(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 690 621 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
29.01.2014 Bulletin 2014/05

(51) Int Cl.:
G10L 19/008 $^{(2013.01)}$ *G10L 19/02* $^{(2013.01)}$

(21) Application number: 12305914.9

(22) Date of filing: 26.07.2012

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME

(71) Applicant: Thomson Licensing
92130 Issy-les-Moulineaux (FR)

(72) Inventors:
- Wuebbolt, Oliver
30559 Hannover (DE)
- Murtaza, Adrian
200082 Craiova Dolj (RO)

(74) Representative: Hartnack, Wolfgang
Deutsche Thomson OHG
European Patent Operations
Karl-Wiechert-Allee 74
30625 Hannover (DE)

(54) Method and Apparatus for downmixing MPEG SAOC-like encoded audio signals at receiver side in a manner different from the manner of downmixing at encoder side

(57) Content providers can add metadata to audio content such that consumers can control down-mix or dynamic range of selected parts of the audio signal. MPEG Spatial Audio Object Coding (SAOC) deals with parametric coding techniques for complex audio scenes at bit rates normally used for mono or stereo sound coding, offering at decoder side an interactive rendering of the audio objects mixed into the audio scene, whereby only a small amount of extra information is added to the audio bit stream. In parametric coding the biggest issue is that a perfect separation between objects is usually not possible. This issue is treated in the MPEG SAOC standard by using residual coding techniques and ensuring better separation only for a small set of objects before encoding. The invention describes how, by adding only a small amount of extra information to SAOC parameters, at receiver side a remixing of a broadcast audio signal is achieved, by using information about the actual mix of the audio signal, audio signal characteristics like correlation, and the desired audio scene rendering.

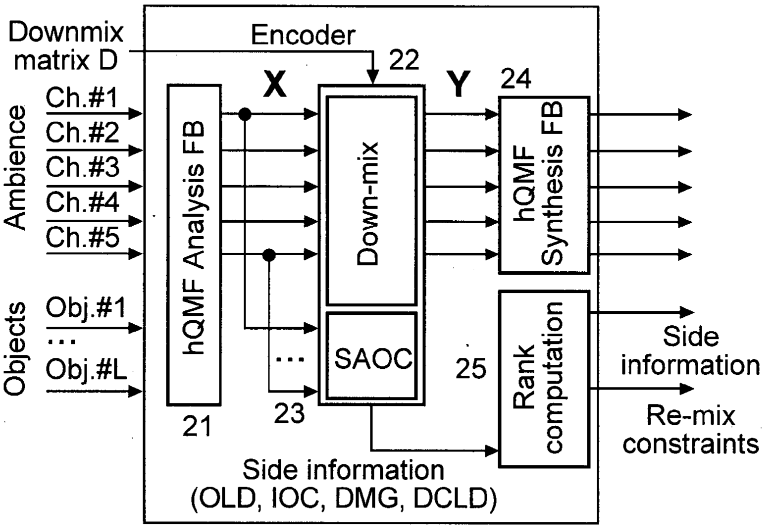

Fig. 2

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] The invention relates to a method and to an apparatus for downmixing MPEG SAOC-like encoded audio signals at receiver side in a manner different from the manner of downmixing at encoder side, wherein the decoder side downmixing is controlled by desired playback configuration data and/or desired object positioning data.

Background

[0002] Content providers are facing consumers with increasingly heterogeneous listening situations, e.g. home theatres, mobile audio, car and in-flight entertainment. Audio content cannot be processed by its creator or broadcaster so as to match every possible consumer listening condition, for example audio/video content played back on a mobile phone. Besides different listening conditions, also different listening experiences can be desirable, for instance in a live soccer broadcast a consumer can control his own virtual position within the sound scene of the stadium (pitch or stands), or can control the virtual position and the predominance of the commentator.

[0003] Content providers can add guiding metadata to the audio content, such that consumers can control down-mix or dynamic range of selected parts of the audio signal and/or assure high speech intelligibility. For incorporation of such metadata into existing broadcasting chains, it is important that the general audio format is not changed (legacy playback) and that only a small amount of extra information (e.g. as ancillary data) is added to the audio bit stream. MPEG Spatial Audio Object Coding (SAOC), ISO/IEC 23003-1:2007, MPEG audio technologies - Part 1: MPEG Surround, and ISO/IEC 23003-2:2010, MPEG audio technologies - Part 2: Spatial Audio Object Coding, deals with parametric coding techniques for complex audio scenes at bit rates normally used for mono or stereo sound coding, offering at decoder side an interactive rendering of the audio objects mixed into the audio scene.

[0004] MPEG SAOC was developed starting from Spatial Audio Coding which is based on a 'channel-oriented' approach, by introducing the concept of audio objects, having the purpose to offer even more flexibility at receiver side. Since it is a parametric multiple object coding technique, the additional cost in terms of bit rate is limited to 2-3 kbit/s for each audio object. Although the bit rate increases with the number of audio objects, it still remains at small values in comparison with the actual audio data transmitted as a mono/stereo downmix.

[0005] In parametric coding the biggest issue is that a perfect separation between objects is usually not possible, especially in an extreme mixing scenario, for example when all sound objects are mixed in all channels, or in case of 'applause-like' sound. This issue is treated in the MPEG SAOC standard by using residual coding techniques and ensuring better separation only for a small set of objects before encoding.

[0006] A standard MPEG SAOC architecture is illustrated in Fig. 1. A variable number N of sound objects Obj.#1, Obj.#2, Obj.#3, Obj.#4 is input to an SAOC encoder 11 that provides an encoded background compatible mono or stereo downmix signal or signals with SAOC parameters and side information. The downmix data signal or signals and the side information data are sent as dedicated bit streams to an SAOC decoder 12. The receiver side processing is carried out basically in two steps: first, using the side information, SAOC decoder 12 reconstructs the original sound objects Obj.#1, Obj.#2, Obj.#3, Obj.#4. Controlled by interaction/control information, these reconstructed sound objects are re-mixed and rendered in an MPEG surround renderer 13. For example, the reconstructed sound objects are output as a stereo signal with channels #1 and #2.

[0007] In practice, in order to avoid changing twice the temporal domain and the frequency domain, the steps are merged, thereby substantially reducing the computational complexity. The processing in SAOC encoder 11, SAOC decoder 12 and renderer 13 is performed in the frequency domain, using a nonuniform scale in order to model most efficiently the human auditory system. To ensure compatibility with MPEG Surround, a hybrid Quadrature Mirror Filter hQMF 21 shown in Fig. 2 was chosen, which filter offers a better resolution for low frequencies. After applying the hQMF filter to all input signals, a time/frequency grid is obtained that maps every time slot or frame of the input signals to a number of processing bands obtained by merging the frequency bands.

[0008] The following SAOC parameters are computed for every time/frequency tile and are transmitted to the decoder side:

- Object Level Differences data OLD describe the amount of energy contained by each sound object with respect to the sound object having the highest energy. The energy level value of the loudest object is described by the Object Energy parameter data NRG, which can be transmitted to the decoder;
- Inter-Object Coherence data IOC are used to describe the amount of similarity between the sound objects and are computed for every pair of two input audio objects;
- Downmix Gains data DMG and Downmix Channel Level Differences data DCLD are used for describing the gains applied to each sound object in the mixing process, and at decoder side they are used for reconstructing the downmixing matrix.

[0009] At decoder side, in order to avoid generation of bad quality audio content due to extreme rendering, a Distortion Control Unit DCU can be used. The final rendering matrix coefficients are computed as a linear combination of user-specified coefficients and the target coefficients which are assumed to be distortion-free.

Invention

[0010] The main drawback of the solution offered by MPEG SAOC is the limitation to a maximum of two down-mix channels. Further, the MPEG SAOC standard is not designed for 'Solo/Karaoke' type of applications, which involve the complete suppression of one or more audio objects. In the MPEG SAOC standard this problem is tackled by using residual coding for specific audio objects, thereby increasing the bit rate.

[0011] A problem to be solved by the invention is to overcome these limitations of MPEG SAOC and to allow for adding of side information for a legacy multi-channel audio broadcasting like 5.1. This problem is solved by the methods disclosed in claims 1 and 2. Apparatuses that utilise these methods are disclosed in claims 3 and 4, respectively.

[0012] The invention describes how, by adding only a small amount of extra bit rate, at decoder or receiver side a re-mixing of a broadcast audio signal is achieved using information about the actual mix of the audio signal, audio signal characteristics like correlation, level differences, and the desired audio scene rendering.

[0013] A second embodiment shows how to determine already at encoder side the suitability of the actual multi-channel audio signal for a remix at decoder side. This feature allows for countermeasures (e.g. changing the mixing matrix used, i.e. how the sound objects are mixed into the different channels) if a decoder side re-mixing without perceivable artefacts, or without problems like the necessary additional transmission of the audio objects itself for a short time, is not possible.

[0014] Advantageously, due to processing the same type of side information parameters, the invention could be used to amend the MPEG SAOC standard correspondingly, based on the same building blocks.

[0015] In principle, the inventive encoding method is suited for downmixing spatial audio signals that can be downmixed at receiver side in a manner different from the manner of downmixing at encoder side, wherein said encoding is based on MPEG SAOC and said downmixing at receiver side can be controlled by desired playback configuration data and/or desired object positioning data, said method including the steps:

- processing *M* correlated sound signals, *M* being greater than '2', and *L* independent sound signals, *L* being '1' or greater, in an analysis filter bank providing corresponding time/frequency domain signals;
- multiplying said time/frequency domain signals with a downmix matrix $D_{l,m}$, followed by processing the resulting signals in a synthesis filter bank that has an inverse operation of said analysis filter bank and that provides *M* time domain output signals;
- determining from said time/frequency domain signals MPEG SAOC side information data including Object Level Differences data *OLD* and Inter-Object Coherence data *IOC,* as well as enhanced Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD,* wherein said *DMG* and *DCLD* data are related to *M* channels.

[0016] In principle, the inventive decoding method is suited downmixing spatial audio signals processed according to the encoding method in a manner different from the manner of downmixing at encoder side, wherein said downmixing at receiver side can be controlled by desired playback configuration data and/or desired object positioning data, said method including the steps:

- receiving said processed spatial audio signals and processing them in an analysis filter bank, providing corresponding time/frequency domain signals;
- determining from said desired playback configuration data and/or said desired object positioning data a rendering matrix $A_{l,m}$;
- determining from the received *OLD, IOC, DMG* and *DCLD* data an estimated covariance matrix $C_{l,m}$ and a reconstructed down-mixing matrix $D_{l,m}$;
- calculating an estimation matrix $T_{l,m} = A_{l,m}C_{l,m}D_{l,m}^H(D_{l,m}C_{l,m}D_{l,m}^H)^{-1}$;
- multiplying said time/frequency domain signals (Y) with said estimation matrix $T_{l,m}$ so as to get desired-remix signals, followed by processing said desired-remix signals in a synthesis filter bank that has an inverse operation of said analysis filter bank.

[0017] In principle, the inventive encoding apparatus is suited for downmixing spatial audio signals that can be downmixed at receiver side in a manner different from the manner of downmixing at encoder side, wherein said encoding is based on MPEG SAOC and said downmixing at receiver side can be controlled by desired playback configuration data and/or desired object positioning data, said apparatus including:

- an analysis filter bank for processing *M* correlated sound signals, *M* being greater than '2', and *L* independent sound signals, *L* being '1' or greater, providing corresponding time/frequency domain signals;
- means being adapted for multiplying said time/frequency domain signals with a downmix matrix $D_{l,m}$ *;*
- a synthesis filter bank for said multiplied time/frequency domain signals that has an inverse operation of said analysis filter bank and that provides *M time* domain output signals;
- means being adapted for determining from said time/frequency domain signals MPEG SAOC side information data including Object Level Differences data *OLD* and Inter-Object Coherence data *IOC,* as well as enhanced Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD,* wherein said *DMG* and *DCLD* data are related to *M* channels.

[0018] In principle, the inventive decoding apparatus is suited for downmixing spatial audio signals processed according to the encoding method in a manner different from the manner of downmixing at encoder side, wherein said downmixing at receiver side can be controlled by desired playback configuration data and/or desired object positioning data, said apparatus including:

- means being adapted for receiving said processed spatial audio signals and for processing them in an analysis filter bank, providing corresponding time/frequency domain signals;
- means being adapted for determining from said desired playback configuration data and/or said desired object positioning data a rendering matrix $A_{l,m}$ *;*
- means being adapted for determining from the received *OLD, IOC, DMG* and *DCLD* data an estimated covariance matrix $C_{l,m}$ and a reconstructed down-mixing matrix $D_{l,m}$*;*
- means (36) being adapted for calculating an estimation matrix $T_{l,m} = A_{l,m} C_{l,m} D^H_{l,m} (D_{l,m} C_{l,m} D^H_{l,m})^{-1}$ **;**
- means being adapted for multiplying said time/frequency domain signals with said estimation matrix $T_{l,m}$ so as to get desired-remix signals, followed by processing said desired-remix signals in a synthesis filter bank that has an inverse operation of said analysis filter bank.

[0019] Advantageous additional embodiments of the invention are disclosed in the respective dependent claims.

Drawings

[0020] Exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in:

Fig. 1	standard MPEG SAOC system;
Fig. 2	enhanced MPEG SAOC encoder;
Fig. 3	enhanced MPEG SAOC decoder.

Exemplary embodiments

[0021] The inventive spatial audio object coding system with five down-mix channels facilitates a backward compatible transmission at bit rates only slighter higher (due to the extended content of the side information: *OLD, IOC, DMG* and *DCLD*) than the bit rates for known 5.1 channel transmission. By using the side information, a guided spatial remix is achieved, as well as the adaption of the audio mix to different listening situations, including guided downmix (or even upmix) of multi-channel audio.

[0022] In the following embodiment, a number of *M*=5 channels containing the ambience signals and a number of *L* audio objects mixed over the ambiance are considered. An example is the stadium ambiance of a soccer match plus specific sound effects (ball kicks, whistle) and one or more commentators. Thus, the encoder input gets *N* = *M*+*L* audio channels.

[0023] In other embodiments, *M is* at least '2' and *L* is '1' or greater.

[0024] At decoder side it is not intended to reconstruct the audio objects, but to offer the possibility of re-mixing, attenuating, totally suppressing, and changing the position of the audio objects in the rendered audio scene.

[0025] For the processing part of the system any time/frequency transform can be used. In this embodiment, hybrid Quadrature Mirror Filter (hQMF) banks are used for better selectivity in the frequency domain. The spatial audio input signals are processed in non-overlapping multiple-sample temporal slots, in particular 64-samples temporal slots. These temporal slots are used for computing the perceptual cues or characteristics for every successive frame, which has a length of a fixed number of temporal slots, in particular 16 temporal slots.

[0026] In the frequency domain, 71 frequency bands are used according to the sensitivity of the human auditory system, and are grouped into *K* processing bands, *K* having a value of '2', '3' or '4', thereby obtaining different levels of accuracy. The hQMF filter bank transforms in each case 64 time samples into 71 frequency samples. The processing band borders are represented in the following table:

Table 2.1: Processing Bands according to different quality levels

	K=4	K=3	K=2
Processing Band 1	0-7	0-7	0-20
Processing Band 2	8-20	8-55	21-70
Processing Band 3	21-29	56-70	-
Processing Band 4	30-70	-	-

[0027] A basic block diagram of the inventive encoder and decoder is shown in Fig.2 and Fig.3, respectively.

[0028] In the encoder in Fig.2, a hybrid Quadrature Mirror Filter analysis filter bank step or stage 21 is applied to all audio input signals, e.g. ambience channels Ch.#1 to Ch.#5 and sound objects Obj.#1 to Obj.#*L* (at least one sound object). The number of ambience channels is not limited to five. In this invention, contrary to MPEG SAOC, the ambience channels are not independent sound objects but are usually correlated sound signals. The corresponding filter bank outputs time/frequency domain signals X, which are fed on one hand to a down-mixer step or stage 22 that multiplies them with a downmix matrix *D* and provides, via a hybrid Quadrature Mirror Filter synthesis filter bank step or stage 24 that has an inverse operation of the analysis filter bank, the audio channels in time domain for transmission, and on the other hand are fed to an enhanced MPEG SAOC parameter calculator step or stage 23. The synthesis filter bank inverses the function of the analysis filter bank. The enhanced SAOC parameters determine the rendering flexibility in a decoder and include, as mentioned above, Object Level Differences data *OLD,* Inter-Object Coherence data *IOC,* Downmix Gains data *DMG,* Downmix Channel Level Differences data *DCLD,* and can comprise Object Energy parameter data *NRG.* Except the *DMG* and *DCLD* data, the other parameters correspond to original MPEG SAOC parameters. From these side information data items a rank can be determined as described below in a rank calculator step or stage 25 (i.e. step/stage 25 is optional), and the side information data items and data items regarding any re-mix constraints (described below) are transmitted.

[0029] The output signals of step/stage 24 together with the output signals of step/stage 25 are used to form an enhanced MPEG SAOC bitstream.

[0030] As mentioned in section 5.1 SAOC overview/Introduction of the MPEG SAOC standard, the object parameters are quantised and coded efficiently, and are correspondingly decoded and inversely quantised at receiver side. Before transmission, the downmix signal can be compressed, and is correspondingly decompressed at receiver side. The SAOC side information is (or can be) encoded according to the MPEG SAOC standard and is transmitted together with *DMG* and *DCLG* data e.g. as ancillary data portion of the downmix bitstream.

[0031] In the receiver-side decoder in Fig.3, a hybrid Quadrature Mirror Filter analysis filter bank step or stage 31 corresponding to filter bank 21 receives and processes the transmitted hQMF synthesised data from the enhanced MPEG SAOC bitstream, and feds them as time/frequency domain data to an enhanced MPEG SAOC decoder or transcoder step or stage 32 that is controlled by an estimation matrix *T.* A rendering matrix step or stage 35 receives user data regarding perceptual cues, e.g. a desired playback configuration and desired object positions, as well as the transmitted re-mix constraint data items, and there from a corresponding rendering matrix *A* is determined. Matrix A has the size of down-mixing matrix *D* and its coefficients are based on the coefficients of matrix *D.* In case e.g. only the 'last' sound object Obj.#L shall not be present in the decoder output signals, the last row of rendering matrix A contains zero values only and all other matrix coefficients are identical to the coefficients in matrix *D.* I.e., each sound object is represented by a different row in matrix *A.*

[0032] Matrix *A,* together with an estimated covariance matrix *C* and a reconstructed down-mixing matrix *D,* are used for determining (as explained below) in an estimation matrix generator step or stage 36 the estimation matrix *T.*

[0033] The estimated covariance matrix *C* and the reconstructed down-mixing matrix *D* are determined from the received side information data in a covariance and down-mixing matrix calculation step or stage 34. The estimation matrix T is used for decoding or transcoding the audio signals of the new audio scene. The downmixed signals of step/stage 32 are output as channel signals (e.g. Ch.#1 to Ch.#5) via a hybrid Quadrature Mirror Filter synthesis filter bank step or stage 33 (corresponding to synthesis filter bank 24).

[0034] In order to obtain a minimum bit rate for the side information encoding, the perceptual cues are used as ancillary data in the main bitstream. "Minimum bit rate" means: such that the resulting audio quality is not affected, i.e. the distortions caused due to slightly less available bit rate for the audio signals are not audible, or at least not annoying. For characterising the input audio objects the Object Level Differences data *OLD* and Inter Object Coherence data *IOC* are used, the values of which are computed in step/stage 23 e.g. according to section Annex D.2 "Calculation of SAOC parameters" of the MPEG SAOC standard, for every frame/frequency processing band tile (*l,m*), i.e. for every non-overlapping 16 temporal slots and every *K* processing bands.

[0035] First the auto-correlation and the cross-correlation of any two objects are computed and saved in a matrix format:

$$nrg_{l,m}(i,j)=\frac{\sum_{t\in l}\sum_{k\in m}X_{t,k}(i)*X^{*}_{t,k}(j)}{\sum_{t\in l}\sum_{k\in m}1}+\varepsilon \quad , \qquad (1)$$

where the indices *i* and *j* stand for the ambience channel number and the audio object number, respectively, *m* is a current frequency processing band, *k* is a running frequency sample index within frequency processing band *m*, *l* is a current frame, *t* is a running temporal slot index within frame *l*, and ε has a small value (e.g. 10^{-9}) and avoids a division by zero in the following computations.

[0036] The desired perceptual cues *OLD* and *IOC* are computed as:

$$\begin{aligned} OLD_{l,m}(i)&=\frac{nrg_{l,m}(i,i)}{NRG_{l,m}} \\ IOC_{l,m}(i,j)&=\frac{nrg_{l,m}(i,j)}{\sqrt{nrg_{l,m}(i,i)nrg_{l,m}(j,j)}} \end{aligned} \quad , \qquad (2)$$

where $NRG_{l,m} = \max_i(nrg_{l,m}(i,i))$ represents the absolute object energy of the object with the highest energy.

[0037] In order to characterise at decoder side the down-mixing matrix coefficients, in step/stage 23 the values for Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD* are calculated, again for every frame *l* with 16 temporal slots and frequency processing band *m*:

$$\begin{aligned} DMG_{l,m}(r)&=10\log_{10}\Big(\sum_{i=1}^{5}D^{2}_{l,m}(i,r)+\varepsilon\Big) \\ DCLD_{l,m}(r,j)&=10\log_{10}\Big(\frac{D^{2}_{l,m}(1,r)+\varepsilon}{D^{2}_{l,m}(j,r))+\varepsilon}\Big) \end{aligned} \quad , \qquad (3)$$

where *r*=1:*N* represents the input signal index and *j*=1:5 (in other embodiments *j*=1:*M*) represents the down-mix channel index. The *DMG* and *DCLD* data or parameters are extended versions of the corresponding MPEG SAOC parameters because they are not limited to two channels like in MPEG SAOC. Depending on the mixing procedure, the time/frequency resolution of *DMG* and *DCLD* can be modified with the moving speed of the audio objects in the audio scene. This time/frequency resolution change will not affect the performance of the inventive processing, and therefore it is assumed for simplicity that the time/frequency resolution at which these parameters are computed is equal to the time/frequency resolution at which the processing is done.

[0038] At decoder side the perceptual cues are used in step/stage 34 for approximating the covariance matrix C of the original input channels. When considering the definition of the Object Level Differences *OLD* and Inter Object Coherence *IOC,* the following form for the estimated covariance matrix C is obtained, computed for every time/frequency (*l,m*) tile:

$$C_{l,m}(i,j)=\sqrt{OLD_{l,m}(i)OLD_{l,m}(j)}IOC_{l,m}(i,j)\quad . \tag{4}$$

[0039] In order to remix the audio objects, besides the new rendering matrix also the original rendering or mixing matrix is required at decoder side. For re-constructing the downmixing matrix *D* in step/stage 34, the Downmix Gains *DMG* values and the Downmix Channel Level Differences *DCLD* values from the additional side information are used. The content of the original down-mix matrix *D* is computed in step/stage 34 as:

$$\begin{aligned} D_{l,m}(1,r) &= 10^{\frac{DMG_{l,m}(r)}{2}}\sqrt{\frac{1}{1+\sum_{j=2}^{5}10^{-0.1*DCLD_{l,m}(r,j)}}} \\ D_{l,m}(i,k) &= 10^{\frac{DMG_{l,m}(k)}{2}}\sqrt{\frac{10^{-0.1*DCLD_{l,m}(r,i)}}{1+\sum_{j=2}^{5}DCLD_{l,m}(r,j)}} \end{aligned}\quad , \tag{5}$$

where *r*=1:*N* represents the input signal index and *i*=2:5 (in other embodiments *i*=2:*M*) represents the down-mix channel index, j represents here a running input signal index within the sum, and *k* is a running frequency sample index within frequency processing band *m*. Matrix *D* is computed differently than according to MPEG SAOC, but its resulting content can assumed to be identical.

[0040] In the transcoding at decoder side, every frame *l* and processing band *m* (cf. the above table), mixing matrix $D_{l,m}(i,j)$ and the rendering matrix $A_{l,m}(i,j)$ used for remixing the *L* objects have a size 5 by *N* in this embodiment. For every time/frequency tile the original down-mixed signals Y at encoder side are, and the desired remixed signals *Z* at decoder side would be:

$$\begin{aligned} Y_t(i,k) &= D_{l,m}(i,j)*X_t(j,k) \\ Z_t(i,k) &= A_{l,m}(i,j)*X_t(j,k) \end{aligned}\quad , \tag{6}$$

where $t \in l, k \in m$ and {i = 1:5, j = 1:N}.

[0041] For simplicity of notation, the *i-j-k* indices are dropped, and $Y_{t,m}$, $Z_{t,m}$ and $X_{t,m}$ are considered to be matrices having a number of columns equal to the number of frequency subbands in each processing band m:

$$\begin{aligned} Y_{t,m} &= D_{l,m} * X_{t,m} \\ Z_{t,m} &= A_{l,m} * X_{t,m} \end{aligned}\quad . \tag{7}$$

[0042] I.e., the calculation of Y is carried out for each temporal slot *t* in a current frame *l* but for all these temporal slots the same downmix matrix *D* is used.

[0043] Using this notation, at encoder side the covariance matrix *C* of the input signals is defined for every time/frequency tile (*l,m*) by:

$$C_{l,m} = E_l[X_{t,m} * X_{t,m}^H]\quad , \tag{8}$$

where $E_l[X_{t,m}]$ represents the expectation of frame *l* of the input signals $X_{t,m}$ and $X_{t,m}^H$ represents the Hermitian notation of matrix $X_{t,m}$, i.e. its conjugate transpose.

[0044] Using equations (7) and (8), the correlation of the down-mixed signals at encoder side can be expressed as:

$$\begin{aligned} G_{l,m} &= E_l[Y_{t,m}Y_{t,m}^H] \\ &= E_l[D_{l,m}X_{t,m}X_{t,m}^H D_{l,m}^H] \\ &= D_{l,m}E_l[X_{t,m}X_{t,m}^H]D_{l,m}^H \\ &= D_{l,m}C_{l,m}D_{l,m}^H \end{aligned} \quad , \qquad (9)$$

and the correlation of the desired remixed signals at decoder side as:

$$\begin{aligned} F_{l,m} &= E_l[Z_{t,m}Z_{t,m}^H] \\ &= E_l[A_{l,m}X_{t,m}X_{t,m}^H A_{l,m}^H] \\ &= A_{l,m}E_l[X_{t,m}X_{t,m}^H]A_{l,m}^H \\ &= A_{l,m}C_{l,m}A_{l,m}^H \end{aligned} \quad . \qquad (10)$$

[0045] However, because there is no access at decoder side to the original input signals *X* in order to remix them using the rendering matrix *A*, the desired remix signals $Z_{t,m}$ are approximated in step/stage 32 as remix signals $\hat{Z}_{t,m}$ from the down-mixed signals *Y*:

$$\hat{Z}_{t,m} = T_{l,m} * Y_{t,m} \quad , \qquad (11)$$

with $t \in l$.

[0046] The estimation matrix $T_{l,m}$ should be chosen such that the squared error is minimised:

$$T_{l,m} = \operatorname{argmin} E_l[(Z_{t,m} - \hat{Z}_{t,m})(Z_{t,m} - \hat{Z}_{t,m})^H] \quad . \qquad (12)$$

Remark: $Z_{t,m}$ is not available at decoder side but is used for the derivation of the following equations, and it finally turns out that knowledge of $Z_{t,m}$ at decoder side is not required.

[0047] Using the 'Orthogonality Principle', it is known that the squared error is minimised when the error is orthogonal on the space spanned by the original down-mix signals. This means that:

$$\begin{aligned} & E_l[(Z_{t,m} - \hat{Z}_{t,m})Y_{t,m}^H] = 0 \\ \Leftrightarrow\ & E_l[Z_{t,m}Y_{t,m}^H] = E_l[\hat{Z}_{t,m}Y_{t,m}^H] \\ \Leftrightarrow\ & E_l[A_{l,m}X_{t,m}X_{t,m}^H D_{l,m}^H] = E_l[T_{l,m}Y_{t,m}Y_{t,m}^H] \\ \Leftrightarrow\ & A_{l,m}E_l[X_{t,m}X_{t,m}^H]D_{l,m}^H = T_{l,m}D_{l,m}E_l[X_{t,m}X_{t,m}^H]D_{l,m}^H \\ \Leftrightarrow\ & A_{l,m}C_{l,m}D_{l,m}^H = T_{l,m}D_{l,m}C_{l,m}D_{l,m}^H \end{aligned} \quad . \qquad (13)$$

[0048] If the matrix $D_{l,m}C_{l,m}D_{l,m}^H$ is not singular, the estimation matrix $T_{l,m}$ can be computed at decoder side by inverting matrix $G_{l,m}$:

$$T_{l,m} = A_{l,m}C_{l,m}D_{l,m}^H(D_{l,m}C_{l,m}D_{l,m}^H)^{-1} \quad , \qquad (14)$$

wherein covariance matrix C is estimated at decoder side according to equation (4).

[0049] Because the expression of $G_{l,m}$ is depending only on parameters known at encoder side and does not depend on the rendering matrix, it can be decided before encoding whether or not remixing is feasible at decoder side. Practice, and the special form of the down-mixing and correlation matrices of the ambiance channels, show that in most cases matrix $G_{l,m}$ will be invertible.

[0050] In order to ensure the functionality of the decoding processing, the rank of this matrix $G_{l,m}$ is computed in step/stage 25 before final encoding, and one can proceed with the encoding of the side information if that matrix has a full rank. The rank of a matrix is the number of independent columns or rows, and this can be hard to determine. Thus, instead of using the rank, the effective rank is used, which is a more stable measure for the rank and is described in section 6.3 "Singular Value Decomposition" of the textbook of Gilbert Strang, "Linear Algebra and its applications", 4th edition, published 19 July 2005. First, the eigenvalues of matrix $G_{l,m}$ are computed. Next, the number of eigenvalues greater than a tolerance value τ is counted and this number is taken as the effective rank. Practice shows that $\tau = 10^{-5}$ (a preferential range for τ is $10^{-4}...10^{-8}$) can be chosen, and for matrices with full effective rank it can be assumed that it is accurate enough for the inverse computation.

[0051] The rank value can be used for controlling the number K of frequency bands applied in the inventive processing, and thereby the accuracy of the side information parameters. The rank value can also be used for switching on or off a residual coding like in the MPEG SAOC standard.

[0052] Measuring the sensitivity entirely by the smallest singular value may not be sufficient, because a simple multiplication of the matrix with factor 10^5 will indicate a much less singular matrix, and an ill-conditioning problem is not solved by a simple re-scaling. Thus, before computing the effective rank as described above, a normalisation of matrix $G_{l,m}$ is carried out.

[0053] Because $C_{l,m}$ represents a correlation matrix of the input audio signals, it will be a symmetric matrix and obviously $G_{l,m} = D_{l,m}C_{l,m}D_{l,m}^H$ will form a symmetric matrix. Therefore, in order to compute the eigenvalues, a Schur decomposition is used for a symmetric matrix:

$$G_{l,m} = Q_{l,m}U_{l,m}Q_{l,m}^{-1} \quad , \qquad (15)$$

where $Q_{l,m}$ is a unitary matrix $(Q_{l,m}^{-1} = Q_{l,m}^H)$ and $U_{l,m}$ is a diagonal matrix having on the main diagonal the eigenvalues of

$$G_{l,m}: diag(U_{l,m}) = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}.$$

[0054] An interesting property of the Schur decomposition is that the inverse of $G_{l,m}$ can be easily computed as:

$$G_{l,m}^{-1} = Q_{l,m}^{-1}\overline{U}_{l,m}Q_{l,m} \quad , \qquad (16)$$

where $\overline{U}_{l,m}$ is a diagonal matrix having on the main diagonal the inverse of the eigenvalues of matrix $G_{l,m}$ $diag(\overline{U}_{l,m})$= $\{1/\lambda_1, 1/\lambda_2, 1/\lambda_3, 1/\lambda_4, 1/\lambda_5\}$. Proof: matrix $U_{l,m}$ is a diagonal matrix and, if the values from the main diagonal are different from zero, the matrix is invertible, whereby the inverse is equal to $\overline{U}_{l,m}$. Thus the processing starts with the following computation:

$$\begin{aligned} I_5 &= Q_{l,m}Q_{l,m}^{-1} \\ &= Q_{l,m}I_5Q_{l,m}^{-1} \\ &= Q_{l,m}U_{l,m}\overline{U}_{l,m}Q_{l,m}^{-1} \\ &= Q_{l,m}U_{l,m}Q_{l,m}^{-1}Q_{l,m}\overline{U}_{l,m}Q_{l,m}^{-1} \\ &= G_{l,m}(Q_{l,m}\overline{U}_{l,m}Q_{l,m}^{-1}) \end{aligned} \quad . \qquad (17)$$

[0055] Thus $G_{l,m}$ is invertible with the inverse being equal to

$$G_{l,m}^{-1} = Q_{l,m}\bar{U}_{l,m}Q_{l,m}^{-1} \quad .$$

[0056] A singular matrix is not very common and it can become non-singular with a small weighting of one coefficient of the singular matrix. Thus, in order to ensure at decoder side that matrix $G_{l,m}$ is invertible, after computing the Schur decomposition and finding eigenvalues smaller than the defined tolerance value τ, these eigenvalues are modified by adding a weight of τ to each one of them. In this way, when computing the inverse of $G_{l,m}$, using the described property of the Schur decomposition, it is sure that this matrix is well-conditioned. The error introduced by this procedure is of order τ and will not affect the remixing processing in step/stage 32.

[0057] The values used for $C_{l,m}$ and $D_{l,m}$ are estimated according to equations (4) and (5).

Claims

1. Method for encoding by downmixing (22) spatial audio signals (Ch.#1 - Ch.#5, Obj.#1 - Obj.#L) that can be downmixed at receiver side in a manner different from the manner of downmixing at encoder side, wherein said encoding is based on MPEG SAOC and said downmixing at receiver side can be controlled (35) by desired playback configuration data and/or desired object positioning data, said method including the steps:

   - processing M correlated sound signals (Ch.#1 - Ch.#5), M being greater than '2', and L independent sound signals (Obj.#1 - Obj.#L), *L* being '1' or greater, in an analysis filter bank (21) providing corresponding time/frequency domain signals (*X*);
   - multiplying (22) said time/frequency domain signals with a downmix matrix $D_{l,m}$, followed by processing the resulting signals (*Y*) in a synthesis filter bank (24) that has an inverse operation of said analysis filter bank (21) and that provides *M* time domain output signals;
   - determining (23) from said time/frequency domain signals (X) MPEG SAOC side information data including Object Level Differences data *OLD* and Inter-Object Coherence data *IOC,* as well as enhanced Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD,* wherein said *DMG* and *DCLD* data are related to *M* channels.

2. Method for downmixing (32) spatial audio signals processed according to claim 1 in a manner different from the manner of downmixing at encoder side, wherein said downmixing at receiver side can be controlled (35) by desired playback configuration data and/or desired object positioning data, said method including the steps:

   - receiving said processed spatial audio signals and processing them in an analysis filter bank (31), providing corresponding time/frequency domain signals (*Y*);
   - determining (35) from said desired playback configuration data and/or said desired object positioning data a rendering matrix $A_{l,m}$;
   - determining (34) from the received *OLD, IOC, DMG* and *DCLD* data an estimated covariance matrix $C_{l,m}$ and a reconstructed down-mixing matrix $D_{l,m}$ *;*
   - calculating (36) an estimation matrix

   $$T_{l,m} = A_{l,m}C_{l,m}D_{l,m}^{H}(D_{l,m}C_{l,m}D_{l,m}^{H})^{-1} \,;$$

   - multiplying (32) said time/frequency domain signals (*Y*) with said estimation matrix $T_{l,m}$ so as to get desired-remix signals ($\hat{Z}$), followed by processing said desired-remix signals in a synthesis filter bank (33) that has an inverse operation of said analysis filter bank (31).

3. Apparatus for encoding by downmixing (22) spatial audio signals (Ch.#1 - Ch.#5, Obj.#1 - Obj.#L) that can be downmixed at receiver side in a manner different from the manner of downmixing at encoder side, wherein said encoding is based on MPEG SAOC and said downmixing at receiver side can be controlled (35) by desired playback configuration data and/or desired object positioning data, said apparatus including:

   - an analysis filter bank (21) for processing *M* correlated sound signals (Ch.#1 - Ch.#5), *M* being greater than '2', and *L* independent sound signals (Obj.#1 - Obj.#L), *L* being '1' or greater, providing corresponding time/

frequency domain signals (*X*);
- means (22) being adapted for multiplying said time/frequency domain signals (*X*) with a downmix matrix $D_{l,m}$;
- a synthesis filter bank (24) for said multiplied time/frequency domain signals (*Y*) that has an inverse operation of said analysis filter bank (21) and that provides *M* time domain output signals;
- means (23) being adapted for determining from said time/frequency domain signals (*X*) MPEG SAOC side information data including Object Level Differences data *OLD* and Inter-Object Coherence data *IOC,* as well as enhanced Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD,* wherein said *DMG* and *DCLD* data are related to *M* channels.

4. Apparatus for downmixing (32) spatial audio signals processed according to claim 1 in a manner different from the manner of downmixing at encoder side, wherein said downmixing at receiver side can be controlled (35) by desired playback configuration data and/or desired object positioning data, said apparatus including:

- means (31) being adapted for receiving said processed spatial audio signals and for processing them in an analysis filter bank, providing corresponding time/frequency domain signals (*Y*);
- means (35) being adapted for determining from said desired playback configuration data and/or said desired object positioning data a rendering matrix $A_{l,m}$;
- means (34) being adapted for determining from the received *OLD, IOC, DMG* and *DCLD* data an estimated covariance matrix $C_{l,m}$ and a reconstructed down-mixing matrix $D_{l,m}$;

- means (36) being adapted for calculating an estimation matrix $T_{l,m} = A_{l,m}C_{l,m}D_{l,m}^{H}(D_{l,m}C_{l,m}D_{l,m}^{H})^{-1}$;
- means (32) being adapted for multiplying said time/frequency domain signals (*Y*) with said estimation matrix $T_{l,m}$ so as to get desired-remix signals ($\hat{Z}$), followed by processing said desired-remix signals in a synthesis filter bank (33) that has an inverse operation of said analysis filter bank (31).

5. Method according to the method of claim 1 or 2, or apparatus according to the apparatus of claim 3 or 4, wherein said spatial input signals are processed in non-overlapping multiple-sample temporal slots, a fixed number of such temporal slots representing a frame *l*, and in *K* frequency processing bands into which the total frequency range is divided, *K* having a value of '2', '3' or '4'.

6. Method according to the method of claim 5, or apparatus according to the apparatus of claim 5, wherein said Downmix Gains data *DMG* and Downmix Channel Level Differences data *DCLD* are calculated for every input signal frame *l* and processing band m according to:

$$DMG_{l,m}(r) = 10\log_{10}(\sum_{i=1}^{5} D_{l,m}^{2}(i,r) + \varepsilon)$$
$$DCLD_{l,m}(r,j) = 10\log_{10}(\frac{D_{l,m}^{2}(1,r) + \varepsilon}{D_{l,m}^{2}(j,r)) + \varepsilon}) \quad ,$$

where *r*=1:(*M*+*L*) represents an spatial audio input signal index, *j*=1:5 represents a down-mix channel index and value ε is used for avoiding a division by zero in other related computations.

7. Method according to the method of claim 5 or 6, or apparatus according to the apparatus of claim 5 or 6, wherein said rendering matrix $A_{l,m}$ has a size of said downmix matrix $D_{l,m}$ and its coefficients are based on the coefficients of matrix $D_{l,m}$, wherein each sound object is represented by a different row in matrix $A_{l,m}$.

8. Method according to the method of one of claims 5 to 7, or apparatus according to the apparatus of one of claims 5 to 7, wherein said estimated covariance matrix $C_{l,m}$ is calculated according to $C_{l,m}(i,j) = \sqrt{OLD_{l,m}(i)OLD_{l,m}(j)}IOC_{l,m}(i,j)$, and said reconstructed down-mixing matrix $D_{l,m}$ is calculated according to

$$D_{l,m}(1,r)=10^{\frac{DMG_{l,m}(r)}{2}}\sqrt{\frac{1}{1+\sum_{j=2}^{5}10^{-0.1*DCLD_{l,m}(r,j)}}}$$

$$D_{l,m}(i,k)=10^{\frac{DMG_{l,m}(k)}{2}}\sqrt{\frac{10^{-0.1*DCLD_{l,m}(r,i)}}{1+\sum_{j=2}^{5}DCLD_{l,m}(r,j)}},$$

where *r*=1:(*M*+*L*) represents a spatial audio input signal index, *i*=2:5 represents a down-mix channel index, and *k* is a running frequency sample index within a current frequency processing band.

9. Method according to the method of one of claims 5 to 8, or apparatus according to the apparatus of one of claims 5 to 8, wherein from said *OLD, IOC, DMG* and *DCLD* data a matrix $G_{l,m} = D_{l,m}C_{l,m}D_{l,m}^{H}$ *is* calculated (25), in which $C_{l,m}$ is a related covariance matrix and $D_{l,m}^{H}$ represents the Hermitian notation of downmix matrix $D_{l,m}$, and wherein there from a rank value is calculated and it is determined whether matrix $G_{l,m}$ has a full rank, and, if true, said side information data is encoded for transmission in said bitstream.

10. Method according to the method of claim 9, or apparatus according to the apparatus of claim 9, wherein said rank value is used to control the number *K* of frequency bands used in the processing.

11. Method according to the method of claim 9 or 10, or apparatus according to the apparatus of claim 9 or 10, wherein said rank value is used to switch on or off a residual coding.

12. Digital audio signal that is encoded according to the method of one of claims 1, 5 and 6.

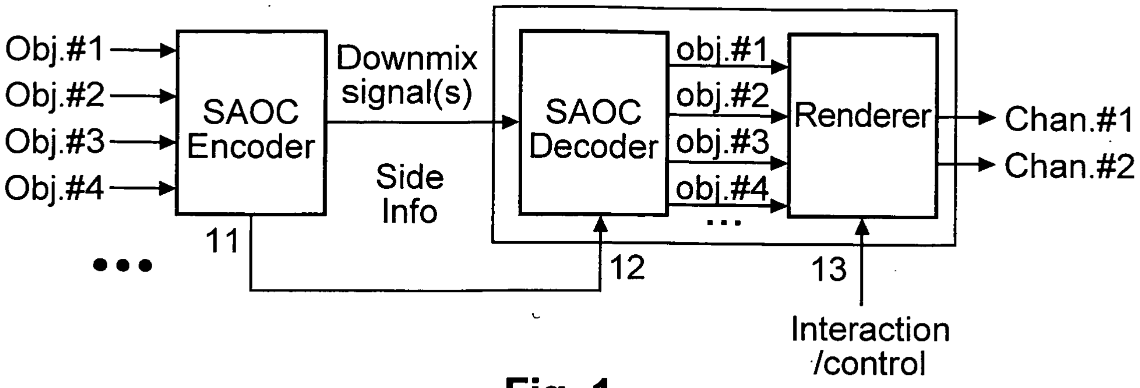
Obj.#1
Obj.#2
Obj.#3
Obj.#4
•••
SAOC Encoder
11
Downmix signal(s)
Side Info
SAOC Decoder
12
obj.#1
obj.#2
obj.#3
obj.#4
...
Renderer
13
Interaction /control
Chan.#1
Chan.#2

Fig. 1

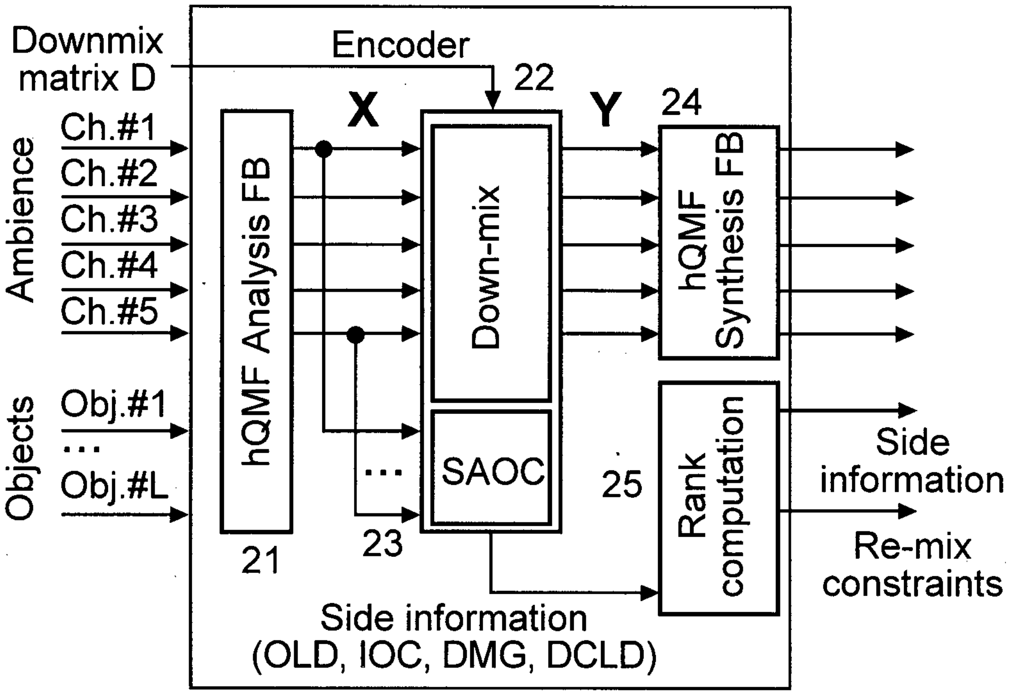
Downmix
matrix D
Encoder
22
X
Y
24
Ambience
Ch.#1
Ch.#2
Ch.#3
Ch.#4
Ch.#5
Objects
Obj.#1
...
Obj.#L
hQMF Analysis FB
Down-mix
SAOC
hQMF
Synthesis FB
Rank
computation
...
21
23
25
Side
information
Re-mix
constraints
Side information
(OLD, IOC, DMG, DCLD)

Fig. 2

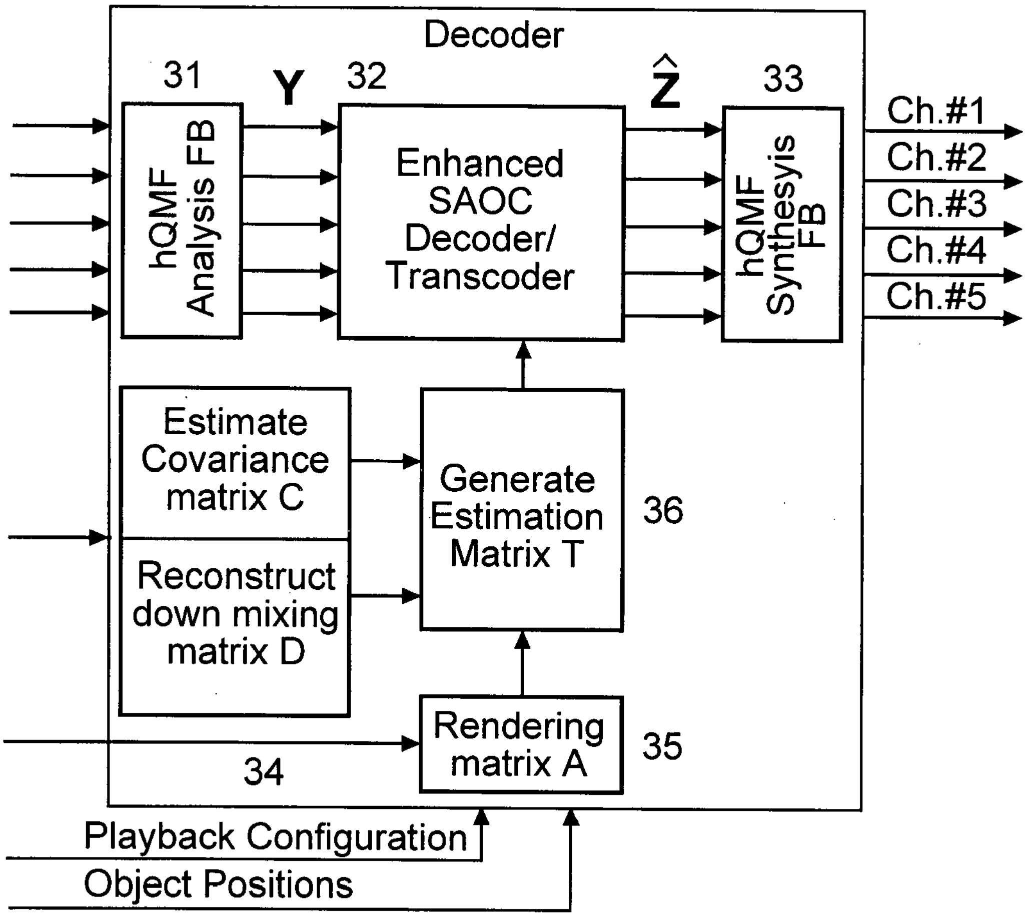
Decoder
31
Y
32
$\hat{Z}$
33
hQMF Analysis FB
Enhanced SAOC Decoder/ Transcoder
hQMF Synthesyis FB
Ch.#1
Ch.#2
Ch.#3
Ch.#4
Ch.#5
Estimate Covariance matrix C
Reconstruct down mixing matrix D
Generate Estimation Matrix T
36
Rendering matrix A
35
34
Playback Configuration
Object Positions

Fig. 3

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 12 30 5914

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X Y	US 2012/078642 A1 (SEO JEONG IL [KR] ET AL) 29 March 2012 (2012-03-29) * paragraphs [0039] - [0041], [0044], [0045], [0050] * * figures 1,3,5 * -----	1-6,12 7-11	INV. G10L19/008 ADD. G10L19/02
X	US 2011/166867 A1 (SEO JEONGIL [KR] ET AL) 7 July 2011 (2011-07-07) * paragraphs [0003], [0032] - [0037], [0058] - [0064], [0089] - [0094] * * figures 1,9 * -----	1-4,12	
X	US 2008/049943 A1 (FALLER CHRISTOF [CH] ET AL) 28 February 2008 (2008-02-28) * paragraphs [0048], [0049], [0052], [0089] - [0092], [0103] * * figure 1A * -----	1-4,12	
Y	US 2012/177204 A1 (HELLMUTH OLIVER [DE] ET AL) 12 July 2012 (2012-07-12) * paragraph [0217] * -----	7-11	TECHNICAL FIELDS SEARCHED (IPC) G10L

The present search report has been drawn up for all claims

1

Place of search	Date of completion of the search	Examiner
Munich	27 November 2012	Geißler, Christian

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
....
& : member of the same patent family, corresponding document

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 12 30 5914

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

27-11-2012

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012078642 A1	29-03-2012	CN 102460571 A	16-05-2012
		EP 2442303 A2	18-04-2012
		KR 20100132913 A	20-12-2010
		US 2012078642 A1	29-03-2012
US 2011166867 A1	07-07-2011	CN 102171751 A	31-08-2011
		EP 2320415 A1	11-05-2011
		KR 20100008755 A	26-01-2010
		US 2011166867 A1	07-07-2011
US 2008049943 A1	28-02-2008	AT 524939 T	15-09-2011
		AT 527833 T	15-10-2011
		AT 528932 T	15-10-2011
		AU 2007247423 A1	15-11-2007
		BR PI0711192 A2	23-08-2011
		CA 2649911 A1	15-11-2007
		CN 101690270 A	31-03-2010
		EP 1853092 A1	07-11-2007
		EP 1853093 A1	07-11-2007
		EP 2291007 A1	02-03-2011
		EP 2291008 A1	02-03-2011
		JP 4902734 B2	21-03-2012
		JP 2010507927 A	11-03-2010
		KR 20090018804 A	23-02-2009
		KR 20110002498 A	07-01-2011
		RU 2008147719 A	10-06-2010
		US 2008049943 A1	28-02-2008
		WO 2007128523 A1	15-11-2007
US 2012177204 A1	12-07-2012	AR 077226 A1	10-08-2011
		AU 2010264736 A1	16-02-2012
		CA 2766727 A1	29-12-2010
		CN 102460573 A	16-05-2012
		CO 6480949 A2	16-07-2012
		EP 2446435 A1	02-05-2012
		EP 2535892 A1	19-12-2012
		KR 20120023826 A	13-03-2012
		SG 177277 A1	28-02-2012
		TW 201108204 A	01-03-2011
		US 2012177204 A1	12-07-2012
		WO 2010149700 A1	29-12-2010

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- Singular Value Decomposition. **GILBERT STRANG.** Linear Algebra and its applications. 19 July 2005 **[0050]**