United States Patent Application Publication	20250266048
Kind Code	A1
Publication Date	August 21, 2025
Inventor(s)	Hirvonen; Toni et al.

RECONSTRUCTION OF AUDIO SCENES FROM A DOWNMIX

Abstract

Audio objects are associated with positional metadata. A received downmix signal comprises downmix channels that are linear combinations of one or more audio objects and are associated with respective positional locators.

In a first aspect, the downmix signal, the positional metadata and frequency-dependent object gains are received. An audio object is reconstructed by applying the object gain to an upmix of the downmix signal in accordance with coefficients based on the positional metadata and the positional locators.

In a second aspect, audio objects have been encoded together with at least one bed channel positioned at a positional locator of a corresponding downmix channel. The decoding system receives the downmix signal and the positional metadata of the audio objects. A bed channel is reconstructed by suppressing the content representing audio objects from the corresponding downmix channel on the basis of the positional locator of the corresponding downmix channel.

Inventors: **Hirvonen; Toni (Helsinki, FI), Purnhagen; Heiko (Sundbyberg, SE), Samuelsson; Leif Jonas (Sundbyberg, SE), Villemoes; Lars (Stockholm, SE)**

Applicant: **DOLBY INTERNATIONAL AB** (Dublin, IE)

Family ID: **1000008576924**

Assignee: **DOLBY INTERNATIONAL AB (Dublin, IE)**

Appl. No.: **19/066143**

Filed: **February 27, 2025**

Related U.S. Application Data

parent US continuation 18540546 20231214 parent-grant-document US 12243542 child US 19066143
parent US continuation 18167204 20230210 parent-grant-document US 11894003 child US 18540546

parent US continuation 17219911 20210401 parent-grant-document US 11580995 child US 18167204
parent US continuation 15584553 20170502 parent-grant-document US 10290304 child US 16380879
parent US continuation 14893377 20151123 parent-grant-document US 9666198 WO continuation PCT/EP2014/060732 20140523 child US 15584553
parent US division 16380879 20190410 parent-grant-document US 10971163 child US 17219911
us-provisional-application US 61827469 20130524

Publication Classification

Int. Cl.: **G10L19/008** (20130101); **G10L19/02** (20130101); **G10L19/20** (20130101); **G10L25/06** (20130101); **H04S3/00** (20060101); **H04S3/02** (20060101); **H04S5/00** (20060101); **H04S7/00** (20060101)

U.S. Cl.:

CPC **G10L19/008** (20130101); **G10L19/0204** (20130101); **G10L19/20** (20130101); **G10L25/06** (20130101); **H04S3/008** (20130101); **H04S3/02** (20130101); **H04S5/00** (20130101); **H04S7/30** (20130101); H04S2400/03 (20130101); H04S2400/11 (20130101); H04S2420/03 (20130101)

Background/Summary

CROSS-REFERENCE TO RELATED APPLICATIONS [0001] This application is a continuation of U.S. application Ser. No. 18/540,546, filed Dec. 14, 2023, which is a continuation of U.S. application Ser. No. 18/167,204, filed Feb. 10, 2023, which is a continuation of U.S. application Ser. No. 17/219,911, filed Apr. 1, 2021 (now Issued U.S. Pat. No. 11,580,995), which is a divisional of U.S. application Ser. No. 16/380,879 filed Apr. 10, 2019, now U.S. Pat. No. 10,971,163 issued on Apr. 6, 2021, which is a continuation of U.S. application Ser. No. 15/584,553 filed May 2, 2017, now U.S. Pat. No. 10,290,304 issued on May 14, 2019, which is a continuation of U.S. patent application Ser. No. 14/893,377 filed Nov. 23, 2015, now U.S. Pat. No. 9,666,198 issued on May 30, 2017, which is a U.S. 371 National Phase entry from PCT/EP2014/060732 filed May 23, 2014, which claims priority to U.S. Provisional Patent Application No. 61/827,469 filed May 24, 2013, which are all hereby incorporated by reference in their entirety.

TECHNICAL FIELD
[0002] The invention disclosed herein generally relates to the field of encoding and decoding of audio. In particular it relates to encoding and decoding of an audio scene comprising audio objects.
[0003] The present disclosure is related to U.S. Provisional application No. 61/827,246 filed on the same date as the present application, entitled “Coding of Audio Scenes”, and naming Heiko Purnhagen et al., as inventors is hereby included by reference in its entirety.
BACKGROUND
[0004] There exist audio coding systems for parametric spatial audio coding. For example, MPEG Surround describes a system for parametric spatial coding of multichannel audio. MPEG SAOC (Spatial Audio Object Coding) describes a system for parametric coding of audio objects.
[0005] On an encoder side these systems typically downmix the channels/objects into a downmix, which typically is a mono (one channel) or a stereo (two channels) downmix, and extract side information describing the properties of the channels/objects by means of parameters like level

differences and cross-correlation. The downmix and the side information are then encoded and sent to a decoder side. At the decoder side, the channels/objects are reconstructed, i.e. approximated, from the downmix under control of the parameters of the side information.
[0006] A drawback of these systems is that the reconstruction is typically mathematically complex and often has to rely on assumptions about properties of the audio content that is not explicitly described by the parameters sent as side information. Such assumptions may for example be that the channels/objects are treated as uncorrelated unless a cross-correlation parameter is sent, or that the downmix of the channels/objects is generated in a specific way.
[0007] In addition to the above, coding efficiency emerges as a key design factor in applications intended for audio distribution, including both network broadcasting and one-to-one file transmission. Coding efficiency is of some relevance also to keep file sizes and required memory limited, at least in non-professional products.

Description

BRIEF DESCRIPTION OF THE DRAWINGS
[0008] In what follows, example embodiments will be described with reference to the accompanying drawings, on which:
[0009] FIG. **1** is a generalized block diagram of an audio encoding system receiving an audio scene with a plurality of audio objects (and possibly bed channels as well) and outputting a downmix bitstream and a metadata bitstream;
[0010] FIG. **2**A illustrates a detail of a method for reconstructing bed channels; more precisely, it is a time-frequency diagram showing different signal portions in which signal energy data are computed in order to accomplish Wiener-type filtering;
[0011] FIG. **2**B is another time-frequency diagram showing different signal portions in which signal energy data are computed in order to accomplish Wiener-type filtering;
[0012] FIG. **2**C is another time-frequency diagram showing different signal portions in which signal energy data are computed in order to accomplish Wiener-type filtering;
[0013] FIG. **3** is a generalized block diagram of an audio decoding system, which reconstructs an audio scene on the basis of a downmix bitstream and a metadata bitstream;
[0014] FIG. **4** shows a detail of an audio encoding system configured to code an audio object by an object gain;
[0015] FIG. **5** shows a detail of an audio encoding system which computes said object gain while taking into account coding distortion;
[0016] FIG. **6** shows example virtual positions of downmix channels ({right arrow over (z)}.sub.1, . . . , {right arrow over (z)}.sub.M), bed channels ({right arrow over (x)}.sub.1, {right arrow over (x)}.sub.2) and audio objects ({right arrow over (x)}.sub.3, . . . , {right arrow over (x)}.sub.7) in relation to a reference listening point; and
[0017] FIG. **7** illustrates an audio decoding system particularly configured for reconstructing a mix of bed channels and audio objects.
[0018] All the figures are schematic and generally show parts to elucidate the subject matter herein, whereas other parts may be omitted or merely suggested. Unless otherwise indicated, like reference numerals refer to like parts in different figures.
DETAILED DESCRIPTION
[0019] As used herein, an audio signal may refer to a pure audio signal, an audio part of a video signal or multimedia signal, or an audio signal part of a complex audio object, wherein an audio object may further comprise or be associated with positional or other metadata. The present disclosure is generally concerned with methods and devices for converting from an audio scene into a bitstream encoding the audio scene (encoding) and back (decoding or reconstruction). The

conversions are typically combined with distribution, whereby decoding takes place at a later point in time than encoding and/or in a different spatial location and/or using different equipment. In the audio scene to be encoded, there is at least one audio object. The audio scene may be considered segmented into frequency bands (e.g., B=11 frequency bands, each of which includes a plurality of frequency samples) and time frames (including, say, 64 samples), whereby one frequency band of one time frame forms a time/frequency tile. A number of time frames, e.g., 24 time frames, may constitute a super frame. A typical way to implement such time and frequency segmentation is by windowed time-frequency analysis (example window length: 640 samples), including well-known discrete harmonic transforms.

1. OVERVIEW—CODING BY OBJECT GAINS

[0020] In an example embodiment within a first aspect, there is provided a method for encoding an audio scene whereby a bitstream is obtained. The bitstream may be partitioned into a downmix bitstream and a metadata bitstream. In this example embodiment, signal content in several (or all) frequency bands in one time frame is encoded by a joint processing operation, wherein intermediate results from one processing step are used in subsequent steps affecting more than one frequency band.

[0021] The audio scene comprises a plurality of audio objects. Each audio object is associated with positional metadata. A downmix signal is generated by forming, for each of a total of M downmix channels, a linear combination of one or more of the audio objects. The downmix channels are associated with respective positional locators.

[0022] For each audio object, the positional metadata associated with the audio object and the spatial locators associated with some or all the downmix channels are used to compute correlation coefficients. The correlation coefficients may coincide with the coefficients which are used in the downmixing operation where the linear combinations in the downmix channels are formed; alternatively, the downmixing operation uses an independent set of coefficients. By collecting all non-zero correlation coefficients relating to the audio object, it is possible to upmix the downmix signal, e.g., as the inner product of a vector of the correlation coefficients and the M downmix channels. In each frequency band, the upmix thus obtained is adjusted by a frequency-dependent object gain, which preferably can be assigned different values with a resolution of one frequency band. This is accomplished by assigning a value to the object gain in such manner that the upmix of the downmix signal rescaled by the gain approximates the audio object in that frequency band; hence, even if the correlation coefficients are used to control the downmixing operation, the object gain may differ between frequency band to improve the fidelity of the encoding. This may be accomplished by comparing the audio object and the upmix of the downmix signal in each frequency band and assigning a value to the object gain that provides a faithful approximation. The bitstream resulting from the above encoding method encodes at least the downmix signal, the positional metadata and the object gains.

[0023] The method according to the above example embodiment is able to encode a complex audio scene with a limited amount of data, and is therefore advantageous in applications where efficient, particularly bandwidth-economical, distribution formats are desired.

[0024] The method according to the above example embodiment preferably omits the correlation coefficients from the bitstream. Instead, it is understood that the correlation coefficients are computed on the decoder side, on the basis of the positional metadata in the bitstreams and the positional locators of the downmix channels, which may be predefined.

[0025] In an example embodiment, the correlation coefficients are computed in accordance with a predefined rule. The rule may be a deterministic algorithm defining how positional metadata (of audio objects) and positional locators (of downmix channels) are processed to obtain the correlation coefficients. Instructions specifying relevant aspects of the algorithm and/or implementing the algorithm in processing equipment may be stored in an encoder system or other entity performing the audio scene encoding. It is advantageous to store an identical or equivalent

copy of the rule on the decoder side, so that the rule can be omitted from the bitstream to be transmitted from the encoder to the decoder side.

[0026] In a further development of the preceding example embodiment, the correlation coefficients may be computed on the basis of the geometric positions of the audio objects, in particular their geometric positions relative to the audio objects. The computation may take into account the Euclidean distance and/or the propagation angle. In particular, the correlation coefficients may be computed on the basis of an energy preserving panning law (or pan law), such as the sine-cosine panning law. Panning laws and particularly stereo panning laws, are well known in the art, where they are used for source positioning. Panning laws notably include assumptions on the conditions for preserving constant power or apparent constant power, so that the loudness (or perceived auditory level) can be kept the same or approximately so when an audio object changes its position.

[0027] In an example embodiment, the correlation coefficients are computed by a model or algorithm using only inputs that are constant with respect to frequency. For instance, the model or algorithm may compute the correlation coefficients based on the spatial metadata and the spatial locators only. Hence, the correlation coefficients will be constant with respect to frequency in each time frame. If frequency-dependent object gains are used, however, it is possible to correct the upmix of the downmix channels at frequency-band resolution so that the upmix of the downmix channels approximates the audio object as faithfully as possible in each frequency band.

[0028] In an example embodiment, the encoding method determines the object gain for at least one audio object by an analysis-by-synthesis approach. More precisely, it includes encoding and decoding the downmix signal, whereby a modified version of the downmix signal is obtained. An encoded version of the downmix signal may already be prepared for the purpose of being included in the bitstream forming the final result of the encoding. In audio distribution systems or audio distribution methods including both encoding of an audio scene as a bitstream and decoding of the bitstream as an audio scene, the decoding of the encoded downmix signal is preferably identical or equivalent to the corresponding processing on the decoder side. In these circumstances, the object gain may be determined in order to rescale the upmix of the reconstructed downmix channels (e.g., an inner product of the correlation coefficients and a decoded encoded downmix signal) so that it faithfully approximates the audio object in the time frame. This makes it possible to assign values to the object gains that reduce the effect of coding-induced distortion.

[0029] In an example embodiment, an audio encoding system comprising at least a downmixer, a downmix encoder, an upmix coefficient analyzer and a metadata encoder is provided. The audio encoding system is configured to encode an audio scene so that a bitstream is obtained, as explained in the preceding paragraphs.

[0030] In an example embodiment, there is provided a method for reconstructing an audio scene with audio objects based on a bitstream containing a downmix signal and, for each audio object, an object gain and positional metadata associated with the audio object. According to the method, correlation coefficients—which may be said to quantify the spatial relatedness of the audio object and each downmix channel—are computed based on the positional metadata and the spatial locators of the downmix channels. As discussed and exemplified above, it is advantageous to compute the correlation coefficients in accordance with a predetermined rule, preferably in a uniform manner on the encoder and decoder side. Likewise, it is advantageous to store the spatial locators of the downmix channels on the decoder side rather than transmitting them in the bitstream. Once the correlation coefficients have been computed, the audio object is reconstructed as an upmix of the downmix signal in accordance with the correlation coefficients (e.g., an inner product of the correlation coefficients and the downmix signal) which is rescaled by the object gain. The audio objects may then optionally be rendered for playback in multi-channel playback equipment.

[0031] Alone, the decoding method according to this example embodiment realizes an efficient decoding process for faithful audio scene reconstruction based on a limited amount of input data.

Together with the encoding method previously discussed, it can be used to define an efficient distribution format for audio data.

[0032] In an example embodiment, the correlation coefficients are computed on the basis only of quantities without frequency variation in a single time frame (e.g., positional metadata of audio objects). Hence, each correlation coefficient will be constant with respect to frequency. Frequency variations in the encoded audio object can be captured by the use of frequency-dependent object gains.

[0033] In an example embodiment, an audio decoding system comprising at least a metadata decoder, a downmix decoder, an upmix coefficient decoder and an upmixer is provided. The audio decoding system is configured to reconstruct an audio scene on the basis of a bitstream, as explained in the preceding paragraphs.

[0034] Further example embodiments include: a computer program for performing an encoding or decoding method as described in the preceding paragraphs; a computer program product comprising a computer-readable medium storing computer-readable instructions for causing a programmable processor to perform an encoding or decoding method as described in the preceding paragraphs; a computer-readable medium storing a bitstream obtainable by an encoding method as described in the preceding paragraphs; a computer-readable medium storing a bitstream, based on which an audio scene can be reconstructed in accordance with a decoding method as described in the preceding paragraphs. It is noted that also features recited in mutually different claims can be combined to advantage unless otherwise stated.

II. OVERVIEW—CODING OF BED CHANNELS

[0035] In an example embodiment within a second aspect, there is provided a method for reconstructing an audio scene on the basis of a bitstream comprising at least a downmix signal with M downmix channels. Downmix channels are associated with positional locators, e.g., virtual positions or directions of preferred channel playback sources. In the audio scene, there is at least one audio object and at least one bed channel. Each audio object is associated with positional metadata, indicating a fixed (for a stationary audio object) or momentary (for a moving audio object) virtual position. A bed channel, in contrast, is associated with one of the downmix channels and may be treated as positionally related to that downmix channel, which will from time to time be referred to as a corresponding downmix channel in what follows. For practical purposes, it may therefore be considered that a bed channel is rendered most faithfully where the positional locator indicates, namely, at the preferred location of a playback source (e.g., loudspeaker) for a downmix channel. As a further practical consequence, there is no particular advantage in defining more bed channels than there are available downmix channels. In summary, the position of an audio object can be defined and possibly modified over time by way of the positional metadata, whereas the position of a bed channel is tied to the corresponding bed channel and thus constant over time.

[0036] It is assumed in this example embodiment that each channel in the downmix signal in the bitstream comprises a linear combination of one or more of the audio object(s) and the bed channel(s), wherein the linear combination has been computed in accordance with downmix coefficients. The bitstream forming the input of the present decoding method comprises, in addition to the downmix signal, either the positional metadata associated with the audio objects (the decoding method can be completed without knowledge of the downmix coefficients) or the downmix coefficients controlling the downmixing operation. To reconstruct a bed channel on the basis of its corresponding downmix channel, said positional metadata (or downmix coefficients) are used in order to suppress that content in the corresponding downmix channel which represents audio objects. After suppression, the downmix channel contains bed channel content only, or is at least dominated by bed channel content. Optionally, after these processing steps, the audio objects may be reconstructed and rendered, along with the bed channels, for playback in multi-channel playback equipment.

[0037] Alone, the decoding method according to this example embodiment realizes an efficient

decoding process for faithful audio scene reconstruction based on a limited amount of input data. Together with the encoding method to be discussed below, it can be used to define an efficient distribution format for audio data.

[0038] In various example embodiments, the object-related content to be suppressed is reconstructed explicitly, so that it would be renderable for playback. Alternatively, the object-related content is obtained by a process designed to return an incomplete representation estimation which is deemed sufficient in order to perform the suppression. The latter may be the case where the corresponding downmix channel is dominated by bed channel content, so that the suppression of the object-related content represents a relatively minor modification. In the case of explicit reconstruction, one or more of the following approaches may be adopted: [0039] a) auxiliary signals capturing at least some of the N audio objects are received at the decoding end, as described in detail in the related U.S. provisional application (titled “Coding of Audio Scenes”) initially referenced, which auxiliary signals can then be suppressed from the corresponding downmix channel; [0040] b) a reconstruction matrix is received at the decoding end, as described in detail in the related U.S. provisional application (titled “Coding of Audio Scenes”) initially referenced, which matrix permits reconstruction of the N audio objects from the M downmix signals, while possibly relying on auxiliary channels as well; [0041] c) the decoding end receives object gains for reconstructing the audio objects based on the downmix signal, as described in this disclosure under the first aspect. The gains can be used together with downmix coefficients extracted from the bitstream, or together with downmix coefficients that are computed on the basis of the positional locators of the downmix channels and the positional metadata associated with the audio objects.

[0042] Various example embodiments may involve suppression of object-related content to different extents. One option is to suppress as much object-related content as possible, preferably all object-related content. Another option is to suppress a subset of the total object-related content, e.g., by an incomplete suppression operation, or by a suppression operation restricted to suppressing content that represents fewer than the full number of audio objects contributing to the corresponding downmix channel. If fewer audio objects than the full number are (attempted to be) suppressed, these may in particular be selected according to their energy content. Specifically, the decoding method may order the objects according to decreasing energy content and select so many of the strongest objects for suppression that a threshold value on the energy of the remaining object-related content is met; the threshold may be a fixed maximal energy of the object-related content or may be expressed as a percentage of the energy of the corresponding downmix channel after suppression has been performed. A still further option is to take the effect of auditory masking into account. Such an approach may include suppression of the perceptually dominating audio objects whereas content emanating from less noticeable audio objects—in particular audio objects that are masked by other audio objects in the signal—may be left in the downmix channel without inconvenience.

[0043] In an example embodiment, the suppression of the object-related content from the downmix channel is accompanied—preferably preceded—by a computation (or estimation) of the downmix coefficients that were applied to the audio objects when the downmix signal—in particular the corresponding downmix channel—was generated. The computation is based on the positional metadata, which are associated with the objects and received in the bitstream, and further on the positional locator of the corresponding downmix channel. (It is noted that in this second aspect, unlike the first aspect, it is assumed that the downmix coefficients that controlled the downmixing operation on the encoder side are obtainable once the positional locators of the downmix channels and the positional metadata of the audio objects are known.) If the downmix coefficients were received as part of the bitstream, there is clearly no need to compute the downmix coefficients in this manner. Next, the energy of the contribution of the audio objects to the corresponding downmix channel, or at least the energy of the contribution of a subset of the audio objects to the corresponding downmix channel, is computed based on the reconstructed audio objects or based on

the downmix coefficients and the downmix signal. The energy is estimated by considering the audio objects jointly, so that the effect of statistical correlation (generally a decrease) is captured Alternatively, if in a given use case it is reasonable to assume that the audio objects are substantially uncorrelated or approximately uncorrelated, the energy of each audio object is estimated separately. The energy estimation may either proceed indirectly, based on the downmix channels and the downmix coefficients together, or directly, by first reconstructing the audio objects. A further way in which the energy of each object could be obtained is as part of the incoming bitstream. After this stage, there is available, for each bed channel, an estimated energy of at least one of those audio objects that provide a non-zero contribution to the corresponding downmix channel, or an estimate of the total energy of two or more contributing audio objects considered jointly. The energy of the corresponding downmix channel is estimated as well. The bed channel is then reconstructed by filtering the corresponding downmix channel, with the estimated energy of at least one audio object as further inputs.

[0044] In an example embodiment, the computation of the downmix coefficients referred to above preferably follows a predefined rule applied in a uniform fashion on the encoder and decoder side. The rule may be a deterministic algorithm defining how positional metadata (of audio objects) and positional locators (of downmix channels) are processed to obtain the downmix coefficients. Instructions specifying relevant aspects of the algorithm and/or implementing the algorithm in processing equipment may be stored in an encoder system or other entity performing the audio scene encoding. It is advantageous to store an identical or equivalent copy of the rule on the decoder side, so that the rule can be omitted from the bitstream to be transmitted from the encoder to the decoder side.

[0045] In a further development of the preceding example embodiment, the downmix coefficients are computed on the basis of the geometric positions of the audio objects, in particular their geometric positions relative to the audio objects. The computation may take into account the Euclidean distance and/or the propagation angle. In particular, the downmix coefficients may be computed on the basis of an energy preserving panning law (or pan law), such as the sine-cosine panning law. As mentioned above, panning laws and stereo panning laws in particular, are well known in the art, where they are used, inter alia, for source positioning. Panning laws notably include assumptions on the conditions for preserving constant power or apparent constant power, so that the perceived auditory level remains the same when an audio object changes its position.

[0046] In an example embodiment, the suppression of the object-related content from the downmix channel is preceded by a computation (or estimation) of the downmix coefficients that were applied to the audio objects when the downmix signal—and the corresponding downmix channel in particular—was generated. The computation is based on the positional metadata, which are associated with the objects and received in the bitstream, and further on the positional locator of the corresponding downmix channel. If the downmix coefficients were received as part of the bitstream, there is clearly no need to compute the downmix coefficients in this manner. Next, the audio objects—or at least each audio object that provides a non-zero contribution to the downmix channels associated with the relevant bed channels to be reconstructed—are reconstructed and their energies are computed. After this stage, there is available, for each bed channel, the energy of each contributing audio object as well as the corresponding downmix channel itself. The energy of the corresponding downmix channel is estimated. The bed channel is then reconstructed by rescaling the corresponding downmix channel, namely by applying a scaling factor which is based on the energies of the audio objects, the energy of the corresponding downmix channel and the downmix coefficients controlling contributions from the audio objects to the corresponding downmix channel. The following is an example way of computing the scaling factor h.sub.n on the basis of the energy (E[Y.sub.n]) of the corresponding downmix channel, the energy (E[S.sub.n.sup.2], n=N.sub.B+1, . . . , N) of each audio object and the downmix coefficients (d.sub.n,N.sub.B.sub.+1, d.sub.n,N.sub.B.sub.+2, . . . , d.sub.n,N applied to the audio objects:

[00001] $h_n = (\max\{\ ,\ 1 - \frac{\sum_{n=N_B+1}^{N} d_{m,n}^2 E[S_n^2]}{E[Y_n^2]}\})$

Here, ε≤0 and γ∈[0.5, 1] are constants. Preferably, ε=0 and γ=0.5. In different example embodiments, the energies may be computed for different sections of the respective signals. Basically, the time resolution of the energies may be one time frame or a fraction (subdivision) of a time frame. The energies may refer to a particular frequency band or collection of frequency bands, or the entire frequency range, i.e., the total energy for all frequency bands. As such, the scaling factor h.sub.n may have one value per time frame (i.e., may be a broadband quantity, cf. FIG. **2**A), or one value per time/frequency tile (cf. FIG. **2**B) or more than one value per time frame, or more than one value per time/frequency tile (cf. FIG. **2**C). It may be advantageous to use a finer granularity (increasing the number of independent values per unit time) for bed channel reconstruction than for audio object reconstruction, wherein the latter may be performed on the basis of object gains assuming one value per time/frequency tile, see above under the first aspect. Similarly, the positional metadata have a granularity of one time frame, i.e., the duration of one time/frequency tile. One such advantage is the improved ability to handle transient signal content, particularly if the relationship between audio objects and bed channels is changing on a short time scale.

[0047] In an example embodiment, the object-related content is suppressed by signal subtraction in the time domain or the frequency domain. Such signal subtraction may be a constant-gain subtraction of the waveform of each audio object from the waveform of the corresponding downmix channel; alternatively, the signal subtraction amounts to subtracting transform coefficients of each audio object from corresponding transform coefficients of the corresponding downmix channel, again with constant gain in each time/frequency tile. Other example embodiments may instead rely on a spectral suppression technique, wherein the energy spectrum (or magnitude spectrum) of the bed channel is substantially equal to the difference of the energy spectrum of the corresponding downmix channel and the energy spectrum of each audio object that is subject to the suppression. Put differently, a spectral suppression technique may leave the phase of the signal unchanged but attenuate its energy. In implementations acting on time-domain or frequency-domain representations of the signals, spectral suppression may require gains that are time- and/or frequency-dependent. Techniques for determining such variable gains are well known in the art and may be based on an estimated phase difference between the respective signals and similar considerations. It is noted that in the art, the term spectral subtraction is sometimes used as a synonym of spectral suppression in the above sense.

[0048] In an example embodiment, an audio decoding system comprising at least a downmix decoder, a metadata decoder and an upmixer is provided. The audio decoding system is configured to reconstruct an audio scene on the basis of a bitstream, as explained in the preceding paragraphs.

[0049] In an example embodiment, there is provided a method for encoding an audio scene, which comprises at least one audio object and at least one bed channel, as a bitstream that encodes a downmix signal and the positional metadata of the audio objects. In this example embodiment, it is preferred to encode at least one time/frequency tile at a time. The downmix signal is generated by forming, for each of a total of M downmix channels, a linear combination of one or more of the audio objects and any bed channel associated with the respective downmix channel. The linear combination is formed in accordance with downmix coefficients, wherein each such downmix coefficients that is to be applied to the audio objects is computed on the basis of a positional locator of a downmix channel and positional metadata associated with an audio object. The computation preferably follows a predefined rule, as discussed above.

[0050] It is understood that the output bitstream comprises data sufficient to reconstruct the audio objects at an accuracy deemed sufficient in the use case concerned, so that the audio objects may be suppressed from the corresponding bed channel. The reconstruction of the object-related content either is explicit, so that the audio objects would in principle be renderable for playback, or is done

by an estimation process returning an incomplete representation sufficient to perform the suppression. Particularly advantageous approaches include: [0051] a) including auxiliary signals, containing at least some of the N audio objects, in the bitstream; [0052] b) including a reconstruction matrix, which permits reconstruction of the N audio objects from the M downmix signals (and optionally from the auxiliary signals as well), in the bitstream; [0053] c) including object gains, as described in this disclosure under the first aspect, in the bitstream.

[0054] The method according to the above example embodiment is able to encode a complex audio scene—such as one including both positionable audio objects and static bed channels—with a limited amount of data, and is therefore advantageous in applications where efficient, particularly bandwidth-economical, distribution formats are desired.

[0055] In an example embodiment, an audio encoding system comprising at least a downmixer, a downmix encoder and a metadata encoder is provided. The audio encoding system is configured to encode an audio scene in such manner that a bitstream is obtained, as explained in the preceding paragraphs.

[0056] Further example embodiments include: a computer program for performing an encoding or decoding method as described in the preceding paragraphs; a computer program product comprising a computer-readable medium storing computer-readable instructions for causing a programmable processor to perform an encoding or decoding method as described in the preceding paragraphs; a computer-readable medium storing a bitstream obtainable by an encoding method as described in the preceding paragraphs; a computer-readable medium storing a bitstream, based on which an audio scene can be reconstructed in accordance with a decoding method as described in the preceding paragraphs. It is noted that also features recited in mutually different claims can be combined to advantage unless otherwise stated.

III. EXAMPLE EMBODIMENTS

[0057] The technological context of the present invention can be understood more fully from the related U.S. provisional application (titled “Coding of Audio Scenes”) initially referenced.

[0058] FIG. **1** schematically shows an audio encoding system **100**, which receives as its input a plurality of audio signals S.sub.n representing audio objects (and bed channels, in some example embodiments) to be encoded and optionally rendering metadata (dashed line), which may include positional metadata. A downmixer **101** produces a downmix signal Y with M>1 downmix channels by forming linear combinations of the audio objects (and bed channels), Y=Σ.sub.n=1.sup.N d.sub.nS.sub.n, wherein the downmix coefficients applied may be variable and more precisely influenced by the rendering metadata. The downmix signal Y is encoded by a downmix encoder (not shown) and the encoded downmix signal Y.sub.c is included in an output bitstream from the encoding system **1**. An encoding format suited for this type of applications is the Dolby Digital Plus™ (or Enhanced AC-3) format, notably its 5.1 mode, and the downmix encoder may be a Dolby Digital Plus™-enabled encoder. Parallel to this, the downmix signal Y is supplied to a time-frequency transform **102** (e.g., a QMF analysis bank), which outputs a frequency-domain representation of the downmix signal, which is then supplied to an up mix coefficient analyzer **104**. The upmix coefficient analyzer **104** further receives a frequency-domain representation of the audio objects S.sub.n(k, l), where k is an index of a frequency sample (which is in turn included in one of B frequency bands) and l is the index of a time frame, which has been prepared by a further time-frequency transform **103** arranged upstream of the upmix coefficient analyzer **104**. The upmix coefficient analyzer **104** determines upmix coefficients for reconstructing the audio objects on the basis of the downmix signal on the decoder side. Doing so, the upmix coefficient analyzer **104** may further take the rendering metadata into account, as the dashed incoming arrow indicates. The upmix coefficients are encoded by an upmix coefficient encoder **106**. Parallel to this, the respective frequency-domain representations of the downmix signal Y and the audio objects are supplied, together with the upmix coefficients and possibly the rendering metadata, to a correlation analyzer **105**, which estimates statistical quantities (e.g., cross-covariance E[S.sub.n(k, l)S.sub.n′(k, l)],

n≠n′) which it is desired to preserve by taking appropriate correction measures at the decoder side. Results of the estimations in the correlation analyzer **105** are fed to a correlation data encoder **107** and combined with the encoded upmix coefficients, by a bitstream multiplexer **108**, into a metadata bitstream P constituting one of the outputs of the encoding system **100**.

[0059] FIG. **4** shows a detail of the audio encoding system **100**, more precisely the inner workings of the upmix coefficients analyzer **104** and its relationship with the downmixer **101**, in an example embodiment within the first aspect. In the example embodiment shown, the encoding system **100** receives N audio objects (and no bed channels), and encodes the N audio objects in terms of the downmix signal Y and, in a further bitstream P, spatial metadata {right arrow over (x)}.sub.n associated with the audio objects and N object gains g.sub.n. The upmix coefficients analyzer **104** includes a memory **401**, which stores spatial locators {right arrow over (z)}.sub.m of the downmix channels, a downmix coefficient computation unit **402** and an object gain computation unit **403**. The downmix coefficient computation unit **402** stores a predefined rule for computing the downmix coefficients (preferably producing the same result as a corresponding rule stored in an intended decoding system) on the basis of the spatial metadata {right arrow over (x)}.sub.n, which the encoding system **100** receives as part of the rendering metadata, and the spatial locators {right arrow over (z)}.sub.m. In normal circumstances, each of the downmix coefficients thus computed is a number less than or equal to one, d.sub.m,n≤1, m=1, . . . , M, n=1 . . . , N, or less than or equal to some other absolute constant. The downmix coefficients may also be computed subject to an energy conservation rule or panning rule, which implies a uniform upper bound on the vector d.sub.n=[d.sub.n,1 d.sub.n,2 . . . d.sub.n,m].sup.T applied to each given audio object S.sub.n, such as ‖d.sub.n‖ C uniformly for all n=1, . . . , N, wherein normalization may ensure ‖d.sub.n‖=C. The downmix coefficients are supplied to both the downmixer **101** and the object gain computation unit **403**. The output of the downmixer **101** may be written as the sum Y=Σ.sub.i=1.sup.Nd.sub.lS.sub.l. In this example embodiment, the downmix coefficients are broadband quantities, whereas the object gains g.sub.n can be assigned an independent value for each frequency band. The object gain computation unit **403** compares each audio object S.sub.n with the estimate that will be obtained from the upmix at the decoder side, namely

[00002] $$d_n^T Y = d_n^T \cdot \sum_{l=1}^{N} d_l S_l = \cdot \sum_{l=1}^{N} (d_n^T d_l) S_l \,.$$

Assuming ‖d.sub.l‖=C for all l=1, . . . , N, then d.sub.n.sup.Td.sub.l≤C.sup.2 with equality for l=n, that is, the dominating coefficient will be the one multiplying S.sub.n. The signal d.sub.n.sup.TY may however include contributions from the other audio objects as well, and the impact of these further contributions may be limited by an appropriate choice of the object gain g.sub.n. More precisely, the object gain computation unit **403** assigns a value to the object gain g.sub.n such that

[00003] $$S_n \approx g_n (C^2 S_n + \cdot \sum_{\substack{l=1 \\ l \neq n}}^{N} (d_n^T d_l) S_l)$$

in the time/frequency tile.

[0060] FIG. **5** shows a further development of the encoder system **100** of FIG. **4**. Here, the object gain computation unit **403** (within the upmix coefficients analyzer **104**) is configured to compute the object gains by comparing each audio objects S.sub.n not with an upmix d.sub.n.sup.TY of the downmix signal Y, but with an upmix d.sub.n.sup.T{tilde over (Y)} of a restored downmix signal {tilde over (Y)}. The restored downmix signal is obtained by using the output of a downmix encoder **501**, which receives the output from the downmixer **101** and prepares the bitstream with the encoded downmix signal. The output Y.sub.c of the downmix encoder **501** is supplied to a downmix decoder **502** mimicking the action of a corresponding downmix decoder on the decoding side. It is advantageous to use an encoder system according to FIG. **5** when the downmix decoder **501** performs lossy encoding, as such encoding will introduce coding noise (including quantization distortion), which can be compensated to some extent by the object gains g.sub.n.

[0061] FIG. **3** schematically shows a decoding system **300** designed to cooperate, on a decoding

side, with an encoding system of any of the types shown in FIG. **1**, **4** or **5**. The decoding system **300** receives a metadata bitstream P and a downmix bitstream Y. Based on the downmix bitstream Y, a time-frequency transform **302** (e.g., a QMF analysis bank) prepares a frequency-domain representation of the downmix signal and supplies this to an upmixer **304**. The operations in the upmixer **304** are controlled by upmix coefficients, which it receives from a chain of metadata processing components. More precisely, an upmix coefficient decoder **306** decodes the metadata bitstream and supplies its output to an arrangement performing interpolation—and possibly transient control—of the upmix coefficients. In some example embodiments, values of the upmix coefficients are given at discrete points in time, and interpolation may be used to obtain values applying for intermediate points in time. The interpolation may be of a linear, quadratic, spline or higher-order type, depending on the requirements in a specific use case. Said interpolation arrangement comprises a buffer **309**, configured to delay the received upmix coefficients by a suitable period of time, and an interpolator **310** for deriving the intermediate values based on a current and a previous given upmix coefficient value. Parallel to this, a correlation control data decoder **307** decodes the statistical quantities estimated by the correlation analyzer **105** and supplies the decoded data to an object correlation controller **305**. To summarize, the downmix signal Y undergoes time—frequency transformation in the time—frequency transform **302**, is upmixed into signals representing audio objects in the upmixer **304**, which signals are then corrected so that the statistical characteristics—as measured by the quantities estimated by the correlation analyzer **105**—are in agreement with those of the audio objects originally encoded. A frequency-time transform **311** provides the final output of the decoding system **300**, namely, a time-domain representation of the decoded audio objects, which may then be rendered for playback.

[0062] FIG. **7** shows a further development of the audio decoding system **300**, notably with an ability to reconstruct an audio scene that includes bed channels S.sub.n, n=1, . . . , N.sub.B in addition to audio objects S.sub.n, n=N.sub.B+1, . . . , N. From an incoming bitstream, a multiplexer **701** extracts and decodes: a downmix signal Y, energies of the audio objects E[S.sub.n.sup.2], n=N.sub.B+1, . . . , N, object gains associated with the audio objects g.sub.n, n=N.sub.B+1, . . . , N, and positional metadata {right arrow over (x)}, n=N.sub.B+1, . . . , N, associated with the audio objects. The bed channels are reconstructed on the basis of their corresponding downmix channel signals by suppressing object-related content therein, in accordance with the second aspect, wherein the audio objects are reconstructed by upmixing the downmix signal using an upmix matrix U determined based on the object gains, according to the first aspect. A downmix coefficient reconstruction unit **703** uses positional locators {right arrow over (z)}.sub.m, m=1, . . . M, of the downmix channels, the positional locators being retrieved from a connected memory **702**, and the positional metadata to compute, according to a predefined rule, the restore the downmix coefficients d.sub.m,n used on the encoding side. The downmix coefficients computed by the downmix coefficient reconstruction unit **703** are used for two purposes. Firstly, they are multiplied row-wise by the object gains and arranged as an upmix matrix

[00004] $$U = \begin{bmatrix} g_1 d_{1,1} & g_1 d_{2,1} & \text{.Math.} & g_1 d_{M,1} \\ g_2 d_{1,2} & g_2 d_{2,2} & \text{.Math.} & g_2 d_{M,2} \\ \text{.Math.} & \text{.Math.} & \ddots & \text{.Math.} \\ g_N d_{1,N} & g_N d_{2,N} & \text{.Math.} & g_N d_{M,N} \end{bmatrix},$$

which is then provided to an upmixer **705**, which applies the elements of matrix U to the downmix channels to reconstruct the audio objects. Parallel to this, the downmix coefficients are supplied from the downmix coefficient reconstruction unit **703** to a Wiener filter **707** after being multiplied by the energies of the audio objects. Between the multiplexer **701** and a further input of the Wiener filter **707**, there is provided an energy estimator **706** for computing the energy E[Y.sub.m.sup.2], m=1, . . . , N.sub.B of each downmix channel that is associated with a bed channel. Based on this

information, the Wiener filter **707** internally computes a scaling factor

[00005] $h_n = (\max\{\ ,1 - \frac{\text{.Math.}_{n=N_B+1}^{N} d_{m,n}^2 E[S_n^2]}{E[Y_n^2]}\})\ \ ,n = L = 1,\ \text{.Math.}\ ,N_B\ ,$

with constant ε≤0 and 0.5≤γ≤1, and applies this to the corresponding downmix channel, so as to reconstruct the bed channel as Ŝ.sub.n=h.sub.nY.sub.n, n=1, . . . , N.sub.B. In summary, the decoding system shown in FIG. **7** outputs reconstructed signals corresponding to all audio objects and all bed channels, which may subsequently be rendered for playback in multichannel equipment. The rendering may additionally rely on the positional metadata associated with the audio objects and the positional locators associated with the downmix channels.

[0063] In comparison with the baseline audio decoding system **300** shown in FIG. **3**, it may be considered that unit **705** in FIG. **7** fulfils the duties of units **302**, **304** and **311** therein, units **702**, **703** and **704** fulfil the duties (but with a different task distribution) of units **306**, **309** and **310**, whereas units **706** and **707** represent functionality not present in the baseline system, and no component corresponding to units **305** and **307** in the baseline system has been drawn explicitly in FIG. **7**. In a variation to the example embodiment shown in FIG. **7**, the energies of the audio objects could be estimated by computing the energies E[Ŝ.sub.n], n=N.sub.B+1, . . . , N, of the reconstructed audio objects output from the upmixer **705**. This way, at the price of a certain amount of additional computational power spent in the decoding system, the bitrate of the transmitted bitstream can be decreased.

[0064] Furthermore, it is recalled that the computation of the energies of the downmix channels and the energies of the audio objects (or reconstructed audio objects) may be performed with a granularity with respect to time/frequency than the time/frequency tiles into which the audio signals are segmented. The granularity may be coarser with respect to frequency (as illustrated by FIG. **2**A), equal to the time/frequency tile segmentation (FIG. **2**B) or finer with respect to time (FIG. **2**C). In FIG. **2**, time frames are denoted T.sub.1, T.sub.2, T.sub.3, . . . and frequency bands denoted F.sub.1, F.sub.2, F.sub.3, . . . , whereby a time/frequency tile may be referred to by the pair (T.sub.l, F.sub.k). In FIG. **2**C, which shows a finer time granularity, a second index is used to refer to subdivisions of a time frame, such as T.sub.4,1, T.sub.4,2, T.sub.4,3, T.sub.4,4 in an example case where time frame T.sub.4 is subdivided into four subframes.

[0065] FIG. **7** illustrates an example geometry of bed channels and audio channels, wherein bed channels are tied to the virtual positions of downmix channels, while it is possible to define (and redefine over time) the positions of audio objects, which are then encoded as positional metadata. FIG. **7** (where (M, N, N.sub.B)=(5, 7, 2)) shows the virtual positions of the downmix channels, in accordance with their respective positional locators {right arrow over (z)}.sub.1, . . . , {right arrow over (z)}.sub.M, which coincide with the positions of bed channels S.sub.1, S.sub.2. The positions of these bed channels have been denoted {right arrow over (x)}.sub.1, {right arrow over (x)}.sub.2, but it is emphasized they do not necessarily form part of the positional metadata; rather, as already discussed above, it is sufficient to transmit the positional metadata associated with the audio objects only. FIG. **7** further shows a snapshot for a given point in time of the positions {right arrow over (x)}.sub.3, . . . , {right arrow over (x)}.sub.7 of the audio objects, as expressed by the positional metadata.

IV. EQUIVALENTS, EXTENSIONS, ALTERNATIVES AND MISCELLANEOUS

[0066] Further example embodiments will become apparent to a person skilled in the art after studying the description above. Even though the present description and drawings disclose embodiments and examples, the scope is not restricted to these specific examples. Numerous modifications and variations can be made without departing from the scope, which is defined by the accompanying claims. Any reference signs appearing in the claims are not to be understood as limiting their scope.

[0067] The systems and methods disclosed hereinabove may be implemented as software, firmware, hardware or a combination thereof. In a hardware implementation, the division of tasks

between functional units referred to in the above description does not necessarily correspond to the division into physical units; to the contrary, one physical component may have multiple functionalities, and one task may be carried out by several physical components in cooperation. Certain components or all components may be implemented as software executed by a digital signal processor or microprocessor, or be implemented as hardware or as an application-specific integrated circuit. Such software may be distributed on computer readable media, which may comprise computer storage media (or non-transitory media) and communication media (or transitory media). As is well known to a person skilled in the art, the term computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by a computer. Further, it is well known to the skilled person that communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media.

Claims

1. A method for reconstructing a time frame of an audio scene with at least a plurality of N audio signals from a bitstream, the method comprising: receiving the bitstream comprising the N audio signals signal, wherein N>1; decoding a downmix signal from the bitstream, the downmix signal comprising M downmix channels, wherein M>1 and each downmix channel is associated with a spatial locator of a plurality of spatial locators; and reconstructing the N audio signals based on as an inner product of a plurality of correlation coefficients and the downmix signal, wherein the correlation coefficients were predetermined.
2. A computer program product comprising a non-transitory computer-readable medium encoded with instructions configured to cause one or more processing devices to perform the method of claim 1.
3. An audio decoding system configured to reconstruct a time frame of an audio scene with at least a plurality of N audio signals from a bitstream, the system comprising: a receiver for receiving the bitstream comprising the N audio signals signal, wherein N>1; a decoder for decoding a downmix signal from the bitstream, the downmix signal comprising M downmix channels, wherein M>1 and each downmix channel is associated with a spatial locator of a plurality of spatial locators; and a reconstructor for reconstructing the N audio signals based on as an inner product of a plurality of correlation coefficients and the downmix signal, wherein the correlation coefficients were predetermined.