threshold is greater than or equal to the difference threshold.

20. The apparatus according to any one of claims 12 to 19, wherein the control unit comprises:

an extraction unit, configured to extract a host voice by using a set voice extraction manner.

21. The apparatus according to claim 20, wherein the extraction unit is configured to:

recognize, based on an energy proportion corresponding to each of the more than one branch voice at each frequency point, a branch voice corresponding to a maximum value of the energy proportion as the host voice;

determine frequency information corresponding to the host voice from the energy information; and

extract the host voice from the communication voice based on the frequency information.

22. The apparatus according to any one of claims 12 to 19, wherein the quantity of the branch voices is one or at least two, and

the control unit controlling the voice communication by setting a communication voice control manner based on the quantity of the branch voices comprises:

in a case that the quantity of the at least one branch voice is one, based on a set single talk echo processing manner, recognizing a return voice of the branch voice, and eliminating a single talk echo of the return voice; and

in a case that the quantity of the branch voices is at least two, based on a set double talk echo processing manner, respectively recognizing a return voice corresponding to each of the branch voices, and eliminating a double talk echo of the return voice.

~~23~~13. A computer-readable storage medium, storing a computer program, the computer program, when executed by a processor, implementing the voice communication control method according to any one of claims 1 to 11.

~~24~~14. An electronic device, comprising:

one or more processors; and

a storage apparatus, configured to store one or more programs, the one or more programs, when executed by the one or more processors, causing the one or more processors to implement the voice communication control method according to any one of claims 1 to 11.