(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 4 398 242 A1

(12) EUROPEAN PATENT APPLICATION
published in accordance with Art. 153(4) EPC

(43) Date of publication:
10.07.2024 Bulletin 2024/28

(21) Application number: **22874757.2**

(22) Date of filing: **22.09.2022**

(51) International Patent Classification (IPC):
G10L 19/008 *(2013.01)*

(52) Cooperative Patent Classification (CPC):
G10L 19/008

(86) International application number:
PCT/CN2022/120507

(87) International publication number:
WO 2023/051370 (06.04.2023 Gazette 2023/14)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(30) Priority: **29.09.2021 CN 202111155355**

(71) Applicant: **Huawei Technologies Co., Ltd.**
Shenzhen, Guangdong 518129 (CN)

(72) Inventors:
- **LIU, Shuai**
  Shenzhen, Guangdong 518129 (CN)
- **GAO, Yuan**
  Shenzhen, Guangdong 518129 (CN)
- **WANG, Bin**
  Shenzhen, Guangdong 518129 (CN)
- **WANG, Zhe**
  Shenzhen, Guangdong 518129 (CN)

(74) Representative: **Pfenning, Meinig & Partner mbB**
Patent- und Rechtsanwälte
Theresienhöhe 11a
80339 München (DE)

(54) ENCODING AND DECODING METHODS AND APPARATUS, DEVICE, STORAGE MEDIUM, AND COMPUTER PROGRAM

(57) Embodiments of this application disclose encoding and decoding methods and apparatuses, devices, a storage medium, and a computer program, and belong to the field of three-dimensional audio encoding and decoding technologies. The method includes: separately performing transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results; determining a global transient state detection result based on the M transient state detection results; converting the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result; performing spatial encoding on the frequency-domain three-dimensional audio signal to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels; encoding the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result; encoding the spatial encoding parameter to obtain a spatial encoding parameter encoding result; and writing the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bitstream. This can reduce encoding complexity and improve encoding efficiency.

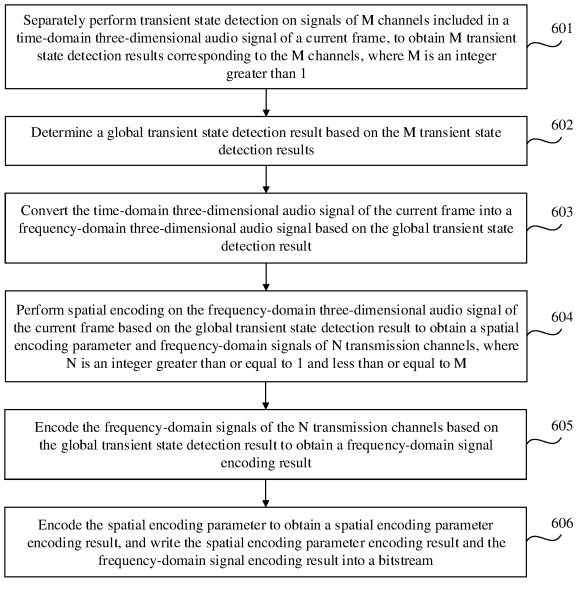

FIG. 6

EP 4 398 242 A1

Description

[0001] This application claims priority to Chinese Patent Application No. 202111155355.4, filed on September 29, 2021, and is hereby incorporated by reference in it entirety.

TECHNICAL FIELD

[0002] Embodiments of this application relate to the field of three-dimensional audio encoding and decoding technologies, and in particular, to encoding and decoding methods and apparatuses, devices, a storage medium, and a computer program.

BACKGROUND

[0003] A three-dimensional audio technology is an audio technology that obtains, processes, transmits, renders, and plays back sound events and three-dimensional sound field information in the real world through computer and signal processing. To achieve better audio auditory effect, a three-dimensional audio signal usually needs to include a large amount of data, to record spatial information of a sound scenario in more detail. However, it is difficult to transmit and store the large amount of data. Therefore, the three-dimensional audio signal needs to be encoded and decoded.

[0004] As a three-dimensional audio technology, a higher order ambisonics (higher order ambisonics, HOA) audio technology has a property irrelevant to a speaker layout in recording, encoding, and playback phases, and data in an HOA format has a feature of rotatable playback. Therefore, an HOA signal has higher flexibility during playback, and attracts more attention.

[0005] A related technology proposes a method for encoding an HOA signal. In the method, time-frequency transform is first performed on a time-domain HOA signal to obtain a frequency-domain HOA signal, and spatial encoding is performed on the frequency-domain HOA signal to obtain frequency-domain signals of a plurality of channels. Then, time-frequency inverse transform is performed on a frequency-domain signal of each channel to obtain a time-domain signal of each channel, and transient state detection is performed on the time-domain signal of each channel to obtain a transient state detection result of each channel. Then, time-frequency transform is performed again on the time-domain signal of each channel to obtain the frequency-domain signal of each channel, and the frequency-domain signal of each channel is encoded based on the transient state detection result of each channel.

[0006] However, in the foregoing method, an audio signal needs to be transformed between a time domain and a frequency domain for a plurality of times. This increases encoding complexity and further reduces encoding efficiency.

SUMMARY

[0007] Embodiments of this application provide encoding and decoding methods and apparatuses, devices, a storage medium, and a computer program, to reduce encoding complexity and improve encoding efficiency. The technical solutions are as follows.

[0008] According to a first aspect, an encoding method is provided, including: separately performing transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1; determining a global transient state detection result based on the M transient state detection results; converting the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result; performing spatial encoding on the frequency-domain three-dimensional audio signal based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M; encoding the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result; encoding the spatial encoding parameter to obtain a spatial encoding parameter encoding result; and writing the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bistream.

[0009] The transient state detection result includes a transient state flag, or the transient state detection result includes a transient state flag and transient state position information. The transient state flag indicates whether a signal of a corresponding channel is a transient state signal. The transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel. The M transient state detection results corresponding to the M channels may be determined in a plurality of manners. The following describes one of the manners. Because the transient state detection result corresponding to each of the M channels is determined in a same manner, the following uses one of the channels as an example to describe a manner of determining the transient state detection result corresponding to the channel. For ease of description, the channel is referred to as a target channel, and the following separately describes a transient state flag and transient state position information of the target channel.

[0010] A manner of determining the transient state flag of the target channel includes: determining, based on a signal of the target channel, a transient state detection parameter corresponding to the target channel; and determining, based on the transient state detection parameter corresponding to the target channel, the transient state flag corresponding to the target channel.

[0011] In an example, the transient state detection pa-

rameter corresponding to the target channel is an absolute value of an inter-frame energy difference. To be specific, energy of the signal of the target channel in the current frame and energy of a signal of a target channel in a previous frame relative to the current frame are determined, and an absolute value of a difference between the energy of the signal of the target channel in the current frame and the energy of the signal of the target channel in the previous frame is determined to obtain the absolute value of the inter-frame energy difference. If the absolute value of the inter-frame energy difference exceeds a first energy difference threshold, it is determined that the transient state flag corresponding to the target channel in the current frame is a first value. Otherwise, it is determined that the transient state flag corresponding to the target channel in the current frame is a second value.

[0012] In another example, the transient state detection parameter corresponding to the target channel is an absolute value of a subframe energy difference. To be specific, the signal of the target channel in the current frame includes signals of a plurality of subframes, and the absolute value of the subframe energy difference corresponding to each of the plurality of subframes is determined to determine a transient state flag corresponding to each subframe. If a subframe whose transient state flag is the first value exists in the plurality of subframes, it is determined that the transient state flag corresponding to the target channel in the current frame is the first value. If no subframe whose transient state flag is the first value exists in the plurality of subframes, it is determined that the transient state flag corresponding to the target channel in the current frame is the second value.

[0013] A manner of determining the transient state position information of the target channel includes: determining, based on the transient state flag corresponding to the target channel, the transient state position information corresponding to the target channel.

[0014] In an example, if the transient state flag corresponding to the target channel is the first value, the transient state position information corresponding to the target channel is determined. If the transient state flag corresponding to the target channel is the second value, it is determined that the target channel does not have corresponding transient state position information, or the transient state position information corresponding to the target channel is set to a preset value, for example, set to -1.

[0015] In some embodiments, the transient state detection result includes the transient state flag. The global transient state detection result includes a global transient state flag. The transient state flag indicates whether a signal of a corresponding channel is a transient state signal. The determining a global transient state detection result based on the M transient state detection results includes: if a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to m, determining that the global transient state flag is the first value, where m is a positive integer greater than 0 and less than M; or if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M channels is greater than or equal to n, determining that the global transient state flag is the first value, where n is a positive integer greater than 0 and less than M.

[0016] Herein, m and n are preset values, and m and n can also be adjusted based on different requirements. If the three-dimensional audio signal is an HOA signal, the first preset condition includes belonging to channels of a first-order ambisonics (first-order ambisonics, FOA) signal. For example, the channels of the FOA signal may include first four channels in the HOA signal. In other words, if the three-dimensional audio signal is the HOA signal, if a quantity of channels whose corresponding transient state flags are the first value in the channels of the FOA signal in the three-dimensional audio signal of the current frame is greater than or equal to n, it is determined that the global transient state flag is the first value. Certainly, the first preset condition may alternatively be another condition.

[0017] In some other embodiments, the transient state detection result further includes transient state position information. The global transient state detection result further includes global transient state position information. The transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel. The determining a global transient state detection result based on the M transient state detection results includes: if only one transient state flag in the M transient state flags is the first value, determining transient state position information corresponding to a channel whose transient state flag is the first value as the global transient state position information; or if at least two transient state flags in the M transient state flags are the first value, determining transient state position information, as the global transient state position information, corresponding to a channel with a largest transient state detection parameter in at least two channels corresponding to the at least two transient state flags.

[0018] Alternatively, if at least two transient state flags in the M transient state flags are the first value, and a difference between transient state position information corresponding to two channels is less than a position difference threshold, an average value of the transient state position information corresponding to the two channels is determined as the global transient state position information. The position difference threshold is preset and can be adjusted based on different requirements.

[0019] Based on the foregoing descriptions, a transient state detection parameter corresponding to a channel is an absolute value of an inter-frame energy difference or an absolute value of a subframe energy difference. If the transient state detection parameter corresponding to the channel is the absolute value of the inter-frame energy difference, one channel corresponds to one absolute value of the inter-frame energy difference. In this case, a

channel corresponding to a largest absolute value of the inter-frame energy difference may be selected from the at least two channels, and then transient state position information corresponding to the selected channel is determined as the global transient state position information. If the transient state detection parameter corresponding to the channel is the absolute value of the subframe energy difference, one channel corresponds to a plurality of absolute values of the subframe energy difference. In this case, a channel corresponding to a largest absolute value of the subframe energy difference may be selected from the at least two channels, and then transient state position information corresponding to the selected channel is determined as the global transient state position information.

[0020] Optionally, the converting the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result includes: determining a target encoding parameter based on the global transient state detection result, where the target encoding parameter includes a window function type of the current frame and/or a frame type of the current frame; and converting the time-domain three-dimensional audio signal into the frequency-domain three-dimensional audio signal based on the target encoding parameter.

[0021] In an example, the global transient state detection result includes the global transient state flag. An implementation process of determining the window function type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining a type of a first preset window function as the window function type of the current frame; or if the global transient state flag is the second value, determining a type of a second preset window function as the window function type of the current frame. A window length of the first preset window function is less than a window length of the second preset window function.

[0022] In another example, the global transient state detection result includes the global transient state flag and the global transient state position information. An implementation process of determining the window function type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining the window function type of the current frame based on the global transient state position information; or if the global transient state flag is the second value, determining a type of a third preset window function as the window function type of the current frame, or determining the window function type of the current frame based on a window function type of a previous frame relative to the current frame.

[0023] The global transient state detection result may include only the global transient state flag, or may include the global transient state flag and the global transient state position information. In addition, the global transient state position information may be transient state position information corresponding to a channel whose transient state flag is the first value, or may be a preset value. If global transient state detection results are different, the frame type of the current frame is determined in different manners. Therefore, the following separately describes the following three cases.

[0024] In a first case, the global transient state detection result includes the global transient state flag. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames; or if the global transient state flag is the second value, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame.

[0025] In a second case, the global transient state detection result includes the global transient state flag and the global transient state position information. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value and the global transient state position information meets a second preset condition, determining that the frame type of the current frame is a third type, where the third type indicates that the current frame includes a plurality of ultra-short frames; if the global transient state flag is the first value and the global transient state position information does not meet the second preset condition, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames; or if the global transient state flag is the second value, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame. A frame length of an ultra-short frame is less than a frame length of a short frame. A frame length of a short frame is less than a frame length of a long frame. The second preset condition may be that a distance between a transient state occurrence position indicated by the global transient state position information and a start position of the current frame is less than the frame length of the ultra-short frame, or a distance between the transient state occurrence position indicated by the global transient state position information and an end position of the current frame is less than the frame length of the ultra-short frame.

[0026] In a third case, the global transient state detection result includes the global transient state position information. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state position information is a preset value, for example, -1, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame; if the global tran-

sient state position information is not the preset value and meets a second preset condition, determining that the frame type of the current frame is a third type, where the third type indicates that the current frame includes a plurality of ultra-short frames; or if the global transient state position information is not the preset value and does not meet the second preset condition, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames. A frame length of an ultra-short frame is less than a frame length of a short frame. A frame length of a short frame is less than a frame length of a long frame. The second preset condition may be that a distance between a transient state occurrence position indicated by the global transient state position information and a start position of the current frame is less than the frame length of the ultra-short frame, or a distance between the transient state occurrence position indicated by the global transient state position information and an end position of the current frame is less than the frame length of the ultra-short frame.

[0027] It should be noted that the window function type of the current frame indicates a shape and a length of a window function corresponding to the current frame. The window function of the current frame is used for performing windowing processing on the time-domain three-dimensional audio signal of the current frame. The frame type of the current frame indicates whether the current frame is an ultra-short frame, a short frame, or a long frame. The ultra-short frame, the short frame, and the long frame may be distinguished based on duration of the frames. The specific duration may be set based on different requirements. This is not limited in this embodiment of this application.

[0028] Based on the foregoing descriptions, the target encoding parameter includes the window function type of the current frame and/or the frame type of the current frame. To be specific, the target encoding parameter includes the window function type of the current frame, or the target encoding parameter includes the frame type of the current frame, or the target encoding parameter includes the window function type and the frame type of the current frame. When parameters included in the target encoding parameter are different, processes of converting the time-domain three-dimensional audio signal of the current frame into the frequency-domain three-dimensional audio signal based on the target encoding parameter are different. Therefore, descriptions are separately provided below.

[0029] In a first case, the target encoding parameter includes the window function type of the current frame. In this case, windowing processing is performed on the time-domain three-dimensional audio signal of the current frame based on the window function indicated by the window function type of the current frame. Then, a three-dimensional audio signal obtained through the windowing processing is converted into the frequency-domain three-dimensional audio signal.

[0030] In a second case, the target encoding parameter includes the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, a time-domain three-dimensional audio signal of each short frame included in the current frame is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, a time-domain three-dimensional audio signal of the long frame included in the current frame is directly converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, a time-domain three-dimensional audio signal of each ultra-short frame included in the current frame is converted into a frequency-domain three-dimensional audio signal.

[0031] In a third case, the target encoding parameter includes the window function type and the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of each short frame included in the current frame, and a time-domain three-dimensional audio signal of each short frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of the long frame included in the current frame, and a time-domain three-dimensional audio signal of the long frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of each ultra-short frame included in the current frame, and a time-domain three-dimensional audio signal of each ultra-short frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal.

[0032] In some embodiments, the target encoding parameter may be further encoded to obtain a target encoding parameter encoding result. The target encoding parameter encoding result is written into the bistream.

[0033] In some embodiments, the performing spatial encoding on the frequency-domain three-dimensional

audio signal based on the global transient state detection result includes: performing spatial encoding on the frequency-domain three-dimensional audio signal based on the frame type.

[0034] When spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame based on the frame type of the current frame, if the frame type of the current frame is the first type, namely if the current frame includes a plurality of short frames, frequency-domain three-dimensional audio signals of the plurality of short frames included in the current frame are interleaved to obtain a frequency-domain three-dimensional audio signal of a long frame, and spatial encoding is performed on the frequency-domain three-dimensional audio signal of the long frame obtained through the interleaving. If the frame type of the current frame is the second type, namely if the current frame includes a long frame, spatial encoding is performed on a frequency-domain three-dimensional audio signal of the long frame. If the frame type of the current frame is the third type, namely if the current frame includes a plurality of ultra-short frames, frequency-domain three-dimensional audio signals of the plurality of ultra-short frames included in the current frame are interleaved to obtain a frequency-domain three-dimensional audio signal of a long frame, and spatial encoding is performed on the frequency-domain three-dimensional audio signal of the long frame obtained through the interleaving.

[0035] In some embodiments, the encoding the frequency-domain signals of the N transmission channels based on the global transient state detection result includes: encoding the frequency-domain signals of the N transmission channels based on the frame type of the current frame.

[0036] In an example, an implementation process of encoding the frequency-domain signals of the N transmission channels includes: performing noise shaping processing on the frequency-domain signals of the N transmission channels based on the frame type of the current frame; performing transmission channel downmixing processing on frequency-domain signals of the N transmission channels obtained through the noise shaping processing, to obtain a downmixed signal; performing quantization and encoding processing on a low-frequency part of the downmixed signal, and writing an encoding result into the bistream; and performing bandwidth expansion and encoding processing on a high-frequency part of the downmixed signal, and writing an encoding result into the bistream.

[0037] Optionally, the method further includes: encoding the global transient state detection result to obtain a global transient state detection result encoding result; and writing the global transient state detection result encoding result into the bistream.

[0038] According to a second aspect, a decoding method is provided, including: parsing a bitstream to obtain a global transient state detection result and a spatial encoding parameter; performing decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels; performing spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal; and determining a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

[0039] Optionally, the determining a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal includes: determining a target encoding parameter based on the global transient state detection result, where the target encoding parameter includes a window function type of a current frame and/or a frame type of the current frame; and converting the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter.

[0040] Based on the foregoing descriptions, the target encoding parameter includes the window function type of the current frame and/or the frame type of the current frame. To be specific, the target encoding parameter includes the window function type of the current frame, or the target encoding parameter includes the frame type of the current frame, or the target encoding parameter includes the window function type and the frame type of the current frame. When parameters included in the target encoding parameter are different, processes of converting the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter are different. Therefore, descriptions are separately provided below.

[0041] In a first case, the target encoding parameter includes the window function type of the current frame. In this case, windowing removal processing is performed on the reconstructed frequency-domain three-dimensional audio signal based on a window function indicated by the window function type of the current frame. Then, a frequency-domain three-dimensional audio signal obtained through the windowing removal processing is converted into the reconstructed time-domain three-dimensional audio signal.

[0042] The windowing removal processing is also referred to as windowing and overlap-add processing.

[0043] In a second case, the target encoding parameter includes the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, a reconstructed frequency-domain three-dimensional audio signal of each short frame is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-

dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, a reconstructed frequency-domain three-dimensional audio signal of the long frame included in the current frame is directly converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, a reconstructed frequency-domain three-dimensional audio signal of each ultra-short frame is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal.

[0044] In a third case, the target encoding parameter includes the window function type and the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a frequency-domain three-dimensional audio signal of each short frame included in the current frame, and a reconstructed frequency-domain three-dimensional audio signal of each short frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a reconstructed frequency-domain three-dimensional audio signal of the long frame included in the current frame, and a frequency-domain three-dimensional audio signal of the long frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a frequency-domain three-dimensional audio signal of each ultra-short frame included in the current frame, and a reconstructed frequency-domain three-dimensional audio signal of each ultra-short frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal.

[0045] Optionally, the global transient state detection result includes a global transient state flag. The target encoding parameter includes the window function type of the current frame. The determining a target encoding parameter based on the global transient state detection result includes: if the global transient state flag is the first value, determining a type of a first preset window function as the window function type of the current frame; or if the global transient state flag is the second value, determining a type of a second preset window function as the window function type of the current frame. A window length of the first preset window function is less than a window length of the second preset window function.

[0046] Optionally, the global transient state detection result includes a global transient state flag and global transient state position information. The target encoding parameter includes the window function type of the current frame. The determining a target encoding parameter based on the global transient state detection result includes: if the global transient state flag is the first value, determining the window function type of the current frame based on the global transient state position information.

[0047] According to a third aspect, an encoding apparatus is provided. The encoding apparatus has a function of implementing behavior of the encoding method in the first aspect. The encoding apparatus includes at least one module. The at least one module is configured to implement the encoding method provided in the first aspect.

[0048] According to a fourth aspect, a decoding apparatus is provided. The decoding apparatus has a function of implementing behavior of the decoding method in the second aspect. The decoding apparatus includes at least one module. The at least one module is configured to implement the decoding method provided in the second aspect.

[0049] According to a fifth aspect, an encoder side device is provided. The encoder side device includes a processor and a memory. The memory is configured to store a program for performing the encoding method provided in the first aspect. The processor is configured to execute the program stored in the memory, to implement the encoding method provided in the first aspect.

[0050] Optionally, the encoder side device may further include a communication bus. The communication bus is configured to establish a connection between the processor and the memory.

[0051] According to a sixth aspect, a decoder side device is provided. The decoder side device includes a processor and a memory. The memory is configured to store a program for performing the decoding method provided in the second aspect. The processor is configured to execute the program stored in the memory, to implement the decoding method provided in the second aspect.

[0052] Optionally, the decoder side device may further include a communication bus. The communication bus is configured to establish a connection between the processor and the memory.

[0053] According to a seventh aspect, a computer-readable storage medium is provided. The storage medium stores instructions. When the instructions run on a computer, the computer is enabled to perform the steps of the encoding method according to the first aspect or

the steps of the decoding method according to the second aspect.

[0054] According to an eighth aspect, a computer program product including instructions is provided. When the instructions run on a computer, the computer is enabled to perform the steps of the encoding method according to the first aspect or the steps of the decoding method according to the second aspect. Alternatively, a computer program is provided. When the computer program is executed, the steps of the encoding method according to the first aspect or the steps of the decoding method according to the second aspect are implemented.

[0055] According to a ninth aspect, a computer-readable storage medium is provided. The computer-readable storage medium includes a bistream obtained by using the encoding method according to the first aspect.

[0056] Technical effects obtained in the third aspect, the fourth aspect, the fifth aspect, the sixth aspect, the seventh aspect, the eighth aspect, and the ninth aspect are similar to technical effects obtained through corresponding technical means in the first aspect or the second aspect. Details are not described herein again.

[0057] The technical solutions provided in embodiments of this application can bring at least the following beneficial effects:

[0058] Transient state detection is performed on the signals of the M channels included in the time-domain three-dimensional audio signal of the current frame, to determine the global transient state detection result. Then, based on the global transient state detection result, time-frequency transform and spatial encoding of the audio signal are sequentially performed, and the frequency-domain signal of each transmission channel is encoded. Especially, when the frequency-domain signal of each transmission channel obtained through the spatial encoding is encoded, the global transient state detection result is used for instructing the encoding of the frequency-domain signal of each transmission channel, and the frequency-domain signal of each transmission channel does not need to be converted into a time domain to determine the transient state detection result corresponding to each transmission channel. Therefore, the three-dimensional audio signal does not need to be transformed between a time domain and a frequency domain for a plurality of times. This can reduce encoding complexity and improve encoding efficiency.

BRIEF DESCRIPTION OF DRAWINGS

[0059]

FIG. 1 is a schematic diagram of an implementation environment according to an embodiment of this application;

FIG. 2 is a schematic diagram of an implementation environment of a terminal scenario according to an embodiment of this application;

FIG. 3 is a schematic diagram of an implementation environment of a transcoding scenario of a wireless or core network device according to an embodiment of this application;

FIG. 4 is a schematic diagram of an implementation environment of a broadcast television scenario according to an embodiment of this application;

FIG. 5 is a schematic diagram of an implementation environment of a virtual reality flow scenario according to an embodiment of this application;

FIG. 6 is a flowchart of a first encoding method according to an embodiment of this application;

FIG. 7 is a first example block diagram of the encoding method shown in FIG. 6 according to an embodiment of this application;

FIG. 8 is a second example block diagram of the encoding method shown in FIG. 6 according to an embodiment of this application;

FIG. 9 is a flowchart of a first decoding method according to an embodiment of this application;

FIG. 10 is a first example block diagram of the decoding method shown in FIG. 9 according to an embodiment of this application;

FIG. 11 is a flowchart of a second encoding method according to an embodiment of this application;

FIG. 12 is a first example block diagram of the encoding method shown in FIG. 11 according to an embodiment of this application;

FIG. 13 is a second example block diagram of the encoding method shown in FIG. 11 according to an embodiment of this application;

FIG. 14 is a flowchart of a second decoding method according to an embodiment of this application;

FIG. 15 is an example block diagram of the decoding method shown in FIG. 14 according to an embodiment of this application;

FIG. 16 is a schematic diagram of a structure of an encoding apparatus according to an embodiment of this application;

FIG. 17 is a schematic diagram of a structure of a decoding apparatus according to an embodiment of this application; and

FIG. 18 is a schematic block diagram of an encoding and decoding apparatus according to an embodiment of this application.

DESCRIPTION OF EMBODIMENTS

[0060] To make the objectives, technical solutions, and advantages of this application clearer, the following further describes the implementations of embodiments of this application in detail with reference to the accompanying drawings.

[0061] Before the encoding and decoding methods provided in embodiments of this application are explained and described in detail, terms and implementation environments in embodiments of this application are first described.

[0062] For ease of understanding, terms in embodiments of this application are first explained.

[0063] **Encoding:** a process of compressing a to-be-encoded audio signal into a bistream. It should be noted that, after the audio signal is compressed into the bistream, the audio signal may be referred to as an encoded audio signal or a compressed audio signal.

[0064] **Decoding:** a process of restoring an encoded bistream to a reconstructed audio signal by using a specific syntax rule and processing method.

[0065] **Three-dimensional audio signal:** includes signals of a plurality of channels, represents a sound field in three-dimensional space, and may be a combination of one or more of an HOA signal, a multi-channel signal, and an object audio signal. For the HOA signal, a quantity of channels of the three-dimensional audio signal is related to an order of the three-dimensional audio signal. For example, if the three-dimensional audio signal is an A-order signal, the quantity of the channels of the three-dimensional audio signal is $(A+1)^2$.

[0066] The three-dimensional audio signal mentioned below may be any three-dimensional audio signal, for example, may be a combination of one or more of an HOA signal, a multi-channel signal, and an object audio signal.

[0067] **Transient state signal:** represents a transient state phenomenon of a signal of a channel corresponding to a three-dimensional audio signal. If a signal of a channel is a transient state signal, it indicates that the signal of the channel is a non-stationary signal, for example, a signal whose energy greatly changes in a short time, such as a drum sound or a sound of a percussion instrument.

[0068] The following describes an implementation environment in embodiments of this application.

[0069] FIG. 1 is a schematic diagram of an implementation environment according to an embodiment of this application. The implementation environment includes a source apparatus 10, a destination apparatus 20, a link 30, and a storage apparatus 40. The source apparatus 10 may generate an encoded three-dimensional audio signal. Therefore, the source apparatus 10 may also be referred to as a three-dimensional audio signal encoding apparatus. The destination apparatus 20 may decode the encoded three-dimensional audio signal generated by the source apparatus 10. Therefore, the destination apparatus 20 may also be referred to as a three-dimensional audio signal decoding apparatus. The link 30 may receive the encoded three-dimensional audio signal generated by the source apparatus 10, and may transmit the encoded three-dimensional audio signal to the destination apparatus 20. The storage apparatus 40 may receive the encoded three-dimensional audio signal generated by the source apparatus 10, and may store the encoded three-dimensional audio signal. In this case, the destination apparatus 20 may directly obtain the encoded three-dimensional audio signal from the storage apparatus 40. Alternatively, the storage apparatus 40 may correspond to a file server or another intermediate storage apparatus that may store the encoded three-dimensional audio signal generated by the source apparatus 10. In this case, the destination apparatus 20 may stream or download the encoded three-dimensional audio signal stored in the storage apparatus 40.

[0070] Both the source apparatus 10 and the destination apparatus 20 may include one or more processors and a memory coupled to the one or more processors. The memory may include a random access memory (random access memory, RAM), a read-only memory (read-only memory, ROM), an electrically erasable programmable read-only memory (electrically erasable programmable read-only memory, EEPROM), a flash memory, any other medium that may be configured to store required program code in a form of instructions or a data structure accessible to a computer, or the like. For example, both the source apparatus 10 and the destination apparatus 20 may include a desktop computer, a mobile computing apparatus, a notebook (for example, a laptop) computer, a tablet computer, a set-top box, a telephone handset such as a so-called "smart" phone, a television set, a camera, a display apparatus, a digital media player, a video game console, an in-vehicle computer, or the like.

[0071] The link 30 may include one or more media or apparatuses that can transmit the encoded three-dimensional audio signal from the source apparatus 10 to the destination apparatus 20. In a possible implementation, the link 30 may include one or more communication media that can enable the source apparatus 10 to directly send the encoded three-dimensional audio signal to the destination apparatus 20 in real time. In this embodiment of this application, the source apparatus 10 may modulate the encoded three-dimensional audio signal based on a communication standard. The communication standard may be a wireless communication protocol or the like, and modulated three-dimensional audio signal may be sent to the destination apparatus 20. The one or more communication media may include wireless and/or wired communication media. For example, the one or more communication media may include a radio frequency (radio frequency, RF) spectrum or one or more physical transmission lines. The one or more communication media may form a part of a packet-based network. The packet-based network may be a local area network, a wide area network, a global network (for example, Internet), or the like. The one or more communication media may include a router, a switch, a base station, another device that facilitates communication from the source apparatus 10 to the destination apparatus 20, or the like. This is not specifically limited in this embodiment of this application.

[0072] In a possible implementation, the storage apparatus 40 may store the received encoded three-dimensional audio signal sent by the source apparatus 10. The destination apparatus 20 may directly obtain the encoded three-dimensional audio signal from the storage apparatus 40. In this case, the storage apparatus 40 may include any one of a plurality of distributed or locally accessed

data storage media. For example, any one of the plurality of distributed or locally accessed data storage media may be a hard disk drive, a Blu-ray disc, a digital versatile disc (digital versatile disc, DVD), a compact disc read-only memory (compact disc read-only memory, CD-ROM), a flash memory, a volatile or non-volatile memory, or any other suitable digital storage medium configured to store the encoded three-dimensional audio signal.

[0073] In a possible implementation, the storage apparatus 40 may correspond to a file server or another intermediate storage apparatus that may store the encoded three-dimensional audio signal generated by the source apparatus 10. The destination apparatus 20 may stream or download the three-dimensional audio signal stored in the storage apparatus 40. The file server may be any type of server that can store the encoded three-dimensional audio signal and send the encoded three-dimensional audio signal to the destination apparatus 20. In a possible implementation, the file server may include a network server, a file transfer protocol (file transfer protocol, FTP) server, a network attached storage (network attached storage, NAS) apparatus, a local disk drive, or the like. The destination apparatus 20 may obtain the encoded three-dimensional audio signal through any standard data connection (including an Internet connection). The any standard data connection may include a wireless channel (for example, a Wi-Fi connection), a wired connection (for example, a digital subscriber line (digital subscriber line, DSL) or a cable modem), or a combination of the wireless channel and the wired connection suitable for obtaining the encoded three-dimensional audio signal stored on the file server. Transmission of the encoded three-dimensional audio signal from the storage apparatus 40 may be streaming transmission, download transmission, or a combination thereof.

[0074] The technology in this embodiment of this application may be applicable to the source apparatus 10 that encodes the three-dimensional audio signal and that is shown in FIG. 1, and may be further applicable to the destination apparatus 20 that decodes the encoded three-dimensional audio signal.

[0075] In the implementation environment shown in FIG. 1, the source apparatus 10 includes a data source 120, an encoder 100, and an output interface 140. In some embodiments, the output interface 140 may include a modulator/demodulator (modem) and/or a sender. The sender may also be referred to as a transmitter. The data source 120 may include an image capture apparatus (for example, a camera), an archive containing a previously captured three-dimensional audio signal, a feed-in interface for receiving the three-dimensional audio signal from a three-dimensional audio signal content provider, and/or a computer graphics system for generating the three-dimensional audio signal, or a combination of these sources of the three-dimensional audio signal.

[0076] The data source 120 may send a three-dimensional audio signal to the encoder 100. The encoder 100 may encode the received three-dimensional audio signal sent by the data source 120, to obtain an encoded three-dimensional audio signal. The encoder may send the encoded three-dimensional audio signal to the output interface. In some embodiments, the source apparatus 10 directly sends the encoded three-dimensional audio signal to the destination apparatus 20 through the output interface 140. In another embodiment, the encoded three-dimensional audio signal may be further stored on the storage apparatus 40 for the destination apparatus 20 to obtain subsequently for decoding and/or displaying.

[0077] In the implementation environment shown in FIG. 1, the destination apparatus 20 includes an input interface 240, a decoder 200, and a display apparatus 220. In some embodiments, the input interface 240 includes a receiver and/or a modem. The input interface 240 may receive an encoded three-dimensional audio signal through the link 30 and/or from the storage apparatus 40, and then send the encoded three-dimensional audio signal to the decoder 200. The decoder 200 may decode the received encoded three-dimensional audio signal to obtain a decoded three-dimensional audio signal. The decoder may send the decoded three-dimensional audio signal to the display apparatus 220. The display apparatus 220 may be integrated with the destination apparatus 20 or may be external to the destination apparatus 20. Generally, the display apparatus 220 displays the decoded three-dimensional audio signal. The display apparatus 220 may be a display apparatus of any one of a plurality of types. For example, the display apparatus 220 may be a liquid crystal display (liquid crystal display, LCD), a plasma display, an organic light-emitting diode (organic light-emitting diode, OLED) display, or another type of display apparatus.

[0078] Although not shown in FIG. 1, in some aspects, the encoder 100 and the decoder 200 may be respectively integrated with an encoder and a decoder, and may include an appropriate multiplexer-demultiplexer (multiplexer-demultiplexer, MUX-DEMUX) unit or other hardware and software for encoding both audio and videos in a shared data stream or separate data streams. In some embodiments, if applicable, the MUX-DEMUX unit may comply with the ITU H.223 multiplexer protocol or another protocol such as a user datagram protocol (user datagram protocol, UDP).

[0079] The encoder 100 and the decoder 200 may each be any one of the following circuits: one or more microprocessors, a digital signal processor (digital signal processor, DSP), an application specific integrated circuit (application specific integrated circuit, ASIC), a field-programmable gate array (field-programmable gate array, FPGA), discrete logic, hardware, or any combination thereof. If the technology in this embodiment of this application is implemented partially in software, the apparatus may store, in an appropriate non-volatile computer-readable storage medium, instructions used for the software, and may use one or more processors to execute the instructions in hardware, to implement the technology in this embodiment of this application. Any of the forego-

ing content (including hardware, software, and a combination of hardware and software) may be considered as one or more processors. Each of the encoder 100 and the decoder 200 may be included in one or more encoders or decoders. Any one of the encoders or the decoders may be integrated as a part of a combined encoder/decoder (codec) in a corresponding apparatus.

[0080] In this embodiment of this application, the encoder 100 may be generally referred to as "signaling" or "sending" some information to another apparatus, for example, the decoder 200. The term "signaling" or "sending" may generally refer to transmission of a syntax element for decoding compressed three-dimensional audio signal and/or other data. Such transmission may occur in real time or almost real time. Alternatively, such communication may occur after a period of time, for example, may occur when a syntax element in an encoded bitstream is stored in a computer-readable storage medium during encoding. The decoding apparatus may then retrieve the syntax element at any time after the syntax element is stored in the medium.

[0081] The encoding and decoding methods provided in embodiments of this application may be applied to a plurality of scenarios. The following separately describes several of the scenarios.

[0082] FIG. 2 is a schematic diagram of an implementation environment in which an encoding and decoding method is applied to a terminal scenario according to an embodiment of this application. The implementation environment includes a first terminal 101 and a second terminal 201. The first terminal 101 and the second terminal 201 perform a communication connection. The communication connection may be a wireless connection, or may be a wired connection. This is not limited in this embodiment of this application.

[0083] The first terminal 101 may be a transmit end device, or may be a receive end device. Similarly, the second terminal 201 may be a receive end device, or may be a transmit end device. If the first terminal 101 is a transmit end device, the second terminal 201 is a receive end device. If the first terminal 101 is a receive end device, the second terminal 201 is a transmit end device.

[0084] The following describes an example in which the first terminal 101 is a transmit end device and the second terminal 201 is a receive end device.

[0085] The first terminal 101 may be the source apparatus 10 in the implementation environment shown in FIG. 1. The second terminal 201 may be the destination apparatus 20 in the implementation environment shown in FIG. 1. Both the first terminal 101 and the second terminal 201 include an audio collection module, an audio playback module, an encoder, a decoder, a channel encoding module, and a channel decoding module.

[0086] An audio collection module in the first terminal 101 collects a three-dimensional audio signal and transmits the three-dimensional audio signal to an encoder. The encoder encodes the three-dimensional audio signal by using the encoding method provided in embodiments of this application. The encoding may be referred to as source encoding. Then, to transmit the three-dimensional audio signal on a channel, the channel encoding module further needs to perform channel encoding, and then transmit a bistream obtained through encoding on a digital channel through a wireless or wired network communication device.

[0087] The second terminal 201 receives, through the wireless or wired network communication device, the bistream transmitted on the digital channel. The channel decoding module performs channel decoding on the bistream. Then, the decoder performs decoding by using the decoding method provided in embodiments of this application to obtain a three-dimensional audio signal, and then plays the three-dimensional audio signal through the audio playback module.

[0088] The first terminal 101 and the second terminal 201 may be any electronic product that can perform human-computer interaction with a user in one or more manners such as a keyboard, a touchpad, a touchscreen, a remote controller, a voice interaction device, or a handwriting device, for example, a personal computer (personal computer, PC), a mobile phone, a smartphone, a personal digital assistant (Personal Digital Assistant, PDA), a wearable device, a pocket personal computer (pocket PC, PPC), a tablet computer, a smart head unit, a smart television, or a smart speaker.

[0089] A person skilled in the art should understand that the foregoing terminal is merely an example. Another existing or future terminal that may be applicable to embodiments of this application should also fall within the protection scope of embodiments of this application, and is included herein by reference.

[0090] FIG. 3 is a schematic diagram of an implementation environment in which an encoding and decoding method is applied to a transcoding scenario of a wireless or core network device according to an embodiment of this application. The implementation environment includes a channel decoding module, an audio decoder, an audio encoder, and a channel encoding module.

[0091] The audio decoder may be a decoder that uses the decoding method provided in embodiments of this application, or may be a decoder that uses another decoding method. The audio encoder may be an encoder that uses the encoding method provided in embodiments of this application, or may be an encoder that uses another encoding method. If the audio decoder is a decoder that uses the decoding method provided in embodiments of this application, the audio encoder is an encoder that uses another encoding method. If the audio decoder is a decoder that uses another decoding method, the audio encoder is an encoder that uses the encoding method provided in embodiments of this application.

[0092] In a first case, the audio decoder is a decoder that uses the decoding method provided in embodiments of this application. The audio encoder is an encoder that uses another encoding method.

[0093] In this case, the channel decoding module is

configured to perform channel decoding on a received bistream. The audio decoder is configured to perform source decoding by using the decoding method provided in embodiments of this application. Then, the audio encoder performs encoding by using the another encoding method. Conversion from one format to another format, namely, transcoding, is implemented. Then, transmission is performed after channel encoding.

[0094] In a second case, the audio decoder is a decoder that uses another decoding method. The audio encoder is an encoder that uses the encoding method provided in embodiments of this application.

[0095] In this case, the channel decoding module is configured to perform channel decoding on a received bistream. The audio decoder is configured to perform source decoding by using the another decoding method. Then, the audio encoder performs encoding by using the encoding method provided in embodiments of this application. Conversion from one format to another format, namely, transcoding, is implemented. Then, transmission is performed after channel encoding.

[0096] The wireless device may be a wireless access point, a wireless router, a wireless connector, or the like. The core network device may be a mobility management entity, a gateway, or the like.

[0097] A person skilled in the art should understand that the foregoing wireless device or core network device is merely an example. Another existing or future wireless or core network device that may be applicable to embodiments of this application should also fall within the protection scope of embodiments of this application, and is included herein by reference.

[0098] FIG. 4 is a schematic diagram of an implementation environment in which an encoding and decoding method is applied to a broadcast television scenario according to an embodiment of this application. The broadcast television scenario includes a live broadcast scenario and a post-production scenario. For the live broadcast scenario, the implementation environment includes a live program three-dimensional sound production module, a three-dimensional sound encoding module, a set-top box, and a speaker group. The set-top box includes a three-dimensional sound decoding module. For the post-production scenario, the implementation environment includes a post-program three-dimensional sound production module, a three-dimensional sound encoding module, a network receiver, a mobile terminal, a headphone, and the like.

[0099] In the live broadcast scenario, the live program three-dimensional sound production module produces a three-dimensional sound signal. The three-dimensional sound signal includes a three-dimensional audio signal. The three-dimensional sound signal is encoded by using the encoding method in embodiments of this application to obtain a bistream. The bistream is transmitted to a user side through a broadcast network. The three-dimensional sound decoder in the set-top box decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The speaker group plays back the reconstructed three-dimensional sound signal. Alternatively, the bistream is transmitted to the user side through the Internet. A three-dimensional sound decoder in a network receiver decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The speaker group plays back the reconstructed three-dimensional sound signal. Alternatively, the bistream is transmitted to the user side through the Internet. A three-dimensional sound decoder in a mobile terminal decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The headphone plays back the reconstructed three-dimensional sound signal.

[0100] In the post-production scenario, the post-program three-dimensional sound production module produces a three-dimensional sound signal. The three-dimensional sound signal is encoded by using the encoding method in embodiments of this application to obtain a bistream. The bistream is transmitted to a user side through a broadcast network. The three-dimensional sound decoder in the set-top box decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The speaker group plays back the reconstructed three-dimensional sound signal. Alternatively, the bistream is transmitted to the user side through the Internet. A three-dimensional sound decoder in a network receiver decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The speaker group plays back the reconstructed three-dimensional sound signal. Alternatively, the bistream is transmitted to the user side through the Internet. A three-dimensional sound decoder in a mobile terminal decodes the bistream by using the decoding method provided in embodiments of this application, to reconstruct a three-dimensional sound signal. The headphone plays back the reconstructed three-dimensional sound signal.

[0101] FIG. 5 is a schematic diagram of an implementation environment in which an encoding and decoding method is applied to a virtual reality flow scenario according to an embodiment of this application. The implementation environment includes an encoder side and a decoder side. The encoder side includes a collection module, a preprocessing module, an encoding module, an encapsulation module, and a delivery module. The decoder side includes a decapsulation module, a decoding module, a rendering module, and a headphone.

[0102] The collection module collects a three-dimensional audio signal. Then, the preprocessing module performs a preprocessing operation. The preprocessing operation includes filtering out a low-frequency part in the signal usually with 20 Hz or 50 Hz being a boundary point, and extracting orientation information and the like in the signal. Then, the encoding module performs encoding

processing by using the encoding method provided in embodiments of this application. After the encoding, the encapsulation module performs encapsulation. Then, the delivery module delivers an encapsulated signal to the decoder side.

[0103] The decapsulation module at the decoder side first performs decapsulation. The decoding module performs decoding by using the decoding method provided in embodiments of this application. Then, the rendering module performs binaural rendering processing on the decoded signal. A signal obtained through the rendering processing is mapped to the headphone of a listener. The headphone may be an independent headphone, or may be a headphone on a glasses device based on virtual reality.

[0104] It should be noted that the system architecture and the service scenario described in embodiments of this application are intended to describe the technical solutions in embodiments of this application more clearly, and do not constitute a limitation on the technical solutions provided in embodiments of this application. A person of ordinary skill in the art may know that: With the evolution of the system architecture and the emergence of new service scenarios, the technical solutions provided in embodiments of this application are also applicable to similar technical problems.

[0105] The following describes in detail the encoding and decoding methods provided in embodiments of this application. It should be noted that, with reference to the implementation environment shown in FIG. 1, any one of the following encoding methods may be performed by the encoder 100 in the source apparatus 10. Any one of the following decoding methods may be performed by the decoder 200 in the destination apparatus 20.

[0106] FIG. 6 is a flowchart of a first encoding method according to an embodiment of this application. The encoding method is applied to an encoder side device, and includes the following steps.

[0107] **Step 601:** Separately perform transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1.

[0108] The M transient state detection results one-to-one correspond to the M channels included in the time-domain three-dimensional audio signal of the current frame. The transient state detection result includes a transient state flag, or the transient state detection result includes a transient state flag and transient state position information. The transient state flag indicates whether a signal of a corresponding channel is a transient state signal. The transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel.

[0109] The M transient state detection results corresponding to the M channels may be determined in a plurality of manners. The following describes one of the manners. Because the transient state detection result corresponding to each of the M channels is determined in a same manner, the following uses one of the channels as an example to describe a manner of determining the transient state detection result corresponding to the channel. For ease of description, the channel is referred to as a target channel, and the following separately describes a transient state flag and transient state position information of the target channel.

Transient state flag of the target channel

[0110] The transient state detection parameter corresponding to the target channel is determined based on the signal of the target channel. The transient state flag corresponding to the target channel is determined based on the transient state detection parameter corresponding to the target channel.

[0111] In an example, the transient state detection parameter corresponding to the target channel is an absolute value of an inter-frame energy difference. To be specific, energy of the signal of the target channel in the current frame and energy of a signal of a target channel in a previous frame relative to the current frame are determined, and an absolute value of a difference between the energy of the signal of the target channel in the current frame and the energy of the signal of the target channel in the previous frame is determined to obtain the absolute value of the inter-frame energy difference. If the absolute value of the inter-frame energy difference exceeds a first energy difference threshold, it is determined that the transient state flag corresponding to the target channel in the current frame is a first value. Otherwise, it is determined that the transient state flag corresponding to the target channel in the current frame is a second value.

[0112] Based on the foregoing descriptions, the transient state flag indicates whether the signal of the corresponding channel is a transient state signal. Therefore, if the absolute value of the inter-frame energy difference exceeds the first energy difference threshold, it indicates that the signal of the target channel in the current frame is a transient state signal. In this case, it is determined that the transient state flag corresponding to the target channel in the current frame is the first value. If the absolute value of the inter-frame energy difference does not exceed the first energy difference threshold, it indicates that the signal of the target channel in the current frame is not a transient state signal. In this case, it is determined that the transient state flag corresponding to the target channel in the current frame is the second value.

[0113] It should be noted that the first value and the second value can be represented in a plurality of manners. For example, the first value is true, and the second value is false. Alternatively, the first value is 1, and the second value is 0. Certainly, the first value and the second value can alternatively be represented in another manner. The first energy difference threshold is preset.

The first energy difference threshold can be adjusted based on different requirements.

[0114] In another example, the transient state detection parameter corresponding to the target channel is an absolute value of a subframe energy difference. To be specific, the signal of the target channel in the current frame includes signals of a plurality of subframes, and the absolute value of the subframe energy difference corresponding to each of the plurality of subframes is determined to determine a transient state flag corresponding to each subframe. If a subframe whose transient state flag is the first value exists in the plurality of subframes, it is determined that the transient state flag corresponding to the target channel in the current frame is the first value. If no subframe whose transient state flag is the first value exists in the plurality of subframes, it is determined that the transient state flag corresponding to the target channel in the current frame is the second value.

[0115] Because the transient state flag of each of the plurality of subframes is determined in a same manner, the following describes an example of an i^{th} subframe in the plurality of subframes, where i is a positive integer. To be specific, energy of a signal in the i^{th} subframe and energy of a signal in the $(i-1)^{th}$ subframe in the plurality of subframes are determined, and an absolute value of a difference between the energy of the signal of the i^{th} subframe and the energy of the signal of the $(i-1)^{th}$ subframe is determined to obtain an absolute value of a subframe energy difference corresponding to the i^{th} subframe. If the absolute value of the subframe energy difference corresponding to the i^{th} subframe exceeds a second energy difference threshold, it is determined that a transient state flag of the i^{th} subframe is the first value. Otherwise, it is determined that the transient state flag of the i^{th} subframe is the second value.

[0116] Based on the foregoing descriptions, the transient state flag indicates whether the signal of the corresponding channel is a transient state signal. Therefore, if the absolute value of the subframe energy difference corresponding to the i^{th} subframe exceeds the second energy difference threshold, it indicates that the signal of the i^{th} subframe is a transient state signal. In this case, it is determined that the transient state flag of the i^{th} subframe is the first value. If the absolute value of the subframe energy difference corresponding to the i^{th} subframe does not exceed the second energy difference threshold, it indicates that the signal of the i^{th} subframe is not a transient state signal. In this case, it is determined that the transient state flag of the i^{th} subframe is the second value.

[0117] It should be noted that, when i = 0, the energy of the signal of the $(i-1)^{th}$ subframe is energy of a signal of a last subframe of the target channel in the previous frame relative to the current frame. The second energy difference threshold is preset. The second energy difference threshold can be adjusted based on different requirements. In addition, the second energy difference threshold may be the same as or different from the first energy difference threshold.

Transient state position information of the target channel

[0118] The transient state position information corresponding to the target channel is determined based on the transient state flag corresponding to the target channel.

[0119] In an example, if the transient state flag corresponding to the target channel is the first value, the transient state position information corresponding to the target channel is determined. If the transient state flag corresponding to the target channel is the second value, it is determined that the target channel does not have corresponding transient state position information, or the transient state position information corresponding to the target channel is set to a preset value, for example, set to -1.

[0120] In other words, if the transient state flag corresponding to the target channel is the second value, it indicates that the signal of the target channel is not a transient state signal. In this case, the transient state detection result of the target channel does not include the transient state position information, or the transient state position information corresponding to the target channel is directly set to a preset value, where the preset value indicates that the signal of the target channel is not a transient state signal. In other words, the transient state detection result of the transient state signal includes the transient state flag and the transient state position information. A transient state detection result of a non-transient state signal may include a transient state flag, or may include a transient state flag and transient state position information.

[0121] It should be noted that if the transient state flag corresponding to the target channel is the first value, the transient state position information corresponding to the target channel is determined in a plurality of manners. In an example, the signal of the target channel in the current frame includes signals of a plurality of subframes. A subframe whose transient state flag is the first value and whose absolute value of the subframe energy difference is the largest is selected from the plurality of subframes. A sequence number of the selected subframe is determined as the transient state position information corresponding to the target channel in the current frame.

[0122] For example, the transient state flag corresponding to the target channel in the current frame is the first value, and the signals of the target channel in the current frame include signals of four subframes, where i = 0, 1, 2, and 3. An absolute value of a subframe energy difference of a 0^{th} subframe is 18. An absolute value of a subframe energy difference of a 1^{st} subframe is 21. An absolute value of a subframe energy difference of a 2^{nd} subframe is 24. An absolute value of a subframe energy difference of a 3^{rd} subframe is 35. Assuming that the preset second energy difference threshold is 20, the signal of the 1^{st} subframe is a transient state signal, the

signal of the 2nd subframe is a transient state signal, and the signal of the 3rd subframe is a transient state signal. In this case, it is determined that transient state flags of the 1st subframe, the 2nd subframe, and the 3rd subframe are all the first value, and a subframe whose absolute value of the subframe energy difference is the largest in the three subframes is the 3rd subframe. In this case, a sequence number 3 of the 3rd subframe is determined as the transient state position information corresponding to the target channel in the current frame.

[0123] **Step 602:** Determine a global transient state detection result based on the M transient state detection results.

[0124] In some embodiments, the global transient state detection result includes a global transient state flag. If a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to m, it is determined that the global transient state flag is the first value, where m is a positive integer greater than 0 and less than M; or if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M channels is greater than or equal to n, it is determined that the global transient state flag is the first value, where n is a positive integer greater than 0 and less than M.

[0125] For example, the three-dimensional audio signal of the current frame is a third-order HOA signal. A quantity of channels of the HOA signal is $(3+1)^2$, namely, 16. Assuming that m is 1, if a quantity of transient state flags that are the first value in the 16 transient state flags is greater than or equal to 1, it is determined that the global transient state flag is the first value. Alternatively, the first preset condition includes belonging to channels of an FOA signal. For example, the channels of the FOA signal may include first four channels of the HOA signal. Assuming that a channel that meets the first preset condition in the M channels is a channel in which the FOA signal in the current frame is located, n is 1. If a quantity of channels that belong to the channels of the FOA signal and whose corresponding transient state flags are the first value in the 16 channels is greater than or equal to 1, it is determined that the global transient state flag is the first value.

[0126] Herein, m and n are preset values, and m and n can also be adjusted based on different requirements. If the three-dimensional audio signal is an HOA signal, the first preset condition includes belonging to channels of an FOA signal. A channel that meets the first preset condition in the M channels is a channel in which the FOA signal in the three-dimensional audio signal of the current frame is located. The FOA signal is signals of the first four channels of the HOA signal. Certainly, the first preset condition may alternatively be another condition.

[0127] In some other embodiments, the global transient state detection result further includes the global transient state position information. If only one transient state flag in the M transient state flags is the first value, transient state position information corresponding to a channel whose transient state flag is the first value is determined as the global transient state position information; or if at least two transient state flags in the M transient state flags are the first value, transient state position information corresponding to a channel with a largest transient state detection parameter in at least two channels corresponding to the at least two transient state flags is determined as the global transient state position information. Alternatively, if at least two transient state flags in the M transient state flags are the first value, and a difference between transient state position information corresponding to two channels is less than a position difference threshold, an average value of the transient state position information corresponding to the two channels is determined as the global transient state position information. The position difference threshold is preset and can be adjusted based on different requirements.

[0128] Based on the foregoing descriptions, a transient state detection parameter corresponding to a channel is an absolute value of an inter-frame energy difference or an absolute value of a subframe energy difference. If the transient state detection parameter corresponding to the channel is the absolute value of the inter-frame energy difference, one channel corresponds to one absolute value of the inter-frame energy difference. In this case, a channel corresponding to a largest absolute value of the inter-frame energy difference may be selected from the at least two channels, and then transient state position information corresponding to the selected channel is determined as the global transient state position information. If the transient state detection parameter corresponding to the channel is the absolute value of the subframe energy difference, one channel corresponds to a plurality of absolute values of the subframe energy difference. In this case, a channel corresponding to a largest absolute value of the subframe energy difference may be selected from the at least two channels, and then transient state position information corresponding to the selected channel is determined as the global transient state position information.

[0129] For example, for the third-order HOA signal, if only a transient state flag corresponding to a third channel in the 16 transient state flags of the HOA signal is the first value, the transient state position information corresponding to the third channel may be directly determined as the global transient state position information.

[0130] If three transient state flags in the 16 transient state flags of the HOA signal are the first value and correspond to a channel 1, a channel 2, and a channel 3, the channel 1 corresponds to transient state position information 1, an absolute value of an inter-frame energy difference corresponding to the channel 1 is 22, the channel 2 corresponds to transient state position information 2, an absolute value of an inter-frame energy difference corresponding to the channel 2 is 23, the channel 3 corresponds to transient state position information 3, and an absolute value of an inter-frame energy difference corresponding to the channel 3 is 28. If a channel with a

largest absolute value of an inter-frame energy difference in the three channels is the channel 3, the transient state position information 3 corresponding to the channel 3 is determined as the global transient state position information.

[0131] For another example, if three transient state flags in the 16 transient state flags of the HOA signal are the first value and correspond to a channel 1, a channel 2, and a channel 3, the channel 1 corresponds to transient state position information 1, a signal of the channel 1 includes three subframes, absolute values of subframe energy differences corresponding to the three subframes are 20, 18, and 22 respectively, the channel 2 corresponds to transient state position information 2, a signal of the channel 2 includes three subframes, absolute values of subframe energy differences corresponding to the three subframes are 20, 23, and 25 respectively, the channel 3 corresponds to transient state position information 3, a signal of the channel 3 includes three subframes, and absolute values of subframe energy differences corresponding to the three subframes are 25, 28, and 30. If a channel with a largest absolute value of a subframe energy difference in the three channels is the channel 3, the transient state position information 3 corresponding to the channel 3 is determined as the global transient state position information.

[0132] If three transient state flags in the 16 transient state flags of the HOA signal are the first value and correspond to a channel 1, a channel 2, and a channel 3, the channel 1 corresponds to transient state position information 1, the channel 2 corresponds to transient state position information 3, and the channel 3 corresponds to transient state position information 6. If a difference 2 between the transient state position information corresponding to the channel 1 and the transient state position information corresponding to the channel 2 in the three channels is less than a preset position difference threshold 3, an average value 2 of the transient state position information corresponding to the channel 1 and the transient state position information corresponding to the channel 2 is determined as the global transient state position information.

[0133] **Step 603:** Convert the time-domain three-dimensional audio signal of the current frame into a frequency-domain three-dimensional audio signal based on the global transient state detection result.

[0134] In some embodiments, a target encoding parameter is determined based on the global transient state detection result. The target encoding parameter includes a window function type of the current frame and/or a frame type of the current frame. The time-domain three-dimensional audio signal of the current frame is converted into the frequency-domain three-dimensional audio signal based on the target encoding parameter.

[0135] In an example, the global transient state detection result includes the global transient state flag. An implementation process of determining the window function type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining a type of a first preset window function as the window function type of the current frame; or if the global transient state flag is the second value, determining a type of a second preset window function as the window function type of the current frame. A window length of the first preset window function is less than a window length of the second preset window function.

[0136] In another example, the global transient state detection result includes the global transient state flag and the global transient state position information. An implementation process of determining the window function type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining the window function type of the current frame based on the global transient state position information; or if the global transient state flag is the second value, determining a type of a third preset window function as the window function type of the current frame, or determining the window function type of the current frame based on a window function type of a previous frame relative to the current frame.

[0137] If the global transient state flag is the first value, the window function type of the current frame is determined based on the global transient state position information in a plurality of manners. For example, a type of a fourth preset window function is adjusted based on the global transient state position information, so that a central position of the fourth preset window function corresponds to a global transient state occurrence position, and then a value of a window function corresponding to the global transient state occurrence position is the largest. Alternatively, a window function corresponding to the global transient state occurrence position is selected from a window function set, and then a type of the selected window function is determined as the window function type of the current frame. In other words, the window function set stores a window function corresponding to each transient state occurrence position. In this way, the window function corresponding to the global transient state occurrence position may be selected.

[0138] In addition, the window function type of the current frame is determined based on the window function type of the previous frame relative to the current frame by using a plurality of methods. For details, refer to a related technology. Details are not described in this embodiment of this application.

[0139] The global transient state detection result may include only the global transient state flag, or may include the global transient state flag and the global transient state position information. In addition, the global transient state position information may be transient state position information corresponding to a channel whose transient state flag is the first value, or may be a preset value. If global transient state detection results are different, the frame type of the current frame is determined in different manners. Therefore, the following separately describes

the following three cases.

[0140] In a first case, the global transient state detection result includes the global transient state flag. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames; or if the global transient state flag is the second value, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame.

[0141] In a second case, the global transient state detection result includes the global transient state flag and the global transient state position information. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state flag is the first value and the global transient state position information meets a second preset condition, determining that the frame type of the current frame is a third type, where the third type indicates that the current frame includes a plurality of ultra-short frames; if the global transient state flag is the first value and the global transient state position information does not meet the second preset condition, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames; or if the global transient state flag is the second value, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame. A frame length of an ultra-short frame is less than a frame length of a short frame. A frame length of a short frame is less than a frame length of a long frame. The second preset condition may be that a distance between a transient state occurrence position indicated by the global transient state position information and a start position of the current frame is less than the frame length of the ultra-short frame, or a distance between the transient state occurrence position indicated by the global transient state position information and an end position of the current frame is less than the frame length of the ultra-short frame.

[0142] In a third case, the global transient state detection result includes the global transient state position information. An implementation process of determining the frame type of the current frame based on the global transient state detection result includes: if the global transient state position information is a preset value, for example, -1, determining that the frame type of the current frame is a second type, where the second type indicates that the current frame includes a long frame; if the global transient state position information is not the preset value and meets a second preset condition, determining that the frame type of the current frame is a third type, where the third type indicates that the current frame includes a plurality of ultra-short frames; or if the global transient state position information is not the preset value and does not meet the second preset condition, determining that the frame type of the current frame is a first type, where the first type indicates that the current frame includes a plurality of short frames. A frame length of an ultra-short frame is less than a frame length of a short frame. A frame length of a short frame is less than a frame length of a long frame. The second preset condition may be that a distance between a transient state occurrence position indicated by the global transient state position information and a start position of the current frame is less than the frame length of the ultra-short frame, or a distance between the transient state occurrence position indicated by the global transient state position information and an end position of the current frame is less than the frame length of the ultra-short frame.

[0143] It should be noted that the window function type of the current frame indicates a shape and a length of a window function corresponding to the current frame. The window function of the current frame is used for performing windowing processing on the time-domain three-dimensional audio signal of the current frame. The frame type of the current frame indicates whether the current frame is an ultra-short frame, a short frame, or a long frame. The ultra-short frame, the short frame, and the long frame may be distinguished based on duration of the frames. The specific duration may be set based on different requirements. This is not limited in this embodiment of this application.

[0144] A manner of converting the time-domain three-dimensional audio signal of the current frame into the frequency-domain three-dimensional audio signal may be modified discrete cosine transform (modified discrete cosine transform, MDCT), modified discrete sine transform (modified discrete sine transform, MDST), or fast Fourier transform (fast Fourier transform, FFT).

[0145] Based on the foregoing descriptions, the target encoding parameter includes the window function type of the current frame and/or the frame type of the current frame. To be specific, the target encoding parameter includes the window function type of the current frame, or the target encoding parameter includes the frame type of the current frame, or the target encoding parameter includes the window function type and the frame type of the current frame. When parameters included in the target encoding parameter are different, processes of converting the time-domain three-dimensional audio signal of the current frame into the frequency-domain three-dimensional audio signal based on the target encoding parameter are different. Therefore, descriptions are separately provided below.

[0146] In a first case, the target encoding parameter includes the window function type of the current frame. In this case, windowing processing is performed on the time-domain three-dimensional audio signal of the current frame based on the window function indicated by the window function type of the current frame. Then, a three-dimensional audio signal obtained through the win-

dowing processing is converted into the frequency-domain three-dimensional audio signal.

[0147] In a second case, the target encoding parameter includes the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, a time-domain three-dimensional audio signal of each short frame included in the current frame is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, a time-domain three-dimensional audio signal of the long frame included in the current frame is directly converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, a time-domain three-dimensional audio signal of each ultra-short frame included in the current frame is converted into a frequency-domain three-dimensional audio signal.

[0148] In a third case, the target encoding parameter includes the window function type and the frame type of the current frame. In this case, if the frame type of the current frame is the first type, it indicates that the current frame includes a plurality of short frames. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of each short frame included in the current frame, and a time-domain three-dimensional audio signal of each short frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the second type, it indicates that the current frame includes a long frame. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of the long frame included in the current frame, and a time-domain three-dimensional audio signal of the long frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal. If the frame type of the current frame is the third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, windowing processing is performed, based on the window function indicated by the window function type of the current frame, on a time-domain three-dimensional audio signal of each ultra-short frame included in the current frame, and a time-domain three-dimensional audio signal of each ultra-short frame obtained through the windowing processing is converted into a frequency-domain three-dimensional audio signal.

[0149] In other words, if the current frame includes a plurality of ultra-short frames and short frames, after the time-domain three-dimensional audio signal of the current frame is converted into the frequency-domain three-dimensional audio signal, the frequency-domain three-dimensional audio signal of each ultra-short frame and short frame included in the current frame is obtained. If the current frame includes a long frame, after the time-domain three-dimensional audio signal of the current frame is converted into the frequency-domain three-dimensional audio signal, the frequency-domain three-dimensional audio signal of the long frame included in the current frame is obtained.

[0150] **Step 604:** Perform spatial encoding on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M.

[0151] In some embodiments, spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame based on the frame type of the current frame to obtain the spatial encoding parameter and the frequency-domain signals of the N transmission channels.

[0152] When spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame based on the frame type of the current frame, if the frame type of the current frame is the first type, namely if the current frame includes a plurality of short frames, frequency-domain three-dimensional audio signals of the plurality of short frames included in the current frame are interleaved to obtain a frequency-domain three-dimensional audio signal of a long frame, and spatial encoding is performed on the frequency-domain three-dimensional audio signal of the long frame obtained through the interleaving. If the frame type of the current frame is the second type, namely if the current frame includes a long frame, spatial encoding is performed on a frequency-domain three-dimensional audio signal of the long frame. If the frame type of the current frame is the third type, namely if the current frame includes a plurality of ultra-short frames, frequency-domain three-dimensional audio signals of the plurality of ultra-short frames included in the current frame are interleaved to obtain a frequency-domain three-dimensional audio signal of a long frame, and spatial encoding is performed on the frequency-domain three-dimensional audio signal of the long frame obtained through the interleaving.

[0153] A spatial encoding method may be any method that can obtain the spatial encoding parameter and the frequency-domain signals of the N transmission channels based on the frequency-domain three-dimensional audio signal of the current frame. For example, a spatial encoding method of matched projection may be used. The spatial encoding method is not limited in this embodiment of this application.

[0154] The spatial encoding parameter is a parameter determined in a process of performing spatial encoding on the frequency-domain three-dimensional audio signal of the current frame, and includes side information, bit

pre-allocation side information, and the like. The frequency-domain signals of the N transmission channels may include virtual speaker signals of one or more channels, and residual signals of one or more channels. In addition, when a quantity of bits for encoding is insufficient, the frequency-domain signals of the N transmission channels may alternatively include only virtual speaker signals of one or more channels.

[0155] **Step 605:** Encode the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result.

[0156] In some embodiments, the frequency-domain signals of the N transmission channels are encoded based on the frame type of the current frame.

[0157] In an example, an implementation process of encoding the frequency-domain signals of the N transmission channels includes: performing noise shaping processing on the frequency-domain signals of the N transmission channels based on the frame type of the current frame; performing transmission channel downmixing processing on frequency-domain signals of the N transmission channels obtained through the noise shaping processing, to obtain a downmixed signal; performing quantization and encoding processing on a low-frequency part of the downmixed signal, and writing an encoding result into the bistream; and performing bandwidth expansion and encoding processing on a high-frequency part of the downmixed signal, and writing an encoding result into the bistream.

[0158] It should be noted that for a manner of performing noise shaping processing based on the frame type of the current frame, refer to a related technology. Details are not described in this embodiment of this application. The noise shaping processing includes temporal noise shaping (temporal noise shaping, TNS) processing and frequency domain noise shaping (frequency domain noise shaping, FDNS) processing.

[0159] When the transmission channel downmixing processing is performed on the frequency-domain signals of the N transmission channels obtained through the noise shaping processing, the N transmission channels obtained through the noise shaping processing may be paired according to a preset criterion, or the frequency-domain signals of the N transmission channels obtained through the noise shaping processing may be paired according to signal correlation. Then, mid side (mid side, MS) downmixing processing is performed based on the two paired frequency-domain signals.

[0160] For example, if the N transmission channels include two virtual speaker signals and four residual signals, the two virtual speaker signals may be grouped into a pair according to a preset criterion for downmixing processing. A correlation between every two residual signals in the four residual signals may be further determined, two residual signals with high correlation are selected to form a pair, and the other two residual signals form a pair, and downmixing processing is separately performed.

[0161] It should be noted that downmixing processing is performed on the two paired frequency-domain signals, and a result of the downmixing processing may be one frequency-domain signal, or may be two frequency-domain signals, depending on an encoding processing process.

[0162] The low-frequency part and the high-frequency part of the signal may be divided in a plurality of manners. For example, 2000 Hz is used as a boundary point, a part of the downmixed signal whose frequency is less than 2000 Hz is used as the low-frequency part of the signal, and a part of the downmixed signal whose frequency is greater than 2000 Hz is used as the high-frequency part of the signal. For another example, 5000 Hz is used as a boundary point, a part of the downmixed signal whose frequency is less than 5000 Hz is used as the low-frequency part of the signal, and a part of the downmixed signal whose frequency is greater than 5000 Hz is used as the high-frequency part of the signal.

[0163] **Step 606:** Encode the spatial encoding parameter to obtain a spatial encoding parameter encoding result, and write the spatial encoding parameter encoding result and the frequency-domain signal encoding result into the bitstream.

[0164] Optionally, the method may further include: encoding the global transient state detection result to obtain a global transient state detection result encoding result, and writing the global transient state detection result encoding result into the bitstream; or encoding the target encoding parameter to obtain a target encoding parameter encoding result, and writing the target encoding parameter encoding result into the bitstream.

[0165] In this embodiment of this application, transient state detection may be performed on the signals of the M channels included in the time-domain three-dimensional audio signal of the current frame, to determine the global transient state detection result. Then, based on the global transient state detection result, time-frequency transform and spatial encoding of the audio signal are sequentially performed, and the frequency-domain signal of each transmission channel is encoded. Especially, when the frequency-domain signal of each transmission channel obtained through the spatial encoding is encoded, the global transient state detection result is used as the transient state detection result of each transmission channel, and the frequency-domain signal of each transmission channel does not need to be converted into a time domain to determine the transient state detection result corresponding to each transmission channel. Therefore, the three-dimensional audio signal does not need to be transformed between a time domain and a frequency domain for a plurality of times. This can reduce encoding complexity and improve encoding efficiency. In addition, in this embodiment of this application, the transient state detection result of each transmission channel does not need to be encoded, and only the global transient state detection result needs to be encoded into

the bitstream. This can reduce a quantity of bits for encoding.

[0166] FIG. 7 and FIG. 8 are both block diagrams of an example encoding method according to an embodiment of this application. FIG. 7 and FIG. 8 mainly describe an example of the encoding method shown in FIG. 6. In FIG. 7, transient state detection is separately performed on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels. A global transient state detection result is determined based on the M transient state detection results. The global transient state detection result is encoded to obtain a global transient state detection result encoding result. The global transient state detection result encoding result is written into a bitstream. A time-domain three-dimensional audio signal of a current frame is converted into a frequency-domain three-dimensional audio signal based on the global transient state detection result. Spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels. The spatial encoding parameter is encoded to obtain a spatial encoding parameter encoding result. The spatial encoding parameter encoding result and the frequency-domain signal encoding result are written into the bitstream. The frequency-domain signals of the N transmission channels are encoded based on the global transient state detection result. Further, in FIG. 8, after the spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame to obtain the spatial encoding parameter and the frequency-domain signals of the N transmission channels, the spatial encoding parameter is encoded to obtain the spatial encoding parameter encoding result, and the spatial encoding parameter encoding result and the frequency-domain signal encoding result are written into the bitstream. Then, noise shaping processing, transmission channel downmixing processing, quantization and encoding processing, and bandwidth expansion processing are performed on the frequency-domain signals of the N transmission channels based on the global transient state detection result, and an encoding result of a signal obtained through the bandwidth expansion processing is written into the bitstream.

[0167] Based on the descriptions in step 606, the encoder side device may encode the global transient state detection result into the bitstream, or may not encode the global transient state detection result into the bitstream. In addition, the encoder side device may encode the target encoding parameter into the bitstream, or may not encode the target encoding parameter into the bitstream. If the encoder side device encodes the global transient state detection result into the bitstream, a decoder side device may perform decoding by using the following method shown in FIG. 9. When the encoder side device encodes the target encoding parameter into the bitstream, the decoder side device may parse the bitstream to obtain the target encoding parameter, and then perform decoding based on a frame type of the current frame included in the target encoding parameter. A specific implementation process is similar to the process in FIG. 9. Certainly, the encoder side device may not encode the global transient state detection result into the bitstream, and may not encode the target encoding parameter into the bitstream. In this case, for a process of decoding the three-dimensional audio signal, refer to a related technology. Details are not described in this embodiment of this application.

[0168] FIG. 9 is a flowchart of a first decoding method according to an embodiment of this application. The method is applied to a decoder side, and includes the following steps.

[0169] **Step 901:** Parse a bitstream to obtain a global transient state detection result and a spatial encoding parameter.

[0170] **Step 902:** Perform decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels.

[0171] In some embodiments, a frame type of a current frame is determined based on the global transient state detection result. Decoding is performed based on the frame type of the current frame and the bitstream to obtain the frequency-domain signals of the N transmission channels.

[0172] For an implementation of determining the frame type of the current frame based on the global transient state detection result, refer to related descriptions in step 603. Details are not described herein again. For an implementation of performing decoding based on the frame type of the current frame and the bitstream, refer to a related technology. Details are not described in this embodiment of this application.

[0173] **Step 903:** Perform spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal.

[0174] In some embodiments, spatial decoding is performed on the frequency-domain signals of the N transmission channels based on the frame type of the current frame and the spatial encoding parameter to obtain the reconstructed frequency-domain three-dimensional audio signal. The frame type of the current frame is determined based on the global transient state detection result. In other words, the frame type of the current frame is determined based on the global transient state detection result, and then spatial decoding is performed on the frequency-domain signals of the N transmission channels based on the frame type of the current frame and the spatial encoding parameter to obtain the reconstructed frequency-domain three-dimensional audio signal.

[0175] For an implementation process of performing

spatial decoding on the frequency-domain signals of the N transmission channels based on the frame type of the current frame and the spatial encoding parameter, refer to a related technology. Details are not described in this embodiment of this application.

[0176] **Step 904:** Determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

[0177] In some embodiments, a target encoding parameter is determined based on the global transient state detection result. The target encoding parameter includes a window function type of the current frame and/or the frame type of the current frame. The reconstructed frequency-domain three-dimensional audio signal is converted into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter.

[0178] For an implementation of determining the target encoding parameter based on the global transient state detection result, refer to related descriptions in step 603. Details are not described herein again.

[0179] Based on the foregoing descriptions, the target encoding parameter includes the window function type of the current frame and/or the frame type of the current frame. To be specific, the target encoding parameter includes the window function type of the current frame, or the target encoding parameter includes the frame type of the current frame, or the target encoding parameter includes the window function type and the frame type of the current frame. When parameters included in the target encoding parameter are different, processes of converting the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter are different. Therefore, descriptions are separately provided below.

[0180] In a first case, the target encoding parameter includes the window function type of the current frame. In this case, windowing removal processing is performed on the reconstructed frequency-domain three-dimensional audio signal based on a window function indicated by the window function type of the current frame. Then, a frequency-domain three-dimensional audio signal obtained through the windowing removal processing is converted into the reconstructed time-domain three-dimensional audio signal.

[0181] The windowing removal processing is also referred to as windowing and overlap-add processing.

[0182] In a second case, the target encoding parameter includes the frame type of the current frame. In this case, if the frame type of the current frame is a first type, it indicates that the current frame includes a plurality of short frames. In this case, a reconstructed frequency-domain three-dimensional audio signal of each short frame is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is a second type, it indicates that the current frame includes a long frame. In this case, a reconstructed frequency-domain three-dimensional audio signal of the long frame included in the current frame is directly converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is a third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, a reconstructed frequency-domain three-dimensional audio signal of each ultra-short frame is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal.

[0183] In a third case, the target encoding parameter includes the window function type and the frame type of the current frame. In this case, if the frame type of the current frame is a first type, it indicates that the current frame includes a plurality of short frames. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a frequency-domain three-dimensional audio signal of each short frame included in the current frame, and a reconstructed frequency-domain three-dimensional audio signal of each short frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is a second type, it indicates that the current frame includes a long frame. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a reconstructed frequency-domain three-dimensional audio signal of the long frame included in the current frame, and a frequency-domain three-dimensional audio signal of the long frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal. If the frame type of the current frame is a third type, it indicates that the current frame includes a plurality of ultra-short frames. In this case, windowing removal processing is performed, based on the window function indicated by the window function type of the current frame, on a frequency-domain three-dimensional audio signal of each ultra-short frame included in the current frame, and a reconstructed frequency-domain three-dimensional audio signal of each ultra-short frame obtained through the windowing removal processing is converted into a time-domain three-dimensional audio signal to obtain a reconstructed time-domain three-dimensional audio signal.

[0184] In this embodiment of this application, a decoder side parses a bitstream to obtain the global transient state detection result and the spatial encoding parameter. In this way, the time-domain three-dimensional audio signal can be reconstructed based on the global transient state detection result and the spatial encoding parameter, and there is no need to parse the bitstream to obtain

a transient state detection result of each transmission channel. This can reduce decoding complexity and improve decoding efficiency. In addition, when the target encoding parameter is not encoded into the bitstream, the target encoding parameter may be directly determined based on the global transient state detection result, to reconstruct the time-domain three-dimensional audio signal.

[0185] FIG. 10 is a block diagram of an example decoding method according to an embodiment of this application. FIG. 10 mainly describes an example of the decoding method shown in FIG. 9. In FIG. 10, a bitstream is parsed to obtain a global transient state detection result and a spatial encoding parameter. Decoding is performed based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels. Spatial decoding is performed on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal. A reconstructed time-domain three-dimensional audio signal is determined through windowing removal processing and time-frequency inverse transform based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

[0186] FIG. 11 is a flowchart of a second encoding method according to an embodiment of this application. The encoding method is applied to an encoder side device, and includes the following steps.

[0187] **Step 1101:** Separately perform transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1.

[0188] For an implementation of determining the M transient state detection results corresponding to the M channels, refer to related descriptions in step 601. Details are not described herein again.

[0189] **Step 1102:** Determine a global transient state detection result based on the M transient state detection results.

[0190] For an implementation of determining global transient state position information based on the M transient state detection results, refer to related descriptions in step 602. Details are not described herein again.

[0191] **Step 1103:** Convert the time-domain three-dimensional audio signal of the current frame into a frequency-domain three-dimensional audio signal based on the global transient state detection result.

[0192] For a manner of converting the time-domain three-dimensional audio signal of the current frame into the frequency-domain three-dimensional audio signal based on the global transient state detection result, refer to related descriptions in step 603. Details are not described herein again.

[0193] **Step 1104:** Perform spatial encoding on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M.

[0194] For an implementation of performing spatial encoding on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result, refer to related descriptions in step 604. Details are not described herein again.

[0195] **Step 1105:** Determine N transient state detection results corresponding to the N transmission channels based on the M transient state detection results.

[0196] In some embodiments, transient state flags of virtual speaker signals of one or more channels included in the N transmission channels are determined based on the M transient state flags by using a first preset rule. Transient state flags of residual signals of one or more channels included in the N transmission channels are determined based on the M transient state flags by using a second preset rule.

[0197] In an example, the first preset rule includes: if a quantity of transient state flags that are a first value in the M transient state flags is greater than or equal to P, all the transient state flags of the virtual speaker signals of the one or more channels included in the N transmission channels are the first value. The second preset rule includes: if a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to Q, all the transient state flags of the residual signals of the one or more channels included in the N transmission channels are the first value.

[0198] Both P and Q are positive integers less than M. P and Q are preset values, and P and Q can also be adjusted based on different requirements. Optionally, because the virtual speaker signal is used for recording a real three-dimensional audio signal and is more important than the residual signal, P is less than Q.

[0199] In another example, the first preset rule includes: if a quantity of transient state flags that are a first value in the M transient state flags is greater than or equal to P, all the transient state flags corresponding to the virtual speaker signals of the one or more channels included in the N transmission channels are the first value. The second preset rule includes: if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M transient state flags is greater than or equal to R, all the transient state flags corresponding to the residual signals of the one or more channels included in the N transmission channels are the first value.

[0200] Both P and R are positive integers less than M. P and R are preset values, and P and R can also be adjusted based on different requirements. If the three-dimensional audio signal is an HOA signal, the first preset condition includes belonging to channels of an FOA sig-

nal. A channel that meets the first preset condition in the M channels is a channel in which the FOA signal in the three-dimensional audio signal of the current frame is located. The FOA signal is signals of first four channels of the HOA signal. Certainly, the first preset condition may alternatively be another condition.

[0201] In some other embodiments, the N transient state flags may be further determined based on the M transient state flags according to a mapping relationship between the M transient state flags and the N transmission channels. The mapping relationship is pre-determined.

[0202] For example, a transmission channel included in the N transmission channels is mapped to a plurality of channels in the M channels. If at least one transient state flag in the M transient state flags is the first value, a transient state flag of the transmission channel in the N transmission channels is the first value.

[0203] It should be noted that step 1105 may be performed on any occasion after step 1101 and before step 1106. An occasion for performing step 1105 is not limited in this embodiment of this application.

[0204] **Step 1106:** Encode the frequency-domain signals of the N transmission channels based on the N transient state detection results to obtain a frequency-domain signal encoding result.

[0205] In some embodiments, the frame type corresponding to each of the N transmission channels is determined based on the N transient state detection results. The frequency-domain signal of the corresponding transmission channel in the N transmission channels is encoded based on the frame type corresponding to each of the N transmission channels.

[0206] Because implementations of determining the frame type corresponding to each of the N transmission channels are the same, the following describes an example of one of the transmission channels. For ease of description, the transmission channel is referred to as a target transmission channel.

[0207] An implementation process of determining a frame type corresponding to the target transmission channel based on a transient state detection result corresponding to the target transmission channel includes: if a transient state flag corresponding to the target transmission channel is the first value, determining that the frame type corresponding to the target transmission channel is a first type, where the first type indicates that a signal of the target transmission channel includes a plurality of short frames; or if the transient state flag corresponding to the target transmission channel is the second value, determining that the frame type corresponding to the target transmission channel is a second type, where the second type indicates that the signal of the target transmission channel includes a long frame.

[0208] It should be noted that the frame type of the current frame indicates whether the current frame is a short frame or a long frame. The short frame and the long frame may be distinguished based on duration of the frames. The specific duration may be set based on different requirements. This is not limited in this embodiment of this application.

[0209] After the frame type corresponding to each transmission channel is determined, noise shaping processing may be performed on the frequency-domain signal of each transmission channel based on the frame type corresponding to each transmission channel. Then, transmission channel downmixing processing is performed on the frequency-domain signals of the N transmission channels obtained through the noise shaping processing, to obtain a downmixed signal. Quantization and encoding processing are performed on a low-frequency part of the downmixed signal, and an encoding result is written into a bitstream. Bandwidth expansion and encoding processing are performed on a high-frequency part of the downmixed signal, and an encoding result is written into the bitstream.

[0210] For related content of the noise shaping processing, the transmission channel downmixing processing, the quantization and encoding processing of the low-frequency part, and the bandwidth expansion and encoding processing, refer to related descriptions in step 605. Details are not described herein again.

[0211] **Step 1107:** Encode the spatial encoding parameter and the N transient state detection results to obtain a spatial encoding parameter encoding result and N transient state detection result encoding results, and write the spatial encoding parameter encoding result and the N transient state detection result encoding results into the bitstream.

[0212] Optionally, the method may further include: encoding the global transient state detection result to obtain a global transient state detection result encoding result, and writing the global transient state detection result encoding result into the bitstream; or encoding the target encoding parameter to obtain a target encoding parameter encoding result, and writing the target encoding parameter encoding result into the bitstream.

[0213] In this embodiment of this application, the virtual speaker signal included in each transmission channel and the transient state detection result corresponding to the residual signal are determined based on the M transient state detection results corresponding to the M channels included in the three-dimensional audio signal. This can improve encoding accuracy when the frequency-domain signal of each transmission channel is encoded. In addition, the transient state detection result corresponding to each transmission channel is determined based on the M transient state detection results, and the frequency-domain signal of each transmission channel does not need to be converted into a time domain to determine the transient state detection result corresponding to each transmission channel. Therefore, the three-dimensional audio signal does not need to be transformed between a time domain and a frequency domain for a plurality of times. This can reduce encoding complexity and improve encoding efficiency.

[0214] FIG. 12 and FIG. 13 are both block diagrams of another example encoding method according to an embodiment of this application. FIG. 12 and FIG. 13 mainly describe an example of the encoding method shown in FIG. 11. In FIG. 12, transient state detection is separately performed on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels. A global transient state detection result is determined based on the M transient state detection results. The global transient state detection result is encoded to obtain a global transient state detection result encoding result. The global transient state detection result encoding result is written into a bitstream. A time-domain three-dimensional audio signal of a current frame is converted into a frequency-domain three-dimensional audio signal based on the global transient state detection result. Spatial encoding is performed on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels. The spatial encoding parameter is encoded to obtain a spatial encoding parameter encoding result. The spatial encoding parameter encoding result is written into the bitstream. N transient state detection results corresponding to the N transmission channels are determined based on the M transient state detection results. The N transient state detection results are encoded to obtain N transient state detection result encoding results. The N transient state detection result encoding results are written into the bitstream. The frequency-domain signals of the N transmission channels are encoded based on the N transient state detection results. Further, in FIG. 13, after the N transient state detection results are determined, noise shaping processing is performed on the frequency-domain signals of the N transmission channels based on the N transient state detection results. Then, transmission channel downmixing processing, quantization and encoding processing, and bandwidth expansion processing are performed on a frequency-domain signal of each transmission channel obtained through the noise shaping processing, and an encoding result of a signal obtained through the bandwidth expansion processing is written into the bitstream.

[0215] Based on the descriptions in step 1107, the encoder side device may encode the global transient state detection result into the bitstream, or may not encode the global transient state detection result into the bitstream. In addition, the encoder side device may encode the target encoding parameter into the bitstream, or may not encode the target encoding parameter into the bitstream. If the encoder side device encodes the global transient state detection result into the bitstream, a decoder side device may perform decoding by using the following method shown in FIG. 14. When the encoder side device encodes the target encoding parameter into the bitstream, the decoder side device may parse the bitstream to obtain the target encoding parameter, and then perform decoding based on the frame type of the current frame included in the target encoding parameter. A specific implementation process is similar to the process in FIG. 14. Certainly, the encoder side device may not encode the global transient state detection result into the bitstream, and may not encode the target encoding parameter into the bitstream. In this case, for a process of decoding the three-dimensional audio signal, refer to a related technology. Details are not described in this embodiment of this application.

[0216] FIG. 14 is a flowchart of a second decoding method according to an embodiment of this application. The method is applied to a decoder side, and includes the following steps.

[0217] **Step 1401:** Parse a bitstream to obtain a global transient state detection result, N transient state detection results corresponding to N transmission channels, and a spatial encoding parameter.

[0218] **Step 1402:** Perform decoding based on the N transient state detection results and the bistream to obtain frequency-domain signals of the N transmission channels.

[0219] In some embodiments, a frame type corresponding to each transmission channel is determined based on the N transient state detection results. Decoding is performed based on the frame type corresponding to each transmission channel and the bitstream to obtain the frequency-domain signals of the N transmission channels.

[0220] For an implementation of determining the frame type corresponding to each transmission channel based on the N transient state detection results, refer to related descriptions in step 1106. Details are not described herein again. For an implementation of performing decoding based on the frame type corresponding to each transmission channel and the bitstream, refer to a related technology. Details are not described in this embodiment of this application.

[0221] **Step 1403:** Perform spatial decoding on the frequency-domain signals of the N transmission channels based on the frequency-domain signals of the N transmission channels and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal.

[0222] In some embodiments, the frame type corresponding to each transmission channel is determined based on the N transient state detection results. Spatial decoding is performed on the frequency-domain signals of the N transmission channels based on the frame type corresponding to each transmission channel and the spatial encoding parameter to obtain the reconstructed frequency-domain three-dimensional audio signal.

[0223] For an implementation process of performing the spatial decoding on the frequency-domain signals of the N transmission channels based on the frame type corresponding to each transmission channel and the spatial encoding parameter, refer to a related technology.

Details are not described in this embodiment of this application.

[0224] **Step 1404:** Determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

[0225] For an implementation of determining the reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal, refer to related descriptions in step 904. Details are not described herein again.

[0226] In this embodiment of this application, the decoder side parses the bitstream to obtain the global transient state detection result, the transient state detection result corresponding to each transmission channel, and the spatial encoding parameter. In this way, when decoding is performed based on the transient state detection result corresponding to each transmission channel, the frequency-domain signal of each transmission channel can be accurately obtained. In addition, when a target encoding parameter is not encoded into the bitstream, the target encoding parameter may be directly determined based on the global transient state detection result, to reconstruct the time-domain three-dimensional audio signal.

[0227] FIG. 15 is a block diagram of another example decoding method according to an embodiment of this application. FIG. 15 mainly describes an example of the decoding method shown in FIG. 14. In FIG. 15, a global transient state detection result, N transient state detection results corresponding to N transmission channels, and a bitstream are parsed to obtain a spatial encoding parameter. Decoding is performed based on the N transient state detection results and the bistream to obtain frequency-domain signals of the N transmission channels. Spatial decoding is performed on the frequency-domain signals of the N transmission channels based on the frequency-domain signals of the N transmission channels and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal. A reconstructed time-domain three-dimensional audio signal is determined based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

[0228] FIG. 16 is a schematic diagram of a structure of an encoding apparatus according to an embodiment of this application. The encoding apparatus may be implemented as a part or entire of an encoder side device by using software, hardware, or a combination thereof. The encoder side device may be the source apparatus shown in FIG. 1. As shown in FIG. 16, the apparatus includes: a transient state detection module 1601, a determining module 1602, a conversion module 1603, a spatial encoding module 1604, a first encoding module 1605, a second encoding module 1606, and a first writing module 1607.

[0229] The transient state detection module 1601 is configured to separately perform transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0230] The determining module 1602 is configured to determine a global transient state detection result based on the M transient state detection results. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0231] The conversion module 1603 is configured to convert the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0232] The spatial encoding module 1604 is configured to perform spatial encoding on the frequency-domain three-dimensional audio signal based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0233] The first encoding module 1605 is configured to encode the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0234] The second encoding module 1606 is configured to encode the spatial encoding parameter to obtain a spatial encoding parameter encoding result. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0235] The first writing module 1607 is configured to write the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bistream. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0236] Optionally, the conversion module 1603 includes:

a determining unit, configured to determine a target encoding parameter based on the global transient state detection result, where the target encoding parameter includes a window function type of a current frame and/or a frame type of the current frame; and

a conversion unit, configured to convert the time-domain three-dimensional audio signal into the frequency-domain three-dimensional audio signal based on the target encoding parameter.

[0237] Optionally, the global transient state detection result includes a global transient state flag. The target encoding parameter includes the window function type of the current frame.

[0238] The determining unit is specifically configured to:

if the global transient state flag is a first value, determine a type of a first preset window function as the window function type of the current frame; or
if the global transient state flag is a second value, determine a type of a second preset window function as the window function type of the current frame.

[0239] A window length of the first preset window function is less than a window length of the second preset window function.

[0240] Optionally, the global transient state detection result includes a global transient state flag and global transient state position information. The target encoding parameter includes the window function type of the current frame.

[0241] The determining unit is specifically configured to:

if the global transient state flag is the first value, determine the window function type of the current frame based on the global transient state position information.

[0242] Optionally, the apparatus further includes:

a third encoding module, configured to encode the target encoding parameter to obtain a target encoding parameter encoding result; and
a second writing module, configured to write the target encoding parameter encoding result into the bitstream.

[0243] Optionally, the spatial encoding module 1604 is specifically configured to:

perform spatial encoding on the frequency-domain three-dimensional audio signal based on the frame type.

[0244] Optionally, the first encoding module 1605 is specifically configured to:

encode the frequency-domain signals of the N transmission channels based on the frame type of the current frame.

[0245] Optionally, the transient state detection result includes the transient state flag. The global transient state detection result includes the global transient state flag. The transient state flag indicates whether a signal of a corresponding channel is a transient state signal.

[0246] The determining module 1602 is specifically configured to:

if a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to m, determine that the global transient state flag is the first value, where m is a positive integer greater than 0 and less than M; or
if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M channels is greater than or equal to n, determine that the global transient state flag is the first value, where n is a positive integer greater than 0 and less than M.

[0247] Optionally, the transient state detection result further includes transient state position information. The global transient state detection result further includes the global transient state position information. The transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel.

[0248] The determining module 1602 is specifically configured to:

if only one transient state flag in the M transient state flags is the first value, determine transient state position information corresponding to a channel whose transient state flag is the first value as the global transient state position information; or
if at least two transient state flags in the M transient state flags are the first value, determine transient state position information, as the global transient state position information, corresponding to a channel with a largest transient state detection parameter in at least two channels corresponding to the at least two transient state flags.

[0249] Optionally, the apparatus further includes:

a fourth encoding module, configured to encode the global transient state detection result to obtain a global transient state detection result encoding result; and
a third writing module, configured to write the global transient state detection result encoding result into the bistream.

[0250] In this embodiment of this application, transient state detection may be performed on the signals of the M channels included in the time-domain three-dimensional audio signal of the current frame, to determine the global transient state detection result. Then, based on the global transient state detection result, time-frequency transform and spatial encoding of the audio signal are sequentially performed, and the frequency-domain signal of each transmission channel is encoded. Especially, when the frequency-domain signal of each transmission channel obtained through the spatial encoding is encoded, the global transient state detection result is used as the transient state detection result of each transmission

channel, and the frequency-domain signal of each transmission channel does not need to be converted into a time domain to determine the transient state detection result corresponding to each transmission channel. Therefore, the three-dimensional audio signal does not need to be transformed between a time domain and a frequency domain for a plurality of times. This can reduce encoding complexity and improve encoding efficiency. In addition, in this embodiment of this application, the transient state detection result of each transmission channel does not need to be encoded, and only the global transient state detection result needs to be encoded into the bitstream. This can reduce a quantity of bits for encoding.

[0251] It should be noted that, when the encoding apparatus provided in the foregoing embodiment performs encoding, division of the foregoing functional modules is merely used as an example for description. In actual application, the foregoing functions may be allocated to different functional modules for implementation based on a requirement. In other words, an internal structure of the apparatus is divided into different functional modules, to implement all or some of the functions described above. In addition, the encoding apparatus provided in the foregoing embodiment has a same concept as the encoding method embodiment. For details about a specific implementation process of the encoding apparatus, refer to the method embodiment. Details are not described herein again.

[0252] FIG. 17 is a schematic diagram of a structure of a decoding apparatus according to an embodiment of this application. The decoding apparatus may be implemented as a part or entire of a decoder side device by using software, hardware, or a combination thereof. The decoder side device may be the destination apparatus shown in FIG. 1. As shown in FIG. 17, the apparatus includes a parsing module 1701, a decoding module 1702, a spatial decoding module 1703, and a determining module 1704.

[0253] The parsing module 1701 is configured to parse a bitstream to obtain a global transient state detection result and a spatial encoding parameter. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0254] The decoding module 1702 is configured to perform decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0255] The spatial decoding module 1703 is configured to perform spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0256] The determining module 1704 is configured to determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal. For a detailed implementation process, refer to corresponding content in the foregoing embodiments. Details are not described herein again.

[0257] Optionally, the determining module 1704 includes:

a determining unit, configured to determine a target encoding parameter based on the global transient state detection result, where the target encoding parameter includes a window function type of a current frame and/or a frame type of the current frame; and

a conversion unit, configured to convert the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter.

[0258] Optionally, the global transient state detection result includes a global transient state flag. The target encoding parameter includes the window function type of the current frame.

[0259] The determining unit is specifically configured to:

if the global transient state flag is a first value, determine a type of a first preset window function as the window function type of the current frame; or

if the global transient state flag is a second value, determine a type of a second preset window function as the window function type of the current frame.

[0260] A window length of the first preset window function is less than a window length of the second preset window function.

[0261] Optionally, the global transient state detection result includes a global transient state flag and global transient state position information. The target encoding parameter includes the window function type of the current frame.

[0262] The determining unit is specifically configured to:

if the global transient state flag is the first value, determine the window function type of the current frame based on the global transient state position information.

[0263] In this embodiment of this application, a decoder side parses the bitstream to obtain the global transient state detection result and the spatial encoding parameter. In this way, the time-domain three-dimensional audio signal can be reconstructed based on the global transient state detection result and the spatial encoding parame-

ter, and there is no need to parse the bitstream to obtain a transient state detection result of each transmission channel. This can reduce decoding complexity and improve decoding efficiency. In addition, when the target encoding parameter is not encoded into the bitstream, the target encoding parameter may be directly determined based on the global transient state detection result, to reconstruct the time-domain three-dimensional audio signal.

[0264] It should be noted that, when the decoding apparatus provided in the foregoing embodiment performs decoding, division of the foregoing functional modules is merely used as an example for description. In actual application, the foregoing functions may be allocated to different functional modules for implementation based on a requirement. In other words, an internal structure of the apparatus is divided into different functional modules, to implement all or some of the functions described above. In addition, the decoding apparatus provided in the foregoing embodiment has a same concept as the decoding method embodiment. For details about a specific implementation process of the decoding apparatus, refer to the method embodiment. Details are not described herein again.

[0265] FIG. 18 is a schematic block diagram of an encoding and decoding apparatus 1800 according to an embodiment of this application. The encoding and decoding apparatus 1800 may include a processor 1801, a memory 1802, and a bus system 1803. The processor 1801 and the memory 1802 are connected through the bus system 1803. The memory 1802 is configured to store instructions. The processor 1801 is configured to execute the instructions stored in the memory 1802, to perform the encoding or decoding methods described in embodiments of this application. To avoid repetition, details are not described herein again.

[0266] In this embodiment of this application, the processor 1801 may be a central processing unit (central processing unit, CPU), or the processor 1801 may be another general-purpose processor, a DSP, an ASIC, an FPGA or another programmable logic device, a discrete gate or a transistor logic device, a discrete hardware component, or the like. The general-purpose processor may be a microprocessor, or the processor may be any conventional processor, or the like.

[0267] The memory 1802 may include a ROM device or a RAM device. Any another suitable type of storage device may also be used as the memory 1802. The memory 1802 may include code and data 18021 accessed by the processor 1801 through the bus 1803. The memory 1802 may further include an operating system 18023 and an application 18022. The application 18022 includes at least one program that enables the processor 1801 to perform the encoding or decoding methods described in embodiments of this application. For example, the application 18022 may include applications 1 to N, and further includes an encoding or decoding application (a coding application for short) that performs the encoding or decoding methods described in embodiments of this application.

[0268] In addition to a data bus, the bus system 1803 may further include a power bus, a control bus, a status signal bus, and the like. However, for clear description, various types of buses in the figure are marked as the bus system 1803.

[0269] Optionally, the encoding and decoding apparatus 1800 may further include one or more output devices, such as a display 1804. In an example, the display 1804 may be a touch-sensitive display that combines the display with a touch-sensitive unit operable to sense a touch input. The display 1804 may be connected to the processor 1801 through the bus 1803.

[0270] It should be noted that the encoding and decoding apparatus 1800 may perform the encoding method in embodiments of this application, and may also perform the decoding method in embodiments of this application.

[0271] A person skilled in the art can understand that functions described with reference to various illustrative logical blocks, modules, and algorithm steps disclosed in this specification may be implemented by using hardware, software, firmware, or any combination thereof. If implemented by using software, the functions described with reference to the illustrative logical blocks, modules, and steps may be stored in or transmitted over a computer-readable medium as one or more instructions or code and executed by a hardware-based processing unit. The computer-readable medium may include a computer-readable storage medium, which corresponds to a tangible medium such as a data storage medium, or a communication medium including any medium that facilitates transfer of a computer program from one place to another (for example, according to a communication protocol). In this manner, the computer-readable medium may generally correspond to a non-transitory tangible computer-readable storage medium (1), or a communication medium (2) such as a signal or a carrier. The data storage medium may be any usable medium that can be accessed by one or more computers or one or more processors to retrieve instructions, code, and/or data structures for implementing the technologies described in this application. A computer program product may include a computer-readable medium.

[0272] By way of example and not limitation, such computer-readable storage media may include a RAM, a ROM, an EEPROM, a CD-ROM or another optical disc storage apparatus, a magnetic disk storage apparatus or another magnetic storage apparatus, a flash memory, or any other medium that can store required program code in a form of instructions or data structures and that can be accessed by a computer. In addition, any connection is properly referred to as a computer-readable medium. For example, if an instruction is transmitted from a website, a server, or another remote source through a coaxial cable, an optical fiber, a twisted pair, a digital subscriber line (digital subscriber line, DSL), or a wireless technology such as infrared, radio, or microwave, the co-

axial cable, the optical fiber, the twisted pair, the DSL, or the wireless technology such as infrared, radio, or microwave is included in a definition of the medium. However, it should be understood that the computer-readable storage medium and the data storage medium do not include connections, carriers, signals, or other transitory media, but actually mean non-transitory tangible storage media. A disk and an optical disc used in this specification include a compact disc (CD), a laser disc, an optical disc, a DVD, and a Blu-ray disc, where the disk generally magnetically reproduces data, and the optical disc optically reproduces data by using a laser. A combination of the foregoing items should also be included in the scope of the computer-readable medium.

[0273] An instruction may be executed by one or more processors such as one or more digital signal processors (DSP), a general microprocessor, an application-specific integrated circuit (ASIC), a field programmable gate array (FPGA), or an equivalent integrated circuit or discrete logic circuits. Therefore, the term "processor" used in this specification may refer to the foregoing structure, or any other structure that may be applied to implementation of the technologies described in this specification. In addition, in some aspects, the functions described with reference to the illustrative logical blocks, modules, and steps described in this specification may be provided within dedicated hardware and/or software modules configured for encoding and decoding, or may be incorporated into a combined codec. In addition, the technologies may be completely implemented in one or more circuits or logic elements. In an example, various illustrative logic blocks, units, and modules in the encoder 100 and the decoder 200 may be understood as corresponding circuit devices or logic elements.

[0274] The technologies in embodiments of this application may be implemented in various apparatuses or devices, including a wireless handset, an integrated circuit (IC), or a set of ICs (for example, a chip set). Various components, modules, or units are described in embodiments of this application to emphasize functional aspects of the apparatuses configured to perform the disclosed technologies, but are not necessarily implemented by different hardware units. Actually, as described above, various units may be combined in an encoder and decoder hardware unit in combination with appropriate software and/or firmware, or may be provided by interoperable hardware units (including one or more processors described above).

[0275] In other words, all or a part of the foregoing embodiments may be implemented by using software, hardware, firmware, or any combination thereof. When the software is used for implementing embodiments, all or some of embodiments may be implemented in a form of a computer program product. The computer program product includes one or more computer instructions. When the computer instructions are loaded and executed on the computer, the procedure or functions according to the embodiments of this application are all or partially generated. The computer may be a general-purpose computer, a dedicated computer, a computer network, or other programmable apparatuses. The computer instructions may be stored in a computer-readable storage medium or may be transmitted from a computer-readable storage medium to another computer-readable storage medium. For example, the computer instructions may be transmitted from a website, computer, server, or data center to another website, computer, server, or data center in a wired (for example, a coaxial cable, an optical fiber, or a digital subscriber line (digital subscriber line, DSL)) or wireless (for example, infrared, radio, or microwave) manner. The computer-readable storage medium may be any usable medium accessible by a computer, or a data storage device, such as a server or a data center, integrating one or more usable media. The usable medium may be a magnetic medium (for example, a floppy disk, a hard disk, or a magnetic tape), an optical medium (for example, a digital versatile disc (digital versatile disc, DVD)), a semiconductor medium (for example, a solid-state drive (solid state drive, SSD)), or the like. It should be noted that the computer-readable storage medium mentioned in this embodiment of this application may be a non-volatile storage medium, or in other words, may be a non-transitory storage medium.

[0276] It should be understood that "a plurality of" in this specification means two or more. In the descriptions of embodiments of this application, "/" means "or" unless otherwise specified. For example, A/B may represent A or B. In this specification, "and/or" describes only an association relationship between associated objects and represents that three relationships may exist. For example, A and/or B may represent the following three cases: Only A exists, both A and B exist, and only B exists. In addition, to clearly describe the technical solutions in embodiments of this application, terms such as "first" and "second" are used in embodiments of this application to distinguish between same items or similar items that provide basically same functions or purposes. A person skilled in the art may understand that the terms such as "first" and "second" do not limit a quantity or an execution sequence, and the terms such as "first" and "second" do not indicate a definite difference.

[0277] The foregoing descriptions are merely embodiments of this application, but are not intended to limit this application. Any modification, equivalent replacement, or improvement made without departing from the spirit and principle of this application should fall within the protection scope of this application.

Claims

1. An encoding method, wherein the method comprises:

   separately performing transient state detection on signals of M channels comprised in a time-

domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, wherein M is an integer greater than 1;
determining a global transient state detection result based on the M transient state detection results;
converting the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result;
performing spatial encoding on the frequency-domain three-dimensional audio signal based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, wherein N is an integer greater than or equal to 1 and less than or equal to M;
encoding the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result;
encoding the spatial encoding parameter to obtain a spatial encoding parameter encoding result; and
writing the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bistream.

2. The method according to claim 1, wherein the converting the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result comprises:

determining a target encoding parameter based on the global transient state detection result, wherein the target encoding parameter comprises a window function type of the current frame and/or a frame type of the current frame; and
converting the time-domain three-dimensional audio signal into the frequency-domain three-dimensional audio signal based on the target encoding parameter.

3. The method according to claim 2, wherein the global transient state detection result comprises a global transient state flag, and the target encoding parameter comprises the window function type of the current frame; and
the determining a target encoding parameter based on the global transient state detection result comprises:

if the global transient state flag is a first value, determining a type of a first preset window function as the window function type of the current frame; or
if the global transient state flag is a second value, determining a type of a second preset window function as the window function type of the current frame, wherein
a window length of the first preset window function is less than a window length of the second preset window function.

4. The method according to claim 2, wherein the global transient state detection result comprises a global transient state flag and global transient state position information, and the target encoding parameter comprises the window function type of the current frame; and
the determining a target encoding parameter based on the global transient state detection result comprises:
if the global transient state flag is a first value, determining the window function type of the current frame based on the global transient state position information.

5. The method according to any one of claims 2 to 4, wherein the method further comprises:

encoding the target encoding parameter to obtain a target encoding parameter encoding result; and
writing the target encoding parameter encoding result into the bistream.

6. The method according to any one of claims 2 to 5, wherein the performing spatial encoding on the frequency-domain three-dimensional audio signal based on the global transient state detection result comprises:
performing spatial encoding on the frequency-domain three-dimensional audio signal based on the frame type.

7. The method according to any one of claims 2 to 6, wherein the encoding the frequency-domain signals of the N transmission channels based on the global transient state detection result comprises:
encoding the frequency-domain signals of the N transmission channels based on the frame type of the current frame.

8. The method according to any one of claims 1 to 7, wherein the transient state detection result comprises a transient state flag, the global transient state detection result comprises the global transient state flag, and the transient state flag indicates whether a signal of a corresponding channel is a transient state signal; and
the determining a global transient state detection result based on the M transient state detection results comprises:

if a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to m, determining that the global transient state flag is the first value, wherein m is a positive integer greater than 0 and less than M; or
if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M channels is greater than or equal to n, determining that the global transient state flag is the first value, wherein n is a positive integer greater than 0 and less than M.

9. The method according to claim 8, wherein the transient state detection result further comprises transient state position information, the global transient state detection result further comprises global transient state position information, and the transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel; and
the determining a global transient state detection result based on the M transient state detection results comprises:

if only one transient state flag in the M transient state flags is the first value, determining transient state position information corresponding to a channel whose transient state flag is the first value as the global transient state position information; or
if at least two transient state flags in the M transient state flags are the first value, determining transient state position information, as the global transient state position information, corresponding to a channel with a largest transient state detection parameter in at least two channels corresponding to the at least two transient state flags.

10. The method according to any one of claims 1 to 9, wherein the method further comprises:

encoding the global transient state detection result to obtain a global transient state detection result encoding result; and
writing the global transient state detection result encoding result into the bistream.

11. A decoding method, wherein the method comprises:

parsing a bitstream to obtain a global transient state detection result and a spatial encoding parameter;
performing decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels;
performing spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal; and
determining a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

12. The method according to claim 11, wherein the determining a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal comprises:

determining a target encoding parameter based on the global transient state detection result, wherein the target encoding parameter comprises a window function type of a current frame and/or a frame type of the current frame; and
converting the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter.

13. The method according to claim 12, wherein the global transient state detection result comprises a global transient state flag, and the target encoding parameter comprises the window function type of the current frame; and
the determining a target encoding parameter based on the global transient state detection result comprises:

if the global transient state flag is a first value, determining a type of a first preset window function as the window function type of the current frame; or
if the global transient state flag is a second value, determining a type of a second preset window function as the window function type of the current frame, wherein
a window length of the first preset window function is less than a window length of the second preset window function.

14. The method according to claim 12, wherein the global transient state detection result comprises a global transient state flag and global transient state position information, and the target encoding parameter comprises the window function type of the current frame; and
the determining a target encoding parameter based

on the global transient state detection result comprises:
if the global transient state flag is a first value, determining the window function type of the current frame based on the global transient state position information.

15. An encoding apparatus, wherein the apparatus comprises:

a transient state detection module, configured to separately perform transient state detection on signals of M channels comprised in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, wherein M is an integer greater than 1;
a determining module, configured to determine a global transient state detection result based on the M transient state detection results;
a conversion module, configured to convert the time-domain three-dimensional audio signal into a frequency-domain three-dimensional audio signal based on the global transient state detection result;
a spatial encoding module, configured to perform spatial encoding on the frequency-domain three-dimensional audio signal based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, wherein N is an integer greater than or equal to 1 and less than or equal to M;
a first encoding module, configured to encode the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result;
a second encoding module, configured to encode the spatial encoding parameter to obtain a spatial encoding parameter encoding result; and
a first writing module, configured to write the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bistream.

16. The apparatus according to claim 15, wherein the conversion module comprises:

a determining unit, configured to determine a target encoding parameter based on the global transient state detection result, wherein the target encoding parameter comprises a window function type of the current frame and/or a frame type of the current frame; and
a conversion unit, configured to convert the time-domain three-dimensional audio signal into the frequency-domain three-dimensional audio signal based on the target encoding parameter.

17. The apparatus according to claim 16, wherein the global transient state detection result comprises a global transient state flag, and the target encoding parameter comprises the window function type of the current frame; and
the determining unit is specifically configured to:

if the global transient state flag is a first value, determine a type of a first preset window function as the window function type of the current frame; or
if the global transient state flag is a second value, determine a type of a second preset window function as the window function type of the current frame, wherein
a window length of the first preset window function is less than a window length of the second preset window function.

18. The apparatus according to claim 16, wherein the global transient state detection result comprises a global transient state flag and global transient state position information, and the target encoding parameter comprises the window function type of the current frame; and
the determining unit is specifically configured to:
if the global transient state flag is a first value, determine the window function type of the current frame based on the global transient state position information.

19. The apparatus according to any one of claims 16 to 18, wherein the apparatus further comprises:

a third encoding module, configured to encode the target encoding parameter to obtain a target encoding parameter encoding result; and
a second writing module, configured to write the target encoding parameter encoding result into the bistream.

20. The apparatus according to any one of claims 16 to 19, wherein the spatial encoding module is specifically configured to:
perform spatial encoding on the frequency-domain three-dimensional audio signal based on the frame type.

21. The apparatus according to any one of claims 16 to 20, wherein the first encoding module is specifically configured to:
encode the frequency-domain signals of the N transmission channels based on the frame type of the current frame.

22. The apparatus according to any one of claims 15 to 21, wherein the transient state detection result comprises a transient state flag, the global transient state detection result comprises the global transient state flag, and the transient state flag indicates whether a signal of a corresponding channel is a transient state signal; and
the determining module is specifically configured to:

if a quantity of transient state flags that are the first value in the M transient state flags is greater than or equal to m, determine that the global transient state flag is the first value, wherein m is a positive integer greater than 0 and less than M; or
if a quantity of channels that meet a first preset condition and whose corresponding transient state flags are the first value in the M channels is greater than or equal to n, determine that the global transient state flag is the first value, wherein n is a positive integer greater than 0 and less than M.

23. The apparatus according to claim 22, wherein the transient state detection result further comprises transient state position information, the global transient state detection result further comprises global transient state position information, and the transient state position information indicates a position in which a transient state occurs in the signal of the corresponding channel; and
the determining module is specifically configured to:

if only one transient state flag in the M transient state flags is the first value, determine transient state position information corresponding to a channel whose transient state flag is the first value as the global transient state position information; or
if at least two transient state flags in the M transient state flags are the first value, determine transient state position information, as the global transient state position information, corresponding to a channel with a largest transient state detection parameter in at least two channels corresponding to the at least two transient state flags.

24. The apparatus according to any one of claims 15 to 23, wherein the apparatus further comprises:

a fourth encoding module, configured to encode the global transient state detection result to obtain a global transient state detection result encoding result; and
a third writing module, configured to write the global transient state detection result encoding result into the bistream.

25. A decoding apparatus, wherein the apparatus comprises:

a parsing module, configured to parse a bitstream to obtain a global transient state detection result and a spatial encoding parameter;
a decoding module, configured to perform decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels;
a spatial decoding module, configured to perform spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal; and
a determining module, configured to determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal.

26. The apparatus according to claim 25, wherein the determining module comprises:

a determining unit, configured to determine a target encoding parameter based on the global transient state detection result, wherein the target encoding parameter comprises a window function type of a current frame and/or a frame type of the current frame; and
a conversion unit, configured to convert the reconstructed frequency-domain three-dimensional audio signal into the reconstructed time-domain three-dimensional audio signal based on the target encoding parameter.

27. The apparatus according to claim 26, wherein the global transient state detection result comprises a global transient state flag, and the target encoding parameter comprises the window function type of the current frame; and
the determining unit is specifically configured to:

if the global transient state flag is a first value, determine a type of a first preset window function as the window function type of the current frame; or
if the global transient state flag is a second value, determine a type of a second preset window function as the window function type of the current frame, wherein
a window length of the first preset window function is less than a window length of the second preset window function.

28. The apparatus according to claim 26, wherein the

global transient state detection result comprises a global transient state flag and global transient state position information, and the target encoding parameter comprises the window function type of the current frame; and
the determining unit is specifically configured to:
if the global transient state flag is a first value, determine the window function type of the current frame based on the global transient state position information.

29. An encoder side device, wherein the encoder side device comprises a memory and a processor; and
the memory is configured to store a computer program, and the processor is configured to execute the computer program stored in the memory, to implement the encoding method according to any one of claims 1 to 10.

30. A decoder side device, wherein the decoder side device comprises a memory and a processor; and
the memory is configured to store a computer program, and the processor is configured to execute the computer program stored in the memory, to implement the decoding method according to any one of claims 11 to 14.

31. A computer-readable storage medium, wherein the storage medium stores instructions, and when the instructions run on a computer, the computer is enabled to perform the method according to any one of claims 1 to 14.

32. A computer-readable storage medium, comprising a bistream obtained by using the encoding method according to any one of claims 1 to 10.

33. A computer program, wherein when the computer program is executed, the method according to any one of claims 1 to 14 is implemented.

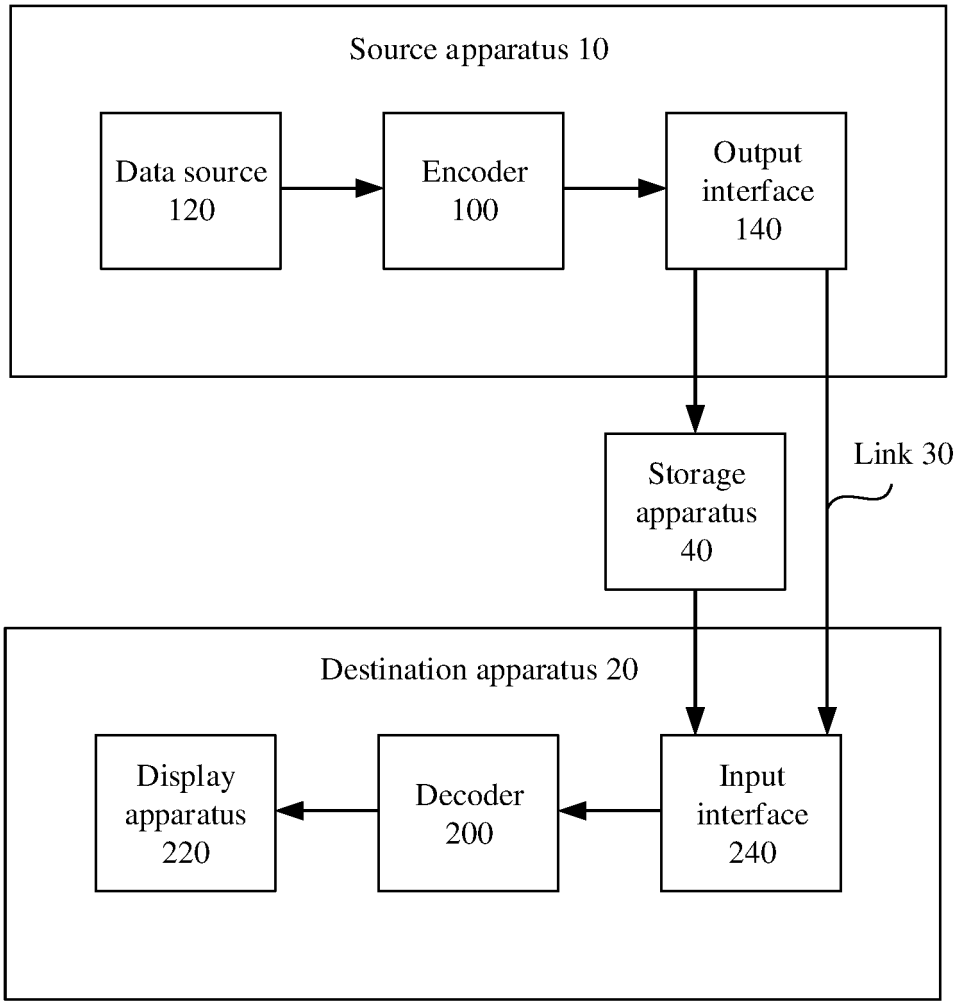
Source apparatus 10
Data source 120
Encoder 100
Output interface 140
Storage apparatus 40
Link 30
Destination apparatus 20
Display apparatus 220
Decoder 200
Input interface 240

FIG. 1

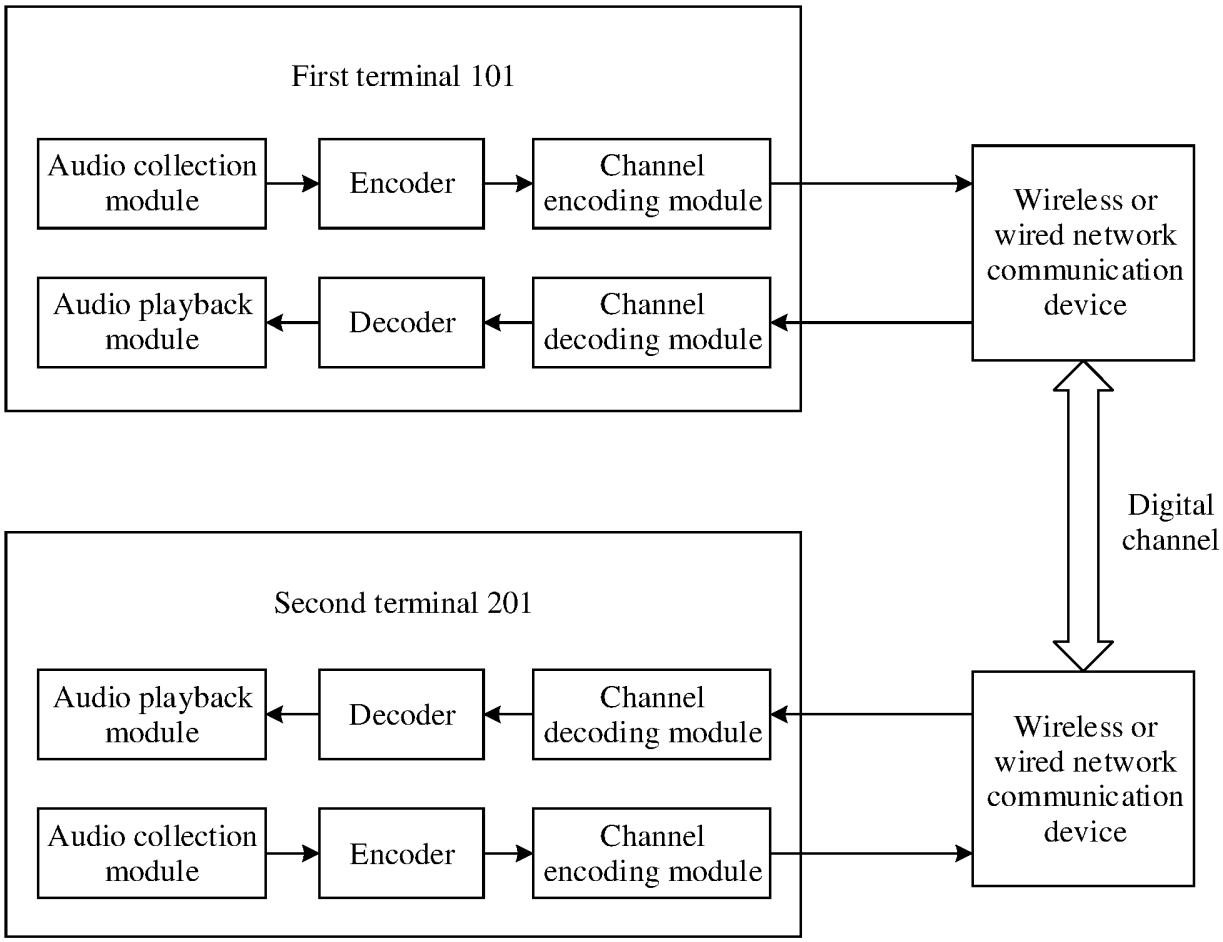
First terminal 101
Audio collection module
Encoder
Channel encoding module
Audio playback module
Decoder
Channel decoding module
Wireless or wired network communication device
Digital channel
Second terminal 201
Audio playback module
Decoder
Channel decoding module
Audio collection module
Encoder
Channel encoding module
Wireless or wired network communication device

FIG. 2

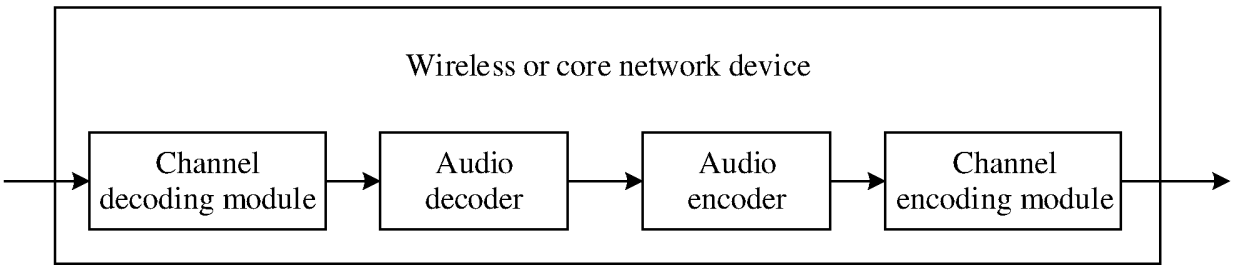
Wireless or core network device
Channel decoding module
Audio decoder
Audio encoder
Channel encoding module

FIG. 3

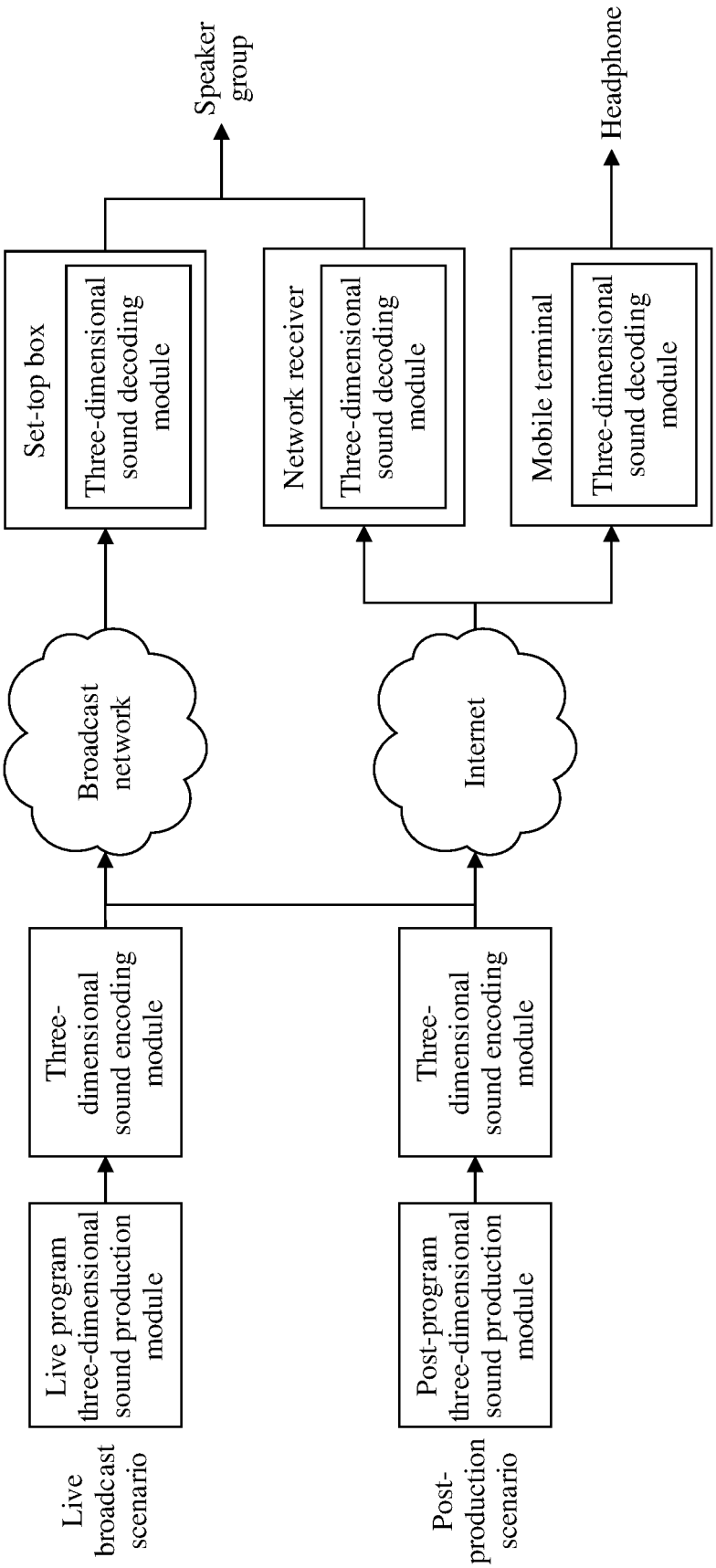
Live broadcast scenario
Post-production scenario
Live program three-dimensional sound production module
Post-program three-dimensional sound production module
Three-dimensional sound encoding module
Three-dimensional sound encoding module
Broadcast network
Internet
Set-top box
Three-dimensional sound decoding module
Network receiver
Three-dimensional sound decoding module
Mobile terminal
Three-dimensional sound decoding module
Speaker group
Headphone

FIG. 4

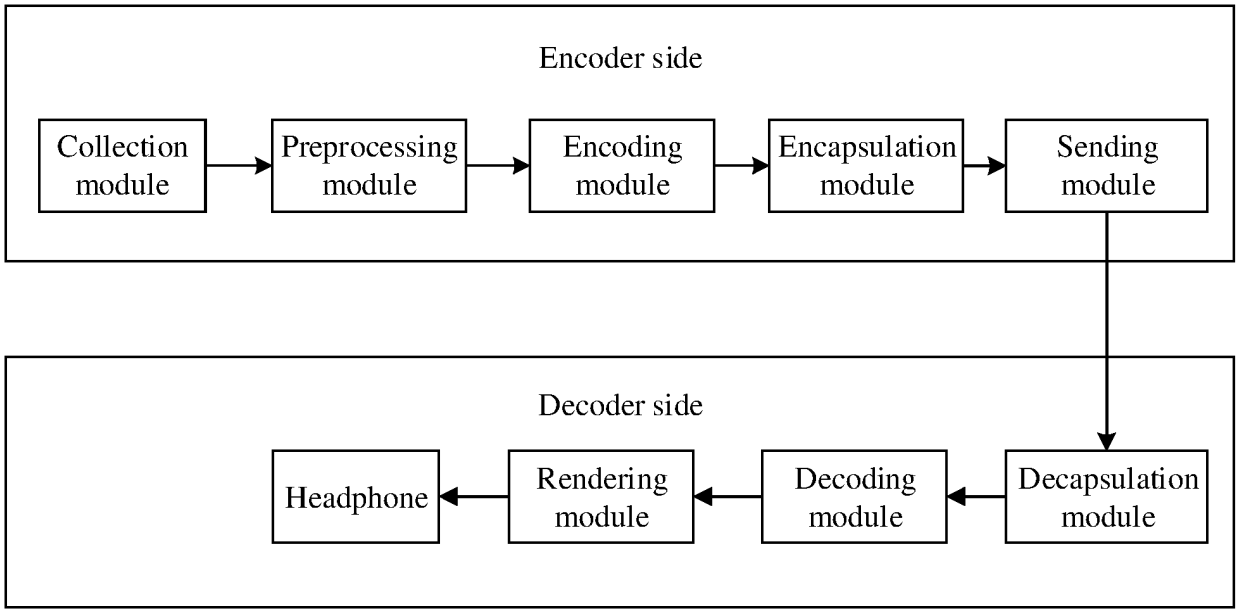
Encoder side
Collection module
Preprocessing module
Encoding module
Encapsulation module
Sending module
Decoder side
Headphone
Rendering module
Decoding module
Decapsulation module

FIG. 5

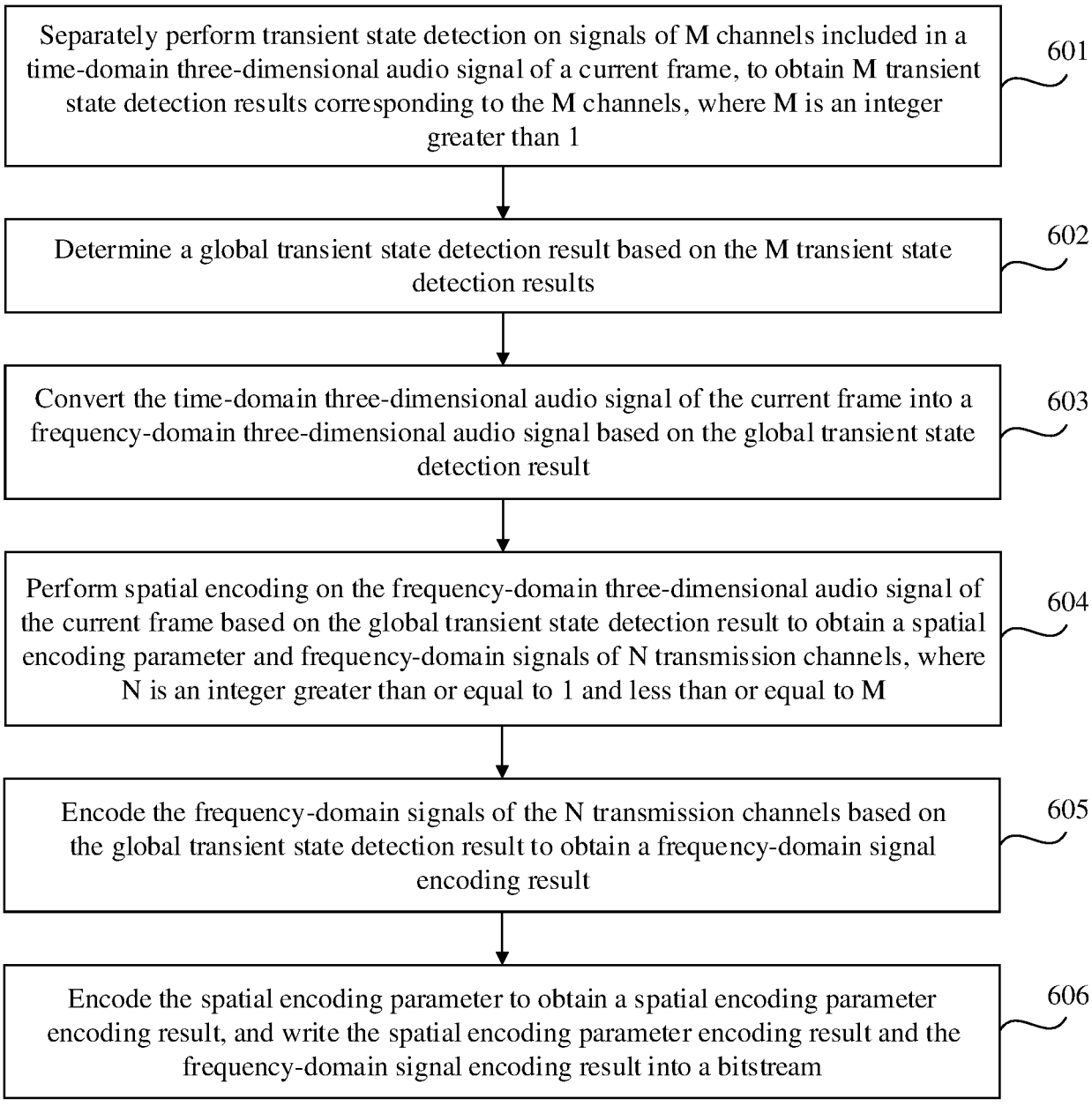
Separately perform transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1
601
Determine a global transient state detection result based on the M transient state detection results
602
Convert the time-domain three-dimensional audio signal of the current frame into a frequency-domain three-dimensional audio signal based on the global transient state detection result
603
Perform spatial encoding on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M
604
Encode the frequency-domain signals of the N transmission channels based on the global transient state detection result to obtain a frequency-domain signal encoding result
605
Encode the spatial encoding parameter to obtain a spatial encoding parameter encoding result, and write the spatial encoding parameter encoding result and the frequency-domain signal encoding result into a bitstream
606

FIG. 6

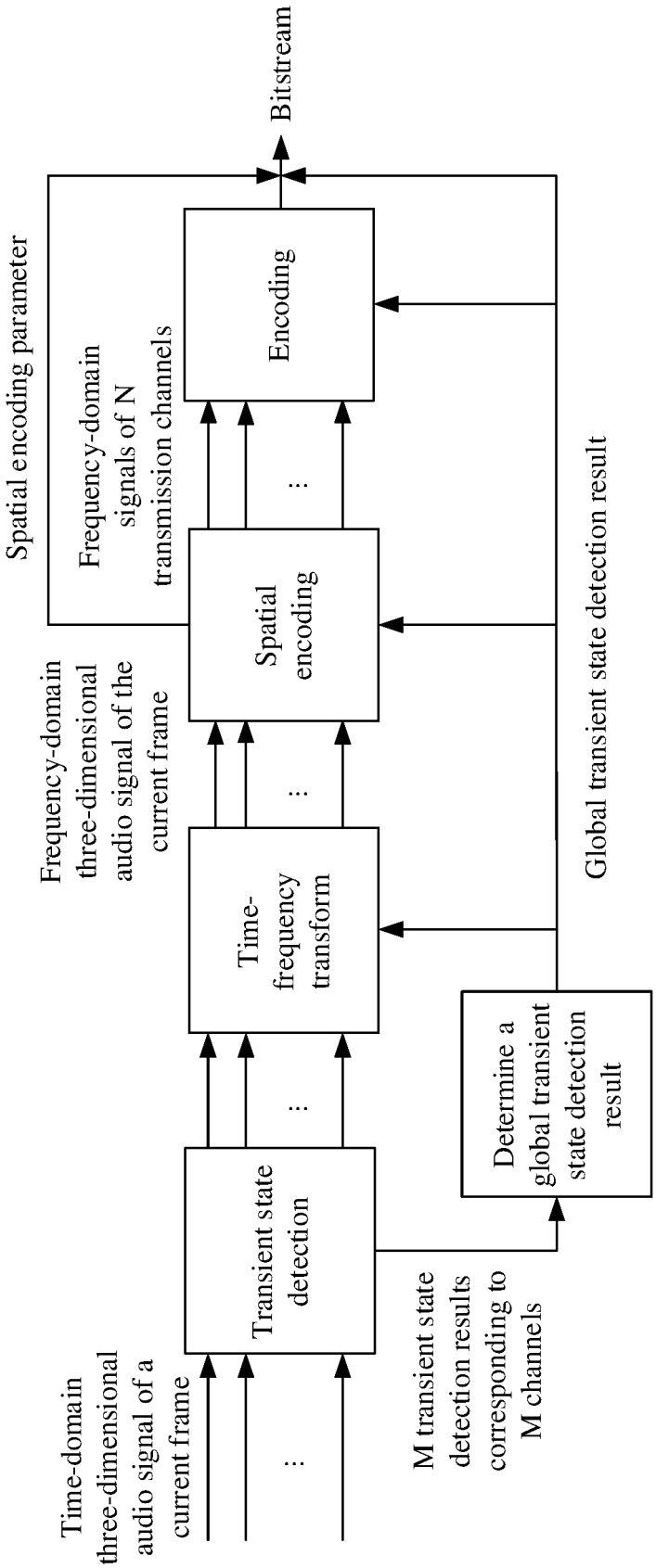
Time-domain three-dimensional audio signal of a current frame
Transient state detection
Time-frequency transform
Frequency-domain three-dimensional audio signal of the current frame
Spatial encoding
Frequency-domain signals of N transmission channels
Encoding
Bitstream
Spatial encoding parameter
M transient state detection results corresponding to M channels
Determine a global transient state detection result
Global transient state detection result

FIG. 7

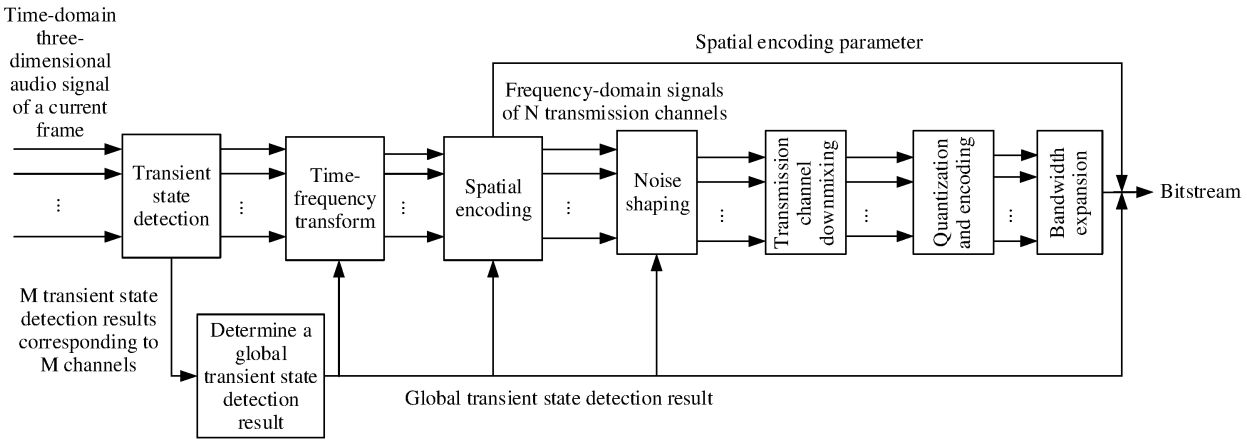
Time-domain three-dimensional audio signal of a current frame
Transient state detection
Time-frequency transform
Spatial encoding
Noise shaping
Transmission channel downmixing
Quantization and encoding
Bandwidth expansion
Bitstream
Spatial encoding parameter
Frequency-domain signals of N transmission channels
M transient state detection results corresponding to M channels
Determine a global transient state detection result
Global transient state detection result

FIG. 8

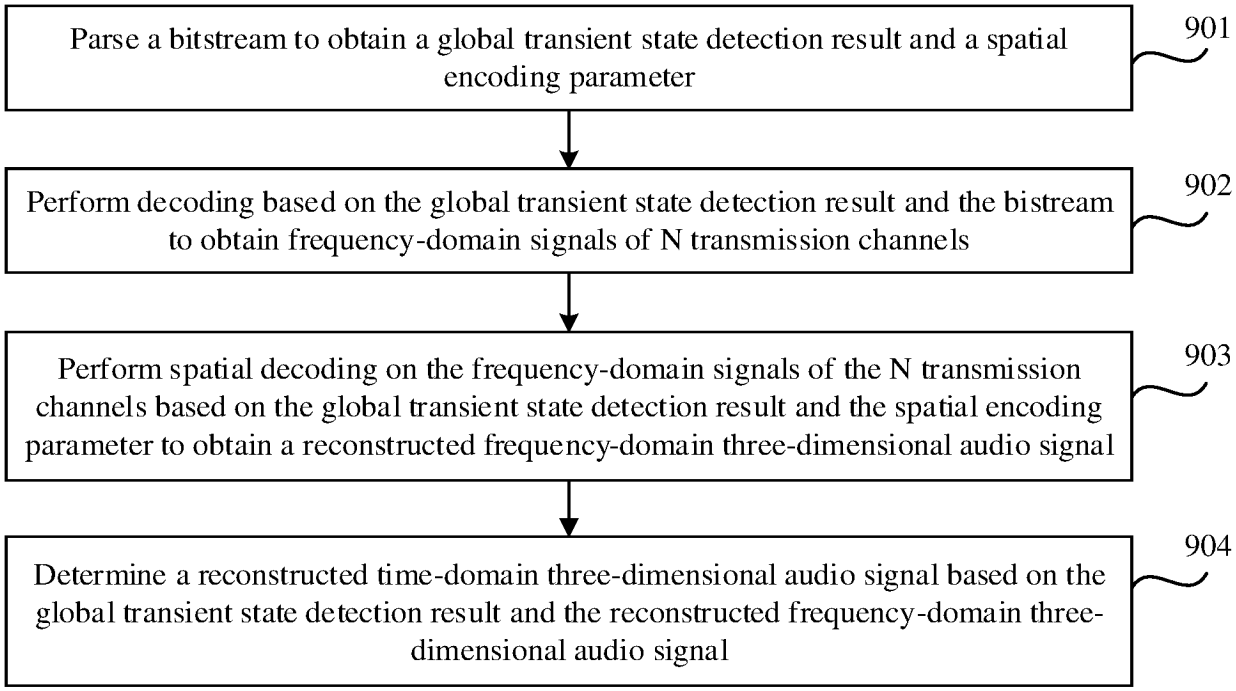
Parse a bitstream to obtain a global transient state detection result and a spatial encoding parameter
901
Perform decoding based on the global transient state detection result and the bistream to obtain frequency-domain signals of N transmission channels
902
Perform spatial decoding on the frequency-domain signals of the N transmission channels based on the global transient state detection result and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal
903
Determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal
904

FIG. 9

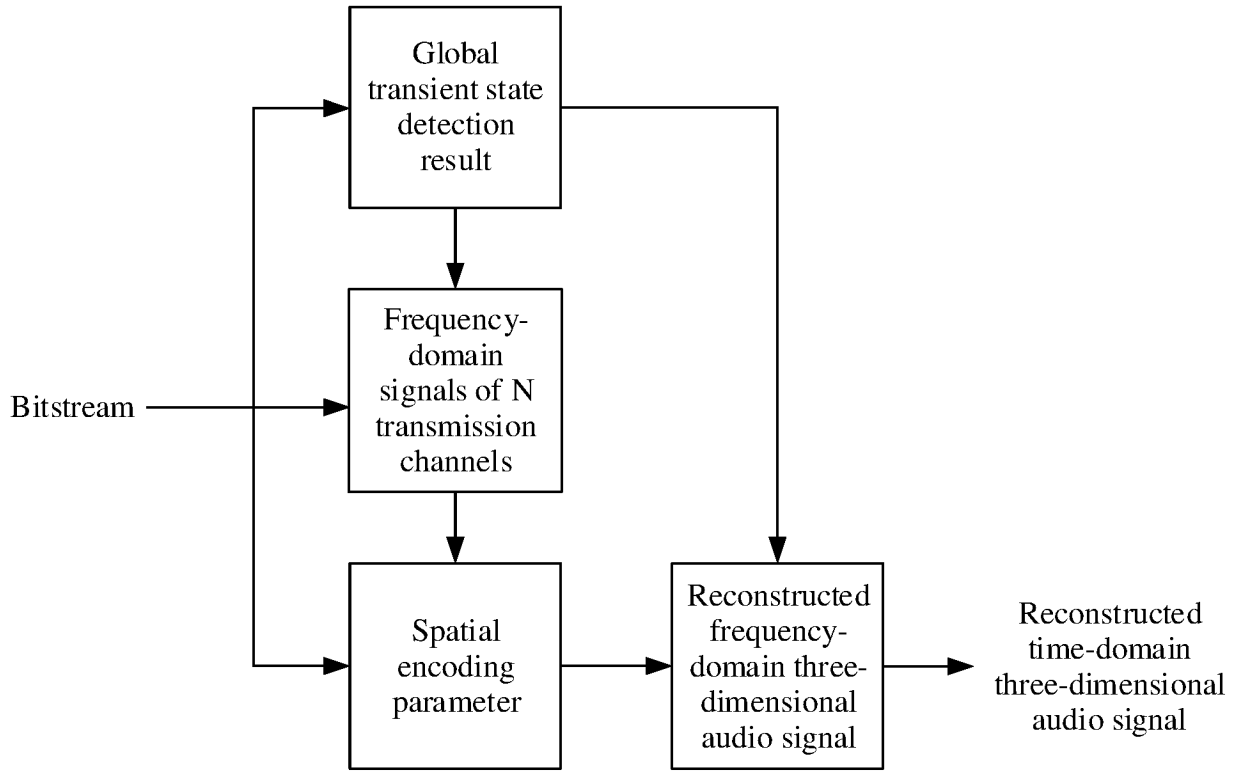
Global transient state detection result
Frequency-domain signals of N transmission channels
Bitstream
Spatial encoding parameter
Reconstructed frequency-domain three-dimensional audio signal
Reconstructed time-domain three-dimensional audio signal

FIG. 10

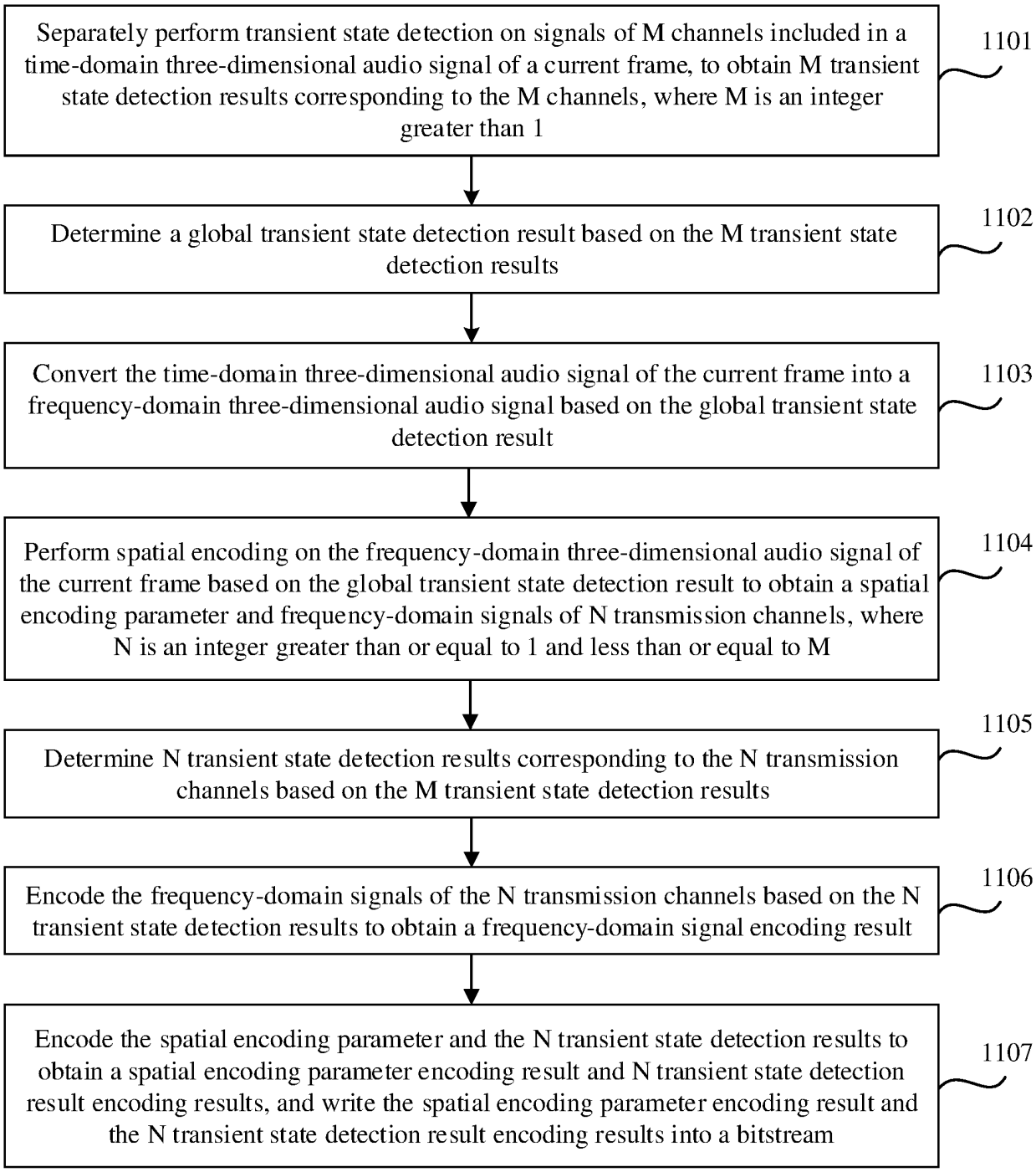
Separately perform transient state detection on signals of M channels included in a time-domain three-dimensional audio signal of a current frame, to obtain M transient state detection results corresponding to the M channels, where M is an integer greater than 1
1101
Determine a global transient state detection result based on the M transient state detection results
1102
Convert the time-domain three-dimensional audio signal of the current frame into a frequency-domain three-dimensional audio signal based on the global transient state detection result
1103
Perform spatial encoding on the frequency-domain three-dimensional audio signal of the current frame based on the global transient state detection result to obtain a spatial encoding parameter and frequency-domain signals of N transmission channels, where N is an integer greater than or equal to 1 and less than or equal to M
1104
Determine N transient state detection results corresponding to the N transmission channels based on the M transient state detection results
1105
Encode the frequency-domain signals of the N transmission channels based on the N transient state detection results to obtain a frequency-domain signal encoding result
1106
Encode the spatial encoding parameter and the N transient state detection results to obtain a spatial encoding parameter encoding result and N transient state detection result encoding results, and write the spatial encoding parameter encoding result and the N transient state detection result encoding results into a bitstream
1107

FIG. 11

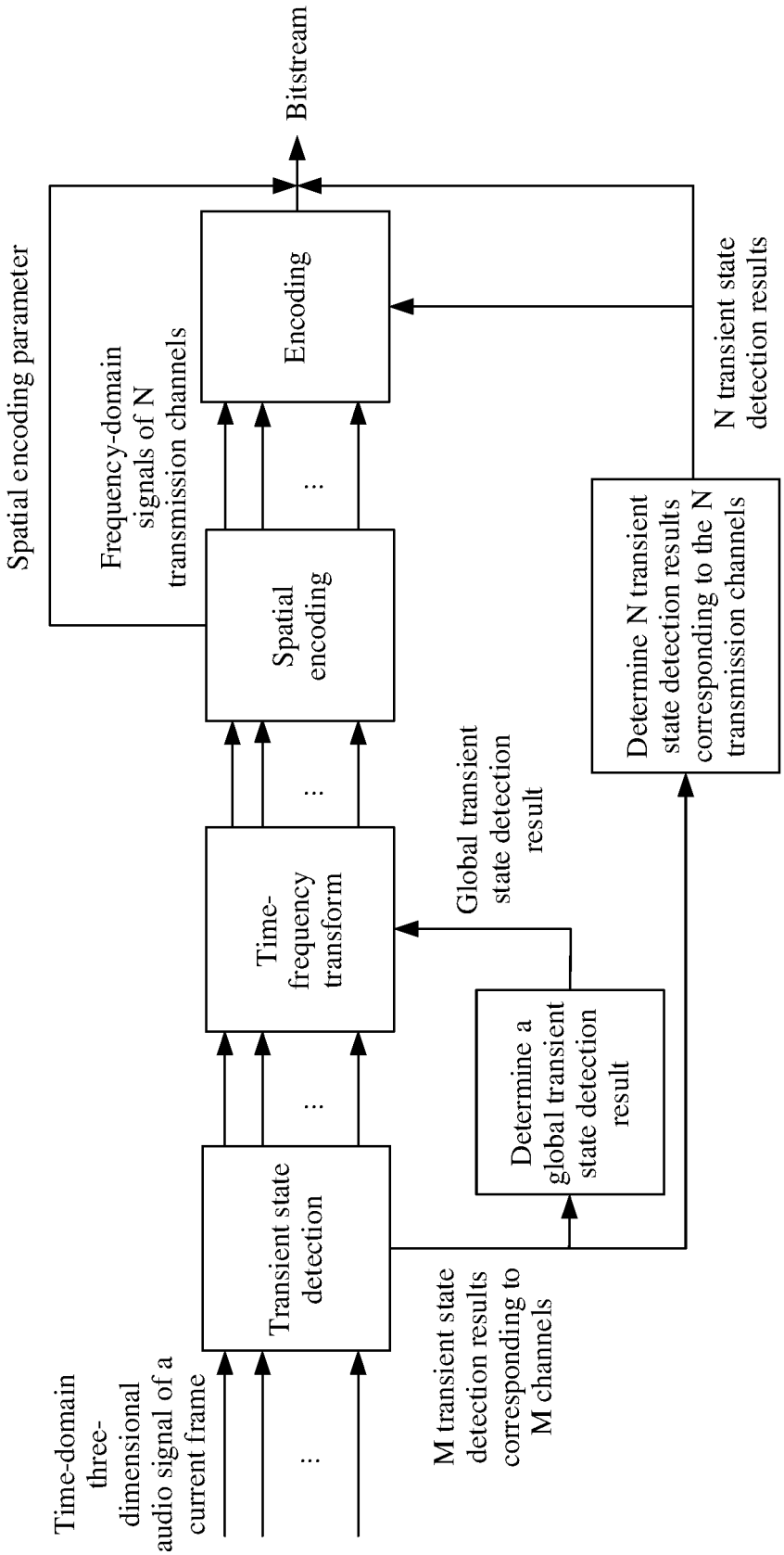
Time-domain three-dimensional audio signal of a current frame
Transient state detection
Time-frequency transform
Spatial encoding
Encoding
Bitstream
Spatial encoding parameter
Frequency-domain signals of N transmission channels
M transient state detection results corresponding to M channels
Determine a global transient state detection result
Global transient state detection result
Determine N transient state detection results corresponding to the N transmission channels
N transient state detection results

FIG. 12

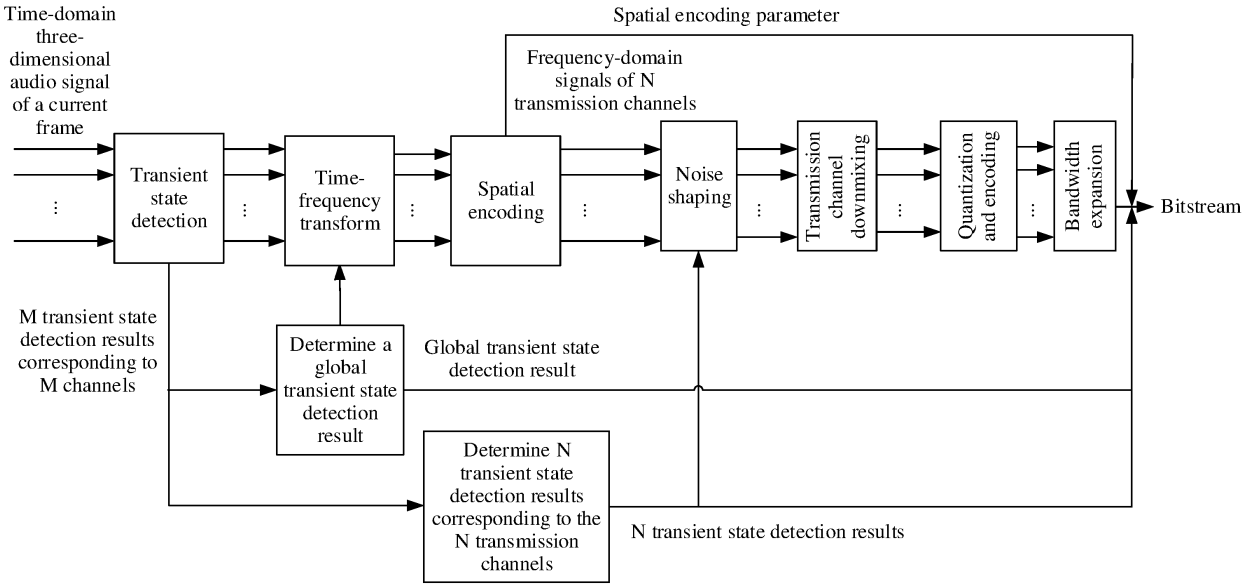
Time-domain three-dimensional audio signal of a current frame
Spatial encoding parameter
Frequency-domain signals of N transmission channels
Transient state detection
Time-frequency transform
Spatial encoding
Noise shaping
Transmission channel downmixing
Quantization and encoding
Bandwidth expansion
Bitstream
M transient state detection results corresponding to M channels
Determine a global transient state detection result
Global transient state detection result
Determine N transient state detection results corresponding to the N transmission channels
N transient state detection results

FIG. 13

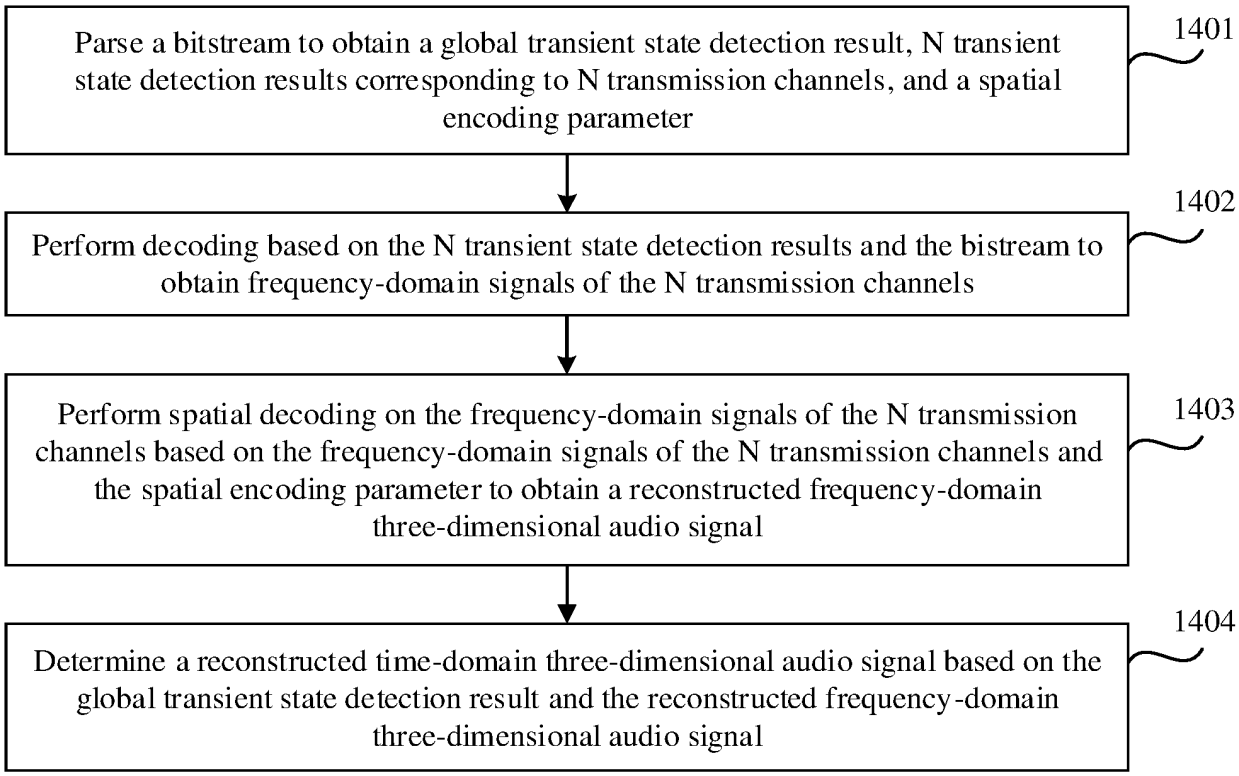
Parse a bitstream to obtain a global transient state detection result, N transient state detection results corresponding to N transmission channels, and a spatial encoding parameter
1401
Perform decoding based on the N transient state detection results and the bistream to obtain frequency-domain signals of the N transmission channels
1402
Perform spatial decoding on the frequency-domain signals of the N transmission channels based on the frequency-domain signals of the N transmission channels and the spatial encoding parameter to obtain a reconstructed frequency-domain three-dimensional audio signal
1403
Determine a reconstructed time-domain three-dimensional audio signal based on the global transient state detection result and the reconstructed frequency-domain three-dimensional audio signal
1404

FIG. 14

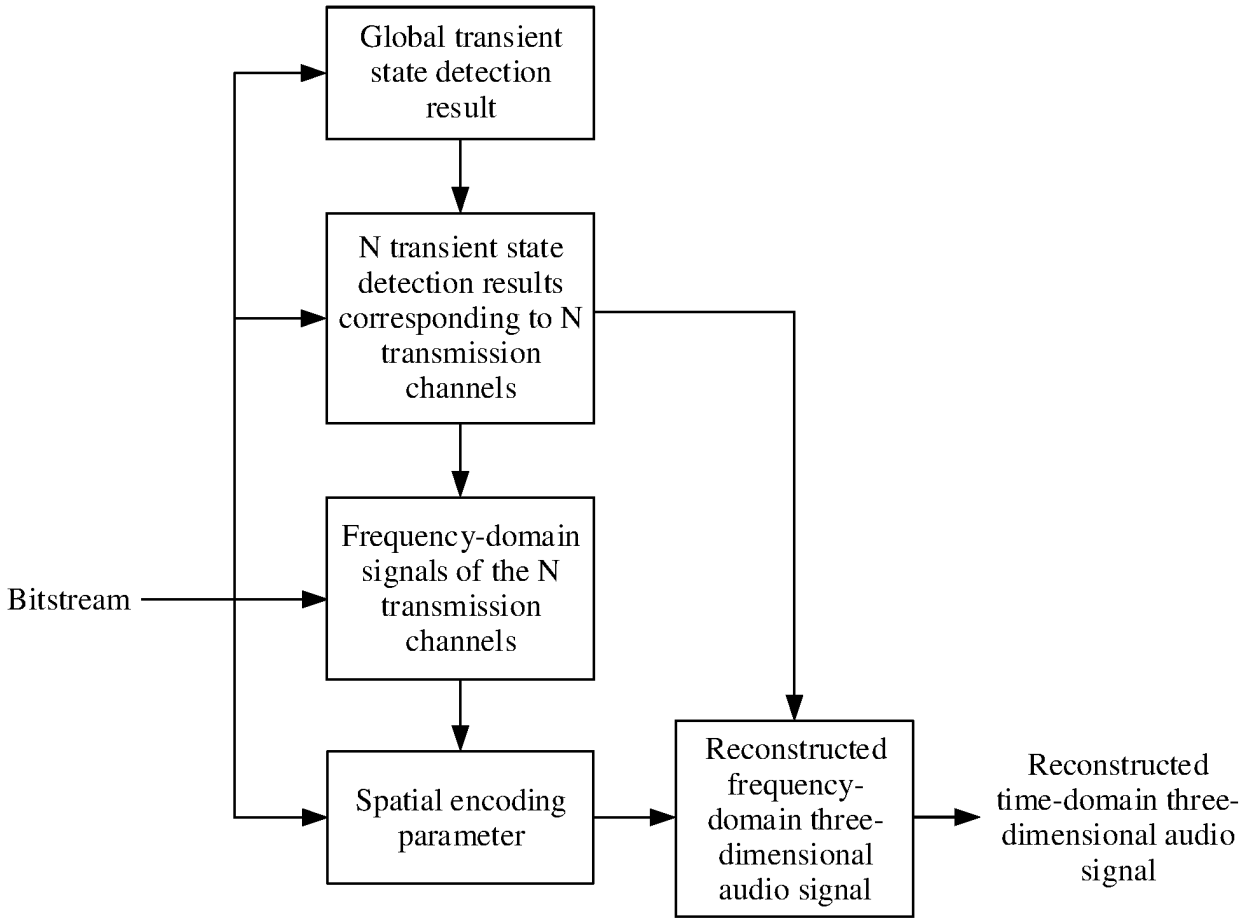
Global transient state detection result
N transient state detection results corresponding to N transmission channels
Frequency-domain signals of the N transmission channels
Bitstream
Spatial encoding parameter
Reconstructed frequency-domain three-dimensional audio signal
Reconstructed time-domain three-dimensional audio signal

FIG. 15

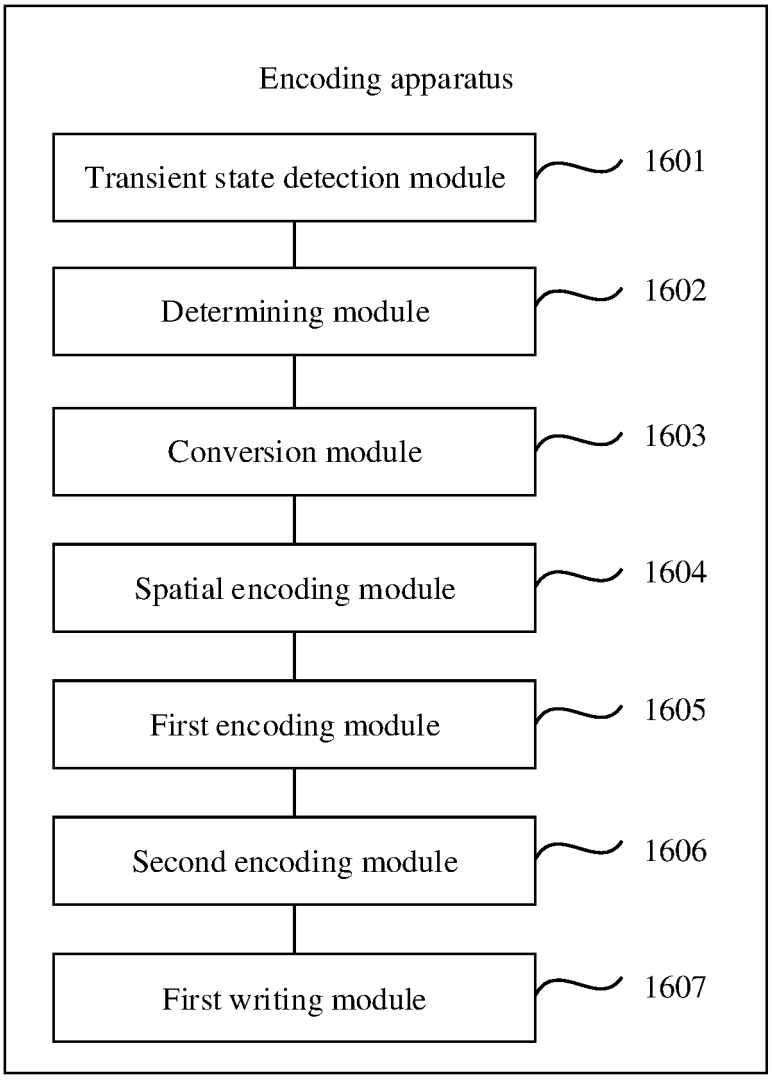
Encoding apparatus
Transient state detection module
1601
Determining module
1602
Conversion module
1603
Spatial encoding module
1604
First encoding module
1605
Second encoding module
1606
First writing module
1607

FIG. 16

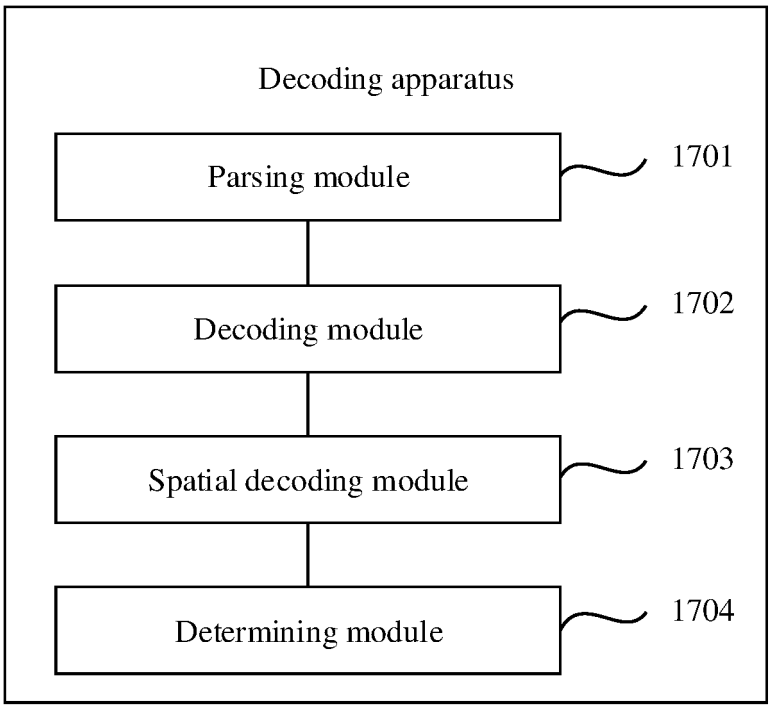
Decoding apparatus
Parsing module
1701
Decoding module
1702
Spatial decoding module
1703
Determining module
1704

FIG. 17

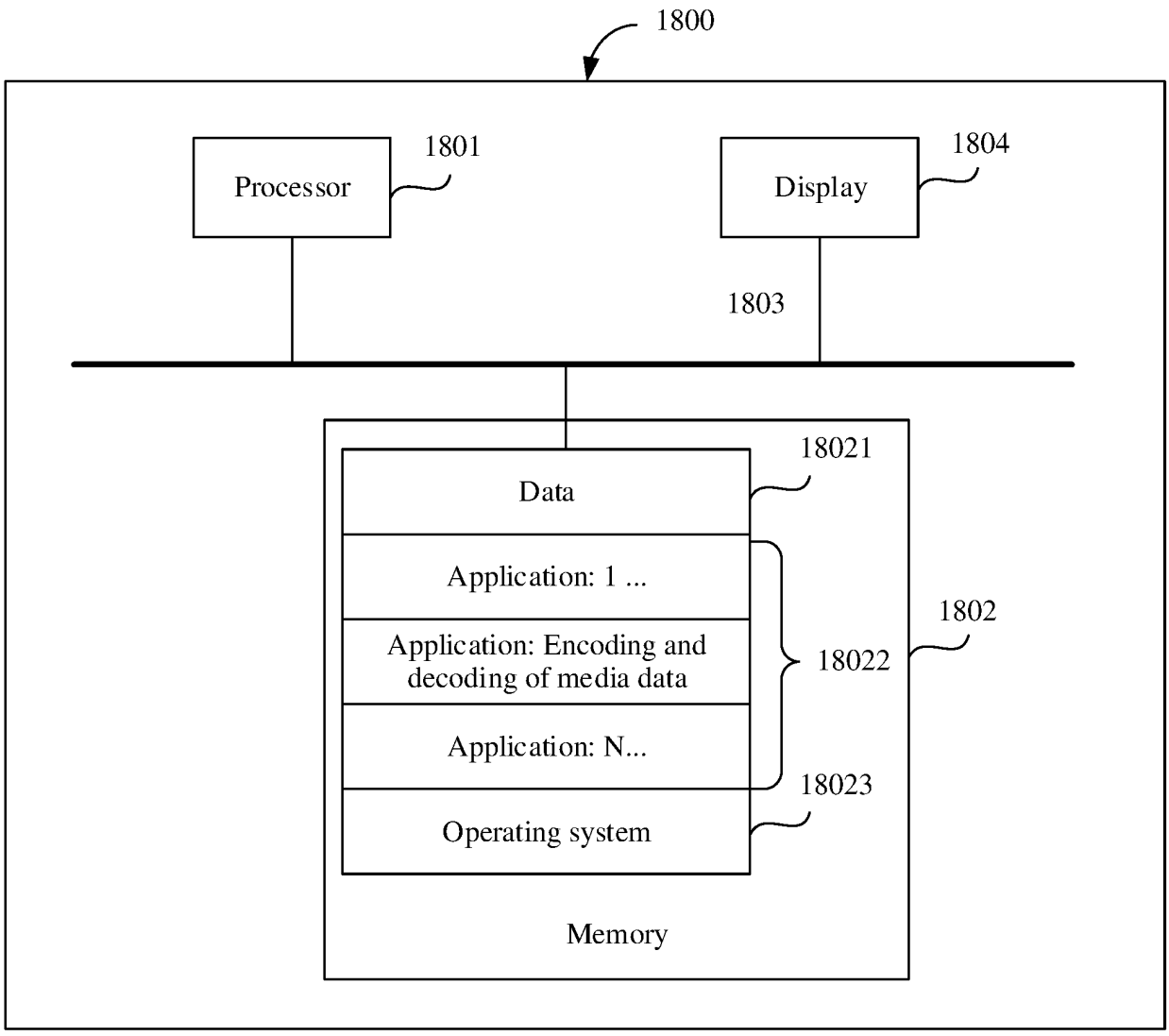
1800
Processor
1801
Display
1804
1803
18021
Data
Application: 1 ...
Application: Encoding and decoding of media data
Application: N...
18022
1802
18023
Operating system
Memory

FIG. 18

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/120507

A. CLASSIFICATION OF SUBJECT MATTER

G10L 19/008(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: 编码, 解码, 三维, 3D, 音频, 通道, 暂态, 检测, 全局, 空间, 频域, 码流, 标志, 窗函数, 类型, code, decode, three-dimension, channel, transient, detection, global, spatial, frequency domain, bit stream, code stream, sign, flag, window, function, type

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 1750407 A (ZHENGYIN DIGITAL TECHNOLOGY CO., LTD., ZHONGSHAN CITY et al.) 22 March 2006 (2006-03-22) description, page 7, line 10 to page 23, line 14, and figures 1-16	1-33
A	CN 101197577 A (SPREADTRUM COMMUNICATIONS SHANGHAI INC.) 11 June 2008 (2008-06-11) entire document	1-33
A	CN 110544484 A (ZHONGKE CHAOYING (BEIJING) MEDIA TECHNOLOGY CO., LTD.) 06 December 2019 (2019-12-06) entire document	1-33
A	CN 103493129 A (FRAUNHOFER-GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG E.V.) 01 January 2014 (2014-01-01) entire document	1-33
A	CN 1783726 A (DIGITAL RISE TECHNOLOGY CO., LTD.) 07 June 2006 (2006-06-07) entire document	1-33
A	US 5687283 A (NEC CORPORATION) 11 November 1997 (1997-11-11) entire document	1-33

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 November 2022	**15 December 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/120507

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 1750407 A	22 March 2006	CN 1750409 A	22 March 2006
		CN 1783727 A	07 June 2006
		CN 1750408 A	22 March 2006
		CN 1750406 A	22 March 2006
		CN 1750405 A	22 March 2006
		CN 1750404 A	22 March 2006
		CN 1750403 A	22 March 2006
		CN 1477872 A	25 February 2004
		CN 1756087 A	05 April 2006
		CN 1750402 A	22 March 2006
		CN 100481736 C	22 April 2009
CN 101197577 A	11 June 2008	None	
CN 110544484 A	06 December 2019	None	
CN 103493129 A	01 January 2014	BR 112013020588 A2	10 July 2018
		AU 2012217216 A1	26 September 2013
		MX 2013009304 A	03 October 2013
		PL 2676270 T3	31 July 2017
		AR 098480 A2	01 June 2016
		MX 327034 B	14 January 2015
		WO 2012110448 A1	23 August 2012
		AR 085217 A1	18 September 2013
		SG 192714 A1	30 September 2013
		CA 2920964 A1	23 August 2012
		PT 2676270 T	02 May 2017
		JP 2014510303 A	24 April 2014
		CA 2827266 A1	23 August 2012
		ZA 201306842 B	28 May 2014
		MY 166006 A	21 May 2018
		KR 20130126708 A	20 November 2013
		RU 2013142072 A	27 March 2015
		KR 20140139630 A	05 December 2014
		ES 2623291 T3	10 July 2017
		TW 201301265 A	01 January 2013
		US 2013332177 A1	12 December 2013
		EP 2676270 A1	25 December 2013
		IN 201302510 P2	06 December 2013
		JP 2014510303 W	24 April 2014
		HK 1192049 A0	08 August 2014
		SG 192714 B	14 August 2014
		KR 101525185 B1	02 June 2015
		AU 2012217216 B2	17 September 2015
		KR 101562281 B1	22 October 2015
		RU 2573231 C2	20 January 2016
		JP 5914527 B2	11 May 2016
		CN 103493129 B	10 August 2016
		EP 2676270 B1	10 February 2017
		CA 2827266 C	28 February 2017
		HK 1192049 A1	09 February 2018
		IN 349720 B	30 October 2020
		BR 112013020588 B1	13 July 2021

Form PCT/ISA/210 (patent family annex) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/120507

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 1783726 A	07 June 2006	None	
US 5687283 A	11 November 1997	JP H08314497 A	29 November 1996
		JP 2728122 B2	18 March 1998

Form PCT/ISA/210 (patent family annex) (January 2015)

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- CN 202111155355 **[0001]**