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;

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 2] The speech enhancement method according to claim 1, 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 3] The speech enhancement method according to claim 2, 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 4] The speech enhancement method according to claim 3, 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 5] The speech enhancement method according to claim 3, 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 6] The speech enhancement method according to claim 3, 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 7] The speech enhancement method according to claim 3, 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 ith second convolutional recurrent block is output data of an (i-1)th second convolutional recurrent block.

[Claim 8] 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;

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 9] The speech recognition method according to claim 8, 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 10] The speech recognition method according to claim 9, 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 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, and 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 11] The speech recognition method according to claim 10, 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 12] The speech recognition method according to claim 10, 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 13] The speech recognition method according to claim 10, 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 14] 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 7 or the speech recognition method according to any one of claims 8 to 13.

[Claim 15] 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 7 or the speech recognition method according to any one of claims 8 to 13.

[Claim 16] 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 7 or the speech recognition method according to any one of claims 8 to 13.