(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 January 2014 (23.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/013070 A1

(51) International Patent Classification:
G10L 19/008 (2013.01)

(21) International Application Number:
PCT/EP2013/065343

(22) International Filing Date:
19 July 2013 (19.07.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12290239.8 19 July 2012 (19.07.2012) EP

(71) Applicant: **THOMSON LICENSING** [FR/FR]; 1-5 rue Jeanne d'Arc, F-92130 Issy-les-Moulineaux (FR).

(72) Inventors: **WUEBBOLT, Oliver**; Saarbrückener Str. 4, 30559 Hannover (DE). **BOEHM, Johannes**; Sieberweg 35, 37081 Göttingen (DE). **JAX, Peter**; St. Ingbert-Weg 13, 30559 Hannover (DE).

(74) Agent: **KÖNIG, Uwe**; Karl-Wiechert-Allee 74, 30625 Hannover (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: METHOD AND DEVICE FOR IMPROVING THE RENDERING OF MULTI-CHANNEL AUDIO SIGNALS

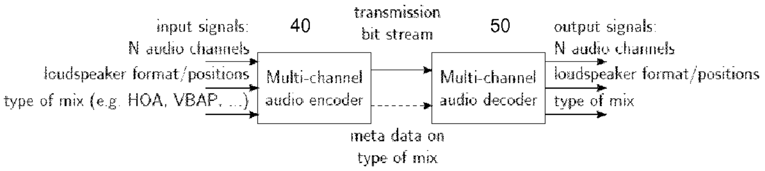
input signals:
N audio channels
loudspeaker format/positions
type of mix (e.g. HOA, VBAP, ...)
40
Multi-channel audio encoder
transmission bit stream
meta data on type of mix
50
Multi-channel audio decoder
output signals:
N audio channels
loudspeaker format/positions
type of mix

Fig.2

(57) **Abstract**: Conventional audio compression technologies perform a standardized signal transformation, independent of the type of the content. Multi-channel signals are decomposed into their signal components, subsequently quantized and encoded. This is disadvantageous due to lack of knowledge on the characteristics of scene composition, especially for e.g. multi-channel audio or Higher-Order Ambisonics (HOA) content. An improved method for encoding pre-processed audio data comprises encoding the pre-processed audio data, and encoding auxiliary data that indicate the particular audio pre- processing. An improved method for decoding encoded audio data comprises determining that the encoded audio data had been pre-processed before encoding, decoding the audio data, extracting from received data information about the pre- processing, and post-processing the decoded audio data according to the extracted pre- processing information.

Method and device for improving the rendering of multi-channel audio signals

Field of the invention

The invention is in the field of Audio Compression, in particular compression of multi-channel audio signals and sound-field-oriented audio scenes, e.g. Higher Order Ambisonics (HOA).

Background of the invention

At present, compression schemes for multi-channel audio signals do not explicitly take into account how the input audio material has been generated or mixed. Thus, known audio compression technologies are not aware of the origin/mixing type of the content they shall compress. In known approaches, a “blind” signal transformation is performed, by which the multi-channel signal is decomposed into its signal components that are subsequently quantized and encoded. A disadvantage of such approaches is that the computation of the above-mentioned signal decomposition is computationally demanding, and it is difficult and error prone to find the best suitable and most efficient signal decomposition for a given segment of the audio scene.

Summary of the invention

The present invention relates to a method and a device for improving multi-channel audio rendering.

It has been found that at least some of the above-mentioned disadvantages are due to the lack of prior knowledge on the characteristics of the scene composition. Especially for spatial audio content, e.g. multichannel-audio or *Higher-Order Ambisonics* (HOA) content, this prior information is useful in order to adapt the compression scheme. For instance, a common pre-processing step in compression algorithms is an audio scene analysis, which targets at extracting directional audio sources or audio objects from the original content or original content mix. Such directional audio sources or audio objects can be coded separately from the residual spatial audio content.

In one embodiment, a method for encoding pre-processed audio data comprises steps of encoding the pre-processed audio data, and encoding auxiliary data that indicate the particular audio pre-processing.

In one embodiment, the invention relates to a method for decoding encoded audio data, comprising steps of determining that the encoded audio data had been pre-processed before encoding, decoding the audio data, extracting from received data information about the pre-processing, and post-processing the decoded audio data according to the extracted pre-processing information. The step of determining that the encoded audio data had been pre-processed before encoding can be achieved by analysis of the audio data, or by analysis of accompanying metadata.

In one embodiment of the invention, an encoder for encoding pre-processed audio data comprises a first encoder for encoding the pre-processed audio data, and a second encoder for encoding auxiliary data that indicate the particular audio pre-processing.

In one embodiment of the invention, a decoder for decoding encoded audio data comprises an analyzer for determining that the encoded audio data had been pre-processed before encoding, a first decoder for decoding the audio data, a data stream parser unit or data stream extraction unit for extracting from received data information about the pre-processing, and a processing unit for post-processing the decoded audio data according to the extracted pre-processing information.

In one embodiment of the invention, a computer readable medium has stored thereon executable instructions to cause a computer to perform a method according to at least one of the above-described methods.

A general idea of the invention is based on at least one of the following extensions of multi-channel audio compression systems:

According to one embodiment, a multi-channel audio compression and/or rendering system has an interface that comprises the multi-channel audio signal stream (e.g. PCM streams), the related spatial positions of the channels or corresponding loudspeakers, and metadata indicating the type of mixing that had been applied to the multi-channel audio signal stream. The mixing type indicate for instance a (previous) use or configuration and/or any details of HOA or VBAP panning, specific recording techniques, or equivalent information. The interface can be an input interface towards a signal transmission chain. In the case of HOA content, the spatial positions of loudspeakers can be positions of virtual loudspeakers.

According to one embodiment, the bit stream of a multi-channel compression codec comprises signaling information in order to transmit the above-mentioned metadata about virtual or real loudspeaker positions and original mixing information to the decoder and subsequent rendering algorithms. Thereby, any applied rendering techniques on the decoding side can be adapted to the specific mixing characteristics on the encoding side of the particular transmitted content.

In one embodiment, the usage of the metadata is optional and can be switched on or off. I.e., the audio content can be decoded and rendered in a simple mode without using the metadata, but the decoding and/or rendering will be not optimized in the simple mode. In an enhanced mode, optimized decoding and/or rendering can be achieved by making use of the metadata. In this embodiment, the decoder/renderer can be switched between the two modes.

Brief description of the drawings

Advantageous exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in

Fig.1 the structure of a known multi-channel transmission system;
Fig.2 the structure of a multi-channel transmission system according to one embodiment of the invention;
Fig.3 a smart decoder according to one embodiment of the invention;
Fig.4 the structure of a multi-channel transmission system for HOA signals;
Fig.5 spatial sampling points of a DSHT;
Fig.6 examples of spherical sampling positions for a codebook used in encoder and decoder building blocks; and
Fig.7 an exemplary embodiment of a particularly improved multi-channel audio encoder.

Detailed description of the invention

Fig. 1 shows a known approach for multi-channel audio coding. Audio data from an audio production stage 10 are encoded in a multi-channel audio encoder 20, transmitted and decoded in a multi-channel audio decoder 30. Metadata may explicitly be transmitted (or their information may be included implicitly) and related to the spatial audio composition. Such conventional metadata are limited to information on the spatial positions of loudspeakers, e.g. in the form of specific formats (e.g. stereo or ITU-R BS.775-1 also known as “5.1 surround sound”) or by tables with loudspeaker positions. No information on *how* a specific spatial audio mix/recording has been produced is communicated to the multi-channel audio encoder 20, and thus such information cannot be exploited or utilized in compressing the signal within the multi-channel audio encoder 20.

However, it has been recognized that knowledge of at least one of origin and mixing type of the content is of particular importance if a multi-channel spatial audio coder processes at least one of content that has been derived from a Higher-Order Ambisonics (HOA) format, a recording with any fixed microphone setup and a multi-channel mix with any specific panning algorithms, because in these cases the specific mixing characteristics can be exploited by the compression scheme. Also original multi-channel audio content can benefit from additional mixing information indication. It is advantageous to indicate e.g. a used panning method such as e.g. *Vector-Based Amplitude Panning* (VBAP), or any details thereof, for improving the encoding efficiency. Advantageously, the signal models for the audio scene analysis, as well as the subsequent encoding steps, can be adapted according to this information. This results in a more efficient compression system with respect to both rate-distortion performance and computational effort.

In the particular case of HOA content, there is the problem that many different conventions exist, e.g. complex-valued vs. real-valued spherical harmonics, multiple/different normalization schemes, etc. In order to avoid incompatibilities between differently produced HOA content, it is useful to define a common format. This can be achieved via a transformation of the HOA time-domain coefficients to its equivalent spatial representation, which is a multi-channel representation, using a transform such as the *Discrete Spherical Harmonics Transform* (DSHT). The DSHT is created from a regular spherical distribution of spatial sampling positions, which can be regarded equivalent to virtual loudspeaker positions. More definitions and details about the DSHT are given below. Any system using another definition of HOA is able to derive its own HOA coefficients representation from this common format defined in the spatial domain. Compression of signals of said common format benefits considerably from the prior knowledge that the virtual loudspeaker signals represent an original HOA signal, as described in more detail below.

Furthermore, this mixing information etc. is also useful for the decoder or renderer. In one embodiment, the mixing information etc. is included in the bit stream. The used rendering algorithm can be adapted to the original mixing e.g. HOA or VBAP, to allow for a better down-mix or rendering to flexible loudspeaker positions.

Fig. 2 shows an extension of the multi-channel audio transmission system according to one embodiment of the invention. The extension is achieved by adding metadata that describe at least one of the type of mixing, type of recording, type of editing, type of synthesizing etc. that has been applied in the production stage 10 of the audio content. This information is carried through to the decoder output and can be used inside the multi-channel compression codec 40,50 in order to improve efficiency. The information on how a specific spatial audio mix/recording has been produced is communicated to the multi-channel audio encoder 40, and thus can be exploited or utilized in compressing the signal.

One example as to how this metadata information can be used is that, depending on the mixing type of the input material, different coding modes can be activated by the multi-channel codec. For instance, in one embodiment, a coding mode is switched to a HOA-specific encoding/decoding principle (HOA mode), as described below (with respect to eq.(3)-(16)) if HOA mixing is indicated at the encoder input, while a different (e.g. more traditional) multi-channel coding technology is used if the mixing type of the input signal is not HOA, or unknown. In the HOA mode, the encoding starts in one embodiment with a DSHT block in which a DSHT regains the original HOA coefficients, before a HOA-specific encoding process is started. In another embodiment, a different discrete transform other than DSHT is used for a comparable purpose.

Fig.3 shows a “smart” rendering system according to one embodiment of the invention, which makes use of the inventive metadata in order to accomplish a flexible down-mix, up-mix or re-mix of the decoded N channels to M loudspeakers that are present at the decoder terminal. The metadata on the type of mixing, recording etc. can be exploited for selecting one of a plurality of modes, so as to accomplish efficient, high-quality rendering. A multi-channel encoder 50 uses optimized encoding, according to metadata on the type of mix in the input audio data, and encodes/provides not only N encoded audio channels and information about loudspeaker positions, but also e.g. “type of mix” information to the decoder 60. The decoder 60 (at the receiving side) uses real loudspeaker positions of loudspeakers available at the receiving side, which are unknown at the transmitting side (i.e. encoder), for generating output signals for M audio channels. In one embodiment, N is different from M. In one embodiment, N equals M or is different from M, but the real loudspeaker positions at the receiving side are different from loudspeaker positions that were assumed in the encoder 50 and in the audio production 10. The encoder 50 or the audio production 10 may assume e.g. standardized loudspeaker positions.

Fig.4 shows how the invention can be used for efficient transmission of HOA content. The input HOA coefficients are transformed into the spatial domain via an inverse DSHT (iDSHT) 410. The resulting N audio channels, their (virtual) spatial positions, as well as an indication (e.g. a flag such as a “HOA mixed” flag) are provided to the multi-channel audio encoder 420, which is a compression encoder. The compression encoder can thus utilize the prior knowledge that its input signals are HOA-derived. An interface between the audio encoder 420 and an audio decoder 430 or audio renderer comprises N audio channels, their (virtual) spatial positions, and said indication. An inverse process is performed at the decoding side, i.e. the HOA representation can be recovered by applying, after decoding 430, a DSHT 440 that uses knowledge of the related operations that had been applied before encoding the content. This knowledge is received through the interface in form of the metadata according to the invention.

Some (but not necessarily all) kinds of metadata that are in particular within the scope of this invention would be, for example, at least one of the following:

- an indication that original content was derived from HOA content, plus at least one of:
    - o an order of the HOA representation
    - o indication of 2D, 3D or hemispherical representation; and
    - o positions of spatial sampling points (adaptive or fixed)
- an indication that original content was mixed synthetically using VBAP, plus an assignment of VBAP tupels (pairs) or triples of loudspeakers; and
- an indication that original content was recorded with fixed, discrete microphones, plus at least one of:
    - o one or more positions and directions of one or more microphones on the recording set; and
    - o one or more kinds of microphones, e.g. cardoid vs. omnidirectional vs. super-cardoid, etc.

Main advantages of the invention are at least the following.
A more efficient compression scheme is obtained through better prior knowledge on the signal characteristics of the input material. The encoder can exploit this prior knowledge for improved audio scene analysis (e.g. a source model of mixed content can be adapted). An example for a source model of mixed content is a case where a signal source has been modified, edited or synthesized in an audio production stage 10. Such audio production stage 10 is usually used to generate the multichannel audio signal, and it is usually located before the multi-channel audio encoder block 20. Such audio production stage 10 is also assumed (but not shown) in Fig.2 before the new encoding block 40. Conventionally, the editing information is lost and not passed to the encoder, and can therefore not be exploited. The present invention enables this information to be preserved. Examples of the audio production stage 10 comprise recording and mixing,

synthetic sound or multi-microphone information, e.g., multiple sound sources that are synthetically mapped to loudspeaker positions.

Another advantage of the invention is that the rendering of transmitted and decoded content can be considerably improved, in particular for ill-conditioned scenarios where a number of available loudspeakers is different from a number of available channels (so-called down-mix and up-mix scenarios), as well as for flexible loudspeaker positioning. The latter requires re-mapping according to the loudspeaker position(s).

Yet another advantage is that audio data in a sound field related format, such as HOA, can be transmitted in channel-based audio transmission systems without losing important data that are required for high-quality rendering.

The transmission of metadata according to the invention allows at the decoding side an optimized decoding and/or rendering, particularly when a spatial decomposition is performed. While a general spatial decomposition can be obtained by various means, e.g. a Karhunen-Loève Transform (KLT), an optimized decomposition (using metadata according to the invention) is less computationally expensive and, at the same time, provides a better quality of the multi-channel output signals (e.g. the single channels can easier be adapted or mapped to loudspeaker positions during the rendering, and the mapping is more exact). This is particularly advantageous if the number of channels is modified (increased or decreased) in a mixing (matrixing) stage during the rendering, or if one or more loudspeaker positions are modified (especially in cases where each channel of the multi-channels is adapted to a particular loudspeaker position).

In the following, the Higher Order Ambisonics (HOA) and the Discrete Spherical Harmonics Transform (DSHT) are described.

HOA signals can be transformed to the spatial domain, e.g. by a Discrete Spherical Harmonics Transform (DSHT), prior to compression with perceptual coders.
The transmission or storage of such multi-channel audio signal representations usually demands for appropriate multi-channel compression techniques. Usually, a channel independent perceptual decoding is performed before finally matrixing the I decoded signals $\hat{\hat{x}}_i(l)$, $i = 1, \dots, I$, into J new signals $\hat{\hat{y}}_j(l)$, $j = 1, \dots, J$. The term matrixing means adding or mixing the decoded signals $\hat{\hat{x}}_i(l)$ in a weighted manner. Arranging all signals $\hat{\hat{x}}_i(l)$, $i = 1, \dots, I$, as well as all new signals $\hat{\hat{y}}_j(l)$, $j = 1, \dots, J$ in vectors according to

$$\hat{\mathrm{x}}(l) := \left[\,\hat{\mathrm{x}}_1(l) \;\; \ldots \;\; \hat{\mathrm{x}}_I(l)\right]^{\mathrm{T}} \tag{1a}$$

$$\hat{\mathrm{y}}(l) := \left[\,\hat{\mathrm{y}}_1(l) \;\; \ldots \;\; \hat{\mathrm{y}}_J(l)\right]^{\mathrm{T}} \tag{1b}$$

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

$$\hat{\mathrm{y}}(l) = \boldsymbol{A}\,\hat{\mathrm{x}}(l) \tag{2}$$

where $\boldsymbol{A}$ 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.

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.

The following section gives a brief introduction to Higher Order Ambisonics (HOA) and defines the signals to be processed (data rate compression).

Higher Order Ambisonics (HOA) is based on the description of a sound field within a compact area of interest, which is assumed to be free of sound sources. In that case the spatiotemporal behavior of the sound pressure $p(t,\boldsymbol{x})$ at time t and position $\boldsymbol{x} = [r,\theta,\phi]^T$ within the area of interest (in spherical coordinates) is physically fully determined by the homogeneous wave equation. It can be shown that the Fourier transform of the sound pressure with respect to time, i.e.,

$$P(\,\omega,\boldsymbol{x}) = \mathcal{F}_t\,\{\;p(t,\boldsymbol{x})\;\} \tag{3}$$

where ω denotes the angular frequency (and $\mathcal{F}_t\,\{\}$ corresponds to $\int_{-\infty}^{\infty} p(t,\boldsymbol{x})\; e^{-\omega t} dt$), may be expanded into the series of Spherical Harmonics (SHs) according to:

$$P(\,k\,c_s,\boldsymbol{x}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^m(k)\;\; j_n(kr)\;\; Y_n^m(\theta,\phi) \tag{4}$$

In eq.(4), c_s denotes the speed of sound and $k = \frac{\omega}{c_s}$ the angular wave number. Further, $j_n(\cdot)$ indicate the spherical Bessel functions of the first kind and order n and $Y_n^m(\cdot)$ denote the Spherical Harmonics (SH) of order n and degree m. The complete information about the sound field is actually contained within the *sound field coefficients* $A_n^m(k)$.

It should be noted that the SHs are complex valued functions in general. However, by an appropriate linear combination of them, it is possible to obtain real valued functions and perform the expansion with respect to these functions.

Related to the pressure *sound field* description in eq.(4), a *source field* can be defined as:

$$D(\, k\, c_s, \mathbf{\Omega}) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_n^m(k)\ Y_n^m(\mathbf{\Omega}\,), \qquad (5)$$

with *the source field* or *amplitude density* [9] $D(\, k\, c_s, \mathbf{\Omega})$ depending on angular wave number and angular direction $\mathbf{\Omega} = [\theta, \phi]^T$. A source field can consist of far-field/ near-field, discrete/ continuous sources [1]. The source field coefficients B_n^m are related to the sound field coefficients A_n^m by [1]:

$$A_n^m = \begin{cases} 4\,\pi\, i^n\, B_n^m & \text{for the far field} \\ -i\, k\, h_n^{(2)}(kr_s)\, B_n^m & \text{for the near field} \end{cases} \qquad (6)$$

where $h_n^{(2)}$ is the spherical Hankel function of the second kind and r_s is the source distance from the origin. Concerning the near field, it is noted that positive frequencies and the spherical Hankel function of second kind $\mathrm{h}_\mathrm{n}^{(2)}$ are used for incoming waves (related to $\mathrm{e}^{-\mathrm{ikr}}$).

Signals in the HOA domain can be represented in frequency domain or in time domain as the inverse Fourier transform of the *source field* or *sound fie*ld coefficients. The following description will assume the use of a time domain representation of source *field coefficients:*

$$b_n^m = i\mathcal{F}_t\{\, B_n^m\, \} \qquad (7)$$

of a finite number: The infinite series in eq.(5) is truncated at $n = N$. Truncation corresponds to a spatial bandwidth limitation. The number of coefficients (or HOA channels) is given by:

$$\mathrm{O_{3D}} = (\mathrm{N} + 1)^2 \ \text{ for 3D} \qquad (8)$$

or by $O_{2D} = 2N + 1$ for 2D only descriptions. The coefficients b_n^m comprise the Audio information of one time sample m for later reproduction by loudspeakers. They can be stored or transmitted and are thus subject to data rate compression. A single time sample m of coefficients can be represented by vector $\boldsymbol{b}(m)$ with O_{3D} elements:

$$\boldsymbol{b}(m) := [b_0^0(m)\,, b_1^{-1}(m), b_1^0(m), b_1^1(m), b_2^{-2}(m),\ \ \ldots, b_N^N(m)]^T \qquad (9)$$

and a block of M time samples by matrix $\boldsymbol{B}$

$$\boldsymbol{B} := [\boldsymbol{b}\,(m_{\mathrm{START}} + 1), \boldsymbol{b}\,(m_{\mathrm{START}} + 2), .., \boldsymbol{b}\,(m_{\mathrm{START}} + M)] \qquad (10)$$

Two dimensional representations of sound fields can be derived by an expansion with circular harmonics. This is can be seen as a special case of the general description presented above using a fixed inclination of $\theta = \frac{\pi}{2}$, different weighting of coefficients and a reduced set to O_{2D} coefficients ($m = \pm n$). Thus all of the following considerations also apply to 2D representations, the term sphere then needs to be substituted by the term circle.

The following describes a transform from HOA coefficient domain to a spatial, channel based, domain and vice versa. Eq.(5) can be rewritten using time domain HOA coefficients for l discrete spatial sample positions $\mathbf{\Omega}_l = [\theta_l, \phi_l]^T$ on the unit sphere:

$$d_{\Omega_l} := \sum_{n=0}^{N} \sum_{m=-n}^{n} b_n^m \; Y_n^m(\mathbf{\Omega}_l), \tag{11}$$

Assuming $L_{sd} = (N+1)^2$ spherical sample positions $\mathbf{\Omega}_l$, this can be rewritten in vector notation for a HOA data block $\boldsymbol{B}$:

$$\boldsymbol{W} = \mathbf{\Psi}_{\mathrm{i}} \; \boldsymbol{B}\,, \tag{12}$$

with $\boldsymbol{W} := [\boldsymbol{w}\,(m_{\mathrm{START}}+1), \boldsymbol{w}\,(m_{\mathrm{START}}+2), .., \boldsymbol{w}\,(m_{\mathrm{START}}+M)]$ and $\boldsymbol{w}(m) = \left[d_{\Omega_1}(m), \ldots, d_{\Omega_{L_{sd}}}(m)\right]^{\mathrm{T}}$ representing a single time-sample of a L_{sd} multichannel signal, and matrix $\mathbf{\Psi}_{\mathrm{i}} = \left[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{L_{sd}}\right]^H$ with vectors $\boldsymbol{y}_l = [Y_0^0(\mathbf{\Omega}_l),\; Y_1^{-1}(\mathbf{\Omega}_l), \ldots, Y_N^N(\mathbf{\Omega}_l)]^T$. If the spherical sample positions are selected very regular, a matrix $\mathbf{\Psi}_{\mathrm{f}}$ exists with

$$\mathbf{\Psi}_{\mathrm{f}}\, \mathbf{\Psi}_{\mathrm{i}} = \boldsymbol{I}, \tag{13}$$

where $\boldsymbol{I}$ is a O_{3D}x O_{3D} identity matrix. Then the corresponding transformation to eq.(12) can be defined by:

$$\boldsymbol{B} = \mathbf{\Psi}_{\mathrm{f}} \; \boldsymbol{W}. \tag{14}$$

Eq.(14) transforms L_{sd} spherical signals into the *coefficient domain* and can be rewritten as a forward transform:

$$\boldsymbol{B} = DSHT\{\boldsymbol{W}\}, \tag{15}$$

where $DSHT\{\quad\}$ denotes the *Discrete Spherical Harmonics Transform*. The corresponding inverse transform, transforms O_{3D} coefficient signals into the *spatial domain* to form L_{sd} channel based signals and eq.(12) becomes:

$$\boldsymbol{W} = iDSHT\{\boldsymbol{B}\}. \tag{16}$$

The DSHT with a number of spherical positions L_{Sd} matching the number of HOA coefficients $\mathrm{O}_{3\mathrm{D}}$ (see eq.(8)) is described below. First, a default spherical sample grid is selected. For a block of M time samples, the spherical sample grid is rotated such that the logarithm of the term

$$\textstyle\sum_{l=1}^{L_{Sd}} \sum_{j=1}^{L_{Sd}} |\, \Sigma_{W_{Sd}\,_{l,j}}| - \sum\left(\sigma_{S_{d_1}}^2, \ldots, \sigma_{S_{d_{\mathrm{L_{Sd}}}}}^2\right) \tag{17}$$

is minimized, where $|\Sigma_{W_{Sd}}{}_{l,j}|$ are the absolute values of the elements of $\boldsymbol{\Sigma}_{W_{Sd}}$ (with matrix row index l and column index j) and $\sigma^2_{S_{d_l}}$ are the diagonal elements of $\boldsymbol{\Sigma}_{W_{Sd}}$. Visualized, this corresponds to the spherical sampling grid of the DSHT as shown in Fig.5.

Suitable spherical sample positions for the DSHT and procedures to derive such positions are well-known. Examples of sampling grids are shown in Fig.6. In particular, Fig.6 shows examples of spherical sampling positions for a codebook used in encoder and decoder building blocks pE, pD, namely in Fig.6 a) for $\boldsymbol{L_{Sd}}$ =4 , in Fig.6 b) for $\boldsymbol{L_{Sd}}$ =9, in Fig.6 c) for $\boldsymbol{L_{Sd}}$ =16 and in Fig.6 d) for $\boldsymbol{L_{Sd}} = 25$. Such codebooks can, *inter alia*, be used for rendering according to pre-defined spatial loudspeaker configurations.

Fig.7 shows an exemplary embodiment of a particularly improved multi-channel audio encoder 420 shown in Fig.4. It comprises a DSHT block 421, which calculates a DSHT that is inverse to the Inverse DSHT of block 410 (in order to reverse the block 410). The purpose of block 421 is to provide at its output 70 signals that are substantially identical to the input of the Inverse DSHT block 410. The processing of this signal 70 can then be further optimized. The signal 70 comprises not only audio components that are provided to an MDCT block 422, but also signal portions 71 that indicate one or more dominant audio signal components, or rather one or more locations of dominant audio signal components. These are then used for detecting 424 at least one strongest source direction and calculating 425 rotation parameters for an adaptive rotation of the iDSHT. In one embodiment, this is time variant, i.e. the detecting 424 and calculating 425 is continuously re-adapted at defined discrete time steps. The adaptive rotation matrix for the iDSHT is calculated and the adaptive iDSHT is performed in the iDSHT block 423. The effect of the rotation is that the sampling grid of the iDSHT 423 is rotated such that one of the sides (i.e. a single spatial sample position) matches the strongest source direction (this may be time variant). This provides a more efficient and therefore better encoding of the audio signal in the iDSHT block 423. The MDCT block 422 is advantageous for compensating the temporal overlapping of audio frame segments. The iDSHT block 423 provides an encoded audio signal 74, and the rotation parameter calculating block 425 provides rotation parameters as (at least a part of) pre-processing information 75. Additionally, the pre-processing information 75 may comprise other information.

Further, the present invention relates to the following embodiments.

In one embodiment, the invention relates to a method for transmitting and/or storing and processing a channel based 3D-audio representation, comprising steps of sending/storing side information (SI) along the channel based audio information, the side information indicating the mixing type and intended speaker position of the channel based audio information, where the mixing type indicates an algorithm according to which the audio content was mixed (e.g.in the mixing studio) in a previous processing stage, where the speaker positions indicate the positions of the speakers (ideal positions e.g. in the mixing studio) or the virtual positions of the previous processing stage. Further processing steps, after receiving said data structure and channel based audio information, utilize the mixing & speaker position information.

In one embodiment, the invention relates to a device for transmitting and/or storing and processing a channel based 3D-audio representation, comprising means for sending (or means for storing) side information (SI) along the channel based Audio information, the side information indicating the mixing type and intended speaker position of the channel based audio information, where the mixing type signals the algorithm according to which the audio content was mixed (e.g.in the mixing studio) in a previous processing stage, where the speaker positions indicate the positions of the speakers (ideal positions e.g. in the mixing studio) or the virtual positions of the previous processing stage. Further, the device comprises a processor that utilizes the mixing & speaker position information after receiving said data structure and channel based audio information.

In one embodiment, the present invention relates to a 3D audio system where the mixing information signals HOA content, the HOA order and virtual speaker position information that relates to an ideal spherical sampling grid that has been used to convert HOA 3D audio to the channel based representation before. After receiving/reading transmitted channel based audio information and accompanying side information (SI), the SI is used to re-encode the channel based audio to HOA format. Said re-encoding is done by calculating a mode-matrix Ψ from said spherical sampling positions and matrix multiplying it with the channel based content (DSHT).

In one embodiment, the system/method is used for circumventing ambiguities of different HOA formats. The HOA 3D audio content in a 1st HOA format at the production side is converted to a related channel based 3D audio representation using the iDSHT related to the 1st format and distributed in the SI. The received channel based audio information is converted to a 2nd HOA format using SI and a DSHT related to the 2nd format. In one embodiment of the system, the 1st HOA format uses a HOA representation with complex values and the 2nd HOA format uses a HOA representation with real values. In one embodiment of the system, the 2nd HOA format uses a complex HOA representation and the 1st HOA format uses a HOA representation with real values.

In one embodiment, the present invention relates to a 3D audio system, wherein the mixing information is used to separate directional 3D audio components (audio object extraction) from the signal used within rate compression, signal enhancement or rendering. In one embodiment, further steps are signaling HOA, the HOA order and the related ideal spherical sampling grid that has been used to convert HOA 3D audio to the channel based representation before, restoring the HOA representation and extracting the directional components by determining main signal directions by use of block based covariance methods. Said directions are used for HOA decoding the directional signals to these directions. In one embodiment, the further steps are signaling Vector Base Amplitude Panning (VBAP) and related speaker position information, where the speaker position information is used to determine the speaker triplets and a covariance method is used to extract a correlated signal out of said triplet channels.
In one embodiment of the 3D audio system, residual signals are generated from the directional signals and the restored signals related to the signal extraction (HOA signals, VBAP triplets (pairs)).

In one embodiment, the present invention relates to a system to perform data rate compression of the residual signals by steps of reducing the order of the HOA residual signal and compressing reduced order signals and directional signals, mixing the residual triplet channels to a mono stream and providing related correlation information, and transmitting said information and the compressed mono signals together with compressed directional signals.

In one embodiment of the system to perform data rate compression, it is used for rendering audio to loudspeakers, wherein the extracted directional signals are panned to loudspeakers using the main signal directions and the de-correlated residual signals in the channel domain.

The invention allows generally a signalization of audio content mixing characteristics. The invention can be used in audio devices, particularly in audio encoding devices, audio mixing devices and audio decoding devices.

It should be noted that although shown simply as a DSHT, other types of transformation may be constructed or applied other than a DSHT, 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. Further, although the HOA format is exemplarily mentioned in the above description, the invention can also be used with other types of soundfield related formats other than Ambisonics, 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.

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 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. 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.

References

[1] T.D. Abhayapala “Generalized framework for spherical microphone arrays: Spatial and frequency decomposition”, In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (accepted) Vol. X, pp. , April 2008, Las Vegas, USA.

[2] James R. Driscoll and Dennis M. Healy Jr.: “Computing Fourier transforms and convolutions on the 2-sphere”, Advances in Applied Mathematics, 15:202–250, 1994

Claims

1. Method for encoding pre-processed audio data, comprising steps of
   - encoding the audio data;
   - encoding auxiliary data that indicate the particular audio pre-processing of the audio data.

2. Method according to claim 1, wherein the audio data are in HOA format.

3. Method according to claim 1 or 2, wherein the encoding comprises using an adaptive Inverse DSHT (423).

4. Method according to one of the claims 1-3, wherein the auxiliary data indicate that the audio content was derived from HOA content, plus at least one of: an order of the HOA content representation, a 2D, 3D or hemispherical representation, and positions of spatial sampling points.

5. Method according to one of the claims 1-4, wherein the auxiliary data indicate that the audio content was mixed synthetically using VBAP, plus an assignment of VBAP tupels or triples of loudspeakers.

6. Method according to one of the claims 1-5 wherein the auxiliary data indicate that the audio content was recorded with fixed, discrete microphones, plus at least one of: one or more positions and directions of one or more microphones on the recording set, and one or more kinds of microphones.

7. Method for decoding encoded audio data, comprising steps of
   - determining that the encoded audio data has been pre-processed before encoding;
   - decoding the audio data;
   - extracting from received data information about the pre-processing; and
   - post-processing the decoded audio data according to the extracted pre-processing information.

8. Method according to claim 7, wherein the information about the pre-processing indicates that the audio content was derived from HOA content, plus at least one of

an order of the HOA content representation, a 2D, 3D or hemispherical representation, and positions of spatial sampling points.

9. Method according to one of the claims 1-8, wherein the information about the pre-processing indicates that the audio content was mixed synthetically using VBAP, plus an assignment of VBAP tupels or triples of loudspeakers.

10. Method according to one of the claims 1-9 wherein the information about the pre-processing indicates that the audio content was recorded with fixed, discrete microphones, plus at least one of: one or more positions and directions of one or more microphones on the recording set, and one or more kinds of microphones.

11. Encoder for encoding pre-processed audio data, comprising
    - first encoder for encoding the audio data;
    - second encoder for encoding auxiliary data that indicate the particular audio pre-processing.

12. Encoder according to claim 11, where the encoder comprises an adaptive Inverse DSHT block.

13. Decoder for decoding encoded audio data, comprising
    - analyzer for determining that the encoded audio data has been pre-processed before encoding;
    - first decoder for decoding the audio data;
    - data stream parser/extraction unit for extracting from received data information about the pre-processing; and
    - processing unit for post-processing the decoded audio data according to the extracted pre-processing information.

14. Decoder according to claim 13, wherein the information about the pre-processing comprises indication of a microphone setup or of a panning algorithm that has been used for mixing the audio data.

15. Audio renderer suitable for rendering HOA signals, the audio renderer including an interface that comprises a plurality of input channels for receiving multi-channel audio data and spatial position information for the input channels, and at least one channel

for receiving metadata, the metadata specifying a type of audio mixing that has been applied to the multi-channel audio data.

16. Audio renderer according to claim 15, wherein the metadata specify a microphone setup or of a panning algorithm that has been used for mixing the audio data.

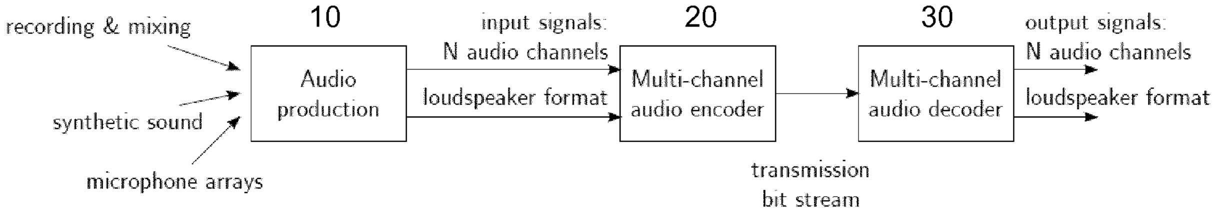

Fig.1

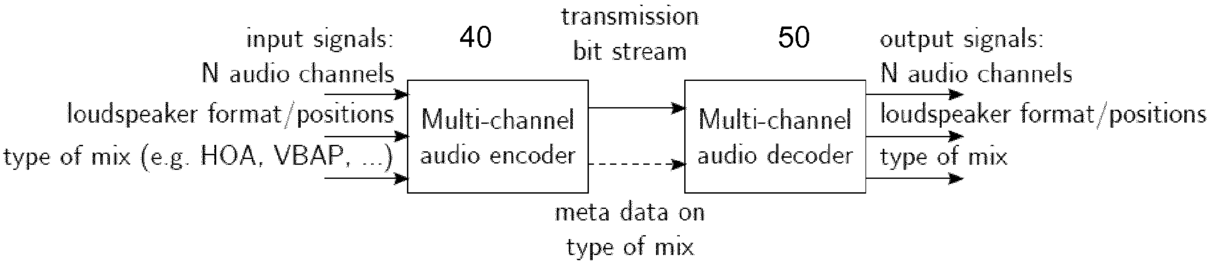
input signals:
N audio channels
loudspeaker format/positions
type of mix (e.g. HOA, VBAP, ...)
40
Multi-channel audio encoder
transmission bit stream
meta data on type of mix
50
Multi-channel audio decoder
output signals:
N audio channels
loudspeaker format/positions
type of mix

Fig.2

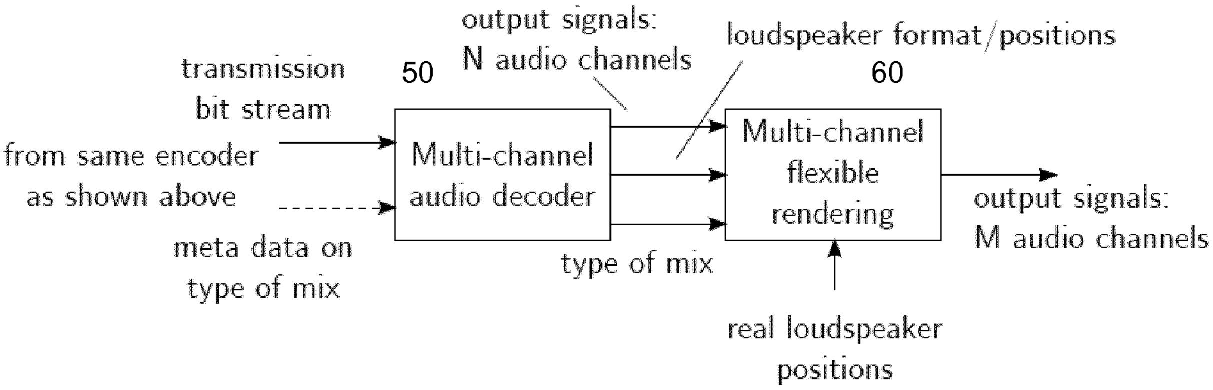
transmission bit stream
from same encoder as shown above
meta data on type of mix
50
Multi-channel audio decoder
output signals:
N audio channels
loudspeaker format/positions
type of mix
60
Multi-channel flexible rendering
real loudspeaker positions
output signals:
M audio channels

Fig.3

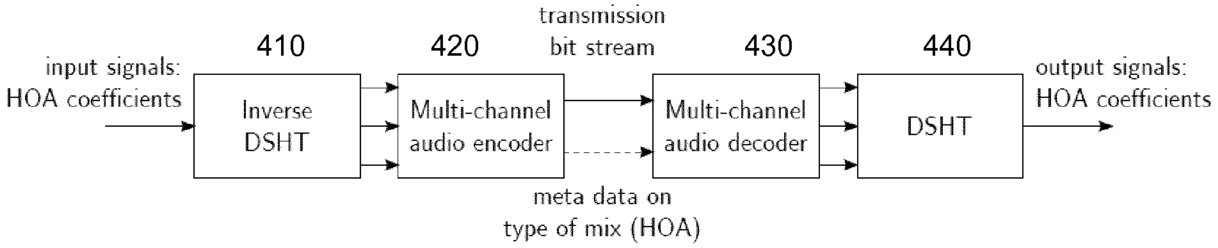
410
420
430
440
transmission
bit stream
input signals:
HOA coefficients
Inverse
DSHT
Multi-channel
audio encoder
Multi-channel
audio decoder
DSHT
output signals:
HOA coefficients
meta data on
type of mix (HOA)

Fig.4

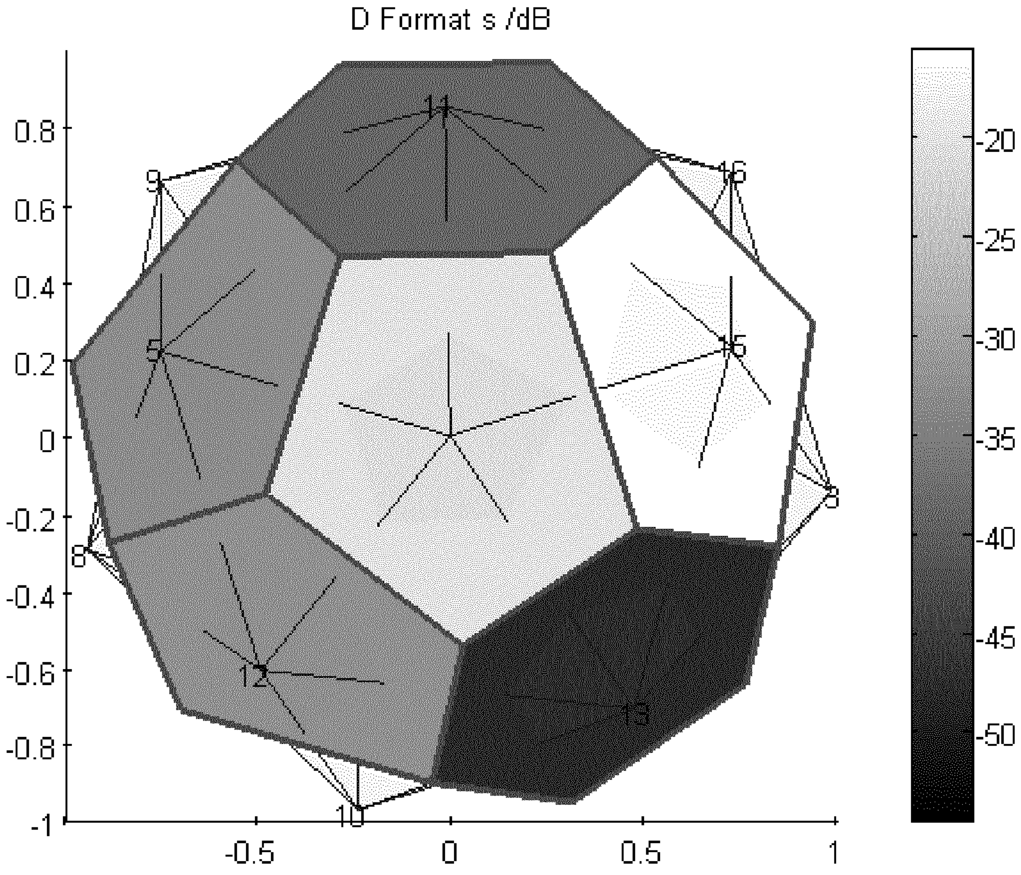
D Format s /dB
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-0.5
0
0.5
1
14
9
16
5
15
3
8
12
13
10
-20
-25
-30
-35
-40
-45
-50

Fig.5

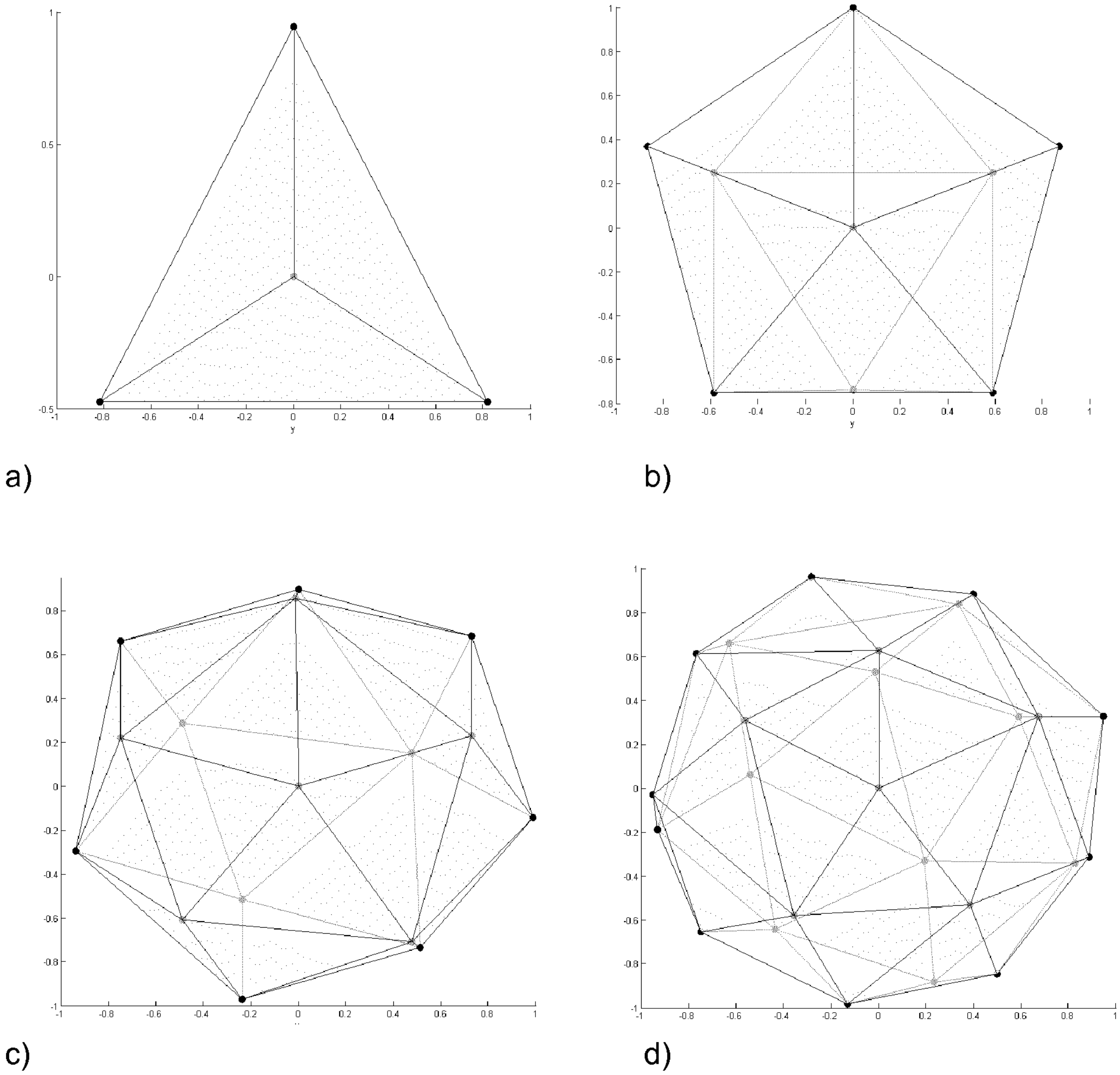

Fig.6

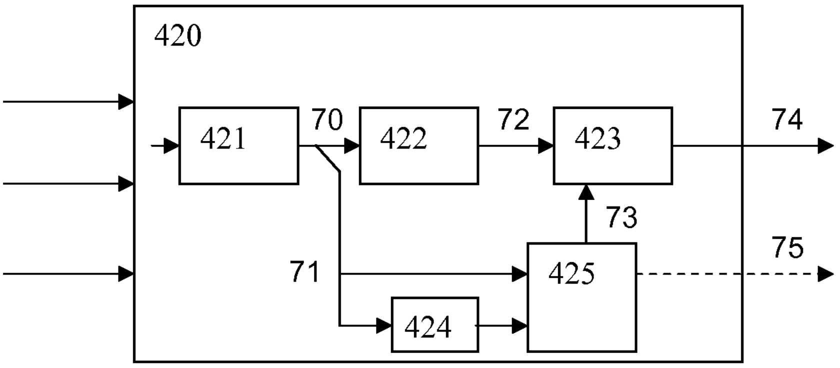

Fig.7

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2013/065343

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L19/008

ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/057715 A1 (JOHNSTON JAMES D [US] ET AL) 8 March 2012 (2012-03-08) paragraphs [0048] - [0051] paragraphs [0056] - [0058] paragraphs [0061] - [0062] paragraphs [0109] - [0130] ----- -/--	1-16

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 6 September 2013	Date of mailing of the international search report 17/09/2013
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Ramos Sánchez, U

1

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	OSAMU SHIMADA ET AL: "A core experiment proposal for an additional SAOC functionality of separating real-environment signals into multiple objects", 83. MPEG MEETING; 14-1-2008 - 18-1-2008; ANTALYA; (MOTION PICTURE EXPERT GROUP OR ISO/IEC JTC1/SC29/WG11),, no. M15110, 9 January 2008 (2008-01-09), XP030043707, section 2.B. Object reconstruction by a SAOC decoder with new functionality; figures 3,4 -----	1-16
X	US 2004/049379 A1 (THUMPUDI NAVEEN [US] ET AL) 11 March 2004 (2004-03-11) paragraphs [0139], [0145]; figure 6 paragraphs [0162], [0171]; figure 7 paragraph [0180] paragraphs [0376] - [0378], [0385] -----	1-16
X	BIN CHENG ET AL: "Encoding Independent Sources in Spatially Squeezed Surround Audio Coding", 11 December 2007 (2007-12-11), ADVANCES IN MULTIMEDIA INFORMATION PROCESSING Â PCM 2007; [LECTURE NOTES IN COMPUTER SCIENCE], SPRINGER BERLIN HEIDELBERG, BERLIN, HEIDELBERG, PAGE(S) 804 - 813, XP019085579, ISBN: 978-3-540-77254-5 page 805, paragraph 2 page 808, last paragraph - page 810, paragraph 1 -----	1-16

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2013/065343

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012057715 A1	08-03-2012	CN 103270508 A	28-08-2013
		EP 2614445 A1	17-07-2013
		US 2012057715 A1	08-03-2012
		US 2012082319 A1	05-04-2012
		WO 2012033950 A1	15-03-2012
US 2004049379 A1	11-03-2004	AT 418137 T	15-01-2009
		EP 1403854 A2	31-03-2004
		EP 2028648 A2	25-02-2009
		ES 2316678 T3	16-04-2009
		JP 4676139 B2	27-04-2011
		JP 5097242 B2	12-12-2012
		JP 2004264810 A	24-09-2004
		JP 2010217900 A	30-09-2010
		US 2004049379 A1	11-03-2004
		US 2008221908 A1	11-09-2008
		US 2011054916 A1	03-03-2011
		US 2011060597 A1	10-03-2011
		US 2012082316 A1	05-04-2012
		US 2012087504 A1	12-04-2012
		US 2013144630 A1	06-06-2013