

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

5 **[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.

10 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

15 **[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
20 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
25 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
30 disclosure.

[0006] According to an aspect of the embodiments of the present disclosure, a voice communication control method is provided, which includes: obtaining a mixed communication voice, the mixed communication voice including 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 each frequency point in total energy at the frequency point; determining a quantity of the at least one branch voice included 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.

[0007] According to an aspect of the embodiments of the present disclosure, a voice communication control apparatus is provided, which includes: an obtaining unit, configured to obtain a mixed communication voice, the mixed communication voice including 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 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 included 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.

[0008] According to an aspect of the embodiments of the present disclosure, a computer-readable medium is provided, which stores a computer program. The computer program, when executed by a processor, implements the voice communication control method according to this embodiment of the present disclosure.

[0009] According to an aspect of this embodiment of the present disclosure, an electronic device is provided, including: 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, cause the one or more processors to implement the voice communication control method according to this embodiment of the present disclosure.

[0010] According to an aspect of this embodiment of the present disclosure, a computer program product or a computer program is provided, the computer program product or the computer

program includes computer instructions, the computer instructions is stored in a computer-readable storage medium. A processor of a computer device reads the computer instructions from the computer-readable storage medium. The processor executes the computer instructions, to cause the computer device to implement the voice communication control method according to various implementations of the present disclosure.

[0011] It is to be understood that the foregoing general descriptions and the following detailed descriptions are merely for illustration and explanation purposes and are not intended to limit the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG. 1 is a schematic diagram of an exemplary system architecture to which a technical solution according to an embodiment of the present disclosure is applicable.

[0013] FIG. 2 is a schematic diagram of a conference system according to some embodiments of the present disclosure.

[0014] FIG. 3 is a schematic flowchart of a voice communication control method according to some embodiments of the present disclosure.

[0015] FIG. 4 is a schematic diagram of a voice separation process according to some embodiments of the present disclosure.

[0016] FIG. 5 is a schematic diagram of voice extraction according to some embodiments of the present disclosure.

[0017] FIG. 6 is a schematic diagram of conference voice extraction according to some embodiments of the present disclosure.

[0018] FIG. 7 is a schematic diagram of an application scenario of echo elimination according to some embodiments of the present disclosure.

[0019] FIG. 8 is a schematic diagram of echo elimination according to some embodiments of the present disclosure.

[0020] FIG. 9 is a schematic flowchart of a process of extracting energy information according to some embodiments of the present disclosure.

[0021] FIG. 10 is a schematic flowchart of a process of extracting energy information according to some embodiments of the present disclosure.

[0022] FIG. 11 is a schematic flowchart of a process of training a separation model according to some embodiments of the present disclosure.

[0023] FIG. 12 is a schematic diagram of an interface for setting a host according to some embodiments of the present disclosure.

5 **[0024]** FIG. 13 is a schematic diagram of an interface for allocating communication resources according to some embodiments of the present disclosure.

[0025] FIG. 14 is a schematic block diagram of a voice communication control apparatus according to some embodiments of the present disclosure.

10 **[0026]** FIG. 15 is a schematic structural diagram of a computer system adapted to implement an electronic device according to an embodiment of the present disclosure.

DESCRIPTION OF EMBODIMENTS

[0027] Cloud computing is a computing mode, in which computing tasks are distributed on a resource pool formed by a large quantity of computers, so that various application systems can obtain computing power, storage space, and information services according to requirements. A network that provides resources is referred to as a "cloud". For a user, resources in a "cloud" seem to be infinitely expandable, and can be obtained readily, used on demand, expanded readily, and paid according to usage. As a basic capability provider of cloud computing, a cloud computing resource pool (which is referred to as a cloud platform for short, and is generally referred to as an Infrastructure as a Service (IaaS)) platform is built, and multiple types of virtual resources are deployed in the resource pool for external customers to choose. The cloud computing resource pool mainly includes: a computing device (a virtualized machine including an operating system), a storage device, and a network device. According to the division of logical functions, a Platform as a Service (PaaS) layer may be deployed on the Infrastructure as a Service (IaaS) layer, and the Software as a Service (SaaS) layer may be deployed on the PaaS layer, or the SaaS layer may be directly deployed on the IaaS layer. PaaS is a platform on which software runs, such as a database and a web container. SaaS is a variety of service software, such as a web portal and an SMS group sender. Generally, SaaS and PaaS are upper layers relative to IaaS.

30 **[0028]** A cloud call center is a call center system based on the cloud computing technology. An enterprise does not need to purchase any software or hardware systems, but only needs to have basic conditions such as personnel and venues to quickly own a call center belonging to the enterprise. Software and hardware platforms, communication resources, routine maintenance and services are provided by service providers. The cloud call center has multiple features such as short construction

period, low investment, low risk, flexible deployment, strong system capacity scalability, low operation and maintenance cost and so on. Whether the cloud call center is a telemarketing center or a customer service center, by renting services on demand, the enterprise can build a call center system with comprehensive, stable and reliable functions, where the seats can be distributed all over the country and to receive calls all over the country supporting national call access.

[0029] In this embodiment of the present disclosure, the cloud call center may be used for calling a center system, and in addition, a voice communication control method may be embedded in the system to implement the voice control in a call process, thereby causing the cloud call center to be more intelligent, and improving the reliability and security of the cloud call center.

[0030] The cloud conference is an efficient, convenient and low-cost conference form based on the cloud computing technology. Users can quickly, efficiently and synchronously share voice, data files and videos with teams and customers around the world by only performing simple and easy operations through Internet interfaces. Complex technologies such as data transmission and processing in conferences are provided by a cloud conference service provider to assist the users to operate. At present, domestic cloud conferences mainly focus on service content of a software as a service (SaaS) mode, including calls, networks, videos and other service forms. Video conferences based on the cloud computing are referred to as cloud conferences. In the cloud conferences, data transmission, processing, and storage are all performed by computer resources of video conference manufacturers providers, and the users do not need to purchase expensive hardware or install cumbersome software at all. The users only need to open browsers and log in to corresponding interfaces to conduct efficient teleconferences. A cloud conference system supports multi-server dynamic cluster deployment and provides multiple high-performance servers, which greatly improves stability, security and availability of conferences. In recent years, because video conferences can greatly improve communication efficiency, continuously reduce communication costs, and bring an upgrade in internal management level, the video conferences are popular among many users and have been widely used in many fields such as government, military, transportation, transmission, finance, operators, education, enterprises and so on. There is no doubt that after using the cloud computing, the video conferences will be more attractive competitive in terms of convenience, speed and ease of use, which will surely stimulate arrival of a new upsurge in video conference applications.

[0031] In the application scenario of the cloud conference, the voice communication control method may be applied to the cloud conference in this embodiment of the present disclosure, so that the voice communication in a process of the cloud conference is smooth, and a voice communication process is intelligent, thereby improving the efficiency of the conference.

[0032] Cloud social is a virtual social application mode for interactive applications of the Internet of Things, cloud computing, and mobile internet, aims to establish a famous "resource sharing relationship map", and develops network social. A main feature of the cloud social is the unified integration and evaluation of a large quantity of social resources to form an effective resource pool to provide services to the user on demand. The more users who participate in the sharing, the greater the utilization value that may be created.

[0033] Artificial intelligence (AI) is a theory, a method, a technology, and an application system that use a digital computer or a machine controlled by the digital computer to simulate, extend, and expand human intelligence, perceive an environment, obtain knowledge, and use knowledge to obtain an optimal result. In other words, the AI is a comprehensive technology of computer sciences, attempts to understand essence of intelligence, and produces a new intelligent machine that can react in a similar manner to human intelligence. The AI is to study the design principles and implementation methods of various intelligent machines, to enable the machines to have the functions of perception, reasoning, and decision-making. The AI technology is a comprehensive discipline, covering a wide range of fields including both a hardware-level technology and a software-level technology. The basic AI technology generally includes a technology such as a sensor, a dedicated AI chip, cloud computing, distributed storage, a big data processing technology, an operation/interaction system, or mechatronics. An AI software technology mainly includes fields such as a computer vision technology, a voice processing technology, a natural language processing technology, and machine learning/deep learning.

[0034] Key technologies of the speech technology include an automatic speech recognition (ASR) technology, a text-to-speech (TTS) technology, and a voice print recognition technology. To make a computer capable of listening, seeing, speaking, and feeling is the future development direction of human-computer interaction, and speech has become one of the most promising human-computer interaction methods in the future. Machine learning (ML) is a multi-disciplinary subject involving multiple disciplines such as probability theory, statistics, approximation theory, convex analysis, and algorithm complexity theory. The machine learning specializes in studying how a computer simulates or implements a human learning behavior to obtain new knowledge or skills, and reorganize an existing knowledge structure, to keep improving its performance. The machine learning is a core of the AI, is a basic way to make the computer intelligent, and is applied to various fields of the AI. The ML and deep learning generally include technologies such as an artificial neural network, a belief network, reinforcement learning, transfer learning, inductive learning, and learning from demonstrations.

[0035] With the research and progress of the AI technology, the AI technology is studied and applied in multiple fields such as a common smart home, a smart wearable device, a virtual assistant, a smart speaker, smart marketing, unmanned driving, automatic driving, an unmanned aerial vehicle, a robot, smart medical care, and smart customer service. It is believed that with the development of technologies, the AI technology will be applied to more fields, and play an increasingly important role.

[0036] 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 is the quantity of speakers. A voice segment is given during testing, and the system predicts the current quantity of speakers. However, in the solution, more context information is required in a detection process to determine the current quantity of speakers. For example, long-term voice segments need to be collected as a data basis to determine the 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.

[0037] The solutions provided in the embodiments of the present disclosure involve technologies such as voice technology and machine learning of artificial intelligence. By using the technologies, the voice communication control method in this embodiment of the present disclosure may be more accurate, which is specifically described by the following embodiments.

[0038] FIG. 1 is a schematic diagram of an exemplary system architecture to which a technical solution according to an embodiment of the present disclosure is applicable.

[0039] As shown in FIG. 1, the system architecture may include a terminal device (which may be any one or more of a smart phone 101, a tablet computer 102, and a portable computer 103 shown in FIG. 1, or certainly may be a desktop computer, and the like), a network 104, and a server 105. The network 104 is configured to provide a medium of a communication link between the terminal device and the server 105. The network 104 may include various connection types, for example, a wired communication link and a wireless communication link.

[0040] It is to be understood that the quantities of terminal devices, networks, and servers in FIG. 1 are merely exemplary. There may be any quantities of terminal devices, networks and servers according to an actual requirement. For example, the server 105 may be a server cluster including multiple servers.

[0041] It is to be noted that each terminal device in this embodiment may be directed to different communication clusters, where the quantity of participants in the communication cluster

may be one, two, or more, or the like. For example, a communication cluster targeted by a portable computer 103 may include multiple participants, a communication cluster targeted by a tablet computer 102 may include other participants, and the user may further participate in a conference through a smart phone 101.

5 **[0042]** For example, during a conference, multiple users or one user may use a terminal device to make a conference communication. In addition, a server 105 may obtain a communication voice between the terminal devices through a network 104, perform frequency domain transformation on the communication voice, and determine energy information corresponding to the communication voice in a frequency domain; perform separation on the energy information based on a neural network
10 to determine an energy proportion of each branch voice included in the communication voice in the communication voice; determine a quantity of branch voices included in the communication voice based on the energy proportion; and control the voice communication by setting a communication voice control manner based on the quantity of the branch voices.

[0043] In the above solution, the communication voice is obtained in real time during the
15 communication, and frequency domain transformation is performed on the communication voice to determine the corresponding energy information of the communication voice in the frequency domain. Then, separation is performed on the energy information based on the neural network, and the energy proportion of each branch voice included in the communication voice in the communication voice is determined, so as to determine the quantity of branch voices included in the
20 communication voice based on the energy proportion. Finally, based on the quantity of branch voices, the voice communication is controlled by setting the communication voice control manner, thereby implementing the real-time detection on the quantity of people during the voice communication and the real-time control of the voice communication. In this way, the accuracy of detecting the quantity of people in the voice communication is ensured, and the efficiency of recognizing the quantity of
25 people in the voice communication and the efficiency of controlling the voice communication are improved.

[0044] In addition, as shown in FIG. 2, in this embodiment, processing may be performed for only one communication cluster, and the communication cluster includes one, two, or more participants. Through the voice communication control method, the real-time quantity of speakers in
30 the communication cluster may be detected, and communication control may be performed to ensure the voice quality during the communication and improve the communication efficiency.

[0045] It is to be noted that the voice communication control method provided in the embodiments of the present disclosure is generally performed by the server 105, and accordingly, the

voice communication control apparatus is generally arranged in the server 105. However, in another embodiment of the present disclosure, the terminal device may have similar functions to the server, so as to perform the voice communication control solution provided in this embodiment of the present disclosure.

5 **[0046]** It is to be noted that the server in this embodiment may be an independent physical server, or may be a server cluster including multiple physical servers or a distributed system, or may be a cloud server providing a cloud computing service. The terminal may be a smartphone, a tablet computer, a notebook computer, a desktop computer, a smart speaker, a smart watch, or the like, but is not limited thereto. The terminal and the server may be directly or indirectly connected in a wired
10 or wireless communication manner, which is not limited in the present disclosure.

[0047] The details of implementing the technical solution of this embodiment of the present disclosure are described in detail as follows.

[0048] FIG. 3 is a schematic flowchart of a voice communication control method according to an embodiment of the present disclosure. The voice communication control method may be
15 performed by a server, and the server may be the server shown in FIG. 1. With reference to FIG. 3, the voice communication control method includes at least step S110 to step S150. A detailed description is as follows.

[0049] In step S110, a mixed communication voice is obtained, the mixed communication voice includes at least one branch voice.

20 **[0050]** In some embodiments of the present disclosure, the mixed communication voice may be obtained during the communication in the communication cluster. A length of the communication voice is not limited in this embodiment, and an audio frame of communication voice, or a communication voice with a duration of one second or one minute, or the like may be obtained in real time.

25 **[0051]** For example, the communication may be a real-time communication conference scenario. In a process of the real-time communication conference, the communication voice may be collected in real time, the collected communication voice is recognized, and subsequent control is performed based on a recognition result, so as to implement the effect of real-time control for the communication voice.

30 **[0052]** In step S120, frequency domain transformation is performed on the communication voice to determine energy information of the communication voice at each frequency point.

[0053] In an embodiment of the present disclosure, the communication voice obtained in this embodiment is in the time domain, which is a voice signal with time as an independent variable and volume as a dependent variable. In this embodiment, after the communication voice is obtained, frequency domain transformation is performed on the communication voice, and a voice signal in the time domain is converted into a voice signal in the frequency domain, so as to reflect the energy information of the communication voice in the frequency domain.

[0054] As shown in FIG. 4, in a speaker scenario classification framework based on voice separation, and in a "signal preprocessing" stage, in this embodiment of the present disclosure, an audio signal is obtained, and the audio signal is preprocessed to extract acoustic features. A logarithmic energy spectrum corresponding to the audio signal is generated, which is used as energy information at each frequency point of the communication voice in the frequency domain.

[0055] Specifically, the energy information in this embodiment may include information such as an energy value and an energy spectrum at each frequency point of the communication voice. In this embodiment, attributes such as energy of each frequency point may be measured by using the energy information, to distinguish the branch voices at each frequency point based on the energy information.

[0056] In step S130, separation is performed 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.

[0057] In an embodiment of the present disclosure, after the energy information at each frequency point is obtained, separation is performed on the energy information by using the neural network. That is, the energy information is inputted into the trained separation model to obtain the energy proportion of each branch voice at the frequency point, which is also referred to as a frequency point coefficient based on voice separation. For example, in FIG. 4, using two branch voices as an example, frequency point coefficients of the two branch voices at one frequency point may be obtained, that is, a frequency point coefficient P_A and a frequency point coefficient P_B . It is to be noted that only two branch voices are used as an example in FIG. 4, and FIG. 4 shows two frequency point coefficients at one frequency point, corresponding to a speaker A and a speaker B respectively. In practice, frequency point coefficients obtained at one frequency point are related to the quantity of people simultaneously speaking, which are not limited to two. In addition, each frequency point whose energy value is not 0 may obtain a frequency point coefficient based on voice separation.

[0058] Specifically, the frequency point coefficient in this embodiment is used for representing a proportion of the energy corresponding to a specific speaker at a frequency point in

the total energy information of the frequency point. The idea of voice separation in this embodiment is based on a manner of each frequency point coefficient in the frequency domain. The energy proportion of a speaker at a specific frequency point is proportional to a value of the frequency point coefficient predicted in the mixed signal. A frequency point coefficient (P) is calculated by dividing
5 a voice energy value of mixed speakers at the frequency point by a voice energy value (E) of a speaker at a frequency point. It is assumed that there are two people (A and B), the frequency point coefficient of the speaker A at a frequency point is calculated by the following formula:

$$P_A = \frac{E_A}{E_A + E_B}$$

[0059] After the frequency point coefficient P_A and the frequency point coefficient P_B at a
10 frequency point are calculated by the formula, when P_A is greater than P_B , the frequency point is dominated by the speaker A, and when P_A is less than P_B , the frequency point is dominated by the speaker B.

[0060] In the method, especially in a scenario in which multiple people simultaneously speaking, by separating the energy information, an energy proportion corresponding to each branch
15 voice at each frequency point may be determined, and the distribution of each branch voice may be determined based on the energy proportion, thereby improving the accuracy and real-time performance of recognizing the quantity of voices.

[0061] In step S140, a quantity of the at least one branch voice included in the communication voice is determined based on the energy proportion of each branch voice at each frequency point.

[0062] In an embodiment of the present disclosure, after the energy proportion of each branch
20 voice at each frequency point in the communication voice is determined, in this embodiment, the quantity of branch voices included in the communication voice is determined by averaging the energy proportions.

[0063] In an embodiment of the present disclosure, for the energy proportion of a voice of the
25 speaker A corresponding to each frequency point in each frame, the energy proportions at each frequency point in each frame is averaged to obtain a stable average value of energy within a time period of one frame. Then, it is determined whether a current average value of energy corresponds to a voice of a single speaker or voices of multiple speakers based on a set threshold, and finally, information of the quantity of speakers in a current frame is outputted. For example, discrete 0 (a
30 voice of a single speaker) or 1 (voices of multiple speakers) corresponding to the current quantity of speakers in each frame may be outputted.

[0064] In an embodiment of the present disclosure, the collected communication voice includes many frames, and there are multiple frequency points in one frame. For example, the quantity of frequency points may be the quantity of points obtained by performing Fourier transform, and the quantity of frequency points in a frame is set as f . Fi is an energy proportion at the i^{th} frequency point of a specific speaker at the frequency point, that is, the frequency point coefficient. An average value of the energy proportions corresponding to the speaker in this frame is obtained as $\frac{1}{f} \sum_{i=0}^{f-1} Fi$ by averaging. Then, the average value corresponding to each branch voice may be compared with a threshold to determine the quantity of speakers simultaneously speaking. For example, using two speakers as an example, because $P_A + P_B = 1$, it is only required to compare one of P_A and P_B with the threshold. For example, using P_A as an example, when the quantity of speakers is determined in a voice communication, and when the value is 0 or 1, it indicates that current voice energy (P_B or P_A) occupies all the voice energy. Therefore, only one speaker is speaking, that is, B or A is speaking, and the quantity of speakers is 1 in this case; and if the value is 0.5, it indicates that two people are simultaneously speaking, the voice energies are the same in this case, and the quantity of speakers is 2. By comparing the calculated average value of the energy proportion with the set threshold, the current quantity of branch voices may be determined. In practice, the threshold may be set according to the application scenario.

[0065] In this embodiment, the quantity of current branch voices is determined by the threshold detection manner, so that a multi-speaker scenario may be determined within a time period of a very short frame level in real time, and a voice stream may be processed in real time; and multi-to-multi labels may be used to make full use of voice information to improve the accuracy of scenario detection.

[0066] In step S150, the voice communication is controlled by setting a communication voice control manner based on the quantity of the at least one branch voice.

[0067] In an embodiment of the present disclosure, after the quantity of the branch voice is determined, a current communication state may be determined with reference to the current quantity of branch voices. Then, the voice communication is controlled through the set communication voice control manner, so that the voice communication can be controlled in real-time, thereby improving the accuracy and real-time performance of the voice control.

[0068] For example, as shown in FIG. 5, when the quantity of branch voices is more than one, based on the quantity of the branch voices, a background speaker is filtered out, and only an audio stream of the host is extracted. In this case, the quantity of current speakers needs to be detected at a front end. If the quantity of current speakers is greater than 1, voice extraction of the host is enabled;

and if it is detected that the current quantity of speakers is 1, the voice extraction is disabled to avoid damage to the voice.

[0069] In this embodiment, in a process of voice extraction, based on an energy proportion of each of the multiple branch voices at each frequency point, a branch voice corresponding to a maximum value of the energy proportion is recognized as the host voice; frequency information corresponding to the host voice is determined from the energy information; and the host voice is extracted from the communication voice based on the frequency information.

[0070] As shown in FIG. 6, in a scenario in which multiple users are speaking, the host may be determined by the frequency detection manner, as shown in the user 4 in FIG. 6, and the voice of the host is extracted, or the voices of other users are filtered out to cause the voice communication to be clear.

[0071] In the manner, the speaking voice of the host may be extracted in a scenario in which multiple people are speaking, so that the voice during the communication is clear, and the quality and effect of the communication are improved.

[0072] In a case of the voice broadcast shown in FIG. 7, one party of the communication hears his own echo from the voice returned by the other party, which leads to a reduced quality of the communication.

[0073] To avoid the situation, as shown in FIG. 8, if it is a single talk scenario in which a communication peer and a current end take turns to speak, that is, the quantity of branch voices is one, based on a set single talk echo processing manner, a return voice of the branch voice is recognized, and a single talk echo of the return voice is eliminated. In a double talk scenario in which the communication peer and the current end simultaneously speak, that is, the quantity of branch voices is at least two, based on a set double talk echo processing manner, a return voice corresponding to the branch voice is respectively recognized, and a double talk echo of the return voice is eliminated. In the communication system, a signal of the current end is ensured to be retained to the greatest extent in an echo elimination process.

[0074] In the above solution, the communication voice is obtained in real time during the communication, and frequency domain transformation is performed on the communication voice to determine the corresponding energy information of the communication voice in the frequency domain. Then, separation is performed on the energy information by using the neural network, and the energy proportion of each branch voice included in the communication voice in the communication voice is determined, so as to determine the quantity of branch voices included in the

communication voice based on the energy proportion. Finally, based on the quantity of branch voices, the voice communication is controlled by setting the communication voice control manner, thereby implementing the real-time detection of the quantity of people during the voice communication and the real-time control of the voice communication. In this way, the accuracy of detecting the quantity of people in the voice communication is ensured, and the efficiency of recognizing the quantity of people in the voice communication and the efficiency of controlling the voice communication are improved.

[0075] In an embodiment of the present disclosure, as shown in FIG. 9, the step S120 of performing the frequency domain transformation on the communication voice to determine a process of the energy information corresponding to the communication voice in the frequency domain includes step S1210 to step S1230, and the details are as follows.

[0076] In step S1210, framing is performed on the communication voice to obtain at least one audio frame of voice information.

[0077] In step S1220, frequency domain transformation is performed on each audio frame of the voice information to obtain a voice energy spectrum of the frequency domain.

[0078] In step S1230, energy information at each frequency point of the communication voice in the frequency domain is determined based on the voice energy spectrum.

[0079] In an embodiment of the present disclosure, the performing frequency domain transformation on each audio frame of voice information to obtain a voice energy spectrum of the frequency domain includes: performing Fourier transform or other operations of converting from a time domain to a frequency domain on each audio frame of the voice information in the frequency domain to obtain a voice energy spectrum of the voice information in the frequency domain.

[0080] In this embodiment, the step S1230 of determining the energy information at each frequency point of the communication voice in the frequency domain based on a voice energy spectrum includes: performing modulo processing on an amplitude corresponding to each frequency point in the voice energy spectrum 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 at each frequency point of the communication voice in the frequency domain.

[0081] As shown in FIG. 10, in an embodiment of the present disclosure, framing and windowing are performed on the collected time domain voice; and N-point Fourier transform is performed on each frame to calculate a Fourier transform coefficient of the N frequency points

obtained by the Fourier transform. Modulo processing is performed on each frame to obtain an amplitude spectrum of the frequency domain, a logarithm of the square of the obtained amplitude spectrum is calculated to obtain a logarithmic energy spectrum, and energy information of the voice at the N frequency points may be obtained.

5 **[0082]** In an embodiment of the present disclosure, the neural network includes a long-short-term memory neural network, and the step S130 of performing separation on the energy information based on the neural network to determine the energy proportion of each branch voice included in the communication voice includes: inputting the energy information into a preset voice separation model, performing convolution processing by using the long short-term memory neural network, and
10 determining the energy proportion of each branch voice included in the communication voice at each frequency point.

[0083] In an embodiment of the present disclosure, as shown in FIG. 11, in a process of training the voice separation model, the method further includes the following steps S1110 to S1140.

[0084] In step S1110, a first voice sample corresponding to a single voice, and a second voice
15 sample corresponding to a mixed voice including the single voice are obtained.

[0085] In step S1120, a first voice feature is extracted from the first voice sample, and a second voice feature is extracted from the second voice sample.

[0086] In step S1130, the second voice feature is inputted into a voice separation model constructed based on the long short-term memory artificial neural network, and a predicted voice
20 separated from the second voice feature is determined, and an energy proportion corresponding to the predicted voice in the second voice sample is predicted.

[0087] In step S1140, parameters of the voice separation model are updated 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.

25 **[0088]** In an embodiment of the present disclosure, first, a dataset used for training is constructed to obtain two audio libraries, a single speaker voice corpus and a multi-speaker voice corpus. The single-speaker voice corpus includes a first voice sample corresponding to a single voice. The multi-speaker voice corpus refers to randomly extracting multiple single speaker voice segments for superposition, extracting logarithmic energy spectrum in the voice from the two databases through
30 signal preprocessing, obtaining each frequency point coefficient through a separation model, and obtaining a separated voice through post-processing. In this embodiment, the extracted logarithmic energy spectrum feature is used as an input, and the feature is inputted into a separation model

including a two-layer long short-term memory (LSTM) network and an output layer to obtain a frequency point coefficient.

[0089] It is to be noted that a reason for using the LSTM network in this embodiment is that the LSTM network not only considers the input at a current moment, but also has a memory function for previous content. In addition, an input gate, an output gate, a forget gate, and a cell state unit added in the network structure in this embodiment cause the time series modeling capability of the LSTM to be significantly improved, so that more information is memorized, and long-term dependencies in data are effectively captured.

[0090] In the training of the whole model, a training set includes voices of multiple speakers and a voice of a single speaker. The voices of multiple speakers are obtained by adding voices of multiple single speakers, and the voice of the single speaker may be regarded as a mixture with silence. A voice feature a is extracted from the mixed voice, a voice feature b is extracted from a corresponding clean voice, a feature vector a is used as a training input, a feature vector b is used as a training target, and frequency point coefficients m and n are outputted by a model. The separated voice is obtained through post-processing, an error between the separated voice and the clean voice is used for guiding the training of the LSTM model, and parameters in the LSTM model are adjusted to obtain a final separation model, thereby obtaining an accurate and complete separation model.

[0091] In the voice separation model obtained in the training process, a multi-speaker scenario may be determined within a very short time period at a frame level in real time, and a voice stream is processed in real time; and in addition, based on a label corresponding to each frequency point in the frame, in a manner of multiple frequency points corresponding to multiple labels, the voice information may be fully used to improve the accuracy of scenario detection.

[0092] Further, in this embodiment, in step S1130, in a process of inputting the second voice feature into a voice separation model constructed based on the long short-term memory artificial neural network and determining a predicted voice separated from the second voice feature, the obtained frequency point coefficient is multiplied by a spectrum of a mixed voice, and a separated voice signal may be obtained through inverse Fourier transform in combination with a phase of a mixed signal.

[0093] The step S140 of determining a quantity of branch voices included in the communication voice based on the energy proportion of each branch voice at each frequency point includes: calculating an average value of the energy proportions of branch voice based on the energy proportion of each branch voice included in the communication voice at each frequency point; and

determining the quantity of branch voices included in the communication voice based on the average value of each branch voice and a set threshold.

[0094] In an embodiment of the present disclosure, the collected communication voice includes many frames, and there are multiple frequency points in one frame. The quantity of frequency points in a frame is set as f . F_i is an energy proportion corresponding to the i^{th} frequency point of a specific speaker in the frequency point, that is, the frequency point coefficient. An average value of the energy proportions corresponding to the speaker in this frame is obtained by averaging according to the following formula: $\frac{1}{f} \sum_{i=0}^{f-1} F_i$.

[0095] In an embodiment of the present disclosure, the determining the quantity of branch voices included in the communication voice based on the average value of each branch voice and the set threshold includes: determining that the quantity of the branch voices 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 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.

[0096] Specifically, in this embodiment, using two people simultaneously speaking as an example, when an average value is closer to 0 or closer to 1, the probability of a single speaker speaking is greater, and the closer the average value is to 0.5, the probability of two speakers simultaneously speaking is greater. A threshold is determined according to a specific task. For example, in an extraction algorithm of a host, to avoid the damage to the voice made by the algorithm in a voice of a single speaker, the false detection rate of the voice of the single speaker needs to be determined to be low, and a threshold may be set close to 0 or 1 in this case.

[0097] As shown in FIG. 12, in the actual conference application scenario, when it is detected that the current quantity of speakers is more than one, the detected speaker or participant may be displayed in the interface, and in response to a user-triggered instruction for setting a host, one or more people among the speakers are set as the host, so as to filter the audio of the other people to ensure the quality of the communication in the conference.

[0098] As shown in FIG. 13, for a session control server, when there are multiple voice sessions simultaneously performed, more communication resources may be allocated to a conference with a large quantity of hosts to ensure the quality of the communication.

[0099] The following describes apparatus embodiments of the present disclosure, and the apparatus embodiments may be used for performing the voice communication control method in the

foregoing embodiment of the present disclosure. It is to be understood that the apparatus may be a computer program (including program code) run in a computer device. For example, the apparatus is an application software; and the apparatus may be configured to perform corresponding steps in the methods provided in this embodiment of the present disclosure. For details not disclosed in the apparatus embodiments of the present disclosure, reference may be made to the foregoing voice communication control method embodiments of the present disclosure.

[0100] FIG. 14 is a schematic block diagram of a voice communication control apparatus according to some embodiments of the present disclosure.

[0101] As shown in FIG. 14, the voice communication control apparatus 1400 according to an embodiment of the present disclosure includes: an obtaining unit 1410, a transform unit 1420, a separation unit 1430, a quantity unit 1440 and a control unit 1450. The obtaining unit 1410 is configured to obtain a mixed communication voice. The transform unit 1420 is configured to perform frequency domain transformation on the communication voice to determine energy information at each frequency point of the communication voice. The separation unit 1430 is configured to perform separation on the energy information at each frequency point based on a neural network to determine an energy proportion of each branch voice included in the communication voice at each frequency point in the communication voice. The quantity unit 1440 is configured to determine a quantity of branch voices included in the communication voice based on the energy proportion of each branch voice at each frequency point. The control unit 1450 is configured to control the voice communication by setting a communication voice control manner based on the quantity of the branch voices.

[0102] In some embodiments of the present disclosure, based on the foregoing solution, the transform unit 1420 includes: a framing unit, a frequency domain transform unit and an energy unit. The framing unit is configured to perform framing on the communication voice to obtain at least one audio frame of voice information. The frequency domain transform unit is configured to perform frequency domain transformation on each audio frame of the voice information to obtain a voice energy spectrum of the frequency domain. The energy unit is configured to determine energy information at each frequency point of the communication voice in the frequency domain based on the voice energy spectrum.

[0103] In some embodiments of the present disclosure, based on the foregoing solution, the frequency domain transform unit is configured to: perform Fourier transform on each audio frame of the voice information in the frequency domain to obtain a voice energy spectrum of each audio frame of the voice information corresponding to the frequency domain.

[0104] In some embodiments of the present disclosure, based on the foregoing solution, the energy unit is configured to: perform modulo processing on an amplitude corresponding to each frequency point in the voice energy spectrum 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 at each frequency point of the communication voice in the frequency domain.

[0105] In some embodiments of the present disclosure, based on the foregoing solution, the neural network includes a long-short-term memory neural network. The separation unit 1430 is configured to: input the energy information into a preset voice separation model, perform convolution processing by using the long short-term memory neural network, and determine a branch voice at each frequency point; and determine an energy proportion of each branch voice at each frequency point based on the energy information of each branch voice at each frequency point.

[0106] In some embodiments of the present disclosure, based on the foregoing solution, the voice communication control apparatus 1400 further includes an update unit. The update unit is configured to: obtain a first voice sample corresponding to a single voice, and a second voice sample corresponding to a mixed voice including the single voice; extract a first voice feature from the first voice sample, and extract 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, and 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.

[0107] In some embodiments of the present disclosure, based on the foregoing solution, the quantity unit 1440 includes: an average unit and a voice quantity unit. The average unit is configured to calculate, based on the energy proportion corresponding to each branch voice at each frequency point, an average value of the energy proportions of the branch voice. The voice quantity unit is configured to determine the quantity of branch voices included in the communication voice based on the average value of each branch voice and a set threshold.

[0108] In some embodiments of the present disclosure, based on the foregoing solution, the voice quantity unit includes: a first quantity unit and a second quantity unit. The first quantity unit is configured to determine that the quantity of the branch voices 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. The second quantity unit is configured to determine that the quantity of the 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.

5 **[0109]** In some embodiments of the present disclosure, based on the foregoing solution, the control unit 1450 includes an extraction unit, configured to extract a host voice based on a set voice extraction manner.

10 **[0110]** In some embodiments of the present disclosure, based on the foregoing solution, the extraction unit is configured to recognize, based on an energy proportion corresponding to each of the multiple branch voices 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.

15 **[0111]** In some embodiments of the present disclosure, based on the foregoing solution, the quantity of the branch voices is one or at least two, and the control unit 1450 is configured to: in a case that the quantity of the branch voices is one, based on a set single talk echo processing manner, recognize a return voice of the branch voice, and eliminate 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
20 voices, and eliminating a double talk echo of the return voice.

[0112] It is to be noted that FIG. 15 is a schematic structural diagram of a computer system adapted to implement an electronic device according to an embodiment of the present disclosure.

25 **[0113]** A computer system 1500 of the electronic device shown in FIG. 15 is merely an example, and does not constitute any limitation on functions and use ranges of the embodiments of the present disclosure.

30 **[0114]** As shown in FIG. 15, the computer system 1500 includes a central processing unit (CPU) 1501, which may perform various suitable actions and processing based on a program stored in a read-only memory (ROM) 1502 or a program loaded from a storage part 1508 into a random access memory (RAM) 1503, for example, perform the method described in the foregoing embodiments. The RAM 1503 further stores various programs and data required for operating the system. The CPU 1501, the ROM 1502, and the RAM 1503 are connected to each other through a bus 1504. An input/output (I/O) interface 1505 is also connected to the bus 1504.

[0115] The following components are connected to the I/O interface 1505: an input part 1506 including a keyboard, a mouse, or the like, an output part 1507 including a cathode ray tube (CRT), a liquid crystal display (LCD), a speaker, or the like, a storage part 1508 including a hard disk, or the like, and a communication part 1509 including a network interface card such as a local area network (LAN) card or a modem. The communication part 1509 performs communication processing by using a network such as the Internet. A driver 1510 is also connected to the I/O interface 1505 as required. A removable medium 1511, such as a magnetic disk, an optical disc, a magneto-optical disk, or a semiconductor memory, is installed on the driver 1510 as required, so that a computer program read therefrom is installed into the storage part 1508 as required.

[0116] Particularly, according to an embodiment of the present disclosure, the processes described above by referring to the flowcharts may be implemented as computer software programs. For example, an embodiment of the present disclosure includes a computer program product. The computer program product includes a computer program stored in a computer-readable medium. The computer program includes a computer program used for performing a method shown in the flowchart. In such an embodiment, by using the communication part 1509, the computer program may be downloaded and installed from a network, and/or installed from the removable medium 1511. When the computer program is executed by the CPU 1501, the various functions defined in the system of the present disclosure are executed.

[0117] It is to be noted that the computer-readable medium shown in the embodiments of the present disclosure may be a computer-readable signal medium or a computer-readable storage medium or any combination thereof. The computer-readable storage medium is not limited to, an electric, magnetic, optical, electromagnetic, infrared, or semi-conductive system, apparatus, or component, or any combination thereof. A more specific example of the computer-readable storage medium may include but is not limited to: an electrical connection having one or more wires, a portable computer magnetic disk, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM), a flash memory, an optical fiber, a compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any appropriate combination thereof. In the present disclosure, the computer-readable storage medium may be any tangible medium containing or storing a program, and the program may be used by or used in combination with an instruction execution system, apparatus, or device. In the present disclosure, the computer-readable signal medium may include a data signal transmitted in a baseband or as part of a carrier, and stores a computer-readable computer program. The data signal propagated in such a way may assume multiple forms, including, but not limited to, an electromagnetic signal, an optical signal, or any appropriate combination thereof. The computer-readable signal medium may

be further any computer-readable medium in addition to a computer-readable storage medium. The computer-readable medium may transmit, propagate, or transmit a program that is used by or used in combination with an instruction execution system, apparatus, or device. The computer program included in the computer-readable medium may be transmitted by using any suitable medium, including but not limited to: a wireless medium, a wire, or the like, or any suitable combination thereof.

[0118] According to an aspect of the present disclosure, a computer program product or a computer program is provided, the computer program product or the computer program includes computer instructions, the computer instructions are stored in a computer-readable storage medium. A processor of a computer device reads the computer instructions from the computer-readable storage medium and executes the computer instructions to cause the computer device to perform the method provided in the foregoing various implementations.

[0119] According to another aspect of the present disclosure, a computer-readable medium is provided. The computer-readable medium may be included in the electronic device described in the foregoing embodiments, or may exist alone and is not disposed in the electronic device. The computer-readable medium carries one or more programs, the one or more programs, when executed by the electronic device, causing the electronic device to implement the method described in the foregoing embodiments.

[0120] It is to be understood that the present disclosure is not limited to the precise structures described above and shown in the accompanying drawings, and various modifications and changes may be made without departing from the scope of the present disclosure. The scope of the present disclosure is topic only to the appended claims.

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
5 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
10 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
20 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:

5 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.

10

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:

15 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
20 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;

25 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
30 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

5 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. 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.

10

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
15 the voice communication control method according to any one of claims 1 to 11.