(54)   Title

Decoder and method for a generalized spatial-audio-object-coding parametric concept for multichannel downmix/upmix cases

(51)   International Patent Classification(s)

G10L 19/008 (2013.01)

(21)   Application No: **2016234987**   (22)   Date of Filing: **2016.09.29**

(43)   Publication Date: **2016.10.20**

(43)   Publication Journal Date: **2016.10.20**

(62)   Divisional of:

2013298463

(71)   Applicant(s)

Fraunhofer-Gesellschaft zur Foerderung der angewandten Forschung e.V.

(72)   Inventor(s)

Kastner, Thorsten;Herre, Jurgen;Terentiv, Leon;Hellmuth, Oliver

(74)   Agent / Attorney

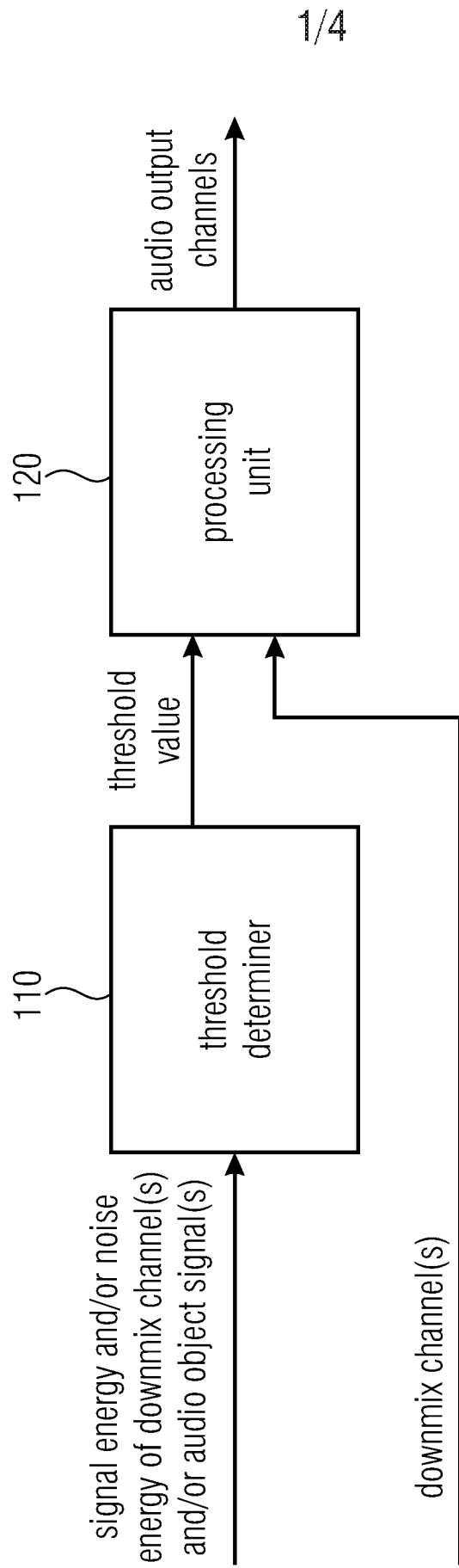**Davies Collison Cave Pty Ltd, GPO Box 3876, SYDNEY, NSW, 2001**

signal energy and/or noise energy of downmix channel(s) and/or audio object signal(s)

110

threshold determiner

threshold value

120

processing unit

audio output channels

downmix channel(s)

FIG 1

Decoder and Method for a Generalized Spatial-Audio-Object-Coding

Parametric Concept for Multichannel Downmix/Upmix Cases

<u>Abstract</u>

5

A decoder for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels is provided. The downmix signal encodes one or more audio object signals. The decoder comprises a threshold determiner (110) for determining a threshold value depending on a signal energy and/or a noise energy of at least one of the of or more audio object signals and/or

10   depending on a signal energy and/or a noise energy of at least one of the one or more downmix channels. Moreover, the decoder comprises a processing unit (120) for generating the one or more audio output channels from the one or more downmix channels depending on the threshold value.

15

Fig. 1

Decoder and Method for a Generalized Spatial-Audio-Object-Coding Parametric Concept for Multichannel Downmix/Upmix Cases

The present invention relates to an apparatus and a method for a generalized spatial-audio-object-coding parametric concept for multichannel downmix/upmix cases.

In modern digital audio systems, it is a major trend to allow for audio-object related modifications of the transmitted content on the receiver side. These modifications include gain modifications of selected parts of the audio signal and/or spatial re-positioning of dedicated audio objects in case of multi-channel playback via spatially distributed speakers. This may be achieved by individually delivering different parts of the audio content to the different speakers.

In other words, in the art of audio processing, audio transmission, and audio storage, there is an increasing desire to allow for user interaction on object-oriented audio content playback and also a demand to utilize the extended possibilities of multi-channel playback to individually render audio contents or parts thereof in order to improve the hearing impression. By this, the usage of multi-channel audio content brings along significant improvements for the user. For example, a three-dimensional hearing impression can be obtained, which brings along an improved user satisfaction in entertainment applications. However, multi-channel audio content is also useful in professional environments, for example, in telephone conferencing applications, because the talker intelligibility can be improved by using a multi-channel audio playback. Another possible application is to offer to a listener of a musical piece to individually adjust playback level and/or spatial position of different parts (also termed as "audio objects") or tracks, such as a vocal part or different instruments. The user may perform such an adjustment for reasons of personal taste, for easier transcribing one or more part(s) from the musical piece, educational purposes, karaoke, rehearsal, etc.

The straightforward discrete transmission of all digital multi-channel or multi-object audio content, e.g., in the form of pulse code modulation (PCM) data or even compressed audio formats, demands very high bitrates. However, it is also desirable to transmit and store audio data in a bitrate efficient way. Therefore, one is willing to accept a reasonable tradeoff between audio quality and bitrate requirements in order to avoid an excessive resource load caused by multi-channel/multi-object applications.

Recently, in the field of audio coding, parametric techniques for the bitrate-efficient transmission/storage of multi-channel/multi-object audio signals have been introduced by,

e.g., the Moving Picture Experts Group (MPEG) and others. One example is MPEG Surround (MPS) as a channel oriented approach [MPS, BCC], or MPEG Spatial Audio Object Coding (SAOC) as an object oriented approach [JSC, SAOC, SAOC1, SAOC2]. Another object–oriented approach is termed as "informed source separation" [ISS1, ISS2,

5   ISS3, ISS4, ISS5, ISS6]. These techniques aim at reconstructing a desired output audio scene or a desired audio source object on the basis of a downmix of channels/objects and additional side information describing the transmitted/stored audio scene and/or the audio source objects in the audio scene.

10   The estimation and the application of channel/object related side information in such systems is done in a time-frequency selective manner. Therefore, such systems employ time-frequency transforms such as the Discrete Fourier Transform (DFT), the Short Time Fourier Transform (STFT) or filter banks like Quadrature Mirror Filter (QMF) banks, etc. The basic principle of such systems is depicted in Fig. 2, using the example of MPEG

15   SAOC.

In case of the STFT, the temporal dimension is represented by the time-block number and the spectral dimension is captured by the spectral coefficient ("bin") number. In case of QMF, the temporal dimension is represented by the time-slot number and the spectral

20   dimension is captured by the sub-band number. If the spectral resolution of the QMF is improved by subsequent application of a second filter stage, the entire filter bank is termed hybrid QMF and the fine resolution sub-bands are termed hybrid sub-bands.

As already mentioned above, in SAOC the general processing is carried out in a time-

25   frequency selective way and can be described as follows within each frequency band, as depicted in Fig. 2:

-       N input audio object signals s_1 ... s_N are mixed down to P channels x_1 ... x_P as part of the encoder processing using a downmix matrix consisting of the elements $d_{1,1}$

30   ... $d_{N,P}$. In addition, the encoder extracts side information describing the characteristics of the input audio objects (side-information-estimator (SIE) module). For MPEG SAOC, the relations of the object powers w.r.t. each other are the most basic form of such a side information.

35   -   Downmix signal(s) and side information are transmitted/stored. To this end, the downmix audio signal(s) may be compressed, e.g., using well-known perceptual audio coders such MPEG-1/2 Layer II or III (aka .mp3), MPEG-2/4 Advanced Audio Coding (AAC) etc.

- On the receiving end, the decoder conceptually tries to restore the original object signals ("object separation") from the (decoded) downmix signals using the transmitted side information. These approximated object signals $\hat{s}_1 \ldots \hat{s}_N$ are then mixed into a target scene represented by M audio output channels $\hat{y}_1 \ldots \hat{y}_M$ using a rendering matrix described by the coefficients $r_{1,1} \ldots r_{N,M}$ in Fig. 2. The desired target scene may be, in the extreme case, the rendering of only one source signal out of the mixture (source separation scenario), but also any other arbitrary acoustic scene consisting of the objects transmitted. For example, the output can be a single-channel, a 2-channel stereo or 5.1 multi-channel target scene.

Increasing bandwidth / storage available and ongoing improvements in the field of audio coding allows the user to select from a steadily increasing choice of multi-channel audio productions. Multi-channel 5.1 audio formats are already standard in DVD and Blue-Ray productions. New audio formats like MPEG-H 3D Audio with even more audio transport channels appear at the horizon, which will provide the end-users a highly immersive audio experience.

Parametric audio object coding schemes are currently restricted to a maximum of two downmix channels. They can only be applied to some extend on multi-channel mixtures, for example on only two selected downmix channels. The flexibility these coding schemes offer to the user to adjust the audio scene to his/her own preferences is thus severely limited, e.g., with respect to changing audio level of the sports commentator and the atmosphere in sports broadcast.

Moreover, current audio object coding schemes offer only a limited variability in the mixing process at the encoder side. The mixing process is limited to time-variant mixing of the audio objects; and frequency-variant mixing is not possible.

It would therefore be highly appreciated if improved concepts for audio object coding would be provided.

The present invention seeks to provide improved concepts for audio object coding.

A decoder for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels is provided. The downmix signal encodes two or more audio object signals. The decoder comprises a threshold determiner for determining a threshold value depending on a signal energy

and/or a noise energy of at least one of the two or more audio object signals and/or depending on a signal energy and/or a noise energy of at least one of the one or more downmix channels. Moreover, the decoder comprises a processing unit for generating the one or more audio output channels from the one or more downmix channels depending on

5 the threshold value. The processing unit is configured to generate the one or more audio output channels from the one or more downmix channels depending on an object covariance matrix (E) of the one or more audio object signals, depending on a downmix matrix (D) for downmixing the two or more audio object signals to obtain the one or more downmix channels, and depending on the threshold value. The processing unit is further

10 configured to generate the one or more audio output channels from the one or more downmix channels by applying the threshold value in a function to inverse a downmix channel cross correlation matrix Q, wherein Q is defined as $Q=DED^*$. D is the downmix matrix for downmixing the two or more audio object signals to obtain the two or more downmix channels. E is the object covariance matrix of the one or more audio object

15 signals. The processing unit is further configured to generate the one or more audio output channels from the one or more downmix channels by computing the eigenvalues of the downmix channel cross correlation matrix Q.

According to an embodiment, the downmix signal may comprise two or more downmix

20 channels, and the threshold determiner may be configured to determine the threshold value depending on a noise energy of each of the two or more downmix channels.

In an embodiment, the threshold determiner may be configured to determine the threshold value depending on the sum of all noise energy in the two or more downmix channels.

25
According to an embodiment, the downmix signal may encode two or more audio object signals, and the threshold determiner may be configured to determine the threshold value depending on a signal energy of the audio object signal of the two or more audio object signals which has the greatest signal energy of the two or more audio object signals.

30
In an embodiment, the downmix signal may comprise two or more downmix channels, and the threshold determiner may be configured to determine the threshold value depending on the sum of all noise energy in the two or more downmix channels.

35 According to an embodiment, the downmix signal may encode the one or more audio object signals for each time-frequency tile of a plurality of time-frequency tiles. The threshold determiner may be configured to determine a threshold value for each time-frequency tile of the plurality of time-frequency tiles depending on the signal energy or the

noise energy of at least one of the of or more audio object signals or depending on the signal energy or the noise energy of at least one of the one or more downmix channels, wherein a first threshold value of a first time-frequency tile of the plurality of time-frequency tiles may differ from a second time-frequency time of the plurality of time-frequency tiles. The processing unit may be configured to generate for each time-frequency tile of the plurality of time-frequency tiles a channel value of each of the one or more audio output channels from the one or more downmix channels depending on the threshold value if said time-frequency tile.

In an embodiment, the decoder may be configured to determine the threshold value T in decibel according to the formula

$$T[dB] = E_{noise}[dB] - E_{ref}[dB] - Z \quad \text{or according to the formula}$$

$$T[dB] = E_{noise}[dB] - E_{ref}[dB],$$

wherein $T[dB]$ indicates the threshold value in decibel, wherein $E_{noise}[dB]$ indicates the sum of all noise energy in the two or more downmix channels in decibel, wherein $E_{ref}[dB]$ indicates the signal energy of one of the audio object signals in decibel, and wherein Z indicates an additional parameter being a number. In an alternative embodiment, $E_{noise}[dB]$ indicates the sum of all noise energy in the two or more downmix channels in decibel divided by the number of the downmix channels.

According to an embodiment, the decoder may be configured to determine the threshold value T according to the formula

$$T = \frac{E_{noise}}{E_{ref} \cdot Z} \quad \text{or according to the formula}$$

$$T = \frac{E_{noise}}{E_{ref}},$$

wherein T indicates the threshold value, wherein E_{noise} indicates the sum of all noise energy in the two or more downmix channels, wherein E_{ref} indicates the signal energy of one of the audio object signals, and wherein Z indicates an additional parameter being a

number. In an alternative embodiment, E_{noise}[dB] indicates the sum of all noise energy in the two or more downmix channels divided by the number of the downmix channels.

According to an embodiment, the processing unit may be configured to generate the one or more audio output channels from the one or more downmix channels depending on an object covariance matrix (E) of the one or more audio object signals, depending on a downmix matrix (D) for downmixing the two or more audio object signals to obtain the two or more downmix channels, and depending on the threshold value.

In an embodiment, the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels by applying the threshold value in a function to inverse a downmix channel cross correlation matrix Q, wherein Q is defined as $Q = DED^*$, wherein D is the downmix matrix for downmixing the two or more audio object signals to obtain the two or more downmix channels, and wherein E is the object covariance matrix of the one or more audio object signals.

For example, the processing unit may be configured to generate the one or more audio output channels from the one or more downmix channels by computing the eigenvalues of the downmix channel cross correlation matrix Q or by calculating the singular values of the downmix channel cross correlation matrix Q.

E.g., the processing unit may be configured to generate the one or more audio output channels from the one or more downmix channels by multiplying the largest eigenvalue of the eigenvalues of the downmix channel cross correlation matrix Q with the threshold value to obtain a relative threshold.

For example, the processing unit may be configured to generate the one or more audio output channels from the one or more downmix channels by generating a modified matrix. The processing unit may be configured to generate the modified matrix depending on only those eigenvectors of the downmix channel cross correlation matrix Q, which have an eigenvalue of the eigenvalues of the downmix channel cross correlation matrix Q, which is greater than or equal to the modified threshold. Moreover, the processing unit may be configured to conduct a matrix inversion of the modified matrix to obtain an inverted matrix. Furthermore, the processing unit may be configured to apply the inverted matrix on one or more of the downmix channels to generate the one or more audio output channels.

Moreover, a method for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels is

provided. The downmix signal encodes two or more audio object signals. The method comprises:

- Determining a threshold value depending on a signal energy or a noise energy of at least one of the two or more audio object signals or depending on a signal energy or a noise energy of at least one of the one or more downmix channels. And:

- Generating the one or more audio output channels from the one or more downmix channels depending on the threshold value.

Generating the one or more audio output channels from the one or more downmix channels depending on an object covariance matrix (E) of the one or more audio object signals is conducted depending on a downmix matrix (D) for downmixing the two or more audio object signals to obtain the one or more downmix channels, and depending on the threshold value. Furthermore, generating the one or more audio output channels from the one or more downmix channels is conducted by applying the threshold value in a function to inverse a downmix channel cross correlation matrix Q, wherein Q is defined as $Q=DED^*$. D is the downmix matrix for downmixing the two or more audio object signals to obtain the two or more downmix channels. E is the object covariance matrix of the one or more audio object signals. Furthermore, generating the one or more audio output channels from the one or more downmix channels is conducted by computing the eigenvalues of the downmix channel cross correlation matrix Q.

Moreover, a computer program for implementing the above-described method when being executed on a computer or signal processor is provided.

In the following, embodiments of the present invention are described in more detail with reference to the figures, in which:

Fig. 1   illustrates a decoder for generating an audio output signal comprising one or more audio output channels according to an embodiment,

Fig. 2   is a SAOC system overview depicting the principle of such systems using the example of MPEG SAOC,

Fig. 3   illustrates an overview of the G-SAOC parametric upmix concept, and

Fig. 4   illustrates a general downmix/upmix concept.

Before describing embodiments of the present invention, more background on state-of-the-art-SAOC systems is provided.

5   Fig. 2 shows a general arrangement of an SAOC encoder 10 and an SAOC decoder 12. The SAOC encoder 10 receives as an input N objects, i.e., audio signals s_1 to s_N. In particular, the encoder 10 comprises a downmixer 16 which receives the audio signals s_1 to s_N and downmixes same to a downmix signal 18. Alternatively, the downmix may be provided externally ("artistic downmix") and the system estimates additional side

10   information to make the provided downmix match the calculated downmix. In Fig. 2, the downmix signal is shown to be a P-channel signal. Thus, any mono (P=1), stereo (P=2) or multi-channel (P>2) downmix signal configuration is conceivable.

In the case of a stereo downmix, the channels of the downmix signal 18 are denoted $L0$

15   and $R0$, in case of a mono downmix same is simply denoted $L0$. In order to enable the SAOC decoder 12 to recover the individual objects s_1 to s_N, side-information estimator 17 provides the SAOC decoder 12 with side information including SAOC-parameters. For example, in case of a stereo downmix, the SAOC parameters comprise object level differences (OLD), inter-object correlations (IOC) (inter-object cross correlation

20   parameters), downmix gain values (DMG) and downmix channel level differences (DCLD). The side information 20, including the SAOC-parameters, along with the downmix signal 18, forms the SAOC output data stream received by the SAOC decoder 12.

The SAOC decoder 12 comprises an up-mixer which receives the downmix signal 18 as

25   well as the side information 20 in order to recover and render the audio signals $\hat{s}_1$ and $\hat{s}_N$ onto any user-selected set of channels $\hat{y}_1$ to $\hat{y}_M$, with the rendering being prescribed by rendering information 26 input into SAOC decoder 12.

The audio signals s_1 to s_N may be input into the encoder 10 in any coding domain, such

30   as, in time or spectral domain. In case the audio signals s_1 to s_N are fed into the encoder 10 in the time domain, such as PCM coded, encoder 10 may use a filter bank, such as a hybrid QMF bank, in order to transfer the signals into a spectral domain, in which the audio signals are represented in several sub-bands associated with different spectral portions, at a specific filter bank resolution. If the audio signals s_1 to s_N are already in the

35   representation expected by encoder 10, same does not have to perform the spectral decomposition.

More flexibility in the mixing process allows an optimal exploitation of signal object characteristics. A downmix can be produced which is optimized for the parametric separation at the decoder side regarding perceived quality.

5   The embodiments extends the parametric part of the SAOC scheme to an arbitrary number of downmix/upmix channels. The following figure provides overview of the Generalized Spatial Audio Object Coding (G-SAOC) parametric upmix concept:

Fig. 3 illustrates an overview of the G-SAOC parametric upmix concept A fully flexible
10   post-mixing (rendering) of the parametrically reconstructed audio objects can be realized.

Inter alia, Fig. 3 illustrates an audio decoder 310, an object separator 320 and a renderer 330.

15   Let us consider the following common notation:

x - input audio object signal   (of size N_{obj})

y - downmix audio signal   (of size N_{dmx})

z - rendered output scene signal   (of size N_{upmix})

20   D - downmix matrix   (of size $N_{obj} \times N_{dmx}$)

R - rendering matrix   (of size $N_{obj} \times N_{upmix}$)

G - parametric upmix matrix   (of size $N_{dmx} \times N_{upmix}$)

E - object covariance matrix   (of size $N_{obj} \times N_{obj}$)

25   All introduced matrices are (in general) time and frequency variant.

In the following, the constitutive relationship for parametric upmixing is provided.

At first, general downmix/upmix concepts are provided with reference to Fig. 4. In
30   particular, Fig. 4 illustrates a general downmix/upmix concept, wherein Fig. 4 illustrates modeled (left) and parametric upmix (right) systems.

More particularly, Fig. 4 illustrates a rendering unit 410, a downmix unit 421 and a parametrix upmix unit 422.

35

The ideal (modeled) rendered output scene signal z is defined as, see Fig (left):

$$\mathbf{R}x = z. \qquad\qquad (1)$$

The downmix audio signal y is determined as, see Fig. 4 (right):

$$\mathbf{D}x = y. \qquad (2)$$

The constitutive relationship (applied to the downmix audio signal) for the parametric output scene signal reconstruction can be represented as, see Fig. 4 (right):

$$\mathbf{G}y = z. \qquad (3)$$

The parametric upmix matrix can be defined from (1) and (2) as the following function of the downmix and rendering matrices $\mathbf{G} = \mathbf{G}(\mathbf{D}, \mathbf{R})$:

$$\mathbf{G} = \mathbf{RED}^{*}\left(\mathbf{DED}^{*}\right)^{-1}. \qquad (4)$$

In the following, improving the stability of the parametric source estimation according to embodiments is considered.

The parametric separation scheme within MPEG SAOC is based on a Least Mean Square (LMS) estimation of the sources in the mixture. The LMS estimation involves the inversion of the parametrically described downmix-channel covariance matrix $\mathbf{Q} = \mathbf{DED}^{*}$. Algorithms for matrix inversion are in general sensitive to ill-conditioned matrices. The inversion of such a matrix can cause unnatural sounds, called artifacts, in the rendered output scene. A heuristically determined fixed threshold $\mathbf{T}$ in MPEG SAOC currently avoids this. Although artifacts are avoided by this method, a sufficient possible separation performance at the decoder side can thereby not be achieved.

Fig. 1 illustrates a decoder for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels according to an embodiment. The downmix signal encodes one or more audio object signals.

The decoder comprises a threshold determiner 110 for determining a threshold value depending on a signal energy and/or a noise energy of at least one of the of or more audio object signals and/or depending on a signal energy and/or a noise energy of at least one of the one or more downmix channels.

Moreover, the decoder comprises a processing unit 120 for generating the one or more audio output channels from the one or more downmix channels depending on the threshold value.

5   In contrast to the state of the art, the threshold value determined by the threshold determiner 110 depends on a signal energy or a noise energy of the one or more downmix channels or of the encoded one or more audio object signals. In embodiments, as the signal and noise energies of the one or more downmix channels and/or of the one or more audio object signal values varies, so varies the threshold value, e.g., from time

10   instance to time instance, or from time-frequency tile to time-frequency tile.

Embodiments provide an adaptive threshold method for matrix inversion to achieve an improved parametric separation of the audio objects at the decoder side. The separation performance is on the average better but never less the currently utilized fixed threshold

15   scheme used in MPEG SAOC in the algorithm for inverting the $\mathbf{Q}$ matrix.

The threshold $\mathbf{T}$ is dynamically adapted to the precision of the data for each processed time-frequency tile. Separation performance is thus improved and artifacts in the rendered output scene caused by inversion of ill-conditioned matrices are avoided.

20

According to an embodiment, the downmix signal may comprise two or more downmix channels, and the threshold determiner 110 may be configured to determine the threshold value depending on a noise energy of each of the two or more downmix channels.

25   In an embodiment, the threshold determiner 110 may be configured to determine the threshold value depending on the sum of all noise energy in the two or more downmix channels.

According to an embodiment, the downmix signal may encode two or more audio object

30   signals, and the threshold determiner 110 may be configured to determine the threshold value depending on a signal energy of the audio object signal of the two or more audio object signals which has the greatest signal energy of the two or more audio object signals.

35   In an embodiment, the downmix signal may comprise two or more downmix channels, and the threshold determiner 110 may be configured to determine the threshold value depending on the sum of all noise energy in the two or more downmix channels.

According to an embodiment, the downmix signal may encode the one or more audio object signals for each time-frequency tile of a plurality of time-frequency tiles. The threshold determiner 110 may be configured to determine a threshold value for each time-frequency tile of the plurality of time-frequency tiles depending on the signal energy or the noise energy of at least one of the of or more audio object signals or depending on the signal energy or the noise energy of at least one of the one or more downmix channels, wherein a first threshold value of a first time-frequency tile of the plurality of time-frequency tiles may differ from a second time-frequency time of the plurality of time-frequency tiles. The processing unit 120 may be configured to generate for each time-frequency tile of the plurality of time-frequency tiles a channel value of each of the one or more audio output channels from the one or more downmix channels depending on the threshold value if said time-frequency tile.

According to an embodiment, the decoder may be configured to determine the threshold value T according to the formula

$$T = \frac{E_{noise}}{E_{ref} \cdot Z}$$ or according to the formula

$$T = \frac{E_{noise}}{E_{ref}},$$

wherein T indicates the threshold value, wherein E_{noise} indicates the sum of all noise energy in the two or more downmix channels, wherein E_{ref} indicates the signal energy of one of the audio object signals, and wherein Z indicates an additional parameter being a number. In an alternative embodiment, E_{noise} indicates the sum of all noise energy in the two or more downmix channels divided by the number of the downmix channels.

In an embodiment, the decoder may be configured to determine the threshold value T in decibel according to the formula

$$T[dB] = E_{noise}[dB] - E_{ref}[dB] - Z$$ or according to the formula

$$T[dB] = E_{noise}[dB] - E_{ref}[dB],$$

wherein $T[dB]$ indicates the threshold value in decibel, wherein $E_{noise}[dB]$ indicates the sum of all noise energy in the two or more downmix channels in decibel, wherein $E_{ref}[dB]$

indicates the signal energy of one of the audio object signals in decibel, and wherein $\mathbf{Z}$ indicates an additional parameter being a number. In an alternative embodiment, $\mathbf{E}_{noise}[dB]$ indicates the sum of all noise energy in the two or more downmix channels in decibel divided by the number of the downmix channels.

5

In particular, a rough estimation of the threshold can be given for each time-frequency tile by:

$$\mathbf{T}[dB] = \mathbf{E}_{noise}[dB] - \mathbf{E}_{ref}[dB] - \mathbf{Z}. \qquad (5)$$

10

$\mathbf{E}_{noise}$ may indicate the noise floor level, e.g., the sum of all noise energy in the downmix channels. The noise floor can be defined by the resolution of the audio data, e.g., a noise floor caused by PCM-coding of the channels. Another possibility is to account for coding noise if the downmix is compressed. For such a case, the noise floor caused by the

15 coding algorithm can be added. In an alternative embodiment, $\mathbf{E}_{noise}[dB]$ indicates the sum of all noise energy in the two or more downmix channels in decibel divided by the number of the downmix channels.

$\mathbf{E}_{ref}$ may indicate a reference signal energy. In the simplest form, this can be the energy

20 of the strongest audio object:

$$\mathbf{E}_{ref} = \max(\mathbf{E}). \qquad (6)$$

$\mathbf{Z}$ may indicate a penalty factor to cope for additional parameters that affect the

25 separation resolution, e.g. the difference of the number of downmix channels and number of source objects. Separation performance decreases with increasing number of audio objects. Moreover, the effects of the quantization of the parametric side info on the separation can also be included.

30 In an embodiment, the processing unit 120 is configured to generate the one or more audio output channels from the one or more downmix channels depending on the object covariance matrix $\mathbf{E}$ of the one or more audio object signals, depending on the downmix matrix $\mathbf{D}$ for downmixing the two or more audio object signals to obtain the two or more downmix channels, and depending on the threshold value.

35

According to an embodiment, for generating the one or more audio output channels from the one or more downmix channels depending on the threshold value, the processing unit 120 may be configured to proceed as follows:

5   The threshold (which may be referred to as a "separation-resolution threshold") is applied at the decoder side in the function to inverse the parametrically estimated downmix channel cross correlation matrix $\mathbf{Q}$.

The singular values of $\mathbf{Q}$ or the eigenvalues of $\mathbf{Q}$ are computed.

10   The largest eigenvalue is taken and multiplied with the threshold $\mathbf{T}$.

All except the largest eigenvalue are compared to this relative threshold and omitted if they are smaller.

15   The matrix inversion is then carried out on a modified matrix, wherein the modified matrix may, for example, be the matrix defined by the reduced set of vectors. It should be noted that for the case that all except the highest eigenvalue are omitted, the highest eigenvalue should be set to the noise floor level if the eigenvalue is below.

20   For example, the processing unit 120 may be configured to generate the one or more audio output channels from the one or more downmix channels by generating the modified matrix. The modified matrix may be generated depending on only those eigenvectors of the downmix channel cross correlation matrix $\mathbf{Q}$, which have an

25   eigenvalue of the eigenvalues of the downmix channel cross correlation matrix $\mathbf{Q}$, which is greater than or equal to the modified threshold. The processing unit 120 may be configured to conduct a matrix inversion of the modified matrix to obtain an inverted matrix. Then, the processing unit 120 may be configured to apply the inverted matrix on one or more of the downmix channels to generate the one or more audio output channels.

30   For example, the inverted matrix may be applied on one or more of the downmix channels in one of the ways as the inverted matrix of the matrix product $\mathbf{D}\mathbf{E}\mathbf{D}^{*}$ is applied on the downmix channels (see, e.g. [SAOC], see, in particular, for example: ISO/IEC, "MPEG audio technologies – Part 2: Spatial Audio Object Coding (SAOC)," ISO/IEC JTC1/SC29/WG11 (MPEG) International Standard 23003-2:2010, in particular, see,

35   chapter "SAOC Processing", more particularly, see subchapter "Transcoding modes" and subchapter "Decoding modes").

The parameters which may be employed for estimating the threshold $\mathbf{T}$ can be either determined at the encoder and embedded in the parametric side information or estimated directly at the decoder side.

5 A simplified version of the threshold estimator can be used at the encoder side to indicate potential instabilities in the source estimation at the decoder side. In its simplest form, neglecting all noise terms, the norm of the downmix matrix can be computed indicating that the full potential of the available downmix channels for parametrically estimating the source signals at the decoder side cannot be exploited. Such an indicator can be used

10 during the mixing process to avoid mixing matrices that are critical for estimating the source signals.

Regarding parameterization of the object covariance matrix, one can see that the described parametric upmix method based on the constitutive relationship (4) is invariant

15 to the sign of off-diagonal entities of the object covariance matrix $\mathbf{E}$. This results in possibility of more efficient (in comparison with SAOC) parameterization (quantization and coding) of the values representing inter-object correlations.

Regarding transport of information representing the downmix matrix, generally, the audio

20 input and downmix signals x, y together with the covariance matrix $\mathbf{E}$ are determined at the encoder side. The coded representation of the audio downmix signal y and information describing covariance matrix $\mathbf{E}$ are transmitted to the decoder side (via bitstream payload). The rendering matrix $\mathbf{R}$ is set and available at the decoder side.

25 The information representing the downmix matrix $\mathbf{D}$ (applied at the encoder and used as the decoder) can be determined (at the encoder) and obtained (at the decoder) using the following principle methods.

The downmix matrix $\mathbf{D}$ can be:

30

- set and applied (at the encoder) and its quantized and coded representation explicitly transmitted (to the decoder) via bitstream payload.

- assigned and applied (at the encoder) and restored (at the decoder) using stored

35 lookup tables (i.e. set of predetermined downmix matrices).

- assigned and applied (at the encoder) and restored (at the decoder) according to the specific algorithm or method (e.g. specially weighted and ordered equidistant placement of audio objects to the available downmix channels).

5 - estimated and applied (at the encoder) and restored (at the decoder) using particular optimization criterion allowing "flexible mixing" of input audio objects (i.e. generation of the downmix matrix that is optimized for the parametric estimation of the audio objects at the decoder side). For example, the encoder generates the downmix matrix in a way to make the parametric upmix more efficient, in terms of 10 special signal property reconstruction, like covariance, inter-signal correlation or improve/ensure numerical stability of the parametric upmix algorithm.

The provided embodiments can be applied on an arbitrary number of downmix / upmix channels. It can be combined with any current and also future audio formats.

15

The flexibility of the inventive method allows bypassing of unaltered channels to reduce computational complexity, reduce bitstream payload / reduced data amount.

An audio encoder, method or computer program for encoding is provided. Moreover, an 20 audio decoder, method or computer program for decoding is provided. Furthermore, an encoded signal is provided.

Although some aspects have been described in the context of an apparatus, it is clear that these aspects also represent a description of the corresponding method, where a block or 25 device corresponds to a method step or a feature of a method step. Analogously, aspects described in the context of a method step also represent a description of a corresponding block or item or feature of a corresponding apparatus.

The inventive decomposed signal can be stored on a digital storage medium or can be 30 transmitted on a transmission medium such as a wireless transmission medium or a wired transmission medium such as the Internet.

Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software. The implementation can be performed using a 35 digital storage medium, for example a floppy disk, a DVD, a CD, a ROM, a PROM, an EPROM, an EEPROM or a FLASH memory, having electronically readable control signals stored thereon, which cooperate (or are capable of cooperating) with a programmable computer system such that the respective method is performed.

2016234987   29 Sep 2016

Some embodiments according to the invention comprise a non-transitory data carrier having electronically readable control signals, which are capable of cooperating with a programmable computer system, such that one of the methods described herein is performed.

Generally, embodiments of the present invention can be implemented as a computer program product with a program code, the program code being operative for performing one of the methods when the computer program product runs on a computer. The program code may for example be stored on a machine readable carrier.

Other embodiments comprise the computer program for performing one of the methods described herein, stored on a machine readable carrier.

In other words, an embodiment of the inventive method is, therefore, a computer program having a program code for performing one of the methods described herein, when the computer program runs on a computer.

A further embodiment of the inventive methods is, therefore, a data carrier (or a digital storage medium, or a computer-readable medium) comprising, recorded thereon, the computer program for performing one of the methods described herein.

A further embodiment of the inventive method is, therefore, a data stream or a sequence of signals representing the computer program for performing one of the methods described herein. The data stream or the sequence of signals may for example be configured to be transferred via a data communication connection, for example via the Internet.

A further embodiment comprises a processing means, for example a computer, or a programmable logic device, configured to or adapted to perform one of the methods described herein.

A further embodiment comprises a computer having installed thereon the computer program for performing one of the methods described herein.

In some embodiments, a programmable logic device (for example a field programmable gate array) may be used to perform some or all of the functionalities of the methods described herein. In some embodiments, a field programmable gate array may cooperate

with a microprocessor in order to perform one of the methods described herein. Generally, the methods are preferably performed by any hardware apparatus.

The above described embodiments are merely illustrative for the principles of the present invention. It is understood that modifications and variations of the arrangements and the details described herein will be apparent to others skilled in the art. It is the intent, therefore, to be limited only by the scope of the impending patent claims and not by the specific details presented by way of description and explanation of the embodiments herein.

Throughout this specification and the claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" or "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as, an acknowledgement or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

References

[MPS] ISO/IEC 23003-1:2007, MPEG-D (MPEG audio technologies), Part 1: MPEG Surround, 2007.

5

[BCC] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, Nov. 2003

[JSC] C. Faller, "Parametric Joint-Coding of Audio Sources", 120th AES Convention,
10   Paris, 2006

[SAOC1] J. Herre, S. Disch, J. Hilpert, O. Hellmuth: "From SAC To SAOC - Recent Developments in Parametric Coding of Spatial Audio", 22nd Regional UK AES Conference, Cambridge, UK, April 2007

15

[SAOC2] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers and W. Oomen: " Spatial Audio Object Coding (SAOC) – The Upcoming MPEG Standard on Parametric Object Based Audio Coding", 124th AES Convention, Amsterdam 2008

20

[SAOC] ISO/IEC, "MPEG audio technologies – Part 2: Spatial Audio Object Coding (SAOC)," ISO/IEC JTC1/SC29/WG11 (MPEG) International Standard 23003-2.

[ISS1] M. Parvaix and L. Girin: "Informed Source Separation of underdetermined
25   instantaneous Stereo Mixtures using Source Index Embedding", IEEE ICASSP, 2010

[ISS2] M. Parvaix, L. Girin, J.-M. Brossier: "A watermarking-based method for informed source separation of audio signals with a single sensor", IEEE Transactions on Audio, Speech and Language Processing, 2010

30

[ISS3] A. Liutkus and J. Pinel and R. Badeau and L. Girin and G. Richard: "Informed source separation through spectrogram coding and data embedding", Signal Processing Journal, 2011

35   [ISS4] A. Ozerov, A. Liutkus, R. Badeau, G. Richard: "Informed source separation: source coding meets source separation", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2011

[ISS5] Shuhua Zhang and Laurent Girin: "An Informed Source Separation System for Speech Signals", INTERSPEECH, 2011

[ISS6] L. Girin and J. Pinel: "Informed Audio Source Separation from Compressed Linear Stereo Mixtures", AES 42nd International Conference: Semantic Audio, 2011

5

The claims defining the invention are as follows:

1. A decoder for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels, wherein the downmix signal encodes two or more audio object signals, wherein the decoder comprises:

    a threshold determiner for determining a threshold value depending on a signal energy or a noise energy of at least one of the two or more audio object signals or depending on a signal energy or a noise energy of at least one of the one or more downmix channels, and

    a processing unit for generating the one or more audio output channels from the one or more downmix channels depending on the threshold value,

    wherein the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels depending on an object covariance matrix ($\mathbf{E}$) of the one or more audio object signals, depending on a downmix matrix ($\mathbf{D}$) for downmixing the two or more audio object signals to obtain the one or more downmix channels, and depending on the threshold value,

    wherein the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels by applying the threshold value in a function to inverse a downmix channel cross correlation matrix $\mathbf{Q}$,

    wherein $\mathbf{Q}$ is defined as $\mathbf{Q=DED}^{*}$,

    wherein $\mathbf{D}$ is the downmix matrix for downmixing the two or more audio object signals to obtain the two or more downmix channels,

    wherein $\mathbf{E}$ is the object covariance matrix of the one or more audio object signals, and

    wherein the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels by computing the eigenvalues of the downmix channel cross correlation matrix $\mathbf{Q}$.

2. A decoder according to claim 1,

wherein the downmix signal comprises two or more downmix channels, and

wherein the threshold determiner is configured to determine the threshold value depending on a noise energy of each of the two or more downmix channels.

3. A decoder according to claim 2, wherein the threshold determiner is configured to determine the threshold value depending on the sum of all noise energy in the two or more downmix channels.

4. A decoder according to any one of the preceding claims,

wherein the threshold determiner is configured to determine the threshold value depending on a signal energy of the audio object signal of the two or more audio object signals which has the greatest signal energy of the two or more audio object signals.

5. A decoder according to any one of the preceding claims,

wherein the downmix signal encodes the two or more audio object signals for each time-frequency tile of a plurality of time-frequency tiles,

wherein the threshold determiner is configured to determine a threshold value for each time-frequency tile of the plurality of time-frequency tiles depending on the signal energy or the noise energy of at least one of the two or more audio object signals or depending on the signal energy or the noise energy of at least one of the one or more downmix channels, wherein a first threshold value of a first time-frequency tile of the plurality of time-frequency tiles differs from a second time-frequency tile of the plurality of time-frequency tiles, and

wherein the processing unit is configured to generate for each time-frequency tile of the plurality of time-frequency tiles a channel value of each of the one or more audio output channels from the one or more downmix channels depending on the threshold value of said time-frequency tile.

6. A decoder according to any one of the preceding claims,

wherein the downmix signal comprises two or more downmix channels,

wherein the decoder is configured to determine the threshold value $\mathbf{T}$ in decibel according to the formula

$$\mathrm{T[dB]} = \mathbf{E}_{\mathrm{noise}}[\mathrm{dB}] - \mathbf{E}_{\mathrm{ref}}[\mathrm{dB}] - \mathbf{Z} \quad \text{or according to the formula}$$

$$\mathrm{T[dB]} = \mathbf{E}_{\mathrm{noise}}[\mathrm{dB}] - \mathbf{E}_{\mathrm{ref}}[\mathrm{dB}],$$

wherein $\mathbf{T}[\mathrm{dB}]$ indicates the threshold value in decibel,

wherein $\mathbf{E}_{\mathrm{noise}}[\mathrm{dB}]$ indicates the sum of all noise energy in the two or more downmix channels in decibel, or $\mathbf{E}_{\mathrm{noise}}[\mathrm{dB}]$ indicates the sum of all noise energy in the two or more downmix channels in decibel divided by the number of the two or more downmix channels,

wherein $\mathbf{E}_{\mathrm{ref}}[\mathrm{dB}]$ indicates the signal energy of one of the audio object signals in decibel, and

wherein $\mathbf{Z}$ indicates an additional parameter being a number.

7.   A decoder according to any one of claims 1 to 5,

wherein the downmix signal comprises two or more downmix channels,

wherein the decoder is configured to determine the threshold value $\mathbf{T}$ according to the formula

$$\mathbf{T} = \frac{\mathbf{E}_{\mathrm{noise}}}{\mathbf{E}_{\mathrm{ref}} \cdot \mathbf{Z}} \quad \text{or according to the formula}$$

$$\mathbf{T} = \frac{\mathbf{E}_{\mathrm{noise}}}{\mathbf{E}_{\mathrm{ref}}},$$

wherein $\mathbf{T}$ indicates the threshold value,

wherein $\mathbf{E}_{\mathrm{noise}}$ indicates the sum of all noise energy in the two or more downmix channels, or $\mathbf{E}_{\mathrm{noise}}$ in decibel indicates the sum of all noise energy in the two or

more downmix channels in decibel divided by the number of the two or more downmix channels,

wherein E_{ref} indicates the signal energy of one of the audio object signals, and

wherein Z indicates an additional parameter being a number.

8. A decoder according to any one of the preceding claims, wherein the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels by multiplying the largest eigenvalue of the eigenvalues of the downmix channel cross correlation matrix Q with the threshold value to obtain a relative threshold.

9. A decoder according to claim 8,

wherein the processing unit is configured to generate the one or more audio output channels from the one or more downmix channels by generating a modified matrix,

wherein the processing unit is configured to generate the modified matrix depending on only those eigenvectors of the downmix channel cross correlation matrix Q, which have an eigenvalue of the eigenvalues of the downmix channel cross correlation matrix Q, which is greater than or equal to the relative threshold,

wherein the processing unit is configured to conduct a matrix inversion of the modified matrix to obtain an inverted matrix, and

wherein the processing unit is configured to apply the inverted matrix on one or more of the downmix channels to generate the one or more audio output channels.

10. A method for generating an audio output signal comprising one or more audio output channels from a downmix signal comprising one or more downmix channels, wherein the downmix signal encodes two or more audio object signals, wherein the method comprises:

determining a threshold value depending on a signal energy or a noise energy of at least one of the two or more audio object signals or depending on a signal energy or a noise energy of at least one of the one or more downmix channels, and

generating the one or more audio output channels from the one or more downmix channels depending on the threshold value,

5   wherein generating the one or more audio output channels from the one or more downmix channels depending on an object covariance matrix (E) of the one or more audio object signals is conducted depending on a downmix matrix (D) for downmixing the two or more audio object signals to obtain the one or more downmix channels, and depending on the threshold value,

10

wherein generating the one or more audio output channels from the one or more downmix channels is conducted by applying the threshold value in a function to inverse a downmix channel cross correlation matrix Q,

15   wherein Q is defined as $Q=DED^*$,

wherein D is the downmix matrix for downmixing the two or more audio object signals to obtain the two or more downmix channels,

20   wherein E is the object covariance matrix of the one or more audio object signals, and

wherein generating the one or more audio output channels from the one or more downmix channels is conducted by computing the eigenvalues of the downmix

25   channel cross correlation matrix Q.

11.   A computer program for implementing the method of claim 10 when being executed on a computer or signal processor.
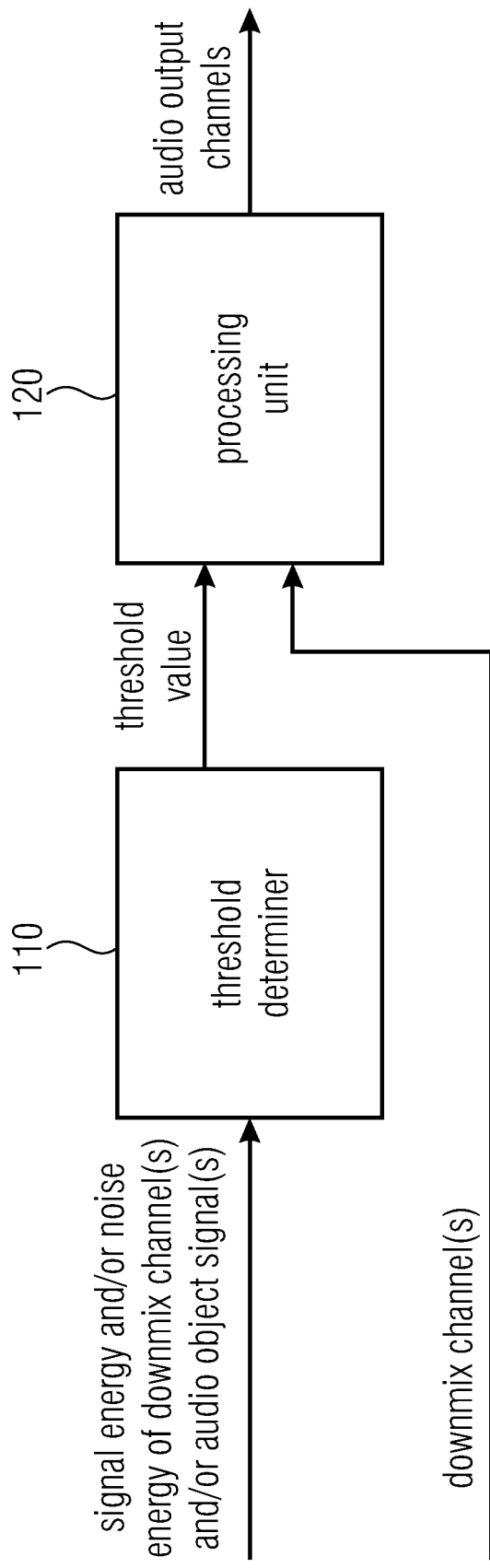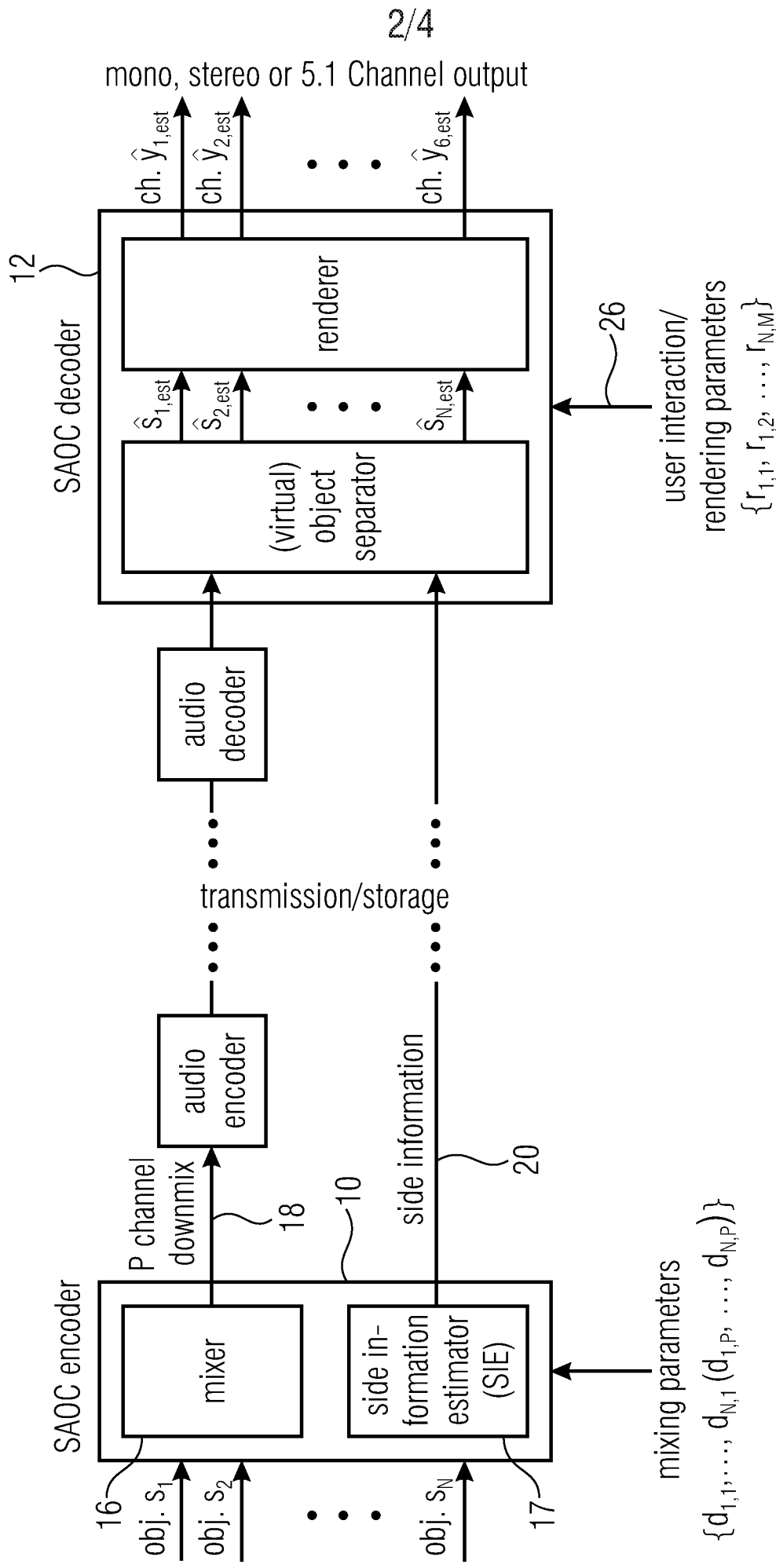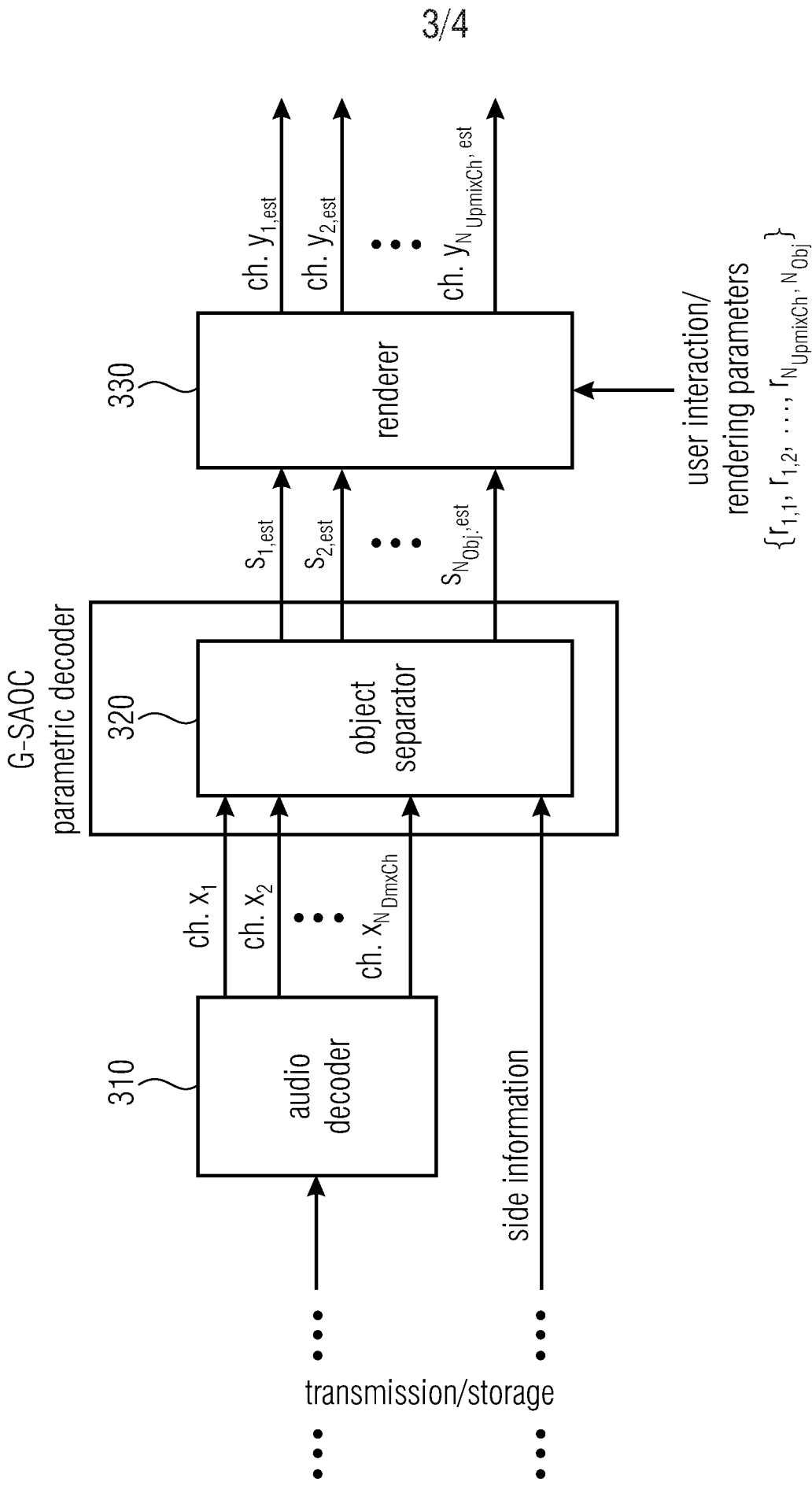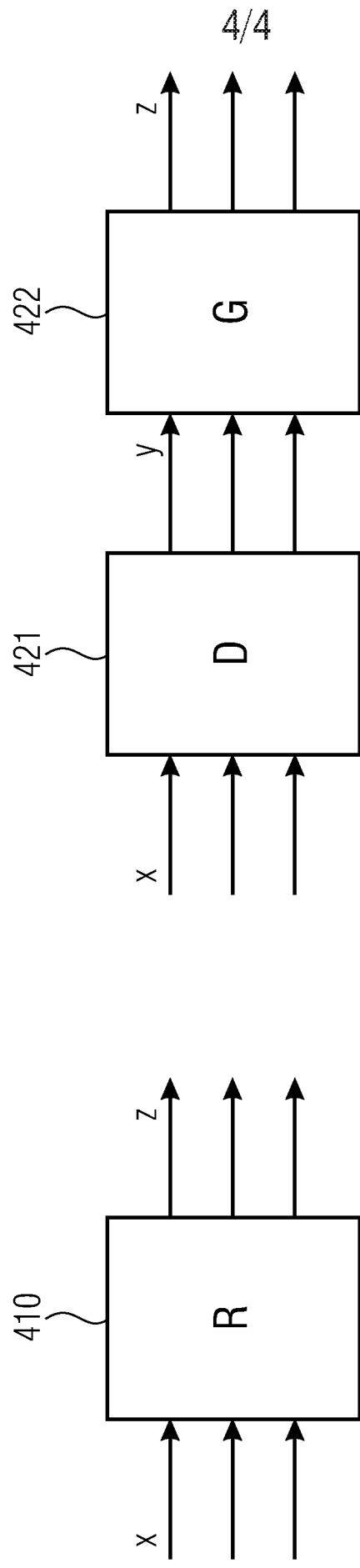
FIG 1

2016234987 29 Sep 2016

mono, stereo or 5.1 Channel output

ch. $\hat{y}_{1,est}$ ch. $\hat{y}_{2,est}$ • • •   ch. $\hat{y}_{6,est}$

SAOC decoder   12

renderer

$\hat{s}_{1,est}$ $\hat{s}_{2,est}$ • • • $\hat{s}_{N,est}$

(virtual) object separator

26

user interaction/
rendering parameters
$\{r_{1,1},\ r_{1,2},\ \ldots,\ r_{N,M}\}$

audio decoder

transmission/storage

audio encoder

P channel downmix   18

10

side information   20

SAOC encoder

mixer

side in-formation estimator (SIE)

16   obj. s_1
obj. s_2
• • •
obj. s_N
17

mixing parameters
$\{d_{1,1},\ \ldots,\ d_{N,1}\ (d_{1,P},\ \ldots,\ d_{N,P})\}$

FIG 2

FIG 3

4/4

x → | R (410) | → z

x → | D (421) | →y → | G (422) | → z

FIG 4