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

(11) **EP 4 471 767 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
04.12.2024 Bulletin 2024/49

(21) Application number: **23176014.1**

(22) Date of filing: **30.05.2023**

(51) International Patent Classification (IPC):
G10L 21/0272 *(2013.01)*

(52) Cooperative Patent Classification (CPC):
G10L 21/0272; G10L 2021/02166

(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:
- **JANSE, Cornelis Pieter**
  Eindhoven (NL)
- **BLOEMENDAL, Brian Brand Antonius Johannes**
  Eindhoven (NL)
- **JANSSEN, Rik Jozef Martinus**
  Eindhoven (NL)

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

(54) **AN AUDIO APPARATUS AND METHOD OF OPERATION THEREFOR**

(57) An audio apparatus comprises a receiver (201) arranged to receive a set of input audio signals. An audio beamformer (101) performs beamforming by combining outputs of filters. A feedback circuit (103) comprising a matching set of filters with complex conjugate frequency responses and a beamform adapter (105) adapts the filters in response to a comparison of the input audio signals and the feedback audio signals. An adaptive coefficient processor (207) determines decorrelation coefficients for a set of spatial decorrelation filters generating decorrelated output signals from the input audio signals where the decorrelation coefficients are adapted based on the input signals. A set of spatial filters (205, 301, 401) are arranged to apply a spatial filtering to the input audio signals or the feedback signals where the set of spatial filters have coefficients that are determined from the decorrelation coefficients.

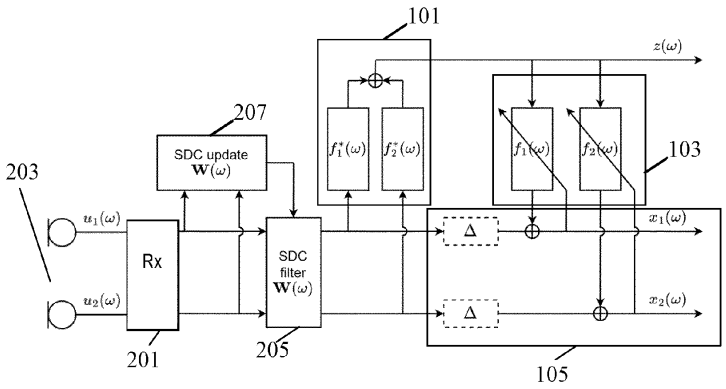

FIG. 2

EP 4 471 767 A1

Description

FIELD OF THE INVENTION

[0001] The invention relates to an apparatus and a method for generating an audio output signals, and in particular, but not exclusively, to extract audio from a wanted audio source, such as a desired speaker.

BACKGROUND OF THE INVENTION

[0002] Capturing audio, and in particularly speech, has become increasingly important in the last decades. For example, capturing speech or other audio has become increasingly important for a variety of applications including telecommunication, teleconferencing, gaming, audio user interfaces, etc. However, a problem in many scenarios and applications is that the desired audio source is typically not the only audio source in the environment. Rather, in typical audio environments there are many other audio/noise sources which are being captured by the microphone. Audio processing is often used to improve the capture of audio, and in particular to post-process the captured audio time interval improves the resulting audio signals.

[0003] In many embodiments, audio may be represented by a plurality of different audio signals that reflect the same audio scene or environment. In particular, in many practical applications, audio is captured by a plurality of microphones at different positions. For example, a linear array of a plurality of microphones is often used to capture audio in an environment, such as in a room. The use of multiple microphones allows spatial information of the audio to be captured. Many different applications may exploit such spatial information allowing improved and/or new services.

[0004] One frequently used approach is to try to separate audio sources by applying beamforming to form beams directed towards the direction of arrival of audio from specific audio sources. However, although this may provide advantageous performance in many scenarios, it is not optimal in all cases. For example, it may not provide optimal source separation in some cases, and indeed in some applications such a spatial beamforming may not provide audio properties that are ideal for further processing to achieve a given effect.

[0005] Thus, whereas spatial audio source separation, and specifically such separation based on audio beamforming, is highly advantageous in many scenarios and applications, there is a desire to improve the performance and operation of such approaches. However, there is typically also a desire for low complexity and/or resource usage (e.g. computational resources) and often these preferences conflict with each other.

[0006] Hence, an improved approach would be advantageous, and in particular an approach allowing reduced complexity, increased flexibility, facilitated implementation, reduced cost, improved audio capture, improved spatial perception/differentiation of audio sources, improved audio source separation, improved audio/speech application support, reduced dependency on known or static acoustic properties, improved flexibility and customization to different audio environments and scenarios, improved audio beamforming, an improved trade-off between performance and complexity/ resource usage, and/or improved performance would be advantageous.

SUMMARY OF THE INVENTION

[0007] Accordingly, the Invention seeks to preferably mitigate, alleviate or eliminate one or more of the above mentioned disadvantages singly or in any combination.

[0008] According to an aspect of the invention there is provided an audio apparatus comprising: a receiver arranged to receive a first set of audio signals, the first set of audio signals comprising audio signals capturing audio of a scene from different positions; an audio beamformer comprising: a first set of filters arranged to filter the first set of audio signals, and a combiner arranged to combine outputs of the first set of filters to generate a beamform output audio signal; a feedback circuit comprising a second set of filters arranged to generate a second set of audio signals from a filtering of the beamform output audio signal, each filter of the second set of filters having a frequency response being a complex conjugate of a filter of the first set of filters; a beamform adapter arranged to adapt the first set of filters and the second set of filters in response to a comparison of the first set of audio signals and the second set of audio signals; an adaptive coefficient processor arranged to determine decorrelation coefficients for a set of spatial decorrelation filters generating decorrelated output signals from the first set of audio signals, the adaptive coefficient processor being arranged to adapt the decorrelation coefficients in response to an update value determined from the first set of audio signals; a first set of spatial filters arranged to apply a first spatial filtering to at least one of the first set of audio signals and the second set of audio signals, the first set of spatial filters having coefficients determined from the decorrelation coefficients.

[0009] The approach may provide improved operation and/or performance in many embodiments. It may in particular allow improved beamforming to focus on a specific audio source in the scene. It may allow improved extraction/separation of audio from a specific source in the presence of other audio and noise sources in the scene.

[0010] The approach may allow the operation of the audio apparatus to effectively adapt to the current conditions, and in

particular the acoustic and spatial properties of the audio sources and the scene. It may provide reduced sensitivity to noise and unwanted audio sources in the scene and captured audio.

[0011] The approach may allow efficient operation and low complexity in many embodiments. The different adaptations may synergistically interwork to provide improved separation of a desired audio source from other captured audio of the scene. Further, the use of multiple adaptations may provide an improved operation while allowing lower complexity adaptation algorithms and criteria to be used.

[0012] Each filter of the first set of filters is linked with a filter of the second set of filters with this filter having a frequency response being a complex conjugate of the frequency response of the filter of the first set of filters.

[0013] The first and second set of filters may have an equal number of linked paired filters having complex conjugate frequency responses. For each audio signal in the first set of audio signals there may be one filter of the first set of filters, one linked/paired filter in the second set of filters (with a complex conjugate frequency response), and one audio signal in the second set of audio signals. The comparison may be a comparison between linked/paired audio signals of the first set of audio signals and the second set of audio signals. The adaptation of a given filter of the first set of filters and a given linked filter of the second set of filters may be in response/dependence on (possibly only) a comparison of a signal of the first set of audio signals filtered by the given filter of the first set of filters and a signal of the second set of audio signals generated by filtering of the beamform audio signal by the given linked filter of the second set of filters. The adaptation may be such that the difference is reduced.

[0014] Each of the second set of audio signals may be an estimate of the contribution to the linked audio signal of the first set of audio signals from the audio captured by the beamform output audio signal, and thus typically may be an estimate of the contribution from a wanted/desired source.

[0015] The set of spatial decorrelation filters may include one spatial filter for each signals of the first set of signals. Each filter of the set of spatial decorrelation filters may generate a filtered/modified version of one audio signal of the first set of audio signals. The set of spatial decorrelation filters may together generate a modified first set of audio signals with a higher degree of decorrelation. The spatial decorrelation filters may perform filtering over the first set of audio signals. The output of a decorrelation filter (the corresponding modified audio signal of the first set of audio signals) may be dependent on a plurality of the (unmodified) first set of audio signals. The spatial decorrelation filters may specifically be frequency domain filters and the decorrelation coefficients may be frequency domain coefficients. For a given spatial decorrelation filter, the output value for a given frequency bin at a given time may be a weighted combination of a plurality of values of the (unmodified) first set of audio signals for the given frequency bin at the given time. The decorrelated output signals may have a reduced normalized cross channel signal correlation with respect to the first set of input signals to the set of spatial decorrelation filters.

[0016] The set of spatial filters may include one spatial filter for each signal of the first set of signals/ second set of signals. Each filter of the set of spatial filters may generate a filtered/modified version of one audio signal of the first or second set of audio signals. The set of spatial filters may together generate a modified first or second set of audio signals. The spatial filters may perform filtering over the first set of audio signals. The output of a spatial filter (the corresponding modified audio signal of the first or second set of audio signals) may be dependent on a plurality of the (unmodified) first or second set of audio signals. The spatial filters may specifically be frequency domain filters and the coefficients may be frequency domain coefficients. For a given spatial filter, the output value for a given frequency bin at a given time may be a weighted combination of a plurality of values of the (unmodified) first or second set of audio signals for the given frequency bin at the given time.

[0017] In accordance with an optional feature of the invention, the audio apparatus further comprises an audio detector arranged to determine a set of active time intervals during which an audio source is active and a set of inactive time intervals during which the audio source is not active; and wherein at least one of the adaption of the first set of filters and the second set of filters and the adaptation the decorrelation coefficients is different for the set of active time intervals and the set of inactive time intervals.

[0018] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation which typically may lead to improved extraction of a desired audio source.

[0019] The active time intervals may be desired/wanted/speech audio source active time intervals and the inactive time intervals may be desired/wanted/speech audio source inactive time intervals.

[0020] In accordance with an optional feature of the invention, the beamform adapter is arranged to adapt the first set of filters and the second set of filters with a higher rate of adaptation during the set of active time intervals than during the set of inactive time intervals.

[0021] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation which typically may lead to improved extraction of a desired audio source.

[0022] In accordance with an optional feature of the invention, the beamform adapter is arranged to adapt the first set of filters and the second set of filters during only one set of time intervals of the set of active time intervals and the set of inactive time intervals.

[0023] This may provide improved performance and/or operation in many embodiments. It may typically provide

improved adaptation which typically may lead to improved extraction of a desired audio source.

[0024] In some embodiments, the beamform adapter is arranged to adapt the first set of filters and the second set of filters during the set of active time intervals but not during the set of inactive time intervals.

[0025] In accordance with an optional feature of the invention, the adaptive coefficient processor is arranged to adapt the decorrelation coefficients with a higher rate of adaptation during the set of inactive time intervals than during the set of active time intervals.

[0026] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation which typically may lead to improved extraction of a desired audio source.

[0027] In accordance with an optional feature of the invention, the adaptive coefficient processor is arranged to adapt the decorrelation coefficients during only one set of time intervals of the set of inactive time intervals and the set of active time intervals.

[0028] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation which typically may lead to improved extraction of a desired audio source.

[0029] In some embodiments, the adaptive coefficient processor is arranged to adapt the decorrelation coefficients during the set of inactive time intervals but not during the set of active time intervals.

[0030] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to filter the first set of audio signals.

[0031] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0032] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to have coefficients set to the decorrelation coefficients determined for the set of spatial decorrelation filters.

[0033] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0034] In accordance with an optional feature of the invention, the first set of filters for filtering is arranged to filter the first set of audio signals after filtering by the first set of spatial filters and the beamform adapter is arranged to perform the comparison using the first set of audio signals before filtering by the first set of spatial filters.

[0035] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0036] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to have coefficients matching coefficients of a spatial filter being a cascade of two of the set of decorrelation filters.

[0037] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0038] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to filter the second set of audio signals.

[0039] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0040] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to have coefficients determined in response to a set of inverse spatial decorrelation filters, the set of inverse spatial decorrelation filters being inverse filters of the set of spatial decorrelation filters.

[0041] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0042] In accordance with an optional feature of the invention, the first set of spatial filters is arranged to have coefficients matching coefficients of a spatial filter being a cascade of two sets of spatial inverse filters each of which comprises inverse filters of the set of spatial decorrelation filters.

[0043] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0044] In accordance with an optional feature of the invention, adaptive coefficient processor is arranged to determine the set of spatial decorrelation filters to generate a set of output audio signals, each output audio signal of the set of the set of output signals being linked with one input audio signal of the first set of audio signals, by performing the steps of: segmenting the first set of audio signals into time segments, and for at least some time segments performing the steps of: generating a frequency bin representation of the first set of audio signals, each frequency bin of the frequency bin representation of the first set of audio signals comprising a frequency bin value for each of the audio signals of the first set of audio signals; generating a frequency bin representation of a set of output signals, each frequency bin of the frequency bin representation of a set of output signals comprising a frequency bin value for each of the output signals, the frequency bin value for a given output signal of the set of output signals for a given frequency bin being generated as a weighted combination of frequency bin values of the first set of audio signals for the given frequency bin, the weighted combination having the decorrelation coefficients as weights; updating a first weight for a contribution to a first frequency bin value of a first frequency bin for a first output signal linked with a first input audio signal from a second frequency bin value of the first

frequency bin for a second input audio signal linked to a second output signal in response to a correlation measure between a first previous frequency bin value of the first output signal for the first frequency bin and a second previous frequency bin value of the second output signal for the first frequency bin.

[0045] This may provide improved performance and/or operation in many embodiments. In many embodiments and scenarios, this may provide particularly attractive performance and/or implementation.

[0046] This may provide an advantageous generation of output audio signals with typically increased decorrelation in comparison to the input signals. The approach may provide an efficient adaptation of the operation resulting in improved decorrelation in many embodiments. The adaptation may typically be implemented with low complexity and/or resource usage. The approach may specifically apply a local adaptation of individual weights yet achieve an efficient global adaptation.

[0047] The generation of the set of output signals may be adapted to provide increased decorrelation relative to the input signals which in many embodiments and for many applications may provide improved audio processing and in particular beamforming.

[0048] The first and second output audio signals may typically be different output audio signals.

[0049] In accordance with an optional feature of the invention, the adaptive coefficient processor is arranged to update the first weight in response to a product of a first value and a second value, the first value being one of the first previous frequency bin value and the second previous frequency bin value and the second value being a complex conjugate of the other of the first previous frequency bin value and the second previous frequency bin value.

[0050] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0051] In some embodiments, the audio apparatus may be arranged to update a second weight being for a contribution to the first frequency bin value from a third frequency bin value being a frequency bin value of the first frequency bin for the first input audio signal in response to a magnitude of the first previous frequency bin value.

[0052] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios. It may in particular provide an improved adaptation of the generated output signals. In many embodiments, the updating of the weight reflecting the contribution to an output signal from the linked input signal may be dependent on the signal magnitude/amplitude of that linked input signal. For example, the updating may seek to compensate the weight for the level of the input signal to generate a normalized output signal.

[0053] The approach may allow a normalization/ signal compensation/level compensation to provide e.g., a desired output level.

[0054] In some embodiments, the audio apparatus may be arranged to set a weight for a contribution to the first frequency bin value from a third frequency bin value being a frequency bin value of the first frequency bin for the first input audio signal to a predetermined value.

[0055] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios. It may in many embodiments provide improved adaptation while ensuring convergence of the adaptation towards a non-zero signal level. It may interwork very efficiently with the adaptation of weights that are not for linked signal pairs.

[0056] In many embodiments, the adapter may be arranged to keep the weight constant and with no adaptation or updating of the weight.

[0057] In some embodiments, the audio apparatus may be arranged to constrain a weight for a contribution to the first frequency bin value from a third frequency bin value being a frequency bin value of the first frequency bin for the first input audio signal to be a real value.

[0058] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0059] The weights between linked input/output signals may advantageously be determined as/ constrained to be a real valued weight. This may lead to improved performance and adaptation ensuring convergence on a non-zero level solution.

[0060] In some embodiments, the audio apparatus may be arranged to set a second weight being a weight for a contribution to a fourth frequency bin value of the first frequency bin for the second output audio signal from the first input audio signal to be a complex conjugate of the first weight.

[0061] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0062] The two weights for two pairs of input/output signals may be complex conjugates of each other in many embodiments.

[0063] In some embodiments, weights of the weighted combination for other input audio signals than the first input audio signal are complex valued weights.

[0064] This may provide improved performance and/or operation in many embodiments. The use of complex values for weights for non-linked input signals provide an improved frequency domain operation.

[0065] In some embodiments, the audio apparatus may be arranged to determine output bin values for the given frequency bin ω from:

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega)$$

where $\mathbf{y}(\omega)$ is a vector comprising the frequency bin values for the output audio signals for the given frequency bin ω; $\mathbf{x}(\omega)$ is a vector comprising the frequency bin values for the input audio signals for the given frequency bin ω; and $\mathbf{W}(\omega)$ is a matrix having rows comprising weights of a weighted combination for the output audio signals.

[0066] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0067] The matrix $\mathbf{W}(\omega)$ may advantageously be Hermitian. In many embodiments, the diagonal of the matrix $\mathbf{W}(\omega)$ may be constrained to be real values, may be set to a predetermined value(s), and/or may not be updated/adapted but may be maintained as a fixed value. The weights/coefficients outside the diagonal may generally be complex values.

[0068] In some embodiments, the audio apparatus may be arranged to adapt weights w_{ij} of the matrix $\mathbf{W}(\omega)$ according to:

$$w_{ij}(k+1,\omega) = w_{ij}(k,\omega) - \eta(k,\omega)\left[y_i(k,\omega)\, y_j^*(k,\omega)\right]$$

where i is a row index of the matrix $\mathbf{W}(\omega)$, j is a column index of the matrix $\mathbf{W}(\omega)$, k is a time segment index, ω represents the frequency bin, and $\eta(k,\omega)$ is a scaling parameter for adapting an adaptation speed.

[0069] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0070] In some embodiments, the audio apparatus may be arranged to compensate the correlation value for a signal level of the first frequency bin.

[0071] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios. It may allow a compensation of the update rate for signal variations.

[0072] In some embodiments, the audio apparatus may be arranged to initialize the weights for the weighted combination to comprise at least one zero value weight and one non-zero value weight.

[0073] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios. It may allow a more efficient and/or quicker adaptation and convergence towards advantageous decorrelation. In many embodiments, the matrix $\mathbf{W}(\omega)$ may be initialized with zero values for weights or coefficients for nonlinked signals and fixed non-zero real values for linked signals. Typically, the weights may be set to e.g., 1 for weights on the diagonal and all other weights may initially be set to zero.

[0074] In some embodiments, the weighted combination comprises applying a time domain windowing to a frequency representation of weights formed by weights for the first input audio signal and the second input audio signal for different frequency bins.

[0075] This may provide improved performance and/or operation in many embodiments. It may typically provide improved adaptation leading to increased decorrelation of the output audio signals in many scenarios.

[0076] Applying the time domain windowing to the frequency representation of weights may comprise: converting the frequency representation of weights to a time domain representation of weights; applying a window to the time domain representation to generate a modified time domain representation; and converting the modified time domain representation to the frequency domain.

[0077] According to an aspect of the invention, there is provided a method of generating an output audio signal, the method comprising: receiving a first set of audio signals, the first set of audio signals comprising audio signals capturing audio of a scene from different positions; an audio beamformer generating an output signal by: a first set of filters filtering the first set of audio signals, and a combiner combining outputs of the first set of filters to generate an output audio signal; a second set of filters generating a second set of audio signals from a filtering of the output audio signal, each filter of the second set of filters having a frequency response being a complex conjugate of a filter of the first set of filters; adapting the first set of filters and the second set of filters in response to a comparison of the first set of audio signals and the second set of audio signals; determining decorrelation coefficients for a set of spatial decorrelation filters generating decorrelated output signals from the first set of audio signals including adapting the decorrelation coefficients in response to an update value determined from the first set of audio signals; a first set of spatial filters applying a first spatial filtering to at least one of the first set of audio signals and the second set of audio signals, the first set of spatial filters having coefficients determined from the decorrelation coefficients.

[0078] 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

[0079] Embodiments of the invention will be described, by way of example only, with reference to the drawings, in which

FIG. 1 illustrates an example of a beamformer;
FIG. 2 illustrates an example of an audio apparatus in accordance with some embodiments of the invention;
FIG. 3 illustrates an example of an audio apparatus in accordance with some embodiments of the invention;
FIG. 4 illustrates an example of an audio apparatus in accordance with some embodiments of the invention;
FIG. 5 illustrates an example of an audio apparatus in accordance with some embodiments of the invention; and
FIG. 6 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

[0080] The following description focuses on embodiments of the invention applicable to audio capturing such as e.g., speech capturing for a teleconferencing apparatus. However, it will be appreciated that the approach is applicable to many other audio signals, audio processing systems, and scenarios for capturing and/or processing audio.

[0081] FIG. 1 illustrates an example of an adaptive audio beamform arrangement. For clarity, the examples of the description is focused on an example where the audio apparatus is arranged to process two audio input signals, but it will be appreciated that the example readily extends to more input signals being processed.

[0082] The arrangement is arranged to generate a first set of input signals which comprises audio signals that capture audio of a scene from different positions. The first set of audio signals may specifically be audio signals generated from a set of microphone signals from a set of microphones being at different positions in an audio environment, such as a microphone array, and in many cases specifically a linear set of microphones.

[0083] The arrangement includes a beamformer 101 which is arranged to receive a set of a plurality of audio signals from which a single signal is generated. The beamformer 101 seeks to combine the input signals such that the contributions from a given audio source are combined constructively. The beamformer 101 specifically combines the signals to perform a beamforming for the set of spatial audio signals capturing audio in an environment, such as for a set of signals from a microphone array. The beamformer 101 is arranged to generate a beamform output audio signal from the first set of audio signals.

[0084] The beamformer 101 specifically comprises a first set of filters $f_1^*(\omega), f_2^*(\omega)$ which filters the first set of audio signals. Each filter is arranged to filter one of the audio signals. Further, the filters are adaptive filters that are dynamically adapted to achieve an adaptive beamforming. The first set of filters will henceforth also be referred as beamformer filters.

[0085] The audio arrangement further comprises a feedback circuit 103 which comprises a second set of filters which generate a second set of audio signals from a filtering of the beamform output audio signal. The second set of filters will also be referred to as feedback filters. The input signals to the beamformer may also be referred to as beamformer input signals and the signals generated by the feedback circuit may also be referred to as feedback signals.

[0086] The first and second set of filters are specifically matched/linked such that for each filter of the first set (i.e. for each beamformer filter) there is a filter of the second set (i.e. a feedback filter) which has a frequency response that is the complex conjugate of the corresponding beamformer filter. Thus, the filters of the beamformer 101 and the feedback circuit 103 are adapted together such that the filter coefficients provide a complex conjugate frequency response (equivalent/-corresponding to time reversed filter impulse responses).

[0087] Thus, from the beamform output audio signal, the feedback circuit 103 generates a set of feedback signals with each feedback signal being generated from the beamform output audio signal by a time reversed/complex conjugate filtering by matching filters.

[0088] The audio arrangement further comprises a beamform adapter 105 arranged to adapt the first set of filters and the second set of filters, i.e. the beamformer filters and the feedback filters. The beamform adapter 105 is arranged to perform the adaptation based on a comparison of the first and the second set of audio signals, and specifically based on the comparison of the beamformer signals and the feedback signals. Specifically, for each matching/linked beamformer signal and feedback signal (signals for which the beamformer filter and the feedback filter have complex conjugate frequency responses), a difference measure may be determined, and the filters of the matching filters may be adapted to reduce/minimize the difference measure.

[0089] The arrangement of the beamformer 101, feedback circuit 103, and beamform adapter 105 may interact to provide a highly efficient adaptive beamforming which for a set of spatial audio signals may adapt the filters to form a beam towards an audio source with the beamform output audio signal providing the audio captured in the formed beam. The adaption of the filters to minimize the difference measures allows the beamformer arrangement to detect and track a given audio source. Specifically, in the ideal case, each of the beamformer filters concentrate all of the energy from a given audio

source that is captured in one of the beamformer signals into a single signal value. In the ideal case, this is achieved by the beamform filter having an impulse response that is the time inverse of the acoustic impulse response from the audio sources to the microphone capturing the signal. In the ideal case, this is achieved for all the signals/filters resulting in all the captured energy from the specific audio source being combined into a single value by a summation of the outputs of the beamformer filters thereby generating a sample value for the beamform output audio signal that maximizes the energy captured from the audio source.

[0090] Further, as each of the feedback filters are the complex conjugate frequency response of the corresponding beamformer filter, it is in the ideal case the same as the acoustic impulse response from the audio source to the microphone (signal). Accordingly, by filtering the beamform output audio signal by a feedback filter, a feedback signal is generated which reflects the signal that would be captured at the microphone from the audio source as represented by the beamform output audio signal. In the ideal case with only one audio source present and the filters being ideally adapted, the difference between the generated feedback signal and the captured microphone signal is zero. In the presence of other audio sources and/or noise, there may be a difference which is reflected in the difference measure. However, for uncorrelated noise/audio, the difference, and thus typically the difference measure, is typically zero on average and accordingly will tend to not prevent efficient adaptation. Thus, by adapting the filters for a given beamformer signal to minimize the difference measure to the corresponding feedback signal, an adaptation towards the optimum beamforming filters can typically be achieved.

[0091] Such a beamforming approach is described in US 7 146 012. In the following a more detailed description and analysis of the arrangement will be provided. This description will be based on the example of FIG. 1 where the adaptive beamformer arrangement consists of a filter part where the microphone signals are filtered by two filters $f_1^*(\omega)$ and $f_2^*(\omega)$ and the outputs are summed to obtain the output signal $z(\omega)$. In the update part the output signal $z(\omega)$ is fed to two adaptive filters $f_1(\omega)$ and $f_2(\omega)$ that have the respective microphone signals as a reference.

[0092] With this configuration we have an adaptive beamformer that maximizes the output power under the constraint: $|f_1(\omega)|^2 + |f_2(\omega)|^2 = 1$. The constraint is implied by the configuration, specifically because the conjugates of the filters $f_1(\omega)$ and $f_2(\omega)$ in the update part are copied to the filter part. The adaptive filters in the update part are "normal" unconstrained adaptive filters.

[0093] That the constraint is implied by the configuration can be seen if we look at the optimal solution, that is when both residual signals are zero. Suppose we have a speech signal $s(\omega)$ and transfer functions $h_1(\omega)$ and $h_2(\omega)$ from the speech source to the microphones. The microphone signals are then given by: $u_1(\omega) = h_1(\omega)s(\omega)$ and $u_2(\omega) = h_2(\omega)s(\omega)$.

[0094] The output signal $z(\omega)$ then is:

$$z(\omega) = u_1(\omega)f_1^*(\omega) + u_2(\omega)f_2^*(\omega) = \\ = s(\omega)(h_1(\omega)f_1^*(\omega) + h_2(\omega)f_2^*(\omega)).$$

[0095] In the update part we get after convergence two equations, where we omit ω for convenience:

$$(h_1 f_1^* + h_2 f_2^*) f_1 = h_1 \quad (1)$$

$$(h_1 f_1^* + h_2 f_2^*) f_2 = h_2 \quad (2)$$

[0096] From (1) and (2) follows (3):

$$\frac{f_1}{f_2} = \frac{h_1}{h_2} \qquad (3)$$

[0097] Inserting (3) into (2) gives:

$$\left(h_1 \frac{h_1^*}{h_2^*} f_2^* + h_2 f_2^*\right) f_2 = h_2$$

$$|h_1|^2 |f_2|^2 + |h_2|^2 |f_2|^2 = |h_2|^2$$

$$|f_2|^2 = \frac{|h_2|^2}{|h_1|^2 + |h_2|^2}$$

[0098] Similarly we get:

$$|f_1|^2 = \frac{|h_1|^2}{|h_1|^2 + |h_2|^2}$$

fulfilling the constraint $|f_1|^2 + |f_2|^2 = 1$.

[0099] The solution for f_1 is given by:

$$f_1 = \frac{a_1 h_1}{\sqrt{|h_1|^2 + |h_2|^2}}$$

,where a_1 is an allpass, i.e. $|a_1| = 1$, phase undetermined.

[0100] Similarly for f_2:

$$f_2 = \frac{a_2 h_2}{\sqrt{|h_1|^2 + |h_2|^2}}$$

[0101] Since (3) still must hold, $a_1 = a_2 = a$, the so-called common all-pass term.

[0102] More general and using matrix notation, with

$$\boldsymbol{f}(\omega) = [f_1(\omega) \dots \dots . f_{nmics}(\omega)]^T$$

where f,;, (ω) is the beamformer filter corresponding to the m^{th} microphone and

$$\boldsymbol{h}(\omega) = [h_1(\omega) \dots \dots . h_{nmics}(\omega)]^T$$

with $h_m(\omega)$ the transfer function from the source $s(\omega)$ to the m^{th} microphone, we get for the output $z(\omega)$:

$$z(\omega) = s(\omega)\, \boldsymbol{h}^T(\omega)\boldsymbol{f}^*(\omega) \quad (4)$$

[0103] In the update part, after convergence we have the equality:

$$z(\omega)\boldsymbol{f}(\omega) = s(\omega)\boldsymbol{h}(\omega), \quad (5)$$

which leads to (since $z(\omega)$ and $s(\omega)$ are scalars):

$$\boldsymbol{f}(\omega) = \alpha(\omega)\boldsymbol{h}(\omega), \quad (6)$$

where $a(\omega)$ is a complex scaler.

[0104] Substituting (4) into (5), using (6) gives :

$$\left(\boldsymbol{h}^T(\omega)\boldsymbol{h}^*(\omega)\right)\alpha^*(\omega)\alpha(\omega)\boldsymbol{h}(\omega) = \boldsymbol{h}(\omega)$$

$$|\alpha(\omega)|^2 = \frac{1}{\boldsymbol{h}^T(\omega)\boldsymbol{h}^*(\omega)}$$

$$\alpha(\omega) = \frac{\mathrm{a}(\omega)}{\sqrt{\left(\boldsymbol{h}^T(\omega)\boldsymbol{h}^*(\omega)\right)}} \quad (7)$$

with a(ω) the common allpass term.

[0105] Substituting (7) into (6) finally gives the solution after convergence (8):

$$\boldsymbol{f}(\omega) = \mathrm{a}(\omega)\frac{\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{h}(\omega)}} \quad (8)$$

where $(.)^H$ denotes complex conjugate transpose.

[0106] For the constraint we then get:

$$\boldsymbol{f}^H(\omega)\boldsymbol{f}(\omega) = \mathbf{1}$$

[0107] To understand that the solution given by Eq. 8 is also the optimal solution that maximizes the output power, we look at the expected value of the update after convergence, which consists of a correlation between the residual signal ($x_1(\omega)$ and $x_2(\omega)$ in FIG. 1) and the conjugate of the input signal of the adaptive filters:

$$\mathbb{E}\{z^*(\omega)\left(\boldsymbol{u}(\omega) - z(\omega)\boldsymbol{f}(\omega)\right)\} = 0,$$

with

$$\boldsymbol{u}(\omega) = [u_1(\omega) \ldots \ldots u_{nmics}(\omega)]^T$$

where $u_m(\omega)$ is the microphone signal of the m^{th} microphone.

[0108] This is equivalent to:

$$\mathbb{E}\{\mathrm{z}^*(\omega)\boldsymbol{u}(\omega) - |\mathrm{z}(\omega)|^2\,\boldsymbol{f}(\omega)\} = 0$$

$$\mathbb{E}\{\mathrm{z}^*(\omega)\boldsymbol{u}(\omega)\} - \mathbb{E}\{|\mathrm{z}(\omega)|^2\}\boldsymbol{f}(\omega)\} = 0$$

[0109] Using $\boldsymbol{z}(\omega) = \boldsymbol{u}^T(\omega)\boldsymbol{f}^*(\omega)$ and thus $\boldsymbol{z}^*(\omega) = \boldsymbol{u}^H\boldsymbol{f}(\omega)$ we get:

$$\mathbb{E}\{\boldsymbol{u}(\omega)\boldsymbol{u}^H(\omega)\}\boldsymbol{f}(\omega) = \mathbb{E}\{|z(\omega)|^2\}\boldsymbol{f}(\omega)$$

$$\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) = \rho_{zz}^2(\omega)\boldsymbol{f}(\omega)$$

$$\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) = \rho_z^2(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{f}(\omega) = \rho_z^2(\omega)$$

with $\boldsymbol{R}_{uu}(\omega)$ the input covariance matrix. From this we learn that $\boldsymbol{f}(\omega)$ and $\rho_z^2(\omega)$ are an eigen vector and eigen value of the input covariance matrix respectively. It can be shown that a stable solution is obtained when $\rho_z^2(\omega)$ is the largest eigen value. From matrix theory, we know that for a Hermitian matrix the maximum value of the Rayleigh coefficient (omitting ω):

$$\frac{\boldsymbol{f}^H\boldsymbol{R}_{uu}\boldsymbol{f}}{\boldsymbol{f}^H\boldsymbol{f}}$$

is obtained for the eigen vector that belongs to the largest eigen value. This proofs that the beamformer maximizes it output under the constraint $\boldsymbol{f}^H\boldsymbol{f} = 1$

[0110] When the microphone signals only contain speech with $\boldsymbol{u}(\omega) = \boldsymbol{s}(\omega)\boldsymbol{h}(\omega)$, then maximization of $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega)$ corresponds to the maximization of the speech in the output. If we consider the case with $\boldsymbol{u}(\omega) = \boldsymbol{s}(\omega)\boldsymbol{h}(\omega) + \boldsymbol{v}(\omega)$, where we assume that $\boldsymbol{v}(\omega)$ is uncorrelated noise with equal variance in all microphones i.e.

$$\mathbb{E}\{v_p^*(\omega)v_q(\omega)\} = \rho_{vv}^2(\omega)\delta_{pq}$$

[0111] $\boldsymbol{R}_{uu}(\omega)$ can be written as:

$$\boldsymbol{R}_{uu}(\omega) = \boldsymbol{R}_{ss}(\omega) + \rho_{vv}^2(\omega)\boldsymbol{I}(\omega)$$

[0112] Maximization of $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega)$ then corresponds to maximization of (using $\boldsymbol{f}^H(\omega)\boldsymbol{f}(\omega) = 1$):

$$\boldsymbol{f}^H(\omega)\boldsymbol{R}_{\mathrm{uu}}(\omega)\boldsymbol{f}(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{f}(\omega) + \rho_{vv}^2(\omega),$$

which means that maximization corresponds to maximization of the speech and maximizing the Signal-to-Noise ratio.

[0113] Now assume that $\boldsymbol{v}(\omega)$ consists of correlated noise such that:

$$\boldsymbol{R}_{uu}(\omega) = \boldsymbol{R}_{ss}(\omega) + \boldsymbol{R}_{nn}(\omega)$$

with $\boldsymbol{R}_{nn}(\omega)$ being the covariance matrix of the noise with off-diagonal elements that are non-zero.

[0114] Maximization of $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega)$ then leads to the maximization of:

$$\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) = \boldsymbol{f}^H(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{f}(\omega),$$

[0115] This means that maximization of $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega)$ does not necessarily lead to a better Signal-to-Noise ratio since the choice of $\boldsymbol{f}$ not only determines the amount of speech in the output, but also the amount of noise in the output.

[0116] In the following an approach will be described that may provide improved performance in many scenarios. In the approach, a beamforming approach such as that described with reference to FIG. 1 is further enhanced by introduction of an adaptive spatial filtering of the signals. The spatial filtering is performed across signals and is based on decorrelation coefficients/weights that are determined based on the input audio signals to adapt to the current audio properties. As will be described in more detail in the following, the spatial filtering may be placed at different places in the loop structure of the beamforming including as an inverse spatial decorrelation filtering in the feedback loop or as a spatial decorrelation filtering directly on the input signals being used by the beamforming. It has been found that the approach may provide substantially improved operation, such as e.g. substantially improved source separation, in many practical scenarios. In relation to the above analysis, the approach may in particular achieve that the noise contribution $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{f}(\omega)$ may be made independent of the choice of $\boldsymbol{f}$.

[0117] FIG. 2 illustrates an audio apparatus that may provide improved performance. In the example, the audio apparatus comprises the beamforming structure as described with reference to FIG. 1. However, the approach has further been enhanced by this structure operating on signals that have been spatially decorrelated by an adaptive spatial decorrelator.

[0118] The audio apparatus of FIG. 2 comprises a receiver 201 which receives a first set of audio signals and in the specific examples receives a set of microphone signals from a set of microphones capturing the audio scene from different positions. The microphones may for example be arranged in a linear array and relatively close to each other. For example, the maximum distance between capture points for the audio signals may in many embodiments not exceed 1 meter, 50 cm, 25 cm, or even in some cases 10 cm.

[0119] The input audio signals may be received from different sources, including internal or external sources. In the following an embodiment will be described where the receiver 201 is coupled to a plurality of microphones, such as a linear array of microphones, providing a set of input audio signals in the form of microphone signals.

[0120] However, rather than the directly performing the operation of FIG. 1 on the microphone signals, the audio apparatus of FIG. 2 is arranged to apply an adaptive spatial decorrelation operation to the audio signals before the beamforming and adaptation operation. The spatial decorrelation is an adaptive decorrelation which is adapted based on the input audio signals so that the decorrelation may be continuously adapted to provide an increased decorrelation.

[0121] Specifically, the audio apparatus of FIG. 2 comprises spatial decorrelator 205 in the form of a set of spatial (decorrelation) filters that apply a spatial filtering to the first set of audio signals. Thus, following the filtering by the spatial filters, the first set of audio signals are modified to have increased decorrelation (for at least one audio source) than before

the spatial decorrelation filtering.

[0122] The spatial filters are based on coefficients that are adapted by an adaptive coefficient processor 207 which dynamically adapts and updates the filter coefficients used by the set of filters of the decorrelator 205. The spatial filters are accordingly set to have coefficients that are determined by the coefficient adapter 207.

[0123] The beamforming, decorrelation, and adaptation may typically be performed in the frequency domain. The receiver 201 may comprise a segmenter which is arranged to segment the set of input audio signals into time segments. In many embodiments, the segmentation may typically be a fixed segmentation into time segments of a fixed and equal duration such as e.g. a division into time segments/intervals with a fixed duration of between 10-20msecs. In some embodiments, the segmentation may be adaptive for the segments to have a varying duration. For example, the input audio signals may have a varying sample rate and the segments may be determined to comprise a fixed number of samples.

[0124] The segmentation may typically be into segments with a given fixed number of time domain samples of the input signals. For example, in many embodiments, the segmenter 203 may be arranged to divide the input signals into consecutive segments of e.g., 256 or 512 samples.

[0125] The receiver 201 may be arranged to generate a frequency bin representation of the input audio signals and the first set of input signals further processed are typically represented in the frequency domain by a frequency bin representation. The audio apparatus may be arranged to perform frequency domain processing of the frequency domain representation of the input audio signals. The signal representation and processing are based on frequency bins and thus the signals are represented by values of frequency bins and these values are processed to generate frequency bin values of the output signals. In many embodiments, the frequency bins have the same size, and thus cover frequency intervals of the same size. However, in other embodiments, frequency bins may have different bandwidths, and for example a perceptually weighted bin frequency interval may be used.

[0126] In some embodiments, the input audio signals may already be provided in a frequency representation and no further processing or operation is required. In some such cases, however, a rearrangement into suitable segment representations may be desired, including e.g. using interpolation between frequency values to align the frequency representation to the time segments.

[0127] In other embodiments, a filter bank, such as a Quadrature Mirror Filter, QMF, may be applied to the time domain input signals to generate the frequency bin representation. However, in many embodiments, a Discrete Fourier Transform (DFT) and specifically a Fast Fourier Transform, (FFT) may be applied to generate the frequency representation.

[0128] In the audio apparatus of FIG. 2, the spatial filters 205 specifically process the audio signals in the frequency domain. In the following description, the first set of audio signals may also be referred to as input audio signals (to the spatial filters) prior to the filtering and the resulting signals may also be referred as output audio signals (from the spatial filters).

[0129] For each frequency bin, an output frequency bin value is generated from one or more input frequency bin values of one or more input signals as will be described in more detail in the following. The output signals are generated to (typically/on average) reduce the correlation between signals relative to the correlation of the input signals, at least for one audio source which for example may be a dominant audio source.

[0130] The spatial filter set is arranged to filter the input audio signals. The filtering is a spatial filtering in that for a given output signal, the output value is determined from a plurality of, and typically all of the input audio signals (for the same time/segment and for the same frequency bin). The spatial filtering is specifically performed on a frequency bin basis such that a frequency bin value for a given frequency bin of an output signal is generated from the frequency bin values of the input signals for that frequency bin. The filtering/weighted combination is across the signals rather than being a typical time/frequency filtering.

[0131] Specifically, the frequency bin value for a given frequency bin is determined as the weighted combination of the frequency bin values of the input signals for that frequency bin. The combination may specifically be a summation and the frequency bin value may be determined as a weighted summation of the frequency bin values of the input signals for that frequency bin. The determination of a bin value for a given frequency bin may be determined as the vector multiplication of a vector of the weights/coefficients of the weighted summation and a vector comprising the bin values of the input signals:

$$[y] = [w_1 \quad w_2 \quad w_3] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where y is the output bin value, w_1- w_3 are the weights of the weighted combination, and x_1- x_3 are the input signal bin values.

[0132] Representing the output bin values for a given frequency bin ω as a vector $\boldsymbol{y}(\omega)$, the determination of the output signals may be determined as:

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega)$$

where the matrix $\boldsymbol{W}(\omega)$ represents the weights/coefficients of the weighted summation for the different output signals and $\boldsymbol{x}(\omega)$ is a vector comprising the input signal values.

[0133] For example, for an example with only three input signals and output signals, the output bin values for the frequency bin ω may be given by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where y_n represents the output bin values, w_{ij} represents the weights of the weighted combinations, and x_m represents the input signal bin values.

[0134] The adaptive coefficient processor 207 may seek to adapt the spatial filters to be spatial decorrelation filters that seek to generate the output signals to correspond to the input signals but with an increased decorrelation of the signals. The output audio signals are generated to have an increased spatial decorrelation with the cross correlation between audio signals being lower for the output audio signals than for the input audio signals. Specifically, the output signals may be generated to have the same combined energy/power as the combined energy/power of the input signals (or have a given scaling of this) but with an increased decorrelation (decreased correlation) between the signals. The output audio signals may be generated to include all the audio/signal components of the input signals in the output audio signals but with a re-distribution to different signals to achieve increased decorrelation.

[0135] The decorrelation filters may specifically be arranged to generate output signals that have a lower coherence/normalized decorrelation than the input signals. The output signals of the decorrelation filters may thus have lower coherence normalized decorrelation than the coherence in the input signals to the decorrelation filters.

[0136] The adapter 207 is arranged to determine update values for the weights of the weighted combination(s) forming the set of decorrelation filters. Thus, specifically, update values may be determined for the matrix $\boldsymbol{W}(\omega)$. The adapter 207 may then update the weights of the weighted combination based on the update values.

[0137] The adapter 207 is arranged to apply an adaptation approach that determines update values which may allow the output signals of the set of decorrelation filters 205 to represent the audio of the input signals to the set of decorrelation filters 205 but with the output signals typically being more decorrelated than the input signals.

[0138] The adapter 207 is arranged to use a specific approach of adapting the weights based on the generated output signals. The operation is based on each output audio signal being linked with one input audio signal. The exact linking between output signals and input signals is not essential and many different (including in principle random) linkings/pairings of each output signal to an input signal may be used. However, the processing is different for a weight that reflects a contribution from an input signal that is linked to/paired with the output signal for the weight than the processing for a weight that reflects a contribution from an input signal that is not linked to/paired with the output signal for the weight. For example, in some embodiments, the weights for linked signals (i.e. for input signals that are linked to the output signal generated by the weighted combination that includes the weight) may be set to a fixed value and not updated, and/or the weights for linked signals may be restricted to be real valued weights whereas other weights are generally complex values.

[0139] The adapter 207 uses an adaptation/update approach where an update value is determined for a given weight that represents the contribution to a bin value for a given output signal from a given non-linked input signal based on a correlation measure between the output bin value for the given output signal and the output bin value for the output signal that is linked with the given (non-linked) input signal. The update value may then be applied to modify the given weight in the subsequent segment, or the update value for weight in a given segment is determined in response to two output bin values of a (and typically the immediate) prior segment, where the two output values represent respectively the input signal and the output signal to which the weight relate.

[0140] The described approach is typically applied to a plurality, and typically all, of the weights used in determining the output bin values based on a non-linked input signal. For weights relating to the input signal that is linked to the output signal for the weight, other considerations may be used, such as e.g. setting the weight to a fixed value as will be described in more detail later.

[0141] Specifically, the update value may be determined in dependence on a product of the output bin value for the weight and the complex conjugate of the output bin value linked to the input signal for the weight, or equivalently in dependence on a product of the complex conjugate of the output bin value for the weight and the output bin value linked to the input signal for the weight.

[0142] As a specific example, an update value for segment k+1 for frequency bin ω may be determined in dependence on the correlation measure given by:

$$\left[y_i(k,\omega)\ y_j^*(k,\omega)\right]$$

or equivalently by:

$$\left[\ y_i^*(k,\omega)y_j(k,\omega)\ \right]$$

where $y_i(k, \omega)$ is the output bin value for the output signal i being determined based on the weight; and $y_j(k, \omega)$ is the output bin value for the output signal j that is linked to the input signal from which the contribution is determined (i.e. the input signal bin value that is multiplied by the weight to determine a contribution to the output bin value for signal i).

[0143] The measure of $\left[y_i(k,\omega)\ y_j^*(k,\omega)\right]$ (or the conjugate value) indicates the correlation of the time domain signal in the given segment. In the specific example, this value may then be used to update and adapt the weight $w_{i,j}(k+1,\omega)$.

[0144] As previously mentioned, the set of decorrelation filters may be arranged to determine the output bin values for the output signals for the given frequency bin ω from:

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega)$$

where $\mathbf{y}(\omega)$ is a vector comprising the frequency bin values for the output signals for the given frequency bin ω; $\mathbf{x}(\omega)$ is a vector comprising the frequency bin values for the input audio signals for the given frequency bin ω; and $\mathbf{W}(\omega)$ is a matrix having rows comprising weights of a weighted combination for the output audio signals.

[0145] In the example, adapter 207 may specifically be arranged to adapt at least some of the weights w_{ij} of the matrix $\mathbf{W}(\omega)$ according to:

$$w_{ij}(k+1,\omega) = w_{ij}(k,\omega) - \eta(k,\omega)\left[y_i(k,\omega)\ y_j^*(k,\omega)\right]$$

where i is a row index of the matrix $\mathbf{W}(\omega)$, j is a column index of the matrix $\mathbf{W}(\omega)$, k is a time segment index, ω represents the frequency bin, and $\eta(k, \omega)$ is a scaling parameter for adapting an adaptation speed. Typically, the adapter 207 may be arranged to adapt all the weights that are not relating the input signal with the linked output signal (i.e. "cross-signal" weights).

[0146] In some embodiments, the adapter 207 may be arranged to adapt the update rate/speed of the adaptation of the weights. For example, in some embodiments, the adapter may be arranged to compensate the correlation measure for a given weight in dependence on the signal level of the output bin value to which a contribution is determined by the weight.

[0147] As a specific example, the compensation value $\left[y_i(k,\omega)\ y_j^*(k,\omega)\right]$ may be compensated by the signal level of the output bin value, $|y_i(k, \omega)|$. The compensation may for example be included to normalize the update step values to be less dependent on the signal level of the generated decorrelated signals.

[0148] In many embodiments, such a compensation or normalization may specifically be performed on a frequency bin basis, i.e. the compensation may be different in different frequency bins. This may in many scenarios improve the operation and may typically result in an improved adaptation of the weights generating the decorrelated signals.

[0149] The compensation may for example be built into the scaling parameter $\eta(k,\omega)$ of the previous update equation. Thus, in many embodiments, the adapter 207 may be arranged to adapt/change the scaling parameter $\eta(k,\omega)$ differently in different frequency bins.

[0150] In many embodiments, the arrangement of the input signal vector $\mathbf{x}(\omega)$ the output signal vector $\mathbf{y}(\omega)$ is such that linked signals are at the same position in the respective vectors, i.e. specifically y_1 is linked with x_1, y_2 with x_2, y_3 is with x_3, etc. In this case, the weights for the linked signals are on the diagonal of the weight matrix $\mathbf{W}(k, \omega)$. The diagonal values may in many embodiments be set to a fixed real value, such as e.g. specifically be set to a constant value of 1.

[0151] In many embodiments, the weights/spatial filters/weighted combinations may be such that the weights for a contribution to a first output signal from a first input signal (not linked to the first output signal) is a complex conjugate of the contribution to a second output signal being linked with the first input signal from a second input signal being linked with the first input signal. Thus, the two weights for two pairs of linked input/output signals are complex conjugates.

[0152] In the example of the weights for linked input and output signals being arranged on the diagonal of the weight matrix $\mathbf{W}(\omega)$, this results in a Hermitian matrix. Indeed, in many embodiments, the weight matrix $\mathbf{W}(\omega)$ is a Hermitian matrix. Specifically, the coefficients/weights of the weight matrix $\mathbf{W}(\omega)$ may meet the criterion:

$$w_{ij} = w_{ji}^{*}.$$

[0153] As previously mentioned, the weights for the contributions to an output signal bin value from the linked input signal (corresponding to the values of the diagonal of the weight matrix **W**(ω) in the specific example) are treated differently than the weights for non-linked input signals. The weights for linked input signals will in the following for brevity also be referred to as linked weights and the weights for non-linked input signals will in the following for brevity also be referred to as non-linked weights, and thus in the specific example the weight matrix **W**(ω) will be a Hermitian matrix comprising the linked weights on the diagonal and the non-linked weights outside the diagonal.

[0154] In many approaches, the adaptation of the non-linked weights is such that it seeks to reduce the correlation measures. Specifically, each update value may be determined to reduce the correlation measure. Overall, the adaptation will accordingly seek to reduce the cross-correlations between the output signals. However, the linked weights are determined differently to ensure that the output signals maintain a suitable audio energy/power/level. Indeed, if the linked weights where instead adapted to seek to reduce the autocorrelation of the output signal for the weight, there is a high risk that the adaptation would converge on a solution where all weights, and thus output signals, are essentially zero (as indeed this would result in the lowest correlations). Further, the audio apparatus is arranged to seek to generate signals with less cross-correlation but do not seek to reduce autocorrelation.

[0155] Thus, in many embodiments, the linked weights may be set to ensure that the output signals are generated to have a desirable (combined) energy/power/level.

[0156] In some cases, the adapter 207 may be arranged to adapt the linked weights, and in other cases the adapter may be arranged to not adapt the linked weights.

[0157] For example, in some embodiments, the linked weights may simply be set to a fixed constant value that is not adapted. For example, in many embodiments, the linked weights may be set to a constant scalar value, such as specifically to the value 1 (i.e. a unitary gain being applied for a linked input signal). For example, the weights on the diagonal of the weight matrix **W**(ω) may be set to 1.

[0158] Thus, in many embodiments, the weight for a contribution to a given output signal frequency bin value from a linked input signal frequency bin value may be set to a predetermined value. This value may in many embodiments be maintained constant without any adaptation.

[0159] Such an approach has been found to provide very efficient performance and may result in an overall adaptation that has been found to provide output signals to be generated that provide a highly accurate representation of the original audio of the input signals but in a set of output signals that have increased decorrelation.

[0160] In some embodiments, the linked weights may also be adapted but will be adapted differently than the non-linked weights. In particular, in many embodiments, the linked weights may be adapted based on the output signals.

[0161] Specifically, in many embodiments, a linked weight for a first input signal and linked output signal may be adapted based on the generated output bin value of the linked audio signal, and specifically based on the magnitude of the output bin value.

[0162] Such an approach may for example allow a normalization and/or setting of the desired energy level for the signals.

[0163] In many embodiments, the linked weights are constrained to be real valued weights whereas the non-linked weights are generally complex values. In particular, in many embodiments, the weight matrix **W**(ω) may be a Hermitian matrix with real values on the diagonal and with complex values outside of the diagonal.

[0164] Such an approach may provide a particular advantageous operation and adaptation in many scenarios and embodiments. It has been found to provide a highly efficient spatial decorrelation while maintaining a relatively low complexity and computational resource.

[0165] The adaptation may gradually adapt the weights to increase the decorrelation between signals. In many embodiments, the adaptation may be arranged to converge towards a suitable weight matrix **W**(ω) regardless of the initial values and indeed in some cases the adaptation may be initialized by random values for the weights.

[0166] However, in many embodiments, the adaptation may be started with advantageous initial values that may e.g. result in faster adaptation and/or result in the adaptation being more likely to converge towards more optimal weights for decorrelating signals.

[0167] In particular, in many embodiments, the weight matrix **W**(ω) may be arranged with a number of weights being zero but at least some weights being non-zero. In many embodiments, the number of weights being substantially zero may be no less than 2,3,5, or 10 times the number of weights that are set to a non-zero value. This has been found to tend to provide improved adaptation in many scenarios.

[0168] In particular, in many embodiments, the adapter 207 may be arranged to initialize the weights with linked weights being set to a non-zero value, such as typically a predetermined non-zero real value, whereas non-linked weights are set to substantially zero. Thus, in the above example where linked signals are arranged at the same positions in the vectors, this will result in an initial weight matrix **W**(ω) having non-zero values on the diagonal and (substantially) zero values outside of

the diagonal.

[0169] Such initializations may provide particularly advantageous performance in many embodiments and scenarios. It may reflect that due to the audio signals typically representing audio at different positions, there is a tendency for the input signals to be somewhat decorrelated. Accordingly, a starting point that assumes the input signals are fully correlated is often advantageous and will lead to a faster and often improved adaptation.

[0170] It will be appreciated that the weights, and specifically the non-linked weights, may not necessarily be exactly zero but may in some embodiments be set to low values close to zero. However, the initial non-zero values may be at least 5,10,20 or 100 times higher than the initially substantially zero values.

[0171] The described approach may provide a highly efficient adaptive spatial decorrelator that may generate output signals that represent the same audio as the input signals but with increased decorrelation. The approach has in practice been found to provide a highly efficient adaptation in a wide variety of scenarios and in many different acoustic environments and for many different audio sources. For example, it has been found to provide a highly efficient decorrelation of speaker signals in environments with multiple speakers.

[0172] The adaptation approach is furthermore computationally efficient and in particular allows localized and individual adaptation of individual weights based only on two signals (and specifically on only two frequency bin values) closely related to the weight, yet the process results in an efficient and often substantially optimized global optimization of the spatial filtering, and specifically the weight matrix $\mathbf{W}(\omega)$. The local adaptation has been found to lead to a highly advantageous global adaptation in many embodiments.

[0173] A particular advantage of the approach is that it may be used to decorrelate convolutive mixtures and is not limited to only decorrelate instantaneous mixtures. For a convolutive mixture, the full impulse response determines how the signals from the different audio sources combine at the microphones (i.e. the delay/timing characteristics are significant) whereas for instantaneous mixes a scalar representation is sufficient to determine how the audio sources combine at the microphones (i.e. the delay/timing characteristics are not significant). By transforming a convolutive mixture into the frequency domain, the mixture can be considered a complex-valued instantaneous mixture per frequency bin.

[0174] The adaptive coefficient processor 207 thus determines the coefficients for the set of spatial decorrelation filters of the decorrelator 205 such that these modify the first set of audio signals to represent the same audio but with an increased decorrelation for at least one of the audio sources. Such a decorrelation of the audio signals to which the previously described beamforming approach is then applied may substantially improve the overall performance in many scenarios. Indeed, it may in many scenarios result in improved separation and selection of a specific audio source, such as a specific speaker. Thus, counterintuitively, the decorrelation of signals may provide improved beamforming performance despite the beamforming inherently being based on exploiting and adapting to correlation between audio signals from different positions to extract/separate an audio source by spatially forming a beam towards the desired source. Indeed, the decorrelation inherently breaks the link between the audio signals and the specific location in the audio scene that is typically exploited by a beamforming operation. However, the Inventors have realized that despite this, the decorrelation may provide highly advantageous effects and improved performance in many scenarios. For example, in the presence of a strong noise source, the approach may facilitate and/or improve the extraction/isolation of a specific desired audio source such as specifically a speaker.

[0175] The specific adaptation described above may provide a highly advantageous approach in many embodiments. It may typically provide a low complexity yet highly accurate adaptation to result in spatial decorrelation filters that generate highly decorrelated signals. In particular, it may allow local adaptation of individual weights/filter coefficients to result in a highly efficient global decorrelation of the first set of audio signals.

[0176] However, it will be appreciated that in other embodiments, other approaches for adapting the spatial filters/decorrelation filters may be used. For example, in some embodiments, the adaptive coefficient processor 207 may comprise a neural network which based on the input samples is arranged to generate update values for modifying the filter coefficients. For example, for each segment, the frequency bin values for all the audio signals may be fed to a trained neural network which as an output generates an update value for each weight. Each weight may then be updated by this update value. The trained network may for example have been trained by training data that includes many different frequency bin values and associated update values that have manually been determined to modify weights towards increased decorrelation.

[0177] As another example, the adaptive coefficient processor 207 may comprise a location processor which, based on visual cues related to (changing) positions, is arranged to generate update values for modifying the filter coefficients. As another example, the adapter 207 may determine the cross-correlation matrix of the input signals and compute its eigenvalue decomposition. The eigenvectors and eigenvalues can be used to construct a decorrelation matrix.

[0178] In the example of FIG. 2, the decorrelation and spatial decorrelation filters are directly applied to the first set of audio signals before these are fed to both the beamformer 101 and the beamform adapter 105. However, other approaches of adapting the operation based on applying a spatial/cross signal filtering using coefficients that are derived from the determined decorrelation coefficients may be used.

[0179] Specifically, FIG. 3 illustrates an example where the audio apparatus is arranged to perform a spatial filtering of

the first set of audio signals before these are filtered by the first set of filters, i.e. the beamforming is based on the first set of audio signals after these have been filtered by a set of spatial filters 301. However, in the example, the beamform adapter 105 receives the first set of audio signals before any filtering by the section set of filters 301, i.e. the second set of filters are applied only to the beamformer path and not to the adaptation path.

[0180] The coefficients for this set of spatial filters 301 are determined from the decorrelation coefficients that have been determined by the coefficient adapter 207. Indeed, in some embodiments, the approach described with reference to the configuration of FIG. 2 may be applied directly and the resulting second set of filters may be applied (only) to the signals of the beamform path.

[0181] However, in many embodiments, the set of spatial filters 301 in the configuration example of FIG. 3 will be different from the set of spatial filters 201 in the configuration example of FIG. 2. In particular, in many embodiments, the set of spatial filters will be modified to correspond to a cascade of two sets of decorrelation filters as determined by the adaptive coefficient processor 207. Thus, filter coefficients of the set of spatial filters 301 have coefficients matching coefficients of a spatial filter that is a cascade of two of the set of decorrelation filters. This can typically be considered equivalent to a double/repeated filtering of the first set of audio signals by the set of decorrelation filters determined by the adaptive coefficient processor 207.

[0182] In particular, the adaptive coefficient processor 207 may determine the weights/coefficients of a weight matrix $\mathbf{W}(\omega)$ for a decorrelation filter that can decorrelate the first set of audio signal in accordance with:

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega)$$

as previously described.

[0183] The set of spatial filters 301 may in this case be generated to correspond to a cascaded application of two such filters, i.e. it may be arranged to correspond to:

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{W}(\omega)\mathbf{x}(\omega) = \mathbf{W}^{\mathbf{2}}(\omega)\mathbf{x}(\omega)$$

[0184] Thus, in the example, the adaptive coefficient processor 207 may proceed to perform the adaptation as described for the example of FIG. 2 to determine suitable coefficients for a set of spatial decorrelation filters that would decorrelate the first set of audio signals. In some cases, such filters may then be applied by the set of spatial filters 301 in FIG. 3. However, in many embodiments, the determined decorrelation filters are not used directly but rather coefficients for the set of spatial filters 301 are determined from the determined coefficients for the decorrelation filters. In such cases, the adaptive coefficient processor 207 may as part of the determination/adaptation of coefficients implement a decorrelation filter and apply this to the first set of signals (e.g. in order to determine update values from $\left[y_i(k,\omega)\, y_j^*(k,\omega)\right]$. However, in the specific example, such filtered signals may only be used for the adaptation process and may not be used further in the beamforming/ processing. Instead, the set of spatial filters 301 that are applied may be generated from the decorrelation coefficients/weight matrix $\mathbf{W}(\omega)$, specifically as $\mathbf{W^2}(\omega)$.

[0185] The approach may provide improved overall performance including an improved beamforming experience. Indeed, it can be shown that for the example where the set of spatial filters 301 of FIG. 3 are set to coefficients corresponding to $\mathbf{W^2}$ then this may result in the same performance, results, and output signal as for the example of FIG. 2 It can be shown that these approaches may result in the same optimum solutions.

[0186] Another possible example of applying a spatial filtering by a set of spatial filters determined from the decorrelation coefficients determined by the adaptive coefficient processor 207 is illustrated in FIG. 4. In this example, the first set of audio signals is not filtered by the set of spatial filters but rather the second set of audio signals generated by the feedback circuit 103 are filtered by the set of spatial filters 401. Thus, in this example, the feedback signals rather than the beamform signals are filtered by the set of spatial filters.

[0187] Further, the set of spatial filters may be determined as a set of a set of inverse spatial decorrelation filters where each of the inverse spatial de correlation filters may include an inverse filtering of a determined decorrelation filter.

[0188] For example, if the spatial decorrelation filter is represented by the weight matrix $\mathbf{W}(\omega)$, an inverse spatial decorrelation filter may be determined by the weight matrix $\mathbf{W^{-1}}(\omega)$. Thus, in many embodiments, the set of spatial filters applied to the second set of audio signals may include a filtering corresponding to $\mathbf{W^{-1}}(\omega)$ where $\mathbf{W}(\omega)$ is the spatial decorrelation determined by the adaptive coefficient processor 207.

[0189] In many embodiments, the set of spatial filters applied to the second set of audio signals may be determined to have filter coefficients that correspond to coefficients of a spatial filter which is a cascade of two sets of spatial inverse filters each of which is inverse filters of the spatial decorrelation filters determined by the adaptive coefficient processor 207.

[0190] In particular, the adaptive coefficient processor 207 may thus determine the weights/coefficients of a weight matrix $\mathbf{W}(\omega)$ for a decorrelation filter that can decorrelate the first set of audio signal in accordance with:

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega)$$

[0191] The set of spatial filters 301 may in this case be generated to correspond to a cascaded application of two such filters, i.e. it may be arranged to correspond to:

$$\mathbf{y}(\omega) = \mathbf{W}^{-1}(\omega)\,\mathbf{W}^{-1}(\omega)\mathbf{s}(\omega) = \mathbf{W}^{-2}(\omega)\mathbf{s}(\omega)$$

where $\mathbf{s}(\omega)$ represents the second set of audio signals generated by the feedback circuit 103.

[0192] Thus, in the example, the adaptive coefficient processor 207 may proceed to perform the adaptation as described for the example of FIG. 2 to determine suitable coefficients for a spatial decorrelation filter that would decorrelate the first set of audio signals. The determined decorrelation filter coefficients are not used directly but rather coefficients for the set of spatial filters 401 is determined from the determined coefficients. In such cases, the adaptive coefficient processor 207 may as part of the determination/adaptation of coefficients implement a decorrelation filter and apply this to the first set of signals (e.g. in order to determine update values from $\left[y_i(k,\omega)\; y_j^*(k,\omega)\right]$. However, in the specific example, such filtered signals may only be used for the adaptation process and may not be used further in the beamforming/ processing.

[0193] The approach may provide improved overall performance including an improved beamforming experience. Indeed, it can be shown that for the example in which the set of spatial filters 401 of FIG. 4 are set to coefficients corresponding to $\mathbf{W}^{-2}$ then this may result in the same performance and operation as for the example of FIG. 2, and specifically that these may result in the same optimum solutions.

[0194] The audio apparatus implements a highly adaptive approach that may be particularly suitable for adapting to extract specific audio sources, such as a desired speaker. The approach may be particularly advantageous when e.g. a strong noise source is present in the audio environment being captured. The adaptive audio apparatus comprises different adaptations of respectively the set of spatial filters (via the adaptation of the decorrelation filters/decorrelation coefficients) and the adaptation of the beamformer filters. The adaptive coefficient processor 207 and beamform adapter 105 synergistically interwork to provide a highly advantageous and high performance adaptation to the current audio properties of the scene.

[0195] In many embodiments, the audio apparatus may be arranged to adapt the different adaptations at different times. Thus, the adaptive coefficient processor 207 and the beamform adapter 105 may be controlled to be adapted at different time periods, and in many embodiments these may be nonoverlapping. Thus, at a given time, either the adaptive coefficient processor 207 or the beamform adapter 105 may be adapting but not both (although for some periods neither may be adapted).

[0196] For example, as illustrated in FIG. 5, the audio apparatus of FIG. 2 (or correspondingly of FIG. 3 or 4) may include an audio detector 501 which is arranged to detect when an audio source is active and when it is not. The audio detector 501 may specifically be arranged to divide the time into (at least) a set of active time intervals during which an audio source is active and a set of inactive time intervals during which the audio source is not active. The audio source may specifically be a desired audio source, such as specifically a desired speaker.

[0197] In some embodiments, a low complexity audio detection may be used to determine whether a source is active or not. For example, the audio apparatus may be used to extract the dominant audio source in the environment, such as for example the strongest speaker. For example, typically for teleconferencing applications, only one person will be speaking at a time and the audio apparatus may be used to extract the speech of the current speaker from background or ambient noise.

[0198] The audio detector 501 may for example in such applications simply detect the audio source activity based on a level/ energy being above a given threshold. If the signal level is above the given threshold (which may e.g. be a dynamic threshold that is set based on a longer time averaged signal level), the audio detector 501 may consider the audio source/speaker to be active and otherwise it may consider the audio source/speaker to not be active. Thus, the audio detector 501 may divide the time into active time intervals when the captured audio is above a threshold (due to the audio source is active) and non-active time intervals when the signal level is below the threshold (due to the audio source not being active).

[0199] In many embodiments, more complex detections may be used and indeed techniques may be used that separate between different audio sources. For example, in many embodiments the detection may include a consideration of whether the audio signal has desired/expected properties matching the specific audio source. For example, the audio detector 501 may differentiate between speech and other types of audio based on evaluating whether the captured audio has properties matching speech.

[0200] It will be appreciated that many different techniques and algorithms are known for speech/voice activity detection, and more generally for detecting that an audio source is active, and that any suitable approach may be used without detracting from the invention.

[0201] In some embodiments, a trained artificial neural network may be used, and it has in practice been found that effective speech/voice activity detection (and/or noise detection) can be performed. The artificial neural network may be trained with speech and all types of non-speech and may for each frame provide an indication whether it is noise or speech.

[0202] In many embodiments, the detection may be relatively fast and the audio detector 501 may be arranged to designate relatively short time intervals as active or non-active time intervals. For example, in some embodiments, the audio detector 501 may be arranged to detect silent pauses/intervals during normal speech and designate such intervals as non-active time intervals. For example, time intervals less than e.g. 5 msec, 10 msec, or 20 msec may in some embodiments be identified/designated as active or non-active time intervals.

[0203] In many embodiments, the audio detector 501 may directly detect activity for an audio source based on the received microphone signals/ the first set of audio signals/ the beamformer signals. For example, the total captured audio energy may be determined and compared to a threshold. However, in other embodiments other signals may be considered, such as e.g. the second set of audio signals. In many embodiments, the audio detector 501 may base the activity detection on the generated beamform output audio signal. For example, level detection or speech detection may be applied directly to the beamform output audio signal. This may in many embodiments provide improved performance as the output signal may specifically be generated to focus on the desired signal, such as the desired speaker.

[0204] The audio detector 501 may be arranged to control the adaptation based on the audio source activity detection and specifically the adaption of the decorrelation coefficients and/or of the beamform filter coefficients may be different in active time intervals than in inactive time intervals.

[0205] Indeed, in some embodiments, the adaptation of the decorrelation coefficients may not be performed during the active time intervals but only during the inactive time intervals. The audio detector 501 may specifically seek to detect whether a desired or wanted audio source/speaker is active. It may further control the adaptive coefficient processor 207 to adapt the decorrelation coefficients only when this audio source is not active. Thus, the adaptive coefficient processor 207 may be arranged to adapt to provide decorrelation filters that seek to decorrelate the unwanted audio. Thus, rather than update the decorrelation filters to be a decorrelation of the entire captured audio, it will be aimed at adapting to reduce decorrelation of the undesired audio. Such an approach may provide a very efficient and improved operation and performance in many situations. The decorrelation approach for the non-desired audio may allow improved beamforming with improved rejection of the undesired audio by the beamforming while allowing the beamforming to provide efficient and high performance extraction of the wanted signal.

[0206] In some embodiments, the adaptive coefficient processor 207 may be arranged to adapt the decorrelation coefficients in both the active time intervals and the inactive time intervals but with the adaption rate being higher during the inactive time intervals than during the active time intervals. Thus, rather than only adapting the decorrelation coefficients during the inactive time interval, there may also be some adaptation during the active time intervals but with this adaptation being slower, e.g., by a factor of no less than 2,5,10 or 100. This may be advantageous in some scenarios, such as for example where the active time intervals are much longer than the inactive time intervals. In some embodiments, the update rate during the active time intervals and/or the inactive time intervals may be dynamically adapted depending on e.g., properties of the time intervals or the input signals. For example, the adaptive coefficient processor 207 may switch between only updating during the inactive time intervals to also updating (at a lower rate) during the active time intervals if the duration of the active time intervals exceed a given duration.

[0207] In many embodiments, the adaptation of the beamforming filters may not be performed during the inactive time intervals but only during the active time intervals. The audio detector 501 may specifically seek to detect whether a desired or wanted audio source/speaker is active and control the beamform adapter 105 to adapt the beamform filters only when this audio source is active. Thus, the beamform adapter 105 is arranged to specifically adapt the beamforming to specifically (virtually) form a beam towards the desired audio source.

[0208] In some embodiments, the beamform adapter 105 may be arranged to adapt the beamformer coefficients in both the active time intervals and the inactive time intervals but with the adaption rate being higher during the active time intervals than during the inactive time intervals. Thus, rather than only adapting the beamform filters during the active time interval, there may also be some adaptation during the inactive time intervals but with this adaptation being slower, e.g. by a factor of no less than 2,5,10 or 100. This may be advantageous in some scenarios, such as for example where the audio detector 501 is set to provide a detection with a very low risk of false detection of the active audio source such that the risk of not detecting the audio source being active is relatively high. In such cases, it may be desirable to still adapt during the inactive time interval but with an update rate that is (typically) much lower. Thus, a low update rate may be used during time intervals when the wanted source may or may not be active and with a high update rate being used during time intervals when the wanted source is almost certainly present.

[0209] In many embodiments the audio apparatus may be arranged to adapt the decorrelation coefficients (only) during the inactive time intervals and to adapt the beamforming filters (only) during the active time intervals. This may provide highly efficient performance and in particularly may provide substantially improved extraction/separation of a desired audio source/speaker in the presence of noise/ undesired audio, such as specifically a dominant and correlated noise

source.

[0210] In many practical applications, such as many speech capture and processing applications, the adaptation control of both the beamforming and decorrelation is important for the effective operation of the audio apparatus. It may be desired for the beamformer to adapt on the speech, which by nature is not always active, and the decorrelation on the noise only part. In use cases where the noise is continuously available, the beamformer can only adapt when there is also noise present. In case the noise is uncorrelated, or the noise is decorrelated by the decorrelation, the noise does not influence the adaptation if the noise is equivariant. If the noise is not equivariant, the beamformer will diverge during noise only, but the beamformer can often quickly (almost unnoticeably) adjust to the desired speaker if near-end becomes active. Depending on the application, either no further adaptation control is used, such that the beamformer can quickly find new sources, or a voice-activity detector is used, which can vary from a simple energy-based detector until more sophisticated detectors that also use speech characteristics based on pitch for example.

[0211] The detector for the noise is typically made more conservative, such that it is not active when speech is also present, and it starts to decorrelate the speech. Depending on the noise type and level of the noise, energy-based detectors can be used that distinguish (stationary) noise from speech, or more sophisticated ones, e.g. neural-net classifiers that are trained to distinguish noise from speech.

[0212] In the above, the description focused on a scenario where the active time intervals are wanted/speech audio source active time intervals and the inactive time intervals are unwanted source/noise/interference active time intervals. In this case, the adaptation rate for the filters is higher during the active time intervals than during the inactive time intervals, and specifically the adaptation rate for the filters may be performed only during the active time intervals and not during during the inactive time intervals. Equivalently, the adaptation rate for the decorrelation coefficients is higher during the inactive time intervals than during the active time intervals, and specifically the adaptation rate for the filters may be performed only during the inactive time intervals and not during the active time intervals.

[0213] Such time intervals may for example be determined directly by the use of a speech detector which detects and determines the speech active time intervals, which may be considered as the active time intervals. The remaining time intervals, i.e. the non- speech active time intervals may be considered as the inactive time intervals.

[0214] In some embodiments, a detection of undesired audio properties may be used, such as specifically a detection of the activity of an interferer or noise source. For example, detection of activity of an undesired audio source may be used, e.g., a music detector, silence detector, specific noises detector. In such cases, the time intervals detected may be considered as the inactive time intervals. The remaining time intervals may be considered as the active time intervals.

[0215] It will be appreciated that if instead, active time intervals were considered to correspond to time intervals in which an undesired audio source is active and inactive time intervals were considered to correspond to time intervals in which the undesired audio source is not active, then the previously described adaptation would be reversed (adaptation rate of filters would be higher during inactive time intervals, adaptation rate of decorrelation coefficients would be higher during active time intervals, and specifically in many embodiments adaptation of filters would only be in inactive time intervals and adaptation of decorrelation coefficients would only be in active time intervals).

[0216] In the following, it will be demonstrated that the specific configurations of FIGs. 2, 3, and 4 can provide the same solutions. For brevity, these configurations and solutions will be referred to as A, B and C respectively. For the different solutions, we will derive the optimum filter coefficients and the corresponding output.

[0217] With configuration A (FIG. 2), the set of spatial filters is placed directly after the microphones, completely outside the beamformer. This has the advantage that nothing has to be changed to the beamformer algorithm: it will still have the constraint $\boldsymbol{f}^H(\omega)\boldsymbol{f}(\omega) = 1$ only the inputs differ. A suitable decorrelator may be as previously described. This decorrelator transforms the noise covariance matrix $\boldsymbol{R}_{nn}(\omega)$ into:

$$\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^H(\omega) = \Lambda(\omega) \quad \text{(Eq. 9)}$$

where $\Lambda(\omega) = diag(\lambda_1(\omega) \ldots . \lambda_{\boldsymbol{N}mics}(\omega))$ is a diagonal matrix and $\boldsymbol{W}(\omega)$ the $nmics \times nmics$ decorrelation matrix of the SDC. Instead of maximizing $\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega)$ the beamformer now maximizes $\boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{W}^H(\omega)\boldsymbol{f}(\omega)$ which can be written as

$$\begin{aligned}\boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^H(\omega)\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^H(\omega)\boldsymbol{f}(\boldsymbol{\omega}) \\ = \boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^H\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\Lambda(\omega)\boldsymbol{f}(\omega)\end{aligned}$$

[0218] If $\Lambda(\omega)$ is a scaled version of the identity matrix and can be written as $\beta(\omega)\boldsymbol{I}$, $\boldsymbol{f}^H(\omega)\Lambda(\omega)\boldsymbol{f}(\omega)$ transforms to $\beta(\omega)$ and the noise contribution is independent of the choice of $\boldsymbol{f}(\omega)$and maximization of $\boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{W}^H(\omega)\boldsymbol{f}(\omega)$ leads to maximization of the Signal-to-Noise ratio in the output.

[0219] The choice of $\beta(\omega)$ in the decorrelator does not influence the SNR but influences the level of the speech in the

output. Suppose we choose $\beta(\omega) = 1$ and the noise level at the input increases by a factor of 2, then all the coefficients in $\boldsymbol{W}(\omega)$ will be scaled by a factor of 0.5, so also decreasing the level of the speech at the output of the SDC. The SDC keeps updating depending on the level of the input (per frequency bin). As a result also the beamformer keeps updating. It can be shown that $\Lambda(\omega) = \boldsymbol{\alpha(\omega)\rho_{nn}^2(\omega)I}$, i.e. linearly related to the power of the noise source, $\boldsymbol{W}(\omega)$ only depends on the acoustic transfer functions between the noise source and the microphones. Since the transfer functions generally vary much more slowly when compared to the noise characteristics, this is also advantageous for the beamformer, since it sees a "stable" path. A solution is specifically where $diag(\boldsymbol{W}(\omega)) = \boldsymbol{I}$. $\Lambda(\omega)$ will have elements $\lambda_i(\omega)$ that are linearly related to the input. A disadvantage is that not all elements of $\Lambda(\omega)$ are equal and as a result, the noise in the output depends on $\boldsymbol{f}(\omega)$. Although this is not desirable, in practice it was found that it is typically more important that $\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^{\boldsymbol{H}}(\omega)$ does not contain off-diagonal elements then that all the elements on the diagonal are equal. The combined solution with a beamformer preceded by a decorrelator with $diag(\boldsymbol{W}(\omega)) = \boldsymbol{I}$ has proven to be very effective.

[0220] The optimal coefficients for the beamformer can be found, using Equation (8) with instead of $\boldsymbol{h}(\omega)$ using the output of the spatial decorrelator: $\boldsymbol{W}(\omega)\boldsymbol{h}(\omega)$.

[0221] Using a decorrelator with $\boldsymbol{W}(\omega)$ Hermitian and thus $\boldsymbol{W}^H(\omega) = \boldsymbol{W}(\omega)$ we get for the optimal solution:

$$\boldsymbol{f}_{\mathrm{opt}}(\omega) = \mathrm{a}(\omega)\frac{\boldsymbol{W}(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}(\omega)\boldsymbol{h}(\omega),}}$$

$$= \mathrm{a}(\omega)\frac{\boldsymbol{W}(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^{\boldsymbol{2}}(\omega)\boldsymbol{h}(\omega)}}$$

[0222] The output of the combination is given by:

$$z(\omega) = \boldsymbol{f}_{opt}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{x}(\omega) \quad \text{(Eq. 10)}$$

$$\boldsymbol{u}(\omega) = s(\omega)\boldsymbol{h}(\omega) \qquad \text{(Eq. 11)}$$

$$\boldsymbol{f}_{\mathrm{opt}}^{\mathrm{H}}(\omega) = \frac{\mathrm{a}^*(\omega)\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}} \quad \text{(Eq. 12)}$$

[0223] Substituting (12) and (11) into (10) gives for the output:

$$\mathrm{z}(\omega) = \mathrm{a}^*(\omega)\frac{\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}}\mathrm{s}(\omega)$$

$$= \mathrm{a}^*(\omega)\sqrt{\boldsymbol{h}^{\mathrm{H}}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}\,\mathrm{s}(\omega)$$

[0224] Note that with equal variance uncorrelated noise at the inputs, using the decorrelator with $diag(\boldsymbol{W}(\omega)) = \boldsymbol{I}$ and thus $\boldsymbol{W}^2(\omega) = \mathbf{W}(\omega) = \mathbf{I}$, we get the same solution as without decorrelator with only the beamformer.

[0225] For configuration B (as illustrated in FIG. 3), it is firstly noted that the decorrelation can transform the noise covariance matrix $\boldsymbol{R}_{nn}(\omega)$ into:

$$\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^H(\omega) = \boldsymbol{\Lambda}(\omega)$$

[0226] With $\boldsymbol{W}(\omega)$ and $\boldsymbol{R}_{nn}(\omega)$ positive definite, their inverse exist and we can write:

$$\boldsymbol{W}^{-1}(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^H(\omega)\left(\boldsymbol{W}^H(\omega)\right)^{-1} = \boldsymbol{W}^{-1}(\omega)\boldsymbol{\Lambda}(\omega)\left(\boldsymbol{W}^H(\omega)\right)^{-1}$$

$$\boldsymbol{R}_{nn}(\omega) = \boldsymbol{W}^{-1}(\omega)\boldsymbol{\Lambda}(\omega)\left(\boldsymbol{W}^{H}(\omega)\right)^{-1}$$

[0227] If $\Lambda(\omega)$ is a scaled version of the identity matrix and can be written as $\rho_{nn}^{2}(\omega)\boldsymbol{I}$ then we get for the inverse of $\boldsymbol{R}_{nn}(\omega)$:

$$\boldsymbol{R}_{nn}^{-1}(\boldsymbol{\omega}) = \frac{\boldsymbol{W}^{H}(\omega)\boldsymbol{W}(\omega)}{\rho_{nn}^{2}(\omega)}$$

[0228] Using the decorrelator with $\boldsymbol{W}^{H}(\omega) = \boldsymbol{W}(\omega)$ we can finally write:

$$\boldsymbol{W}^{2}(\omega) = \rho_{nn}^{2}(\omega)\boldsymbol{R}_{nn}^{-1}(\omega)$$

$\boldsymbol{W}^{2}(\omega)$ can be seen as the inverse of a normalized covariance matrix. Note that in the adaptive coefficient processor 207, $\boldsymbol{W}(\omega)$ is calculated, whereas in the filtering part the squared matrix $\boldsymbol{W}^{2}(\omega)$ is used.

[0229] We will now proof that this combination maximizes the speech-to-noise ratio in the output. First we will derive the filter coefficients after convergence like we did before for the beamformer.

[0230] At the output of the filtering part we have:

$$z(\omega) = \boldsymbol{f}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{u}(\omega) = s(\omega)\boldsymbol{f}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{h}(\omega) \quad \text{(Eq. 13)}$$

[0231] In the update part we have:

$$z(\omega)\boldsymbol{f}(\omega) = s(\omega)\boldsymbol{h}(\omega) \qquad \text{(Eq 14)}$$

[0232] Since $z(\omega)$ and $s(\omega)$ are scalars we can write:

$$\mathbf{f}(\omega) = \alpha(\omega)\boldsymbol{h}(\omega)$$

[0233] Substitution in Eq 14 using Eq 13 gives:

$$\alpha^{*}(\omega)\boldsymbol{h}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{h}(\omega)\alpha(\omega)\boldsymbol{h}(\omega) = \boldsymbol{h}(\omega)$$

$$|\alpha(\omega)|^{2}\left(\boldsymbol{h}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{h}(\omega)\right) = 1$$

$$\alpha(\omega) = \frac{a(\omega)}{\sqrt{\mathbf{h}^{H}(\omega)\mathbf{W}^{2}(\omega)\mathbf{h}(\omega)}}$$

with $a(\omega)$ the common all-pass. Finally we get:

$$\boldsymbol{f}_{opt}(\omega) = \frac{a(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^{\mathrm{H}}(\omega)\mathbf{W}^{2}(\omega)\boldsymbol{h}(\omega)}}$$

[0234] Substitution of $\boldsymbol{f}_{opt}(\omega)$ into $\boldsymbol{f}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{f}(\omega)$ gives:

$$\boldsymbol{f}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{f}(\omega) = \frac{\boldsymbol{h}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{f}(\omega)}{\boldsymbol{h}^{H}(\omega)\boldsymbol{W}^{2}(\omega)\boldsymbol{f}(\omega)} = 1$$

so the constraint is met.

[0235] Next, we look at the expected value of the update after convergence:

$$\mathbb{E}\{z^*(\omega)\big(\boldsymbol{u}(\omega) - z(\omega)\boldsymbol{f}(\omega)\big)\} = 0,$$

$$\mathbb{E}\{z^*(\omega)\boldsymbol{u}(\omega)\} = \mathbb{E}\{|z(\omega)|^2\}\boldsymbol{f}(\omega) \quad \text{(Eq. 15)}$$

[0236] With

$$\mathrm{z}(\omega) = \big(\boldsymbol{W}^2(\omega)\boldsymbol{u}(\omega)\big)^T\boldsymbol{f}^*(\omega)$$

$$z^*(\omega) = \big(\boldsymbol{W}^2(\omega)\boldsymbol{u}(\omega)\big)^H\boldsymbol{f}(\omega)$$

$$z^*(\omega) = \boldsymbol{u}^H(\omega)\big(\boldsymbol{W}^2(\omega)\big)^H\boldsymbol{f}(\omega) = \boldsymbol{u}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega)$$

where we use $\mathbf{W}(\omega) = \boldsymbol{W}^H(\omega)$ and $(\boldsymbol{W}^2(\omega))^H = \boldsymbol{W}^H(\omega)\,\boldsymbol{W}^H(\omega) = \mathbf{W}(\omega)\mathbf{W}(\omega) = \boldsymbol{W}^2(\omega)$. We then get for the update equation (Eq. 15):

$$\begin{aligned}\mathbb{E}\{z^*(\omega)\boldsymbol{u}(\omega)\} &= \mathbb{E}\{\boldsymbol{u}(\omega)z^*(\omega)\}\\ &= \mathbb{E}\{\boldsymbol{u}(\omega)\boldsymbol{u}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega)\}\\ &= \boldsymbol{R}_{uu}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) = \rho_{zz}^2(\omega)\boldsymbol{f}(\omega) \quad \text{(Eq. 16)}\end{aligned}$$

with $\boldsymbol{R}_{\mathrm{uu}}(\omega) = \mathbb{E}\{\boldsymbol{u}(\omega)\boldsymbol{u}^H(\omega)\}$ and $\rho_{zz}^2(\omega) = \mathbb{E}\{|z(\omega)|^2\}$.

[0237] What Equation 16 shows is that after convergence $\boldsymbol{f}(\omega)$ is an eigen vector and $\rho_{zz}^2(\omega)$ the corresponding eigen value of the matrix $\boldsymbol{R}_{uu}(\omega)\boldsymbol{W}^2(\omega)$. As before it can be shown that the solution is only stable if $\rho_{zz}^2(\omega)$ is the largest eigen value.

[0238] Multiplying left and right-hand side with $\boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)$we get:

$$\begin{aligned}\boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) &= \rho_{zz}^2(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega)\\ &= \rho_{zz}^2(\omega)\end{aligned}$$

[0239] Note that the left-hand side contains the output power of the beamformer.

[0240] Now consider: $\boldsymbol{R}_{uu}(\omega) = \boldsymbol{R}_{ss}(\omega) + \boldsymbol{R}_{nn}(\omega)$,

where the adaptive decorrelator has adapted on the noise source. We then have for the output power $\mathbb{E}\{|z|^2(\omega)\}$:

$$\begin{aligned}\mathbb{E}\{|z|^2(\omega)\} &= \boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega)\\ &= \boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}(\omega)\boldsymbol{f}(\omega) =\\ &\boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) + \boldsymbol{\rho}_{nn}^2(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{W}(\omega)\boldsymbol{W}(\omega)\boldsymbol{f}(\omega)\\ &= \boldsymbol{f}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{f}(\omega) + \boldsymbol{\rho}_{nn}^2(\omega)\end{aligned}$$

[0241] The noise contribution is independent from the choice of $\boldsymbol{f}(\omega)$ and thus maximizing the output of the FSB corresponds to maximizing the speech-to-noise ratio in the output, not necessarily the speech output itself.

[0242] The optimal solutions after convergence for solutions A and B are related by:

$$\boldsymbol{f}_A(\omega) = \mathbf{W}(\omega)\boldsymbol{f}_B(\omega),$$

[0243] This means that the outputs for both solutions will be equal.

[0244] For solution C of FIG. 4, a matrix block may be placed after the filters in the feedback circuit 103. Here the constraint equals $\boldsymbol{f}^H(\omega)\Gamma(\omega)\boldsymbol{f}(\omega)=1$, where $\Gamma(\omega)$ is the coherence matrix, which is equal to a normalized covariance matrix. We want to use the decorrelator again and we use

$$\boldsymbol{W}(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{W}^H(\omega) = \boldsymbol{\Lambda}(\omega) = \boldsymbol{\rho}_{nn}^2(\omega)\boldsymbol{I}$$

$$\boldsymbol{R}_{nn}(\omega) = \boldsymbol{\rho}_{nn}^2(\omega)\boldsymbol{W}^{-1}(\omega)\left(\boldsymbol{W}^H(\omega)\right)^{-1}$$

$$\frac{\boldsymbol{R}_{nn}(\omega)}{\boldsymbol{\rho}_{nn}^2(\omega)} = \boldsymbol{W}^{-2}(\omega)$$

[0245] Thus $\boldsymbol{W}^{-2}(\omega)$ equals the normalized covariance matrix and we can use $\boldsymbol{W}^{-2}(\omega)$ instead of $\Gamma(\omega)$ and the constraint becomes $\boldsymbol{f}^H(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega) = 1$.

[0246] Like before we find the optimal solution for the filter coefficients $\boldsymbol{f}(\omega)$ by considering the signal flow in the filtering part and update part.

[0247] In the filtering part we have:

$$z(\omega) = s(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{h}(\omega) \quad \text{(Eq. 17)}$$

and in the update part:

$$z(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega) = s(\omega)\boldsymbol{h}(\omega) \quad \text{(Eq. 18)}$$

$$\boldsymbol{f}(\omega) = \alpha(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega) \quad \text{(Eq. 19)}$$

where we use that $z(\omega)$ and $s(\omega)$ are complex scalers.

[0248] Substitution of (19) into (18), using (17) gives:

$$\alpha^*(\omega)\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{W}^{-2}(\omega)\alpha(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega) = \boldsymbol{h}(\omega)$$

$$|\alpha(\omega)|^2 = \frac{1}{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}$$

[0249] Using (Eq. 19) we get:

$$\boldsymbol{f}_{opt}(\omega) = \frac{\mathrm{a}(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}}$$

with $\mathrm{a}(\omega)$ allpass.

[0250] Substitution of $\boldsymbol{f}_{opt}(\omega)$ into $\boldsymbol{f}^H(\omega)\ \boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)$ gives:

$$\boldsymbol{f}^H(\omega)\ \boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega) = \frac{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\mathbf{W}^{-2}(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)} = 1$$

meeting the constraint.

[0251] Finally we look at the expected value of the update after convergence

$$\mathbb{E}\{z^*(\omega)\big(\boldsymbol{u}(\omega) - z(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)\big)\} = 0$$

$$\mathbb{E}\{z^*(\omega)\boldsymbol{u}(\omega)\} = \mathbb{E}\{|z(\omega)|^2\}\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)$$

[0252] With $z^*(\omega) = \boldsymbol{u}^H\boldsymbol{f}(\omega)$:

$$\mathbb{E}\{\boldsymbol{u}(\omega)\boldsymbol{u}^H(\omega)\}\boldsymbol{f}(\omega) = \mathbb{E}\{|z(\omega)|^2\}\,\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)$$

$$\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) = \rho_{zz}^2(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega) \quad \text{(Eq. 20)}$$

with $\boldsymbol{R}_{uu}(\omega) = \mathbb{E}\{\boldsymbol{u}(\omega)\boldsymbol{u}^H(\omega)\}$ and $\rho_{zz}^2(\omega) = \mathbb{E}\{|z(\omega)|^2\}$.

[0253] Multiplying both sides with $\mathbf{W}^2(\omega)$ gives:

$$\begin{aligned}\boldsymbol{W}^2(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) &= \rho_{zz}^2(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)\\ &= \rho_{zz}^2(\omega)\boldsymbol{f}(\omega)\end{aligned}$$

[0254] After convergence $\boldsymbol{f}(\omega)$ is an eigen vector and $\rho_{zz}^2(\omega)$ the corresponding eigen value of the matrix $\boldsymbol{W}^2(\omega)\boldsymbol{R}_{uu}(\omega)$. As before, it can be shown that the solution is only stable if $\rho_{zz}^2(\omega)$ is the largest eigen value.

[0255] Multiplying Eq. 20 left and right with $\boldsymbol{f}^H(\omega)$ gives:

$$\boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) = \rho_{zz}^2(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega) = \rho_{zz}^2(\omega)$$

[0256] Note that the left-hand side contains the output power of the FSB.

[0257] Now consider: $\boldsymbol{R}_{uu}(\omega) = \boldsymbol{R}_{ss}(\omega) + \boldsymbol{R}_{nn}(\omega)$, where the adaptive decorrelator has adapted on the noise source. We then have for the output power $\mathbb{E}\{|z|^2(\omega)\}$:

$$\begin{aligned}\mathbb{E}\{|z|^2(\omega)\} = \boldsymbol{f}^H(\omega)\boldsymbol{R}_{uu}(\omega)\boldsymbol{f}(\omega) &= \boldsymbol{f}^H(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{f}(\omega) + \boldsymbol{f}^H(\omega)\boldsymbol{R}_{nn}(\omega)\boldsymbol{f}(\omega)\\ &= \boldsymbol{f}^H(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{f}(\omega) + \rho_{nn}^2(\omega)\boldsymbol{f}^H(\omega)\boldsymbol{W}^{-2}(\omega)\boldsymbol{f}(\omega)\\ &= \boldsymbol{f}^H(\omega)\boldsymbol{R}_{ss}(\omega)\boldsymbol{f}(\omega) + \rho_{nn}^2(\omega)\end{aligned}$$

[0258] Maximizing the output of the FSB corresponds to maximizing the speech-to-noise ratio in the output, not necessarily the speech output itself.

[0259] The optimal solutions after convergence for solutions C and B are related by $\boldsymbol{f}_C(\omega) = \mathbf{W}^2(\omega)\boldsymbol{f}_B(\omega)$.

[0260] This means that the outputs for both solutions will be equal.

[0261] Thus, the different configurations/solutions can be summarized by:

Solution	System output $z(\omega)$	Optimal coefficients $\boldsymbol{f}_{opt}(\omega)$
A	$a^*(\omega)\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}\,s(\omega)$	$\frac{a(\omega)\boldsymbol{W}(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}}$
B	$a^*(\omega)\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}\,s(\omega)$	$\frac{a(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}}$

(continued)

Solution	System output $z(\omega)$	Optimal coefficients $\boldsymbol{f}_{opt}(\omega)$
C	$a^*(\omega)\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{W}^2(\omega)\boldsymbol{h}(\omega)}\, s(\omega)$	$\frac{a(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\mathbf{W}^2(\omega)\boldsymbol{h}(\omega)}}$

[0262] As can be seen, by selecting the coefficients appropriately, the exact same (optimal) beamform output audio signal can be generated.

[0263] However, as can be seen, the coefficients that are selected for the set of spatial filters to achieve this output may be different.

[0264] The choice of the approach to use will depend on the preferences and requirements of the individual embodiment.

[0265] An advantage of solution A is that the interaction between the decorrelator and beamformer is reduced in the sense that the synergistic effect can be achieved with reduced modifications of the beamformer, feedback circuit, and beamform adapter.

[0266] An advantage of solution B is that if you want to calculate the Direction of Arrival (DOA) of an audio source using the beam former coefficients, then this is easier to do for this configuration.

[0267] For example, an approach for calculating a DOA is described in US 6774934 for a "normal" beamformer with a speech source with possibly uncorrelated equivariant noise added. For this use case the optimal coefficients are given by:

$$\boldsymbol{f}_{\mathrm{opt}}(\omega) = \frac{a(\omega)\boldsymbol{h}(\omega)}{\sqrt{\boldsymbol{h}^H(\omega)\boldsymbol{h}(\omega)}}$$

[0268] Part of the procedure is that the cross-power spectrum is calculated for at least one pair of filter coefficients, e.g. $f_1(\omega)$ and $f_2(\omega)$. This gives:

$$f_1(\omega)f_2^*(\omega) = \frac{h_1(\omega)h_2^*(\omega)}{\boldsymbol{h}^H(\omega)\boldsymbol{h}(\omega)}$$

where we use $a(\omega)a^*(\omega) = 1$, since a is an allpass.

[0269] Important is the phase difference, not the amplitude. This means that for solution B we can directly use the beamformer coefficients, but for methods A and C we must pre-multiply the coefficients with $\boldsymbol{W}^{-1}(\omega)$ and $\boldsymbol{W}^{-2}(\omega)$ respectively.

[0270] The audio apparatus(s) may specifically be implemented in one or more suitably programmed processors. The different functional blocks 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.

[0271] FIG. 6 is a block diagram illustrating an example processor 600 according to embodiments of the disclosure. Processor 600 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 600 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.

[0272] The processor 600 may include one or more cores 602. The core 602 may include one or more Arithmetic Logic Units (ALU) 604. In some embodiments, the core 602 may include a Floating Point Logic Unit (FPLU) 606 and/or a Digital Signal Processing Unit (DSPU) 608 in addition to or instead of the ALU 604.

[0273] The processor 600 may include one or more registers 612 communicatively coupled to the core 602. The registers 612 may be implemented using dedicated logic gate circuits (e.g., flip-flops) and/or any memory technology. In some embodiments the registers 612 may be implemented using static memory. The register may provide data, instructions and addresses to the core 602.

[0274] In some embodiments, processor 600 may include one or more levels of cache memory 610 communicatively coupled to the core 602. The cache memory 610 may provide computer-readable instructions to the core 602 for execution. The cache memory 610 may provide data for processing by the core 602. In some embodiments, the computer-readable instructions may have been provided to the cache memory 610 by a local memory, for example, local memory

attached to the external bus 616. The cache memory 610 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.

[0275] The processor 600 may include a controller 614, which may control input to the processor 600 from other processors and/or components included in a system and/or outputs from the processor 600 to other processors and/or components included in the system. Controller 614 may control the data paths in the ALU 604, FPLU 606 and/or DSPU 608. Controller 614 may be implemented as one or more state machines, data paths and/or dedicated control logic. The gates of controller 614 may be implemented as standalone gates, FPGA, ASIC or any other suitable technology.

[0276] The registers 612 and the cache 610 may communicate with controller 614 and core 602 via internal connections 620A, 620B, 620C and 620D. Internal connections may be implemented as a bus, multiplexer, crossbar switch, and/or any other suitable connection technology.

[0277] Inputs and outputs for the processor 600 may be provided via a bus 616, which may include one or more conductive lines. The bus 616 may be communicatively coupled to one or more components of processor 600, for example the controller 614, cache 610, and/or register 612. The bus 616 may be coupled to one or more components of the system.

[0278] The bus 616 may be coupled to one or more external memories. The external memories may include Read Only Memory (ROM) 632. ROM 632 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) 633. RAM 633 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) 635. The external memory may include Flash memory 634. The External memory may include a magnetic storage device such as disc 636. In some embodiments, the external memories may be included in a system.

[0279] It will be appreciated that the above description for clarity has described embodiments of the invention with reference to different functional circuits, units and processors. However, it will be apparent that any suitable distribution of functionality between different functional circuits, units or processors may be used without detracting from the invention. For example, functionality illustrated to be performed by separate processors or controllers may be performed by the same processor or controllers. Hence, references to specific functional units or circuits are only to be seen as references to suitable means for providing the described functionality rather than indicative of a strict logical or physical structure or organization.

[0280] 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.

[0281] 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.

[0282] 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 comprising:

a receiver (201) arranged to receive a first set of audio signals, the first set of audio signals comprising audio signals capturing audio of a scene from different positions;
an audio beamformer (101) comprising:

a first set of filters arranged to filter the first set of audio signals, and
a combiner arranged to combine outputs of the first set of filters to generate a beamform output audio signal;

a feedback circuit (103) comprising a second set of filters arranged to generate a second set of audio signals from a filtering of the beamform output audio signal, each filter of the second set of filters having a frequency response being a complex conjugate of a filter of the first set of filters;
a beamform adapter (105) arranged to adapt the first set of filters and the second set of filters in response to a comparison of the first set of audio signals and the second set of audio signals;
an adaptive coefficient processor (207) arranged to determine decorrelation coefficients for a set of spatial decorrelation filters generating decorrelated output signals from the first set of audio signals, the adaptive coefficient processor being arranged to adapt the decorrelation coefficients in response to an update value determined from the first set of audio signals;
a first set of spatial filters (205, 301, 401) arranged to apply a first spatial filtering to at least one of the first set of audio signals and the second set of audio signals, the first set of spatial filters (205, 301, 401) having coefficients determined from the decorrelation coefficients.

2. The audio apparatus of claim 1 further comprising an audio detector (501) arranged to determine a set of active time intervals during which an audio source is active and a set of inactive time intervals during which the audio source is not active; and wherein at least one of the adaption of the first set of filters and the second set of filters and the adaptation the decorrelation coefficients is different for the set of active time intervals and the set of inactive time intervals.

3. The audio apparatus of claim 2 wherein the beamform adapter (105) is arranged to adapt the first set of filters and the second set of filters with a higher rate of adaptation during the set of active time intervals than during the set of inactive time intervals.

4. The audio apparatus of claim 3 wherein the beamform adapter (105) is arranged to adapt the first set of filters and the second set of filters during only one set of time intervals of the set of active time intervals and the set of inactive time intervals.

5. The audio apparatus of any of claims 2 to 4 wherein the adaptive coefficient processor (207) is arranged to adapt the decorrelation coefficients with a higher rate of adaptation during the set of inactive time intervals than during the set of active time intervals.

6. The audio apparatus of claim 5 wherein the adaptive coefficient processor (207) is arranged to adapt the decorrelation coefficients during only one set of time intervals of the set of inactive time intervals and the set of active time intervals.

7. The audio apparatus of any previous claim wherein the first set of spatial filters is arranged to filter the first set of audio signals.

8. The audio apparatus of claim 7 wherein the first set of spatial filters is arranged to have coefficients set to the decorrelation coefficients determined for the set of spatial decorrelation filters.

9. The audio apparatus of claim 1 to 6 wherein the first set of filters for filtering is arranged to filter the first set of audio signals after filtering by the first set of spatial filters and the beamform adapter (105) is arranged to perform the comparison using the first set of audio signals before filtering by the first set of spatial filters.

10. The audio apparatus of claim 9 wherein the first set of spatial filters is arranged to have coefficients matching coefficients of a spatial filter being a cascade of two of the set of decorrelation filters.

11. The audio apparatus of any of claim 1 to 6 wherein the first set of spatial filters is arranged to filter the second set of audio signals.

12. The audio apparatus of claim 11 wherein the first set of spatial filters is arranged to have coefficients determined in response to a set of inverse spatial decorrelation filters, the set of inverse spatial decorrelation filters being inverse filters of the set of spatial decorrelation filters.

13. The audio apparatus of claim 11 wherein the first set of spatial filters is arranged to have coefficients matching coefficients of a spatial filter being a cascade of two sets of spatial inverse filters each of which comprises inverse filters

of the set of spatial decorrelation filters.

14. The audio apparatus of any previous claim wherein the adaptive coefficient processor (207) is arranged to determine the set of spatial decorrelation filters to generate a set of output audio signals, each output audio signal of the set of the set of output signals being linked with one input audio signal of the first set of audio signals, by performing the steps of: segmenting the first set of audio signals into time segments, and for at least some time segments performing the steps of:

generating a frequency bin representation of the first set of audio signals, each frequency bin of the frequency bin representation of the first set of audio signals comprising a frequency bin value for each of the audio signals of the first set of audio signals;
generating a frequency bin representation of a set of output signals, each frequency bin of the frequency bin representation of a set of output signals comprising a frequency bin value for each of the output signals, the frequency bin value for a given output signal of the set of output signals for a given frequency bin being generated as a weighted combination of frequency bin values of the first set of audio signals for the given frequency bin, the weighted combination having the decorrelation coefficients as weights;
updating a first weight for a contribution to a first frequency bin value of a first frequency bin for a first output signal linked with a first input audio signal from a second frequency bin value of the first frequency bin for a second input audio signal linked to a second output signal in response to a correlation measure between a first previous frequency bin value of the first output signal for the first frequency bin and a second previous frequency bin value of the second output signal for the first frequency bin.

15. The audio apparatus of claim 14 wherein the adaptive coefficient processor (207) is arranged to update the first weight in response to a product of a first value and a second value, the first value being one of the first previous frequency bin value and the second previous frequency bin value and the second value being a complex conjugate of the other of the first previous frequency bin value and the second previous frequency bin value.

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

receiving a first set of audio signals, the first set of audio signals comprising audio signals capturing audio of a scene from different positions;
an audio beamformer (101) generating an output signal by:

a first set of filters filtering the first set of audio signals, and
a combiner combining outputs of the first set of filters to generate an output audio signal;

a second set of filters generating a second set of audio signals from a filtering of the output audio signal, each filter of the second set of filters having a frequency response being a complex conjugate of a filter of the first set of filters;
adapting the first set of filters and the second set of filters in response to a comparison of the first set of audio signals and the second set of audio signals;
determining decorrelation coefficients for a set of spatial decorrelation filters generating decorrelated output signals from the first set of audio signals including adapting the decorrelation coefficients in response to an update value determined from the first set of audio signals;
a first set of spatial filters (205, 301, 401) applying a first spatial filtering to at least one of the first set of audio signals and the second set of audio signals, the first set of spatial filters (205, 301, 401) having coefficients determined from the decorrelation coefficients.

17. A computer program product comprising computer program code means adapted to perform all the steps of claim 16 when said program is run on a computer.

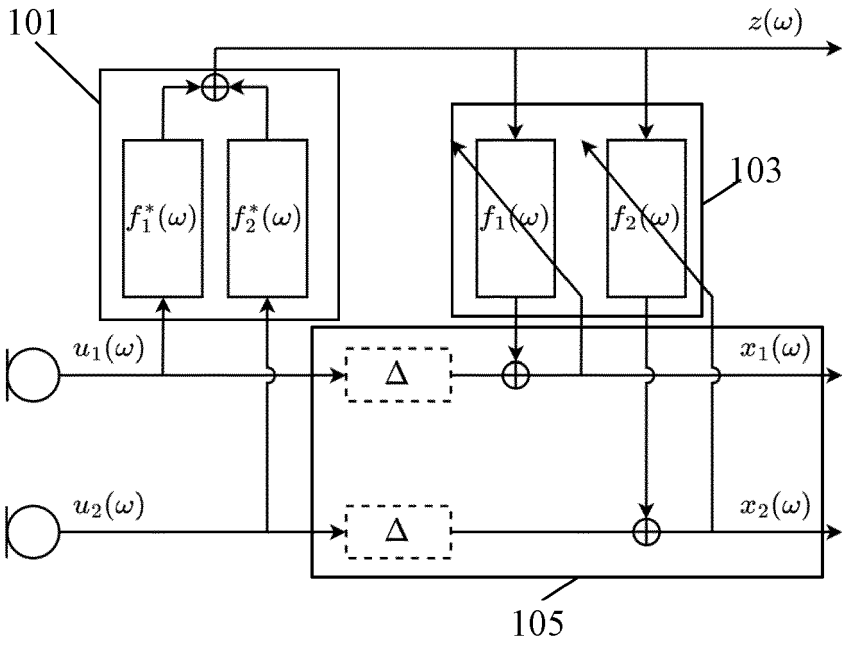
101
$z(\omega)$
$f_1^*(\omega)$
$f_2^*(\omega)$
$f_1(\omega)$
$f_2(\omega)$
103
$u_1(\omega)$
Δ
$x_1(\omega)$
$u_2(\omega)$
Δ
$x_2(\omega)$
105

FIG. 1

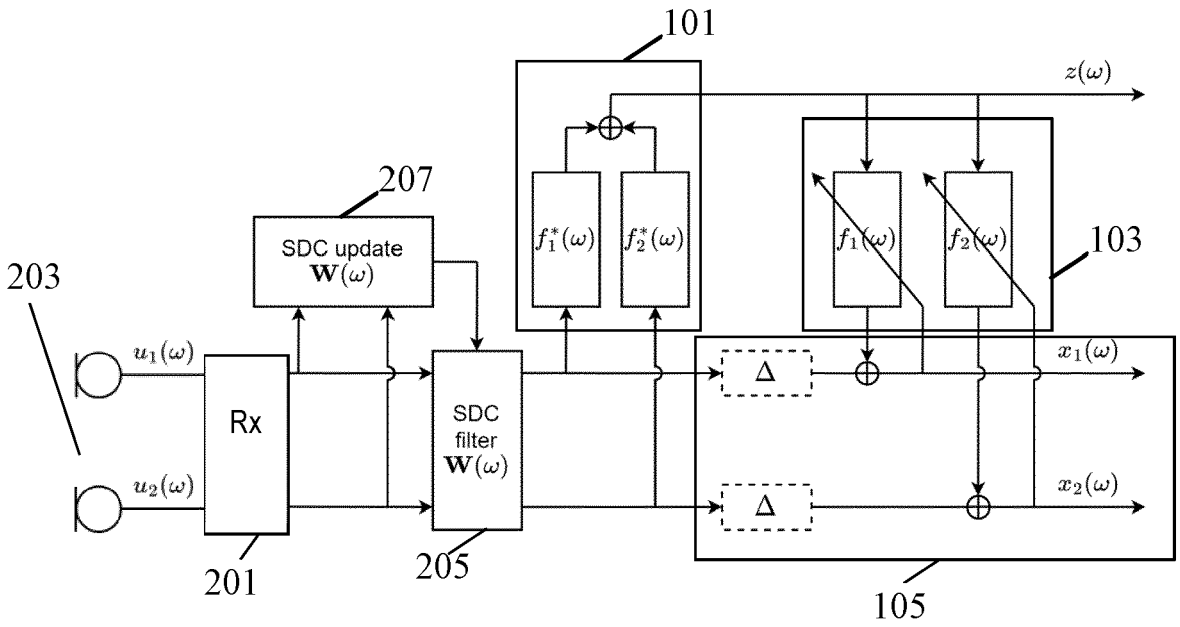
101
$z(\omega)$
207
SDC update
$\mathbf{W}(\omega)$
$f_1^*(\omega)$
$f_2^*(\omega)$
$f_1(\omega)$
$f_2(\omega)$
103
203
$u_1(\omega)$
Rx
SDC
filter
$\mathbf{W}(\omega)$
Δ
$x_1(\omega)$
$u_2(\omega)$
Δ
$x_2(\omega)$
201
205
105

FIG. 2

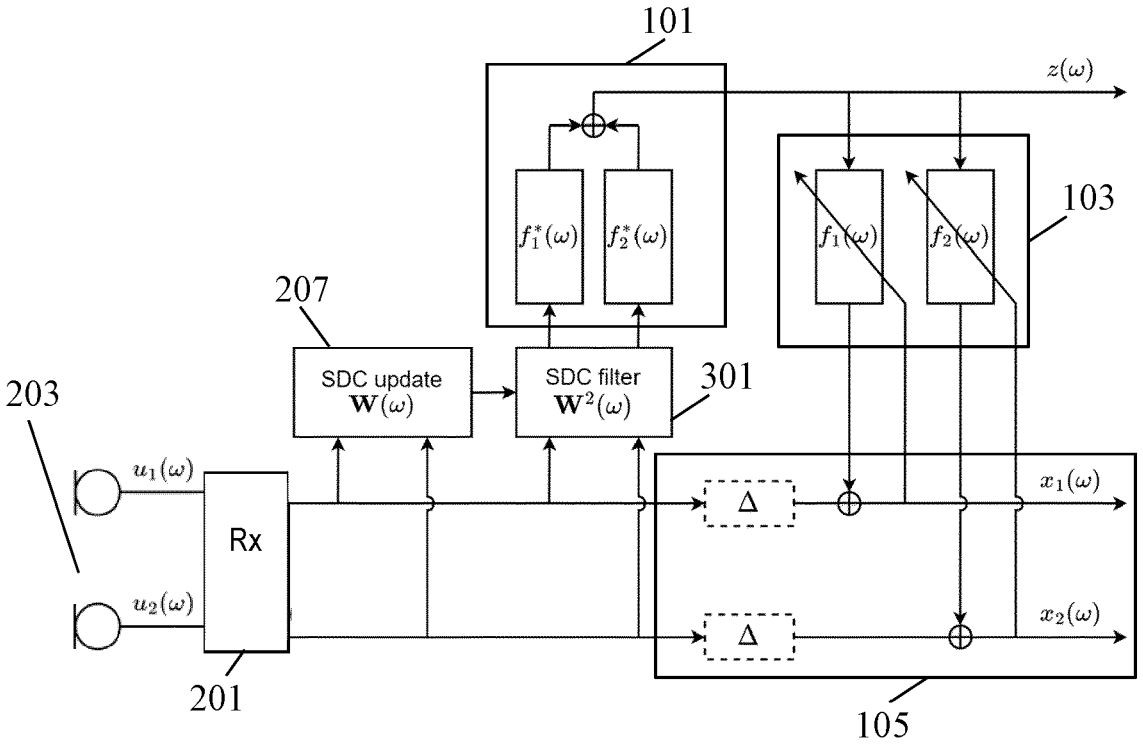
101
$z(\omega)$
$f_1^*(\omega)$
$f_2^*(\omega)$
$f_1(\omega)$
$f_2(\omega)$
103
207
SDC update
$\mathbf{W}(\omega)$
SDC filter
$\mathbf{W}^2(\omega)$
301
203
$u_1(\omega)$
$u_2(\omega)$
Rx
201
Δ
Δ
$x_1(\omega)$
$x_2(\omega)$
105

FIG. 3

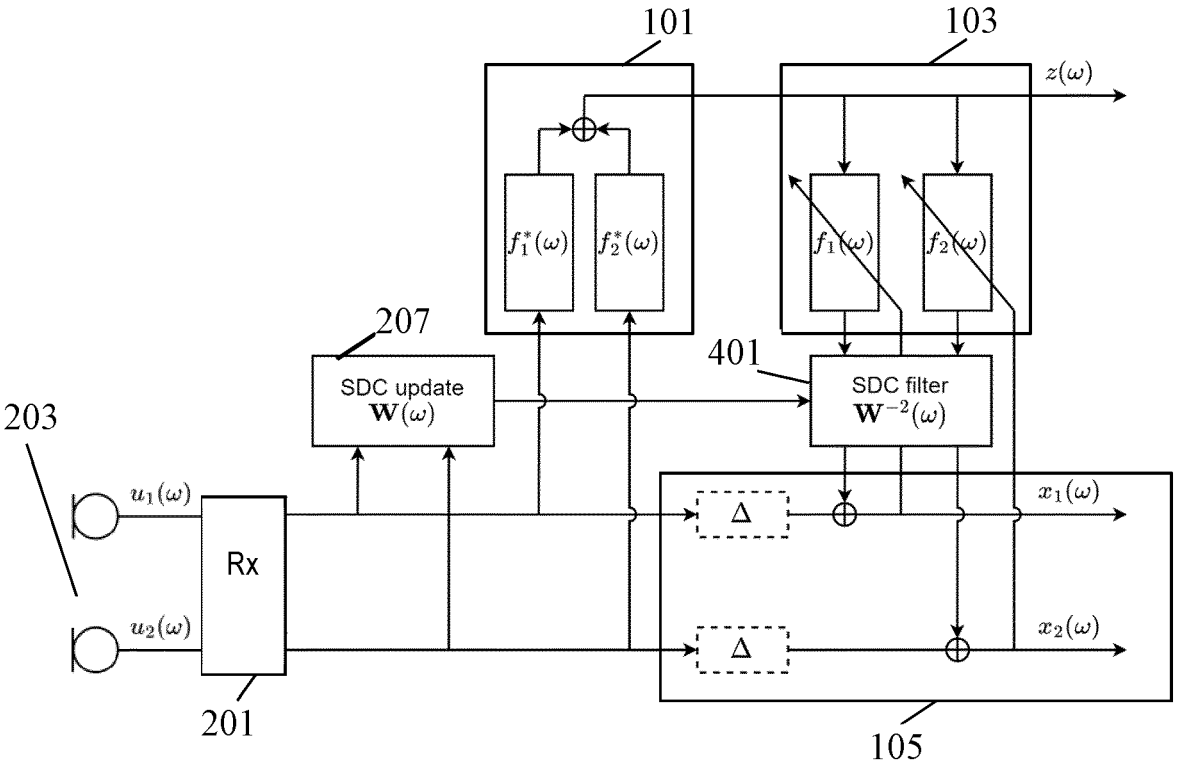
101
103
$z(\omega)$
$f_1^*(\omega)$
$f_2^*(\omega)$
$f_1(\omega)$
$f_2(\omega)$
207
401
SDC update
$\mathbf{W}(\omega)$
SDC filter
$\mathbf{W}^{-2}(\omega)$
203
$u_1(\omega)$
$u_2(\omega)$
Rx
Δ
Δ
$x_1(\omega)$
$x_2(\omega)$
201
105

FIG. 4

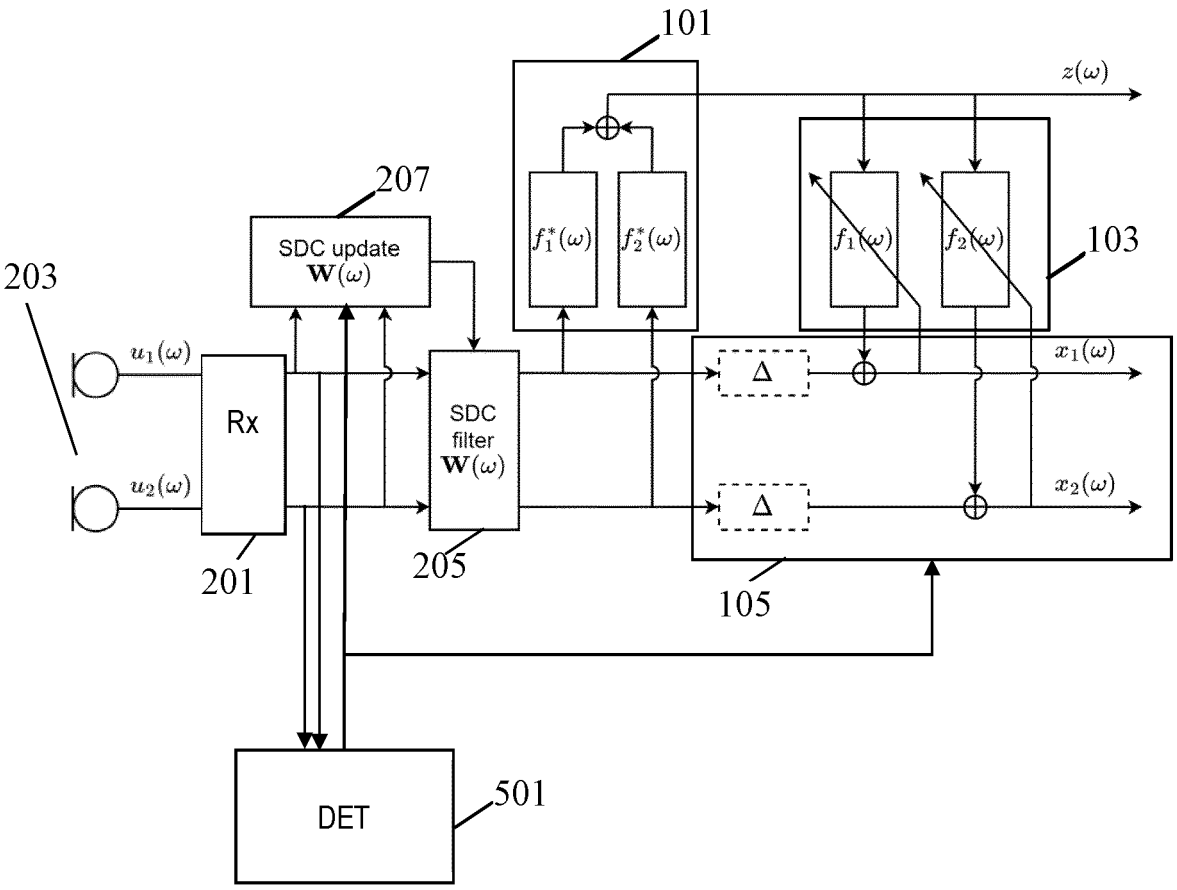
101
$z(\omega)$
207
SDC update
$\mathbf{W}(\omega)$
$f_1^*(\omega)$
$f_2^*(\omega)$
$f_1(\omega)$
$f_2(\omega)$
103
203
$u_1(\omega)$
$u_2(\omega)$
Rx
SDC
filter
$\mathbf{W}(\omega)$
Δ
Δ
$x_1(\omega)$
$x_2(\omega)$
201
205
105
DET
501

FIG. 5

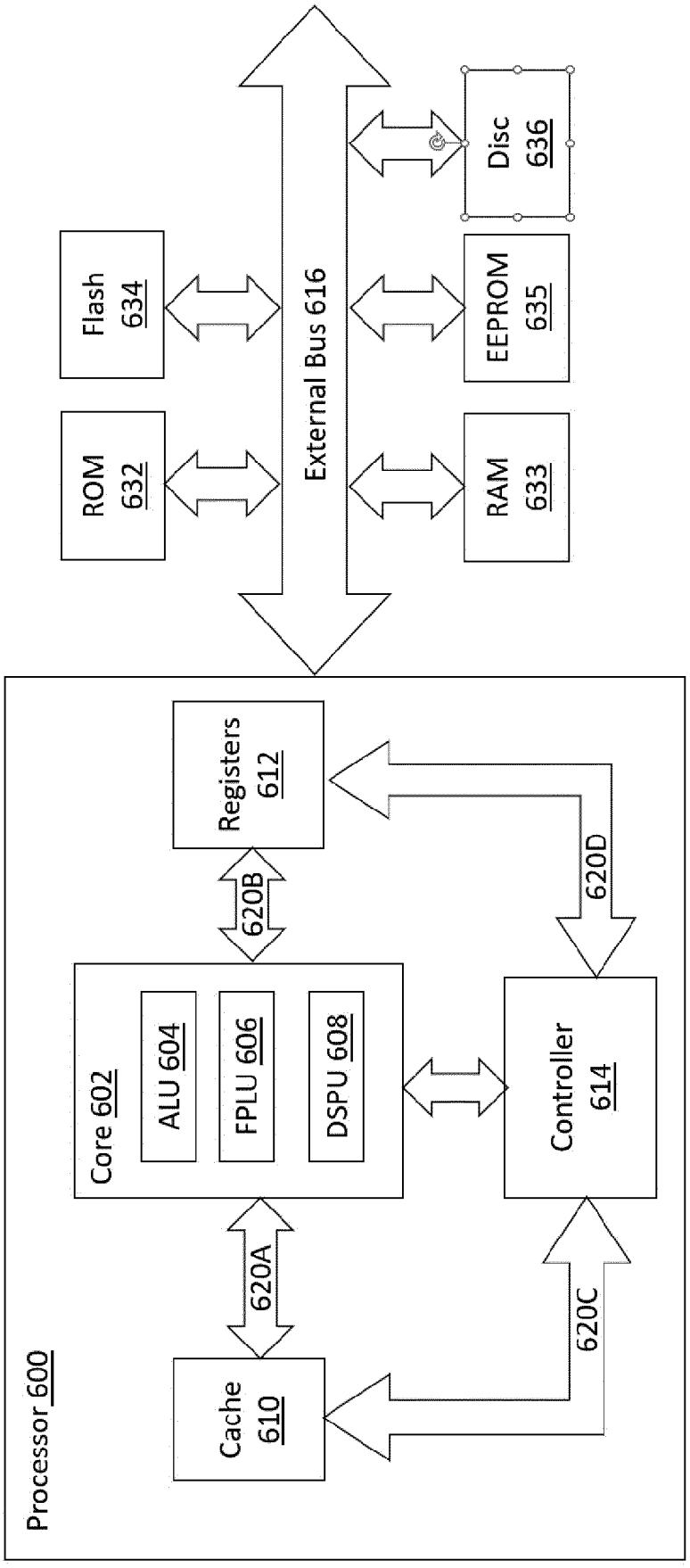
Processor 600
Cache 610
620A
Core 602
ALU 604
FPLU 606
DSPU 608
620B
Registers 612
620C
620D
Controller 614
External Bus 616
ROM 632
Flash 634
RAM 633
EEPROM 635
Disc 636

FIG. 6

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 23 17 6014

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	EP 1 923 866 A1 (ASAHI CHEMICAL IND [JP]) 21 May 2008 (2008-05-21)	1,7,16	INV. G10L21/0272
Y	* abstract; figures 1-3 *	2-6	
A	* paragraphs [0050] - [0054] * * paragraphs [0007] - [0010] * * paragraphs [0022] - [0026] *	8-15,17	
X	WO 99/27522 A2 (KONINKL PHILIPS ELECTRONICS NV [NL]; PHILIPS SVENSKA AB [SE]) 3 June 1999 (1999-06-03)	1,7,16	
Y	* abstract; figures 1-5 *	2-6	
A	* page 4 * * pages 9-12 *	8-15,17	
Y	WO 2012/109384 A1 (DOLBY LAB LICENSING CORP [US]; DICKINS GLENN N [AU] ET AL.) 16 August 2012 (2012-08-16)	2-6	
A	* abstract; figures 1-4 * * paragraphs [0239] - [0243] *	1,7-17	
A	INES HAFIZOVIC ET AL: "Decorrelation for adaptive beamforming applied to arbitrarily sampled spherical microphone arrays", APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS (WASPAA), 2011 IEEE WORKSHOP ON, IEEE, 16 October 2011 (2011-10-16), pages 233-236, XP032011509, DOI: 10.1109/ASPAA.2011.6082300 ISBN: 978-1-4577-0692-9 * abstract; figure 1 * * columns 1-2 *	1-17	TECHNICAL FIELDS SEARCHED (IPC) G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	24 November 2023	Képesi, Marián

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 17 6014

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.

24-11-2023

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 1923866 A1	21-05-2008	CN 101238511 A	06-08-2008
		EP 1923866 A1	21-05-2008
		JP 4225430 B2	18-02-2009
		JP WO2007018293 A1	19-02-2009
		KR 20080009211 A	25-01-2008
		US 2009055170 A1	26-02-2009
		WO 2007018293 A1	15-02-2007
WO 9927522 A2	03-06-1999	CN 1251192 A	19-04-2000
		DE 69822128 T2	20-01-2005
		EP 0954850 A2	10-11-1999
		JP 4372232 B2	25-11-2009
		JP 2001510001 A	24-07-2001
		KR 20000070387 A	25-11-2000
		US 7146012 B1	05-12-2006
		US 7454023 B1	18-11-2008
		WO 9927522 A2	03-06-1999
WO 2012109384 A1	16-08-2012	CN 103348408 A	09-10-2013
		CN 103354937 A	16-10-2013
		EP 2673777 A1	18-12-2013
		EP 2673778 A1	18-12-2013
		JP 6002690 B2	05-10-2016
		JP 2014510452 A	24-04-2014
		WO 2012109384 A1	16-08-2012
		WO 2012109385 A1	16-08-2012

EPO FORM P0459

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

REFERENCES CITED IN THE DESCRIPTION

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

Patent documents cited in the description

- US 7146012 B **[0091]**
- US 6774934 B **[0267]**