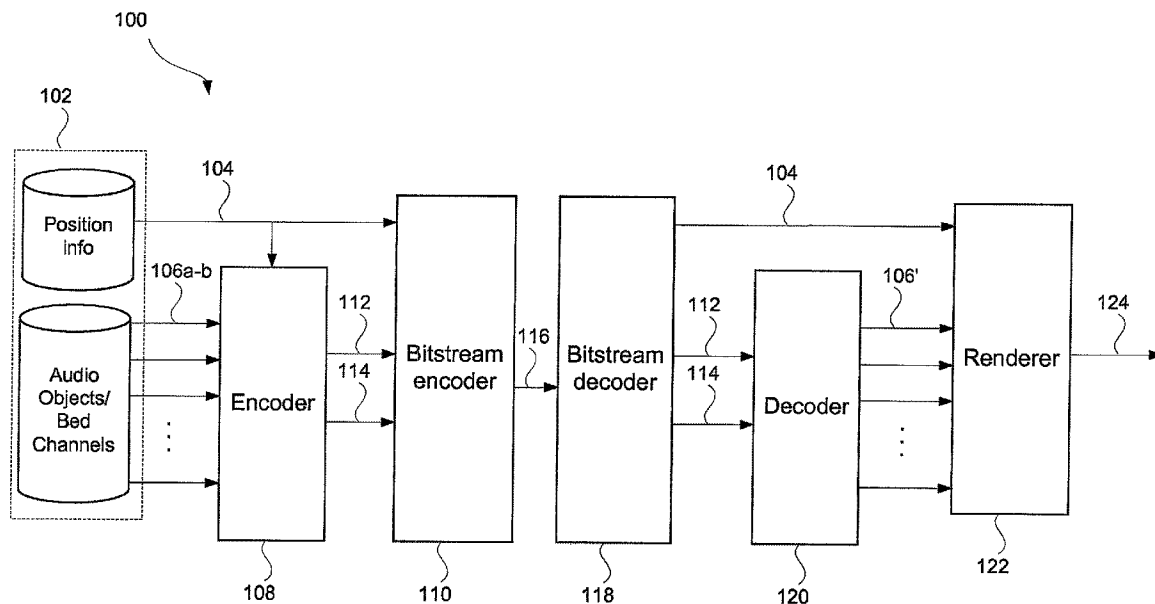




(22) Date de dépôt/Filing Date: 2014/05/23
(41) Mise à la disp. pub./Open to Public Insp.: 2014/11/27
(45) Date de délivrance/Issue Date: 2024/01/02
(62) Demande originale/Original Application: 3 017 077
(30) Priorité/Priority: 2013/05/24 (US61/827246)

(51) Cl.Int./Int.Cl. *G10L 19/008* (2013.01),
G10L 19/032 (2013.01)
(72) Inventeurs/Inventors:
PURNHAGEN, HEIKO, SE;
VILLEMOES, LARS, SE;
SAMUELSSON, LEIF JONAS, SE;
HIRVONEN, TONI, SE
(73) Propriétaire/Owner:
DOLBY INTERNATIONAL AB, NL
(74) Agent: SMART & BIGGAR LP

(54) Titre : CODAGE DE SCENES AUDIO
(54) Title: CODING OF AUDIO SCENES



(57) Abrégé/Abstract:

Exemplary embodiments provide decoding methods, and associated decoders, for decoding of an audio scene which at least comprises receiving a bit stream comprising information for determining M downmix signals and a reconstruction matrix; generating the reconstruction matrix; and reconstructing N audio objects from the M downmix signals using the reconstruction matrix, wherein the reconstructing takes place in a frequency domain, wherein matrix elements of the reconstruction matrix are applied as coefficients in linear combinations to at least M downmix signals, and wherein the matrix elements are based on the N audio objects.

ABSTRACT

Exemplary embodiments provide decoding methods, and associated decoders, for decoding of an audio scene which at least comprises receiving a bit stream comprising information for determining M downmix signals and a

- 5 reconstruction matrix; generating the reconstruction matrix; and reconstructing N audio objects from the M downmix signals using the reconstruction matrix, wherein the reconstructing takes place in a frequency domain, wherein matrix elements of the reconstruction matrix are applied as coefficients in linear combinations to at least M downmix signals, and wherein the matrix elements are based on the N audio objects.

CODING OF AUDIO SCENES

This application is a divisional of Canadian Patent Application No. 3,017,077, which is a divisional of Canadian Patent Application No. 2,910,755, filed May 23, 2014.

Technical field

- 5 The invention disclosed herein generally relates to the field of encoding and decoding of audio. In particular it relates to encoding and decoding of an audio scene comprising audio objects.

Background

- 10 There exist audio coding systems for parametric spatial audio coding. For example, MPEG Surround describes a system for parametric spatial coding of multichannel audio. MPEG SAOC (Spatial Audio Object Coding) describes a system for parametric coding of audio objects.

- 15 On an encoder side these systems typically downmix the channels/objects into a downmix, which typically is a mono (one channel) or a stereo (two channels) downmix, and extract side information describing the properties of the channels/objects by means of parameters like level differences and cross-correlation. The downmix and the side information are then encoded and sent to a decoder side. At the decoder side, the channels/objects are reconstructed, i.e. approximated, from the downmix under control of the parameters of the side
20 information.

- 25 A drawback of these systems is that the reconstruction is typically mathematically complex and often has to rely on assumptions about properties of the audio content that is not explicitly described by the parameters sent as side information. Such assumptions may for example be that the channels/objects are considered to be uncorrelated unless a cross-correlation parameter is sent, or that the downmix of the channels/objects is generated in a specific way. Further, the mathematical complexity and the need for additional assumptions increase dramatically as the number of channels of the downmix increases.

Furthermore, the required assumptions are inherently reflected in algorithmic details of the processing applied on the decoder side. This implies that quite a lot of intelligence has to be included on the decoder side. This is a drawback in that it may be difficult to upgrade or modify the algorithms once the decoders are deployed in
5 e.g. consumer devices that are difficult or even impossible to upgrade.

Brief description of the drawings

In what follows, example embodiments will be described in greater detail and with reference to the accompanying drawings, on which:

10 Fig. 1 is a schematic drawing of an audio encoding/decoding system according to example embodiments;

Fig. 2 is a schematic drawing of an audio encoding/decoding system having a legacy decoder according to example embodiments;

15 Fig. 3 is a schematic drawing of an encoding side of an audio encoding/decoding system according to example embodiments;

Fig. 4 is a flow chart of an encoding method according to example embodiments;

Fig. 5 is a schematic drawing of an encoder according to example embodiments;

20 Fig. 6 is a schematic drawing of a decoder side of an audio encoding/decoding system according to example embodiments;

Fig. 7 is a flow chart of a decoding method according to example embodiments;

25 Fig. 8 is a schematic drawing of a decoder side of an audio encoding/decoding system according to example embodiments; and

Fig. 9 is a schematic drawing of time/frequency transformations carried out on a decoder side of an audio encoding/decoding system according to example embodiments.

30 All the figures are schematic and generally only show parts which are necessary in order to elucidate the invention, whereas other parts may be omitted or merely suggested. Unless otherwise indicated, like reference numerals refer to like parts in different figures.

Detailed description

In view of the above it is an object to provide an encoder and a decoder and
5 associated methods which provide less complex and more flexible reconstruction of
audio objects.

I. Overview - Encoder

10 According to a first aspect, example embodiments propose encoding
methods, encoders, and computer program products for encoding. The proposed
methods, encoders and computer program products may generally have the same
features and advantages.

According to example embodiments there is provided a method for encoding a
15 time/frequency tile of an audio scene which at least comprises N audio objects. The
method comprises: receiving the N audio objects; generating M downmix signals
based on at least the N audio objects; generating a reconstruction matrix with matrix
elements that enables reconstruction of at least the N audio objects from the M
downmix signals; and generating a bit stream comprising the M downmix signals and
20 at least some of the matrix elements of the reconstruction matrix.

The number N of audio objects may be equal to or greater than one. The
number M of downmix signals may be equal to or greater than one.

With this method a bit stream is thus generated which comprises M downmix
signals and at least some of the matrix elements of a reconstruction matrix as side
25 information. By including individual matrix elements of the reconstruction matrix in
the bit stream, very little intelligence is required on the decoder side. For example,
there is no need on the decoder side for complex computation of the reconstruction
matrix based on the transmitted object parameters and additional assumptions.
Thus, the mathematical complexity at the decoder side is significantly reduced.
30 Moreover, the flexibility concerning the number of downmix signals is increased
compared to prior art methods since the complexity of the method is not dependent
on the number of downmix signals used.

As used herein *audio scene* generally refers to a three-dimensional audio environment which comprises audio elements being associated with positions in a three-dimensional space that can be rendered for playback on an audio system.

As used herein *audio object* refers to an element of an audio scene. An audio
5 object typically comprises an audio signal and additional information such as the position of the object in a three-dimensional space. The additional information is typically used to optimally render the audio object on a given playback system.

As used herein a *downmix signal* refers to a signal which is a combination of at least the N audio objects. Other signals of the audio scene, such as bed channels
10 (to be described below), may also be combined into the downmix signal. For example, the M downmix signals may correspond to a rendering of the audio scene to a given loudspeaker configuration, e.g. a standard 5.1 configuration. The number of downmix signals, here denoted by M, is typically (but not necessarily) less than the sum of the number of audio objects and bed channels, explaining why the M
15 downmix signals are referred to as a downmix.

Audio encoding/decoding systems typically divide the time-frequency space into time/frequency tiles, e.g. by applying suitable filter banks to the input audio signals. By a time/frequency tile is generally meant a portion of the time-frequency space corresponding to a time interval and a frequency sub-band. The time interval
20 may typically correspond to the duration of a time frame used in the audio encoding/decoding system. The frequency sub-band may typically correspond to one or several neighboring frequency sub-bands defined by the filter bank used in the encoding/decoding system. In the case the frequency sub-band corresponds to several neighboring frequency sub-bands defined by the filter bank, this allows for
25 having non-uniform frequency sub-bands in the decoding process of the audio signal, for example wider frequency sub-bands for higher frequencies of the audio signal. In a broadband case, where the audio encoding/decoding system operates on the whole frequency range, the frequency sub-band of the time/frequency tile may correspond to the whole frequency range. The above method discloses the encoding
30 steps for encoding an audio scene during one such time/frequency tile. However, it is to be understood that the method may be repeated for each time/frequency tile of the audio encoding/decoding system. Also it is to be understood that several time/frequency tiles may be encoded simultaneously. Typically, neighboring time/frequency tiles may overlap a bit in time and/or frequency. For example, an

overlap in time may be equivalent to a linear interpolation of the elements of the reconstruction matrix in time, i.e. from one time interval to the next. However, this disclosure targets other parts of encoding/decoding system and any overlap in time and/or frequency between neighboring time/frequency tiles is left for the skilled
5 person to implement.

According to exemplary embodiments the M downmix signals are arranged in a first field of the bit stream using a first format, and the matrix elements are arranged in a second field of the bit stream using a second format, thereby allowing a decoder that only supports the first format to decode and playback the M downmix
10 signals in the first field and to discard the matrix elements in the second field. This is advantageous in that the M downmix signals in the bit stream are backwards compatible with legacy decoders that do not implement audio object reconstruction. In other words, legacy decoders may still decode and playback the M downmix signals of the bitstream, for example by mapping each downmix signal to a channel
15 output of the decoder.

According to exemplary embodiments, the method may further comprise the step of receiving positional data corresponding to each of the N audio objects, wherein the M downmix signals are generated based on the positional data. The positional data typically associates each audio object with a position in a three-
20 dimensional space. The position of the audio object may vary with time. By using the positional data when downmixing the audio objects, the audio objects will be mixed in the M downmix signals in such a way that if the M downmix signals for example are listened to on a system with M output channels, the audio objects will sound as if they were approximately placed at their respective positions. This is for example
25 advantageous if the M downmix signals are to be backwards compatible with a legacy decoder.

According to exemplary embodiments, the matrix elements of the reconstruction matrix are time and frequency variant. In other words, the matrix elements of the reconstruction matrix may be different for different time/frequency
30 tiles. In this way a great flexibility in the reconstruction of the audio objects is achieved.

According to exemplary embodiments the audio scene further comprises a plurality of bed channels. This is for example common in cinema audio applications where the audio content comprises bed channels in addition to audio objects. In

such cases the M downmix signals may be generated based on at least the N audio objects and the plurality of bed channels. By a bed channel is generally meant an audio signal which corresponds to a fixed position in the three-dimensional space. For example, a bed channel may correspond to one of the output channels of the audio encoding/decoding system. As such, a bed channel may be interpreted as an audio object having an associated position in a three-dimensional space being equal to the position of one of the output speakers of the audio encoding/decoding system. A bed channel may therefore be associated with a label which merely indicates the position of the corresponding output speaker.

10 When the audio scene comprises bed channels, the reconstruction matrix may comprise matrix elements which enable reconstruction of the bed channels from the M downmix signals.

15 In some situations, the audio scene may comprise a vast number of objects. In order to reduce the complexity and the amount of data required to represent the audio scene, the audio scene may be simplified by reducing the number of audio objects. Thus, if the audio scene originally comprises K audio objects, wherein $K > N$, the method may further comprise the steps of receiving the K audio objects, and reducing the K audio objects into the N audio objects by clustering the K objects into N clusters and representing each cluster by one audio object.

20 In order to simplify the scene the method may further comprise the step of receiving positional data corresponding to each of the K audio objects, wherein the clustering of the K objects into N clusters is based on a positional distance between the K objects as given by the positional data of the K audio objects. For example, audio objects which are close to each other in terms of position in the three-
25 dimensional space may be clustered together.

30 As discussed above, exemplary embodiments of the method are flexible with respect to the number of downmix signals used. In particular, the method may advantageously be used when there are more than two downmix signals, i.e. when M is larger than two. For example, five or seven downmix signals corresponding to conventional 5.1 or 7.1 audio setups may be used. This is advantageous since, in contrast to prior art systems, the mathematical complexity of the proposed coding principles remains the same regardless of the number of downmix signals used.

 In order to further enable improved reconstruction of the N audio objects, the method may further comprise: forming L auxiliary signals from the N audio objects;

including matrix elements in the reconstruction matrix that enable reconstruction of at least the N audio objects from the M downmix signals and the L auxiliary signals; and including the L auxiliary signals in the bit stream. The auxiliary signals thus serves as help signals that for example may capture aspects of the audio objects
5 that is difficult to reconstruct from the downmix signals. The auxiliary signals may further be based on the bed channels. The number of auxiliary signals may be equal to or greater than one.

According to one exemplary embodiment, the auxiliary signals may correspond to particularly important audio objects, such as an audio object
10 representing dialogue. Thus at least one of the L auxiliary signals may be equal to one of the N audio objects. This allows the important objects to be rendered at higher quality than if they would have to be reconstructed from the M downmix channels only. In practice, some of the audio objects may have been prioritized and/or labeled by a audio content creator as the audio objects that preferably are
15 individually included as auxiliary objects. Furthermore, this makes modification/ processing of these objects prior to rendering less prone to artifacts. As a compromise between bit rate and quality, it is also possible to send a mix of two or more audio objects as an auxiliary signal. In other words, at least one of the L auxiliary signals may be formed as a combination of at least two of the N audio
20 objects.

According to one exemplary embodiment, the auxiliary signals represent signal dimensions of the audio objects that got lost in the process of generating the M downmix signals, e.g. since the number of independent objects typically is higher than the number of downmix channels or since two objects are associated with such
25 positions that they are mixed in the same downmix signal. An example of the latter case is a situation where two objects are only vertically separated but share the same position when projected on the horizontal plane, which means that they typically will be rendered to the same downmix channel(s) of a standard 5.1 surround loudspeaker set-up, where all speakers are in the same horizontal plane.
30 Specifically, the M downmix signals span a hyperplane in a signal space. By forming linear combinations of the M downmix signals only audio signals that lie in the hyperplane may be reconstructed. In order to improve the reconstruction, auxiliary signals may be included that do not lie in the hyperplane, thereby also allowing reconstruction of signals that do not lie in the hyperplane. In other words, according

to exemplary embodiments, at least one of the plurality of auxiliary signals does not lie in the hyperplane spanned by the M downmix signals. For example, at least one of the plurality of auxiliary signals may be orthogonal to the hyperplane spanned by the M downmix signals.

- 5 According to example embodiments there is provided a computer-readable medium comprising computer code instructions adapted to carry out any method of the first aspect when executed on a device having processing capability.

- According to example embodiments there is provided an encoder for encoding a time/frequency tile of an audio scene which at least comprises N audio objects,
- 10 comprising: a receiving component configured to receive the N audio objects; a downmix generating component configured to receive the N audio objects from the receiving component and to generate M downmix signals based on at least the N audio objects; an analyzing component configured to generate a reconstruction matrix with matrix elements that enables reconstruction of at least the N audio
- 15 objects from the M downmix signals; and a bit stream generating component configured to receive the M downmix signals from the downmix generating component and the reconstruction matrix from the analyzing component and to generate a bit stream comprising the M downmix signals and at least some of the matrix elements of the reconstruction matrix.

20

II. Overview - Decoder

- According to a second aspect, example embodiments propose decoding methods, decoding devices, and computer program products for decoding. The proposed methods, devices and computer program products may generally have the
- 25 same features and advantages.

 Advantages regarding features and setups as presented in the overview of the encoder above may generally be valid for the corresponding features and setups for the decoder.

- According to exemplary embodiments, there is provided a method for
- 30 decoding a time-frequency tile of an audio scene which at least comprises N audio objects, the method comprising the steps of: receiving a bit stream comprising M downmix signals and at least some matrix elements of a reconstruction matrix; generating the reconstruction matrix using the matrix elements; and reconstructing the N audio objects from the M downmix signals using the reconstruction matrix.

According to exemplary embodiments, the M downmix signals are arranged in a first field of the bit stream using a first format, and the matrix elements are arranged in a second field of the bit stream using a second format, thereby allowing a decoder that only supports the first format to decode and playback the M downmix
5 signals in the first field and to discard the matrix elements in the second field.

According to exemplary embodiments the matrix elements of the reconstruction matrix are time and frequency variant.

According to exemplary embodiments the audio scene further comprises a plurality of bed channels, the method further comprising reconstructing the bed
10 channels from the M downmix signals using the reconstruction matrix.

According to exemplary embodiments the number M of downmix signals is larger than two.

According to exemplary embodiments, the method further comprises: receiving L auxiliary signals being formed from the N audio objects; reconstructing
15 the N audio objects from the M downmix signals and the L auxiliary signals using the reconstruction matrix, wherein the reconstruction matrix comprises matrix elements that enable reconstruction of at least the N audio objects from the M downmix signals and the L auxiliary signals.

According to exemplary embodiments at least one of the L auxiliary signals is
20 equal to one of the N audio objects.

According to exemplary embodiments at least one of the L auxiliary signals is a combination of the N audio objects.

According to exemplary embodiments, the M downmix signals span a hyperplane, and wherein at least one of the plurality of auxiliary signals does not lie
25 in the hyperplane spanned by the M downmix signals.

According to exemplary embodiments, the at least one of the plurality of auxiliary signals that does not lie in the hyperplane is orthogonal to the hyperplane spanned by the M downmix signals.

As discussed above, audio encoding/decoding systems typically operate in
30 the frequency domain. Thus, audio encoding/decoding systems perform time/frequency transforms of audio signals using filter banks. Different types of time/frequency transforms may be used. For example the M downmix signals may be represented with respect to a first frequency domain and the reconstruction matrix may be represented with respect to a second frequency

domain. In order to reduce the computational burden in the decoder, it is advantageous to choose the first and the second frequency domains in a clever manner. For example, the first and the second frequency domain could be chosen as the same frequency domain, such as a Modified Discrete Cosine Transform (MDCT) domain. In this way one can avoid transforming the M downmix signals from the first frequency domain to the time domain followed by a transformation to the second frequency domain in the decoder. Alternatively it may be possible to choose the first and the second frequency domains in such a way that the transform from the first frequency domain to the second frequency domain can be implemented jointly such that it is not necessary to go all the way via the time domain in between.

The method may further comprise receiving positional data corresponding to the N audio objects, and rendering the N audio objects using the positional data to create at least one output audio channel. In this way the reconstructed N audio objects are mapped on the output channels of the audio encoder/decoder system based on their position in the three-dimensional space.

The rendering is preferably performed in a frequency domain. In order to reduce the computational burden in the decoder, the frequency domain of the rendering is preferably chosen in a clever way with respect to the frequency domain in which the audio objects are reconstructed. For example, if the reconstruction matrix is represented with respect to a second frequency domain corresponding to a second filter bank, and the rendering is performed in a third frequency domain corresponding to a third filter bank, the second and the third filter banks are preferably chosen to at least partly be the same filter bank. For example, the second and the third filter bank may comprise a Quadrature Mirror Filter (QMF) domain. Alternatively, the second and the third frequency domain may comprise an MDCT filter bank. According to an example embodiment, the third filter bank may be composed of a sequence of filter banks, such as a QMF filter bank followed by a Nyquist filter bank. If so, at least one of the filter banks of the sequence (the first filter bank of the sequence) is equal to the second filter bank. In this way, the second and the third filter bank may be said to at least partly be the same filter bank.

According to exemplary embodiments, there is provided a computer-readable medium comprising computer code instructions adapted to carry out any method of the second aspect when executed on a device having processing capability.

According to exemplary embodiments, there is provided a decoder for decoding a time-frequency tile of an audio scene which at least comprises N audio objects, comprising: a receiving component configured to receive a bit stream comprising M downmix signals and at least some matrix elements of a reconstruction matrix; a reconstruction matrix generating component configured to receive the matrix elements from the receiving component and based thereupon generate the reconstruction matrix; and a reconstructing component configured to receive the reconstruction matrix from the reconstruction matrix generating component and to reconstruct the N audio objects from the M downmix signals using the reconstruction matrix.

According to one aspect of the present invention, there is provided a method for decoding an audio scene represented by N audio signals, the method comprising: receiving a bit stream comprising M downmix signals and matrix elements of a reconstruction matrix; generating the reconstruction matrix using the matrix elements; and reconstructing the N audio signals from the M downmix signals using the reconstruction matrix, wherein approximations of the N audio signals are obtained as linear combinations of the M downmix signals with the matrix elements of the reconstruction matrix as coefficients in the linear combinations, wherein M is less than N, and M is equal or greater than one.

According to another aspect of the present invention, there is provided an audio decoder that decodes an audio scene represented by N audio signals, the audio decoder comprising: a receiver that receives a bit stream comprising M downmix signals and matrix elements of a reconstruction matrix; a reconstruction matrix generator that receives the matrix elements from the receiver and based thereupon generates the reconstruction matrix; and a reconstructor that receives the reconstruction matrix from the reconstruction matrix generator and reconstructs the N audio signals from the M downmix signals using the reconstruction matrix, wherein approximations of the N audio signals are obtained as linear combinations of the M downmix signals with the matrix elements of the reconstruction matrix as coefficients

in the linear combinations, wherein M is less than N, and M is equal or greater than one.

III. Example embodiments

Fig. 1 illustrates an encoding/decoding system 100 for encoding/decoding of
5 an audio scene 102. The encoding/decoding system 100 comprises an encoder 108,
a bit stream generating component 110, a bit stream decoding component 118, a
decoder 120, and a renderer 122.

The audio scene 102 is represented by one or more audio objects 106a, i.e.
audio signals, such as N audio objects. The audio scene 102 may further comprise
10 one or more bed channels 106b, i.e. signals that directly correspond to one of the
output channels of the renderer 122. The audio scene 102 is further represented by
metadata comprising positional information 104. The positional information 104 is for
example used by the renderer 122 when rendering the audio scene 102. The
positional information 104 may associate the audio objects 106a, and possibly also
15 the bed channels 106b, with a spatial position in a three dimensional space as a
function of time. The metadata may further comprise other type of data which is
useful in order to render the audio scene 102.

The encoding part of the system 100 comprises the encoder 108 and the bit
stream generating component 110. The encoder 108 receives the audio objects
20 106a, the bed channels 106b if present, and the metadata comprising positional
information 104. Based thereupon, the encoder 108 generates one or more downmix
signals 112, such as M downmix signals. By way of example, the downmix signals
112 may correspond to the channels [*Lf Rf Cf Ls Rs LFE*] of a 5.1 audio system. ("L"
stands for left, "R" stands for right, "C" stands for center, "f" stands for front, "s"
25 stands for surround, and "LFE" for low frequency effects).

The encoder 108 further generates side information. The side information comprises a reconstruction matrix. The reconstruction matrix comprises matrix elements 114 that enable reconstruction of at least the audio objects 106a from the downmix signals 112. The reconstruction matrix may further enable reconstruction of the bed channels 106b.

The encoder 108 transmits the M downmix signals 112, and at least some of the matrix elements 114 to the bit stream generating component 110. The bit stream generating component 110 generates a bit stream 116 comprising the M downmix signals 112 and at least some of the matrix elements 114 by performing quantization and encoding. The bit stream generating component 110 further receives the metadata comprising positional information 104 for inclusion in the bit stream 116.

The decoding part of the system comprises the bit stream decoding component 118 and the decoder 120. The bit stream decoding component 118 receives the bit stream 116 and performs decoding and dequantization in order to extract the M downmix signals 112 and the side information comprising at least some of the matrix elements 114 of the reconstruction matrix. The M downmix signals 112 and the matrix elements 114 are then input to the decoder 120 which based thereupon generates a reconstruction 106' of the N audio objects 106a and possibly also the bed channels 106b. The reconstruction 106' of the N audio objects is hence an approximation of the N audio objects 106a and possibly also of the bed channels 106b.

By way of example, if the downmix signals 112 correspond to the channels [Lf Rf Cf Ls Rs LFE] of a 5.1 configuration, the decoder 120 may reconstruct the objects 106' using only the full-band channels [Lf Rf Cf Ls Rs], thus ignoring the LFE. This also applies to other channel configurations. The LFE channel of the downmix 112 may be sent (basically unmodified) to the renderer 122.

The reconstructed audio objects 106', together with the positional information 104, are then input to the renderer 122. Based on the reconstructed audio objects 106' and the positional information 104, the renderer 122 renders an output signal 124 having a format which is suitable for playback on a desired loudspeaker or headphones configuration. Typical output formats are a standard 5.1 surround setup (3 front loudspeakers, 2 surround loud speakers, and 1 low frequency effects, LFE, loudspeaker) or a 7.1 + 4 setup (3 front loudspeakers, 4 surround loud speakers, 1 LFE loudspeaker, and 4 elevated speakers).

In some embodiments, the original audio scene may comprise a large number of audio objects. Processing of a large number of audio objects comes at the cost of high computational complexity. Also the amount of side information (the positional information 104 and the reconstruction matrix elements 114) to be embedded in the bit stream 116 depends on the number of audio objects. Typically the amount of side information grows linearly with the number of audio objects. Thus, in order to save computational complexity and/or to reduce the bitrate needed to encode the audio scene, it may be advantageous to reduce the number of audio objects prior to encoding. For this purpose the audio encoder/decoder system 100 may further comprise a scene simplification module (not shown) arranged upstreams of the encoder 108. The scene simplification module takes the original audio objects and possibly also the bed channels as input and performs processing in order to output the audio objects 106a. The scene simplification module reduces the number, K say, of original audio objects to a more feasible number N of audio objects 106a by performing clustering. More precisely, the scene simplification module organizes the K original audio objects and possibly also the bed channels into N clusters. Typically, the clusters are defined based on spatial proximity in the audio scene of the K original audio objects/bed channels. In order to determine the spatial proximity, the scene simplification module may take positional information of the original audio objects/bed channels as input. When the scene simplification module has formed the N clusters, it proceeds to represent each cluster by one audio object. For example, an audio object representing a cluster may be formed as a sum of the audio objects/bed channels forming part of the cluster. More specifically, the audio content of the audio objects/bed channels may be added to generate the audio content of the representative audio object. Further, the positions of the audio objects/bed channels in the cluster may be averaged to give a position of the representative audio object. The scene simplification module includes the positions of the representative audio objects in the positional data 104. Further, the scene simplification module outputs the representative audio objects which constitute the N audio objects 106a of Fig. 1.

The M downmix signals 112 may be arranged in a first field of the bit stream 116 using a first format. The matrix elements 114 may be arranged in a second field of the bit stream 116 using a second format. In this way, a decoder that only supports the first format is able to decode and playback the M downmix signals 112 in the first field and to discard the matrix elements 114 in the second field.

The audio encoder/decoder system 100 of Fig. 1 supports both the first and the second format. More precisely, the decoder 120 is configured to interpret the first and the second formats, meaning that it is capable of reconstructing the objects 106' based on the M downmix signals 112 and the matrix elements 114.

5 Fig. 2 illustrates an audio encoder/decoder system 200. The encoding part 108, 110 of the system 200 corresponds to that of Fig. 1. However, the decoding part of the audio encoder/decoder system 200 differs from that of the audio encoder/decoder system 100 of Fig. 1. The audio encoder/decoder system 200 comprises a legacy decoder 230 which supports the first format but not the second
10 format. Thus, the legacy decoder 230 of the audio encoder/decoder system 200 is not capable of reconstructing the audio objects/bed channels 106a-b. However, since the legacy decoder 230 supports the first format, it may still decode the M downmix signals 112 in order to generate an output 224 which is a channel based representation, such as a 5.1 representation, suitable for direct playback over a
15 corresponding multichannel loudspeaker setup. This property of the downmix signals is referred to as backwards compatibility meaning that also a legacy decoder which does not support the second format, i.e. is incapable of interpreting the side information comprising the matrix elements 114, may still decode and playback the M downmix signals 112.

20 The operation on the encoder side of the audio encoding/decoding system 100 will now be described in more detail with reference to Fig. 3 and the flowchart of Fig. 4.

Fig. 4 illustrates the encoder 108 and the bit stream generating component 110 of Fig. 1 in more detail. The encoder 108 has a receiving component (not
25 shown), a downmix generating component 318 and an analyzing component 328.

In step E02, the receiving component of the encoder 108 receives the N audio objects 106a and the bed channels 106b if present. The encoder 108 may further receive the positional data 104. Using vector notation the N audio objects may be denoted by a vector $S = [S1 \ S2 \ ... \ SN]^T$, and the bed channels by a vector B . The N
30 audio objects and the bed channels may together be represented by a vector $A = [B^T \ S^T]^T$.

In step E04, the downmix generating component 318 generates M downmix signals 112 from the N audio objects 106a and the bed channels 106b if present. Using vector notation, the M downmix signals may be represented by a vector

$D = [D1 D2 \dots DM]^T$ comprising the M downmix signals. Generally a downmix of a plurality of signals is a combination of the signals, such as a linear combination of the signals. By way of example, the M downmix signals may correspond to a particular loudspeaker configuration, such as the configuration of the loudspeakers
5 $[Lf Rf Cf Ls Rs LFE]$ in a 5.1 loudspeaker configuration.

The downmix generating component 318 may use the positional information 104 when generating the M downmix signals, such that the objects will be combined into the different downmix signals based on their position in a three-dimensional space. This is particularly relevant when the M downmix signals themselves
10 correspond to a specific loudspeaker configuration as in the above example. By way of example, the downmix generating component 318 may derive a presentation matrix Pd (corresponding to a presentation matrix applied in the renderer 122 of Fig. 1) based on the positional information and use it to generate the downmix according to $D = Pd * [B^T S^T]^T$.

15 The N audio objects 106a and the bed channels 106b if present are also input to the analyzing component 328. The analyzing component 328 typically operates on individual time/frequency tiles of the input audio signals 106a-b. For this purpose, the N audio objects 106a and the bed channels 106b may be fed through a filter bank 338, e.g. a QMF bank, which performs a time to frequency transform of the input
20 audio signals 106a-b. In particular, the filter bank 338 is associated with a plurality of frequency sub-bands. The frequency resolution of a time/frequency tile corresponds to one or more of these frequency sub-bands. The frequency resolution of the time/frequency tiles may be non-uniform, i.e. it may vary with frequency. For example, a lower frequency resolution may be used for high frequencies, meaning
25 that a time/frequency tile in the high frequency range may corresponds to several frequency sub-bands as defined by the filter bank 338.

In step E06, the analyzing component 328 generates a reconstruction matrix, here denoted by $R1$. The generated reconstruction matrix is composed of a plurality of matrix elements. The reconstruction matrix $R1$ is such that it allows reconstruction
30 of (an approximation) of the audio objects N 106a and possibly also the bed channels 106b from the M downmix signals 112 in the decoder.

The analyzing component 328 may take different approaches to generate the reconstruction matrix. For example, a Minimum Mean Squared Error (MMSE) predictive approach can be used which takes both the N audio objects/bed channels

106a-b as input as well as the M downmix signals 112 as input. This can be described as an approach which aims at finding the reconstruction matrix that minimizes the mean squared error of the reconstructed audio objects/bed channels. Particularly, the approach reconstructs the N audio objects/bed channels using a candidate reconstruction matrix and compares them to the input audio objects/bed channels 106a-b in terms of the mean squared error. The candidate reconstruction matrix that minimizes the mean squared error is selected as the reconstruction matrix and its matrix elements 114 are output of the analyzing component 328.

The MMSE approach requires estimates of correlation and covariance matrices of the N audio objects/bed channels 106a-b and the M downmix signals 112. According to the above approach, these correlations and covariances are measured based on the N audio objects/bed channels 106a-b and the M downmix signals 112. In an alternative, model-based, approach the analyzing component 328 takes the positional data 104 as input instead of the M downmix signals 112. By making certain assumptions, e.g. assuming that the N audio objects are mutually uncorrelated, and using this assumption in combination with the downmix rules applied in the downmix generating component 318, the analyzing component 328 may compute the required correlations and covariances needed to carry out the MMSE method described above.

The elements of the reconstruction matrix 114 and the M downmix signals 112 are then input to the bit stream generating component 110. In step E08, the bit stream generating component 110 quantizes and encodes the M downmix signals 112 and at least some of the matrix elements 114 of the reconstruction matrix and arranges them in the bit stream 116. In particular, the bit stream generating component 110 may arrange the M downmix signals 112 in a first field of the bit stream 116 using a first format. Further, the bit stream generating component 110 may arrange the matrix elements 114 in a second field of the bit stream 116 using a second format. As previously described with reference to Fig. 2, this allows a legacy decoder that only supports the first format to decode and playback the M downmix signals 112 and to discard the matrix elements 114 in the second field.

Fig. 5 illustrates an alternative embodiment of the encoder 108. Compared to the encoder shown in Fig. 3, the encoder 508 of Fig. 5 further allows one or more auxiliary signals to be included in the bit stream 116.

For this purpose, the encoder 508 comprises an auxiliary signals generating component 548. The auxiliary signals generating component 548 receives the audio objects/bed channels 106a-b and based thereupon one or more auxiliary signals 512 are generated. The auxiliary signals generating component 548 may for example
5 generate the auxiliary signals 512 as a combination of the audio objects/bed channels 106a-b. Denoting the auxiliary signals by the vector $C = [C1 \ C2 \ ... \ CL]^T$, the auxiliary signals may be generated as $C = Q * [B^T \ S^T]^T$, where Q is a matrix which can be time and frequency variant. This includes the case where the auxiliary signals equals one or more of the audio objects and where the auxiliary signals are linear
10 combinations of the audio objects. For example, the auxiliary signal could represent be a particularly important object, such as dialogue.

The role of the auxiliary signals 512 is to improve the reconstruction of the audio objects/bed channels 106a-b in the decoder. More precisely, on the decoder side, the audio objects/bed channels 106a-b may be reconstructed based on the M
15 downmix signals 112 as well as the L auxiliary signals 512. The reconstruction matrix will therefore comprises matrix elements 114 which allow reconstruction of the audio objects/bed channels from the M downmix signals 112 as well as the L auxiliary signals.

The L auxiliary signals 512 may therefore be input to the analyzing component
20 328 such that they are taken into account when generating the reconstruction matrix. The analyzing component 328 may also send a control signal to the auxiliary signals generating component 548. For example the analyzing component 328 may control which audio objects/bed channels to include in the auxiliary signals and how they are to be included. In particular, the analyzing component 328 may control the choice of
25 the Q -matrix. The control may for example be based on the MMSE approach described above such that the auxiliary signals are selected such that the reconstructed audio objects/bed channels are as close as possible to the audio objects/bed channels 106a-b.

The operation of the decoder side of the audio encoding/decoding system 100
30 will now be described in more detail with reference to Fig. 6 and the flowchart of Fig. 7.

Fig. 6 illustrates the bit stream decoding component 118 and the decoder 120 of Fig. 1 in more detail. The decoder 120 comprises a reconstruction matrix generating component 622 and a reconstructing component 624.

In step D02 the bit stream decoding component 118 receives the bit stream 116. The bit stream decoding component 118 decodes and dequantizes the information in the bit stream 116 in order to extract the M downmix signals 112 and at least some of the matrix elements 114 of the reconstruction matrix.

5 The reconstruction matrix generating component 622 receives the matrix elements 114 and proceeds to generate a reconstruction matrix 614 in step D04. The reconstruction matrix generating component 622 generates the reconstruction matrix 614 by arranging the matrix elements 114 at appropriate positions in the matrix. If not all matrix elements of the reconstruction matrix are received, the
10 reconstruction matrix generating component 622 may for example insert zeros instead of the missing elements.

The reconstruction matrix 614 and the M downmix signals are then input to the reconstructing component 624. The reconstructing component 624 then, in step D06, reconstructs the N audio objects and, if applicable, the bed channels. In other
15 words, the reconstructing component 624 generates an approximation 106' of the N audio objects/bed channels 106a-b.

By way of example, the M downmix signals may correspond to a particular loudspeaker configuration, such as the configuration of the loudspeakers [*Lf Rf Cf Ls Rs LFE*] in a 5.1 loudspeaker configuration. If so, the reconstructing
20 component 624 may base the reconstruction of the objects 106' only on the downmix signals corresponding to the full-band channels of the loudspeaker configuration. As explained above, the band-limited signal (the low-frequency LFE signal) may be sent basically unmodified to the renderer.

The reconstructing component 624 typically operates in a frequency domain.
25 More precisely, the reconstructing component 624 operates on individual time/frequency tiles of the input signals. Therefore the M downmix signals 112 are typically subject to a time to frequency transform 623 before being input to the reconstructing component 624. The time to frequency transform 623 is typically the same or similar to the transform 338 applied on the encoder side. For example, the
30 time to frequency transform 623 may be a QMF transform.

In order to reconstruct the audio objects/bed channels 106', the reconstructing component 624 applies a matrixing operation. More specifically, using the previously introduced notation, the reconstructing component 624 may generate an approximation A' of the audio object/bed channels as $A' = R1 * D$. The

reconstruction matrix $R1$ may vary as a function of time and frequency. Thus, the reconstruction matrix may vary between different time/frequency files processed by the reconstructing component 624.

5 The reconstructed audio objects/bed channels 106' are typically transformed back to the time domain 625 prior to being output from the decoder 120.

Fig. 8 illustrates the situation when the bit stream 116 additionally comprises auxiliary signals. Compared to the embodiment of Fig. 7, the bit stream decoding component 118 now additionally decodes one or more auxiliary signals 512 from the bit stream 116. The auxiliary signals 512 are input to the reconstructing component 10 624 where they are included in the reconstruction of the audio objects/bed channels. More particularly, the reconstructing component 624 generates the audio objects/bed channels by applying the matrix operation $A' = R1 * [D^T C^T]^T$.

Fig. 9 illustrates the different time/frequency transforms used on the decoder side in the audio encoding/decoding system 100 of Fig. 1. The bit stream decoding 15 component 118 receives the bit stream 116. A decoding and dequantizing component 918 decodes and dequantizes the bit stream 116 in order to extract positional information 104, the M downmix signals 112, and matrix elements 114 of a reconstruction matrix.

At this stage, the M downmix signals 112 are typically represented in a first 20 frequency domain, corresponding to a first set of time/frequency filter banks here denoted by T/F_C and F/T_C for transformation from the time domain to the first frequency domain and from the first frequency domain to the time domain, respectively. Typically, the filter banks corresponding to the first frequency domain may implement an overlapping window transform, such as an MDCT and an inverse 25 MDCT. The bit stream decoding component 118 may comprise a transforming component 901 which transforms the M downmix signals 112 to the time domain by using the filter bank F/T_C .

The decoder 120, and in particular the reconstructing component 624, typically processes signals with respect to a second frequency domain. The second 30 frequency domain corresponds to a second set of time/frequency filter banks here denoted by T/F_U and F/T_U for transformation from the time domain to the second frequency domain and from the second frequency domain to the time domain, respectively. The decoder 120 may therefore comprise a transforming component 903 which transforms the M downmix signals 112, which are represented in the time

domain, to the second frequency domain by using the filter bank T/F_U . When the reconstructing component 624 has reconstructed the objects 106' based on the M downmix signals by performing processing in the second frequency domain, a transforming component 905 may transform the reconstructed objects 106' back to the time domain by using the filter bank F/T_U .

The renderer 122 typically processes signals with respect to a third frequency domain. The third frequency domain corresponds to a third set of time/frequency filter banks here denoted by T/F_R and F/T_R for transformation from the time domain to the third frequency domain and from the third frequency domain to the time domain, respectively. The renderer 122 may therefore comprise a transform component 907 which transforms the reconstructed audio objects 106' from the time domain to the third frequency domain by using the filter bank T/F_R . Once the renderer 122, by means of a rendering component 922, has rendered the output channels 124, the output channels may be transformed to the time domain by a transforming component 909 by using the filter bank F/T_R .

As is evident from the above description, the decoder side of the audio encoding/decoding system includes a number of time/frequency transformation steps. However, if the first, the second, and the third frequency domains are selected in certain ways, some of the time/frequency transformation steps become redundant.

For example, some of the first, the second, and the third frequency domains could be chosen to be the same or could be implemented jointly to go directly from one frequency domain to the other without going all the way to the time-domain in between. An example of the latter is the case where the only difference between the second and the third frequency domain is that the transform component 907 in the renderer 122 uses a Nyquist filter bank for increased frequency resolution at low frequencies in addition to a QMF filter bank that is common to both transformation components 905 and 907. In such case, the transform components 905 and 907 can be implemented jointly in the form of a Nyquist filter bank, thus saving computational complexity.

In another example, the second and the third frequency domain are the same. For example, the second and the third frequency domain may both be a QMF frequency domain. In such case, the transform components 905 and 907 are redundant and may be removed, thus saving computational complexity.

According to another example, the first and the second frequency domains may be the same. For example the first and the second frequency domains may both be a MDCT domain. In such case, the first and the second transform components 901 and 903 may be removed, thus saving computational complexity.

5

Equivalents, extensions, alternatives and miscellaneous

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

15 Additionally, variations to the disclosed embodiments can be understood and effected by the skilled person in practicing the disclosure, from a study of the drawings, the disclosure, and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality. The mere fact that certain measures are recited
20 in mutually different dependent claims does not indicate that a combination of these measured cannot be used to advantage.

The systems and methods disclosed hereinabove may be implemented as software, firmware, hardware or a combination thereof. In a hardware
implementation, the division of tasks between functional units referred to in the
25 above description does not necessarily correspond to the division into physical units; to the contrary, one physical component may have multiple functionalities, and one task may be carried out by several physical components in cooperation. Certain components or all components may be implemented as software executed by a digital signal processor or microprocessor, or be implemented as hardware or as an
30 application-specific integrated circuit. Such software may be distributed on computer readable media, which may comprise computer storage media (or non-transitory media) and communication media (or transitory media). As is well known to a person skilled in the art, the term computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or

technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic
5 cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by a computer. Further, it is well known to the skilled person that communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a
10 carrier wave or other transport mechanism and includes any information delivery media.

CLAIMS:

1. A method for decoding an audio scene represented by N audio signals, the method comprising:

receiving a bit stream comprising M downmix signals and matrix elements of a
5 reconstruction matrix;

generating the reconstruction matrix using the matrix elements; and

reconstructing the N audio signals from the M downmix signals using the
reconstruction matrix, wherein approximations of the N audio signals are obtained as linear
combinations of the M downmix signals with the matrix elements of the reconstruction
10 matrix as coefficients in the linear combinations,

wherein M is less than N, and M is equal or greater than one.

2. The method of claim 1 further comprising receiving in the bit stream L auxiliary
signals and reconstructing the N audio signals from the M downmix signals and the L
auxiliary signals using the reconstruction matrix.

3. The method of claim 1 wherein at least some of the M downmix signals are formed
15 from two or more of the N audio signals.

4. The method of claim 1 wherein at least some of the N audio signals are rendered to
generate a three-dimensional audio environment.

5. The method of claim 1 wherein the audio scene comprises a three-dimensional
20 audio environment which includes audio elements being associated with positions in a
three-dimensional space that can be rendered for playback on an audio system.

6. The method of claim 1 wherein the M downmix signals are arranged in a first portion
of the bit stream using a first format and the matrix elements are arranged in a second field
of the bit stream using a second format.

7. The method of claim 1 wherein the matrix elements are transmitted as side information in the bit stream.
8. The method of claim 1 wherein at least some of the matrix elements are zero valued.
- 5 9. The method of claim 1 wherein the linear combinations are formed by multiplying a matrix of the M downmix signals with the reconstruction matrix.
- 10 10. The method of claim 1 further comprising receiving L auxiliary signals and wherein the linear combinations are formed by multiplying a matrix of the M downmix signals and the L auxiliary signals with the reconstruction matrix.
- 10 11. The method of claim 1 wherein the M downmix signals are decoded before the reconstructing.
12. The method of claim 1 further comprising receiving in the bit stream one or more bed channels and reconstructing the N audio signals from the M downmix signals and the bed channels using the reconstruction matrix.
- 15 13. The method of claim 12 further comprising receiving in the bit stream L auxiliary signals and reconstructing the N audio signals from the M downmix signals, the L auxiliary signals, and the one or more bed channels using the reconstruction matrix.
14. The method of claim 13 wherein the one or more bed channels represent audio elements having a fixed location in the audio scene.
- 20 15. An audio decoder that decodes an audio scene represented by N audio signals, the audio decoder comprising:

a receiver that receives a bit stream comprising M downmix signals and matrix elements of a reconstruction matrix;

a reconstruction matrix generator that receives the matrix elements from the
25 receiver and based thereupon generates the reconstruction matrix; and

a reconstructor that receives the reconstruction matrix from the reconstruction matrix generator and reconstructs the N audio signals from the M downmix signals using the reconstruction matrix, wherein approximations of the N audio signals are obtained as linear combinations of the M downmix signals with the matrix elements of the reconstruction matrix as coefficients in the linear combinations,

wherein M is less than N, and M is equal or greater than one.

16. The audio decoder of claim 15 wherein the receiver further receives in the bit stream L auxiliary signals and the reconstructor reconstructs the N audio signals from the M downmix signals and the L auxiliary signals using the reconstruction matrix.

17. The audio decoder of claim 15 wherein at least some of the M downmix signals are formed from two or more of the N audio signals.

18. The audio decoder of claim 15 wherein the M downmix signals are arranged in a first portion of the bit stream using a first format and the matrix elements are arranged in a second field of the bit stream using a second format.

19. The audio decoder of claim 15 wherein the matrix elements are transmitted as side information in the bit stream and at least some of the matrix elements are zero valued.

20. A non-transitory computer-readable medium including instructions, which when executed by a processor of an information processing system, cause the information processing system to perform the method of claim 1.

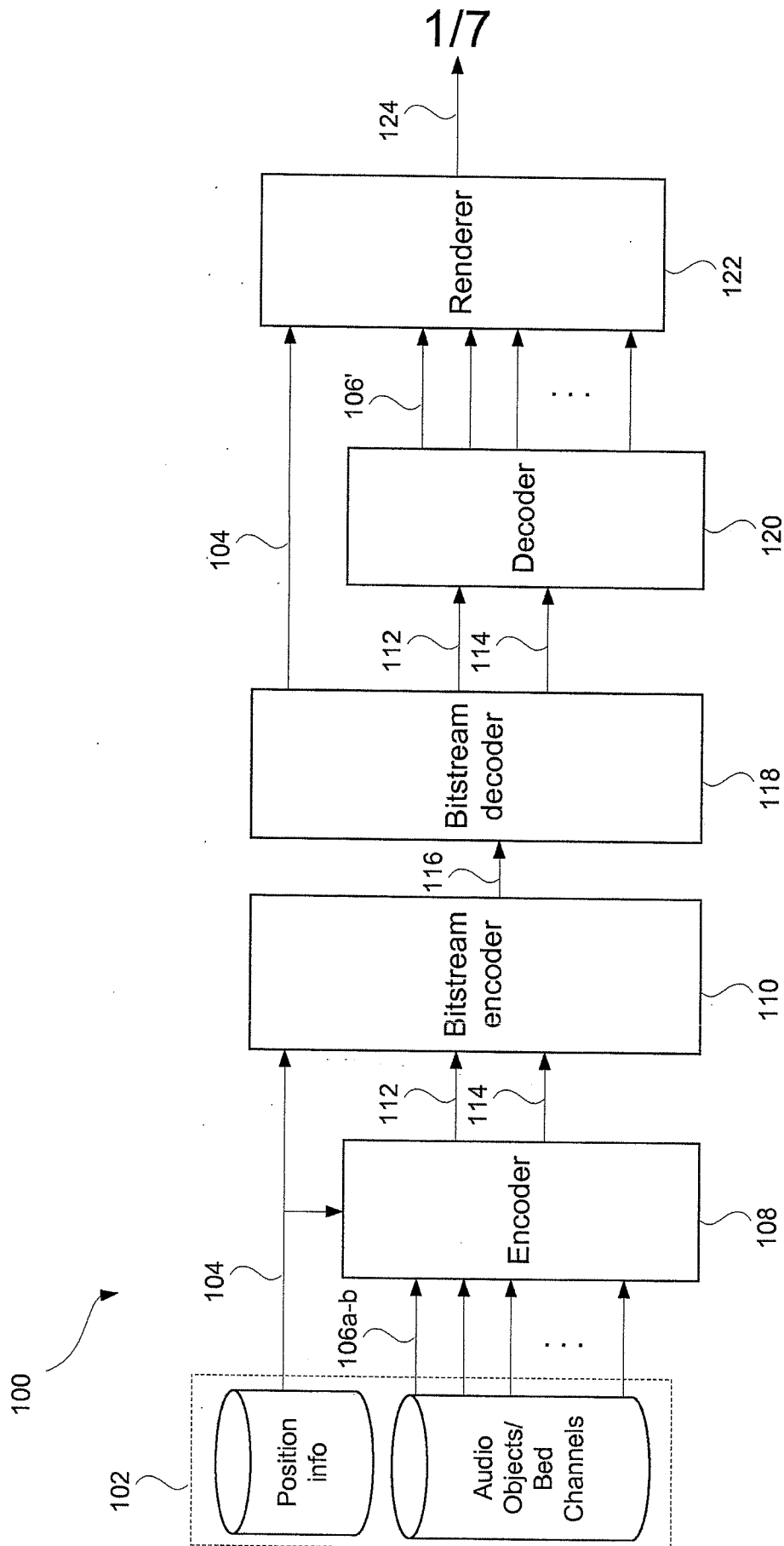


Fig. 1

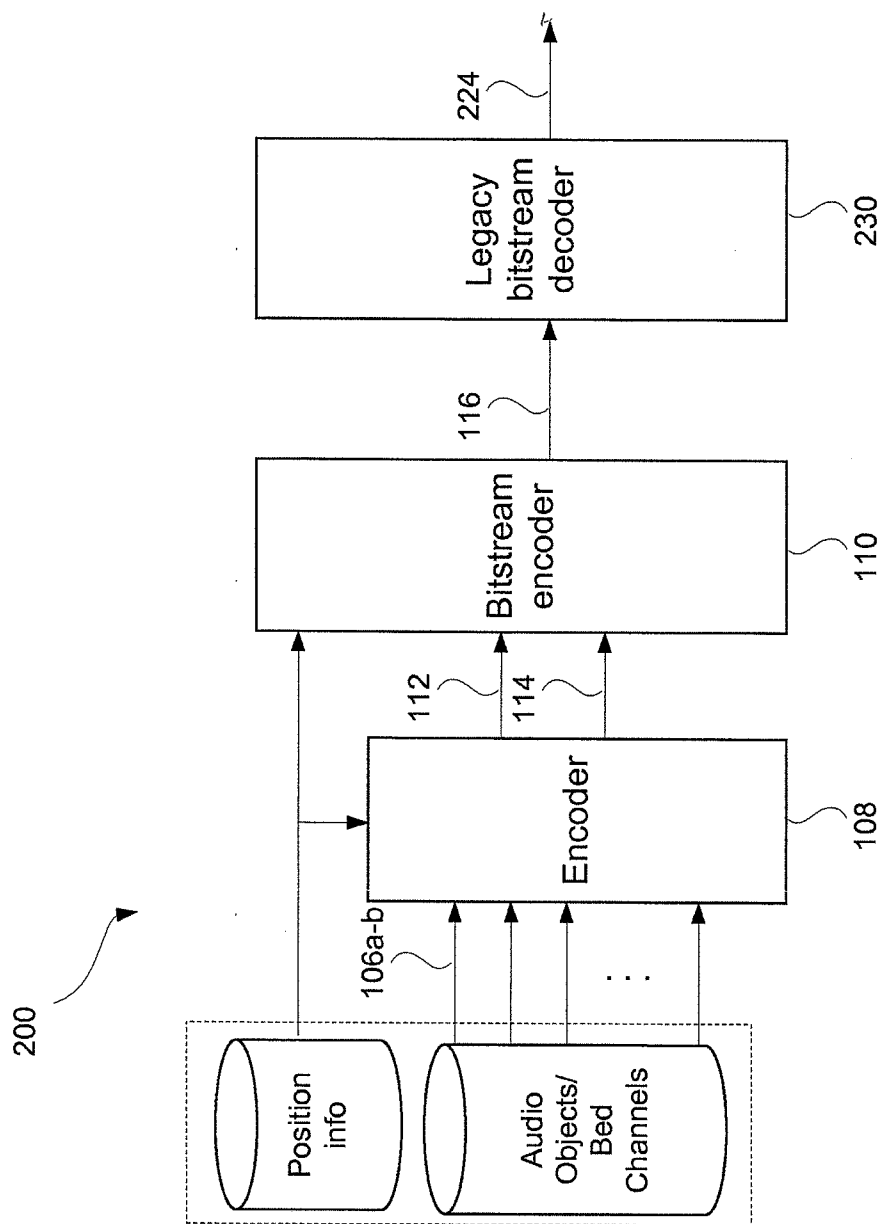


Fig. 2

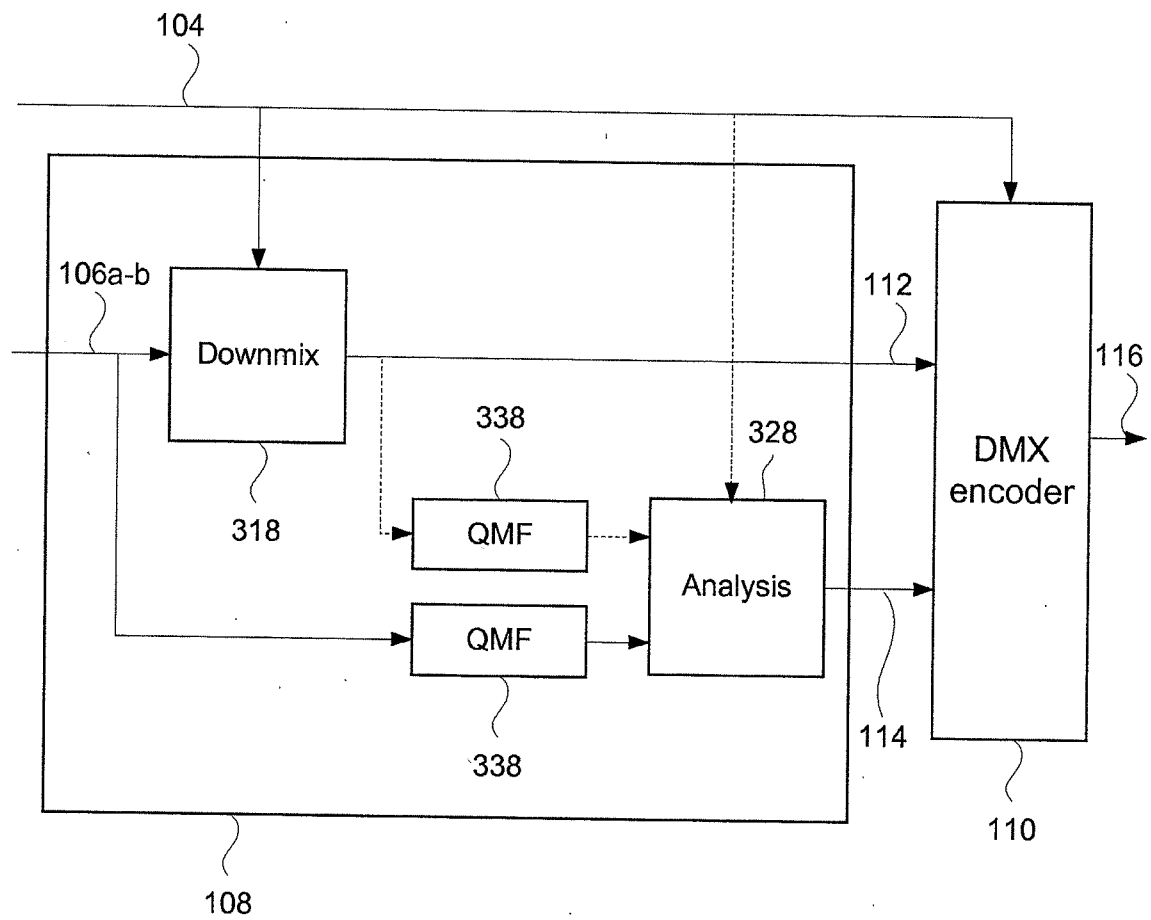


Fig. 3

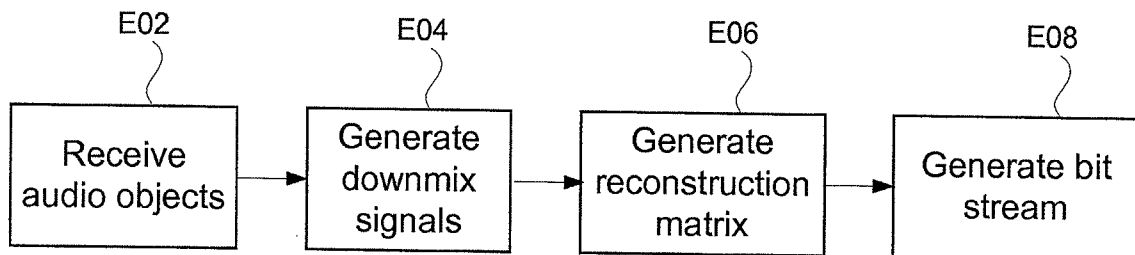
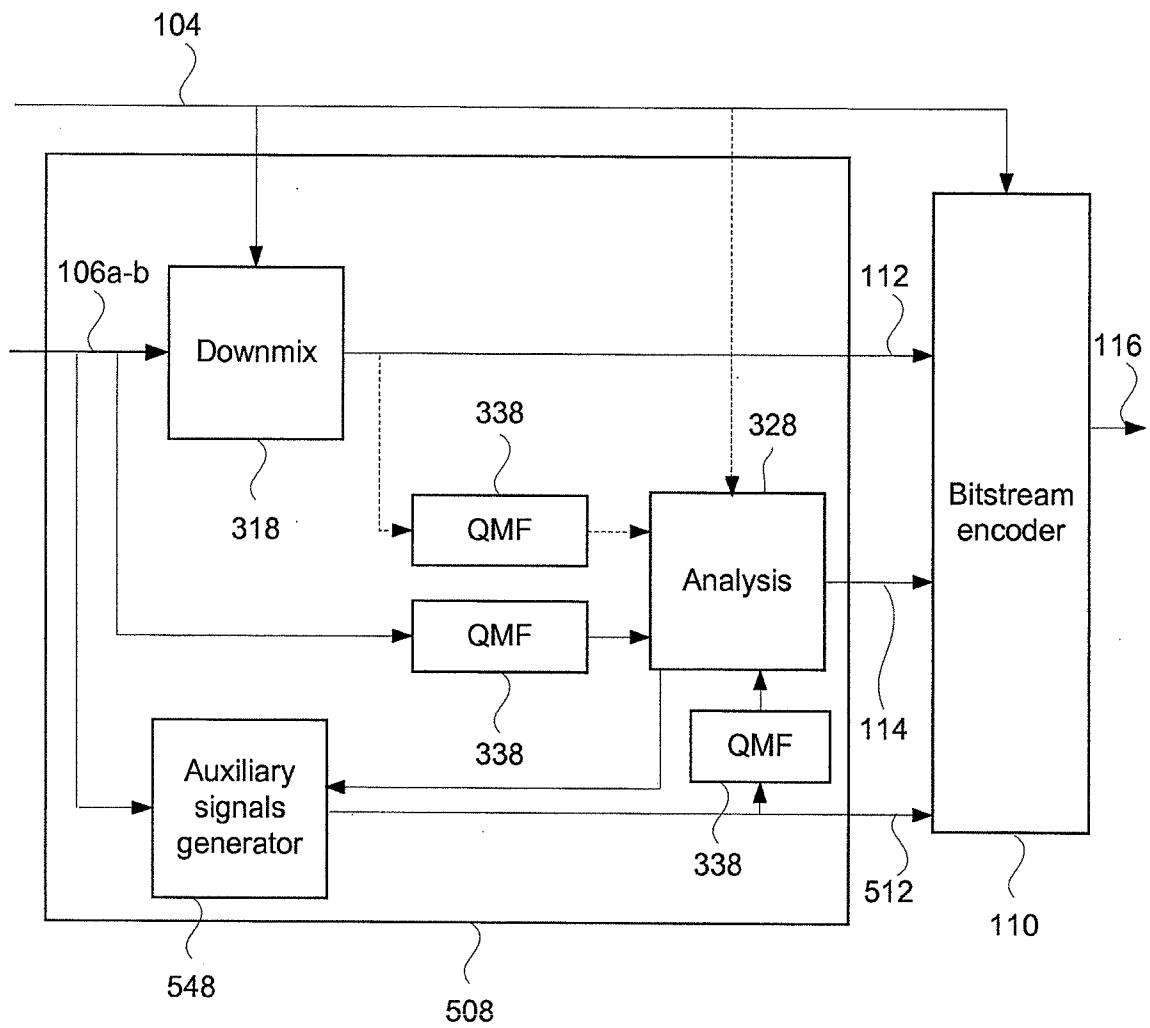


Fig. 4

*Fig. 5*

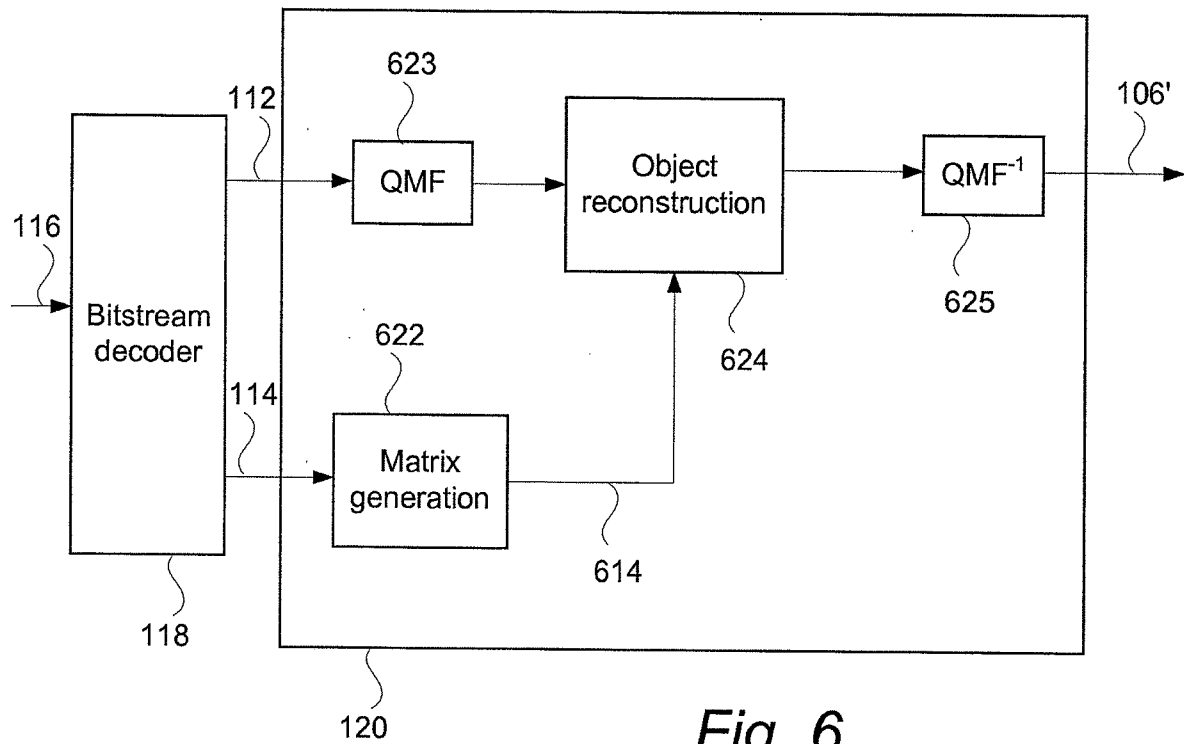


Fig. 6

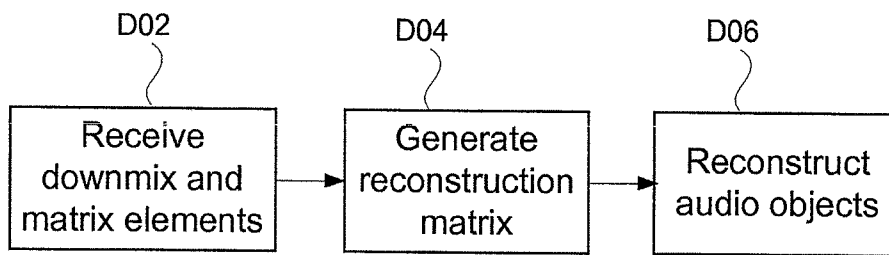
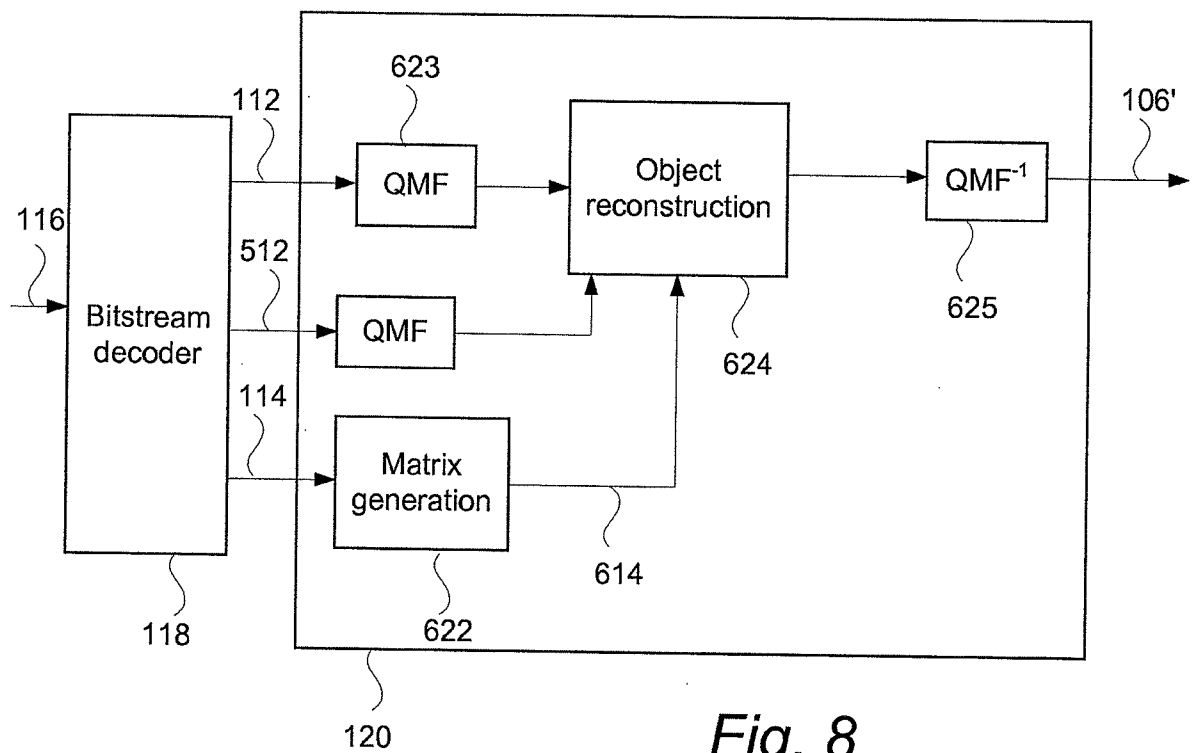


Fig. 7



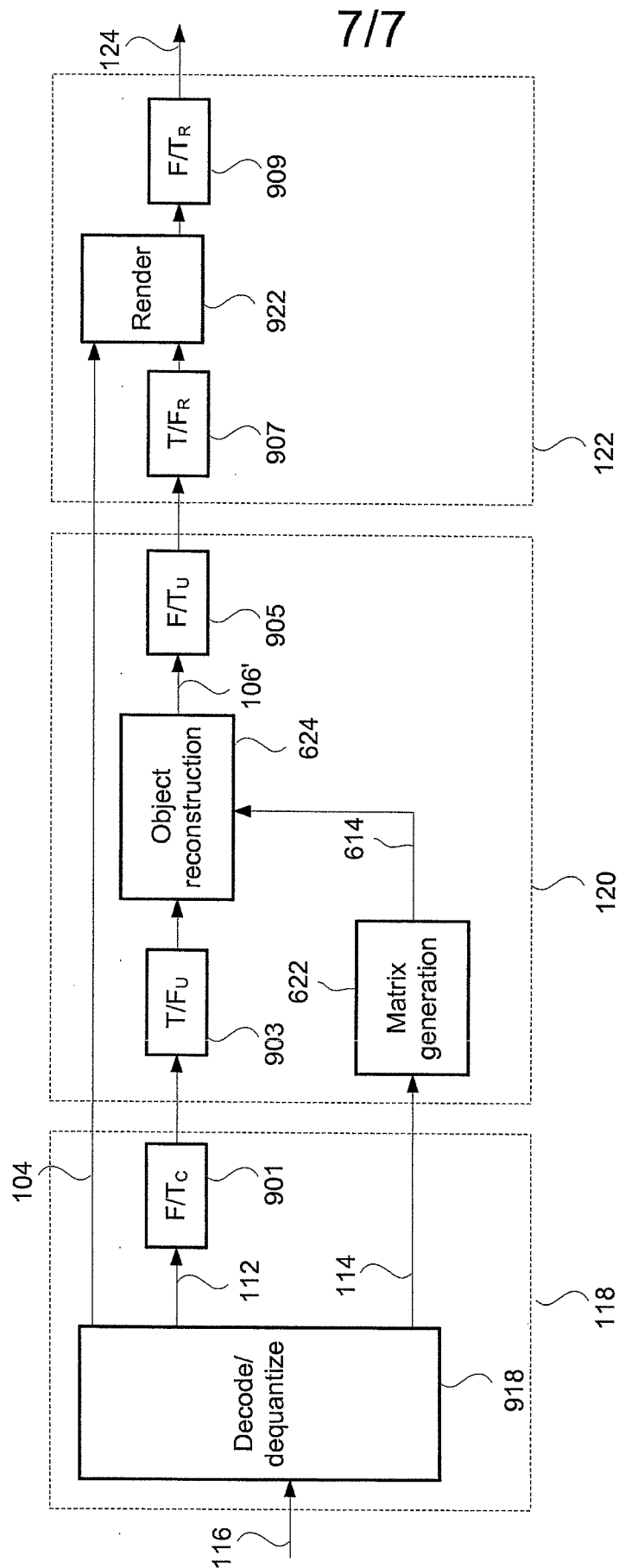


Fig. 9

