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

(11) EP 4 531 039 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
02.04.2025 Bulletin 2025/14

(21) Application number: **23199838.6**

(22) Date of filing: **26.09.2023**

(51) International Patent Classification (IPC):
G10L 19/008 *(2013.01)* *G10L 19/025* *(2013.01)*

(52) Cooperative Patent Classification (CPC):
G10L 19/008; G10L 19/025

(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 ME MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA
Designated Validation States:
KH MA MD TN

(71) Applicant: **Koninklijke Philips N.V.**
5656 AG Eindhoven (NL)

(72) Inventors:
- **SCHUIJERS, Erik Gosuinus Petrus**
  Eindhoven (NL)
- **KECHICHIAN, Patrick**
  Eindhoven (NL)
- **RAVI, Akshaya**
  Eindhoven (NL)

(74) Representative: **Philips Intellectual Property & Standards**
High Tech Campus 52
5656 AG Eindhoven (NL)

(54) GENERATION OF MULTICHANNEL AUDIO SIGNAL AND AUDIO DATA SIGNAL REPRESENTING A MULTICHANNEL AUDIO SIGNAL

(57) A decoder audio apparatus comprises a receiver (101) receiving an audio data signal comprising a downmix audio signal being a downmix of a multichannel audio signal, sets of upmix parameters including a level difference parameter, a correlation parameter, and a phase difference parameter as well as at least one transient upmix parameter indicative of an interchannel level difference for a transient. A first upmixer (103) generates an upmixed multichannel signal by upmixing the downmix audio signal in dependence on the upmix parameters and a second upmixer (105) generate an transient audio component by upmixing of the a transient audio component in dependence on the transient upmix parameter. A generator (107) generates an output multichannel signal from a combination of the upmixed multichannel signal and the transient audio component. An encoder audio apparatus may from a received multichannel audio signal generate the audio data signal comprising the downmix audio signal, the sets of upmix parameters, and the transient parameter.

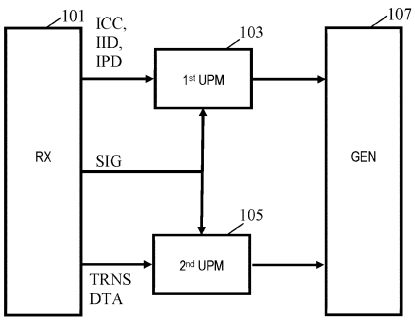


FIG. 1

EP 4 531 039 A1

Description

FIELD OF THE INVENTION

[0001] The invention relates to generation of a multichannel audio signals and/or an audio data signal representing a multichannel audio signal, and in particular, but not exclusively, to encoding and/or decoding of stereo signals.

BACKGROUND OF THE INVENTION

[0002] Spatial audio applications have become numerous and widespread and increasingly form at least part of many audiovisual experiences. Indeed, new and improved spatial experiences and applications are continuously being developed which result in increased demands on the audio processing and rendering.

[0003] For example, in recent years, Virtual Reality (VR) and Augmented Reality (AR) have received increasing interest and a number of implementations and applications are reaching the consumer market. Indeed, equipment is being developed for both rendering the experience as well as for capturing or recording suitable data for such applications. For example, relatively low cost equipment is being developed for allowing gaming consoles to provide a full VR experience. It is expected that this trend will continue and indeed will increase in speed with the market for VR and AR reaching a substantial size within a short time scale. In the audio domain, a prominent field explores the reproduction and synthesis of realistic and natural spatial audio. The ideal aim is to produce natural audio sources such that the user cannot recognize the difference between a synthetic or an original one.

[0004] A lot of research and development effort has focused on providing efficient and high quality audio encoding and audio decoding for spatial audio. A frequently used spatial audio representation is multichannel audio representations, including stereo representation, and efficient encoding of such multichannel audio based on downmixing multichannel audio signals to downmix channels with fewer channels have been developed. One of the main advances in low bit-rate audio coding has been the use of parametric multichannel coding where a downmix signal is generated together with parametric data that can be used to upmix the downmix signal to recreate the multichannel audio signal.

[0005] In particular, instead of traditional mid-side or intensity coding, in parametric multichannel audio coding a multichannel input signal is downmixed to a lower number of channels (e.g. two to one) and multichannel image (stereo) parameters are extracted. Then the downmix signal is encoded using a more traditional audio coder (e.g. a mono audio encoder). The bitstream of the downmix is multiplexed with the encoded multichannel image parameter bitstream. This bitstream is then transmitted to the decoder, where the process is inverted. First the downmix audio signal is decoded, after which the multichannel audio signal is reconstructed guided by the encoded multichannel image/ upmix parameters.

[0006] An example of stereo coding is described in E. Schuijers, W. Oomen, B. den Brinker, J. Breebaart, "Advances in Parametric Coding for High-Quality Audio", 114th AES Convention, Amsterdam, The Netherlands, 2003, Preprint 5852. In the described approach, the downmixed mono signal is parametrized by exploiting the natural separation of the signal into three components (objects): transients, sinusoids, and noise. In E. Schuijers, J. Breebaart, H. Pumhagen, J. Engdegård, "Low Complexity Parametric Stereo Coding", 116th AES, Berlin, Germany, 2004, Preprint 6073 more details are provided describing how parametric stereo was realized with a low (decoder) complexity when combining it with Spectral Band Replication (SBR).

[0007] In the described approaches, the decoding is based on the use of the so-called decorrelation process. The decorrelation process generates a decorrelated helper signal from the monaural signal. In the stereo reconstruction process, both the monaural signal and the decorrelated helper signal are used to generate the upmixed stereo signal based on the upmix parameters. Specifically, the two signals may be multiplied by a time- and frequency-dependent 2x2 matrix having coefficients determined from the upmix parameters to provide the output stereo signal.

[0008] However, although Parametric Stereo (PS) and similar downmix encoding/ decoding approaches were a leap forward from traditional stereo and multichannel coding, the approach is not optimal in all scenarios. In particular, known encoding and decoding approaches tend to introduce some distortion, changes, artefacts etc. that may introduce differences between the (original) multichannel audio signal provided to the encoder and the multichannel audio signal recreated at the decoder. Typically, the audio quality may be degraded and imperfect recreation of the multichannel audio signal occurs. Further, the data rate may still be higher than desired and/or the complexity/ resource usage of the processing may be higher than preferred.

[0009] A further issue is that it is often particularly desirable to reduce the complexity and computational load, especially at the decoder side.

[0010] Hence, an improved approach would be advantageous. In particular an approach allowing increased flexibility, improved adaptability, an improved performance, increased audio quality, improved audio quality to data rate trade-off, reduced complexity and/or resource usage, improved encoder side input on decoder side operation/processing, reduced computational load, facilitated implementation and/or an improved spatial audio experience would be advantageous.

SUMMARY OF THE INVENTION

[0011] Accordingly, the Invention seeks to preferably mitigate, alleviate or eliminate one or more of the above mentioned disadvantages singly or in any combination.

[0012] According to an aspect of the invention there is provided an audio apparatus for generating an output multichannel signal, the audio apparatus comprising: a receiver arranged to receive an audio data signal, the audio data signal comprising (data describing): a downmix audio signal being a downmix of a first multichannel signal; sets of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least: a level difference parameter indicative of a level difference between channels of the first multichannel signal; a correlation parameter indicative of a coherence between channels of the first multichannel signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel signal; and at least one transient upmix parameter for at least one transient audio component of the first multichannel signal, the transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal; a first upmixer arranged to generate an upmixed multichannel signal by upmixing of the downmix audio signal and in dependence on the sets of upmix parameters; a second upmixer arranged to generate an upmixed multichannel transient audio component by upmixing of the at least one transient audio component in dependence on the at least one transient upmix parameter; and a generator arranged to generate the output multichannel signal from a combination of the upmixed multichannel signal and the upmixed multichannel transient audio component.

[0013] The approach may provide an improved audio experience in many embodiments. For many signals and scenarios, the approach may provide improved generation/ reconstruction of a multichannel audio signal with an improved perceived audio quality. The approach may not only provide substantially improved representation of the transients of the original first multichannel audio signal in the generated output multichannel audio signal but may also provide improved upmixing with parameters more closely representing relationships between channels of the multichannel audio signals as the impact of transient behavior may be compensated.

[0014] The approach may provide a particularly advantageous arrangement which in many embodiments and scenarios may allow a facilitated and/or improved representation of both longer term audio properties as well as short term properties, such as for example both background applause and foreground clapping.

[0015] The approach may in many embodiments allow an improved multichannel audio signal to be generated by allowing the encoder to adapt/ modify the processing of the apparatus generating the multichannel audio signal based on transient properties.

[0016] The approach may provide an efficient implementation and may in many embodiments allow a reduced complexity and/or resource usage. The approach may in many scenarios allow a reduced data rate for data representing a multichannel audio signal using a downmix signal. Indeed, in many embodiments, substantially improved audio quality may be achieved with very little increase in the overall data rate.

[0017] The approach may in many embodiments allow reduced complexity and/or resource usage at the decoder/ generating/ reconstruction side.

[0018] The samples of the downmix audio signal may be time domain samples or may be frequency domain samples (specifically subband samples). The samples may span a particular time and frequency range.

[0019] The first upmixer may be arranged to generate the upmixed multichannel signal by applying a matrix multiplication to the downmix signal and an auxiliary audio signal with the coefficients of the matrix being determined as a function of parameters of the sets of upmix parameters. The matrix may be time- and frequency-dependent.

[0020] The audio apparatus may specifically be an audio decoder apparatus.

[0021] The processing may be time frequency segments or tiles that may be different time intervals and frequency intervals. Each time frequency segment/tile may represent a frequency interval in a time interval. In many embodiments, the first multichannel audio signal may be divided into time segments/intervals and a frequency representation of the signal in the time segment/interval may be provided by signal values representing different frequency segments of the signal in the time segment/interval.

[0022] The transient parameter may have a different time frequency resolution than the sets of upmix parameters which may be a different time resolution and/or a different frequency resolution. In many embodiments, the transient parameter and the upmix parameters may have the same time resolution but have different frequency resolution. In particular, the transient parameter may typically have a coarser frequency resolution than the upmix parameters. For at least some frequency intervals for which the upmix parameters provide separate values, the transient parameter may provide only a single parameter value. In many cases, the transient parameter may provide only a single value for the entire frequency band. Thus, in some embodiments, the frequency spectrum is not divided and the time frequency segments/tiles may be time segments/ intervals.

[0023] The first upmixer may specifically be arranged to generate the output multichannel audio signal by applying a matrix multiplication to (samples of) the downmix audio signal and a decorrelated signal with the matrix coefficients being determined from the sets of upmix parameters.

[0024] The upmixed multichannel audio signal, the upmixed multichannel transient audio component, and the output multichannel audio signal may have the same number of channels.

[0025] According to an optional feature of the invention, the receiver is arranged to extract a first transient audio component from the downmix audio signal and the second upmixer is arranged to upmix the first transient audio component in dependence on the transient upmix parameter.

[0026] This may provide an advantageous approach for many scenarios, including e.g. providing an advantageous trade-off between complexity, computational resources and/or the perceived audio quality of the generated multichannel audio signal.

[0027] According to an optional feature of the invention, the receiver is arranged to generate a residual downmix audio signal resulting from extracting a set of transient audio components from the downmix audio signal including the audio data for the at least one transient audio component, and the first upmixer is arranged to upmix the residual downmix audio signal in dependence on the sets of upmix parameters.

[0028] This may provide an advantageous approach for many scenarios and may provide an advantageous trade-off between performance, data rate and computational resource/complexity.

[0029] According to an optional feature of the invention, the first upmixer is arranged to decorrelate the residual downmix audio signal to generate a decorrelated residual downmix audio signal, and to generate the upmixed multichannel signal by upmixing of the residual downmix audio signal and the decorrelated residual downmix audio signal in dependence on the sets of upmix parameters.

[0030] This may provide an advantageous approach in many scenarios.

[0031] According to an optional feature of the invention, the audio data signal comprises audio data for the at least one transient audio component and the second upmixer is arranged to upmix the audio data for the at least one transient audio component in dependence on the transient upmix parameter.

[0032] This may provide an advantageous approach in many scenarios and may often provide improved audio quality.

[0033] In some embodiments, the downmix audio signal is a residual downmix audio signal representing the first multichannel signal following extraction of a set of transient audio components, the set of transient audio components including the at least one transient audio component.

[0034] According to an optional feature of the invention, the second upmixer is arranged to perform a panning of the least one transient audio component between two channels of the upmixed multichannel transient audio component in dependence on the transient upmix parameter.

[0035] This may provide an advantageous approach in many scenarios and may often provide improved audio quality.

[0036] In some embodiments, upmixing of the at least one transient audio component by the second upmixer is dependent on no other interchannel property parameter than the transient upmix parameter.

[0037] In some embodiments, the audio data signal comprises no other upmix parameter for the at least one transient audio component than the transient upmix parameter.

[0038] In some embodiments, the audio data signal comprises no correlation data, no coherence data, no phase data for the at least one transient audio component.

[0039] According to an optional feature of the invention, the first upmixer is arranged to perform a subband domain upmixing of the downmix audio signal; and the second upmixer is arranged to perform a time domain upmixing of the at least one transient audio component.

[0040] This may provide particularly advantageous performance, operation, reduced complexity and/or implementation in many embodiments and scenarios.

[0041] In some embodiments, the audio data signal comprises transient indications for time segments of the downmix audio signal, each transient indication being indicative of a number of transients in a time segment.

[0042] According to an optional feature of the invention, the audio data signal comprises a timing indication for the at least one transient audio component, and the combination is dependent on the timing indication.

[0043] This may provide particularly advantageous performance, operation and/or implementation in many embodiments and scenarios.

[0044] According to an optional feature of the invention, the frequency resolution for the at least one transient upmix parameter is coarser than a frequency resolution of the sets of upmix parameters.

[0045] This may provide advantageous operation and/or implementation and/or performance in many embodiments.

[0046] According to an aspect of the invention, there is provided an audio apparatus for generating an audio data signal, the audio apparatus comprising: a receiver arranged to receive a first multichannel signal; a downmixer arranged to generate a mono downmix audio signal from the first multichannel signal and to determine a set of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least: a level difference parameter indicative of a level difference between channels of the first multichannel signal; a correlation parameter indicative of a coherence between channels of the first multichannel signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel signal; a transient detector arranged to detect at least one transient audio component of the first multichannel signal and to generate at least one transient upmix parameter for the at least one

transient audio component of the first multichannel signal, the at least one transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal; a data generator arranged to generate the audio data signal to comprise the mono downmix audio signal, the sets of upmix parameters and the transient upmix parameter.

[0047] According to an optional feature of the invention, the transient detector is arranged to detect the at least one transient audio component by applying transient detection to the downmix audio signal.

[0048] This may provide advantageous operation and/or implementation and/or performance in many embodiments.

[0049] According to an optional feature of the invention, the transient detector is arranged to detect the at least one transient audio component by applying transient detection to channels of the first multichannel signal.

[0050] This may provide advantageous operation and/or implementation and/or performance in many embodiments.

[0051] In some embodiments, the data generator is arranged to include audio data describing the at least one transient audio component in the audio data signal.

[0052] In some embodiments, the transient detector is arranged to remove the at least one transient audio component from the downmix audio signal prior to the downmix audio signal being included in the audio data signal.

[0053] According to an aspect of the invention, there is provided a method of generating an output multichannel signal, the method comprising: receiving an audio data signal, the audio data signal comprising (data describing): a downmix audio signal being a downmix of a first multichannel signal; sets of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least: a level difference parameter indicative of a level difference between channels of the first multichannel signal; a correlation parameter indicative of a coherence between channels of the first multichannel signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel signal; and at least one transient upmix parameter for at least one transient audio component of the first multichannel signal, the transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal; generating an upmixed multichannel signal by upmixing of the downmix audio signal and in dependence on the sets of upmix parameters; generating an upmixed multichannel transient audio component by upmixing of the at least one transient audio component in dependence on the at least one transient upmix parameter; and generating the output multichannel signal from a combination of the upmixed multichannel signal and the upmixed multichannel transient audio component.

[0054] According to an aspect of the invention, there is provided a method of generating an output audio signal, the method comprising: receiving a first multichannel signal; generating a mono downmix audio signal from the first multichannel signal and to determine a set of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least: a level difference parameter indicative of a level difference between channels of the first multichannel signal; a correlation parameter indicative of a coherence between channels of the first multichannel signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel signal; detecting at least one transient audio component of the first multichannel signal and to generate at least one transient upmix parameter for the at least one transient audio component of the first multichannel signal, the at least one transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal; and generating the audio data signal to comprise the mono downmix audio signal, the sets of upmix parameters and the transient upmix parameter.

[0055] These and other aspects, features and advantages of the invention will be apparent from and elucidated with reference to the embodiment(s) described hereinafter.

BRIEF DESCRIPTION OF THE DRAWINGS

[0056] Embodiments of the invention will be described, by way of example only, with reference to the drawings, in which

FIG. 1 illustrates some elements of an example of an audio apparatus in accordance with some embodiments of the invention;

FIG. 2 illustrates some elements of an example of an audio apparatus in accordance with some embodiments of the invention;

FIG. 3 illustrates an example of a transient representation for a frame of an audio signal;

FIG. 4 illustrates an example of a transient representation for a frame of an audio signal;

FIG. 5 illustrates an example of elements of a transient detector in accordance with some embodiments of the invention;

FIG. 6 illustrates an example of a transient representation for a frame of an audio signal;

FIG. 7 illustrates an example of a transient representation for a frame of an audio signal;

FIG. 8 illustrates an example of a stereo audio signal, a stereo transient audio signal, and a stereo residual audio signal;

FIG. 9 illustrates an example of elements of an arrangement for training a neural network;

FIG. 10 illustrates an example of a time frequency representation (spectrogram) of a stereo applause signal;
FIG. 11 illustrates an example of a time frequency representation (spectrogram) of a stereo applause signal;
FIG. 12 illustrates an example of time segments of a stereo signal comprising a transient;
FIG. 13 illustrates some elements of an example of an audio apparatus in accordance with some embodiments of the invention;
FIG. 14 illustrates some elements of an example of an audio apparatus in accordance with some embodiments of the invention;
FIG. 15 illustrates some elements of an example of an audio apparatus in accordance with some embodiments of the invention; and
FIG. 16 illustrates some elements of a possible arrangement of a processor for implementing elements of an audio apparatus in accordance with some embodiments of the invention.

DETAILED DESCRIPTION OF SOME EMBODIMENTS OF THE INVENTION

[0057] FIG. 1 illustrates some elements of an audio apparatus for generating an output multichannel audio signal in accordance with some embodiments of the invention. FIG. 2 illustrates an example of an audio apparatus arranged to generate an audio data signal representing a multichannel audio signal henceforth referred to as the first multichannel audio signal. The audio data signal generated by the audio apparatus of FIG. 2 may specifically be fed to the audio apparatus of FIG. 1 which may be arranged to generate the output multichannel audio signal as a replica of the first multichannel audio signal. The audio apparatus of FIG. 1 will also be referred to as a decoder audio apparatus (or just as a decoder) and the audio apparatus of FIG. 2 will also be referred to as an encoder audio apparatus (or just as an encoder).

[0058] The decoder audio apparatus comprises a receiver 101 which is arranged to receive a data signal/ bitstream comprising a downmix audio signal which is a downmix of a multichannel audio signal. The data signal/ bitstream may specifically be one generated by the encoder audio apparatus to represent the first multichannel audio signal.

[0059] The following description will focus on cases where the multichannel audio signal is a stereo signal and the downmix signal is a mono signal, but it will be appreciated that the described approach and principles are equally applicable to the multichannel audio signal having more than two channels and to the downmix signal having more than a single channel (albeit fewer channels than the multichannel audio signal).

[0060] In addition to the downmix audio signal, the received data signal includes upmix parametric data which comprises sets of upmix parameters for upmixing the downmix audio signal. The upmix parameters may specifically be parameters that indicate relationships between the signals of different audio channels of the multichannel audio signal (specifically the stereo signal) and/or between the downmix signal and audio channels of the multichannel audio signal. Typically, the upmix parameters may be indicative of time differences, phase differences, level/intensity differences and/or a measure of similarity, such as correlation.

[0061] A set of upmix parameters comprises at least the following:

- A level difference parameter being indicative of a level difference between channels (and specifically two channels) of the first multichannel audio signal. The level difference parameter may specifically be an Interaural Intensity Difference (IID) and/or an Interaural Level Difference (ILD) as e.g. known from ISO/IEC 23003-3:2020 Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding.
- A correlation parameter being indicative of a coherence between channels (and specifically two channels) of the first multichannel audio signal. The correlation parameter may specifically be an Inter-channel Cross Correlation (ICC) parameter as e.g. known from ISO/IEC 23003-3:2020 Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding.
- A phase difference parameter being indicative of a phase difference between channels (and specifically two channels) of the first multichannel audio signal. The phase difference parameter may specifically be an Inter-channel Phase Difference (IPD), Overall Phase Difference (OPD), or Channel Phase Difference (CPD) parameter as e.g. known from ISO/IEC 23003-3:2020 Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding.

[0062] A set of parameters may specifically include an IID, ICC, and IPD parameter determined in accordance with ISO/IEC 23003-3:2020 Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding. In particular, the upmix parameters IID, ICC, and IPD may be determined as:

$$IID = \frac{\langle l, l \rangle}{\langle r, r \rangle}$$

$$ICC = \left| \frac{\langle l, r \rangle}{\sqrt{\langle l, l \rangle \langle r, r \rangle}} \right|$$

$$IPD = \angle \left\{ \frac{\langle l, r \rangle}{\sqrt{\langle l, l \rangle \langle r, r \rangle}} \right\}$$

where l and r represent the signal values of two channels/signals of the first multichannel audio signal (specifically the left and right channel signal of a stereo signal) and <*a, b*> represents the complex-valued inner product between the vectors *a* and *b.*

[0063] Typically, the upmix parameters are provided on a per time and per frequency basis (time frequency tiles). For example, new parameters may periodically be provided for each of a set of subbands.

[0064] The encoder audio apparatus is arranged to receive a first multichannel audio signal and to generate an audio data signal that represents the first multichannel audio signal with the representation including a downmix audio signal and the sets of upmix parameters. Specifically, the encoder audio apparatus may be a Parametric Stereo (PS) encoder that receives a stereo signal and encodes it as a mono audio signal with associated upmix parametric data.

[0065] Typically, the downmix audio signal is encoded and the receiver 101 is arranged to decode the downmix audio signal to provide the downmix audio signal, i.e. the mono signal in the specific example as well as the sets of upmix parameters and any other required data.

[0066] The decoder audio apparatus comprises a first upmixer 103 and a second upmixer 105 arranged to generate upmixed multichannel signals which are fed to a generator 107 which is arranged to combine these signals to generate an output multichannel signal.

[0067] The first upmixer 103 is arranged to generate an upmixed multichannel signal by upmixing of the downmix audio signal based on the sets of upmix parameters and specifically based on the level difference parameters, the correlation parameter, and the phase difference parameters (e.g. the IID, ICC, IPD parameters). In some embodiments, a conventional upmixing of a downmix signal may be performed. For example, for a stereo case, the first upmixer 103 may be arranged to perform a Parametric Stereo upmixing.

[0068] In the approach, the audio data signal further comprises transient data indicative of one or more properties of one or more transients in the first multichannel signal that is represented by the audio data signal. Thus, the audio data signal includes data that reflects one or more properties of transients in the original multichannel signal that the decoder audio apparatus is seeking to reproduce. The second upmixer 105 is arranged to generate an upmixed multichannel transient audio component by upmixing one or more transient audio components in dependence on the at least one transient upmix parameter. The second upmixer 105 may accordingly generate transient components for the output multichannel signal based on transient data that is provided in the audio data signal.

[0069] The first upmixer 103 and the second upmixer 105 are coupled to a generator 107 which is arranged to receive the upmixed multichannel signal and the upmixed multichannel transient audio components and combine them into an output multichannel signal.

[0070] In the approach, multiple and separate upmix operations are accordingly employed using different upmix data extracted from the audio data signal and separately upmixing different elements of the audio signal.

[0071] The encoder audio apparatus comprises a receiver 201 which receives the first multichannel audio signal from an internal or external source. The receiver 201 is coupled to a downmixer 203 that is arranged to downmix the first multichannel audio signal to generate a downmix audio signal which is a signal that has fewer channels than the first multichannel audio signal. In addition to the downmix audio signal, the downmixer 203 proceeds to generate sets of upmix parameters where each set of upmix parameters as described previously with respect to the audio data signal comprises at least a level difference parameter indicative of a level difference between channels of the multichannel audio signal; a correlation parameter indicative of a coherence between channels of the multichannel audio signal; and a phase difference parameter indicative of a phase difference between channels of the multichannel audio signal.

[0072] It will be appreciated that a number of approaches for generating such a downmix audio signal and associated upmix parameters are known and that any approach may be used as appropriate without detracting from the invention.

[0073] In many embodiments, the first multichannel audio signal may specifically be a stereo signal and the upmix parameters may be generated from the samples of a left and right channel signal of the input stereo signal. In such cases, the downmix audio signal is a mono downmix audio signal.

[0074] In addition, the encoder audio apparatus comprises a transient detector 205 which is arranged to detect transients in the first multichannel audio signal (and/or equivalently in the downmix audio signal). The transient detector 205 is arranged to detect transient audio components in the first multichannel signal and to generate at least one transient upmix parameter for each of the detected transient audio components. The upmix parameter indicates a level difference of

the transient audio component between channels of the first multichannel signal. Specifically, the transient upmix parameter may be an Interchannel Intensity Difference (IID) that is determined specifically for the transient audio component (e.g. it is determined for a small duration corresponding to the duration of the transient audio component).

[0075] In many embodiments, the transient upmix parameter may be a composite transient upmix parameter that includes a plurality of parameter/property values for the transient (or equivalently a plurality of transient upmix parameters may be provided for a given transient audio component). In particular, in many embodiments, the transient upmix parameter may also be indicative of timing property of the transient audio component, such as specifically the timing of when the transient occurs and/or an indication of the duration of the transient.

[0076] In general, the specific transient parameter(s) and property(ies) that are transmitted may be different in different embodiments. For example, in many cases the transient data may include an indication of one or more of a presence, number, time, duration, amplitude, interchannel level difference, interchannel phase difference, interchannel coherence for one or more transients that are present in the first multichannel audio signal (and often in the downmix audio signal).

[0077] The encoder audio apparatus further comprises a data signal generator 207 which generates the audio data signal to include data representing the downmix audio signal, the upmix parameters, and the transient parameter.

[0078] In many embodiments, the encoder audio apparatus and the decoder audio apparatus are arranged to perform subband processing. In particular, the upmix parameters may be generated for different (frequency) subbands of the first multichannel audio signal and the downmix audio signal.

[0079] Specifically, the receiver 201 or the downmixer 203 may comprise a filter bank which is arranged to generate a frequency subband representation of the downmix audio signal. Typically, the receiver 201 or the downmixer 203 may comprise a filter bank that is applied to all the channels of the first multichannel audio signal such that each channel signal is divided into subbands. The downmixing may then be performed on a per subband basis with upmix parameters being determined for each subband and a subband downmix signal being generated. The subband downmix audio signal may then in some cases be included directly in the audio data signal as a subband downmix audio signal or may be transformed to the time domain to provide a time domain signal.

[0080] The filter bank may be Quadrature Mirror Filter (QMF) bank or may e.g. be implemented by a Fast Fourier Transform (FFT), but it will be appreciated that many other filter banks and approaches for dividing an audio signal into a plurality of subband signals are known and may be used. The filterbank may specifically be a complex-valued pseudo QMF bank, resulting in e.g. 32 or 64 complex-valued sub-band signals.

[0081] The processing is furthermore typically performed in time segments. In most embodiments, the first multichannel audio signal is divided into time intervals/segments with a conversion to the frequency/subband domain by applying e.g. an FFT or QMF filtering to the samples of each signal. For example, each channel of the multichannel audio signal may be divided into time segments of 2048, 1024, or 512 samples which may then be processed to generate a plurality (e.g. 32, 16, 8) of subband signals of e.g. 64, 32 or 16 subband samples. Thus, a set of samples may be determined for each subband of the downmix audio signal. Further, for each time segment/ interval and frequency interval/subband, a set of upmix parameters may be generated.

[0082] It should be noted that the number of time domain samples is not directly coupled to the number of subbands. Typically, for a so-called critically sampled filterbank of N bands, every N input samples will lead to N sub-band samples (one for every sub-band). An oversampled filterbank will produce more output samples. E.g. for every N input samples, it would generate k*N output samples, i.e., k consecutive samples for every band.

[0083] Thus, sets of upmix parameters may be generated with each set being provided for a given time interval and a given frequency interval, also referred to as a given time frequency tile or segment.

[0084] Each set of upmix parameters may as previously described specifically include an IID, ICC, and IPD value and thus these parameters are provided with a given time resolution and a given frequency resolution. The time intervals may vary but typically have a fixed duration in many embodiments. In some embodiments, the subband size/ frequency resolution may also be fixed/constant for all subbands but in many embodiments the subbands may have different resolutions/ sizes. In many embodiments, the filterbank may be arranged to generate subband signals for subbands having equal bandwidth, and in many other embodiments, the filterbank may be arranged to generate subband signals with subbands having different bandwidths. For example, a higher frequency subbands may have a higher bandwidth than a lower frequency subband. Also, subbands may be grouped together to form a higher bandwidth sub-band.

[0085] Typically, the subbands may have a bandwidth in the range from 10Hz to 10000Hz.

[0086] However, in many embodiments, the time frequency resolution of the transient upmix parameter may be different from that of the set of upmix parameters.

[0087] The transient data/parameter is typically provided with a different time frequency resolution than the sets of upmix parameters, and indeed in many cases may be provided with a finer timing resolution or coarser frequency resolution than the sets of upmix parameters, and indeed in many cases with both a finer time resolution and a coarser frequency resolution. For example, a timing of a transient may be indicated with a timing granularity that is finer than the (processing) time segments/intervals and e.g. a non-frequency dependent interchannel level difference may be provided.

[0088] The sets of upmix parameters may be provided for time frequency tiles corresponding to the subbands and time

segments of the processing as previously described (e.g. with a fixed number of samples per time segment and subband). However, the transient parameter values may in some cases be provided with a finer time resolution. For example, the timing or duration of a transient may be provided with a higher time resolution than the sampling times of the subband samples.

[0089] Further, in many embodiments, the transient property may be provided with a coarser frequency resolution than for the sets of upmix parameters. In particular, in some embodiments, a set of upmix parameters may be provided for each subband whereas a single common transient parameter value is provided for a plurality, and possibly all, subbands.

[0090] In many embodiments, a set of one or more parameters may be provided for each of a set of detected transients. The parameters may specifically include an indication of a channel level difference between channels of the first multichannel audio signal.

[0091] A specific example of a time segment/ frame FRM in which three transients are detected is shown in FIG. 3. Here, for a given time segment/ frame three transients are detected at positions p0, p1 and p2. These positions/ time instants may be encoded and included in the audio data signal and accordingly transmitted to the decoder audio apparatus. Furthermore, for each transient parameter position p, an amplitude level difference is determined for the transient and is included in the audio data signal. For example, an IID (Interchannel Intensity Difference) may be determined and included in the audio data signal. In this case, a positive IID can correspond to a left panning, and a negative IID can correspond to a right panning of the transient signal. In some embodiments, the individual amplitudes a0, a1 and a2 can additionally or alternatively be transmitted. Thus, the transient data may be used to encode a representation of the detected transients.

[0092] In some embodiments, as illustrated in FIG. 4, the duration of a transient may be determined, encoded, and communicated to the decoder audio apparatus in the audio data signal.

[0093] In many embodiments, the parameter values may be quantized into relatively few levels, and thus a relatively low number of bits may be used for each value. In many embodiments, word lengths may be no more than 1, 2, 3, or 4 bits. For example, amplitude values may be quantized into a few levels (e.g. 5 or 7 discrete levels) using only a few bits.

[0094] The encoding of the parameter values in the audio data signal may for example use absolute or differential encoding. Specifically, the three IID values corresponding to positions p0, p1 and p2 may be coded differentially to the (average over the frequency bands) IID transmitted (per band) for the whole frame.

[0095] The encoder audio apparatus may accordingly generate transient data indicative of transients in the first multichannel audio signal and provide it to the decoder audio apparatus where it is used to perform a separate upmixing to the one used for the set of upmix parameters and thus an additional transient upmixed signal is generated. The encoder audio apparatus may provide this transient data with different time frequency resolution than the upmix parameters thereby allowing the transient data to be optimized independently. In particular, a coarser frequency resolution can be employed, and in many scenarios the transient data may not include any frequency dependency but rather the same parameter value may be provided for all subbands of the downmix audio signal. In many embodiments, coarse quantization of the parameter values into few discrete levels may also be achieved. Accordingly, a very low data overhead may in many embodiments be achieved. However, it has been found that the provision of this transient data/information and using it for a second separate transient upmixing operation can result in a substantially improved perceived audio quality, and in particular may very significantly improve the perceived audio realism for some scenarios and environments.

[0096] It will be appreciated that different approaches may be used for detecting transients in the first multichannel audio signal and/or in the downmix audio signal (indeed these operations can be considered equivalent as the transients of the first multichannel audio signal are also present in the downmix, and thus detecting a transient in the first multichannel audio signal also detects a transient in the downmix audio signal and vice versa).

[0097] FIG. 5 illustrates an example of elements of a transient detector 205 that may be used in the encoder audio apparatus. In the example, the transient detector 205 may be arranged to detect transients in a stereo signal.

[0098] The transient detection may be performed independently for the different channels, and in the example specifically the left and right channels, with the detections being combined thereafter. In other embodiments, it may be based on information from both channels directly, such as e.g. by considering the downmix audio signal. Such an approach may be beneficial as it can make use of the interchannel level (e.g. IID) parameters. The following examples will mainly consider such approaches.

[0099] In many embodiments, as will be described in the following, the transient detector 205 may detect a transient in response to a detection that a first level difference measure indicative of a level difference between channels of the first multichannel audio signal (specifically a first IID measure) differs from a second level difference measure indicative of a level difference between the channels (specifically a second IID measure for the same channels) by more than a threshold where the first level difference measure is determined for a shorter time interval than the second level difference. In many embodiments, the shorter time interval may not exceed 10%, 20%, 30%, or 50% of the time interval for the second level difference.

[0100] In the example of FIG. 5, the transient detector 205 includes an analysis filterbank 501 which typically decomposes the left and right stereo channels into a time-frequency (TF) representation where the distribution of

center-frequencies e.g. follows the logarithmically-spaced critical bands of the human auditory system (inner ear). The spectral decomposition may be performed using a hybrid quadrature mirror filterbank (QMF) that produces fine resolution at low frequencies, with the resolution decreasing (bandwidth increasing) as the frequency increases. It will be appreciated that in many embodiments such an analysis filterbank 501 may equivalently be part of the receiver 201 and the generated subband representation may also be used for the downmixing, upmix parameter estimation etc.

[0101] The QMF decomposition may produce complex outputs and the transient detector 205 comprises an envelope circuit 503 which determines the real envelope of both left and right channels. A simple envelope is given by,

$$e_i(m,k) = \sqrt{\Re_i(m,k)^2 + \Im_i(m,k)^2}$$

where $\Re_i(m,k)$ and $\Im_i(m,k)$ are the real and imaginary parts of the time frequency samples for time slice m and frequency bin *k* for $i \in \{left, right\}$. The square-root operation may be omitted to reduce complexity.

[0102] It should be noted that, depending on the embodiment, it is not always necessary to compute the real envelope, since the IID calculation already incorporates calculating the squared magnitude of the complex signal.

[0103] The transient detector 205 comprises a detection circuit 505 which is coupled to the envelope circuit 503 and which in the example processes the current and previous frames of time frequency samples. In the example, the IID over the windowed frames is determined (e.g. using an approach similar to a legacy PS coder which incorporates the symmetric Hanning window). This IID value serves as a baseline to predict the perceptual effect of detected transients later in the processing and will be denoted by μ.

[0104] Positive baseline values of μ indicate a left panning over the frames, values close to zero indicate center panning (no panning) and negative baseline IIDs indicate a right panning.

[0105] A short sliding window w corresponding to approximately 10 ms is used to compute the IID between left and right channels:

$$IID_w = 10\log_{10}|l_w|^2/|r_w|^2$$

[0106] Values of IID_w that deviate from μ can be considered as transient candidates and these may have their IID values encoded separately from the baseline IID μ. It should be noted that IID values may be computed at each QMF time instant per frequency band or aggregated across frequencies either globally or according to a customized binning scheme that may or may not omit certain frequencies that are not relevant to detecting certain transients.

[0107] Next, a perceptually-motivated step may evaluate the perceptual effect of the (PS) coder on the detected transients. It may compare the set of IID_w, to μ and filter out those transients that are not affected by the baseline (legacy) IID reconstruction in the decoder. This helps to reduce the number of parameters that has to be sent in the bitstream, thus keeping the resulting bit-rate under control.

[0108] It is known that humans perceive slowly varying stereo parameters as a moving source but can only detect rapidly changing parameters as either an increase or decrease in the stereo image width.

[0109] The perceptual filtering step has two objectives:

1. Decides whether an IID parameter should be separately calculated and transmitted for a given transient.
2. Group together transients depending on their IID properties, i.e., group transients that originate from the same source/location.

[0110] The first objective is perceptually motivated and is based on the deviation of the transient's IID parameter from the overall estimated frame parameter. If this deviation exceeds a certain threshold, then the transient is included as part of the stereo transient parameters.

[0111] The second objective can further reduce the bit rate, but bundling transients based on their stereo properties and assigning them to a virtual source (object) in the stereo image. This way only timing information and not both timing and IID features have to be transmitted for the same source if the IID value for the given source is stable over time.

[0112] FIG. 6 illustrates the same stereo transient representation as in FIG. 3 and 4 but with an average IID of the frame also being shown (μ) along with an IID range around the average IID given by $\pm\varepsilon$. In this case, the transients that fall outside of the indicated range may be represented by parameters that are included in the audio data signal.

[0113] The value of ε can be tuned based on perceptual (listening) tests or models which may e.g. determined the Just Noticeable Difference (JND) between the transient and average frame IID.

[0114] However, since it is known that the JND is also a function of the IID level - typically, the larger the IID, the larger the JND, the region between $\mu \pm \varepsilon$ can be replaced with a region of exclusion around the IID_w level itself as shown in FIG. 7. If

the average IID of the frame falls within this range, then the transient can be ignored and will not be encoded (as in the example is the case for transient p_1).

[0115] The perceptual filtering step may further include a model of masking. Similar to the region of exclusion, the masking model may indicate that a given transient cannot be sufficiently perceived by the user based on the background noise, for example.

[0116] In the example, short term properties are accordingly compared to longer term properties and a transient is detected in terms of these differing sufficiently.

[0117] Many suitable transient detection approaches are based on tracking fast changes in a signal envelope (either wideband or per frequency spectrum) relative to a slowly changing change or residual to detect onsets (and offsets) of transients.

[0118] In another approach, the magnitude of the time-frequency representation of a signal (e.g. a channel signal of the first multichannel audio signal or of the downmix audio signal) is determined, and the resulting frequency envelopes are summed across frequencies. Two smoothed versions of this envelope may then be created with one tracking the envelope more slowly than the other. Thus, a slowly varying residual envelope value is determined, and the other value is determined with a time-constant that tracks the envelope much faster. A first order exponential smoothing or e.g. a smoothing moving average filter can be applied to create the smoothed envelope. In the case of the fast-tracking envelope, the instantaneous envelope can also be used.

$$\widetilde{m_s}(n) = \alpha_s \tilde{m}(n) + (1-\alpha_s)\widetilde{m_s}(n-1)$$

$$\widetilde{m_f}(n) = \alpha_f \tilde{m}(n) + \left(1-\alpha_f\right)\widetilde{m_f}(n-1)$$

where $m_s(n)$ and $m_f(n)$ are the slow and fast envelopes respective with α_s, $\alpha_f \in (0,1]$ and $\alpha_f \gg \alpha_s$. The ratio between the fast and slow-tracking envelopes can then be used to indicate sharp changes of transients with respect to the residual signal and serves as a time-domain (wideband) gain function with

$$g(n) = \max\left(1 - \frac{\widetilde{m_s}(n)}{\widetilde{m_f}(n)},\ 0\right),$$

being an indication of a presence of transients as for transients, $m_f(n) \gg m_s(n)$, and $g(n) \to 1$. Such an approach is e.g. described in Adami, A., Herzog, A., Disch, S. and Herre, J., 2017, October. "Transient-to-noise ratio restoration of coded applause-like signals", 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 349-353). IEEE.

[0119] The measure may be used to detect the transients and the timing of these. Specifically, if g(n) exceeds a threshold, it may be considered that a transient has been detected. In some cases, the measure g(n) may then be compared to a second threshold (which may specifically be the same as the first threshold) and if it falls below this second threshold than the end of the transient may be considered to have been detected. Thus, the approach may detect both the beginning and the end of a transient, and accordingly may also detect the duration.

[0120] The intervals between onset and offset of the transients in the first multichannel audio signal and/or the downmix audio signal can further be used to filter out non-transient components that have longer durations and thus the transient detection may only detect relatively short transients rather than longer duration step changes.

[0121] In some embodiments, the encoder audio apparatus may in some cases separate the downmix audio signal into a set of transients and a residual signal having these transients removed. For example, a part of the downmix audio signal between the detection of the onset of a transition and the detection of the end of a transition may be extracted and represented as a separate transient with the resulting downmix audio signal representing a residual signal.

[0122] In some cases, a softer separation into transients and a residual signal may be performed by using a weighted rather than binary selection. For example, a transient signal t(n) may be generated by multiplying the first multichannel audio signal and/or the downmix audio signal by the detection signal g(n), i.e.

$$t(n) = g(n)m(n)$$

where m(n) represents e.g. a channel signal of the first multichannel audio signal or the downmix audio signal.

[0123] Similarly, a residual signal may be generated, e.g. as:

$$r(n) = 1 - g(n)m(n).$$

[0124] FIG. 8 illustrates an example of a stereo input signal $m(n)$ being separated into a transient signal $t(n)$ and a residual signal r(n).

[0125] Another approach to separate transients is to track the residual signal r(n) using a minimum tracking of the envelopes (based on the minimum statistics approach for stationary noise tracking e.g. described in Martin, Rainer. "Noise power spectral density estimation based on optimal smoothing and minimum statistics." IEEE Transactions on speech and audio processing 9.5 (2001): 504-512.) per frequency band. The transient signal can in such a case be written in the frequency domain as,

$$T(f) = \frac{|M(f)| - \gamma|\hat{R}(f)|}{|M(f)|} M(f),$$

where $\hat{R}(f)$ is the frequency-domain estimate of the residual signal using minimum tracking and γ is an over-subtraction factor (≥ 1.0) to account for under-estimating the residual signal.

[0126] As another example, source separation techniques employing neural networks may be employed. An example of such technology can be found in Daniel Stoller, Sebastian Ewert, Simon Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation", http://arxiv.org/abs/1806.03185, 2018.

[0127] Transient detection may be performed using a neural network model that has been trained to detect the location of transients using various neural network embodiments such as fully-connected, convolutional, or recurrent layers. A block diagram of a neural network and its training is illustrated in FIG. 9.

[0128] In the example, the input corresponds to either the current frame or a combination of the current (F) and previous (F-1) frame, where time blocks of the previous frame can serve as additional padding in case a transient occurs at the beginning of the current block F. Furthermore, the samples can correspond to all time frequency samples or a range of frequency samples relevant for transient detection. (for example, for clapping, most of the energy lies between 1 and 3 kHz). Assuming a $2 \times f \times n$ block of stereo data is used as input, the training data can consist of $2 \times f \times n$ input blocks, and the labels corresponding to a $1 \times n$ vector of 1s or 0s indicating the transient position of the training data. Additionally a transient presence flag can be added that indicates whether the frame includes at least a single transient, to the training labels, producing a $1 \times (n+1)$ vector label.

[0129] The employed loss function can correspond to an aggregated cross-entropy loss (assuming a frame length of n samples):

$$\mathcal{L} = -\sum_{i=1}^{n+1} [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

where y_i corresponds to the ground-truth label of the 7^{th} time instance for the given frame, and p_i is the predicted probability that a transient exists at time instant i by the neural network.

[0130] An alternative loss function can be based on the mean-squared error between the true ground-truth and predicted labels,

$$\mathcal{L} = \frac{1}{n+1} \sum_{i=1}^{n+1} (y_i - p_i)^2$$

[0131] Manually annotating applause signals with foreground claps is possible, although time consuming. Therefore, for the purposes of this invention, training can be performed on synthesized data using several available transient synthesizer models (e.g., clap/applause). These models can generate realistic sounding clap signals using signal processing techniques. During inference, the model estimates the locations of the detected stereo transients which can then be used to calculate the corresponding IID values.

[0132] In the described approach, the encoder audio apparatus is thus arranged to generate an audio data signal that not only includes a downmix and associated sets of upmix data but in addition generates transient parameters that reflect the interchannel level difference for transient audio components of the multichannel audio signal. The decoder audio apparatus uses this data in the upmixing to generate the output multichannel audio signal by employing two different upmixers and generating two different upmix signals that are then combined together.

[0133] The first upmixer 103 upmixes the downmix audio signal based on the set of upmix parameters and may specifically for a mono downmix signal being upmixed to a stereo signal proceed to apply a matrix multiplication to the samples of the mono downmix audio signal and a decorrelated signal as e.g. known from traditional PS upmixing:

$$\begin{pmatrix} l' \\ r' \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \begin{pmatrix} m \\ d \end{pmatrix}$$

where a *m* represents the mono downmix audio signal and *d* represents the decorrelated signal. This upmix procedure is typically operated in a time- and frequency dependent way with the upmix parameters being time and frequency dependent. For example, a set of upmix coefficients h_{xy} are typically determined for each time segment/frame and each subband. The upmix coefficients are determined from the received sets of upmix parameters. The exact dependency of the upmix coefficients on the sets of upmix parameters will be dependent on the specific embodiment. For example, in many embodiments, the relationships defined for conventional PS may be used as defined in ISO/IEC 23003-3:2020 Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding:

$$\begin{bmatrix} l' \\ r' \end{bmatrix} = \frac{1}{2c} \begin{bmatrix} 1+\alpha & \beta \\ 1-\alpha & -\beta \end{bmatrix} \begin{bmatrix} m \\ d \end{bmatrix}$$

$$\alpha = \frac{IID - 1 + 2j \cdot \sin(IPD) \cdot ICC \cdot \sqrt{IID}}{IID + 1 + 2 \cdot \cos(IPD) \cdot ICC \cdot \sqrt{IID}}$$

$$\beta = \frac{2 \cdot \sqrt{IID \cdot (1 - ICC^2)}}{IID + 1 + 2 \cdot \cos(IPD) \cdot ICC \cdot \sqrt{IID}}$$

$$c = \min\left(\sqrt{\frac{IID + 1}{IID + 1 + 2 \cdot \cos(IPD) \cdot ICC \cdot \sqrt{IID}}}, c_{max} \right)$$

[0134] The second upmixer 105 is arranged to generate upmixed multichannel transient audio components for each of the transients for which the audio data signal comprise transient parameters.

[0135] In some embodiments, the second upmixer 105 may be arranged to operate similarly to the first upmixer 105 in that it may also use a matrix multiplication of the downmix audio signal, and possibly also one or more auxiliary signals, to generate the upmixed multichannel transient audio component. However, rather than determine coefficients based on the set of upmix parameters, they are determined based on the transient parameters.

[0136] For example, a two by two matrix multiplication of a mono downmix audio signal and a decorrelated signal may be used to generate the upmixed multichannel transient audio component similarly to the approach used in a traditional PS upmix approach. Indeed, in some embodiments, the same upmix approach may be used as for the first upmixer 105 but with the upmix coefficients being determined from the specific transient parameters that are provided for the transient audio component. In some embodiments, the transient data may for a given transient audio component even specify an IID, IPD, and ICC and the coefficients for the transient upmix may be determined using similar formulas as for the first upmixer 103. However, rather than applying such an upmix operation to the entire downmix audio signal, the transient upmix operation is only applied to the corresponding transient audio component.

[0137] In some embodiments, a transient audio component may for example be defined as a time interval of the downmix audio signal in which a transient has been detected. Thus, in some embodiments, the second upmixer 105 may be arranged to specifically upmix a time interval (and typically a very short time interval) of the downmix audio signal using dedicated upmix coefficients determined from dedicated transient parameters. Thus, the upmixing specifically reflects the transient properties and not the properties of the whole audio segment for which the set of upmix parameters are provided.

[0138] Thus, even in such a case, a separate and dedicated upmix process is used to generate upmixed multichannel transient audio components that specifically reflect and reproduce the transient properties of the first multichannel audio signal.

[0139] However, in most embodiments, the type of transient parameters also differ from those used for conventional upmix of the first upmixer 103. In particular, the transient parameters typically provide less information than the sets of upmix parameters in that they are typically less frequency dependent and typically represent fewer interchannel signal properties.

[0140] In most embodiments, the transient parameter may have a coarse frequency dependency and indeed in many embodiments may not have any frequency dependency. Often, the same parameter value is provided for the entire audio frequency band. Thus, rather than a separate value being provided for each frequency subband (as is typically the case for

the sets of upmix parameters), a single value may be provided for the entire frequency range. Accordingly, the data overhead of the transient data may be reduced substantially in comparison to that of the sets of upmix parameters.

[0141] In many embodiments, the type of information/ parameters provided for the transient upmix may also be reduced with respect to the sets of upmix parameters. Indeed, in many embodiments, the transient parameters for a given transient audio component may not include any correlation/ coherence and/or phase difference information for the channels of the multichannel audio signal. Indeed, in many embodiments, the only relative interchannel properties that are represented by the transient parameter(s) for a given transient audio component may be the interchannel level difference. In many embodiments, the transient parameter(s) may for a given transient audio component include an indication of interchannel level difference(s) and optionally timing properties for the transient audio component. For example, for a given transient audio component, a single full band IID value may be provided together with an indication of the timing (and often duration) of a given transient audio component.

[0142] In such a case, the upmixing performed by the second upmixer 105 may simply scale the transient audio component to the different channels of the upmixed multichannel transient audio component such that the resulting upmixed multichannel transient audio component has an interchannel level difference that matches that indicated by the transient parameter.

[0143] In many embodiments, the second upmixer 105 may be arranged to perform a panning of the least one transient audio component between two channels of the upmixed multichannel transient audio component in dependence on the transient upmix parameter.

[0144] In a panning operation, the individual channel signals of the upmixed multichannel transient audio component may be generated by scaling the transient audio component by weights with the weights or different channels being dependent on the interchannel level difference. Thus, in such an operation, the channel signals are scaled copies of the transient audio component but with the scaling being dependent on the transient parameter. Effectively, such an operation may generate a upmixed multichannel transient audio component which corresponds directly to the transient audio component but spatially positioned (by the relative scaling) to have a desired position as represented by the transient parameter.

[0145] For example, for a stereo situation, a transient audio component in the form of a defined segment of a mono downmix audio signal may be panned between the left and right signals, and thus be positioned in the stereo image, by determining the scalings/ weights for the left and right signal to correspond to the interchannel level difference indicated by the transient parameter.

[0146] As a specific example, the second upmixer 105 may be arranged to perform a simple matrix multiplication to the samples of a mono downmix audio signal, such as:

$$\begin{pmatrix} l' \\ r' \end{pmatrix} = (h_1 \quad h_2)m$$

where h_1 and h_2 are scalar values determined as a function of the interchannel level difference indicated by the transient parameter (e.g. as an IID), and in many cases may only be dependent on the interchannel level difference.

[0147] As a specific example, the second upmixer 105 can implement a panning operation as follows:

$$\begin{bmatrix} l_t \\ r_t \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{iid}}{\sqrt{iid+1}} & \frac{1}{\sqrt{iid+1}} \end{bmatrix} [t]$$

where the *iid* value represents the linear inter-channel intensity difference between the left and right transient signal and t is the transient signal component of the mono downmix audio signal.

[0148] The approach may exploit the Inventors' realization that in contrast to a diffuse background signal, transients are typically well localized spatially. Furthermore, due to the typically short duration of the transients, it is often not necessary to reconstruct the full (diffuse) spatial image of a transient. Instead, it typically suffices to pan the transient signals to the correct position.

[0149] The transient parameters may accordingly in many embodiments provide substantially less interchannel information than the sets of upmix parameters and accordingly the overhead of providing the transient data may be very low, and often insignificant compared to the overhead imposed by the sets of upmix parameters. Further, it has been found that despite the low overhead, a significantly improved audio experience and perceived quality can be achieved by the parallel and separate upmixing of transient audio components based on the provided transient data.

[0150] It has been found that a particularly advantageous generation of an upmixed multichannel signal can be achieved in many scenarios and for many signals (and embodiments) using the described approach. It has been found that the inclusion of the transient data and information allows for information that may otherwise be lost (due to not being sufficiently represented by the upmix parameters) to be taken into account by the decoder audio apparatus. Further, the representa-

tion of the transient information using a different time frequency resolution of the transient information than used for the upmix parameters has been found to allow a substantially improved upmixing and audio quality while only introducing a small overhead. In particular, it has been found that a very coarse frequency resolution, including having no frequency dependency of the transient parameter values, may still allow very accurate and substantially improved upmixing that includes transient components. Further, it has been found that a different time resolution, including in particular allowing a finer time resolution than the segments/time intervals used for upmix parameters allow improved audio quality, and in particular allows much better representation of some audio components and sounds.

[0151] Further, the Inventors have realized that such an improved audio quality and user experience can be achieved while only providing reduced information compared to the conventional upmixing. In particular, they have realized that substantial improvements can typically be achieved by only providing interchannel level difference information and without necessarily providing interchannel coherence or phase difference information.

[0152] To illustrate the considerations, an audio signal representing an applause from a group of people may be considered. Such applause signals tend to consist of a seemingly random (spatial) superposition of individual claps, i.e. short bursts of energy in time. On the one hand, estimating a set of stereo parameters and interpolating these between frames does not lead to an accurate reconstruction of the stereo signal at the decoder. On the other hand, fine-grained estimates of parameters may substantially increase the overall bit-rate.

[0153] To further appreciate the effects of a low (per-frame) parameter update rate, the signal of FIG. 10 may be considered. The figure illustrates an example of the time-frequency decomposition of an applause stereo signal sampled at 44.1 kHz (spectrogram). The signal consists of background and foreground applause, where the background applause is dominant below 3 kHz. The foreground claps are clearly panned to the left as they do not appear as prominently in the right channel. The time-frequency energy for the background clapping is quite random and noise-like.

[0154] The output of a conventional parametric stereo decoder for the same segment is shown in FIG. 11. It should be noted that this approach results in a smeared-out background applause and foreground clapping. The conventional approach of stereo parameter interpolation does not work very well here. For the background applause, the smearing effect results in a musical tones-like effect, with clear generation of harmonic components. The foreground claps are also slightly smeared out in time and lose their panning characteristics (IID): some foreground claps now also appear more prominently in the right channel. The latter effects are a result of the stereo parameter estimator's frame-by-frame update rate.

[0155] To understand the effect on the stereo parameters better, consider a depiction of the left channel's spectrogram for two frames of data in FIG. 12. A foreground clap 1201 occurs in the middle of the previous frame. The stereo parameters are estimated by first windowing the two frames such that the energy near the beginning of the previous frame and end of the current frame is attenuated.

[0156] If the IID, IPD, and ICC are calculated over these two frames, the distribution of phase, intensity and coherence from the background applause between left and right channels will largely determine the estimated stereo parameters, even though for the foreground clap, the parameters should be different since at least for the IID, it is clear that the signal is panned to the left.

[0157] Therefore, providing transient data with a higher time resolution allows for additional information that can be included in the upmixing allowing an improved output audio signal to be generated. In the described approach, such information is used to perform an additional and separate upmixing of transient parameters for one or more transient audio components to generate upmixed multichannel transient audio components that are combined with a more conventionally generated upmixed signal.

[0158] The approach may in particular allow the transient information to be adapted to the specific importance of information of the transients. Indeed, for the transients, it has been found that upmixing is much less sensitive to frequency dependencies than to timing accuracy and in the described approach, the transient data may be adapted/generated accordingly and is not limited to follow the same resolutions as for the upmix data.

[0159] Further, in many embodiments, typically in addition to timing information indicating a timing of the transients, interchannel level differences may be significant and e.g. allow an accurate representation of the spatial position (e.g. in a stereo image) of the transients.

[0160] In many embodiments, the only interchannel information provided for the transient may be an interchannel level difference indication. In particular, in many embodiments, the transient data may not include any interchannel phase difference or interchannel coherence. Indeed, such information provides substantially less relevant information and typically has significantly less impact on the resulting audio quality of the generated output multichannel signal.

[0161] The second upmixer 105 is arranged to upmix a transient audio component based on transient parameter values provided for the transient audio component. The exact signal component that is upmixed to generate the upmixed multichannel transient audio component will depend on the individual scenario and embodiment.

[0162] In many embodiments, the transient audio component may be determined by the decoder audio apparatus based on parameter data which is provided by the encoder audio apparatus, and specifically a transient audio component may be generated as a specific time interval of the downmix audio signal which is indicated by received transient timing values. For

example, the transient detection at the encoder audio apparatus may detect a transient occurring and determine the start time and the end time of a corresponding transient time interval. These times may be communicated to the decoder audio apparatus which may be arranged to generate a transient audio component by extracting the segment of the received downmix audio signal corresponding to the indicated time interval. This, transient audio component, i.e. the identified time segment of the downmix audio signal may then be upmixed based on the transient parameter to generate a corresponding upmixed multichannel transient audio component. The first upmixer 103 may upmix the full downmix audio signal to generate the upmixed multichannel signal which in the generator 107 is then combined with the generated upmixed multichannel transient audio component. In many embodiments, the combination may simply add the two multi channel signals/components together, e.g. with potentially a scaling/ weighting of each signal.

[0163] In many embodiments, the downmix audio signal is used to perform the transient upmixing and is also used for the upmixing by the first upmixer 103. In such cases, two contributions may be generated for a given part of the downmix audio signal that corresponds to a transient, i.e. in the previous example the indicated segment of the downmix audio signal are both upmixed based on the sets of upmix parameters (by the first upmixer 103) and based on the transient parameters (by the second upmixer 105). This may in some cases cause some degradation which however may be acceptable (or possibly even desirable) in some cases. In some embodiments, the parameters may also be adapted to include a compensation for such combined upmixing. For example, the interchannel level difference provided for the transient may not directly indicate the actual level difference but may indicate the relative level difference with respect to the level difference indicated by a set of upmix parameters for the segment including the transient. Specifically, the IID value provided for the transient may be relative to the IID value of the corresponding set of upmix parameters and thus include both a contribution that will cancel the contribution from the first upmixer 103 as well as a contribution representing the desired level difference for the transient.

[0164] However, in other embodiments, the first upmixer 103 may be performed on a residual downmix audio signal that results from extracting the transient audio components from the downmix (or at least one transient audio component). For example, in the previous example, a residual downmix audio signal may be generated simply be deleting the time intervals indicated as transients from the downmix audio signal. In such a case, the downmix audio signal may then be upmixed by the first upmixer 103 based on the sets of upmix parameters except for transient segments that are upmixed by the second upmixer 105 based on the transient parameters.

[0165] As another example, the separation of the downmix audio signal into a transient signal/transient audio components and a residual downmix audio signal may be performed by analyzing the downmix audio signal (or at the encoder audio apparatus the first multichannel audio signal) as previously described in connection with transient detection. E.g. the transient signal can be determined as:

$$t(n) = g(n)m(n)$$

and the residual signal as:

$$r(n) = 1 - g(n)m(n).$$

where m(n) represents e.g. a channel signal of the first multichannel audio signal or the downmix audio signal and g(n) is the detection function previously described.

[0166] In some embodiments, the decoder audio apparatus may be arranged to perform a transient detection on the downmix audio signal to detect the transients. The previous description of transient detection at the encoder audio apparatus may in some embodiments apply equally (mutatis mutandis) to the decoder audio apparatus, and specifically to the receiver 101 of the decoder audio apparatus which may identify transients in the downmix audio signal.

[0167] The second upmixer 105 may then be arranged to upmix the identified parts of the downmix audio signal using the received transient parameters.

[0168] In many embodiments, the receiver 101 may be arranged to generate the transient signal/ transient audio components as previously described as well as the residual downmix audio signal. Thus, as described above, the decoder audio apparatus may be arranged to receive a downmix audio signal, such as a mono downmix audio signal, and perform transient detection to generate a number of transient audio components and a residual downmix audio signal with these transient audio components being extracted. The transient audio components are then upmixed by the second upmixer 105 based on received transient parameters and the residual downmix audio signal is upmixed by the first upmixer 103 with the resulting upmixed signals being combined by the generator 107 to generate an output multichannel audio signal.

[0169] In such embodiments, the decoder audio apparatus may itself detect the transients for which transient data is provided from the encoder audio apparatus. The linking of the received transient parameters to the detected transients may be done in different ways. For example, the same transient detection may be performed in the encoder and decoder based on the same signal (e.g. the transient detection at the encoder audio apparatus may also be based on the downmix

audio signal (e.g. after an encoding and decoding process that matches that performed at the decoder audio apparatus)). In such cases, there will be a strong correspondence between the detected transients and a simple allocation may be performed (e.g. the first received transient parameter is assigned to the first detected transient, the second received transient parameter is assigned to the second detected transient, etc.). In some cases, the audio data signal may for example for each set of upmix parameter time segment indicate the number of transient parameters, and the receiver 101 may be arranged to detect the corresponding number of transients and assign the transient parameters provided for the segment. Thus, in many embodiments, the decoder audio apparatus further comprises functionality for generating the transient audio component(s) from the received downmix audio signal.

[0170] The approach of providing a full downmix audio signal and allowing a decoder audio apparatus detection and extraction of transients may provide a number of advantages. It may typically reduce the data rate, and often to a significant extent. It may for example also facilitate backwards compatibility as a full downmix will be available for upmixing by legacy decoders which do not include a dedicated transient upmix path.

[0171] In some embodiments, the receiver 101 may further be arranged to extract the transient audio component(s) based on the provided transient parameter. As previously indicated, if the transient parameter comprises a timing indication, the extraction may be performed by extracting a section of the downmix audio signal that corresponds to the indicated time interval.

[0172] In some embodiments, the encoder audio apparatus may be arranged to generate the audio data signal to include audio data for one or more of the transient audio components. For example, dedicated audio data that describes the transient may be provided. In such cases, the receiver 301 of the decoder audio apparatus may also be arranged to recreate a local replica of the transient audio component from the received audio signal and upmix this in the second upmixer 105 using the received transient parameter.

[0173] Such an approach may in many cases increase the required data rate and bandwidth but often only by a relatively small amount. However, it may typically allow a substantially reduced complexity and computational resource at the decoder audio apparatus which is a substantial benefit in many practical applications. Further, it may allow an improved audio quality as the transient may be more accurately represented.

[0174] In many embodiments, the encoder downmix audio signal may still provide a full downmix audio signal that includes transients. However, in other embodiments, the encoder may be arranged to generate a residual downmix audio signal and to transmit this as the downmix audio signal. In such cases, the decoder audio apparatus may simply decode the residual downmix audio signal and feed it to the first upmixer 103 for upmixing, and decode the transient audio components and feed these to the second upmixer 105 for upmixing. Thus, an efficient and relatively low complexity and resource demanding operation is required by the decoder audio apparatus.

[0175] Thus, in some embodiments, the transient detector 205 may be arranged to remove the transient audio component(s) from the downmix audio signal prior to the downmix audio signal being included in the audio data signal. The transient detector 205 may for example perform the previously described operations to generate a transient signal and a residual downmix signal based on the transient detection function g(n).

[0176] As a specific example, in some embodiments the decoder audio apparatus mays specifically be a parametric stereo decoder which receives a mono downmix audio signal and sets of upmix parameters comprising parametric stereo parameters. The mono downmix audio signal may be separated into a transient signal *t* and a residual signal *r* where the transient signal is upmixed to a stereo signal (l_t, r_t) by a second upmixer 105 and the residual signal is upmixed to a stereo signal (l_r, r_r) by the first upmixer 103. An output stereo pair (l', r') is then constructed from the summation of the stereo pair signals.

[0177] The residual upmixer (the first upmixer 103) will comprise a decorrelation module to produce a signal substantially decorrelated to the residual signal r and the residual stereo pair is constructed from a mixture of the residual signal r and its decorrelated signal, controlled by a set of PS parameters.

[0178] The transient upmixer (the second upmixer 105) however in the specific example only consist of a panning module, controlled by a set of PS parameters to pan the transient signal to the stereo pair (l_t, r_t).

[0179] A specific example of a corresponding encoder audio apparatus is shown in FIG. 13. This encoder may generate a downmix signal and estimate transient (foreground) and residual (background) parameters. For both the input left and right signals (l, r), the signals are split into left transient (ti), left residual (n), right transient (t_r) and right residual (r_r) signals by transient separators 1301. The left and right transient signals are fed to a transient parameter estimator 1303, resulting in transient parameters. The left and right residual signals are fed to a residual parameter estimator 1305, resulting in residual parameters. The residual parameter estimator 1305 can be a traditional PS parameter estimator, estimating IID, ICC and IPD values. The transient parameter estimator 1303 can be a simplified PS parameter estimator, only estimating an intensity difference (panning / IID) between the left and right transient signals. The transient signals, the residual signals and the transient and residual parameters are fed to a downmix module 1307 that downmix the incoming signal. In addition, the transient and residual parameters may be used to normalize the power of the downmix signal with regard to the incoming signals. The resulting downmix and the transient and residual parameters are further encoded (not shown in FIG. 13) and transmitted to a decoder.

[0180] In some embodiments, to ensure similarity to the decoder transient separation, the encoder may include a traditional PS encoder, generating a downmix after which the same transient separation module as the decoder is run to generate the mono transient and residual signals. The transient signal can then be used to control the detection of stereo transients in the encoder since it is clear where the decoder would generate transient signals from the mono signal.

[0181] Since not all frames may contain transient information, it may be beneficial to signal to the decoder whether a frame contains transient information. In that case the overhead is kept lower on average as no transient parameters need to be transmitted each frame, and the decoder transient separation only needs to be run upon presence of this flag. Instead of this binary signaling, the signaling may also consist of transmission of the number of transients within a frame, possibly entropy encoded to provide a shorter word length for lower number of transients. In that case the transient parameter data may consists of individual parameters, e.g. one IID per transient.

[0182] As previously mentioned, at least some of the processing is typically performed on subband signals and in the subband domain. In particular, for the decoder audio apparatus, the downmix audio signal may be converted to (or received in) a subband representation and all the processing may be in the subband domain until the left and right output signals are generated and converted to the time domain. FIG. 14 illustrates an example of such a decoder audio apparatus.

[0183] In other embodiments, the downmix audio signal may be received in the time domain and all of the processing may be performed in the time domain.

[0184] In some embodiments, some of the processing may advantageously be performed in the subband domain on subband samples/signals and other processing may advantageously be performed in the time domain on time domain samples/signals. Specifically, advantageously in many scenarios, the first upmixer 103 may be arranged to perform a subband domain upmixing of the downmix audio signal whereas the second upmixer 105 may be arranged to perform a time domain upmixing of the at least one transient audio component.

[0185] In some embodiments, the downmix audio signal/residual downmix audio signal fed to the first upmixer 103 may be a subband domain signal and the upmixing may be performed in subbands using sets of upmix parameters that are provided for the different subbands. In contrast, the transient audio components fed to the second upmixer 105 may be a time domain signal and the upmixing may be performed on the time domain samples using transient parameters that are also provided in the time domain and which may have no time dependency. FIG. 15 illustrates an example of such a decoder audio apparatus.

[0186] It is often beneficial to operate the described processing in different frequency bands. Efficient means of operating in the frequency domain may e.g. be a sub-band processing by means of (hybrid) Quadrature Mirror Filtering (QMF) banks. From an efficiency point of view, it is often beneficial that all processing blocks operate in the same domain (as in FIG. 14). Therefore, the incoming mono signal m is processed, e.g., by a hybrid QMF bank to result in a set of subband domain signals representative of the mono signal m. All subsequent processing (transient separation, upmixing) then operates in the same domain. Finally, after the transient and residual stereo (subband domain) signals have been reconstructed and added, they are synthesized back to the time domain.

[0187] In some embodiments, it may be desirable for especially the transient signals to be processed separately in the time domain (as in FIG. 15). This has the advantage that the complexity of the transient upmix processing is even lower. It is noted that the transient separation may still (partially) be performed in the frequency domain.

[0188] The audio apparatus(s) may specifically be implemented in one or more suitably programmed processors. In particular, the artificial neural networks may be implemented in one more such suitably programmed processors. The different functional blocks, and in particular the artificial neural networks, may be implemented in separate processors and/or may e.g. be implemented in the same processor. An example of a suitable processor is provided in the following.

[0189] FIG. 16 is a block diagram illustrating an example processor 1600 according to embodiments of the disclosure. Processor 1600 may be used to implement one or more processors implementing an apparatus as previously described or elements thereof (including in particular one more artificial neural network). Processor 1600 may be any suitable processor type including, but not limited to, a microprocessor, a microcontroller, a Digital Signal Processor (DSP), a Field ProGrammable Array (FPGA) where the FPGA has been programmed to form a processor, a Graphical Processing Unit (GPU), an Application Specific Integrated Circuit (ASIC) where the ASIC has been designed to form a processor, or a combination thereof.

[0190] The processor 1600 may include one or more cores 1602. The core 1602 may include one or more Arithmetic Logic Units (ALU) 1604. In some embodiments, the core 1602 may include a Floating Point Logic Unit (FPLU) 1606 and/or a Digital Signal Processing Unit (DSPU) 1608 in addition to or instead of the ALU 1604.

[0191] The processor 1600 may include one or more registers 1612 communicatively coupled to the core 1602. The registers 1612 may be implemented using dedicated logic gate circuits (e.g., flip-flops) and/or any memory technology. In some embodiments the registers 1612 may be implemented using static memory. The register may provide data, instructions and addresses to the core 1602.

[0192] In some embodiments, processor 1600 may include one or more levels of cache memory 1610 communicatively coupled to the core 1602. The cache memory 1610 may provide computer-readable instructions to the core 1602 for execution. The cache memory 1610 may provide data for processing by the core 1602. In some embodiments, the

computer-readable instructions may have been provided to the cache memory 1610 by a local memory, for example, local memory attached to the external bus 1616. The cache memory 1610 may be implemented with any suitable cache memory type, for example, Metal-Oxide Semiconductor (MOS) memory such as Static Random Access Memory (SRAM), Dynamic Random Access Memory (DRAM), and/or any other suitable memory technology.

[0193] The processor 1600 may include a controller 1614, which may control input to the processor 1600 from other processors and/or components included in a system and/or outputs from the processor 1600 to other processors and/or components included in the system. Controller 1614 may control the data paths in the ALU 1604, FPLU 1606 and/or DSPU 1608. Controller 1614 may be implemented as one or more state machines, data paths and/or dedicated control logic. The gates of controller 1614 may be implemented as standalone gates, FPGA, ASIC or any other suitable technology.

[0194] The registers 1612 and the cache 1610 may communicate with controller 1614 and core 1602 via internal connections 1620A, 1620B, 1620C and 1620D. Internal connections may be implemented as a bus, multiplexer, crossbar switch, and/or any other suitable connection technology.

[0195] Inputs and outputs for the processor 1600 may be provided via a bus 1616, which may include one or more conductive lines. The bus 1616 may be communicatively coupled to one or more components of processor 1600, for example the controller 1614, cache 1610, and/or register 1612. The bus 1616 may be coupled to one or more components of the system.

[0196] The bus 1616 may be coupled to one or more external memories. The external memories may include Read Only Memory (ROM) 1632. ROM 1632 may be a masked ROM, Electronically Programmable Read Only Memory (EPROM) or any other suitable technology. The external memory may include Random Access Memory (RAM) 1633. RAM 1633 may be a static RAM, battery backed up static RAM, Dynamic RAM (DRAM) or any other suitable technology. The external memory may include Electrically Erasable Programmable Read Only Memory (EEPROM) 1635. The external memory may include Flash memory 1634. The External memory may include a magnetic storage device such as disc 1636. In some embodiments, the external memories may be included in a system.

[0197] The invention can be implemented in any suitable form including hardware, software, firmware, or any combination of these. The invention may optionally be implemented at least partly as computer software running on one or more data processors and/or digital signal processors. The elements and components of an embodiment of the invention may be physically, functionally and logically implemented in any suitable way. Indeed the functionality may be implemented in a single unit, in a plurality of units or as part of other functional units. As such, the invention may be implemented in a single unit or may be physically and functionally distributed between different units, circuits and processors.

[0198] Although the present invention has been described in connection with some embodiments, it is not intended to be limited to the specific form set forth herein. Rather, the scope of the present invention is limited only by the accompanying claims. Additionally, although a feature may appear to be described in connection with particular embodiments, one skilled in the art would recognize that various features of the described embodiments may be combined in accordance with the invention. In the claims, the term comprising does not exclude the presence of other elements or steps.

[0199] Furthermore, although individually listed, a plurality of means, elements, circuits or method steps may be implemented by e.g. a single circuit, unit or processor. Additionally, although individual features may be included in different claims, these may possibly be advantageously combined, and the inclusion in different claims does not imply that a combination of features is not feasible and/or advantageous. Also the inclusion of a feature in one category of claims does not imply a limitation to this category but rather indicates that the feature is equally applicable to other claim categories as appropriate. Furthermore, the order of features in the claims do not imply any specific order in which the features must be worked and in particular the order of individual steps in a method claim does not imply that the steps must be performed in this order. Rather, the steps may be performed in any suitable order. In addition, singular references do not exclude a plurality. Thus references to "a", "an", "first", "second" etc. do not preclude a plurality. Reference signs in the claims are provided merely as a clarifying example shall not be construed as limiting the scope of the claims in any way.

Claims

1. An audio apparatus for generating an output multichannel signal, the audio apparatus comprising:
   a receiver (101) arranged to receive an audio data signal, the audio data signal comprising:

   a downmix audio signal being a downmix of a first multichannel signal;
   sets of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least:

   a level difference parameter indicative of a level difference between channels of the first multichannel signal;
   a correlation parameter indicative of a coherence between channels of the first multichannel signal; and
   a phase difference parameter indicative of a phase difference between channels of the first multichannel

signal; and
at least one transient upmix parameter for at least one transient audio component of the first multichannel signal, the transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal;

a first upmixer (103) arranged to generate an upmixed multichannel signal by upmixing of the downmix audio signal and in dependence on the sets of upmix parameters;
a second upmixer (105) arranged to generate an upmixed multichannel transient audio component by upmixing of the at least one transient audio component in dependence on the at least one transient upmix parameter; and
a generator (107) arranged to generate the output multichannel signal from a combination of the upmixed multichannel signal and the upmixed multichannel transient audio component.

2. The audio apparatus of claim 1 wherein the receiver (101) is arranged to extract a first transient audio component from the downmix audio signal and the second upmixer is arranged to upmix the first transient audio component in dependence on the transient upmix parameter.

3. The audio apparatus of claim 2 wherein the receiver (101) is arranged to generate a residual downmix audio signal resulting from extracting a set of transient audio components from the downmix audio signal including the audio data for the at least one transient audio component, and the first upmixer (103) is arranged to upmix the residual downmix audio signal in dependence on the sets of upmix parameters.

4. The apparatus of claim 3 wherein the first upmixer (103) is arranged to decorrelate the residual downmix audio signal to generate a decorrelated residual downmix audio signal, and to generate the upmixed multichannel signal by upmixing of the residual downmix audio signal and the decorrelated residual downmix audio signal in dependence on the sets of upmix parameters.

5. The audio apparatus of any previous claim wherein the audio data signal comprises audio data for the at least one transient audio component and the second upmixer (105) is arranged to upmix the audio data for the at least one transient audio component in dependence on the transient upmix parameter.

6. The audio apparatus of any previous claim wherein the second upmixer (105) is arranged to perform a panning of the least one transient audio component between two channels of the upmixed multichannel transient audio component in dependence on the transient upmix parameter.

7. The audio apparatus of any previous claim wherein the first upmixer (103) is arranged to perform a subband domain upmixing of the downmix audio signal; and the second upmixer (105) is arranged to perform a time domain upmixing of the at least one transient audio component.

8. The audio apparatus of any previous claim wherein the audio data signal comprises a timing indication for the at least one transient audio component, and the combination is dependent on the timing indication.

9. The audio apparatus of any previous claim wherein a frequency resolution for the at least one transient upmix parameter is coarser than a frequency resolution of the sets of upmix parameters.

10. An audio apparatus for generating an audio data signal, the audio apparatus comprising:

a receiver (201) arranged to receive a first multichannel signal;
a downmixer (203) arranged to generate a mono downmix audio signal from the first multichannel signal and to determine a set of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least:

a level difference parameter indicative of a level difference between channels of the first multichannel signal;
a correlation parameter indicative of a coherence between channels of the first multichannel signal; and
a phase difference parameter indicative of a phase difference between channels of the first multichannel signal;

a transient detector (205) arranged to detect at least one transient audio component of the first multichannel signal and to generate at least one transient upmix parameter for the at least one transient audio component of the first

multichannel signal, the at least one transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal;
a data generator (207) arranged to generate the audio data signal to comprise the mono downmix audio signal, the sets of upmix parameters and the transient upmix parameter.

11. The audio apparatus of claim 10 wherein the transient detector (205) is arranged to detect the at least one transient audio component by applying transient detection to the downmix audio signal.

12. The audio apparatus of claim 10 or 11 wherein the transient detector (205) is arranged to detect the at least one transient audio component by applying transient detection to channels of the first multichannel signal.

13. A method of generating an output multichannel signal, the method comprising:
receiving an audio data signal, the audio data signal comprising:

a downmix audio signal being a downmix of a first multichannel signal;
sets of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least:

a level difference parameter indicative of a level difference between channels of the first multichannel signal;
a correlation parameter indicative of a coherence between channels of the first multichannel signal; and
a phase difference parameter indicative of a phase difference between channels of the first multichannel signal; and
at least one transient upmix parameter for at least one transient audio component of the first multichannel signal, the transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal;

generating an upmixed multichannel signal by upmixing of the downmix audio signal and in dependence on the sets of upmix parameters;
generating an upmixed multichannel transient audio component by upmixing of the at least one transient audio component in dependence on the at least one transient upmix parameter; and
generating the output multichannel signal from a combination of the upmixed multichannel signal and the upmixed multichannel transient audio component.

14. A method of generating an output audio signal, the method comprising:

receiving a first multichannel signal;
generating a mono downmix audio signal from the first multichannel signal and to determine a set of upmix parameters for the downmix audio signal, each set of upmix parameters comprising at least:

a level difference parameter indicative of a level difference between channels of the first multichannel signal;
a correlation parameter indicative of a coherence between channels of the first multichannel signal; and
a phase difference parameter indicative of a phase difference between channels of the first multichannel signal;

detecting at least one transient audio component of the first multichannel signal and to generate at least one transient upmix parameter for the at least one transient audio component of the first multichannel signal, the at least one transient upmix parameter being indicative of a level difference of the at least one transient audio component between channels of the first multichannel signal; and
generating the audio data signal to comprise the mono downmix audio signal, the sets of upmix parameters and the transient upmix parameter.

15. A computer program product comprising computer program code means adapted to perform all the steps of claims 13 or 14 when said program is run on a computer.

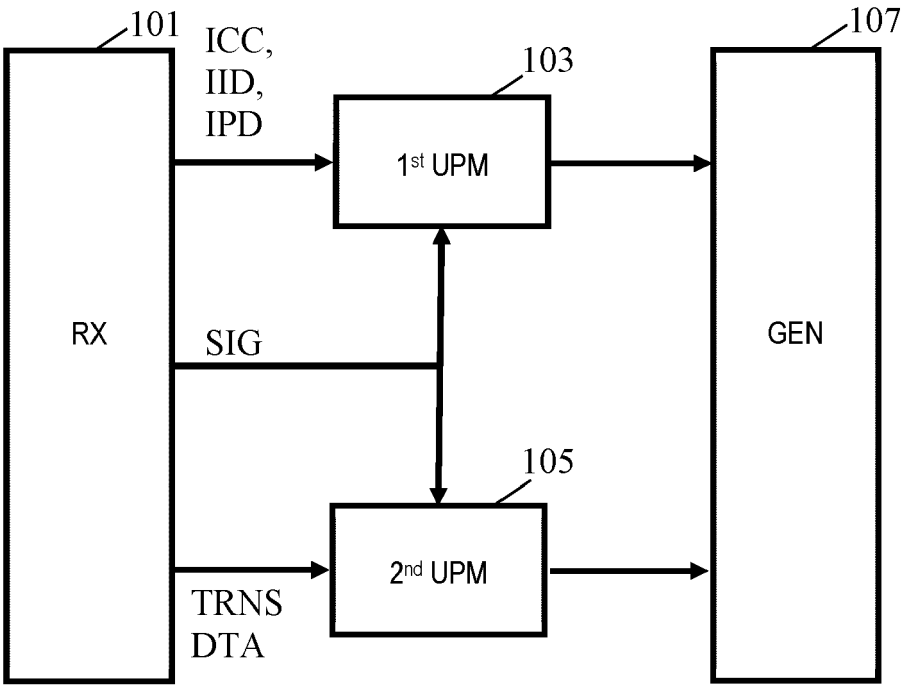


FIG. 1

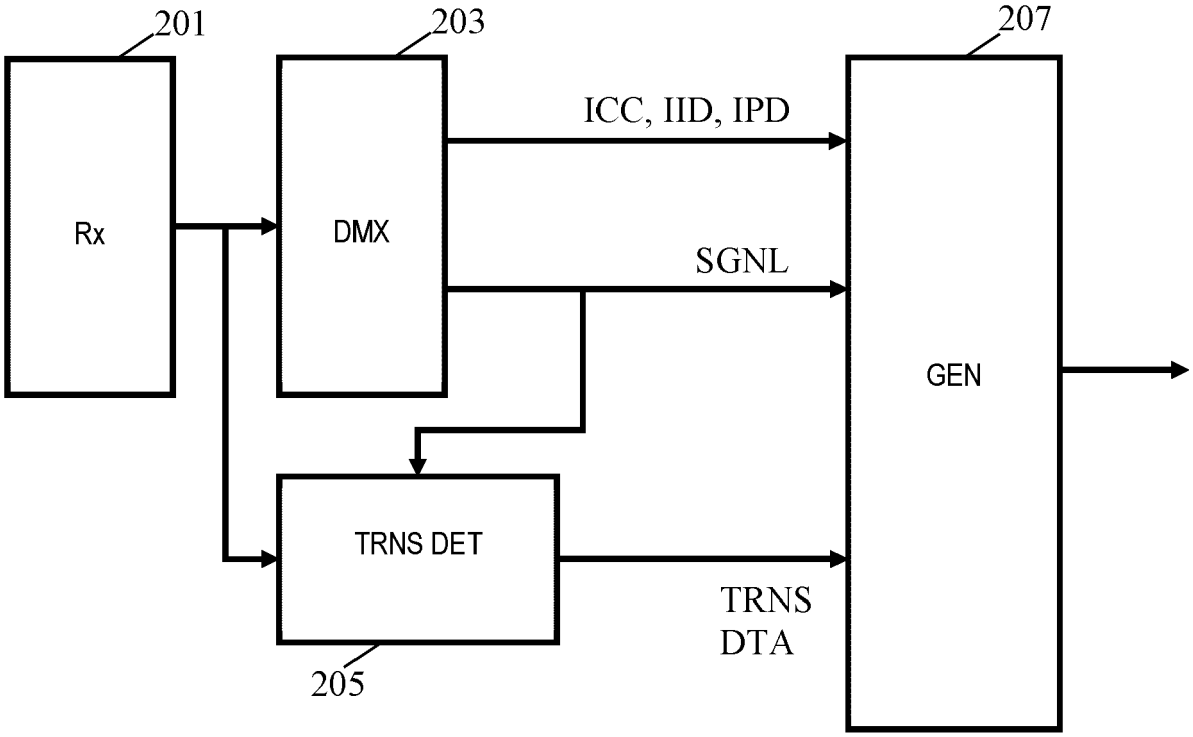


FIG. 2

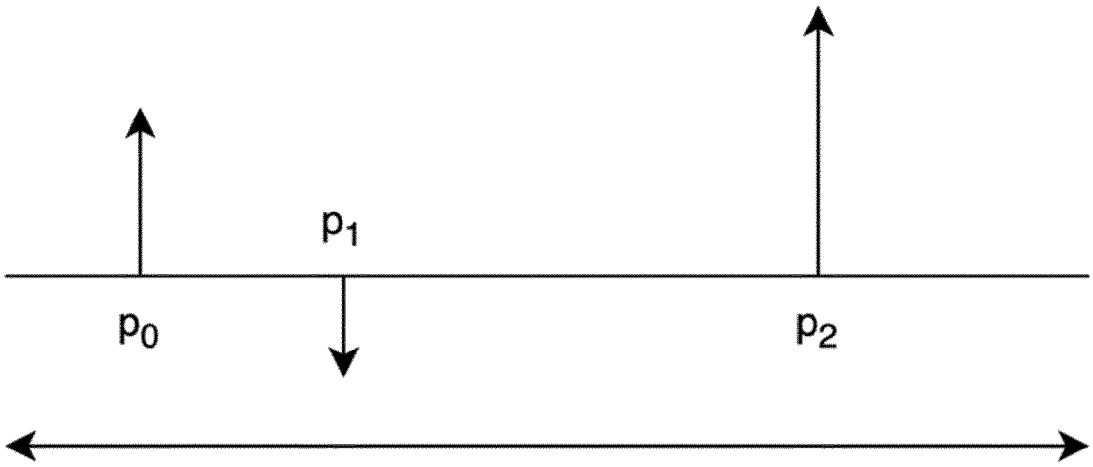


FIG. 3

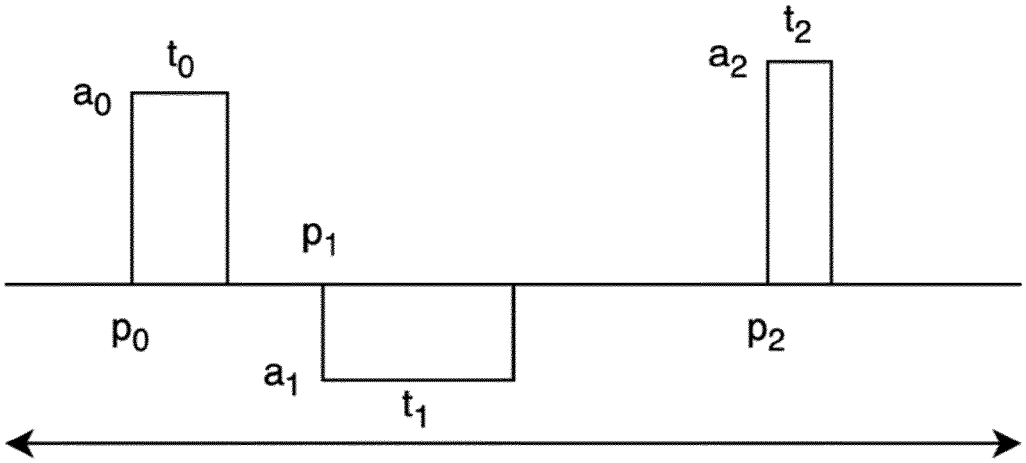
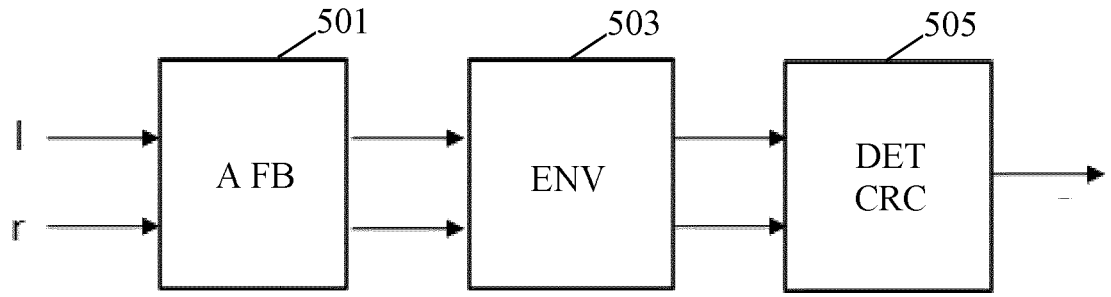


FIG. 4



207

FIG. 5

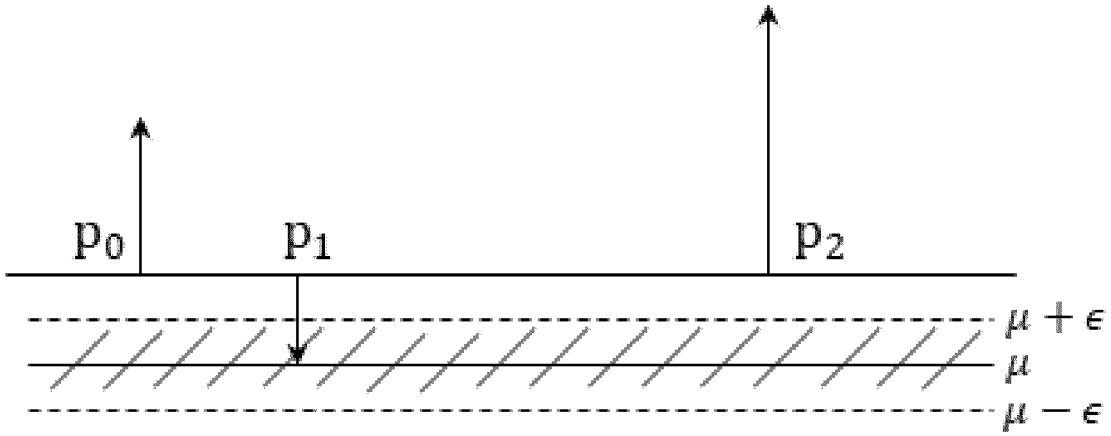


FIG. 6

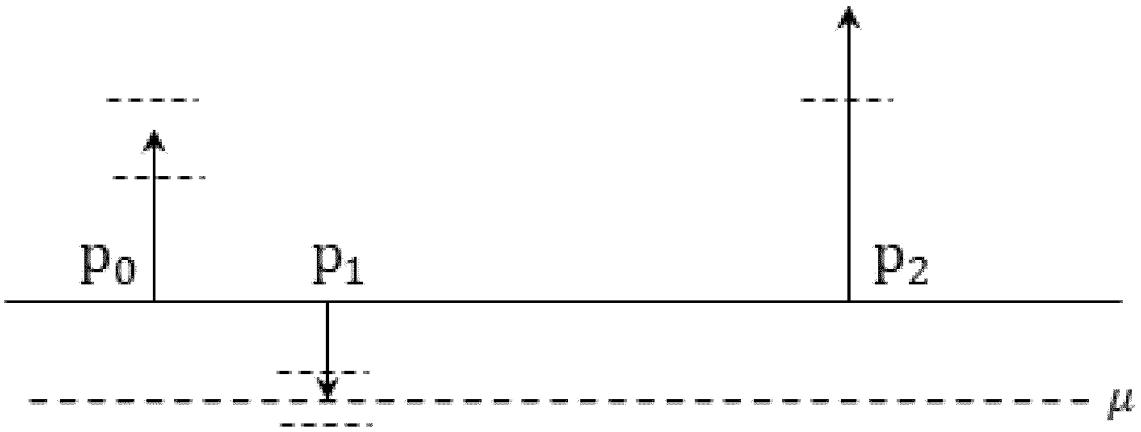


FIG. 7

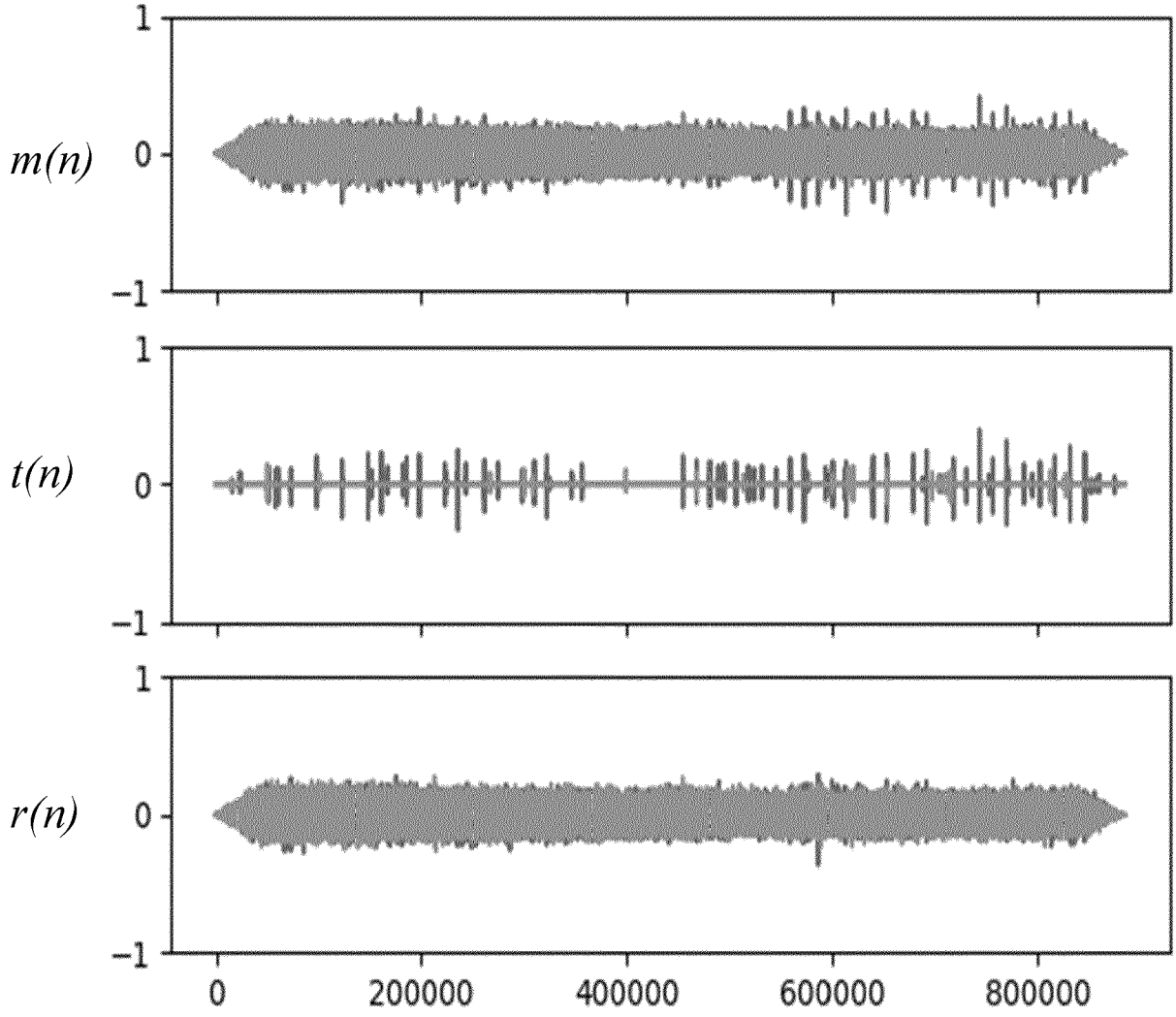


FIG. 8

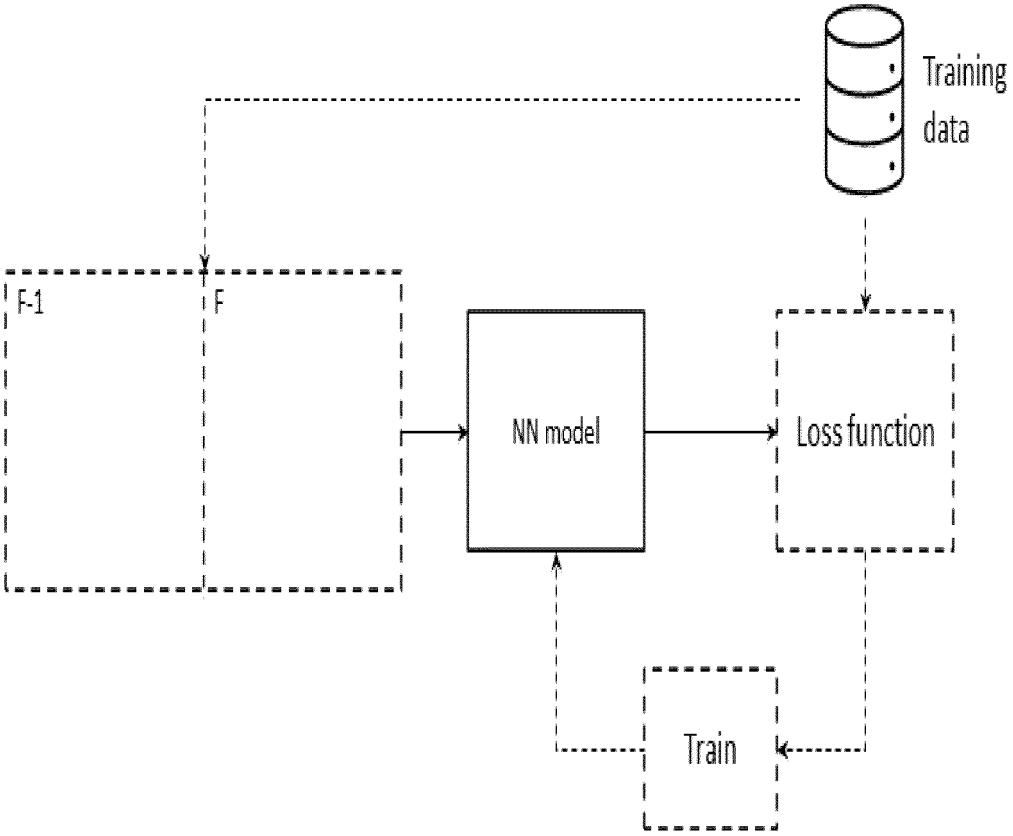


FIG. 9

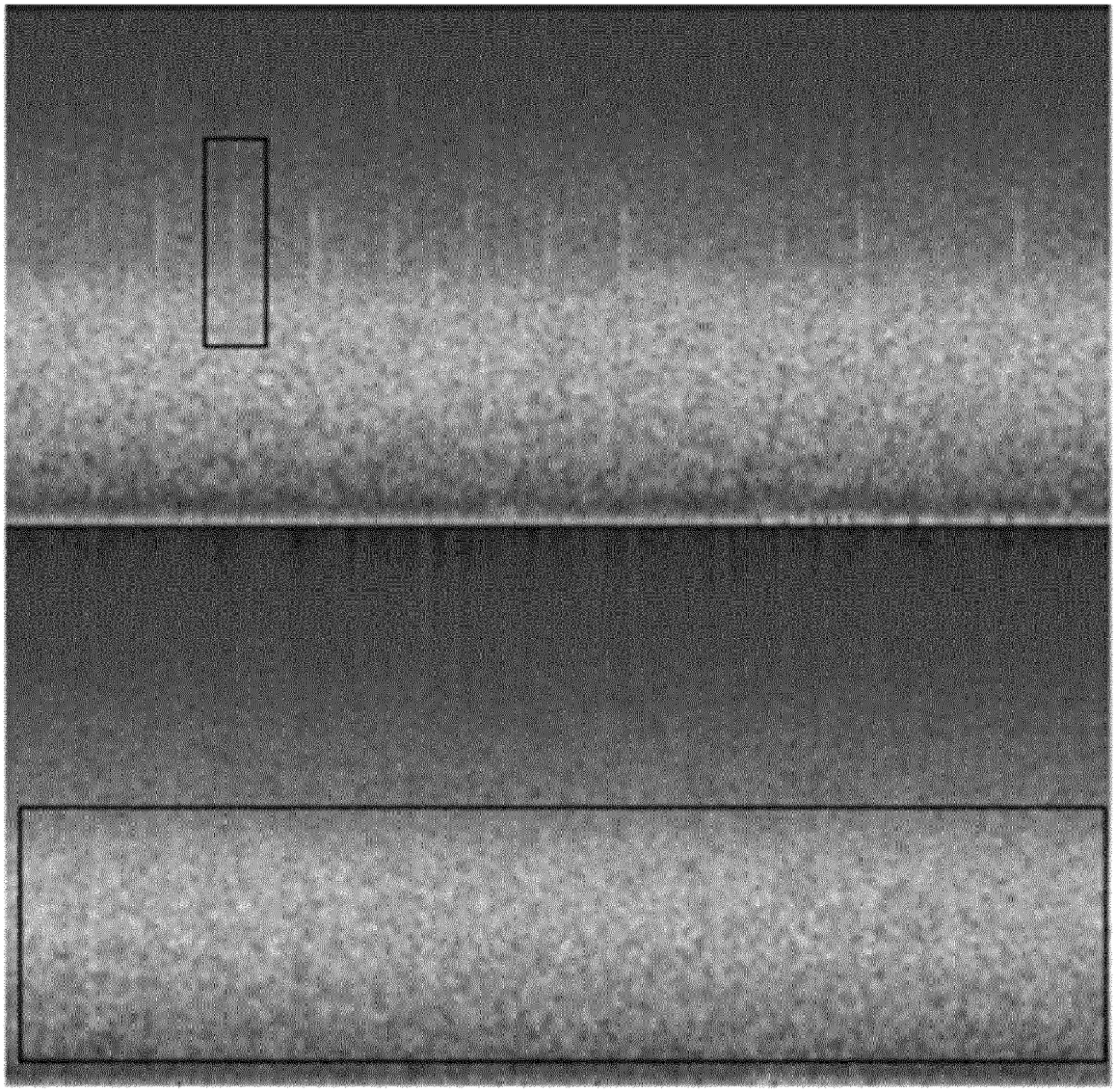


FIG. 10

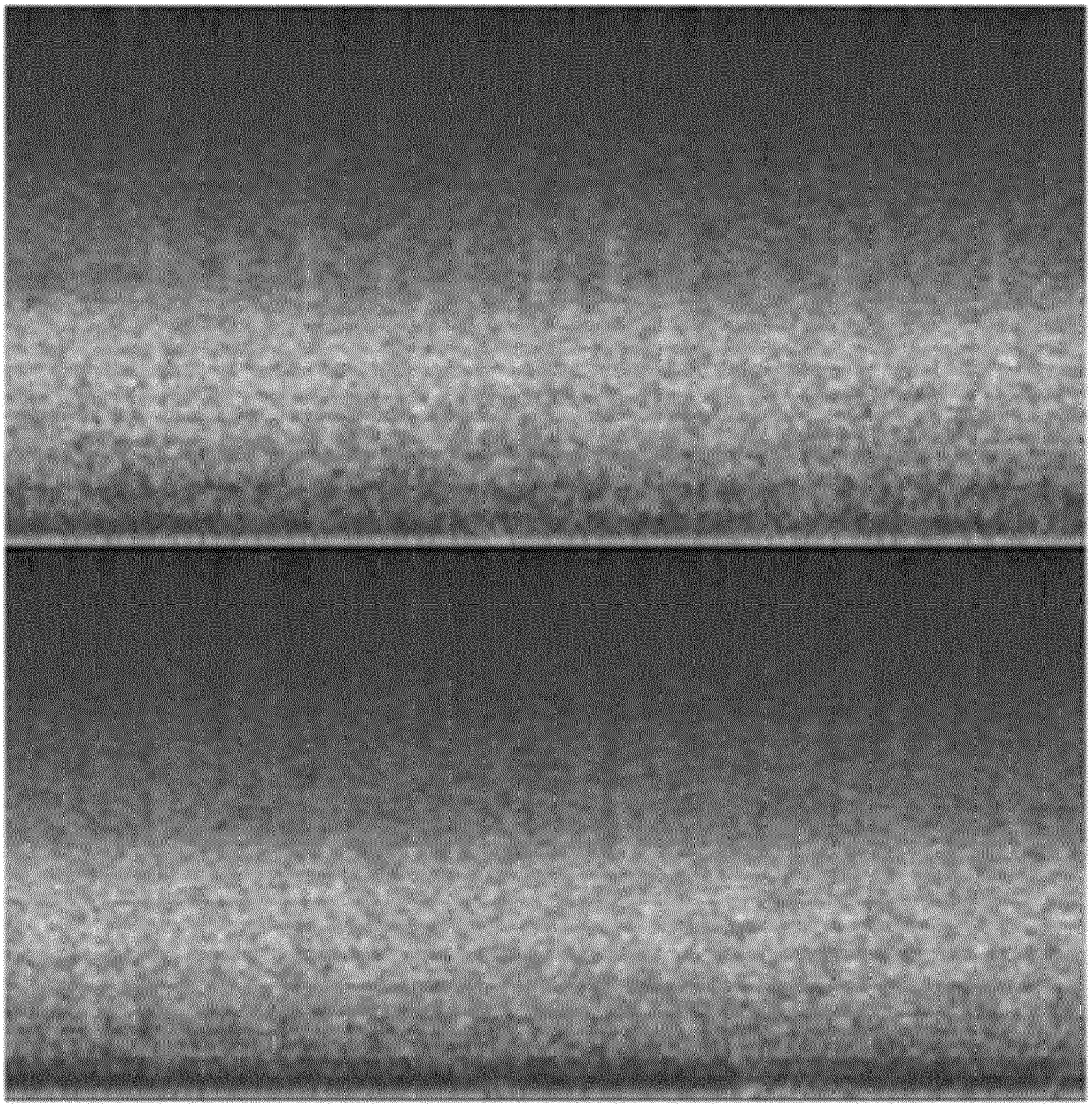


FIG. 11

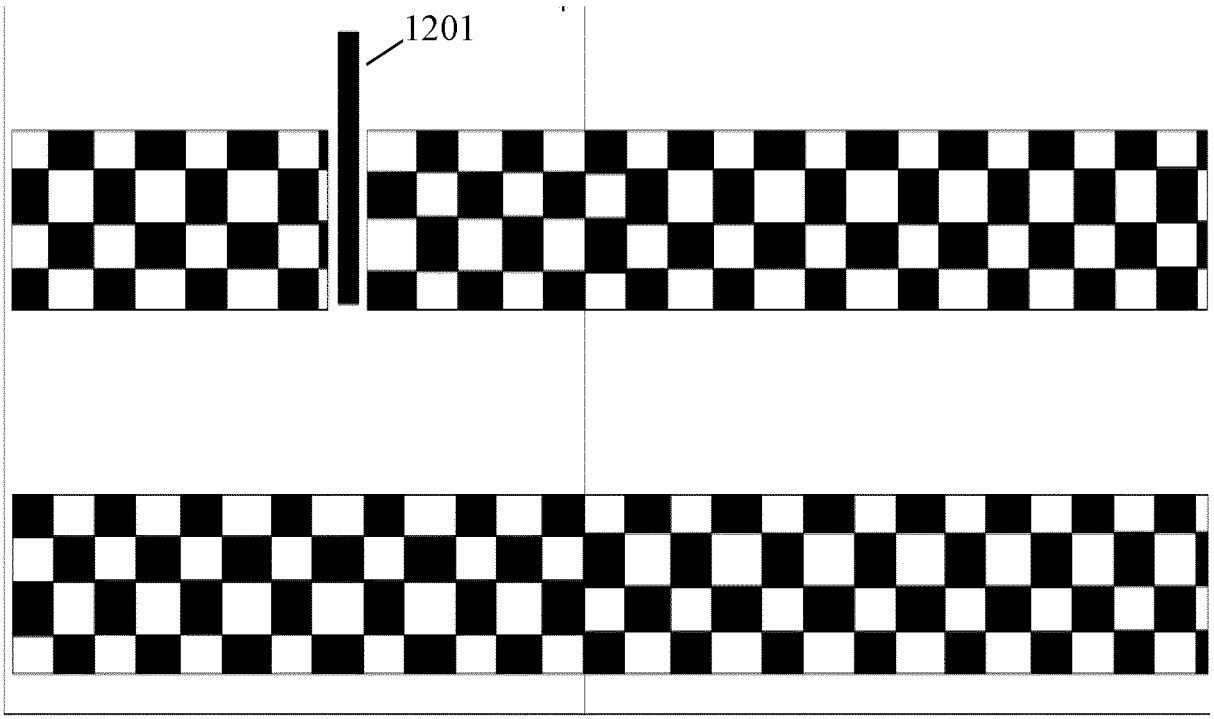


FIG. 12

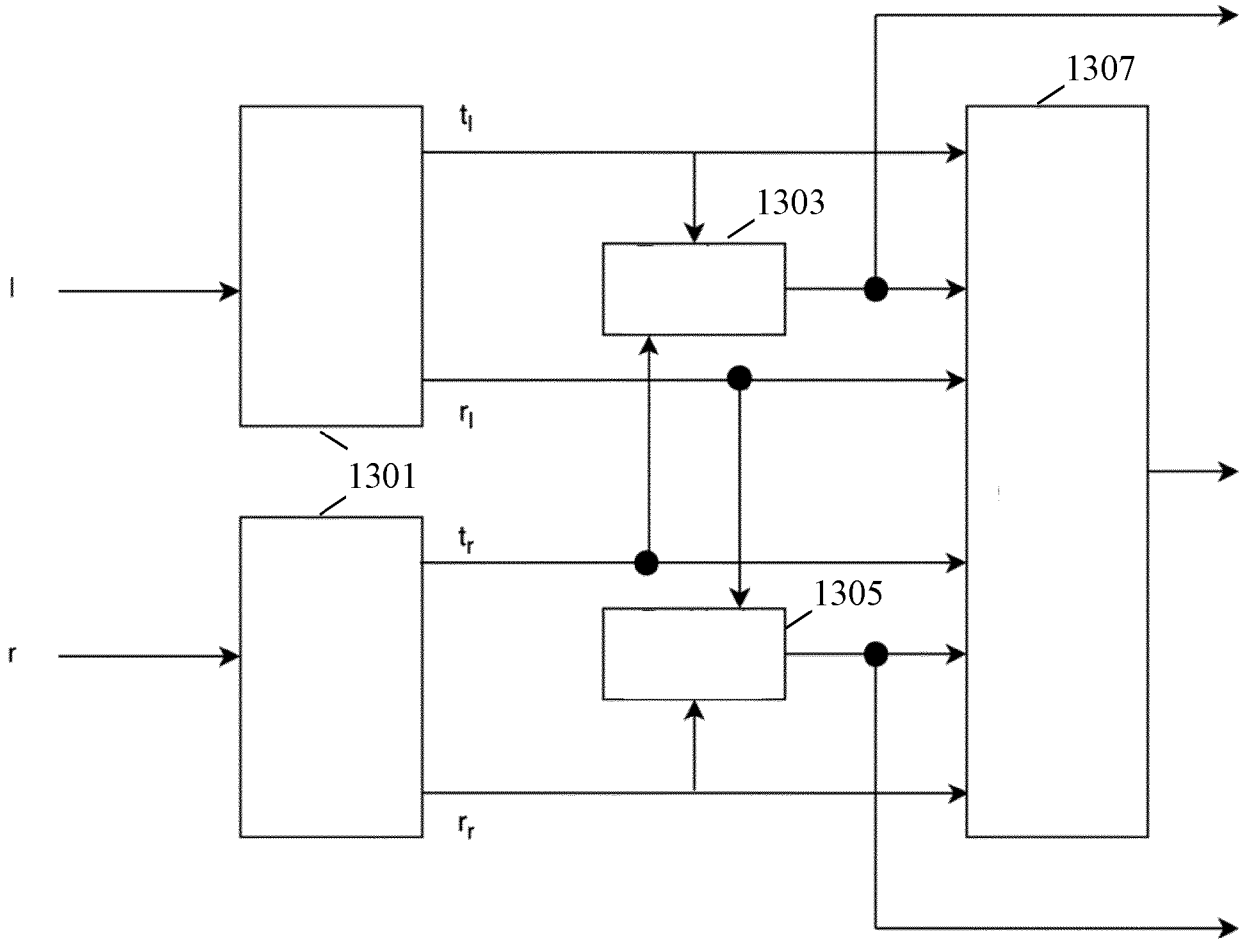


FIG. 13

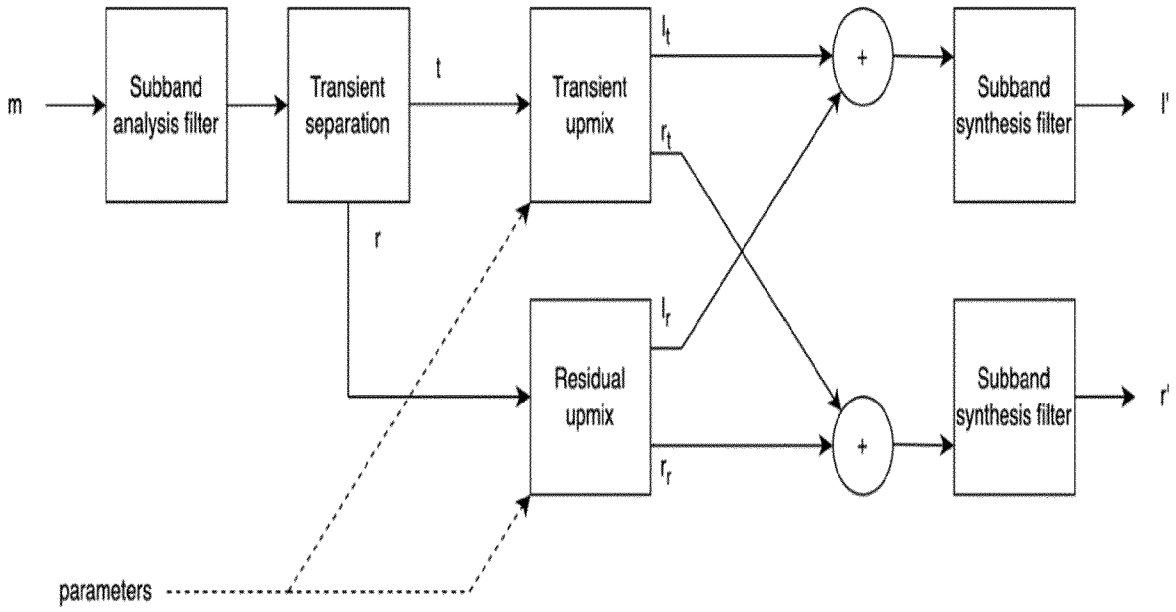

m
Subband analysis filter
Transient separation
t
r
Transient upmix
Residual upmix
l_t
r_t
l_r
r_r
+
+
Subband synthesis filter
Subband synthesis filter
l'
r'
parameters

FIG. 14

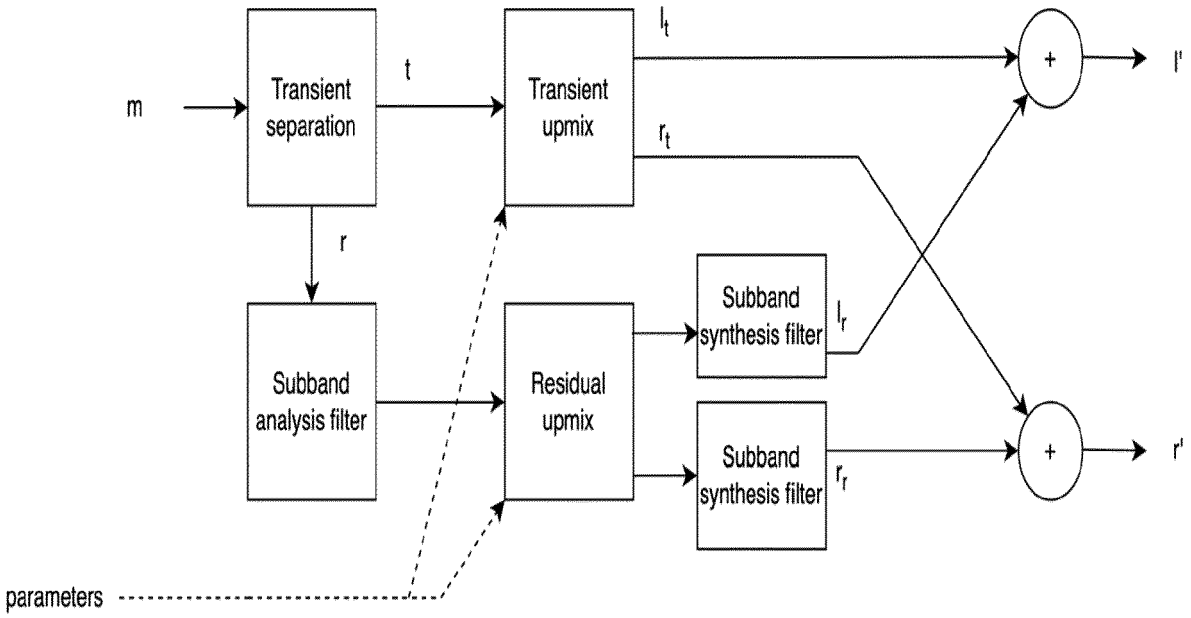

m
Transient separation
t
Transient upmix
l_t
+
l'
r_t
r
Subband synthesis filter
l_r
Subband analysis filter
Residual upmix
Subband synthesis filter
r_r
+
r'
parameters

FIG. 15

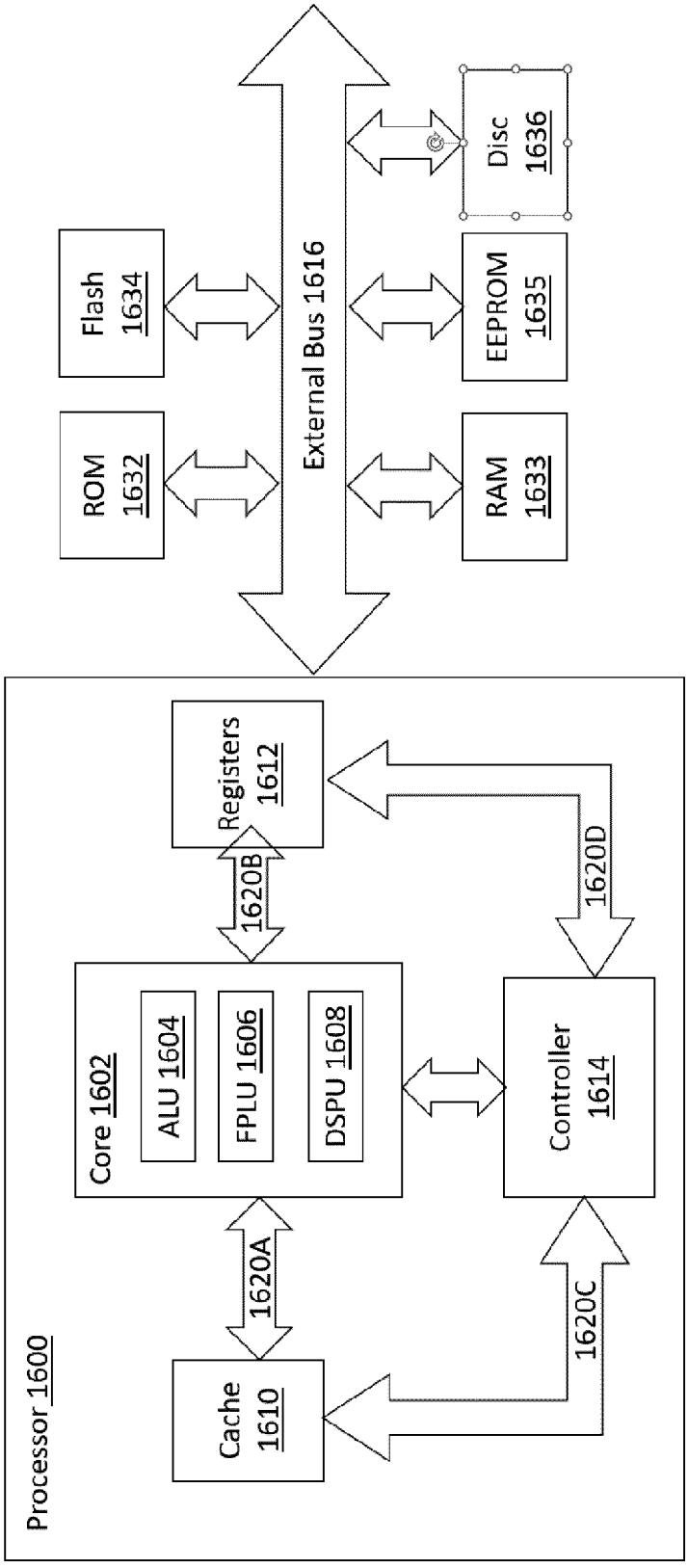

Processor 1600
Core 1602
ALU 1604
FPLU 1606
DSPU 1608
Registers 1612
Cache 1610
Controller 1614
1620A
1620B
1620C
1620D
External Bus 1616
ROM 1632
Flash 1634
RAM 1633
EEPROM 1635
Disc 1636

FIG. 16

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 23 19 9838

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	US 2009/319282 A1 (ALLAMANCHE ERIC [DE] ET AL) 24 December 2009 (2009-12-24) * paragraphs [0023], [0043], [0049] - [0054]; claim 4; figure 2 * * paragraphs [0102], [0103], [0104], [0133] - [0138] *	1,10, 13-15	INV. G10L19/008 ADD. G10L19/025
A	KUNTZ ACHIM ET AL: "The Transient Steering Decorrelator Tool in the Upcoming MPEG Unified Speech and Audio Coding Standard", AES CONVENTION 131; OCTOBER 2011, AES, 60 EAST 42ND STREET, ROOM 2520 NEW YORK 10165-2520, USA, 19 October 2011 (2011-10-19), XP040567607, * paragraph [0004]; figures 1,2 *	1-15	
A	JULIEN CAPOBIANCO ET AL: "Dynamic strategy for window splitting, parameters estimation and interpolation in spatial parametric audio coders", 2012 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP 2012) : KYOTO, JAPAN, 25 - 30 MARCH 2012 ; [PROCEEDINGS], IEEE, PISCATAWAY, NJ, 25 March 2012 (2012-03-25), pages 397-400, XP032227144, DOI: 10.1109/ICASSP.2012.6287900 ISBN: 978-1-4673-0045-2 * paragraphs [0003], [03.1], [03.2], [03.4] * * paragraph [03.4] *	1-9,13, 15	TECHNICAL FIELDS SEARCHED (IPC) G10L
A	US 2013/279702 A1 (HUAWEI TECH CO LTD [CN]) 24 October 2013 (2013-10-24) * paragraphs [0005], [0012], [0083], [0086] * * paragraph [0215]; figure 7 *	1,13	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	19 February 2024	Krembel, Luc

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 23 19 9838

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

19-02-2024

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009319282 A1	24-12-2009	AT E413792 T1	15-11-2008
		AU 2005299070 A1	04-05-2006
		BR PI0516392 A	02-09-2008
		CA 2583146 A1	04-05-2006
		CN 101044794 A	26-09-2007
		CN 101853660 A	06-10-2010
		EP 1803325 A1	04-07-2007
		ES 2317297 T3	16-04-2009
		HK 1104412 A1	11-01-2008
		IL 182235 A	31-10-2011
		JP 4625084 B2	02-02-2011
		JP 2008517334 A	22-05-2008
		KR 20070061882 A	14-06-2007
		NO 339587 B1	09-01-2017
		PL 1803325 T3	30-04-2009
		PT 1803325 E	13-02-2009
		RU 2384014 C2	10-03-2010
		TW I330827 B	21-09-2010
		US 2006085200 A1	20-04-2006
		US 2009319282 A1	24-12-2009
		WO 2006045373 A1	04-05-2006
US 2013279702 A1	24-10-2013	CN 103262158 A	21-08-2013
		EP 2612321 A1	10-07-2013
		JP 5681290 B2	04-03-2015
		JP 2013540283 A	31-10-2013
		US 2013279702 A1	24-10-2013
		WO 2012040898 A1	05-04-2012

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **E. SCHUIJERS** ; **W. OOMEN** ; **B. DEN BRINKER** ; **J. BREEBAART**. Advances in Parametric Coding for High-Quality Audio. *114th AES Convention, Amsterdam, The Netherlands*, 2003 **[0006]**
- **E. SCHUIJERS** ; **J. BREEBAART** ; **H. PUMHAGEN** ; **J. ENGDEGÅRD**. Low Complexity Parametric Stereo Coding. *116th AES, Berlin, Germany*, 2004 **[0006]**
- Transient-to-noise ratio restoration of coded applause-like signals. **ADAMI, A** ; **HERZOG, A** ; **DISCH, S** ; **HERRE, J.** 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, October 2017, 349-353 **[0118]**
- **MARTIN, RAINER**. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on speech and audio processing*, 2001, vol. 9 (5), 504-512 **[0125]**
- **DANIEL STOLLER** ; **SEBASTIAN EWERT** ; **SIMON DIXON**. *Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation*, 2018, http://arxiv.org/abs/1806.03185 **[0126]**