(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
08 October 2020 (08.10.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/201619 A1

(51) International Patent Classification:
G10L 19/22 (2013.01) *H04S 3/00* (2006.01)
G10L 19/16 (2013.01) *G10L 19/02* (2013.01)
G10L 19/008 (2013.01) *G10L 19/00* (2013.01)

(21) International Application Number:
PCT/FI2020/050171

(22) International Filing Date:
19 March 2020 (19.03.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1904819.8 05 April 2019 (05.04.2019) GB

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI]; Karakaari 7, 02610 Espoo (FI).

(72) Inventor: **LAAKSONEN, Lasse**; Näsilinnankatu 23 B 28, 33210 Tampere (FI).

(74) Agent: **NOKIA TECHNOLOGIES OY** et al.; Ari Aarnio, IPR Department, Karakaari 7, 02610 Espoo (FI).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

WO 2020/201619 A1

(54) Title: SPATIAL AUDIO REPRESENTATION AND ASSOCIATED RENDERING

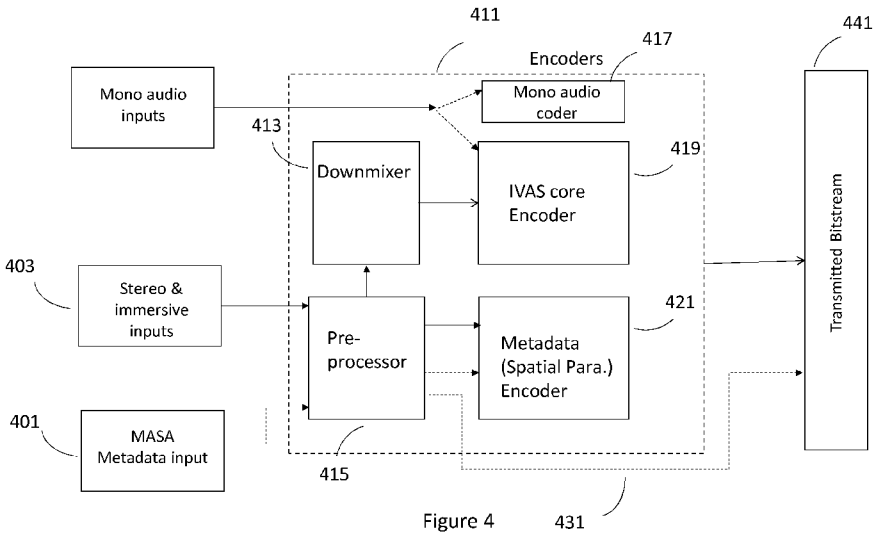

Figure 4

(57) Abstract: There is inter alia disclosed an apparatus for determining whether to quantise spatial audio parameters associated with a plurality of channel audio signals, wherein the spatial audio parameters are comprised in a metadata structure; quantising the spatial audio parameters in accordance with the determination; and adding to a data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised.

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

SPATIAL AUDIO REPRESENTATION AND ASSOCIATED RENDERING

Field

The present application relates to apparatus and methods for sound-field related audio representation and associated rendering, but not exclusively for audio representation for an audio encoder and decoder.

Background

Immersive audio codecs are being implemented supporting a multitude of operating points ranging from a low bit rate operation to transparency. An example of such a codec is the immersive voice and audio services (IVAS) codec which is being designed to be suitable for use over a communications network such as a 3GPP 4G/5G network. Such immersive services include uses for example in immersive voice and audio for applications such as virtual reality (VR), augmented reality (AR) and mixed reality (MR). This audio codec is expected to handle the encoding, decoding and rendering of speech, music and generic audio. It is furthermore expected to support channel-based audio and scene-based audio inputs including spatial information about the sound field and sound sources.

Furthermore, parametric spatial audio processing is a field of audio signal processing where the spatial aspect of the sound is described using a set of parameters. For example, in parametric spatial audio capture from microphone arrays, it is a typical and an effective choice to estimate from the microphone array signals a set of parameters such as directions of the sound in frequency bands, and the energy ratios between the directional and non-directional parts of the captured sound in frequency bands. These parameters are known to well describe the perceptual spatial properties of the captured sound at the position of the microphone array. These parameters can be utilized in synthesis of the spatial sound accordingly, for headphones binaurally, for loudspeakers, or to other formats, such

as Ambisonics. Furthermore, these parameters can be packaged as Metadata and transported as part of the encoded bitstream.

The IVAS codec is expected to operate with low latency to enable high-quality immersive conversational services and preferably in a tandem free mode of operation in order to avoid degradation from transcoding. Furthermore, the IVAS codec may be foreseen to be interoperable with the 3GPP Enhanced Voice Service in order to facilitate greater interoperability.

Summary

There is provided according to a first aspect an apparatus comprising means for: determining whether to quantise spatial audio parameters associated with a plurality of channel audio signals, wherein the spatial audio parameters are comprised in a metadata structure; quantising the spatial audio parameters in accordance with the determination; and adding to a data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised.

The indication may be a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag may indicate that the spatial audio parameters are quantised and wherein a further state of the binary flag may indicate that the spatial audio parameters have not been quantised, wherein the means for adding to the data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised may comprise: means for changing the state of the binary flag in accordance with the determination of whether to quantise the spatial audio parameters.

Alternatively the indication may be a variable comprised in the data filed of the metadata structure, wherein a value of the variable may indicate either that the spatial audio parameters have not been quantised and a further value may indicate that the spatial audio parameters have been quantised to one of a plurality of coding rates, wherein the means for adding to the data field metadata structure an

indication as to whether the spatial audio parameters have been quantised may comprise means for setting the value of the variable to indicate either: the spatial audio parameters have not been quantised; or a coding rate of the plurality of coding rates to which the spatial audio parameters have been quantised.

The spatial audio parameters may be been derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the apparatus may further comprise: means for adding to a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

The energy value of the sub band of the time frame associated with the plurality of channel audio signals may be at least one of: an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

The apparatus may further comprise: means for adding to a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band may be dependent the energy value of the sub band and an energy value of the further sub band respectively.

According to a second aspect there is an apparatus comprising means for: reading from a data field in a metadata structure comprising spatial audio parameters associated with a plurality of channel audio signals an indication as to whether the spatial audio parameters have been quantised; and dequantizing the spatial audio parameters in accordance with the indication.

The indication may be a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag may indicate that the spatial audio

parameters are quantised and wherein a further state of the binary flag may indicate that the spatial audio parameters have not been quantised.

Alternatively, the indication may be a variable comprised in the data filed of the metadata structure, wherein a value of the variable may indicate either that the spatial audio parameters have not been quantised and a further value may indicate that the spatial audio parameters have been quantised to one of a plurality of coding rates.

The spatial audio parameters may be derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the apparatus may further comprise: means for reading from a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

The energy value of the sub band of the time frame associated with the plurality of channel audio signals may be at least one of: an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

The apparatus may further comprise: means for reading from a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band may be dependent the energy value of the sub band and an energy value of the further sub band respectively.

According to a third aspect there is provided a method comprising means: determining whether to quantise spatial audio parameters associated with a plurality of channel audio signals, wherein the spatial audio parameters are comprised in a

metadata structure; quantising the spatial audio parameters in accordance with the determination; and adding to a data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised.

The indication may be a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag may indicate that the spatial audio parameters are quantised and wherein a further state of the binary flag may indicate that the spatial audio parameters have not been quantised, wherein adding to the data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised may comprise: changing the state of the binary flag in accordance with the determination of whether to quantise the spatial audio parameters.

Alternatively the indication may be a variable comprised in the data filed of the metadata structure, wherein a value of the variable indicates either that the spatial audio parameters have not been quantised and a further value indicates that the spatial audio parameters have been quantised to one of a plurality of coding rates, wherein the adding to the data field metadata structure an indication as to whether the spatial audio parameters have been quantised may comprise setting the value of the variable to indicate either: the spatial audio parameters have not been quantised; a coding rate of the plurality of coding rates to which the spatial audio parameters have been quantised.

The spatial audio parameters may have been derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the method may further comprise: adding to a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

The energy value of the sub band of the time frame associated with the plurality of channel audio signals may be at least one of: an energy value of all channel audio

signals in the sub band of the time frame associated with the plurality of channel audio signals; and an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

The method may further comprise: adding to a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band may be dependent the energy value of the sub band and an energy value of the further sub band respectively.

According to a fourth aspect there is a method comprising: reading from a data field in a metadata structure comprising spatial audio parameters associated with a plurality of channel audio signals an indication as to whether the spatial audio parameters have been quantised; and dequantizing the spatial audio parameters in accordance with the indication.

The indication may be a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag may indicate that the spatial audio parameters are quantised and wherein a further state of the binary flag may indicate that the spatial audio parameters have not been quantised.

Alternatively, the indication may be a variable comprised in the data filed of the metadata structure, wherein a value of the variable may indicate either that the spatial audio parameters have not been quantised and a further value may indicate that the spatial audio parameters have been quantised to one of a plurality of coding rates.

The spatial audio parameters may be derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the method may further comprise: means for reading from a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

The energy value of the sub band of the time frame associated with the plurality of channel audio signals may be at least one of: an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

The method may further comprise: reading from a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band may be dependent the energy value of the sub band and an energy value of the further sub band respectively.

A computer program product stored on a medium may cause an apparatus to perform a method as described herein.

An electronic device may comprise an apparatus as described herein.

A chipset may comprise an apparatus as described herein.

Summary of the Figures

For a better understanding of the present application, reference will now be made by way of example to the accompanying drawings in which:

Figure 1 shows schematically a system of apparatus suitable for implementing some embodiments;

Figure 2 shows schematically the metadata encoder according to some embodiments;

Figure 3 shows schematically a system of apparatus for an IVAS encoder architecture of Figure 1;

Figure 4 shows schematically an IVAS encoder architecture of Figure 1 including a MASA metadata input; and

Figure 5 shows an example device suitable for implementing the apparatus shown.

Embodiments of the Application

The following describes in further detail suitable apparatus and possible mechanisms for the provision of efficient representation of audio in immersive systems which implement a parametric spatial audio encoding mode of operation and support tandem free operation (or pass through mode) for the metadata of spatial audio parameters. The examples enable the immersive audio encoding system to operate in a tandem free operating mode in which the spatial audio parameters are not subjected to transcoding. It is generally understood that the process of transcoding, in the context of immersive audio coding, would typically involve the re-quantization of the spatial audio parameters. The process of transcoding would therefore not only result in a loss of audio quality but would also consume more computational processing power (or instruction cycles) than would be used if the spatial audio parameters were allowed to pass through an intermediate processing/network node in their original encoded state. This factor may be especially prevalent when the intermediate processing mode is an AR/VR server or conferencing bridge designed to handle multiple encoded spatial audio parameter streams.

Although the examples shown herein are described with respect to the IVAS codec, any other codec or coding can implement some embodiments as described herein. For example, a pass-through mode for spatial audio extension can be of great interest for interoperability and conformance reasons (particularly in managed services with quality of service (QoS) requirements).

The concept as discussed in further detail hereafter is to enable the MASA parameter set to pass through audio coding nodes without having to be either dequantized and then quantised/transformed into a different coding format in the case of a transcoding node or be re-quantised in the case the MASA parameter set is presented to another IVAS encoding node (generally a spatial audio encoding node). The Metadata-assisted spatial audio (MASA) parameter set can be understood (at least) as an ingest format intended for the 3GPP IVAS audio codec. Additional functionality can be added which would allow the spatial audio parameters to pass through audio encoding nodes in a tandem free manner which support immersive spatial audio coding or more specifically supports the 3GPP IVAS audio codec. To be clear the MASA parameter set comprises the metadata of spatial audio parameters or coefficients which define the sound field including associated sound sources within an audio scene. The MASA parameter set typically accompanies an audio stream comprising one or more audio channels such as a L-R stereo pair for example. The audio stream is thus a combination of the various sound sources within the audio scene. For example, in some embodiments the audio stream may comprise two channels which when decoded in conjunction with the MASA parameter set produces an immersive audio scene to the end user. It is to be noted that the audio channels accompanying the MASA parameter set may be formed by either downmixing the sound source channels into fewer audio channels or by simply selecting a number of the sound source channels without the active process of downmixing. An example of the latter technique may comprise selecting the L-R channels from a four microphone capture device such as that found in a mobile device.

Before further discussing the embodiments we initially discuss the systems for obtaining and rendering spatial audio signals which may be used in some embodiments.

With respect to Figure 1 is shown an example apparatus and system for implementing the obtaining and encoding an audio signal (in the form of audio capture in this example) and rendering (the encoded audio signals).

The system 100 is shown with an 'analysis' part 121 and a 'synthesis' part 131. The 'analysis' part 121 is the part from receiving the multi-channel signals up to an encoding of the metadata of spatial audio parameters and downmix signal and the 'synthesis' part 131 is the part from a decoding of the encoded metadata of spatial audio parameters and downmix signal to the presentation of the re-generated signal (for example in multi-channel loudspeaker form).

The input to the system 100 and the 'analysis' part 121 is the multi-channel signals 102. In the following examples a microphone channel signal input is described, however any suitable input (or synthetic multi-channel) format may be implemented in other embodiments. For example, in some embodiments the spatial analyser and the spatial analysis may be implemented external to the encoder. For example, in some embodiments the metadata of spatial audio parameters associated with the audio signals may be a provided to an encoder as a separate bit-stream.

The multi-channel signals are passed to a downmixer 103 and to an analysis processor 105.

In some embodiments the downmixer 103 is configured to receive the multi-channel signals and generate a suitable transport signal comprising a determined number of channels and output the downmix signals 104. For example, the downmixer 103 may be configured to generate a 2 audio channel downmix of the multi-channel signals. The determined number of channels may be any suitable number of channels. The downmixer 103 in some embodiments is configured to otherwise select or combine, for example, by beamforming techniques the input audio signals to the determined number of channels and output these as transport signals.

In some embodiments the downmixer 103 is optional and the multi-channel signals are passed unprocessed to an encoder 107 in the same manner as the transport signal are in this example.

In some embodiments the analysis processor 105 is also configured to receive the multi-channel signals and analyse the signals to produce the metadata of spatial audio parameters 106 associated with the multi-channel signals and thus associated with the downmix signals 104. The analysis processor 105 may be configured to generate the metadata which may comprise for example, for each time-frequency analysis interval, a direction parameter 108 and an energy ratio parameter 110 (and/or diffuseness parameter) and a coherence parameter 112. The direction, energy ratio and coherence parameters may in some embodiments be considered to be spatial audio parameters. In other words, the metadata of spatial audio parameters comprise parameters which aim to characterize the sound-field represented by the multi-channel signals (or two or more playback audio signals in general).

In some embodiments the spatial audio parameters generated may differ from frequency band to frequency band. Thus, for example in band X all of the parameters are generated and transmitted, whereas in band Y only one of the parameters is generated and transmitted, and furthermore in band Z no parameters are generated or transmitted. A practical example of this may be that for some frequency bands such as the highest band some of the parameters are not required for perceptual reasons. The downmix signals 104 and the metadata of spatial audio parameters 106 may be passed to an encoder 107.

In some embodiments, the spatial audio parameters may be grouped or separated into directional and non-directional (such as, e.g., diffuse) parameters.

The encoder 107 may comprise an audio encoder core 109 which is configured to receive the transport (for example downmix) signals 104 and generate a suitable

encoding of these audio signals. The encoder 107 can in some embodiments be a computer (running suitable software stored on memory and on at least one processor), or alternatively a specific device utilizing, for example, FPGAs or ASICs. The encoding may be implemented using any suitable scheme. The encoder 107 may furthermore comprise a metadata encoder/quantizer 111 which is configured to receive the metadata of spatial audio parameters and output an encoded or compressed form of the information. In some embodiments the encoder 107 may further interleave, multiplex to a single data stream or embed the metadata of spatial audio parameters within encoded downmix signals before transmission or storage shown in Figure 1 by the dashed line. The multiplexing may be implemented using any suitable scheme.

In the decoder side, the received or retrieved data (stream) may be received by a decoder/demultiplexer 133. The decoder/demultiplexer 133 may demultiplex the encoded streams and pass the audio encoded stream to a transport extractor 135 which is configured to decode the audio signals to obtain the transport signals. Similarly, the decoder/demultiplexer 133 may comprise a metadata extractor 137 which is configured to receive the encoded metadata and generate metadata. The decoder/demultiplexer 133 can in some embodiments be a computer (running suitable software stored on memory and on at least one processor), or alternatively a specific device utilizing, for example, FPGAs or ASICs.

The decoded metadata and transport audio signals may be passed to a synthesis processor 139.

The system 100 ‘synthesis’ part 131 further shows a synthesis processor 139 configured to receive the downmix signals and the metadata of spatial audio parameters and re-creates in any suitable format a synthesized spatial audio in the form of multi-channel signals 110 (these may be multichannel loudspeaker format or in some embodiments any suitable output format such as binaural signals for

headphone listening or Ambisonics signals, depending on the use case) based on the downmix signals and the metadata of spatial audio parameters.

Therefore, in summary first the system (analysis part) is configured to receive multi-channel audio signals.

Then the system (analysis part) is configured to generate a suitable transport audio signal (for example by selecting or downmixing some of the audio signal channels).

The system is then configured to encode for storage/transmission the downmix signal and the metadata of spatial audio parameters.

After this the system may store/transmit the encoded downmix signal and metadata of spatial audio parameters.

The system may retrieve/receive the encoded downmix signals and metadata of spatial audio parameters. The system may then be configured to extract the downmix signal and metadata of spatial audio parameters from encoded downmix signal and metadata of spatial audio parameters, for example demultiplex and decode the encoded downmix signal and metadata of spatial audio parameters.

The system (synthesis part) is configured to synthesize an output multi-channel audio signal based on extracted downmix of multi-channel audio signals and metadata of spatial audio parameters.

With respect to Figure 2 an example analysis processor 105 (as shown in Figure 1) according to some embodiments is shown.

The analysis processor 105 in some embodiments comprises a time-frequency domain transformer 201.

In some embodiments the time-frequency domain transformer 201 is configured to receive the multi-channel signals 102 and apply a suitable time to frequency domain transform such as a Short Time Fourier Transform (STFT) in order to convert the input time domain signals into a suitable time-frequency signals. These time-frequency signals may be passed to a spatial analyser 203 and to the direction index generator 205.

Thus, for example the time-frequency signals 202 may be represented in the time-frequency domain representation by

$$s_i(b, n),$$

where b is the frequency bin index and n is the time-frequency block (frame) index and i is the channel index. In another expression, n can be considered as a time index with a lower sampling rate than that of the original time-domain signals. These frequency bins can be grouped into subbands that group one or more of the bins into a subband of a band index k = 0,..., K-1. Each subband k has a lowest bin $b_{k,low}$ and a highest bin $b_{k,high}$, and the subband contains all bins from $b_{k,low}$ to $b_{k,high}$. The widths of the subbands can approximate any suitable distribution. For example the Equivalent rectangular bandwidth (ERB) scale or the Bark scale.

A time frequency (TF) tile (or block) is thus a specific sub band within a subframe of the frame.

In some embodiments the analysis processor 105 comprises a spatial analyser 203. The spatial analyser 203 may be configured to receive the time-frequency signals 202 and based on these signals estimate direction parameters 108. The direction parameters may be determined based on any audio based 'direction' determination.

For example, in some embodiments the spatial analyser 203 is configured to estimate at least one direction for the TF tile with two or more signal inputs. This

represents the simplest configuration to estimate a 'direction', more complex processing may be performed with even more signals.

The spatial analyser 203 may thus be configured to provide at least one azimuth and elevation for each frequency band and temporal time-frequency block within a frame of an audio signal, denoted as azimuth $\varphi(k,n)$ and elevation $\theta(k,n)$. The direction parameters 108 may also be passed to a direction index generator 205.

The spatial analyser 203 may also be configured to determine an energy ratio parameter 110. The energy ratio may be considered to be a determination of the energy of the audio signal which can be considered to arrive from a direction. The direct-to-total energy ratio r(k,n) can be estimated, e.g., using a stability measure of the directional estimate, or using any correlation measure, or any other suitable method to obtain a ratio parameter. The energy ratio may be passed to an energy ratio analyser 221 and an energy ratio encoder 223.

Therefore, in summary the analysis processor is configured to receive time domain multichannel or other format such as microphone or ambisonics audio signals.

Following this the analysis processor may apply a time domain to frequency domain transform (e.g. STFT) to generate suitable time-frequency domain signals for analysis and then apply direction analysis to determine direction and energy ratio parameters.

The analysis processor may then be configured to output the determined parameters.

Although directions and ratios are here expressed for each time index *n,* in some embodiments the parameters may be combined over several time indices. Same applies for the frequency axis, as has been expressed, the direction of several frequency bins *b* could be expressed by one direction parameter in band *k* consisting

of several frequency bins *b*. The same applies for all of the discussed spatial parameters herein.

The energy ratio analyser 221 may be configured to receive the energy ratios and from the analysis generate a quantization resolution for the direction parameters (in other words a quantization resolution for elevation and azimuth values) for all of the time-frequency (TF) blocks in the frame.

The direction index generator 205 may be configured to receive the direction parameters (such as the azimuth $\varphi(k,n)$ and elevation $\theta(k,n)$) 108 and the quantization bit allocation and from this generate a quantized output in the form of indexes to various tables and codebooks which represent the quantized direction parameters.

Figure 2 also depicts a combiner 207 which can be configured to combine the direction indices from the direction index generator 205 and energy ratios from the energy ratio encoder 223. The output (the unquantized spatial audio parameters) of the combiner 207 may be passed to the Metadata Encoder/quantizer 111 for quantization of the aforementioned parameters.

With respect to Figure 3 is shown a high-level view of an example IVAS encoder including the various inputs which may, as non-exclusive examples, be expected for the codec. The underlying idea is a spatial or immersive input is handled by the IVAS core tools for audio waveform encoding complemented by a metadata encoder.

The system as shown in Figure 3 can comprise input format generators 303. The input format generators 303 may be considered in some examples to be the same as the downmixer 103 and the analysis processor 105 from Figure 1. The input format generators 303 may be configured to generate suitable audio signals and

metadata for capturing the audio and spatial audio qualities of input signals such as those from a channel-based music file (such as a 5.1 music mix file).

The input format generators 303 can comprise a multichannel or spatial format generator 307.

The multichannel or spatial format generator 307 in some embodiments comprises a metadata-assisted spatial audio (MASA) generator 309. The metadata-assisted spatial audio generator 309 is configured to generate audio signals (such as the downmix signals in the form of a stereo-channel audio signal) and metadata of spatial audio parameters associated with the audio signals.

The multichannel or spatial format generator 307 in some embodiments comprises a multichannel format generator 311 configured to generate suitable multichannel audio signals (for example stereo channel format audio signals and/or 5.1 channel format audio signals).

The multichannel or spatial format generator 307 in some embodiments comprises an ambisonics generator 313 configured to generate a suitable ambisonics format audio signal (which may comprise first order ambisonics and/or higher order ambisonics).

The multichannel or spatial format generator 307 in some embodiments can comprise an independent mono streams with metadata generator 315 configured to generate mono audio signals and metadata of spatial audio parameters.

In some embodiments the apparatus comprises encoders 321. The encoder(s) is configured to receive the output of the input format generators 303 and encode these into a suitable format for storage and/or transmission. The encoders may be considered to be the same as the encoder 107.

In some embodiments the encoders 321 may comprise IVAS core encoder 325. The IVAS core encoder 325 may be configured to receive the audio signals generated by the input format generators 303 and encode these according to the IVAS standard.

In some embodiments the encoders comprise a metadata encoder 327. The metadata encoder is configured to receive the metadata of spatial audio parameters and encode it or compress it in any suitable manner.

The encoders 321 in some embodiments can be configured to combine or multiplex the datastreams generated by the encoders prior to being transmitted and/or stored.

The system furthermore comprises a transmitter configured to transmit or store the bitstream 331.

With respect to Figure 4 a further view of the encoding system according to Figure 3 is shown.

In this example there comprises a mono audio input 402 and stereo (and immersive audio) input 403 Additionally, there is shown a MASA Metadata Input 401 which represents spatial parameters of the audio scene which are in the form of metadata, in other words this input is the metadata of spatial audio parameters which accompanies either the mono audio input 402 or the stereo audio input 403.

The stereo & immersive inputs 403 and the MASA metadata input 401 are both passed to the encoder 411 via a pre-processor 415. The pre-processor 415 may be configured to receive the stereo and immersive inputs and pre-process the signal before being passed to the downmixer 413 and to the IVAS core encoder 419. The metadata output of the pre-processor 415 can be passed to the metadata encoder 421. The pre-processor 415 can be configured to process the MASA metadata input

401 and decide whether to process the spatial audio parameters either via the Metadata encoder 421, or to "pass through" in which the audio spatial parameters of the Metadata input are not subjected to metadata encoding but instead are passed directly to the transmitted bitstream 431. This "path" through the encoder 417 is depicted as 431 in Figure 4.

The mono audio input 402 may be passed to a mono audio coder 417 for encoding such as the 3GPP EVS codec for instance. Alternatively, the mono audio input 402 may be encoded as a single channel by the IVAS core encoder 419.

In some embodiments the encoder 411 may comprise the IVAS core encoder 419. The IVAS core encoder 419 may be configured to receive the downmixed audio signals generated by the downmixer 413 and encode these according to a suitable encoding algorithm.

In some embodiments the encoder 411 comprise the metadata encoder 421. The metadata encoder 421 may be configured to receive spatial metadata from the downmixer 413 and/or pre-processor 415 and encode it or compress it in any suitable manner.

According to the 3GPP technical document Tdoc S4-180462 "On spatial metadata for IVAS spatial audio input format", Nokia Corporation, SA4#98, Kista, Sweden 9-13 April 2018, the metadata assisted spatial audio parameters may have the high-level data structure according to Table 1 below.

Field	Bits	Description
Version	x	Spatial metadata version number.
Number of directions	x	Number of sound source directions analysed.
Channel configuration	x	Channel configuration index.
Reserved	x	(For example, fill byte alignment if needed.)
TF tile Parameters		Given by Table 2 for each TF tile

Table 1

From Table 1 it can be seen that the number of bits encompassed by the MASA metadata structure may be dependent at least in part on the TF (time-frequency) tile resolution (i.e., the number of TF subframes or tiles). For example, a 20ms audio frame may be divided into 4 time-domain subframes of 5ms a piece, and each time-domain subframe may have up to 24 frequency subbands divided in the frequency domain according to a Bark scale or any other suitable division. In this particular example the audio frame may be divided into 96 TF subframes/tiles, in other words 4 time-domain subframes with 24 frequency subbands. Therefore, the number of bits required to represent the spatial audio parameters for an audio frame can be dependent on the TF tile resolution. For example, Table 2 below shows a distribution of bits for each TF tile (with one sound source direction per TF tile) according to the above cited example with a TF tile resolution of 96.

Field	Bits	Description
Direction index	16	Direction of arrival of the sound at a time-frequency parameter interval. Spherical representation at about 1-degree accuracy. Range of values: “covers all directions at about 1° accuracy”
Directional energy ratio	8	Energy ratio for the direction index (i.e., time-frequency subframe). Calculated as energy in direction / total energy. The remainder of the energy is non-directional. Range of values: [0.0, 1.0]
Spread coherence	8	Spread of energy for the direction index (i.e., time-frequency subframe). Defines the direction to be reproduced as a point source or coherently around the direction. Range of values: [0.0, 1.0]
Surround coherence	8	Coherence of the non-directional sound over the surrounding directions. Range of values: [0.0, 1.0] (Parameter is independent of number of directions provided.)
Distance	8	Distance of the sound originating from the distance index (i.e., time-frequency subframes) in meters on a logarithmic scale. Range of values: for example, 0 to 100 m. (Feature intended mainly for future extensions, e.g., 6DoF audio.)

Table 2

It is to be appreciated that in some deployments of an immersive spatial audio system, such as that depicted by Figure 1, there may be a mode of operation in which there is no quantization of the spatial audio parameters contained within the MASA metadata structure. In other words, the spatial audio parameters 106 in Figure 1 may not be subjected to a quantization process by the Metadata encoder/quantizer 111.

With reference to Figure 1, the encoder 107 may be arranged to not quantize the spatial audio parameters 106 with the Metadata encoder/quantizer 111. In other words, it may be a switchable function of the Metadata encoder/quantizer 111 as to whether to quantize the spatial audio parameters 106. The decision to quantize the metadata of spatial audio parameters 106 may be driven by either a user input or by network dependent feedback mechanism. For instance, the decision to quantize the metadata of spatial audio parameters 106 may be determined by the available bandwidth of the communication travel going forward from the encoding system 100.

The encoder of Figure 4 may be viewed from the perspective of being part of a network node in which the MASA metadata input 401 presented to the pre-processor 415 may be in the form of having the spatial audio parameters in either a quantized state or in a unquantized raw state. For the case that the spatial audio parameters are in an unquantized raw state then the pre-processor 415 can be arranged to direct the spatial audio parameters to the Metadata encoder 421 in order to be quantised. Alternatively, in some particular operating circumstances the pre-processor 415 may allow the unquantized raw spatial audio parameters to bypass quantization and to pass through the encoder along the pass- through path 431. For the case that the spatial audio parameters are already in a quantised state the pre-processor 415 can be arranged to by-pass the quantization process in the Metadata encoder 421 to avoid re-quantization of the spatial audio parameters and subsequent loss of spatial reproduction accuracy and quality. This particular circumstance may arise when Figure 4 forms part of a spatial conferencing system.

In order to enable the above operating scenarios of the IVAS coding system the MASA metadata structure may be extended to include an additional data field in order to indicate whether spatial parameters such as those listed in Table 2 have been quantised according to a Metadata encoder/quantizer such as that depicted by 111 in Figure 1, or whether the spatial parameters are in an unquantized state or "raw" state. In other words, spatial parameters which are in an unquantized state have not been quantised by a Metadata encoder/quantizer such as 111. In embodiments the additional data field may be termed a "spatial parameter status" field.

In some embodiments the "spatial parameter status" field may be a simple binary flag indicating whether the spatial audio parameters contained in the MASA metadata structure are in a quantised state or a non-quantised state.

In further embodiments the "spatial parameter status" field may comprise several bits in order to not only indicate whether the spatial audio parameters have been quantised but to also indicate the bit rate at which the spatial audio parameters have been quantised. For example, the "spatial parameter status" field may have the following format.

Field	Bits	Description
Spatial Parameter status	x	Defines the quantization applied to the metadata. Example values: "0" – "raw" IVAS spatial parameters with no encoder quantization "1" – IVAS spatial parameters quantized at mode / bit rate X "2" – IVAS spatial parameters quantized at mode / bit rate Y … "z" – IVAS spatial parameters quantized at mode / bit rate Z

Table 3

From the example of table 3 it may be seen that the value of the "Spatial Parameter Status" indicates the coding or quantization rate of the spatial audio parameters contained in the MASA metadata structure. Furthermore, it can be seen that one of the values of the "Spatial Parameter Status" is reserved to indicate that there has been no quantization performed on the IVAS spatial audio parameters, in other words the parameters are in their "raw" state.

Table 4 below depicts how the above "Spatial parameter status" field may be incorporated into the MASA metadata structure.

Field	Bits	Description
Version	x	Spatial metadata version number.
Spatial Parameter Status	x	Quantization rate applied to the TF tile Parameters
Number of directions	x	Number of sound source directions analysed.
Channel configuration	x	Channel configuration index.
Reserved	x	(For example, fill byte alignment if needed.)
TF tile Parameters		Given by Table 2 for each TF tile

Table 4

In accordance with embodiments a Metadata Encoder/quantizer such as those depicted as 111 in Figure 1, 327 in Figure 3 or 421 in Figure 4 may be arranged to set the above Spatial Parameter Status field of the MASA metadata structure in accordance with the above examples. For instance, if the Spatial Parameter Status data field comprises a binary flag then the Metadata Encoder/quantizer can be arranged to set the state of the flag in accordance with whether the quantisation has been performed on the metadata of spatial audio parameters. Alternatively, if the Spatial Parameter Status data field comprises a numerical value then the Metadata Encoder/quantizer can be arranged to set to a specific value which reflects whether the spatial audio parameters have been quantised, and if they have been quantised the actual value may reflect the coding rate to which the parameters have been quantised. Accordingly, with this particular use of the Spatial Parameter Status data

field one of the values will be reserved to indicate that the spatial audio parameters are not quantised.

Furthermore, encoding systems such as those depicted in Figures 1, 3 and 4 can be arranged to accept a MASA metadata input in which the data fields of the MASA metadata structure can be parsed in order to at least determine the state of the spatial audio parameters. More specifically the encoding systems can be arranged to parse the Spatial Parameter Status field in the MASA metadata structure in order to determine whether the spatial audio parameters have been quantised. The Spatial Parameter Status field may therefore be used by the encoder to determine how the spatial audio parameters are handled. For instance, if the Spatial Parameter Status field indicates that the spatial audio parameters are in a quantised form then the encoder may enable a “pass through” mode in which the audio spatial parameters of the MASA metadata structure are not subjected to metadata encoding but instead are passed directly along the connection 431 to a transmitted bitstream 441. This functionality may be enabled in order to avoid re-quantisation of the spatial audio parameters. However, if the Spatial Parameter Status field indicates that the spatial audio parameters are in the “raw” unquantized form, then the encoder may subject spatial audio parameters to metadata encoding such as that depicted by 421. In such a scenario the Spatial parameter metadata encoder 421 may not only be configured to quantize the spatial audio parameters but also to change the Spatial Parameter Status field to indicate that the coding rate at which the spatial audio parameters have been quantized.

Furthermore, an IVAS decoding instance such as 131 in Figure 1 can be arranged to read the Spatial Parameter Status field in the in the MASA metadata structure in order to determine whether the spatial audio parameters have been quantised, and if quantised what quantisation coding rate was used. For instance, the parsing of the Spatial Parameter filed may take place in the Metadata extractor 137.

In embodiments, the IVAS decoder 131 can dequantize the spatial audio parameters via the Metadata extractor 137. In some instances, the decoder may write out the dequantized spatial audio parameters in the MASA metadata structure format. In which case, the Metadata extractor 137 may set the Spatial Parameter Status field to show that the spatial audio parameters are in an unquantized “raw” state.

In some embodiments the analysis processor 105 may be configured to generate metadata that describes the energy for each time-frequency analysis interval. This energy parameter may be a signal energy (e.g., the sum of the energies of the individual channels) or there may be an energy parameter for each individual audio signal or audio channel. In further embodiments the analysis processor 105 may be configured to furthermore determine an importance ranking order of the time-frequency analysis intervals (in each frame, i.e., metadata block update interval). Such importance ranking order may also take the form of a priority index in some embodiments. The importance ranking or priority index information generation may be based at least on the signal energy in each time-frequency analysis interval. However, psychoacoustics may further be considered. The intention to include this type of parameter (signal energy, channel energy, importance ranking, priority index, or any other suitable parameter) may be to allow analysis of the spatial audio in the analysis processor 105 using a different filterbank (and particularly a different filterbank time-frequency resolution) than used in an audio encoder that ingests the audio format while retaining the capability of the audio encoder to quantize or otherwise reduce the amount of metadata for transmission based on energy characteristics observed for each of the time-frequency analysis intervals. In other words, the audio waveform signal encoding and the metadata quantization can be maintained separate while allowing for metadata quantization based on selected audio waveform signal properties even though a different filterbank (TF resolution) would be used.

This metadata which describes the energy for each time-frequency analysis interval may be incorporated into the existing metadata assisted spatial audio parameters data structure as part of the TF tile Parameters in Table 1. For example, Table 2 may be expanded to include at least one of the following additional metadata fields shown in Table 5 below.

Field	Bits	Description
Energy	8	Energy of the signal (corresponding to the time-frequency subframe).
Channel energy	8	Energy of the channel (corresponding to the time-frequency subframe).
Subframe ranking index	8	Importance ranking order of time-frequency subframes.
Subframe priority index	8	Priority value, where a smaller value indicates a higher priority.

Table 5

In embodiments the metadata fields may be read at a function in a decoder 131 such as a metadata extractor 137.

With respect to Figure 5 an example electronic device which may be used as the analysis or synthesis device is shown. The device may be any suitable electronics device or apparatus. For example, in some embodiments the device 1400 is a mobile device, user equipment, tablet computer, computer, audio playback apparatus, etc.

In some embodiments the device 1400 comprises at least one processor or central processing unit 1407. The processor 1407 can be configured to execute various program codes such as the methods such as described herein.

In some embodiments the device 1400 comprises a memory 1411. In some embodiments the at least one processor 1407 is coupled to the memory 1411. The memory 1411 can be any suitable storage means. In some embodiments the memory 1411 comprises a program code section for storing program codes implementable upon the processor 1407. Furthermore, in some embodiments the memory 1411 can further comprise a stored data section for storing data, for

example data that has been processed or to be processed in accordance with the embodiments as described herein. The implemented program code stored within the program code section and the data stored within the stored data section can be retrieved by the processor 1407 whenever needed via the memory-processor coupling.

In some embodiments the device 1400 comprises a user interface 1405. The user interface 1405 can be coupled in some embodiments to the processor 1407. In some embodiments the processor 1407 can control the operation of the user interface 1405 and receive inputs from the user interface 1405. In some embodiments the user interface 1405 can enable a user to input commands to the device 1400, for example via a keypad. In some embodiments the user interface 1405 can enable the user to obtain information from the device 1400. For example the user interface 1405 may comprise a display configured to display information from the device 1400 to the user. The user interface 1405 can in some embodiments comprise a touch screen or touch interface capable of both enabling information to be entered to the device 1400 and further displaying information to the user of the device 1400. In some embodiments the user interface 1405 may be the user interface for communicating with the position determiner as described herein.

In some embodiments the device 1400 comprises an input/output port 1409. The input/output port 1409 in some embodiments comprises a transceiver. The transceiver in such embodiments can be coupled to the processor 1407 and configured to enable a communication with other apparatus or electronic devices, for example via a wireless communications network. The transceiver or any suitable transceiver or transmitter and/or receiver means can in some embodiments be configured to communicate with other electronic devices or apparatus via a wire or wired coupling.

The transceiver can communicate with further apparatus by any suitable known communications protocol. For example, in some embodiments the transceiver can use a suitable universal mobile telecommunications system (UMTS) protocol, a

wireless local area network (WLAN) protocol such as for example IEEE 802.X, a suitable short-range radio frequency communication protocol such as Bluetooth, or infrared data communication pathway (IRDA).

The transceiver input/output port 1409 may be configured to receive the signals and in some embodiments determine the parameters as described herein by using the processor 1407 executing suitable code. Furthermore, the device may generate a suitable downmix signal and parameter output to be transmitted to the synthesis device.

In some embodiments the device 1400 may be employed as at least part of the synthesis device. As such the input/output port 1409 may be configured to receive the downmix signals and in some embodiments the parameters determined at the capture device or processing device as described herein, and generate a suitable audio signal format output by using the processor 1407 executing suitable code. The input/output port 1409 may be coupled to any suitable audio output for example to a multichannel speaker system and/or headphones (which may be a headtracked or a non-tracked headphones) or similar.

In general, the various embodiments of the invention may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. For example, some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device, although the invention is not limited thereto. While various aspects of the invention may be illustrated and described as block diagrams, flow charts, or using some other pictorial representation, it is well understood that these blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

The embodiments of this invention may be implemented by computer software executable by a data processor of the mobile device, such as in the processor entity, or by hardware, or by a combination of software and hardware. Further in this regard it should be noted that any blocks of the logic flow as in the Figures may represent program steps, or interconnected logic circuits, blocks and functions, or a combination of program steps and logic circuits, blocks and functions. The software may be stored on such physical media as memory chips, or memory blocks implemented within the processor, magnetic media such as hard disk or floppy disks, and optical media such as for example DVD and the data variants thereof, CD.

The memory may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor-based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The data processors may be of any type suitable to the local technical environment, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs), application specific integrated circuits (ASIC), gate level circuits and processors based on multi-core processor architecture, as non-limiting examples.

Embodiments of the inventions may be practiced in various components such as integrated circuit modules. The design of integrated circuits is by and large a highly automated process. Complex and powerful software tools are available for converting a logic level design into a semiconductor circuit design ready to be etched and formed on a semiconductor substrate.

Programs can automatically route conductors and locate components on a semiconductor chip using well established rules of design as well as libraries of pre-stored design modules. Once the design for a semiconductor circuit has been

completed, the resultant design, in a standardized electronic format may be transmitted to a semiconductor fabrication facility or "fab" for fabrication.

The foregoing description has provided by way of exemplary and non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those skilled in the relevant arts in view of the foregoing description, when read in conjunction with the accompanying drawings and the appended claims. However, all such and similar modifications of the teachings of this invention will still fall within the scope of this invention as defined in the appended claims.

CLAIMS:

1. An apparatus comprising means for:

determining whether to quantise spatial audio parameters associated with a plurality of channel audio signals, wherein the spatial audio parameters are comprised in a metadata structure;

quantising the spatial audio parameters in accordance with the determination; and

adding to a data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised.

2. The apparatus as claimed in Claim 1, wherein the indication is a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag indicates that the spatial audio parameters are quantised and wherein a further state of the binary flag indicates that the spatial audio parameters have not been quantised, wherein the means for adding to the data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised comprises:

means for changing the state of the binary flag in accordance with the determination of whether to quantise the spatial audio parameters.

3. The apparatus as claimed in Claim 1, wherein the indication is a variable comprised in the data filed of the metadata structure, wherein a value of the variable indicates either that the spatial audio parameters have not been quantised and a further value indicates that the spatial audio parameters have been quantised to one of a plurality of coding rates, wherein the means for adding to the data field metadata structure an indication as to whether the spatial audio parameters have been quantised comprises means for setting the value of the variable to indicate either:

the spatial audio parameters have not been quantised; or

a coding rate of the plurality of coding rates to which the spatial audio parameters have been quantised.

4. The apparatus as claimed in Claims 1 to 3, wherein the spatial audio parameters are derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the apparatus further comprises:

means for adding to a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

5. The apparatus as claimed in Claim 4, wherein the energy value of the sub band of the time frame associated with the plurality of channel audio signals is at least one of:

an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and

an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

6. The apparatus as claimed in Claim 4 and 5, wherein the apparatus further comprises:

means for adding to a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band is dependent the energy value of the sub band and an energy value of the further sub band respectively.

7. An apparatus comprising means for:

reading from a data field in a metadata structure comprising spatial audio parameters associated with a plurality of channel audio signals an indication as to whether the spatial audio parameters have been quantised; and

dequantizing the spatial audio parameters in accordance with the indication.

8. The apparatus as claimed in Claim 7, wherein the indication is a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag indicates that the spatial audio parameters are quantised and wherein a further state of the binary flag indicates that the spatial audio parameters have not been quantised.

9. The apparatus as claimed in Claim 7, wherein the indication is a variable comprised in the data filed of the metadata structure, wherein a value of the variable indicates either that the spatial audio parameters have not been quantised and a further value indicates that the spatial audio parameters have been quantised to one of a plurality of coding rates.

10. The apparatus as claimed in Claims 7 to 9, wherein the spatial audio parameters are derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the apparatus further comprises:

means for reading from a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

11. The apparatus as claimed in Claim 10, wherein the energy value of the sub band of the time frame associated with the plurality of channel audio signals is at least one of:

an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and

an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

12. The apparatus as claimed in Claim 10 and 11, wherein the apparatus further comprises:

means for reading from a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the

priority if the further sub band is dependent the energy value of the sub band and an energy value of the further sub band respectively.

13. A method comprising means:

determining whether to quantise spatial audio parameters associated with a plurality of channel audio signals, wherein the spatial audio parameters are comprised in a metadata structure;

quantising the spatial audio parameters in accordance with the determination; and

adding to a data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised.

14. The method as claimed in Claim 13, wherein the indication is a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag indicates that the spatial audio parameters are quantised and wherein a further state of the binary flag indicates that the spatial audio parameters have not been quantised, wherein adding to the data field in the metadata structure an indication as to whether the spatial audio parameters have been quantised comprises:

changing the state of the binary flag in accordance with the determination of whether to quantise the spatial audio parameters.

15. The method as claimed in Claim 13, wherein the indication is a variable comprised in the data filed of the metadata structure, wherein a value of the variable indicates either that the spatial audio parameters have not been quantised and a further value indicates that the spatial audio parameters have been quantised to one of a plurality of coding rates, wherein the adding to the data field metadata structure an indication as to whether the spatial audio parameters have been quantised comprises setting the value of the variable to indicate either:

the spatial audio parameters have not been quantised; or

a coding rate of the plurality of coding rates to which the spatial audio parameters have been quantised.

16. The method as claimed in Claims 13 to 15, wherein the spatial audio parameters are derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the method further comprises:

adding to a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

17. The method as claimed in Claim 16, wherein the energy value of the sub band of the time frame associated with the plurality of channel audio signals is at least one of:

an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and

an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

18. The method as claimed in Claim 16 and 17, wherein the method further comprises:

adding to a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the further sub band is dependent the energy value of the sub band and an energy value of the further sub band respectively.

19. A method comprising:

reading from a data field in a metadata structure comprising spatial audio parameters associated with a plurality of channel audio signals an indication as to whether the spatial audio parameters have been quantised; and

dequantizing the spatial audio parameters in accordance with the indication.

20. The method as claimed in Claim 19, wherein the indication is a binary flag comprised in the data field of the metadata structure, wherein one state of the binary flag indicates that the spatial audio parameters are quantised and wherein a further state of the binary flag indicates that the spatial audio parameters have not been quantised.

21. The method as claimed in Claim 19, wherein the indication is a variable comprised in the data filed of the metadata structure, wherein a value of the variable indicates either that the spatial audio parameters have not been quantised and a further value indicates that the spatial audio parameters have been quantised to one of a plurality of coding rates.

22. The method as claimed in Claims 19 to 21, wherein the spatial audio parameters are derived on a sub band basis for a time frame associated with the plurality of channel audio signals, wherein the method further comprises:

means for reading from a further data field of the metadata structure an energy value of the sub band of the time frame associated with the plurality of channel audio signals for the time frame.

23. The method as claimed in Claim 22, wherein the energy value of the sub band of the time frame associated with the plurality of channel audio signals is at least one of:

an energy value of all channel audio signals in the sub band of the time frame associated with the plurality of channel audio signals; and

an energy value of a channel audio signal in the sub band of the time frame associated with the plurality of channel audio signals.

24. The method as claimed in Claim 22 and 23, wherein the method further comprises:

reading from a yet further data field of the metadata structure an indication of the priority of the sub band of the time frame in relation to a priority of a further sub band of the time frame, wherein the priority of the sub band and the priority if the

further sub band is dependent the energy value of the sub band and an energy value of the further sub band respectively.

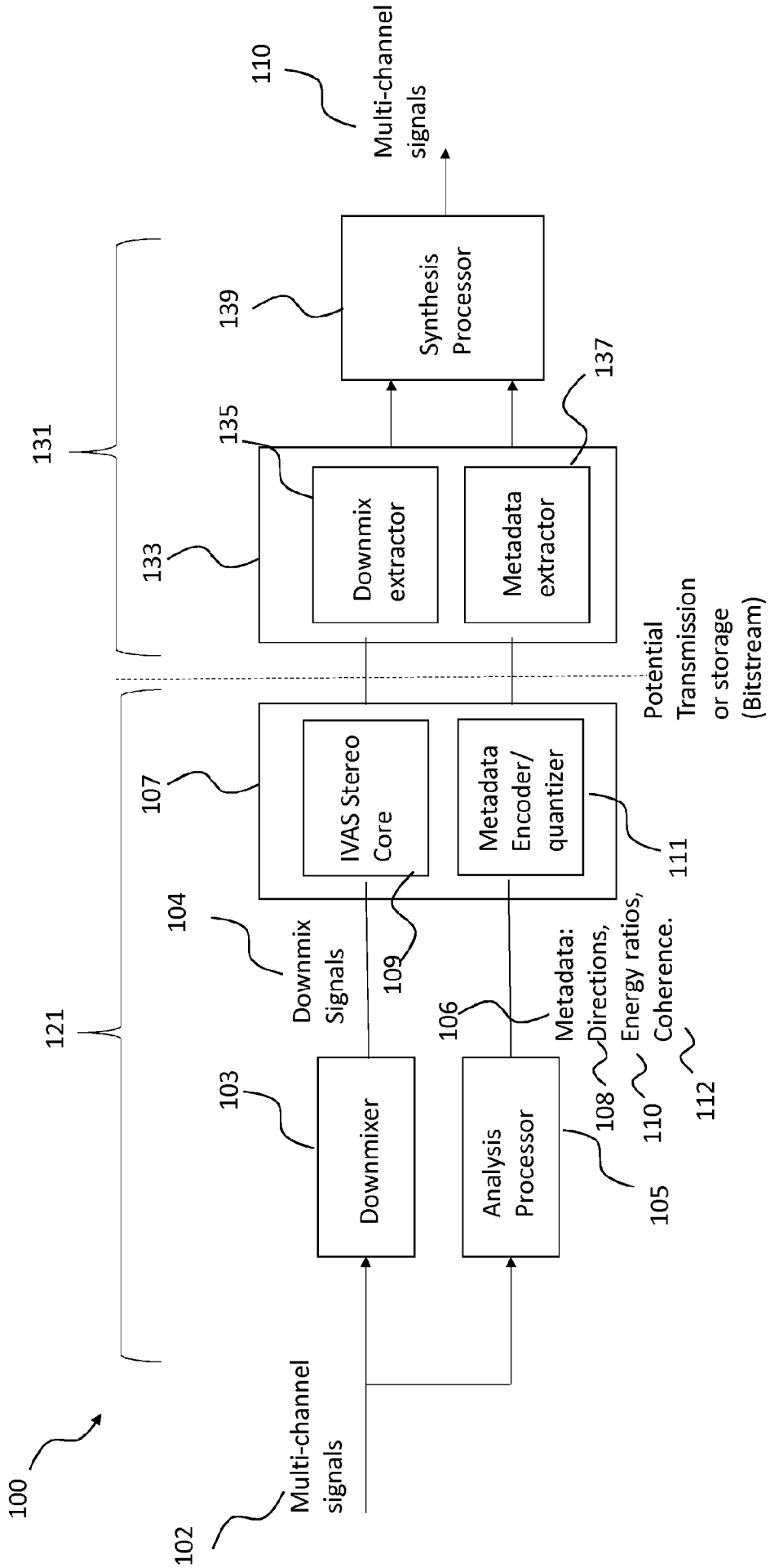
100
102
Multi-channel signals
121
103
Downmixer
104
Downmix Signals
107
IVAS Stereo Core
109
106
Analysis Processor
Metadata: Directions,
108
Energy ratios,
110
Coherence.
112
105
Metadata Encoder/ quantizer
111
Potential Transmission or storage (Bitstream)
131
133
Downmix extractor
135
Metadata extractor
137
139
Synthesis Processor
110
Multi-channel signals

Figure 1

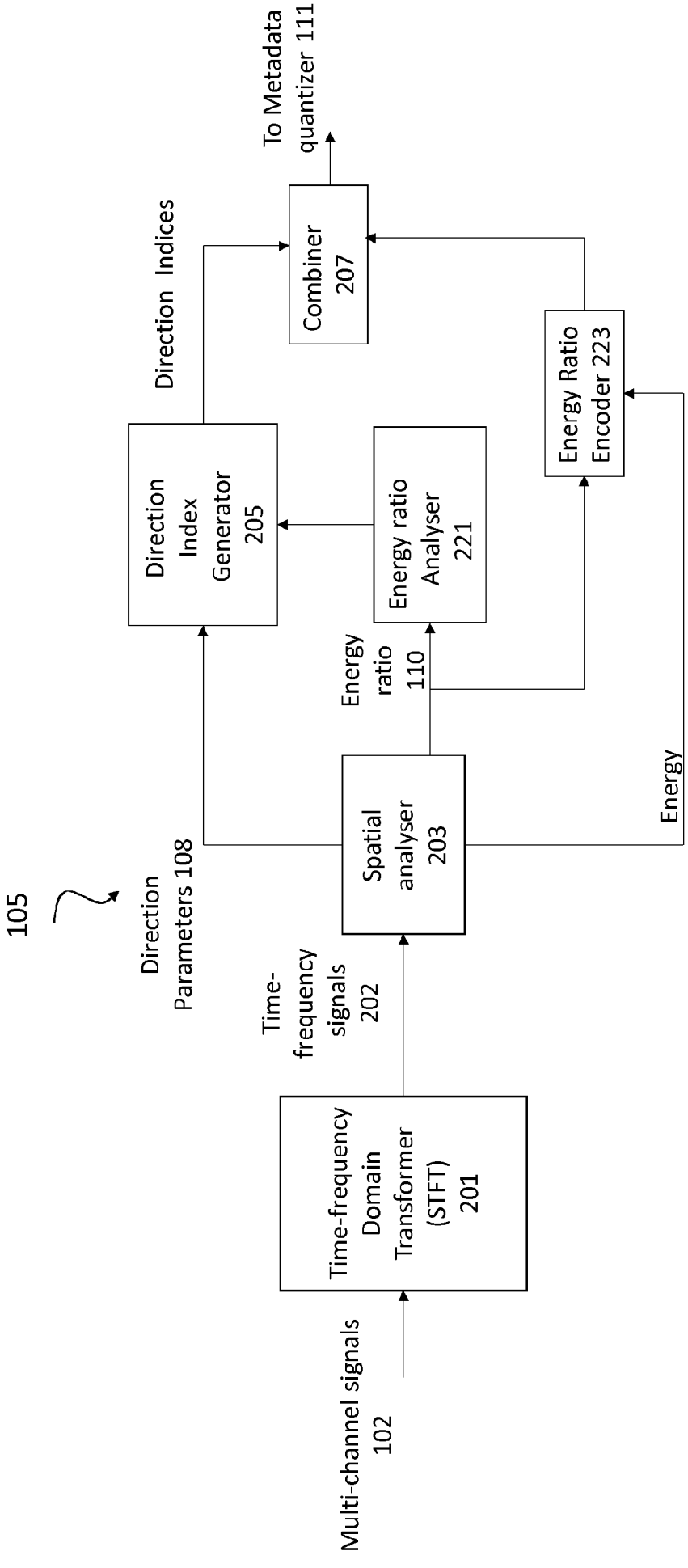
Multi-channel signals 102
Time-frequency Domain Transformer (STFT) 201
Time-frequency signals 202
105
Direction Parameters 108
Spatial analyser 203
Energy ratio 110
Energy 250
Direction Index Generator 205
Energy ratio Analyser 221
Energy Ratio Encoder 223
Direction Indices
Combiner 207
To Metadata quantizer 111

Figure 2

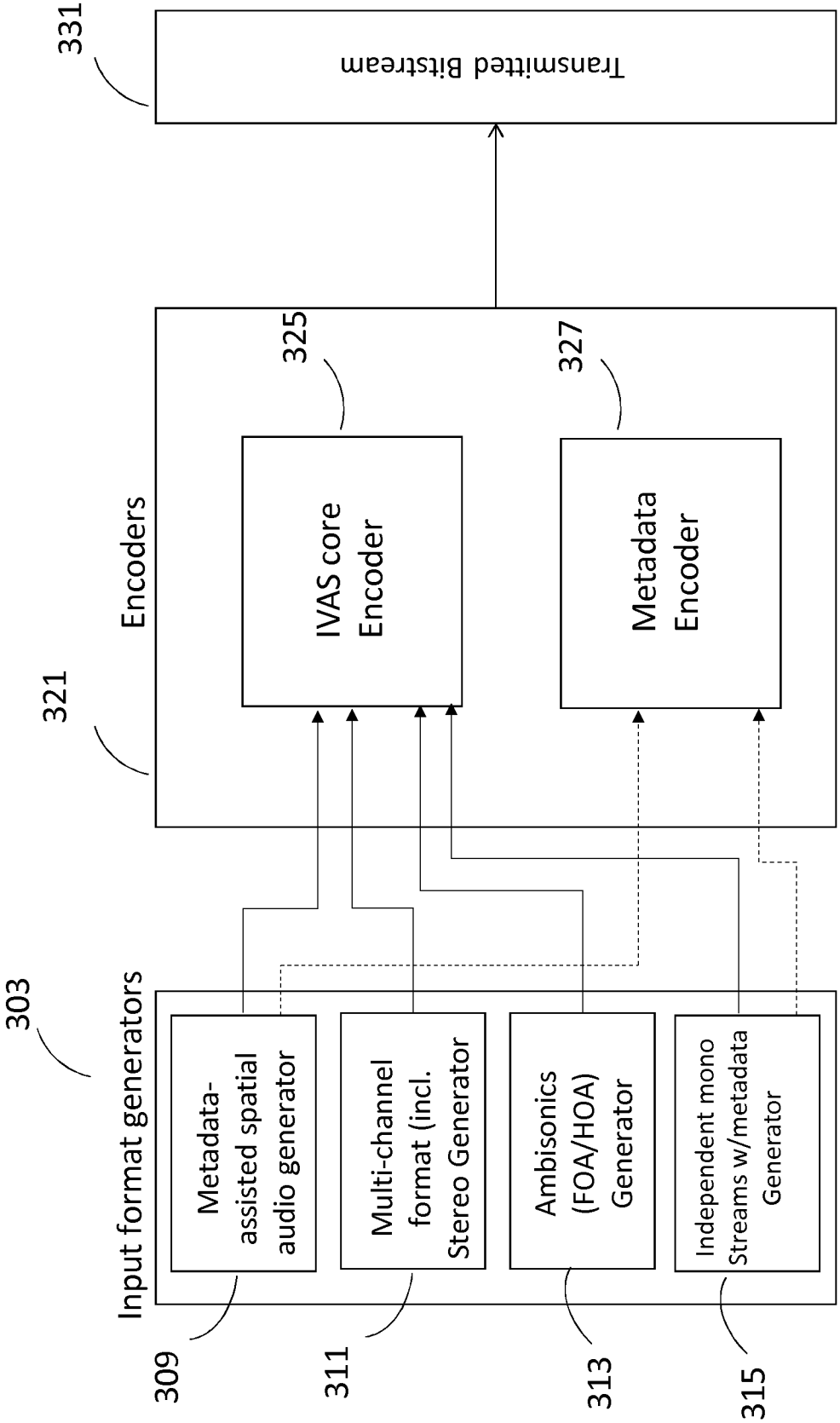
331
Transmitted Bitstream
325
327
Encoders
IVAS core Encoder
Metadata Encoder
321
303
Input format generators
Metadata-assisted spatial audio generator
Multi-channel format (incl. Stereo Generator
Ambisonics (FOA/HOA) Generator
Independent mono Streams w/metadata Generator
309
311
313
315

Figure 3

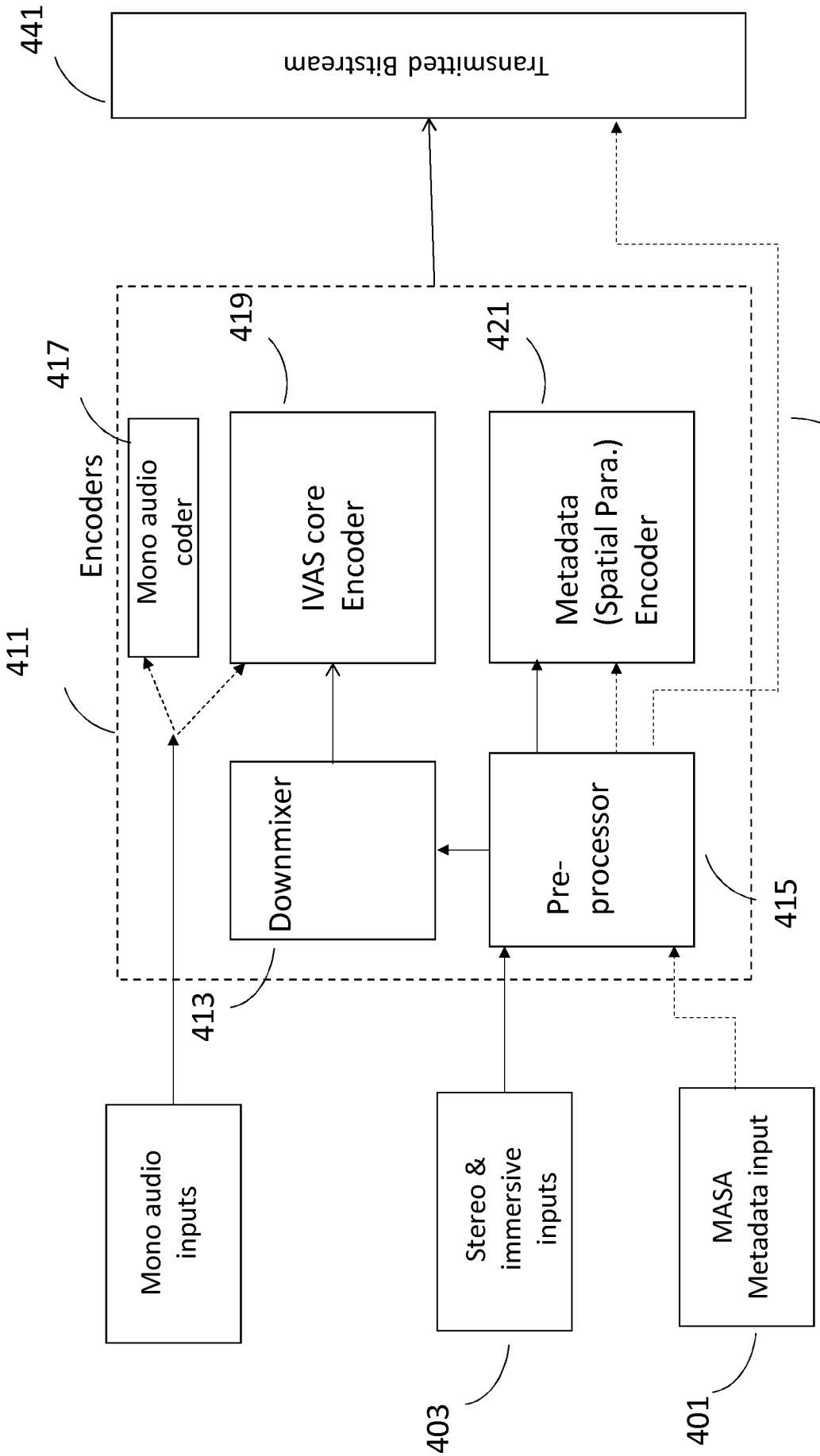
441
Transmitted Bitstream
411
Encoders
417
Mono audio coder
419
IVAS core Encoder
421
Metadata (Spatial Para.) Encoder
431
Downmixer
413
Pre-processor
415
Mono audio inputs
Stereo & immersive inputs
403
MASA Metadata input
401

Figure 4

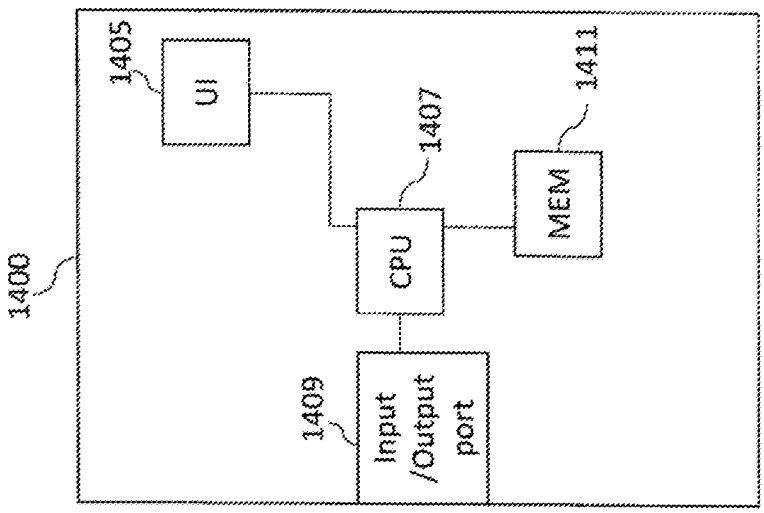

Figure 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2020/050171

A. **CLASSIFICATION OF SUBJECT MATTER**

See extra sheet

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G10L, H04N, H04S

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

FI, SE, NO, DK

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

EPODOC, EPO-Internal full-text databases, Full-text translation databases from Asian languages, WPIAP, COMPDX, INSPEC, TDB, NPL, XP3GPP, XPAIP, XPCPVO, XPESP, XPETSI, XPI3E, XPIEE, XPIETF, XPIOP, XPIPCOM, XPJPEG, XPMISC, XPOAC, XPRD, XPSPRNG, XPTK, 3GPP Documents, Internet (Google)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 3007168 A1 (SONY CORP [JP]) 13 April 2016 (13.04.2016) abstract; pars. [0007]-[0008], [0044]-[0045], [0067], [0069], [0347]-[0348], [0351], [0353]	1-24
X	US 2016005407 A1 (FRIEDRICH TOBIAS [DE] et al.) 07 January 2016 (07.01.2016) abstract; pars. [0013], [0033], [0083]-[0084], [0088], [0091], [0103], [0108]-[0110], [0119]; Fig. 6	1-24
A	US 2003115041 A1 (CHEN WEI-GE [US] et al.) 19 June 2003 (19.06.2003) abstract; pars. [0014], [0155], [0194]	6, 12, 18, 24
A	US 2009125315 A1 (KOISHIDA KAZUHITO [US] et al.) 14 May 2009 (14.05.2009) abstract; pars. [0029], [0046], [0052]	1-24

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

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 May 2020 (18.05.2020)	03 June 2020 (03.06.2020)

Name and mailing address of the ISA/FI
Finnish Patent and Registration Office
FI-00091 PRH, FINLAND
Facsimile No. +358 29 509 5328

Authorized officer
Janne Viljas
Telephone No. +358 29 509 5000

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2020/050171

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2019023488 A1 (DOLBY LABORATORIES LICENSING CORP [US]) 31 January 2019 (31.01.2019) abstract; page 33, line 11	1-24
A	US 2016064005 A1 (PETERS NILS GÜNTHER [US] et al.) 03 March 2016 (03.03.2016) par. [0108]	1-24
A	EP 2146344 A1 (FRAUNHOFER GES FORSCHUNG [DE]) 20 January 2010 (20.01.2010) abstract; pars. [0077], [0093], [0105], [0114]; Figs. 3A-3E	1-24
A	DOLBY LABORATORIES, Inc. IVAS Immersive Conferencing and Example Usage Scenarios, S4-190094, 3GPP Draft. In: 3rd Generation Partnership Project (3GPP) TSGS4 102, Bruges, Belgium [online], 2019-01-22, Vol. SA WG4, pages 1-7, [retrieved on 2020-05-15]. Retrieved from <https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_102_Bruges/>. XP051611907 the whole document	1-24

EP 3007168 A1	13/04/2016	CN 105229734 A	06/01/2016
		CN 105229734 B	20/08/2019
		WO 2014192602	23/02/2017
		JP 6380389 B2	29/08/2018
		TW 201503113 A	16/01/2015
		TW I615834 B	21/02/2018
		US 2016133261 A1	12/05/2016
		US 9805729 B2	31/10/2017
		WO 2014192602 A1	04/12/2014
US 2016005407 A1	07/01/2016	US 9715880 B2	25/07/2017
		CN 105074818 A	18/11/2015
		CN 105074818 B	13/08/2019
		CN 110379434 A	25/10/2019
		EP 2959479 A1	30/12/2015
		EP 2959479 B1	03/07/2019
		EP 3582218 A1	18/12/2019
		JP 2016509260 A	24/03/2016
		JP 6250071 B2	20/12/2017
		JP 2018049287 A	29/03/2018
		JP 6472863 B2	20/02/2019
		JP 2019080347 A	23/05/2019
		US 2017309280 A1	26/10/2017
		US 10360919 B2	23/07/2019
		US 2019348052 A1	14/11/2019
		US 10643626 B2	05/05/2020
		WO 2014128275 A1	28/08/2014
US 2003115041 A1	19/06/2003	US 7240001 B2	03/07/2007
		US 2007185706 A1	09/08/2007
		US 7917369 B2	29/03/2011
		US 2009326962 A1	31/12/2009
		US 8554569 B2	08/10/2013
		US 2014039884 A1	06/02/2014
		US 8805696 B2	12/08/2014
		US 2014316788 A1	23/10/2014
		US 9443525 B2	13/09/2016
US 2009125315 A1	14/05/2009	US 8457958 B2	04/06/2013

WO 2019023488 A1	31/01/2019	CN 110945494 A	31/03/2020
US 2016064005 A1	03/03/2016	US 9847088 B2	19/12/2017
		CN 106575506 A	19/04/2017
		WO 2016033480 A2	03/03/2016
EP 2146344 A1	20/01/2010	EP 2146344 B1	06/07/2016
		AR 072551 A1	08/09/2010
		AU 2009270524 A1	21/01/2010
		AU 2009270524 B2	15/03/2012
		PI 0910999	14/05/2019
		PI 0910999	10/03/2020
		CA 2727883 A1	21/04/2010
		CA 2727883 C	02/09/2014
		CN 102099856 A	15/06/2011
		CN 102099856 B	07/11/2012
		EP 2301024 A1	30/03/2011
		EP 2301024 B1	25/07/2012
		ES 2391715 T3	29/11/2012
		ES 2592416 T3	30/11/2016
		HK 1138673 A1	18/08/2017
		HK 1156143 A1	03/05/2013
		JP 2011528129 A	10/11/2011
		JP 5613157 B2	22/10/2014
		KR 20110055515 A	25/05/2011
		KR 101224884 B1	06/02/2013
		MX 2011000534 A	27/04/2011
		PL 2146344 T3	31/01/2017
		PL 2301024 T3	31/12/2012
		PT 2146344 T	13/10/2016
		RU 2010154749 A	10/07/2012
		RU 2483364 C2	27/05/2013
		TW 201009814 A	01/03/2010
		TW I441167 B	11/06/2014
		US 2011202355 A1	18/08/2011
		US 8321210 B2	27/11/2012
		US 2013066640 A1	14/03/2013
		US 8959017 B2	17/02/2015
		WO 2010006717 A1	21/01/2010

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2020/050171

CLASSIFICATION OF SUBJECT MATTER

IPC
G10L 19/22 (2013.01)
G10L 19/16 (2013.01)
G10L 19/008 (2013.01)
H04S 3/00 (2006.01)
G10L 19/02 (2013.01)
G10L 19/00 (2013.01)