SPEECH/AUDIO BITSTREAM DECODING METHOD AND APPARATUS

[0001] .

TECHNICAL FIELD

[0002] The present invention relates to audio decoding technologies, and specifically, to a speech/audio bitstream decoding method and apparatus.

BACKGROUND

[0003] In a system based on Voice over Internet Protocol (VoIP, Voice over Internet Protocol), a packet may need to pass through multiple routers in a transmission process, but because these routers may change in a call process, a transmission delay in the call process may change. In addition, when two or more users attempt to enter a network by using a same gateway, a routing delay may change, and such a delay change is called a delay jitter (delay jitter). Similarly, a delay jitter may also be caused when a receiver, a transmitter, a gateway, and the like use a non-real-time operating system, and in a severe situation, a data packet loss occurs, resulting in speech/audio distortion and deterioration of VoIP quality.

[0004] Currently, many technologies have been used at different layers of a communication system to reduce a delay, smooth a delay jitter, and perform packet loss compensation. A receiver may use a high-efficiency jitter buffer processing (JBM, Jitter Buffer Management) algorithm to compensate for a network delay jitter to some extent. However, in a case of a relatively high packet loss rate, apparently, a high-quality communication requirement cannot be met only by using the JBM technology.

[0005] To help avoid the quality deterioration problem caused by a delay jitter of a speech/audio frame, a redundancy coding algorithm is introduced. That is, in addition to encoding current speech/audio frame information at a particular bit rate, an encoder encodes other speech/audio frame information than the current speech/audio

frame at a lower bit rate, and transmits a relatively low bit rate bitstream of the other speech/audio frame information, as redundancy information, to a decoder together with a bitstream of the current speech/audio frame information. When a speech/audio frame is lost, if a jitter buffer buffers or a received bitstream includes redundancy information of the lost speech/audio frame, the decoder recovers the lost speech/audio frame according to the redundancy information, thereby improving speech/audio quality.

[0006] In an existing redundancy coding algorithm, in addition to including speech/audio frame information of the N^{th} frame, a bitstream of the N^{th} frame includes speech/audio frame information of the $(N-M)^{th}$ frame at lower bit rate. In a transmission process, if the $(N-M)^{th}$ frame is lost, decoding processing is performed according to the speech/audio frame information that is of the $(N-M)^{th}$ frame and is included in the bitstream of the N^{th} frame, to recover a speech/audio signal of the $(N-M)^{th}$ frame.

[0007] It can be learned from the foregoing description that, in the existing redundancy coding algorithm, redundancy bitstream information is obtained by means of encoding at a lower bit rate, which is therefore highly likely to cause signal instability and further cause low quality of an output speech/audio signal.

SUMMARY

[0008] Embodiments of the present invention provide a speech/audio bitstream decoding method and apparatus, which help improve quality of an output speech/audio signal.

[0009] A first aspect of the embodiments of the present invention provides a speech/audio bitstream decoding method, which may include:

acquiring a speech/audio decoding parameter of a current speech/audio frame, where the current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame;

performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio dccoding parameter of the current speech/audio frame, where the X speech/audio frames include M speech/audio frames

previous to the current speech/audio frame and/or N speech/audio frames next to the current speech/audio frame, and M and N are positive integers; and

recovering a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame.

[0010] With reference to the first aspect, in a first possible implementation manner of the first aspect, the speech/audio decoding parameter of the current speech/audio frame includes a spectral pair parameter of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame includes:

performing post processing on the spectral pair parameter of the current speech/audio frame according to at least one of a signal class, a spectrum tilt factor, an adaptive codebook gain, or a spectral pair parameter of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the current speech/audio frame.

[0011] With reference to the first possible implementation manner of the first aspect, in a second possible implementation manner of the first aspect, the performing post processing on the spectral pair parameter of the current speech/audio frame according to at least one of a signal class, a spectrum tilt factor, an adaptive codebook gain, or a spectral pair parameter of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the current speech/audio frame includes:

if the current speech/audio frame is a normal decoded frame, the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is unvoiced, and a signal class of the speech/audio frame previous to the current speech/audio frame is not unvoiced, using the spectral pair parameter of the current speech/audio frame as the post-processed spectral pair parameter of the current speech/audio frame, or obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame; or

if the current speech/audio frame is a normal decoded frame, the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is unvoiced, and a signal class of the speech/audio frame previous to the current speech/audio frame is not unvoiced,

obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the current speech/audio frame is unvoiced, using a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame as the post-processed spectral pair parameter of the current speech/audio frame, or obtaining the post-processed spectral pair parameter of the current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the current speech/audio frame is unvoiced, obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, a maximum value of an adaptive codebook gain of a subframe in a speech/audio frame next to the current speech/audio frame is less than or equal to a first threshold, and a spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a second threshold, using a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame as the post-processed spectral pair parameter of the current speech/audio frame, or obtaining the post-processed spectral pair parameter of the current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, a maximum value of an adaptive codebook gain of a subframe in a speech/audio frame next to the current speech/audio frame is less than or equal to a first threshold, and a spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a second threshold, obtaining the post-processed spectral pair parameter of

the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, a speech/audio frame next to the current speech/audio frame is unvoiced, a maximum value of an adaptive codebook gain of a subframe in the speech/audio frame next to the current speech/audio frame is less than or equal to a third threshold, and a spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a fourth threshold, using a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame as the post-processed spectral pair parameter of the current speech/audio frame, or obtaining the post-processed spectral pair parameter of the current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, a signal class of a speech/audio frame next to the current speech/audio frame is unvoiced, a maximum value of an adaptive codebook gain of a subframe in the speech/audio frame next to the current speech/audio frame is less than or equal to a third threshold, and a spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a fourth threshold, obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame.

[0012] With reference to the second possible implementation manner of the first aspect, in a third possible implementation manner of the first aspect, the obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame includes: obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame and by using the following formula:

$$lsp[k] = \alpha * lsp_old[k] + \beta * lsp_mid[k] + \delta * lsp_new[k] \quad 0 \le k \le L$$,

where

$lsp[k]$ is the post-processed spectral pair parameter of the current speech/audio frame, $lsp_old[k]$ is the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, $lsp_mid[k]$ is a middle value of the spectral pair parameter of the current speech/audio frame, $lsp_new[k]$ is the spectral pair parameter of the current speech/audio frame, L is an order of a spectral pair parameter, α is a weight of the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, β is a weight of the middle value of the spectral pair parameter of the current speech/audio frame, δ is a weight of the spectral pair parameter of the current speech/audio frame, $\alpha \ge 0$, $\beta \ge 0$, $\delta \ge 0$, and $\alpha + \beta + \delta = 1$, where

if the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, α is equal to 0 or α is less than or equal to a fifth threshold; or

if the current speech/audio frame is a redundant decoded frame, β is equal to 0 or β is less than or equal to a sixth threshold; or

if the current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than or equal to a seventh threshold; or

if the current speech/audio frame is a redundant decoded frame, β is equal to 0 or β is less than or equal to a sixth threshold, and δ is equal to 0 or δ is less than or equal to a seventh threshold.

[0013] With reference to the second possible implementation manner of the first aspect, in a fourth possible implementation manner of the first aspect, the obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame includes: obtaining the post-processed spectral pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and the spectral pair parameter of the speech/audio frame previous to the

current speech/audio frame and by using the following formula:

$lsp[k] = \alpha * lsp_old[k] + \delta * lsp_new[k] \quad 0 \le k \le L$, where

$lsp[k]$ is the post-processed spectral pair parameter of the current speech/audio frame, $lsp_old[k]$ is the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, $lsp_new[k]$ is the spectral pair parameter of the current speech/audio frame, L is an order of a spectral pair parameter, α is a weight of the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, δ is a weight of the spectral pair parameter of the current speech/audio frame, $\alpha \ge 0$, $\delta \ge 0$, and $\alpha + \delta = 1$, where

if the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, α is equal to 0 or α is less than or equal to a fifth threshold; or

if the current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than or equal to a seventh threshold.

[0014] With reference to the first aspect, or any one of the first to the fourth possible implementation manner of the first aspect, in a fifth possible implementation manner of the first aspect,

the speech/audio decoding parameter of the current speech/audio frame includes an adaptive codebook gain of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame includes:

performing post processing on the adaptive codebook gain of the current speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames, to obtain a post-processed adaptive codebook gain of the current speech/audio frame.

[0015] With reference to the fifth possible implementation manner of the first aspect, in a sixth possible implementation manner of the first aspect, the performing post processing on the adaptive codebook gain of the current speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames includes:

if the current speech/audio frame is a redundant decoded frame, the signal class of the current speech/audio frame is not unvoiced, a signal class of at least one of two speech/audio frames next to the current speech/audio frame is unvoiced, and an algebraic codebook gain of a current subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the current speech/audio frame, attenuating an adaptive codebook gain of the current subframe; or

if the current speech/audio frame is a redundant decoded frame, the signal class of the current speech/audio frame is not unvoiced, a signal class of at least one of two speech/audio frames next to the current speech/audio frame is unvoiced, and an algebraic codebook gain of a current subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the current subframe, attenuating an adaptive codebook gain of the current subframe; or

if the current speech/audio frame is a redundant decoded frame, or the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, and if the signal class of the current speech/audio frame is generic, the signal class of the speech/audio frame next to the current speech/audio frame is voiced, and an algebraic codebook gain of a subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the subframe, adjusting an adaptive codebook gain of a current subframe of the current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the current speech/audio frame to that of a subframe adjacent to the current subframe, a ratio of the adaptive codebook gain of the current subframe of the current speech/audio frame to that of the subframe adjacent to the current subframe, or a ratio of the algebraic codebook gain of the current subframe of the current speech/audio frame to that of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, or the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, and if the signal class of the current speech/audio frame is generic, the signal class of the speech/audio frame next to the current speech/audio frame is voiced, and an algebraic codebook gain of a subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the current

speech/audio frame, adjusting an adaptive codebook gain of a current subframe of the current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the current speech/audio frame to that of a subframe adjacent to the current subframe, a ratio of the adaptive codebook gain of the current subframe of the current speech/audio frame to that of the subframe adjacent to the current subframe, or a ratio of the algebraic codebook gain of the current subframe of the current speech/audio frame to that of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, or the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, and if the current speech/audio frame is voiced, the signal class of the speech/audio frame previous to the current speech/audio frame is generic, and an algebraic codebook gain of a subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the subframe, adjusting an adaptive codebook gain of a current subframe of the current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the current speech/audio frame to that of a subframe adjacent to the current subframe, a ratio of the adaptive codebook gain of the current subframe of the current speech/audio frame to that of the subframe adjacent to the current subframe, or a ratio of the algebraic codebook gain of the current subframe of the current speech/audio frame to that of the speech/audio frame previous to the current speech/audio frame; or

if the current speech/audio frame is a redundant decoded frame, or the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, and if the signal class of the current speech/audio frame is voiced, the signal class of the speech/audio frame previous to the current speech/audio frame is generic, and an algebraic codebook gain of a subframe of the current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the current speech/audio frame, adjusting an adaptive codebook gain of a current subframe of the current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the current speech/audio frame to that of a subframe adjacent to the current subframe, a ratio of the adaptive codebook gain of the current subframe of the current speech/audio frame to that of the subframe

adjacent to the current subframe, or a ratio of the algebraic codebook gain of the current subframe of the current speech/audio frame to that of the speech/audio frame previous to the current speech/audio frame.

[0016] With reference to the first aspect, or any one of the first to the sixth possible implementation manner of the first aspect, in a seventh possible implementation manner of the first aspect, the speech/audio decoding parameter of the current speech/audio frame includes an algebraic codebook of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame includes: performing post processing on the algebraic codebook of the current speech/audio frame according to at least one of the signal class, an algebraic codebook, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed algebraic codebook of the current speech/audio frame.

[0017] With reference to the seventh possible implementation manner of the first aspect, in an eighth possible implementation manner of the first aspect, the performing post processing on the algebraic codebook of the current speech/audio frame according to at least one of the signal class, an algebraic codebook, or the spectrum tilt factor of the X speech/audio frames includes:

if the current speech/audio frame is a redundant decoded frame, the signal class of the speech/audio frame next to the current speech/audio frame is unvoiced, the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to an eighth threshold, and an algebraic codebook of a subframe of the current speech/audio frame is 0 or is less than or equal to a ninth threshold, using an algebraic codebook or a random noise of a subframe previous to the current speech/audio frame as an algebraic codebook of the current subframe.

[0018] With reference to the first aspect, or any one of the first to the eighth possible implementation manner of the first aspect, in a ninth possible implementation manner of the first aspect, the speech/audio decoding parameter of the current speech/audio frame includes a bandwidth extension envelope of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter

of the current speech/audio frame includes: performing post processing on the bandwidth extension envelope of the current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the current speech/audio frame.

[0019] With reference to the ninth possible implementation manner of the first aspect, in a tenth possible implementation manner of the first aspect,

the performing post processing on the bandwidth extension envelope of the current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the current speech/audio frame includes:

if the speech/audio frame previous to the current speech/audio frame is a normal decoded frame, and the signal class of the speech/audio frame previous to the current speech/audio frame is the same as that of the speech/audio frame next to the current speech/audio frame, obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

if the current speech/audio frame is a prediction form of redundancy decoding, obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

if the signal class of the current speech/audio frame is not unvoiced, the signal class of the speech/audio frame next to the current speech/audio frame is unvoiced, and the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a tenth threshold, modifying the bandwidth extension envelope of the current speech/audio frame according to a bandwidth extension envelope or the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, to obtain the post-processed bandwidth extension envelope of the current speech/audio frame.

[0020] With reference to the tenth possible implementation manner of the first aspect, in an eleventh possible implementation manner of the first aspect,

the obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame includes: obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame and by using the following formula:

$GainFrame = fac1 * GainFrame_old + fac2 * GainFrame_new$, where

$GainFrame$ is the post-processed bandwidth extension envelope of the current speech/audio frame, $GainFrame_old$ is the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame, $GainFrame_new$ is the bandwidth extension envelope of the current speech/audio frame, $fac1$ is a weight of the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame, $fac2$ is a weight of the bandwidth extension envelope of the current speech/audio frame, $fac1 \geq 0$, $fac2 \geq 0$, and $fac1 + fac2 = 1$.

[0021] With reference to the tenth possible implementation manner of the first aspect, in a twelfth possible implementation manner of the first aspect, a modification factor for modifying the bandwidth extension envelope of the current speech/audio frame is inversely proportional to the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, and is proportional to a ratio of the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame to the bandwidth extension envelope of the current speech/audio frame.

[0022] With reference to the first aspect, or any one of the first to the twelfth possible implementation manner of the first aspect, in a thirteenth possible implementation manner of the first aspect, the speech/audio decoding parameter of the current speech/audio frame includes a pitch period of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current

speech/audio frame includes: performing post processing on the pitch period of the current speech/audio frame according to the signal classes and/or pitch periods of the X speech/audio frames, to obtain a post-processed pitch period of the current speech/audio frame.

[0023] A second aspect of the embodiments of the present invention provides a decoder for decoding a speech/audio bitstream, including:

a parameter acquiring unit, configured to acquire a speech/audio decoding parameter of a current speech/audio frame, where the current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame;

a post processing unit, configured to perform post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame, where the X speech/audio frames include M speech/audio frames previous to the current speech/audio frame and/or N speech/audio frames next to the current speech/audio frame, and M and N are positive integers; and

a recovery unit, configured to recover a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame.

[0024] With reference to the second aspect, in a first possible implementation manner of the second aspect, the post processing unit is specifically configured to perform some or all steps of any method described in any one of the first to the twelfth possible implementation manner of the first aspect.

[0025] if the current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than or equal to a seventh threshold; or if the current speech/audio frame is a redundant decoded frame, β is equal to 0 or β is less than or equal to a sixth threshold, an pair parameter of the current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame and by using the following formula:

$lsp[k] = \alpha * lsp_old[k] + \delta * lsp_new[k] \quad 0 \le k \le L$, where

$lsp[k]$ is the post-processed spectral pair parameter of the current

speech/audio frame, $lsp_old[k]$ is the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, $lsp_new[k]$ is the spectral pair parameter of the current speech/audio frame, L is an order of a spectral pair parameter, α is a weight of the spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, δ is a weight of the spectral pair parameter of the current speech/audio frame, $\alpha \geq 0$, $\delta \geq 0$, and $\alpha + \delta = 1$, where if the current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, α is equal to 0 or α is less than or equal to a fifth threshold; or if the current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than.

[0037] A third aspect of the embodiments of the present invention provides a computer storage medium, where the computer storage medium may store a program, and when being executed, the program includes some or all steps of any speech/audio bitstream decoding method described in the embodiments of the present invention.

[0038] It can be learned that in some embodiments of the present invention, in a scenario in which a current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, after obtaining a speech/audio decoding parameter of the current speech/audio frame, a decoder performs post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and recovers a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame, which ensures stable quality of a decoded signal during transition between a redundant decoded frame and a normal decoded frame or between a redundant decoded frame and a frame erasure concealment (FEC, Frame erasure concealment) recovered frame, thereby improving quality of an output speech/audio signal.

BRIEF DESCRIPTION OF DRAWINGS

[0039] To describe the technical solutions in the embodiments of the present invention more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. Apparently, the accompanying drawings in the following description show merely some embodiments of the present invention, and persons of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

[0040] FIG. 1 is a schematic flowchart of a speech/audio bitstream decoding method according to an embodiment of the present invention;

[0041] FIG. 2 is a schematic flowchart of another speech/audio bitstream decoding method according to an embodiment of the present invention;

[0042] FIG. 3 is a schematic diagram of a decoder according to an embodiment of the present invention;

[0043] FIG. 4 is a schematic diagram of another decoder according to an embodiment of the present invention; and

[0044] FIG. 5 is a schematic diagram of another decoder according to an embodiment of the present invention.

DESCRIPTION OF EMBODIMENTS

[0045] Embodiments of the present invention provide a speech/audio bitstream decoding method and apparatus, which help improve quality of an output speech/audio signal.

[0046] To make the invention objectives, features, and advantages of the present invention clearer and more comprehensible, the following clearly describes the technical solutions in the embodiments of the present invention with reference to the accompanying drawings in the embodiments of the present invention. Apparently, the embodiments described in the following are merely a part rather than all of the embodiments of the present invention. All other embodiments obtained by persons of ordinary skill in the art based on the embodiments of the present invention without creative efforts shall fall within the protection scope of the present invention.

[0047] In the specification, claims, and accompanying drawings of the present invention, the terms "first", "second", "third", "fourth", and so on are intended to

distinguish between different objects but not to indicate a particular order. In addition, the terms "including", "including", or any other variant thereof, are intended to cover a non-exclusive inclusion. For example, a process, a method, a system, a product, or a device including a series of steps or units is not limited to the listed steps or units, and may include steps or units that are not listed.

[0048] The following gives respective descriptions in details.

[0049] The speech/audio bitstream decoding method provided in the embodiments of the present invention is first described. The speech/audio bitstream decoding method provided in the embodiments of the present invention is executed by a decoder, where the decoder may be any apparatus that needs to output speeches, for example, a device such as a mobile phone, a notebook computer, a tablet computer, or a personal computer.

[0050] In an embodiment of the speech/audio bitstream decoding method in the present invention, the speech/audio bitstream decoding method may include: acquiring a speech/audio decoding parameter of a current speech/audio frame, where the foregoing current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame; performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and M and N are positive integers; and recovering a speech/audio signal of the foregoing current speech/audio frame by using the post-processed speech/audio decoding parameter of the foregoing current speech/audio frame.

[0051] FIG. 1 is a schematic flowchart of a speech/audio bitstream decoding method according to an embodiment of the present invention. The speech/audio bitstream decoding method provided in this embodiment of the present invention may include the following content:

[0052] 101. Acquire a speech/audio decoding parameter of a current speech/audio frame.

[0053] The foregoing current speech/audio frame is a redundant decoded frame or

a speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame.

[0054] When the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, the current speech/audio frame may be a normal decoded frame, an FEC recovered frame, or a redundant decoded frame, where if the current speech/audio frame is an FEC recovered frame, the speech/audio decoding parameter of the current speech/audio frame may be predicated based on an FEC algorithm.

[0055] 102. Perform post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and M and N are positive integers.

[0056] That a speech/audio frame (for example, the current speech/audio frame or the speech/audio frame previous to the current speech/audio frame) is a normal decoded frame means that a speech/audio parameter of the foregoing speech/audio frame can be directly obtained from a bitstream of the speech/audio frame by means of decoding. That a speech/audio frame (for example, a current speech/audio frame or a speech/audio frame previous to a current speech/audio frame) is a redundant decoded frame means that a speech/audio parameter of the speech/audio frame cannot be directly obtained from a bitstream of the speech/audio frame by means of decoding, but redundant bitstream information of the speech/audio frame can be obtained from a bitstream of another speech/audio frame.

[0057] The M speech/audio frames previous to the current speech/audio frame refer to M speech/audio frames preceding the current speech/audio frame and immediately adjacent to the current speech/audio frame in a time domain.

[0058] For example, M may be equal to 1, 2, 3, or another value. When M=1, the M speech/audio frames previous to the current speech/audio frame are the speech/audio frame previous to the current speech/audio frame, and the speech/audio frame previous to the current speech/audio frame and the current speech/audio frame are two immediately adjacent speech/audio frames; when M=2, the M speech/audio frames previous to the current speech/audio frame are the speech/audio frame

previous to the current speech/audio frame and a speech/audio frame previous to the speech/audio frame previous to the current speech/audio frame, and the speech/audio frame previous to the current speech/audio frame, the speech/audio frame previous to the speech/audio frame previous to the current speech/audio frame, and the current speech/audio frame are three immediately adjacent speech/audio frames; and so on.

[0059] The N speech/audio frames next to the current speech/audio frame refer to N speech/audio frames following the current speech/audio frame and immediately adjacent to the current speech/audio frame in a time domain.

[0060] For example, N may be equal to 1, 2, 3, 4, or another value. When N=1, the N speech/audio frames next to the current speech/audio frame are a speech/audio frame next to the current speech/audio frame, and the speech/audio frame next to the current speech/audio frame and the current speech/audio frame are two immediately adjacent speech/audio frames; when N=2, the N speech/audio frames next to the current speech/audio frame are a speech/audio frame next to the current speech/audio frame and a speech/audio frame next to the speech/audio frame next to the current speech/audio frame, and the speech/audio frame next to the current speech/audio frame, the speech/audio frame next to the speech/audio frame next to the current speech/audio frame, and the current speech/audio frame are three immediately adjacent speech/audio frames; and so on.

[0061] The speech/audio decoding parameter may include at least one of the following parameters:

a bandwidth extension envelope, an adaptive codebook gain (gain_pit), an algebraic codebook, a pitch period, a spectrum tilt factor, a spectral pair parameter, and the like.

[0062] The speech/audio parameter may include a speech/audio decoding parameter, a signal class, and the like.

[0063] A signal class of a speech/audio frame may be unvoiced (UNVOICED), voiced (VOICED), generic (GENERIC), transient (TRANSIENT), inactive (INACTIVE), or the like.

[0064] The spectral pair parameter may be, for example, at least one of a line spectral pair (LSP: Line Spectral Pair) parameter or an immittance spectral pair (ISP: Immittance Spectral Pair) parameter.

[0065] It may be understood that in this embodiment of the present invention, post processing may be performed on at least one speech/audio decoding parameter of a

bandwidth extension envelope, an adaptive codebook gain, an algebraic codebook, a pitch period, or a spectral pair parameter of the current speech/audio frame. Specifically, how many parameters are selected and which parameters are selected for post processing may be determined according to an application scenario and an application environment, which is not limited in this embodiment of the present invention.

[0066] Different post processing may be performed on different speech/audio decoding parameters. For example, post processing performed on the spectral pair parameter of the current speech/audio frame may be adaptive weighting performed by using the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, to obtain a post-processed spectral pair parameter of the current speech/audio frame, and post processing performed on the adaptive codebook gain of the current speech/audio frame may be adjustment such as attenuation performed on the adaptive codebook gain.

[0067] A specific post processing manner is not limited in this embodiment of the present invention, and specific post processing may be set according to a requirement or according to an application environment and an application scenario.

[0068] 103. Recover a speech/audio signal of the foregoing current speech/audio frame by using the post-processed speech/audio decoding parameter of the foregoing current speech/audio frame.

[0069] It can be learned from the foregoing description that in this embodiment, in a scenario in which a current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, after obtaining a speech/audio decoding parameter of the current speech/audio frame, a decoder performs post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and recovers a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame, which ensures stable quality of a decoded signal

during transition between a redundant decoded frame and a normal decoded frame or between a redundant decoded frame and an FEC recovered frame, thereby improving quality of an output speech/audio signal.

[0070] In some embodiments of the present invention, the speech/audio decoding parameter of the foregoing current speech/audio frame includes the spectral pair parameter of the foregoing current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, for example, may include: performing post processing on the spectral pair parameter of the foregoing current speech/audio frame according to at least one of a signal class, a spectrum tilt factor, an adaptive codebook gain, or a spectral pair parameter of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the foregoing current speech/audio frame.

[0071] For example, the performing post processing on the spectral pair parameter of the foregoing current speech/audio frame according to at least one of a signal class, a spectrum tilt factor, an adaptive codebook gain, or a spectral pair parameter of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the foregoing current speech/audio frame may include:

if the foregoing current speech/audio frame is a normal decoded frame, the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is unvoiced, and a signal class of the speech/audio frame previous to the foregoing current speech/audio frame is not unvoiced, using the spectral pair parameter of the foregoing current speech/audio frame as the post-processed spectral pair parameter of the foregoing current speech/audio frame, or obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a normal decoded frame, the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is unvoiced, and a signal class of the speech/audio frame previous to the foregoing current speech/audio frame is not unvoiced, obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair

parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the foregoing current speech/audio frame is unvoiced, using a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame as the post-processed spectral pair parameter of the foregoing current speech/audio frame, or obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the foregoing current speech/audio frame is unvoiced, obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, a maximum value of an adaptive codebook gain of a subframe in a speech/audio frame next to the foregoing current speech/audio frame is less than or equal to a first threshold, and a spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a second threshold, using a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame as the post-processed spectral pair parameter of the foregoing current speech/audio frame, or obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, a maximum value of an adaptive codebook gain of a subframe in a speech/audio frame next to the foregoing current speech/audio frame is less than or equal to a first threshold, and a spectrum tilt factor of the speech/audio frame previous to the

foregoing current speech/audio frame is less than or equal to a second threshold, obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, a speech/audio frame next to the foregoing current speech/audio frame is unvoiced, a maximum value of an adaptive codebook gain of a subframe in the speech/audio frame next to the foregoing current speech/audio frame is less than or equal to a third threshold, and a spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a fourth threshold, using a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame as the post-processed spectral pair parameter of the foregoing current speech/audio frame, or obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a redundant decoded frame, a signal class of the foregoing current speech/audio frame is not unvoiced, a signal class of a speech/audio frame next to the foregoing current speech/audio frame is unvoiced, a maximum value of an adaptive codebook gain of a subframe in the speech/audio frame next to the foregoing current speech/audio frame is less than or equal to a third threshold, and a spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a fourth threshold, obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame.

[0072] There may be various manners for obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame.

[0073] For example, the obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the

foregoing current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame may include: specifically obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame and by using the following formula:

$$lsp[k] = \alpha * lsp_old[k] + \beta * lsp_mid[k] + \delta * lsp_new[k] \quad 0 \le k \le L \quad ,$$

where

$lsp[k]$ is the post-processed spectral pair parameter of the foregoing current speech/audio frame, $lsp_old[k]$ is the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame, $lsp_mid[k]$ is a middle value of the spectral pair parameter of the foregoing current speech/audio frame, $lsp_new[k]$ is the spectral pair parameter of the foregoing current speech/audio frame, L is an order of a spectral pair parameter, α is a weight of the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame, β is a weight of the middle value of the spectral pair parameter of the foregoing current speech/audio frame, δ is a weight of the spectral pair parameter of the foregoing current speech/audio frame, $\alpha \ge 0$, $\beta \ge 0$, $\delta \ge 0$, and $\alpha + \beta + \delta = 1$, where

if the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, α is equal to 0 or α is less than or equal to a fifth threshold; or if the foregoing current speech/audio frame is a redundant decoded frame, β is equal to 0 or β is less than or equal to a sixth threshold; or if the foregoing current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than or equal to a seventh threshold; or if the foregoing current speech/audio frame is a redundant decoded frame, β is equal to 0 or β is less than or equal to a sixth threshold, and δ is equal to 0 or δ is less than or equal to a seventh threshold.

[0074] For another example, the obtaining the post-processed spectral pair

parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame may include: specifically obtaining the post-processed spectral pair parameter of the foregoing current speech/audio frame based on the spectral pair parameter of the foregoing current speech/audio frame and the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame and by using the following formula:

$lsp[k] = \alpha * lsp_old[k] + \delta * lsp_new[k] \quad 0 \le k \le L$, where

$lsp[k]$ is the post-processed spectral pair parameter of the foregoing current speech/audio frame, $lsp_old[k]$ is the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame, $lsp_new[k]$ is the spectral pair parameter of the foregoing current speech/audio frame, L is an order of a spectral pair parameter, α is a weight of the spectral pair parameter of the speech/audio frame previous to the foregoing current speech/audio frame, δ is a weight of the spectral pair parameter of the foregoing current speech/audio frame, $\alpha \ge 0$, $\delta \ge 0$, and $\alpha + \delta = 1$, where

if the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, α is equal to 0 or α is less than or equal to a fifth threshold; or if the foregoing current speech/audio frame is a redundant decoded frame, δ is equal to 0 or δ is less than or equal to a seventh threshold.

[0075] The fifth threshold, the sixth threshold, and the seventh threshold each may be set to different values according to different application environments or scenarios. For example, a value of the fifth threshold may be close to 0, where for example, the fifth threshold may be equal to 0.001, 0.002, 0.01, 0.1, or another value close to 0; a value of the sixth threshold may be close to 0, where for example, the sixth threshold may be equal to 0.001, 0.002, 0.01, 0.1, or another value close to 0; and a value of the seventh threshold may be close to 0, where for example, the seventh threshold may be equal to 0.001, 0.002, 0.01, 0.1, or another value close to 0.

[0076] The first threshold, the second threshold, the third threshold, and the fourth threshold each may be set to different values according to different application

environments or scenarios.

[0077] For example, the first threshold may be set to 0.9, 0.8, 0.85, 0.7, 0.89, or 0.91.

[0078] For example, the second threshold may be set to 0.16, 0.15, 0.165, 0.1, 0.161, or 0.159.

[0079] For example, the third threshold may be set to 0.9, 0.8, 0.85, 0.7, 0.89, or 0.91.

[0080] For example, the fourth threshold may be set to 0.16, 0.15, 0.165, 0.1, 0.161, or 0.159.

[0081] The first threshold may be equal to or not equal to the third threshold, and the second threshold may be equal to or not equal to the fourth threshold.

[0082] In other embodiments of the present invention, the speech/audio decoding parameter of the foregoing current speech/audio frame includes the adaptive codebook gain of the foregoing current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame may include: performing post processing on the adaptive codebook gain of the foregoing current speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames, to obtain a post-processed adaptive codebook gain of the foregoing current speech/audio frame.

[0083] For example, the performing post processing on the adaptive codebook gain of the foregoing current speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames may include:

if the foregoing current speech/audio frame is a redundant decoded frame, the signal class of the foregoing current speech/audio frame is not unvoiced, a signal class of at least one of two speech/audio frames next to the foregoing current speech/audio frame is unvoiced, and an algebraic codebook gain of a current subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame (for example, the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame is 1 or more than 1 time, for example, 1,

1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame, attenuating an adaptive codebook gain of the foregoing current subframe; or

if the foregoing current speech/audio frame is a redundant decoded frame, the signal class of the foregoing current speech/audio frame is not unvoiced, a signal class of at least one of the speech/audio frame next to the foregoing current speech/audio frame or a speech/audio frame next to the next speech/audio frame is unvoiced, and an algebraic codebook gain of a current subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the foregoing current subframe (for example, the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame is 1 or more than 1 time, for example, 1, 1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the subframe previous to the foregoing current subframe), attenuating an adaptive codebook gain of the foregoing current subframe; or

if the foregoing current speech/audio frame is a redundant decoded frame, or the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, and if the signal class of the foregoing current speech/audio frame is generic, the signal class of the speech/audio frame next to the foregoing current speech/audio frame is voiced, and an algebraic codebook gain of a subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the foregoing subframe (for example, the algebraic codebook gain of the subframe of the foregoing current speech/audio frame may be 1 or more than 1 time, for example, 1, 1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the subframe previous to the foregoing subframe), adjusting (for example, augmenting or attenuating) an adaptive codebook gain of a current subframe of the foregoing current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of a subframe adjacent to the foregoing current subframe, a ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe, or a ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame (for example, if the ratio of the algebraic codebook gain of

the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to an eleventh threshold (where the eleventh threshold may be equal to, for example, 2, 2.1, 2.5, 3, or another value), the ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to a twelfth threshold (where the twelfth threshold may be equal to, for example, 1, 1.1, 1.5, 2, 2.1, or another value), and the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a thirteenth threshold (where the thirteenth threshold may be equal to, for example, 1, 1.1, 1.5, 2, or another value), the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame may be augmented); or

if the foregoing current speech/audio frame is a redundant decoded frame, or the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, and if the signal class of the foregoing current speech/audio frame is generic, the signal class of the speech/audio frame next to the foregoing current speech/audio frame is voiced, and an algebraic codebook gain of a subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame (where the algebraic codebook gain of the subframe of the foregoing current speech/audio frame is 1 or more than 1 time, for example, 1, 1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame), adjusting (attenuating or augmenting) an adaptive codebook gain of a current subframe of the foregoing current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of a subframe adjacent to the foregoing current subframe, a ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe, or a ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame (for example, if the ratio of the algebraic codebook gain of the current subframe of the

foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to an eleventh threshold (where the eleventh threshold may be equal to, for example, 2, 2.1, 2.5, 3, or another value), the ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to a twelfth threshold (where the twelfth threshold may be equal to, for example, 1, 1.1, 1.5, 2, 2.1, or another value), and the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a thirteenth threshold (where the thirteenth threshold may be equal to, for example, 1, 1.1, 1.5, 2, or another value), the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame may be augmented); or

if the foregoing current speech/audio frame is a redundant decoded frame, or the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, and if the foregoing current speech/audio frame is voiced, the signal class of the speech/audio frame previous to the foregoing current speech/audio frame is generic, and an algebraic codebook gain of a subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of a subframe previous to the foregoing subframe (for example, the algebraic codebook gain of the subframe of the foregoing current speech/audio frame may be 1 or more than 1 time, for example, 1, 1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the subframe previous to the foregoing subframe), adjusting (attenuating or augmenting) an adaptive codebook gain of a current subframe of the foregoing current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of a subframe adjacent to the foregoing current subframe, a ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe, or a ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame (for example, if the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the

subframe adjacent to the foregoing current subframe is greater than or equal to an eleventh threshold (where the eleventh threshold is equal to, for example, 2, 2.1, 2.5, 3, or another value), the ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to a twelfth threshold (where the twelfth threshold is equal to, for example, 1, 1.1, 1.5, 2, 2.1, or another value), and the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a thirteenth threshold (where the thirteenth threshold may be equal to, for example, 1, 1.1, 1.5, 2, or another value), the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame may be augmented; or

if the foregoing current speech/audio frame is a redundant decoded frame, or the foregoing current speech/audio frame is a normal decoded frame, and the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, and if the signal class of the foregoing current speech/audio frame is voiced, the signal class of the speech/audio frame previous to the foregoing current speech/audio frame is generic, and an algebraic codebook gain of a subframe of the foregoing current speech/audio frame is greater than or equal to an algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame (for example, the algebraic codebook gain of the subframe of the foregoing current speech/audio frame is 1 or more than 1 time, for example, 1, 1.5, 2, 2.5, 3, 3.4, or 4 times, the algebraic codebook gain of the speech/audio frame previous to the foregoing current speech/audio frame), adjusting (attenuating or augmenting) an adaptive codebook gain of a current subframe of the foregoing current speech/audio frame based on at least one of a ratio of an algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of a subframe adjacent to the foregoing current subframe, a ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe, or a ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame (for example, if the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing

current subframe is greater than or equal to an eleventh threshold (where the eleventh threshold may be equal to, for example, 2, 2.1, 2.5, 3, or another value), the ratio of the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame to that of the subframe adjacent to the foregoing current subframe is greater than or equal to a twelfth threshold (where the twelfth threshold may be equal to, for example, 1, 1.1, 1.5, 2, 2.1, or another value), and the ratio of the algebraic codebook gain of the current subframe of the foregoing current speech/audio frame to that of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a thirteenth threshold (where the thirteenth threshold is equal to, for example, 1, 1.1, 1.5, 2, or another value), the adaptive codebook gain of the current subframe of the foregoing current speech/audio frame may be augmented.

[0084] In other embodiments of the present invention, the speech/audio decoding parameter of the foregoing current speech/audio frame includes the algebraic codebook of the foregoing current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame may include: performing post processing on the algebraic codebook of the foregoing current speech/audio frame according to at least one of the signal class, an algebraic codebook, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed algebraic codebook of the foregoing current speech/audio frame.

[0085] For example, the performing post processing on the algebraic codebook of the foregoing current speech/audio frame according to at least one of the signal class, an algebraic codebook, or the spectrum tilt factor of the X speech/audio frames may include: if the foregoing current speech/audio frame is a redundant decoded frame, the signal class of the speech/audio frame next to the foregoing current speech/audio frame is unvoiced, the spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to an eighth threshold, and an algebraic codebook of a subframe of the foregoing current speech/audio frame is 0 or is less than or equal to a ninth threshold, using an algebraic codebook or a random noise of a subframe previous to the foregoing current speech/audio frame as an algebraic codebook of the foregoing current subframe.

[0086] The eighth threshold and the ninth threshold each may be set to different values according to different application environments or scenarios.

[0087] For example, the eighth threshold may be set to 0.16, 0.15, 0.165, 0.1, 0.161, or 0.159.

[0088] For example, the ninth threshold may be set to 0.1, 0.09, 0.11, 0.07, 0.101, 0.099, or another value close to 0.

[0089] The eighth threshold may be equal to or not equal to the second threshold.

[0090] In other embodiments of the present invention, the speech/audio decoding parameter of the foregoing current speech/audio frame includes a bandwidth extension envelope of the foregoing current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame may include: performing post processing on the bandwidth extension envelope of the foregoing current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the foregoing current speech/audio frame.

[0091] For example, the performing post processing on the bandwidth extension envelope of the foregoing current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the foregoing current speech/audio frame may include:

if the speech/audio frame previous to the foregoing current speech/audio frame is a normal decoded frame, and the signal class of the speech/audio frame previous to the foregoing current speech/audio frame is the same as that of the speech/audio frame next to the current speech/audio frame, obtaining the post-processed bandwidth extension envelope of the foregoing current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame and the bandwidth extension envelope of the foregoing current speech/audio frame; or

if the foregoing current speech/audio frame is a prediction form of redundancy decoding, obtaining the post-processed bandwidth extension envelope of the foregoing current speech/audio frame based on a bandwidth extension envelope of

the speech/audio frame previous to the foregoing current speech/audio frame and the bandwidth extension envelope of the foregoing current speech/audio frame; or

if the signal class of the foregoing current speech/audio frame is not unvoiced, the signal class of the speech/audio frame next to the foregoing current speech/audio frame is unvoiced, the spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame is less than or equal to a tenth threshold, modifying the bandwidth extension envelope of the foregoing current speech/audio frame according to a bandwidth extension envelope or the spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame, to obtain the post-processed bandwidth extension envelope of the foregoing current speech/audio frame.

[0092] The tenth threshold may be set to different values according to different application environments or scenarios. For example, the tenth threshold may be set to 0.16, 0.15, 0.165, 0.1, 0.161, or 0.159.

[0093] For example, the obtaining the post-processed bandwidth extension envelope of the foregoing current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame and the bandwidth extension envelope of the foregoing current speech/audio frame may include: specifically obtaining the post-processed bandwidth extension envelope of the foregoing current speech/audio frame based on the bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame and the bandwidth extension envelope of the foregoing current speech/audio frame and by using the following formula:

$GainFrame = fac1 * GainFrame_old + fac2 * GainFrame_new$, where

$GainFrame$ is the post-processed bandwidth extension envelope of the foregoing current speech/audio frame, $GainFrame_old$ is the bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame, $GainFrame_new$ is the bandwidth extension envelope of the foregoing current speech/audio frame, $fac1$ is a weight of the bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame, $fac2$ is a weight of the bandwidth extension envelope of the foregoing current speech/audio frame, $fac1 \geq 0$, $fac2 \geq 0$, and $fac1 + fac2 = 1$.

[0094] For another example, a modification factor for modifying the bandwidth extension envelope of the foregoing current speech/audio frame is inversely proportional to the spectrum tilt factor of the speech/audio frame previous to the foregoing current speech/audio frame, and is proportional to a ratio of the bandwidth extension envelope of the speech/audio frame previous to the foregoing current speech/audio frame to the bandwidth extension envelope of the foregoing current speech/audio frame.

[0095] In other embodiments of the present invention, the speech/audio decoding parameter of the foregoing current speech/audio frame includes a pitch period of the foregoing current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame may include: performing post processing on the pitch period of the foregoing current speech/audio frame according to the signal classes and/or pitch periods of the X speech/audio frames (for example, post processing such as augmentation or attenuation may be performed on the pitch period of the foregoing current speech/audio frame according to the signal classes and/or the pitch periods of the X speech/audio frames), to obtain a post-processed pitch period of the foregoing current speech/audio frame.

[0096] It can be learned from the foregoing description that in some embodiments of the present invention, during transition between an unvoiced speech/audio frame and a non-unvoiced speech/audio frame (for example, when a current speech/audio frame is of an unvoiced signal class and is a redundant decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a non unvoiced signal type and is a normal decoded frame, or when a current speech/audio frame is of a non unvoiced signal class and is a normal decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of an unvoiced signal class and is a redundant decoded frame), post processing is performed on a speech/audio decoding parameter of the current speech/audio frame, which helps avoid a click (click) phenomenon caused during the interframe transition between the unvoiced speech/audio frame and the non-unvoiced speech/audio frame, thereby improving quality of an output speech/audio signal.

[0097] In other embodiments of the present invention, during transition between a

generic speech/audio frame and a voiced speech/audio frame (when a current speech/audio frame is a generic frame and is a redundant decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a voiced signal class and is a normal decoded frame, or when a current speech/audio frame is of a voiced signal class and is a normal decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a generic signal class and is a redundant decoded frame), post processing is performed on a speech/audio decoding parameter of the current speech/audio frame, which helps rectify an energy instability phenomenon caused during the transition between a generic frame and a voiced frame, thereby improving quality of an output speech/audio signal.

[0098] In still other embodiments of the present invention, when a current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the current speech/audio frame is unvoiced, a bandwidth extension envelope of the current frame is adjusted, to rectify an energy instability phenomenon in time-domain bandwidth extension, and improve quality of an output speech/audio signal.

[0099] To help better understand and implement the foregoing solution in this embodiment of the present invention, some specific application scenarios are used as examples in the following description.

[0100] Referring to FIG. 2, FIG. 2 is a schematic flowchart of another speech/audio bitstream decoding method according to another embodiment of the present invention. The another speech/audio bitstream decoding method provided in the another embodiment of the present invention may include the following content:

[0101] 201. Determine a decoding status of a current speech/audio frame.

[0102] Specifically, for example, it may be determined, based on a JBM algorithm or another algorithm, that the current speech/audio frame is a normal decoded frame, a redundant decoded frame, or an FEC recovered frame.

[0103] If the current speech/audio frame is a normal decoded frame, and a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame, step 202 is executed.

[0104] If the current speech/audio frame is a redundant decoded frame, step 203 is executed.

[0105] If the current speech/audio frame is an FEC recovered frame, and a speech/audio frame previous to the foregoing current speech/audio frame is a

redundant decoded frame, step 204 is executed.

[0106] 202. Obtain a speech/audio decoding parameter of the current speech/audio frame based on a bitstream of the current speech/audio frame, and jump to step 205.

[0107] 203. Obtain a speech/audio decoding parameter of the foregoing current speech/audio frame based on a redundant bitstream of the current speech/audio frame, and jump to step 205.

[0108] 204. Obtain a speech/audio decoding parameter of the current speech/audio frame by means of prediction based on an FEC algorithm, and jump to step 205.

[0109] 205. Perform post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and M and N are positive integers.

[0110] 206. Recover a speech/audio signal of the foregoing current speech/audio frame by using the post-processed speech/audio decoding parameter of the foregoing current speech/audio frame.

[0111] Different post processing may be performed on different speech/audio decoding parameters. For example, post processing performed on a spectral pair parameter of the current speech/audio frame may be adaptive weighting performed by using the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, to obtain a post-processed spectral pair parameter of the current speech/audio frame, and post processing performed on an adaptive codebook gain of the current speech/audio frame may be adjustment such as attenuation performed on the adaptive codebook gain.

[0138] It can be learned from the foregoing description that in this embodiment, in a scenario in which a current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, after obtaining a speech/audio decoding parameter of the current speech/audio frame, a decoder performs post processing on the speech/audio

decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and recovers a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame, which ensures stable quality of a decoded signal during transition between a redundant decoded frame and a normal decoded frame or between a redundant decoded frame and an FEC recovered frame, thereby improving quality of an output speech/audio signal.

[0139] It can be learned from the foregoing description that in some embodiments of the present invention, during transition between an unvoiced speech/audio frame and a non-unvoiced speech/audio frame (for example, when a current speech/audio frame is of an unvoiced signal class and is a redundant decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a non unvoiced signal type and is a normal decoded frame, or when a current speech/audio frame is of a non unvoiced signal class and is a normal decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of an unvoiced signal class and is a redundant decoded frame), post processing is performed on a speech/audio decoding parameter of the current speech/audio frame, which helps avoid a click (click) phenomenon caused during the interframe transition between the unvoiced speech/audio frame and the non-unvoiced speech/audio frame, thereby improving quality of an output speech/audio signal.

[0140] In other embodiments of the present invention, during transition between a generic speech/audio frame and a voiced speech/audio frame (when a current speech/audio frame is a generic frame and is a redundant decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a voiced signal class and is a normal decoded frame, or when a current speech/audio frame is of a voiced signal class and is a normal decoded frame, and a speech/audio frame previous or next to the current speech/audio frame is of a generic signal class and is a redundant decoded frame), post processing is performed on a speech/audio decoding parameter of the current speech/audio frame, which helps rectify an energy instability phenomenon caused during the transition between a generic frame and a voiced frame,

thereby improving quality of an output speech/audio signal.

[0141] In still other embodiments of the present invention, when a current speech/audio frame is a redundant decoded frame, a signal class of the current speech/audio frame is not unvoiced, and a signal class of a speech/audio frame next to the current speech/audio frame is unvoiced, a bandwidth extension envelope of the current frame is adjusted, to rectify an energy instability phenomenon in time-domain bandwidth extension, and improve quality of an output speech/audio signal.

[0142] An embodiment of the present invention further provides a related apparatus for implementing the foregoing solution.

[0143] Referring to FIG. 3, an embodiment of the present invention provides a decoder 300 for decoding a speech/audio bitstream, which may include: a parameter acquiring unit 310, a post processing unit 320, and a recovery unit 330.

[0144] The parameter acquiring unit 310 is configured to acquire a speech/audio decoding parameter of a current speech/audio frame, where the foregoing current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame.

[0145] When the speech/audio frame previous to the foregoing current speech/audio frame is a redundant decoded frame, the current speech/audio frame may be a normal decoded frame, a redundant decoded frame, or an FEC recovery frame.

[0146] The post processing unit 320 is configured to perform post processing on the speech/audio decoding parameter of the foregoing current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the foregoing current speech/audio frame, where the foregoing X speech/audio frames include M speech/audio frames previous to the foregoing current speech/audio frame and/or N speech/audio frames next to the foregoing current speech/audio frame, and M and N are positive integers.

[0147] The recovery unit 330 is configured to recover a speech/audio signal of the foregoing current speech/audio frame by using the post-processed speech/audio decoding parameter of the foregoing current speech/audio frame.

[0148] That a speech/audio frame (for example, the current speech/audio frame or the speech/audio frame previous to the current speech/audio frame) is a normal decoded frame means that a speech/audio parameter, and the like of the foregoing

speech/audio frame can be directly obtained from a bitstream of the speech/audio frame by means of decoding. That a speech/audio frame (for example, the current speech/audio frame or the speech/audio frame previous to the current speech/audio frame) is a redundant decoded frame means that a speech/audio parameter, and the like of the speech/audio frame cannot be directly obtained from a bitstream of the speech/audio frame by means of decoding, but redundant bitstream information of the speech/audio frame can be obtained from a bitstream of another speech/audio frame.

[0149] The M speech/audio frames previous to the current speech/audio frame refer to M speech/audio frames preceding the current speech/audio frame and immediately adjacent to the current speech/audio frame in a time domain.

[0150] For example, M may be equal to 1, 2, 3, or another value. When M=1, the M speech/audio frames previous to the current speech/audio frame are the speech/audio frame previous to the current speech/audio frame, and the speech/audio frame previous to the current speech/audio frame and the current speech/audio frame are two immediately adjacent speech/audio frames; when M=2, the M speech/audio frames previous to the current speech/audio frame are the speech/audio frame previous to the current speech/audio frame and a speech/audio frame previous to the speech/audio frame previous to the current speech/audio frame, and the speech/audio frame previous to the current speech/audio frame, the speech/audio frame previous to the speech/audio frame previous to the current speech/audio frame, and the current speech/audio frame are three immediately adjacent speech/audio frames; and so on.

[0151] The N speech/audio frames next to the current speech/audio frame refer to N speech/audio frames following the current speech/audio frame and immediately adjacent to the current speech/audio frame in a time domain.

[0152] For example, N may be equal to 1, 2, 3, 4, or another value. When N=1, the N speech/audio frames next to the current speech/audio frame are a speech/audio frame next to the current speech/audio frame, and the speech/audio frame next to the current speech/audio frame and the current speech/audio frame are two immediately adjacent speech/audio frames; when N=2, the N speech/audio frames next to the current speech/audio frame are a speech/audio frame next to the current speech/audio frame and a speech/audio frame next to the speech/audio frame next to the current speech/audio frame, and the speech/audio frame next to the current speech/audio frame, the speech/audio frame next to the speech/audio frame next to the current speech/audio frame, and the current speech/audio frame are three immediately

adjacent speech/audio frames; and so on.

[0153] The speech/audio decoding parameter may include at least one of the following parameters:

a bandwidth extension envelope, an adaptive codebook gain (gain_pit), an algebraic codebook, a pitch period, a spectrum tilt factor, a spectral pair parameter, and the like.

[0154] The speech/audio parameter may include a speech/audio decoding parameter, a signal class, and the like.

[0155] A signal class of a speech/audio frame may be unvoiced, voiced, generic, transient, inactive, or the like.

[0156] The spectral pair parameter may be, for example, at least one of a line spectral pair (LSP) parameter or an immittance spectral pair (ISP) parameter.

[0157] It may be understood that in this embodiment of the present invention, the post processing unit 320 may perform post processing on at least one speech/audio decoding parameter of a bandwidth extension envelope, an adaptive codebook gain, an algebraic codebook, a pitch period, or a spectral pair parameter of the current speech/audio frame. Specifically, how many parameters are selected and which parameters are selected for post processing may be determined according to an application scenario and an application environment, which is not limited in this embodiment of the present invention.

[0158] The post processing unit 320 may perform different post processing on different speech/audio decoding parameters. For example, post processing performed by the post processing unit 320 on the spectral pair parameter of the current speech/audio frame may be adaptive weighting performed by using the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, to obtain a post-processed spectral pair parameter of the current speech/audio frame, and post processing performed by the post processing unit 320 on the adaptive codebook gain of the current speech/audio frame may be adjustment such as attenuation performed on the adaptive codebook gain.

[0184] It may be understood that functions of function modules of the decoder 300 in this embodiment may be specifically implemented according to the method in the foregoing method embodiment. For a specific implementation process, refer to related descriptions of the foregoing method embodiment. Details are not described

herein. The decoder 300 may be any apparatus that needs to output speeches, for example, a device such as a notebook computer, a tablet computer, or a personal computer, or a mobile phone.

[0189] FIG. 4 is a schematic diagram of a decoder 400 according to an embodiment of the present invention. The decoder 400 may include at least one bus 401, at least one processor 402 connected to the bus 401, and at least one memory 403 connected to the bus 401.

[0190] By invoking, by using the bus 401, code stored in the memory 403, the processor 402 is configured to perform the steps as described in the previous method embodiments, and the specific implementation process of the processor 402 can refer to related descriptions of the foregoing method embodiments. Details are not described herein.

[0199] It may be understood that in this embodiment of the present invention, by invoking the code stored in the memory 403, the processor 402 may be configured to perform post processing on at least one speech/audio decoding parameter of a bandwidth extension envelope, an adaptive codebook gain, an algebraic codebook, a pitch period, or a spectral pair parameter of the current speech/audio frame. Specifically, how many parameters are selected and which parameters are selected for post processing may be determined according to an application scenario and an application environment, which is not limited in this embodiment of the present invention.

[0200] Different post processing may be performed on different speech/audio decoding parameters. For example, post processing performed on the spectral pair parameter of the current speech/audio frame may be adaptive weighting performed by using the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, to obtain a post-processed spectral pair parameter of the current speech/audio frame, and post processing performed on the adaptive codebook gain of the current speech/audio frame may be adjustment such as attenuation performed on the adaptive codebook gain.

[0201] A specific post processing manner is not limited in this embodiment of the present invention, and specific post processing may be set according to a requirement or according to an application environment and an application scenario.

[0233] Referring to FIG. 5, FIG. 5 is a structural block diagram of a decoder 500

according to another embodiment of the present invention. The decoder 500 may include at least one processor 501, at least one network interface 504 or user interface 503, a memory 505, and at least one communications bus 502. The communication bus 502 is configured to implement connection and communication between these components. The decoder 500 may optionally include the user interface 503, which includes a display (for example, a touchscreen, an LCD, a CRT, a holographic device, or a projector (Projector)), a click/tap device (for example, a mouse, a trackball (trackball), a touchpad, or a touchscreen), a camera and/or a pickup apparatus, and the like.

[0234] The memory 505 may include a read-only memory and a random access memory, and provide an instruction and data for the processor 501. A part of the memory 505 may further include a nonvolatile random access memory (NVRAM).

[0235] In some implementation manners, the memory 505 stores the following elements, an executable module or a data structure, or a subset thereof, or an extended set thereof:

an operating system 5051, including various system programs, and used to implement various basic services and process hardware-based tasks; and

an application program module 5052, including various application programs, and configured to implement various application services.

[0236] The application program module 5052 includes but is not limited to a parameter acquiring unit 310, a post processing unit 320, a recovery unit 330, and the like.

In this embodiment of the present invention, by invoking a program or an instruction stored in the memory 505, the processor 501 may be configured to perform the steps as described in the previous method embodiments.

[0246] It may be understood that in this embodiment, by invoking the program or the instruction stored in the memory 505, the processor 501 may perform post processing on at least one speech/audio decoding parameter of a bandwidth extension envelope, an adaptive codebook gain, an algebraic codebook, a pitch period, or a spectral pair parameter of the current speech/audio frame. Specifically, how many parameters are selected and which parameters are selected for post processing may be determined according to an application scenario and an application environment, which is not limited in this embodiment of the present invention.

[0247] Different post processing may be performed on different speech/audio decoding parameters. For example, post processing performed on the spectral pair parameter of the current speech/audio frame may be adaptive weighting performed by

using the spectral pair parameter of the current speech/audio frame and a spectral pair parameter of the speech/audio frame previous to the current speech/audio frame, to obtain a post-processed spectral pair parameter of the current speech/audio frame, and post processing performed on the adaptive codebook gain of the current speech/audio frame may be adjustment such as attenuation performed on the adaptive codebook gain. The specific implementation details about the post processing can refer to related descriptions of the foregoing method embodiments.

[0280] An embodiment of the present invention further provides a computer storage medium, where the computer storage medium may store a program. When being executed, the program includes some or all steps of any speech/audio bitstream decoding method described in the foregoing method embodiments.

[0281] It should be noted that, to make the description brief, the foregoing method embodiments are expressed as a series of actions. However, persons skilled in the art should appreciate that the present invention is not limited to the described action sequence, because according to the present invention, some steps may be performed in other sequences or performed simultaneously.

[0282] In the foregoing embodiments, the description of each embodiment has respective focuses. For a part that is not described in detail in an embodiment, refer to related descriptions in other embodiments.

[0283] In the several embodiments provided in this application, it should be understood that the disclosed apparatus may be implemented in another manner. For example, the described apparatus embodiment is merely exemplary. For example, the unit division is merely logical function division and may be other division in actual implementation. For example, multiple units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented through some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic or other forms.

[0284] The units described as separate parts may or may not be physically separate, and parts displayed as units may or may not be physical units, may be located in one position, or may be distributed on multiple network units. Some or all of the units may be selected according to actual needs to achieve the objectives of the solutions of the embodiments.

[0285] In addition, functional units in the embodiments of the present invention may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit. The integrated unit may be implemented in a form of hardware, or may be implemented in a form of a software functional unit.

[0286] When the integrated unit is implemented in the form of a software functional unit and sold or used as an independent product, the integrated unit may be stored in a computer-readable storage medium. Based on such an understanding, the technical solutions of the present invention essentially, or the part contributing to the prior art, or all or a part of the technical solutions may be implemented in the form of a software product. The software product is stored in a storage medium and includes several instructions for instructing a computer device (which may be a personal computer, a server, or a network device, and may specifically be a processor in a computer device) to perform all or a part of the steps of the foregoing methods described in the embodiments of the present invention. The foregoing storage medium may include: any medium that can store program code, such as a USB flash drive, a magnetic disk, a random access memory (RAM, random access memory), a read-only memory (ROM, read-only memory), a removable hard disk, or an optical disc.

The foregoing embodiments are merely intended for describing the technical solutions of the present invention, but not for limiting the present invention. Although the present invention is described in detail with reference to the foregoing embodiments, persons of ordinary skill in the art should understand that they may still make modifications to the technical solutions described in the foregoing embodiments or make equivalent replacements to some technical features thereof, without departing from the scope of the technical solutions of the embodiments of the present invention.

CLAIMS

1. A speech/audio bitstream decoding method, comprising:

acquiring a speech/audio decoding parameter of a current speech/audio frame, wherein the current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame;

performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame, wherein the X speech/audio frames comprise M speech/audio frames previous to the current speech/audio frame and/or N speech/audio frames next to the current speech/audio frame, and M and N are positive integers; and

recovering a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame.

2. The method according to claim 1, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a spectral pair parameter of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame comprises:

performing post processing on the spectral pair parameter of the current speech/audio frame according to at least one of a signal class, a spectrum tilt factor, an adaptive codebook gain, or a spectral pair parameter of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the current speech/audio frame.

3. The method according to claims 1 or 2, wherein

the speech/audio decoding parameter of the current speech/audio frame comprises an adaptive codebook gain of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame comprises:

performing post processing on the adaptive codebook gain of the current

speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames, to obtain a post-processed adaptive codebook gain of the current speech/audio frame.

4. The method according to any one of claims 1 to 3, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a bandwidth extension envelope of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame comprises: performing post processing on the bandwidth extension envelope of the current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the current speech/audio frame.

5. The method according to claim 4, wherein the performing post processing on the bandwidth extension envelope of the current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames comprises:

when the speech/audio frame previous to the current speech/audio frame is a normal decoded frame, and the signal class of the speech/audio frame previous to the current speech/audio frame is the same as that of the speech/audio frame next to the current speech/audio frame, obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

when the current speech/audio frame is a prediction form of redundancy decoding, obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

when the signal class of the current speech/audio frame is not unvoiced, the signal class of the speech/audio frame next to the current speech/audio frame is unvoiced, and the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a tenth threshold, modifying the bandwidth extension envelope of the current speech/audio frame according to a

bandwidth extension envelope or the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, to obtain the post-processed bandwidth extension envelope of the current speech/audio frame.

6. The method according to claim 5, wherein the obtaining the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame comprises: obtaining the post-processed bandwidth extension envelope of the current speech/audio frame by using the following formula:

$GainFrame = fac1 * GainFrame_old + fac2 * GainFrame_new$, Wherein $GainFrame$ is the post-processed bandwidth extension envelope of the current speech/audio frame, $GainFrame_old$ is the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame, $GainFrame_new$ is the bandwidth extension envelope of the current speech/audio frame, $fac1$ is a weight of the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame, $fac2$ is a weight of the bandwidth extension envelope of the current speech/audio frame, $fac1 \geq 0$, $fac2 \geq 0$, and $fac1 + fac2 = 1$.

7. The method according to claim 5, wherein a modification factor for modifying the bandwidth extension envelope of the current speech/audio frame is inversely proportional to the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, and is proportional to a ratio of the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame to the bandwidth extension envelope of the current speech/audio frame.

8. The method according to any one of claims 1 to 7, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a pitch period of the current speech/audio frame, and the performing post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame comprises: performing post processing on the pitch period of the current speech/audio frame according to the signal classes and/or pitch periods of the X speech/audio frames, to obtain a post-processed pitch period of the current speech/audio frame.

9. A decoder for decoding a speech/audio bitstream, comprising:

a parameter acquiring unit, configured to acquire a speech/audio decoding parameter of a current speech/audio frame, wherein the current speech/audio frame is a redundant decoded frame or a speech/audio frame previous to the current speech/audio frame is a redundant decoded frame;

a post processing unit, configured to perform post processing on the speech/audio decoding parameter of the current speech/audio frame according to speech/audio parameters of X speech/audio frames, to obtain a post-processed speech/audio decoding parameter of the current speech/audio frame, wherein the X speech/audio frames comprise M speech/audio frames previous to the current speech/audio frame and/or N speech/audio frames next to the current speech/audio frame, and M and N are positive integers; and

a recovery unit, configured to recover a speech/audio signal of the current speech/audio frame by using the post-processed speech/audio decoding parameter of the current speech/audio frame.

10. The decoder according to claim 9, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a spectral pair parameter of the current speech/audio frame; and

the post processing unit is configured to: perform post processing on the spectral pair parameter of the current speech/audio frame according to at least one of a spectral pair parameter, an adaptive codebook gain, a spectrum tilt factor, or a signal class of the X speech/audio frames, to obtain a post-processed spectral pair parameter of the current speech/audio frame.

11. The decoder according to claims 9 or 10, wherein the speech/audio decoding parameter of the current speech/audio frame comprises an adaptive codebook gain of the current speech/audio frame;

the post processing unit is configured to: perform post processing on the adaptive codebook gain of the current speech/audio frame according to at least one of the signal class, an algebraic codebook gain, or the adaptive codebook gain of the X speech/audio frames, to obtain a post-processed adaptive codebook gain of the current speech/audio frame.

12. The decoder according to any one of claims 9 to 11, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a bandwidth extension envelope of the current speech/audio frame, and the post processing unit is configured to: perform post processing on the bandwidth extension

envelope of the current speech/audio frame according to at least one of the signal class, a bandwidth extension envelope, or the spectrum tilt factor of the X speech/audio frames, to obtain a post-processed bandwidth extension envelope of the current speech/audio frame.

13. The decoder according to claim 12, wherein in the aspect of performing post processing on the bandwidth extension envelope of the current speech/audio frame, the post processing unit is specifically configured to:

when the speech/audio frame previous to the current speech/audio frame is a normal decoded frame, and the signal class of the speech/audio frame previous to the current speech/audio frame is the same as that of the speech/audio frame next to the current speech/audio frame, obtain the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

when the current speech/audio frame is a prediction form of redundancy decoding, obtain the post-processed bandwidth extension envelope of the current speech/audio frame based on a bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame and the bandwidth extension envelope of the current speech/audio frame; or

when the signal class of the current speech/audio frame is not unvoiced, the signal class of the speech/audio frame next to the current speech/audio frame is unvoiced, and the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame is less than or equal to a tenth threshold, modify the bandwidth extension envelope of the current speech/audio frame according to a bandwidth extension envelope or the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, to obtain the post-processed bandwidth extension envelope of the current speech/audio frame.

14. The decoder according to claim 13, wherein a modification factor used by the post processing unit for modifying the bandwidth extension envelope of the current speech/audio frame is inversely proportional to the spectrum tilt factor of the speech/audio frame previous to the current speech/audio frame, and is proportional to a ratio of the bandwidth extension envelope of the speech/audio frame previous to the current speech/audio frame to the bandwidth extension envelope of the current speech/audio frame.

15. The decoder according to any one of claims 9 to 14, wherein the speech/audio decoding parameter of the current speech/audio frame comprises a pitch period of the current speech/audio frame, and the post processing unit is specifically configured to: perform post processing on the pitch period of the current speech/audio frame according to at least one of the signal class or a pitch period of the X speech/audio frames, to obtain a post-processed pitch period of the current speech/audio frame.