

SPEECH ENHANCEMENT MODEL, ELECTRONIC DEVICE, STORAGE MEDIUM, AND RELATED METHODS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The disclosure claims the benefits of priority to Chinese Application No. 202210022926.5, filed January 10, 2022, which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to speech processing and, in particular, to a speech enhancement model, an electronic device, a storage medium, and related methods.

BACKGROUND

[0003] Speech enhancement is a speech processing technique that extracts useful speech signals from a noisy background after speech signals have been interfered with or annihilated by various noises, to suppress and reduce noise interference. Speech enhancement is widely used in scenarios of real-time communication (RTC), such as audio/video conferencing, special delivery classroom, and the like.

[0004] Currently, to perform speech enhancement on speech data, the time sequence features along the time direction and the short-term frequency features along the frequency direction are extracted from the speech data. Then, speech enhancement is performed on the speech data according to the time sequence features and the short-term frequency features, so as to suppress noise in the speech data.

[0005] However, based on the time sequence features and the short-term frequency features, it is impossible to effectively identify noise with large frequency differences and useful speech signals in the speech data, so it is impossible to filter out the noise with large frequency differences from the useful speech signals, resulting in poor effect of speech enhancement on the speech signals.

SUMMARY OF THE DISCLOSURE

[0006] Embodiments of the present disclosure provide a speech enhancement method. The speech enhancement method includes: converting noisy speech data into time-frequency domain feature data; generating a mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in a frequency direction; and generating enhanced speech data for the noisy speech data according to the mask value and the time-frequency domain feature data.

[0007] Embodiments of the present disclosure provide a speech enhancement model. The speech enhancement model includes an encoder, an attention module, a recurrent module, and a decoder, wherein the encoder is configured to process time-frequency domain feature data corresponding to noisy speech data; the attention module is configured to process output data of the encoder; the recurrent module is configured to process the output data of the encoder; and the decoder is configured to process output data of the attention module and the recurrent module to obtain correlation feature data for characterizing a long-range correlation of the time-frequency domain feature data in a frequency direction.

[0008] Embodiments of the present disclosure provide a speech recognition method. The speech recognition method includes acquiring speech data to be processed, wherein the speech data to be processed comprises noise, and the speech data to be processed comprises one of the following: audio/video conference speech data, online education speech data, and live webcast speech data; converting the speech data to be processed into time-frequency domain feature data; generating a mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in a frequency direction; generating enhanced speech data for the speech data to be

processed according to the mask value and the time-frequency domain feature data; and performing speech recognition on the enhanced speech data to obtain a recognition result.

[0009] Embodiments of the present disclosure provide an electronic device. The electronic device includes a processor, a memory, a communication interface, and a communication bus, the processor, the memory, and the communication interface communicating with each other via the communication bus, wherein the memory is configured to store at least one executable instruction that causes the processor to perform the above-mentioned speech enhancement method or the above-mentioned speech recognition method.

[0010] Embodiments of the present disclosure provide a computer storage medium having a computer program stored thereon. The program, when executed by a processor, implements the above-mentioned speech enhancement method or the above-mentioned speech recognition method.

[0011] Embodiments of the present disclosure provide a computer program product. The computer program product includes computer instructions, wherein the computer instructions instruct a computing device to perform the above-mentioned speech enhancement method or the above-mentioned speech recognition method.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Embodiments and various aspects of the present disclosure are illustrated in the following detailed description and the accompanying figures. Various features shown in the figures are not drawn to scale.

[0013] **FIG. 1** is a schematic diagram of an exemplary system, according to some embodiments of the present disclosure.

[0014] **FIG. 2** is a flow chart of a speech enhancement method, according to some embodiments of the present disclosure.

[0015] **FIG. 3** is a schematic diagram of a speech enhancement model, according to some embodiments of the present disclosure.

[0016] **FIG. 4** is a schematic diagram of another speech enhancement model, according to some embodiments of the present disclosure.

[0017] **FIG. 5** is a schematic diagram of a first convolutional recurrent block, according to some embodiments of the present disclosure.

[0018] **FIG. 6** is a flow chart of a model training method, according to some embodiments of the present disclosure.

[0019] **FIG. 7** is a schematic diagram of an auxiliary network, according to some embodiments of the present disclosure.

[0020] **FIG. 8** is a flow chart of a speech recognition method, according to some embodiments of the present disclosure.

[0021] **FIG. 9** is a schematic diagram of a speech enhancement apparatus, according to some embodiments of the present disclosure.

[0022] **FIG. 10** is a schematic diagram of an electronic device, according to some embodiments of the present disclosure.

DETAILED DESCRIPTION

[0023] Reference will now be made in detail to exemplary embodiments, examples of which are illustrated in the accompanying drawings. The following description refers to the accompanying drawings in which the same numbers in different drawings represent the same or similar elements unless otherwise represented. The implementations set forth in the following description of exemplary embodiments do not represent all implementations consistent with the invention. Instead, they are merely examples of apparatuses and methods consistent with aspects related to the invention as recited in the appended claims. Particular aspects of the present disclosure are described in greater detail below. The terms

and definitions provided herein control, if in conflict with terms and/or definitions incorporated by reference.

[0024] In the embodiments of the present disclosure, in order to improve the effect of speech enhancement, noisy speech data is converted into time-frequency domain feature data, and a mask value of the noisy speech data is generated according to a long-range correlation of the time-frequency domain feature data in a frequency direction. Then, enhanced speech data for the noisy speech data is generated according to the mask value and the time-frequency domain feature data. The long-range correlation of the time-frequency domain feature data in the frequency direction can characterize the correlation between noise signals with large frequency differences and useful speech signals in the speech data. The noise with large frequency differences can be effectively filtered out from the useful speech signals based on the long-range correlation in the frequency direction. Thus, a mask value is generated based on the long-range correlation of the time-frequency domain feature data in the frequency direction, and then the enhanced speech data is generated based on the mask value, so that the noise in the noisy speech data can be filtered out more effectively, thus improving the effect of speech enhancement.

[0025] In a specific implementation, the speech enhancement method provided in some embodiments of the present disclosure can be used in a variety of application scenarios. For example, speech enhancement services may be provided in a cloud service system, which can be implemented by the solution provided in the embodiments of the present disclosure. Specifically, the above-mentioned cloud service system provides a speech enhancement model and provides a cloud speech enhancement interface to users. Multiple users can invoke the interface in their respective application systems. After receiving the invocation, the cloud service system runs relevant processing programs to achieve speech enhancement through the speech enhancement model, and returns the

enhanced speech data. Alternatively, the speech enhancement method provided in the embodiments of the present disclosure may also be used in a localized device. For example, the speech enhancement method provided by the embodiments of the present disclosure may be implemented in a local device such as an audio/video conference terminal, an online education terminal, and the like.

[0026] **FIG. 1** illustrates an exemplary system applicable to the speech enhancement method, according to some embodiments of the present disclosure. As shown in **FIG. 1**, the system 100 includes a server 102, a communication network 104, or one or more user devices 106 (illustrated as a plurality of user devices in **FIG. 1**).

[0027] The server 102 may be any suitable server for storing information, data, programs, or any other suitable type of content. In some embodiments, the server 102 may perform any suitable function. For example, in some embodiments, the server 102 may be used for speech enhancement. In some embodiments, the server 102 may be used to perform speech enhancement via a speech enhancement model. In some embodiments, the server 102 may be used to send the result of speech enhancement to the user device.

[0028] In some embodiments, the communication network 104 may be any suitable combination of one or more wired or wireless networks. For example, the communication network 104 can include any one or more of: the Internet, an Intranet, a wide area network (WAN), a local area network (LAN), a wireless network, a digital subscriber line (DSL) network, a frame relay network, an asynchronous transfer mode (ATM) network, a virtual private network (VPN), or any other suitable communication network. The user device 106 can be connected via one or more communication links (e.g., communication link 112) to the communication network 104, and the communication network 104 can be linked to the server 102 via one or more communication links (e.g., communication link 114). The communication link may be any communication link suitable for transmitting data between

the user device 106 and the server 102, such as a network link, a dial-up link, a wireless link, a hardwired link, any other suitable communication link, or any suitable combination of such links.

[0029] The user device 106 may include any one or more user devices suitable for receiving speech data and collecting speech data. In some embodiments, the user device 106 may include any suitable type of device. For example, in some embodiments, the user device 106 may include a mobile device, a tablet computer, a laptop computer, a desktop computer, a wearable computer, a gaming console, a media player, a vehicle entertainment system, or any other suitable type of user device.

[0030] Although the server 102 is illustrated as a single device, in some embodiments, any suitable number of devices may be used to perform the functions performed by the server 102. For example, in some embodiments, multiple devices may be used to implement the functions performed by the server 102. Alternatively, a cloud service may be used to implement the functions performed by the server 102.

[0031] Based on the system described above, some embodiments of the present disclosure provide a speech enhancement method, which is illustrated by a plurality of embodiments below.

[0032] **FIG. 2** is a schematic flow chart of a speech enhancement method provided, according to some embodiments of the present disclosure. As shown in **FIG. 2**, the method 200 includes steps 201 to 203.

[0033] At step 201, noisy speech data is converted into time-frequency domain feature data.

[0034] The noisy speech data may be expressed as $x = y + z$, where x is used to denote the noisy speech data, y is used to denote the clean speech signal, and z is used to denote the noise signal. The noisy speech data describes the time-domain features of the

speech signal, that is, characterizes the change of the amplitude of the speech signal over time. In order to perform speech enhancement on the noisy speech data, the time-frequency domain features of the noisy speech signal are needed. The time-frequency domain features can characterize the time-domain features and the frequency-domain features of the noisy speech signal. Thus, the noisy speech data need to be converted into time-frequency domain feature data, and the time-frequency domain feature data indicate the intensity of each frequency signal in the speech signal at different times.

[0035] In some embodiments, the noisy speech data is converted into the time-frequency domain feature data by means of short-time Fourier transform (STFT).

[0036] At step 202, a mask value of the noisy speech data is generated according to a long-range correlation of the time-frequency domain feature data in a frequency direction.

[0037] The long-range correlation of the noisy speech data along the frequency axis direction can be captured from the time-frequency domain feature data. The mask value of the noisy speech data is generated based on the captured long-range correlation. The mask value indicates the useful speech signal and the noise speech signal in the time-frequency domain feature data.

[0038] The time-frequency domain feature data contains the long-range correlation of the noisy speech data along the frequency axis direction. The time-frequency domain feature data can be input into the speech enhancement model. The long-range correlation of the noisy speech data along the frequency axis direction can be captured by the speech enhancement model. The useful speech signal and the noise speech signal in the time-frequency domain feature data can be determined based on the captured long-range correlation, so as to generate the mask value for indicating the useful speech signal and the noise speech signal in the time-frequency domain feature data.

[0039] At step 203, enhanced speech data with the noisy speech data is generated according to the mask value and the time-frequency domain feature data.

[0040] Since the mask value of the time-frequency domain feature data may indicate the useful speech signal and the noise speech signal in the time-frequency domain feature data, the noise speech signal in the time-frequency domain feature data can be filtered out according to the mask value, so as to obtain the enhanced speech data that mainly includes the useful speech signal in the time-frequency domain feature data.

[0041] In order to facilitate subsequent processing such as speech recognition and speech play, the enhanced speech data in the time-frequency domain needs to be converted to the time domain. In some embodiments, the enhanced speech data can be converted to the time domain by means of inverse short-time Fourier transform (ISTFT).

[0042] In the embodiments of the present disclosure, after the noisy speech data is converted into time-frequency domain feature data, since the long-range correlation of the time-frequency domain feature data in the frequency direction can characterize the correlation between noise signals with large frequency differences and useful speech signals in the noisy speech data, the noise signals with large frequency differences and the useful speech signals in the time-frequency domain feature data can be determined based on the long-range correlation of the time-frequency domain feature data in the frequency direction, so as to generate a mask value of the noisy speech data and generate enhanced speech data for the noisy speech data according to the mask value and the time-frequency domain feature data, where the noise signals with large frequency differences are filtered out from the generated enhanced speech data, thereby improving the effect of speech enhancement.

[0043] In some embodiments, the time-frequency domain feature data can be inputted into the speech enhancement model. Then, based on the long-range correlation of

the time-frequency domain feature data in the frequency direction, the speech enhancement model can be used to identify the noise signals and the useful speech signals in the time-frequency domain feature data and obtain the correlation feature data that can indicate the noise signals and the useful speech signals in the time-frequency domain feature data. The correlation feature data can then be processed by activation functions to obtain the mask value.

[0044] When converting the noisy speech data into the time-frequency domain feature data, the noisy speech data may be converted into the time-frequency domain feature data in complex form. The corresponding speech enhancement model is also a neural network model based on complex values. For example, the time-frequency domain feature data in complex form may be denoted as $V = V_r + jV_i \in \mathbb{C}^{C \times T \times F}$, where V is used to characterize the time-frequency domain feature data, V_r is used to characterize the real part of the time-frequency domain feature data, V_i is used to characterize the imaginary part of the time-frequency domain feature data, j is an imaginary unit, $j^2 = -1$, $\mathbb{C}$ is used to characterize the complex domain, C is used to characterize the channel dimension of the time-frequency domain feature data, T is used to characterize the frame dimension of the time-frequency domain feature data, and F is used to characterize the frequency dimension of the time-frequency domain feature data.

[0045] Since the time-frequency domain feature data is in complex form, and the speech enhancement model used to process the time-frequency domain feature data is also a neural network model based on complex values, the correlation feature data that is output from the speech enhancement model is also in complex form. When the correlation feature data is processed by the activation functions, the real part of the correlation feature data can be processed by a real-part activation function, and the imaginary part of the correlation feature data can be processed by an imaginary-part activation function. Then,

the outputs of the real-part activation function and the imaginary-part activation function form the mask value.

[0046] In some embodiments, the real-part activation function and the imaginary-part activation function are both hyperbolic tangent functions ($\tanh$). The real-part activation function and the imaginary-part activation function may also be other types of functions, and the embodiments of the present disclosure do not limit the types of the real-part activation function and the imaginary-part activation function.

[0047] In the embodiments of the present disclosure, the noisy speech data is converted into the time-frequency domain feature data in complex form. The correlation feature data in complex form is obtained by processing the time-frequency domain feature data using the speech enhancement model based on the complex-valued neural network. Then the real part and the imaginary part of the correlation feature data are processed by the activation functions separately to obtain the mask value. Speech enhancement is performed on the noisy speech data by the speech enhancement model of the complex-valued neural network. By means of special operations of complex numbers, the time-frequency domain feature data can be processed from multiple dimensions such as the channels, frames and frequency of the time-frequency domain feature data, so as to extract the high-level feature representation of the time-frequency domain feature data. Then, speech enhancement is performed on the noisy speech data based on this high-level feature representation, thereby ensuring the effect of speech enhancement.

[0048] In some embodiments, the speech enhancement model for speech enhancement includes an encoder, an attention module, a recurrent module, and a decoder. After the time-frequency domain feature data is input into the speech enhancement model, the encoder is first used to process the time-frequency domain feature data. Then, the attention module and the recurrent module are used to process the output data of the

encoder separately. Then, the decoder is used to process the output data of the attention module and the output data of the recurrent module to obtain correlation feature data for indicating the noise signals and the useful speech signals in the time-frequency domain feature data.

[0049] In the embodiments of the present disclosure, the encoder-decoder is implemented by a convolutional neural network (CNN) and the recurrent module is implemented by a recurrent neural network (RNN). High-level features of the noisy speech data can be extracted by the encoder-decoder and the long-term time dependence of the noisy speech data can be modeled by the recurrent module, so that with the speech enhancement model having a convolutional recurrent structure, the noise signals in the time-frequency domain feature data can be effectively filtered out, thus ensuring the effect of speech enhancement on the noisy speech data.

[0050] After the encoder processes the time-frequency domain feature data, the attention module processes the output data of the encoder by filtering the interference information in the output data of the encoder, and then the filtered data is inputted into the decoder for processing. The attention module processes the output data of the encoder based on the attention mechanism, therefore, the interference information in the output data of the encoder can be filtered. Then, the decoder processes the filtered data, which can more accurately identify the noise signals and the useful speech signals in the time-frequency domain feature data, thereby further improving the effect of speech enhancement.

[0051] In some embodiments, the encoder includes M first convolutional recurrent blocks, M being a natural number greater than or equal to 2, where the first convolutional recurrent block includes a first convolution unit for first convolution processing and a first recurrent unit for first recurrent processing. For the i th first convolutional recurrent block

in the encoder, first convolution processing is performed on input data using the first convolution unit in that first convolutional recurrent block, and first recurrent processing is performed on a result of the first convolution processing using the first recurrent unit in that first convolutional recurrent block, so as to obtain output data of that first convolutional recurrent block, wherein $1 \leq i \leq M$; and output data of each first convolutional recurrent block is input to the attention module and output data of the Mth first convolutional recurrent block is input to the recurrent module.

[0052] In the embodiments of the present disclosure, the encoder includes at least two first convolutional recurrent blocks. Each first convolutional recurrent block includes a first convolution unit and a first recurrent unit. The first convolution unit is used to perform first convolution processing on the data that is input to the first convolutional recurrent block, and the first recurrent unit is used to perform first recurrent processing on the result of the first convolution processing. After each first convolutional recurrent block performs convolutional recurrent processing on the input data, output data of each first convolutional recurrent block is input to the attention module and output data of the Mth first convolutional recurrent block is input to the recurrent module. By performing convolutional recurrent processing on the time-frequency domain feature data using the plurality of first convolutional recurrent blocks, high-level features of the time-frequency domain feature data can be extracted more effectively and comprehensively. Then, speech enhancement can be performed on the noisy speech data based on the extracted high-level features, thereby ensuring the effect of speech enhancement. The output data of each first convolutional recurrent block is input to the attention module, and the attention module is used to filter the interference information to ensure that the input to the decoder includes less interference information, so as to ensure that the decoder can more accurately identify

the noise signals and the useful speech information, and then the correlation feature data that can accurately denote the noise signals and the useful speech signals is generated.

[0053] For each first convolutional recurrent block, after the first convolution unit performs convolution processing on the input data, the first recurrent unit performs recurrent processing on the output data of the first convolution unit. The first recurrent unit can realize the recurrence in the frequency direction to achieve frequency recursion. The long-range correlation of the time-frequency domain feature data along the frequency direction can be extracted. When the decoder subsequently processes the output data of the encoder, the decoder can identify the noise signals and the useful speech signals based on the long-range correlation of the time-frequency domain feature data along the frequency direction, and the correlation feature data that can accurately indicate the noise signals and the useful speech signals is generated, thereby ensuring the effect of speech enhancement on the noisy speech data.

[0054] In some embodiments, when the value of i is 1, the input data of the i th first convolutional recurrent block is the time-frequency domain feature data, and when the value of i is 2 to M , the input data of the i th first convolutional recurrent block is output data of the $(i-1)$ th first convolutional recurrent block.

[0055] In the embodiments of the present disclosure, the first first convolutional recurrent block takes the time-frequency domain feature data as input data, and all the first convolutional recurrent blocks other than the first first convolutional recurrent block take the output data of the previous first convolutional recurrent block as input data. Starting from the second first convolutional recurrent block, each first convolutional recurrent block performs convolutional recurrent processing on the output data of the previous first convolutional recurrent block and inputs the result of the convolutional recurrent processing to the next first convolutional recurrent block, and the last (M th) first

convolutional recurrent block inputs the result of the convolutional recurrent processing to the recurrent module. The encoder includes a plurality of first convolutional recurrent blocks, and the output data of the previous first convolutional recurrent block is used as the input data by the following first convolutional recurrent block. Through the convolutional recurrent processing of each first convolutional recurrent block, the high-level features in the time-frequency domain feature data can be extracted accurately and comprehensively, and thus speech enhancement can be performed more accurately on the noisy speech data.

[0056] In some embodiments, when processing the output data of the encoder using the attention module, the output data of each first convolutional recurrent block can be processed separately using the attention module, and the output data of the attention module can be input to the decoder.

[0057] The encoder includes a plurality of first convolutional recurrent blocks. Each first convolutional recurrent block, after performing convolutional recurrent processing on the input data, inputs the output data to the attention module. The attention module processes the output data of each first convolutional recurrent block separately to filter the interference information in the output data of each first convolutional recurrent block separately. The attention module, after processing the output data of each first convolutional recurrent block, sends the result of the processing to the decoder. The decoder generates correlation feature data according to the output results of the attention module. Since the data received by the decoder includes less interference information, the noise signals and the useful speech signals in the time-frequency domain feature data can be identified more accurately according to the received data, so as to generate correlation feature data that can accurately indicate the noise signals and the useful speech signals in the time-frequency domain feature data, thus improving the effect of speech enhancement on the noisy speech data.

[0058] In some embodiments, the recurrent module includes one or more recurrent units, and the recurrent units are implemented by a recurrent neural network. When the recurrent module includes a plurality of recurrent units, the first recurrent unit takes the output data of the Mth first convolutional recurrent block as input data, and all the recurrent units other than the first recurrent unit take the output data of the previous recurrent unit as input data. By performing recurrent processing on the output data of the Mth first convolutional recurrent block by the plurality of recurrent units, the long-term time dependence of the time-frequency domain feature data can be effectively extracted. The decoder can generate correlation feature data according to the long-term time dependence of the time-frequency domain feature data, which ensures that the generated correlation feature data can accurately indicate the noise signals and the useful speech signals in the time-frequency domain feature data, thus improving the effect of speech enhancement on the noisy speech data.

[0059] In some embodiments, the recurrent units included in the recurrent module can be implemented by feedforward sequential memory networks (FSMN). When the time-frequency domain feature data is in complex form and the speech enhancement model is a speech enhancement model based on a complex-valued neural network, the recurrent unit is an FSMN that can process complex values.

[0060] In some embodiments, corresponding to that the encoder includes M first convolutional recurrent blocks, the decoder includes M second convolutional recurrent blocks, wherein each second convolutional recurrent block includes a second convolution unit for second convolution processing and a second recurrent unit for second recurrent processing. When the output data of the attention module and the output data of the recurrent module are processed using the decoder to obtain the correlation feature data, for the ith second convolutional recurrent block among the M second convolutional recurrent

blocks, second convolution processing is performed on first input data and second input data using the second convolution unit in that second convolutional recurrent block, and second recurrent processing is performed on a result of the second convolution processing using the second recurrent unit in that second convolutional recurrent block, so as to obtain output data of that second convolutional recurrent block. The output data of the M th second convolutional recurrent block is the correlation feature data. If the value of i is 1, the first input data of the second convolution unit in the i th second convolutional recurrent block is output data of the attention module after processing the output data of the M th first convolutional recurrent block, and the second input data of the second convolution unit in the i th second convolutional recurrent block is the output data of the recurrent module. If the value of i is 2 to M , the first input data of the second convolution unit in the i th second convolutional recurrent block is output data of the attention module after processing output data of the $(M+1-i)$ th first convolutional recurrent block, and the second input data of the second convolution unit in the i th second convolutional recurrent block is output data of the $(i-1)$ th second convolutional recurrent block.

[0061] In the embodiments of the present disclosure, the encoder and the decoder are in a symmetrical structure. The encoder includes M first convolutional recurrent blocks, and the decoder includes M second convolutional recurrent blocks. The output data of the i th first convolutional recurrent block is processed by the attention module and then input to the $(M+1-i)$ th second convolutional recurrent block as the first input data, and the output data of the recurrent module is input to the 1st second convolutional recurrent block as the second input data of the 1st second convolutional recurrent block, and starting from the second second convolutional recurrent block, each second convolutional recurrent block takes the output data of the previous second convolutional recurrent block as the second input data. The encoder and the decoder are in a symmetrical structure to ensure

that the time-frequency domain feature data that is input to the speech enhancement model has the same data structure as the correlation feature data that is output from the speech enhancement model, so as to ensure that enhanced speech data for the noisy speech data can be obtained based on the correlation feature data that is output from the decoder, thereby ensuring the effect of speech noise reduction.

[0062] In some embodiments, the first recurrent units in the first convolutional recurrent blocks and the second recurrent units in the second convolutional recurrent blocks can all be implemented by feedforward sequential memory networks. Since the feedforward sequential memory network can effectively perform frequency recursion to capture the long-range correlation of the time-frequency domain feature data in the frequency direction, the encoder and the decoder that are composed of the convolution units and the feedforward sequential memory networks can capture the local time-frequency structure of the noisy speech data and can also capture the long-range correlation of the frequency spectrum, thus enabling better identification of noise signals and useful speech signals, thus improving the effect of speech enhancement.

[0063] A speech enhancement model that can implement the speech enhancement method in the embodiments of the present disclosure will be described in detail below.

FIG. 3 is a schematic diagram of a speech enhancement model, according to some embodiments of the present disclosure. As shown in **FIG. 3**, the speech enhancement model 31 includes an encoder 311, a decoder 312, an attention module 313, and a recurrent module 314. The encoder 311 is configured to process time-frequency domain feature data corresponding to noisy speech data. The attention module 313 is configured to process output data of the encoder. The recurrent module 314 is configured to process the output data of the encoder. The decoder 312 is configured to process output data of the attention module 313 and the recurrent module 314 to obtain correlation feature data for

characterizing a long-range correlation of the time-frequency domain feature data in a frequency direction.

[0064] It should be noted that the speech enhancement model 31 can be used in the speech enhancement method in the above embodiments for generating the correlation feature data based on the time-frequency domain feature data. For the processes of data processing by the modules in the speech enhancement model 31, reference can be made to the description in the above method embodiments, which will not be repeated here.

[0065] **FIG. 4** is a schematic diagram of another speech enhancement model, according to some embodiments of the present disclosure. As shown in **FIG. 4**, the encoder 311 includes M first convolutional recurrent blocks, where the M first convolutional recurrent blocks are the first convolutional recurrent block 3111 to the first convolutional recurrent block 311M in sequence, and the *i*th first convolutional recurrent block is the first convolutional recurrent block 311*i*. The decoder 312 includes M second convolutional recurrent blocks, where the M second convolutional recurrent blocks are the second convolutional recurrent block 3121 to the second convolutional recurrent block 312M in sequence, and the *i*th second convolutional recurrent block is the second convolutional recurrent block 312*i*, where M is a natural number greater than or equal to 2. The recurrent module 314 includes N complex feedforward sequential memory networks (CFSMNs) connected in series, where N is a natural number, and when N is a natural number greater than or equal to 2, the N CFSMNs are CFSMN 3141 to CFSMN 314N in sequence, the *i*th CFSMN is CFSMN 314*i*, the 1st CFSMN 3141 is connected to the Mth first convolutional recurrent block, and the Nth CFSMN 314N is connected to the 1st second convolutional recurrent block 3121.

[0066] The noisy speech data *x* is converted into the time-frequency domain feature data *X* in a complex form via the short-time Fourier transform (STFT) 32, the time-

frequency domain feature data X including the real part X_r and the imaginary part X_i . After the time-frequency domain feature data X is input to the speech enhancement model 31, the speech enhancement model 31 processes the time-frequency domain feature data X and outputs the correlation feature data. The real part of the correlation feature data is input to the real-part activation function (tanh) 331 to obtain the real part of the mask value $\hat{M}$, and the imaginary part of the correlation feature data is input to the imaginary-part activation function (tanh) 332 to obtain the imaginary part of the mask value $\hat{M}$. The real part and the imaginary part of the mask value $\hat{M}$ form the mask value $\hat{M}$ in complex form. The time-frequency domain feature data X and the mask value $\hat{M}$ are subjected to element-by-element complex multiplication to obtain the enhanced speech data $\hat{Y}$ in the time-frequency domain. The enhanced speech data $\hat{Y}$ in the time-frequency domain is converted to enhanced speech data $\hat{y}$ in the time domain by the inverse short-time Fourier transform (ISTFT) 34.

[0067] **FIG. 5** is a schematic diagram of a first convolutional recurrent block, according to some embodiments of the present disclosure. As shown in **FIG. 5**, the first convolutional recurrent block 311i includes a convolution unit 501, a normalization unit 502, an activation function 503, and a recurrent unit 504. The convolution unit 501 is configured to process a complex two-dimensional convolutional neural network. The activation function 503 is a Leaky ReLU (Rectified Linear Unit) function. The recurrent unit 504 is implemented by a CFMNM.

[0068] The input data of the convolution unit 501 is a three-dimensional feature matrix $V = V_r + jV_i \in \mathbb{C}^{C \times T \times F}$, where V_r is used to characterize the real part, V_i is used to characterize the imaginary part, j is an imaginary unit, $j^2 = -1$, $\mathbb{C}$ is used to characterize the complex domain, C is used to characterize the channel dimension, T is used to characterize the frame dimension, and F is used to characterize the frequency

dimension. The convolution kernel of the convolution unit 501 is

$W = W_r + jW_i \in \mathbb{C}^{C' \times T' \times F'}$, where W_r is used to characterize the real part, W_i is used to characterize the imaginary part, C' is used to characterize the number of convolution kernels, and $T' \times F'$ is used to characterize the size of the convolution kernel. The output data of the convolution unit 501 may be expressed as $U = U_r + jU_i \in \mathbb{C}^{C' \times T \times F''}$, and the real part U_r and the imaginary part U_i of the output of the convolution unit 501 may be expressed as the following **Equation (1)**:

$$\begin{aligned} U_r &= V_r * W_r - V_i * W_i \\ U_i &= V_r * W_i + V_i * W_r \end{aligned} \quad (1)$$

where $*$ is used to characterize the convolutional filtering of the real numbers. The convolution unit 501 performs causal convolution, where in the time direction, $T' = 2$, the stride is equal to 1, and zero padding is used; and in the frequency direction, $F' = 5$, the stride is equal to 2, and zero padding is not required, which makes the feature map corresponding to the time-frequency domain feature data in the encoder 311 be halved block by block.

[0069] The above processing process is applicable to the convolution unit 501 in each first convolutional recurrent block, and the same C' is used in the convolution unit 501 included in each first convolutional recurrent block, so as to ensure that each first convolutional recurrent block has the feature map with the same data, for example $C' = 128$.

[0070] The output data of the convolution unit 501 is processed by the normalization unit 502 and the activation function 503, and then input to the recurrent unit 504. In the recurrent unit 504, the same processing is performed on the real part and the imaginary part of the data $U = U_r + jU_i \in \mathbb{C}^{C' \times T \times F''}$. Only the processing of the real part of the data $U = U_r + jU_i \in \mathbb{C}^{C' \times T \times F''}$ will be explained below. For the process of

processing the imaginary part, reference can be made to the process of processing the real part.

[0071] The dimension of the real part U_r of the data $U = U_r + jU_i \in \mathbb{C}^{C' \times T \times F''}$ is converted from $C \times T \times F''$ to $T \times F'' \times C$. For the current frame t in the data $U_r \in \mathbb{R}^{T \times F'' \times C}$, a frequency series $\mathbf{S}_r(t) \in \mathbb{R}^{F'' \times C'} = \{\mathbf{s}_{f_1}(t), \mathbf{s}_{f_2}(t), \dots, \mathbf{s}_{F''}(t) \in \mathbb{R}^{C' \times 1}\}$ can be formed, and the above frequency series is input into the real-part recurrent unit in the recurrent unit 504 for processing. For the i th component in the above frequency series, the real-part recurrent unit performs processing by the following **Equations (2) to (4)**:

$$\mathbf{h}_{f_i}^l = \delta(\mathbf{W}_{f_i}^l \mathbf{s}_{f_i}^{l-1} + \mathbf{b}_{f_i}^l) \quad (2)$$

$$\mathbf{p}_{f_i}^l = \mathbf{V}_{f_i}^l \mathbf{h}_{f_i}^l + \mathbf{v}_{f_i}^l \quad (3)$$

$$\mathbf{s}_{f_i}^l = \mathbf{s}_{f_i}^{l-1} + \mathbf{p}_{f_i}^l + \sum_{\tau=0}^{N_L} \mathbf{a}_{\tau}^l \odot \mathbf{p}_{f_i-\tau}^l + \sum_{\kappa=0}^{N_R} \mathbf{c}_{\kappa}^l \odot \mathbf{p}_{f_i+\kappa}^l \quad (4)$$

where $f_i = f_1, f_2, \dots, F''$, and t is omitted in the above **Equations (2) to (4)**, δ characterizes the activation function ReLU, N_L and N_R respectively denote the look-back and lookahead sequences of the i th component, such as $N_L=20$ and $N_L=0$, and other parameters are all the model parameters of the recurrent unit 504. Since the recurrent unit 504 in the embodiments of the present disclosure includes only one CFMNN, the value of l is 1, $\mathbf{s}_{f_i}^0$ is equivalent to $\mathbf{s}_{f_i}$, and $\mathbf{s}_{f_i}^1$ is the output of the current real-part recurrent unit and is the input to the next real-part recurrent unit. For the first real-part recurrent unit, $\mathbf{S}_r(t)$ is the output of the activation function 503. In other embodiments, the value of l may be other positive integers, such as 2, 3, 4, etc.

[0072] The output $\mathbf{S}_{\text{out}} \in \mathbb{C}^{F'' \times C'}$ of the recurrent unit 504 may be expressed as the following **Equation (5)**:

$$\mathbf{S}_{\text{out}} = \text{FSMN}_r(\mathbf{S}_r) - \text{FSMN}_i(\mathbf{S}_i) + j(\text{FSMN}_r(\mathbf{S}_i) + \text{FSMN}_i(\mathbf{S}_r)) \quad (5)$$

where FSMN_r and FSMN_i are used to characterize the real-part recurrent unit and the imaginary-part recurrent unit of the recurrent unit 504, respectively, and S_r and S_i are used to characterize the real part and the imaginary part of the frequency series, respectively.

[0073] After the output $\mathbf{S}_{\text{out}} \in \mathbb{C}^{F'' \times C'}$ of the recurrent unit 504 in the first convolutional recurrent block 311M is input to the recurrent module 314, the real part of the data $\mathbf{S}_{\text{out}} \in \mathbb{C}^{F'' \times C'}$ is shaped as $U_r \in \mathbb{R}^{T \times H}$, where $H = F'' \times C'$, so as to process $U_r \in \mathbb{R}^{T \times H}$ in two dimensions by the CFSMN. The time series $\mathbf{Q}_r = \{\mathbf{q}_{t_1}, \mathbf{q}_{t_2}, \dots, \mathbf{q}_T \in \mathbb{R}^{H \times 1}\}$ is formed based on $U_r \in \mathbb{R}^{T \times H}$, and then, each CFSMN in the recurrent module 314 applies the way of processing of the above **Equations (2) to (4)** to $\mathbf{Q}_r$ so as to fit the time dynamics. Reference can be made to the above **Equation (5)** for the output of the recurrent module 314.

[0074] It should be noted that the processes of processing the real part and the imaginary part in the input data by the first convolutional recurrent block are similar, and the process of processing the input data by the recurrent module 314 is similar to the process of processing the input data by the recurrent unit in the first convolutional recurrent block. Therefore, the embodiments of the present disclosure only describe in detail the process of processing the real part in the input data by the first convolutional recurrent block, and the process of processing the imaginary part in the input data by the first convolutional recurrent block and the process of processing the input data by the recurrent module 314 will not be repeated here, and for the specific processes, reference can be made to the description of the processing of the real part in the input data by the first convolutional recurrent block described above.

[0075] It should also be noted that the structure of the decoder 312 is symmetrical with the structure of the encoder 311, so the process of processing the input data by the second convolutional recurrent block 312i in the decoder 312 is similar to that of

processing the input data by the first convolutional recurrent block 311i. Therefore, the data processing process by the second convolutional recurrent block 312i will not be described in detail, and for the detailed process, reference can be made to the description of the first convolutional recurrent block 311i above.

[0076] For the speech enhancement model shown in **FIG. 3** above, the embodiments of the present disclosure further provide a method for training the speech enhancement model. **FIG. 6** is a schematic flow chart of a model training method 600 for training the speech enhancement model of the above embodiments, according to some embodiments of the present disclosure. As shown in **FIG. 6**, the model training method 600 includes the steps 601 to 605.

[0077] At step 601, time-frequency domain feature data of a training sample is inputted into a model to be trained.

[0078] The training sample may be formed of clean speech data and noise speech data. The training sample in the time domain is converted to the time-frequency domain to obtain the time-frequency domain feature data of the training sample. For example, by converting the training sample to the time-frequency domain through the short-time Fourier transform (STFT), the time-frequency domain feature data of the training sample is obtained.

[0079] At step 602, a first loss value of the model to be trained is determined.

[0080] After the time-frequency domain feature data of the training sample is input into the model to be trained, the output result of the model to be trained is obtained, and the first loss value is obtained according to the output result of the model to be trained. The first loss value of the model to be trained is determined based on at least one of a signal-to-noise ratio loss (SI-SNR loss), a mask loss, and a real-imaginary part spectral loss.

[0081] In some embodiments, the first loss value of the model to be trained can be expressed as **Equation (6)**:

$$\mathcal{L}_{Step1} = \mathcal{L}(y, \hat{y}) + \lambda_1 \mathcal{L}(M, \hat{M}) + \lambda_2 \mathcal{L}(Y, \hat{Y}), \quad (6)$$

where $\mathcal{L}_{Step1}$ is used to characterize the first loss value, $\mathcal{L}(y, \hat{y})$ is used to characterize the signal-to-noise ratio loss, $\mathcal{L}(M, \hat{M})$ is used to characterize the mask loss, and $\mathcal{L}(Y, \hat{Y})$ is used to characterize the real-imaginary part spectral loss; y is used to characterize the time-domain feature data for the clean speech in the training sample, $\hat{y}$ is used to characterize the enhanced speech data in the time domain determined according to the output of the speech enhancement model, M is used to characterize the real mask value determined according to the clean speech data in the training sample, $\hat{M}$ is used to characterize the mask value determined according to the output of the speech enhancement model, Y is used to characterize the time-frequency domain feature data for the clean speech in the training sample, and $\hat{Y}$ is used to characterize the enhanced speech data in the time-frequency domain determined according to the correlation feature data that is output from the speech enhancement model; and λ_1 and λ_2 are used to characterize the preset coefficients.

[0082] The signal-to-noise ratio loss $\mathcal{L}(y, \hat{y})$ may be calculated by the equation $\mathcal{L}(y, \hat{y}) = 10 \log_{10} \left(\frac{\|y_{\text{target}}\|_2^2}{\|e_{\text{noise}}\|_2^2} \right)$, where $\|\cdot\|_2$ denotes the L2 norm, $y_{\text{target}} = (\langle \hat{y}, y \rangle \cdot y) / \|y\|_2^2$ and $e_{\text{noise}} = \hat{y} - y_{\text{target}}$, and $\langle \cdot, \cdot \rangle$ denotes the dot product. The mask loss $\mathcal{L}(M, \hat{M})$ may be calculated by the equation $\mathcal{L}(M, \hat{M}) = \sum_{t,f} [(\hat{M}_r - M_r)^2 + (\hat{M}_i - M_i)^2]$. The real-imaginary part spectral loss $\mathcal{L}(Y, \hat{Y})$ may be calculated by the equation $\mathcal{L}(Y, \hat{Y}) = \sum_{t,f} [(\hat{Y}_r - Y_r)^2 + (\hat{Y}_i - Y_i)^2]$.

[0083] At step 603, a model parameter of the model to be trained is adjusted according to the first loss value until the first loss value satisfies a preset first condition.

[0084] After each time the time-frequency domain feature data of the training sample is input to the model to be trained, the first loss value of the model to be trained is acquired, and it is judged whether the first loss value satisfies the preset first condition. If the first loss value satisfies the preset first condition, step 604 is executed; and if the first loss value does not satisfy the preset first condition, the model parameter of the model to be trained is adjusted according to the first loss value, and then step 601 is continued to be executed, until the first loss value satisfies the preset first condition.

[0085] The first condition may be that the first loss value is less than a preset first threshold, or the first loss value is no longer decreasing, etc.

[0086] At step 604, a second loss value of the model to be trained is determined.

[0087] After the first loss value satisfies the preset first condition, the second loss value of the model to be trained is determined based on the output result of the model to be trained. The second loss value of the model to be trained is determined according to at least one of the first loss value, an adversarial loss, and a depth feature loss.

[0088] In some embodiments, the second loss value of the model to be trained can be expressed as **Equation (7)**:

$$\mathcal{L}_{Step2} = \mathcal{L}_{Step1} + \lambda_3 \mathcal{L}_{AL} + \lambda_4 \mathcal{L}_{DFL}. \quad (7)$$

where $\mathcal{L}_{Step2}$ is used to characterize the second loss value, $\mathcal{L}_{AL}$ is used to characterize the adversarial loss, $\mathcal{L}_{DFL}$ is used to characterize the depth feature loss, and λ_3 and λ_4 are used to characterize the preset coefficients.

[0089] To determine the second loss value, an auxiliary network with the same model structure as the model to be trained is constructed, where this auxiliary network has the structure as shown in **FIG. 7**, and the auxiliary network 700 includes an encoder 701, a recurrent module 702, a decoder 703, and a linear projection module 704. The adversarial loss and the depth feature loss may be determined.

[0090] The encoder 701 and the linear projection module 704 included in the auxiliary network constitute a discriminator D. The enhanced speech data $\hat{Y}$ in the time-frequency domain determined according to the correlation feature data that is output by the speech enhancement model is input into the discriminator D, and the adversarial loss $\mathcal{L}_{AL}$ can be determined according to the output result of the linear projection module 704.

Specifically, the adversarial loss $\mathcal{L}_{AL}$ can be calculated by the Equation

$\mathcal{L}_{AL} = 1/2E[(1 - D(\hat{Y}))^2]$, where $E()$ is used to characterize the operation of obtaining an expected value and $D(\hat{Y})$ is used to characterize the output result of the linear projection module 704. The time-frequency domain feature data Y used to characterize the clean speech in the training sample as well as $\hat{Y}$ are input to the discriminator D, and the depth feature loss $\mathcal{L}_{DFL}$ is determined according to the output result of the encoder 701.

Specifically, the depth feature loss $\mathcal{L}_{DFL}$ can be calculated by the Equation

$\mathcal{L}_{DFL} = \sum_{l=1}^{L_D} 1/L_D \|D_l(\hat{Y}) - D_l(Y)\|_2^2$, where L_D is used to characterize the number of the first convolutional recurrent blocks in the encoder 701, $D_l(\hat{Y})$ is used to characterize the output of the l th first convolutional recurrent block in the encoder 701 with respect to $\hat{Y}$, and $D_l(Y)$ is used to characterize the output of the l th first convolutional recurrent block in the encoder 701 with respect to Y .

[0091] To ensure that the discriminator D has sufficient spectral information, the encoder 701, the recurrent module 702, and the decoder 703 included in the auxiliary network form a reconstruction network C, where the loss value of the reconstruction network C is defined as $\mathcal{L}_C = 1/2\|Y - C(Y)\|_2^2$ and the loss value of the discriminator D is defined as the following **Equation (8)**:

$$\mathcal{L}_D = \frac{1}{2} [E[(1 - D(Y))^2] + E[D(\hat{Y})^2]] + \lambda_5 \mathcal{L}_C. \quad (8)$$

where $E()$ is used to characterize the operation of obtaining an expected value and λ_5 is used to characterize the preset coefficient.

[0092] During the training of the model to be trained, the model parameters of the reconstruction network C, the discriminator D, and the model to be trained are trained alternatively.

[0093] At step 605, the model parameter of the model to be trained is adjusted according to the second loss value until the second loss value satisfies a preset second condition to obtain the speech enhancement model.

[0094] Each time the second loss value of the model to be trained is acquired, it is judged whether the second loss value satisfies the preset second condition. If the second loss value satisfies the preset second condition, the training is completed and the speech enhancement model is obtained. If the second loss value does not satisfy the preset second condition, the model parameter of the model to be trained is adjusted according to the second loss value, and then step 601 is continued to be executed until the second loss value satisfies the preset second condition, the training of the model is completed, and the speech enhancement model is obtained.

[0095] In the embodiments of the present disclosure, the training process of the speech enhancement model is divided into two stages, where in the first stage, the speech enhancement model is trained with the first loss value as a reference, and in the second stage, the speech enhancement model is trained with the second loss value as a reference. The first loss value is determined according to at least one of the signal-to-noise ratio loss, the mask loss, and the real-imaginary part spectral loss, and the first loss value indicates the signal-level loss of the speech signal itself, and training the speech enhancement model based on the first loss value can enable the speech enhancement model to converge quickly, thus improving the effect of model training. The second loss value is determined

according to at least one of the first loss value, the adversarial loss, and the depth feature loss, and training the speech enhancement model according to the second loss value can improve the perceptiveness of the speech enhancement model, thus ensuring the effect of speech enhancement by the speech enhancement model.

[0096] For the application scenario of the speech enhancement solution provided in the embodiments of the present disclosure in speech recognition, the embodiments of the present disclosure provide a speech recognition method. As shown in **FIG. 8**, the speech recognition method 800 includes the steps 801 to 805.

[0097] At step 801, speech data to be processed is acquired, wherein the speech data to be processed includes noise, and the speech data to be processed comprises one of the following: audio/video conference speech data, online education speech data, and live webcast speech data.

[0098] At step 802, the speech data to be processed is converted into time-frequency domain feature data.

[0099] At step 803, a mask value of the noisy speech data is generated according to a long-range correlation of the time-frequency domain feature data in a frequency direction.

[00100] At step 804, enhanced speech data for the speech data to be processed is generated according to the mask value and the time-frequency domain feature data.

[00101] At step 805, speech recognition is performed on the enhanced speech data to obtain a recognition result.

[00102] It should be noted that the method shown in **FIG. 8** is a specific application of the speech enhancement solution according to some embodiments of the present disclosure, and for the specific speech enhancement process, reference can be made to the description in the preceding embodiment, and will not be repeated here.

[00103] Corresponding to the method embodiment described above, **FIG. 9** illustrates a schematic diagram of a speech enhancement apparatus 900, according to some embodiments of the present disclosure. As shown in **FIG. 9**, the speech enhancement apparatus 900 includes a conversion unit 901, a processing unit 902, and a generation unit 903.

[00104] The conversion unit 901 is configured to convert noisy speech data into time-frequency domain feature data.

[00105] The processing unit 902 is configured to generate a mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in a frequency direction.

[00106] The generation unit 903 is configured to generate enhanced speech data for the noisy speech data according to the mask value and the time-frequency domain feature data.

[00107] It should be noted that the speech enhancement apparatus of this embodiment is used to implement the corresponding speech enhancement method in the preceding method embodiments and has the beneficial effect of the corresponding method embodiments, which will not be repeated herein.

[00108] **FIG. 10** is a schematic block diagram of an electronic device, according to some embodiments of the present disclosure, and specific embodiments of the present disclosure do not limit the specific implementation of the electronic device. As shown in **FIG. 10**, the electronic device 1000 includes a processor 1002, a communication interface 1004, a memory 1006, and a communication bus 1008.

[00109] The processor 1002, the communication interface 1004, and the memory 1006 communicate with each other via the communication bus 1008.

[00110] The communication interface 1004 is configured to communicate with other electronic devices or servers.

[00111] The processor 1002 is configured to execute a program 1010, and may specifically perform the relevant steps in any of the preceding speech enhancement method embodiments (e.g., speech enhancement method 200 in **FIG. 2**).

[00112] In some embodiments, the program 1010 may include program code, which includes computer operating instructions.

[00113] The processor 1002 may be a CPU, or an application specific integrated circuit (ASIC), or one or more integrated circuits configured to implement the embodiments of the present disclosure. One or more processors included in a smart device may be the same type of processor, such as one or more CPUs; or may be different types of processors, such as one or more CPUs and one or more ASICs.

[00114] The memory 1006 is configured to store the program 1010. The memory 1006 may include a high-speed RAM memory and may also include a non-volatile memory, for example, at least one disk memory.

[00115] The program 1010 may be used specifically to cause the processor 1002 to perform the speech enhancement method in any of the preceding embodiments (e.g., speech enhancement method 200 in **FIG. 2**).

[00116] For the specific implementation of the steps in the program 1010, reference can be made to the corresponding descriptions of the corresponding steps and units in any of the preceding speech enhancement method embodiments and will not be repeated herein. It will be clear to those skilled in the field that, for the convenience and brevity of the description, for the specific working processes of the devices and modules described above, reference can be made to the corresponding process descriptions in the preceding method embodiments and will not be repeated here.

[00117] With the electronic device of the embodiments of the present disclosure, after the noisy speech data is converted into time-frequency domain feature data, since the long-range correlation of the time-frequency domain feature data in the frequency direction can characterize the correlation between noise signals with large frequency differences and useful speech signals in the noisy speech data, the noise signals with large frequency differences and the useful speech signals in the time-frequency domain feature data can be determined based on the long-range correlation of the time-frequency domain feature data in the frequency direction, so as to generate a mask value of the noisy speech data and generate enhanced speech data for the noisy speech data according to the mask value and the time-frequency domain feature data, where the noise signals with large frequency differences are filtered out from the generated enhanced speech data, thereby improving the effect of speech enhancement.

[00118] Embodiments of the present disclosure further provide a computer-readable storage medium storing instructions for causing a machine to perform instructions of the speech enhancement method as described herein. Specifically, a system or apparatus equipped with a storage medium may be provided, on which software program code for implementing the functionality of any of the above embodiments is stored, and a computer (or CPU or MPU) of the system or device is caused to read and execute the program code stored in the storage medium.

[00119] In this case, the program code read from the storage medium may itself implement the functionality of any of the above embodiments, and thus the program code and the storage medium storing the program code form part of the present disclosure.

[00120] Embodiments of the storage medium for providing the program code include floppy disks, hard disks, magnetic disks, optical disks (e.g., CD-ROM, CD-R, CD-RW, DVD-ROM, DVD-RAM, DVD-RW, DVD+RW), magnetic tapes, non-volatile

memory cards, and ROMs. Optionally, program code can be downloaded from the server computer via the communication network.

[00121] Embodiments of the present disclosure further provide a computer program product including computer instructions that instruct the computing device to perform any of the corresponding operations of the plurality of method embodiments described above.

[00122] It is noted that, depending on the needs of the implementation, the individual components/steps described in the embodiments of the present disclosure may be split into more components/steps, or two or more components/steps or part of the operations of the components/steps may be combined into a new component/step to achieve the purpose of the embodiments of the present disclosure.

[00123] The methods according to the embodiments of the present disclosure described above may be implemented in hardware or firmware, or be implemented as software or computer code that may be stored in a recording medium (such as a CD ROM, RAM, floppy disk, hard disk, or magnetic disk), or be implemented as computer code downloaded over a network that was originally stored in a remote recording medium or non-transitory machine-readable medium and will be stored in a local recording medium, such that the methods described herein may be processed by such software stored on a recording medium using a general-purpose computer, a special purpose processor, or programmable or dedicated hardware (such as an ASIC or FPGA). It can be understood that the computer, processor, microprocessor controller, or programmable hardware includes a storage component (e.g., RAM, ROM, flash memory, etc.) that can store or receive software or computer code, where the software or computer code, when accessed and executed by the computer, the processor, or the hardware, implements the methods described herein. Further, when a general-purpose computer accesses the code used to

implement the methods illustrated herein, the execution of the code converts the general-purpose computer to a dedicated computer for performing the methods illustrated herein.

[00124] Those of ordinary skill in the art can realize that the units and method steps of the examples described in conjunction with the embodiments disclosed herein are capable of being implemented with electronic hardware, or a combination of computer software and electronic hardware. Whether these functions are performed in hardware or software depends on the specific application and design constraints of the technical solution. A skilled professional may use different methods to implement the described functions for each particular application, but such implementations should not be considered outside the scope of the embodiments of the present disclosure.

[00125] The above implementation is only for the purpose of illustrating the embodiments of the present disclosure, and is not a limitation of the embodiments of the present disclosure. A person of ordinary skill in the art concerned may make various variations and variants without departing from the spirit and scope of the embodiments of the present disclosure, so that all equivalent technical solutions also fall within the scope of the embodiments of the present disclosure, and the scope of patent protection of the embodiments of the present disclosure shall be limited by the claims.

[00126] In some embodiments, a non-transitory computer-readable storage medium including instructions is also provided, and the instructions may be executed by a device, for performing the above-described methods. Common forms of non-transitory media include, for example, a floppy disk, a flexible disk, hard disk, solid state drive, magnetic tape, or any other magnetic data storage medium, a CD-ROM, any other optical data storage medium, any physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM or any other flash memory, NVRAM, a cache, a register, any other memory chip or cartridge, and networked versions of the same. The device may

include one or more processors (CPUs), an input/output interface, a network interface, or a memory.

[00127] It should be noted that, the relational terms herein such as “first” and “second” are used only to differentiate an entity or operation from another entity or operation, and do not require or imply any actual relationship or sequence between these entities or operations. Moreover, the words “comprising,” “having,” “containing,” and “including,” and other similar forms are intended to be equivalent in meaning and be open ended in that an item or items following any one of these words is not meant to be an exhaustive listing of such item or items, or meant to be limited to only the listed item or items.

[00128] As used herein, unless specifically stated otherwise, the term “or” encompasses all possible combinations, except where infeasible. For example, if it is stated that a database may include A or B, then, unless specifically stated otherwise or infeasible, the database may include A, or B, or A and B. As a second example, if it is stated that a database may include A, B, or C, then, unless specifically stated otherwise or infeasible, the database may include A, or B, or C, or A and B, or A and C, or B and C, or A and B and C.

[00129] It is appreciated that the above-described embodiments can be implemented by hardware, or software (program codes), or a combination of hardware and software. If implemented by software, it may be stored in the above-described computer-readable media. The software, when executed by the processor can perform the disclosed methods. The computing units and other functional units described in this disclosure can be implemented by hardware, or software, or a combination of hardware and software. One of ordinary skill in the art will also understand that multiple ones of the above-

described modules/units may be combined as one module/unit, and each of the above-described modules/units may be further divided into a plurality of sub-modules/sub-units.

[00130] In the foregoing specification, embodiments have been described with reference to numerous specific details that can vary from implementation to implementation. Certain adaptations and modifications of the described embodiments can be made. Other embodiments can be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed herein. It is intended that the specification and examples be considered as exemplary only, with a true scope and spirit of the invention being indicated by the following claims. It is also intended that the sequence of steps shown in figures are only for illustrative purposes and are not intended to be limited to any particular sequence of steps. As such, those skilled in the art can appreciate that these steps can be performed in a different order while implementing the same method.

[00131] In the drawings and specification, there have been disclosed exemplary embodiments. However, many variations and modifications can be made to these embodiments. Accordingly, although specific terms are employed, they are used in a generic and descriptive sense only and not for purposes of limitation.

WHAT IS CLAIMED IS:

[Claim 1] A speech enhancement method, comprising:

converting noisy speech data into time-frequency domain feature data;

generating a mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in a frequency direction; and

generating enhanced speech data for the noisy speech data according to the mask value and the time-frequency domain feature data.

[Claim 2] The speech enhancement method according to claim 1, wherein generating the mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in the frequency direction comprises:

inputting the time-frequency domain feature data into a speech enhancement model to obtain correlation feature data that is output from the speech enhancement model; and

processing a real part and an imaginary part of the correlation feature data separately by activation functions to obtain the mask value.

[Claim 3] The speech enhancement method according to claim 2, wherein the speech enhancement model comprises an encoder, an attention module, a recurrent module, and a decoder, and

the inputting the time-frequency domain feature data into the speech enhancement model to obtain correlation feature data that is output from the speech enhancement model further comprises:

processing the time-frequency domain feature data using the encoder;

processing output data of the encoder separately using the attention module and the recurrent module, wherein the attention module is configured to filter interference information in the output data of the encoder; and

processing output data of the attention module and output data of the recurrent

module using the decoder to obtain the correlation feature data.

[Claim 4] The speech enhancement method according to claim 3, wherein the encoder comprises M first convolutional recurrent blocks, M being a natural number greater than or equal to 2, the first convolutional recurrent block comprising a first convolution unit for first convolution processing and a first recurrent unit for first recurrent processing, and processing the time-frequency domain feature data using the encoder further comprises:

for an ith first convolutional recurrent block, performing the first convolution processing on input data using the first convolution unit in the ith first convolutional recurrent block; performing the first recurrent processing on a result of the first convolution processing using the first recurrent unit in the ith first convolutional recurrent block, so as to obtain output data of the ith first convolutional recurrent block, wherein $1 \leq i \leq M$; and

inputting output data of each first convolutional recurrent block to the attention module and inputting output data of an Mth first convolutional recurrent block to the recurrent module.

[Claim 5] The speech enhancement method according to claim 4, wherein

if a value of i is 1, the input data of the ith first convolutional recurrent block is the time-frequency domain feature data; and

if the value of i is 2 to M, the input data of the ith first convolutional recurrent block is output data of the (i-1)th first convolutional recurrent block.

[Claim 6] The speech enhancement method according to claim 4, wherein processing the output data of the encoder using the attention module further comprises:

processing output data of each of the first convolutional recurrent blocks separately using the attention module and inputting the output data of the attention module to the decoder.

[Claim 7] The speech enhancement method according to claim 4, wherein processing the output data of the encoder using the recurrent module further comprises:

processing the output data of the Mth first convolutional recurrent block using the recurrent module; and

inputting the output data of the recurrent module to the decoder.

[Claim 8] The speech enhancement method according to claim 4, wherein the decoder comprises M second convolutional recurrent blocks, the second convolutional recurrent block comprising a second convolution unit for second convolution processing and a second recurrent unit for second recurrent processing, and

processing output data of the attention module and output data of the recurrent module using the decoder to obtain the correlation feature data further comprises:

for an ith second convolutional recurrent block, performing second convolution processing on first input data and second input data using the second convolution unit in the ith second convolutional recurrent block, and performing second recurrent processing on a result of the second convolution processing using the second recurrent unit in the ith second convolutional recurrent block, to obtain output data of the ith second convolutional recurrent block, wherein output data of an Mth second convolutional recurrent block is the correlation feature data;

if the value of i is 1, the first input data of the second convolution unit in the ith second convolutional recurrent block is output data of the attention module after processing the output data of the Mth first convolutional recurrent block, and the second input data of the second convolution unit in the ith second convolutional recurrent block is the output data of the recurrent module; and

if the value of i is 2 to M, the first input data of the second convolution unit in the ith second convolutional recurrent block is output data of the attention module after processing

output data of an $(M+1-i)$ th first convolutional recurrent block, and the second input data of the second convolution unit in the i th second convolutional recurrent block is output data of an $(i-1)$ th second convolutional recurrent block.

[Claim 9] A speech enhancement model, comprising:

an encoder configured to process time-frequency domain feature data corresponding to noisy speech data;

an attention module configured to process output data of the encoder;

a recurrent module configured to process the output data of the encoder; and

a decoder configured to process output data of the attention module and the recurrent module to obtain correlation feature data for characterizing a long-range correlation of the time-frequency domain feature data in a frequency direction.

[Claim 10] The speech enhancement model according to claim 9, wherein the speech enhancement model is obtained through training in a following manner:

time-frequency domain feature data of a training sample is inputted into a model to be trained;

a first loss value of the model to be trained is determined according to at least one of a signal-to-noise ratio loss, a mask loss, and a real-imaginary part spectral loss;

a model parameter of the model to be trained is adjusted according to the first loss value until the first loss value satisfies a preset first condition;

a second loss value of the model to be trained is determined according to at least one of the first loss value, an adversarial loss, and a depth feature loss; and

the model parameter of the model to be trained is adjusted according to the second loss value until the second loss value satisfies a preset second condition to obtain the speech enhancement model.

[Claim 11] A speech recognition method, comprising:

acquiring speech data to be processed, wherein the speech data to be processed comprises noise, and the speech data to be processed comprises one of the following: audio/video conference speech data, online education speech data, and live webcast speech data;

converting the speech data to be processed into time-frequency domain feature data;

generating a mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in a frequency direction;

generating enhanced speech data for the speech data to be processed according to the mask value and the time-frequency domain feature data; and

performing speech recognition on the enhanced speech data to obtain a recognition result.

[Claim 12] The speech recognition method according to claim 11, wherein the generating the mask value of the noisy speech data according to a long-range correlation of the time-frequency domain feature data in the frequency direction comprises:

inputting the time-frequency domain feature data into a speech enhancement model to obtain correlation feature data that is output from the speech enhancement model; and

processing a real part and an imaginary part of the correlation feature data separately by activation functions to obtain the mask value.

[Claim 13] The speech recognition method according to claim 12, wherein the speech enhancement model comprises an encoder, an attention module, a recurrent module, and a decoder, and

the inputting the time-frequency domain feature data into the speech enhancement model to obtain correlation feature data that is output from the speech enhancement model further comprises:

processing the time-frequency domain feature data using the encoder;

processing output data of the encoder separately using the attention module and the recurrent module, wherein the attention module is configured to filter interference information in the output data of the encoder; and

processing output data of the attention module and output data of the recurrent module using the decoder to obtain the correlation feature data.

[Claim 14] The speech recognition method according to claim 13, wherein the encoder comprises M first convolutional recurrent blocks, M being a natural number greater than or equal to 2, the first convolutional recurrent block comprising a first convolution unit for first convolution processing and a first recurrent unit for first recurrent processing, and

the processing the time-frequency domain feature data using the encoder further comprises:

for an i th first convolutional recurrent block, performing the first convolution processing on input data using the first convolution unit in the i th first convolutional recurrent block, and performing the first recurrent processing on a result of the first convolution processing using the first recurrent unit in the i th first convolutional recurrent block, so as to obtain output data of the i th first convolutional recurrent block, wherein $1 \leq i \leq M$; and

inputting output data of each first convolutional recurrent block to the attention module and inputting output data of an M th first convolutional recurrent block to the recurrent module.

[Claim 15] The speech recognition method according to claim 14, wherein

if a value of i is 1, the input data of the i th first convolutional recurrent block is the time-frequency domain feature data; and

if the value of i is 2 to M , the input data of the i th first convolutional recurrent block is output data of the $(i-1)$ th first convolutional recurrent block.

[Claim 16] The speech recognition method according to claim 14, wherein the processing the output data of the encoder using the attention module further comprises:

processing output data of each of the first convolutional recurrent blocks separately using the attention module and inputting the output data of the attention module to the decoder.

[Claim 17] The speech recognition method according to claim 14, wherein the processing the output data of the encoder using the recurrent module further comprises:

processing the output data of the Mth first convolutional recurrent block using the recurrent module; and

inputting the output data of the recurrent module to the decoder.

[Claim 18] An electronic device, comprising: a processor, a memory, a communication interface, and a communication bus, the processor, the memory, and the communication interface communicating with each other via the communication bus, wherein

the memory is configured to store at least one executable instruction that causes the processor to perform the speech enhancement method according to any one of claims 1 to 8 or the speech recognition method according to any one of claims 11 to 17.

[Claim 19] A computer storage medium having a computer program stored thereon, wherein the program, when executed by a processor, implements the speech enhancement method according to any one of claims 1 to 8 or the speech recognition method according to any one of claims 11 to 17.

[Claim 20] A computer program product comprising computer instructions, wherein the computer instructions instruct a computing device to perform the speech enhancement method according to any one of claims 1 to 8 or the speech recognition method according to any one of claims 11 to 17.