

**ENCODING OF MULTI-CHANNEL AUDIO SIGNALS COMPRISING DOWNMIXING OF A
PRIMARY AND TWO OR MORE SCALED NON-PRIMARY INPUT CHANNELS**

5 **CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims priority to United States Provisional Patent Application No. 63/037,635, filed June 11, 2020, and United States Provisional Patent Application No. 63/193,926, filed May 27, 2021, each of which is hereby incorporated by reference in its entirety.

10 **TECHNICAL FIELD**

[0002] This disclosure relates generally to audio coding, and in particular to coding of multi-channel audio signals.

BACKGROUND

15 [0003] When an input audio signal is to be stored or transmitted for later use (e.g., to be played back to a listener) it is often desirable to encode the audio signal with a reduced amount of data. The process of data reduction, as applied to an input audio signal, is commonly referred to as “audio encoding” (or “encoding”), and the apparatus used for encoding is commonly referred to an “audio encoder” (or “encoder”). The process of regeneration of an
20 output audio signal from the reduced data is commonly referred to as “audio decoding” (or “decoding”), and the apparatus used for the decoding is commonly referred to as an “audio decoder” (or “decoder”). Audio encoders and decoders may be adapted to operate on input signals that are composed of a single audio channel or multiple audio channels. When an input signal is composed of multiple audio channels, the audio encoder and audio decoder is referred
25 to as a multi-channel audio encoder and a multi-channel audio decoder, respectively.

SUMMARY

30 [0004] Implementations are disclosed for adaptive downmixing of audio signals with improved continuity.

[0005] In some embodiments, an audio encoding method comprises: receiving, with at least one processor, an input multi-channel audio signal comprising a primary input audio channel and L non-primary input audio channels; determining, with the at least one processor, a set of L input gains, where L is a positive integer greater than one; for each of the L non-
35 primary input audio channels and L input gains, forming a respective scaled non-primary input audio channel from the respective non-primary input audio channel scaled according to the

input gain; forming a primary output audio channel from the sum of the primary input audio channel and the scaled non-primary input audio channels; determining, with the at least one processor, a set of L prediction gains: for each of the L prediction gains, forming, with the at least one processor, a prediction channel from the primary output audio channel scaled
 5 according to the prediction gain; forming, with the at least one processor, L non-primary output audio channels from the difference of the respective non-primary input audio channel and the respective prediction signal; forming, with the at least one processor, an output multi-channel audio signal from the primary output audio channel and the L non-primary output audio channels; encoding, with an audio encoder, the output multi-channel audio signal; and
 10 transmitting or storing, with the at least one processor, the encoded output multi-channel audio signal.

[0006] In some embodiments, wherein determining the set of L input gains, comprises: determining a set of L mixing coefficients; determining an input mixture strength coefficient; and determining the L input gains by scaling the L mixing coefficients by the input mixture
 15 strength coefficient.

[0007] In some embodiments, determining the set of L prediction gains, comprises: determining a set of L mixing coefficients; determining a prediction mixture strength coefficient; and determining the L prediction gains by scaling the L mixing coefficients by the prediction mixture strength coefficient.

20 [0008] In some embodiments, the input mixture strength coefficient, h , is determined by a pre-prediction constraint equation, $h=fg$, where f is a pre-determined constant value greater than zero and less than or equal to one, and g is the prediction mixture strength coefficient.

[0009] In some embodiments, the prediction mixture strength coefficient, g , is a largest real value solution to: $\beta f^2 g^3 + 2\alpha f g^2 - \beta f g - \alpha + g w = 0$, where $\beta = u^H \times E \times u$,
 25 $u = \frac{1}{\alpha} v$, $\alpha = |v|_2 = \sqrt{\sum_{n=1}^N v_n^2}$, and quantity w , column vector v and matrix E are components of a covariance matrix for an intermediate signal that has a dominant channel.

[0010] In some embodiments, the covariance matrix of the intermediate signal is computed from a covariance matrix of the multi-channel input audio signal.

[0011] In some embodiments, two or more input multi-channel audio channels are processed according to a mixing matrix to produce the primary input audio channel and the L
 30 non-primary input audio channels.

[0012] In some embodiments, the primary input audio channel is determined by a dominant eigen-vector of an expected covariance of a typical input multi-channel audio signal.

[0013] In some embodiments, each of the L mixing coefficients are determined based on a correlation of a respective one of the non-primary input audio channels and the primary input audio channel.

[0014] In some embodiments, the encoding includes allocating more bits to the primary output audio channel than to the L non-primary output audio channels, or discarding one or more of the L non-primary output audio channels.

[0015] Other implementations disclosed herein are directed to a system, apparatus and computer-readable medium. The details of the disclosed implementations are set forth in the accompanying drawings and the description below. Other features, objects and advantages are apparent from the description, drawings and claims.

[0016] Particular implementations disclosed herein provide one or more of the following advantages. An input multi-channel audio signal is processed by an audio encoder pre-mixer to form an output multi-channel audio signal that has two desirable attributes for efficient encoding. The first attribute is that at least one dominant audio channel of the output multi-channel audio signal contains most or all of the sonic elements of the input multi-channel audio signal. The second attribute is that each of the audio channels of the output multi-channel audio signal are largely uncorrelated to each of the other audio channels. The simple encoder may provide data to a simple decoder to assist in the regeneration of audio channels that were discarded by the simple encoder.

[0017] The two attributes described above allow the output multi-channel audio signal to be efficiently encoded by a simple encoder by allocating fewer bits to the encoding of less dominant channels or choosing to discard less dominant audio channels entirely.

DESCRIPTION OF DRAWINGS

[0018] In the drawings, specific arrangements or orderings of schematic elements, such as those representing devices, units, instruction blocks and data elements, are shown for ease of description. However, it should be understood by those skilled in the art that the specific ordering or arrangement of the schematic elements in the drawings is not meant to imply that a particular order or sequence of processing, or separation of processes, is required. Further, the inclusion of a schematic element in a drawing is not meant to imply that such element is required in all embodiments or that the features represented by such element may not be included in or combined with other elements in some implementations.

[0019] Further, in the drawings, where connecting elements, such as solid or dashed lines or arrows, are used to illustrate a connection, relationship, or association between or

among two or more other schematic elements, the absence of any such connecting elements is not meant to imply that no connection, relationship, or association can exist. In other words, some connections, relationships, or associations between elements are not shown in the drawings so as not to obscure the disclosure. In addition, for ease of illustration, a single
5 connecting element is used to represent multiple connections, relationships or associations between elements. For example, where a connecting element represents a communication of signals, data, or instructions, it should be understood by those skilled in the art that such element represents one or multiple signal paths, as may be needed, to affect the communication.

[0020] FIG. 1 is a block diagram of an arrangement of a simple audio encoder and
10 simple audio decoder intended to form an output multi-channel audio signal that is a facsimile of an input multi-channel audio signal, according to some embodiments.

[0021] FIG. 2 is a block diagram of audio codec system that includes an audio encoder, audio decoder 106, encoder pre-mixer and decoder post-mixer, according to some embodiments.

15 [0022] FIG. 3 illustrates an arrangement of processing elements whereby an input multi-channel audio signal is split by a filterbank into subband signals, where each subband is processed by a mixing matrix to produce a remixed subband signal, according to some embodiments.

[0023] FIG. 4 is a block diagram of an arrangement of two mixing operations intended
20 to implement the function of the encoder pre-mixer of FIG. 2 or the encoder pre-mixer of FIG. 3, according to some embodiments.

[0024] FIG. 5 is a block diagram of a prediction mixer, according to some embodiments.

[0025] FIG. 6 shows an arrangement of processing elements that implement the
25 decoder post-mixer of FIG. 2, according to some embodiments.

[0026] FIG. 7 is a flow diagram of a process of adaptive downmixing of audio signals with improved continuity, according to some embodiments.

[0027] FIG. 8 is a block diagram of a system for implementing the features and processes described in reference to FIGS. 1-7, according to some embodiments.

30 [0028] The same reference symbol used in various drawings indicates like elements.

DETAILED DESCRIPTION

[0029] In the following detailed description, numerous specific details are set forth to provide a thorough understanding of the various described embodiments. It will be apparent to one of ordinary skill in the art that the various described implementations may be practiced without these specific details. In other instances, well-known methods, procedures, components, and circuits, have not been described in detail so as not to unnecessarily obscure aspects of the embodiments. Several features are described hereafter that can each be used independently of one another or with any combination of other features.

Nomenclature

10 [0030] As used herein, the term “includes” and its variants are to be read as open-ended terms that mean “includes, but is not limited to.” The term “or” is to be read as “and/or” unless the context clearly indicates otherwise. The term “based on” is to be read as “based at least in part on.” The term “one example implementation” and “an example implementation” are to be read as “at least one example implementation.” The term “another implementation” is to be read as “at least one other implementation.” The terms “determined,” “determines,” or “determining” are to be read as obtaining, receiving, computing, calculating, estimating, predicting or deriving. In addition, in the following description and claims, unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skills in the art to which this disclosure belongs.

20 [0031] FIG. 1 is a block diagram of an arrangement 10 of a simple audio encoder and simple audio decoder, intended to form a multi-channel audio signal 17 (Z') that is a facsimile of multi-channel audio signal 13 (Z). Multi-channel audio signal 13 is processed by simple audio encoder 14 to produce encoded representation 15, which may be stored and/or transmitted 20 to simple audio decoder 16 which produces multi-channel audio signal 17. Preferably, the data size of encoded representation 15 is minimized whilst minimizing the difference between multi-channel audio signal 13 and multi-channel audio signal 17. Furthermore, the difference between multi-channel audio signal 13 and multi-channel audio signal 17 may be measured according to similarity as perceived by a human listener. The measure of human-perceived similarity between audio signal 13 and audio signal 17 is based on a reference playback method (that is, the assumed default means by which the audio channels of multi-channel audio signals 13, 17 are presented as an auditory experience to the listener).

30 [0032] The efficiency of simple audio encoder 14 and decoder 16 may be defined in terms of the data rate (measured in bits per second) of the encoded representation 15 required

to provide multi-channel audio signal 17 that will be judged by a listener to match multi-channel audio signal 13 with a particular perceived quality level. Simple audio encoder 14 and decoder 16 may achieve greater efficiency (that is, a lower data rate) when the multi-channel audio signal 13 is known to possess particular attributes. In particular, greater efficiency may be achieved when it is known that multi-channel audio signal 13 possesses the following attributes (DD1 and DD2):

[0033] DD1: One or more channels of the multi-channel audio signal are generally more dominant than others, where a more dominant audio channel is one that will contain substantial elements of most (or all) of the sonic elements in the scene. That is, a dominant audio signal, when presented as a single audio channel to a listener, will contain most (or all) of the sonic elements of the multi-channel signal, when the multi-channel audio signal is presented to a listener through a reference playback method.

[0034] DD2: Each of the audio channels of the multi-channel audio signal is largely uncorrelated to each of the other audio channels

[0035] Given the knowledge that multi-channel audio signal 13 possesses attributes DD1 and DD2, simple audio encoder 14 may achieve improved efficiency using several techniques including, but not limited to: allocating fewer bits to the encoding of less dominant channels or choosing to discard less dominant channels entirely. Simple audio encoder 14 may provide data to simple audio decoder 16 to assist in the regeneration of channels that were discarded by simple encoder audio encoder 14. Preferably, a multi-channel audio signal that does not possess attributes DD1 and DD2 may be processed by an encoder pre-mixer to form, e.g., to calculate, to determine, to construct or to generate, a multi-channel audio signal that does possess attributes DD1 and DD2, as described further in reference to FIG. 2. A corresponding decoder post-mixer may be applied to the simple decoder output to form an output multi-channel audio signal, such that the decoder post-mixer performs an approximate inverse operation relative to the operation of the encoder pre-mixer.

[0036] FIG. 2 is a block diagram of audio codec system 100 that includes audio encoder 104 and audio decoder 106, encoder pre-mixer 102 and decoder post-mixer 108. Audio encoder 104 and audio decoder 106 form a multi-channel audio signal 109 (X') that is a facsimile of multi-channel audio signal 101 (X). Preferably, the data size of encoded representation 105 is minimized whilst minimizing the difference between multi-channel audio signal 101 and multi-channel audio signal 109. Furthermore, the difference between multi-channel audio signal 101 and multi-channel audio signal 109 may be measured according to similarity as perceived by a human listener.

[0037] The measure of human-perceived similarity between multi-channel audio signal 101 and multi-channel audio signal 109 is based on a reference playback method (that is, the assumed default means by which the audio channels of audio signals 101, 109 are presented as an auditory experience to the listener). The efficiency of multi-channel audio encoder 104 and
 5 multi-channel audio decoder 106 may be defined in terms of the data rate (measured in bits per second) of encoded representation 105 that provides a multi-channel audio signal 109 that will be judged by a listener to match multi-channel audio signal 101 with a particular perceived quality level.

[0038] Referring to FIG. 2, input multi-channel audio signal 101 is mixed according to
 10 encoder pre-mixer 102 (R) to produce output multi-channel audio signal 103 (Z) which is processed by simple audio encoder 104 to produce encoded representation 105, which may be stored and/or transmitted 110 to simple audio decoder 106, which produces multi-channel audio signal 107 (Z'). Multi-channel audio signal 107 is processed by decoder post-mixer 108 (R') to produce decoded multi-channel audio signal 109. Encoder pre-mixer 102 provides
 15 metadata 112 (Q) that includes necessary information to determine a behavior of decoder post-mixer 108. Metadata 112 may be stored and/or transmitted 110 with encoded representation 105. Measurement of the efficiency of multi-channel audio encoder 104 and multi-channel audio decoder 106 may include the size of the metadata 112 (commonly measured in bits per second), as will be appreciated by those skilled in the art.

[0039] Multi-channel audio signal 101 may be composed of N audio channels wherein
 20 significant correlations may exist between some pairs of channels, and wherein no single channel may be considered to be a dominant channel. That is, multi-channel audio signal 101 may not possess the attributes DD1 and DD2, and hence multi-channel audio signal 101 might not be a suitable signal for encoding and decoding using simple audio encoder 104 and decoder
 25 106, respectively.

[0040] Preferably, encoder pre-mixer 102 is adapted to process input multi-channel audio signal 101 to produce output multi-channel audio signal 103, where output multi-channel audio signal 103 possesses attributes DD1 and DD2. Given input multi-channel audio signal X composed of N channels:

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_N(t) \end{pmatrix} \quad [1]$$

the output multi-channel audio signal Z is computed as:

$$Z(t) = \begin{pmatrix} Z_1(t) \\ Z_2(t) \\ \vdots \\ Z_N(t) \end{pmatrix} \quad [2]$$

$$= \mathbf{R}(t) \times X(t). \quad [3]$$

[0041] The coefficients of encoder pre-mixer matrix $\mathbf{R}$ may vary over time, and $\mathbf{R}$ may thus be considered to be a function of time. The values of the elements of $\mathbf{R}$ may be computed at regular intervals (e.g., where the interval may be 20 ms, or a value between 1 ms and 100 ms) or at irregular intervals. When the values of the elements of $\mathbf{R}$ are changed, the change may be smoothly interpolated. In the following discussion, references to $\mathbf{R}$ should be treated as references to a time-varying encoder pre-mixer $\mathbf{R}(t)$ and references to $\mathbf{R}'$ should be treated as references to a time-varying decoder pre-mixer $\mathbf{R}'(t)$.

[0042] In an embodiment, encoder pre-mixer 102 may make use of mixing coefficients, $\mathbf{R}_b(t)$ for processing the components of the audio signals in a band b , where $1 \leq b \leq B$. FIG. 4 illustrates an arrangement of processing elements 150 whereby multi-channel audio signal 151 (X) is split by filterbank 152 into B sub-band signals, $X^{[1]}(t), X^{[2]}(t), \dots, X^{[B]}(t)$, with each sub-band signal (for example 153 ($X^{[1]}(t)$)) is processed by a mixing matrix (for example 154 ($\mathbf{R}_1$)) to produce a remixed subband signal (for example 155 ($Z^{[1]}(t)$)). Remixed sub-band signals, $Z^{[1]}(t), Z^{[2]}(t), \dots, Z^{[B]}(t)$, are recombined by combiner 156 to form multi-channel audio signal 157 (Z).

[0043] For the purpose of the following discussion, references to the matrix $\mathbf{R}(t)$ may be interpreted as references to $\mathbf{R}_b(t)$, where b refers to a subband. It will be appreciated that the discussion that follows may be applied to signals that are processed in subbands, or to signals that are processed without subband treatment. It will be appreciated by those skilled in the art that many methods may be used to process audio signals according to sub-bands, and the discussion of the matrix $\mathbf{R}$ will apply to those methods.

[0044] Referring to FIG. 2, $\mathbf{R}$ mixes the channels of multi-channel audio signal 101 to produce multi-channel audio signal 103 that possesses the attributes, DD1 and DD2, as described above, thus enabling encoder 106 to achieve improved data efficiency. Decoder pre-mixer 108 ($\mathbf{R}'$) provides a mixing operation that is the inverse of mixer $\mathbf{R}$, such that:

$$X'(t) = \mathbf{R}'(t) \times Z'(t) \quad . \quad [4]$$

[0045] FIG. 3 is a block diagram of an arrangement 200 of two mixing operations intended to implement the function of encoder pre-mixer 102 ($\mathbf{R}$) of FIG. 2 or encoder pre-mixer $\mathbf{R}_b$ of FIG. 4. N -channel multi-channel input signal 201 (X) is mixed by mixing matrix 202 ($\mathbf{M}$) to produce the N -channel intermediate signal 203 (Y), which is then processed by mixer 204 ($\mathbf{P}$) to produce the N -channel signal 205 (Z). The signals 201 (X) and 205 (Z) in FIG. 3 are intended to correspond respectively with input signal 101(X) and 103 (Z) in FIG. 2, or to sub-band signals 153 ($X_b(t)$) and 155 ($Z_b(t)$) in FIG. 4.

[0046] Analysis block 210 (A) takes input from signal 201, and computes the coefficients 212 to be used to adapt the operation of the mixer 204. Analysis block 210 also produces the metadata 211 (Q), corresponding to the metadata 112 of FIG. 2, which will be provided to the decoder, as 113 (Q), to be used by decoder post-mixer 108.

[0047] It will be appreciated from the arrangement of the mixers 202 and 204 in FIG. 4, that the matrix $\mathbf{R}$ will be:

$$R(t) = \mathbf{P}(t) \times \mathbf{M} \quad [5]$$

wherein the matrix $\mathbf{P}(t)$ may vary with time.

[0048] Hence:

$$\begin{aligned} Y(t) &= \mathbf{M} \times X(t) \\ Z(t) &= \mathbf{P}(t) \times Y(t) \\ &= \mathbf{P}(t) \times \mathbf{M} \times X(t) \\ &= \mathbf{R}(t) \times X(t) \end{aligned} \quad [6]\text{--}[9]$$

[0049] The matrix $\mathbf{M}$ is adapted to ensure that the intermediate signal 203 (Y) possesses attribute DD1. That is the N -channel signal 203 (Y) contains one channel that may be considered to be a dominant channel. Without loss of generality, the matrix $\mathbf{M}$ is adapted to ensure that the first channel, $Y_1(t)$ is a dominant channel. Hereinafter, when the first channel of a multi-channel signal is a dominant channel, this first channel will be referred to as a primary channel. The primary channel may also be referred to as an “eigen channel” in some contexts.

[0050] The $[N \times N]$ matrix $\mathbf{M}$ may be determined from the $[N \times N]$ expected covariance matrix $\mathbf{Cov}$ of the N -channel input signal, $X(t)$:

$$\begin{aligned}
\mathbf{Cov} &= E(X(t) \times X(t)^H) \\
&= \begin{pmatrix} E(X_1(t)\overline{X_1(t)}) & E(X_1(t)\overline{X_2(t)}) & \cdots & E(X_1(t)\overline{X_N(t)}) \\ E(X_2(t)\overline{X_1(t)}) & E(X_2(t)\overline{X_2(t)}) & \cdots & E(X_2(t)\overline{X_N(t)}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_N(t)\overline{X_1(t)}) & E(X_N(t)\overline{X_2(t)}) & \cdots & E(X_N(t)\overline{X_N(t)}) \end{pmatrix} \\
\mathbf{Cov} &= E(X(t) \times X(t)^H) \\
&= \begin{pmatrix} E(X_1(t)\overline{X_1(t)}) & E(X_1(t)\overline{X_2(t)}) & \cdots & E(X_1(t)\overline{X_N(t)}) \\ E(X_2(t)\overline{X_1(t)}) & E(X_2(t)\overline{X_2(t)}) & \cdots & E(X_2(t)\overline{X_N(t)}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_N(t)\overline{X_1(t)}) & E(X_N(t)\overline{X_2(t)}) & \cdots & E(X_N(t)\overline{X_N(t)}) \end{pmatrix} \quad [10]\text{--}[11]
\end{aligned}$$

where the $X(t)^H$ operation indicates the Hermitian Transpose of the N -length column vector $X(t)$, and the $E()$ operation indicates the expected value of a variable quantity.

[0051] The expected values, as used in Equation [10], may be estimated based on the assumed characteristics of typical input multi-channel audio signals, or they may be estimated by statistical analysis of a set of typical input multi-channel audio signals.

[0052] The covariance matrix, $\mathbf{Cov}$, may be factored according to eigen-analysis, as will be familiar to those skilled in the art:

$$\mathbf{Cov} = \mathbf{V} \times \mathbf{D} \times \mathbf{V}^H \quad [12]$$

where the matrix $\mathbf{V}$ is a unitary matrix and the matrix $\mathbf{D}$ is a diagonal matrix with the diagonal elements being non-negative real values sorted in descending order.

[0053] The matrix $\mathbf{M}$ may be chosen to be:

$$\mathbf{M} = \mathbf{V}^H \quad [13]$$

[0054] It will be appreciated by those skilled in the art that the covariance matrix, $\mathbf{Cov}$, will be dependent on the panning methods used to form the original input signal $X(t)$, as well as the typical use of the panning methods as used by the creators of typical signals.

[0055] By way of example, when the original input signal is a 2-channel stereo signal intended for playback on stereo speakers, the typical panning rules used by content creators will result in some audio objects being panned to the first channel (in this context, this is often referred to as the Left channel), some audio objects being panned to the second channel (in this context, this is often referred to as the Right channel), and some objects being panned simultaneously to both channels. In this case, the covariance matrix may be similar to:

$$\text{for L/R stereo: } \mathbf{Cov} = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix} \quad [14]$$

and according to Equations [12] and [13]:

$$\text{for L/R stereo: } \mathbf{M} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \quad [15]$$

[0056] The matrix $\mathbf{M}$ in Equation [15] will be familiar to those skilled in the art as a mixing matrix suitable for converting the original input audio signal X in L/R stereo format to an intermediate signal Z that will be in Mid/Side format. It will also be appreciated by those skilled in the art that the first channel of Z (often referred to as the Mid signal in this case) is a dominant audio signal (the primary channel), having the property that most audio elements in a stereo mix will be present in the Mid signal.

[0057] By way of an alternative example, when the original input signal is a 5-channel surround signal intended for playback on a common arrangement of five speakers, the typical panning rules used by content creators will result in some audio objects being panned to the one of the five channels, and some objects being panned simultaneously to two or more channels. In this case, the covariance matrix may be similar to:

$$\text{for 5-channels: } \mathbf{Cov} = \begin{pmatrix} 1.500 & 0.595 & 1.155 & 1.155 & 0.595 \\ 0.595 & 1.500 & 1.155 & 0.595 & 1.155 \\ 1.155 & 1.155 & 1.500 & 0.595 & 0.595 \\ 1.155 & 0.595 & 0.595 & 1.500 & 1.155 \\ 0.595 & 1.155 & 0.595 & 1.155 & 1.500 \end{pmatrix}, \quad [16]$$

and according to equations [12] and [13]:

$$\text{for 5-channels: } \mathbf{M} = \begin{pmatrix} 0.447 & 0.447 & 0.447 & 0.447 & 0.447 \\ -0.195 & -0.195 & -0.632 & 0.512 & 0.512 \\ 0.602 & -0.602 & 0.000 & 0.372 & -0.372 \\ -0.512 & -0.512 & 0.632 & 0.195 & 0.195 \\ -0.372 & 0.372 & 0.000 & 0.602 & -0.602 \end{pmatrix}. \quad [17]$$

[0058] It will be appreciated that the top row of matrix $\mathbf{M}$ of Equation [17] is made up of similar (or identical) positive values. This means that, according to Equation [6], the first channel of the intermediate signal $Y(t)$ will be formed by the sum of the five channels of the original input audio signal, $X(t)$, and this ensures that all sonic elements that are panned in the

original input audio signal will be present in $Y_1(t)$ (the first channel of the N -channel signal $Y(t)$). Hence, this choice of the matrix $\mathbf{M}$ ensures that the intermediate signal Y possesses the attribute DD1 ($Y_1(t)$ is a primary channel).

[0059] In a further alternative example, when the input multi-channel audio signal, $X(t)$, already contains a dominant channel (and, without loss of generality, it is assumed the first channel, $X_1(t)$ is dominant), the matrix $\mathbf{M}$ may be an $[N \times N]$ identity matrix. In a more specific example of an input multi-channel audio signal with a dominant/primary first channel, the input multi-channel audio signal may represent an acoustic scene encoded in an Ambisonic format (a means for encoding acoustic scenes that will be familiar to those skilled in the art).

[0060] The matrix 212 ($\mathbf{P}(t)$) is computed by the analysis block 210 ($\mathbf{A}$) in FIG. 4, at time t , according to the following process:

1. Determine the covariance of the intermediate signal $Y(t)$ at time t . An example of a method for computing the covariance is:

$$Cov_Y(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} Y(t) \times Y(t)^H \quad [18]$$

[0061] Alternatively, the covariance of the intermediate signal $Y(t)$ may be computed from the covariance of the input multi-channel audio signal $X(t)$, as:

$$Cov_Y(t) = \mathbf{M} \times Cov_X(t) \times \mathbf{M}^H, \quad [19]$$

where

$$Cov_X(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} X(t) \times X(t)^H. \quad [20]$$

2. From the $[L \times L]$ covariance matrix, $Cov_Y(t)$, extract the scalar quantity $w = [Cov_Y(t)]_{1,1}$, the $[N \times 1]$ column vector $v = [Cov_Y(t)]_{2..L,1}$ and the $[N \times N]$ matrix $E = [Cov_Y(t)]_{2..L,2..L}$, where $N = L - 1$, and:

$$Cov_Y(t) = \begin{pmatrix} w & v^H \\ v & E \end{pmatrix}. \quad [21]$$

3. Determine the quantities α , β and the $[N \times 1]$ vector of mixing coefficients u :

$$\alpha = |v|_2 = \sqrt{\sum_{n=1}^N v_n^2} \quad [22]$$

$$u = \frac{1}{\alpha} v \quad [23]$$

$$\beta = u^H \times E \times u \quad [24]$$

[0062] 4. Given the quantities w , α and β , solve Equation [25], to determine the input mixture strength coefficient h and the prediction mixture strength coefficient g :

$$\beta h^2 g + 2\alpha h g - \beta h - \alpha + gw = 0 \quad [25]$$

where the solutions to this equation will also satisfy a pre-prediction constraint equation. One example of a pre-prediction constraint equation is:

$$\text{PPC1: } h = fg, \quad [26]$$

where f is a pre-determined constant value satisfying $0 < f \leq 1$.

[0063] When the pre-prediction constraint PPC1 is used, Equation [25] can be modified to be:

$$\beta f^2 g^3 + 2\alpha f g^2 - \beta f g - \alpha + gw = 0 \quad [27]$$

and Equation [27] can be solved for the largest real value of g , and hence the value of h may be determined using Equation [26].

25

5. Form the $[L \times L]$ matrix $\mathbf{Q}$ as:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \dots & 0 \\ u & \vdots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \quad [28]$$

30

6. Form the $[L \times L]$ matrix $\mathbf{P}(t)$ as:

$$\mathbf{P}(t) = (\mathbf{I}_L - g\mathbf{Q}) \times (\mathbf{I}_L + h\mathbf{Q}^H) \quad [29]$$

where $\mathbf{I}_L$ is the $[L \times L]$ identity matrix.

5 [0064] The metadata 211 (Q) in FIG. 4 may convey information that will allow the unit-vector u and the coefficients g and h to be determined by the decoder post-mixer 113 of FIG. 2.

[0065] The solution for g of Equation [27] may be approximated by choosing an initial estimate $g_1 = 1$ and iterating (according to Newton's method, as is known in the art) a number
10 of times:

$$g_{k+1} = g_k - \frac{f^2 g_k^3 + 2\alpha f g_k^2 - \beta f g_k - \alpha + g_k w}{3f^2 g_k^2 + 4\alpha f g_k - \beta f + w}, \quad [30]$$

such that a reasonable approximation for the solution may be found from $g = g_5$. It will be appreciated that other methods are known in the art for finding approximate solutions to the
15 cubic Equation [27].

[0066] According to an alternative embodiment, the $[L \times L]$ matrix $\mathbf{P}(t)$ may be determined, at time t , by determining a $[N \times 1]$ vector u indicative of the correlation between the primary channel of the intermediate signal $Y(t)$ and the remaining N non-primary channels, and determining the input mixture strength coefficient h and the prediction mixture
20 strength coefficient g to form $\mathbf{P}(t)$ according to Equation [28], such that the signal $Z(t) = \mathbf{P}(t) \times Y(t)$ will possess the attributes DD1 and DD2.

[0067] The determination of coefficients g and h may be governed by a pre-prediction constraint equation. An example of a pre-prediction constraint equation is given (PPC1) in Equation [26]. A preferred choice for the coefficient f may be $f = 0.5$, but values of f in the
25 range $0.2 \leq f \leq 1$ may be appropriate for use.

[0068] In an alternative embodiment, the following pre-prediction constraints may be used:

$$\text{PPC2: } g = \begin{cases} \frac{\alpha}{w} & \text{when: } \frac{\alpha}{w} < c \\ c & \text{otherwise} \end{cases} \quad [31]$$

30 where c is a pre-determined constant. A typical value may be $c = 1$, but values of c may be chosen in the range $0.25 \leq c \leq 4$.

[0069] According to the constraint PPC2 in Equation [31], the solution to Equation [25] is:

$$\begin{aligned} \text{when: } \frac{\alpha}{w} < c \\ g &= \frac{\alpha}{w}, \\ h &= 0 \end{aligned} \quad [32]\text{--}[33]$$

otherwise:

$$\begin{aligned} g &= c \\ h &= \frac{\beta - 2c\alpha + \sqrt{\beta^2 + 4\alpha^2 c^2 - 4c^2 \beta w}}{2c\beta}. \end{aligned} \quad [34]\text{--}[35]$$

[0070] FIG. 5 is a block diagram of a prediction mixer 300, according to some embodiments. The matrix terms, $(\mathbf{I}_L - g\mathbf{Q})$ and $(\mathbf{I}_L + h\mathbf{Q}^H)$ of Equation [29] may be implemented by prediction mixer 300, wherein, in this example, the signal $Y(t)$ is composed of 4 channels ($L = 4$), the first channel 301 (Y_1) is a primary channel and the remaining 3 non-
 10 primary channels 302 (e.g., Y_2, Y_3, Y_4) are scaled according to the three input gains 312 (H_2, H_3 and H_4) to form the scaled input signal components (e.g., 304). The scaled input signal components are summed 305 with the primary input channel 301 (Y_1) to form the primary output 306 (Z_1). Primary output 306 (Z_1) is scaled by the three prediction gains 313 (G_2, G_3 and G_4) to form three prediction signals (e.g., 311). Each prediction signal is subtracted (e.g.
 15 308 and 309) from the respective input (e.g., Y_2 302) to form the respective non-dominant output 310 (Z_2).

[0071] The three input gains 312 (H_2, H_3 and H_4) may be determined from the mixing coefficients u (determined as per Equation [23]) and the input mixture strength coefficient h (as per the solution to Equation [25]), where:

$$\begin{pmatrix} H_2 \\ H_3 \\ H_4 \end{pmatrix} = hu. \quad [36]$$

[0072] The three prediction gains 313 (G_2, G_3 and G_4) may be determined from the mixing coefficients u (determined as per Equation [23]) and the prediction mixture strength coefficient g (as per the solution to Equation [25]), where:

$$\begin{pmatrix} G_2 \\ G_3 \\ G_4 \end{pmatrix} = gu. \quad [37]$$

[0073] It will be appreciated, by those skilled in the art, that the arrangement of linear matrix operations M 202 and P 204 of FIG. 4 may be implemented using a single matrix $R = P \times M$.

[0074] It will be appreciated, by those skilled in the art, that the decoder matrix R' of FIG. 2 may be formed from the matrices M' , the inverse of M and P' (the inverse of P):

$$R'(t) = M' \times P'(t), \quad [38]$$

and M' may be pre-computed (not varying as a function of time) and P' may be formed by the method:

$$P' = (I_L - hQ^H) \times (I_L + gQ). \quad [39]$$

[0075] FIG. 6 shows an arrangement 400 of processing elements that implement a decoder post-mixer, 108 in FIG. 2. The metadata 402 (Q) provides information to the inverse-prediction determination block 403 (B) which computes the coefficients necessary to determine the operation of inverse-predictor 405 (P'). The signal 401 (Z') is processed by inverse-predictor 405 (P') to produce the intermediate signal 406 (Y'), which is then processed by matrix 407 (M') to produce the output signal 408 X' .

Example Process

[0076] FIG. 7 is a flow diagram of a process 700 of adaptive downmixing of audio signals with improved continuity, according to some embodiments. Process 700 can be implemented by, for example, system 800 shown in FIG. 8.

[0077] Process 700 includes the steps of: receiving an input multi-channel audio signal comprising a primary input audio channel and L non-primary input audio channels (701); determining a set of L input gains, where L is a positive integer greater than one (702); for each of the L non-primary input audio channels and L input gains, forming a respective scaled non-primary input audio channel from the respective non-primary input audio channel scaled according to the input gain (703); forming a primary output audio channel from the sum of the primary input audio channel and the scaled non-primary input audio channels (704); determining a set of L prediction gains for each of the L prediction gains (705), forming a prediction channel from the primary output audio channel scaled according to the prediction gain (706); forming L non-primary output audio channels from the difference of the respective

non-primary input audio channel and the respective prediction signal (707); forming an output multi-channel audio signal from the primary output audio channel and the L non-primary output audio channels (708); encoding the output multi-channel audio signal (709); and transmitting or storing the encoded output multi-channel audio signal (710). Each of these steps are described more fully in reference to FIGS. 1-6.

Example System Architecture

[0078] FIG. 8 shows a block diagram of an example system 800 for implementing the features and processes described in reference to FIGS. 1-7, according to an embodiment. System 800 includes any devices that are capable of playing audio, including but not limited to: smart phones, tablet computers, wearable computers, vehicle computers, game consoles, surround systems, kiosks.

[0079] As shown, the system 800 includes a central processing unit (CPU) 801 which is capable of performing various processes in accordance with a program stored in, for example, a read only memory (ROM) 802 or a program loaded from, for example, a storage unit 808 to a random access memory (RAM) 803. In the RAM 803, the data required when the CPU 801 performs the various processes is also stored, as required. The CPU 801, the ROM 802 and the RAM 803 are connected to one another via a bus 809. An input/output (I/O) interface 805 is also connected to the bus 804.

[0080] The following components are connected to the I/O interface 805: an input unit 806, that may include a keyboard, a mouse, or the like; an output unit 807 that may include a display such as a liquid crystal display (LCD) and one or more speakers; the storage unit 808 including a hard disk, or another suitable storage device; and a communication unit 809 including a network interface card such as a network card (e.g., wired or wireless).

[0081] In some implementations, the input unit 806 includes one or more microphones in different positions (depending on the host device) enabling capture of audio signals in various formats (e.g., mono, stereo, spatial, immersive, and other suitable formats).

[0082] In some implementations, the output unit 807 include systems with various number of speakers. As illustrated in FIG. 8, the output unit 807 (depending on the capabilities of the host device) can render audio signals in various formats (e.g., mono, stereo, immersive, binaural, and other suitable formats).

[0083] The communication unit 809 is configured to communicate with other devices (e.g., via a network). A drive 810 is also connected to the I/O interface 805, as required. A

removable medium 811, such as a magnetic disk, an optical disk, a magneto-optical disk, a flash drive or another suitable removable medium is mounted on the drive 810, so that a computer program read therefrom is installed into the storage unit 808, as required. A person skilled in the art would understand that although the system 800 is described as including the
5 above-described components, in real applications, it is possible to add, remove, and/or replace some of these components and all these modifications or alteration all fall within the scope of the present disclosure.

[0084] Aspects of the systems described herein may be implemented in an appropriate computer-based sound processing network environment for processing digital or digitized
10 audio files. Portions of the adaptive audio system may include one or more networks that comprise any desired number of individual machines, including one or more routers (not shown) that serve to buffer and route the data transmitted among the computers. Such a network may be built on various different network protocols, and may be the Internet, a Wide Area Network (WAN), a Local Area Network (LAN), or any combination thereof.

[0085] In accordance with example embodiments of the present disclosure, the processes described above may be implemented as computer software programs or on a computer-readable storage medium. For example, embodiments of the present disclosure include a computer program product including a computer program tangibly embodied on a machine readable medium, the computer program including program code for performing
20 methods. In such embodiments, the computer program may be downloaded and mounted from the network via the communication unit 809, and/or installed from the removable medium 811, as shown in FIG. 8.

[0086] Generally, various example embodiments of the present disclosure may be implemented in hardware or special purpose circuits (e.g., control circuitry), software, logic or
25 any combination thereof. For example, the units discussed above can be executed by control circuitry (e.g., a CPU in combination with other components of FIG. 8), thus, the control circuitry may be performing the actions described in this disclosure. Some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device (e.g., control
30 circuitry). While various aspects of the example embodiments of the present disclosure are illustrated and described as block diagrams, flowcharts, or using some other pictorial representation, it will be appreciated that the blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software,

firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[0087] Additionally, various blocks shown in the flowcharts may be viewed as method steps, and/or as operations that result from operation of computer program code, and/or as a plurality of coupled logic circuit elements constructed to carry out the associated function(s). For example, embodiments of the present disclosure include a computer program product including a computer program tangibly embodied on a machine readable medium, the computer program containing program codes configured to carry out the methods as described above.

[0088] In the context of the disclosure, a machine readable medium may be any tangible medium that may contain, or store a program for use by or in connection with an instruction execution system, apparatus, or device. The machine readable medium may be a machine readable signal medium or a machine readable storage medium. A machine readable medium may be non-transitory and may include but not limited to an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples of the machine readable storage medium would include an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[0089] Computer program code for carrying out methods of the present disclosure may be written in any combination of one or more programming languages. These computer program codes may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus that has control circuitry, such that the program codes, when executed by the processor of the computer or other programmable data processing apparatus, cause the functions/operations specified in the flowcharts and/or block diagrams to be implemented. The program code may execute entirely on a computer, partly on the computer, as a stand-alone software package, partly on the computer and partly on a remote computer or entirely on the remote computer or server or distributed over one or more remote computers and/or servers.

[0090] While this document contains many specific implementation details, these should not be construed as limitations on the scope of what may be claimed, but rather as descriptions of features that may be specific to particular embodiments. Certain features that are described in this specification in the context of separate embodiments can also be

implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable sub combination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or
5 more features from a claimed combination can, in some cases, be excised from the combination, and the claimed combination may be directed to a sub combination or variation of a sub combination. Logic flows depicted in the figures do not require the particular order shown, or sequential order, to achieve desirable results. In addition, other steps may be provided, or steps may be eliminated, from the described flows, and other components may be
10 added to, or removed from, the described systems. Accordingly, other implementations are within the scope of the following claims.

What is claimed is:

CLAIMS

- 5 1. An audio encoding method comprising:
 - receiving, with at least one processor, an input multi-channel audio signal comprising a primary input audio channel and L non-primary input audio channels;
 - determining, with the at least one processor, a set of L input gains, where L is a positive integer greater than one, wherein the set of L input gains are determined by scaling a set of L
 - 10 mixing coefficients by an input mixture strength coefficient;
 - for each of the L non-primary input audio channels and L input gains, forming a respective scaled non-primary input audio channel from the respective non-primary input audio channel scaled according to the input gain;
 - forming a primary output audio channel from a sum of the primary input audio channel
 - 15 and the scaled non-primary input audio channels;
 - determining, with the at least one processor, a set of L prediction gains, wherein the set of L prediction gains is determined by scaling the set of L mixing coefficients by a prediction mixture strength coefficient;
 - for each of the L prediction gains, forming, with the at least one processor, a prediction
 - 20 channel from the primary output audio channel scaled according to the prediction gain;
 - forming, with the at least one processor, L non-primary output audio channels from a difference of the respective non-primary input audio channel and the respective prediction channel;
 - forming, with the at least one processor, an output multi-channel audio signal from the
 - 25 primary output audio channel and the L non-primary output audio channels;
 - encoding, with an audio encoder, the output multi-channel audio signal; and
 - transmitting or storing, with the at least one processor, the encoded output multi-channel audio signal.
- 30 2. A system comprising:
 - one or more computer processors; and
 - a non-transitory computer-readable medium storing instructions that, when executed by

the one or more computer processors, cause the one or more computer processors to perform operations of claim 1.

3. A non-transitory computer-readable medium storing instructions that, when
5 executed by one or more computer processors, cause the one or more computer processors to perform operations of claim 1.