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

(11) EP 4 531 038 A1

(12) EUROPEAN PATENT APPLICATION

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

(21) Application number: **23199816.2**

(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:
- **KECHICHIAN, Patrick**
  Eindhoven (NL)
- **RAVI, Akshaya**
  Eindhoven (NL)
- **SCHUIJERS, Erik Gosuinus Petrus**
  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 first multichannel audio signal, sets of upmix parameters for time frequency segments of the downmix audio signal including a level difference parameter, a correlation parameter, and a phase difference parameter as well as at least one transient parameter indicative of a transient property for the first multichannel audio signal. An audio signal generator (103) generates an output multichannel audio signal by upmixing the downmix audio signal in dependence on the upmix parameters and the transient parameter. It uses an artificial neural network (105) having input nodes receiving the upmix parameters and the transient parameter. The transient parameter has a different time frequency resolution than the upmix parameters, and typically with a much coarser frequency resolution. An improved multichannel audio signal may be generated with little additional overhead in terms of processing complexity, resource, and data rate.

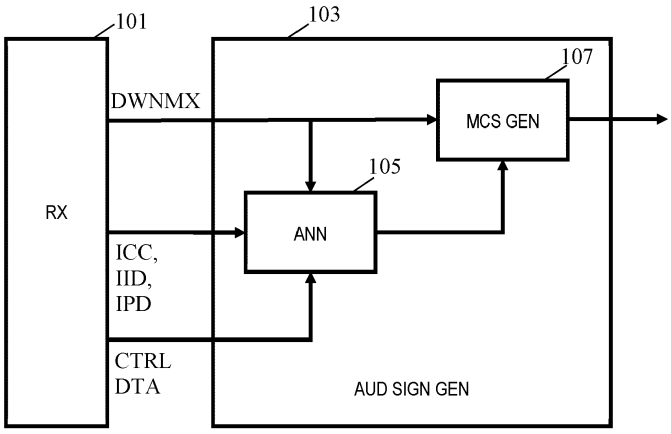


FIG. 1

EP 4 531 038 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 de-correlation process. The de-correlation 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 multi-channel audio 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 audio signal; sets of upmix parameters for time frequency segments of 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 audio signal (in a time frequency segment); a correlation parameter indicative of a coherence between channels of the first multichannel audio signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel audio signal; and neural network control data including: at least one transient parameter indicative of a transient property for the first multichannel audio signal, the transient parameter having a different time frequency resolution than the sets of upmix parameters; an audio signal generator arranged to generate the output multichannel audio signal by upmixing the downmix audio signal in dependence on the set of upmix parameters and the neural network control data, the audio signal generator comprising an artificial neural network having input nodes receiving the set of upmix parameters and input nodes receiving the transient parameter.

[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 possibility of utilizing artificial neural networks in audio processing, including typically multichannel audio decoding/reproduction. The approach may allow an advantageous employment of an artificial neural network in generating a multichannel audio signal from a downmix audio signal.

[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. The approach may enable or facilitate guidance by an encoder by providing a particular approach that is compatible with and exploits advantages that can be achieved by employing an artificial neural network as part of the audio processing.

[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 artificial neural network is a trained artificial neural network.

[0020] The artificial neural network may be a trained artificial neural network trained by training data including training downmix audio signals and training upmix parameters and neural network control data generated from training multi-channel audio signals; the training employing a cost function comparing the training multichannel audio signals to output multi-channel signals generated by the audio signal generator, using the training upmix parameter and neural network control data, from the training downmix signals. The artificial neural network may be a trained artificial neural network(s) trained by training data including training data representing a range of relevant audio sources including recording of videos, movies, telecommunications, etc.

[0021] The generator may be arranged to generate the multichannel audio 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 upmix parameters. The matrix may be time- and frequency-dependent.

[0022] The audio apparatus may specifically be an audio decoder apparatus.

[0023] The transient parameter data values may be referred to as metadata, conditioning features, latent representations, conditioning variables, and/or composites.

[0024] Each time frequency segment may consist in/correspond to a time interval and a frequency interval. Each time frequency segment may be a time frequency tile. The time frequency segments may be disjoint, and may be disjoint in both the frequency domain and/or time domain.

[0025] The time frequency segments or tiles 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.

[0026] The transient parameter having a different time frequency resolution than the sets of upmix parameters 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 (for the neural network control data).

[0027] According to an optional feature of the invention, the audio signal generator is arranged to generate the output multichannel audio signal by applying upmix coefficients to the downmix audio signal and a decorrelated signal generated from the downmix audio signal, and the artificial neural network is arranged to generate the upmix coefficients.

[0028] This may provide an advantageous approach for many scenarios and may provide an implementation highly suitable for employing artificial neural networks, including e.g. providing an advantageous trade-off between complexity, computational resources and/or the perceived audio quality of the generated multichannel audio signal. The audio signal generator may specifically be arranged to generate the output multichannel audio signal by applying a matrix multiplication to (samples of) the downmix audio signal and the decorrelated signal with the matrix coefficients being determined by the artificial neural network.

[0029] According to an optional feature of the invention, the audio signal generator is arranged to generate a decorrelated signal from the downmix audio signal and to generate at least one channel of the output multichannel audio signal by upmixing the downmix audio signal and the decorrelated signal, and the artificial neural network is arranged to control the generation of the decorrelated signal.

[0030] This may provide an advantageous approach for many scenarios and may provide an implementation highly suitable for employing artificial neural networks, including e.g. providing an advantageous trade-off between complexity, computational resources and/or the perceived audio quality of the generated multichannel audio signal. The artificial neural network may in many embodiments be arranged to generate signal samples of the decorrelated signal. In some embodiments, the artificial neural network may be arranged to generate parameter values for a decorrelator to which the downmix audio signal is applied to generate the decorrelated signal.

[0031] According to an optional feature of the invention, the artificial neural network comprises inputs for a segment of samples of the downmix audio signal and outputs providing samples of a segment of the output multichannel audio signal.

[0032] This may provide an advantageous approach for many scenarios and may provide an implementation highly suitable for employing artificial neural networks, including e.g. providing an advantageous trade-off between complexity, computational resources and/or the perceived audio quality of the generated multichannel audio signal.

[0033] According to an optional feature of the invention, the neural network control data comprises an interchannel level difference for each of a plurality of transients.

[0034] This may provide particularly advantageous neural network control data that may provide particularly suitable information on relevant transient properties. In many embodiments, the feature may allow improved audio quality to data rate trade-off.

[0035] According to an optional feature of the invention, the neural network control data comprises a timing parameter indicative of a timing of at least one transient.

[0036] This may provide particularly advantageous neural network control data that may provide particularly suitable information on relevant transient properties. In many embodiments, the feature may allow improved audio quality to data rate trade-off.

[0037] According to an optional feature of the invention, the neural network control data comprises no inter-channel correlation or inter-channel phase difference data for at least some transients of the first multichannel audio signal.

[0038] This may provide particularly advantageous neural network control data that may provide particularly suitable information on relevant transient properties. In many embodiments, the feature may allow improved audio quality to data rate trade-off.

[0039] According to an optional feature of the invention, the neural network control data has a lower frequency resolution than the upmix parameters.

[0040] This may provide particularly advantageous operation and may typically allow improved audio quality to data rate trade-off. In many embodiments, the at least one transient parameter may not be frequency dependent. E.g. a transient parameter value may apply to the entire frequency range of the downmix audio signal/multichannel audio signal. In some embodiments, a transient parameter value may be common to a plurality, and possibly all, subbands.

[0041] According to an optional feature of the invention, the neural network control data comprises data indicative of a

probability distribution property for transients of the first multichannel audio signal.

[0042] This may provide advantageous operation and/or implementation and/or performance in many embodiments.

[0043] According to an aspect of the invention, audio apparatus for generating an audio data signal, the audio apparatus comprising: a receiver receiving a first multichannel audio signal; a downmixer arranged to downmix the first multichannel audio signal to a downmix audio signal and determining sets of upmix parameters for time frequency segments of 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 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; a transient detector arranged to determine at least one transient parameter indicative of a transient property of the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix parameters; and a generator arranged to generate the audio data signal to comprise the downmix audio signal, the sets of upmix parameters, and neural network control data comprising the at least one transient parameter.

[0044] According to an optional feature of the invention, the transient detector is arranged to 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 differs from a second level difference measure indicative of a level difference between the channels by more than a threshold, the first level difference measure being determined for a shorter time interval than the second level difference, and to determine the at least one transient parameter to be indicative of the first level difference measure.

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

[0046] According to an aspect of the invention, method of generating an output multichannel audio 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 audio signal; sets of upmix parameters for time frequency segments of 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 audio signal; a correlation parameter indicative of a coherence between channels of the first multichannel audio signal; and a phase difference parameter indicative of a phase difference between channels of the first multichannel audio signal; and neural network control data including: at least one transient parameter indicative of a transient property for the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix parameters; generating the output multichannel audio signal by upmixing the downmix audio signal in dependence on the set of upmix parameters and the neural network control data, the output multichannel audio signal being generated in dependence on an output from an artificial neural network having input nodes receiving the set of upmix parameters and input nodes receiving the transient parameter.

[0047] According to an aspect of the invention, there is provided a method of operation of generating an audio data signal, the method comprises: receiving a first multichannel audio signal; downmixing the first multichannel audio signal to a downmix audio signal and determining sets of upmix parameters for time frequency segments of 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 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; determining at least one transient parameter indicative of a transient property of the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix parameters; and generating the audio data signal to comprise the downmix audio signal, the sets of upmix parameters and neural network control data comprising the at least one transient parameter.

[0048] 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

[0049] 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 structure of an artificial neural network;

FIG. 4 illustrates an example of a node of an artificial neural network;

FIG. 5 illustrates an example of a transient representation for a frame of an audio signal;

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

FIG. 7 illustrates an example of elements of a transient detector in accordance with some embodiments of the

invention;
FIG. 8 illustrates an example of a transient representation for a frame of an audio signal;
FIG. 9 illustrates an example of a transient representation for a frame of an audio signal;
FIG. 10 illustrates an example of a stereo audio signal, a stereo transient audio signal, and a stereo residual audio signal;
FIG. 11 illustrates an example of elements of an arrangement for training a neural network;
FIG. 12 illustrates an example of transients for a frame of an audio signal;
FIG. 13 illustrates an example of a structure of an artificial neural network;
FIG. 14 illustrates an example of a time frequency representation (spectrogram) of a stereo applause signal;
FIG. 15 illustrates an example of a time frequency representation (spectrogram) of a stereo applause signal;
FIG. 16 illustrates an example of time segments of a stereo signal comprising a transient; and
FIG. 17 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

[0050] 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).

[0051] 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.

[0052] 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).

[0053] 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.

[0054] 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.

[0055] 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.*

[0056] 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.

[0057] The encoder audio apparatus accordingly 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.

[0058] 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.

[0059] The receiver 101 is coupled to an audio signal generator 103 which generates the multichannel audio signal from the downmix signal and based on the upmix parameters. The audio signal generator 103 comprises an artificial neural network 105 that is coupled to a multichannel audio signal generator 107 which provides the output multichannel audio signal. The artificial neural network 105 may generate output samples/values based on the upmix parameters being provided as input values to the artificial neural network. In many embodiments, the downmix audio samples may also be provided to the artificial neural network 105. The multichannel audio signal generator 107 is arranged to generate the output multichannel audio signal from the output of the artificial neural network 105 and in many cases also from the downmix audio signal. It will be appreciated that the specific function of the multichannel audio signal generator 107 and the artificial neural network 105 (and the training thereof) will be different in different embodiments and a number of approaches will be described later.

[0060] 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.

[0061] 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.

[0062] 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.

[0063] The encoder audio apparatus further comprises a data signal generator 205 which generates the audio data signal to include data representing the downmix audio signal and the upmix parameters.

[0064] 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.

[0065] 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.

[0066] 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.

[0067] 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 e.g. 2048, 1024, or 512 samples. These signals may then be processed to generate samples for e.g. 64, 32 or 16 subbands. 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.

[0068] 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.

[0069] 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.

[0070] 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.

[0071] Typically, the subbands may have a bandwidth in the range from 10Hz to 10000Hz.

[0072] The audio signal generator 103 as mentioned comprises an artificial neural network 105 that is part of the generation of the output multichannel audio signal from the downmix audio signal and the sets of upmix parameters. The artificial neural network 105 may for example in various embodiments be arranged to e.g. generate parameter values/ weights for an upmixing of the downmix audio signal, to generate a decorrelation auxiliary audio signal corresponding to the downmix audio signal, to directly generate upmixed channel signals for an output multichannel audio signal etc.

[0073] An artificial neural network as used in the described functions may be a network of nodes arranged in layers and with each node holding a node value. FIG. 3 illustrates an example of a section of an artificial neural network.

[0074] The node value for a given node may be calculated to include contributions from some or often all nodes of a previous layer of the artificial neural network. Specifically, the node value for a node may be calculated as a weighted summation of the node values of all the nodes output of the previous layer. Typically, a bias may be added and the result may be subjected to an activation function. The activation function provides an essential part of each neuron by typically providing a non-linearity. Such non-linearities and activation functions provides a significant effect in the learning and adaptation process of the neural network. Thus, the node value is generated as a function of the node values of the previous layer.

[0075] The artificial neural network may specifically comprise an input layer 301 comprising a plurality of nodes receiving the input data values for the artificial neural network. Thus, the node values for nodes of the input layer may typically directly be the input data values to the artificial neural network and thus may not be calculated from other node values.

[0076] The artificial neural network may further comprise none, one, or more hidden layers 303, 305 or processing layers. For each of such layers, the node values are typically generated as a function of the node values of the nodes of the previous layer, and specifically a weighted combination and added bias followed by an activation function (such as a sigmoid, ReLu, or tanh function may be applied).

[0077] Specifically, as shown in FIG. 3, each node, which may also be referred to as a neuron, may receive input values (from nodes of a previous layer) and therefrom calculate a node value as a function of these values. Often, this includes first generating a value as a linear combination of the input values with each of these weighted by a weight:

$$k = \sum_{n} w_n x_n$$

where w refers to weights, x refers to the nodes of the previous layer and n is an index referring to the different nodes of the previous layer.

[0078] An activation function may then by applied to the resulting combination. For example, the node value l may be determined as:

$$l = f(k)$$

where the function may for example be a Rectified Linear Unit function as described in Xavier Glorot, Antoine Bordes, Yoshua Bengio Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR 15:315-323, 2011):

$$f(k) = ReLU(k) = \max(0, k)$$

[0079] Other often used functions include a sigmoid function or a tanh function. In many embodiments, the node output or value may be calculated using a plurality of functions. For example, both a ReLU and Sigmoid function may be combined using an activation function such as:

$$f(k) = ReLU(k) + \sigma(k)$$

[0080] Such operations may be performed by each node of the artificial neural network (except for typically the input nodes).

[0081] The artificial neural network further comprises an output layer 307 which provides the output from the artificial neural network, i.e. the output data of the artificial neural network is the node values of the output layer. As for the hidden/ processing layers, the output node values are generated by a function of the node values of the previous layer. However, in contrast to the hidden/ processing layers where the node values are typically not accessible or used further, the node values of the output layer are accessible and provide the result of the operation of the artificial neural network.

[0082] A number of different networks structures and toolboxes for artificial neural network have been developed and in many embodiments the artificial neural network may be based on adapting and customizing such a network. An example of a network architecture that may be suitable for the applications mentioned above is WaveNet by van den Oord et al which is described in Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." arXiv preprint arXiv: 1609.03499 (2016).

[0083] WaveNet is an architecture used for the synthesis of time domain signals using dilated causal convolution, and has been successfully applied to audio signals. For WaveNet the following activation function is commonly used:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right),$$

where * denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W represents the weights of the learned artificial neural network. The filter product of the equation may typically provide a filtering effect with the gating product providing a weighting of the result which may in many cases effectively allow the contribution of the node to be reduced to substantially zero (i.e. it may allow or "cutoff' the node providing a contribution to other nodes thereby providing a "gate" function). In different circumstances, the gate function may result in the output of that node being negligible, whereas in other cases it would contribute substantially to the output. Such a function may substantially assist in allowing the neural network to effectively learn and be trained.

[0084] An artificial neural network may in some cases further be arranged to include additional contributions that allow the artificial neural network to be dynamically adapted or customized for a specific desired property or characteristics of the generated output. For example, a set of values may be provided to adapt the artificial neural network. These values may be included by providing a contribution to some nodes of the artificial neural network. These nodes may be specifically input nodes but may typically be nodes of a hidden or processing layer. Such adaptation values may for example be weighted and added as a contribution to the weighted summation/ correlation value for a given node. For example, for WaveNet such adaptation values may be included in the activation function. For example, the output of the activation function may be given as:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}\right)$$

where **y** is a vector representing the adaptation values and V represents suitable weights for these values.

[0085] The above description relates to a neural network approach that may be suitable for many embodiments and implementations. However, it will be appreciated that many other types and structures of neural network may be used. Indeed, many different approaches for generating a neural network have been, and are being, developed including neural

networks using complex structures and processes that differ from the ones described above. The approach is not limited to any specific neural network approach and any suitable approach may be used without detracting from the invention.

[0086] In the approach, the encoder audio apparatus is arranged to generate control data for the artificial neural network 105 which is included in the audio data signal and transmitted to the decoder audio apparatus where it is provided as an input to the artificial neural network. This neural network control data is thus data that is processed by the artificial neural network 105 and with the output of the artificial neural network 105 being dependent on the neural network control data.

[0087] The encoder audio apparatus is arranged to determine the neural network control data to include at least one transient parameter that is dependent on/ represents a transient property for the first multichannel audio signal.

[0088] The specific transient parameter(s) and property(ies) that are transmitted and input to the artificial neural network 105 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).

[0089] The transient data/parameter is 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.

[0090] 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.

[0091] 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.

[0092] 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.

[0093] A specific example of a time segment/ frame FRM in which three transients are detected is shown in FIG. 6. Here, for a given time segment/ frame three transients are detected at positions p_0, p_1, and pz. 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 a_0, a_1 and a_2 can additionally or alternatively be transmitted. Thus, the transient data may be used to encode a representation of the detected transients.

[0094] In some embodiments, as illustrated in FIG. 6, the duration of a transient may be determined, encoded, and communicated to the decoder audio apparatus in the audio data signal.

[0095] 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.

[0096] 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 p_0, p_1 and p_2 may be coded differentially to the (average over the frequency bands) IID transmitted (per band) for the whole frame.

[0097] 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 input to the artificial neural network 105. The encoder audio apparatus may provide this 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 to the artificial neural network 105 of an upmixer 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.

[0098] 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 transients in the first multichannel audio signal also detects transients in the downmix audio signal and vice versa).

[0099] FIG. 7 illustrates an example of elements of a transient detector 207 that may be used in the encoder audio apparatus. In the example, the transient detector 207 may be arranged to detect transients in a stereo signal.

[0100] 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.

[0101] In many embodiments, as will be described in the following, the transient detector 207 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.

[0102] In the example of FIG. 7, the transient detector 207 includes an analysis filterbank 701 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 701 may equivalently be part of the receiver 201 and the generated subband representation may also be used for the downmixing, upmix parameter estimation etc.

[0103] The QMF decomposition may produce complex-valued outputs and the transient detector 207 comprises an envelope circuit 703 which determines the real envelope of both left and right channels. An example of an 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. This envelope is a type of (absolute) spectrogram with there being no summation over the different bands to generate a time-domain envelope.

[0104] It should be noted thought 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.

[0105] The transient detector 207 comprises a detection circuit 705 which is coupled to the envelope circuit 703 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.

[0106] 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.

[0107] A short sliding window w corresponding to approximately 10 ms is used to compute the IID between left and right channels:

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

[0108] 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.

[0109] 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.

[0110] 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.

[0111] 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.

[0112] 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.

[0113] 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.

[0114] FIG. 8 illustrates the same stereo transient representation as in FIG. 5 and 6 but with an average IID of the frame also being shown (μ) along with an IID range around the average IID given by $\pm\epsilon$. 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.

[0115] 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.

[0116] 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. 9. 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).

[0117] 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.

[0118] In the example, short term properties are accordingly compared to longer term properties and a transient is detected in terms of these differing sufficiently.

[0119] 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.

[0120] 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.

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

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

where $m_s(n)$ and $m_f(n)$ are the slow and fast envelopes respective with α_s, $\alpha_f \in (0,1]$ and $\alpha_f \gg \alpha_s$.

[0121] 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.

[0122] 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.

[0123] 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.

[0124] 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.

[0125] 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.

[0126] Similarly, a residual signal may be generated, e.g. as:

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

[0127] FIG. 10 illustrates an example of a stereo input signal $m(n)$ (with the two channels being represented overlaid by each other and with different shades of grey) being separated into a stereo transient signal $t(n)$ and a stereo residual signal r(n).

[0128] 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.

[0129] 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.

[0130] 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.

[0131] 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. 11.

[0132] 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.

[0133] 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 i^{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.

[0134] An alternative loss function can be based on the mean-squared error between the true ground-truth and predicted labels,

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

[0135] 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.

[0136] As previously mentioned, the audio signal generator 103 may generate the output multichannel audio signal using an artificial neural network as part of the processing and with the artificial neural network having an input comprising both the sets of upmix parameters and the transition parameters. Different approaches may be used in different embodiments, and in particular different signals or parameters that are used as part of the upmixing/ multichannel audio signal generation may be generated by the artificial neural network in different embodiments.

[0137] In some embodiments, the artificial neural network 105 comprises inputs for a segment of samples of the downmix audio signal. In some embodiments, the samples may be time domain samples of the downmix audio signal. In other embodiments, the samples may e.g. be time frequency samples, such as specifically subband samples of a frame/ segment.

[0138] In some cases, the artificial neural network 105 may be arranged to directly generate the output multichannel audio signal. Thus, the artificial neural network 105 may have output nodes that directly provide samples of the output multichannel audio signal. The samples may in some embodiments directly be time domain samples of the channel signals of the output multichannel audio signal. In other examples, the artificial neural network 105 may generate subband samples of the output multichannel audio signals. The generated output multichannel audio signal samples may be fed to the multichannel audio signal generator 107 which in the former example e.g. may simply forward the samples or e.g. may perform post processing of the generated samples of the output multichannel audio signal. In the latter case, the multichannel audio signal generator 107 may perform a processing that e.g. includes a frequency to time domain transformation to convert subband samples into time domain samples.

[0139] In some such cases, the decoder audio apparatus may be arranged to generate a decorrelated version of the downmix audio signal by inputting the downmix audio signal to a suitable decorrelator. In such cases, samples of both the downmix audio signal and the decorrelated signal may e.g. be provided to the artificial neural network 105 for the generation of the samples of the output multichannel audio signal.

[0140] In such an approach, the training of the artificial neural network 105 may for example be performed by generating a large number of training input multichannel audio signals that are processed by an encoder audio apparatus in accordance with e.g. prescribed specifications or standards. The data may then be provided to the decoder audio apparatus which may from these generate an output multichannel audio signal. A cost function for the training may then be determined by a comparison of the generated output multichannel audio signal and the original training input multichannel audio signal.

[0141] In some embodiments, the decoder audio apparatus may be arranged to generate the output multichannel audio signal in response to an upmixing of the downmix audio signal and a decorrelated signal which is a decorrelated version of the downmix audio signal. The decorrelated signal may have the same overall properties as the downmix audio signal in terms of frequency envelope, average energy etc., but being decorrelated with the downmix audio signal.

[0142] For example, for a traditional Parametric Stereo, PS, processing an upmixing of a mono downmix audio signal *m* may be based on a decorrelated signal *d* by using the following approach for upmixing:

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

[0143] 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:

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

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

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

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

[0144] In some embodiments, the artificial neural network 105 may be arranged to generate the decorrelated signal, and thus based on an input including samples of the downmix audio signal, the sets of upmix parameters, and the transient parameter(s) the artificial neural network 105 proceeds to generate samples of the decorrelated signal. These samples (and thus the decorrelated signal generated by the artificial neural network 105) may then be fed to the audio signal generator 103 where the matrix multiplication upmix is performed.

[0145] In some embodiments, the audio signal generator 103 is arranged to generate the output multichannel audio signal from the downmix audio signal as well as from a decorrelated audio signal in dependence on the parametric upmix data. The generator may specifically for the stereo case generate the output multichannel audio signal by applying a time- and frequency-dependent 2x2 matrix multiplication to the samples of the downmix audio signal and the decorrelated signal. The coefficients of the 2x2 matrix are determined from the upmix parameters of the upmix parametric data, typically on a time and frequency band basis. For other upmix operations, such as from a mono or stereo downmix signal to a five channel multichannel audio signal, the audio signal generator 103 may apply matrix multiplications with matrices of suitable dimensions.

[0146] It will be appreciated that many different approaches of generating such a multichannel audio signal from a downmix audio signal and a decorrelated signal, and for determining suitable matrix coefficients from upmix parametric data, will be known to skilled person and that any suitable approach may be used. Specifically, various approaches for PS upmixing that are based on downmix and auxiliary audio signals are well known to the skilled person.

[0147] In conventional systems, the upmixing includes generating a decorrelated signal of the mono audio signal determined by applying the downmix audio signal to a decorrelator function. It has been found that by generating a decorrelated signal and mixing this with the mono audio signal, an improved quality of the upmix signal is perceived and therefore decoders have been developed to exploit this. The decorrelated signal is typically generated by a decorrelator in the form of an all-phase filter that is applied to the mono audio signal. However, whereas the use of such an all pass filter tends to result in a multichannel audio signal being generated that is perceived to be of improved quality, it is still not ideal, and some audio quality degradation may often be perceived.

[0148] In some embodiments, the decorrelated signal is not generated by a straightforward filtering of the downmix/ mono audio signal, but rather a decorrelated audio signal is generated by the artificial neural network 105 with the decorrelated signal being used by the multichannel audio signal generator 107 to generate the multichannel audio signal based on the upmix parameters.

[0149] In some embodiments, the artificial neural network 105 may thus directly be trained to generate the decorrelated signal. It has been found that this may in many embodiments provide a substantially improved performance with a more realistic sounding output multichannel audio signal typically be perceived.

[0150] In some embodiments, the artificial neural network 105 may not directly generate the decorrelated signal but may for example generate one or more parameters for a decorrelator which is applied to the downmix audio signal to generate the decorrelated signal. For example, the artificial neural network 105 may as an output generate parameters, such as filter coefficients, for an all pass filter that is applied to the downmix audio signal to generate the decorrelated signal.

[0151] Similarly to the previously described approaches, the training of the artificial neural network 105 may for example be performed by generating a large number of training input multichannel audio signals that are processed by an encoder audio apparatus in accordance with e.g. prescribed specifications or standards. The generated data may then be provided to the decoder audio apparatus which may from these generate an output multichannel audio signal. A cost function for the training may then be determined by a comparison of the generated output multichannel audio signal and the original

training input multichannel audio signal. Thus, the training of the artificial neural network 105 may be based on an end to end cost function that includes the upmixing etc. Such a trained artificial neural network has been found to result in improved audio quality in many scenarios and for many signals.

[0152] In some embodiments, the artificial neural network 105 may be arranged to directly generate the upmix coefficients for upmixing the downmix audio signal and one or more auxiliary signals, such as specifically for upmixing the downmix audio signal and a decorrelated signal. In the example above, the artificial neural network 105 may thus directly generate the coefficients of the upmix matrix:

$$\begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}$$

[0153] In the approach, the upmix procedure may include a pre-trained network determining upmix parameters that are applied to the downmix audio signal and at least one decorrelated signal to generate the upmixed output multichannel audio signal.

[0154] In the approach, the artificial neural network 105 may be trained to directly generate the coefficients h_{xx}, for a given set of legacy PS parameters (represented by the sets of upmix parameters) and a given set of stereo transient sequences (represented by the transient parameters), while minimizing the (perceptual) loss of the synthesized stereo output signal compared to the original stereo signal. An example architecture could be a fully connected network, receiving the PS parameters for the current frame, the PS parameters of the previous frame, the current sample index n and the value of the stereo transient sequence at that position. An example of the input parameters temporal relations are illustrated in FIG. 12.

[0155] FIG. 13 illustrates elements of a simplified diagram of a fully connected network which has input nodes for the PS parameters, the stereo transient sequence s[n] and the relative sample position n inside the frame. For simplicity only a few connections and only the real-valued part of the upmix entry *h* are shown.

[0156] In some embodiments, instead of feeding the artificial neural network 105 with (legacy) PS parameters of the previous frame F-1 and the current frame F, the legacy upmix equations may be employed first to pre-calculate the upmix entries at frame F-1 and frame F. In some cases, such coefficients calculated by predetermined formulas/equations may then be input to the artificial neural network 105 and modified coefficients may be calculated. This may result in a lower complexity network as no network capacity needs to be applied in modelling the original upmix equations.

[0157] An alternative network architecture may employ so-called gated activation units, as e.g. also employed in WaveNet. A gated activation unit can be described as:

$$z = \tanh(W_f * x) \cdot \sigma(W_g * s),$$

where x is the input signal (at a certain layer/position of the network), s is the stereo transient signal, W_f is the filter (convolution) that is applied to the input signal (at a certain layer/position in the network), W_g is the gate (convolution) that is applied to the stereo transient signal (at a certain layer/position in the network), $\sigma()$ is a non-linear sigmoid function and tanh() a non-linear tangent function. In case of zero stereo transients, the combination of all left parts in the above equation will effectively mimic traditional stereo upmixing, whereas in case the gate function is activated, the combination of the left and right parts of the above equation will generate proper stereo output taking into account the stereo image parameters.

[0158] In some embodiments, the neural network control data comprises data indicative of a probability distribution property for transients of the first multichannel audio signal. Thus, rather than the transient data (only) providing a stereo transient representation which is fully deterministic, the transient data may (alternatively or additionally) provide an indication of a probability distribution for the transients. The transient data may include stochastic components where e.g. some parameters like amplitude and or duration may be synthesized using a stochastic (noise-like) process.

[0159] In more detail, for example, if we assume that the transient amplitude IID parameters follow a Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

[0160] At the decoder side we could generate values that follow that same distribution using a zero bias, unity variance Gaussian noise generator $g(n)$:

$$E[g] = 0$$

$$E[(g - E[g])^2] = 1$$

[0161] The resulting transient amplitude IID parameters can then be generated as:

$$g'(n) = \sigma g(n) + \mu$$

[0162] This requires that the parameters σ and μ are determined on the encoder side and transmitted to the decoder. This can be realized by simple measuring the mean and variance of the collected transient amplitude IID parameters, followed by simple quantization into a limited number of discrete values:

$$\mu = \sum_{\forall i} IID(i)$$

$$\sigma^2 = \sum_{\forall i} (IID(i) - \mu)^2$$

[0163] Above the example of a Gaussian distribution is provided. Other probability density functions may be applicable, e.g. for generating temporal positions of the transients a uniform distribution may be employed.

[0164] In the described approach, the upmixing to generate the output multichannel audio signal employs an artificial neural network which in addition to the normal signals and parameters that are used to perform upmixing (e.g. the downmix audio signal and upmix parameters relating properties between different channels of the original multichannel audio signal) is provided with control data that is determined in dependence on transient properties of the original multichannel signal/ downmix audio signal.

[0165] 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). 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 artificial neural network thus allowing a better training and output of the artificial neural network. Further, the representation 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.

[0166] To illustrate this, 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 transient-specific parameters may also increase the overall bit-rate.

[0167] To further appreciate the effects of a low (per-frame) parameter update rate, the signal of FIG. 14 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.

[0168] The output of a conventional parametric stereo decoder for the same segment is shown in FIG. 15. 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.

[0169] 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. 16. A foreground clap 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.

[0170] 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.

[0171] Therefore, providing transient data with a higher time resolution allows for additional information that can be included in the artificial neural network 105 by a training process thereby allowing an improved output audio signal to be generated.

[0172] The approach may in particular allow such transient information to be adapted to the specific importance of information of the transients. Indeed, for the transients, it has been found that the artificial neural network is much less sensitive to frequency dependencies than to timing accuracy and 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.

[0173] Further, it has been found that 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.

[0174] 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 to the artificial neural network and typically has significantly less impact on the resulting audio quality of the generated output multichannel signal.

[0175] The artificial neural network 105 is specifically trained to provide suitable output data by employing a training process e.g. as part of the manufacturing or design phase. The result of the training process, such as coefficients etc. for the different nodes, may be performed once with the results than being used for all manufactured apparatuses.

[0176] Artificial neural networks are adapted to specific purposes by a training process which are used to adapt/ tune/ modify the weights and other parameters (e.g. bias) of the artificial neural network. It will be appreciated that many different training processes and algorithms are known for training artificial neural networks. Typically, training is based on large training sets where a large number of examples of input data are provided to the network. Further, the output of the artificial neural network is typically (directly or indirectly) compared to an expected or ideal result. A cost function may be generated to reflect the desired outcome of the training process. In a typical scenario known as supervised learning, the cost function often represents the distance between the prediction and the ground truth for a particular input data. Based on the cost function, the weights may be changed and by reiterating the process for the modified weights, the artificial neural network may be adapted towards a state for which the cost function is minimized.

[0177] In more detail, during a training step the neural network may have two different flows of information from input to output (forward pass) and from output to input (backward pass). In the forward pass, the data is processed by the neural network as described above while in the backward pass the weights are updated to minimize the cost function. Typically, such a backward propagation follows the gradient direction of the cost function landscape. In other words, by comparing the predicted output with the ground truth for a batch of data input, one can estimate the direction in which the cost function is minimized and propagate backward, by updating the weights accordingly. Other approaches known for training artificial neural networks include for example Levenberg-Marquardt algorithm, the conjugate gradient method, and the Newton method etc.

[0178] In the present case, training may specifically include a training set comprising a potentially large number of multichannel audio signals. In some embodiments, training data may be multichannel audio signals in time segments corresponding to the processing time intervals of the artificial neural networks being trained, e.g. the number of samples in a training multichannel audio signal may correspond to a number of samples corresponding to the input nodes of the artificial neural network(s) being trained. Each training example may thus correspond to one operation of the artificial neural network(s) being trained. Usually, however, a batch of training samples is considered for each step to speed up the training process. Furthermore, many upgrades to gradient descent are possible also to speed up convergence or avoid local minima in the cost function landscape.

[0179] For each training multichannel audio signal, a training processor may perform a downmix operation to generate a downmix audio signal and corresponding upmix parametric and transient data. Thus, the encoding process that is applied to the multichannel audio signal during normal operation may also be applied to the training multichannel audio signal thereby generating a downmix and the upmix parametric data.

[0180] Specifically, for a stereo multichannel audio signal, the training processor may use a Parametric Stereo scheme (e.g. in accordance with a suitable standardized approach). Such an encoding will apply a frequency- and time-dependent matrix operation, e.g. a rotation operation to the input stereo signal to generate a downmix signal and a residual signal. For example, typically a 2x2 matrix multiplication/ complex value multiplication is applied to the input stereo signals to e.g. substantially align one of the rotated channel signals to have a maximum signal value. This channel may be used as the mono-signal and the rotation is typically performed on a frame basis. The rotation value may be stored as part of the upmix

parametric data (or a parameter allowing this to be determined may be included in the upmix parametric data). Thus, in a synthesis apparatus, the opposite rotation may be performed to reconstruct the stereo signal. The rotation of the stereo signal results in another stereo signal of which one channel is accordingly aligned with the maximum intensity. The other channel is typically discarded in a Parametric Stereo encoder in order to reduce the data rate. In conventional PS decoding, a decorrelated signal is typically generated at the decoder and used for the upmixing process. In the current training approach this second signal may be used as a residual signal for the downmixing as it may represent the information discarded in the encoder, and thus it represents the ideal signal to be reconstructed in the decoder as part of an upmixing process.

[0181] Thus, in some embodiments, a training processor may from training multichannel audio signals generate training downmix signals and sets of upmix parameters and transient parameters (and possibly a training residual signal). This training data may be fed to an decoder audio apparatus comprising the artificial neural network 105 to generate an output multichannel audio signal. A cost function is applied to determine a cost value for each training downmix audio signal and/or for the combined set of training downmix audio signals (e.g. an average cost value for the training sets is determined). The cost function may include various components.

[0182] Typically, the cost function will include at least one component that reflects how close a generated signal is to a reference signal, i.e. a so-called reconstruction error. In some embodiments the cost function will include at least one component that reflects how close a generated signal is to a reference signal from a perceptual point of view.

[0183] Typically, the generated multichannel audio signal may be compared to the original multichannel audio signal input to the encoder audio apparatus and a difference measure may be determined and used as a cost function. This process may be generated for all training sets to generate an overall cost function.

[0184] It will be appreciated that many different approaches may be used to determine the cost value reflecting difference between the signals. For example, a correlation may be performed with the cost value having a monotonically decreasing value for the increasing correlation value. As another example, the two signals may be subtracted from each other and a power measure for the difference signal may be used as a cost value. It will be appreciated that many other approaches are available and may be used.

[0185] Thus, in the example, the cost function generates a cost value that reflects how closely the generated multi-channel audio signal match the corresponding original training multichannel audio signals.

[0186] Based on the cost value, the training processor may adapt the weights of the artificial neural network 105. For example, a back-propagation approach may be used. In particular, the training processor may adjust the weights of the artificial neural network 105 based on the cost value. For example, given the derivative (representing the slope) of the weights with respect to the cost function the weights values are modified to go in the direction of the slope. For a simple/minima account one can refer to the training of the perceptron (single neuron) in case of backward pass of a single data input.

[0187] The process may be iterated until the artificial neural network is considered to be trained. For example, training may be performed for a predetermined number of iterations. As another example, training may be continued until the weights change be less than a predetermined amount. Also very common, a validation stop is implemented where the network is tested again a validation metric and stopped when reaching the expected outcome.

[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. 17 is a block diagram illustrating an example processor 1700 according to embodiments of the disclosure. Processor 1700 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 1700 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 1700 may include one or more cores 1702. The core 1702 may include one or more Arithmetic Logic Units (ALU) 1704. In some embodiments, the core 1702 may include a Floating Point Logic Unit (FPLU) 1706 and/or a Digital Signal Processing Unit (DSPU) 1708 in addition to or instead of the ALU 1704.

[0191] The processor 1700 may include one or more registers 1712 communicatively coupled to the core 1702. The registers 1712 may be implemented using dedicated logic gate circuits (e.g., flip-flops) and/or any memory technology. In some embodiments the registers 1712 may be implemented using static memory. The register may provide data, instructions and addresses to the core 1702.

[0192] In some embodiments, processor 1700 may include one or more levels of cache memory 1710 communicatively coupled to the core 1702. The cache memory 1710 may provide computer-readable instructions to the core 1702 for execution. The cache memory 1710 may provide data for processing by the core 1702. In some embodiments, the

computer-readable instructions may have been provided to the cache memory 1710 by a local memory, for example, local memory attached to the external bus 1716. The cache memory 1710 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 1700 may include a controller 1714, which may control input to the processor 1700 from other processors and/or components included in a system and/or outputs from the processor 1700 to other processors and/or components included in the system. Controller 1714 may control the data paths in the ALU 1704, FPLU 1706 and/or DSPU 1708. Controller 1714 may be implemented as one or more state machines, data paths and/or dedicated control logic. The gates of controller 1714 may be implemented as standalone gates, FPGA, ASIC or any other suitable technology.

[0194] The registers 1712 and the cache 1710 may communicate with controller 1714 and core 1702 via internal connections 1720A, 1720B, 1720C and 1720D. 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 1700 may be provided via a bus 1716, which may include one or more conductive lines. The bus 1716 may be communicatively coupled to one or more components of processor 1700, for example the controller 1714, cache 1710, and/or register 1712. The bus 1716 may be coupled to one or more components of the system.

[0196] The bus 1716 may be coupled to one or more external memories. The external memories may include Read Only Memory (ROM) 1732. ROM 1732 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) 1733. RAM 1733 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) 1735. The external memory may include Flash memory 1734. The External memory may include a magnetic storage device such as disc 1736. 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 audio 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 audio signal;
sets of upmix parameters for time frequency segments of 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

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

neural network control data including:

at least one transient parameter indicative of a transient property for the first multichannel audio signal, the transient parameter having a different time frequency resolution than the sets of upmix parameters;

an audio signal generator (103) arranged to generate the output multichannel audio signal by upmixing the downmix audio signal in dependence on the set of upmix parameters and the neural network control data, the audio signal generator (103) comprising an artificial neural network (105) having input nodes receiving the set of upmix parameters and input nodes receiving the transient parameter.

2. The audio apparatus of claim 1 wherein the audio signal generator (103) is arranged to generate the output multichannel audio signal by applying upmix coefficients to the downmix audio signal and a decorrelated signal generated from the downmix audio signal, and the artificial neural network (105) is arranged to generate the upmix coefficients.

3. The audio apparatus of any previous claim wherein the audio signal generator (103) is arranged to generate a decorrelated signal from the downmix audio signal and to generate at least one channel of the output multichannel audio signal by upmixing the downmix audio signal and the decorrelated signal, and the artificial neural network (105) is arranged to control the generation of the decorrelated signal.

4. The audio apparatus of any previous claim wherein the artificial neural network (105) comprises inputs for a segment of samples of the downmix audio signal and outputs providing samples of a segment of the output multichannel audio signal.

5. The audio apparatus of any previous claim wherein the neural network control data comprises an interchannel level difference for each of a plurality of transients.

6. The audio apparatus of any previous claim wherein the neural network control data comprises a timing parameter indicative of a timing of at least one transient.

7. The audio apparatus of any previous claim wherein the neural network control data comprises no inter-channel correlation or inter-channel phase difference data for at least some transients of the first multichannel audio signal.

8. The audio apparatus of any previous claim wherein the neural network control data has a lower frequency resolution than the upmix parameters.

9. The audio apparatus of any previous claim wherein the neural network control data comprises data indicative of a probability distribution property for transients of the first multichannel audio signal.

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

a receiver (201) receiving a first multichannel audio signal;
a downmixer (203) arranged to downmix the first multichannel audio signal to a downmix audio signal and determining sets of upmix parameters for time frequency segments of 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 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;

a transient detector (207) arranged to determine at least one transient parameter indicative of a transient property of the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix parameters; and
a generator (205) arranged to generate the audio data signal to comprise the downmix audio signal, the sets of upmix parameters, and neural network control data comprising the at least one transient parameter.

11. The audio apparatus of claim 10 wherein the transient detector (207) is arranged to 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 differs from a second level difference measure indicative of a level difference between the channels by more than a threshold, the first level difference measure being determined for a shorter time interval than the second level difference, and to determine the at least one transient parameter to be indicative of the first level difference measure.

12. An audio system comprising an audio apparatus for generating an audio data signal as claimed in claim 10 or 11 and an audio apparatus for generating an output multichannel audio signal from the audio data signal as claimed in any of claims 1 to 9.

13. A method of generating an output multichannel audio 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 audio signal;
sets of upmix parameters for time frequency segments of 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 audio signal;
a correlation parameter indicative of a coherence between channels of the first multichannel audio signal; and
a phase difference parameter indicative of a phase difference between channels of the first multichannel audio signal; and

neural network control data including:

at least one transient parameter indicative of a transient property for the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix parameters;

generating the output multichannel audio signal by upmixing the downmix audio signal in dependence on the set of upmix parameters and the neural network control data, the output multichannel audio signal being generated in dependence on an output from an artificial neural network (207) having input nodes receiving the set of upmix parameters and input nodes receiving the transient parameter.

14. A method of operation of generating an audio data signal, the method comprises:

receiving a first multichannel audio signal;
downmixing the first multichannel audio signal to a downmix audio signal and determining sets of upmix parameters for time frequency segments of 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 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;

determining at least one transient parameter indicative of a transient property of the first multichannel audio signal, the at least one transient parameter having a different time frequency resolution than the sets of upmix

parameters; and
generating the audio data signal to comprise the downmix audio signal, the sets of upmix parameters and neural network control data comprising the at least one transient 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.

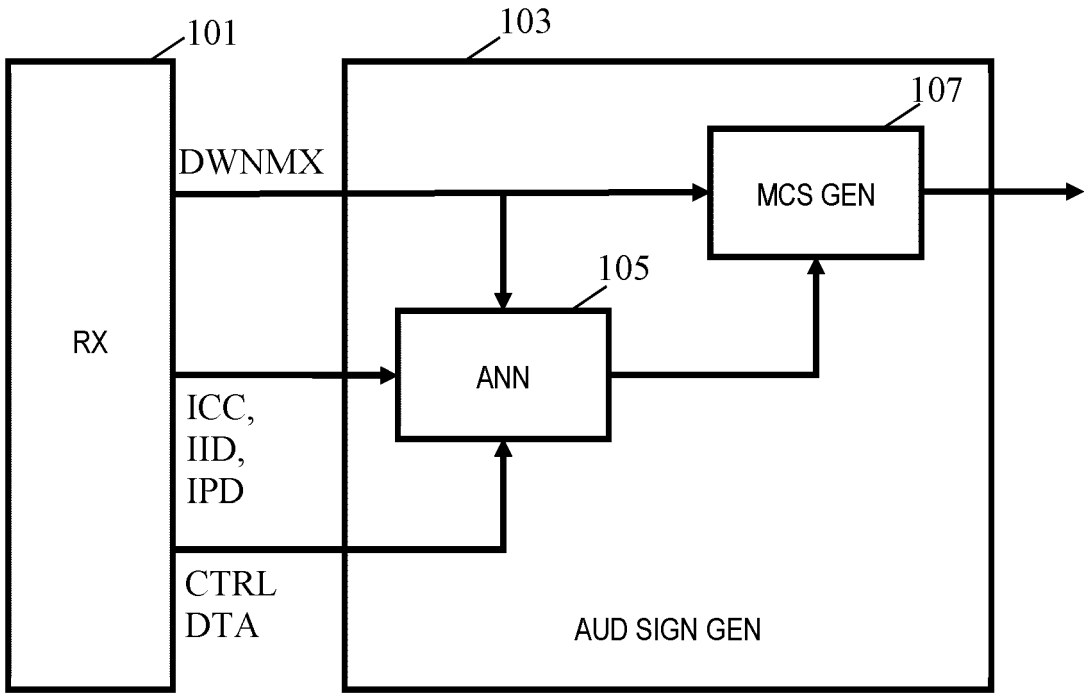

101
103
107
DWNMX
MCS GEN
105
RX
ANN
ICC,
IID,
IPD
CTRL
DTA
AUD SIGN GEN

FIG. 1

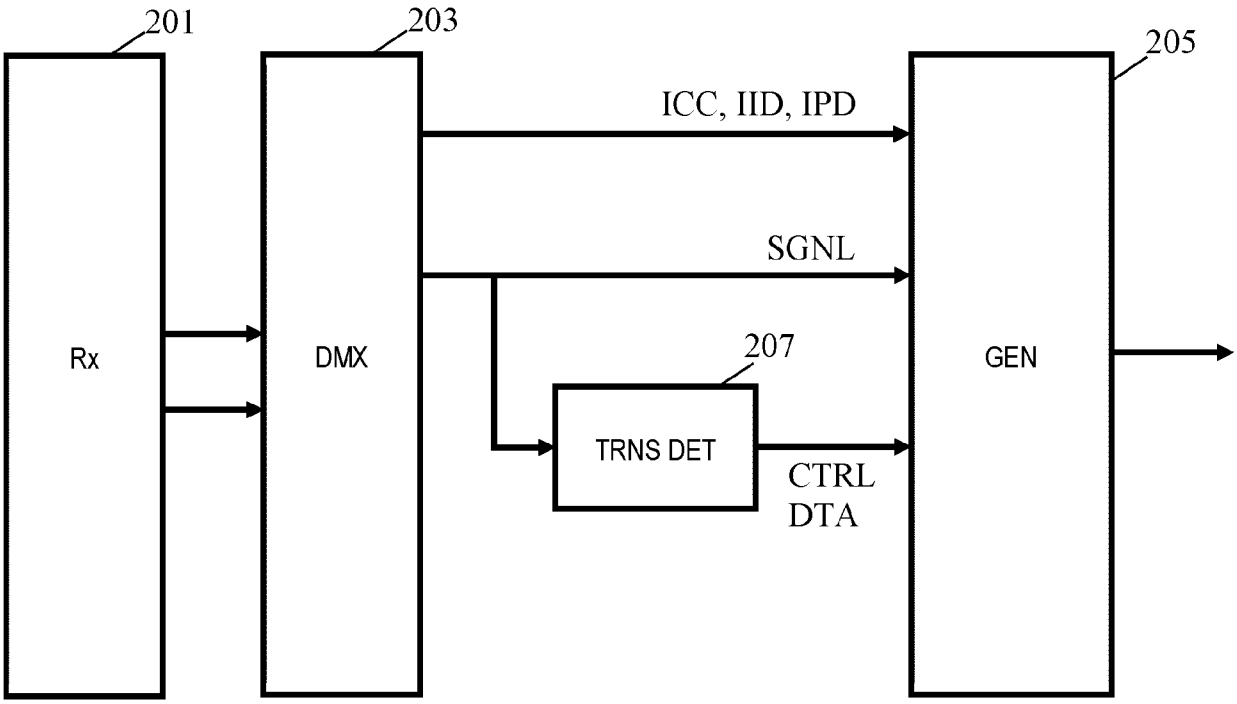


FIG. 2

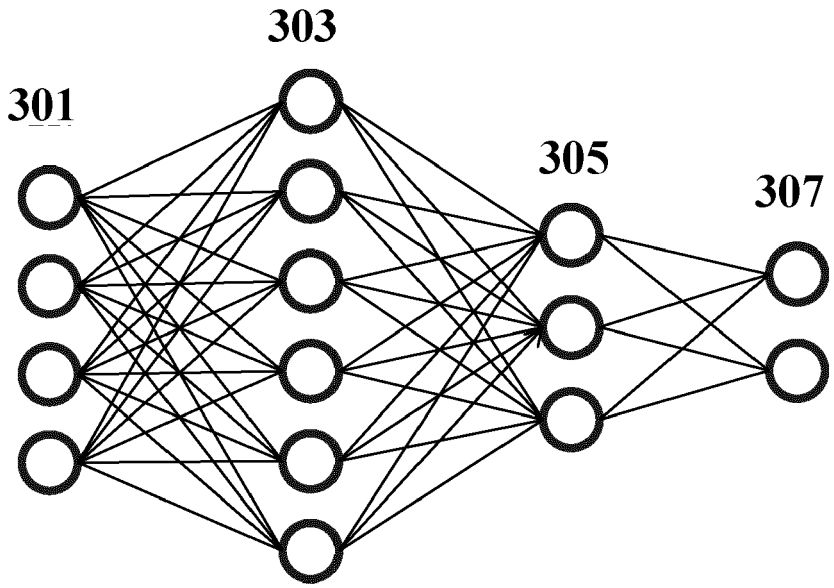
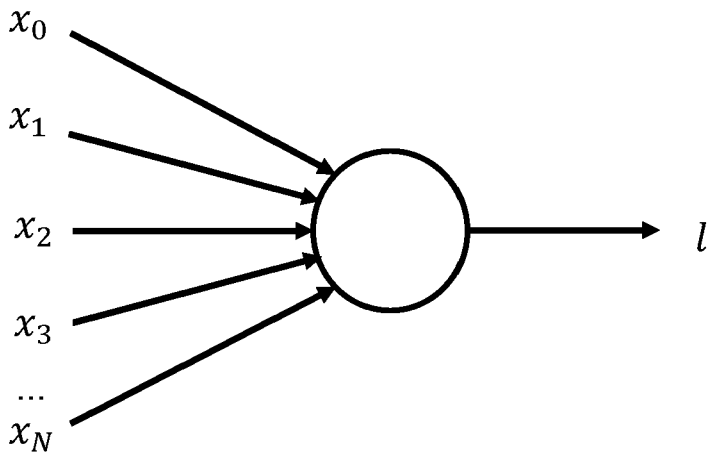


FIG. 3



$$\mathrm{l} = \max\left(0, \sum_{n} w_n x_n\right)$$

FIG. 4

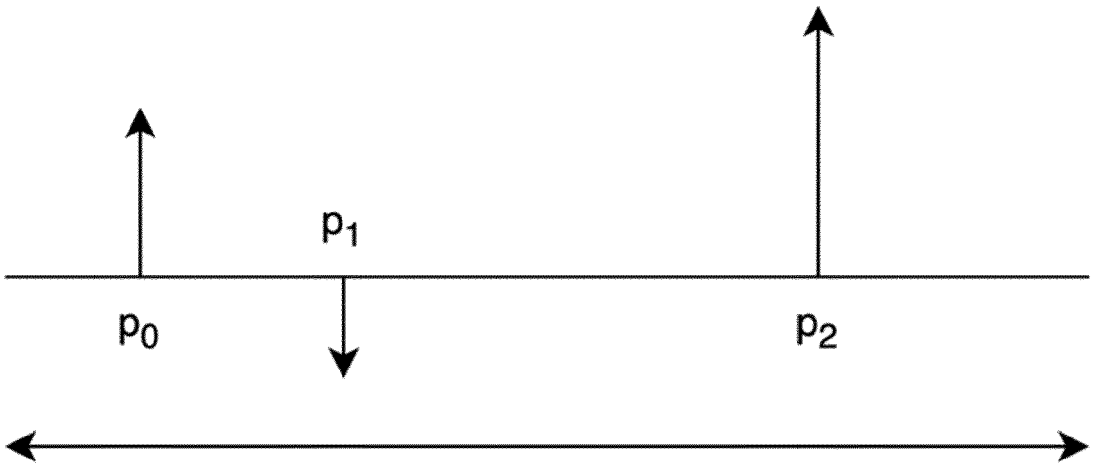


FIG. 5

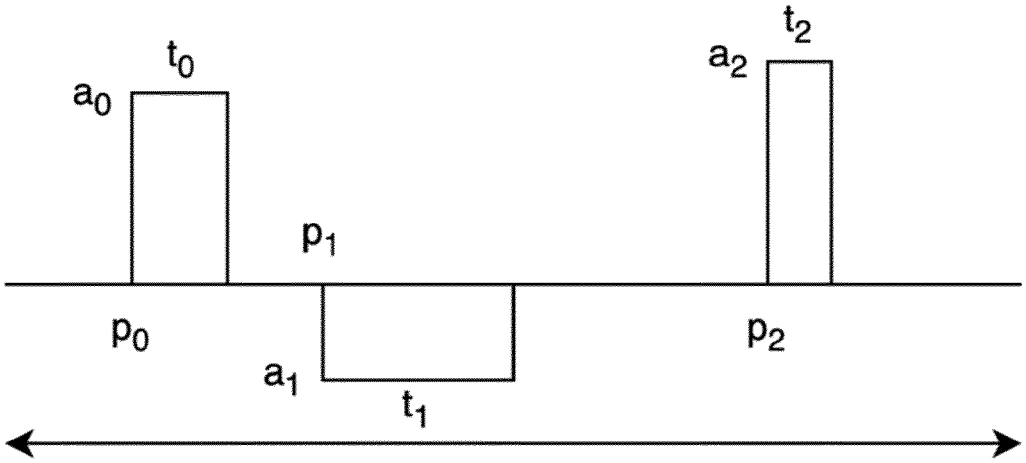
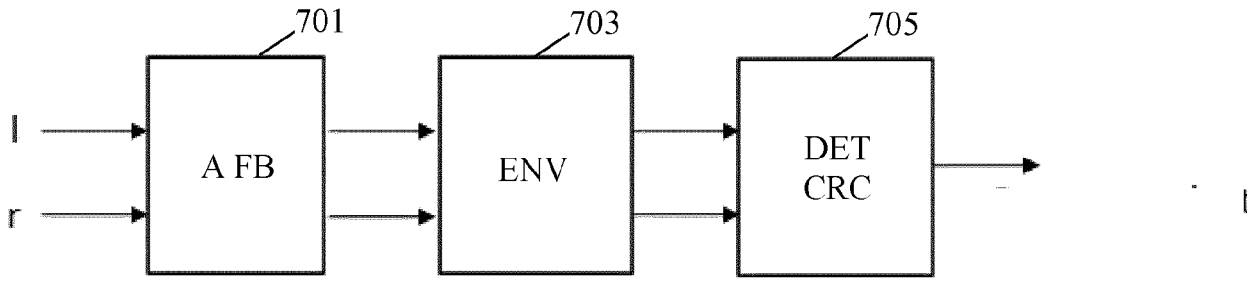


FIG. 6



207

FIG. 7

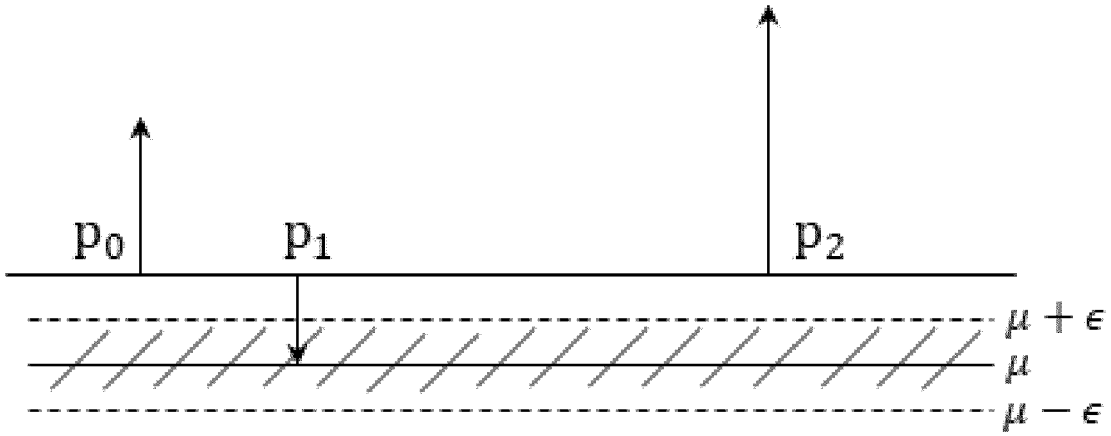


FIG. 8

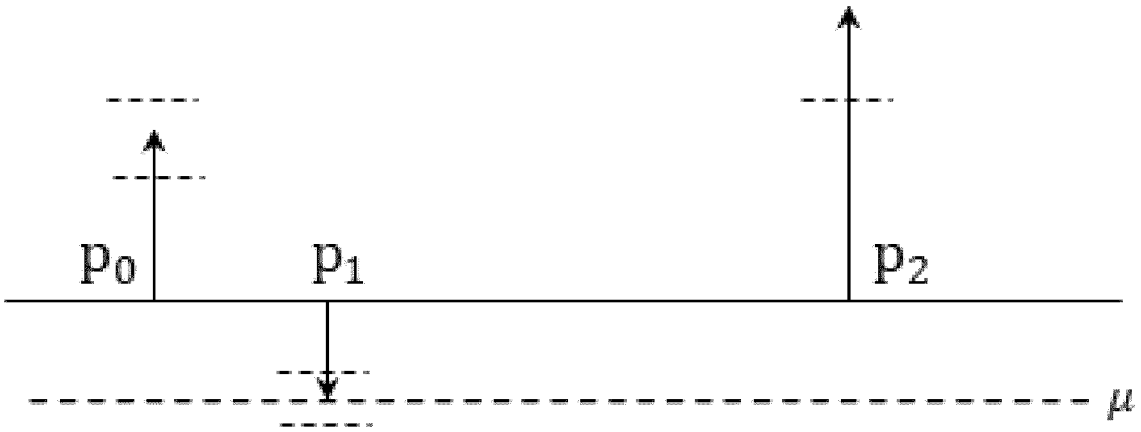


FIG. 9

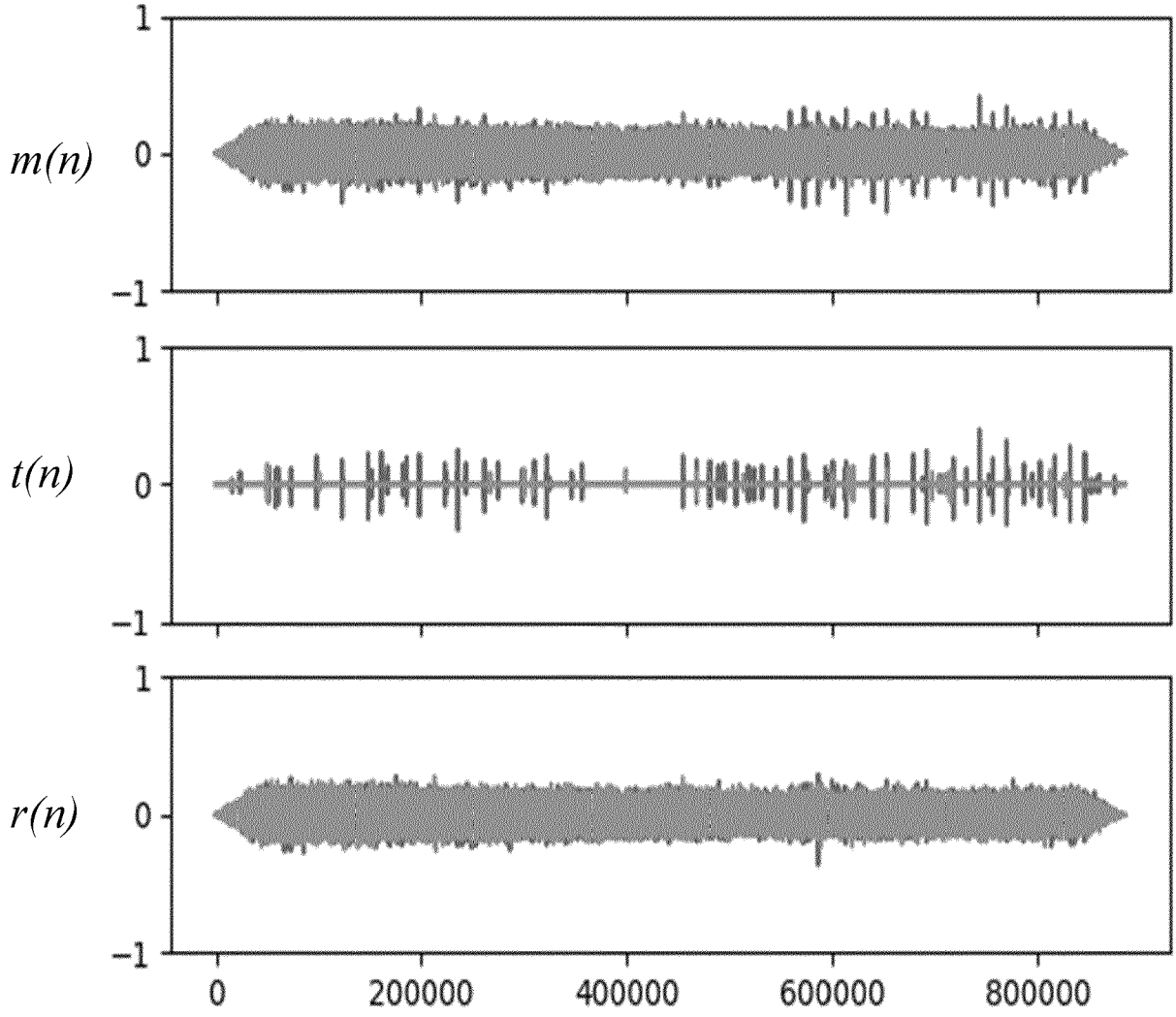


FIG. 10

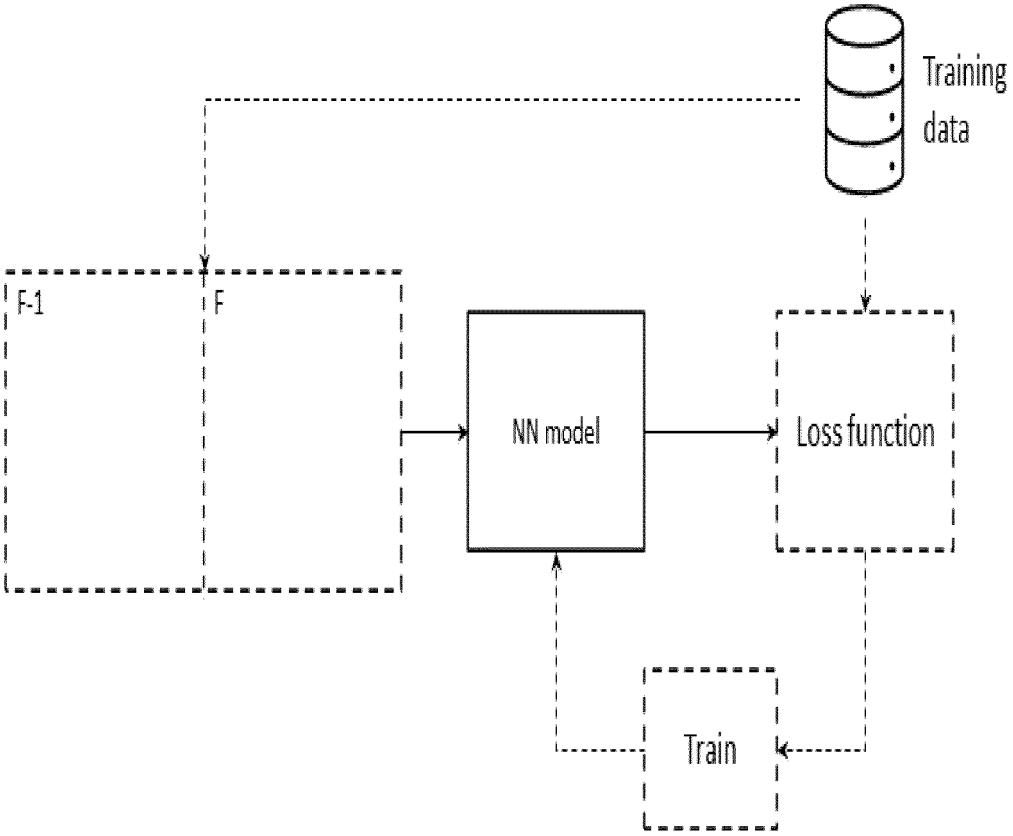


FIG. 11

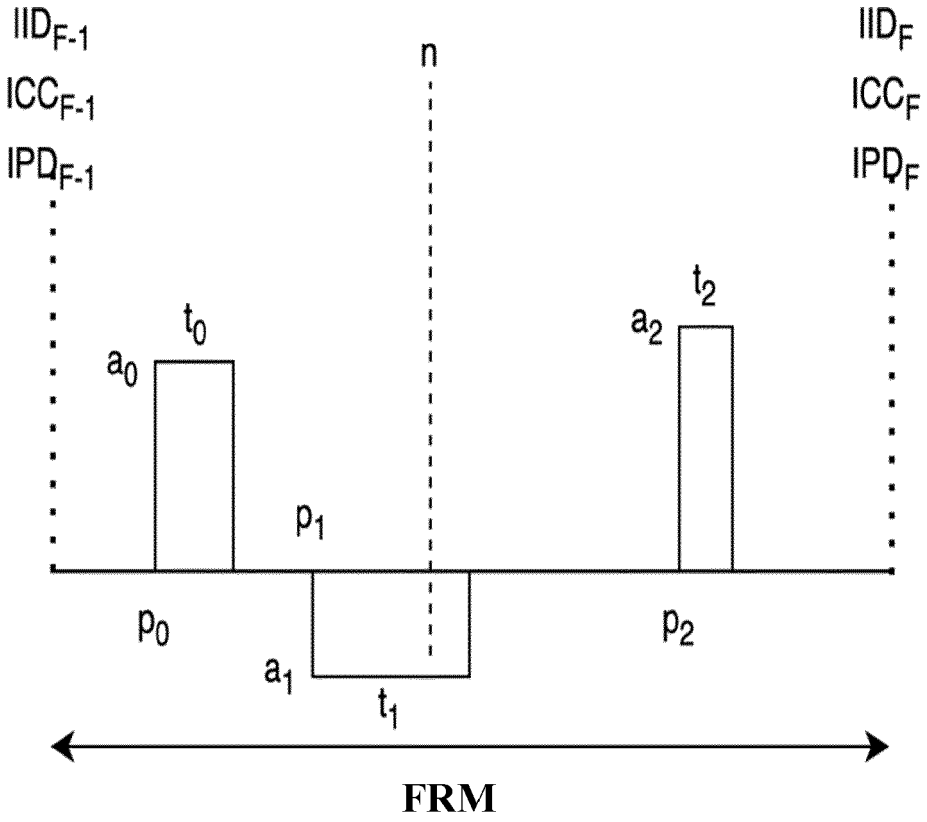


FIG. 12

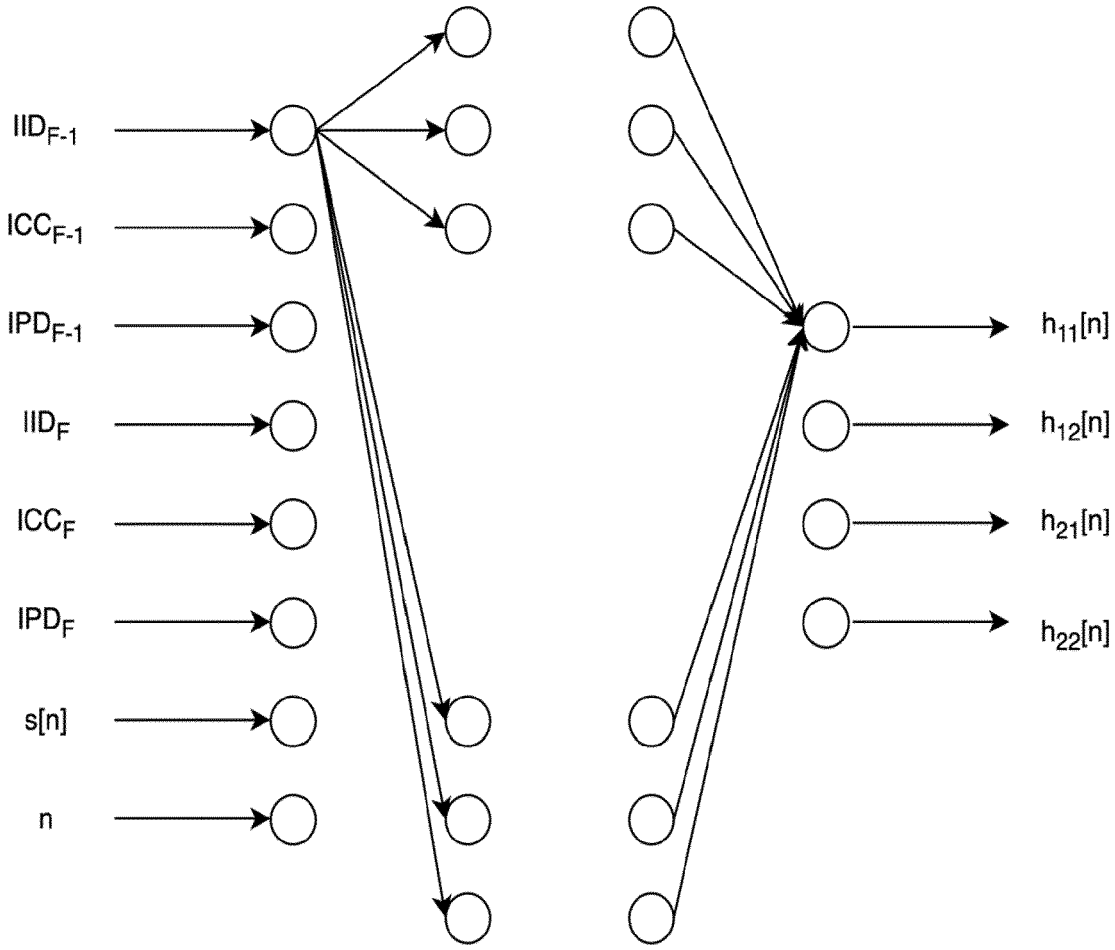


FIG. 13

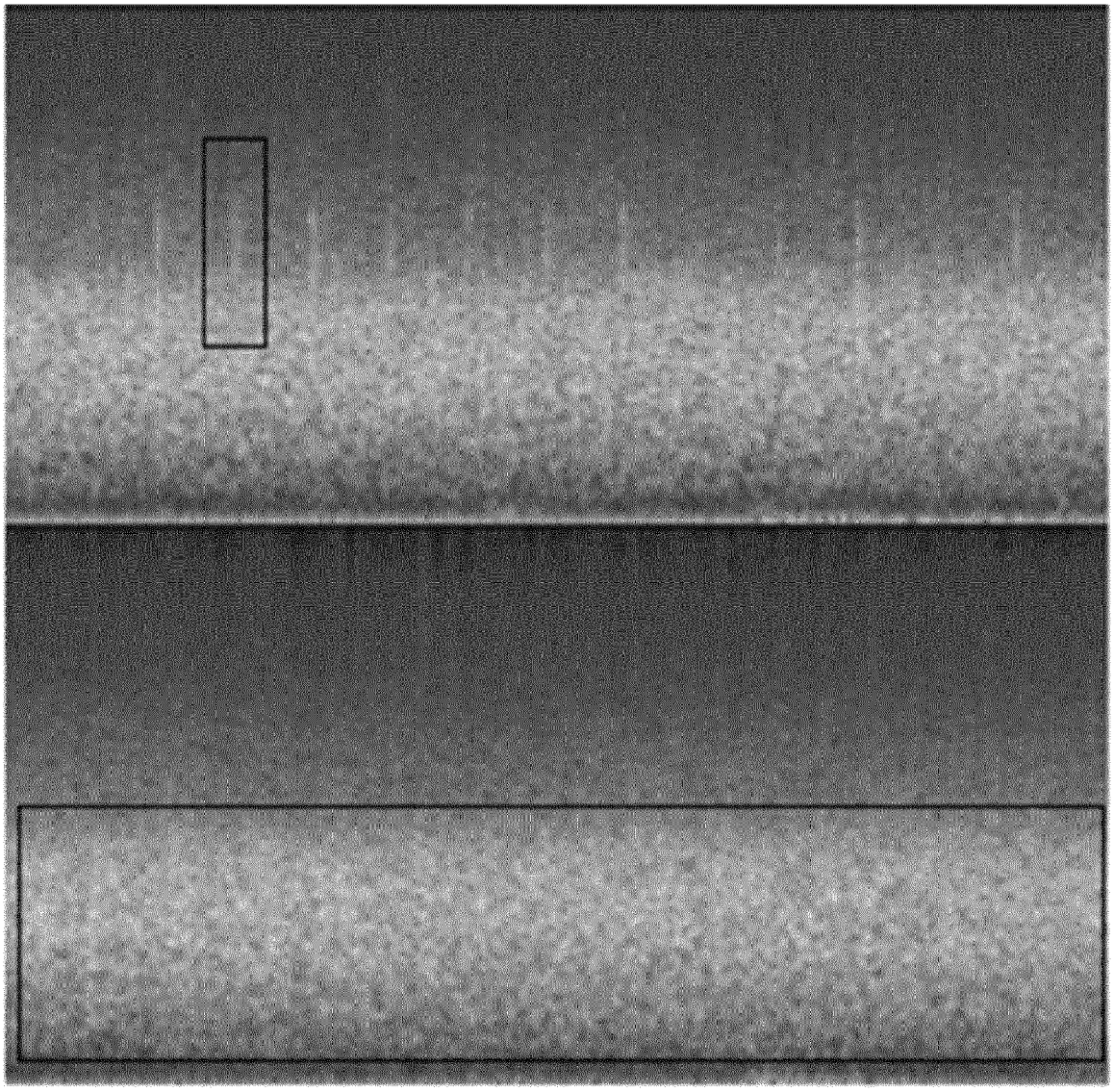


FIG. 14

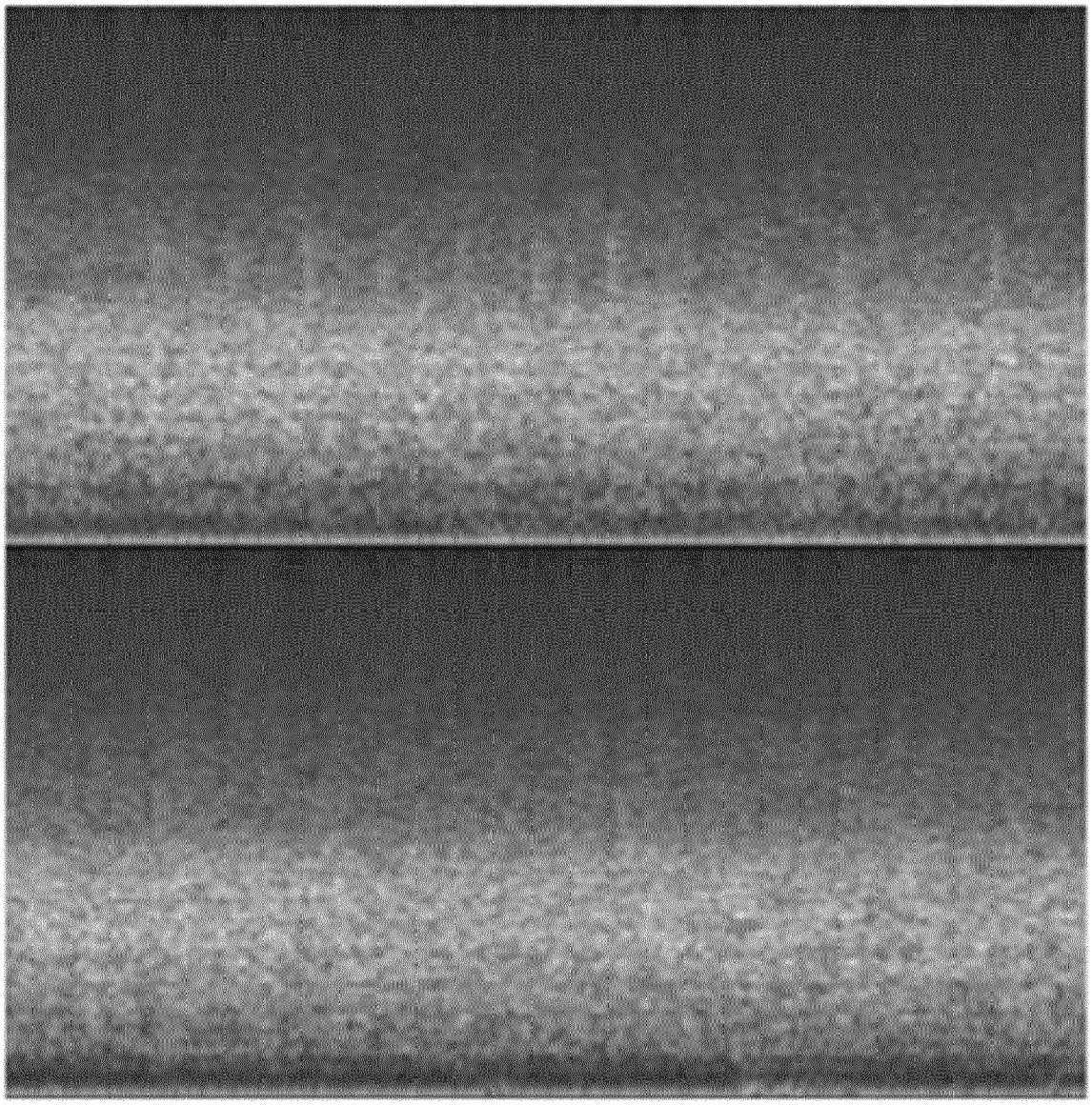


FIG. 15

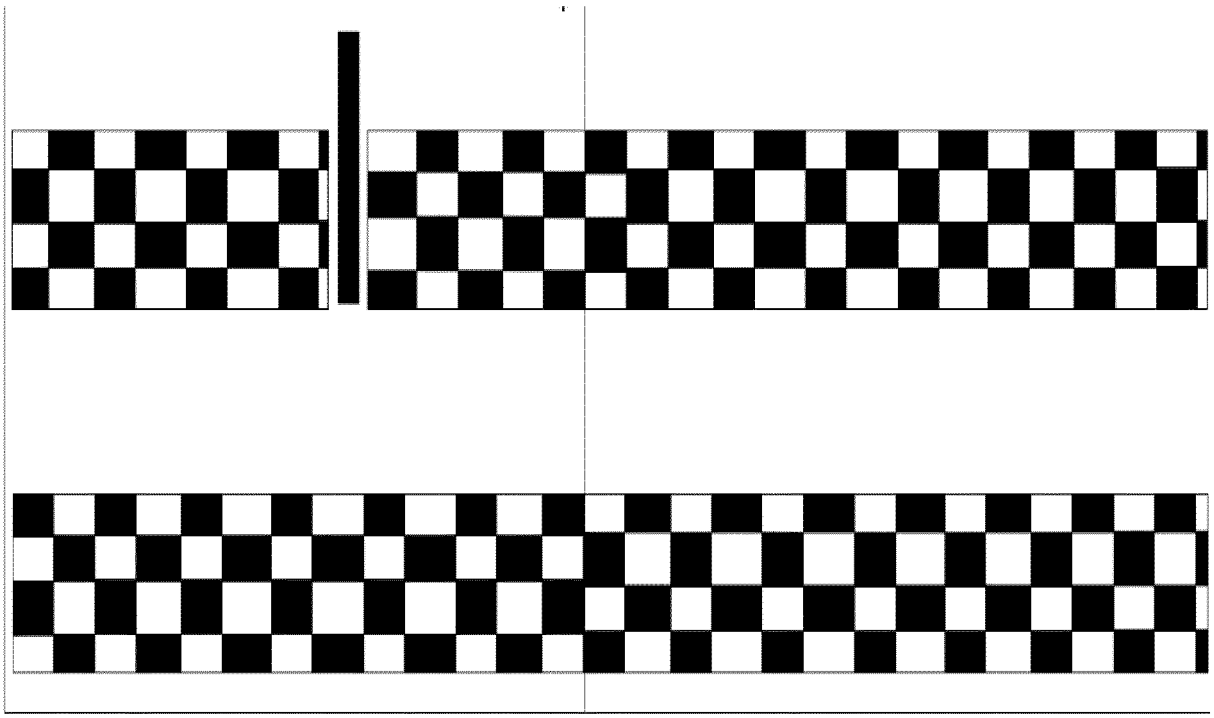


FIG. 16

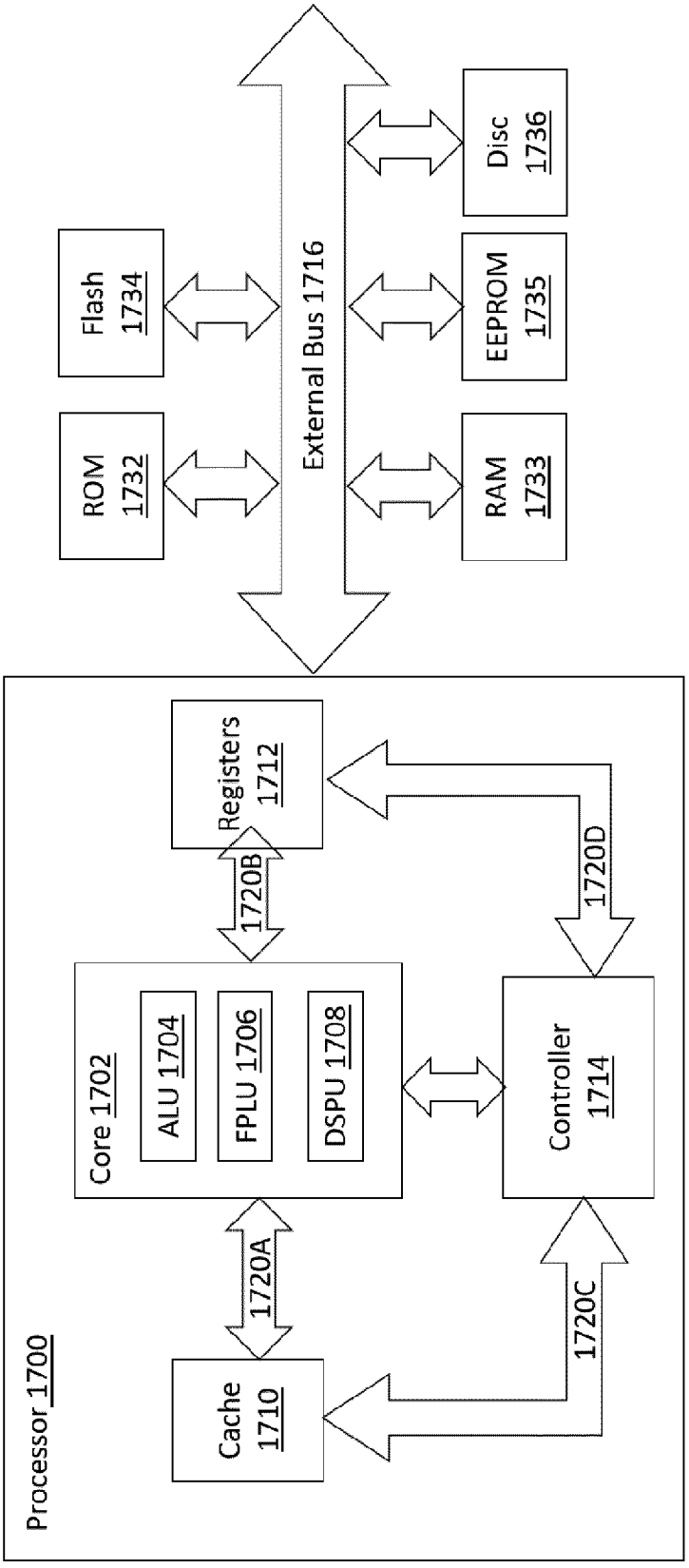

Processor 1700
Cache 1710
Core 1702
ALU 1704
FPLU 1706
DSPU 1708
Registers 1712
Controller 1714
1720A
1720B
1720C
1720D
External Bus 1716
ROM 1732
Flash 1734
RAM 1733
EEPROM 1735
Disc 1736

FIG. 17

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 23 19 9816

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 [0050] - [0056]; figure 2 * * paragraphs [0103] - [0106], [0133] - [0138] * * claim 4 *	1-15	INV. G10L19/008 ADD. G10L19/025
A	JOAN SERR\`A ET AL: "Mono-to-stereo through parametric stereo generation", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 26 June 2023 (2023-06-26), XP091547784, * paragraphs [02.1], [02.2], [03.3] *	1-15	
A	PARK SU YEON ET AL: "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks", 2016 INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY CONVERGENCE (ICTC), IEEE, 19 October 2016 (2016-10-19), pages 377-380, XP033015750, DOI: 10.1109/ICTC.2016.7763500 [retrieved on 2016-11-30] * figure 2 * * abstract *	1-15	TECHNICAL FIELDS SEARCHED (IPC) G10L
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, * paragraphs [04.1], [04.4]; figure 2 *	1-15	

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 9816

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

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]**
- **XAVIER GLOROT** ; **ANTOINE BORDES** ; **YOSHUA BENGIO**. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. *PMLR*, 2011, vol. 15, 315-323 **[0078]**
- **OORD** ; **AARON VAN DEN** ; **SANDER DIELEMAN** ; **HEIGA ZEN** ; **KAREN SIMONYAN** ; **ORIOL VINYALS** ; **ALEX GRAVES** ; **NAL KALCHBRENNER** ; **ANDREW SENIOR** ; **KORAY KAVUKCUOGLU**. Wavenet: A generative model for raw audio. *arXiv preprint arXiv: 1609.03499*, 2016 **[0082]**
- 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 **[0121]**
- **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 **[0128]**
- **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 **[0130]**