

**Audio Signal Decoder, Method for Decoding an Audio Signal and Computer Program
using Cascaded Audio Object Processing Stages**

Description

5

Technical Field

Embodiments according to the invention are related to an audio signal decoder for
providing an upmix signal representation in dependence on a downmix signal
10 representation and an object-related parametric information.

Further embodiments according to the invention are related to a method for providing an
upmix signal representation in dependence on a downmix signal representation and an
object-related parametric information.

15

Further embodiments according to the invention are related to a computer program.

Some embodiments according to the invention are related to an enhanced Karaoke/Solo
SAOC system.

20

Background of the Invention

In modern audio systems, it is desired to transfer and store audio information in a bitrate-
efficient way. In addition, it is often desired to reproduce an audio content using a plurality
25 of two or even more speakers, which are spatially distributed in a room. In such cases, it is
desired to exploit the capabilities of such a multi-speaker arrangement to allow for a user
to spatially identify different audio contents or different items of a single audio content.
This may be achieved by individually distributing the different audio contents to the
different speakers.

30

In other words, in the art of audio processing, audio transmission and audio storage, there
is an increasing desire to handle multi-channel contents in order to improve the hearing
impression. Usage of multi-channel audio content brings along significant improvements
for the user. For example, a 3-dimensional hearing impression can be obtained, which
35 brings along an improved user satisfaction in entertainment applications. However, multi-
channel audio contents are also useful in professional environments, for example in
telephone conferencing applications, because the speaker intelligibility can be improved by
using a multi-channel audio playback.

However, it is also desirable to have a good tradeoff between audio quality and bitrate requirements in order to avoid an excessive resource load caused by multi-channel applications.

5

Recently, parametric techniques for the bitrate-efficient transmission and/or storage of audio scenes containing multiple audio objects has been proposed, for example, Binaural Cue Coding (Type I) (see, for example reference [BCC]), Joint Source Coding (see, for example, reference [JSC]), and MPEG Spatial Audio Object Coding (SAOC) (see, for example, references [SAOC1], [SAOC2]).

10

These techniques aim at perceptually reconstructing the desired output audio scene rather than by a waveform match.

Fig. 8 shows a system overview of such a system (here: MPEG SAOC). The MPEG SAOC system 800 shown in Fig. 8 comprises an SAOC encoder 810 and an SAOC decoder 820. The SAOC encoder 810 receives a plurality of object signals x_1 to x_N , which may be represented, for example, as time-domain signals or as time-frequency-domain signals (for example, in the form of a set of transform coefficients of a Fourier-type transform, or in the form of QMF subband signals). The SAOC encoder 810 typically also receives downmix coefficients d_1 to d_N , which are associated with the object signals x_1 to x_N . Separate sets of downmix coefficients may be available for each channel of the downmix signal. The SAOC encoder 810 is typically configured to obtain a channel of the downmix signal by combining the object signals x_1 to x_N in accordance with the associated downmix coefficients d_1 to d_N . Typically, there are less downmix channels than object signals x_1 to x_N . In order to allow (at least approximately) for a separation (or separate treatment) of the object signals at the side of the SAOC decoder 820, the SAOC encoder 810 provides both the one or more downmix signals (designated as downmix channels) 812 and a side information 814. The side information 814 describes characteristics of the object signals x_1 to x_N , in order to allow for a decoder-sided object-specific processing.

20

25

30

The SAOC decoder 820 is configured to receive both the one or more downmix signals 812 and the side information 814. Also, the SAOC decoder 820 is typically configured to receive a user interaction information and/or a user control information 822, which describes a desired rendering setup. For example, the user interaction information/user control information 822 may describe a speaker setup and the desired spatial placement of the objects provided by the object signals x_1 to x_N .

35

The SAOC decoder 820 is configured to provide, for example, a plurality of decoded upmix channel signals $\hat{y}_1$ to $\hat{y}_M$. The upmix channel signals may for example be associated with individual speakers of a multi-speaker rendering arrangement. The SAOC decoder 820 may, for example, comprise an object separator 820a, which is configured to reconstruct, at least approximately, the object signals x_1 to x_N on the basis of the one or more downmix signals 812 and the side information 814, thereby obtaining reconstructed object signals 820b. However, the reconstructed object signals 820b may deviate somewhat from the original object signals x_1 to x_N , for example, because the side information 814 is not quite sufficient for a perfect reconstruction due to the bitrate constraints. The SAOC decoder 820 may further comprise a mixer 820c, which may be configured to receive the reconstructed object signals 820b and the user interaction information/user control information 822, and to provide, on the basis thereof, the upmix channel signals $\hat{y}_1$ to $\hat{y}_M$. The mixer 820c may be configured to use the user interaction information /user control information 822 to determine the contribution of the individual reconstructed object signals 820b to the upmix channel signals $\hat{y}_1$ to $\hat{y}_M$. The user interaction information/user control information 822 may, for example, comprise rendering parameters (also designated as rendering coefficients), which determine the contribution of the individual reconstructed object signals 820b to the upmix channel signals $\hat{y}_1$ to $\hat{y}_M$.

However, it should be noted that in many embodiments, the object separation, which is indicated by the object separator 820a in Fig. 8, and the mixing, which is indicated by the mixer 820c in Fig. 8, are performed in one single step. For this purpose, overall parameters may be computed which describe a direct mapping of the one or more downmix signals 812 onto the upmix channel signals $\hat{y}_1$ to $\hat{y}_M$. These parameters may be computed on the basis of the side information 814 and the user interaction information/user control information 822.

Taking reference now to Figs. 9a, 9b and 9c, different apparatus for obtaining an upmix signal representation on the basis of a downmix signal representation and object-related side information will be described. Fig. 9a shows a block schematic diagram of an MPEG SAOC system 900 comprising an SAOC decoder 920. The SAOC decoder 920 comprises, as separate functional blocks, an object decoder 922 and a mixer/renderer 926. The object decoder 922 provides a plurality of reconstructed object signals 924 in dependence on the downmix signal representation (for example, in the form of one or more downmix signals represented in the time domain or in the time-frequency-domain) and object-related side information (for example, in the form of object meta data). The mixer/renderer 926 receives the reconstructed object signals 924 associated with a plurality of N objects and provides, on the basis thereof, one or more upmix channel signals 928. In the SAOC

decoder 920, the extraction of the object signals 924 is performed separately from the mixing/rendering which allows for a separation of the object decoding functionality from the mixing/rendering functionality but brings along a relatively high computational complexity.

5

Taking reference now to Fig. 9b, another MPEG SAOC system 930 will be briefly discussed, which comprises an SAOC decoder 950. The SAOC decoder 950 provides a plurality of upmix channel signals 958 in dependence on a downmix signal representation (for example, in the form of one or more downmix signals) and an object-related side information (for example, in the form of object meta data). The SAOC decoder 950 comprises a combined object decoder and mixer/renderer, which is configured to obtain the upmix channel signals 958 in a joint mixing process without a separation of the object decoding and the mixing/rendering, wherein the parameters for said joint upmix process are dependent on both, the object-related side information and the rendering information. The joint upmix process also depends on the downmix information, which is considered to be part of the object-related side information.

20

To summarize the above, the provision of the upmix channel signals 928, 958 can be performed in a one step process or a two-step process.

Taking reference now to Fig. 9c, an MPEG SAOC system 960 will be described. The SAOC system 960 comprises an SAOC to MPEG Surround transcoder 980, rather than an SAOC decoder.

The SAOC to MPEG Surround transcoder comprises a side information transcoder 982, which is configured to receive the object-related side information (for example, in the form of object meta data) and, optionally, information on the one or more downmix signals and the rendering information. The side information transcoder is also configured to provide an MPEG Surround side information 984 (for example, in the form of an MPEG Surround bitstream) on the basis of a received data. Accordingly, the side information transcoder 982 is configured to transform an object-related (parametric) side information, which is relieved from the object encoder, into a channel-related (parametric) side information 984, taking into consideration the rendering information and, optionally, the information about the content of the one or more downmix signals.

35

Optionally, the SAOC to MPEG Surround transcoder 980 may be configured to manipulate the one or more downmix signals, described, for example, by the downmix signal representation, to obtain a manipulated downmix signal representation 988. However, the

downmix signal manipulator 986 may be omitted, such that the output downmix signal representation 988 of the SAOC to MPEG Surround transcoder 980 is identical to the input downmix signal representation of the SAOC to MPEG Surround transcoder. The downmix signal manipulator 986 may, for example, be used if the channel-related MPEG Surround side information 984 would not allow to provide a desired hearing impression on the basis of the input downmix signal representation of the SAOC to MPEG Surround transcoder 980, which may be the case in some rendering constellations.

Accordingly, the SAOC to MPEG Surround transcoder 980 provides the downmix signal representation 988 and the MPEG Surround bitstream 984 such that a plurality of upmix channel signals, which represent the audio objects in accordance with the rendering information input to the SAOC to MPEG Surround transcoder 980 can be generated using an MPEG Surround decoder which receives the MPEG Surround bitstream 984 and the downmix signal representation 988.

To summarize the above, different concepts for decoding SAOC-encoded audio signals can be used. In some cases, an SAOC decoder is used, which provides upmix channel signals (for example, upmix channel signals 928, 958) in dependence on the downmix signal representation and the object-related parametric side information. Examples for this concept can be seen in Figs. 9a and 9b. Alternatively, the SAOC-encoded audio information may be transcoded to obtain a downmix signal representation (for example, a downmix signal representation 988) and a channel-related side information (for example, the channel-related MPEG Surround bitstream 984), which can be used by an MPEG Surround decoder to provide the desired upmix channel signals.

In the MPEG SAOC system 800, a system overview of which is given in Fig. 8, the general processing is carried out in a frequency selective way and can be described as follows within each frequency band:

- N input audio object signals x_1 to x_N are downmixed as part of the SAOC encoder processing. For a mono downmix, the downmix coefficients are denoted by d_1 to d_N . In addition, the SAOC encoder 810 extracts side information 814 describing the characteristics of the input audio objects. For MPEG SAOC, the relations of the object powers with respect to each other are the most basic form of such a side information.
- Downmix signal (or signals) 812 and side information 814 are transmitted and/or stored. To this end, the downmix audio signal may be compressed using well-known

perceptual audio coders such as MPEG-1 Layer II or III (also known as “.mp3”), MPEG Advanced Audio Coding (AAC), or any other audio coder.

- 5 • On the receiving end, the SAOC decoder 820 conceptually tries to restore the original object signal (“object separation”) using the transmitted side information 814 (and, naturally, the one or more downmix signals 812). These approximated object signals (also designated as reconstructed object signals 820b) are then mixed into a target scene represented by M audio output channels (which may, for example, be represented by the upmix channel signals $\hat{y}_1$ to $\hat{y}_M$) using a rendering matrix. For a mono output, the rendering matrix coefficients are given by r_1 to r_N .
- 10 • Effectively, the separation of the object signals is rarely executed (or even never executed), since both the separation step (indicated by the object separator 820a) and the mixing step (indicated by the mixer 820c) are combined into a single transcoding step, which often results in an enormous reduction in computational complexity.
- 15

It has been found that such a scheme is tremendously efficient, both in terms of transmission bitrate (it is only necessary to transmit a few downmix channels plus some side information instead of N discrete object audio signals or a discrete system) and
20 computational complexity (the processing complexity relates mainly to the number of output channels rather than the number of audio objects). Further advantages for the user on the receiving end include the freedom of choosing a rendering setup of his/her choice (mono, stereo, surround, virtualized headphone playback, and so on) and the feature of user interactivity: the rendering matrix, and thus the output scene, can be set and changed
25 interactively by the user according to will, personal preference or other criteria. For example, it is possible to locate the talkers from one group together in one spatial area to maximize discrimination from other remaining talkers. This interactivity is achieved by providing a decoder user interface.

30 For each transmitted sound object, its relative level and (for non-mono rendering) spatial position of rendering can be adjusted. This may happen in real-time as the user changes the position of the associated graphical user interface (GUI) sliders (for example: object level = +5dB, object position = -30deg).

35 However, it has been found that it is difficult to handle audio objects of different audio object types in such a system. In particular, it has been found that it is difficult to process audio objects of different audio object types, for example, audio objects to which different

side information is associated, if the total number of audio objects to be processed is not predetermined.

In view of this situation, it is an objective of the present invention to create a concept, which allows for a computationally-efficient and flexible decoding of an audio signal comprising a downmix signal representation and an object-related parametric information, wherein the object-related parametric information describes audio objects of two or more different audio object types.

10 Summary of the Invention

This objective is achieved by an audio signal decoder for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information, a method for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information, and a computer program, as defined by the independent claims.

An embodiment according to the invention creates an audio signal decoder for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information. The audio signal decoder comprises an object separator configured to decompose the downmix signal representation, to provide a first audio information describing a first set of one or more audio objects of a first audio object type and a second audio information describing a second set of one or more audio objects of a second audio object type in dependence on the downmix signal representation and using at least a part of the object-related parametric information. The audio signal decoder also comprises an audio signal processor configured to receive the second audio information and to process the second audio information in dependence on the object-related parametric information, to obtain a processed version of the second audio information. The audio signal decoder also comprises an audio signal combiner configured to combine the first audio information with the processed version of the second audio information to obtain the upmix signal representation.

It is a key idea of the present invention that an efficient processing of different types of audio objects can be obtained in a cascaded structure, which allows for a separation of the different types of audio objects using at least a part of the object-related parametric information in a first processing step performed by the object separator, and which allows for an additional spatial processing in a second processing step performed in dependence on at least a part of the object-related parametric information by the audio signal processor.

It has been found that extracting a second audio information, which comprises audio objects of the second audio object type, from a downmix signal representation can be performed with a moderate complexity even if there is a larger number of audio objects of the second audio object type. In addition, it has been found that a spatial processing of the audio objects of the second audio type can be performed efficiently once the second audio information is separated from the first audio information describing the audio objects of the first audio object type.

Additionally, it has been found that the processing algorithm performed by the object separator for separating the first audio information and the second audio information can be performed with comparatively small complexity if the object-individual processing of the audio objects of the second audio object type is postponed to the audio signal processor and not performed at the same time as the separation of the first audio information and the second audio information.

In a preferred embodiment, the audio signal decoder is configured to provide the upmix signal representation in dependence on the downmix signal representation, the object-related parametric information and a residual information associated to a sub-set of audio objects represented by the downmix signal representation. In this case, the object separator is configured to decompose the downmix signal representation to provide the first audio information describing the first set of one or more audio objects (for example, foreground objects FGO) of the first audio object type to which residual information is associated and the second audio information describing the second set of one or more audio objects (for example, background objects BGO) of the second audio object type to which no residual information is associated in dependence on the downmix signal representation and using at least part of the object-related parametric information and the residual information.

This embodiment is based on the finding that a particularly accurate separation between the first audio information describing the first set of audio objects of the first audio object type and the second audio information describing the second set of audio objects of the second audio object type can be obtained by using a residual information in addition to the object-related parametric information. It has been found that the mere use of the object-related parametric information would result in distortions in many cases, which can be reduced significantly or even entirely eliminated by the use of residual information. The residual information describes, for example, a residual distortion, which is expected to remain if an audio object of the first audio object type is isolated merely using the object-related parametric information. The residual information is typically estimated by an audio signal encoder. By applying the residual information, the separation between the audio

objects of the first audio object type and the audio objects of the second audio object type can be improved.

This allows to obtain the first audio information and the second audio information with particularly good separation between the audio objects of the first audio object type and the audio objects of the second audio object type, which, in turn, allows to achieve a high-quality spatial processing of the audio objects of the second audio object type when processing the second audio information in the audio signal processor.

In a preferred embodiment, the object separator is therefore configured to provide the first audio information such that audio objects of the first audio object type are emphasized over audio objects of the second audio object type in the first audio information. The object separator is also configured to provide the second audio information such that audio objects of the second audio object type are emphasized over audio objects of the first audio object type in the second audio information.

In a preferred embodiment, the audio signal decoder is configured to perform a two-step processing, such that a processing of the second audio information in the audio signal processor is performed subsequently to a separation between the first audio information describing the first set of one or more audio objects of the first audio object type and the second audio information describing the second set of one or more audio objects of the second audio object type.

In a preferred embodiment, the audio signal processor is configured to process the second audio information in dependence on the object-related parametric information associated with the audio objects of the second audio object type and independent from the object-related parametric information associated with the audio objects of the first audio object type. Accordingly, a separate processing of the audio objects of the first audio object type and the audio objects of the second audio object type can be obtained.

In a preferred embodiment, the object separator is configured to obtain the first audio information and the second audio information using a linear combination of one or more downmix channels and one or more residual channels. In this case, the object separator is configured to obtain combination parameters for performing the linear combination in dependence on downmix parameters associated with the audio objects of the first audio object type and in dependence on channel prediction coefficients of the audio objects of the first audio object type. The computation of the channel prediction coefficients of the audio objects of the first audio object type may, for example, take into consideration the audio

objects of the second audio object type as a single, common audio object. Accordingly, a separation process can be performed with sufficiently small computational complexity, which may, for example, be almost independent from the number of audio objects of the second audio object type.

5

In a preferred embodiment, the object separator is configured to apply a rendering matrix to the first audio information to map object signals of the first audio information onto audio channels of the upmix audio signal representation. This can be done, because the object separator may be capable of extracting separate audio signals individually
10 representing the audio objects of the first audio object type. Accordingly, it is possible to map the object signals of the first audio information directly onto the audio channels of the upmix audio signal representation.

In a preferred embodiment, the audio processor is configured to perform a stereo
15 processing of the second audio information in dependence on a rendering information, an object-related covariance information and a downmix information, to obtain audio channels of the upmix audio signal representation.

Accordingly, the stereo processing of the audio objects of the second audio object type is
20 separated from the separation between the audio objects of the first audio object type and the audio objects of the second audio object type. Thus, the efficient separation between audio objects of the first audio object type and audio objects of the second audio object type is not affected (or degraded) by the stereo processing, which typically leads to a distribution of audio objects over a plurality of audio channels without providing the high
25 degree of object separation, which can be obtained in the object separator, for example, using the residual information.

In another preferred embodiment, the audio processor is configured to perform a post-
processing of the second audio information in dependence on a rendering information, an
30 object-related covariance information and a downmix information. This form of post-processing allows for a spatial placement of the audio objects of the second audio object type within an audio scene. Nevertheless, due to the cascaded concept, the computational complexity of the audio processor can be kept sufficiently small, because the audio processor does not need to consider the object-related parametric information associated
35 with the audio objects of the first audio object type.

In addition, different types of processing can be performed by the audio processor, like, for example, a mono-to-binaural processing, a mono-to-stereo processing, a stereo-to-binaural processing or a stereo-to-stereo processing.

5 In a preferred embodiment, the object separator is configured to treat audio objects of the second audio object type, to which no residual information is associated, as a single audio object. In addition, the audio signal processor is configured to consider object-specific rendering parameters to adjust contributions of the objects of the second audio object type to the upmix signal representation. Thus, the audio objects of the second audio object type
10 are considered as a single audio object by the object separator, which significantly reduces the complexity of the object separator and also allows to have a unique residual information, which is independent from the rendering parameters associated with the audio objects of the second audio object type.

15 In a preferred embodiment, the object separator is configured to obtain a common object-level difference value for a plurality of audio objects of the second audio object type. The object separator is configured to use the common object-level difference value for a computation of channel prediction coefficients. In addition, the object separator is configured to use the channel prediction coefficients to obtain one or two audio channels
20 representing the second audio information. For obtaining a common object-level difference value, the audio objects of the second audio object type can be handled efficiently as a single audio object by the object separator.

In a preferred embodiment, the object separator is configured to obtain a common object
25 level difference value for a plurality of audio objects of the second audio object type and the object separator is configured to use the common object-level difference value for a computation of entries of an energy-mode mapping matrix. The object separator is configured to use the energy-mode mapping matrix to obtain the one or more audio channels representing the second audio information. Again, the common object level
30 difference value allows for a computationally efficient common treating of the audio objects of the second audio object type by the object separator.

In a preferred embodiment, the object separator is configured to selectively obtain a common inter-object correlation value associated to the audio objects of the second audio
35 object type in dependence on the object-related parametric information if it is found that there are two audio objects of the second audio object type and to set the inter-object correlation value associated to the audio objects of the second audio object type to zero if it is found that there are more or less than two audio objects of the second audio object type.

The object separator is configured to use the common inter-object correlation value associated to the audio objects of the second audio object type to obtain the one or more audio channels representing the second audio information. Using this approach, the inter-object correlation value is exploited if it is obtainable with high computational efficiency, i.e. if there are two audio objects of the second audio object type. Otherwise, it would be computationally demanding to obtain inter-object correlation values. Accordingly, it has been found to be a good compromise in terms of hearing impression and computational complexity to set the inter-object correlation value associated to the audio objects of the second audio object type to zero if there are more or less than two audio objects of the second object type.

In a preferred embodiment, the audio signal processor is configured to render the second audio information in dependence on (at least a part of) the object-related parametric information, to obtain a rendered representation of the audio objects of the second audio object type as a processed version of the second audio information. In this case, the rendering can be made independent from the audio objects of the first audio object type.

In a preferred embodiment, the object separator is configured to provide the second audio information such that the second audio information describes more than two audio objects of the second audio object type. Embodiments according to the invention allow for a flexible adjustment of the number of audio objects of the second audio object type, which is significantly facilitated by the cascaded structure of the processing.

In a preferred embodiment, the object separator is configured to obtain, as the second audio information, a one-channel audio signal representation or a two-channel audio signal representation representing more than two audio objects of the second audio object type. Extracting one or two audio signal channels can be performed by the object separator with low computational complexity. In particular, the complexity of the object separator can be kept significantly smaller when compared to a case in which the object separator would need to deal with more than two audio objects of the second audio object type. Nevertheless, it has been found that it is a computationally efficient representation of the audio objects of the second audio object type to use one or two channels of an audio signal.

In a preferred embodiment, the audio signal processor is configured to receive the second audio information and to process the second audio information in dependence on (at least a part of) the object-related parametric information, taking into consideration object-related parametric information associated with more than two audio objects of the second audio object type. Accordingly, an object-individual processing is performed by the audio

processor, while such an object-individual processing is not performed for audio objects of the second audio object type by the object separator.

5 In a preferred embodiment, the audio decoder is configured to extract a total object number information and a foreground object number information from a configuration information related to the object-related parametric information. The audio decoder is also configured to determine a number of audio objects of the second audio object type by forming a difference between the total object number information and the foreground object number information. Accordingly, efficient signalling of the number of audio objects of the second
10 audio object type is achieved. In addition, this concept provides for a high degree of flexibility regarding the number of audio objects of the second audio object type.

In a preferred embodiment, the object separator is configured to use object-related parametric information associated with N_{eao} audio objects of the first audio object type to
15 obtain, as the first audio information, N_{eao} audio signals representing (preferably, individually) the N_{eao} audio objects of the first audio object type, and to obtain, as the second audio information, one or two audio signals representing the $N - N_{\text{eao}}$ audio objects of the second audio object type, treating the $N - N_{\text{eao}}$ audio objects of the second audio object type as a single one-channel or two-channel audio object. The audio signal
20 processor is configured to individually render the $N - N_{\text{eao}}$ audio objects represented by the one or two audio signals of the second audio information using the object-related parametric information associated with the $N - N_{\text{eao}}$ audio objects of the second audio object type. Accordingly, the audio object separation between the audio objects of the first audio object type and the audio objects of the second audio object type is separated from the
25 subsequent processing of the audio objects of the second audio object type.

An embodiment according to the invention creates a method for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information.
30

Another embodiment according to the invention creates a computer program for performing said method.

Brief Description of the Figs.

35 Embodiments according to the invention will subsequently be described taking reference to the enclosed Figs., in which:

- Fig. 1 shows a block schematic diagram of an audio signal decoder, according to an embodiment of the invention;
- 5 Fig. 2 shows a block schematic diagram of another audio signal decoder, according to an embodiment of the invention;
- Figs. 3a and 3b show a block schematic diagrams of a residual processor, which can be used as an object separator in an embodiment of the invention;
- 10 Figs. 4a to 4e show block schematic diagrams of audio signal processors, which can be used in an audio signal decoder according to an embodiment of the invention:
- 15 Fig. 4f shows a block diagram of an SAOC transcoder processing mode;
- Fig. 4g shows a block diagram of an SAOC decoder processing mode;
- Fig. 5a shows a block schematic diagram of an audio signal decoder, according to an embodiment of the invention;
- 20 Fig. 5b shows a block schematic diagram of another audio signal decoder, according to an embodiment of the invention;
- Fig. 6a shows a Table representing a listening test design description;
- 25 Fig. 6b shows a Table representing systems under test;
- Fig. 6c shows a Table representing the listening test items and rendering matrices;
- 30 Fig. 6d shows a graphical representation of average MUSHRA scores for a Karaoke/Solo type rendering listening test;
- Fig. 6e shows a graphical representation of average MUSHRA scores for a classic rendering listening test;
- 35 Fig. 7 shows a flow chart of a method for providing an upmix signal representation, according to an embodiment of the invention;

Fig. 8 shows a block schematic diagram of a reference MPEG SAOC system;

Fig. 9a shows a block schematic diagram of a reference SAOC system using a separate decoder and mixer;

5

Fig. 9b shows a block schematic diagram of a reference SAOC system using an integrated decoder and mixer; and

Fig. 9c shows a block schematic diagram of a reference SAOC system using an SAOC-to-MPEG transcoder.

10

Detailed Description of the Embodiments

1. Audio signal decoder according to Fig. 1

15

Fig. 1 shows a block schematic diagram of an audio signal decoder 100 according to an embodiment of the invention.

20

The audio signal decoder 100 is configured to receive an object-related parametric information 110 and a downmix signal representation 112. The audio signal decoder 100 is configured to provide an upmix signal representation 120 in dependence on the downmix signal representation and the object-related parametric information 110. The audio signal decoder 100 comprises an object separator 130, which is configured to decompose the downmix signal representation 112 to provide a first audio information 132 describing a first set of one or more audio objects of a first audio object type and a second audio information 134 describing a second set of one or more audio objects of a second audio object type in dependence on the downmix signal representation 112 and using at least a part of the object-related parametric information 110. The audio signal decoder 100 also comprises an audio signal processor 140, which is configured to receive the second audio information 134 and to process the second audio information in dependence on at least a part of the object-related parametric information 112, to obtain a processed version 142 of the second audio information 134. The audio signal decoder 100 also comprises an audio signal combiner 150 configured to combine the first audio information 132 with the processed version 142 of the second audio information 134, to obtain the upmix signal representation 120.

25

30

35

The audio signal decoder 100 implements a cascaded processing of the downmix signal representation, which represents audio objects of the first audio object type and audio objects of the second audio object type in a combined manner.

5 In a first processing step, which is performed by the object separator 130, the second audio information describing a second set of audio objects of the second audio object type is separated from the first audio information 132 describing a first set of audio objects of a first audio object type using the object-related parametric information 110. However, the
10 second audio information 134 is typically an audio information (for example, a one-channel audio signal or a two-channel audio signal) describing the audio objects of the second audio object type in a combined manner.

In the second processing step, the audio signal processor 140 processes the second audio information 134 in dependence on the object-related parametric information. Accordingly,
15 the audio signal processor 140 is capable of performing an object-individual processing or rendering of the audio objects of the second audio object type, which are described by the second audio information 134, and which is typically not performed by the object separator 130.

20 Thus, while the audio objects of the second audio object type are preferably not processed in an object-individual manner by the object separator 130, the audio objects of the second audio object type are, indeed, processed in an object-individual manner (for example, rendered in an object-individual manner) in the second processing step, which is performed by the audio signal processor 140. Thus, the separation between the audio objects of the
25 first audio object type and the audio objects of the second audio object type, which is performed by the object separator 130, is separated from the object-individual processing of the audio objects of the second audio object type, which is performed afterwards by the audio signal processor 140. Accordingly, the processing which is performed by the object separator 130 is substantially independent from a number of audio objects of the second
30 audio object type. In addition, the format (for example, one-channel audio signal or the two-channel audio signal) of the second audio information 134 is typically independent from the number of audio objects of the second audio object type. Thus, the number of audio objects of the second audio object type can be varied without having the need to modify the structure of the object separator 130. In other words, the audio objects of the
35 second audio object type are treated as a single (for example, one-channel or two-channel) audio object for which a common object-related parametric information (for example, a common object-level-difference value associated with one or two audio channels) is obtained by the object separator 140.

Accordingly, the audio signal decoder 100 according to Fig. 1 is capable to handle a variable number of audio objects of the second audio object type without a structural modification of the object separator 130. In addition, different audio object processing algorithms can be applied by the object separator 130 and the audio signal processor 140. Accordingly, for example, it is possible to perform an audio object separation using a residual information by the object separator 130, which allows for a particularly good separation of different audio objects, making use of the residual information, which constitutes a side information for improving the quality of an object separation. In contrast, the audio signal processor 140 may perform an object-individual processing without using a residual information. For example, the audio signal processor 140 may be configured to perform a conventional spatial-audio-object-coding (SAOC) type audio signal processing to render the different audio objects.

2. Audio Signal Decoder according to Fig. 2

In the following, an audio signal decoder 200 according to an embodiment of the invention will be described. A block-schematic diagram of this audio signal decoder 200 shown in Fig. 2.

The audio decoder 200 is configured to receive a downmix signal 210, a so-called SAOC bitstream 212, rendering matrix information 214 and, optionally, head-related-transfer-function (HRTF) parameters 216. The audio signal decoder 200 is also configured to provide an output/MPS downmix signal 220 and (optionally) a MPS bitstream 222.

2.1. Input signals and output signals of the audio signal decoder 200

In the following, various details regarding input signals and output signals of the audio decoder 200 will be described.

The downmix signal 200 may, for example, be a one-channel audio signal or a two-channel audio signal. The downmix signal 210 may, for example, be derived from an encoded representation of the downmix signal.

The spatial-audio-object-coding bitstream (SAOC bitstream) 212 may, for example, comprise object-related parametric information. For example, the SAOC bitstream 212 may comprise object-level-difference information, for example, in the form of object-level-

difference parameters OLD, an inter-object-correlation information, for example, in the form of inter-object-correlation parameters IOC.

In addition, the SAOC bitstream 212 may comprise a downmix information describing how the downmix signals have been provided on the basis of a plurality of audio object signals using a downmix process. For example, the SAOC bitstream may comprise a downmix gain parameter DMG and (optionally) downmix-channel-level difference parameters DCLD.

The rendering matrix information 214 may, for example, describe how the different audio objects should be rendered by the audio decoder. For example, the rendering matrix information 214 may describe an allocation of an audio object to one or more channels of the output/MPS downmix signal 220.

The optional head-related-transfer-function (HRTF) parameter information 216 may further describe a transfer function for deriving a binaural headphone signal.

The output/MPEG-Surround downmix signal (also briefly designated with “output/MPS downmix signal”) 220 represents one or more audio channels, for example, in the form of a time domain audio signal representation or a frequency-domain audio signal representation. Alone or in combination with the optional MPEG-Surround bitstream (MPS bitstream) 222, which comprises MPEG-Surround parameters describing a mapping of the output/MPS downmix signal 220 onto a plurality of audio channels, an upmix signal representation is formed.

2.2. Structure and functionality of the audio signal decoder 200

In the following, the structure of the audio signal decoder 200, which may fulfill the functionality of an SAOC transcoder or the functionality of a SAOC decoder, will be described in more detail.

The audio signal decoder 200 comprises a downmix processor 230, which is configured to receive the downmix signal 210 and to provide, on the basis thereof, the output/MPS downmix signal 220. The downmix processor 230 is also configured to receive at least a part of the SAOC bitstream information 212 and at least a part of the rendering matrix information 214. In addition, the downmix processor 230 may also receive a processed SAOC parameter information 240 from a parameter processor 250.

The parameter processor 250 is configured to receive the SAOC bitstream information 212, the rendering matrix information 214 and, optionally, the head-related-transfer-function parameter information 260, and to provide, on the basis thereof, the MPEG Surround bitstream 222 carrying the MPEG surround parameters (if the MPEG surround parameters are required, which is, for example, true in the transcoding mode of operation). In addition, the parameter processor 250 provides the processed SAOC information 240 (if this processed SAOC information is required).

In the following, the structure and functionality of the downmix processor 230 will be described in more detail.

The downmix processor 230 comprises a residual processor 260, which is configured to receive the downmix signal 210 and to provide, on the basis thereof, a first audio object signal 262 describing so-called enhanced audio objects (EAOs), which may be considered as audio objects of a first audio object type. The first audio object signal may comprise one or more audio channels and may be considered as a first audio information. The residual processor 260 is also configured to provide a second audio object signal 264, which describes audio objects of a second audio object type and may be considered as a second audio information. The second audio object signal 264 may comprise one or more channels and may typically comprise one or two audio channels describing a plurality of audio objects. Typically, the second audio object signal may describe even more than two audio objects of the second audio object type.

The downmix processor 230 also comprises an SAOC downmix pre-processor 270, which is configured to receive the second audio object signal 264 and to provide, on the basis thereof, a processed version 272 of the second audio object signal 264, which may be considered as a processed version of the second audio information.

The downmix processor 230 also comprises an audio signal combiner 280, which is configured to receive the first audio object signal 262 and the processed version 272 of the second audio object signal 264, and to provide, on the basis thereof, the output/MPS downmix signal 220, which may be considered, alone or together with the (optional) corresponding MPEG-Surround bitstream 222, as an upmix signal representation.

In the following, the functionality of the individual units of the downmix processor 230 will be discussed in more detail.

The residual processor 260 is configured to separately provide the first audio object signal 262 and the second audio object signal 264. For this purpose, the residual processor 260 may be configured to apply at least a part of the SAOC bitstream information 212. For example, the residual processor 260 may be configured to evaluate an object-related parametric information associated with the audio objects of the first audio object type, i.e. the so-called “enhanced audio objects” EAO. In addition, the residual processor 260 may be configured to obtain an overall information describing the audio objects of the second audio object type, for example, the so-called “non-enhanced audio objects”, commonly.

The residual processor 260 may also be configured to evaluate a residual information, which is provided in the SAOC bitstream information 212, for a separation between enhanced audio objects (audio objects of the first audio object type) and non-enhanced audio objects (audio objects of the second audio object type). The residual information may, for example, encode a time domain residual signal, which is applied to obtain a particularly clean separation between the enhanced audio objects and the non-enhanced audio objects. In addition, the residual processor 260 may, optionally, evaluate at least a part of the rendering matrix information 214, for example, in order to determine a distribution of the enhanced audio objects to the audio channels of the first audio object signal 262.

The SAOC downmix pre-processor 270 comprises a channel re-distributor 274, which is configured to receive the one or more audio channels of the second audio object signal 264 and to provide, on the basis thereof, one or more (typically two) audio channels of the processed second audio object signal 272. In addition, the SAOC downmix pre-processor 270 comprises a decorrelated-signal-provider 276, which is configured to receive the one or more audio channels of the second audio object signal 264 and to provide, on the basis thereof, one or more decorrelated signals 278a, 278b, which are added to the signals provided by the channel re-distributor 274 in order to obtain the processed version 272 of the second audio object signal 264.

Further details regarding the SAOC downmix processor will be discussed below.

The audio signal combiner 280 combines the first audio object signal 262 with the processed version 272 of the second audio object signal. For this purpose, a channel-wise combination may be performed. Accordingly, the output/MPS downmix signal 220 is obtained.

The parameter processor 250 is configured to obtain the (optional) MPEG-Surround parameters, which make up the MPEG-Surround bitstream 222 of the upmix signal representation, on the basis of the SAOC bitstream, taking into consideration the rendering matrix information 214 and, optionally, the HRTF parameter information 216. In other words, the SAOC parameter processor 252 is configured to translate the object-related parameter information, which is described by the SAOC bitstream information 212, into a channel-related parametric information, which is described by the MPEG Surround bitstream 222.

In the following, a short overview of the structure of the SAOC transcoder/decoder architecture shown in Fig. 2 will be given. Spatial audio object coding (SAOC) is a parametric multiple object coding technique. It is designed to transmit a number of audio objects in an audio signal (for example the downmix audio signal 210) that comprises M channels. Together with this backward compatible downmix signal, object parameters are transmitted (for example, using the SAOC bitstream information 212) that allow for recreation and manipulation of the original object signals. An SAOC encoder (not shown here) produces a downmix of the object signals at its input and extracts these object parameters. The number of objects that can be handled is in principle not limited. The object parameters are quantized and coded efficiently into the SAOC bitstream 212. The downmix signal 210 can be compressed and transmitted without the need to update existing coders and infrastructures. The object parameters, or SAOC side information, are transmitted in a low bit rate side channel, for example, the ancillary data portion of the downmix bitstream.

On the decoder side, the input objects are reconstructed and rendered to a certain number of playback channels. The rendering information containing reproduction level and panning position for each object is user-supplied or can be extracted from the SAOC bitstream (for example, as a preset information). The rendering information can be time-variant. Output scenarios can range from mono to multi-channel (for example, 5.1) and are independent from both, the number of input objects and the number of downmix channels. Binaural rendering of objects is possible including azimuth and elevation of virtual object positions. An optional effect interface allows for advanced manipulation of object signals, besides level and panning modification.

The objects themselves can be mono signals, stereophonic signals, as well as a multi-channel signals (for example 5.1 channels). Typical downmix configurations are mono and stereo.

In the following, the basic structure of the SAOC transcoder/decoder, which is shown in Fig. 2, will be explained. The SAOC transcoder/decoder module described herein may act either as a stand-alone decoder or as a transcoder from an SAOC to an MPEG-surround bitstream, depending on the intended output channel configuration. In a first mode of operation, the output signal configuration is mono, stereo or binaural, and two output channels are used. In this first case, the SAOC module may operate in a decoder mode, and the SAOC module output is a pulse-code-modulated output (PCM output). In the first case, an MPEG surround decoder is not required. Rather, the upmix signal representation may only comprise the output signal 220, while the provision of the MPEG surround bit stream 222 may be omitted. In a second case, the output signal configuration is a multi-channel configuration with more than two output channels. The SAOC module may be operational in a transcoder mode. The SAOC module output may comprise both a downmix signal 220 and an MPEG surround bit stream 222 in this case, as shown in Fig. 2. Accordingly, an MPEG surround decoder is required in order to obtain a final audio signal representation for output by the speakers.

Fig. 2 shows the basic structure of the SAOC transcoder/decoder architecture. The residual processor 216 extracts the enhanced audio object from the incoming downmix signal 210 using the residual information contained in the SAOC bit stream 212. The downmix preprocessor 270 processes the regular audio objects (which are, for example, non-enhanced audio objects, i.e., audio objects for which no residual information is transmitted in the SAOC bit stream 212). The enhanced audio objects (represented by the first audio object signal 262) and the processed regular audio objects (represented, for example, by the processed version 272 of the second audio object signal 264) are combined to the output signal 220 for the SAOC decoder mode or to the MPEG surround downmix signal 220 for the SAOC transcoder mode. Detailed descriptions of the processing blocks are given below.

3. Architecture and functionality of Residual Processor and Energy Mode Processor

In the following, details regarding a residual processor will be described, which may, for example, take over the functionality of the object separator 130 of the audio signal decoder 100 or of the residual processor 260 of the audio signal decoder 200. For this purpose, Figs. 3a and 3b show block schematic diagrams of such a residual processor 300, which may take the place of the object separator 130 or of the residual processor 260. Fig. 3a shows less details than Fig. 3b. However, the following description applies to the residual processor 300 according to Fig. 3a and also to the residual processor 380 according to Fig. 3b.

The residual processor 300 is configured to receive an SAOC downmix signal 310, which may be equivalent to the downmix signal representation 112 of Fig. 1 or the downmix signal representation 210 of Fig. 2. The residual processor 300 is configured to provide, on the basis thereof, a first audio information 320 describing one or more enhanced audio objects, which may, for example, be equivalent to the first audio information 132 or to the first audio object signal 262. Also, the residual processor 300 may provide a second audio information 322 describing one or more other audio objects (for example, non-enhanced audio objects, for which no residual information is available), wherein the second audio information 322 may be equivalent to the second audio information 134 or to the second audio object signal 264.

The residual processor 300 comprises a 1-to-N/2-to-N unit (OTN/TTN unit) 330, which receives the SAOC downmix signal 310 and which also receives SAOC data and residuals 332. The 1-to-N/2-to-N unit 330 also provides an enhanced-audio-object signal 334, which describes the enhanced audio objects (EAO) contained in the SAOC downmix signal 310. Also, the 1-to-N/2-to-N unit 330 provides the second audio information 322. The residual processor 300 also comprises a rendering unit 340, which receives the enhanced-audio-object signal 334 and a rendering matrix information 342 and provides, on the basis thereof, the first audio information 320.

In the following, the enhanced audio object processing (EAO processing), which is performed by the residual processor 300, will be described in more detail.

3.1. Introduction into the Operation of the Residual Processor 300

Regarding the functionality of the residual processor 300, it should be noted that the SAOC technology allows for the individual manipulation of a number of audio objects in terms of their level amplification/attenuation without significant decrease in the resulting sound quality only in a very limited way. A special “karaoke-type” application scenario requires a total (or almost total) suppression of the specific objects, typically the lead vocal, keeping the perceptual quality of the background sound scene unharmed.

A typical application case contains up to four enhanced audio objects (EAO) signals, which can, for example, represent two independent stereo objects (for example, two independent stereo objects which are prepared to be removed at the side of the decoder).

It should be noted that the (one or more) quality enhanced audio objects (or, more precisely, the audio signal contributions associated with the enhanced audio objects) are included in the SAOC downmix signal 310. Typically, the audio signal contributions associated with the (one or more) enhanced audio objects are mixed, by the downmix processing performed by the audio signal encoder, with audio signal contributions of other audio objects, which are not enhanced audio objects. Also, it should be noted that audio signal contributions of a plurality of enhanced audio objects are also typically overlapped or mixed by the downmix processing performed by the audio signal encoder.

3.2 SOAC Architecture Supporting Enhanced Audio Objects

In the following, details regarding the residual processor 300 will be described. Enhanced audio object processing incorporates the 1-to-N or 2-to-N units, depending on the SAOC downmix mode. The 1-to-N processing unit is dedicated to a mono downmix signal and the 2-to-N processing unit is dedicated to a stereo downmix signal 310. Both these units represent a generalized and enhanced modification of the 2-to-2 box (TTT box) known from ISO/IEC 23003-1:2007. In the encoder, regular and EAO signals are combined into the downmix. The OTN⁻¹/TTN⁻¹ processing units (which are inverse one-to-N processing units or inverse 2-to-N processing units) are employed to produce and encode the corresponding residual signals.

The EAO and regular signals are recovered from the downmix 310 by the OTN/TTN units 330 using the SAOC side information and incorporated residual signals. The recovered EAOs (which are described by the enhanced audio object signal 334) are fed into the rendering unit 340 which represents (or provides) the product of the corresponding rendering matrix (described by the rendering matrix information 342) and the resulting output of the OTN/TTN unit. The regular audio objects (which are described by the second audio information 322) are delivered to the SAOC downmix pre-processor, for example, the SAOC downmix preprocessor 270, for further processing. Figs. 3a and 3b depict the general structure of the residual processor, i.e., the architecture of the residual processor.

The residual processor output signals 320,322 are computed as

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ} \mathbf{X}_{res}$$

$$\mathbf{X}_{EAO} = \mathbf{A}_{EAO} \mathbf{M}_{EAO} \mathbf{X}_{res}$$

where $\mathbf{X}_{OBJ}$ represents the downmix signal of the regular audio objects (i.e. non-EAOs) and $\mathbf{X}_{EAO}$ is the rendered EAO output signal for the SAOC decoding mode or the corresponding EAO downmix signal for the SAOC transcoding mode.

- 5 The residual processor can operate in prediction (using residual information) mode or energy (without residual information) mode. The extended input signal $\mathbf{X}_{res}$ is defined accordingly:

$$\mathbf{X}_{res} = \begin{cases} \begin{pmatrix} \mathbf{X} \\ \mathbf{res} \end{pmatrix}, & \text{for prediction mode,} \\ \mathbf{X}, & \text{for energy mode.} \end{cases}$$

- 10 Here, $\mathbf{X}$ may, for example, represent the one or more channels of the downmix signal representation 310, which may be transported in the bitstream representing the multi-channel audio content. $\mathbf{res}$ may designate one or more residual signals, which may be described by the bitstream representing the multi-channel audio content.
- 15 The OTN/TTN processing is represented by matrix $\mathbf{M}$ and EAO processor by matrix $\mathbf{A}_{EAO}$.

The OTN/TTN processing matrix $\mathbf{M}$ is defined according to the EAO operation mode (i.e. prediction or energy) as

$$20 \quad \mathbf{M} = \begin{cases} \mathbf{M}_{\text{Prediction}}, & \text{for prediction mode,} \\ \mathbf{M}_{\text{Energy}}, & \text{for energy mode.} \end{cases}$$

The OTN/TTN processing matrix $\mathbf{M}$ is represented as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{OBJ} \\ \mathbf{M}_{EAO} \end{pmatrix},$$

where the matrix $\mathbf{M}_{OBJ}$ relates to the regular audio objects (i.e. non-EAOs) and $\mathbf{M}_{EAO}$ to the enhanced audio objects (EAOs).

In some embodiments, one or more multichannel background objects (MBO) may be treated the same way by the residual processor 300.

A Multi-channel Background Object (MBO) is an MPS mono or stereo downmix that is part of the SAOC downmix. As opposed to using individual SAOC objects for each channel in a multi-channel signal, an MBO can be used enabling SAOC to more efficiently handle a multi-channel object. In the MBO case, the SAOC overhead gets lower as the MBO's SAOC parameters only are related to the downmix channels rather than all the upmix channels.

3.3 Further Definitions

3.3.1 Dimensionality of Signals and Parameters

In the following, the dimensionality of the signals and parameters will be briefly discussed in order to provide an understanding how often the different calculations are performed.

The audio signals are defined for every time slot n and every hybrid subband (which may be a frequency subband) k . The corresponding SAOC parameters are defined for each parameter time slot l and processing band m . A Subsequent mapping between the hybrid and parameter domain is specified by table A.31 ISO/IEC 23003-1:2007. Hence, all calculations are performed with respect to the certain time/band indices and the corresponding dimensionalities are implied for each introduced variable.

However, in the following, the time and frequency band indices will be omitted sometimes to keep the notation concise.

3.3.2 Calculation of the matrix $\mathbf{A}_{EAO}$

The EAO pre-rendering matrix $\mathbf{A}_{EAO}$ is defined according to the number of output channels (i.e. mono, stereo or binaural) as

$$\mathbf{A}_{EAO} = \begin{cases} \mathbf{A}_1^{EAO}, & \text{for mono case,} \\ \mathbf{A}_2^{EAO}, & \text{for other cases.} \end{cases}$$

The matrices $\mathbf{A}_1^{EAO}$ of size $1 \times N_{EAO}$ and $\mathbf{A}_2^{EAO}$ of size $2 \times N_{EAO}$ are defined as

$$\mathbf{A}_1^{EAO} = \mathbf{D}_{16}^{EAO} \mathbf{M}_{ren}^{EAO}, \quad \mathbf{D}_{16}^{EAO} = \begin{pmatrix} w_1^{EAO} & w_2^{EAO} & w_3^{EAO} & w_3^{EAO} & w_1^{EAO} & w_2^{EAO} \end{pmatrix},$$

$$5 \quad \mathbf{A}_2^{EAO} = \mathbf{D}_{26}^{EAO} \mathbf{M}_{ren}^{EAO}, \quad \mathbf{D}_{26}^{EAO} = \begin{pmatrix} w_1^{EAO} & 0 & \frac{w_3^{EAO}}{\sqrt{2}} & \frac{w_3^{EAO}}{\sqrt{2}} & w_1^{EAO} & 0 \\ 0 & w_2^{EAO} & \frac{w_3^{EAO}}{\sqrt{2}} & \frac{w_3^{EAO}}{\sqrt{2}} & 0 & w_2^{EAO} \end{pmatrix},$$

where the rendering sub-matrix M_{ren}^{EAO} corresponds to the EAO rendering (and describes a desired mapping of enhanced audio objects onto channels of the upmix signal representation).

10

The values w_i^{EAO} are computed in dependence on rendering information associated with the enhanced audio objects using the corresponding EAO elements and using the equations of section 4.2.2.1.

15 In case of binaural rendering the matrix A_2^{EAO} is defined by equations given in section 4.1.2, for which the corresponding target binaural rendering matrix contains only EAO related elements.

3.4 Calculation of the OTN/TTN Elements in the Residual Mode

20

In the following, it will be discussed how the SAOC downmix signal 310, which typically comprises one or two audio channels, is mapped onto the enhanced audio object signal 334, which typically comprises one or more enhanced audio object channels, and the second audio information 322, which typically comprises one or two regular audio object channels.

25

The functionality of the 1-to-N unit or 2-to-N unit 330 may, for example, be implemented using a matrix vector multiplication, such that a vector describing both the channels of the enhanced audio object signal 334 and the channels of the second audio information 322 is obtained by multiplying a vector describing the channels of the SAOC downmix signal 310 and (optionally) one or more residual signals with a matrix $\mathbf{M}_{\text{Prediction}}$ or $\mathbf{M}_{\text{Energy}}$. Accordingly, the determination of the matrix $\mathbf{M}_{\text{Prediction}}$ or $\mathbf{M}_{\text{Energy}}$ is an important step in the derivation of the first audio information 320 and the second audio information 322 from the SAOC downmix 310.

- 10 To summarize, the OTN/TTN upmix process is presented by either a matrix $\mathbf{M}_{\text{Prediction}}$ for a prediction mode or $\mathbf{M}_{\text{Energy}}$ for an energy mode.

The energy based encoding/decoding procedure is designed for non-waveform preserving coding of the downmix signal. Thus the OTN/TTN upmix matrix for the corresponding energy mode does not rely on specific waveforms, but only describe the relative energy distribution of the input audio objects, as will be discussed in more detail below.

3.4.1 Prediction mode

- 20 For the prediction mode the matrix $\mathbf{M}_{\text{Prediction}}$ is defined exploiting the downmix information contained in the matrix $\tilde{\mathbf{D}}^{-1}$ and the CPC data from matrix $\mathbf{C}$:

$$\mathbf{M}_{\text{Prediction}} = \tilde{\mathbf{D}}^{-1} \mathbf{C}.$$

- 25 With respect to the several SAOC modes, the extended downmix matrix $\tilde{\mathbf{D}}$ and CPC matrix $\mathbf{C}$ exhibit the following dimensions and structures:

3.4.1.1 Stereo downmix modes (TTN):

- 30 For stereo downmix modes (TTN) (for example, for the case of a stereo downmix on the basis of two regular-audio-object channels and N_{EAO} enhanced-audio-object-channels), the (extended) downmix matrix $\tilde{\mathbf{D}}$ and the CPC matrix $\mathbf{C}$ can be obtained as follows:

$$\tilde{\mathbf{D}} = \left(\begin{array}{cc|ccc} 1 & 0 & m_0 & \dots & m_{N_{EAO}-1} \\ 0 & 1 & n_0 & \dots & n_{N_{EAO}-1} \\ \hline m_0 & n_0 & -1 & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ m_{N_{EAO}-1} & n_{N_{EAO}-1} & 0 & \dots & -1 \end{array} \right),$$

$$\mathbf{C} = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \hline c_{0,0} & c_{0,1} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N_{EAO}-1,0} & c_{N_{EAO}-1,1} & 0 & \dots & 1 \end{array} \right).$$

- 5 With a stereo downmix, each EAO j holds two CPCs $c_{j,0}$ and $c_{j,1}$ yielding matrix $\mathbf{C}$.

The residual processor output signals are computed as

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Prediction}} \left(\begin{array}{c} l_0 \\ r_0 \\ \hline res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{array} \right),$$

10

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Prediction}} \left(\begin{array}{c} l_0 \\ r_0 \\ \hline res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{array} \right).$$

Accordingly, two signals y_L, y_R (which are represented by $\mathbf{X}_{OBJ}$) are obtained, which represent one or two or even more than two regular audio objects (also designated as non-extended audio objects). Also, N_{EAO} signals (represented by $\mathbf{X}_{EAO}$) representing N_{EAO} enhanced audio objects are obtained. These signals are obtained on the basis of two SAOC downmix signals l_0, r_0 and N_{EAO} residual signals res_0 to $res_{N_{EAO}-1}$, which will be encoded in the SAOC side information, for example, as a part as the object-related parametric information.

20

It should be noted that the signals y_L and y_R may be equivalent to the signal 322, and that the signals $y_{0,EAO}$ to $y_{N_{EAO}-1, EAO}$ (which are represented by $\mathbf{X}_{EAO}$) may be equivalent to the signals 320.

- 5 The matrix $\mathbf{A}^{EAO}$ is a rendering matrix. Entries of the matrix $\mathbf{A}^{EAO}$ may describe, for example, a mapping of enhanced audio objects to the channels of the enhanced audio object signal 334 ($\mathbf{X}_{EAO}$).

Accordingly, an appropriate choice of the matrix $\mathbf{A}^{EAO}$ may allow for an optional
 10 integration of the functionality of the rendering unit 340, such that the multiplication of the vector describing the channels (l_0, r_0) of the SAOC downmix signal 310 and one or more residual signals ($res_0, \dots, res_{N_{EAO}-1}$) with the matrix $\mathbf{A}^{EAO} \mathbf{M}_{EAO}^{Prediction}$ may directly result in a representation $\mathbf{X}_{EAO}$ of the first audio information 320.

15

3.4.1.2 Mono downmix modes (OTN):

In the following, the derivation of the enhanced audio object signals 320 (or, alternatively, of the enhanced audio object signals 334) and of the regular audio object signal 322 will be
 20 described for the case in which the SAOC downmix signal 310 comprises a signal channel only.

For mono downmix modes (OTN) (e.g., a mono downmix on the basis of one regular-audio-object channel and N_{EAO} enhanced-audio-object channels), the (extended) downmix
 25 matrix $\tilde{\mathbf{D}}$ and the CPC matrix $\mathbf{C}$ can be obtained as follows:

$$\tilde{\mathbf{D}} = \left(\begin{array}{c|ccc} 1 & m_0 & \dots & m_{N_{EAO}-1} \\ \hline m_0 & -1 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ m_{N_{EAO}-1} & 0 & \dots & -1 \end{array} \right),$$

$$\mathbf{C} = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline c_{0,0} & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ c_{N_{EAO}-1,0} & 0 & \dots & 1 \end{array} \right).$$

30

With a mono downmix, one EAO j is predicted by only one coefficient c_j yielding the matrix $\mathbf{C}$. All matrix elements c_j are obtained, for example, from the SAOC parameters (for example, from the SAOC data 322) according to the relationships provided below (section 3.4.1.4).

5

The residual processor output signals are computed as

$$\begin{aligned} \mathbf{X}_{OBJ} &= \mathbf{M}_{OBJ}^{\text{Prediction}} \begin{pmatrix} d_0 \\ \text{---} \\ res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{pmatrix}, \\ 10 \quad \mathbf{X}_{EAO} &= \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Prediction}} \begin{pmatrix} d_0 \\ \text{---} \\ res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{pmatrix}. \end{aligned}$$

The output signal $\mathbf{X}_{OBJ}$ comprises, for example, one channel describing the regular audio objects (non-enhanced audio objects). The output signal $\mathbf{X}_{EAO}$ comprises, for example, one, two, or even more channels describing the enhanced audio objects (preferably N_{EAO} channels describing the enhanced audio objects). Again, said signals are equivalent to the signals 320, 322.

20

3.4.1.3 Calculation of the inverse extended downmix matrix

The matrix $\tilde{\mathbf{D}}^{-1}$ is the inverse of the extended downmix matrix $\tilde{\mathbf{D}}$ and $\mathbf{C}$ implies the CPCs.

25 The matrix $\tilde{\mathbf{D}}^{-1}$ is the inverse of the extended downmix matrix $\tilde{\mathbf{D}}$ and can be calculated as

$$\tilde{\mathbf{D}}^{-1} = \frac{\tilde{d}_{i,j}}{den}.$$

The elements $\tilde{d}_{i,j}$ (for example, of the inverse $\tilde{\mathbf{D}}^{-1}$ of the extended downmix matrix $\tilde{\mathbf{D}}$ of size 6×6) are derived using the following values:

$$\begin{aligned}
 \tilde{d}_{1,1} &= 1 + \sum_{j=1}^4 n_j^2, \\
 5 \quad \tilde{d}_{1,2} &= - \left(\sum_{j=1}^4 m_j n_j \right), \\
 \tilde{d}_{1,3} &= m_1 + m_1 n_2^2 + m_1 n_3^2 + m_1 n_4^2 - m_2 n_1 n_2 - m_3 n_1 n_3 - m_4 n_1 n_4, \\
 \tilde{d}_{1,4} &= m_2 + m_2 n_1^2 + m_2 n_3^2 + m_2 n_4^2 - m_1 n_2 n_1 - m_3 n_2 n_3 - m_4 n_2 n_4, \\
 \tilde{d}_{1,5} &= m_3 + m_3 n_1^2 + m_3 n_2^2 + m_3 n_4^2 - m_1 n_3 n_1 - m_2 n_3 n_2 - m_4 n_3 n_4, \\
 \tilde{d}_{1,6} &= m_4 + m_4 n_1^2 + m_4 n_2^2 + m_4 n_3^2 - m_1 n_4 n_1 - m_2 n_4 n_2 - m_3 n_4 n_3, \\
 10 \quad \tilde{d}_{2,2} &= 1 + \sum_{j=1}^4 m_j^2, \\
 \tilde{d}_{2,3} &= n_1 + n_1 m_2^2 + n_1 m_3^2 + n_1 m_4^2 - m_1 m_2 n_2 - m_1 m_3 n_3 - m_1 m_4 n_4, \\
 \tilde{d}_{2,4} &= n_2 + n_2 m_1^2 + n_2 m_3^2 + n_2 m_4^2 - m_2 m_1 n_1 - m_2 m_3 n_3 - m_2 m_4 n_4, \\
 \tilde{d}_{2,5} &= n_3 + n_3 m_1^2 + n_3 m_2^2 + n_3 m_4^2 - m_3 m_1 n_1 - m_3 m_2 n_2 - m_3 m_4 n_4, \\
 \tilde{d}_{2,6} &= n_4 + n_4 m_1^2 + n_4 m_2^2 + n_4 m_3^2 - m_4 m_1 n_1 - m_4 m_2 n_2 - m_4 m_3 n_3, \\
 15 \quad \tilde{d}_{3,3} &= -1 - \sum_{j=2}^4 m_j^2 - \sum_{j=2}^4 n_j^2 - m_3^2 n_2^2 - m_4^2 n_2^2 - m_2^2 n_3^2 - m_4^2 n_3^2 - m_2^2 n_4^2 - m_3^2 n_4^2 + 2m_2 m_3 n_2 n_3 + 2m_2 m_4 n_2 n_4 + 2m_3 m_4 n_3 n_4, \\
 \tilde{d}_{3,4} &= m_1 m_2 + n_1 n_2 + m_2^2 n_1 n_2 + m_4^2 n_1 n_2 + m_1 m_2 n_3^2 + m_1 m_2 n_4^2 - m_2 m_3 n_1 n_3 - m_1 m_3 n_2 n_3 - m_2 m_4 n_1 n_4 - m_1 m_4 n_2 n_4, \\
 \tilde{d}_{3,5} &= m_1 m_3 + n_1 n_3 + m_2^2 n_1 n_3 + m_4^2 n_1 n_3 + m_1 m_3 n_2^2 + m_1 m_3 n_4^2 - m_2 m_3 n_1 n_2 - m_1 m_2 n_2 n_3 - m_3 m_4 n_1 n_4 - m_1 m_4 n_3 n_4, \\
 \tilde{d}_{3,6} &= m_1 m_4 + n_1 n_4 + m_2^2 n_1 n_4 + m_3^2 n_1 n_4 + m_1 m_4 n_2^2 + m_1 m_4 n_3^2 - m_2 m_4 n_1 n_2 - m_3 m_4 n_1 n_3 - m_1 m_2 n_2 n_4 - m_1 m_3 n_4 n_4, \\
 \tilde{d}_{4,4} &= -1 - \sum_{\substack{j=1 \\ j \neq 2}}^4 m_j^2 - \sum_{\substack{j=1 \\ j \neq 2}}^4 n_j^2 - m_3^2 n_1^2 - m_4^2 n_1^2 - m_1^2 n_3^2 - m_4^2 n_3^2 - m_1^2 n_4^2 - m_3^2 n_4^2 + 2m_1 m_3 n_1 n_3 + 2m_1 m_4 n_1 n_4 + 2m_3 m_4 n_3 n_4, \\
 20 \quad \tilde{d}_{4,5} &= m_2 m_3 + n_2 n_3 + m_1^2 n_2 n_3 + m_4^2 n_2 n_3 + m_2 m_3 n_1^2 + m_2 m_3 n_4^2 - m_1 m_3 n_1 n_2 - m_1 m_2 n_1 n_3 - m_3 m_4 n_2 n_4 - m_2 m_4 n_3 n_4, \\
 \tilde{d}_{4,6} &= m_2 m_4 + n_2 n_4 + m_1^2 n_2 n_4 + m_3^2 n_2 n_4 + m_2 m_4 n_1^2 + m_2 m_4 n_3^2 - m_1 m_4 n_1 n_2 - m_3 m_4 n_2 n_3 - m_1 m_2 n_1 n_4 - m_2 m_3 n_3 n_4, \\
 \tilde{d}_{5,5} &= -1 - \sum_{\substack{j=1 \\ j \neq 3}}^4 m_j^2 - \sum_{\substack{j=1 \\ j \neq 3}}^4 n_j^2 - m_2^2 n_1^2 - m_4^2 n_1^2 - m_1^2 n_2^2 - m_4^2 n_2^2 - m_1^2 n_4^2 - m_2^2 n_4^2 + 2m_1 m_2 n_1 n_2 + 2m_1 m_4 n_1 n_4 + 2m_2 m_4 n_2 n_4, \\
 \tilde{d}_{5,6} &= m_3 m_4 + n_3 n_4 + m_1^2 n_3 n_4 + m_2^2 n_3 n_4 + m_3 m_4 n_1^2 + m_3 m_4 n_2^2 - m_1 m_4 n_1 n_3 - m_2 m_4 n_2 n_3 - m_1 m_3 n_1 n_4 - m_2 m_3 n_2 n_4, \\
 \tilde{d}_{6,6} &= -1 - \sum_{j=1}^3 m_j^2 - \sum_{j=1}^3 n_j^2 - m_2^2 n_1^2 - m_3^2 n_1^2 - m_1^2 n_2^2 - m_3^2 n_2^2 - m_1^2 n_3^2 - m_2^2 n_3^2 + 2m_1 m_2 n_1 n_2 + 2m_1 m_3 n_1 n_3 + 2m_2 m_3 n_2 n_3, \\
 den &= 1 + \sum_{j=1}^4 m_j^2 + \sum_{j=1}^4 n_j^2 + m_2^2 n_1^2 + m_3^2 n_1^2 + m_4^2 n_1^2 + m_1^2 n_2^2 + m_3^2 n_2^2 + m_4^2 n_2^2 + m_1^2 n_3^2 + m_2^2 n_3^2 + m_4^2 n_3^2 + m_1^2 n_4^2 + m_2^2 n_4^2 + \\
 25 \quad &+ m_3^2 n_4^2 - 2m_1 m_2 n_1 n_2 - 2m_1 m_3 n_1 n_3 - 2m_2 m_3 n_2 n_3 - 2m_1 m_4 n_1 n_4 - 2m_2 m_4 n_2 n_4 - 2m_3 m_4 n_3 n_4.
 \end{aligned}$$

The coefficients m_j and n_j of the extended downmix matrix $\tilde{\mathbf{D}}$ denote the downmix values for every EAO j for the right and left downmix channel as

5

$$m_j = d_{0, EAO(j)}, \quad n_j = d_{1, EAO(j)}.$$

The elements $d_{i,j}$ of the downmix matrix $\mathbf{D}$ are obtained using the downmix gain information DMG and the (optional) downmix channel level different information DCLD, which is included in the SAOC information 332, which is represented, for example, by the object-related parametric information 110 or the SAOC bitstream information 212.

For the stereo downmix case the downmix matrix $\mathbf{D}$ of size $2 \times N$ with elements $d_{i,j}$ ($i = 0, 1; j = 0, \dots, N-1$) is obtained from the DMG and DCLD parameters as

15

$$d_{0,j} = 10^{0.05 DMG_j} \sqrt{\frac{10^{0.1 DCLD_j}}{1 + 10^{0.1 DCLD_j}}}, \quad d_{1,j} = 10^{0.05 DMG_j} \sqrt{\frac{1}{1 + 10^{0.1 DCLD_j}}}.$$

For the mono downmix case the downmix matrix $\mathbf{D}$ of size $1 \times N$ with elements $d_{i,j}$ ($i = 0; j = 0, \dots, N-1$) is obtained from the DMG parameters as

20

$$d_{0,j} = 10^{0.05 DMG_j}.$$

Here, the dequantized downmix parameters DMG_j and $DCLD_j$ are obtained, for example, from the parametric side information 110 or from the SAOC bitstream 212.

25

The function $EAO(j)$ determines mapping between indices of input audio object channels and EAO signals:

30

$$EAO(j) = N - 1 - j, \quad j = 0, \dots, N_{EAO} - 1.$$

3.4.1.4 Calculation of the matrix C

The matrix $\mathbf{C}$ implies the CPCs and is derived from the transmitted SAOC parameters (i.e. the OLDs, IOC, DMGs and DCLDs) as

35

$$c_{j,0} = (1 - \lambda) \tilde{c}_{j,0} + \lambda \gamma_{j,0}, \quad c_{j,1} = (1 - \lambda) \tilde{c}_{j,1} + \lambda \gamma_{j,1}.$$

In other words, the constrained CPCs are obtained in accordance with the above equations, which may be considered as a constraining algorithm. However, the constrained CPCs may also be derived from the values $\tilde{c}_{j,0}, \tilde{c}_{j,1}$ using a different limitation approach (constraining algorithm), or can be set to be equal to the values $\tilde{c}_{j,0}, \tilde{c}_{j,1}$.

It should be noted, that matrix entries $c_{j,1}$ (and the intermediate quantities on the basis of which the matrix entries $c_{j,1}$ are computed) are typically only required if the downmix signal is a stereo downmix signal.

The CPCs are constrained by the subsequent limiting functions:

$$\gamma_{j,1} = \frac{m_j OLD_L + n_j e_{L,R} - \sum_{i=0}^{N_{EAO}-1} m_i e_{i,j}}{2 \left(OLD_L + \sum_{i=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} m_i m_k e_{i,k} \right)}, \quad \gamma_{j,2} = \frac{n_j OLD_R + m_j e_{L,R} - \sum_{i=0}^{N_{EAO}-1} n_i e_{i,j}}{2 \left(OLD_R + \sum_{i=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} n_i n_k e_{i,k} \right)},$$

with the weighting factor λ determined as

$$\lambda = \left(\frac{P_{LoRo}^2}{P_{Lo} P_{Ro}} \right)^8.$$

For one specific EAO channel $j = 0 \dots N_{EAO} - 1$ the unconstrained CPCs are estimated by

$$\tilde{c}_{j,0} = \frac{P_{LoCo,j} P_{Ro} - P_{RoCo,j} P_{LoRo}}{P_{Lo} P_{Ro} - P_{LoRo}^2}, \quad \tilde{c}_{j,1} = \frac{P_{RoCo,j} P_{Lo} - P_{LoCo,j} P_{LoRo}}{P_{Lo} P_{Ro} - P_{LoRo}^2}.$$

The energy quantities P_{Lo} , P_{Ro} , P_{LoRo} , $P_{LoCo,j}$ and $P_{RoCo,j}$ are computed as

$$P_{Lo} = OLD_L + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} m_j m_k e_{j,k},$$

$$P_{Ro} = OLD_R + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} n_j n_k e_{j,k},$$

$$P_{LoRo} = e_{L,R} + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} m_j n_k e_{j,k} ,$$

$$P_{LoCo,j} = m_j OLD_L + n_j e_{L,R} - m_j OLD_j - \sum_{\substack{i=0 \\ i \neq j}}^{N_{EAO}-1} m_i e_{i,j} ,$$

5

$$P_{RoCo,j} = n_j OLD_R + m_j e_{L,R} - n_j OLD_j - \sum_{\substack{i=0 \\ i \neq j}}^{N_{EAO}-1} n_i e_{i,j} .$$

The covariance matrix $e_{i,j}$ is defined in the following way: The covariance matrix $\mathbf{E}$ of size $N \times N$ with elements $e_{i,j}$ represents an approximation of the original signal

10 covariance matrix $\mathbf{E} \approx \mathbf{SS}^*$ and is obtained from the OLD and IOC parameters as

$$e_{i,j} = \sqrt{OLD_i OLD_j} IOC_{i,j} .$$

Here, the dequantized object parameters OLD_i , $IOC_{i,j}$ are obtained, for example, from the
15 parametric side information 110 or from the SAOC bitstream 212.

In addition, $e_{L,R}$ may, for example, be obtained as

$$e_{L,R} = \sqrt{OLD_L OLD_R} IOC_{L,R} .$$

20

The parameters OLD_L , OLD_R and $IOC_{L,R}$ correspond to the regular (audio) objects and can be derived using the downmix information:

$$OLD_L = \sum_{i=0}^{N-N_{EAO}-1} d_{0,i}^2 OLD_i ,$$

$$OLD_R = \sum_{i=0}^{N-N_{EAO}-1} d_{1,i}^2 OLD_i ,$$

$$IOC_{L,R} = \begin{cases} IOC_{0,1}, & N - N_{EAO} = 2, \\ 0, & \text{otherwise.} \end{cases}$$

25

As can be seen, two common object-level-different values OLD_L and OLD_R are computed
30 for the regular audio objects in the case of a stereo downmix signal (which preferably

implies a two-channel regular audio object signal). In contrast, only one common object-level-difference value OLD_L is computed for the regular audio objects in the case of a one-channel (mono) downmix signal (which preferably implies a one-channel regular audio object signal).

5

As can be seen, the first (in the case of a two-channel downmix signal) or sole (in the case of a one-channel downmix signal) common object-level-difference value OLD_L is obtained by summing contributions of the regular audio objects having audio object index (or indices) i to the left channel (or sole channel) of the SAOC downmix signal 310.

10

The second common object-level-difference value OLD_R (which is used in the case of a two-channel downmix signal) is obtained by summing the contributions of the regular audio objects having the audio object index (or indices) i to the right channel of the SAOC downmix signal 310.

15

The contribution OLD_L of the regular audio objects (having audio objects indices $i=0$ to $i=N-N_{EAO}-1$) onto the left channel signal (or sole channel signal) of the SAOC downmix signal 710 is computed, for example, taking into consideration the downmix gain $d_{0,i}$, describing the downmix gain applied to the regular audio object having audio object index i when obtaining the left channel signal of the SAOC downmix signal 310, and also the object level of the regular audio object having the audio object i , which is represented by the value OLD_i .

20

Similarly, the common object level difference value OLD_R is obtained using the downmix coefficients $d_{1,i}$, describing the downmix gain which is applied to the regular audio object having the audio object index i when forming the right channel signal of the SAOC downmix signal 310, and the level information OLD_i associated with the regular audio object having the audio object index i .

25

As can be seen, the equations for the calculation of the quantities P_{Lo} , P_{Ro} , P_{LoRo} , P_{LoCoj} and P_{RoCoj} do not distinguish between the individual regular audio objects, but merely make use of the common object level difference values OLD_L , OLD_R , thereby considering the regular audio objects (having audio object indices i) as a single audio object.

30

Also, the inter-object-correlation value $IOC_{L,R}$, which is associated with the regular audio objects, is set to 0 unless there are two regular audio objects.

35

The covariance matrix $e_{i,j}$ (and $e_{L,R}$) is defined as follows:

The covariance matrix $\mathbf{E}$ of size $N \times N$ with elements $e_{i,j}$ represents an approximation of the original signal covariance matrix $\mathbf{E} \approx \mathbf{S}\mathbf{S}^*$ and is obtained from the OLD and IOC parameters as

$$e_{i,j} = \sqrt{OLD_i OLD_j} IOC_{i,j}.$$

For example,

$$e_{L,R} = \sqrt{OLD_L OLD_R} IOC_{L,R},$$

wherein OLD_L and OLD_R and $IOC_{L,R}$ are computed as described above.

Here, the dequantized object parameters are obtained as

$$OLD_i = \mathbf{D}_{OLD}(i, l, m), \quad IOC_{i,j} = \mathbf{D}_{IOC}(i, j, l, m),$$

wherein $\mathbf{D}_{OLD}$ and $\mathbf{D}_{IOC}$ are matrices comprising objects-level-difference parameters and inter-object-correlation parameters.

3.4.2. Energy Mode

In the following, another concept will be described, which can be used to separate the extended-audio-object signals 320 and the regular-audio-object (non-extended audio object) signals 322, and which can be used in combination with a non-waveform-preserving audio coding of the SAOC downmix channels 310.

In other words, the energy based encoding/decoding procedure is designed for non-waveform preserving coding of the downmix signal. Thus the OTN/TTN upmix matrix for the corresponding energy mode does not rely on specific waveforms, but only describe the relative energy distribution of the input audio objects.

Also, the concept discussed here, which is designated as an “energy mode” concept, can be used without transmitting a residual signal information. Again, the regular audio objects (non-enhanced audio objects) are treated as a single one-channel or two-channel audio object having one or two common object-level-difference values OLD_L, OLD_R.

For the energy mode the matrix $\mathbf{M}_{\text{Energy}}$ is defined exploiting the downmix information and the OLDs, as will be described in the following.

3.4.2.1. Energy Mode for Stereo Downmix Modes (TTN)

In case of a stereo (for example, a stereo downmix on the basis of two regular-audio-object channels and N_{EAO} enhanced-audio-object channels), the matrices $\mathbf{M}_{\text{OBJ}}^{\text{Energy}}$ and $\mathbf{M}_{\text{EAO}}^{\text{Energy}}$ are obtained from the corresponding OLDs according to

$$\mathbf{M}_{\text{OBJ}}^{\text{Energy}} = \begin{pmatrix} \sqrt{\frac{\text{OLD}_L}{\text{OLD}_L + \sum_{i=0}^{N_{\text{EAO}}-1} m_i^2 \text{OLD}_i}} & 0 \\ 0 & \sqrt{\frac{\text{OLD}_R}{\text{OLD}_R + \sum_{i=0}^{N_{\text{EAO}}-1} n_i^2 \text{OLD}_i}} \end{pmatrix}$$

$$\mathbf{M}_{\text{EAO}}^{\text{Energy}} = \begin{pmatrix} \sqrt{\frac{m_0^2 \text{OLD}_0}{\text{OLD}_L + \sum_{i=0}^{N_{\text{EAO}}-1} m_i^2 \text{OLD}_i}} & \sqrt{\frac{n_0^2 \text{OLD}_0}{\text{OLD}_R + \sum_{i=0}^{N_{\text{EAO}}-1} n_i^2 \text{OLD}_i}} \\ \vdots & \vdots \\ \sqrt{\frac{m_{N_{\text{EAO}}-1}^2 \text{OLD}_{N_{\text{EAO}}-1}}{\text{OLD}_L + \sum_{i=0}^{N_{\text{EAO}}-1} m_i^2 \text{OLD}_i}} & \sqrt{\frac{n_{N_{\text{EAO}}-1}^2 \text{OLD}_{N_{\text{EAO}}-1}}{\text{OLD}_R + \sum_{i=0}^{N_{\text{EAO}}-1} n_i^2 \text{OLD}_i}} \end{pmatrix}$$

The residual processor output signals are computed as

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Energy}} \begin{pmatrix} l_0 \\ r_0 \end{pmatrix},$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Energy}} \begin{pmatrix} l_0 \\ r_0 \end{pmatrix}.$$

- 5 The signals y_L , y_R , which are represented by the signal $\mathbf{X}_{OBJ}$, describe the regular audio objects (and may be equivalent to the signal 322) , and the signals $\mathbf{y}_{0,EAO}$ to $\mathbf{y}_{N_{EAO}-1,EAO}$, which are described by the signal $\mathbf{X}_{EAO}$, describe the enhanced audio objects (and may be equivalent to the signal 334 or to the signal 320).
- 10 If a mono upmix signal is desired for the case of a stereo downmix signal, a 2-to-1 processing may be performed, for example, by the pre-processor 270 on the basis of the two-channel signal $\mathbf{X}_{OBJ}$.

15 3.4.2.2. Energy Mode for Mono Downmix Modes (OTN)

For the mono case (for example, a mono downmix on the basis of one regular-audio-object channel and N_{EAO} enhanced-audio-object channels), the matrices $\mathbf{M}_{OBJ}^{\text{Energy}}$ and $\mathbf{M}_{EAO}^{\text{Energy}}$ are obtained from the corresponding OLDs according to

20

$$\mathbf{M}_{OBJ}^{\text{Energy}} = \begin{pmatrix} \sqrt{\frac{OLD_L}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \end{pmatrix},$$

$$\mathbf{M}_{EAO}^{\text{Energy}} = \begin{pmatrix} \sqrt{\frac{m_0^2 OLD_0}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \\ \vdots \\ \sqrt{\frac{m_{N_{EAO}-1}^2 OLD_{N_{EAO}-1}}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \end{pmatrix}.$$

The residual processor output signals are computed as

25

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Energy}}(d_0),$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Energy}}(d_0).$$

5

A single regular-audio-object channel 322 (represented by $\mathbf{X}_{OBJ}$) and N_{EAO} enhanced-audio-object channels 320 (represented by $\mathbf{X}_{EAO}$) can be obtained by applying the matrices $\mathbf{M}_{OBJ}^{\text{Energy}}$ and $\mathbf{M}_{EAO}^{\text{Energy}}$ to a representation of a single channel SAOC downmix signal 310 (represented here by d_0).

If a two-channel (stereo) upmix signal is desired for the case of a one-channel (mono) downmix signal, a 1-to-2 processing may be performed, for example, by the pre-processor 270 on the basis of the one-channel signal $\mathbf{X}_{OBJ}$.

15

4. Architecture and operation of the SAOC Downmix Pre-Processor

In the following, the operation of the SAOC downmix pre-processor 270 will be described both for some decoding modes of operation and for some transcoding modes of operation.

20

4.1 Operation in the Decoding Modes

4.1.1 Introduction

25

In the following, a method for obtaining an output signal using SAOC parameters and panning information (or rendering information) associated with each audio object is described. The SAOC decoder 495 is depicted in Fig. 4g and consists of the SAOC parameter processor 496 and the downmix processor 497.

30

It should be noted that the SAOC decoder 494 may be used to process the regular audio objects, and may therefore receive, as the downmix signal 497a, the second audio object signal 264 or the regular-audio-object signal 322 or the second audio information 134. Accordingly, the downmix processor 497 may provide, as its output signals 497b, the processed version 272 of the second audio object signal 264 or the processed version 142 of the second audio information 134. Accordingly, the downmix processor 497 may take

35

the role of the SAOC downmix pre-processor 270, or the role of the audio signal processor 140.

The SAOC parameter processor 496 may take the role of the SAOC parameter processor 252 and consequently provides downmix information 496a.

4.1.2 Downmix Processor

In the following, the downmix processor, which is part of the audio signal processor 140, and which is designated as a “SAOC downmix pre-processor” 270 in the embodiment of Fig. 2, and which is designated with 497 in the SAOC decoder 495, will be described in more detail.

For the decoder mode of the SAOC system, the output signal 142, 272, 497b of the downmix processor (represented in the hybrid QMF domain) is fed into the corresponding synthesis filterbank (not shown in Figs. 1 and 2) as described in ISO/IEC 23003-1: 2007 yielding the final output PCM signal. Nevertheless, the output signal 142, 272, 497b of the downmix processor is typically combined with one or more audio signals 132, 262 representing the enhanced audio objects. This combination may be performed before the corresponding synthesis filterbank (such that a combined signal combining the output of the downmix processor and the one or more signals representing the enhanced audio objects is input to the synthesis filterbank). Alternatively, the output signal of the downmix processor may be combined with one or more audio signals representing the enhanced audio objects only after the synthesis filterbank processing. Accordingly, the upmix signal representation 120, 220 may be either a QMF domain representation or a PCM domain representation (or any other appropriate representation). The downmix processing incorporates, for example, the mono processing, the stereo processing and, if required, the subsequent binaural processing.

The output signal $\hat{\mathbf{X}}$ of the downmix processor 270, 497 (also designated with 142, 272, 497b) is computed from the mono downmix signal $\mathbf{X}$ (also designated with 134, 264, 497a) and the decorrelated mono downmix signal $\mathbf{X}_d$ as

$$\hat{\mathbf{X}} = \mathbf{G}\mathbf{X} + \mathbf{P}_2\mathbf{X}_d.$$

The decorrelated mono downmix signal $\mathbf{X}_d$ is computed as

$$\mathbf{X}_d = \text{decorrFunc}(\mathbf{X}).$$

The decorrelated signals $\mathbf{X}_d$ are created from the decorrelator described in ISO/IEC 23003-1:2007, subclause 6.6.2. Following this scheme, the bsDecorrConfig == 0 configuration should be used with a decorrelator index, $X = 8$, according to Table A.26 to Table A.29 in ISO/IEC 23003-1:2007. Hence, the $\text{decorrFunc}()$ denotes the decorrelation process:

$$\mathbf{X}_d = \begin{pmatrix} x_{1d} \\ x_{2d} \end{pmatrix} = \begin{pmatrix} \text{decorrFunc}((1 \ 0) \mathbf{P}_1 \mathbf{X}) \\ \text{decorrFunc}((0 \ 1) \mathbf{P}_1 \mathbf{X}) \end{pmatrix}.$$

In case of binaural output the upmix parameters $\mathbf{G}$ and $\mathbf{P}_2$ derived from the SAOC data, rendering information $\mathbf{M}_{\text{ren}}^{l,m}$ and HRTF parameters are applied to the downmix signal $\mathbf{X}$ (and $\mathbf{X}_d$) yielding the binaural output $\hat{\mathbf{X}}$, see Fig. 2, reference numeral 270, where the basic structure of the downmix processor is shown.

The target binaural rendering matrix $\mathbf{A}^{l,m}$ of size $2 \times N$ consists of the elements $a_{x,y}^{l,m}$. Each element $a_{x,y}^{l,m}$ is derived from HRTF parameters and rendering matrix $\mathbf{M}_{\text{ren}}^{l,m}$ with elements $m_{y,i}^{l,m}$, for example, by the SAOC parameter processor. The target binaural rendering matrix $\mathbf{A}^{l,m}$ represents the relation between all audio input objects y and the desired binaural output.

$$a_{y,1}^{l,m} = \sum_{i=0}^{N_{\text{HRTF}}-1} m_{y,i}^{l,m} H_{i,L}^m \exp\left(j \frac{\phi_i^m}{2}\right), \quad a_{y,2}^{l,m} = \sum_{i=0}^{N_{\text{HRTF}}-1} m_{y,i}^{l,m} H_{i,R}^m \exp\left(-j \frac{\phi_i^m}{2}\right).$$

The HRTF parameters are given by $H_{i,L}^m$, $H_{i,R}^m$ and ϕ_i^m for each processing band m . The spatial positions for which HRTF parameters are available are characterized by the index i . These parameters are described in ISO/IEC 23003-1:2007.

4.1.2.1 Overview

In the following, an overview over the downmix processing will be given taking reference to Figs. 4a and 4b, which show a block representation of the downmix processing, which may be performed by the audio signal processor 140 or by the combination of the SAOC

parameter processor 252 and the SAOC downmix pre-processor 270, or by the combination of the SAOC parameter processor 496 and the downmix processor 497.

Taking reference now to Fig. 4a, the downmix processing receives a rendering matrix **M**, an object level difference information OLD, an inter-object-correlation information IOC, a downmix gain information DMG and (optionally) a downmix channel level difference information DCLD. The downmix processing 400 according to Fig. 4a obtains a rendering matrix **A** on the basis of the rendering matrix **M**, for example, using a parameter adjuster and a **M**-to-**A** mapping. Also, entries of a covariance matrix **E** are obtained in dependence on the object level difference information OLD and the inter-object correlation information IOC, for example, as discussed above. Similarly, entries of a downmix matrix **D** are obtained in dependence on the downmix gain information DMG and the downmix channel level difference information DCLD.

Entries *f* of a desired covariance matrix **F** are obtained in dependence on the rendering matrix **A** and the covariance matrix **E**. Also, a scalar value *v* is obtained in dependence on the covariance matrix **E** and the downmix matrix **D** (or in dependence on the entries thereof).

Gain values P_L , P_R for two channels are obtained in dependence on entries of the desired covariance matrix **F** and the scalar value *v*. Also, an inter-channel phase difference value φ_C is obtained in dependence entries *f* of the desired covariance matrix **F**. A rotation angle α is also obtained in dependence on entries *f* of the desired covariance matrix **F**, taking into consideration, for example, a constant *c*. In addition, a second rotation angle β is obtained, for example, in dependence on the channel gains P_L , P_R and the first rotation angle α . Entries of a matrix **G** are obtained, for example, in dependence on the two channel gain values P_L, P_R and also in dependence on the inter-channel phase difference φ_C and, optionally, the rotation angles α , β . Similarly, entries of a matrix **P**₂ are determined in dependence on some or all of said values P_L , P_R , φ_C , α , β .

In the following, it will be described how the matrix **G** and/or **P**₂ (or the entries thereof), which may be applied by the downmix processor as discussed above, can be obtained for different processing modes.

4.1.2.2 Mono to Binaural “x-1-b” Processing Mode

In the following, a processing mode will be discussed in which the regular audio objects are represented by a single channel downmix signal 134, 264, 322, 497a and in which a binaural rendering is desired.

- 5 The upmix parameters $\mathbf{G}^{l,m}$ and $\mathbf{P}_2^{l,m}$ are computed as

$$\mathbf{G}^{l,m} = \begin{pmatrix} P_L^{l,m} \exp\left(j \frac{\phi_C^{l,m}}{2}\right) \cos(\beta^{l,m} + \alpha^{l,m}) \\ P_R^{l,m} \exp\left(-j \frac{\phi_C^{l,m}}{2}\right) \cos(\beta^{l,m} - \alpha^{l,m}) \end{pmatrix},$$

$$\mathbf{P}_2^{l,m} = \begin{pmatrix} P_L^{l,m} \exp\left(j \frac{\phi_C^{l,m}}{2}\right) \sin(\beta^{l,m} + \alpha^{l,m}) \\ P_R^{l,m} \exp\left(-j \frac{\phi_C^{l,m}}{2}\right) \sin(\beta^{l,m} - \alpha^{l,m}) \end{pmatrix}.$$

The gains $P_L^{l,m}$ and $P_R^{l,m}$ for the left and right output channels are

$$P_L^{l,m} = \sqrt{\max\left(\frac{f_{1,1}^{l,m}}{v^{l,m}}, \varepsilon^2\right)}, \quad P_R^{l,m} = \sqrt{\max\left(\frac{f_{2,2}^{l,m}}{v^{l,m}}, \varepsilon^2\right)}.$$

- 10 The desired covariance matrix $\mathbf{F}^{l,m}$ of size 2×2 with elements $f_{i,j}^{l,m}$ is given as

$$\mathbf{F}^{l,m} = \mathbf{A}^{l,m} \mathbf{E}^{l,m} (\mathbf{A}^{l,m})^*.$$

The scalar $v^{l,m}$ is computed as

$$v^{l,m} = \mathbf{D}^l \mathbf{E}^{l,m} (\mathbf{D}^l)^* + \varepsilon^2.$$

The inter channel phase difference $\phi_C^{l,m}$ is given as

- 15
$$\phi_C^{l,m} = \begin{cases} \arg(f_{1,2}^{l,m}), & 0 \leq m \leq 11, \quad \rho_C^{l,m} \geq 0.6, \\ 0, & \text{otherwise.} \end{cases}$$

The inter channel coherence $\rho_C^{l,m}$ is computed as

$$\rho_C^{l,m} = \min\left(\frac{|f_{1,2}^{l,m}|}{\sqrt{\max(f_{1,1}^{l,m}, f_{2,2}^{l,m}, \varepsilon^2)}}, 1\right).$$

The rotation angles $\alpha^{l,m}$ and $\beta^{l,m}$ are given as

$$\alpha^{l,m} = \begin{cases} \frac{1}{2} \arccos(\rho_C^{l,m} \cos(\arg(f_{1,2}^{l,m}))), & 0 \leq m \leq 11, \quad \rho_C^{l,m} < 0.6, \\ \frac{1}{2} \arccos(\rho_C^{l,m}), & \text{otherwise.} \end{cases}$$

$$\beta^{l,m} = \arctan\left(\tan(\alpha^{l,m}) \frac{P_R^{l,m} - P_L^{l,m}}{P_L^{l,m} + P_R^{l,m} + \varepsilon}\right).$$

5

4.1.2.3 Mono-to-Stereo "x-1-2" Processing Mode

In the following, a processing mode will be described in which the regular audio objects are represented by a single-channel signal 134, 264, 222, and in which a stereo rendering is desired.

10

In case of stereo output the "x-1-b" processing mode can be applied without using HRTF information. This can be done by deriving all elements $a_{x,y}^{l,m}$ of the rendering matrix $\mathbf{A}$, yielding:

$$15 \quad a_{1,y}^{l,m} = m_{lf,y}^{l,m}, \quad a_{2,y}^{l,m} = m_{rf,y}^{l,m}.$$

4.1.2.4 Mono-to-Mono "x-1-1" Processing Mode

In the following, a processing mode will be described in which the regular audio objects are represented by a signal channel 134, 264, 322, 497a and in which a two-channel rendering of the regular audio objects is desired.

20

In case of mono output the "x-1-2" processing mode can be applied with the following entries:

25

$$a_{1,y}^{l,m} = m_{c,y}^{l,m}, \quad a_{2,y}^{l,m} = 0$$

4.1.2.5 Stereo-to-binaural "x-2-b" processing mode

30

In the following, a processing mode will be described in which regular audio objects are represented by a two-channel signal 134, 264, 322, 497a, and in which a binaural rendering of the regular audio objects is desired.

- 5 The upmix parameters $\mathbf{G}^{l,m}$ and $\mathbf{P}_2^{l,m}$ are computed as

$$\mathbf{G}^{l,m} = \begin{pmatrix} P_L^{l,m,1} \exp\left(j \frac{\phi^{l,m,1}}{2}\right) \cos(\beta^{l,m} + \alpha^{l,m}) & P_L^{l,m,2} \exp\left(j \frac{\phi^{l,m,2}}{2}\right) \cos(\beta^{l,m} + \alpha^{l,m}) \\ P_R^{l,m,1} \exp\left(-j \frac{\phi^{l,m,1}}{2}\right) \cos(\beta^{l,m} - \alpha^{l,m}) & P_R^{l,m,2} \exp\left(-j \frac{\phi^{l,m,2}}{2}\right) \cos(\beta^{l,m} - \alpha^{l,m}) \end{pmatrix},$$

$$\mathbf{P}_2^{l,m} = \begin{pmatrix} P_L^{l,m} \exp\left(j \frac{\arg(c_{1,2}^{l,m})}{2}\right) \sin(\beta^{l,m} + \alpha^{l,m}) \\ P_R^{l,m} \exp\left(-j \frac{\arg(c_{1,2}^{l,m})}{2}\right) \sin(\beta^{l,m} - \alpha^{l,m}) \end{pmatrix}.$$

- 10 The corresponding gains $P_L^{l,m,x}$, $P_R^{l,m,x}$ and $P_L^{l,m}$, $P_R^{l,m}$ for the left and right output channels are

$$P_L^{l,m,x} = \sqrt{\max\left(\frac{f_{1,1}^{l,m,x}}{v^{l,m,x}}, \varepsilon^2\right)}, \quad P_R^{l,m,x} = \sqrt{\max\left(\frac{f_{2,2}^{l,m,x}}{v^{l,m,x}}, \varepsilon^2\right)},$$

$$P_L^{l,m} = \sqrt{\max\left(\frac{c_{1,1}^{l,m}}{v^{l,m}}, \varepsilon^2\right)}, \quad P_R^{l,m} = \sqrt{\max\left(\frac{c_{2,2}^{l,m}}{v^{l,m}}, \varepsilon^2\right)}.$$

The desired covariance matrix $\mathbf{F}^{l,m,x}$ of size 2×2 with elements $f_{u,v}^{l,m,x}$ is given as

$$\mathbf{F}^{l,m,x} = \mathbf{A}^{l,m} \mathbf{E}^{l,m,x} (\mathbf{A}^{l,m})^*.$$

- 15 The covariance matrix $\mathbf{C}^{l,m}$ of size 2×2 with elements $c_{u,v}^{l,m}$ of the “dry” binaural signal is estimated as

$$\mathbf{C}^{l,m} = \tilde{\mathbf{G}}^{l,m} \mathbf{D}^l \mathbf{E}^{l,m} (\mathbf{D}^l)^* (\tilde{\mathbf{G}}^{l,m})^*,$$

where

$$\tilde{\mathbf{G}}^{l,m} = \begin{pmatrix} P_L^{l,m,1} \exp\left(j \frac{\phi^{l,m,1}}{2}\right) & P_L^{l,m,2} \exp\left(j \frac{\phi^{l,m,2}}{2}\right) \\ P_R^{l,m,1} \exp\left(-j \frac{\phi^{l,m,1}}{2}\right) & P_R^{l,m,2} \exp\left(-j \frac{\phi^{l,m,2}}{2}\right) \end{pmatrix}.$$

The corresponding scalars $v^{l,m,x}$ and $v^{l,m}$ are computed as

$$v^{l,m,x} = \mathbf{D}^{l,x} \mathbf{E}^{l,m} (\mathbf{D}^{l,x})^* + \varepsilon^2, \quad v^{l,m} = (\mathbf{D}^{l,1} + \mathbf{D}^{l,2}) \mathbf{E}^{l,m} (\mathbf{D}^{l,1} + \mathbf{D}^{l,2})^* + \varepsilon^2.$$

The downmix matrix $\mathbf{D}^{l,x}$ of size $1 \times N$ with elements $d_i^{l,x}$ can be found as

$$d_i^{l,1} = 10^{0.05 DMG_i^l} \sqrt{\frac{10^{0.1 DCLD_i^l}}{1 + 10^{0.1 DCLD_i^l}}}, \quad d_i^{l,2} = 10^{0.05 DMG_i^l} \sqrt{\frac{1}{1 + 10^{0.1 DCLD_i^l}}}.$$

5 The stereo downmix matrix $\mathbf{D}^l$ of size $2 \times N$ with elements $d_{x,i}^l$ can be found as

$$d_{x,i}^l = d_i^{l,x}.$$

The matrix $\mathbf{E}^{l,m,x}$ with elements $e_{i,j}^{l,m,x}$ are derived from the following relationship

$$e_{i,j}^{l,m,x} = e_{i,j}^{l,m} \left(\frac{d_i^{l,x}}{d_i^{l,1} + d_i^{l,2}} \right) \left(\frac{d_j^{l,x}}{d_j^{l,1} + d_j^{l,2}} \right).$$

The inter channel phase differences $\phi_C^{l,m}$ are given as

$$10 \quad \phi_{l,m,x}^{l,m} = \begin{cases} \arg(f_{1,2}^{l,m,x}), & 0 \leq m \leq 11, \\ 0, & \text{otherwise.} \end{cases} \quad \rho_C^{l,m} > 0.6,$$

The ICCs $\rho_C^{l,m}$ and $\rho_T^{l,m}$ are computed as

$$\rho_T^{l,m} = \min \left(\frac{|f_{1,2}^{l,m}|}{\sqrt{\max(f_{1,1}^{l,m} f_{2,2}^{l,m}, \varepsilon^2)}}, 1 \right), \quad \rho_C^{l,m} = \min \left(\frac{|c_{1,2}^{l,m}|}{\sqrt{\max(c_{1,1}^{l,m} c_{2,2}^{l,m}, \varepsilon^2)}}, 1 \right).$$

The rotation angles $\alpha^{l,m}$ and $\beta^{l,m}$ are given as

$$\alpha^{l,m} = \frac{1}{2} (\arccos(\rho_T^{l,m}) - \arccos(\rho_C^{l,m})), \quad \beta^{l,m} = \arctan \left(\tan(\alpha^{l,m}) \frac{P_R^{l,m} - P_L^{l,m}}{P_L^{l,m} + P_R^{l,m}} \right).$$

15

4.1.2.6 Stereo-to-stereo “x-2-2” processing mode

In the following, a processing mode will be described in which the regular audio objects are described by a two-channel (stereo) signal 134, 264, 322, 497a and in which a 2-channel (stereo) rendering is desired.

20

In case of stereo output, the stereo preprocessing is directly applied, which will be described below in Section 4.2.2.3.

25 4.1.2.7 Stereo-to-mono “x-2-1” processing mode

In the following, a processing mode will be described in which the regular audio objects are represented by a two-channel (stereo) signal 134, 264, 322, 497a, and in which a one-channel (mono) rendering is desired.

5

In case of mono output, the stereo preprocessing is applied with a single active rendering matrix entry, as described below in Section 4.2.2.3.

4.1.2.8 Conclusion

10

Taking reference again to Figs. 4a and 4b, a processing has been described which can be applied to a 1-channel or a two-channel signal 134, 264, 322, 497a representing the regular audio objects subsequent to a separation between the extended audio objects and the regular audio objects. Figs. 4a and 4b illustrate the processing, wherein the processing of Figs. 4a and 4b differs in that an optional parameter adjustment is introduced in different stages of the processing.

15

4.2. Operation in the transcoding modes

20

4.2.1 Introduction

In the following, a method for combining SAOC parameters and panning information (or rendering information) associated with each audio object (or, preferably, with each regular audio object) in a standard compliant MPEG surround bitstream (MPS bitstream) is explained.

25

The SAOC transcoder 490 is depicted in Fig. 4f and consists of an SAOC parameter processor 491 and a downmix processor 492 applied for a stereo downmix.

30

The SAOC transcoder 490 may, for example, take over the functionality of the audio signal processor 140. Alternatively, the SAOC transcoder 490 may take over the functionality of the SAOC downmix pre-processor 270 when taken in combination with the SAOC parameter processor 252.

35

For example, the SAOC parameter processor 491 may receive an SAOC bitstream 491a, which is equivalent to the object-related parametric information 110 or the SAOC bitstream 212. Also, the SAOC parameter processor 491 may receive a rendering matrix

information 491b, which may be included in the object-related parametric information 110, or which may be equivalent to the rendering matrix information 214. The SAOC parameter processor 491 may also provide downmix processing information 491c to the downmix processor 492, which may be equivalent to the information 240. Moreover, the SAOC parameter processor 491 may provide an MPEG surround bitstream (or MPEG surround parameter bitstream) 491d, which comprises a parametric surround information which is compatible with the MPEG surround standard. The MPEG surround bitstream 491d may, for example, be part of the processed version 142 of the second audio information, or may, for example be part of or take the place of the MPS bitstream 222.

The downmix processor 492 is configured to receive a downmix signal 492a, which is preferably a one-channel downmix signal or a two-channel downmix signal, and which is preferably equivalent to the second audio information 134, or to the second audio object signal 264, 322. The downmix processor 492 may also provide an MPEG surround downmix signal 492b, which is equivalent to (or part of) the processed version 142 of the second audio information 134, or equivalent to (or part of) the processed version 272 of the second audio object signal 264.

However, there are different ways of combining the MPEG surround downmix signal 492b with the enhanced audio object signal 132, 262. The combination may be performed in the MPEG surround domain.

Alternatively, however, the MPEG surround representation, comprising the MPEG surround parameter bitstream 491d and the MPEG surround downmix signal 492b, of the regular audio objects may be converted back to a multi-channel time domain representation or a multi-channel frequency domain representation (individually representing different audio channels) by an MPEG surround decoder and may be subsequently combined with the enhanced audio object signals.

It should be noted that the transcoding modes comprise both one or more mono downmix processing modes and one or more stereo downmix processing modes. However, in the following only the stereo downmix processing mode will be described, because the processing of the regular audio object signals is more elaborate in the stereo downmix processing mode.

4.2.2 Downmix processing in the stereo downmix ("x-2-5") processing mode

4.2.2.1 Introduction

In the following section, a description of the SAOC transcoding mode for the stereo downmix case will be given.

- 5 The object parameters (object level difference OLD, inter-object correlation IOC, downmix gain DMG and downmix channel level difference DCMD) from the SAOC bitstream are transcoded into spatial (preferably channel-related) parameters (channel level difference CLD, inter-channel-correlation ICC, channel prediction coefficient CPC) for the MPEG surround bitstream according to the rendering information. The downmix is
10 modified according to object parameters and a rendering matrix.

Taking reference now to Figs. 4c, 4d and 4e, an overview of the processing, and in particular of the downmix modification, will be given.

- 15 Fig. 4c shows a block representation of a processing which is performed for modifying the downmix signal, for example the downmix signal 134, 264, 322,492a describing the one or, preferably, more regular audio objects. As can be seen from Figs. 4c, 4d and 4e, the processing receives a rendering matrix $\mathbf{M}_{\text{ren}}$, a downmix gain information DMG, a downmix channel level difference information DCLD, an object level difference
20 information OLD, and an inter-object-correlation information IOC. The rendering matrix may optionally be modified by a parameter adjustment, as it is shown in Fig. 4c. Entries of a downmix matrix $\mathbf{D}$ are obtained in dependence on the downmix gain information DMG and the downmix channel level difference information DCLD. Entries of a coherence matrix $\mathbf{E}$ are obtained in dependence on the object level difference information OLD and
25 the inter-object correlation information IOC. In addition, a matrix $\mathbf{J}$ may be obtained in dependence on the downmix matrix $\mathbf{D}$ and the coherence matrix $\mathbf{E}$, or in dependence on the entries thereof. Subsequently, a matrix $\mathbf{C}_3$ may be obtained in dependence on the rendering matrix $\mathbf{M}_{\text{ren}}$, the downmix matrix $\mathbf{D}$, the coherence matrix $\mathbf{E}$ and the matrix $\mathbf{J}$. A matrix $\mathbf{G}$ may be obtained in dependence on a matrix $\mathbf{D}_{\text{TTT}}$, which may be a matrix having
30 predetermined entries, and also in dependence on the matrix $\mathbf{C}_3$. The matrix $\mathbf{G}$ may, optionally, be modified, to obtain a modified matrix $\mathbf{G}_{\text{mod}}$. The matrix $\mathbf{G}$ or the modified version $\mathbf{G}_{\text{mod}}$ thereof may be used to derive the processed version 142, 272,492b of the second audio information 134, 264 from the second audio information 134, 264,492a (wherein the second audio information 134, 264 is designed with $\mathbf{X}$, and wherein the
35 processed version 142, 272 thereof is designated with $\hat{\mathbf{X}}$).

In the following, the rendering of the object energy, which is performed in order to obtain the MPEG surround parameters, will be discussed. Also, the stereo preprocessing, which is

performed in order to obtain the processed version 142, 272,492b of the second audio information 134, 264,492a representing the regular audio objects will be described.

4.2.2.2 Rendering of object energies

5

The transcoder determines the parameters for the MPS decoder according to the target rendering as described by the rendering matrix $\mathbf{M}_{\text{ren}}$. The six channel target covariance is denoted with $\mathbf{F}$ and given by

$$10 \quad \mathbf{F} = \mathbf{Y}\mathbf{Y}^* = \mathbf{M}_{\text{ren}} \mathbf{S}(\mathbf{M}_{\text{ren}} \mathbf{S})^* = \mathbf{M}_{\text{ren}} (\mathbf{S}\mathbf{S}^*) \mathbf{M}_{\text{ren}}^* = \mathbf{M}_{\text{ren}} \mathbf{E} \mathbf{M}_{\text{ren}}^* .$$

The transcoding process can conceptually be divided into two parts. In one part a three channel rendering is performed to a left, right and center channel. In this stage the parameters for the downmix modification as well as the prediction parameters for the TTT box for the MPS decoder are obtained. In the other part the CLD and ICC parameters for the rendering between the front and surround channels (OTT parameters, left front – left surround, right front – right surround) are determined.

4.2.2.2.1 Rendering to left, right and center channel

20

In this stage the spatial parameters are determined that control the rendering to a left and right channel, consisting of front and surround signals. These parameters describe the prediction matrix of the TTT box for the MPS decoding $\mathbf{C}_{\text{TTT}}$ (CPC parameters for the MPS decoder) and the downmix converter matrix $\mathbf{G}$.

25 $\mathbf{C}_{\text{TTT}}$ is the prediction matrix to obtain the target rendering from the modified downmix $\hat{\mathbf{X}} = \mathbf{G}\mathbf{X}$:

$$\mathbf{C}_{\text{TTT}} \hat{\mathbf{X}} = \mathbf{C}_{\text{TTT}} \mathbf{G}\mathbf{X} \approx \mathbf{A}_3 \mathbf{S} .$$

$\mathbf{A}_3$ is a reduced rendering matrix of size $3 \times N$, describing the rendering to the left, right and center channel respectively. It is obtained as $\mathbf{A}_3 = \mathbf{D}_{36} \mathbf{M}_{\text{ren}}$ with the 6 to 3 partial downmix matrix $\mathbf{D}_{36}$ defined by

30

$$\mathbf{D}_{36} = \begin{pmatrix} w_1 & 0 & 0 & 0 & w_1 & 0 \\ 0 & w_2 & 0 & 0 & 0 & w_2 \\ 0 & 0 & w_3 & w_3 & 0 & 0 \end{pmatrix} .$$

The partial downmix weights w_p , $p=1,2,3$ are adjusted such that the energy of $w_p(y_{2p-1} + y_{2p})$ is equal to the sum of energies $\|y_{2p-1}\|^2 + \|y_{2p}\|^2$ up to a limit factor.

$$w_1 = \frac{f_{1,1} + f_{5,5}}{f_{1,1} + f_{5,5} + 2f_{1,5}}, \quad w_2 = \frac{f_{2,2} + f_{6,6}}{f_{2,2} + f_{6,6} + 2f_{2,6}}, \quad w_3 = 0.5,$$

where $f_{i,j}$ denote the elements of $\mathbf{F}$.

- 5 For the estimation of the desired prediction matrix $\mathbf{C}_{\text{TTT}}$ and the downmix preprocessing matrix $\mathbf{G}$ we define a prediction matrix $\mathbf{C}_3$ of size 3×2 , that leads to the target rendering

$$\mathbf{C}_3 \mathbf{X} \approx \mathbf{A}_3 \mathbf{S}.$$

Such a matrix is derived by considering the normal equations

$$\mathbf{C}_3 (\mathbf{D} \mathbf{E} \mathbf{D}^*) \approx \mathbf{A}_3 \mathbf{E} \mathbf{D}^*.$$

- 10 The solution to the normal equations yields the best possible waveform match for the target output given the object covariance model. $\mathbf{G}$ and $\mathbf{C}_{\text{TTT}}$ are now obtained by solving the system of equations

$$\mathbf{C}_{\text{TTT}} \mathbf{G} = \mathbf{C}_3.$$

- To avoid numerical problems when calculating the term $\mathbf{J} = (\mathbf{D} \mathbf{E} \mathbf{D}^*)^{-1}$, $\mathbf{J}$ is modified. First
15 the eigenvalues $\lambda_{1,2}$ of $\mathbf{J}$ are calculated, solving $\det(\mathbf{J} - \lambda_{1,2} \mathbf{I}) = 0$.

Eigenvalues are sorted in descending ($\lambda_1 \geq \lambda_2$) order and the eigenvector corresponding to the larger eigenvalue is calculated according to the equation above. It is assured to lie in the positive x-plane (first element has to be positive). The second eigenvector is obtained from the first by a -90 degrees rotation:

$$20 \quad \mathbf{J} = (\mathbf{v}_1 \mathbf{v}_2) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} (\mathbf{v}_1 \mathbf{v}_2)^*.$$

A weighting matrix is computed from the downmix matrix $\mathbf{D}$ and the prediction matrix $\mathbf{C}_3$,
 $\mathbf{W} = (\mathbf{D} \text{diag}(\mathbf{C}_3))$.

- Since $\mathbf{C}_{\text{TTT}}$ is a function of the MPS prediction parameters c_1 and c_2 (as defined in ISO/IEC 23003-1:2007), $\mathbf{C}_{\text{TTT}} \mathbf{G} = \mathbf{C}_3$ is rewritten in the following way, to find the
25 stationary point or points of the function,

$$\mathbf{\Gamma} \begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix} = \mathbf{b},$$

with $\mathbf{\Gamma} = (\mathbf{D}_{\text{TTT}} \mathbf{C}_3) \mathbf{W} (\mathbf{D}_{\text{TTT}} \mathbf{C}_3)^*$ and $\mathbf{b} = \mathbf{GWC}_3 \mathbf{v}$,

where $\mathbf{D}_{\text{TTT}} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ and $\mathbf{v} = (1 \ 1 \ -1)$.

If $\mathbf{\Gamma}$ does not provide a unique solution ($\det(\mathbf{\Gamma}) < 10^{-3}$), the point is chosen that lies closest

5 to the point resulting in a TTT pass through. As a first step, the row i of $\mathbf{\Gamma}$ is chosen $\boldsymbol{\gamma} = [\gamma_{i,1} \ \gamma_{i,2}]$ where the elements contain most energy, thus $\gamma_{i,1}^2 + \gamma_{i,2}^2 \geq \gamma_{j,1}^2 + \gamma_{j,2}^2$, $j = 1, 2$.

Then a solution is determined such that

$$\begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 3\mathbf{y} \text{ with } \mathbf{y} = \frac{b_{i,3}}{\left(\sum_{j=1,2} (\gamma_{i,j})^2 \right) + \varepsilon} \boldsymbol{\gamma}^T.$$

10 If the obtained solution for $\tilde{c}_1$ and $\tilde{c}_2$ is outside the allowed range for prediction coefficients that is defined as $-2 \leq \tilde{c}_j \leq 3$ (as defined in ISO/IEC 23003-1:2007), $\tilde{c}_j$ shall be calculated according to below.

First define the set of points, $\mathbf{x}_p$ as:

$$\mathbf{x}_p \in \left[\begin{pmatrix} \min \left(3, \max \left(-2, -\frac{-2\gamma_{1,2} - b_1}{\gamma_{1,1} + \varepsilon} \right) \right) \\ -2 \end{pmatrix}, \begin{pmatrix} \min \left(3, \max \left(-2, -\frac{3\gamma_{1,2} - b_1}{\gamma_{1,1} + \varepsilon} \right) \right) \\ 3 \end{pmatrix} \right], \left[\begin{pmatrix} \min \left(3, \max \left(-2, -\frac{-2\gamma_{2,1} - b_2}{\gamma_{2,2} + \varepsilon} \right) \right) \\ -2 \end{pmatrix}, \begin{pmatrix} \min \left(3, \max \left(-2, -\frac{3\gamma_{2,1} - b_2}{\gamma_{2,2} + \varepsilon} \right) \right) \\ 3 \end{pmatrix} \right],$$

and the distance function,

$$15 \quad \text{distFunc}(\mathbf{x}_p) = \mathbf{x}_p^* \mathbf{\Gamma} \mathbf{x}_{p1} - 2\mathbf{b} \mathbf{x}_p.$$

Then the prediction parameters are defined according to:

$$\begin{pmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{pmatrix} = \arg \min_{\mathbf{x} \in \mathbf{x}_p} (\text{distFunc}(\mathbf{x})).$$

The prediction parameters are constrained according to:

$$c_1 = (1 - \lambda) \tilde{c}_1 + \lambda \gamma_1, \quad c_2 = (1 - \lambda) \tilde{c}_2 + \lambda \gamma_2,$$

where λ , γ_1 and γ_2 are defined as

$$\begin{aligned} \gamma_1 &= \frac{2f_{1,1} + 2f_{5,5} - f_{3,3} + f_{1,3} + f_{5,3}}{2f_{1,1} + 2f_{5,5} + 2f_{3,3} + 4f_{1,3} + 4f_{5,3}}, \\ \gamma_2 &= \frac{2f_{2,2} + 2f_{6,6} - f_{3,3} + f_{2,3} + f_{6,3}}{2f_{2,2} + 2f_{6,6} + 2f_{3,3} + 4f_{2,3} + 4f_{6,3}}, \\ 5 \quad \lambda &= \left(\frac{(f_{1,2} + f_{1,6} + f_{5,2} + f_{5,6} + f_{1,3} + f_{5,3} + f_{2,3} + f_{6,3} + f_{3,3})^2}{(f_{1,1} + f_{5,5} + f_{3,3} + 2f_{1,3} + 2f_{5,3})(f_{2,2} + f_{6,6} + f_{3,3} + 2f_{2,3} + 2f_{6,3})} \right)^8. \end{aligned}$$

For the MPS decoder, the CPCs and corresponding ICC_{TTT} are provided as follows

$$\mathbf{D}_{\text{CPC}_1} = c_1(l, m), \quad \mathbf{D}_{\text{CPC}_2} = c_2(l, m) \text{ and } \mathbf{D}_{\text{ICC}_{\text{TTT}}} = 1.$$

4.2.2.2.2 Rendering between front and surround channels

10

The parameters that determine the rendering between front and surround channels can be estimated directly from the target covariance matrix $\mathbf{F}$

$$CLD_{a,b} = 10 \log_{10} \left(\frac{\max(f_{a,a}, \varepsilon^2)}{\max(f_{b,b}, \varepsilon^2)} \right), \quad ICC_{a,b} = \frac{\max(f_{a,b}, \varepsilon^2)}{\sqrt{\max(f_{a,a}, \varepsilon^2) \max(f_{b,b}, \varepsilon^2)}},$$

15

with $(a, b) = (1, 2)$ and $(3, 4)$.

The MPS parameters are provided in the form

$$20 \quad CLD_h^{l,m} = \mathbf{D}_{\text{CLD}}(h, l, m) \text{ and } ICC_h^{l,m} = \mathbf{D}_{\text{ICC}}(h, l, m),$$

for every OTT box h .

4.2.2.3 Stereo processing

25

In the following, a stereo processing of the regular audio object signal 134 to 64, 322 will be described. The stereo processing is used to derive a process to general representation 142, 272 on the basis of a two-channel representation of the regular audio objects.

The stereo downmix $\mathbf{X}$, which is represented by the regular audio object signals 134, 264, 492a is processed into the modified downmix signal $\hat{\mathbf{X}}$, which is represented by the processed regular audio object signals 142, 272:

$$\hat{\mathbf{X}} = \mathbf{G}\mathbf{X},$$

where

$$\mathbf{G} = \mathbf{D}_{\text{TTT}}\mathbf{C}_3 = \mathbf{D}_{\text{TTT}}\mathbf{M}_{\text{ren}}\mathbf{E}\mathbf{D}^*\mathbf{J}.$$

The final stereo output from the SAOC transcoder $\hat{\mathbf{X}}$ is produced by mixing $\mathbf{X}$ with a decorrelated signal component according to:

$$\hat{\mathbf{X}} = \mathbf{G}_{\text{Mod}}\mathbf{X} + \mathbf{P}_2\mathbf{X}_d,$$

where the decorrelated signal $\mathbf{X}_d$ is calculated as described above, and the mix matrices $\mathbf{G}_{\text{Mod}}$ and $\mathbf{P}_2$ according to below.

First, define the render upmix error matrix as

$$\mathbf{R} = \mathbf{A}_{\text{diff}}\mathbf{E}\mathbf{A}_{\text{diff}}^*,$$

where

$$\mathbf{A}_{\text{diff}} = \mathbf{D}_{\text{TTT}}\mathbf{A}_3 - \mathbf{G}\mathbf{D},$$

and moreover define the covariance matrix of the predicted signal $\hat{\mathbf{R}}$ as

$$\hat{\mathbf{R}} = \begin{pmatrix} \hat{r}_{1,1} & \hat{r}_{1,2} \\ \hat{r}_{2,1} & \hat{r}_{2,2} \end{pmatrix} = \mathbf{G}\mathbf{D}\mathbf{E}\mathbf{D}^*\mathbf{G}^*.$$

The gain vector $\mathbf{g}_{\text{vec}}$ can subsequently be calculated as:

$$\mathbf{g}_{\text{vec}} = \left(\min \left(\sqrt{\max \left(\frac{\hat{r}_{1,1} + r_{1,1} + \varepsilon^2}{r_{1,1} + \varepsilon^2}, 0 \right)}, 1.5 \right), \min \left(\sqrt{\max \left(\frac{\hat{r}_{2,2} + r_{2,2} + \varepsilon^2}{r_{2,2} + \varepsilon^2}, 0 \right)}, 1.5 \right) \right),$$

and the mix matrix $\mathbf{G}_{\text{Mod}}$ is given as:

$$\mathbf{G}_{\text{Mod}} = \begin{cases} \text{diag}(\mathbf{g}_{\text{vec}})\mathbf{G}, & r_{1,2} > 0, \\ \mathbf{G}, & \text{otherwise.} \end{cases}$$

5

Similarly, the mix matrix $\mathbf{P}_2$ is given as:

$$\mathbf{P}_2 = \begin{cases} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & r_{1,2} > 0, \\ \mathbf{v}_R \text{diag}(\mathbf{W}_d), & \text{otherwise.} \end{cases}$$

10 To derive $\mathbf{v}_R$ and $\mathbf{W}_d$, the characteristic equation of $\mathbf{R}$ needs to be solved:

$$\det(\mathbf{R} - \lambda_{1,2}\mathbf{I}) = 0, \text{ giving the eigenvalues, } \lambda_1 \text{ and } \lambda_2.$$

The corresponding eigenvectors $\mathbf{v}_{R1}$ and $\mathbf{v}_{R2}$ of $\mathbf{R}$ can be calculated solving the equation system:

15

$$(\mathbf{R} - \lambda_{1,2}\mathbf{I})\mathbf{v}_{R1,R2} = 0.$$

Eigenvalues are sorted in descending ($\lambda_1 \geq \lambda_2$) order and the eigenvector corresponding to the larger eigenvalue is calculated according to the equation above. It is assured to lie in the positive x-plane (first element has to be positive). The second eigenvector is obtained from the first by a -90 degrees rotation:

20

$$\mathbf{R} = (\mathbf{v}_{R1} \mathbf{v}_{R2}) \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} (\mathbf{v}_{R1} \mathbf{v}_{R2})^*.$$

25

Incorporating $\mathbf{P}_1 = \begin{pmatrix} 1 & 1 \end{pmatrix} \mathbf{G}$, $\mathbf{R}_d$ can be calculated according to:

$$\mathbf{R}_d = \begin{pmatrix} \mathbf{r}_{d11} & \mathbf{r}_{d12} \\ \mathbf{r}_{d21} & \mathbf{r}_{d22} \end{pmatrix} = \text{diag}(\mathbf{P}_1 (\mathbf{D} \mathbf{E} \mathbf{D}^*) \mathbf{P}_1^*),$$

30 which gives

$$\mathbf{w}_{d1} = \min \left(\sqrt{\frac{\lambda_1}{r_{d1} + \varepsilon}}, 2 \right), \quad \mathbf{w}_{d2} = \min \left(\sqrt{\frac{\lambda_2}{r_{d2} + \varepsilon}}, 2 \right),$$

and finally the mix matrix,

$$\mathbf{P}_2 = (\mathbf{v}_{R1} \quad \mathbf{v}_{R2}) \begin{pmatrix} \mathbf{w}_{d1} & 0 \\ 0 & \mathbf{w}_{d2} \end{pmatrix}.$$

5

4.2.2.4 Dual mode

The SAOC transcoder can let the mix matrices $\mathbf{P}_1$, $\mathbf{P}_2$ and the prediction matrix $\mathbf{C}_3$ be calculated according to an alternative scheme for the upper frequency range. This alternative scheme is particularly useful for downmix signals where the upper frequency range is coded by a non-waveform preserving coding algorithm e.g. SBR in High Efficiency AAC.

For the upper parameter bands, defined by $\mathbf{bsTttBandsLow} \leq pb < numBands$, $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{C}_3$ should be calculated according to the alternative scheme described below:

$$\begin{cases} \mathbf{P}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \\ \mathbf{P}_2 = \mathbf{G}. \end{cases}$$

Define the energy downmix and energy target vectors, respectively:

$$\begin{cases} \mathbf{e}_{dmx} = \begin{pmatrix} e_{dmx1} \\ e_{dmx2} \end{pmatrix} = \text{diag}(\mathbf{D}\mathbf{E}\mathbf{D}^*) + \varepsilon\mathbf{I}, \\ \mathbf{e}_{tar} = \begin{pmatrix} e_{tar1} \\ e_{tar2} \\ e_{tar3} \end{pmatrix} = \text{diag}(\mathbf{A}_3\mathbf{E}\mathbf{A}_3^*), \end{cases}$$

and the help matrix

25

$$\mathbf{T} = \begin{pmatrix} t_{1,1} & t_{1,2} \\ t_{2,1} & t_{2,2} \\ t_{3,1} & t_{3,2} \end{pmatrix} = \mathbf{A}_3\mathbf{D}^* + \varepsilon\mathbf{I}.$$

Then calculate the gain vector

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix} = \begin{pmatrix} \frac{e_{\text{tar1}}}{\sqrt{t_{1,1}^2 e_{\text{dmx1}} + t_{1,2}^2 e_{\text{dmx2}}}} \\ \frac{e_{\text{tar2}}}{\sqrt{t_{2,1}^2 e_{\text{dmx1}} + t_{2,2}^2 e_{\text{dmx2}}}} \\ \frac{e_{\text{tar3}}}{\sqrt{t_{3,1}^2 e_{\text{dmx1}} + t_{3,2}^2 e_{\text{dmx2}}}} \end{pmatrix},$$

which finally gives the new prediction matrix

$$\mathbf{C}_3 = \begin{pmatrix} g_1 t_{1,1} & g_1 t_{1,2} \\ g_2 t_{2,1} & g_2 t_{2,2} \\ g_3 t_{3,1} & g_3 t_{3,2} \end{pmatrix}.$$

5. Combined EKS SAOC decoding/transcoding mode, encoder according to Fig. 10 and systems according to Figs. 5a, 5b

In the following, a brief description of the combined EKS SAOC processing scheme will be given. A preferred “combined EKS SAOC” processing scheme is proposed, where the EKS processing is integrated into the regular SAOC decoding/transcoding chain by a cascaded scheme.

5.1. Audio signal Encoder according to Fig. 5

In a first step, objects dedicated to EKS processing (enhanced Karaoke/solo processing) are identified as foreground objects (FGO) and their number N_{FGO} (also designated as N_{EAO}) is determined by a bitstream variable “bsNumGroupsFGO”. Said bitstream variable may, for example, be included in an SAOC bitstream, as described above.

For the generation of the bitstream (in an audio signal encoder), the parameters of all input objects N_{obj} are reordered such that the foreground objects FGO comprise the last N_{FGO} (or alternatively, N_{EAO}) parameters in each case, for example, OLD_i for $[N_{\text{obj}} - N_{\text{FGO}} \leq i \leq N_{\text{obj}} - 1]$.

From the remaining objects which are, for example, background objects BGO or non-enhanced audio objects, a downmix signal in the “regular SAOC style” is generated which

at the same time serves as a background object BGO. Next, the background object and the foreground objects are downmixed in the “EKS processing style” and residual information is extracted from each foreground object. This way, no extra processing steps need to be introduced. Thus, no change of the bitstream syntax is required.

5

In other words, at the encoder side, non-enhanced audio objects are distinguished from enhanced audio objects. A one-channel or two-channels regular audio objects downmix signal is provided which represents the regular audio objects (non-enhanced audio objects), wherein there may be one, two or even more regular audio objects (non-enhanced audio objects). The one-channel or two-channel regular audio object downmix signal is then
10 combined with one or more enhanced audio object signals (which may, for example, be one-channel signals or two-channel signals), to obtain a common downmix signal (which may, for example, be a one-channel downmix signal or a two-channel downmix signal) combining the audio signals of the enhanced audio objects and the regular audio object
15 downmix signal.

In the following, the basic structure of such a cascaded encoder will be briefly described taking reference to Fig. 10, which shows a block schematic representation of an SAOC
20 encoder 1000, according to an embodiment of the invention. The SAOC encoder 1000 comprises a first SAOC downmixer 1010, which is typically an SAOC downmixer which does not provide a residual information. The SAOC downmixer 1010 is configured to receive a plurality of N_{BGO} audio object signals 1012 from regular (non-enhanced) audio objects. Also, the SAOC downmixer 1010 is configured to provide a regular audio object
25 downmix signal 1014 on the basis of the regular audio objects 1012, such that the regular audio object downmix signal 1014 combines the regular audio objects signals 1012 in accordance with downmix parameters. The SAOC downmixer 1010 also provides a regular audio object SAOC information 1016, which describes the regular audio object signals and the downmix. For example, the regular audio object SAOC information 1016 may
30 comprise a downmix gain information DMG and a downmix channel level difference information DCLD describing the downmix performed by the SAOC downmixer 1010. In addition, the regular audio object SAOC information 1016 may comprise an object level difference information and an inter-object correlation information describing a relationship between the regular audio objects described by the regular audio object signal 1012.

35

The encoder 1000 also comprises a second SAOC downmixer 1020, which is typically configured to provide a residual information. The second SAOC downmixer 1020 is

preferably configured to receive one or more enhanced audio object signals 1022 and also to receive the regular audio object downmix signal 1014.

The second SAOC downmixer 1020 is also configured to provide a common SAOC downmix signal 1024 on the basis of the enhanced audio object signals 1022 and the regular audio object downmix signal 1014. When providing the common SAOC downmix signal, the second SAOC downmixer 1020 typically treats the regular audio object downmix signal 1014 as a single one-channel or two-channel object signal.

The second SAOC downmixer 1020 is also configured to provide an enhanced audio object SAOC information which describes, for example, downmix channel level difference values DCLD associated with the enhanced audio objects, object level difference values OLD associated with the enhanced audio objects and inter-object correlation values IOC associated with the enhanced audio objects. In addition, the second SAOC 1020 is preferably configured to provide residual information associated with each of the enhanced audio objects, such that the residual information associated with the enhanced audio objects describes the difference between an original individual enhanced audio object signal and an expected individual enhanced audio object signal which can be extracted from the downmix signal using the downmix information DMG, DCLD and the object information OLD, IOC.

The audio encoder 1000 is well-suited for cooperation with the audio decoder described herein.

5.2. Audio signal decoder according to Fig. 5a

In the following, the basic structure of a combined EKS SAOC decoder 500, a block schematic diagram of which is shown in Fig. 5a will be described.

The audio decoder 500 according to Fig. 5a is configured to receive a downmix signal 510, an SAOC bitstream information 512 and a rendering matrix information 514. The audio decoder 500 comprises an enhanced Karaoke/Solo processing and a foreground object rendering 520, which is configured to provide a first audio object signal 562, which describes rendered foreground objects, and a second audio object signal 564, which describes the background objects. The foreground objects may, for example, be so-called "enhanced audio objects" and the background objects may, for example, be so-called "regular audio objects" or "non-enhanced audio objects". The audio decoder 500 also comprises regular SAOC decoding 570, which is configured to receive the second audio

object signal 562 and to provide, on the basis thereof, a processed version 572 of the second audio object signal 564. The audio decoder 500 also comprises a combiner 580, which is configured to combine the first audio object signal 562 and the processed version 572 of the second audio object signal 564, to obtain an output signal 520.

5

In the following, the functionality of the audio decoder 500 will be discussed in some more detail. At the SAOC decoding/transcoding side, the upmix process results in a cascaded scheme comprising firstly an enhanced Karaoke-Solo processing (EKS processing) to decompose the downmix signal into the background object (BGO) and foreground objects (FGOs). The required object level differences (OLDs) and inter-object correlations (IOCs) for the background object are derived from the object and downmix information (which is both object-related parametric information, and which is both typically included in the SAOC bitstream):

15

$$OLD_L = \sum_{i=0}^{N-N_{FGO}-1} d_{0,i}^2 OLD_i$$

$$OLD_R = \sum_{i=0}^{N-N_{FGO}-1} d_{1,i}^2 OLD_i,$$

20

$$IOC_{LR} = \begin{cases} IOC_{0,1}, & N - N_{FGO} = 2, \\ 0, & otherwise. \end{cases}$$

25

30

In addition, this step (which is typically executed by the EKS processing and foreground object rendering 520) includes mapping the foreground objects to the final output channels (such that, for example, the first audio object signal 562 is a multi-channel signal in which the foreground objects are mapped to one or more channels each). The background object (which typically comprises a plurality of so-called “regular audio objects”) is rendered to the corresponding output channels by a regular SAOC decoding process (or, alternatively, in some cases by an SAOC transcoding process). This process may, for example, be performed by the regular SAOC decoding 570. The final mixing stage (for example, the combiner 580) provides a desired combination of rendered foreground objects and background object signals at the output.

This combined EKS SAOC system represents a combination of all beneficial properties of the regular SAOC system and its EKS mode. This approach allows to achieve the corresponding performance using the proposed system with the same bitstream for both classic (moderate rendering) and Karaoke/Solo-similar (extreme rendering) playback scenarios.

5.3. Generalized Structure according to Fig. 5b

10 In the following, a generalized structure of a combined EKS SAOC system 590 will be described taking reference to Fig. 5b, which shows a block schematic diagram of such a generalized combined EKS SAOC system. The combined EKS SAOC system 590 of Fig. 5b may also be considered as an audio decoder.

15 The combined EKS SAOC system 590 is configured to receive a downmix signal 510a, an SAOC bitstream information 512a and the rendering matrix information 514a. Also, the combined EKS SAOC system 590 is configured to provide an output signal 520a on the basis thereof.

20 The combined EKS SAOC system 590 comprises an SAOC type processing stage I 520a, which receives the downmix signal 510a, the SAOC bitstream information 512a (or at least a part thereof) and the rendering matrix information 514a (or at least a part thereof). In particular, the SAOC type processing stage I 520a receives first stage object level difference values (OLD_s). The SAOC type processing stage I 520a provides one or more
25 signals 562a describing a first set of objects (for example, audio objects of a first audio object type). The SAOC type processing stage I 520a also provides one or more signal 564a describing a second set of objects.

The combined EKS SAOC system also comprises an SAOC type processing stage II 570a,
30 which is configured to receive the one or more signals 564a describing the second set of objects and to provide, on the basis thereof, one or more signals 572a describing a third set of objects using second stage object level differences, which are included in the SAOC bitstream information 512a, and also at least a part of the rendering matrix information 514. The combined EKS SAOC system also comprises a combiner 580a, which may, for
35 example, be a summer, to provide the output signals 520a by combining the one or more signals 562a describing the first set of objects and the one or more signals 570a describing the third set of objects (wherein the third set of objects may be a processed version of the second set of objects).

To summarize the above, Fig. 5b shows a generalized form of the basic structure described with reference to Fig. 5a above in a further embodiment of the invention.

5

6. Perceptual Evaluation of the Combined EKS SAOC Processing Scheme

6.1 Test Methodology , Design and Items

10 This subjective listening tests were conducted in an acoustically isolated listening room that is designed to permit high-quality listening. The playback was done using headphones (STAX SR Lambda Pro with Lake-People D/A-Converter and STAX SRM-Monitor). The test method followed the standard procedures used in the spatial audio verification tests, based on the “multiple stimulus with hidden reference and anchors” (MUSHRA) method
15 for the subjective assessment of intermediate quality audio (see reference [7]).

A total of eight listeners participated in the performed test. All subjects can be considered experienced listeners. In accordance with the MUSHRA methodology, the listeners were instructed to compare all test conditions against the reference. The test conditions were
20 randomized automatically for each test item and for each listener. The subjective responses were recorded by a computer-based MUSHRA program on a scale ranging from 0 to 100. An instantaneous switching between the items under test was allowed. The MUSHRA test has been conducted in order to assess the perceptual performance of the considered SAOC modes and the proposed system described in the table of Fig. 6a, which provides a
25 listening test design description.

The corresponding downmix signals were coded using an AAC core-coder with a bitrate of 128 kbps. In order to assess the perceptual quality of the proposed combined EKS SAOC system, it is compared against the regular SAOC RM system (SAOC reference model
30 system) and the current EKS mode (enhanced-Karaoke-Solo mode) for two different rendering test scenarios described in the table of Fig. 6b, which describes the systems under test.

Residual coding with a bit rate of 20 kbps was applied for the current EKS mode and a
35 proposed combined EKS SAOC system. It should be noted that for the current EKS mode it is necessary to generate a stereo background object (BGO) prior to the actual encoding/decoding procedure, since this mode has limitations on the number and type of input objects.

The listening test material and the corresponding downmix and rendering parameters used in the performed tests have been selected from the set of the call-for-proposals (CfP) audio items described in the document [2]. The corresponding data for “Karaoke” and “Classic” rendering application scenarios can be found in the table of Fig. 6c, which describes listening test items and rendering matrices.

6.2 Listening Test Results

A short overview in terms of the diagrams demonstrating the obtained listening test results can be found in Figs. 6d and 6e, wherein Fig. 6d shows average MUSHRA scores for the Karaoke/Solo type rendering listening test, and Fig. 6e shows average MUSHRA scores for the classic rendering listening test. The plots show the average MUSHRA grading per item over all listeners and the statistical mean value over all evaluated items together with the associated 95% confidence intervals.

The following conclusions can be drawn based upon the results of the conducted listening tests:

- Fig. 6d represents the comparison for the current EKS mode with the combined EKS SAOC system for Karaoke-type of applications. For all tested items no significant difference (in the statistical sense) in performance between these two systems can be observed. From this observation it can be concluded that the combined EKS SAOC system is able to efficiently exploit the residual information reaching the performance of the EKS mode. One can also note that the performance of the regular SAOC system (without residual) is below both other systems.
- Fig. 6e represents the comparison of the current regular SAOC with the combined EKS SAOC system for classic rendering scenarios. For all tested items the performance of these two systems is statistically the same. This demonstrates the proper functionality of the combined EKS SAOC system for a classic rendering scenario.

Therefore, it can be concluded that the proposed unified system combining the EKS mode with the regular SAOC preserves the advantages in subjective audio quality for the corresponding types of a rendering.

Taking into account the fact that the proposed combined EKS SAOC system has no longer restrictions on the BGO object, but has entirely flexible rendering capability of the regular SAOC mode and can use the same bitstream for all types of rendering, it appears to be advantageous to incorporate it into the MPEG SAOC standard.

5

7. Method According to Fig. 7

In the following, a method for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information will be described with reference to Fig. 7, which shows a flowchart of such a method.

The method 700 comprises a step 710 of decomposing a downmix signal representation, to provide a first audio information describing a first set of one or more audio objects of a first audio object type and a second audio information describing a second set of one or more audio objects of a second audio object type in dependence on the downmix signal representation and at least a part of the object-related parametric information. The method 700 also comprises a step 720 of processing the second audio information in dependence on the object-related parametric information, to obtain a processed version of the second audio information.

The method 700 also comprises a step 730 of combining the first audio information with the processed version of the second audio information, to obtain the upmix signal representation.

25

The method 700 according to Fig. 7 may be supplemented by any of the features and functionalities which are discussed herein with respect to the inventive apparatus. Also, the method 700 brings along the advantages discussed with respect to the inventive apparatus.

8. Implementation Alternatives

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 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. Some or all of the method steps may be executed by (or using) a hardware apparatus, like for example, a microprocessor, a

35

programmable computer or an electronic circuit. In some embodiments, some one or more of the most important method steps may be executed by such an apparatus.

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

10 Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software. The implementation can be performed using a digital storage medium, for example a floppy disk, a DVD, a Blue-Ray, 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. Therefore, the digital storage medium may be computer readable.

15 Some embodiments according to the invention comprise a 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.

20 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.

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

30 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.

35 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. The data carrier, the digital storage medium or the recorded medium are typically tangible and/or non-transmitting.

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.

5

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.

10 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
15 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
20 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.

25

9. Conclusions

In the following, some aspects and advantages of the combined EKS SAOC system
30 according to the present invention will be briefly summarized. For Karaoke and Solo playback scenarios, the SAOC EKS processing mode supports both reproduction of the background objects/foreground objects exclusively and an arbitrary mixture (defined by the rendering matrix) of these object groups.

35 Also, the first mode is considered to be the main objective of EKS processing, the latter provides additional flexibility.

It has been found that a generalization of the EKS functionality consequently involves the effort of combining EKS with the regular SAOC processing mode to obtain one unified system. The potentials of such a unified system are:

- 5 • One single clear SAOC decoding/transcoding structure;
- One bitstream for both EKS and regular SAOC mode;
- 10 • No limitation to the number of input objects comprising the background object (BGO), such that there is no need to generate the background object prior to the SAOC encoding stage ; and
- Support of a residual coding for foreground objects yielding enhanced perceptual quality in demanding Karaoke/Solo playback situations.

15

These advantages can be obtained by the unified system described herein.

References

- [1] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N8853, "Call for Proposals on Spatial Audio Object Coding", 79th MPEG Meeting, Marrakech, January 2007.
- 5 [2] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N9099, "Final Spatial Audio Object Coding Evaluation Procedures and Criterion", 80th MPEG Meeting, San Jose, April 2007.
- [3] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N9250, "Report on Spatial Audio Object Coding RM0 Selection", 81st MPEG Meeting, Lausanne, July 2007.
- 10 [4] ISO/IEC JTC1/SC29/WG11 (MPEG), Document M15123, "Information and Verification Results for CE on Karaoke/Solo system improving the performance of MPEG SAOC RM0", 83rd MPEG Meeting, Antalya, Turkey, January 2008.
- [5] ISO/IEC JTC1/SC29/WG11 (MPEG), Document N10659, "Study on ISO/IEC 23003-2:200x Spatial Audio Object Coding (SAOC)", 88th MPEG Meeting, Maui, USA, April 2009.
- 15 [6] ISO/IEC JTC1/SC29/WG11 (MPEG), Document M10660, "Status and Workplan on SAOC Core Experiments", 88th MPEG Meeting, Maui, USA, April 2009.
- [7] EBU Technical recommendation: "MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality", Doc. B/AIM022, October 1999.
- 20 [8] ISO/IEC 23003-1:2007, Information technology – MPEG audio technologies – Part 1: MPEG Surround.

Claims

1. An audio signal decoder (100; 200; 500; 590) for providing an upmix signal representation in dependence on a downmix signal and representation (112; 210; 510; 510a), an object-related parametric information (110; 212; 512; 512a) the audio signal decoder comprising:
 - an object separator (130; 260; 520; 520a) configured to decompose the downmix signal representation, to provide a first audio information (132; 262; 562; 562a) describing a first set of one or more audio objects of a first audio object type, and a second audio information (134; 264; 564; 564a) describing a second set of one or more audio objects of a second audio object type in dependence on the downmix signal representation;
 - an audio signal processor configured to receive the second audio information (134; 264; 564; 564a) and to process the second audio information in dependence on the object-related parametric information, to obtain a processed version (142; 272; 572; 572a) of the second audio information; and
 - an audio signal combiner (150; 280; 580; 580a) configured to combine the first audio information with the processed version of the second audio information, to obtain the upmix signal representation.
2. The audio signal decoder according to claim 1, wherein the audio signal decoder is configured to provide the upmix signal representation in dependence on a residual information associated to a subset of audio objects represented by the downmix signal representation,
 - wherein the object separator is configured to decompose the downmix signal representation to provide the first audio information describing a first set of one or more audio objects of a first audio object type to which residual information is associated, and the second audio information describing a second set of one or more audio objects of a second audio object type, to which no residual information is associated, in dependence on the downmix signal representation and using the residual information.
3. The audio signal decoder (100; 200; 500; 590) according to claim 1 or 2, wherein the object separator is configured to provide the first audio information such that

one or more audio objects of the first audio object type are emphasized over audio objects of the second audio object type in the first audio information, and

wherein the object separator is configured to provide the second audio information such that audio objects of the second audio object type are emphasized over audio objects of the first audio object type in the second audio information.

4. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 3, wherein the audio signal decoder is configured to perform a 2-step processing, such that a processing of the second audio information in the audio signal processor (140; 270; 570; 570a) is performed subsequent to a separation between the first audio information, describing the first set of one or more audio objects of the first audio object type, and the second audio information describing the second set of one or more audio objects of the second audio object type.
5. The audio signal decoder (100; 200; 500; 570) according to one of claims 1 to 4, wherein the audio signal processor is configured to process the second audio information (134; 264; 564; 564a) in dependence on the object-related parametric information (110; 212; 512; 512a) associated with the audio objects of the second audio object type and independent from the object-related parametric information (110; 212; 512; 512a) associated with the audio objects of the first audio object type.
6. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 5, wherein the object separator is configured to obtain the first audio information (132; 262; 562; 562a, $\mathbf{X}_{EAO}$) and the second audio information (134; 264; 564; 564a, $\mathbf{X}_{OBJ}$) using a linear combination of one or more downmix signal channels of the downmix signal representation and one or more residual channels, wherein the object separator is configured to obtain combination parameters for performing the linear combination in dependence on downmix parameters associated with the audio objects of the first audio object type ($m_0 \dots m_{NEAO-1}$; $n_0 \dots n_{NEAO-1}$) and in dependence on channel prediction coefficients ($c_{j,0}$, $c_{j,1}$) of the audio objects of the first audio object type.
7. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 6, wherein the object separator is configured to obtain the first audio information and the second audio information according to

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Prediction}} \begin{pmatrix} l_0 \\ r_0 \\ \hline res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{pmatrix}$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Prediction}} \begin{pmatrix} l_0 \\ r_0 \\ \hline res_0 \\ \vdots \\ res_{N_{EAO}-1} \end{pmatrix}$$

5 wherein

$$\mathbf{M}_{\text{Prediction}} = \tilde{\mathbf{D}}^{-1} \mathbf{C}$$

wherein $\mathbf{X}_{OBJ}$ represent channels of the second audio information;

10

wherein $\mathbf{X}_{EAO}$ represent object signals of the first audio information;

wherein $\tilde{\mathbf{D}}^{-1}$ represents a matrix which is an inverse of an extended downmix matrix;

15

wherein $\mathbf{C}$ describes a matrix representing a plurality of channel prediction coefficients, $\tilde{c}_{j,0}$, $\tilde{c}_{j,1}$; wherein l_0 and r_0 represent channels of the downmix signal representation;

20

wherein $\mathbf{res}_0$ to $\mathbf{res}_{N_{EAO}-1}$ represent residual channels; and

wherein $\mathbf{A}^{EAO}$ is a EAO pre-rendering matrix.

8. The audio signal decoder according to claim 7, wherein the object separator is
 25 configured to obtain the inverse downmix matrix $\tilde{\mathbf{D}}^{-1}$ as an inverse of an extended downmix matrix $\tilde{\mathbf{D}}$ which is defined as

$$\tilde{\mathbf{D}} = \left(\begin{array}{cc|ccc} 1 & 0 & m_0 & \dots & m_{N_{EAO}-1} \\ 0 & 1 & n_0 & \dots & n_{N_{EAO}-1} \\ \hline m_0 & n_0 & -1 & \dots & 0 \\ \vdots & \vdots & 0 & \ddots & \vdots \\ m_{N_{EAO}-1} & n_{N_{EAO}-1} & 0 & \dots & -1 \end{array} \right)$$

wherein the object separator is configured to obtain the matrix C as

5

$$\mathbf{C} = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \hline c_{0,0} & c_{0,1} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N_{EAO}-1,0} & c_{N_{EAO}-1,1} & 0 & \dots & 1 \end{array} \right)$$

wherein m_0 to $m_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type;

10

wherein n_0 to $n_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type.

9. Audio signal decoder according to claim 8, wherein the object separator is configured to compute the prediction coefficients $\tilde{c}_{j,0}$ and $\tilde{c}_{j,1}$ as

15

$$\tilde{c}_{j,0} = \frac{P_{LoCo,j}P_{Ro} - P_{RoCo,j}P_{LoRo}}{P_{Lo}P_{Ro} - P_{LoRo}^2}$$

$$\tilde{c}_{j,1} = \frac{P_{RoCo,j}P_{Lo} - P_{LoCo,j}P_{LoRo}}{P_{Lo}P_{Ro} - P_{LoRo}^2}; \text{ and}$$

20

wherein the object separator is configured to derive constrained prediction coefficients $c_{j,0}$ and $c_{j,1}$ from the prediction coefficients $\tilde{c}_{j,0}$ and $\tilde{c}_{j,1}$ using a constraining algorithm, or to use the prediction coefficients $\tilde{c}_{j,0}$ and $\tilde{c}_{j,1}$ as the prediction coefficients $c_{j,0}$ and $c_{j,1}$;

wherein energy quantities P_{Lo} , P_{Ro} , P_{LoRo} , $P_{LoCo,j}$ and $P_{RoCo,j}$ are defined as

$$P_{Lo} = OLD_L + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} m_j m_k e_{j,k}$$

$$P_{Ro} = OLD_R + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} n_j n_k e_{j,k}$$

$$P_{LoRo} = e_{L,R} + \sum_{j=0}^{N_{EAO}-1} \sum_{k=0}^{N_{EAO}-1} m_j n_k e_{j,k}$$

5

$$P_{LoCo,j} = m_j OLD_L + n_j e_{L,R} - m_j OLD_j - \sum_{\substack{i=0 \\ i \neq j}}^{N_{EAO}-1} m_i e_{i,j}$$

$$P_{RoCo,j} = n_j OLD_R + m_j e_{L,R} - n_j OLD_j - \sum_{\substack{i=0 \\ i \neq j}}^{N_{EAO}-1} n_i e_{i,j}$$

wherein parameters OLD_L , OLD_R and $IOC_{L,R}$ correspond to audio objects of the second audio object type and are defined according to

10

$$OLD_L = \sum_{i=0}^{N-N_{EAO}-1} d_{0,i}^2 OLD_i,$$

$$OLD_R = \sum_{i=0}^{N-N_{EAO}-1} d_{1,i}^2 OLD_i,$$

$$IOC_{L,R} = \begin{cases} IOC_{0,1}, & N - N_{EAO} = 2, \\ 0, & otherwise. \end{cases}$$

wherein $d_{0,i}$ and $d_{1,i}$ are downmix values associated with the audio objects of the second audio object type;

15

wherein OLD_i are object level difference values associated with the audio objects of the second audio object type;

wherein N is a total number of audio objects;

20

wherein N_{EAO} is a number of audio objects of the first audio object type;

wherein $IOC_{0,1}$ is an inter-object-correlation value associated with a pair of audio objects of the second audio object type;

25

wherein $e_{i,j}$ and $e_{L,R}$ are covariance values derived from object-level-difference parameters and inter-object-correlation parameters; and

wherein $e_{i,j}$ are associated with a pair of audio objects of the 1st audio object type and $e_{L,R}$ is associated with a pair of audio objects of the second audio object type.

10. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 6, wherein the object separator is configured to obtain the first audio information and the second audio information according to

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Prediction}} \begin{pmatrix} d_0 \\ \text{res}_0 \\ \vdots \\ \text{res}_{N_{EAO}-1} \end{pmatrix}$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Prediction}} \begin{pmatrix} d_0 \\ \text{res}_0 \\ \vdots \\ \text{res}_{N_{EAO}-1} \end{pmatrix}$$

wherein

$$\mathbf{M}_{\text{Prediction}} = \tilde{\mathbf{D}}^{-1} \mathbf{C}$$

wherein $\mathbf{X}_{OBJ}$ represents a channel of the second audio information;

wherein $\mathbf{X}_{EAO}$ represent object signals of the first audio information;

wherein $\tilde{\mathbf{D}}^{-1}$ represents a matrix which is an inverse of an extended downmix matrix;

wherein $\mathbf{C}$ describes a matrix representing a plurality of channel prediction coefficients, $\tilde{c}_{j,0}$, $\tilde{c}_{j,1}$;

wherein d_0 represents a channel of the downmix signal representation; and

wherein res_0 to $\text{res}_{N_{EAO}-1}$ represent residual channels; and

wherein $\mathbf{A}^{EAO}$ is a EAO pre-rendering matrix.

11. The audio signal decoder according to claim 10, wherein the object separator is configured to obtain the inverse downmix matrix $\tilde{\mathbf{D}}^{-1}$ is an inverse of an extended downmix matrix $\tilde{\mathbf{D}}$ which is defined as

$$\tilde{\mathbf{D}} = \left(\begin{array}{c|ccc} 1 & m_0 & \dots & m_{N_{EAO}-1} \\ \hline m_0 & -1 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ m_{N_{EAO}-1} & 0 & \dots & -1 \end{array} \right)$$

wherein the object separator is configured to obtain the matrix C as

$$\mathbf{C} = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline c_0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ c_{N_{EAO}-1} & 0 & \dots & 1 \end{array} \right);$$

wherein m_0 to $m_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type.

12. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 6, wherein the object separator is configured to obtain the first audio information and the second audio information according to

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{\text{Energy}} \begin{pmatrix} l_o \\ r_o \end{pmatrix}$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{\text{Energy}} \begin{pmatrix} l_o \\ r_o \end{pmatrix}$$

wherein $\mathbf{X}_{OBJ}$ represent channels of the second audio information;

wherein $\mathbf{X}_{EAO}$ represent object signals of the first audio information;

wherein

$$\mathbf{M}_{OBJ}^{Energy} = \begin{pmatrix} \sqrt{\frac{OLD_L}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} & 0 \\ 0 & \sqrt{\frac{OLD_R}{OLD_R + \sum_{i=0}^{N_{EAO}-1} n_i^2 OLD_i}} \end{pmatrix}$$

$$\mathbf{M}_{EAO}^{Energy} = \begin{pmatrix} \sqrt{\frac{m_0^2 OLD_0}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} & \sqrt{\frac{n_0^2 OLD_0}{OLD_R + \sum_{i=0}^{N_{EAO}-1} n_i^2 OLD_i}} \\ \vdots & \vdots \\ \sqrt{\frac{m_{N_{EAO}-1}^2 OLD_{N_{EAO}-1}}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} & \sqrt{\frac{n_{N_{EAO}-1}^2 OLD_{N_{EAO}-1}}{OLD_R + \sum_{i=0}^{N_{EAO}-1} n_i^2 OLD_i}} \end{pmatrix}$$

wherein m_0 to $m_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type;

wherein n_0 to $n_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type;

wherein OLD_i are object level difference values associated with the audio objects of the first audio object type;

wherein OLD_L and OLD_R are common object level difference values associated with the audio objects of the second audio object type; and

wherein $\mathbf{A}^{EAO}$ is a EAO pre-rendering matrix.

13. The audio signal decoder according to one of claims 1 to 5, wherein the object separator is configured to obtain the first audio information and the second audio information according to

$$\mathbf{X}_{OBJ} = \mathbf{M}_{OBJ}^{Energy} (d_0)$$

$$\mathbf{X}_{EAO} = \mathbf{A}^{EAO} \mathbf{M}_{EAO}^{Energy} (d_0)$$

wherein $\mathbf{X}_{OBJ}$ represents a channel of the second audio information;

wherein $\mathbf{X}_{EAO}$ represent object signals of the first audio information;

wherein $\mathbf{d}_0$ represents a channel of the downmix signal representation;

wherein

$$\mathbf{M}_{OBJ}^{Energy} = \left(\sqrt{\frac{OLD_L}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \right)$$

$$\mathbf{M}_{EAO}^{Energy} = \begin{pmatrix} \sqrt{\frac{m_0^2 OLD_0}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \\ \vdots \\ \sqrt{\frac{m_{N_{EAO}-1}^2 OLD_{N_{EAO}-1}}{OLD_L + \sum_{i=0}^{N_{EAO}-1} m_i^2 OLD_i}} \end{pmatrix}$$

wherein m_0 to $m_{N_{EAO}-1}$ are downmix values associated with the audio objects of the first audio object type;

wherein OLD_i are object level difference values associated with the audio objects of the first audio object type;

wherein OLD_L is a common object level difference value associated with the audio objects of the second audio object type; and

wherein $\mathbf{A}^{EAO}$ is a EAO pre-rendering matrix.

14. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 13, wherein the object separator is configured to apply a rendering matrix to the first audio information (132; 262; 562; 562a) to map object signals of the first audio information onto audio channels of the upmix audio signal representation (120; 220, 222; 562; 562a).

15. The audio signal decoder (100; 200; 500; 590) according to one of claims 1 to 14, wherein the audio signal processor (140; 270; 570; 570a) is configured to perform a stereo preprocessing of the second audio information (134; 264; 564; 564a) in dependence on a rendering information (M_{ren}), an object-related covariance information (**E**), a downmix information (**D**), to obtain audio channels of the processed version of the second audio information;
16. The audio signal decoder (100; 200; 500; 590) according to claim 15, wherein the audio signal processor (140; 270; 570; 570a) is configured to perform the stereo processing to map an estimated audio object contribution (ED^*JX) of the second audio information (134; 264; 564; 564a) onto a plurality of channels of the upmix audio signal representation in dependence on a rendering information and a covariance information.
17. The audio signal decoder according to claim 15 or claim 16, wherein the audio signal processor is configured to add a decorrelated audio signal contribution (P_2X_d) to the second audio information, or an information derived from the second audio information, in dependence on a render upmix error information (**R**) and one or more decorrelated-signal-intensity scaling values (w_{d1} , w_{d2}).
18. The audio signal decoder according to one of claims 1 to 14, wherein the audio signal processor (140; 270; 570; 570a) is configured to perform a postprocessing of the second audio information (134; 264; 564; 564a) in dependence on a rendering information (**A**), an object-related covariance information (**E**) and a downmix information (**D**).
19. The audio signal decoder according to claim 18, wherein the audio signal processor is configured to perform a mono-to-binaural processing of the second audio information, to map a single channel of the second audio information onto two channels of the upmix signal representation, taking into consideration a head-related transfer function.
20. The audio signal decoder according to claim 18, wherein the audio signal processor is configured to perform a mono-to-stereo processing of the second audio information, to map a single channel of the second audio information onto two channels of the upmix signal representation.

21. The audio signal decoder according to claim 18, wherein the audio signal processor is configured to perform a stereo-to-binaural processing of the second audio information, to map two channels of the second audio information onto two channels of the upmix signal representation, taking into consideration a head-related transfer function.
22. The audio signal decoder according to claim 18, wherein the audio signal processor is configured to perform a stereo-to-stereo processing of the second audio information, to map two channels of the second audio information onto two channels of the upmix signal representation.
23. The audio signal decoder according to one of claims 1 to 22, wherein the object separator is configured to treat audio objects of the second audio object type, to which no residual information is associated, as a single audio object, and
- wherein the audio signal processor (140; 270; 570; 570a) is configured to consider object-specific rendering parameters associated to the audio objects of the second audio object type to adjust contributions of the audio objects of the second audio object type to the upmix signal representation.
24. The audio signal decoder according to one of claims 1 to 23, wherein the object separator is configured to obtain one or two common object level difference values (OLD_L , OLD_R) for a plurality of audio objects of the second audio object type; and
- wherein the object separator is configured to use the common object level difference value for a computation of channel prediction coefficients (CPC); and
- wherein the object separator is configured to use the channel prediction coefficients to obtain one or two audio channels representing the second audio information.
25. The audio signal decoder according to one of claims 1 to 24, wherein the object separator is configured to obtain one or two common object level difference values (OLD_L , OLD_R) for a plurality of audio objects of the second audio object type, and
- wherein the object separator is configured to use the common object level difference value for a computation of entries of an matrix ($\mathbf{M}$); and

wherein the object separator is configured to use the matrix (**M**) to obtain one or more audio channels representing the second audio information.

26. The audio signal decoder according to one of claims 1 to 25, wherein the object separator is configured to selectively obtain a common inter-object correlation value ($IOC_{L,R}$) associated to the audio object of the second audio object type in dependence on the object-related parametric information if it is found that there are two audio objects of the second audio object type, and to set the inter-object correlation value associated to the audio objects of the second audio object type to zero if it is found that there are more or less than two audio objects of the second audio object type; and

wherein the object separator is configured to use the common inter-object correlation value for a computation of entries of an matrix (**M**); and

wherein the object separator is configured to use the common inter-object correlation value associated to the audio objects of the second audio object type to obtain the one or more audio channels representing the second audio information.

27. The audio signal decoder according to one of claims 1 to 26, wherein the audio signal processor is configured to render the second audio information in dependence on the object-related parametric information, to obtain a rendered representation of the audio objects of the second audio object type as the processed version of the second audio information.

28. The audio signal decoder according to one of claims 1 to 27, wherein the object separator is configured to provide the second audio information such that the second audio information describes more than two audio objects of the second audio object type.

29. The audio signal decoder according to claim 28, wherein the object separator is configured to obtain, as the second audio information, a one-channel audio signal representation or a two-channel audio signal representation representing more than two audio objects of the second audio object type.

30. The audio signal decoder according to one of claims 1 to 29, wherein the audio signal processor is configured to receive the second audio information and to process the second audio information in dependence of the object-related

parametric information, taking into consideration object-related parametric information associated with more than two audio objects of the second audio object type.

5 31. The audio signal decoder according to one of claims 1 to 30, wherein the audio signal decoder is configured to extract a total object number information (bsNumObjects) and a foreground object number information (bsNumGroupsFGO) from a configuration information (SAOCSpecificConfig) of the object-related
10 parametric information, and to determine the number of audio objects of the second audio object type by forming a difference between the total object number information and the foreground object number information.

15 32. The audio signal decoder according to one of claims 1 to 31, wherein the object separator is configured to use object-related parametric information associated with N_{EAO} audio objects of the first audio object type to obtain, as the first audio information, N_{EAO} audio signals ($\mathbf{X}_{\text{EAO}}$) representing the N_{EAO} audio objects of the first audio object type and to obtain, as the second audio information, one or two audio signals ($\mathbf{X}_{\text{OBJ}}$) representing the $N - N_{\text{EAO}}$ audio objects of the second audio object type, treating the $N - N_{\text{EAO}}$ audio objects of the second audio object type as a
20 single one-channel or a two-channel audio object; and

wherein the audio signal processor is configured to individually render the $N - N_{\text{EAO}}$ audio objects represented by the one or two audio signals of the second audio information using the object-related parametric information associated with the $N - N_{\text{EAO}}$ audio objects of the second audio object type.
25

30 33. A method for providing an upmix signal representation in dependence on a downmix signal representation and an object-related parametric information, the method comprising:

decomposing the downmix signal representation, to provide a first audio information describing a first set of one or more audio objects of a first audio object type, and a second audio information describing a second set of one or more audio objects of a second audio object type in dependence on the downmix signal representation and using at least a part of the object-related parametric information;
35 and

processing the second audio information in dependence on the object-related parametric information, to obtain a processed version of the second audio information; and

5 combining the first audio information with the processed version of the second audio information, to obtain the upmix signal representation.

34. A computer program for performing the method according to claim 33 when the computer program runs on a computer.