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

(11) **EP 2 688 065 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
22.01.2014 Bulletin 2014/04

(51) Int Cl.:
G10L 19/008 *(2013.01)* ***H04S 3/02*** *(2006.01)*

(21) Application number: **12305860.4**

(22) Date of filing: **16.07.2012**

(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:
- **Krüger, Alexander**
  30655 Hannover (DE)
- **Kordon, Sven**
  31515 Wunstorf (DE)
- **Boehm, Johannes**
  37081 Göttingen (DE)
- **Jax, Peter**
  30171 Hannover (DE)

(74) Representative: **König, Uwe**
Deutsche Thomson OHG
Karl-Wiechert-Allee 74
30625 Hannover (DE)

(54) **Method and apparatus for avoiding unmasking of coding noise when mixing perceptually coded multi-channel audio signals**

(57) Playback of certain multi-channel audio signal representations on a particular loudspeaker set-up requires a special rendering, which usually consists of a matrixing operation. If a particular individual loudspeaker set-up on which the matrix of the matrixing operation depends is not known at the perceptual coding stage, a noise unmasking problem occurs. According to the invention, this problem can be solved by decorrelating the individual audio channel signals before perceptual encoding and recorrelating them after the perceptual decoding.

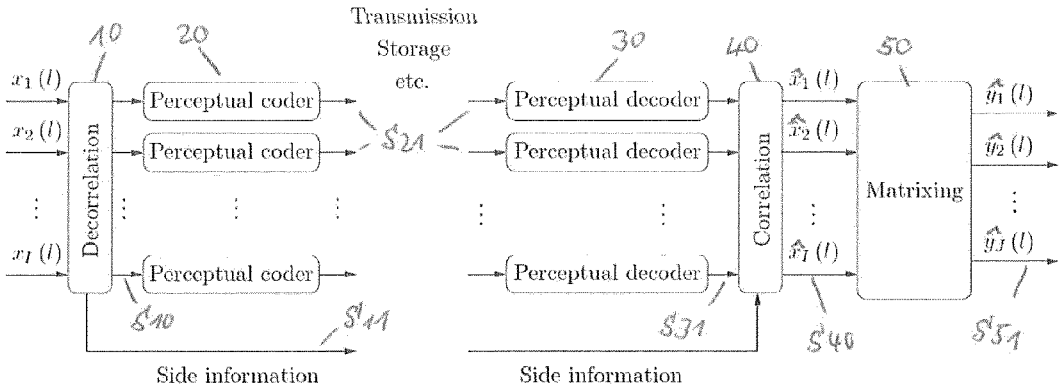

Fig.2

EP 2 688 065 A1

Description

Field of the invention

[0001] This invention relates to a method and an apparatus for avoiding an unmasking of coding noise when mixing perceptually coded and decoded multi-channel audio signals.

Background

[0002] The playback of certain multi-channel audio signal representations on a particular loudspeaker set-up requires a special rendering, which usually consists of a matrixing operation. A prominent example is Higher Order Ambisonics (HOA), which is a multi-channel sound field representation [4]. The transmission or storage of such multi-channel audio signal representations usually demands for appropriate multi-channel compression techniques. A known approach, which is e.g. pursued in [1] with respect to HOA representations, is illustrated in Fig. 1. In a first step the *I* individual channel audio signals $x_i(l)$, $i = 1, ..., I$, are independently perceptually encoded for transmission or storage using e.g. mp3 (MPEG-1 Audio Layer III) or MPEG AAC. In a second step, a channel independent perceptual decoding is performed before finally matrixing the *I* decoded signals $\hat{\hat{\mathrm{x}}}_i(l)$, $i = 1, ..., I$, into *J* new signals $\hat{\hat{\mathrm{y}}}_j(l)$, $j = 1,..., J$. The term matrixing means adding or mixing the decoded signals $\hat{\hat{\mathrm{x}}}_i(l)$ in a weighted manner. Arranging all signals $\hat{\hat{\mathrm{x}}}_i(l)$, $i = 1, ..., I$, as well as all new signals $\hat{\hat{\mathrm{y}}}_j(l)$, $j = 1, ..., J$ in vectors according to

$$\hat{\hat{\mathrm{x}}}(l) := \left[\, \hat{\hat{\mathrm{x}}}_1(l) \;\; ... \;\; \hat{\hat{\mathrm{x}}}_I(l) \right]^{\mathrm{T}}$$

$$\hat{\hat{\mathrm{y}}}(l) := \left[\, \hat{\hat{\mathrm{y}}}_1(l) \;\; ... \;\; \hat{\hat{\mathrm{y}}}_J(l) \right]^{\mathrm{T}}$$

the term "matrixing" origins from the fact that $\hat{\hat{\mathrm{y}}}(l)$ is, mathematically, obtained from $\hat{\hat{\mathrm{x}}}(l)$ through a matrix operation

$$\hat{\hat{\mathrm{y}}}(l) = \boldsymbol{A}\, \hat{\hat{\mathrm{x}}}(l)$$

where A denotes a mixing matrix composed of mixing weights. The terms "mixing" and "matrixing" are used synonymously herein. Mixing/matrixing is used for the purpose of rendering audio signals for any particular loudspeaker setups.

[0003] It is important to note that the particular individual loudspeaker set-up on which the matrix depends, and thus the maxtrix that is used for matrixing during the rendering, is usually not known at the perceptual coding stage.

[0004] The problem of noise unmasking when matrixing perceptually decoded multi-channel signals was addressed in [7] and [8]. In principle, the scenario considered in those publications is similar to that displayed in Fig.1. The input signals $\hat{x}_i(l)$ for perceptual coding are assumed to be a downmix of multi-channel audio signals, e.g. stereo or surround signals. The goal for matrixing of the perceptually decoded signals $\hat{\hat{\mathrm{x}}}_i(l)$ is the reconstruction of the original multi-channel audio signal from the downmix signals. The proposed two step technique should solve the noise unmasking problem, which consisted of an appropriate quantization and choice of the masking thresholds. However, an elementary assumption for the technique to work is that the mixing matrix is known in advance, i.e., already at the perceptual encoding stage. This, however, might be a significant restriction, e.g. when rendering multi-channel representations to a particular loudspeaker set-up, and at least in cases like rendering of HOA (Higher Order Ambisonics) representations to a particular loudspeaker set-up, where the loudspeaker set-up is not known in advance at the perceptual coding stage.

Summary of the Invention

[0005] There is a fundamental problem with the prior art compression approach illustrated in Fig. 1: channel independent perceptual encoding 20 and successive decoding 30 introduces quantization noise into the individual channel audio signals. Considering each decoded audio signal $\hat{\hat{\mathrm{x}}}_i(l)$ independently, the introduced quantization noise remains unnoticed by the human ear because of a masking effect that is caused by the audio signal $x_i(l)$. However, when matrixing is applied to such perceptually coded audio signals $\hat{\hat{\mathrm{x}}}_i(l)$, the audio signal part and the quantization noise can get separated and partially routed to different new signals $\hat{\hat{\mathrm{y}}}_j(l)$. This may result in the appearance of audible artifacts. This phenomenon is called "unmasking of coding noise" herein or, short, "noise unmasking". It has been found that the elementary reason for the appearance of noise unmasking is that for multi-channel representations like those given by $\hat{\hat{\mathrm{x}}}_i(l), i = 1, \ldots, I,$ there is a high cross correlation between the individual channels. During matrixing, these cross correlations between the channel signals may cause the desired audio signal to be cancelled while the coding noise may be constructively amplified.

[0006] The present invention is based on the recognition of the fact that at least the above-mentioned problems can be solved by decorrelating the individual audio channel signals before perceptual encoding and recorrelating them after the perceptual decoding, as shown in Fig.2. Such decorrelation and recorrelation has the advantage that noise unmasking at matrixing is avoided, even for arbitrary mixing matrices which are not yet known at the perceptual encoding stage. That is, the disclosed decorrelating of individual audio channel signals before perceptual encoding and recorrelating them after the perceptual decoding is suitable for achieving any arbitrary matrixing/mixing without unmasking of noise that is masked during the perceptual encoding and perceprual decoding.

[0007] An apparatus that utilizes the method is disclosed in claim 15.

[0008] Different from the systems described in [7] and [8], the method and apparatus proposed herein does not suffer from the restriction that the loudspeaker set-up must be known in advance at the perceptual coding stage.

[0009] According to the invention, a method for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals comprises steps of de-correlating the multi-channel audio signals, wherein de-correlated multi-channel audio signals and correlation information are obtained, perceptually encoding the de-correlated multi-channel audio signals, perceptually decoding the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals are obtained, re-correlating the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information, wherein re-correlated, perceptually decoded multi-channel audio signals are obtained, and mixing the re-correlated, perceptually decoded multi-channel audio signals, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal, and wherein a mixing scheme is used that is independent from (e.g. unknown in) said step of encoding.

[0010] In one embodiment, a corresponding decoding method comprises a method for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals within a decoder, comprising steps of receiving through a connection, or retrieving from a storage, perceptually encoded, de-correlated multi-channel audio signals and correlation information, perceptually decoding the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals are obtained, re-correlating the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information, wherein re-correlated, perceptually decoded multi-channel audio signals are obtained, and mixing the re-correlated, perceptually decoded multi-channel audio signals, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal.

[0011] In one embodiment, the method relates to perceptually coded time-frequency transformed multi-channel audio signals. A method for avoiding any unmasking of coding noise when mixing multi-channel audio signals that have been perceptually coded by using a time-frequency transform comprises perceptual encoding, perceptual decoding and mixing, and the perceptual encoding comprising steps of time-frequency transform framing and time-frequency transforming the multi-channel audio signals, wherein time-frequency transformed multi-channel audio signals are obtained, segmenting (i.e. performing a transform segmentation) on the time-frequency transformed multi-channel audio signals, and de-correlating with a Karhunen-Loève Transform (KLT) the segments obtained in the transform segmentation, and the perceptual decoding comprises steps of

channel independent decoding without inverse time-frequency transform, wherein a channel independent decoded signal is obtained,

re-correlating the channel independent decoded signal with a KLT, wherein re-correlated signals are obtained,

performing a transform de-segmentation on the re-correlated signals, wherein time-frequency transformed multi-channel audio signals are obtained,
performing inverse time-frequency transform with overlap add on the time-frequency transformed multi-channel audio signals, wherein framed time-domain multi-channel audio signals are obtained, and
de-framing the framed time-domain multi-channel audio signals, wherein the re-correlated, perceptually decoded multi-channel audio signals are obtained;
and the mixing comprises
mixing the re-correlated, perceptually decoded multi-channel audio signals, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal, and wherein a mixing scheme is used,
wherein said step of de-correlating with a KLT the segments obtained in the transform segmentation during encoding is independent from the mixing scheme that is used in the step of mixing.

[0012] In one embodiment, a corresponding decoding method comprises a method for avoiding the unmasking of coding noise when mixing, in a perceptual decoder, multi-channel audio signals that have been perceptually coded by using a time-frequency transform, comprising steps of channel independent decoding without inverse time-frequency transform, wherein channel independent decoded signals are obtained, re-correlating the channel independent decoded signals with a KLT, wherein re-correlated signals are obtained, performing a transform de-segmentation of the re-correlated signals, wherein time-frequency transformed multi-channel audio signals are obtained, performing inverse time-frequency transform with overlap add on the time-frequency transformed multi-channel audio signals, wherein framed time-domain multi-channel audio signals are obtained, de-framing the framed time-domain multi-channel audio signals, wherein the re-correlated, perceptually decoded multi-channel audio signals are obtained, and mixing the re-correlated, perceptually decoded multi-channel audio signals, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal, and wherein a mixing scheme is used.

[0013] In one aspect of the invention, a computer readable storage medium has executable instructions to cause a computer to perform a method comprising steps as disclosed in one of the claims 1-14.

[0014] Advantageous embodiments of the invention are disclosed in the dependent claims, the following description and the figures.

Brief description of the drawings

[0015] Exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in

Fig.1 a conventional process of perceptual encoding and perceptual decoding of multi-channel audio signal representations that are mixed during rendering;
Fig.2 an improved process of perceptual encoding and perceptual decoding of multi-channel audio signal representations that are mixed during rendering;
Fig.3 a KLT based perceptual encoder for multi-channel audio signals;
Fig.4 a KLT based perceptual decoder for multi-channel audio signals;
Fig.5 an exemplary framing scheme used for frequency domain transforms; and
Fig.6 the structure of an improved KLT based encoding and decoding scheme for TFT based perceptual coding.

Detailed description of the invention

[0016] In order to avoid noise unmasking at matrixing, even for arbitrary mixing matrices that are unknown at the perceptual encoding stage, the present invention decorrelates the individual audio channel signals before perceptual encoding, as shown in Fig.2. In Fig.2, an improved process of perceptual encoding and perceptual decoding of multi-channel audio signal representations that are mixed during rendering comprises a de-correlation 10 of audio signals $x_i(l)$, $i = 1,..., I$, perceptual encoding 20 of the de-correlated multi-channel audio signals, transmission and/or storage (not shown) of the perceptually encoded de-correlated multi-channel audio signals S21, perceptually decoding 30 the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals S31 are obtained, re-correlating 40 the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information S11, wherein re-correlated, perceptually decoded multi-channel audio signals S40 are obtained, and mixing 50 the re-correlated, perceptually decoded multi-channel audio signals S40, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal $\hat{y}_1(1)$. Advantageously, a mixing scheme that is used in the step of mixing needs not be known during the encoding, since the de-correlation is independent from the mixing scheme.

[0017] The decorrelation is accomplished with the Karhunen-Loève Transform (KLT), which is also known as Principle Component Analysis (PCA) [3]. After perceptual decoding, the decorrelated audio channel signals are correlated (to revert the decorrelation in the encoder) before any arbitrary matrixing is performed. To be able to revert to decorrelation in the decoder, some extra side information S11 is transmitted alongside with the perceptually coded audio channels S21. An analytical justification for this approach is provided further below.

[0018] It should be noted that applying a KLT for the decorrelation of multi-channel signals *before* the independent perceptual coding of individual channel signals is generally known, e.g. in [6] and [5], only for the purpose of reducing the data rate and, hence, for achieving an improved coding efficiency.

[0019] Fig.3 shows, in an embodiment of the present invention, a conceptual scheme for a KLT based perceptual encoder for multi-channel audio signals. It comprises steps of performing a segmentation 110 of the input signals, decorrelation 120 with a KLT, and channel independent perceptual coding 130. The decorrelation comprises following steps: A first step is computing 1210 the inter-channel correlation matrix for each segment. Subsequently, the inter-channel correlation matrix is quantized 1220, rescaled (e.g. inverse quantized) 1230, and the rescaled version of the correlation matrix is employed for the computation 1240 of the KLT transform matrix $B_{KLT}(m)$. Finally, the KLT is applied 1250 to the signals within each respective segment, before the individual transformed signals Z(m) are independently perceptually coded 130.

[0020] In order to be able to invert the KLT in the decoder, the coded quantized correlation matrix is transmitted as side information. For this purpose, it is losslessly encoded 1260. The losslessly encoded quantized correlation matrix is transmitted instead of the KLT transform matrix, since it is always symmetric, and therefore only a portion thereof (the upper or lower half, including the diagonal) needs to be transmitted. The quantized and encoded inter-channel correlation matrix is used for transmission, in order to reduce the amount of side information.

[0021] The reason why the KLT matrix is computed from the rescaled inter-channel correlation matrix, not from the original correlation matrix, is to achieve a perfect KLT inversion in the decoder. This requires the additional coding 1260 of the quantized inter-channel correlation matrix to be lossless.

[0022] In the following, the individual steps of the encoding are described in detail. Note that the steps can also be understood as functions of parts of an encoder.

[0023] The first step 110 concerns the segmentation. The multi-channel signal consisting of *I* channels is segmented into frames according to

$$\mathbf{X}(m) = \begin{bmatrix} x_1(l_{m-1}+1) & \dots & x_1(l_m) \\ \vdots & \ddots & \vdots \\ x_I(l_{m-1}+1) & \dots & x_I(l_m) \end{bmatrix} \qquad (37)$$

[0024] The indices $l_{m-1} + 1$ and l_m corresponding to the left and right border of the m-th segment are chosen to obtain constant inter-channel correlations within the segment.

[0025] Concerning the computation of the inter-channel correlation matrix 1210, it is computed by

$$\mathbf{\Sigma}_{\mathbf{X}}(m) = \mathbf{X}(m)\mathbf{X}^H(m), \qquad (38)$$

where $(\cdot)^H$ denotes the joint transposition and complex conjugation.

[0026] Then, the inter-channel correlation matrix $\Sigma_{\mathbf{X}}(m)$ is quantized (e.g. element-wise) 1220 to obtain the quantized inter-channel correlation matrix $\Sigma_{\mathbf{X}}(m)$. The quantization can be accomplished e.g. as proposed in [6].

[0027] Concerning rescaling of quantized inter-channel correlation matrix, the quantized inter-channel correlation matrix $\Sigma_{\mathbf{X}}(m)$ is rescaled 1230 to obtain the approximate correlation matrix $\hat{\Sigma}_{\mathbf{X}}(m)$. It has to be ensured that the hermitian symmetry of the correlation matrix is retained after the rescaling.

[0028] The KLT matrix $\mathbf{B}_{KLT}(m)$ may e.g. be obtained from an eigenvalue decomposition of the quantized correlation matrix

$$\hat{\mathbf{\Sigma}}_{\mathbf{X}}(m) = \mathbf{V}_{\mathbf{X}}(m)\mathbf{\Lambda}_{\mathbf{X}}(m)\mathbf{V}_{\mathbf{X}}^H(m), \qquad (39)$$

where $\mathbf{V}_{\mathbf{X}}(m)$ denotes the matrix composed from the eigenvectors of $\hat{\Sigma}_{\mathbf{X}}(m)$ as its columns and where $\Lambda_{\mathbf{X}}(m)$ denotes a diagonal matrix with the corresponding eigenvalues of $\hat{\Sigma}_{\mathbf{X}}(m)$ on its diagonal. The KLT matrix is then computed by

$$\mathbf{B}_{\mathrm{KLT}}(m) = \mathbf{V}_{\mathbf{X}}^{H}(m). \qquad (40)$$

[0029] For the KLT, the multi-channel signal segments $\mathbf{X}(m)$ are transformed to segments Z(m) of decorrelated channel signals $z_i(l)$, $i = 1, ..., I$, i.e.,

$$\mathbf{Z}(m) = \begin{bmatrix} z_1(l_{m-1}+1) & \dots & z_1(l_m) \\ \vdots & \ddots & \vdots \\ z_I(l_{m-1}+1) & \dots & z_I(l_m) \end{bmatrix}, \qquad (41)$$

by multiplication with the KLT matrix as

$$\mathbf{Z}(m) = \mathbf{B}_{\mathrm{KLT}}(m)\mathbf{X}(m) \qquad (42)$$

[0030] In the channel independent perceptual coding, the segments of the individual decorrelated signals $z_i(l)$, $i = 1,...,I$ are individually perceptually coded 130. The coded representation of all channels within the segment m is denoted by $\breve{\mathbf{Z}}(m)$.

[0031] The quantized inter-channel correlation matrix $\overline{\Sigma}_{\mathbf{X}}(m)$ is losslessly encoded 1260 to obtain the coded representation

$$\breve{\overline{\mathbf{\Sigma}}}_{\mathbf{X}}(m).$$

At coding, the symmetry of the quantized inter-channel correlation matrix $\overline{\Sigma}_{\mathbf{X}}(m)$ can be exploited by coding only the upper (or lower) triangular part and the diagonal.

[0032] In the following, a proposed KLT based perceptual decoding method is described. A conceptual scheme of the KLT based perceptual decoder for multi-channel signals is depicted in Fig.4, and comprises channel independent perceptual decoding 210, re-correlation with an IKLT 220 and desegmentation 230. First, the inter-channel correlation matrix is decoded 2210 and the decorrelated signals are decoded 210, using a channel independent perceptual decoding. Second, the KL transform matrix is computed 2220 in the same way as in the encoder. Finally, the inverse KL transformation is applied 2230 to the decoded decorrelated signals, and the segments are de-segmented (i.e. combined) 230 to continuous streams of individual channel signals.

[0033] In the following, the individual parts of the decoding are described in detail.

[0034] In a lossless decoding and rescaling of inter-channel correlation matrix step 2210, the coded quantized inter-channel correlation matrix

$$\breve{\overline{\mathbf{\Sigma}}}_{\mathbf{X}}(m)$$

is decoded and rescaled (e.g. inverse quantized) to obtain $\hat{\overline{\Sigma}}_{\mathbf{X}}(m)$. Note that since the coding of the correlation matrix is lossless, the decoded matrix $\hat{\overline{\Sigma}}_{\mathbf{X}}(m)$ corresponds exactly to the one which was coded.

[0035] In the channel independent perceptual decoding step 210, the segments $\breve{\mathbf{Z}}(m)$ of the coded decorrelated signals are perceptually decoded, where each channel is decoded independently. The decoded representation of all channel signals within the segment m is denoted by $\hat{\mathbf{Z}}(m)$.

[0036] The KLT matrix $\mathbf{B}_{\mathrm{KLT}}(m)$ is computed 2220 from the inter-channel correlation matrix $\hat{\overline{\Sigma}}_{\mathrm{X}}(m)$ in the same way as it is done in the KLT based perceptual encoder.

[0037] In the inverse KLT (IKLT), the multi-channel signal segments Z(m) are transformed to segments $\hat{\mathbf{X}}(m)$ of correlated channel signals $\hat{x}_i(l)$, $i = 1, ..., I$, i.e.,

$$\hat{\mathbf{X}}(m) = \begin{bmatrix} \hat{x}_1(l_{m-1}+1) & \dots & \hat{x}_1(l_m) \\ \vdots & \ddots & \vdots \\ \hat{x}_I(l_{m-1}+1) & \dots & \hat{x}_I(l_m) \end{bmatrix} \quad (43)$$

by multiplication with the inverse KLT matrix $\mathbf{B}_{\mathrm{KLT}}^{-1}(m) = \mathbf{B}_{\mathrm{KLT}}^{H}(m)$ as

$$\hat{\mathbf{X}}(m) = \mathbf{B}_{\mathrm{KLT}}^{H}(m)\hat{\mathbf{Z}}(m) \quad (44)$$

[0038] In the Desegmentation step 230, the segments $\hat{\mathbf{X}}(m)$ are combined to continuous streams of individual channel signals $\hat{x}_i(l)$, $i = 1, ..., I$.

[0039] In the following, an embodiment is described that is particularly suitable for being used together with transform coders. Many perceptual coders transform the time domain signal to the frequency domain to better adapt the coding to the human psychoacoustics. A prominent example is Advanced Audio Coding (AAC) where the transform is accomplished using a Modified Discrete Fourier Transform (MDCT). For that purpose, the time domain signal is usually processed frame-wise, where successive frames overlap. For this kind of transform coders the proposed encoders and decoders may be slightly modified in order to solve a problem that has been found to occur for the following reason (cf. Fig.5).

[0040] Consider a segmentation of the multi-channel signal as proposed in the KLT based perceptual encoder, as shown in Fig.5, so that several overlapping segments F1,...,F5 are obtained. At the transition between two successive segments, e.g. m and m + 1, the properties within the individual decorrelated signals $z_i(l)$ may abruptly change due to the change of the KLT matrix from $\mathbf{B}_{\mathrm{KLT}}(m)$ to $\mathbf{B}_{\mathrm{KLT}}(m + 1)$. This situation is illustrated in Fig. 5, where for simplicity it has been assumed that the length of the segment used for the KLT is three times as large as the frame shift used for the frequency transform. In Fig.5, the frequency transform frame with changing signal properties F3 is drawn in bold lines.

[0041] There are two major problems resulting from changing signal properties within a single frame used for the time-frequency transform (TFT). First, since the perceptual coder assumes the signal within a single frame to be stationary, this assumption is violated. Consequently, at decoding pre- and post-echoes may occur. Second, including two different signal properties within a single frame makes the corresponding spectrum become more dense. As a result, the coding efficiency is reduced since a more dense spectrum has to be encoded.

[0042] A further aspect is that the TFT can often be regarded as a linear mapping to a space of minor dimension. For example, the MDCT maps a time domain frame of N samples to a frame of $N/2$ frequency bin values. Since this mapping, with respect to this single frame, is not invertible, the cross-correlations between the time domain and frequency domain representations of the respective frame are different. Consequently, using a KLT matrix computed based on the time domain signals does not assure that the frequency domain signals, which are actually coded, are properly decorrelated.

[0043] At least one embodiment of the presently disclosed solution is suitable for preventing all the above mentioned problems for TFT coders. For this purpose, the TFT is performed before carrying out the KLT segmentation in the encoder. Respectively, in the decoder the inverse TFT (ITFT) is performed after the inverse KLT (IKLT), and not before the IKLT. With this approach, changing signal conditions within a single frame employed for the TFT, which are due to changing KLT matrices, are avoided. It is further guaranteed that those signals are decorrelated which are actually used for perceptual encoding. The scheme of the modified KLT based encoding and decoding for TFT based perceptual encoders is illustrated in Fig.6.

[0044] In the KLT based encoder according to one embodiment, the individual time domain signals $x_i(l)$, $i = 1,..., I$ are framed 310 into overlapping frames of N samples before each frame is transformed 320 to the frequency domain by a time-frequency transform (TFT). The resulting short-time spectra S320 are denoted by $X_i(k, n)$, where k denotes the frequency bin index and n denotes the TFT frame index. If, for simplicity, we assume that e.g. an MDCT is used as TFT, there are $N/2$ frequency bins and $k \in \{0,...,N/2 - 1\}$.

[0045] Successively, the short-time spectra are segmented 110 into KLT segments $\mathbf{W}(m)$ as

$$\mathbf{W}(m) := \begin{bmatrix} X_1(0, n_{m-1}) & \dots & X_1\left(\frac{N}{2}-1, n_{m-1}\right) & \dots & X_1(0, n_m) & \dots & X_1\left(\frac{N}{2}-1, n_m\right) \\ \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ X_I(0, n_{m-1}) & \dots & X_I\left(\frac{N}{2}-1, n_{m-1}\right) & \dots & X_I(0, n_m) & \dots & X_I\left(\frac{N}{2}-1, n_m\right) \end{bmatrix} \quad (45)$$

where m denotes the KLT segment index and where n_{m-1} and n_m indicate the indices of the first and last TFT frame included into the m-th KLT segment. The short-time spectra of the individual channels within the KLT segment are decorrelated 120 using the KLT to obtain the segments V(m) of decorrelated short-time spectra S122 and the respective losslessly encoded quantized inter-channel correlation matrix

$$\breve{\boldsymbol{\Sigma}}_{\mathbf{W}}(m)$$

S121. The individual decorrelated short-time spectra within the segment V(m) S122 are individually perceptually coded 130 directly in the frequency domain (i.e., without using a TFT) to obtain the coded representation

$$\breve{\mathbf{V}}(m)$$

S130. The KLT based encoder outputs the losslessly encoded quantized inter-channel correlation matrix

$$\breve{\boldsymbol{\Sigma}}_{\mathbf{W}}(m)$$

S121 and the coded segment

$$\breve{\mathbf{V}}(m)$$

S130 of decorrelated short-time spectra, which can then be transmitted or stored, and subsequently received or retrieved from storage (not shown in Fig.6).

[0046] In the KLT based decoder 400, the frames

$$\breve{\mathbf{V}}(m)$$

with the encoded short-time spectra are perceptually decoded 210 to the frequency domain without using an inverse TFT (ITFT) to obtain the segments $\hat{\mathrm{V}}(m)$ S210 of decoded short-time spectra. These are correlated 220 with the inverse KLT (IKLT), using the coded quantized inter-channel correlation matrix

$$\breve{\boldsymbol{\Sigma}}_{\mathbf{W}}(m)$$

S121. The resulting segments $\hat{\mathrm{V}}(m)$ S220 are desegmented 230 to obtain the decoded short-time spectra $\hat{X}_i(k, n)$, $i = 1, I$ S230, which are successively transformed to the time domain with an ITFT 440 (individually for each channel $\hat{X}_i(k, n)$). After that, successive time domain frames are merged 440 using the overlap add technique and deframed 450 to obtain continuous sequences of time domain signals $\hat{x}_i(l)$, $i = 1,\dots, I$.

[0047] It should be noted that the above description relates to broadband signals, i.e. full bandwidth. However, in an alternative embodiment, the decorrelation and respective correlation may be performed on frequency bands that are related to e.g. the human perception, rather than on the broadband signals. In that case, instead of a single losslessly encoded quantized inter-channel correlation matrix, a number of losslessly encoded quantized frequency-band related correlation matrices are included into the side information.

[0048] In the following, it will be shown analytically why noise unmasking may occur after matrixing the perceptually decoded individual channel signals. Further, it will be shown that the signal decorrelation applied before perceptual encoding is suitable for avoiding this problem. A mathematical problem formulation is as follows.

[0049] Assume that a discrete-time multi-channel signal consisting of I signals $x_i(l)$, $i = 1,\ldots, I$ is given, where l denotes the sample index. The individual signals may be real or complex valued. We consider a frame of L samples beginning at the sample index $l_{START} + 1$, in which the individual signals are assumed to be stationary. The corresponding samples are arranged within the matrix $\mathbf{X} \in C^{I \times L}$ according to

$$\mathbf{X} := [\mathbf{x}(l_{\mathrm{START}} + 1) \quad \ldots \quad \mathbf{x}(l_{\mathrm{START}} + L)] \tag{1}$$

where

$$\mathbf{x}(l) := [x_1(l) \quad \ldots \quad x_I(l)]^T \tag{2}$$

with $(\cdot)^T$ denoting transposition. The corresponding empirical correlation matrix is given by

$$\boldsymbol{\Sigma}_{\mathbf{X}} := \mathbf{X}\mathbf{X}^H \tag{3}$$

where $(\cdot)^H$ denotes the joint complex conjugation and transposition.

[0050] Now we assume that the multi-channel signal frame is coded, thereby introducing coding error noise at reconstruction. Thus the matrix of the reconstructed frame samples, which is denoted by $\hat{\hat{\mathbf{X}}}$, is composed of the true sample matrix **X** and a coding noise component E according to

$$\hat{\hat{\mathbf{X}}} = \mathbf{X} + \mathbf{E} \tag{4}$$

with

$$\mathbf{E} := [\mathbf{e}(l_{\mathrm{START}} + 1) \quad \ldots \quad \mathbf{e}(l_{\mathrm{START}} + L)] \tag{5}$$

and

$$\mathbf{e}(l) := [e_1(l) \quad \ldots \quad e_I(l)]^T \tag{6}$$

[0051] Since it can be assumed that each channel has been coded independently, the coding noise signals $e_i(l)$ can be assumed to be independent of each other for $i = 1,\ldots, I$. Exploiting this property and the assumption, that the noise signals are zero-mean, the empirical correlation matrix of the noise signals is given by a diagonal matrix as

$$\boldsymbol{\Sigma}_{\mathbf{E}} = \mathrm{diag}\left(\sigma_{e_1}^2, \ldots, \sigma_{e_I}^2\right). \tag{7}$$

[0052] Here, $\mathrm{diag}\left(\sigma_{e_1}^2, \ldots, \sigma_{e_I}^2\right)$ denotes a diagonal matrix with the empirical noise signal powers

$$\sigma_{e_i}^2 = \frac{1}{L}\sum_{l=l_{\mathrm{START}}+1}^{l_{\mathrm{START}}+L} |e_i(l)|^2 \tag{8}$$

on its diagonal. A further essential assumption is that the coding is performed such that a predefined signal-to-noise ratio (SNR) is satisfied for each channel. Without loss of generality, we assume that the predefined SNR is equal for each channel, i.e.,

$$\mathrm{SNR}_x = \frac{\sigma_{x_i}^2}{\sigma_{e_i}^2} \quad for\ all\ \ i = 1, \ldots, I \tag{9}$$

with

$$\sigma_{x_i}^2 := \frac{1}{L}\sum_{l=l_{\mathrm{START}}+1}^{l_{\mathrm{START}}+L} |x_i(l)|^2 \tag{10}$$

[0053] From now on we consider the matrixing of the reconstructed signals into J new signals $y_j(l)$, $j = 1, \ldots, J$. Without introducing any coding error the sample matrix of the matrixed signals may be expressed by

$$\mathbf{Y} = \mathbf{AX}, \tag{11}$$

where $A \in C^{J \times I}$ denotes the mixing matrix and where

$$\mathbf{Y} := [\mathbf{y}(l_{\mathrm{START}} + 1) \quad \ldots \quad \mathbf{y}(l_{\mathrm{START}} + L)] \tag{12}$$

with

$$\mathbf{y}(l) := [y_1(l) \quad \ldots \quad y_J(l)]^T \tag{13}$$

[0054] However, due to coding noise the sample matrix of the matrixed signals is given by

$$\hat{\hat{\mathbf{Y}}} := \mathbf{Y} + \mathbf{N} \tag{14}$$

with N being the matrix containing the samples of the matrixed noise signals. It can be expressed as

$$\mathbf{N} = \mathbf{AE} \tag{15}$$

$$= [\mathbf{n}(l_{\mathrm{START}} + 1) \quad \ldots \quad \mathbf{n}(l_{\mathrm{START}} + L)] \tag{16}$$

where

$$\mathbf{n}(l) := [n_1(l) \quad \dots \quad n_J(l)]^T \tag{17}$$

is the vector of all matrixed noise signals at the sample index *l*.

[0055] Exploiting equation (11), the empirical correlation matrix of the matrixed noise-free signals can be formulated as

$$\mathbf{\Sigma_Y} = \mathbf{A}\mathbf{\Sigma_X}\mathbf{A}^H \tag{18}$$

[0056] Thus, the empirical power of the j-th matrixed noise-free signal, which is the j-th element on the diagonal of $\Sigma_{\mathbf{Y}}$, may be written as

$$\sigma_{y_j}^2 = \mathbf{a}_j{}^H \mathbf{\Sigma_X} \mathbf{a}_j, \tag{19}$$

where $\mathbf{a}_j$ is the j-th column of $\mathbf{A}^H$ according to

$$\mathbf{A}^H = [\mathbf{a}_1 \quad \dots \quad \mathbf{a}_J] \tag{20}$$

[0057] Similarly, with equation (15), the empirical correlation matrix of the matrixed noise signals can be written as

$$\mathbf{\Sigma_N} = \mathbf{A}\mathbf{\Sigma_E}\mathbf{A}^H \tag{21}$$

[0058] The empirical power of the j-th matrixed noise signal, which is the j-th element on the diagonal of $\Sigma_{\mathbf{N}}$, is given by

$$\sigma_{n_j}^2 = \mathbf{a}_j{}^H \mathbf{\Sigma_E} \mathbf{a}_j \tag{22}$$

[0059] Consequently, the empirical SNR of the matrixed signals, which is defined by

$$\mathrm{SNR}_{y_j} := \frac{\sigma_{y_j}^2}{\sigma_{n_j}^2} \tag{23}$$

[0060] can be reformulated using equations (19) and (22) as

$$\mathrm{SNR}_{y_j} = \frac{\mathbf{a}_j{}^H \mathbf{\Sigma_X} \mathbf{a}_j}{\mathbf{a}_j{}^H \mathbf{\Sigma_E} \mathbf{a}_j} \tag{24}$$

[0061] By decomposing $\Sigma_{\mathbf{X}}$ into its diagonal and non-diagonal component as

$$\mathbf{\Sigma_X} = \mathrm{diag}\left(\sigma_{x_1}^2, \dots, \sigma_{x_I}^2\right) + \mathbf{\Sigma_{X,NG}} \tag{25}$$

with

$$\boldsymbol{\Sigma}_{\mathbf{X},\mathrm{NG}} := \boldsymbol{\Sigma}_{\mathbf{X}} - \mathrm{diag}\left(\sigma_{x_1}^2, \ldots, \sigma_{x_I}^2\right) \tag{26}$$

and by exploiting the property

$$\mathrm{diag}\left(\sigma_{x_1}^2, \ldots, \sigma_{x_I}^2\right) = \mathrm{SNR}_x \cdot \mathrm{diag}\left(\sigma_{e_1}^2, \ldots, \sigma_{e_I}^2\right) \tag{27}$$

resulting from the assumptions (7) and (9) (recall that the predefined coding SNR has been assumed to be equal for each channel), we finally obtain the desired expression for the empirical SNR of the matrixed signals:

$$\mathrm{SNR}_{y_j} = \frac{\mathbf{a}_j{}^H \mathrm{diag}\left(\sigma_{x_1}^2, \ldots, \sigma_{x_I}^2\right)\mathbf{a}_j}{\mathbf{a}_j{}^H \boldsymbol{\Sigma}_{\mathbf{E}} \mathbf{a}_j} + \frac{\mathbf{a}_j{}^H \boldsymbol{\Sigma}_{\mathbf{X},\mathrm{NG}} \mathbf{a}_j}{\mathbf{a}_j{}^H \boldsymbol{\Sigma}_{\mathbf{E}} \mathbf{a}_j} \tag{28}$$

$$= \mathrm{SNR}_x \left(1 + \frac{\mathbf{a}_j{}^H \boldsymbol{\Sigma}_{\mathbf{X},\mathrm{NG}} \mathbf{a}_j}{\mathbf{a}_j{}^H \mathrm{diag}\left(\sigma_{x_1}^2, \ldots, \sigma_{x_I}^2\right)\mathbf{a}_j}\right) \tag{29}$$

[0062] From this expression it can be seen that this SNR is obtained from the predefined SNR, SNR_x, by the multiplication with a term, which is dependent on the diagonal and non-diagonal component of the signal correlation matrix $\Sigma_{\mathbf{X}}$. Thus, the empirical SNR of the matrixed signals is equal to the predefined SNR if the signals $x_i(l)$ are uncorrelated to each other such that $\Sigma_{\mathbf{X},\mathrm{NG}}$ becomes a zero matrix, i.e.,

$$\mathrm{SNR}_{y_j} = \mathrm{SNR}_x \quad for\ all\ \ j = 1, \ldots, J, \quad if\ \ \boldsymbol{\Sigma}_{\mathbf{X},\mathrm{NG}} = \mathbf{0}_{I \times I} \tag{30}$$

with $\mathbf{0}_{I \times I}$ denoting a zero matrix with I rows and columns.

[0063] However, if the signals $x_i(l)$ are correlated, the empirical SNR of the matrixed signals may deviate from the predefined SNR. In the worst case, SNR_{yj} can be much lower than SNR_x, which means a potential unmasking of coding noise.

[0064] The proposed solution according to the present invention is based on the following. As has been shown above, noise unmasking at matrixing happens if the signals $x_i(l)$ are correlated. Hence, in one embodiment of the present invention, a Karhunen-Loève Transform (KLT) is applied for decorrelating the signals $x_i(l)$ before coding, and an inverse KLT for re-correlating the decorrelated signals is applied after decoding. In particular, the matrix Z containing the samples of the transformed signals $z_i(l)$, $i = 1, \ldots, I$ as

$$\mathbf{Z} = \begin{bmatrix} z_1(l_{\mathrm{START}} + 1) & \ldots & z_1(l_{\mathrm{START}} + L) \\ \vdots & \ddots & \vdots \\ z_I(l_{\mathrm{START}} + 1) & \ldots & z_I(l_{\mathrm{START}} + L) \end{bmatrix} \tag{31}$$

can be expressed by

$$\mathbf{Z} = \mathbf{B}_{\mathrm{KLT}}\mathbf{X} \tag{32}$$

where $\mathbf{B}_{\mathrm{KLT}}$ denotes the KLT matrix. It can be obtained from the following eigenvalue decomposition of the correlation matrix $\Sigma_{\mathbf{X}}$

$$\mathbf{\Sigma}_{\mathbf{X}} = \mathbf{V}_{\mathbf{X}}\mathbf{\Lambda}_{\mathbf{X}}{\mathbf{V}_{\mathbf{X}}}^{H} \tag{33}$$

[0065] In (33), $\mathbf{V}_{\mathbf{X}}$ denotes a unitary matrix containing the eigenvectors as columns and satisfying the property

$${\mathbf{V}_{\mathbf{X}}}^{H}\mathbf{V}_{\mathbf{X}} = \mathbf{I} \tag{34}$$

with **I** being the identity matrix. Further, $\Lambda_{\mathbf{X}}$ denotes a diagonal matrix containing the respective eigenvalues on its diagonal. The KLT matrix is then given by

$$\mathbf{B}_{\mathrm{KLT}} = {\mathbf{V}_{\mathbf{X}}}^{H} \tag{35}$$

[0066] It can be verified with equations (32), (35) and (34) that the correlation matrix of the transformed signals is given by

$$\mathbf{\Sigma}_{\mathbf{Z}} = \mathbf{Z}\mathbf{Z}^{H} = \mathbf{B}_{\mathrm{KLT}}\mathbf{\Sigma}_{\mathbf{X}}{\mathbf{B}_{\mathrm{KLT}}}^{H} = {\mathbf{V}_{\mathbf{X}}}^{H}\mathbf{V}_{\mathbf{X}}\mathbf{\Lambda}_{\mathbf{X}}{\mathbf{V}_{\mathbf{X}}}^{H}\mathbf{V}_{\mathbf{X}} = \mathbf{\Lambda}_{\mathbf{X}} \tag{36}$$

[0067] In particular, the correlation matrix is diagonal, i.e. all non-diagonal elements are zero. Hence, as shown in the previous subsection, if the transformed signals $z_i(l)$ are coded instead of the original signals $x_i(1)$, any matrixing of the decoded transformed signals will not change the SNR in the matrixed signals. That means that any noise unmasking can be avoided.

[0068] Perceptual coding of audio signals means a coding that is adapted to the human perception of audio. It should be noted that when perceptually coding the audio signals, a quantization is usually performed not on the broad-band audio signal samples, but rather in individual frequency bands related to the human perception. Hence, the ratio between the signal power and the quantization noise may vary between the individual frequency bands. To be precise, in order to completely avoid noise unmasking in these cases, it would be necessary to decorrelate the band-pass signals rather than the broad-band audio channel signals. This, however, increases the amount of side information to be transmitted by a factor equal to the number of frequency bands considered. Assuming a fixed given data rate, this reduces the total rate for the coded audio signals. If the decorrelation effect compared to the case of a single broad-band signal decorrelation matrix is distinctly better, than it might be reasonable to prefer this decorrelation method. However, which of the decorrelation methods (either broadband signal decorrelation or decorrelation in frequency bands,) is most beneficial depends on the correlation properties of the input audio signals $x_i(l)$.

[0069] In one embodiment, an apparatus for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals ($x_1(l)$,..., $x_I(l)$), comprises a de-correlator 10 for de-correlating the multi-channel audio signals, wherein de-correlated multi-channel audio signals S10 and correlation information S11 are obtained;
a perceptual encoder 20 for perceptually encoding the de-correlated multi-channel audio signals;
a perceptual decoder 30 for perceptually decoding the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals S31 are obtained;
a re-correlator 40 for re-correlating the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information S11, wherein re-correlated, perceptually decoded multi-channel audio signals S40 are obtained; and a mixer 50 for mixing the re-correlated, perceptually decoded multi-channel audio signals S40, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal $\hat{y}_1(l)$, and wherein a mixing scheme is used that is unknown in said perceptual encoder.

[0070] In one embodiment, an apparatus for avoiding the unmasking of coding noise when mixing perceptually coded time-frequency transformed multi-channel audio signals, comprising a perceptual encoding unit and a perceptual de-

coding unit, and the perceptual encoding unit comprises

[0071] TFT framing units 310 for time-frequency transform framing and TFT units 320 for time-frequency transforming the multi-channel audio signals $x_1(l),..., x_I(l)$, wherein time-frequency transformed multi-channel audio signals $X_1(k,n),..., X_I(k,n)$ S320 are obtained; a segmentation unit 110 for performing a transform segmentation on the time-frequency transformed multi-channel audio signals S320; and a de-correlator 120 for de-correlating with a KLT the segments obtained in the transform segmentation;

and the perceptual decoding unit 400 comprises

a decoder 210 for channel independent decoding without inverse time-frequency transform, wherein a channel independent decoded signal S21 0 is obtained;

a correlator 220 for re-correlating the channel independent decoded signal S21 0 with a KLT, wherein re-correlated signals S220 are obtained;

a de-segmentation unit 230 performing a transform de-segmentation on the re-correlated signals S220, wherein time-frequency transformed multi-channel audio signals S230 are obtained;

[0072] ITFT units 440 for performing inverse time-frequency transform with overlap add on the time-frequency transformed multi-channel audio signals S230, wherein framed time-domain multi-channel audio signals S440 are obtained; and a de-framer 450 for de-framing the framed time-domain multi-channel audio signals S440, wherein the re-correlated, perceptually decoded multi-channel audio signals S40 are obtained; and

a mixer 50 for mixing the re-correlated, perceptually decoded multi-channel audio signals S40, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal $\hat{y}_1(l)$, and wherein a mixing scheme is used,

wherein in said de-correlator 120 the de-correlating the segments obtained in the transform segmentation during encoding is independent from the mixing scheme that is used in the mixer 50.

[0073] In one embodiment, an apparatus for avoiding the unmasking of coding noise when mixing, in a perceptual decoder, multi-channel audio signals that have been perceptually coded by using a time-frequency transform, comprises

decoder 210 for channel independent decoding without inverse time-frequency transform, wherein channel independent decoded signals S21 0 are obtained;

a correlation unit 220 for re-correlating the channel independent decoded signal S210 with a KLT, wherein re-correlated signals S220 are obtained;

a de-segmenter 230 for performing a transform de-segmentation of the re-correlated signals S220, wherein time-frequency transformed multi-channel audio signals S230 are obtained; ITFT units 440 for performing inverse time-frequency transform with overlap add on the time-frequency transformed multi-channel audio signals S230, wherein framed time-domain multi-channel audio signals S440 are obtained; and a de-framing unit 450 for de-framing the framed time-domain multi-channel audio signals S440, wherein the re-correlated, perceptually decoded multi-channel audio signals S40 are obtained; and a mixer 50 for mixing the re-correlated, perceptually decoded multi-channel audio signals S40 according to a mixing scheme, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal $\hat{y}_1(l)$.

[0074] In one embodiment similar to the one shown in Fig.2, an encoder and decoder are adapted for automatically switching between channel-based audio signals (i.e. loudspeaker channel related) and sound field audio signals, such as HOA. Both required different matrices in the matrixing block at the decoder. While input signals $x_i(l)$ have one particular format assigned (such as e.g. loudspeaker channel based format or HOA format), such information is transmitted or stored as side information S11, received or retrieved on the decoding side and used in the matrixing unit 50 to automatically adapt the matrix for the respective signal format (e.g. switch between loudspeaker channel based and HOA formats).

[0075] It should be noted that although shown simply as KLT, other types of transforms may be constructed other than KLT, as would be apparent to those of ordinary skill in the art, all of which are contemplated within the spirit and scope of the invention.

[0076] While there has been shown, described, and pointed out fundamental novel features of the present invention as applied to preferred embodiments thereof, it will be understood that various omissions and substitutions and changes in the apparatus and method described, in the form and details of the devices disclosed, and in their operation, may be made by those skilled in the art without departing from the spirit of the present invention. It is expressly intended that all combinations of those elements that perform substantially the same function in substantially the same way to achieve the same results are within the scope of the invention. Substitutions of elements from one described embodiment to another are also fully intended and contemplated.

[0077] It will be understood that the present invention has been described purely by way of example, and modifications of detail can be made without departing from the scope of the invention.

[0078] Each feature disclosed in the description and (where appropriate) the claims and drawings may be provided independently or in any appropriate combination. Features may, where appropriate be implemented in hardware, software, or a combination of the two. Connections may, where applicable, be implemented as wireless connections or wired, not necessarily direct or dedicated, connections. Reference numerals appearing in the claims are by way of

illustration only and shall have no limiting effect on the scope of the claims.

Cited References

[0079]

[1] Erik Hellerud and Ian Burnett and Audun Solvang and U. Peter Svensson. Encoding Higher Order Ambisonics with AAC. 124th AES Convention, Amsterdam, 2008.

[2] Peter Jax and Jan-Mark Batke and Johannes Boehm and Sven Kordon. Perceptual Coding of HOA Signals in Spatial Domain. Patent application (Technicolor Internal Reference: PD100051), 2010.

[3] Jolliffe, I. T. Principal Component Analysis. Springer, Second edition, 2002.

[4] M. A. Poletti. Three-Dimensional Surround Sound Systems Based on Spherical Harmonics. J. Audio Eng. Soc., 53(11):1004-1025, 2005.

[5] Soledad Torres-Guijarro and Jon A. Beracoechea-Álava and Luis I. Ortiz-Berenguer and F. Javier Casajús-Quirós. Inter-channel de-correlation for perceptual audio coding. Applied Acoustics, 66(8):889 - 901, 2005.

[6] Yang, Dai and Ai, Hongmei and Kyriakakis, C. and Kuo, C. -C. J. High-fidelity multichannel audio coding with Karhunen-Loeve transform. IEEE Transactions on Speech and Audio Processing, 11 (4):365--380, 2003.

[7] Kate, W.R.T. ten ; Boers, P.M. ; Makivirta, A. ; Kuusama, J. ; Christensen, K.E. ; Sorensen, E.: Matrixing of bit rate reduced audio signals. Proc. of the ICASSP, (2), 205-208, March 1992.

[8] Kate, Warner R. ten: Compatibility Matrixing of Multichannel Bit-Rate Reduced Audio Signals. 96th Convention of the Audio Eng. Soc., Feb. 1994.

Claims

1. A method for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals ($x_1(l)$,..., $x_I(l)$), comprising steps of

   - de-correlating (10) the multi-channel audio signals, wherein de-correlated multi-channel audio signals (S10) and correlation information (S11) are obtained;
   - perceptually encoding (20) the de-correlated multi-channel audio signals;
   - perceptually decoding (30) the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals (S31) are obtained;
   - re-correlating (40) the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information (S11), wherein re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
   - mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40), wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$), and wherein a mixing scheme is used that is unknown in said step of encoding.

2. Method according to claim 1, wherein the mixing (50) of the re-correlated, perceptually decoded multi-channel audio signals (S40) comprises mapping multi-channel audio signals (S40) of a first multi-channel audio format to multi-channel audio signals (S51) of a second multi-channel audio format.

3. Method according to claim 1 or 2, wherein a plurality (J) of output signals (S51) are obtained in the step of mixing (50), at least one of the output signals (S51) being a combination of at least two of the re-correlated, perceptually decoded multi-channel audio signals (S40), and wherein the step of mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40) comprises modifying the number of channels.

4. Method according to any of the claims 1-3, wherein the mixing is performed by a matrix operation according to loudspeaker positions.

5. Method according to any of the claims 1-4, wherein a KLT is used in the step of de-correlating (10), and an iKLT is used in the step of re-correlating (40), and said correlation information (S11) comprises coefficients of a symmetric inter-channel correlation matrix.

6. Method for avoiding the unmasking of coding noise when mixing multi-channel audio signals that have been perceptually coded by using a time-frequency transform, comprising perceptual encoding and perceptual decoding, and the perceptual encoding comprising steps of

- time-frequency transform framing (310) and time-frequency transforming (320) the multi-channel audio signals ($x_1(l)$,..., $x_I(l)$), wherein time-frequency transformed multi-channel audio signals ($X_1(k,n)$,...,$X_I(k,n)$) (S320) are obtained;
- performing a transform segmentation (110) on the time-frequency transformed multi-channel audio signals (S320); and
- de-correlating (120) with a KLT the segments obtained in the transform segmentation;

and the perceptual decoding (400) comprising steps of
- channel independent decoding without inverse time-frequency transform (210), wherein channel independent decoded signals (S21 0) are obtained;
- re-correlating (220) the channel independent decoded signals (S21 0) with a KLT, wherein re-correlated signals (S220) are obtained;
- performing a transform de-segmentation (230) on the re-correlated signals (S220), wherein time-frequency transformed multi-channel audio signals (S230) are obtained;
- performing inverse time-frequency transform with overlap add (440) on the time-frequency transformed multi-channel audio signals (S230), wherein framed time-domain multi-channel audio signals (S440) are obtained; and
- de-framing (450) the framed time-domain multi-channel audio signals (S440), wherein the re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
- mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40), wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$), and wherein a mixing scheme is used,

wherein said step of de-correlating (120) with a KLT the segments obtained in the transform segmentation during encoding is independent from the mixing scheme that is used in the step of mixing (50).

7. Method according to one of the claims 1-6, wherein the step of mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40) comprises mapping the encoded multi-channel audio signal to a channel-based audio format with a different number of channels.

8. Method according to any of claims 1-6, wherein the multi-channel audio signal is a soundfield description, and the step of mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40) comprises rendering the soundfield description to a loudspeaker related channel-based audio format according to loudspeaker positions.

9. Method according to claim 8, wherein the soundfield description is given by a Higher Order Ambisonics representation, and the step of mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40) comprises mapping the soundfield description to a plurality of loudspeaker signals.

10. Method according to any of claims 1-9, further comprising a step of transmitting, and/or steps of storing and retrieving, the perceptually encoded multi-channel audio signals (S21) and the correlation information (S11) before the perceptual decoding (30, 400).

11. A method for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals ($x_1(l)$,...,$x_I(l)$) within a decoder, comprising steps of

- receiving through a connection, or retrieving from a storage, perceptually encoded, de-correlated multi-channel audio signals (S21) and correlation information (S11);
- perceptually decoding (30) the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals (S31) are obtained;
- re-correlating (40) the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information (S11), wherein re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
- mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40), wherein at least two of

the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$).

12. A method (400) for avoiding the unmasking of coding noise when mixing in a perceptual decoder multi-channel audio signals that have been perceptually coded by using a time-frequency transform, comprising steps of

- channel independent decoding without inverse time-frequency transform (210), wherein channel independent decoded signals (S21 0) are obtained;
- re-correlating (220) the channel independent decoded signals (S21 0) with a KLT, wherein re-correlated signals (S220) are obtained;
- performing a transform de-segmentation (230) the re-correlated signals (S220), wherein time-frequency transformed multi-channel audio signals (S230) are obtained;
- performing inverse time-frequency transform with overlap add (440) on the time-frequency transformed multi-channel audio signals (S230), wherein framed time-domain multi-channel audio signals (S440) are obtained; and
- de-framing (450) the framed time-domain multi-channel audio signals (S440), wherein the re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
- mixing (50) the re-correlated, perceptually decoded multi-channel audio signals (S40) according to a mixing scheme, wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$).

13. Method according to claim 12, wherein in the step of re-correlating (220) the channel independent decoded signal (S21 0) with a KLT, re-correlation is independent from the mixing scheme that is used in the step of mixing (50).

14. Method according to any one of the claims 1-13, wherein the decoding receives side information from the encoding, the side information defining an encoding mode of the encoded multi-channel audio signals, wherein the encoding mode comprises at least one of channel-based mode and soundfield representation mode, and wherein a decoding mode according to the side information is set for the decoding.

15. An apparatus being adapted for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals ($x_1(l)$,...,$x_l(l)$), comprising

- de-correlator (10) for de-correlating the multi-channel audio signals,
wherein de-correlated multi-channel audio signals (S10) and correlation information (S11) are obtained; and
- perceptual encoder (20) for perceptually encoding the de-correlated multi-channel audio signals;
- perceptual decoder (30) for perceptually decoding the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals (S31) are obtained;
- re-correlator (40) for re-correlating the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information (S11), wherein re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
- mixer (50) for mixing the re-correlated, perceptually decoded multi-channel audio signals (S40), wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$).

16. Decoder being adapted for avoiding the unmasking of coding noise when mixing perceptually coded multi-channel audio signals ($x_1(l)$,...,$x_l(l)$), comprising

- interface for receiving through a connection, or retrieving from a storage, perceptually encoded, de-correlated multi-channel audio signals (S21) and correlation information (S11);
- perceptual decoder (30) for perceptually decoding the perceptually encoded, de-correlated multi-channel audio signals, wherein perceptually decoded, de-correlated multi-channel audio signals (S31) are obtained;
- re-correlator (40) for re-correlating the perceptually decoded, de-correlated multi-channel audio signals according to the correlation information (S11), wherein re-correlated, perceptually decoded multi-channel audio signals (S40) are obtained; and
- mixer (50) for mixing the re-correlated, perceptually decoded multi-channel audio signals (S40), wherein at least two of the re-correlated, perceptually decoded multi-channel audio signals are combined to obtain at least one audio output signal ($\hat{y}_1(l)$).

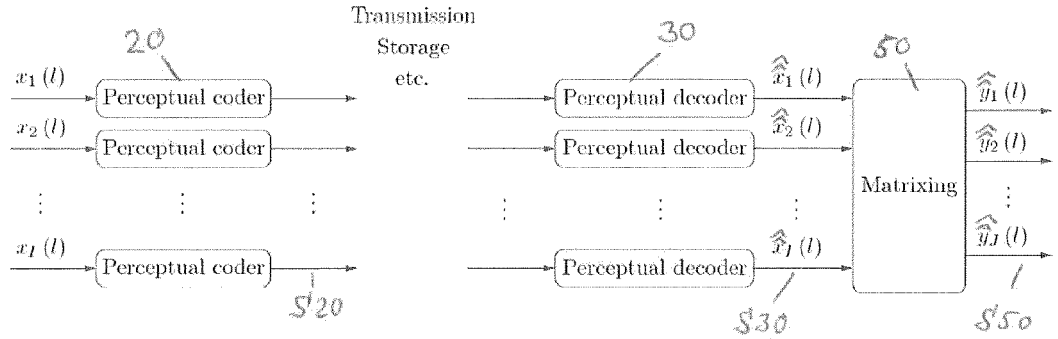
Transmission
Storage
etc.
20
30
50
$x_1(l)$
$x_2(l)$
$x_I(l)$
Perceptual coder
Perceptual coder
Perceptual coder
Perceptual decoder
Perceptual decoder
Perceptual decoder
$\hat{\hat{x}}_1(l)$
$\hat{\hat{x}}_2(l)$
$\hat{\hat{x}}_I(l)$
Matrixing
$\hat{\hat{y}}_1(l)$
$\hat{\hat{y}}_2(l)$
$\hat{\hat{y}}_J(l)$
S20
S30
S50

Fig.1

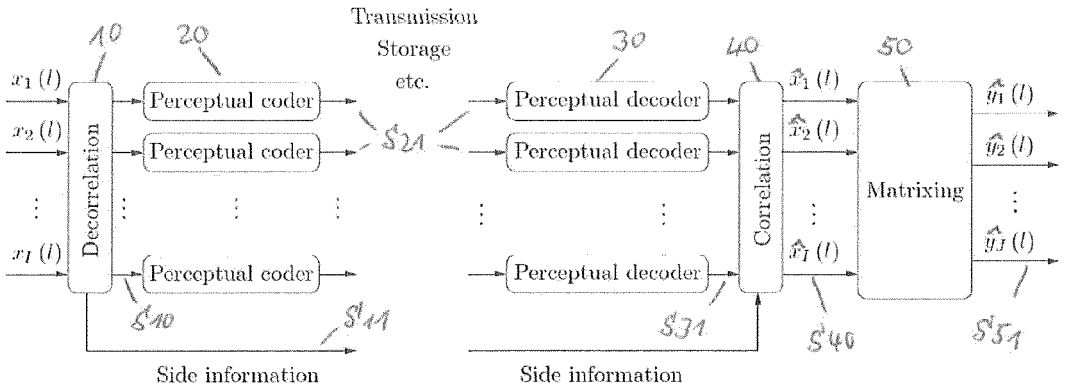
Transmission
Storage
etc.
10
20
30
40
50
$x_1(l)$
$x_2(l)$
$x_I(l)$
Decorrelation
Perceptual coder
Perceptual coder
Perceptual coder
S21
Perceptual decoder
Perceptual decoder
Perceptual decoder
Correlation
$\hat{x}_1(l)$
$\hat{x}_2(l)$
$\hat{x}_I(l)$
Matrixing
$\hat{y}_1(l)$
$\hat{y}_2(l)$
$\hat{y}_J(l)$
S10
S11
S31
S40
S51
Side information
Side information

Fig.2

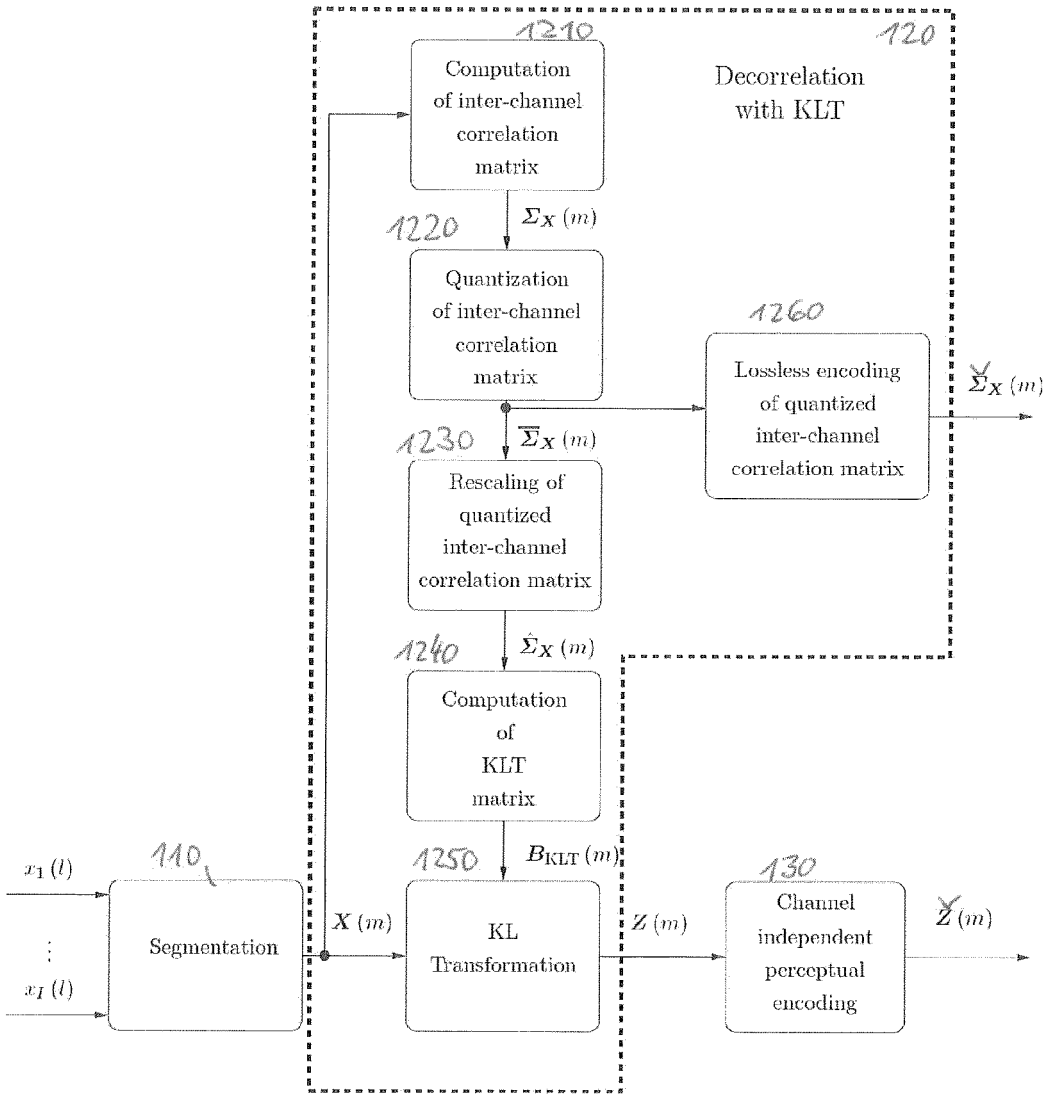
1210
120
Computation of inter-channel correlation matrix
Decorrelation with KLT
$\Sigma_X(m)$
1220
Quantization of inter-channel correlation matrix
1260
Lossless encoding of quantized inter-channel correlation matrix
$\hat{\Sigma}_X(m)$
1230
$\overline{\Sigma}_X(m)$
Rescaling of quantized inter-channel correlation matrix
1240
$\hat{\Sigma}_X(m)$
Computation of KLT matrix
1250
$B_{KLT}(m)$
$x_1(l)$
$x_I(l)$
110
Segmentation
$X(m)$
KL Transformation
$Z(m)$
130
Channel independent perceptual encoding
$\hat{Z}(m)$

Fig.3

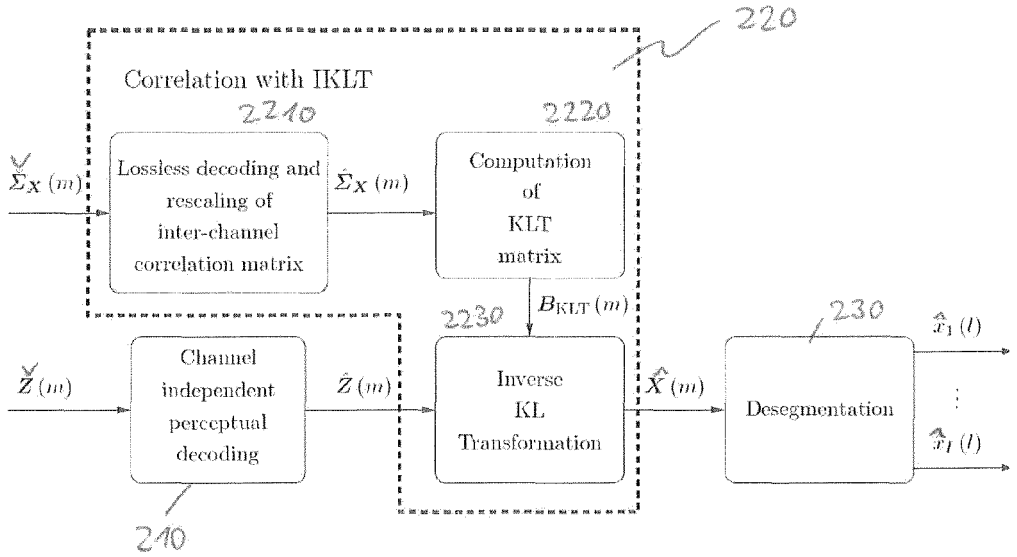
220
Correlation with IKLT
2210
2220
Lossless decoding and rescaling of inter-channel correlation matrix
Computation of KLT matrix
$\check{\Sigma}_X(m)$
$\hat{\Sigma}_X(m)$
$B_{KLT}(m)$
2230
230
$\hat{x}_1(l)$
$\hat{x}_I(l)$
$\check{Z}(m)$
Channel independent perceptual decoding
$\hat{Z}(m)$
Inverse KL Transformation
$\hat{X}(m)$
Desegmentation
210

Fig.4

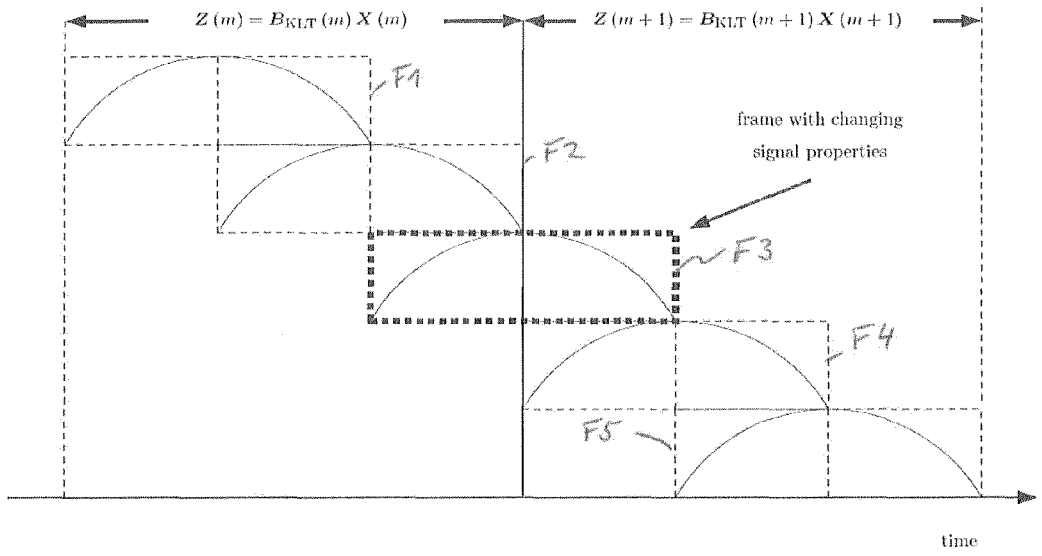
$Z(m) = B_{KLT}(m)\,X(m)$
$Z(m+1) = B_{KLT}(m+1)\,X(m+1)$
F1
F2
frame with changing signal properties
F3
F4
F5
time

Fig.5

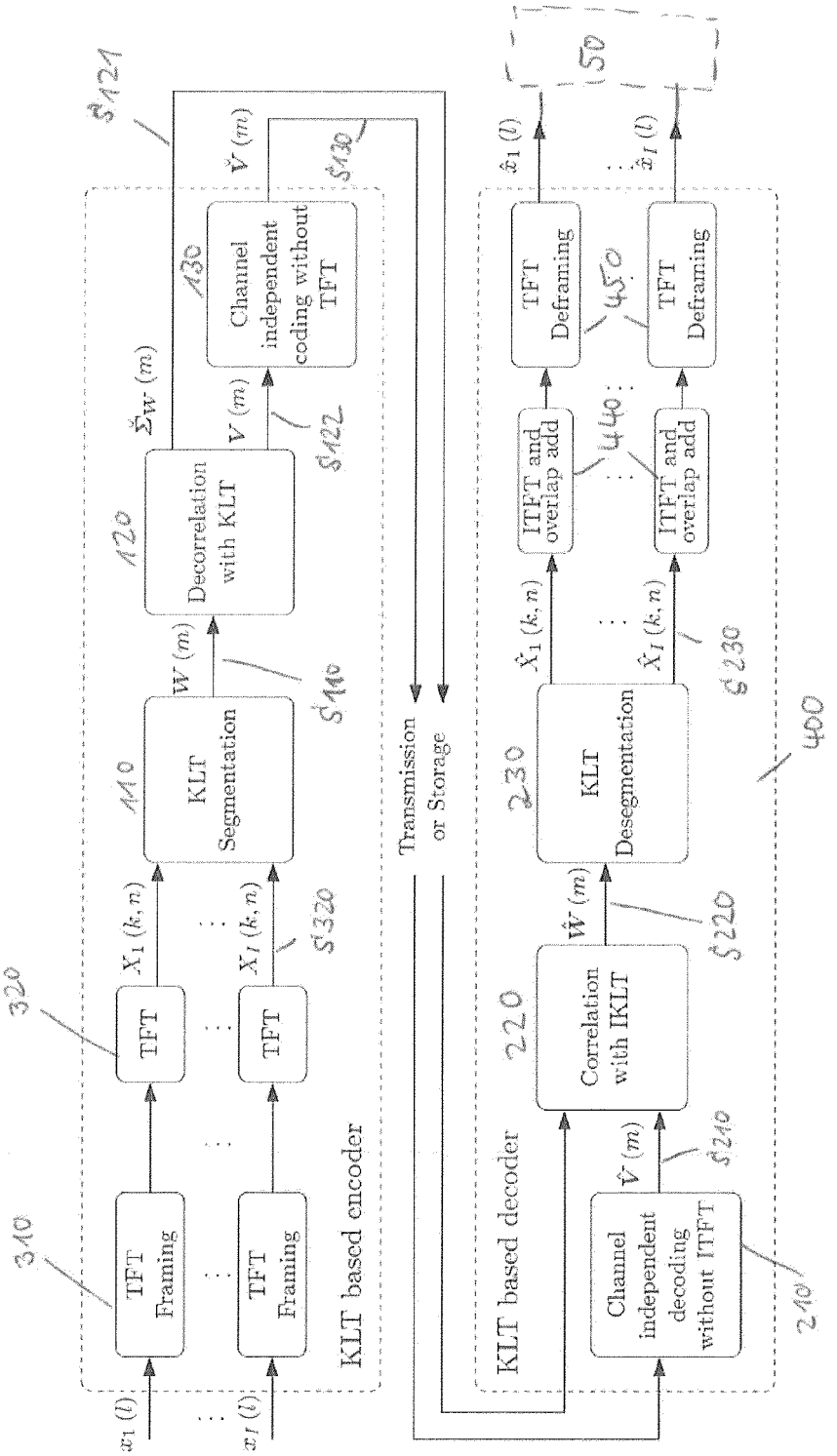
KLT based encoder
TFT Framing
TFT
KLT Segmentation
Decorrelation with KLT
Channel independent coding without TFT
Transmission or Storage
KLT based decoder
Channel independent decoding without ITFT
Correlation with IKLT
KLT Desegmentation
ITFT and overlap add
TFT Deframing

Fig.6

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 12 30 5860

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	MAURI VÄÄNÄNEN: "Robustness Issues in Multi-View Audio Coding", AES CONVENTION 125; OCTOBER 2008, AES, 60 EAST 42ND STREET, ROOM 2520 NEW YORK 10165-2520, USA, 1 October 2008 (2008-10-01), XP040508860, * section 1; section 2; section 3; section 4; section 4.1; section 4.2; section 4.3 *	1-16	INV. G10L19/008 H04S3/02
A	YANG DAI ET AL: "An Inter-Channel Redundancy Removal Approach for High-Quality Multichannel Audio Compression", AES 109TH CONVENTION, LOS ANGELES, , 22 September 2000 (2000-09-22), pages 1-14, XP002517098, Retrieved from the Internet: URL:http://www.aes.org/tmpFiles/elib/20090227/9100.pdf [retrieved on 2000-09-01] * section II; section III *	1,5,6, 10-13, 15,16	TECHNICAL FIELDS SEARCHED (IPC) G10L H04S

-/--

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	31 January 2013	De Meuleneire, M

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

EPO FORM 1503 03.82 (P04C01)

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 12 30 5860

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A,D	POLETTI M A: "THREE-DIMENSIONAL SURROUND SOUND SYSTEMS", JOURNAL OF THE AUDIO ENGINEERING SOCIETY, AUDIO ENGINEERING SOCIETY, NEW YORK, NY, US, vol. 53, no. 11, 1 November 2005 (2005-11-01), pages 1004-1025, XP001243400, ISSN: 1549-4950 * the whole document *	1-16	
A,D	BURNETT IAN ET AL: "Encoding Higher Order Ambisonics with AAC", AES CONVENTION 124; MAY 2008, AES, 60 EAST 42ND STREET, ROOM 2520 NEW YORK 10165-2520, USA, 1 May 2008 (2008-05-01), XP040508582, * the whole document *	1-16	
			TECHNICAL FIELDS SEARCHED (IPC)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	31 January 2013	De Meuleneire, M

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

EPO FORM 1503 03.82 (P04C01)

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

- **ERIK HELLERUD ; LAN BURNETT ; AUDUN SOLVANG ; U. PETER SVENSSON.** Encoding Higher Order Ambisonics with AAC. *124th AES Convention,* 2008 **[0079]**
- **PETER JAX ; JAN-MARK BATKE ; JOHANNES BOEHM ; SVEN KORDON.** Perceptual Coding of HOA Signals in Spatial Domain. *Patent application (Technicolor Internal Reference: PD100051,* 2010 **[0079]**
- **JOLLIFFE, I. T.** Principal Component Analysis. Springer, 2002 **[0079]**
- **M. A. POLETTI.** Three-Dimensional Surround Sound Systems Based on Spherical Harmonics. *J. Audio Eng. Soc.,* 2005, vol. 53 (11), 1004-1025 **[0079]**
- **SOLEDAD TORRES-GUIJARRO ; JON A. BERACOECHEA-ÁLAVA ; LUIS I. ORTIZ-BERENGUER ; F. JAVIER CASAJÚS-QUIRÓS.** Inter-channel de-correlation for perceptual audio coding. *Applied Acoustics,* 2005, vol. 66 (8), 889-901 **[0079]**
- **YANG, DAI ; AI, HONGMEI ; KYRIAKAKIS, C. ; KUO, C. -C. J.** High-fidelity multichannel audio coding with Karhunen-Loeve transform. *IEEE Transactions on Speech and Audio Processing,* 2003, vol. 11 (4), 365-380 **[0079]**
- **KATE, W.R.T. TEN ; BOERS, P.M. ; MAKIVIRTA, A. ; KUUSAMA, J. ; CHRISTENSEN, K.E. ; SORENSEN, E.** Matrixing of bit rate reduced audio signals. *Proc. of the ICASSP,* March 1992, vol. 2, 205-208 **[0079]**
- **KATE, WARNER R. TEN.** Compatibility Matrixing of Multichannel Bit-Rate Reduced Audio Signals. *96th Convention of the Audio Eng. Soc.,* February 1994 **[0079]**