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

(11) **EP 3 745 744 A2**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
02.12.2020 Bulletin 2020/49

(51) Int Cl.:
H04S 3/00 *(2006.01)*

(21) Application number: **20176223.4**

(22) Date of filing: **25.05.2020**

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

(30) Priority: **29.05.2019 GB 201907601**

(71) Applicant: **Nokia Technologies Oy**
02610 Espoo (FI)

(72) Inventors:
- **VÄÄNÄNEN, Riitta**
  00250 Helsinki (FI)
- **VESA, Sampo**
  00510 Helsinki (FI)
- **LAITINEN, Mikko-Ville**
  02130 Espoo (FI)
- **VIROLAINEN, Jussi**
  02210 Espoo (FI)

(74) Representative: **Nokia EPO representatives**
Nokia Technologies Oy
Karakaari 7
02610 Espoo (FI)

(54) **AUDIO PROCESSING**

(57) An apparatus for processing an input audio signal comprising multiple channels, the apparatus comprising: means for deriving, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion;
cross-channel mixing means for cross-channel mixing of a plurality of input channels; means for directing the second signal component to the cross-channel mixing means for cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component;
bypass means for enabling the first signal component to bypass the cross-channel mixing means; and
means for combining the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

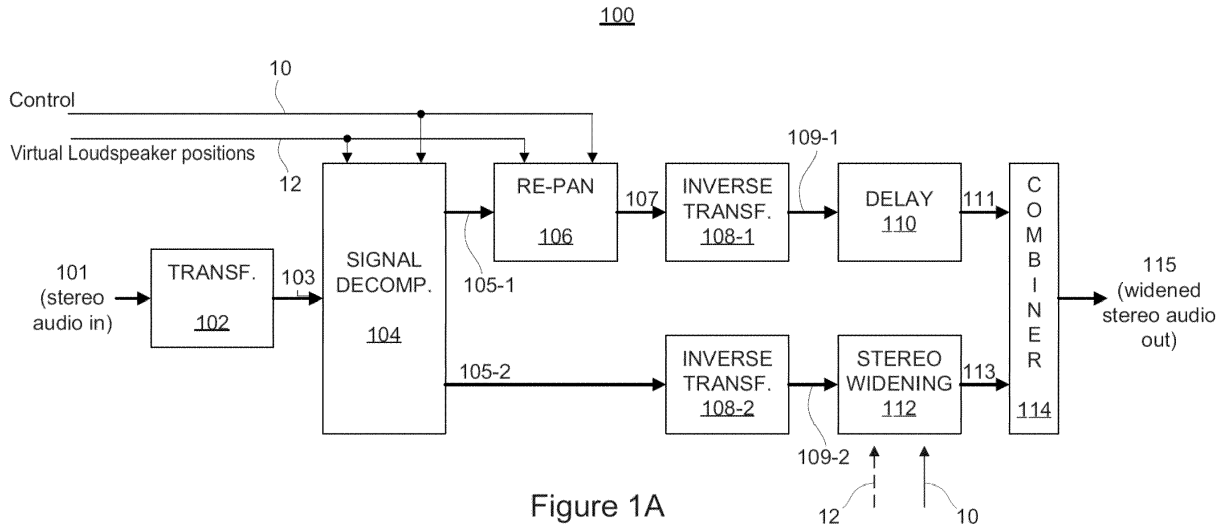

Figure 1A

EP 3 745 744 A2

(Cont. next page)

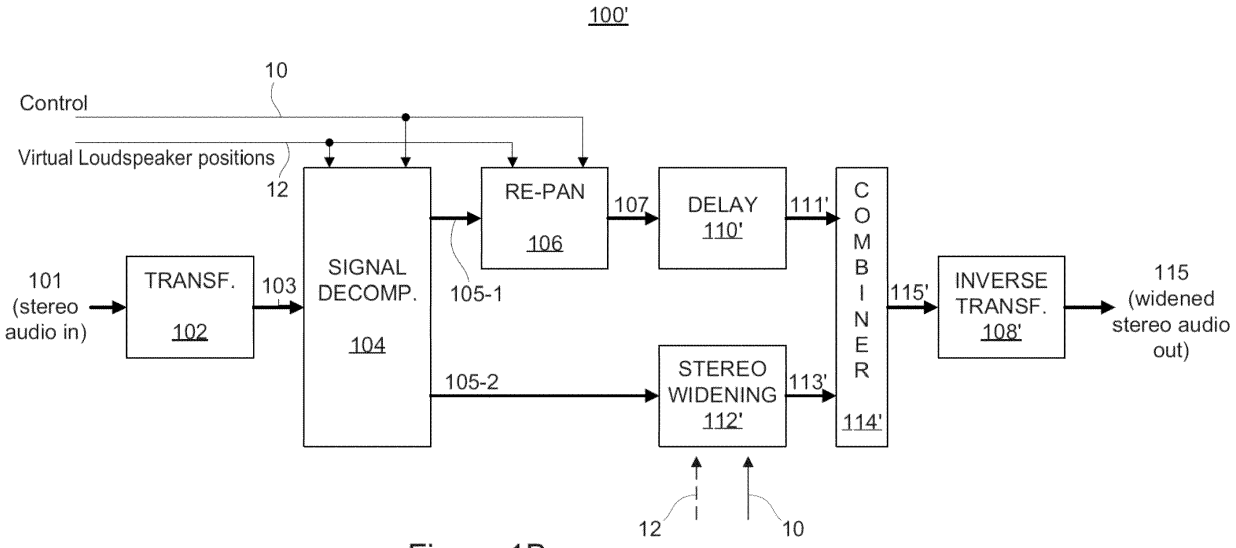
100'
10
Control
Virtual Loudspeaker positions
12
101
(stereo
audio in)
TRANSF.
102
103
SIGNAL
DECOMP.
104
105-1
105-2
RE-PAN
106
107
DELAY
110'
111'
STEREO
WIDENING
112'
113'
COMBINER
114'
115'
INVERSE
TRANSF.
108'
115
(widened
stereo audio
out)
12
10

Figure 1B

Description

TECHNICAL FIELD

[0001] The example and non-limiting embodiments of the present invention relate to processing of audio signals. In particular, various embodiments of the present invention relate to modification of a spatial image represented by a multi-channel audio signal, such as a two-channel stereo signal.

BACKGROUND

[0002] So-called stereo widening is a technique known in the art for enhancing the perceivable spatial audio image of a stereophonic audio signal when reproduced via audio output device. Such a technique aims at processing a stereophonic audio signal such that reproduced sound is not only perceived as originating from directions that are localized between the audio output devices but at least part of the sound field is perceived as if it originated from directions that are not localized between the audio output devices, thereby widening the perceivable width of spatial audio image from that conveyed in the stereophonic audio signal. Herein, we refer to such spatial audio image as a widened or enlarged spatial audio image.

[0003] While outlined above via references to a two-channel stereophonic audio signal, stereo widening may be applied to multi-channel audio signals that have more than two channels, such as 5.1-channel or 7.1-channel surround sound for playback via a pair of audio output devices. In some contexts, the term virtual surround is applied to refer to a processed audio signal that conveys a spatial audio image originally conveyed in a multi-channel surround audio signal. Hence, even though the term stereo widening is predominantly applied throughout this disclosure, this term should be construed broadly, encompassing a technique for processing the spatial audio image conveyed in a multi-channel audio signal (i.e. a two-channel stereophonic audio signal or a surround sound of more than two channels) to provide audio playback at widened spatial audio image.

[0004] For brevity and clarity of description, in this disclosure we use the term multi-channel audio signal to refer to audio signals that have two or more channels. Moreover, the term stereo signal is used to refer to a stereophonic audio signal and the term surround signal is used to refer to a multi-channel audio signal having more than two channels.

[0005] When applied to a stereo signal, stereo widening techniques known in the art typically involve adding a processed (e.g. filtered) version of a contralateral channel signal to each of the left and right channel signals of the stereo signal in order to derive an output stereo signal having a widened spatial audio image (referred to in the following as a widened stereo signal). In other words, a processed version of the right channel signal of the stereo signal is added to the left channel signal of the stereo signal to create the left channel of a widened stereo signal and a processed version of the left channel signal of the stereo signal is added to the right channel signal of the stereo signal to create the right channel of the widened stereo signal. Moreover, the procedure of deriving the widened stereo signal may further involve pre-filtering (or otherwise processing) each of the left and right channel signals of the stereo signal prior to adding the respective processed contralateral signals thereto in order to preserve desired frequency response in the widened stereo signal.

[0006] Along the lines described above, stereo widening readily generalizes into widening the spatial audio image of a multi-channel input audio signal, thereby deriving an output multi-channel audio signal having a widened spatial audio image (referred to in the following as a widened multi-channel signal). In this regard, the processing involves creating the left channel of the widened multi-channel audio signal as a sum of (first) filtered versions of channels of the multi-channel input audio signal and creating the right channel of the widened multi-channel audio signal as a sum of (second) filtered versions of channels of the multi-channel input audio signal. Herein, a dedicated predefined filter may be provided for each pair of an input channel (channels of the multi-channel input signal) and an output channel (left and right) . As an example in this regard, the left and right channel signals of the widened multi-channel signal $S_{out,left}$ and $S_{out,right}$, respectively, may be defined on basis of channels of a multi-channel audio signal S according to the equation (1):

$$\begin{aligned} S_{out,left}(b,n) &= \sum_i S(i,b,n) H_{left}(i,b) \\ S_{out,right}(b,n) &= \sum_i S(i,b,n) H_{right}(i,b) \end{aligned} \quad (1)$$

where $S(i, b, n)$ denotes frequency bin b in time frame n of channel i of the multi-channel signal S, $H_{left}(i, b)$ denotes a filter for filtering frequency bin b of channel i of the multi-channel signal S to create a respective channel component for creation of the left channel signal $S_{out,left}(b, n)$, and $H_{right}(i, b)$ denotes a filter for filtering frequency bin b of channel i of the multi-channel signal S to create a respective channel component for creation of the right channel signal $S_{out,right}(b, n)$.

[0007] A challenge involved in stereo widening is degraded timbre in the central part of the spatial audio image. In

many real-life stereo signals the central part of the spatial audio image includes perceptually important audio content, e.g. in case of music the voice of the vocalist is typically rendered in the center of the spatial audio image. A sound component that is in the center of the spatial audio image is rendered by reproducing the same signal in both channels of the stereo signal and hence via both audio output devices. When stereo widening is applied to such an input stereo signal (e.g. according to the equation (1) above), each channel of the resulting widened stereo signal involves outcome of two filtering operations carried out for the channels of the input stereo signal. This may result in a comb filtering effect, which may cause differences in the perceived timbre, which may be referred to as 'coloration' of the sound. Moreover, the comb filtering effect may further result in degradation of the engagement of the sound source.

[0008] In some circumstances, the audio output devices are part of a headphone apparatus that comprises a left audio output device that is worn at, over or in a left ear of a user and a right audio output device that is worn at, over or in a right ear of a user.

[0009] Normal playback of stereo audio via headphones may cause the sound to be perceived by a user inside the user's head. The stereo panning cues position the sound in between the ears, inside the head.

[0010] To address this loudspeaker virtualization methods are used to process the audio signals so that the perception to the user listening via headphones is similar to the perception to a user who is listening via loudspeakers. This can be achieved by filtering the audio signals using appropriate head-related transfer functions (HRTF) or binaural room impulse responses (BRIR).

SUMMARY

[0011] According to various, but not necessarily all, examples there is provided an apparatus for processing an input audio signal comprising multiple channels, the apparatus comprising: means for deriving, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion; cross-channel mixing means for cross-channel mixing of a plurality of input channels; means for directing the second signal component to the cross-channel mixing means for cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component; bypass means for enabling the first signal component to bypass the cross-channel mixing means; and means for combining the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

[0012] In some but not necessarily all examples, the cross-channel mixing means for cross-channel mixing of a plurality of input channels comprises means for applying head related transfer functions to each one of the plurality of input channels before mixing those channels to produce a modified second signal component comprising two output channels, wherein the head related transfer function applied to an input channel that is mixed to provide an output channel is dependent upon an identity of the input channel and an identify of the output channel.

[0013] In some but not necessarily all examples, the cross-channel mixing means for cross-channel mixing of a plurality of input channels comprises means for applying a headphone filter to each one of the plurality of input channels before mixing those channels to produce a modified second signal component comprising two output channels, wherein the headphone filter applied to an input channel that is mixed to provide an output channel is dependent upon an identity of the input channel and an identify of the output channel, wherein the headphone filter for an input channel mixes a direct version of the input channel with an ambient version of the input channel.

[0014] In some but not necessarily all examples, the relative gain of the direct version of the input channel compared to the ambient version of the input channel in a mix in the headphone filter is a user-controllable parameter.

[0015] In some but not necessarily all examples, the headphone filter for an input channel mixes a single-path direct version of the input channel with a multiple-path ambient version of the input channel; wherein a head related transfer function is used to form the single-path direct version of the input channel; wherein, an indirect path filter is used in combination with a head related transfer function for each path of the multiple paths, to form the multiple-path ambient version of the input channel. In some but not necessarily all examples, the indirect path filter comprises decorrelation means or reverberation means.

[0016] In some but not necessarily all examples, the cross-channel mixing means is configured to cause stereo-widening for headphone apparatus such that a width of a spatial audio image associated with the modified second signal component is greater than a width of a spatial audio image associated with the second signal component before cross-channel mixing of the second signal component.

[0017] In some but not necessarily all examples, the first portion is front and central relative to a user of the headphone apparatus, and the second portion is peripheral relative to the user of headphone apparatus and does not overlap the first portion.

[0018] In some but not necessarily all examples, the first and second portions are contiguous.

[0019] In some but not necessarily all examples, the bypass means enables components of the input audio signal that represent a sound source that is coherent between two stereo channels and is positioned to front and center, to bypass the cross-channel mixing means.

[0020] In some but not necessarily all examples, a control input controls one or more of:

control the first portion and/or the second portion;
control decomposition of input signal to first component and second component;
control relative gain of the first component and the second component;
control widening of the second component;
control ratio of direct to ambient gain during widening of second component;
control panning of first component;
control whether there is or is not panning of the first component;
control panning extent of first component; and
control energy-based temporal smoothing.

[0021] In some but not necessarily all examples, when the input audio signal comprises a same sound source that is repeated at different positions, and that is rendered at the headphone apparatus without interaural time difference and without frequency dependent interaural level differences, when the sound source of the input audio signal is positioned at a first position that is relatively front and central to a user of the headphone apparatus, then the sound source is rendered at the headphone apparatus with interaural time differences and with frequency dependent interaural level differences when the sound source of the input audio signal is repeated at a second position that is relatively peripheral and is not front and central to a user of the headphone apparatus.

[0022] In some but not necessarily all examples, there is provided a system comprising the apparatus and a headphone apparatus configured for receiving and rendering the output audio signal.

[0023] In some but not necessarily all examples, the apparatus is configured as a headphone apparatus for rendering the output audio signal.

[0024] According to various, but not necessarily all, examples there is provided a method for processing an input audio signal comprising a at least one input channel/multiple input channels, the method comprising:

deriving, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein
the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion;
cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component while enabling the first signal component to bypass cross-channel mixing; and
combining the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

[0025] According to various, but not necessarily all, examples there is provided an apparatus for processing an input audio signal comprising a at least one input channel/multiple input channels, the apparatus comprising at least one processor; and at least one memory including computer program code, which when executed by the at least one processor, causes the apparatus to:

derive, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein
the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion;
perform cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component while enabling the first signal component to bypass cross-channel mixing; and
combine the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

[0026] According to various, but not necessarily all, examples there is provided a computer program comprising computer readable program code configured to cause a computer to:
derive, based on an input audio signal, a first signal component, comprising at least one input channel, and a second

signal component, comprising multiple input channels, wherein the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion; perform cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component while enabling the first signal component to bypass cross-channel mixing.

[0027] According to various, but not necessarily all, examples there is provided an apparatus for processing an input audio signal comprising multiple channels to produce a two-channel output audio signal configured for rendering by headphone apparatus to produce a spatial audio image, the apparatus comprising:

means for processing an input audio signal comprising multiple channels to produce a two-channel output audio signal configured for rendering by headphone apparatus;
means for spatially processing the input audio signal to add at peripheral positions, but not at central positions, of the spatial audio image positionally-dependent interaural time differences measurable between coherent audio events in both of the channels of the output audio signal and frequency-dependent and positionally-dependent interaural level differences measurable between coherent audio events in both of the channels of the output audio signal.

[0028] In some but not necessarily all examples, the means for deriving the first and second signal components is arranged to

derive, on basis of the input audio signal, the first signal component that represents coherent sounds of the spatial audio image that reside within the first portion of the spatial audio image; and
derive, on basis of the input audio signal, the second signal component that represents coherent sounds of the spatial audio image that reside within the second portion of the spatial audio image and outside the first portion of the spatial audio image and non-coherent sounds of the spatial audio image.

[0029] In some but not necessarily all examples, the first portion of the spatial audio image comprises one or more angular ranges that define a set of sound arrival directions within the spatial audio image.

[0030] In some but not necessarily all examples, said one or more angular ranges comprise an angular range that defines a range of sound arrival directions centered around a front direction of the spatial audio image.

[0031] In some but not necessarily all examples, the means for deriving the first and second signal components comprises

a means for deriving, on basis of the input audio signal, for a plurality of frequency sub-bands, a respective coherence value that is descriptive of coherence between channels of the input audio signal in the respective frequency sub-band;
a means for deriving, on basis of estimated sound arrival directions in view of the first portion of the spatial audio image, for said plurality of frequency sub-bands, a respective directional coefficient that is indicative of a relationship between the estimated sound arrival direction and the first portion of the spatial audio image in the respective frequency sub-band;
a means for deriving, on basis of said coherence values and directional coefficients, for said plurality of frequency sub-bands, respective decomposition coefficients; and
a means for decomposing the input audio signal into the first and second signal components using said decomposition coefficients.

[0032] In some but not necessarily all examples, the means for deriving the directional coefficients is arranged to, for said plurality of frequency sub-bands,

set the directional coefficient for a frequency sub-band to a non-zero value in response to the estimated sound arrival direction for said frequency sub-band residing within the first portion of the spatial audio image, and
set the directional coefficient for a frequency sub-band to a zero value in response to the estimated sound arrival direction for said frequency sub-band residing within the second portion of the spatial audio image.

[0033] In some but not necessarily all examples, the means for determining the decomposition coefficients is arranged to derive, for said plurality of frequency sub-bands, the respective decomposition coefficient as the product of the coherence value and the directional coefficient derived for the respective frequency sub-band.

[0034] In some but not necessarily all examples, the means for decomposing the input audio signal is arranged to, for said plurality of frequency sub-bands,

derive the first signal component in each frequency sub-band as a product of the input audio signal in the respective frequency sub-band and a first scaling coefficient that increases with increasing value of the decomposition coefficient derived for the respective frequency sub-band; and

derive the second signal component in each frequency sub-band as a product of the input audio signal in the respective frequency sub-band and a second scaling coefficient that decreases with increasing value of the decomposition coefficient derived for the respective frequency sub-band.

[0035] In some but not necessarily all examples, the apparatus comprises a means for delaying the first signal component by a predefined time delay prior to combining the first signal component with the modified second signal component, thereby creating a delayed first signal component that is temporally aligned with the modified second signal component.

[0036] In some but not necessarily all examples, the apparatus comprises a means for modifying the first signal component prior to combining the first signal component with the modified second signal component, wherein the modification comprises generating, on basis of the first signal component, a modified first signal component wherein one or more sound sources represented by the first signal component are panned in the spatial audio image,

[0037] In some but not necessarily all examples, each of said the multiple input channels comprise two channels.

[0038] According to various, but not necessarily all, embodiments there is provided examples as claimed in the appended claims.

[0039] According to an example embodiment, a computer program is provided, the computer program comprising computer readable program code configured to cause performing at least a method according to the example embodiment described in the foregoing when said program code is executed on a computing apparatus.

[0040] The computer program according to an example embodiment may be embodied on a volatile or a non-volatile computer-readable record medium, for example as a computer program product comprising at least one computer readable non-transitory medium having program code stored thereon, the program which when executed by an apparatus cause the apparatus at least to perform the operations described hereinbefore for the computer program according to an example embodiment of the invention.

[0041] The exemplifying embodiments of the invention presented in this patent application are not to be interpreted to pose limitations to the applicability of the appended claims. The verb "to comprise" and its derivatives are used in this patent application as an open limitation that does not exclude the existence of also unrecited features. The features described hereinafter are mutually freely combinable unless explicitly stated otherwise.

[0042] Some features of the invention are set forth in the appended claims. Aspects of the invention, however, both as to its construction and its method of operation, together with additional objects and advantages thereof, will be best understood from the following description of some example embodiments when read in connection with the accompanying drawings.

DEFINITIONS

[0043] A headphone apparatus is an apparatus that has a left audio output device that is worn at, over or in a left ear of a user and a right audio output device that is worn at, over or in a right ear of a user. The audio heard in the left ear by the user is dependent upon audio output by the left audio output device and is not dependent upon audio output by the right audio output device. The audio heard in the right ear by the user is dependent upon audio output by the right audio output device and is not dependent upon audio output by the left audio output device. The headphone receives input signals wirelessly or over a wired connection. In some but not necessarily all examples, the headphone apparatus comprises acoustic isolators that isolate the ears of the user from external environmental sounds. In some examples, the headphone apparatus can comprise 'cans' that cover the user's ears and provide at least some acoustic isolation. In some examples, the headphone apparatus can comprise deformable 'buds' that fit snugly inside the user's ears and provide at least some acoustic isolation. Each audio output device comprises a transducer that converts a received electrical signal to an acoustic pressure wave or a vibration.

multi-channel audio signal: in this disclosure we use the term multi-channel audio signal to refer to audio signals that have two or more channels.

stereo signal: the term stereo signal is used to refer to a stereophonic audio signal.

surround sound signal: the term surround signal is used to refer to a multi-channel audio signal having more than two channels.

BRIEF DESCRIPTION OF FIGURES

[0044] The embodiments of the invention are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings, where

Figure 1A illustrates a block diagram of some elements of an audio processing system for headphones according to an example;
Figure 1B illustrates a block diagram of some elements of an audio processing system for headphones according to an example;
Figure 2 illustrates a block diagram of some elements of a device that be applied to implement the audio processing system for headphones according to an example;
Figure 3 illustrates a block diagram of some elements of a signal decomposer according to an example;
Figure 4 illustrates a block diagram of some elements of a re-panner for headphones according to an example;
Figure 5 illustrates a block diagram of some elements of a stereo widening processor for headphones according to an example;
Figure 6 illustrates a flow chart depicting a method for audio processing for headphones according to an example; and
Figure 7 illustrates a block diagram of some elements of an apparatus according to an example.

DESCRIPTION OF SOME EMBODIMENTS

[0045] In the following examples there is disclosed an apparatus 100, 100', 50 for processing an input audio signal 101 comprising multiple channels, the apparatus 100, 100', 50 comprising: means 104 for deriving, based on the input audio signal 101, a first signal component 105-1, comprising at least one input channel, and a second signal component 105-2, comprising multiple input channels, wherein the first signal component 105-1 is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal 101, and the second signal component 105-2 is dependent upon at least a second portion of the spatial audio image that is different to the first portion; cross-channel mixing means 112, 112' for cross-channel mixing of a plurality of input channels; means 104 for directing the second signal component 105-2 to the cross-channel mixing means 112, 112' for cross-channel mixing of at least some of the multiple input channels of the second signal component 105-2 to produce a modified second signal component 113, 113'; bypass means 104, 106 for enabling the first signal component 105-1 to bypass the cross-channel mixing means 112, 112'; and means 114, 114' for combining the first signal component 111, 111' and the modified second signal component 113, 113' into an output audio signal 115 comprising two output channels configured for rendering by headphone apparatus 20.

[0046] Figure 1A illustrates a block diagram of some components and/or entities of an audio processing system 100 that may serve as framework for various embodiments of the audio processing technique described in the present disclosure. The audio processing system 100 obtains a stereophonic audio signal as an input signal 101 and provides a stereophonic audio signal having at least partially widened spatial audio image as an output signal 115. The input signal 101 and the output signal 115 are referred to in the following as a stereo signal 101 and a widened stereo signal 115, respectively. In the following examples that pertain to the audio processing system 100, each of these signals is assumed to be a respective two-channel stereophonic audio signal unless explicitly stated otherwise. Moreover, also each of the intermediate audio signals derived on basis of the input signal 101 are likewise respective two-channel audio signals unless explicitly state otherwise.

[0047] Nevertheless, the audio processing system 100 readily generalizes into a one that enables processing of a spatial audio signal (i.e. a multi-channel audio signal with more than two channels, such as a 5.1-channel spatial audio signal or a 7.1-channel spatial audio signal), some aspects of which are also described in the examples provided in the following.

[0048] The audio processing system 100 may further receive a control input 10 and an indication 12 of target sound source (virtual loudspeaker) positions.

[0049] The audio processing system 100 according to the example illustrated in Figure 1A comprises a transform entity (or a transformer) 102 for converting the stereo audio signal 101 from time domain into a transform domain stereo signal 103, a signal decomposer 104 for deriving, based on the transform-domain stereo signal 103, a first signal component 105-1 that represents a focus portion of the spatial audio image and a second signal component 105-2 that represents a non-focus portion of the spatial audio image, a re-panner 106 for generating, on basis of the first signal component 105-1, a modified first signal component 107, where one or more sound sources represented in the focus portion of the spatial audio image are repositioned in dependence of the target configuration, an inverse transform entity 108-1 for converting the modified first signal component 107 from the transform domain to a time-domain modified first signal component 109-1, an inverse transform entity 108-2 for converting the second signal component 105-2 from the transform domain to a time-domain second signal component 109-2, a delay element 110 for delaying the modified first

signal component 109-1 by a predefined time delay, a stereo widening (for headphones) processor 112 for generating, on basis of the second signal component 109-2, a modified second signal component 113 where the width of a spatial audio image is extended from that of the second signal component 109-2, and a signal combiner 114 for combining the delayed first signal component 111 and the modified second signal component 113 into a widened stereo signal 115 that conveys a partially extended spatial audio image.

[0050] Figure 1B illustrates a block diagram of some components and/or entities of an audio processing system 100', which is a variation of the audio processing system 100 illustrated in Figure 1A. In the audio processing system 100', differences to the audio processing system 100 are that the inverse transform entities 108-1 and 108-2 are omitted, the delay element 100 is replaced with the optional delay element 110' for delaying the modified first signal component 107 into delayed modified first signal component 111', the stereo widening processor 112 is replaced with a stereo widening processor 112' for generating, on basis of the transform-domain second signal component 105-2, a modified (transform-domain) second signal component 113', and the signal combiner 114 is replaced with a signal combiner 114' for combining the delayed modified first signal component 111' and the modified second signal component 113' into a widened stereo signal 115' in the transform domain. Moreover, the audio processing system 100' comprises a transform entity 108' for converting the widened stereo signal 115' from the transform domain into a time-domain widened stereo signal 115. In case the optional delay element 110' is omitted, the signal combiner 114' receives the modified first signal component 107 (instead of the delayed version thereof) and operates to combine modified first signal component 107 with the modified second signal component 113' to create the transform-domain widened stereo signal 115'.

[0051] In the following, the audio processing technique described in the present disclosure is predominantly described via examples that pertain to the audio processing system 100 according to the example of Figure 1A and entities thereof, whereas the audio processing system 100' and entities thereof are separately described where applicable. In further examples, the audio processing system 100 or the audio processing system 100' may include further entities and/or some entities depicted in Figures 1A and 1B may be omitted or combined with other entities. In particular, Figures 1A and 1B, as well as the subsequent Figures 2 to 5 serve to illustrate logical components of a respective entity and hence do not impose structural limitations concerning implementation of the respective entity but, for example, respective hardware means, respective software means or a respective combination of hardware means and software means may be applied to implement any of the logical components of an entity separately from the other logical components of that entity, to implement any sub-combination of two or more logical components of an entity, or to implement all logical components of an entity in combination.

[0052] The audio processing system 100, 100' may be implemented by one or more computing devices and the resulting widened stereo signal 115 may be provided for playback via headphone apparatus. Typically, the audio processing system 100, 100' is implemented in a computing device of any type, e.g. a portable handheld device, a desktop computer, a server device, etc. Examples of portable handheld devices include a mobile phone, a media player device, a tablet computer, a laptop computer, etc. The computing device can also be used to play back the widened stereo signal 115 via headphone apparatus. In another example, the audio processing system 100, 100' is provided in the headphone apparatus and the playback of the widened stereo signal 115 is provided in the headphone apparatus. In a further example, a first part of the audio processing system 100, 100' is provided in a first device, whereas a second part of the audio processing system 100, 100' and the playback of the widened stereo signal 115 is provided in the headphone apparatus.

[0053] Figure 2 illustrates a block diagram of some components and/or entities of a portable handheld device 50 that implements the audio processing system 100 or the audio processing system 100'. For brevity and clarity of description, in the following description it is assumed that the elements of the audio processing system 100, 100' and the playback of the resulting widened stereo signal are provided in the device 50. The device 50 further comprises a memory device 52 for storing information, e.g. the stereo signal 101, and a communication interface 54 for communicating with other devices and possibly receiving the stereo signal 101 therefrom. The device 50, optionally, further comprises an audio preprocessor 56 that may be useable for preprocessing the stereo signal 101 read from the memory 52 or received via the communication interface 54 before providing it to the audio processing system 100, 100'. The audio preprocessor 56 may, for example, carry out decoding of an audio signal stored in an encoded format into a time domain stereo audio signal 101.

[0054] Still referring to Figure 2, the audio processing system 100, 100' may further receive the first control input 10 and indication 12 together with the stereo signal 101 from or via the audio preprocessor 56.

[0055] The control input 12 is used to control signal de-composition 104 and/or re-panning 106 and/or stereo-widening 112, 112'. More details are provided in the following description.

[0056] The indication 12 indicates the target sound source (virtual loudspeaker) positions. Effectively this means the positions of loudspeakers if the input audio signal would be reproduced by loudspeakers.

[0057] The virtual loudspeaker positions match typically with the loudspeaker format of input audio signals. For stereo input signals the virtual loudspeaker positions could, e.g., correspond to loudspeaker angles of +/-30 degrees with respect to front direction. For multichannel audio signals, e.g. for 5.1 these angles are typically 0, +/-30 and +/-110

degrees. However, in practice, the virtual loudspeaker positions may have any meaningful values. Target sound source position indication may also be provided by other means (via user interface), be a hardcoded value or be omitted. In at least some examples, the indication 12 is used to control signal decomposition 104. In some but not necessarily all examples, it can be used for stereo-widening 112.

[0058] The audio processing system 100, 100' provides the widened stereo signal 115 derived therein to an interface for communicating to headphone apparatus 20 for rendering.

[0059] The headphone apparatus 20 is an apparatus that has a left audio output device 21 that is worn at, over or in a left ear of a user and a right audio output device 22 that is worn at, over or in a right ear of a user. The audio heard in the left ear by the user is dependent upon audio output by the left audio output device 21 and is not dependent upon audio output by the right audio output device 22. The audio heard in the right ear by the user is dependent upon audio output by the right audio output device 22 and is not dependent upon audio output by the left audio output device 21. The headphone apparatus 20 receives input signals wirelessly or over a wired connection. In some but not necessarily all examples, the headphone apparatus 20 comprises acoustic isolators 23 that isolate the ears of the user from external environmental sounds. In some examples, the headphone apparatus can comprise left and right 'cans' 23 that cover the user's ears, house the respective audio output devices 21, 22 and provide at least some acoustic isolation. In some examples, the headphone apparatus can comprise a deformable 'buds' that fit snugly inside the respective left and right ears of the user, surround the respective audio output devices 21, 22 and provide at least some acoustic isolation.

[0060] Each audio output device 21, 22 comprises a transducer that converts a received electrical signal to an acoustic pressure wave or a vibration.

[0061] The stereo signal 101 may be received at the signal processing system 100, 100' e.g. by reading the stereo signal from a memory or from a mass storage device in the device 50. In another example, the stereo signal is obtained via communication interface (such as a network interface) from another device that stores the stereo signal in a memory or from a mass storage device provided therein. The widened stereo signal 115 may be provided for rendering by headphone apparatus 20. Additionally or alternatively, the widened stereo signal 115 may be stored in the memory or the mass storage device in the device 50 and/or provided via a communication interface to another device for storage therein.

[0062] The information 12 that defines the virtual loudspeaker positions may be used to control stereo widening processing such that audio sources are perceived at desired positions, which may also be at positions outside the physical locations of the headphones. The processing may include maintaining some portions (such as the focus portion of the spatial audio image) in between the physical locations of the headphones.

[0063] The audio processing system 100, 100' may be arranged to process the stereo signal 101 arranged into a sequence of input frames, each input frame including a respective segment of digital audio signal for each of the channels, provided as a respective time series of input samples at a predefined sampling frequency. In typical example, the audio processing system 100, 100' employs a fixed predefined frame length. In other examples, the frame length may be a selectable frame length that may be selected from a plurality of predefined frame lengths, or the frame length may be an adjustable frame length that may be selected from a predefined range of frame lengths. A frame length may be defined as number samples L included in the frame for each channel of the stereo signal 101, which at the predefined sampling frequency maps to a corresponding duration in time. As an example, in this regard, the audio processing system 100, 100' may employ a fixed frame length of 20 milliseconds (ms), which at a sampling frequency of 8, 16, 32 or 48 kHz results in a frame of L=160, L=320, L=640 and L=960 samples per channel, respectively. The frames may be non-overlapping or they may be partially overlapping. These values, however, serve as non-limiting examples and frame lengths and/or sampling frequencies different from these examples may be employed instead, depending e.g. on the desired audio bandwidth, on desired framing delay and/or on available processing capacity.

[0064] Referring back to Figures 1A and 1B, the audio processing system 100, 100' may comprise the transform entity 102 that is arranged to convert the stereo signal 101 from time domain into a transform-domain stereo signal 103. Typically, the transform domain involves a frequency domain. In an example, the transform entity 102 employs short-time discrete Fourier transform (STFT) to convert each channel of the stereo signal 101 into a respective channel of the transform-domain stereo signal 103 using a predefined analysis window length (e.g. 20 milliseconds). In another example, the transform entity 102 employs an (analysis) complex-modulated quadrature-mirror filter (QMF) bank for time-to-frequency-domain conversion. The STFT and QMF bank serve as non-limiting examples in this regard and in further examples any suitable transform technique known in the art may be employed for creating the transform-domain stereo signal 103.

[0065] The transform entity 102 may further divide each of the channels into a plurality of frequency sub-bands, thereby resulting in the transform-domain stereo signal 103 that provides a respective time-frequency representation for each channel of the stereo signal 101. A given frequency band in a given frame may be referred to as a time-frequency tile. The number of frequency sub-bands and respective bandwidths of the frequency sub-bands may be selected e.g. in accordance with the desired frequency resolution and/or available computing power. In an example, the sub-band structure involves 24 frequency sub-bands according to the Bark scale, an equivalent rectangular band (ERB) scale or

3rd octave band scale known in the art. In other examples, different number of frequency sub-bands that have the same or different bandwidths may be employed. A specific example in this regard is a single frequency sub-band that covers the input spectrum in its entirety or a continuous subset thereof.

[0066] A time-frequency tile that represents frequency bin *b* in time frame *n* of channel *i* of the transform-domain stereo signal 103 may be denoted as $S(i, b, n)$. The channel *i* represents a single virtual loudspeaker or an input channel. The transform-domain stereo signal 103, e.g. the time-frequency tiles $S(i, b, n)$, are passed to the signal decomposer 104 for decomposition into the first signal component 105-1 and the second signal component 105-2 therein. As described in the foregoing, a plurality of consecutive frequency bins may be grouped into a frequency sub-band, thereby providing a plurality of frequency sub-bands $k = 0, ..., K-1$. For each frequency sub-band *k*, the lowest bin (i.e. a frequency bin that represents the lowest frequency in that frequency sub-band) may be denoted as $b_{k,low}$ and the highest bin (i.e. a frequency bin that represents the highest frequency in that frequency sub-band) may be denoted as $b_{k,high}$.

[0067] Referring back to Figures 1A and 1B, the audio processing system 100, 100' may comprise the signal decomposer 104 that is arranged to derive, based on the transform-domain stereo signal 103, the first signal component 105-1 and the second signal component 105-2. In the following, the first signal component 105-1 is referred to as a signal component that represents the focus portion of the spatial audio image and the second signal component 105-2 is referred to a signal component that represents the non-focus portion of the spatial audio image. The focus portion represents those parts of the audio image that are front and central and can be considered as 'frontness'. The non-focus portion represents those parts of the audio image that are not represented by the focus portion (not front and central) and may be hence referred to as a 'peripheral' portion of the spatial audio image. Herein, the decomposition procedure does not change the number of channels and hence in the present example each of the first signal component 105-1 and the second signal component 105-2 is provided as a respective two-channel audio signal. It should be noted that the terms focus portion and non-focus portion as used in this disclosure are designations assigned to spatial sub-portions of the spatial audio image represented by the stereo signal 101, while these designation as such do not imply any specific processing to be applied (or having been applied) to the underlying stereo signal 101 or the transform-domain stereo signal 103 e.g. to actively emphasize or de-emphasize any portion of the spatial audio image represented by the stereo signal 101.

[0068] The signal decomposer 104 may derive, on basis of the transform-domain stereo signal 103, the first signal component 105 that represents those coherent sounds of the spatial audio image that are within a predefined focus range, such sounds hence constituting the focus portion of the spatial audio image. The focus range can be defined by the control input 10.

[0069] In contrast, the signal decomposer 104 may derive, on basis of the transform-domain stereo signal 103, the second signal component 105 that represents coherent sound sources or sound components of the spatial audio image that are outside the predefined focus range and all non-coherent sound sources of the spatial audio image, such sound sources or components hence constituting the non-focus portion of the spatial audio image. Hence, the signal decomposer 104 decomposes the sound field represented by the stereo signal 101 into the first signal component 105-1 that is excluded from subsequent stereo widening processing and into the second signal component 105-2 that is subsequently subjected to the stereo widening processing.

[0070] Figure 3 illustrates a block diagram of some components and/or entities of the signal decomposer 104 according to an example. The signal decomposer 104 may be, conceptually, divided into a decomposition analyzer 104a and a signal divider 126, as illustrated in Figure 3. In the following, entities of the signal decomposer 104 according to the example of Figure 3 are described in more detail. In other examples, the signal decomposer 104 may include further entities and/or some entities depicted in Figure 3 may be omitted or combined with other entities

[0071] The signal decomposer 104 may comprise a coherence analyzer 116 for estimating, on basis of the transform-domain stereo signal 103, coherence values 117 that are descriptive of coherence between the channels of the transform-domain stereo signal 103. The coherence values 117 are provided for a decomposition coefficient determiner 124 for further processing therein.

[0072] Computation of the coherence values 117 may involve deriving a respective coherence value $\gamma(k, n)$ for a plurality of frequency sub-bands *k* in a plurality of time frames *n* based on the time-frequency tiles $S(i, b, n)$ that represent the transform domain stereo signal 103. As an example, the coherence values 117 may be computed e.g. according to the equation (3):

$$\gamma(k,n) = \frac{\sum_{b=b_{k,low}}^{b_{k,high}} Re(S^*(1,b,n)S(2,b,n))}{\sum_{b=b_{k,low}}^{b_{k,high}} (|S(1,b,n)||S(2,b,n)|)}, \quad (3)$$

where Re denotes the real part operator and * denotes the complex conjugate.

[0073] The term $\gamma(k, n)$ has a large value when the audio of the channels is dominated by an audio event that is common to both channels. A common audio event will typically cause a complex phasor distribution across the frequency bins *b*. For all frequency bins inside a frequency band, the phase is the same in both channels in the case of full coherence (i.e., $\gamma(k, n) = 1$)..

[0074] Still referring to Figure 3, the signal decomposer 104 may comprise the energy estimator 118 for estimating energy of the transform-domain stereo signal 103 on basis of the transform-domain stereo signal 103. The energy values 119 are provided for a direction estimator 120 for direction angle estimation therein.

[0075] Computation of the energy values 119 may involve deriving a respective energy value $E(i, k, n)$ for a plurality of frequency sub-bands *k* in plurality of audio channels *i* in a plurality of time frames *n* based on the time-frequency tiles $S(i, b, n)$. As an example, the energy values $E(i, k, n)$ may be computed e.g. according to the equation (4):

$$E(i,k,n) = \sum_{b_{k,\mathrm{low}}}^{b_{k,\mathrm{high}}} |S(i,b,n)|^2. \quad (4)$$

[0076] Still referring to Figure 3, the signal decomposer 104 may comprise the direction estimator 120 for estimating perceivable arrival direction of the sound represented by the stereo signal 101 based on the energy values 119 in view of a target virtual loudspeaker configuration applied in the stereo signal 101. The direction estimation may comprise computation of direction angles 121 based on the energy values in view of the target virtual loudspeaker positions, which direction angles 121 are provided for a focus estimator 122 for further analysis therein.

[0077] The target sound source (virtual loudspeaker) configuration may also be referred to as channel configuration (of the stereo signal 101). This information may be obtained, for example, from metadata 12 that accompanies the stereo signal 101, e.g. metadata included in an audio container within which the stereo signal 101 is stored. In another example, the information defining the target virtual loudspeaker configuration applied in the stereo signal 101 may be received (as user input) 12 via a user interface of the device 50. The target virtual loudspeaker configuration may be defined by indicating, for each channel of the stereo signal 101, a respective target virtual loudspeaker position with respect to an assumed listening point. As an example, a target position for a virtual loudspeaker may comprise a target direction, which may be defined as an angle with respect to a reference direction (e.g. a front direction). Hence, for example in case of a two-channel stereo signal the target virtual loudspeaker configuration may be defined as respective target angles α_{in} (1) and α_{in} (2) with respect to the front direction for the left and right virtual loudspeakers. The target angles α_{in} (*i*) with respect to the front direction may be, alternatively, indicated by a single target angle α_{in} which defines the absolute value of the target angles with respect to the front direction e.g. such that α_{in} (1) = α_{in} and α_{in} (2) = -α_{in}.

[0078] In a further example, no indication 12 is received in the audio processing system 100, 100' and the elements of the audio processing system 100, 100' that make use of the information that defines the target virtual loudspeaker configuration applied in the stereo signal 101 (the signal decomposer 104, the re-panner 106) apply predefined information in this regard instead. An example in this regard involves applying a fixed predefined target virtual loudspeaker configuration. Another example involves selecting one of a plurality of predefined target virtual loudspeaker configurations in dependence of the number of audio channels in the received stereo signal 101. Non-limiting examples in this regard include selecting, in response to a two-channel signal 101 (which is hence assumed as a two-channel stereophonic audio signal), a target virtual loudspeaker configuration where the channels are positioned ± 30 degrees with respect to the front direction and/or selecting, in response to a six-channel signal (that is hence assumed to represent a 5.1-channel surround signal), a target virtual loudspeaker configuration where the channels are positioned at target angles α_{in} (*i*) of 0 degrees, ± 30 degrees and ± 110 degrees with respect to the front direction and complemented with a low frequency effects (LFE) channel.

[0079] The direction estimator 120 is configured to estimate perceivable arrival direction of the sound represented by the stereo signal 101. The direction estimation may involve deriving a respective direction angle 121, $\theta(k, n)$, for a plurality of frequency sub-bands *k* in a plurality of time frames *n* based on the estimated energies $E(i, k, n)$ and the target virtual loudspeaker positions α_{in} (*i*), the direction angles 121, $\theta(k, n)$, thereby indicating the estimated perceived arrival direction of the sound in frequency sub-bands of input frames. The direction estimation may be carried out, for example, using the tangent law according to the equations (5) and (6), where an underlying assumption is that sound sources in the sound field represented by the stereo signal 101 are arranged (to a significant extent) in their desired spatial positions using amplitude panning:

$$\theta(k,n) = \arctan\left(\tan \propto_{in} \frac{g_1 - g_2}{g_1 + g_2}\right), \quad (5)$$

where

$$g_1 = \sqrt{E(1,k,n)}$$
$$g_2 = \sqrt{E(2,k,n)} \quad (6)$$

where α_{in} denotes the absolute value of the target angles α_{in} (1) and α_{in} (2) that define, respectively, the target positions of the left and right virtual loudspeakers with respect to the front direction, which in this example are positioned symmetrically (and equidistantly) with respect to the front direction. In other examples, the target positions of the left and right virtual loudspeakers may be positioned non-symmetrically with respect to the front direction (e.g. such that $|\alpha\_in\ (1)| \neq |\alpha\_in\ (2)|$). Modification of the equation (5) such that it accounts for this aspect is a straightforward task for a person skilled in the art.

[0080] For example, the modification of the equation (5) in the case of non-symmetric (virtual) loudspeaker positions can be performed as follows. First, a half of the angle between the loudspeakers is computed

$$\propto_o = \frac{|\propto_{in}(1) - \propto_{in}(2)|}{2}.$$

[0081] Next, the center point between the loudspeakers is computed

$$\propto_c = \frac{\propto_{in}(1) + \propto_{in}(2)}{2}.$$

[0082] Using these values, the equation (5) can represented for non-symmetric cases as

$$\theta(k,n) = \arctan\left(\tan \propto_o \frac{g_1 - g_2}{g_1 + g_2}\right) + \propto_c,$$

where g_1 and g_2 are computed as in equation (6).

[0083] Still referring to Figure 3, the signal decomposer 104 may comprise the focus estimator 122 for determining one or more focus coefficients 123 based on the estimated perceivable arrival direction of the sound represented by the stereo signal 101 (directions angles 121) in view of a defined focus range within the spatial audio image, where the focus coefficients 123 are indicative of the relationship between the estimated arrival direction of the sound (direction angles 121) and the focus range. The focus range may be defined, for example, as a single angular range or as two or more angular sub-ranges in the spatial audio image. In other words, the focus range may be defined as a set of arrival directions of the sound within the spatial audio image. The focus range can be defined by the control input 10.

[0084] The focus coefficients 123 may be derived by the focus estimator 122 based at least in part on the direction angles 121. The focus estimator 122 may optionally further receive the indication 12 of the target virtual loudspeaker configuration applied in the stereo signal 101, and compute the focus coefficients 123 further in view of this information. The focus coefficients 123 are provided for the decomposition coefficient determiner 124 for further processing therein.

[0085] Typically, the one or more angular ranges of the focus range define a set of arrival directions that cover a defined portion around the center of the spatial audio image, thereby rendering the focus estimation as a 'frontness' estimation. The focus estimation may involve deriving a respective focus (frontness) coefficient $\chi(k, n)$ for a plurality of frequency sub-bands k in a plurality of time frames n based on the direction angles 121, $\theta(k, n)$, e.g. according to the equation (7):

$$\chi(k,n) = \begin{cases} 1, |\theta(k,n)| < \theta_{Th1} \\ 1 - \frac{|\theta(k,n)| - \theta_{Th1}}{(\theta_{Th2} - \theta_{Th1})}, \theta_{Th1} \leq |\theta(k,n)| \leq \theta_{Th2}. \\ 0, |\theta(k,n)| > \theta_{Th2} \end{cases} \quad (7)$$

[0086] In the equation (7), the first threshold value θ_{Th1} and the second threshold value θ_{Th2}, where $\theta_{Th1} < \theta_{Th2}$, serve

to define a primary (center) angular focus range (between angles $-\theta_{Th1}$ to θ_{Th1} around the front direction), a secondary angular focus range (from $-\theta_{Th2}$ to $-\theta_{Th1}$ and from θ_{Th1} to θ_{Th2} with respect to the front direction) and a non-focus range (outside $-\theta_{Th2}$ and θ_{Th2} with respect to the front direction). The coefficients defining the focus range θ_{Th1}, θ_{Th2}, can be defined by the control input 10.

[0087] As a non-limiting example, the first and second threshold values may be set to $\theta_{Th1} = 5°$ and $\theta_{Th2} = 15°$, whereas in other examples different threshold values θ_{Th1} and θ_{Th2} may be applied instead. Focus estimation according to the equation (7) hence applies a focus range that includes two angular ranges (i.e. the primary angular focus range and the secondary angular focus range) and sets the focus coefficient $\chi(k, n)$ to unity in response to a sound source direction residing within the primary angular focus range and sets the focus coefficient $\chi(k, n)$ to zero in response to the sound source direction residing outside the focus range, whereas a predefined function of sound source direction is applied to set the focus coefficient $\chi(k, n)$ to a value between unity and zero in response to the sound source direction residing within the secondary angular focus range. In general, the focus coefficient $\chi(k, n)$ is set to a non-zero value in response to the sound source direction residing within the focus range and the focus coefficient $\chi(k, n)$ is set to zero value in response to the perceived sound source direction, direction angles 121, $\theta(k, n)$, residing outside the focus range. In an example, the equation (7) may be modified such that no secondary angular focus range is applied and hence only a single threshold may be applied to define the limit(s) between the focus range and the non-focus range.

[0088] Along the lines described in the foregoing, the focus range may be defined as one or more contiguous, non-overlapping angular focus ranges. As an example, the focus range may include a single defined angular range or two or more defined angular ranges.

[0089] According to another example, at least one of the focus ranges is selectable, e.g. such that an angular focus range may be selected or adjusted (e.g. via selection or adjustment of one or more threshold values that define the respective angular focus range) in dependence of the target (or assumed) virtual loudspeaker configuration associated with the stereo input signal 12, and the focus range parameter present in control input 10. For example, the control information could be used to control how large a portion (or which angles) of the sound image will be sent to widening.

[0090] Still referring to Figure 3, the signal decomposer 104 may comprise the decomposition coefficient determiner 124 for deriving decomposition coefficients 125 based on the coherence values 117 and the focus coefficients 123. The decomposition coefficients 125 are provided for the signal divider 126 for decomposition of the transform-domain stereo signal 103 therein.

[0091] The signal divider 126 is configured to derive, based on the transform-domain stereo signal 103 and the decomposition coefficients 125, the first signal component 105-1 that represents the focus portion of the spatial audio image and the second signal component 105-2 that represents the non-focus portion (e.g. a 'peripheral' portion) of the spatial audio image.

[0092] The decomposition coefficient determination aims at providing a high value for a decomposition coefficient $\beta(k, n)$ for a frequency sub-band k and frame n that exhibits relatively high coherence between the channels of the stereo signal 101 and that conveys a directional sound component that is within the focus portion of the spatial audio image (see description of the focus estimator 122 in the foregoing). In this regard, the decomposition coefficient determination may involve deriving a respective decomposition coefficient $\beta(k, n)$ for a plurality of frequency sub-bands k in a plurality of time frames n based on the respective coherence value $\gamma(k, n)$ and the respective focus coefficient $\chi(k, n)$ e.g. according to the equation (8):

$$\beta(k,n) = \gamma(k,n)\chi(k,n). \tag{8}$$

[0093] In an example, the decomposition coefficients $\beta(k, n)$ may be applied as such as the decomposition coefficients 125 that are provided for the signal divider 126 for decomposition of the transform-domain stereo signal 103 therein.

[0094] In another example, energy-based temporal smoothing is applied to the decomposition coefficient $\beta(k, n)$ obtained from the equation (8) in order to derive smoothed decomposition coefficients $\beta'(k, n)$, which may be provided for the signal divider 126 to be applied for decomposition of the transform-domain stereo signal 103 therein. Smoothing of the decomposition coefficients results in slower variations over time in sub-portions of the spatial audio image assigned to the first signal component 105-1 and the second signal component 105-2, which may enable improved perceivable quality in the resulting widened stereo signal 115 via avoidance of small-scale fluctuances in the spatial audio image therein. A weighting that provides the energy-based temporal smoothing may be provided, for example, according to the equation (9a):

$$\beta'(k,n) = A(k,n) \,/\, B(k,n), \tag{9a}$$

where

$$A(k,n) = aE(k,n)\beta(k,n) + bA(k,n-1)$$
$$B(k,n) = aE(k,n) + bB(k,n-1) \quad \text{, and} \tag{9b}$$

where $E(k, n)$ denotes the total energy of the transform-domain stereo signal 103 for a frequency sub-band k in time frames n (derivable e.g. based on the energies $E(i, k, n)$ derived using the equation (4)) and a and b (where, preferably, $a + b = 1$) denote predefined weighting factors. The weighting factors for energy-based temporal smoothing (a and b) can be defined by the control input 10. As a non-limiting example, values $a = 0.2$ and $b = 0.8$ may be applied, whereas in other examples other values in the range from 0 to 1 may be applied instead.

[0095] Still referring to Figure 3, the signal decomposer 104 may comprise the signal divider 126 for deriving, based on the transform-domain stereo signal 103 and the decomposition coefficients 125, the first signal component 105-1 that represents the focus portion of the spatial audio image and the second signal component 105-2 that represents the non-focus portion (e.g. a 'peripheral' portion) of the spatial audio image.

[0096] As an example, the signal decomposition may be carried out for a plurality of frequency sub-bands k in a plurality of channels i in a plurality of time frames n based on the time-frequency tiles $S(i, b, n)$, according the equation (10a):

$$S_{sw}(i,b,n) = S(i,b,n)(1-\beta(b,n))^p$$
$$S_{dr}(i,b,n) = S(i,b,n)\beta(b,n)^p \tag{10a}$$

where

$S_{dr}(i, b, n)$ denotes frequency bin b in time frame n of channel i of the first signal component 105-1 that represents the focus portion of the spatial audio image,

$S_{sw}(i, b, n)$ denotes frequency bin b in time frame n of channel i of the second signal component 105-2 the non-focus portion (e.g. a 'peripheral' portion) of the spatial audio image, p denotes predefined constant parameter (e.g. $p = 0.5$, or 1), and

β(b, n) is equal to the decomposition coefficients $\beta(k, n)$ for each frequency bin b within the frequency sub-band k.

[0097] The signal divider 126 creates the first signal component 105-1 that represents the focus portion of the spatial audio image and the second signal component 105-2 that represents the non-focus portion (e.g. a 'peripheral' portion) of the spatial audio image but it does not necessarily place a time-frequency tile $S(i, b, n)$ into either the first signal component 105-1 or the second signal component 105-2. It can, as in this example, scale or weight the contribution of a time-frequency tile $S(i, b, n)$ more heavily in one of the first signal component 105-1 or the second signal component 105-2 dependent upon the decomposition coefficients $\beta(k, n)$.

[0098] The scaling coefficient $\beta(b, n)^p$ in the equation (9) may be replaced with another scaling coefficient that increases with increasing value of the decomposition coefficient $\beta(b, n)$ (and decreases with decreasing value of the decomposition coefficient $\beta(b, n)$) and the scaling coefficient $(1 - \beta(b, n))^p$ in the equation (10a) may be replaced with another scaling coefficient that decreases with increasing value of the decomposition coefficient $\beta(b, n)$ (and increases with decreasing value of the decomposition coefficient $\beta(b, n)$).

[0099] In another example, the signal decomposition may be carried out for a plurality of frequency sub-bands k in a plurality of channels i in a plurality of time frames n based on the time-frequency tiles $S(i, b, n)$, according the equation (10b):

$$S_{sw}(i,b,n) = \begin{cases} S(i,b,n), \beta(b,n) \le \beta_{Th} \\ 0, \beta(b,n) > \beta_{Th} \end{cases}$$
$$S_{dr}(i,b,n) = \begin{cases} 0, \beta(b,n) \le \beta_{Th} \\ S(i,b,n), \beta(b,n) > \beta_{Th} \end{cases}, \tag{10b}$$

wherein β_{Th} denotes a defined threshold value that has value in the range from 0 to 1, e.g. $\beta_{Th} = 0.5$. The signal decomposition parameter β_{Th} can be defined by the control input 10. If applying the equation (10b) the temporal smoothing of the decomposition coefficients 125 described in the foregoing and/or temporal smoothing of the resulting signal

components $S_{sw}(i, b, n)$ and $S_{dr}(i, b, n)$ may be advantageous for improved perceivable quality of the resulting widened stereo signal 115.

[0100] The decomposition coefficients $\beta(k, n)$ according to the equation (8) are derived on time-frequency tile basis, whereas the equations (10a) and (10b) apply the decomposition coefficients $\beta(b, n)$ on frequency bin basis. In this regard, the decomposition coefficients $\beta(k, n)$ derived for a frequency sub-band k may be applied for each frequency bin b within the frequency sub-band k.

[0101] Consequently, the transform-domain stereo signal 103 is divided, in each time-frequency tile $S(i, b, n)$, into the first signal component 105-1 that represents sound components positioned in the focus portion of the spatial audio image represented by the stereo signal 101 and into the second signal component 105-2 that represents sound components positioned outside the focus portion of the spatial audio image represented by the stereo signal 101. The first signal component 105-1 is subsequently provided for playback without applying stereo widening thereto, whereas the second signal component 105-2 is subsequently provided for playback after being subjected to stereo widening.

[0102] Referring back to Figures 1A and 1B, the audio processing system 100, 100' may comprise the re-panner 106 that is arranged to generate a modified first signal component 107 on basis of the first signal component 105-1, wherein one or more sound sources represented by the first signal component 105-1 are repositioned in the spatial audio image.

[0103] Figure 4 illustrates a block diagram of some components and/or entities of the re-panner 106 according to an example. In the following, entities of the re-panner 106 according to the example of Figure 4 are described in more detail. In other examples, the re-panner 106 may include further entities and/or some entities depicted in Figure 4 may be omitted or combined with other entities

[0104] The re-panner 106 may comprise an energy estimator 128 for estimating energy of the first signal component 105-1. The energy values 129 are provided for a direction estimator 130 and for a re-panning gain determiner 136 for further processing therein. The energy value computation may involve deriving a respective energy value $E_{dr}(i, k, n)$ for a plurality of frequency sub-bands k in plurality of audio channels i (plurality of virtual loudspeakers) in a plurality of time frames n based on the time-frequency tiles $S_{dr}(i, b, n)$. As an example, the energy values $E_{dr}(i, k, n)$ may be computed e.g. according to the equation (11):

$$E_{dr}(i,k,n) = \sum_{b_{k,\mathrm{low}}}^{b_{k,\mathrm{high}}} S_{dr}|(i,b,n)|^2 \tag{11}$$

[0105] In another example, the energy values 119 computed in the energy estimator 118 (e.g. according to the equation (4)) may be re-used in the re-panner 106, thereby dispensing with a dedicated energy estimator 128 in the re-panner 106. Even though the energy estimator 118 of the signal decomposer 104 estimates the energy values 119 based on the transform-domain stereo signal 103 instead of the first signal component 105-1, the energy values 119 enable correct operation of the direction estimator 130 and the re-panning gain determiner 136.

[0106] Still referring to Figure 4, the re-panner 106 may comprise the direction estimator 130 for estimating perceivable arrival direction of the sound represented by the first signal component 105-1 based on the energy values 129 in view of the target virtual loudspeaker configuration applied in the stereo signal 101. The direction estimation may comprise computation of direction angles 131 based on the energy values 129 in view of the target virtual loudspeaker positions, which direction angles 131 are provided for a direction adjuster 132 for further processing therein.

[0107] The direction estimation may involve deriving a respective direction angle 131, $\theta_{dr}(k, n)$, for a plurality of frequency sub-bands k in a plurality of time frames n based on the estimated energies $E_{dr}(i, k, n)$ and the positions α_{in} (i) of the target virtual loudspeakers. The direction angles 131, $\theta_{dr}(k, n)$, indicate the estimated perceived arrival direction (direction angle 131) of the sound in frequency sub-bands of first signal component 105-1. The direction estimation may be carried out, for example, according to the equations (12) and (13):

$$\theta_{dr}(k,n) = \arctan\left(\tan \propto_{in} \frac{g_{1,dr} - g_{2,dr}}{g_{1,dr} + g_{2,dr}}\right), \tag{12}$$

where

$$\begin{aligned} g_{1,dr} &= \sqrt{E_{dr}(1,k,n)}, \quad \text{for a first virtual loudspeaker} \\ g_{2,dr} &= \sqrt{E_{dr}(2,k,n)}, \quad \text{for a second virtual loudspeaker} \end{aligned} \tag{13}$$

[0108] In another example, the direction angles 121 computed in the energy estimator 128 (e.g. according to the

equations (5) and (6)) may be re-used in the re-panner 106, thereby dispensing with a dedicated direction estimator 130 in the re-panner 106. Even though the direction estimator 120 of the signal decomposer 104 estimates the direction angles 121 based on the energy values 119 derived from the transform-domain stereo signal 103 instead of the first signal component 105-1, the sound source positions are the same or substantially the same and hence the direction angles 121 enable correct operation of the direction adjuster 132.

[0109] Still referring to Figure 4, the re-panner 106 may comprise the direction adjuster 132 for modifying the estimated perceivable arrival direction (direction angle 131) of the sound represented by the first signal component 105-1. The direction adjuster 132 may derive modified direction angles 133 based on the direction angles 131. The modified direction angles 133 are provided for a panning gain determiner 134 for further processing therein.

[0110] The direction adjustment may comprise mapping the currently estimated perceivable arrival direction, direction angles 131, into respective modified direction angles 133 that represent new adjusted perceivable arrival direction of the sound in view of the control information 10.

[0111] The mapping between the currently estimated perceivable arrival direction, direction angles 131, and the new adjusted perceivable arrival directions, modified direction angles 132, may be provided by determining a mapping coefficient μ

which may be applied for deriving a respective modified direction angle $\theta'(k, n)$ for a plurality of frequency sub-bands k in a plurality of time frames n e.g. according to the equation (15):

$$\theta'(k,n) = \mu\,\theta(k,n). \tag{15}$$

[0112] The value of the mapping coefficient μ for panning can be defined explicitly by the control input 10.

[0113] If stereo widening 112 "widens" the signal 105-2 by a certain amount then, the re-panner 106 widens the signal 105-1 via re-panning by the same amount. As a practical example, the stereo widening 112 may widen the signal so that a sound source originally at the location of 5 degrees is perceived after the widening at the location corresponding to 10 degrees in the original signals. Hence, the control information 10 may have information saying that re-panning by the factor 2 (μ=2) is needed, so that the positions of the re-panned audio 107 match with the positions of the stereo widened audio 113.

[0114] The determination of the mapping coefficient μ and derivation of the modified direction angles $\theta'(k, n)$ according to the equations (14) and (15) serves as a non-limiting example and a different procedure for deriving the modified direction angles 133 may be applied instead.

[0115] Still referring to Figure 4, the re-panner 106 may comprise the panning gain determiner 134 for computing a set of panning gains 135 on basis of the modified direction angles 133. The panning gain determination may comprise, for example, using vector base amplitude panning (VBAP) technique known in the art to compute a respective panning gain $g'(i, k, n)$ for a plurality of frequency sub-bands k in plurality of audio channels i in a plurality of time frames n based on the modified direction angles $\theta'(k, n)$.

[0116] For example, the panning gains $g'(i, k, n)$ may be derived based on the tangent law

$$A = \frac{\tan\theta'(k,n)}{\tan\alpha_{in}}$$

$$B = \frac{1+A}{1-A}$$

$$g'(1,k,n) = \sqrt{\frac{B^2}{1+B^2}}$$

$$g'(2,k,n) = \sqrt{1-g'(1,k,n)^2}.$$

[0117] Still referring to Figure 4, the re-panner 106 may comprise the re-panning gain determiner 136 for deriving re-panning gains 137 based on the panning gains 135 and the energy values 129. The re-panning gains 137 are provided

for a re-panning processor 138 for derivation of a modified first signal component 107 therein.

[0118] The re-panning gain determination procedure may comprise computing a respective total energy $E_s(k, n)$ for a plurality of frequency sub-bands k in a plurality of time frames n e.g. according to the equation (18):

$$E_s(k,n) = \sum_i E_{dr}(i,k,n). \quad (18)$$

[0119] The re-panning gain determination may further comprise computing a respective target energy $E_t(i, k, n)$ for a plurality of frequency sub-bands k in plurality of audio channels i in a plurality of time frames n based on the total energies $E_s(k, n)$ and the panning gains $g'(i, k, n)$, e.g. according to the equation (19):

$$E_t(i,k,n) = g'(i,k,n)^2 E_s(k,n). \quad (19)$$

[0120] The target energies $E_t(i, k, n)$ may be applied with the energy values $E_{dr}(i, k, n)$ to derive a respective re-panning gain $g_r(i, k, n)$ for a plurality of frequency sub-bands k in plurality of audio channels i in a plurality of time frames n, e.g. according to the equation (20):

$$g_r(i,k,n) = \sqrt{E_t(i,k,n)/E_{dr}(i,k,n)}. \quad (20)$$

[0121] In an example, the re-panning gains $g_r(i, k, n)$ obtained from the equation (20) may be applied as such as the re-panning gains 137 that are provided for the re-panning processor 138 for derivation of the modified first signal component 107 therein. In another example, energy-based temporal smoothing is applied to the re-panning gains $g_r(i, k, n)$ obtained from the equation (20) in order to derive smoothed re-panning gains $g'_r(i, k, n)$, which may be provided for the re-panning processor 138 to be applied for re-panning therein. Smoothing of the re-panning gains $g_r(i, k, n)$ results in slower variations over time within the sub-portion of the spatial audio image assigned to the first signal component 105-1, which may enable improved perceivable quality in the resulting widened stereo signal 115 via avoidance of small-scale fluctuances in the respective portion of the widened spatial audio image therein.

[0122] Still referring to Figure 4, the re-panner 106 may comprise the re-panning processor 138 for deriving the modified first signal component 107 on basis of the first signal component 105-1 in dependence of the re-panning gains 137. In the resulting modified first signal component 107 the sound sources in the focus portion of the spatial audio image are repositioned (i.e. re-panned) in accordance with the modified direction angles 132 derived in the direction adjuster 132 to account for (possible) differences between direct reproduction of stereo signals over headphones and reproduction of stereo widening 112 processed stereo signals over headphones.,. The channels of the modified first signal component 107 are provided to an inverse transform entity 108-1 for conversion from the transform domain to the time domain therein.

[0123] The procedure for deriving the modified first signal component 107 may comprise deriving a respective time-frequency tile $S_{dr,rp}(i, b, n)$ for a plurality of frequency bins b in plurality of audio channels i in a plurality of time frames n based on a corresponding time-frequency tiles $S_{dr}(i, b, n)$ of the first signal component 105-1 in dependence of the re-panning gains $g_r(i, b, n)$, e.g. according to the equation (21):

$$S_{dr,rp}(i,b,n) = g_r(i,b,n) S_{dr}(i,b,n). \quad (21)$$

[0124] The re-panning gains $g_r(i, k, n)$ according to the equation (20) are derived on time-frequency tile basis, whereas the equation (21) applies the re-panning gains $g_r(i, k, n)$ on frequency bin basis. In this regard, the re-panning gain $g_r(i, k, n)$ derived for a frequency sub-band k may be applied for each frequency bin b within the frequency sub-band k.

[0125] In other examples, panning can apply to each time-frequency tile $S(i, b, n)$ different combinations of controlled gain $g_r(i, b, n)$, controlled reverberation or decorrelation and, optionally, controlled delays to produce the channels of the modified first signal component 107. The reverberation or decorrelation is typically added only at a low level.

[0126] In some embodiments, the modified first signal component 107 may be divided to two paths (e.g., using a variable received in the control information 10). The signal in the second path is processed using reverberation or decorrelation. The signal in the first path is passed forward without processing and without any cross-channel mixing. The signals in the two paths are combined, e.g., by summing them.

[0127] Referring back to Figure 1A, the audio processing system may comprise the inverse transform entity 108-1 that is arranged to transform the channels of the modified first signal component 107 from the transform-domain (back) to the time domain, thereby providing a time-domain modified first signal component 109-1. Along similar lines, the audio

processing system 100 may comprise an inverse transform entity 108-2 that is arranged to transform channels of the second signal component 105-2 from the transform-domain (back) to the time domain, thereby providing a time-domain second signal component 109-2. Both the inverse transform entity 108-1 and the inverse transform entity 108-2 make use of an applicable inverse transform that inverts the time-to-transform-domain conversion carried out in the transform entity 102. As non-limiting examples in this regard, the inverse transform entities 108-1, 108-2 may apply an inverse STFT or a (synthesis) QMF bank to provide the inverse transform. The resulting time-domain modified first signal component 109-1 may be denoted as $s_{dr}(i, m)$ and the resulting time-domain second signal component 109-2 may be denoted as $s_{sw}(i, m)$, where i denotes the channel and m denotes a time index (i.e. a sample index).

[0128] Referring back to Figure 1B, as described in the foregoing, in the audio processing system 100' the inverse transform entities 108-1, 108-2 are omitted, and the modified first signal component 107 is provided as a transform-domain signal to the (optional) delay element 110' and the transform-domain second signal component 105-2 is provided as a transform-domain signal to the stereo widening processor 112'.

[0129] Referring back to Figure 1A, the audio processing system 100 may comprise the stereo widening processor 112 that is arranged to generate, on basis of the second signal component 109-2, the modified second signal component 113 where the width of a spatial audio image is extended from that represented by the second signal component 109-2. The stereo widening processor 112 may apply any stereo widening technique known in the art to extend the width of the spatial audio image. In an example, the stereo widening processor 112 processes the second signal component $s_{sw}(i, m)$ into the modified second signal component $s'_{sw}(i, m)$. where the second signal component $s_{sw}(i, m)$ and the modified second signal component $s'_{sw}(i, m)$ are respective time-domain signals.

[0130] Stereo widening techniques can involve adding a processed (e.g. filtered) version of a contralateral channel signal to each of the left and right channel signals of the stereo signal in order to derive an output stereo signal having a widened spatial audio image (a widened stereo signal). In other words, a processed version of the right channel signal of the stereo signal is added to the left channel signal of the stereo signal to create the left channel of a widened stereo signal and a processed version of the left channel signal of the stereo signal is added to the right channel signal of the stereo signal to create the right channel of the widened stereo signal. The procedure of deriving the widened stereo signal may further involve pre-filtering (or otherwise processing) each of the left and right channel signals of the stereo signal prior to adding the respective processed contralateral signals thereto in order to preserve desired frequency response in the widened stereo signal.

[0131] Along the lines described above, stereo widening readily generalizes into widening the spatial audio image of a multi-channel input audio signal, thereby deriving an output multi-channel audio signal having a widened spatial audio image (a widened multi-channel signal). In this regard, the processing involves creating the left channel of the widened multi-channel audio signal as a sum of (first) filtered versions of channels of the multi-channel input audio signal and creating the right channel of the widened multi-channel audio signal as a sum of (second) filtered versions of channels of the multi-channel input audio signal. A dedicated predefined filter may be provided for each pair of an input channel (channels of the multi-channel input signal) and an output channel (left and right). As an example in this regard, the left and right channel signals of the widened multi-channel signal $S_{out,left}$ and $S_{out,right}$, respectively, may be defined on basis of channels of a multi-channel audio signal S according to the equation (1):

$$\begin{aligned} S_{out,left}(b,n) &= \sum_i S(i,b,n) H_{left}(i,b) \\ S_{out,right}(b,n) &= \sum_i S(i,b,n) H_{right}(i,b)' \end{aligned} \quad (1)$$

where $S(i, b, n)$ denotes frequency bin b in time frame n of channel i of the multi-channel signal S, $H_{left}(i, b)$ denotes a filter for filtering frequency bin b of channel i of the multi-channel signal S to create a respective channel component for creation of the left channel signal $S_{out,left}(b, n)$, and $H_{right}(i, b)$ denotes a filter for filtering frequency bin b of channel i of the multi-channel signal S to create a respective channel component for creation of the right channel signal $S_{out,right}(b, n)$. $H_{left}(i, b)$ and $H_{right}(i, b)$ are a directional filter pair.

[0132] In stereo widening for headphones, the filters $H_{left}(i, b)$ and $H_{right}(i, b)$ can include HRTFs, or HRTFs (or BRIRs) can be used later in the processing chain. In stereo widening for headphones, the filter $H_{left}(i, b)$ could be HRTFs to 90 degrees (i.e. to left). The filter $H_{right}(i, b)$ could be HRTFs to -90 degrees (i.e. to right).

[0133] In stereo widening for headphones, the filter $H_{left}(i, b)$ can comprise a direct (dry) part and an ambient part comprising one or more indirect (wet) paths.

$$H_{left}(i,b) = r^{\frac{1}{2}} H_{left,direct}(i,b) + (1-r)^{\frac{1}{2}} H_{left,ambient}(i,b)$$

where *r* is the ratio between direct and ambient parts.

[0134] The direct to ambient ratio *r* can be defined by the control input 10.

[0135] The direct part filter $H_{left,direct}(i, b)$ can be HRTFs to 90 degrees (i.e. to left).

[0136] The indirect part filter $H_{left,ambient}(i, b)$ can represent, for each time-frequency tile $S(i, b, n)$, different indirect paths that each has a controlled gain, a controlled reverberation or decorrelation and, optionally, a controlled delay. Each different indirect path is processed using a respective HRTF. The directions of the HRTFs are typically selected so that they cover several directions around the listener, creating a perception of envelopment and/or spaciousness. The filters of the different indirect paths are typically combined to the single filter $H_{left,ambient}(i, b)$ before they are applied.

[0137] Likewise, the filter $H_{right}(i, b)$ can comprise a direct (dry) part and an ambient part comprising one or more indirect (wet) paths.

$$H_{right}(i,b) = r^{\frac{1}{2}} H_{right,direct}(i,b) + (1-r)^{\frac{1}{2}} H_{right,ambient}(i,b)$$

where *r* is the ratio between direct and ambient parts.

[0138] The direct part filter $H_{right,direct}(i, b)$ can be HRTFs to -90 degrees (i.e. to right).

[0139] The indirect part filter $H_{right,ambient}(i, b)$ can represent, for each time-frequency tile $S(i, b, n)$, different indirect paths that each has a controlled gain, a controlled reverberation or decorrelation and, optionally, a controlled delay. Each different indirect path is processed using a respective HRTF. The directions of the HRTFs are typically selected so that they cover several directions around the listener, creating a perception of envelopment and/or spaciousness. The filters of the different indirect paths are typically combined to the single filter $H_{right,\ ambient}(i, b)$ before they are applied.

[0140] The target virtual loudspeaker position indication 12 may be optionally provided to the stereo widening block 112. The indicated virtual loudspeaker positions may then be used to select corresponding HRTFs for H_{left} and H_{right} filters, e.g. for a stereo signal HRTFs to +/-30 degrees were selected by default. However, in order to produce maximally strong widening effect for a stereo signal, HRTFs to +/-90 may be selected instead. To generalize, the stereo widening block 112 may map the indicated virtual loudspeaker positions to modified positions (for stronger widening effect) which are then used to derive the filters H_{lett} and H_{right}.

[0141] Figure 5 illustrates a block diagram of some components and/or entities of the stereo widening processor 112 according to a non-limiting example.

[0142] The stereo widening processor 112 is configured to provide cross-channel mixing means for applying a headphone filter H_{LL}, H_{RL}, H_{LR} and H_{RR} to each one of the plurality of input channels before mixing those channels to produce a modified second signal component 113 comprising two output channels (LEFT, RIGHT), wherein the headphone filter H_{mn} applied to an input channel that is mixed to provide an output channel is dependent upon an identity of the output channel m and an identify of the input channel n.

[0143] The headphone filter H_{mn} can comprise a head related transfer function dependent upon an identity of the output channel m and an identify of the input channel n.

[0144] The headphone filter H_{mn} for an input channel n can be configured to mix a direct-rendering version of the input channel with an ambient-rendering version of the input channel. The relative gain of the direct version of the input channel compared to the ambient version of the input channel in a mix in the headphone filter can be controlled via a user-controllable parameter r. The headphone filter for an input channel can be configured to mix a single-path direct version of the input channel with a multiple-path ambient version of the input channel, where a head related transfer function is used to form the single-path direct version of the input channel and an indirect path filter is used in combination with a head related transfer function for each path of the multiple paths, to form the multiple-path ambient version of the input channel. The indirect path filter can comprise decorrelation means or reverberation means.

[0145] The cross-channel mixing causes stereo-widening for headphone apparatus such that a width of a spatial audio image associated with the modified second signal component is greater than a width of a spatial audio image associated with the second signal component before cross-channel mixing of the second signal component.

[0146] In this example, four filters H_{LL}, H_{RL}, H_{LR} and H_{RR} are applied to create the widened spatial audio image: the left channel of the modified second signal component 113 is created as a sum of the left channel of the second signal component 109-2 filtered by the filter H_{LL} and the right channel of the second signal component 109-2 filtered by the filter H_{LR}, whereas the right channel of the modified second signal component 113 is created as a sum of the left channel of the second signal component 109-2 filtered by the filter H_{RL} and the right channel of the second signal component 109-2 filtered by the filter H_{RR}. In the example of Figure 5, the stereo widening procedure is carried out on basis of the time-domain second signal component 109-2. In other examples, the stereo widening procedure (e.g. one that makes use of the filtering structure of Figure 5) may be carried out in the transform domain. In this alternative example, the order of the inverse transform entity 108-2 and the stereo widening processor 112 is changed.

[0147] In an example, the stereo widening processor 112 may be provided with a dedicated set of filters H_{LL}, H_{RL},

H_{LR} and H_{RR} that is designed to produce a desired extent of stereo widening for a target virtual loudspeaker configuration. In another example, the stereo widening processor 112 may be provided with a plurality of sets of filters H_{LL}, H_{RL}, H_{LR} and H_{RR}, each set designed to produce a desired extent of stereo widening for a target virtual loudspeaker configuration. In the latter example, the set of filters is selected in dependence of the indicated target virtual loudspeaker configuration. In a scenario with a plurality of sets of filters, the stereo widening processor 112 may dynamically switch between sets of filters e.g. in response to a change in the indicated virtual loudspeaker positions. There are various ways for designing a set of filters H_{LL}, H_{RL}, H_{LR} and H_{RR}.

[0148] In stereo widening for headphones, the filter H_{LL} can be filter $H_{left}(left, b)$ described above, the filter H_{LR} can be filter $H_{left}(right, b)$ described above, the filter H_{RR} can be filter $H_{right}(right, b)$ described above, the filter H_{RL} can be filter $H_{right}(left, b)$ described above.

[0149] The stereo-widening performed by the spatial audio processor 112, can be performed in the time domain (FIG 1A) or the transform domain (FIG 1B).

[0150] Referring back to Figure 1A, the audio processing system 100 may comprise the delay element 110 that is arranged to delay the modified first signal component 109-1 by a predefined time delay, thereby creating a delayed first signal component 111. The time delay is selected such that it matches or substantially matches the delay resulting from stereo widening processing applied in the stereo widening processor 112, thereby keeping the delayed first signal component 111 temporally aligned with the modified second signal component 113. In an example, the delay element 110 processes the modified first signal component $s_{dr}(i, m)$ into the delayed first signal component $sd_r(i, m)$. In the example of Figure 1A, the time delay is applied in the time domain. In alternative example, the order of the inverse transform entity 108-1 and the delay element 110 may be changed, thereby resulting in application of the predefined time delay in the transform domain.

[0151] Referring back to Figure 1B, as described in the foregoing, in the audio processing system 100' the delay element 110' is optional and, if included, it is arranged to operate in the transform-domain, in other words to apply the predefined time delay to the modified first signal component 107 to create the delayed modified first signal component 111' in the transform-domain for provision to the combiner signal 114' as a transform-domain signal. It will be appreciated from the foregoing that if one wants to create a perception of a sound source outside the headphones, stereo widening 112 is needed (using, e.g., HRTFs). However, in between the headphones, the sound can be positioned without stereo widening. e.g., re-panning can be used to position sound sources in between the headphones (You cannot position sounds outside the headphones with this method). However, the focus portion contains sounds only near the center, so positioning them in between the headphones is sufficient. The peripheral portion 113 may contain sound sources perceived also outside the headphone positions. The focus portion 111 does not contain sound sources perceived outside the headphone positions, but still they may be wider than they originally were.

[0152] Referring back to Figure 1A, the audio processing system 100 may comprise the signal combiner 114 that is arranged to combine the delayed first signal component 111 and the modified second signal component 113 into the widened stereo signal 115, where the width of spatial audio image is partially extended (in the peripheral but not necessarily the front focus portions) from that of the stereo signal 101. As examples in this regard, the widened stereo signal 115 may be derived as a sum, as an average or as another linear combination of the delayed first signal component 111 and the modified second signal component 113, e.g. according to the equation (22):

$$s_{out}(i,m) = s'_{sw}(i,m) + s'_{dr}(i,m), \qquad (22)$$

where $s_{out}(i, m)$ denotes the widened stereo signal 115.

[0153] Referring back to Figure 1B, as described in the foregoing, in the audio processing system 100' the signal combiner 114' is arranged to operate in the transform-domain, in other words to combine the (transform-domain) delayed modified first signal component 113' with the (transform-domain) modified second signal component 113' into the (transform-domain) widened stereo signal 115' for provision to the inverse transform entity 108'. The inverse transform entity 108' is arranged to convert the (transform-domain) widened stereo signal 115' from the transform domain into the (time-domain) widened stereo signal 115. The transform entity 108' may carry out the conversion in a similar manner as described in the foregoing in context of the transform entities 108-1, 108-2.

[0154] Each of the exemplifying audio processing systems 100, 100' described in the foregoing via a number of examples may further varied in a number of ways. In the following, non-limiting examples in this regard are described.

[0155] In the foregoing, description of elements of the audio processing systems 100, 100' refer to processing of relevant audio signals in a plurality of frequency sub-bands k. In an example, the processing of the audio signal in each element of the audio processing systems 100, 100' is carried out across (all) frequency sub-bands *k*. In other examples, in at least some elements of the audio processing systems 100, 100' the processing of the audio signal is carried out in a limited number of frequency sub-bands *k*. As examples in this regard, the processing in a certain element of the audio

processing system 100, 100' may be carried out for a predefined number of lowest frequency sub-bands *k*, for a predefined number of highest frequency sub-bands *k,* or for a predefined subset of frequency sub-bands *k* in the middle of the frequency range such that a first predefined number of lowest frequency sub-bands *k* and a second predefined number of highest frequency sub-bands *k* is excluded from the processing. The frequency sub-bands *k* excluded from the processing (e.g. ones at the lower end of the frequency range and/or ones at the higher end of the frequency range) may be passed unmodified from an input to an output of the respective element. As a non-limiting example concerning elements of the audio processing systems 100, 100' where the processing may be carried out only for a limited subset of frequency sub-bands *k*, involves one or both of the re-panner 116 and the stereo widening processor 112, 112', which may only process the respective input signal in a respective desired sub-range of frequencies, e.g. in a predefined number of lowest frequency sub-bands *k* or in a predefined subset of frequency sub-bands *k* in the middle of the frequency range.

[0156] In another example, as already described in the foregoing, the input audio signal 101 may comprise a multi-channel signal different from a two-channel stereophonic audio signal, e.g. surround signal. For example in case the input audio signal 101 comprises a 5.1--channel surround signal, the audio processing technique(s) described in the foregoing with references to the left and right channels of the stereo signal 101 may be applied to the front left and front right channels of the 5.1-channel surround signal to derive the left and right channels of the output audio signal 115. The other channels of the 5.1-channel surround signal may be processed e.g. such that the center channel of the 5.1-channels surround signal scaled by a predefined gain factor (e.g. by one having value $\sqrt{0.5}$) is added to the left and right channels of the output audio signal 115 obtained from the audio processing system 100, 100', whereas the rear left and right channels of the 5.1-channel surround signal may be processed using a conventional stereo widening technique that makes use of widening filter(s) (utilizing, e.g., HRTFs or BRIRs)) that correspond(s) to respective target positions of the left and right rear loudspeakers (e.g. $\pm$110 degrees with respect to the front direction). The LFE channel of the 5.1-channel surround signal may be added to the center signal of the 5.1-channel surround signal prior to adding the scaled version thereof to the left and right channels of the output audio signal 115.

[0157] In another example, as already described in the foregoing, the input audio signal 101 may comprise N spatially distributed channels that are processed to produce a two-channel audio signal 115 processed specifically for playback via headphone apparatus. The mixing of M channels to produce a first signal component 111, 111' of the two-channel stereophonic audio signal 115 can occur at re-panner 106. The mixing of M' channels to produce a second signal component 113, 113' of the two-channel stereophonic audio signal 115 can occur at the stereo widening processor for headphone apparatus 112.

[0158] Audio events (sound objects) may move within the sound image. When an audio event (sound object) is positioned within the focus range the audio event is rendered via the first signal component 111, 111' of the two-channel stereophonic audio signal 115. When an audio event is positioned within the non-focus, peripheral range the audio event is rendered via the second signal component 113, 113' of the two-channel stereophonic audio signal 115.

[0159] In another example, additionally or alternatively, the audio processing system 100, 100' may enable adjusting balance between the contribution from the first signal component 105-1 and the second signal component 105-2 in the resulting widened stereo signal 115. This may be provided, for example, by applying respective different scaling gains to the first signal component 105-1 (or a derivative thereof) and to the second signal component 105-2 (or a derivative thereof). In this regard, respective scaling gains may be applied e.g. in the signal combiner 114, 114' to scale the signal components derived from the first and second signal components 105-1, 105-2 accordingly, or in the signal divider 126 to scale the first and second signal components 105-1, 105-2 accordingly. A single respective scaling gain may be defined for scaling the first and second signal components 105-1, 105-2 (or a respective derivative thereof) across all frequency sub-bands or in predefined sub-set of frequency sub-bands. Alternatively or additionally, different scaling gains may be applied across the frequency sub-bands, thereby enabling adjustment of the balance between the contribution from the first and second signal components 105-1, 105-2 only on some of the frequency sub-bands and/or adjusting the balance differently at different frequency sub-bands.

[0160] In a further example, alternatively or additionally, the audio processing system 100, 100' may enable scaling of one or both of the first signal component 105-1 and the second signal component 105-2 (or respective derivatives thereof) independently of each other, thereby enabling equalization (across frequency sub-bands) for one or both of the first and second signal components. This may be provided, for example, by applying respective equalization gains to the first signal component 105-1 (or a derivative thereof) and to the second signal component 105-2 (or a derivative thereof). A dedicated equalization gain may be defined for one or more frequency sub-bands for the first signal component 105-1 and/or for the second signal component 105-2. In this regard, for each of the first and second signal components 105-1, 105-2, a respective equalization gain may be applied e.g. in the signal divider 126 or in the signal combiner 114, 114' to scale a respective frequency sub-band of the respective one of the first and second signal components 105-1, 105-2 (or a respective derivative thereof). For a certain frequency sub-band, the equalization gain may be the same for both the first and second signal components 105-1, 105-2 or different equalization gains be applied for the first and

second signal component 105-1, 105-2.

[0161] Operation of the audio processing system 100, 100' described in the foregoing via multiple examples enables adaptively decomposing the stereo signal 101 into the first signal component 105-1 that represents the focus portion of the spatial audio image and that is provided for playback without application of stereo widening thereto and into the second signal component 105-2 that represents peripheral (non-focus) portion of the spatial audio image that is subjected to the stereo widening processing. In particular, since the decomposition is carried out on basis of audio content conveyed by the stereo signal 101 on frame by frame basis, the audio processing system 100, 100' enables both adaptation for relatively static spatial audio images of different characteristics and adaptation to changes in the spatial audio image over time.

[0162] The disclosed stereo widening technique that relies on excluding coherent sound sources within the focus portion of the spatial audio image from the stereo widening processing and applies the stereo widening processing predominantly to coherent sounds that are outside the focus portion and to non-coherent sounds (such as ambience) enables improved timbre and reduced 'coloration' of sounds that are within the focus portion while still providing a large extent of perceivable stereo widening.

[0163] In the previous examples, the control input 10 can have one or more different functions.:
The parameters of the decomposition process can be defined by the control input. The control input 10 can for example define the focus range used in the analysis for dividing the signals to focus (i.e. front center) and non-focus (i.e. side) signals. The focus range can, for example, be defined via θ_{Th1} and θ_{Th2} or β_{Th} The signal decomposition parameter β_{Th} can, for example, be defined by the control input 10

[0164] The control input 10 can for example control relative gains between the peripheral signals 113, 113' that are widened and the frontal signals 111, 111' that are not. For example, it can in some examples control a relative gain ratio of peripheral to frontal.

[0165] The parameters of the widening process can for example be defined by the control input 10. The control input 10 can, for example, control the direct to ambient ratio r used in widening. The parameters may include for example the directions to which the non-focus sounds are processed (for example with the help of HRTF processing), and/or the amount of ambience (for example reverb) added to sound for increasing the "widening" effect or the perceived externalization. Processing the non-focus sounds to different virtual directions is not necessary, one embodiment of the invention can be such that the non-focus sounds are processed only using reverb, decorrelator or other methods which increase the externalization of the non-focus sounds.

[0166] The control input 10 can for example control explicitly or implicitly whether or not panning occurs. For example, panning may not occur if the focus range is narrow. For example, panning may not occur if the relative gain ratio of peripheral to frontal is small.

[0167] The value of the mapping coefficient μ that controls panning extent can, for example, be defined explicitly by the control input 10 or can be controlled via definition of the focus range. The overpan factor μ can be used for modifying the front center sector (i.e. focus sounds) within which the focus signal is perceived (for example, it can be made sound wider than in the original signal). The control input 10 can be also another parameter or a set of parameters which modify where the focus sounds are heard in the left - right panning dimension.

[0168] The weighting factors for energy-based temporal smoothing (*a* and *b*) can, for example, be defined by the control input 10.

[0169] All, part or none of the control input can, for example, be controlled by user input.

[0170] The control input 10 can for example comprise parameters for controlling the focus sounds (e.g. for adding ambience to produce better externalization to front sounds).

[0171] The control input 10 can for example comprise parameters that define multiple analysis sectors (for the decomposition part) and multiple virtual speaker directions (used in the stereo widening block). Non-focus sounds may be divided to more sectors than just left and right (outside of the focus range). There may be several angular regions outside of the focus range, which may be processed separately to e.g. different directions or different amounts of ambience in the invention report).

[0172] Components of the audio processing system 100, 100' may be arranged to operate, for example, in accordance with a method 200 illustrated by a flowchart depicted in Figure 6. The method 200 serves as a method for processing an input audio signal comprising a multi-channel audio signal that represents a spatial audio image.

[0173] The method 200 comprises:

at block 202: deriving, based on the input audio signal 101, a first signal component 105-1 comprising a at least one input channel and a second signal component 105-2 comprising multiple input channels, wherein the first signal component 105-1 is dependent upon at least a first (focus) portion of a spatial audio image conveyed by the input audio signal 101, and the second signal component 105-1 is dependent upon at least a second (non-focus) portion of the spatial audio image that is different to the first (focus) portion.

[0174] The method 200 further comprises, at block 204, cross-channel mixing of at least some of the multiple input channels of the second signal component 105-2 to produce a modified second signal component 113 while enabling the first signal component to bypass cross-channel mixing

[0175] The method 200 further comprises, at block 206, combining the first signal component 105-2 and the modified second signal component 113 into an output audio signal 115 comprising two output channels configured for rendering by headphone apparatus.

The method 200 may be varied in a number of ways, for example in view of the examples concerning operation of the audio processing system 100 and/or the audio processing system 100' described in the foregoing.

The cross-channel mixing enables a width of the spatial audio image to be extended from that of the second signal component 105-2

[0176] Figure 7 illustrates a block diagram of some components of an exemplifying apparatus 300. The apparatus 300 may comprise further components, elements or portions that are not depicted in Figure 7. The apparatus 300 may be employed e.g. in implementing one or more components described in the foregoing in context of the audio processing system 100, 100'. The apparatus 300 may implement, for example, the device 50 or one or more components thereof.

[0177] The apparatus 300 comprises a processor 316 and a memory 315 for storing data and computer program code 317. The memory 315 and a portion of the computer program code 317 stored therein may be further arranged to, with the processor 316, to implement at least some of the operations, procedures and/or functions described in the foregoing in context of the audio processing system 100, 100'.

[0178] The apparatus 300 comprises a communication portion 312 for communication with other devices. The communication portion 312 comprises at least one communication apparatus that enables wired or wireless communication with other apparatuses. A communication apparatus of the communication portion 312 may also be referred to as a respective communication means.

[0179] The apparatus 300 may further comprise user I/O (input/output) components 318 that may be arranged, possibly together with the processor 316 and a portion of the computer program code 317, to provide a user interface for receiving input from a user of the apparatus 300 and/or providing output to the user of the apparatus 300 to control at least some aspects of operation of the audio processing system 100, 100' implemented by the apparatus 300. The user I/O components 318 may comprise hardware components such as a display, a touchscreen, a touchpad, a mouse, a keyboard, and/or an arrangement of one or more keys or buttons, etc. The user I/O components 318 may be also referred to as peripherals. The processor 316 may be arranged to control operation of the apparatus 300 e.g. in accordance with a portion of the computer program code 317 and possibly further in accordance with the user input received via the user I/O components 318 and/or in accordance with information received via the communication portion 312.

[0180] Although the processor 316 is depicted as a single component, it may be implemented as one or more separate processing components. Similarly, although the memory 315 is depicted as a single component, it may be implemented as one or more separate components, some or all of which may be integrated/removable and/or may provide permanent / semi-permanent/ dynamic/cached storage.

[0181] The computer program code 317 stored in the memory 315, may comprise computer-executable instructions that control one or more aspects of operation of the apparatus 300 when loaded into the processor 316. As an example, the computer-executable instructions may be provided as one or more sequences of one or more instructions. The processor 316 is able to load and execute the computer program code 317 by reading the one or more sequences of one or more instructions included therein from the memory 315. The one or more sequences of one or more instructions may be configured to, when executed by the processor 316, cause the apparatus 300 to carry out at least some of the operations, procedures and/or functions described in the foregoing in context of the audio processing system 100, 100'.

[0182] Hence, the apparatus 300 may comprise at least one processor 316 and at least one memory 315 including the computer program code 317 for one or more programs, the at least one memory 315 and the computer program code 317 configured to, with the at least one processor 316, cause the apparatus 300 to perform at least some of the operations, procedures and/or functions described in the foregoing in context of the audio processing system 100, 100'.

[0183] The computer program(s) stored in the memory 315 may be provided e.g. as a respective computer program product comprising at least one computer-readable non-transitory medium having the computer program code 317 stored thereon, the computer program code, when executed by the apparatus 300, causes the apparatus 300 at least to perform at least some of the operations, procedures and/or functions described in the foregoing in context of the audio processing system 100, 100'. The computer-readable non-transitory medium may comprise a memory device or a record medium such as a CD-ROM, a DVD, a Blu-ray disc or another article of manufacture that tangibly embodies the computer program. As another example, the computer program may be provided as a signal configured to reliably transfer the computer program.

[0184] Reference(s) to a processor should not be understood to encompass only programmable processors, but also dedicated circuits such as field-programmable gate arrays (FPGA), application specific circuits (ASIC), signal processors, etc. Features described in the preceding description may be used in combinations other than the combinations explicitly described.

[0185] In at least some of the preceding examples, when the input audio signal 101 comprises a same sound source that is repeated at different positions, and
that is rendered at the headphone apparatus 20 without interaural time difference and without frequency dependent interaural level differences, when the sound source of the input audio signal 101 is positioned at a first position that is relatively front and central to a user of the headphone apparatus 30,
then the sound source is rendered at the headphone apparatus 30 with interaural time differences and with frequency dependent interaural level differences when the sound source of the input audio signal is repeated at a second position that is relatively peripheral and is not front and central to a user of the headphone apparatus 30.

[0186] The stereo-widening (for headphones) processor 112, 112' spatially processes the input audio signal 101 to add at peripheral positions, but not at central positions, of the spatial audio image positionally-dependent interaural time differences measurable between coherent audio events in both of the channels of the output audio signal and frequency-dependent and positionally-dependent interaural level differences measurable between coherent audio events in both of the channels of the output audio signal.

[0187] In the foregoing examples, there is a bypass initiated by the signal decomposer 104 and provided via a bypass route comprising the re-panner 106 thus enabling the first signal component 105-1 to bypass the stereo-widening (for headphones) processor 112, 112'. In some but not necessarily all examples, the bypass enables components of the input audio signal 101 that represent a sound source that is coherent between two stereo channels and is positioned to front and center, to bypass cross-channel mixing at the stereo-widening (for headphones) processor 112, 112'.

[0188] In at least some of the above examples, first focus portion is front and central relative to a user of the headphone apparatus, and the second portion is peripheral relative to a user of headphone apparatus. In at least some of the above examples, first focus portion does not overlap the first portion. In at least some of the above examples, the first focus portion and second non-focus portions are contiguous.

[0189] Although the above description discusses an implementation in which there is a central first focus portion and two second focus portions to left and right split by the first focus portion, other arrangements of the first focus portion and the second focus portion are possible. Reference to a portion may, for example, reference a single portion or multiple portions.

[0190] Where the second portion comprises multiple portions, then different spatial audio processing can be applied to each of the second portions. For example, different control inputs may be used for different second portions. The same control inputs may be used for different second portions that are disposed symmetrically either side of a central direction. For example, different cross-channel mixing may be used for different second portions to achieve different widening effects. The same cross-channel mixing may be used for different second portions that are disposed symmetrically either side of a central direction. For example, different direct to ambient rations r may be used for different second portions to achieve different effects. The same direct to ambient ratio r may be used for different second portions that are disposed symmetrically either side of a central direction.

[0191] Where the first portion comprises multiple portions, then different processing e.g. re-panning can be applied to each of the second portions.

[0192] In the foregoing examples, the first (focus) portion is fixed in the audio image when the headphone apparatus move and the audio image is oriented with respect to the headphone apparatus. In the other examples, the audio image is oriented with respect to the 'world' headphone apparatus and is processed to rotated when the headphones rotate. In this example, the first (focus) portion can be fixed in the audio image when the headphone apparatus move or alternatively can rotate with the headphone apparatus. The headphone apparatus 20 can comprise circuitry for tracking it's orientation.

[0193] In some examples the apparatus 100,100' is separate to the headphone apparatus 20, for example as illustrated in Figure 3. In other examples, the apparatus 100, 100' is part of the headphone apparatus 20.In at least some of the examples described above, audio is divided into two paths, central and side sound. For central sounds, timbre is important, so the processing is designed in order to keep that good. HRTF processing is avoided. The central sounds can be widened by, for example, "re-panning" which does not degrade timbre, and does some widening, even though it cannot create sources outside the headphones. For side sounds, having very wide perception is the most important thing. Hence, HRTFs are used to get that effect (and provide sound sources outside the headphones). This degrades the timbre, but that is accepted as a trade-off in order to get that maximal wideness. While one keeps timbre for central sounds, it is desirable to make them wide. Side sounds are made very wide.

[0194] Although in the foregoing some functions have been described with reference to certain features and/or elements, those functions may be performable by other features and/or elements whether described or not. Although features have been described with reference to certain embodiments, those features may also be present in other embodiments whether described or not.

Claims

1. An apparatus for processing an input audio signal comprising multiple channels, the apparatus comprising:

   means for deriving, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion;
   cross-channel mixing means for cross-channel mixing of a plurality of input channels;
   means for directing the second signal component to the cross-channel mixing means for cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component;
   bypass means for enabling the first signal component to bypass the cross-channel mixing means; and
   means for combining the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

2. An apparatus as claimed in claim 1, wherein the cross-channel mixing means for cross-channel mixing of a plurality of input channels comprises means for applying head related transfer functions to each one of the plurality of input channels before mixing those channels to produce a modified second signal component comprising two output channels, wherein the head related transfer function applied to an input channel that is mixed to provide an output channel is dependent upon an identity of the input channel and an identify of the output channel.

3. An apparatus as claimed in claim 1 or 2, wherein the cross-channel mixing means for cross-channel mixing of a plurality of input channels comprises means for applying a headphone filter to each one of the plurality of input channels before mixing those channels to produce a modified second signal component comprising two output channels, wherein the headphone filter applied to an input channel that is mixed to provide an output channel is dependent upon an identity of the input channel and an identify of the output channel, wherein the headphone filter for an input channel mixes a direct version of the input channel with an ambient version of the input channel.

4. An apparatus as claimed in claim 3, wherein the relative gain of the direct version of the input channel compared to the ambient version of the input channel in a mix in the headphone filter is a user-controllable parameter.

5. An apparatus as claimed in claim 3 or 4, wherein the headphone filter for an input channel mixes a single-path direct version of the input channel with a multiple-path ambient version of the input channel;

   wherein a head related transfer function is used to form the single-path direct version of the input channel;
   wherein, an indirect path filter is used in combination with a head related transfer function for each path of the multiple paths, to form the multiple-path ambient version of the input channel.

6. An apparatus as claimed in claim 5, wherein the indirect path filter comprises decorrelation means or reverberation means.

7. An apparatus as claimed in any preceding claim, wherein the cross-channel mixing causes stereo-widening for headphone apparatus such that a width of a spatial audio image associated with the modified second signal component is greater than a width of a spatial audio image associated with the second signal component before cross-channel mixing of the second signal component.

8. An apparatus as claimed in any preceding claim, wherein the first portion is front and central relative to a user of the headphone apparatus, and the second portion is peripheral relative to the user of headphone apparatus and does not overlap the first portion.

9. An apparatus as claimed in any preceding claim, wherein the first and second portions are contiguous.

10. An apparatus as claimed in any preceding claim, wherein the bypass means enables components of the input audio signal that represent a sound source that is coherent between two stereo channels and is positioned to front and center, to bypass the cross-channel mixing means.

11. An apparatus as claimed in any preceding claim, wherein a control input controls one or more of:

control the first portion and/or the second portion;
control decomposition of input signal to first component and second component;
control relative gain of the first component and the second component;
control widening of the second component;
control ratio of direct to ambient gain during widening of second component;
control panning of first component;
control whether there is or is not panning of the first component;
control panning extent of first component; and
control energy-based temporal smoothing.

12. An apparatus as claimed in any preceding claim, wherein when the input audio signal comprises a same sound source that is repeated at different positions, and
that is rendered at the headphone apparatus without interaural time difference and without frequency dependent interaural level differences, when the sound source of the input audio signal is positioned at a first position that is relatively front and central to a user of the headphone apparatus, then the sound source is rendered at the headphone apparatus with interaural time differences and with frequency dependent interaural level differences when the sound source of the input audio signal is repeated at a second position that is relatively peripheral and is not front and central to a user of the headphone apparatus.

13. An apparatus as claimed in any of claims 1 to 12, configured as a headphone apparatus for rendering the output audio signal.

14. An apparatus as claimed in claim 13, wherein the headphone apparatus is configured to produce a spatial audio image, and further comprising:

means for processing the input audio signal comprising the multiple channels to produce a two-channel output audio signal configured for rendering;
means for spatially processing the input audio signal to add at peripheral positions, but not at central positions, of the spatial audio image positionally-dependent interaural time differences measurable between coherent audio events in both of the channels of the output audio signal and frequency-dependent and positionally-dependent interaural level differences measurable between coherent audio events in both of the channels of the output audio signal.

15. A method for processing an input audio signal comprising a at least one input channel/multiple input channels, the method comprising:

deriving, based on the input audio signal, a first signal component, comprising at least one input channel, and a second signal component, comprising multiple input channels, wherein
the first signal component is dependent upon at least a first portion of a spatial audio image conveyed by the input audio signal, and the second signal component is dependent upon at least a second portion of the spatial audio image that is different to the first portion;
cross-channel mixing of at least some of the multiple input channels of the second signal component to produce a modified second signal component while enabling the first signal component to bypass cross-channel mixing; and
combining the first signal component and the modified second signal component into an output audio signal comprising two output channels configured for rendering by headphone apparatus.

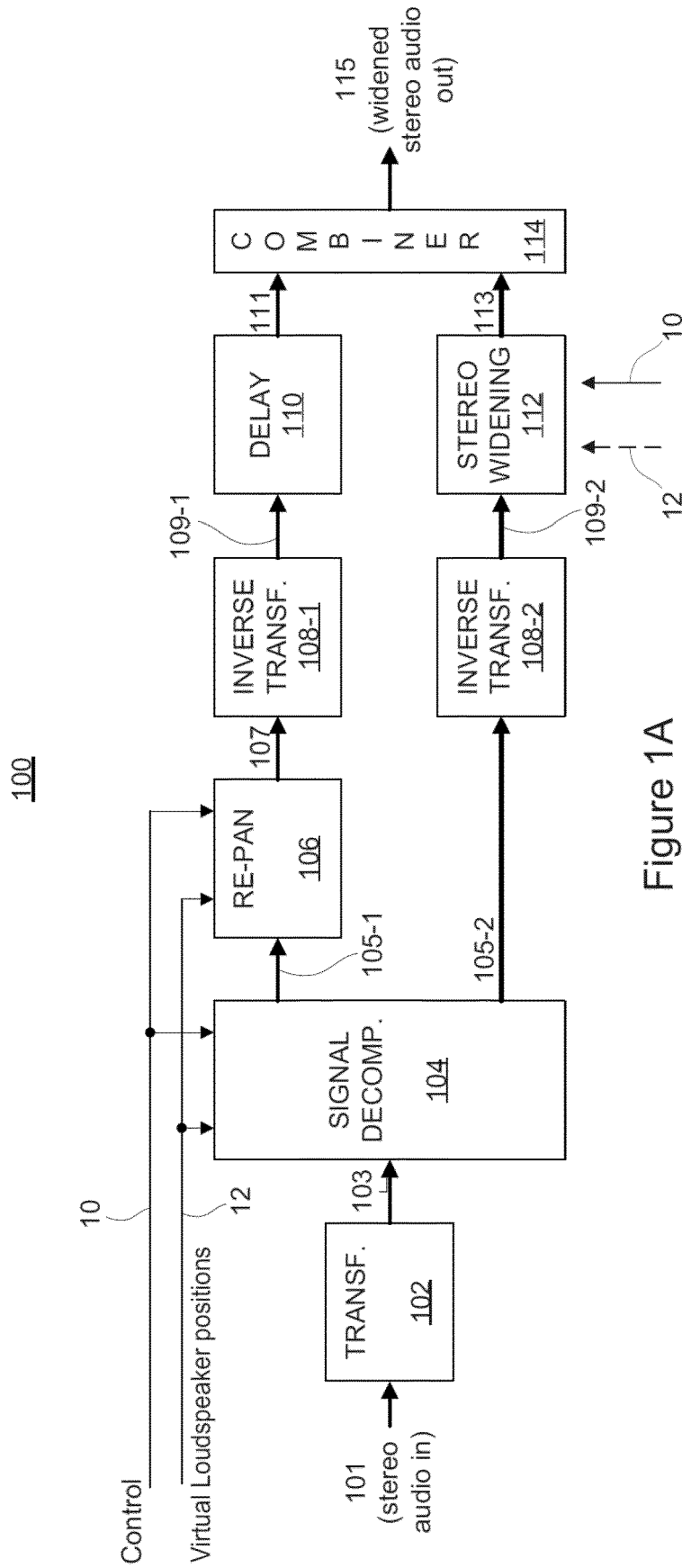
100
101
(stereo
audio in)
TRANSF.
102
103
SIGNAL
DECOMP.
104
105-1
105-2
RE-PAN
106
107
INVERSE
TRANSF.
108-1
INVERSE
TRANSF.
108-2
109-1
109-2
DELAY
110
111
STEREO
WIDENING
112
113
C O M B I N E R
114
115
(widened
stereo audio
out)
Control
10
Virtual Loudspeaker positions
12
10
12

Figure 1A

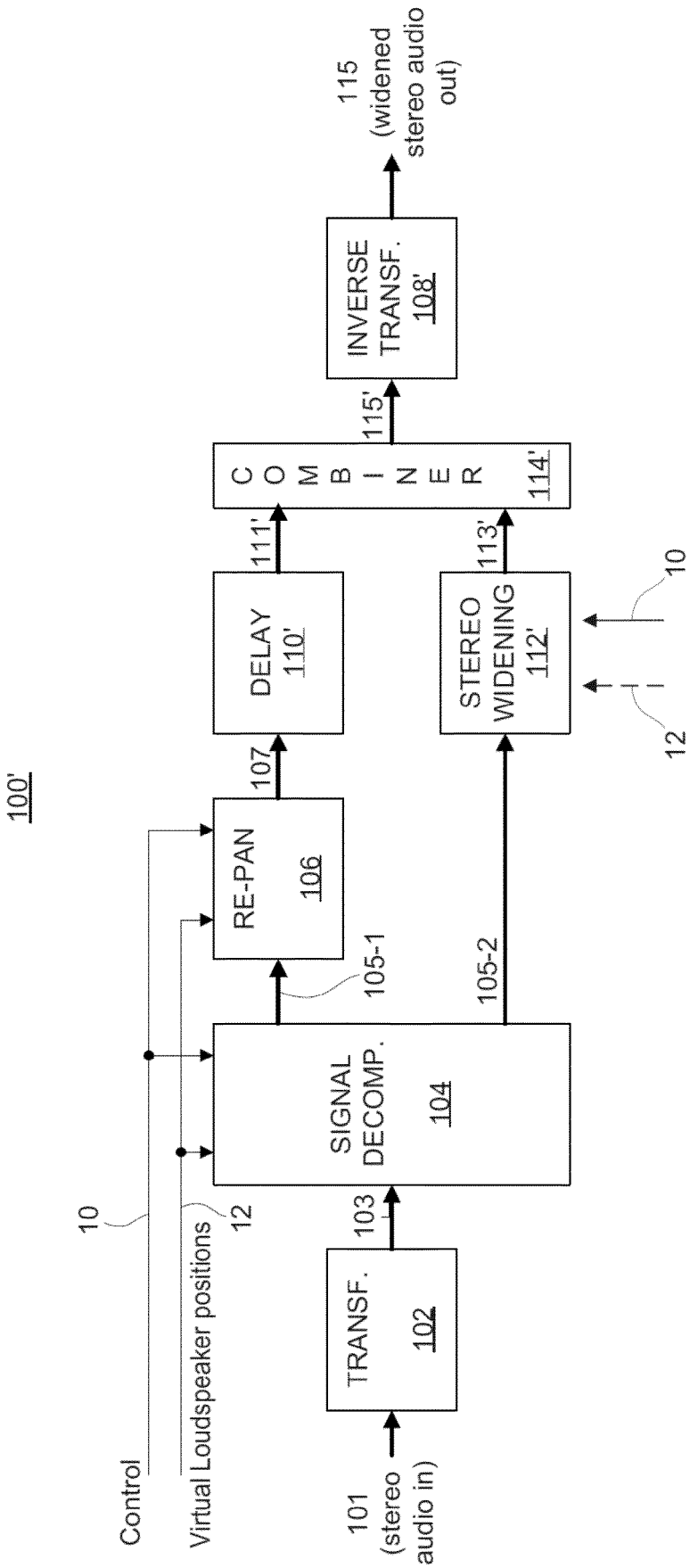
100'
101 (stereo audio in)
TRANSF. 102
103
SIGNAL DECOMP. 104
105-1
105-2
RE-PAN 106
107
DELAY 110'
111'
STEREO WIDENING 112'
113'
COMBINER 114'
115'
INVERSE TRANSF. 108'
115 (widened stereo audio out)
Control
10
Virtual Loudspeaker positions
12
10
12

Figure 1B

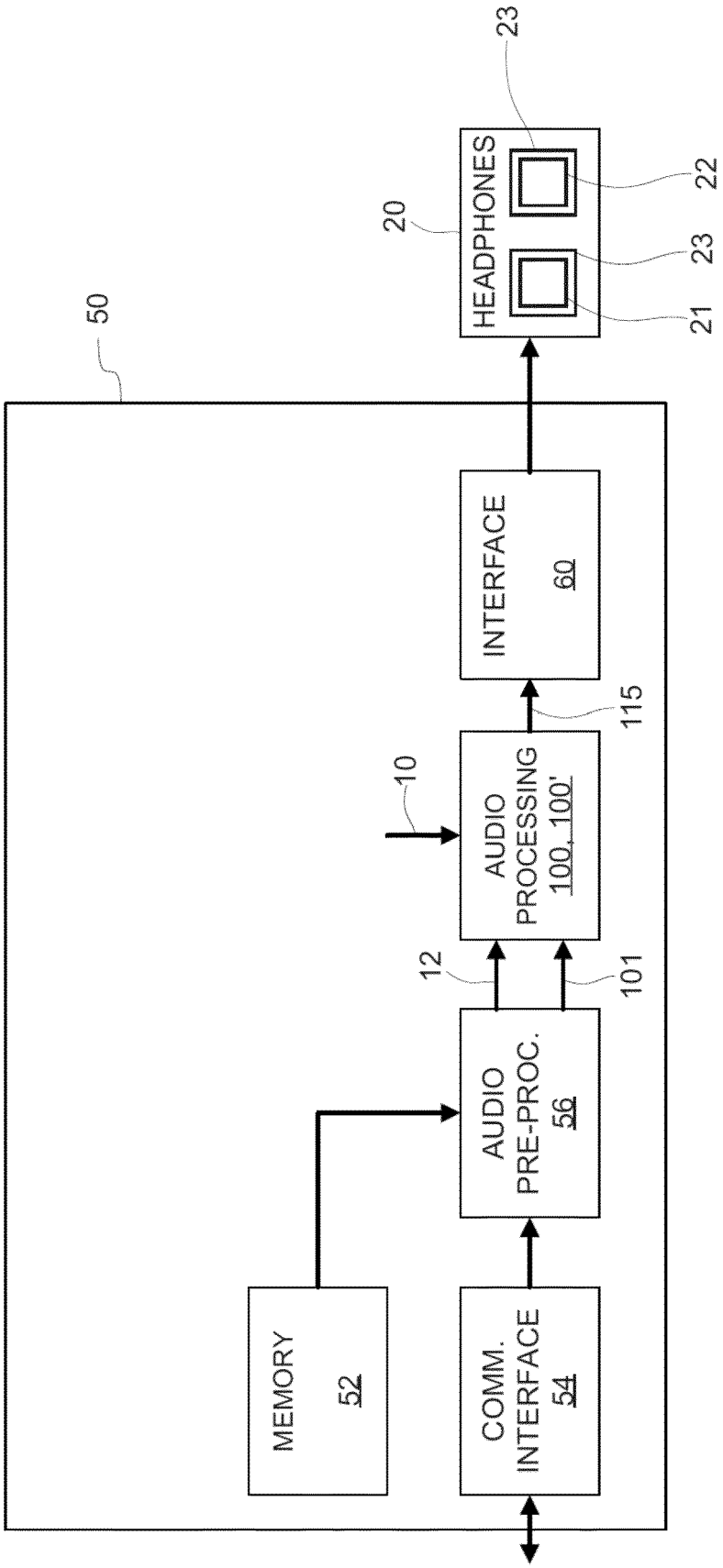
50
HEADPHONES
20
23
22
23
21
INTERFACE
60
115
10
AUDIO
PROCESSING
100, 100'
12
101
AUDIO
PRE-PROC.
56
MEMORY
52
COMM.
INTERFACE
54

Figure 2

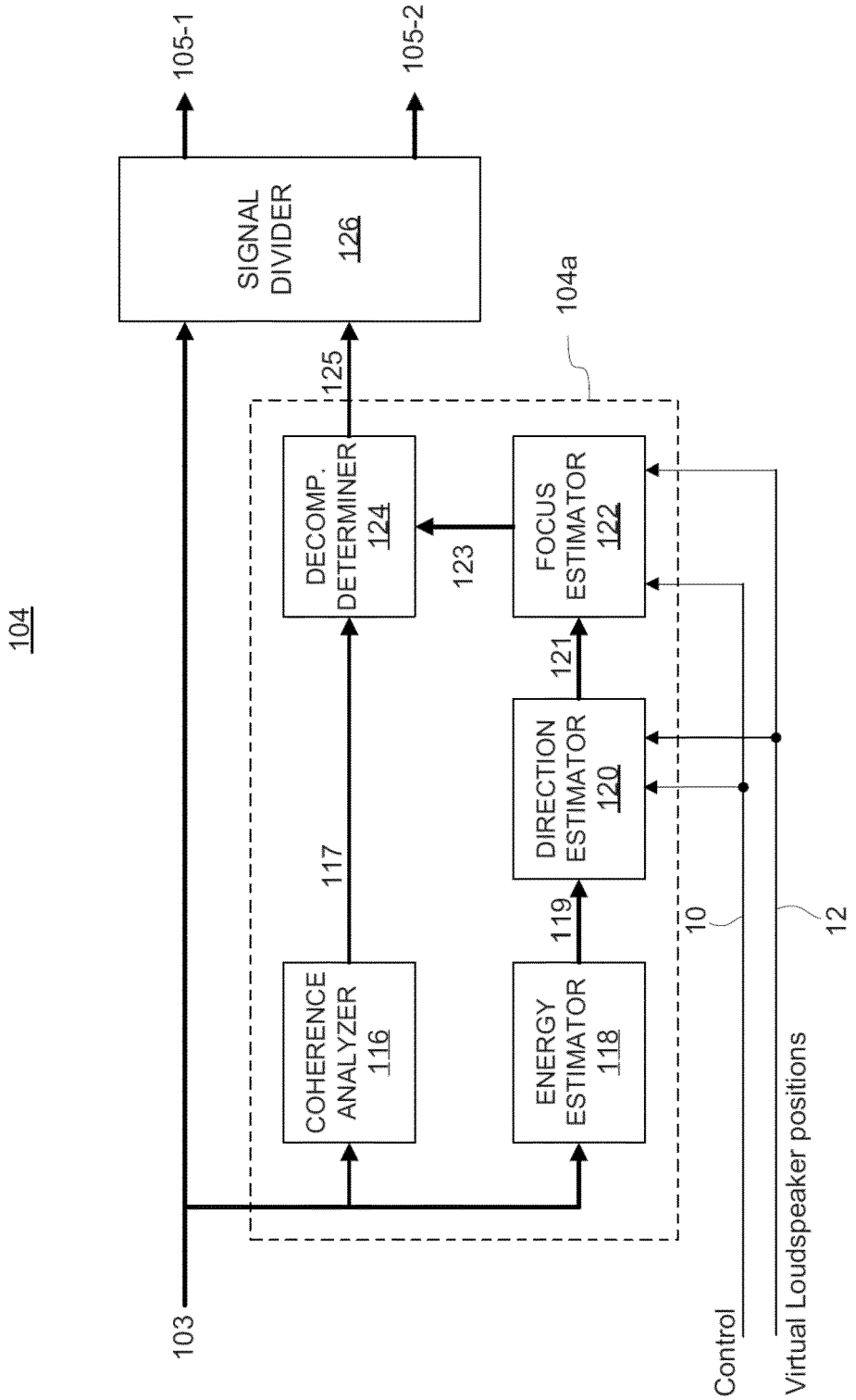
104
105-1
105-2
SIGNAL DIVIDER 126
104a
125
DECOMP. DETERMINER 124
123
FOCUS ESTIMATOR 122
121
DIRECTION ESTIMATOR 120
117
119
COHERENCE ANALYZER 116
ENERGY ESTIMATOR 118
10
12
103
Control
Virtual Loudspeaker positions

Figure 3

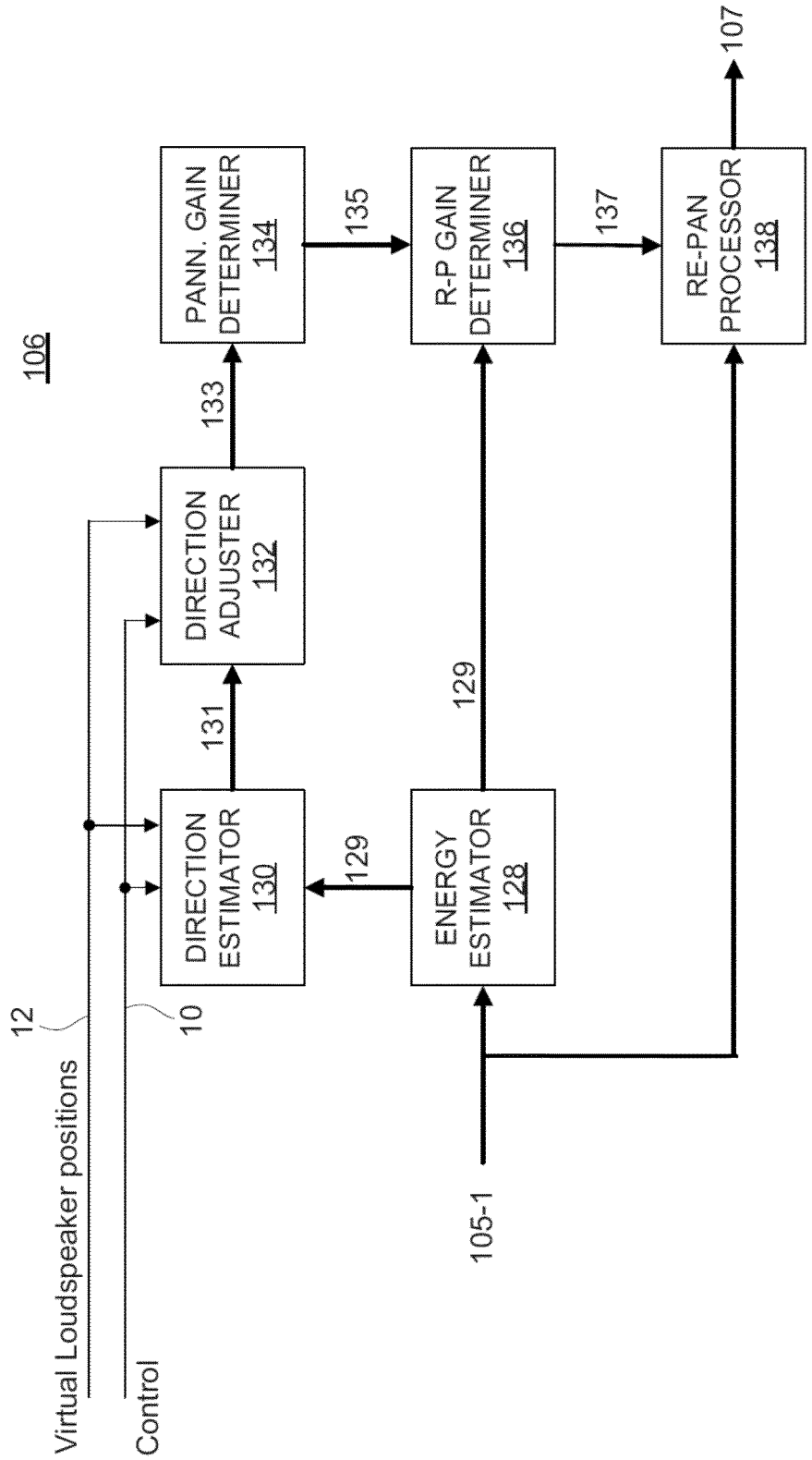
106
PANN. GAIN DETERMINER 134
135
R-P GAIN DETERMINER 136
137
RE-PAN PROCESSOR 138
107
133
DIRECTION ADJUSTER 132
131
129
DIRECTION ESTIMATOR 130
129
ENERGY ESTIMATOR 128
105-1
12
Virtual Loudspeaker positions
10
Control

Figure 4

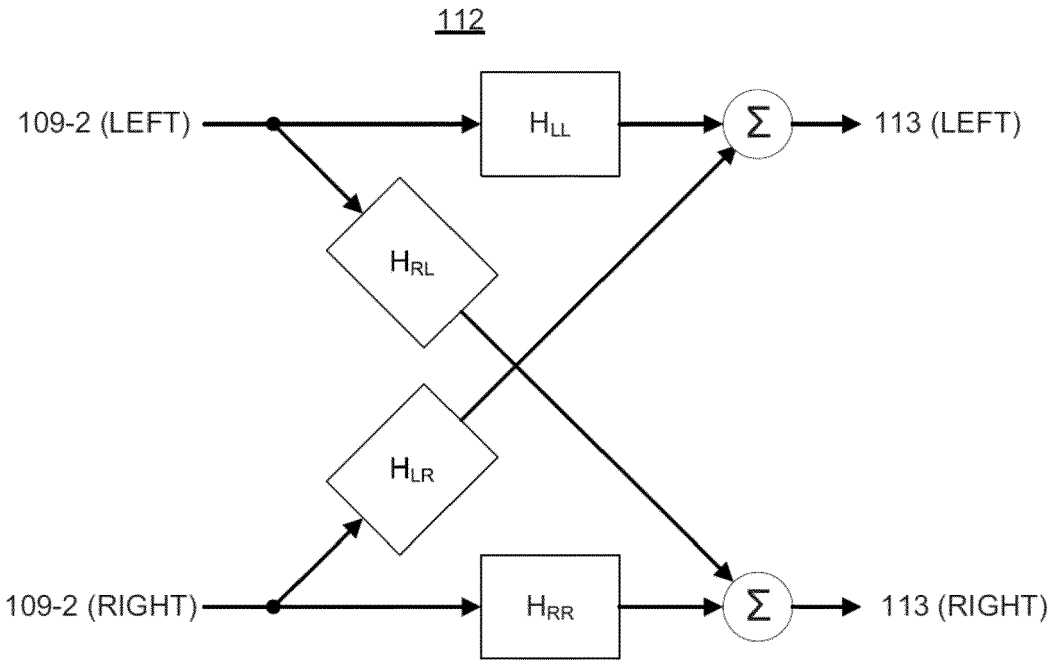

Figure 5

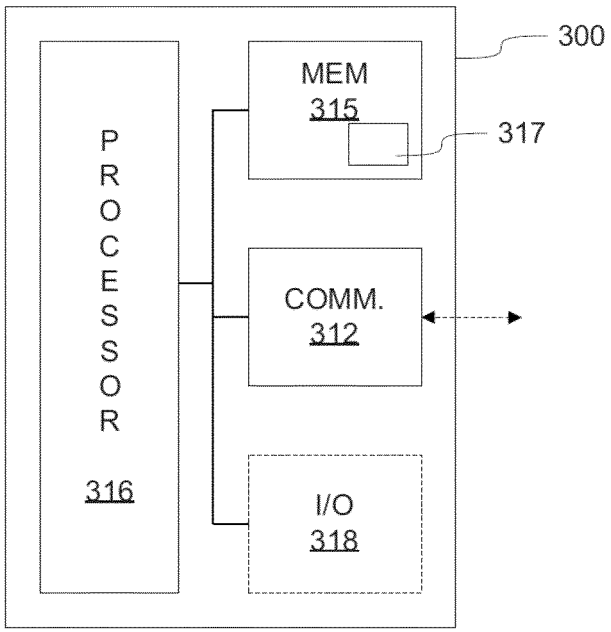

Figure 7

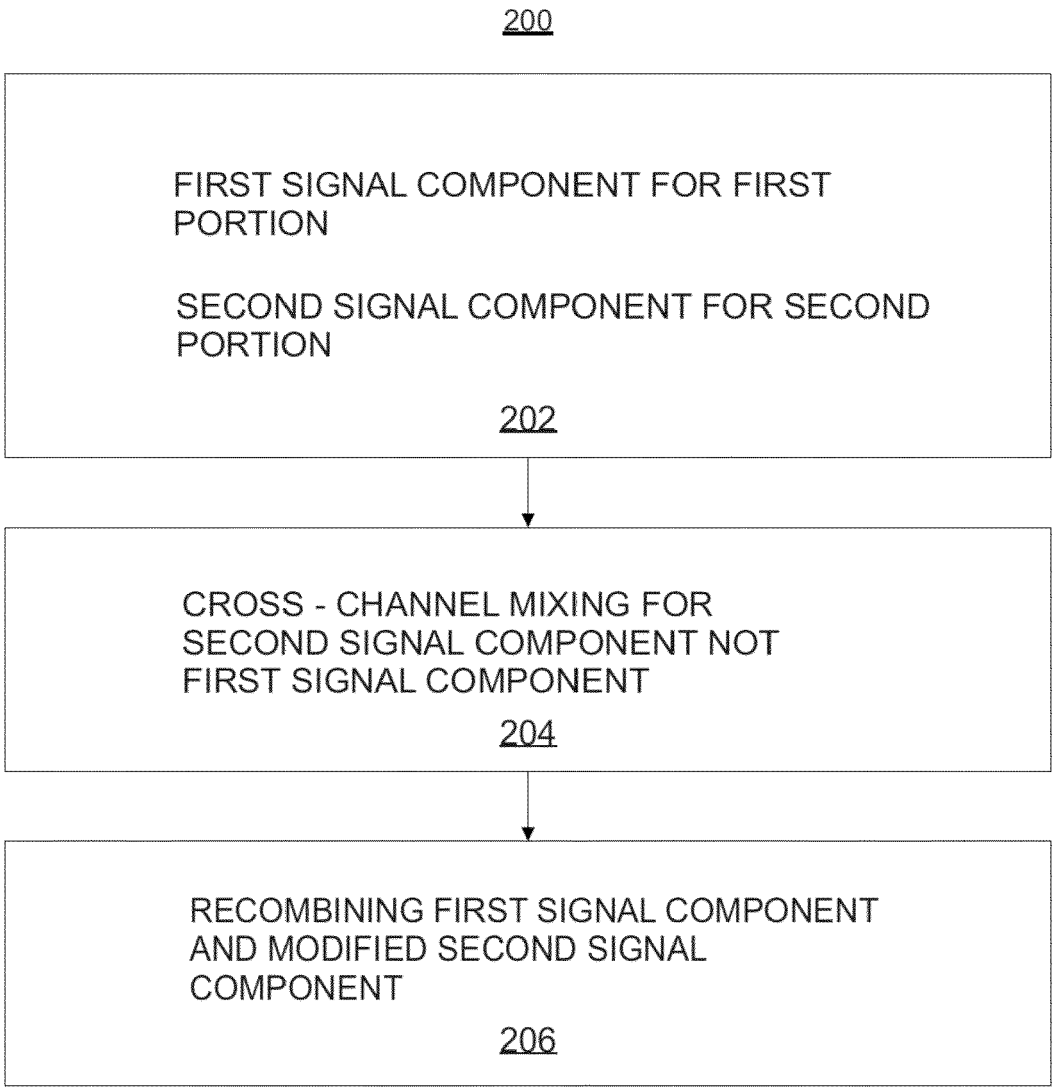
200
FIRST SIGNAL COMPONENT FOR FIRST PORTION
SECOND SIGNAL COMPONENT FOR SECOND PORTION
202
CROSS - CHANNEL MIXING FOR SECOND SIGNAL COMPONENT NOT FIRST SIGNAL COMPONENT
204
RECOMBINING FIRST SIGNAL COMPONENT AND MODIFIED SECOND SIGNAL COMPONENT
206

Figure 6