



(11) **EP 3 762 923 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
10.07.2024 Bulletin 2024/28

(51) International Patent Classification (IPC):
G10L 21/028 ^(2013.01) **G10L 19/008** ^(2013.01)
G10L 19/02 ^(2013.01)

(21) Application number: **18723570.0**

(52) Cooperative Patent Classification (CPC):
G10L 21/028; G10L 19/008; G10L 19/0204

(22) Date of filing: **08.03.2018**

(86) International application number:
PCT/FI2018/050171

(87) International publication number:
WO 2019/170955 (12.09.2019 Gazette 2019/37)

(54) **AUDIO CODING**
AUDIOSIGNALKODIERUNG
CODAGE AUDIO

(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**

(43) Date of publication of application:
13.01.2021 Bulletin 2021/02

(73) Proprietor: **Nokia Technologies Oy**
02610 Espoo (FI)

(72) Inventors:
• **VILERMO, Miikka**
37200 Siuro (FI)
• **LAAKSONEN, Lasse Juhani**
33210 Tampere (FI)
• **TAMMI, Mikko**
33310 Tampere (FI)
• **VILKAMO, Juha Tapio**
00120 Helsinki (FI)
• **LAITINEN, Mikko-Ville**
00210 Helsinki (FI)

• **VASILACHE, Adriana**
33580 Tampere (FI)

(74) Representative: **Nokia EPO representatives**
Nokia Technologies Oy
Karakaari 7
02610 Espoo (FI)

(56) References cited:
EP-A1- 2 249 334 US-A1- 2007 269 063
US-A1- 2012 314 878 US-A1- 2013 044 884

• **NIKUNEN JOONAS ET AL: "Binaural rendering of
microphone array captures based on source
separation", SPEECH COMMUNICATION,
ELSEVIER SCIENCE PUBLISHERS,
AMSTERDAM, NL, vol. 76, 21 September 2015
(2015-09-21), pages 157 - 169, XP029400149,
ISSN: 0167-6393, DOI:
10.1016/J.SPECOM.2015.09.005**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description

TECHNICAL FIELD

[0001] The example and embodiments of the present invention relate to processing of audio signals. In particular, various embodiments of the present invention relate to aspects of encoding and/or decoding of audio signals that represent a spatial audio image, i.e. an audio scene that involves one or more directional sound components possibly together with an ambient sound component.

BACKGROUND

[0002] In many applications, digital audio signals representing audio content such as speech or music are encoded to enable, for example, efficient transmission and/or storage of the audio signals. In this regard, audio encoders and audio decoders (also known as audio codecs) are typically employed to encode and/or decode audio-based signals, such as music, ambient sounds or a combination thereof. These types of audio codecs typically do not assume an audio input of certain characteristics and, in particular, do not utilize a speech model for the encoding-decoding process but rather make use of encoding and decoding procedures that are suitable for representing all types of audio signals, including speech. On the other hand, in this type of audio codecs the encoding procedure typically makes use of a hearing model in order to allocate bits for only those parts of the signal that are actually audible for a human listener. In contrast, speech encoders and speech decoders (also known as speech codecs) can be considered as audio codecs that are optimized for speech signals via utilization of a speech production model in the encoding-decoding process. Relying on the speech production model enables, for speech signals, a lower bit rate at perceivable sound quality comparable to that achievable by an audio codec or an improved perceivable sound quality at a bit rate comparable to that of an audio codec. On the other hand, since e.g. music and ambient sounds are typically a poor match with the speech production model, for a speech codec such signals typically represent background noise. An audio codec or a speech codec may operate at either a fixed or variable bit rate.

[0003] A multi-channel audio signal may convey an audio scene that represents both directional sound components at specific positions of the audio scene as well as the ambience of the audio scene. In this regard, directional sound components represent distinct sound sources that have certain position within the audio scene (e.g. a certain direction of arrival and a certain relative intensity with respect to a listening point), whereas the ambience represents environmental sounds within the audio scene. Listening to such an audio scene enables the listener to experience the audio environment as if he or she was at the location the audio scene serves to represent. The audio scene may also be referred to as a spatial audio

image. An audio scene may be stored into a predefined spatial format that enables rendering the audio scene for the listener via headphones and/or via a loudspeaker arrangement. Non-limiting examples of applicable spatial audio formats include a multi-channel audio signal according to a predefined loudspeaker configuration (such as two-channel stereo, 5.1 surround sound, 7.1 surround sound, 22.2 surround sound, etc.), a multi-channel audio signal from a microphone array, an ambisonics audio signal, a binaural audio signal for headphone listening.

[0004] An audio scene may be obtained by using a microphone arrangement that includes a plurality of microphones to capture a respective plurality of audio signals and processing the audio signals into a desired spatial audio format that represents the audio scene. Alternatively, the audio scene may be created on basis of one or more arbitrary source signals by processing them into a desired spatial audio format that represents the audio scene of desired characteristics (e.g. with respect to directionality of sound sources and ambience of the audio scene). As a further example, a combination of a captured and artificially generated audio scene may be provided e.g. by complementing an audio scene captured by a plurality of microphones via introduction of one or more further sound sources at desired spatial positions of the audio scene.

[0005] Some recently developed audio codecs are able to encode a multi-channel input audio signal into an encoded audio signal that is accompanied by spatial information and to decode the encoded audio signal with the aid of the spatial information into reconstructed audio signal such that the spatial audio image represented by the input audio signal is re-created in the reconstructed audio signal. In this disclosure, such audio codec is referred to as a spatial audio codec, which may include a spatial audio encoder and a spatial audio decoder. A spatial audio encoder may provide the encoded audio signal on basis of a single-channel or multi-channel intermediate audio signal derived on basis of one or more channels of the input audio signal. Such an intermediate audio signal has a smaller number of channels than the input audio signal (typically one or two channels) and it is commonly referred to as a downmix audio signal. The spatial information may also be referred to, for example, as spatial data, as spatial metadata, as spatial parameters or as spatial attributes. An example spatial audio codec is disclosed e.g. in the US patent application US2007/0269063, M. Goodwin et al., "Spatial Audio Coding Based On Universal Spatial Cues", 22.11.2007.

[0006] Figure 1 illustrates a block diagram of some elements of a spatial audio encoder 100 according to an example. The spatial audio encoder 100 includes a downmix entity 101 for creating a downmix signal 112 on basis of a multi-channel input audio signal 111 and an audio encoder 102 for processing the downmix signal 112 into an encoded audio signal 113. The spatial audio encoder 100 further includes a metadata derivation entity 103 for deriving spatial metadata 114 on basis of the

multi-channel input audio signal 111 and a metadata encoder 104 for processing the spatial metadata 114 into compressed spatial metadata 115. The spatial audio encoder 100 further includes a multiplexer entity 105 for arranging the encoded audio signal 113 and the compressed spatial metadata 115 into an audio bitstream 116 for storage in a memory and/or for transmission over a communication channel (e.g. a communication network) to a spatial audio decoder for generation of the reconstructed multi-channel audio signal therein.

[0007] The multi-channel input audio signal 111 to the spatial audio encoder 100 may be provided in a first spatial audio format, whereas the spatial audio decoder may provide the reconstructed audio signal in a second spatial audio format. The second spatial audio format may be the same as the first spatial audio format, or the second spatial audio format may be different from the first spatial audio format.

[0008] The spatial audio encoder 100 typically processes the multi-channel input audio signal 111 arranged into a sequence of input frames, each input frame including a respective segment of digital audio signal for each of the channels of the multi-channel input audio signal 111, provided as a respective time series of input samples at a predefined sampling frequency. Consequently, the spatial audio encoder 100 operates each input frame into a respective frame of encoded audio signal 113 and into a respective frame of compressed spatial metadata 115 for inclusion into a respective frame of the audio bitstream 116 by the multiplexer entity 115. Moreover, some components of the spatial audio encoder 100, e.g. the audio encoder 102, the metadata derivation entity 103 and/or the metadata encoder 104, may process the audio information separately in a plurality of frequency sub-bands. A given frequency sub-band in a given frame may be referred to as a time-frequency tile.

[0009] The spatial metadata 113 derived for a single input frame may comprise a plurality of spatial parameters. As an example, the spatial parameters for a given input frame may comprise one or more direction parameters that serve to indicate a perceivable direction of arrival of a sound represented by the respective input frame and one or more directionality parameters that serve to indicate a relative strength of a directional sound component in the respective input frame. In order to enable high-quality reconstruction of the audio scene in the spatial audio decoder on basis of the encoded audio signal 113 and the compressed spatial metadata 115, the spatial parameters included in the compressed spatial metadata 115 should be derived and encoded at a sufficient accuracy. On the other hand, accurate representation of the compressed spatial metadata 115 may require an excessive bit-rate, which may not be feasible in scenarios where the bandwidth available for the audio bitstream 116 is limited.

SUMMARY

[0010] The invention is set out in the appended set of claims.

BRIEF DESCRIPTION OF FIGURES

[0011] The embodiments of the invention are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings, where

Figure 1 illustrates a block diagram of some elements of a spatial audio encoder according to an example;

Figure 2 illustrates a block diagram of some elements of an audio processing system according to an example;

Figure 3 illustrates a block diagram of some elements of a spatial audio encoder according to an example;

Figure 4 illustrates a block diagram of some elements of a spatial audio encoder according to an example;

Figure 5 illustrates a block diagram of some elements of a spatial audio decoder according to an example;

Figure 6 illustrates a block diagram of some elements of a spatial audio decoder according to an example;

Figure 7A illustrates a flow chart depicting a method for spatial audio encoding according to an example;

Figure 7B illustrates a flow chart depicting a method for spatial audio encoding according to an example; and

Figure 8 illustrates a block diagram of some elements of an apparatus according to an example.

DESCRIPTION OF SOME EMBODIMENTS

[0012] Figure 2 illustrates a block diagram of some components and/or entities of an audio processing system 200 that may serve as framework for various embodiments of the audio coding technique described in the present disclosure. The audio processing system 200 comprises an audio capturing entity 210 for recording an input audio signal 215 that represents at least one sound, an audio encoding entity 220 for encoding the input audio signal 215 into an audio bitstream 225, an audio decoding entity 230 for decoding the audio bitstream 225 obtained from the audio encoding entity into a reconstructed audio

signal 235, and an audio reproduction entity 240 for playing back the reconstructed audio signal 235.

[0013] The audio capturing entity 210 serves to produce the input audio signal 215 as a multi-channel audio signal. In this regard, the audio capturing entity 210 comprises a microphone assembly comprising a plurality of (i.e. two or more) microphones. The microphone assembly may be provided e.g. as a microphone array or an arrangement of a plurality of microphones of other type. The audio capturing entity 210 may further include processing means for recording a plurality of digital audio signals that represent the sound captured by respective microphones of the microphone assembly that thereby constitute respective channels of the multi-channel input audio signal 215. The audio capturing entity 210 provides the input audio signal 215 so obtained to the audio encoding entity 220 and/or for storage in a storage means for subsequent use.

[0014] The audio encoding entity 220 employs an audio encoding algorithm, referred herein to as an audio encoder, to process the input audio signal 215 into the audio bitstream 225. The audio encoding entity 220 may further include a pre-processing entity for processing the input audio signal 215 from a format in which it is received from the audio capturing entity 210 into a format suited for the audio encoder. This pre-processing may involve, for example, level control of the input audio signal 215 and/or modification of frequency characteristics of the input audio signal 215 (e.g. low-pass, high-pass or band-pass filtering). The pre-processing may be provided as a pre-processing entity that is separate from the audio encoder, as a sub-entity of the audio encoder or as a processing entity whose functionality is shared between a separate pre-processing and the audio encoder.

[0015] The audio decoding entity 230 employs an audio decoding algorithm, referred herein to as an audio decoder, to process the audio bitstream 225 into the reconstructed audio signal 235. The audio decoding entity 230 may further include a post-processing entity for processing the reconstructed audio signal 235 from a format in which it is received from the audio decoder into a format suited for the audio reproduction entity 240. This post-processing may involve, for example, level control of the reconstructed audio signal 235 and/or modification of frequency characteristics of the reconstructed audio signal 235 (e.g. low-pass, high-pass or bandpass filtering). The post-processing may be provided as a post-processing entity that is separate from the audio decoder, as a sub-entity of the audio decoder or as a processing entity whose functionality is shared between a separate post-processing and the audio decoder.

[0016] The audio reproduction entity 240 may comprise, for example, headphones, a headset, a loudspeaker or an arrangement of one or more loudspeakers.

[0017] Instead of an arrangement where the audio encoding entity 220 receives the input audio signal 215 (directly) from the audio capturing entity 210, the audio processing system 200 may include a storage means for

storing pre-captured or pre-created audio signals, among which the audio input signal 215 for provision to the audio encoding entity 220 may be selected. As another variation to the audio processing system 200 in this regard, the audio encoding entity 220 may receive the input audio signal 215 from another entity via a communication channel (e.g. via a communication network) instead of receiving it from the audio capturing entity 210.

[0018] Instead of an arrangement where the audio decoding entity 230 provides the reconstructed audio signal 235 (directly) to the audio reproduction entity 240, the audio processing system 200 may comprise a storage means for storing the reconstructed audio signal 235 provided by the audio decoding entity 230 for subsequent analysis, processing, playback and/or transmission to a further entity. As another variation to the audio processing system 200 in this regard, the audio decoding entity 230 may transmit the reconstructed audio signal 235 to a further entity via a communication channel (e.g. via a communication network) instead of providing it for playback by the audio reproduction entity 240.

[0019] The dotted vertical line in Figure 2 serves to denote that, typically, the audio encoding entity 220 and the audio decoding entity 230 are provided in separate devices that may be connected to each other via a network or via a transmission channel. The network/channel may provide a wireless connection, a wired connection or a combination of the two between the audio encoding entity 220 and the audio decoding entity 230. As an example in this regard, the audio encoding entity 220 may further comprise a (first) network interface for encapsulating the audio bitstream 225 into a sequence of protocol data units (PDUs) for transfer to the decoding entity 230 over a network/channel, whereas the audio decoding entity 230 may further comprise a (second) network interface for decapsulating the audio bitstream 225 from the sequence of PDUs received from the audio encoding entity 220 over the network/channel.

[0020] In the following, some aspects of a spatial audio encoding technique are described in a framework of an exemplifying spatial audio encoder 300 that may serve as the audio encoding entity 220 of the audio processing system 200 or an audio encoder thereof. In this regard, Figure 3 illustrates a block diagram of some components and/or entities of the spatial audio encoder 300 that is arranged to carry out encoding of the multi-channel input audio signal 215 into the audio bitstream 225.

[0021] The multi-channel input audio signal 215 serves to represent an audio scene, e.g. one captured by the microphone assembly of the audio capturing entity 210. The audio scene may also be referred to as a spatial audio image. Along the lines described in the foregoing, the audio scene conveyed by the multi-channel input audio signal 215 may represent both one or more directional sound components as well as the ambience of the audio scene, where a directional sound component represents a respective distinct sound source that has certain position within the audio scene whereas the ambience rep-

resents environmental sounds within the audio scene. The spatial audio encoder is arranged to process the multi-channel input audio signal 215 into an encoded audio signal 305 and spatial metadata that are descriptive of the audio scene represented by the input audio signal 215.

[0022] The spatial audio encoder 300 may be arranged to process the multi-channel input audio signal 215 arranged into a sequence of input frames, each input frame including a respective segment of digital audio signal for each of the channels, provided as a respective time series of input samples at a predefined sampling frequency. In typical example, the spatial audio encoder 300 employs a fixed predefined frame length. In other examples, the frame length may be a selectable frame length that may be selected from a plurality of predefined frame lengths, or the frame length may be an adjustable frame length that may be selected from a predefined range of frame lengths. A frame length may be defined as number samples L included in the frame for channel of the input audio signal 215, which at the predefined sampling frequency maps to a corresponding duration in time. As an example in this regard, the audio encoder 220 may employ a fixed frame length of 20 milliseconds (ms), which at a sampling frequency of 8, 16, 32 or 48 kHz results in a frame of $L=160$, $L=320$, $L=640$ and $L=960$ samples per channel, respectively. The frames may be non-overlapping or they may be partially overlapping. These values, however, serve as non-limiting examples and frame lengths and/or sampling frequencies different from these examples may be employed instead, depending e.g. on the desired audio bandwidth, on desired framing delay and/or on available processing capacity.

[0023] The spatial audio encoder 300 includes a downmixing entity 302 for creating a downmix signal 303 on basis of the multi-channel input audio signal 215. As described in the foregoing, the downmix signal 303 serves as an intermediate signal audio signal derived on basis of one or more channels of the input audio signal 215. The downmix signal 303 typically has a smaller number of channels than the input audio signal 215, typically one or two channels. In some examples, the downmix signal 303 may have the same number of channels as in the input audio signal 215. Various techniques for creating the downmix signal 303 are known in the art and a technique suitable for the intended usage of the spatial audio encoder 300 may be selected. As a few non-limiting examples in this regard, a channel of the downmix signal may be created, for example, as a linear combination (e.g. a sum, a difference, an average, etc.) of two or more channels of the input audio signal 215 or by selecting or processing one of the channels of the input audio signal 215 into a respective channel of the downmix signal 303. In some example scenarios the multi-channel input audio signal 215 may be processed by the downmixing entity 302 into one or more first signals that represent respective directional components of the audio scene conveyed by the input audio signal 215 and into a second signal

that represents the ambient component of the audio scene, the first and second signals thereby constituting the respective channels of the downmix signal 302.

[0024] The spatial audio encoder 300 includes an audio encoder 304 for processing the downmix signal 303 into the encoded audio signal 305. The operation of the audio encoder 304 typically aims at reducing the information content present in the downmix signal 303 by ignoring inaudible or perceptually less important aspects of the audio content carried in the downmix signal 303 while retaining perceptually important aspects of the audio content thereof, thereby enabling subsequent reconstruction of an audio signal that is perceptually similar to that represented by the downmix signal 303 by an audio decoder. Such lossy encoding of the downmix signal 303 enables significant reduction in the number of bits required to represent the audio content of the downmix signal 303.

[0025] Operation of the audio encoder 304 typically results in a set of audio parameters that represent a frame of the audio signal, which set of audio parameters is provided as (a component of) the encoded audio signal 305 that enables reconstruction of a perceptually similar audio signal by an audio decoder. In case the downmix signal 303 includes two or more channels, the audio encoder 304 may process each channel of the downmix signal 303 separately into a respective set of audio parameters or it may process two or more channels of the downmix signal 303 jointly into a single set of audio parameters, depending on the characteristics of the downmix signal 303. Various audio encoding techniques are known in the art, and a technique suitable for the intended usage of the spatial audio encoder 300 may be employed. Non-limiting examples in this regard include MPEG Advanced Audio Coding (AAC) encoder, Enhanced Voice Service (EVS) encoder, Adaptive Multi Rate (AMR) encoder, Adaptive Multi Rate Wide Band (AMR-WB) encoder, etc.

[0026] The spatial audio encoder 300 further includes a transform entity 306 for transforming the multi-channel input audio signal 215 from time domain into a respective multi-channel transform-domain audio signal 307. Typically, the transform domain involves a frequency domain. In an example, the transform entity 306 employs short-time discrete Fourier transform (STFT) to convert each channel of the input audio signals 215 of an input frame into a respective channel of the transform-domain signal 307 using a predefined analysis window length (e.g. 20 milliseconds). In another example, the transform entity 306 employs complex-modulated quadrature-mirror filter (QMF) bank for time-to-frequency-domain conversion. The STFT and QMF bank serve as non-limiting examples in this regard and in further examples any suitable technique known in the art may be employed for creating the transform-domain audio signal 307.

[0027] The transform entity 306 may further divide each of the channels into a plurality of frequency subbands, thereby resulting in the transform-domain audio

signal 307 that provides a respective time-frequency representation for each channel of the input audio signal 215. A given frequency band in a given frame may be referred to as a time-frequency tile. The number of frequency sub-bands and respective bandwidths of the frequency sub-bands may be selected e.g. in accordance with the desired frequency resolution and/or available computing power. In an example, the sub-band structure involves 24 frequency sub-bands according to the Bark scale, an equivalent rectangular band (ERB) scale or 3rd octave band scale known in the art. In other examples, different number of frequency sub-bands that have the same or different bandwidths may be employed. A specific example in this regard is a single frequency sub-band that covers the input spectrum in its entirety or a continuous subset thereof.

[0028] The spatial audio encoder 300 further includes a spatial analysis entity 308 for estimation of spatial audio parameters on basis of the multi-channel transform-domain audio signal 307. In an example, the spatial analysis entity 308 derives the spatial audio parameters (i.e. carries out a spatial analysis) for all time-frequency tiles, whereas in other examples the spatial audio parameters are derived in each frame for a predefined time-frequency tiles, e.g. for those time-frequency tiles that represent a predefined sub-range of frequencies. According to an example, the spatial audio parameters include at least one or more direction of arrival (DOA) parameters 309 for each time-frequency tile considered in the spatial analysis. The spatial audio parameters may further include, for each time-frequency tile considered in the analysis, one or more energy (EN) parameters 310 and/or one or more energy ratio (ER) parameters 311.

[0029] A DOA parameter 309 indicates a spatial position of a directional sound component that represents a sound source in the audio scene in a given time-frequency tile. As an example, a DOA parameter 309 may indicate, for example, an azimuth angle that defines the estimated direction of arrival of a sound (with respect to a respective predefined reference direction) in the respective time-frequency tile or an elevation angle that defines the estimated direction of arrival of a sound (with respect to a respective predefined reference direction) in the respective time-frequency tile. As a few non-limiting examples in this regard, the one or more DOA parameters 309 for a given time-frequency tile pertain to a single directional sound component and include e.g. a single azimuth angle, a single elevation angle or a single pair of an azimuth angle and an elevation angle. In further examples, there are two or more DOA parameters 309 for a given time-frequency tile and they pertain to two or more directional sound components and include respective azimuth angles for the two or more directional sound components, respective elevation angles for the two or more directional sound components or respective pairs of an azimuth angle and an elevation angle for the two or more directional sound components.

[0030] An EN parameter 310 may be applied to indi-

cate estimated or computed overall signal energy (or total signal energy) in the given time-frequency tile. An ER parameter 311 may be applied to indicate relative energy of a sound source in the audio scene in a given time-frequency tile. As an example, an ER parameter 311 may indicate estimated or computed ratio of the energy of a directional sound component and the overall signal energy in the given time-frequency tile (referred to in the following as a direct-to-total-energy ratio). In an example, there is a single ER parameter 311 for a given time-frequency tile and it pertains to a single directional sound component. In other examples, there may be two or more ER parameters 311 for a given time-frequency tile and they pertain to respective two or more directional sound components. Hence, an ER parameter 311 serves to indicate relative energy level for a given directional sound components in a given time-frequency tile, whereas the EN parameter described in the foregoing serves to indicate an absolute energy level for the given time-frequency tile.

[0031] Various methods for deriving the DOA, EN and/or ER parameters 309, 310, 311 on basis of multi-channel transform-domain audio signals 307 are known in the art and the most suitable approach to be employed by the spatial analysis entity 308 may be chosen e.g. in view of the characteristics of the spatial audio format represented by the transform-domain audio signals 307, in view of the desired accuracy of the spatial parameter modeling and/or in view of the available computational resources. An exemplifying technique in this regard is described in US patent no. 9,313,599 B2.

[0032] In some examples the spatial audio encoder 300 further includes an energy estimator 308' for estimating the overall signal energy of a reconstructed transform-domain downmix derived on basis of the encoded audio signal 305. In this regard, the energy estimator 308' may employ a (local) audio decoder to derive a local copy of a reconstructed downmix signal on basis of the encoded audio signal 305, which is further transformed into a transform-domain downmix signal and divided into a plurality of frequency sub-bands. The energy estimator 308' further operates to derive, for a plurality of time-frequency tiles, a respective quantized energy (QEN) parameter 310' that indicates the overall signal energy (e.g. total signal energy) in a respective time-frequency tile of the transform-domain downmix audio signal. The transform and frequency sub-band division applied in the energy estimator 308' are preferably the same applied by the transform entity 306.

[0033] The spatial audio parameters derived for a given frame constitute the spatial metadata available for the given frame. The spatial metadata actually provided for inclusion in the audio bitstream 225 may comprise the DOA and/or ER parameters for a plurality of frequency sub-bands. The spatial metadata may include further parameters in addition to the DOA and/or ER parameters. As a few non-limiting examples in this regard, such further spatial audio parameters may include, for a plurality of

time-frequency tiles, e.g. respective indications of a distance from the assumed listening point for one or more sound sources and/or respective indications of a spatial coherence for one or more sound sources. Some of the spatial audio parameters are arranged in the audio bitstream 225 together with the encoded audio signal 305 to enable subsequent reconstruction of a perceptually similar audio signal by an audio decoder.

[0034] The spatial audio encoder 300 further includes a ratio encoder 312 for quantizing and encoding the ER parameters 311 and a direction encoder 314 for quantizing and encoding the DOA parameters 309. For a given frame, the ratio encoder 312 operates to encode one or more ER parameters 311 derived by the spatial analysis entity 308 into respective one or more encoded ER parameters 315 whereas the direction encoder 314 operates to encode zero or more DOA parameters derived by the spatial analysis entity 308 into respective one or more encoded DOA parameters 313. The encoded ER parameter(s) 315 and possible encoded DOA parameter(s) 313 are provided as (part of) the spatial metadata for provision in the audio bitstream 225 to the spatial decoder. In the following, non-limiting examples of deriving the quantized and encoded ER parameters 315 and the DOA parameters 313 are described.

[0035] In various non-limiting examples described in more detail in the following, quantization and encoding of the DOA parameters is dependent on energy levels of one or more directional sound components represented by the input audio signal 215. As an example in this regard, a DOA parameter for a given time-frequency tile may be quantized in dependence of the EN parameter 310 obtained for the given time-frequency tile, where applicable EN parameters 310 may be obtained from the spatial analysis entity 308. In another example, a given DOA parameter for a given time-frequency tile may be quantized in dependence of a directional energy (DEN) parameter that indicates the absolute energy level for the directional sound source corresponding to the given DOA parameter in the given time-frequency tile. As described in the foregoing, there may be one or more directional sound components in a given time-frequency tile with respective one or more DEN parameters indicating their energy levels.

[0036] As an example, a DEN parameter for a given directional sound component in given a time-frequency tile may be derived directly on basis of an ER parameter 311 that indicates estimated or computed direct-to-total-energy ratio for the given directional sound component in the given time-frequency tile and a QEN parameter that indicates the overall signal energy in the given time-frequency tile of the (transform-domain) reconstructed audio signal. In this regard, the applicable ER parameters may be the ones obtained from the spatial analysis entity 308 or they may comprise respective quantized ER parameters 316 derived in the ratio encoder 312, whereas the applicable QEN parameters 310' may be obtained from the energy estimator 308'. In an exemplifying sce-

nario, the DEN parameter for the given directional sound component in the given time-frequency tile may be computed as a product of the direct-to-total-energy ratio indicated by the ER parameter for the given directional sound component in the given time-frequency tile and the overall signal energy indicated by the QEN parameter 310' for the given time-frequency tile.

[0037] According to a first example for encoding the DOA parameters 309, the ratio encoder 312 operates to quantize and encode the one or more ER parameters 311 using a suitable quantizer known in the art. This quantizer employed by the ratio encoder 312 may be referred to as an ER quantizer, which may serve to encode the quantized value of an ER parameter 311 using a fixed predefined number of bits or using a variable number of bits in dependence of the value of the ER parameter 311. In an example, the ER quantizer may comprise a variable bit-rate quantizer that assigns shorter codewords for those ER parameter values that represent relatively high values of the ER parameter 311 (e.g. relatively high values of the direct-to-total-energy ratio) and assigns longer codewords for those ER parameter values that represent relatively low values of the ER parameter 311 (e.g. relatively low values of the direct-to-total-energy ratio). In another example, the ER quantizer may comprise a variable bit-rate quantizer that assigns shorter codewords for those ER parameter values that occur more frequently and assigns longer codewords for those ER parameter values that occur less frequently. The codewords and their lengths may be pre-assigned based on experimental data using techniques known in the art. The quantization and encoding of an ER parameter 311 may rely, for example, on an ER quantization table that maps a plurality of table entries that each store a pair of a quantized ER parameter value and a codeword (e.g. a bit-pattern) assigned thereto. If using such an ER quantization table, the ratio encoder 312 operates to identify the table entry that holds quantized ER parameter value that is closest to the value of the ER parameter 311 under quantization/encoding and sets the value of the quantized ER parameter 316 and the value of the encoded ER parameter 315 to values found in the identified table entry. The ratio encoder 312 provides the encoded ER parameters 315 to a multiplexer 318 for inclusion in the audio bitstream 225 and provides the quantized ER parameters 316 for the direction encoder 314 to serve as control information in quantization and encoding of the DOA parameters 309 therein.

[0038] Still referring to the first example, the direction encoder 314 operates to quantize the DOA parameters 309 in dependence of respective (absolute) energy levels of one or more sound components of the multi-channel transform-domain audio signal 307 in the corresponding time-frequency tile. The direction encoder 314 operates to quantize one or more DOA parameters 309 using a suitable quantizer known in the art. This quantizer employed by the direction encoder 314 may be referred to as a DOA quantizer, which may serve to encode the

quantized value of a DOA parameter 309 using a fixed predefined number of bits or using a variable number of bits in dependence of the value of the DOA parameter. The quantization and encoding of a DOA parameter 309 may rely, for example, on a DOA quantization table that maps a plurality of table entries that each store a pair of a quantized DOA parameter value and a codeword (e.g. a bit-pattern) assigned thereto. If using such a DOA quantization table, the direction encoder 314 operates to identify the table entry that holds quantized DOA parameter value that is closest to the value of the DOA parameter 309 under quantization/encoding and sets the value of the quantized DOA parameter and the value of the encoded DOA parameter 313 to respective values found in the identified table entry. The direction encoder 314 provides the encoded DOA parameters 313 to the multiplexer 318 for inclusion in the audio bitstream 225 therein.

[0039] In the first example, it is assumed that there the DOA parameters 309 for a given time-frequency tile pertain to a single sound source, in other words there is at most a single directional sound component in the given time-frequency tile. As described in the foregoing, also in this scenario there may be one or more DOA parameters 309 derived for the given time-frequency tile, e.g. a DOA parameter that indicates an azimuth angle derived for the single direction sound component and/or a DOA parameter that indicates an elevation angle derived for the single directional sound component in the given time-frequency tile.

[0040] Still referring to the first example, in a first exemplifying scenario the direction encoder 314 operates to make a decision, for a plurality of time-frequency tiles considered in the spatial analysis, between including and omitting the respective encoded DOA parameter(s) 313 in/from the audio bitstream 225. For each considered time-frequency tile, the decision is made in dependence of one or more criteria that pertain to the respective (absolute) energy levels of one or more sound components of the multi-channel transform-domain audio signal 307 in the respective time-frequency tile:

- if, for a given time-frequency tile, the one or more criteria are met, the direction encoder 314 operates to (quantize and) encode the DOA parameter(s) 309 derived for the given time-frequency tile using a predefined DOA quantizer and provides the encoded DOA parameter(s) 313 for inclusion in the audio bitstream 225 by the multiplexer 318;
- if, for the given time-frequency tile, the one or more criteria are not met, the direction encoder 314 omits (quantization and) encoding of the DOA parameter(s) 309 for the given time-frequency tile and, consequently, no DOA parameters concerning the given time-frequency tile are provided for the spatial audio decoder in the audio bitstream 225.

[0041] In a variation of the first example described above, the direction encoder 314 may respond to a failure

to meet the one or more criteria by using the DOA quantizer therein to quantize and encode predefined default value(s) for the DOA parameters instead of completely omitting the encoded DOA parameters 313 for the given time-frequency tile from the audio bitstream 225. The default DOA parameters may serve to indicate, for example, zero azimuth and/or zero elevation (i.e. a sound source positioned directly in front of the assumed listening point). In such a variation the DOA quantizer employed for processing (e.g. quantizing and/or encoding) the derived DOA parameters 309 into respective encoded DOA parameters 313 preferably employs a variable bit rate such that the predefined default value(s) are assigned a codeword that is relatively short (i.e. a codeword that employs a relatively small number of bits).

[0042] According to a second example, the ratio encoder 312 operates in a manner similar to that described in context of the first example, whereas the operation of the direction encoder 314 is different. Also in the second example the direction encoder 314 operates to quantize the DOA parameters 309 in dependence of the respective (absolute) energy levels of one or more sound components of the multi-channel transform-domain audio signal 307 in the corresponding time-frequency tile, while the difference to the first example is that herein the direction encoder 314 employs one of a plurality of DOA quantizers to quantize and encode the DOA parameters 309 for the time-frequency tiles considered in the spatial analysis. The plurality of DOA quantizers provide different bit-rates, thereby providing a respective different tradeoff between accuracy of the quantization and the number of bits employed to define quantization codewords. Each of the plurality of DOA quantizers operates to quantize a value of a DOA parameter 309 using a suitable quantizer known in the art using a fixed predefined number of bits or using a variable number of bits in dependence of the value of the DOA parameter. Along the lines described in context of the first example for a single DOA quantizer, each of the plurality of DOA quantizers may rely, for example, on a respective DOA quantization table that maps each of a plurality quantized DOA parameter values to a respective one of a plurality codewords assigned thereto.

[0043] Also in the second example, it is assumed that there the DOA parameters 309 for a given time-frequency tile pertain to a single sound source, in other words that there is at most a single directional sound component in the given time-frequency tile. As described in the foregoing, also in this scenario there may be one or more DOA parameters 309 derived for the given time-frequency tile, e.g. a DOA parameter that indicates an azimuth angle derived for the single directional sound component and/or a DOA parameter that indicates an elevation angle derived for the single directional sound component in the given time-frequency tile.

[0044] In the following, more detailed description of the second example is provided with reference to two DOA quantizers, where a first DOA quantizer employs a higher

number of bits for encoding the quantized value of a DOA parameter 309 at a higher precision to provide a smaller (average) quantization error and where a second DOA quantizer employs a lower number of bits for encoding the quantized value of the DOA parameter 309 at a lower precision to provide a larger (average) quantization error. Selection of one of the first and second DOA quantizers enables choosing the more appropriate tradeoff between the number of bits used for encoding the value of the DOA parameter 309 and the precision (or accuracy) of the quantization. The approach that involves two DOA quantizers is chosen here for clarity and brevity of description, whereas the described approach readily generalizes into an approach where more than two DOA quantizers at different bit-rates are available for selection to enable choosing the most suitable tradeoff between the number of bits used for encoding the value of the DOA parameter 309 and the precision (or accuracy) of the quantization.

[0045] Still referring to the second example, the direction encoder 314 operates to make a decision, for a plurality of time-frequency tiles considered in the spatial analysis, between using the first DOA quantizer or the second DOA quantizer for quantizing and encoding the DOA parameter(s) 309 for a given time-frequency tile. For each considered time-frequency tile, the decision is made on basis of one or more criteria that pertain to the respective (absolute) energy level(s) of one or more sound components of the multi-channel transform-domain audio signal 307 in the respective time-frequency tile:

- if, for a given time-frequency tile, the one or more criteria are met, the direction encoder 314 operates to quantize and encode the DOA parameter(s) 309 derived for the given time-frequency tile using the first DOA quantizer and provides the encoded DOA parameter(s) 313 derived by the first DOA quantizer for inclusion in the audio bitstream 225 by the multiplexer 318;
- if, for the given time-frequency tile, the one or more criteria are not met, the direction encoder 314 operates to quantize and encode the DOA parameter(s) 309 derived for the given time-frequency tile using the second DOA quantizer and provides the encoded DOA parameter(s) 313 derived by the second DOA quantizer for inclusion in the audio bitstream 225 by the multiplexer 318.

[0046] In an example, the one or more criteria that pertain to respective absolute energy levels of one or more sound components of the multi-channel transform-domain audio signal 307 in a given time-frequency tile involves consideration of directional energy level of the single directional sound component of the audio scene represented by the multi-channel transform-domain audio signal 307 (and hence by the multi-channel input audio signal 215).

[0047] A first criterion in this regard may be provided by evaluating whether the DEN parameter derived for the given time-frequency tile indicates energy level of the directional sound component that exceeds a first threshold. In other words, the first criterion is met in case the energy level of the directional sound component exceeds the first threshold and the first criterion is not met in case the energy level of the directional sound component fails to exceed the first threshold.

[0048] The first threshold may be the same across frequency sub-bands, or the first threshold may be set to a different value from one frequency sub-band to another. In an example, the first threshold is set to represent a threshold value for the energy of a directional sound component above which the strength of the arriving sound is considered to provide sufficient improvement to the reconstructed audio scene in the respective frequency sub-band e.g. in view of the bits required for encoding the respective DOA parameters and/or in view of additional value provided by accurate reconstruction of the arrival direction of the respective directional sound component of the audio scene. In this example, the first threshold may have a respective predefined value for each of the frequency sub-bands.

[0049] In another example, the first threshold comprises a masking threshold derived on basis of (the local copy of) the reconstructed downmix signal derived by the energy estimator 308'. In this regard, typically a dedicated masking threshold is derived for each time-frequency tile, thereby leading to a scenario where the first threshold is different across the frequency sub-bands of a frame. The masking threshold for a given time-frequency tile may be derived, for example, on basis of the overall energy level of the reconstructed downmix signal in the respective time-frequency tile. In another example, the masking threshold derivation further considers tonality and/or spectral flatness of the reconstructed downmix signal in the given time-frequency tile. The masking thresholds may be computed by the energy estimator 308' and passed to the direction encoder 314 along with the QEN parameters 310'. Alternatively, the energy estimator 308' may pass the reconstructed downmix signal to the direction encoder 314 along with the QEN parameters 310' to the direction encoder 314, which may apply this signal for derivation of the masking thresholds. Various techniques for deriving the masking threshold(s) are known in the art and any suitable approach may be applied.

[0050] In a further example, the first threshold for a given time-frequency tile is set to an adaptive value that is defined, for example, in dependence of the energies of the directional sound components in those time-frequency tiles that are adjacent in time and/or frequency to the given time-frequency tile. In another scenario, evaluation of the first exemplifying criterion in the given time frequency tile may depend on the corresponding evaluation carried out in an adjacent time-frequency tile. As an example of the latter, if the energy of a directional

sound component in an adjacent time-frequency tile exceeds the first threshold, the DOA parameter of the given time-frequency tile may be encoded using first/predefined DOA quantizer even if the energy of the directional sound component in the given time-frequency tile fails to exceed the first threshold. In this scenario, the first threshold may comprise the masking threshold described in the foregoing or it may comprise another threshold, e.g. a derivative of the masking threshold derived by adding a predefined margin to the masking threshold, where the predefined margin may be e.g. 6 dB.

[0051] A second criterion comprises evaluating whether the estimated overall signal energy indicated in the QEN parameters 310' for a given time-frequency tile indicates energy level that exceeds a second threshold. In other words, the second criterion is met in case the estimated overall signal energy exceeds the second threshold and the second criterion is not met in case the estimated overall signal energy fails to exceed the second threshold. Similar considerations as provided in the foregoing for the first threshold apply to the second threshold as well.

[0052] The first and second criteria described in the foregoing serve as non-limiting examples of criteria applied by the direction encoder 314 in determining quantization to be applied to the DOA parameters 310 (or omission thereof) for a given time-frequency tile in dependence of energy level(s) of one or more sound components of the multi-channel transform-domain audio signal 307 in the given time-frequency tile. In this regard the direction encoder 314 may apply any one of the first and second criteria (or a further sound-component-energy-related criterion) in deciding on how to quantize or whether to transmit or omit encoded DOA parameters 313 for a given time-frequency tile. In other examples, the direction encoder 314 may apply any combination or sub-combination of the first, the second and further criteria in deciding the manner of DOA quantization (or lack thereof) for a given time-frequency tile, e.g. such that the encoded DOA parameters 313 for the given time-frequency tile are included in the audio bitstream 225 or quantized and encoded using the first DOA quantizer only in case all of the applied criteria are met or such that the encoded DOA parameters 313 for the given time-frequency tile are included in the audio bitstream 225 or quantized and encoded using the first DOA quantizer in case any of the applied criteria is met.

[0053] A third example proceeds from the assumption that at least for some time-frequency tiles considered in the spatial analysis the DOA parameters 309 for a given time-frequency tile pertain to two or more (simultaneous) directional sound components and, hence, for such time-frequency tiles there is at least one DOA parameter 309 derived (by the spatial analysis entity 308) for each of the two more directional sound components. Consequently, for such time-frequency tiles there are also respective two or more ER parameters 311. In other words, the spatial audio parameters for such time-frequency tile

include a respective pair of ER parameter(s) 311 and DOA parameter(s) 309 for each identified directional sound component. Along the lines described in the foregoing, also in this scenario there may be one or more DOA parameters 309 derived for each directional sound component in a given time-frequency tile, e.g. two or more DOA parameters that indicate respective azimuth angles derived for the two or more directional sound components and/or two or more DOA parameters that indicate respective elevation angles derived for the two or more directional sound components in the given time-frequency tile.

[0054] According to the third example, the ratio encoder 312 and the direction encoder 314 operate to quantize and encode the ER parameter(s) 311 and the DOA parameter(s) 309 for a given time-frequency tile using at most a predetermined total number of bits. In a first scenario, this may involve quantizing and encoding each pair of the ER parameter(s) 311 and the DOA parameter(s) 309 in a given time-frequency tile using at most a respective predetermined number of bits. In this example, the ratio encoder 312 operates to quantize and encode the ER parameter(s) 311 using a suitable quantizer known in the art. This quantizer employed by the ratio encoder 312 may be referred to as an ER quantizer, which serves to encode the quantized value of an ER parameter 311 using a variable number of bits in dependence of the value of the ER parameter 311. The ER quantizer may be a variable bit-rate quantizer that assigns shorter codewords for those ER parameter values that represent relatively high values of the ER parameter 311 (e.g. relatively high values of the direct-to-total-energy ratio) and assigns longer codewords for those ER parameter values that represent relatively low values of the ER parameter 311 (e.g. relatively low values of the direct-to-total-energy ratio). The ER quantizer may be provided, for example, as a quantization table as described in the foregoing.

[0055] With fixed number of bits available to represent a pair of the encoded ER parameter(s) 315 and the encoded DOA parameter(s) 313 that pertain to a respective directional sound component for a given time-frequency tile using the shorter codewords (in terms of number of bits) for the high values of the ER parameter 311 leaves more bits for quantization of the DOA parameter(s) 313 in such frames. This is beneficial since accurate reconstruction of the arrival direction of a sound in the reconstructed audio scene is typically perceptually important for those directional sound components that have a relatively high energy, whereas for those directional sound components that have a relatively low energy accurate reconstruction of the arrival direction of a sound is typically perceptually less important.

[0056] In the third example, the ratio encoder 312 uses the variable bit-rate ER quantizer to derive the quantized ER parameter 316 and the encoded ER parameter 315 for each of the directional sound components of a given time-frequency tile. In this regard, instead of or in addition to the quantized ER parameters 316 the ratio encoder

312 provides to the direction encoder 314 with an indication of the number of bits employed for the encoded ER parameters 315 for each directional sound component in the time-frequency tiles considered in the spatial analysis or an indication of the number of bits available for DOA quantization for each directional sound component in the time-frequency tiles considered in the spatial analysis.

[0057] The direction encoder 314 employs one of a plurality of DOA quantizers available therein to quantize and encode the DOA parameters 309 for each of the directional sound components in the time-frequency tiles considered in the spatial analysis. In this regard, the direction encoder 314 selects, for a given directional sound component in a given time-frequency tile, one of the plurality of DOA quantizers in accordance with the number of bits available for quantization of those DOA parameter(s) 309 in the given time-frequency tile. The plurality of DOA quantizers provide different (fixed) bit-rates, thereby providing a respective different tradeoff between accuracy of the quantization and the number of bits employed to define quantization codewords such that a higher quantization bit-rate provides lower (average) quantization error while a lower quantization bit-rate provides higher (average) quantization error. Thus, the direction encoder 314 may select, for a given directional sound component in a given time-frequency tile, the DOA quantizer that uses the highest number of bits that does not exceed the number of bits available for DOA quantization for the respective directional sound component in the given time-frequency tile.

[0058] In the first scenario for the third example described above, an underlying assumption is that the total number of bits available for quantizing the ER parameters 311 and the DOA parameters 309 for a given time-frequency tile are evenly allocated for the two or more directional sound components of the given time-frequency tile, in other words such that each pair of the ER parameter(s) 311 and the DOA parameter(s) 309 that pertain to a given directional sound component in the given time-frequency tile is assigned the same number of bits.

[0059] In a second scenario of the third example, the bit allocation for quantizing the DOA parameters 309 for two or more directional sound components of a given time-frequency tile is dependent on one or more criteria that pertain to the respective energy levels of one or more sound components of the multi-channel transform-domain audio signal 307 in the given time-frequency tile. In this regard, there may be a (predefined) first number of bits for encoding the two or more ER parameters 311 and the two or more DOA parameters 309 of the given time-frequency tile: in this second scenario of the third example, each of the two or more ER parameters 311 of the given time-frequency tile are encoded using the variable bit-rate ER quantizer described in the foregoing, whereas the remaining bits are available for encoding the two or more DOA parameters 309 of the given time-frequency tile. These remaining bits constitute a second

number of bits and they are allocated for encoding the DOA parameters 309 pertaining to respective two or more directional sound components of the given time-frequency tile such that a first directional sound component (of the given time-frequency tile) that has a higher energy level indicated therefor is assigned a larger share of the second number of bits whereas a second directional sound component (of the given time-frequency tile) that has a lower energy level indicated therefor is assigned a smaller share of the second number of bits.

[0060] The bit-rate assignment for the two or more directional sound components of the given time-frequency tile may be carried out by using a first predefined bit allocation rule that defines a respective (maximum) number of bits available for encoding the respective DOA parameter(s) 309 in accordance with respective energy levels indicated for each of the two or more directional sound components. The direction encoder 314 may use the bit allocation so obtained to select, for each of the directional sound components of the given time-frequency tile, a respective one of the plurality of DOA quantizers (described in the foregoing) in accordance with the number of bits allocated for quantization of those DOA parameter(s) 309 in the given time-frequency tile. Consequently, the DOA parameter(s) 309 for directional sound components of the given time-frequency tile that have a high(er) (absolute) energy may be encoded at a higher precision than the DOA parameter(s) 309 for directional sound components of the given time-frequency tile that have a low(er) (absolute) energy.

[0061] As an example, the comparison of energy levels may comprise comparison of directional energy levels indicated by the DEN respective parameters derived for the two or more directional sound components in a given time-frequency tile. In another example, the comparison of energy levels may comprise comparison of direct-to-total-energy ratios indicated by the respective ER parameters 311 obtained for the given time-frequency tile. In the latter example, the underlying overall signal energy is the same for all directional sound components of the given time-frequency tile and hence comparison of the respective ER parameter is sufficient. Consequently, in such a scenario the comparison of energy levels may not necessarily require derivation of the QEN parameters and hence in such a scenario the spatial audio encoder 300 may be provided with the energy estimator 308'.

[0062] As non-limiting examples, the allocation of bits for encoding the DOA parameters 309 of two or more respective directional sound components of a given time-frequency tile in accordance with the second scenario of the third example may involve one or more of the following:

- In case the energy level indicated for a first directional sound component in the given time-frequency tile exceeds the energy level indicated for a second directional sound component in the given time-frequency tile by more than a first predefined margin,

the DOA parameters pertaining to the first directional sound component are assigned a first number of bits and the DOA parameters pertaining to the second directional sound component are assigned a second number of bits, where the first number of bits is higher than the second number of bits. Consequently, the DOA parameters pertaining to the first directional sound component are encoded using a first DOA quantizer that uses at most the first number of bits and the DOA parameters pertaining to the second directional sound component are encoded using a second DOA quantizer that uses at most the second number of bits, where the first DOA quantizer (that employs a higher number of bits) enables encoding the quantized value of a DOA parameter 309 at a higher precision than the second DOA quantizer (that employs a lower number of bits).

- In case the energy level indicated for a first directional sound component in the given time-frequency tile exceeds the energy level indicated for a second directional sound component in the given time-frequency tile by more than a second predefined margin, the DOA parameter(s) 309 pertaining to the first directional sound component are assigned a first number of bits and, consequently, encoded using a DOA quantizer available for the direction encoder 314 and the DOA parameter(s) 309 pertaining to the second directional sound source are not encoded but are omitted from the audio bitstream 225.

[0063] In the latter example above, instead of completely omitting the DOA parameter(s) pertaining to the second directional sound component from the audio bitstream 225, they may be replaced with default value(s) for the DOA parameters, along the lines described in the foregoing in context of the first example.

[0064] In a fourth example, which is described herein as a variation of the second scenario of the third example, the directional energy-level dependent bit allocation for quantizing the DOA parameters 309 for two or more directional sound components of a given time-frequency tile is provided further in view of a third threshold: the direction encoder 314 may apply a predetermined fixed number of bits for encoding the two or more DOA parameters 309 of the given time-frequency tile, and the bits are allocated for encoding the DOA parameters 309 pertaining to respective two or more directional sound components in view of their respective relationships with the third threshold.

[0065] In the fourth example, the bit-rate assignment for the two or more directional sound components of the given time-frequency tile may be carried by using a second predefined bit allocation rule that defines a respective (maximum) number of bits available for encoding the respective DOA parameter(s) 309 in accordance with respective energy levels indicated for each of the two or more directional sound components in view of their relationship with the third threshold. The direction encoder

314 may use the bit allocation so obtained to select, for each of the directional sound components of the given time-frequency tile, a respective one of the plurality of DOA quantizers (described in the foregoing) in accordance with the number of bits allocated for quantization of those DOA parameter(s) 309 in the given time-frequency tile.

[0066] The relationship between an energy level and the third threshold may be expressed, for example, as a comparison value defined by a ratio of the energy level and the third threshold (e.g. by dividing the energy level by the third threshold) or by a difference between the energy level and the third threshold (e.g. by subtracting the third threshold from the energy level). Hence, the direction encoder 314 may derive a respective comparison value for each of the two or more directional sound components of a given time-frequency tile on basis of the respective energy levels indicated therefor and encode the DOA parameters 309 for the given time-frequency tile such that the DOA parameter(s) 309 for directional sound components of the given time-frequency tile that have a high(er) comparison value may be encoded at a higher precision than the DOA parameter(s) 309 for directional sound components of the given time-frequency tile that have a low(er) comparison value. In an example, the energy levels applied for deriving the comparison values for the two or more directional sound components of the given time-frequency tile may comprise directional energy levels indicated by the respective DEN parameters derived for the two or more directional sound components in the given time-frequency tile. In another example, the energy levels applied for deriving the comparison values for the two or more directional sound components of the given time-frequency tile may comprise direct-to-total-energy ratios indicated by the respective ER parameters 311 obtained for the given time-frequency tile.

[0067] As non-limiting examples, the allocation of bits for encoding the DOA parameters 309 of two or more respective directional sound components of a given time-frequency tile in accordance with the fourth example may involve one or more of the following:

- In case the comparison value derived for a first directional sound component in the given time-frequency tile exceeds the comparison value derived for a second directional sound component in the given time-frequency tile by more than a third predefined margin, the DOA parameters pertaining to the first directional sound component are assigned a first number of bits and the DOA parameters pertaining to the second directional sound component are assigned a second number of bits, where the first number of bits is higher than the second number of bits. Consequently, the DOA parameters pertaining to the first directional sound component are encoded using a first DOA quantizer that uses at most the first number of bits and the DOA parameters pertaining

to the second directional sound component are encoded using a second DOA quantizer that uses at most the second number of bits, where the first DOA quantizer (that employs a higher number of bits) enables encoding the quantized value of a DOA parameter 309 at a higher precision than the second DOA quantizer (that employs a lower number of bits).

- In case the comparison value derived for a first directional sound component in the given time-frequency tile exceeds the comparison value derived for a second directional sound component in the given time-frequency tile by more than a fourth predefined margin, the DOA parameter(s) 309 pertaining to the first directional sound component are assigned a first number bits and are, consequently, encoded using a DOA quantizer available for the direction encoder 314 and the DOA parameter(s) 309 pertaining to the second directional sound source are not encoded but are omitted from the audio bitstream 225.

[0068] In the latter example above, instead of completely omitting the DOA parameter(s) pertaining to the second directional sound component from the audio bitstream 225, they may be replaced with default value(s) for the DOA parameters, along the lines described in the foregoing in context of the first example.

[0069] Similar considerations as provided in the foregoing for the first threshold apply to the third threshold as well. Hence, in one example the third threshold may be the same across frequency sub-bands, or the third threshold may be set to a different value from one frequency sub-band to another, where a respective predefined value for the third threshold is used in each of the frequency sub-bands. In another example the third threshold comprises a masking threshold derived on basis of (the local copy of) the reconstructed downmix signal derived by the energy estimator 308'. In this regard, typically a dedicated masking threshold is derived for each time-frequency tile, thereby leading to a scenario where the first threshold is different across the frequency sub-bands of a frame, as described in more detail in the foregoing in context of the first threshold.

[0070] The spatial audio encoder 300 further includes the multiplexer 318 for creation of a segment of audio bitstream 225 for a given input frame. In this regard, the multiplexer 318 operates to combine the encoded audio signal 305 derived for a given input frame with the spatial metadata that may include the encoded ER parameter(s) 315 and/or the encoded DOA parameter(s) 313 derived for the given input frame for one or more frequency sub-bands, depending on the operation of the ratio encoder 312 and the direction encoder 314. Along the lines described in the foregoing, the audio bitstream 225 may be transmitted over a communication channel (e.g. via a communication network) to a spatial audio decoder and/or it may be stored a memory for subsequent use.

[0071] In the foregoing, various aspects related to the spatial audio encoder 300 of Figure 3 were described

with a number of examples. The spatial audio encoder 300 operates to derive the encoded ER parameters 315 and the encoded DOA parameters 313 separately from the audio encoder. As another example, Figure 4 depicts another exemplifying spatial audio encoder 400, where the processing involved in derivation of the encoded ER parameters 315 and the encoded DOA parameters 313 shares some components with the processing applied for derivation of the encoded audio signal 305 and possibly also at least partially makes use of information derived in the process of deriving the encoded audio signal 305.

[0072] The spatial audio encoder 400 includes the transform entity 306 for transforming the multi-channel input audio signal 215 from time domain into the respective multi-channel transform-domain audio signal 307. Non-limiting examples pertaining to operation and characteristics of the transform entity 306 are provided in the foregoing as part of description of the spatial audio encoder 300. In the spatial audio encoder 400 the transform entity provides the transform-domain audio signal 307 for further processing in a downmix entity 302' and in a spatial analysis entity 308a.

[0073] The spatial audio encoder 400 includes a downmixing entity 402 for creating a transform-domain downmix signal 403 on basis of the multi-channel transform-domain audio signal 307. Along the lines described in the foregoing, the transform-domain downmix signal 403 serves as an intermediate signal audio signal derived on basis of the multi-channel transform-domain audio signal 307 such that it has a smaller number of channels than the multi-channel transform-domain audio signal 307, typically one or two channels. The downmixing entity 402 operates on a transform-domain signal, while otherwise its operating principle is similar to that of the downmixing entity 302 of the spatial audio encoder 300.

[0074] The spatial audio encoder 400 includes an audio quantizer 404 for processing the transform-domain downmix signal 403 into an encoded audio signal 405. The operation of the audio quantizer 404 aims at producing an encoded audio signal 405 that represents the transform-domain downmix signal 403 at an accuracy that enables subsequent reconstruction of an audio signal that is perceptually as similar as possible to that represented by the transform-domain downmix signal 403 in view of the available bit-rate.

[0075] The spatial audio encoder 400 further includes an audio dequantizer 404' for reconstructing a local copy of a reconstructed transform-domain downmix signal 403' on basis of the encoded audio signal 405 to enable estimation of the overall signal energy. Due to this role, the audio dequantizer 404' may also be referred to as a local audio dequantizer.

[0076] The spatial audio encoder 400 further includes an energy estimator 308b for estimating the overall signal energy of the reconstructed transform-domain downmix signal 403'. In this regard, the energy estimator 308b may derive, for each time-frequency tile considered in the spa-

tial analysis, a respective QEN parameter 410 that indicates estimated overall signal energy in a respective time-frequency tile.

[0077] The spatial audio encoder 400 further includes a spatial analysis entity 308a for estimation of spatial audio parameters on basis of the multi-channel transform-domain audio signal 307. The spatial analysis entity 308a is similar to the spatial analysis entity 308 described in the foregoing apart from the fact that in the context of the spatial audio encoder 400 the EN parameter estimation is carried out by the energy estimator 308b. According to an example, the spatial audio parameters derived by the spatial analysis entity 308a include at least one or more DOA parameters 309 for each time-frequency tile considered in the spatial analysis. The spatial audio parameters derived by the spatial analysis entity 308a may further include, for each time-frequency tile considered in the analysis, one or more energy ratio (ER) parameters 311.

[0078] The spatial audio encoder 400 further includes the ratio encoder 312 for quantizing and encoding the ER parameters 311 and the direction encoder 314 for quantizing and encoding the DOA parameters 309. The operation and characteristics of these entities are described in detail in the foregoing via a plurality of examples that pertain to their usage as part of the spatial audio encoder 300.

[0079] The spatial audio encoder 400 further includes a bitstream packer 418 for creation of a segment of audio bitstream 225 for a given input frame. In this regard, the bitstream packer 418 operates to combine the encoded audio signal 405 derived for a given input frame with the spatial metadata that may include the encoded ER parameter(s) 315 and/or the encoded DOA parameter(s) 313 derived for the given input frame for one or more frequency sub-bands, depending on the operation of the ratio encoder 312 and the direction encoder 314. Along the lines described in the foregoing, the audio bitstream 225 may be transmitted over a communication channel (e.g. via a communication network) to a spatial audio decoder and/or it may be stored in a memory for subsequent use.

[0080] In the following, some aspects of a spatial audio decoding technique are described in a framework of an exemplifying spatial audio decoder 350 that may serve as the audio decoding entity 230 of the audio processing system 200 or an audio decoder thereof. In this regard, Figure 5 illustrates a block diagram of some components and/or entities of the spatial audio decoder 350 that is arranged to carry out decoding of the audio bitstream 225 generated using the spatial audio encoder 300 into the reconstructed audio signal 235.

[0081] The spatial audio decoder 350 may be arranged to receive the audio bitstream 225 as a sequence of frames and to process each frame of the audio bitstream 225 into a corresponding frame of reconstructed audio signal 235, provided as a respective time series of output samples at a predefined sampling frequency.

[0082] The spatial audio decoder 350 includes a demultiplexer 368 for extracting the encoded audio signal 305 and the spatial metadata from the audio bitstream 225 for a given input frame. The spatial metadata may comprise the encoded ER parameter(s) 315 and/or the encoded DOA parameter(s) 313 for one or more frequency sub-bands derived in the spatial audio encoder 300 for the given input frame.

[0083] The spatial audio decoder 350 includes an audio decoder 354 for processing the encoded audio signal 305 into a reconstructed downmix signal 303'. The audio decoder 354 operates to invert the operation of the audio encoder 304 of the spatial audio encoder 300, thereby creating the reconstructed downmix signal that is perceptually similar to the downmix signal 303 derived in the spatial audio encoder 300. As described in context of the audio encoder 304, various audio encoding techniques are known in the art, and a technique applied by the audio decoder 354 is the one that inverts the audio encoding procedure applied in the audio encoder 304.

[0084] The spatial audio decoder 350 further includes a ratio decoder 362 for decoding the encoded ER parameters 315 into corresponding quantized ER parameters 316 and a direction decoder 364 for decoding the encoded DOA parameters 313 into corresponding quantized DOA parameters 309'. The quantized DOA parameters may not be available for all frequency sub-bands, depending on the operation of the spatial audio encoder 300 described in the foregoing. The quantized ER parameters 316 and possible quantized DOA parameters 309' are provided for an audio synthesizer 370 for derivation of the reconstructed audio signal 235 therein.

[0085] The ratio decoder 362 operates to invert the encoding procedure applied in the ratio encoder 312 to derive the quantized ER parameters 316 that match those derived in the ratio encoder 312. The direction decoder 364 operates to invert the encoding procedure applied in the direction encoder 314 to derive the quantized DOA parameters 309' to match the corresponding values derived in the direction encoder 314. Hence, decoding of the encoded DOA parameters 313 is carried out in dependence on energy levels of one or more directional sound components represented in the audio bitstream 225 and hence in the input audio signal 215. According to various non-limiting examples, the decoding of the encoded DOA parameter(s) 313 for a given time-frequency tile may be carried out in dependence of a DEN parameter that indicates the absolute energy level for the directional sound source corresponding to the given encoded DOA parameter in the given time-frequency tile and/or in dependence of a QEN parameter that indicates overall signal energy in the given time-frequency tile. Derivation of the DEN parameters and QEN parameters is described in the foregoing. The QEN parameters may be obtained e.g. on basis of the reconstructed downmix signal 303' according to procedure described in the foregoing in context of the energy estimator 308'. As described in the foregoing in context of the spatial audio encoder

300, there may be one or more directional sound components in a given time-frequency tile with respective one or more DEN parameters and/or ER parameters indicating their energy levels.

[0086] A first example for deriving the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more time-frequency tiles by the ratio decoder 362 and the direction decoder 364 involves inverting the encoding procedure according to the first example for encoding the DOA parameters 309 described in the foregoing. In this regard, the ratio decoder 362 employs the same ER quantizer used for encoding by the ratio encoder 312 to determine the corresponding quantized ER parameter 316 for each received encoded ER parameter 315, e.g. by using the ER quantization table described in the foregoing.

[0087] In the first example, it is assumed that the encoded DOA parameter(s) 313 for a given time-frequency tile pertain to a single sound source. However, as described in the foregoing, also in this scenario there may be one or more encoded DOA parameters 313 derived for the given time-frequency tile, e.g. a DOA parameter that indicates an azimuth angle derived for the single direction sound component and/or a DOA parameter that indicates an elevation angle derived for the single directional sound component in the given time-frequency tile.

[0088] Still referring to the first example, for each considered time-frequency tile, one or more criteria that pertain to the respective (absolute) energy levels of one or more sound components of the audio signal represented by the audio bitstream 225 in the respective time-frequency tile are evaluated in order to determine whether the audio bitstream 225 includes encoded DOA parameters 313 for the corresponding time-frequency tile:

- if, for a given time-frequency tile, the one or more criteria are met, the audio bitstream 225 includes the encoded DOA parameter(s) 313 for the given time-frequency tile and the direction decoder 364 operates to determine the corresponding quantized DOA parameter(s) 309' for the given time-frequency tile using the predefined DOA quantizer and provides the quantized DOA parameter(s) 309' for an audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235;
- if, for the given time-frequency tile, the one or more criteria are not met, the audio bitstream 225 does not include the encoded DOA parameter(s) 313 for the given time-frequency tile and, consequently, the corresponding directional sound component is omitted (or excluded) from the reconstructed audio signal.

[0089] In a variation of the first example described above, the direction decoder 364 may respond to a failure to meet the one or more criteria by introducing default DOA parameter(s) for the given time-frequency tile and

providing the introduced DOA parameter(s) 313 for the audio synthesizer for creation of the corresponding directional sound component for the reconstructed audio signal 235. As described in the foregoing in context of the spatial audio encoder 300, the default DOA parameter(s) may serve to indicate, for example, zero azimuth and/or zero elevation (i.e. a sound source positioned directly in front of the assumed listening point).

[0090] According to a second example for deriving the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more time-frequency tiles, the ratio decoder 362 operates in a manner similar to that described in context of the first example, whereas the operation of the direction decoder 364 is different. Also in the second example it is assumed that the encoded DOA parameter(s) 313 for a given time-frequency tile pertain to a single sound source and the direction decoder 364 operates to find the quantized DOA parameters 309' in dependence of (absolute) energy levels of one or more sound components of the audio signal represented by the audio bitstream 225 in the respective time-frequency tile.

[0091] As described in the foregoing in context of the direction encoder 314, in the second example the direction decoder 364 has a plurality of DOA quantizers at different bit-rates available for the DOA quantization and the DOA quantizer applied for decoding a given encoded DOA parameter 313 is selected in dependence of (absolute) energy indicated for the corresponding directional sound component. Assuming the example described in context of the direction encoder 314 involving the first and second DOA quantizers, the direction decoder 364 evaluates one or more criteria that pertain to the respective (absolute) energy levels of one or more sound components of the audio signal represented by the audio bitstream 225 in the respective time-frequency tile in order to determine whether the first DOA quantizer or the second DOA quantizer is to be applied for decoding the respective encoded DOA parameter 313:

- if, for a given time-frequency tile, the one or more criteria are met, the direction decoder 364 operates to determine the corresponding quantized DOA parameter(s) 309' for the given time-frequency tile using the first DOA quantizer and provides the quantized DOA parameter(s) 309' for an audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235;
- if, for a given time-frequency tile, the one or more criteria are not met, the direction decoder 364 operates to determine the corresponding quantized DOA parameter(s) 309' for the given time-frequency tile using the second DOA quantizer and provides the quantized DOA parameter(s) 309' for an audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235.

[0092] In an example, the one or more criteria that pertain to respective absolute energy levels of one or more sound components of the audio signal represented by the audio bitstream 225 in a given time-frequency tile involves consideration of directional energy level of the single directional sound component of the audio scene represented by the audio bitstream 225 (and hence by the multi-channel input audio signal 215). In this regard, the first and/or second exemplifying criteria described in the foregoing in context of the direction encoder 314 may be applied, i.e. the one or more criteria may pertain to DEN parameters and/or QEN parameters, where the QEN parameters may be derived for example based on the reconstructed downmix signal 303' obtained in the audio decoder 354.

[0093] A third example for deriving the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more time-frequency tiles proceeds from the assumption that at least for some time-frequency tiles considered in the spatial analysis in the spatial audio encoder 300 the DOA parameters 309 for a given time-frequency tile pertain to two or more (simultaneous) directional sound components and, hence, for such time-frequency tiles there is at least one encoded DOA parameter 313 available for each of the two more directional sound components.

[0094] In the third example, along the lines described in the foregoing in context of the ratio quantizer 312 and the direction quantizer 314, there may be predefined total number of bits available for encoding the ER parameters 311 and DOA parameters 309 for a given time-frequency tile. As described in the foregoing in context of the spatial audio encoder 300, in a first scenario this total number of bits may be evenly allocated for the two or more directional sound components of a time-frequency tile such that ER parameter 311 and the DOA parameter(s) 309 pertaining to a given directional sound component of a given time-frequency tile are encoded using at most a respective predefined number of bits: the encoded ER parameter 316 for the given directional sound component is represented in the audio bitstream 225 using a variable number of bits resulting from operation of the variable bit-rate ER quantizer, whereas the bits remaining after the ER parameter encoding are used for encoding the DOA parameter(s) for the given directional sound component by using selected one of a plurality of fixed bit-rate DOA quantizers. The selected DOA quantizer is the one that uses the highest number of bits that does not exceed the predetermined number of bits available for encoding the ER parameter 311 and the DOA parameter(s) 309 for the given directional sound component. Thus, the direction decoder 364 may detect the selected DOA quantizer based on knowledge of the predefined total number of bits available for encoding the ER and DOA parameters for the given directional sound component and the bits employed for representing the encoded ER parameter 316 via usage of the variable rate ER quantizer.

[0095] In a second scenario of the third example, there is the predefined first number of bits for encoding the two or more ER parameters 311 and the two or more DOA parameters 309 of the given time-frequency tile: each of the two or more ER parameters 311 of the given time-frequency tile are encoded using the variable bit-rate ER quantizer described in the foregoing, whereas the remaining bits are available for encoding the two or more DOA parameters 309 of the given time-frequency tile. As described in the foregoing in context of the spatial audio encoder 300, these remaining bits constitute a second number of bits and they are allocated for encoding the DOA parameters 309 pertaining to respective two or more directional sound components of the given time-frequency tile in dependence of respective energy levels indicated for the two or more directional sound components of the given time-frequency tile.

[0096] As an example in this regard, the direction decoder 364 may apply the first predefined bit-allocation rule to find the respective (maximum) number of bits available for encoding the respective DOA parameter(s) 309 in accordance with respective energy levels indicated for each of the two or more directional sound components, which enables determining the one of the plurality of fixed bit-rate DOA quantizer employed for deriving the encoded DOA parameter(s) 313, i.e. the one of the available DOA quantizers having the highest bit-rate that does not exceed the number of bits available for encoding the respective DOA parameter(s) 309.

[0097] As described in the foregoing in context of the spatial audio encoder 300, in non-limiting examples the comparison of energy levels may rely on comparison of directional energy levels indicated by the DEN respective parameters derived for the two or more directional sound components in a given time-frequency tile or the comparison may rely on comparison of direct-to-total-energy ratios indicated by the respective quantized ER parameters 316 obtained for the given time-frequency tile. As non-limiting examples in this regard, determination of the distribution of bits for the encoded DOA parameters 313 of two or more respective directional sound components of a given time-frequency tile in the direction decoder 364 may involve one or more of the following:

- In case the energy level indicated for a first directional sound component in the given time-frequency tile exceeds the energy level indicated for a second directional sound component in the given time-frequency tile by more than the first predefined margin, the DOA parameter(s) pertaining to the first directional sound component are assigned a first number of bits and the DOA parameter(s) pertaining to the second directional sound component are assigned a second number of bits, where the first number of bits is higher than the second number of bits. Consequently, the encoded DOA parameter(s) 313 pertaining to the first directional sound component are decoded using a first DOA quantizer that uses at

most the first number of bits and the encoded DOA parameter(s) 313 pertaining to the second directional sound component are decoded using a second DOA quantizer that uses at most the second number of bits, where the first DOA quantizer (that employs a higher number of bits) enables representing the quantized DOA parameter 309' at a higher precision than the second DOA quantizer (that employs a lower number of bits). The quantized DOA parameters 309' are provided for the audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235.

- In case the energy level indicated for a first directional sound component in the given time-frequency tile exceeds the energy level indicated for a second directional sound component in the given time-frequency tile by more than the second predefined margin, the DOA parameter(s) pertaining to the first directional sound component are assigned a first number of bits and, consequently, the corresponding encoded DOA parameter(s) 313 are decoded using a DOA quantizer available for the direction decoder 364 whereas the encoded DOA parameter(s) pertaining to the second directional sound source are not received in the audio bitstream 225. The quantized DOA parameters 309', if available, are provided for the audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235.

[0098] In the latter example above, instead of completely omitting or excluding the corresponding directional sound component from the reconstructed audio signal, the direction decoder 364 may respond to a failure to meet the one or more criteria by introducing the default DOA parameter(s) for the given directional sound component in the given time-frequency tile and providing the introduced DOA parameter(s) 313 for the audio synthesizer for creation of the corresponding directional sound component for the reconstructed audio signal 235.

[0099] In a fourth example for deriving the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more time-frequency tiles, which is described herein as a variation of the second scenario of the third example, the directional energy-level dependent bit allocation for quantizing the DOA parameters 309 for two or more directional sound components of a given time-frequency tile is provided further in view of the third threshold: the predetermined fixed number of bits available for encoding the two or more DOA parameters 309 of the given time-frequency tile is allocated for the two or more encoded DOA parameters 313 of the given time-frequency tile in view of their respective relationships with the third threshold.

[0100] The direction decoder 364 may derive the bit allocation across the two or more encoded DOA parameters 309 of the given time-frequency tile by applying the second predefined bit allocation rule that defines a re-

spective (maximum) number of bits available for encoding the respective DOA parameter(s) 309 in accordance with respective energy levels indicated for each of the two or more directional sound components in view of their relationship with the third threshold. The direction decoder 364 may use the bit allocation so obtained to select, for each of the directional sound components of the given time-frequency tile, a respective one of the plurality of DOA quantizers (described in the foregoing) in accordance with the number of bits allocated for encoding those DOA parameter(s) 309 in the given time-frequency tile. The relationship between an energy level and the third threshold may be derived, for example, via the comparison value derived by a ratio of the energy level and the third threshold (e.g. by dividing the energy level by the third threshold) or by a difference between the energy level and the third threshold (e.g. by subtracting the third threshold from the energy level). The respective comparison values for the two or more directional sound components may be derived along the lines described in the foregoing in context of the spatial audio encoder 300.

[0101] As non-limiting examples in this regard, determination of the distribution of bits for the encoded DOA parameters 313 of two or more respective directional sound components of a given time-frequency tile in the direction decoder 364 may involve one or more of the following:

- In case the comparison value derived for a first directional sound component in the given time-frequency tile exceeds the comparison value derived for a second directional sound component in the given time-frequency tile by more than the third predefined margin, the DOA parameters pertaining to the first directional sound component are assigned a first number of bits and the DOA parameters pertaining to the second directional sound component are assigned a second number of bits, where the first number of bits is higher than the second number of bits. Consequently, the encoded DOA parameter(s) 313 pertaining to the first directional sound component are decoded using a first DOA quantizer that uses at most the first number of bits and the encoded DOA parameter(s) 313 pertaining to the second directional sound component are decoded using a second DOA quantizer that uses at most the second number of bits, where the first DOA quantizer (that employs a higher number of bits) enables encoding the quantized value of a DOA parameter 309 at a higher precision than the second DOA quantizer (that employs a lower number of bits). The quantized DOA parameters 309' are provided for the audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235.
- In case the comparison value derived for a first directional sound component in the given time-frequency tile exceeds the comparison value derived

for a second directional sound component in the given time-frequency tile by more than the fourth predefined margin, the DOA parameter(s) 309 pertaining to the first directional sound component are assigned a first number bits and are, consequently, the corresponding encoded DOA parameters(s) 313 are decoded using a DOA quantizer available for the direction decoder 364 whereas encoded DOA parameter(s) pertaining to the second directional sound source are not received in the audio bitstream 225. The quantized DOA parameters 309', if available, are provided for the audio synthesizer 370 for creation of the corresponding directional sound component for the reconstructed audio signal 235.

[0102] In the latter example above, instead of completely omitting or excluding the corresponding directional sound component from the reconstructed audio signal, the direction decoder 364 may respond to a failure to meet the one or more criteria by introducing the default DOA parameter(s) for the given directional sound component in the given time-frequency tile and providing the introduced DOA parameter(s) 313 for the audio synthesizer for creation of the corresponding directional sound component for the reconstructed audio signal 235.

[0103] The audio synthesizer 370 receives the reconstructed downmix signal 303', the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more frequency sub-bands and derives the reconstructed audio signal 235 based on this information. Typically, in order to reconstruct the audio scene represented by the input audio signal 215, the reconstructed audio signal 235 is provided as a multi-channel spatial audio signal that includes two or more channels. The number of channels in the reconstructed audio signal 235 may be different from that of the input audio signal 215. The reconstructed audio signal 235 may be provided, for example, as a multi-channel signal according to a predefined loudspeaker configuration (such as two-channel stereo, 5.1 surround sound, 7.1 surround sound, 22.2 surround sound, etc.) or as a binaural audio signal for headphone listening.

[0104] The reconstructed downmix signal 303' may be provided as a transform-domain (e.g. a frequency-domain) signal, depending on the characteristics of the audio coding technique employed by the audio encoder 304 (in the spatial audio encoder 300) and the audio decoder 354 (in the spatial audio decoder 350). If the reconstructed downmix signal 303' is not provided as a transform-domain (e.g. a frequency-domain) signal, the audio synthesizer 370 may apply a suitable transform technique to convert the reconstructed downmix signal 303' into a corresponding transform-domain signal. The transform-domain reconstructed downmix signal may be divided into a plurality of frequency sub-bands. The transform and the frequency sub-band division are preferably the same applied in the transform entity 306 of the spatial audio encoder 310 to enable direct application of the

quantized ER parameters 316 and the quantized DOA parameters 309' for deriving a reconstructed transform-domain audio signal in the respective one or more frequency sub-bands.

[0105] The audio synthesis may be provided using a suitable spatial audio synthesis technique known in the art. Non-limiting examples of applicable techniques are described for example in the following publications:

- 10 - Pulkki, Ville, "Spatial sound reproduction with directional audio coding", *Journal of the Audio Engineering Society* 55, no. 6 (2007), pp. 503-516 discloses a technique for rendering spatial sound for loudspeaker listening;
- 15 - Laitinen, Mikko-Ville; Pulkki, Ville, "Binaural reproduction for directional audio coding", *Applications of Signal Processing to Audio and Acoustics*, 2009. WASPAA'09, pp. 337-340. IEEE, 2009 discloses a technique for rendering spatial sound for headphone listening;
- 20 - Vilkkamo, Juha; Pulkki, Ville, "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering", *Journal of the Audio Engineering Society* 61, no. 9 (2013), pp. 637-646 discloses a technique for reproducing spatial sound based on direction and energy ratio metadata; and
- 25 - Politis, Archontis; Vilkkamo, Juha; Pulkki, Ville, "Sector-based parametric sound field reproduction in the spherical harmonic domain", *IEEE Journal of Selected Topics in Signal Processing* 9, no. 5 (2015), pp. 852-866 discloses a technique for reproducing spatial sound based on multiple simultaneous directions and energy ratios.
- 30

[0106] After the spatial audio synthesis in the transform domain, the audio synthesizer 370 may further apply an applicable inverse transform to convert the reconstructed transform-domain audio signal into time domain for provision as the reconstructed audio signal 235 (e.g.) for the audio reproduction entity 240.

[0107] Figure 4 depicts a spatial audio decoder 450, where a difference to the spatial audio decoder 350 is that the spatial audio decoder 450 is arranged for derivation of the reconstructed audio signal 235 on basis of the audio bitstream 225 generated by the spatial audio encoder where metadata derivation is at least partially integrated to the encoding and/or quantization of the audio signal.

[0108] The spatial audio decoder 450 includes a bitstream unpacker 468 for extracting the encoded audio signal 405 and the spatial metadata from the audio bitstream 225 for a given input frame. The spatial metadata may comprise the encoded ER parameter(s) 315 and/or the encoded DOA parameter(s) 313 for one or more frequency sub-bands derived in the spatial audio encoder 400 for the given input frame.

[0109] The spatial audio decoder 450 includes an audio dequantizer 454 for processing the encoded audio

signal 405 into a reconstructed (transform-domain) downmix signal 403' and an energy estimator 308b for estimating the overall signal energy of the reconstructed transform-domain downmix signal 403'. In this regard, the energy estimator 308b may derive, for each time-frequency tile considered in the spatial synthesis, a respective QEN parameter 410 that indicates estimated overall signal energy in a respective time-frequency tile.

[0110] The spatial audio decoder 450 further includes the ratio decoder 362 for decoding the encoded ER parameters 315 into corresponding quantized ER parameters 316 and the direction decoder 364 for decoding the encoded DOA parameters 313 into corresponding quantized DOA parameters 309'. The operation and characteristics of these entities are described in detail in the foregoing via a plurality of examples that pertain to their usage as part of the spatial audio decoder 350.

[0111] The spatial audio decoder 450 further comprises an audio synthesizer 370 receives the reconstructed (transform-domain) downmix signal 403', the quantized ER parameters 316 and the quantized DOA parameters 309' for one or more frequency sub-bands and derives the reconstructed audio signal 235 based on this information. The operation and characteristics of the audio synthesizer 470 is similar to those of the audio synthesizer 370 described in the foregoing in context of the spatial audio decoder 350.

[0112] Components of the spatial audio encoder 300, 400 may be arranged to operate, for example, in accordance with a method 500 illustrated by a flowchart depicted in Figure 7A. The method 500 serves as a method for encoding a multi-channel input audio signal that represents an audio scene as an encoded audio signal and spatial audio parameters, wherein the spatial audio parameters are descriptive of said audio scene.

[0113] The method 500 comprises encoding a frame of a downmix signal into a frame of the encoded audio signal, wherein the downmix signal is generated from the multi-channel input audio signal 215 into a frame of the encoded audio signal 305, 405, as indicated in block 502. The method 500 further comprises deriving, from the frame of the multi-channel input audio signal 215, a plurality of spatial audio parameters that are descriptive of the audio scene in the corresponding frame, the spatial audio parameters comprising a plurality of DOA parameters 311, wherein a DOA parameter indicates a spatial position of a given directional sound component of the audio scene in a given frequency sub-band, as indicated in block 504. The method 500 further comprises encoding the spatial audio parameters, comprising encoding a DOA parameter 313 for a given directional sound component in a given frequency sub-band in dependence of an energy level of the given directional sound component in the given frequency sub-band meeting one or more criteria.

[0114] Components of the spatial audio decoder 350, 450 may be arranged to operate, for example, in accordance with a method 550 illustrated by a flowchart depicted

in Figure 7B. The method 550 serves as a method for reconstructing a spatial audio signal that represents an audio scene on basis of an encoded audio signal and encoded spatial audio parameters that are descriptive of said audio scene.

[0115] The method 550 comprises decoding a frame of the encoded audio signal 305, 405 into a frame of a reconstructed downmix signal 303', 403', as indicated in block 552. The method 550 further comprises receiving a plurality of encoded spatial audio parameters that are descriptive of the audio scene in a frame of the reconstructed audio signal 235, the encoded spatial audio parameters comprising a plurality of DOA parameters 313, wherein a DOA parameter 313 indicates a spatial position of a given directional sound component of the audio scene in a given frequency sub-band, as indicated in block 554. The method 550 further comprises decoding the encoded spatial audio parameters, comprising decoding a DOA parameter 313 for a given directional sound component in a given frequency sub-band in dependence of an energy level of the given directional sound component in the given frequency sub-band meeting one or more criteria, as indicated in block 556.

[0116] The method 500 may be varied in a number of ways in view of the examples concerning operation of the spatial audio encoder 300, 400 described in the foregoing, whereas the method 550 may be varied in a number of ways in view of the examples concerning operation of the spatial audio decoder 350, 450 described in the foregoing.

[0117] Figure 8 illustrates a block diagram of some components of an exemplifying apparatus 600. The apparatus 600 may comprise further components, elements or portions that are not depicted in Figure 8. The apparatus 600 may be employed e.g. in implementing one or more components described in the foregoing in context of the spatial audio encoder 300, 400 and/or one or more components described in the foregoing in context of the spatial audio decoder 350, 450.

[0118] The apparatus 600 comprises a processor 616 and a memory 615 for storing data and computer program code 617. The memory 615 and a portion of the computer program code 617 stored therein may be further arranged to, with the processor 616, to implement at least some of the operations, procedures and/or functions described in the foregoing in context of the spatial audio encoder 300, 400 and/or in context of the spatial audio decoder 350, 450.

[0119] The apparatus 600 comprises a communication portion 612 for communication with other devices. The communication portion 612 comprises at least one communication apparatus that enables wired or wireless communication with other apparatuses. A communication apparatus of the communication portion 612 may also be referred to as a respective communication means.

[0120] The apparatus 600 may further comprise user I/O (input/output) components 618 that may be arranged,

possibly together with the processor 616 and a portion of the computer program code 617, to provide a user interface for receiving input from a user of the apparatus 600 and/or providing output to the user of the apparatus 600 to control at least some aspects of operation of the spatial audio encoder 300, 400 and/or spatial audio decoder 350, 450 that are implemented by the apparatus 600. The user I/O components 618 may comprise hardware components such as a display, a touchscreen, a touchpad, a mouse, a keyboard, and/or an arrangement of one or more keys or buttons, etc. The user I/O components 618 may be also referred to as peripherals. The processor 616 may be arranged to control operation of the apparatus 600 e.g. in accordance with a portion of the computer program code 617 and possibly further in accordance with the user input received via the user I/O components 618 and/or in accordance with information received via the communication portion 612.

[0121] Although the processor 616 is depicted as a single component, it may be implemented as one or more separate processing components. Similarly, although the memory 615 is depicted as a single component, it may be implemented as one or more separate components, some or all of which may be integrated/removable and/or may provide permanent / semi-permanent! dynamic/cached storage.

[0122] The computer program code 617 stored in the memory 615, may comprise computer-executable instructions that control one or more aspects of operation of the apparatus 600 when loaded into the processor 616. As an example, the computer-executable instructions may be provided as one or more sequences of one or more instructions. The processor 616 is able to load and execute the computer program code 617 by reading the one or more sequences of one or more instructions included therein from the memory 615. The one or more sequences of one or more instructions may be configured to, when executed by the processor 616, cause the apparatus 600 to carry out at least some of the operations, procedures and/or functions described in the foregoing in context of the spatial audio encoder 300, 400 and/or the spatial audio decoder 350, 450.

[0123] Hence, the apparatus 600 may comprise at least one processor 616 and at least one memory 615 including the computer program code 617 for one or more programs, the at least one memory 615 and the computer program code 617 configured to, with the at least one processor 616, cause the apparatus 600 to perform at least some of the operations, procedures and/or functions described in the foregoing in context of the spatial audio encoder 300, 400 and/or the spatial audio decoder 350, 450.

[0124] The computer programs stored in the memory 615 may be provided e.g. as a respective computer program product comprising at least one computer-readable non-transitory medium having the computer program code 617 stored thereon, the computer program code, when executed by the apparatus 600, causes the appa-

ratus 600 at least to perform at least some of the operations, procedures and/or functions described in the foregoing in context of the spatial audio encoder 300, 400 and/or the spatial audio decoder 350, 450. The computer-readable non-transitory medium may comprise a memory device or a record medium such as a CD-ROM, a DVD, a Blu-ray disc or another article of manufacture that tangibly embodies the computer program. As another example, the computer program may be provided as a signal configured to reliably transfer the computer program.

[0125] Reference(s) to a processor should not be understood to encompass only programmable processors, but also dedicated circuits such as field-programmable gate arrays (FPGA), application specific circuits (ASIC), signal processors, etc. Features described in the preceding description may be used in combinations other than the combinations explicitly described.

[0126] Although functions have been described with reference to certain features, those functions may be performable by other features whether described or not. Although features have been described with reference to certain embodiments, those features may also be present in other embodiments whether described or not.

Claims

1. An apparatus for encoding a multi-channel input audio signal that represents an audio scene as an encoded audio signal and spatial audio parameters, wherein the spatial audio parameters are descriptive of said audio scene, the apparatus comprising:

means for encoding a frame of a downmix signal into a frame of the encoded audio signal, wherein the downmix signal is generated from the multi-channel input audio signal;

means for deriving, from the frame of the multi-channel input audio signal, a plurality of spatial audio parameters that are descriptive of the audio scene, said spatial audio parameters comprising a plurality of direction of arrival (DOA) parameters, wherein a DOA parameter indicates a spatial position of a given directional sound component of the audio scene in a given frequency sub-band; and

means for encoding the spatial audio parameters, comprising means for encoding a DOA parameter for a given directional sound component in a given frequency sub-band, wherein the apparatus is **characterized in that** the means for encoding a DOA parameter for a given directional sound component in a given frequency sub band is in dependence of an absolute energy level of the given directional sound component in the given frequency sub band exceeding a first threshold and in dependence of a total energy of said multi-channel input audio signal

in the given frequency sub band exceeding a second threshold.

2. An apparatus according to claim 1,

wherein said spatial audio parameters comprise a plurality of energy ratio (ER) parameters, wherein an ER parameter indicates a relative energy level of a given directional sound component of the audio scene in a given frequency sub-band, and wherein the means for encoding a DOA parameter comprises:

means for deriving a plurality of directional energy (DEN) parameters, wherein a DEN parameter that indicates said absolute energy level of a given directional sound component in a given frequency sub-band is derived on basis of a total energy of said multi-channel input audio signal in the given frequency sub-band and the ER parameter obtained for the given directional sound component in the given frequency sub-band, and

means for encoding said DOA parameter in dependence of said DEN parameter derived for the given directional sound component in the given frequency sub-band exceeding said first threshold.

3. An apparatus according to claim 2, comprising means for computing the total energy of said multi-channel input audio signal for the given frequency sub-band on basis of a frame of reconstructed audio signal derived by decoding the frame of the encoded audio signal.

4. An apparatus according to claim 2 or 3, wherein

the apparatus comprises means for encoding said plurality of ER parameters, arranged to derive a respective plurality of quantized ER parameters; and

the means for deriving the plurality of DEN parameters is arranged to derive a DEN parameter that indicates the absolute energy level of the given directional sound component in the given frequency sub-band dependent on the quantized ER parameter obtained for the given directional sound component for the given frequency sub-band.

5. An apparatus according to any of claims 2 to 4, wherein an ER parameter indicates a ratio between the energy level of a given directional sound component in a given frequency sub-band and the total energy of said multi-channel input audio signal in the

given frequency sub-band.

6. An apparatus according to any of claims 1 to 5, wherein at least one of the first and second thresholds is a respective predefined threshold assigned for the given frequency sub-band.

7. An apparatus for reconstructing a spatial audio signal that represents an audio scene based on an encoded audio signal and encoded spatial audio parameters, wherein the spatial audio parameters are descriptive of said audio scene, the apparatus comprising:

means for decoding a frame of the encoded audio signal into a frame of a reconstructed down-mix signal;

means for receiving a plurality of encoded spatial audio parameters that are descriptive of said audio scene in a frame of the spatial audio signal, said encoded spatial audio parameters comprising a plurality of direction of arrival, DOA, parameters, wherein a DOA parameter indicates a spatial position of a given directional sound component of the audio scene in a given frequency sub-band; and

means for decoding said encoded spatial audio parameters, comprising means for decoding a DOA parameter for a given directional sound component in a given frequency sub-band, wherein the apparatus is **characterized in that** the means for decoding a DOA parameter for a given directional sound component in a given frequency sub band is in dependence of an absolute energy level of the given directional sound component in the given frequency sub band exceeding a first threshold and in dependence of a total energy of said multi-channel input audio signal in the given frequency sub band exceeding a second threshold.

8. An apparatus according to claim 7,

wherein said spatial audio parameters comprise a plurality of energy ratio, ER, parameters, wherein an ER parameter indicates a relative energy level of a given directional sound component of the audio scene in a given frequency sub-band, wherein the means for decoding of a DOA parameter comprises

means for deriving a plurality of directional energy, DEN, parameters, wherein a DEN parameter that indicates said absolute energy level of a given directional sound component in a given frequency sub-band is derived on basis of a total energy of the frame

of the spatial audio signal in the given frequency sub-band and the ER parameter received for the given directional sound component in the given frequency sub-band, and

means for decoding said DOA parameter in dependence of said DEN parameter derived for the given directional sound component in the given frequency sub-band exceeding said first threshold.

9. An apparatus according to claim 8, further comprising means for computing the total energy of said frame of the spatial audio signal for the given frequency sub-band on basis of said frame of reconstructed downmix signal.

10. An apparatus according to claim 8 or 9, wherein an ER parameter indicates a ratio between the energy level of a given directional sound component in a given frequency sub-band and the total energy in the given frequency sub-band.

11. An apparatus according to any of claims 7 to 10, wherein at least one of the first and second thresholds is a respective predefined threshold assigned for the given frequency sub-band.

Patentansprüche

1. Vorrichtung zum Codieren eines Mehrkanalaudioeingangssignals, das eine Audioszene repräsentiert, als ein codiertes Audiosignal und räumliche Audioparameter, wobei die räumlichen Audioparameter die Audioszene beschreiben, wobei die Vorrichtung Folgendes umfasst:

Mittel zum Codieren eines Frames eines Abwärtsmischsignals in einen Frame des codierten Audiosignals, wobei das Abwärtsmischsignal aus dem Mehrkanalaudioeingangssignal erzeugt wird;

Mittel zum Ableiten einer Vielzahl von räumlichen Audioparametern, die die Audioszene beschreiben, aus dem Frame des Mehrkanalaudioeingangssignals, wobei die räumlichen Audioparameter eine Vielzahl von Ankunftsrichtungs(DOA)-Parametern umfassen, wobei ein DOA-Parameter eine räumliche Position einer gegebenen Richtungsschallkomponente der Audioszene in einem gegebenen Frequenzunterband anzeigt; und

Mittel zum Codieren der räumlichen Audioparameter, die Mittel zum Codieren eines DOA-Parameters für eine gegebene Richtungsschallkomponente in einem gegebenen Frequenzunterband umfassen, wobei die Vorrichtung da-

durch gekennzeichnet ist, dass die Mittel zum Codieren eines DOA-Parameters für eine gegebene Richtungsschallkomponente in einem gegebenen Frequenzunterband davon abhängig ist, dass ein absoluter Energiepegel der gegebenen Richtungsschallkomponente im gegebenen Frequenzunterband einen ersten Schwellwert überschreitet, und davon, dass eine Gesamtenergie des Mehrkanalaudioeingangssignals im gegebenen Frequenzunterband einen zweiten Schwellwert überschreitet.

2. Vorrichtung nach Anspruch 1,

wobei die räumlichen Audioparameter eine Vielzahl von Energieverhältnis(ER)-Parametern umfassen, wobei ein ER-Parameter einen relativen Energiepegel einer gegebenen Richtungsschallkomponente der Audioszene in einem gegebenen Frequenzunterband anzeigt, und wobei die Mittel zum Codieren eines DOA-Parameters Folgendes umfassen:

Mittel zum Ableiten einer Vielzahl von Richtungsenergie(DEN)-Parametern, wobei ein DEN-Parameter, der den absoluten Energiepegel einer gegebenen Richtungsschallkomponente in einem gegebenen Frequenzunterband anzeigt, auf Basis einer Gesamtenergie des Mehrkanalaudioeingangssignals im gegebenen Frequenzunterband und vom ER-Parameter, der für die gegebene Richtungsschallkomponente im Frequenzunterband erhalten wird, abgeleitet wird, und

Mittel zum Codieren des DOA-Parameters in Abhängigkeit davon, dass der DEN-Parameter, der für die gegebene Richtungsschallkomponente im gegebenen Frequenzunterband abgeleitet wird, den ersten Schwellwert überschreitet.

3. Vorrichtung nach Anspruch 2, die Folgendes umfasst

Mittel zum Berechnen der Gesamtenergie des Mehrkanalaudioeingangssignals für das gegebene Frequenzunterband auf Basis eines Frames eines rekonstruierten Audiosignals, das durch Decodieren des Frames des codierten Audiosignals abgeleitet wird.

4. Vorrichtung nach Anspruch 2 oder 3, wobei

die Vorrichtung Mittel zum Codieren der Vielzahl von ER-Parametern umfasst, die angeordnet sind, eine jeweilige Vielzahl von quantisierten ER-Parametern abzuleiten; und die Mittel zum Ableiten der Vielzahl von DEN-

Parametern angeordnet sind, einen DEN-Parameter, der den absoluten Energiepegel der gegebenen Richtungsschallkomponente im gegebenen Frequenzunterband in Abhängigkeit vom quantisierten ER-Parameter ableitet, der für die gegebene Richtungsschallkomponente für das gegebene Frequenzunterband erhalten wird.

5. Vorrichtung nach einem der Ansprüche 2 bis 4, wobei ein ER-Parameter ein Verhältnis zwischen dem Energiepegel einer gegebenen Richtungsschallkomponente in einem gegebenen Frequenzunterband und der Gesamtenergie des Mehrkanalaudioeingangssignals im gegebenen Frequenzunterband anzeigt. 5 10
6. Vorrichtung nach einem der Ansprüche 1 bis 5, wobei mindestens einer des ersten und des zweiten Schwellwerts ein jeweiliger vordefinierter Schwellwert ist, der dem gegebenen Frequenzunterband zugewiesen ist. 20
7. Vorrichtung zum Rekonstruieren eines räumlichen Audiosignals, das eine Audioszene repräsentiert, auf Basis eines codierten Audiosignals und codierten räumlichen Audioparametern, wobei die räumlichen Audioparameter die Audioszene beschreiben, wobei die Vorrichtung Folgendes umfasst: 25

Mittel zum Decodieren eines Frames des codierten Audiosignals in einen Frame eines rekonstruierten Abwärtsmischsignals; Mittel zum Empfangen einer Vielzahl von codierten räumlichen Audioparametern, die die Audioszene beschreiben, in einem Frame des räumlichen Audiosignals, wobei die codierten räumlichen Audioparameter eine Vielzahl von Ankunftsrichtungen(DOA)-Parametern umfassen, wobei ein DOA-Parameter eine räumliche Position einer gegebenen Richtungsschallkomponente der Audioszene in einem gegebenen Frequenzunterband anzeigt; und Mittel zum Decodieren der codierten räumlichen Audioparameter, die Mittel zum Decodieren eines DOA-Parameters für eine gegebene Richtungsschallkomponente in einem gegebenen Frequenzunterband umfassen, wobei die Vorrichtung **dadurch gekennzeichnet ist, dass** die Mittel zum Decodieren eines DOA-Parameters für eine gegebene Richtungsschallkomponente in einem gegebenen Frequenzunterband davon abhängig ist, dass ein absoluter Energiepegel der gegebenen Richtungsschallkomponente im gegebenen Frequenzunterband einen ersten Schwellwert überschreitet, und davon, dass eine Gesamtenergie des Mehrkanalaudioeingangssignals im gegebenen Frequenzunterband einen zweiten Schwellwert

überschreitet.

8. Vorrichtung nach Anspruch 7,

wobei die räumlichen Audioparameter eine Vielzahl von Energieverhältnis(ER)-Parametern umfassen, wobei ein ER-Parameter einen relativen Energiepegel einer gegebenen Richtungsschallkomponente der Audioszene in einem gegebenen Frequenzunterband anzeigt, wobei die Mittel zum Decodieren eines DOA-Parameters Folgendes umfassen Mittel zum Ableiten einer Vielzahl von Richtungsenergie(DEN)-Parametern, wobei ein DEN-Parameter, der den absoluten Energiepegel einer gegebenen Richtungsschallkomponente in einem gegebenen Frequenzunterband anzeigt, auf Basis einer Gesamtenergie des Frames des räumlichen Audiosignals im gegebenen Frequenzunterband und vom ER-Parameter, der für die gegebene Richtungsschallkomponente im Frequenzunterband empfangen wird, abgeleitet wird, und Mittel zum Decodieren des DOA-Parameters in Abhängigkeit davon, dass der DEN-Parameter, der für die gegebene Richtungsschallkomponente im gegebenen Frequenzunterband abgeleitet wird, den ersten Schwellwert überschreitet.

9. Vorrichtung nach Anspruch 8, die ferner Folgendes umfasst

Mittel zum Berechnen der Gesamtenergie des Frames des räumlichen Audiosignals für das gegebene Frequenzunterband auf Basis des Frames des rekonstruierten Abwärtsmischsignals.

10. Vorrichtung nach Anspruch 8 oder 9, wobei ein ER-Parameter ein Verhältnis zwischen dem Energiepegel einer gegebenen Richtungsschallkomponente in einem gegebenen Frequenzunterband und der Gesamtenergie im gegebenen Frequenzunterband anzeigt.

11. Vorrichtung nach einem der Ansprüche 7 bis 10, wobei mindestens einer des ersten und des zweiten Schwellwerts ein jeweiliger vordefinierter Schwellwert ist, der dem gegebenen Frequenzunterband zugewiesen ist.

Revendications

1. Appareil de codage d'un signal audio d'entrée multicanal qui représente une scène audio en signal audio codé et paramètres audio spatiaux, dans lequel les paramètres audio spatiaux décrivent ladite scène audio, l'appareil comprenant :

un moyen pour coder une trame d'un signal de mixage réducteur en une trame du signal audio codé, dans lequel le signal de mixage réducteur est généré à partir du signal audio d'entrée multicanal ;

un moyen pour dériver, à partir de la trame du signal audio d'entrée multicanal, une pluralité de paramètres audio spatiaux qui décrivent la scène audio, lesdits paramètres audio spatiaux comprenant une pluralité de paramètres de direction d'arrivée (DOA), dans lequel un paramètre DOA indique une position spatiale d'une composante sonore directionnelle donnée de la scène audio dans une sous-bande de fréquence donnée ; et

un moyen pour coder les paramètres audio spatiaux, comprenant un moyen pour coder un paramètre DOA pour une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée, dans lequel l'appareil est **caractérisé en ce que** le moyen pour coder un paramètre DOA pour une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée dépend d'un niveau d'énergie absolu de la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée dépassant un premier seuil, et dépend d'une énergie totale dudit signal audio d'entrée multicanal dans la sous-bande de fréquence donnée dépassant un deuxième seuil.

2. Appareil selon la revendication 1,

dans lequel lesdits paramètres audio spatiaux comprennent une pluralité de paramètres de rapport énergétique (ER), dans lequel un paramètre ER indique un niveau d'énergie relatif d'une composante sonore directionnelle donnée du scène audio dans un sous-bande de fréquence donnée, et

dans lequel le moyen pour coder un paramètre DOA comprend :

un moyen pour dériver une pluralité de paramètres d'énergie directionnelle (DEN), dans lequel un paramètre DEN indiquant ledit niveau d'énergie absolu d'une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée est dérivé sur la base d'une énergie totale dudit signal audio d'entrée multicanal dans la sous-bande de fréquence donnée et du paramètre ER obtenu pour la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée, et

un moyen pour coder ledit paramètre DOA en fonction dudit paramètre DEN dérivé pour la composante sonore directionnelle

donnée dans la sous-bande de fréquence donnée dépassant ledit premier seuil.

3. Appareil selon la revendication 2, comprenant un moyen pour calculer l'énergie totale dudit signal audio d'entrée multicanal pour la sous-bande de fréquence donnée sur la base d'une trame du signal audio reconstruit dérivé en décodant la trame du signal audio codé.

4. Appareil selon la revendication 2 ou 3, dans lequel l'appareil comprend un moyen pour coder ladite pluralité de paramètres ER, agencé pour dériver une pluralité respective de paramètres ER quantifiés ; et le moyen pour dériver la pluralité de paramètres DEN est agencé pour dériver un paramètre DEN qui indique le niveau d'énergie absolu de la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée en fonction du paramètre ER quantifié obtenu pour la composante sonore directionnelle donnée pour la sous-bande de fréquence donnée.

5. Appareil selon l'une des revendications 2 à 4, dans lequel un paramètre ER indique un rapport entre le niveau d'énergie d'une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée et l'énergie totale dudit signal audio d'entrée multicanal dans la sous-bande de fréquence donnée.

6. Appareil selon l'une des revendications 1 à 5, dans lequel au moins l'un des premier et deuxième seuils est un seuil prédéfini respectif attribué à la sous-bande de fréquence donnée.

7. Appareil permettant de reconstruire un signal audio spatial qui représente une scène audio sur la base d'un signal audio codé et de paramètres audio spatiaux codés, dans lequel les paramètres audio spatiaux décrivent ladite scène audio, l'appareil comprenant :

un moyen pour décoder une trame du signal audio codé en une trame d'un signal de mixage réducteur reconstruit ;

un moyen pour recevoir une pluralité de paramètres audio spatiaux codés qui décrivent ladite scène audio dans une trame du signal audio spatial, lesdits paramètres audio spatiaux codés comprenant une pluralité de paramètres de direction d'arrivée, DOA, dans lequel un paramètre DOA indique une position spatiale d'une composante sonore directionnelle donnée de la scène audio dans une sous-bande de fréquence donnée ; et

un moyen pour décoder lesdits paramètres audio spatiaux codés, comprenant un moyen pour décoder un paramètre DOA pour une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée, dans lequel l'appareil est **caractérisé en ce que** le moyen pour décoder un paramètre DOA pour une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée dépend d'un niveau d'énergie absolu de la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée dépassant un premier seuil, et dépend d'une énergie totale dudit signal audio d'entrée multicanal dans la sous-bande de fréquence donnée dépassant un deuxième seuil.

lequel au moins l'un des premier et deuxième seuils est un seuil prédéfini respectif attribué à la sous-bande de fréquence donnée.

8. Appareil selon la revendication 7,

dans lequel lesdits paramètres audio spatiaux comprennent une pluralité de paramètres de rapport énergétique, ER, dans lequel un paramètre ER indique un niveau d'énergie relatif d'une composante sonore directionnelle donnée du scène audio dans un sous-bande de fréquence donnée, dans lequel le moyen pour décoder un paramètre DOA comprend un moyen pour dériver une pluralité de paramètres d'énergie directionnelle, DEN, dans lequel un paramètre DEN indiquant ledit niveau d'énergie absolu d'une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée est dérivé sur la base d'une énergie totale de la trame du signal audio spatial dans la sous-bande de fréquence donnée et du paramètre ER reçu pour la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée, et un moyen pour décoder ledit paramètre DOA en fonction dudit paramètre DEN dérivé pour la composante sonore directionnelle donnée dans la sous-bande de fréquence donnée dépassant ledit premier seuil.

9. Appareil selon la revendication 8, comprenant en outre un moyen pour calculer l'énergie totale de ladite trame du signal audio spatial pour la sous-bande de fréquence donnée sur la base de ladite trame du signal de mixage réducteur reconstruit.

10. Appareil selon la revendication 8 ou 9, dans lequel un paramètre ER indique un rapport entre le niveau d'énergie d'une composante sonore directionnelle donnée dans une sous-bande de fréquence donnée et l'énergie totale dans la sous-bande de fréquence donnée.

11. Appareil selon l'une des revendications 7 à 10, dans

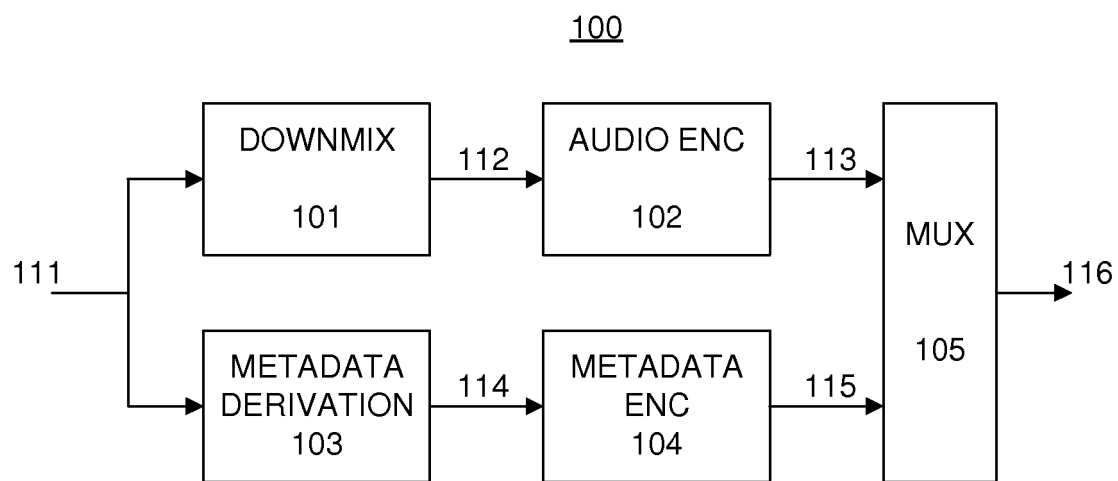


Figure 1

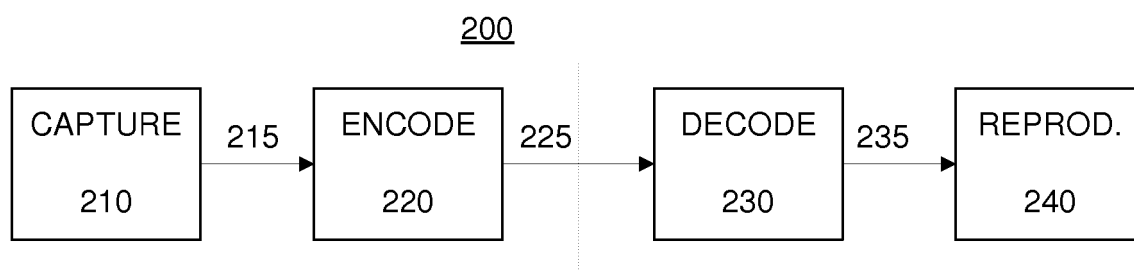


Figure 2

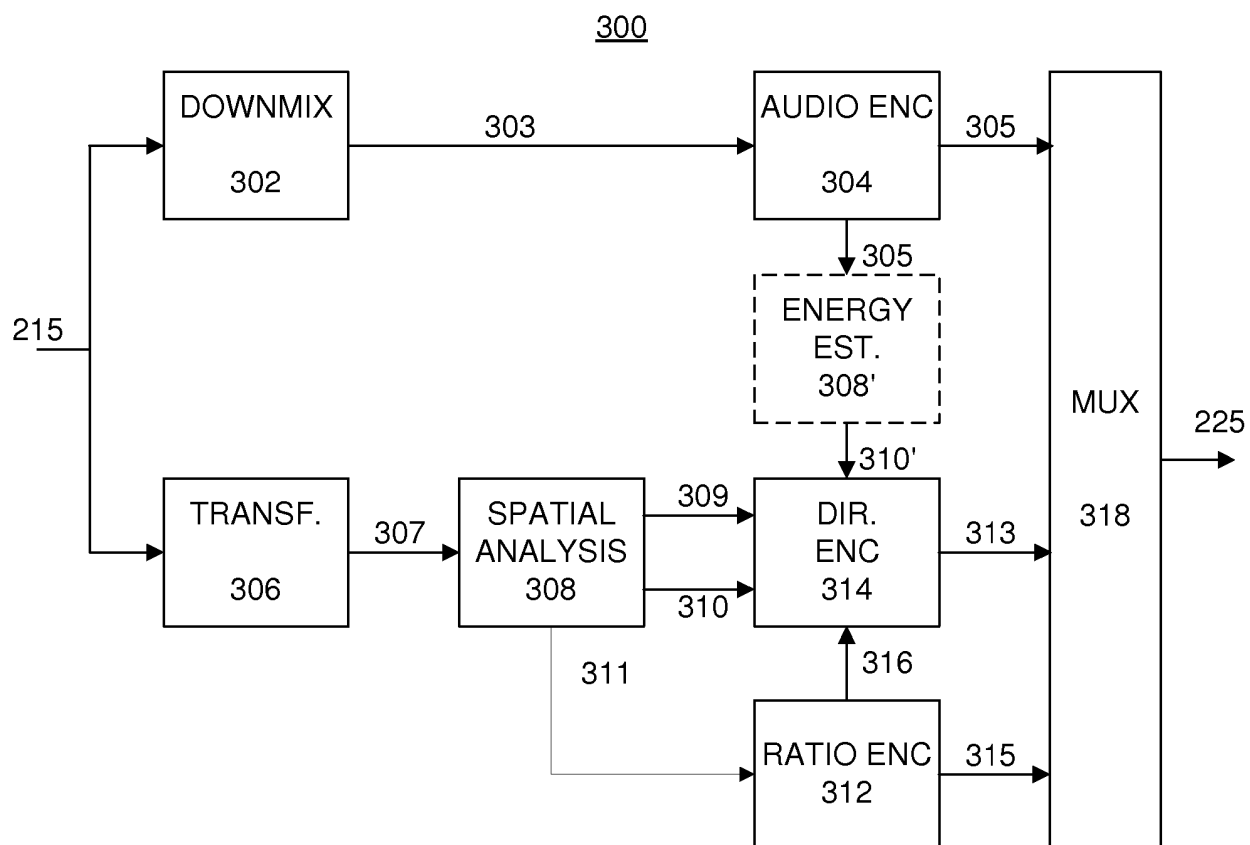


Figure 3

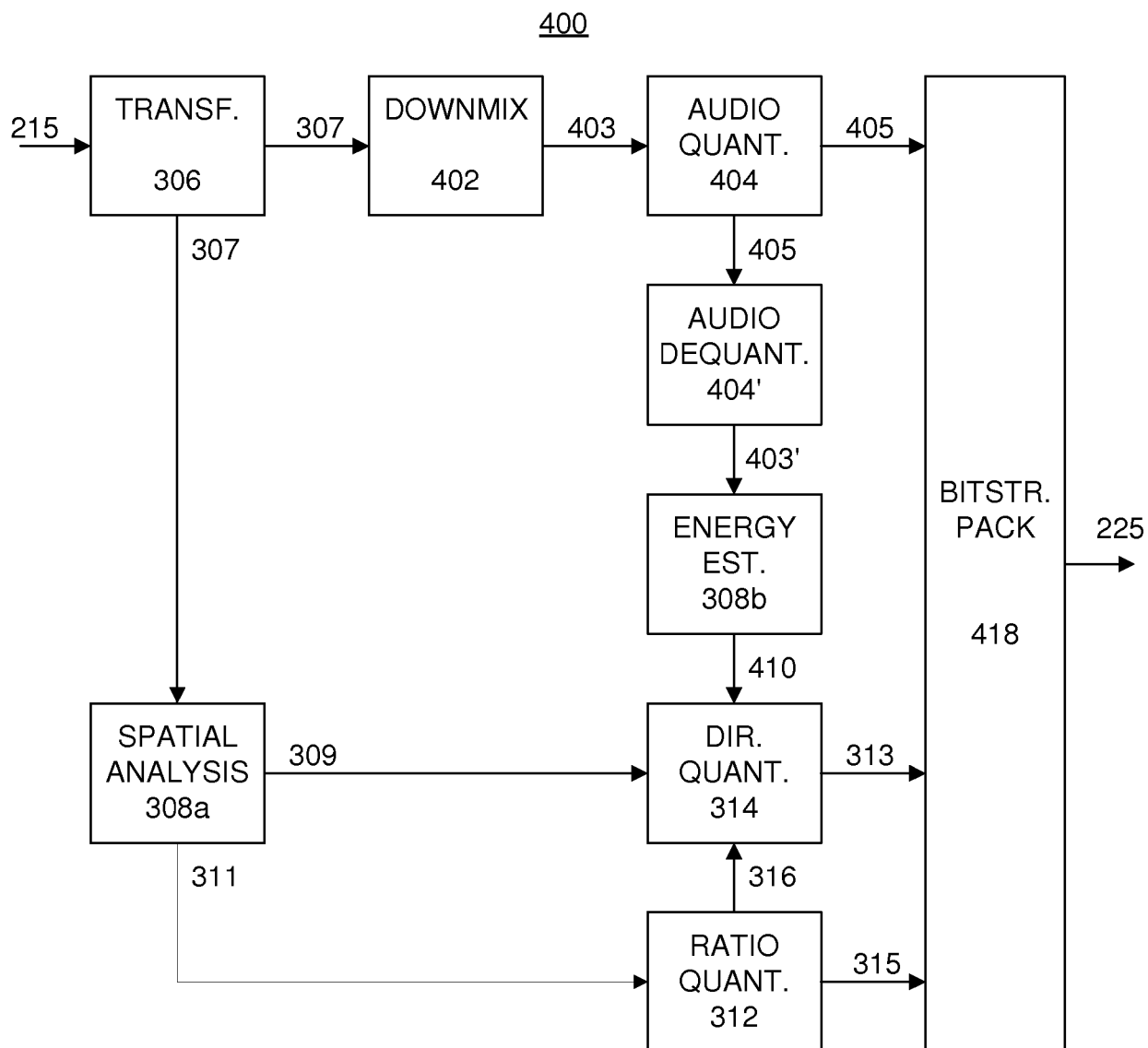


Figure 4

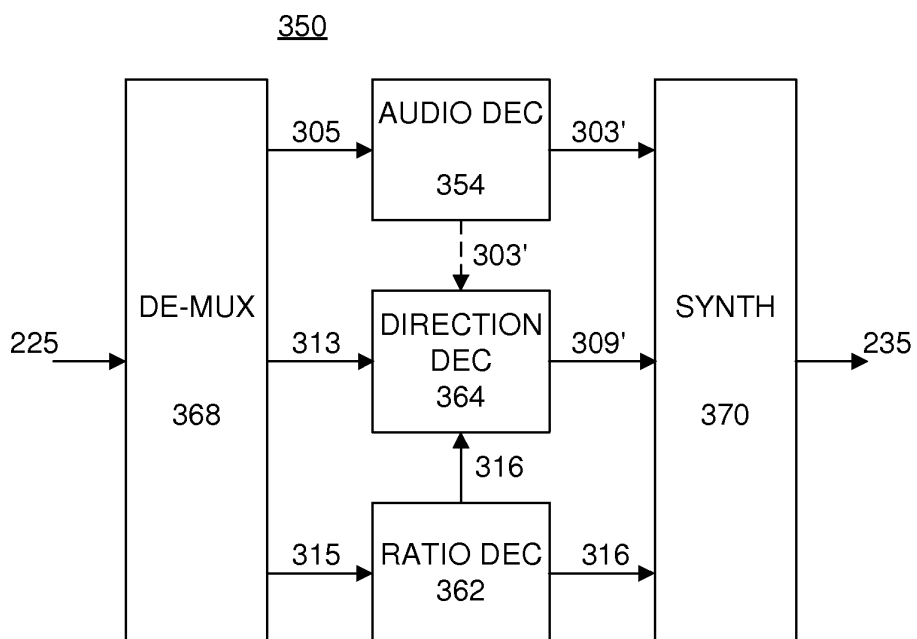


Figure 5

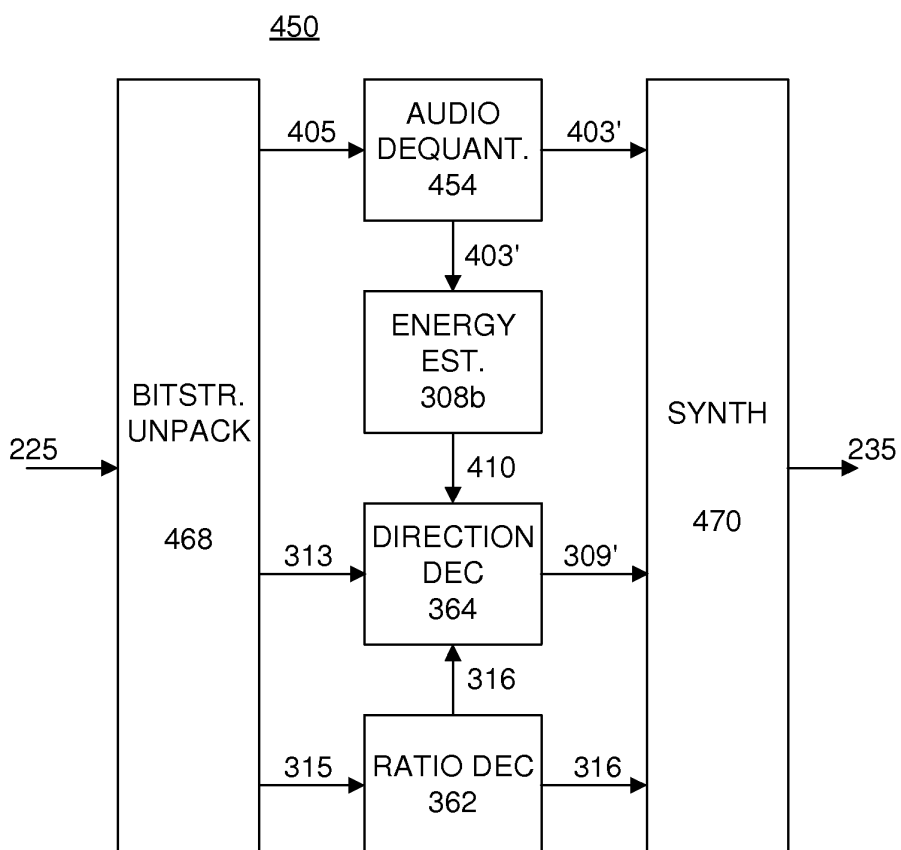


Figure 6

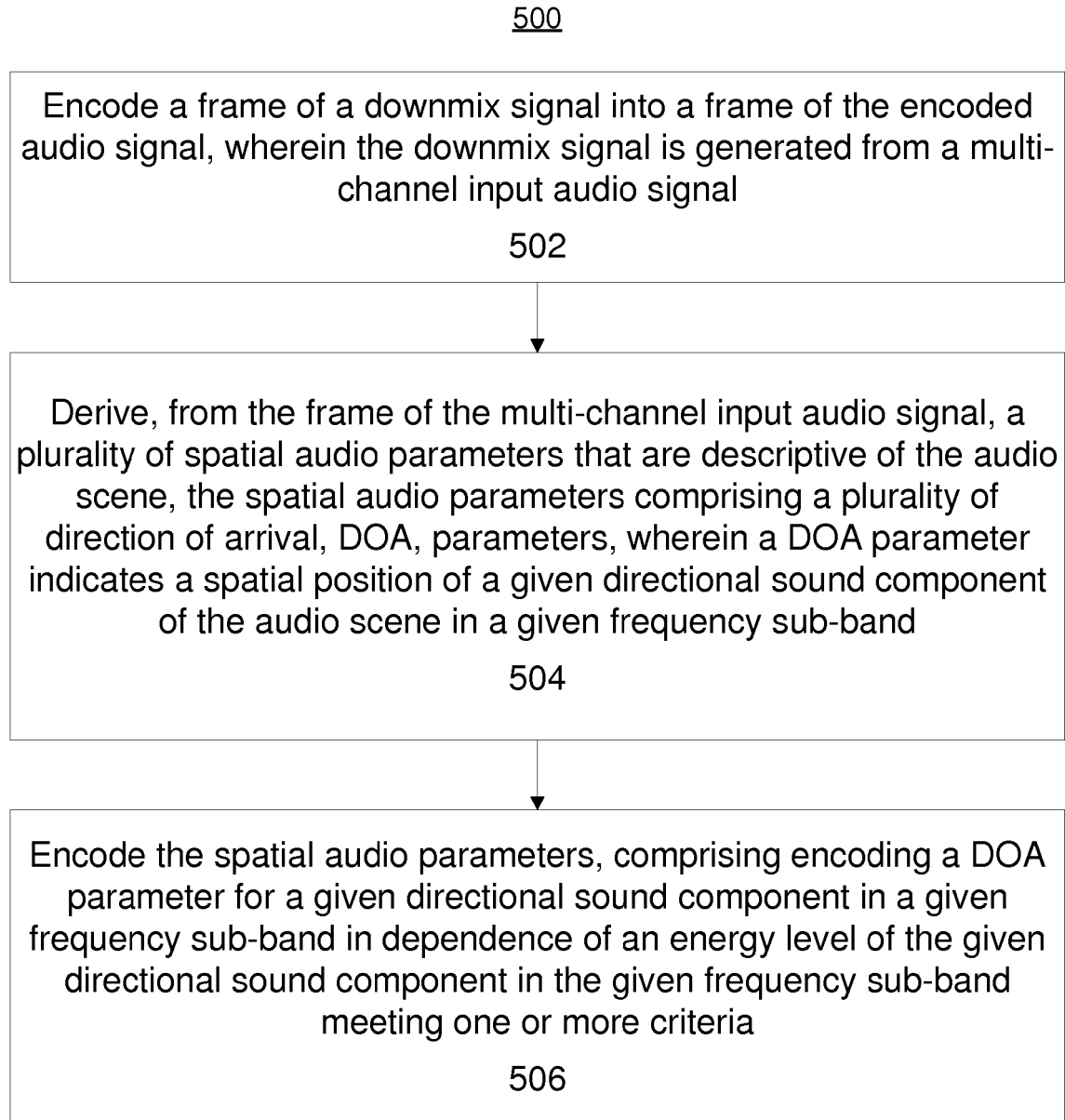


Figure 7A

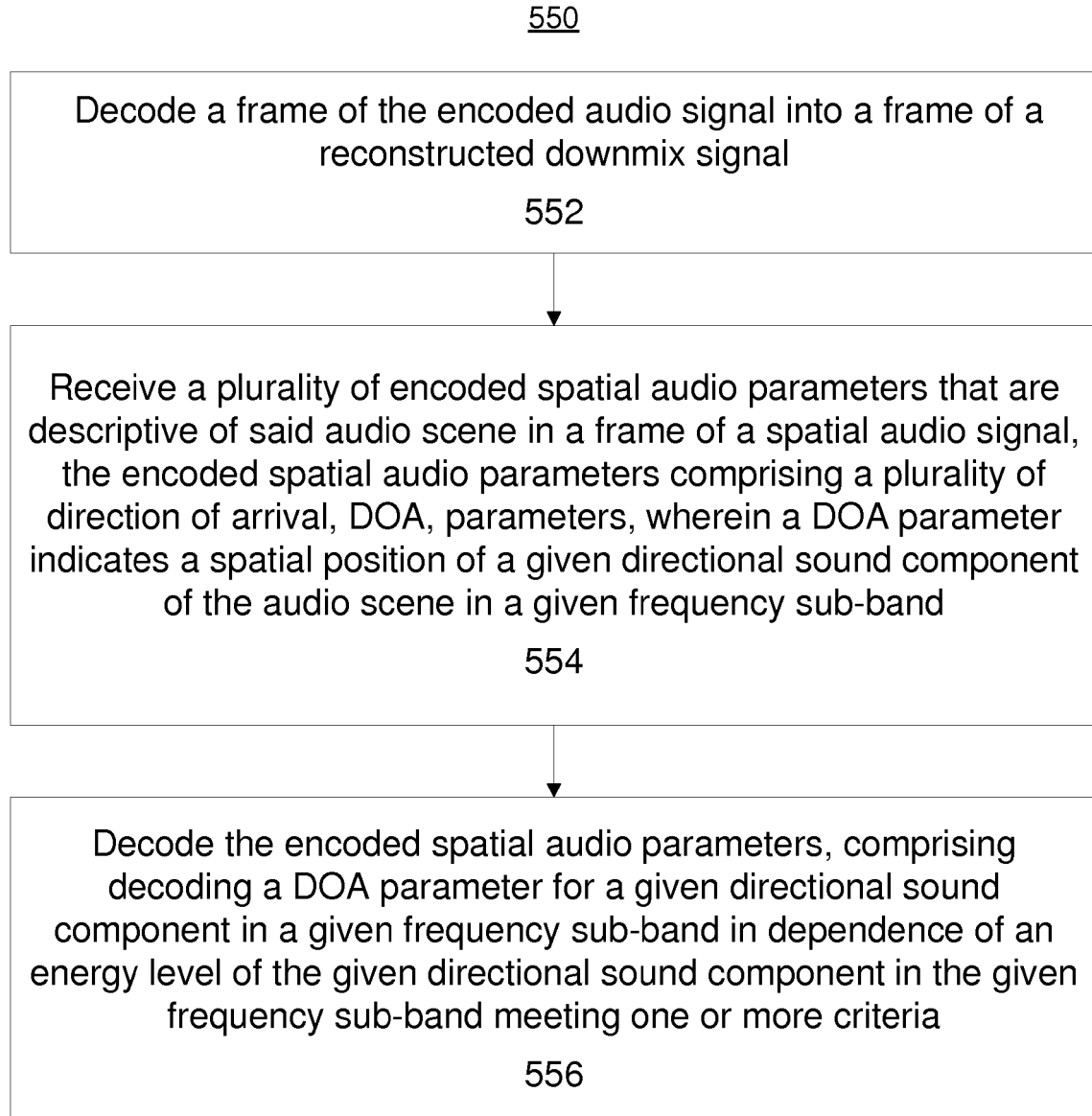


Figure 7B

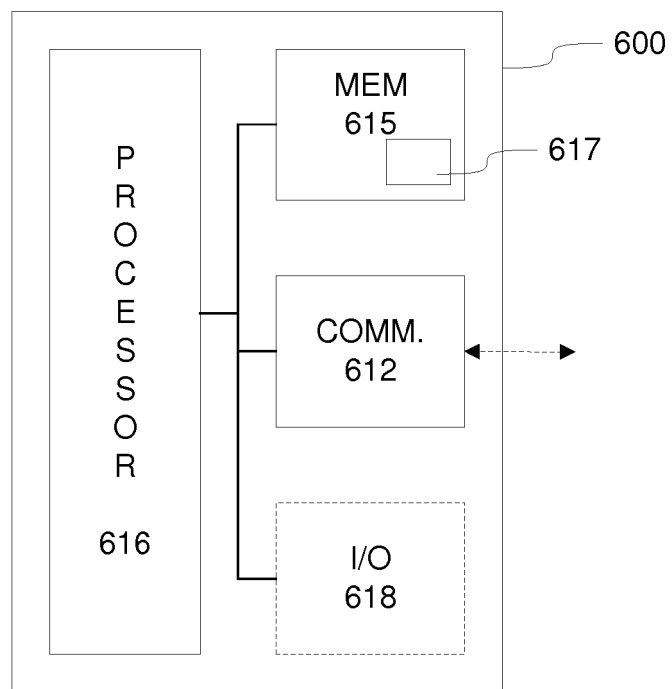


Figure 8

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

- US 20070269063 A [0005]
- US 9313599 B2 [0031]

Non-patent literature cited in the description

- **M. GOODWIN et al.** *Spatial Audio Coding Based On Universal Spatial Cues*, 22 November 2007 [0005]
- **PULKKI, VILLE.** Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 2007, vol. 55 (6), 503-516 [0105]
- Binaural reproduction for directional audio coding. **LAITINEN, MIKKO-VILLE ; PULKKI, VILLE.** *Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, 337-340 [0105]
- **VILKAMO, JUHA ; PULKKI, VILLE.** Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering. *Journal of the Audio Engineering Society*, 2013, vol. 61 (9), 637-646 [0105]
- **POLITIS, ARCHONTIS ; VILKAMO, JUHA ; PULKKI, VILLE.** Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 2015, vol. 9 (5), 852-866 [0105]