(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 October 2016 (27.10.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/169591 A1

(51) International Patent Classification:
H04S 7/00 (2006.01) *H04M 3/56* (2006.01)

(21) International Application Number:
PCT/EP2015/058694

(22) International Filing Date:
22 April 2015 (22.04.2015)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HUAWEI TECHNOLOGIES CO., LTD.** [CN/CN]; Huawei Administration Building Bantian, Longgang District, Shenzhen, Guangdong 518129 (CN).

(72) Inventors; and
(71) Applicants *(for US only)*: **PANG, Liyun** [CN/DE]; c/o Huawei Technologies Duesseldorf GmbH Riesstr.25, 80992 Munich (DE). **HOFFMANN, Pablo** [CL/DE]; c/o Huawei Technologies Duesseldorf GmbH, Riesstr.25, 80992 Munich (DE).

(74) Agent: **KREUZ, Georg**; c/o Huawei Technologies Duesseldorf GmbH, Riesstr. 8, 80992 Munich (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: AN AUDIO SIGNAL PROCESSING APPARATUS AND METHOD

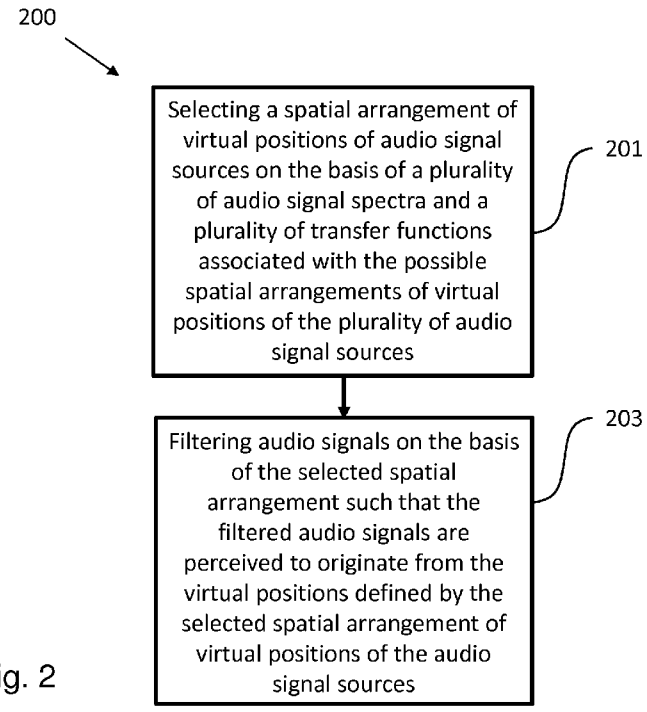

Fig. 2

(57) Abstract: The invention relates to an audio signal processing apparatus (100) for processing a plurality of audio signals (105) defining a plurality of audio signal spectra, the audio signals to be transmitted to a listener in such a way that the listener perceives the audio signals to originate from virtual positions of a plurality of audio signal sources. The audio signal processing apparatus comprises a selector (101) configured to select a spatial arrangement of the virtual positions of the audio signal sources relative to the listener from a plurality of possible spatial arrangements, and a filter (103) configured to filter the plurality of audio signals on the basis of the selected spatial arrangement.

WO 2016/169591 A1

Published:

— *with international search report (Art. 21(3))*

AN AUDIO SIGNAL PROCESSING APPARATUS AND METHOD

TECHNICAL FIELD

The present invention relates to an audio signal processing apparatus and method. In particular, the present invention relates to an audio signal processing apparatus and method for a virtual spatial audio conference system.

BACKGROUND

In the past voices of speakers in a multi-party audio conference system typically have been rendered to the listeners as a monaural audio stream - essentially overlaid on top of each other and usually presented to the listener "within the head" when headphones are used.

A virtual spatial audio conference system, which is a special form of a multiparty telemeeting as defined by the ITU-T recommendation P.1301 "Subjective quality evaluation of audio and audiovisual multiparty telemeetings", enables a 3D audio rendering of the voices of the participants. That is, the participants' voices are placed at different "virtual" locations in space by using spatial filters derived from head-related impulse responses (HRIR) or their corresponding frequency-domain representations, i.e. head-related transfer functions (HRTFs), and/or binaural room impulse responses (BRIR) or their corresponding frequency-domain representations, i.e. binaural room transfer functions (BRTF). These filters encode the auditory cues humans use for spatial sound perception, namely interaural time difference (ITD), interaural level difference (ILD), spectral cues, and also room acoustic information, such as reverberation in the case of BRIRs. The beneficial effect of 3D audio rendering relative to a monaural audio stream of the voices of the participants is not only that the conference experience is more natural, but that also speech intelligibility is substantially enhanced. It has been shown that this psychoacoustic effect, scientifically known as spatial release from masking, can improve speech intelligibility by up to 12-13 dB when a target speaker and competing speakers, typically referred to as maskers, are (virtually) spatially separated.

US7391877 describes a spatial sound processor that virtually distributes speakers over non-equidistant positions along a circle centered at the listener’s position. Based on results from psychoacoustic tests on speech identification the system starts with a

relatively small virtual spatial separation for speakers placed in front of the listener. The virtual spatial separation between speakers is then increased as speakers are placed at more lateral positions. For directions ±90 degrees in azimuth, two virtual speaker locations are proposed, one in the far-field and one in the near-field. Similar solutions based on either equidistant or non-equidistant speakers are described in WO2013/142641 and WO2013/142668.

There have been some attempts to use the information contained in the voice signals themselves to enhance speech intelligibility. These attempts, i.e. the use of voice information to separate maskers from speakers, relies heavily on the amount of spectral overlap that exists between a target speaker and maskers, i.e. energetic masking. Ideal time-frequency binary masks have been proposed, for instance in Brungart et al "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation", J. Acoust. Soc. Am., volume 120, no. 6, 2006, in order to remove time-frequency regions where masker(s) energy dominates and preserve only those time-frequency regions where the energy of the target's voice dominates. They are ideal because access to the clean (original) speech signals from target speaker and masker(s) speaker(s) is required. More specifically, a priori knowledge about the target speaker and masker speakers is required so that those time-frequency regions of the acoustic mixture dominated by the target speaker can be preserved. In practice, however, sometimes the target speaker is not known a priori or variable. In a virtual spatial audio conference, for instance, each participant can be the target speaker for a certain period of time.

Thus, there is a need for an improved audio signal processing apparatus and method, in particular an audio signal processing apparatus and method improving speech intelligibility in a virtual spatial audio conference system.

SUMMARY

It is an objective of the invention to provide an audio signal processing apparatus and method improving speech intelligibility in a virtual spatial audio conference system.

This objective is achieved by the subject matter of the independent claims. Further implementation forms are provided in the dependent claims, the description and the figures.

According to a first aspect the invention relates to an audio signal processing apparatus for processing a plurality of audio signals defining a plurality of audio signal spectra, the plurality of audio signals to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals to originate from virtual positions of a plurality of audio signal sources. The audio signal processing apparatus comprises a selector configured to select a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, wherein each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions, and wherein the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources, and a filter configured to filter the plurality of audio signals on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener, wherein the plurality of filtered audio signals are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

Thus, an audio signal processing apparatus is provided allowing improving, for instance, the speech intelligibility in a virtual spatial audio conference system using both voice (i.e. audio signal spectra) and directional (i.e. transfer functions) information for selecting an improved spatial arrangement.

The plurality of audio signals can comprise N audio signals and the virtual positions of the plurality of audio signal sources can comprise L virtual positions. The transfer functions can be head related transfer functions (HRTFs) or binaural room transfer functions (BRTFs).

In a first possible implementation form of the first aspect of the invention, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources to obtain a plurality of directional-speaker spectral profiles associated with each possible spatial arrangement of the virtual positions

of the plurality of audio signal sources and to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of directional-speaker spectral profiles.

In this implementation form voice and directional information is combined into directional-speaker spectral profiles for selecting an improved spatial arrangement.

In a second possible implementation form of the first possible implementation form of the first aspect of the invention, the selector is configured to combine the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources to obtain a plurality of directional-speaker spectral profiles associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by multiplying the plurality of audio signal spectra by the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources.

This implementation form provides a computationally efficient form for combining voice and directional information into a directional-speaker spectral profile by multiplying the spectra.

In a third possible implementation form of the first or second implementation form of the first aspect of the invention, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which a spectral difference between the plurality of directional-speaker spectral profiles is larger than a predefined threshold value, preferably a maximum.

This implementation form provides for a good speech intelligibility using the spectral difference to determine advantageous spatial arrangements. On the basis of the spectral difference this implementation form allows determining the optimal spatial arrangement.

In a fourth possible implementation form of the third implementation form of the first aspect of the invention, the selector is configured to determine the spectral difference between the directional-speaker spectral profiles associated with the m-th spatial

arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$S_m = \frac{1}{K}\sum_{k=1}^{K} w_k \sigma_{m,k},$$

$$\sigma_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(Y_{n,m,k} - \bar{\mathbf{Y}}_{m,k})^2, \text{ and}$$

$$Y_{n,m,k} = X_{n,k} H_{m,k},$$

wherein S_m denotes a scalar value representing the spectral difference between the plurality of directional-speaker spectral profiles associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, K denotes the total number of frequency bands, w_k denotes a weighting factor, $\sigma_{m,k}$ denotes the variance across the directional-speaker spectral profiles for the k-th frequency band, N denotes the total number of audio signal spectra, $Y_{n,m,k}$ denotes the value of the n-th directional-speaker spectral profile in the k-th frequency band, $\bar{\mathbf{Y}}_{m,k}$ denotes the mean of the directional speaker profiles in the k-th frequency band, $X_{n,k}$ denotes the value of the audio signal spectrum of the n-th audio signal in the k-th frequency band and $H_{m,k}$ denotes the value of the transfer function associated with the virtual position of the audio signal source associated with the n-th audio signal in the k-th frequency band.

In a fifth possible implementation form of the fourth implementation form of the first aspect of the invention, the selector is configured to determine the value of the audio signal spectrum of the n-th audio signal in the k-th frequency band, i.e. $X_{n,k}$, and/or the value of the transfer function associated with the virtual position of the audio signal source associated with the n-th audio signal in the k-th frequency band, i.e. $H_{m,k}$, by performing an averaging operation over a plurality of frequency bins (used for a discrete Fourier transform) on the basis of the following equations:

$$X_{n,k} = \frac{1}{J(k)}\sum_{i=j(k)}^{j(k+1)-1} |\mathcal{X}(i)|, \text{ and}$$

$$H_{m,k} = \frac{1}{J(k)}\sum_{i=j(k)}^{j(k+1)-1} |\mathcal{H}(i)|,$$

wherein $\mathcal{X}(i)$ denotes the value of the discrete Fourier transform of the n-th audio signal in the i-th frequency bin, $\mathcal{H}(i)$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual position of the audio

signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

In a sixth possible implementation form of the third to fifth implementation form of the first aspect of the invention, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and a plurality of left ear transfer functions associated with the virtual positions of the audio signal sources relative to the left ear of the listener to obtain a plurality of left ear directional-speaker spectral profiles and the plurality of audio signal spectra and a plurality of right ear transfer functions associated with the virtual positions of the audio signal sources relative to the right ear of the listener to obtain a plurality of right ear directional-speaker spectral profiles and by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which a spectral difference between the left ear directional-speaker spectral profiles and the right ear directional-speaker spectral profiles is smaller than a predefined threshold, in particular a minimum.

In a seventh possible implementation form of the first aspect of the invention as such, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources from the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, the virtual positions of the plurality of audio signal sources being arranged on a circle centered at the listener and having a constant angular separation on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by determining one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which the spectral difference between the plurality of transfer functions is larger than a predefined threshold value, preferably a maximum.

In an eighth possible implementation form of the seventh implementation form of the first aspect of the invention, the selector is configured to determine the spectral difference between the transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$\hat{S}_m = \frac{1}{K}\sum_{k=1}^{K} w_k \hat{\sigma}_{m,k}, \text{ and}$$

$$\hat{\sigma}_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(H_{n,m,k} - \bar{\mathbf{H}}_{m,k})^2,$$

wherein $\hat{S}_m$ denotes a scalar value representing the spectral difference between the plurality of transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, K denotes the total number of frequency bands, w_k denotes a weighting factor, $\hat{\sigma}_{m,k}$ denotes the variance across the plurality of transfer functions for the k-th frequency band, N denotes the total number of audio signal spectra, $H_{n,m,k}$ denotes the value of the n-th transfer function in the k-th frequency band, and $\bar{\mathbf{H}}_{m,k}$ denotes the mean of the transfer functions in the k-th frequency band.

In a ninth possible implementation form of the seventh or eighth implementation form of the first aspect of the invention, wherein the selector is configured to determine the value of the n-th transfer function in the k-th frequency band, i.e. $H_{n,m,k}$, is determined by performing an averaging operation over a plurality of frequency bins used for a discrete Fourier transform on the basis of the following equation:

$$H_{n,m,k} = \frac{1}{J(k)}\sum_{i=j(k)}^{j(k+1)-1}|\mathcal{H}_n(i)|,$$

wherein $\mathcal{H}_n$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual position of the audio signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

In a tenth possible implementation form of the seventh or eighth implementation form of the first aspect of the invention, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by ranking the plurality of audio signal spectra according to a similarity value of the plurality of audio signal spectra.

In an eleventh possible implementation form of the tenth implementation form of the first aspect of the invention, the selector is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial

arrangement of the virtual positions of the plurality of audio signal sources by assigning the ranked plurality of audio signal spectra to the virtual positions of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources in such a way that the angular separation between audio signal spectra having a large similarity value is maximized.

In a twelfth possible implementation form of the tenth or eleventh implementation form of the first aspect of the invention, the selector is configured to compute the similarity value for the plurality of audio signal spectra by (i) computing an average audio signal spectrum and the spectral differences between each audio signal spectrum and the average audio signal spectrum or (ii) by computing the correlation functions between the audio signal spectra.

According to a second aspect the invention relates to a signal processing method for processing a plurality of audio signals defining a plurality of audio signal spectra, the plurality of audio signals to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals to originate from virtual positions of a plurality of audio signal sources. The audio signal processing method comprises a step of selecting a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, wherein each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions, wherein the spatial arrangement of the virtual positions of the plurality of audio signal sources is selected on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources, and the step of filtering the plurality of audio signals on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener, wherein the plurality of filtered audio signals are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

The audio signal processing method according to the second aspect of the invention can be performed by the audio signal processing apparatus according to the first aspect of the invention. Further features of the audio signal processing method according to the second

aspect of the invention result directly from the functionality of the audio signal processing apparatus according to the first aspect of the invention and its different implementation forms.

According to a third aspect the invention relates to a computer program comprising program code for performing the method according to the second aspect of the invention when executed on a computer.

The invention can be implemented in hardware and/or software.

BRIEF DESCRIPTION OF THE DRAWINGS

Further embodiments of the invention will be described with respect to the following figures, in which:

Fig. 1 shows a schematic diagram of an audio signal processing apparatus according to an embodiment;

Fig. 2 shows a schematic diagram of an audio signal processing method according to an embodiment;

Fig. 3 shows exemplary left, right and average binaural room transfer functions that can be used with an audio signal processing apparatus and method according to an embodiment;

Fig. 4 shows an exemplary audio signal spectrum that can be used with an audio signal processing apparatus and method according to an embodiment;

Fig. 5 shows an exemplary directional-speaker spectral profile that can be obtained and used with an audio signal processing apparatus and method according to an embodiment;

Fig. 6A shows exemplary directional-speaker spectral profiles for the case of five speakers that can be used with an audio signal processing apparatus and method according to an embodiment;

Fig. 6B shows the variance of the exemplary directional-speaker spectral profiles of figure 6A;

Fig. 6C shows exemplary weighting factors used to integrate human hearing sensitivity in an audio signal processing apparatus and method according to an embodiment;

Fig. 7 shows four exemplary spatial arrangements of virtual positions of a plurality of audio signal sources relative to a listener according to an embodiment; and

Figs. 8A and 8B illustrate how to select the optimal spatial arrangement of virtual positions of a plurality of audio signal sources relative to a listener according to an embodiment.

DETAILED DESCRIPTION OF EMBODIMENTS

In the following detailed description, reference is made to the accompanying drawings, which form a part of the disclosure, and in which are shown, by way of illustration, specific aspects in which the disclosure may be practiced. It is understood that other aspects may be utilized and structural or logical changes may be made without departing from the scope of the present disclosure. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims.

It is understood that a disclosure in connection with a described method may also hold true for a corresponding device or system configured to perform the method and vice versa. For example, if a specific method step is described, a corresponding device or apparatus may include a unit to perform the described method step, even if such unit is not explicitly described or illustrated in the figures. Further, it is understood that the features of the various exemplary aspects described herein may be combined with each other, unless specifically noted otherwise.

Figure 1 shows a schematic diagram of an audio signal processing apparatus 100 according to an embodiment. The audio signal processing apparatus 100 is configured to process a plurality of audio signals 105 defining a plurality of audio signal spectra. The plurality of audio signals 105 are to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals to originate from virtual positions of a plurality of audio signal sources. In an embodiment, the audio signal processing apparatus

is part of a virtual spatial audio conference system and the audio signals are the voice signals of the participants of the virtual spatial audio conference.

The audio signal processing apparatus 100 comprises a selector 101 configured to select a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener.

Each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions, in particular head-related transfer functions (HTRF) and/or binaural room transfer functions (BTRF). As known to the person skilled in the art, there is a direct correspondence between the HTRFs/BTRFs and their impulse responses, namely the head-related impulse responses (HRIRs) and the binaural room impulse responses (BRIRs).

Moreover, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources.

The term "virtual position" is well known to the person skilled in the art of audio processing. By choosing suitable transfer functions the position, a listener perceives to receive an audio signal emitted by an (virtual) audio signal source. This position is the "virtual position" used herein, and may include techniques in which sources/speakers presented over headphones appear to originate from any desired direction (i.e., a virtual position) in space.

The audio signal processing apparatus 100 further comprises a filter 103 configured to filter the plurality of audio signals 105 on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener and to produce a plurality of filtered audio signals 107. The plurality of filtered audio signals 107 are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

Figure 2 shows a schematic diagram of an embodiment of an audio signal processing method 200 for processing a plurality of audio signals 105 defining a plurality of audio signal spectra, the plurality of audio signals to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals to originate from virtual positions of a plurality of audio signal sources.

The audio signal processing method 200 comprises a step 201 of selecting a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, wherein each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions. The spatial arrangement of the virtual positions of the plurality of audio signal sources is selected on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources.

Moreover, the audio signal processing apparatus 200 comprises a step 203 of filtering the plurality of audio signals 105 on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener to obtain a plurality of filtered audio signals 107. The plurality of filtered audio signals 107 are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

The audio signal processing method 200 can be performed, for instance, by the audio signal processing apparatus 100 according to the first aspect of the invention.

In the following, further implementation forms and embodiments of the audio signal processing apparatus 100 and the audio signal processing method 200 are described.

In an embodiment, the selector 101 of the audio signal processing apparatus 100 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources. In an embodiment, the plurality of audio signal spectra and the plurality of transfer functions are combined by multiplying the

plurality of audio signal spectra and the plurality of transfer functions to obtain a plurality of directional-speaker spectral profiles associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources.

Figure 3 shows an exemplary transfer function obtained by deriving the average of a left BRTF and a right BRTF. For deriving the exemplary transfer function the left BRTF and the right BRTF are averaged in respective frequency bands. This subband analysis can be done in various ways, such as using quadrature mirror filters (QMF), gammatone filters, or octave or third-octave bands. For the example shown in figure 3 the spectra, profiles and transfer functions are computed using a sixth-octave-band analysis, i.e. 1/n-octave bands with n=6 representing the bandwidth of the filter bank. The analysis approximates a constant-Q filter bank by averaging across magnitude bins of a Discrete Fourier Transform (DFT) which is computed using the Fast Fourier Transform (FFT) algorithm. A constant-Q filter bank means that the ratio between the center frequency and bandwidth of the filter remains the same across filters. In an embodiment, the subband analysis is performed over a frequency range relevant for speech and is set to frequencies between 500 and 6300 Hz. This frequency range results in a subband analysis with a total of 21 different 1/6-octave bands. Other options for the upper frequency limit may be 7000 or 8000 Hz.

The person skilled in the art will appreciate that taking the average between left and right HRTF is just one approach to derive a transfer function that can be used in the context of the audio signal processing apparatus 100 and the audio signal processing method 200. For example, either the left or right HRTF/BRTF can be used as the transfer function. The transfer functions, for instance, the HRTF and/or the BRTF, can be computed once and stored for posterior use.

Figure 4 shows an exemplary audio signal spectrum that can be used with the audio signal processing apparatus 100 and method 200 according to an embodiment. The thin line in figure 4 shows the discrete Fourier transform of an exemplary speech audio signal, i.e. an exemplary audio signal spectrum. The thick line in figure 4 shows an averaged or subband representation of the audio signal spectrum that is used, in an embodiment, for computational purposes.

In an embodiment, the value of the audio signal spectrum of the n-th audio signal in the k-th frequency band, i.e. $X_{n,k}$, and/or the value of the transfer function associated with the

virtual position of the m-th spatial arrangement of the audio signal source associated with the n-th audio signal in the k-th frequency band, i.e. $H_{m,k}$, is determined by performing an averaging operation over a plurality of frequency bins used for a discrete Fourier transform on the basis of the following equations:

$$X_{n,k} = \frac{1}{J(k)} \sum_{i=j(k)}^{j(k+1)-1} |\mathcal{X}(i)|, \text{ and}$$

$$H_{m,k} = \frac{1}{J(k)} \sum_{i=j(k)}^{j(k+1)-1} |\mathcal{H}(i)|,$$

wherein $\mathcal{X}(i)$ denotes the value of the discrete Fourier transform of the n-th audio signal in the i-th frequency bin, $\mathcal{H}(i)$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual positon of the audio signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

Figure 5 shows how a transfer function, such as the transfer function shown in figure 3, and an audio signal spectrum, such as the audio signal spectrum shown in figure 4, can be combined by the selector 101 in order to obtain a directional-speaker spectral profile. As can be taken from figure 5, the directional-speaker spectral profile is obtained by multiplying the (subband averaged) transfer function with the (subband averaged) audio signal spectrum, or alternatively, by summing their corresponding log-magnitude responses. In the context of the present invention, multiplying the transfer function with the audio signal spectrum is the point-wise multiplication of the two vectors defined by the averaged or discretized transfer function and the averaged or discretized audio signal spectrum, respectively. Mathematically, the selector 101 is configured to compute

$$Y_{n,m,k} = X_{n,k} H_{m,k},$$

wherein $Y_{n,m,k}$ denotes the value of the n-th directional-speaker spectral profile associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources in the k-th frequency band.

In an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of directional-speaker spectral profiles. In an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal

sources by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which a spectral difference between the plurality of directional-speaker spectral profiles is larger than a predefined threshold value, preferably a maximum.

In an embodiment, the selector 101 is configured to determine the spectral difference between the directional-speaker spectral profiles associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$S_m = \frac{1}{K}\sum_{k=1}^{K} w_k \sigma_{m,k}, \text{ and}$$

$$\sigma_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(Y_{n,m,k} - \bar{\mathbf{Y}}_{m,k})^2,$$

wherein S_m denotes a scalar value representing the spectral difference between the plurality of directional-speaker spectral profiles associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, K denotes the total number of frequency bands, w_k denotes a weighting factor, $\sigma_{m,k}$ denotes the variance across the directional-speaker spectral profiles for the k-th frequency band, N denotes the total number of audio signal spectra, and $\bar{\mathbf{Y}}_{m,k}$ denotes the mean of the directional speaker profiles in the k-th frequency band.

Figure 6A shows exemplary directional-speaker spectral profiles for the case of five speakers that can be used with the audio signal processing apparatus 100 and the audio signal processing method 200 according to an embodiment. Figure 6B shows the variance $\sigma_{m,k}$ for the five exemplary directional-speaker spectral profiles shown in figure 6A for the different frequency bands.

In an embodiment, the weighting factors w_k used to compute S_m, i.e. the spectral difference between the plurality of directional-speaker spectral profiles, can be all set to one. Alternatively, the weighting factors w_k can represent the human auditory sensitivity at the center frequencies of the different frequency bands. In this case, the weighting factors w_k may be computed as the reciprocal of the absolute threshold of hearing normalized by the minimum threshold, i.e. the threshold of the frequency band at which average human audibility is most sensitive. These exemplary weighting factors w_k, as derived from the absolute human threshold of hearing, are shown in figure 6C.

In order to deal with the possibility that the selector 101 determines at least two spatial arrangements of the virtual positions of the plurality of audio signal sources having the same maximal spectral difference, in an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and a plurality of left ear transfer functions associated with the virtual positions of the audio signal sources relative to the left ear of the listener to obtain a plurality of left ear directional-speaker spectral profiles and the plurality of audio signal spectra and a plurality of right ear transfer functions associated with the virtual positions of the audio signal sources relative to the right ear of the listener to obtain a plurality of right ear directional-speaker spectral profiles and by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which the spectral difference between the left ear directional-speaker spectral profiles and the right ear directional-speaker spectral profiles is smaller than a predefined threshold, in particular a minimum.

Figure 7 shows four exemplary spatial arrangements of virtual positions of a plurality of audio signal sources for the case of three speakers, i.e. audio signals, and twelve possible virtual positions, i.e. transfer functions. With N speakers in a virtual spatial conference capable of rendering a total of L different virtual locations, i.e. L different transfer functions, the total number of possible spatial arrangements M is given by

$$M = \binom{L}{N} \cdot N! = \frac{L!}{(L-N)! \cdot N!} \cdot N! = \frac{L!}{(L-N)!}$$

Thus, for example, if N = 3 speakers and L = 12 spatial locations then there are M = 1320 possible spatial arrangements. For the example shown in figure 7 all four arrangements provide a maximal spectral difference on the basis of a plurality of averaged transfer functions. By using left ear transfer functions and right ear transfer functions an embodiment of the present invention allows to select arrangement 2 as the optimal spatial arrangement of the virtual positions of the plurality of audio signal sources that minimizes the spectral difference between the left ear directional-speaker spectral profiles and the right ear directional-speaker spectral profiles.

In an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources from the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, the virtual positions of the plurality of audio signal sources being arranged on a

circle centered at the position of the listener and having a constant angular separation, on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by determining one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which the spectral difference between the plurality of transfer functions is larger than a predefined threshold value, preferably a maximum.

In an embodiment, the selector 101 is configured to determine the spectral difference between the transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$\hat{S}_m = \frac{1}{K}\sum_{k=1}^{K} w_k \hat{\sigma}_{m,k}, \text{ and}$$

$$\hat{\sigma}_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(H_{n,m,k} - \bar{\mathbf{H}}_{m,k})^2,$$

wherein $\hat{S}_m$ denotes a scalar value representing the spectral difference between the plurality of transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, K denotes the total number of frequency bands, w_k denotes a weighting factor, $\hat{\sigma}_{m,k}$ denotes the variance across the plurality of transfer functions for the k-th frequency band, N denotes the total number of audio signal spectra, $H_{n,m,k}$ denotes the value of the n-th transfer function in the k-th frequency band, and $\bar{\mathbf{H}}_{m,k}$ denotes the mean of the transfer functions in the k-th frequency band.

In an embodiment, the value of the n-th transfer function in the k-th frequency band, i.e. $H_{n,m,k}$, is determined by performing an averaging operation over a plurality of frequency bins used for a discrete Fourier transform on the basis of the following equation:

$$H_{n,m,k} = \frac{1}{J(k)}\sum_{i=j(k)}^{j(k+1)-1}|\mathcal{H}_n(i)|,$$

wherein $\mathcal{H}_n$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual positon of the audio signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

In an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by ranking the plurality of audio signal spectra according to the similarity of the plurality of audio signal spectra. In an embodiment, the selector 101 is configured to compute the similarity value for the plurality of audio signal spectra by (i) computing an average audio signal spectrum and the spectral differences between each audio signal spectrum and the average audio signal spectrum or (ii) by computation the correlation functions between the audio signal spectra.

In an embodiment, the selector 101 is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by assigning the ranked plurality of audio signal spectra to the virtual positions of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources in such a way that the angular separation between audio signal spectra having a small spectral difference, i.e. "similar" audio signal spectra, is maximized.

Figures 8A and 8B illustrate and example of how to select the optimal spatial arrangement of virtual positions of a plurality of speakers, i.e. audio signal sources, relative to a listener according to an embodiment. A given speaker is arbitrarily selected from the N speakers and a correlation is computed between the audio signal spectrum of the selected speaker and each of the audio signal spectra of the other N-1 speakers. The speaker audio signal spectrum that results in the highest correlation is then selected. The same process is repeated on the newly selected speaker until all speaker audio signal spectra have been ranked.

In the example illustrated in figures 8A and 8B there are N=5 speakers (ordered from 1 to 5 according to the time they first entered in to the virtual spatial audio conference), and the optimal spatial arrangement is formed by the 5 directions labeled A, B, C, D and E. The ranking of speakers according to similarity in audio signal spectra ranks them as the sequence 5, 1, 3, 2 and 4. The assignment of transfer functions starts by arbitrarily assigning the first speaker in the speaker list, i.e. speaker 5, to the first direction in the direction list, i.e. direction A. The next speaker, i.e. speaker 1, whose audio signal

spectrum is more similar to speaker 5's audio signal spectrum than to the other speakers, is assigned to the direction with the largest angular separation from direction A. In this particular example there are two options, namely directions C and D. This dual alternative is a consequence of the constraint that the directions have a constant angular separation. Here, an anticlockwise search is chosen and direction C is selected as indicated by the arrow connecting A and C. The process continues by assigning speaker 3 to direction E, because this direction gives the largest angular separation from C. The same process is repeated for speaker 2 (arrow connecting directions E and B) and speaker 4 (arrow connecting directions B and D) until all available directions are occupied.

The person skilled in the art will appreciate that embodiments of the present invention can be used for computing an optimal spatial arrangement, i.e. spatial arrangement, for loudspeaker reproduction as well, which includes but is not limited to stereo playback, 5.1., 7.1, and 22.2 channels. Independent of the number of loudspeakers and their spatial locations, these embodiments make use of the audio signal spectra to rank speakers according to spectral differences in a way that is equivalent to the procedure described above. Depending on the number of loudspeakers, their spatial locations and the maximum angular span Θ they cover, the assignment of location to the different speakers can be done in two ways.

In an embodiment, speakers are spatially separated based on simple angular distances. That is, speakers with most similar audio signal spectra are placed at locations with largest angular distance, and speakers with most dissimilar audio signal spectra are placed at locations with smallest angular distance. These locations may be at the exact positions of real loudspeakers or at positions in between loudspeakers which are then created by panning techniques or other sound field rendering technologies, e.g. wavefield synthesis.

In an alternative embodiment, speakers are spatially separated based on directional-speaker spectral profiles, as described above, or based on transfer functions, as described above. In the particular case of crosstalk cancellation systems, the above embodiments can be implemented in the exact same way as for headphone reproduction. Once the optimal spatial arrangement is found, panning techniques or soundfield rendering techniques can be used to place speakers on their optimal positions.

The person skilled in the art will appreciate that the claimed invention covers also embodiments where the audio signals and their spectra are not analyzed on the fly, but rather where a plurality of audio signal spectra of a user define a user profile, which in, turn, is represented by a profile audio signal spectrum derived therefrom, for instance, an average of audio signal spectra of a user.

Embodiments of the invention may be implemented in a computer program for running on a computer system, at least including code portions for performing steps of a method according to the invention when run on a programmable apparatus, such as a computer system or enabling a programmable apparatus to perform functions of a device or system according to the invention.

A computer program is a list of instructions such as a particular application program and/or an operating system. The computer program may for instance include one or more of: a subroutine, a function, a procedure, an object method, an object implementation, an executable application, an applet, a servlet, a source code, an object code, a shared library/dynamic load library and/or other sequence of instructions designed for execution on a computer system.

The computer program may be stored internally on computer readable storage medium or transmitted to the computer system via a computer readable transmission medium. All or some of the computer program may be provided on transitory or non-transitory computer readable media permanently, removably or remotely coupled to an information processing system. The computer readable media may include, for example and without limitation, any number of the following: magnetic storage media including disk and tape storage media; optical storage media such as compact disk media (e.g., CD-ROM, CD-R, etc.) and digital video disk storage media; nonvolatile memory storage media including semiconductor-based memory units such as FLASH memory, EEPROM, EPROM, ROM; ferromagnetic digital memories; MRAM; volatile storage media including registers, buffers or caches, main memory, RAM, etc.; and data transmission media including computer networks, point-to-point telecommunication equipment, and carrier wave transmission media, just to name a few.

A computer process typically includes an executing (running) program or portion of a program, current program values and state information, and the resources used by the operating system to manage the execution of the process. An operating system (OS) is

the software that manages the sharing of the resources of a computer and provides programmers with an interface used to access those resources. An operating system processes system data and user input, and responds by allocating and managing tasks and internal system resources as a service to users and programs of the system.

The computer system may for instance include at least one processing unit, associated memory and a number of input/output (I/O) devices. When executing the computer program, the computer system processes information according to the computer program and produces resultant output information via I/O devices.

The connections as discussed herein may be any type of connection suitable to transfer signals from or to the respective nodes, units or devices, for example via intermediate devices. Accordingly, unless implied or stated otherwise, the connections may for example be direct connections or indirect connections. The connections may be illustrated or described in reference to being a single connection, a plurality of connections, unidirectional connections, or bidirectional connections. However, different embodiments may vary the implementation of the connections. For example, separate unidirectional connections may be used rather than bidirectional connections and vice versa. Also, plurality of connections may be replaced with a single connection that transfers multiple signals serially or in a time multiplexed manner. Likewise, single connections carrying multiple signals may be separated out into various different connections carrying subsets of these signals. Therefore, many options exist for transferring signals.

Those skilled in the art will recognize that the boundaries between logic blocks are merely illustrative and that alternative embodiments may merge logic blocks or circuit elements or impose an alternate decomposition of functionality upon various logic blocks or circuit elements. Thus, it is to be understood that the architectures depicted herein are merely exemplary, and that in fact many other architectures can be implemented which achieve the same functionality.

Thus, any arrangement of components to achieve the same functionality is effectively "associated" such that the desired functionality is achieved. Hence, any two components herein combined to achieve a particular functionality can be seen as "associated with" each other such that the desired functionality is achieved, irrespective of architectures or intermediate components. Likewise, any two components so associated can also be

viewed as being "operably connected," or "operably coupled," to each other to achieve the desired functionality.

Furthermore, those skilled in the art will recognize that boundaries between the above described operations merely illustrative. The multiple operations may be combined into a single operation, a single operation may be distributed in additional operations and operations may be executed at least partially overlapping in time. Moreover, alternative embodiments may include multiple instances of a particular operation, and the order of operations may be altered in various other embodiments.

Also for example, the examples, or portions thereof, may implemented as soft or code representations of physical circuitry or of logical representations convertible into physical circuitry, such as in a hardware description language of any appropriate type.

Also, the invention is not limited to physical devices or units implemented in nonprogrammable hardware but can also be applied in programmable devices or units able to perform the desired device functions by operating in accordance with suitable program code, such as mainframes, minicomputers, servers, workstations, personal computers, notepads, personal digital assistants, electronic games, automotive and other embedded systems, cell phones and various other wireless devices, commonly denoted in this application as 'computer systems'.

However, other modifications, variations and alternatives are also possible. The specifications and drawings are, accordingly, to be regarded in an illustrative rather than in a restrictive sense.

CLAIMS:

1. An audio signal processing apparatus (100) for processing a plurality of audio signals (105) defining a plurality of audio signal spectra, the plurality of audio signals (105) to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals (105) to originate from virtual positions of a plurality of audio signal sources, the audio signal processing apparatus (100) comprising:

a selector (101) configured to select a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, wherein each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions, and wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources; and

a filter (103) configured to filter the plurality of audio signals (105) on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener, wherein the plurality of filtered audio signals (107) are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

2. The audio signal processing apparatus (100) of claim 1, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources to obtain a plurality of directional-speaker spectral profiles associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources and to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of directional-speaker spectral profiles.

3. The audio signal processing apparatus (100) of claim 1, wherein the selector (101) is configured to combine the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the

plurality of audio signal sources to obtain a plurality of directional-speaker spectral profiles associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by multiplying the plurality of input audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources.

4. The audio signal processing apparatus (100) of claim 2 or 3, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which a spectral difference between the plurality of directional-speaker spectral profiles is larger than a predefined threshold value, in particular a maximum.

5. The audio signal processing apparatus (100) of claim 4, wherein the selector (101) is configured to determine the spectral difference between the directional-speaker spectral profiles associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$S_m = \frac{1}{K}\sum_{k=1}^{K} w_k \sigma_{m,k},$$

$$\sigma_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(Y_{n,m,k} - \bar{\mathbf{Y}}_{m,k})^2, \text{ and}$$

$$Y_{n,m,k} = X_{n,k} H_{m,k},$$

wherein S_m is the spectral difference between the plurality of directional-speaker spectral profiles associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, w_k are weighting factors, $\sigma_{m,k}$ is the variance across the directional-speaker spectral profiles for a frequency band k, $\bar{\mathbf{Y}}_{m,k}$ is the frequency band average across the plurality of directional-speaker spectral profiles, $Y_{n,k,m}$ is the magnitude of a *n*th directional-speaker spectral profile in a frequency band k, $X_{n,k}$ denotes the value of the audio signal spectrum of the n-th audio signal in the k-th frequency band and $H_{m,k}$ denotes the value of the transfer function associated with the virtual position of the audio signal source associated with the n-th audio signal in the k-th frequency band.

6. The audio signal processing apparatus (100) of claim 5, wherein the selector (101) is configured to determine the value of the audio signal spectrum of the n-th audio signal in the k-th frequency band and/or the value of the transfer function associated with the

virtual position of the audio signal source associated with the n-th audio signal in the k-th frequency band by performing an averaging operation over a plurality of frequency bins on the basis of the following equations:

$$X_{n,k} = \frac{1}{J(k)} \sum_{i=j(k)}^{j(k+1)-1} |\mathcal{X}(i)|, \text{ and}$$

$$H_{m,k} = \frac{1}{J(k)} \sum_{i=j(k)}^{j(k+1)-1} |\mathcal{H}(i)|,$$

wherein $\mathcal{X}(i)$ denotes the value of the discrete Fourier transform of the n-th audio signal in the i-th frequency bin, $\mathcal{H}(i)$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual positon of the audio signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

7. The audio signal processing apparatus (100) of any one of claims 4 to 6, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources by combining the plurality of audio signal spectra and a plurality of left ear transfer functions associated with the virtual positions of the audio signal sources relative to the left ear of the listener to obtain a plurality of left ear directional-speaker spectral profiles and the plurality of audio signal spectra and a plurality of right ear transfer functions associated with the virtual positions of the audio signal sources relative to the right ear of the listener to obtain a plurality of right ear directional-speaker spectral profiles and by selecting one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which a spectral difference between the left ear directional-speaker spectral profiles and the right ear directional-speaker spectral profiles is smaller than a predefined threshold, in particular a minimum.

8. The audio signal processing apparatus (100) of claim 1, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources from the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, the virtual positions of the plurality of audio signal sources being arranged on a circle centered at the listener and having a constant angular separation on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by determining

one of the plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources for which the spectral difference between the plurality of transfer functions is larger than a predefined threshold value, in particular a maximum.

9. The audio signal processing apparatus (100) of claim 8, wherein the selector (101) is configured to determine the spectral difference between the transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources using the following equations:

$$\hat{S}_m = \frac{1}{K}\sum_{k=1}^{K} w_k \hat{\sigma}_{m,k}, \text{ and}$$

$$\hat{\sigma}_{m,k} = \frac{1}{N}\sum_{n=1}^{N}(H_{n,m,k} - \overline{\mathbf{H}}_{m,k})^2,$$

wherein $\hat{S}_m$ denotes a scalar value representing the spectral difference between the plurality of transfer functions associated with the m-th spatial arrangement of the virtual positions of the plurality of audio signal sources, K denotes the total number of frequency bands, w_k denotes a weighting factor, $\hat{\sigma}_{m,k}$ denotes the variance across the plurality of transfer functions for the k-th frequency band, N denotes the total number of audio signal spectra, $H_{n,m,k}$ denotes the value of the n-th transfer function in the k-th frequency band, and $\overline{\mathbf{H}}_{m,k}$ denotes the mean of the transfer functions in the k-th frequency band.

10. The audio signal processing apparatus (100) of claim 9, wherein the selector (101) is configured to determine the value of the n-th transfer function in the k-th frequency band by performing an averaging operation over a plurality of frequency bins on the basis of the following equation:

$$H_{n,m,k} = \frac{1}{J(k)}\sum_{i=j(k)}^{j(k+1)-1}|\mathcal{H}_n(i)|,$$

wherein $\mathcal{H}_n$ denotes the value of the discrete Fourier transform of the impulse response of the transfer function associated with the virtual position of the audio signal source associated with the n-th audio signal in the i-th frequency bin and $J(k)$ denotes the number of frequency bins of the k-th frequency band.

11. The audio signal processing apparatus (100) of claim 8, or 9, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the

virtual positions of the plurality of audio signal sources by ranking the plurality of audio signal spectra according to a similarity value of the plurality of audio signal spectra.

12. The audio signal processing apparatus (100) of claim 11, wherein the selector (101) is configured to select the spatial arrangement of the virtual positions of the plurality of audio signal sources on the basis of the plurality of audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources by assigning the ranked plurality of audio signal spectra to the virtual positions of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources in such a way that the angular separation between audio signal spectra having a large similarity value is maximized.

13. The audio signal processing apparatus (100) of claim 11 or 12, wherein the selector (101) is configured to compute the similarity value for the plurality of audio signal spectra by (i) computing an average audio signal spectrum and the spectral differences between each audio signal spectrum and the average audio signal spectrum or (ii) by computing the correlation functions between the audio signal spectra.

14. A signal processing method (200) for processing a plurality of audio signals (105) defining a plurality of audio signal spectra, the plurality of audio signals (105) to be transmitted to a listener in such a way that the listener perceives the plurality of audio signals to originate from virtual positions of a plurality of audio signal sources, the audio signal processing method (200) comprising the following steps:

selecting (201) a spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener from a plurality of possible spatial arrangements of the virtual positions of the plurality of audio signal sources relative to the listener, wherein each possible spatial arrangement of the virtual positions of the plurality of audio signal sources is associated with a plurality of transfer functions, and wherein the spatial arrangement of the virtual positions of the plurality of audio signal sources is selected on the basis of the plurality of input audio signal spectra and the plurality of transfer functions associated with each possible spatial arrangement of the virtual positions of the plurality of audio signal sources; and

filtering (203) the plurality of audio signals (105) on the basis of the selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener, wherein the plurality of filtered audio signals (107) are perceived by the listener to originate from the virtual positions of the plurality of audio signal sources defined by the

selected spatial arrangement of the virtual positions of the plurality of audio signal sources relative to the listener.

15. A computer program comprising a program code for performing the audio signal processing method (200) of claim 14 when executed on a computer.

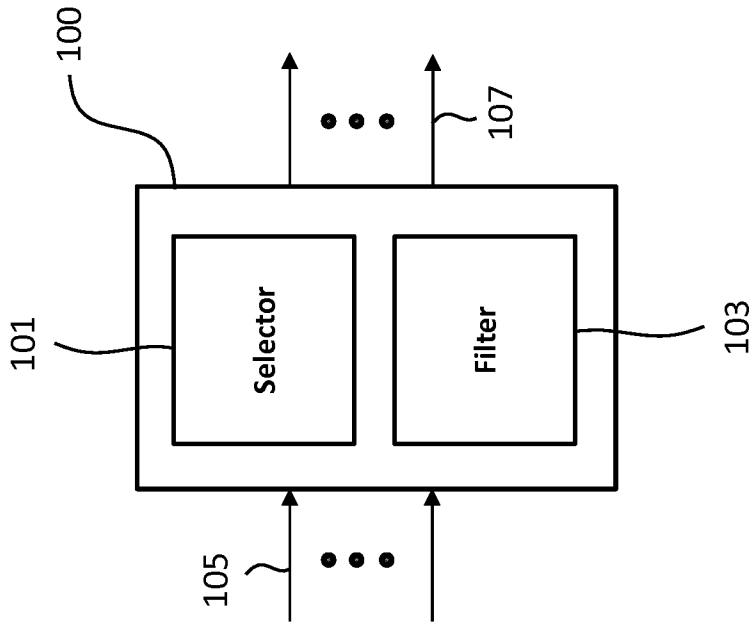

Fig. 1

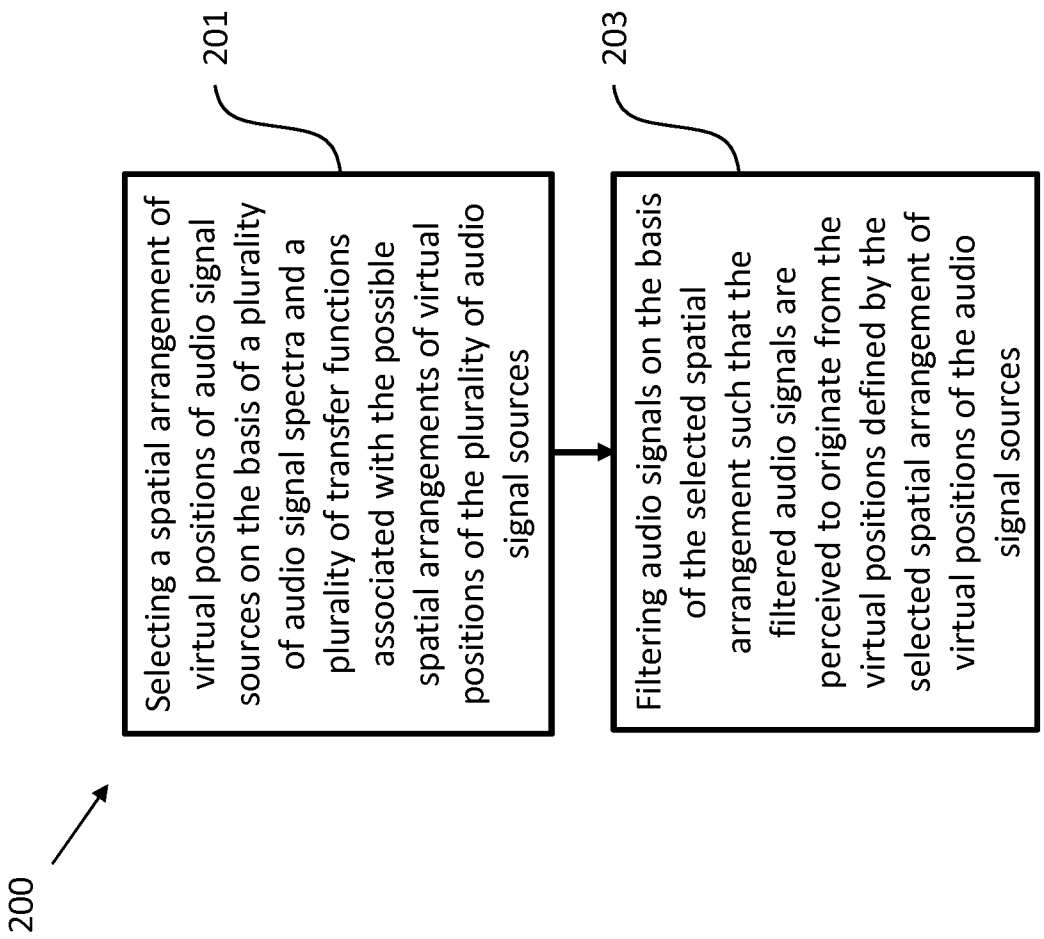
200
201
Selecting a spatial arrangement of virtual positions of audio signal sources on the basis of a plurality of audio signal spectra and a plurality of transfer functions associated with the possible spatial arrangements of virtual positions of the plurality of audio signal sources
203
Filtering audio signals on the basis of the selected spatial arrangement such that the filtered audio signals are perceived to originate from the virtual positions defined by the selected spatial arrangement of virtual positions of the audio signal sources

Fig. 2

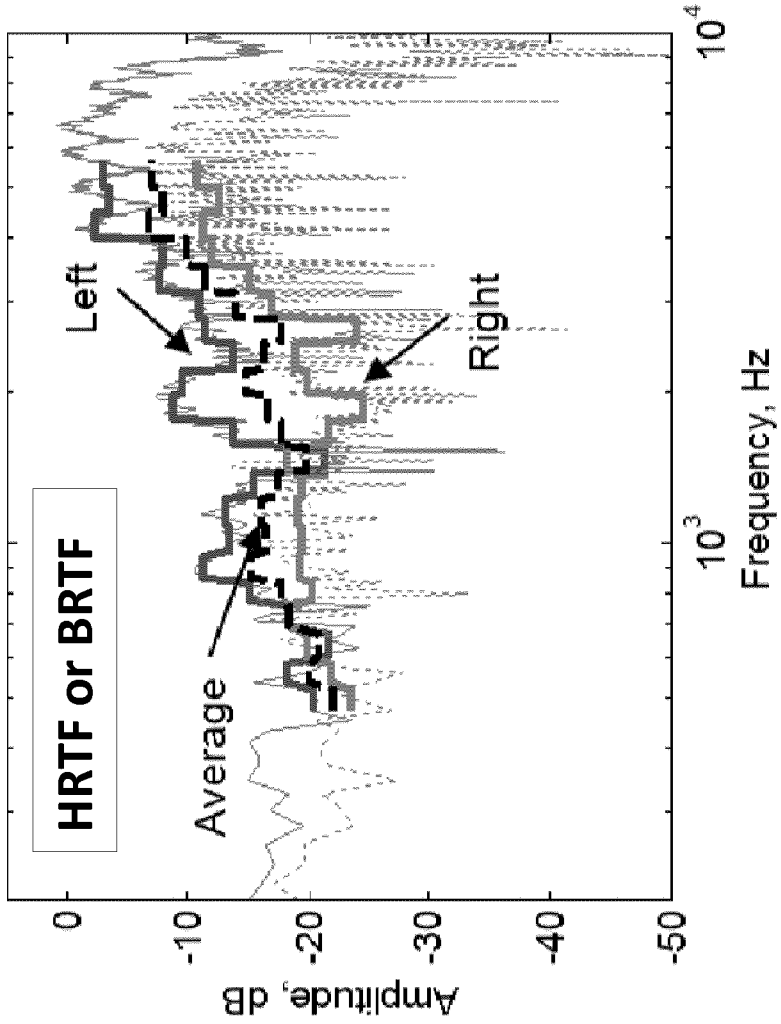
HRTF or BRTF
Left
Right
Average
Amplitude, dB
0
-10
-20
-30
-40
-50
10^3
10^4
Frequency, Hz

Fig. 3

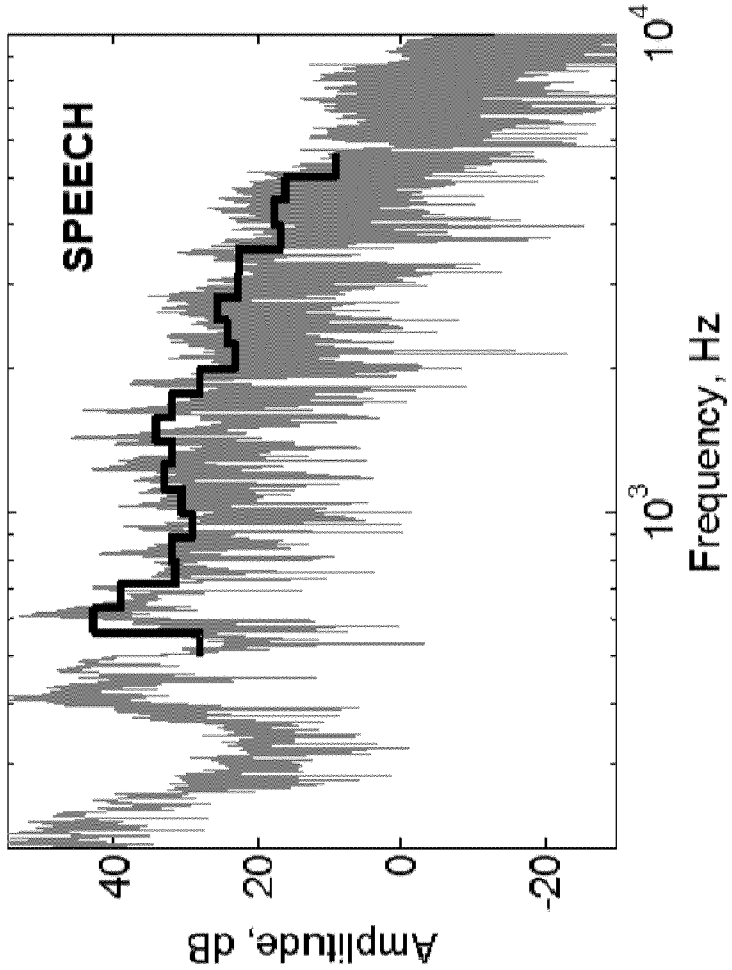

Fig. 4

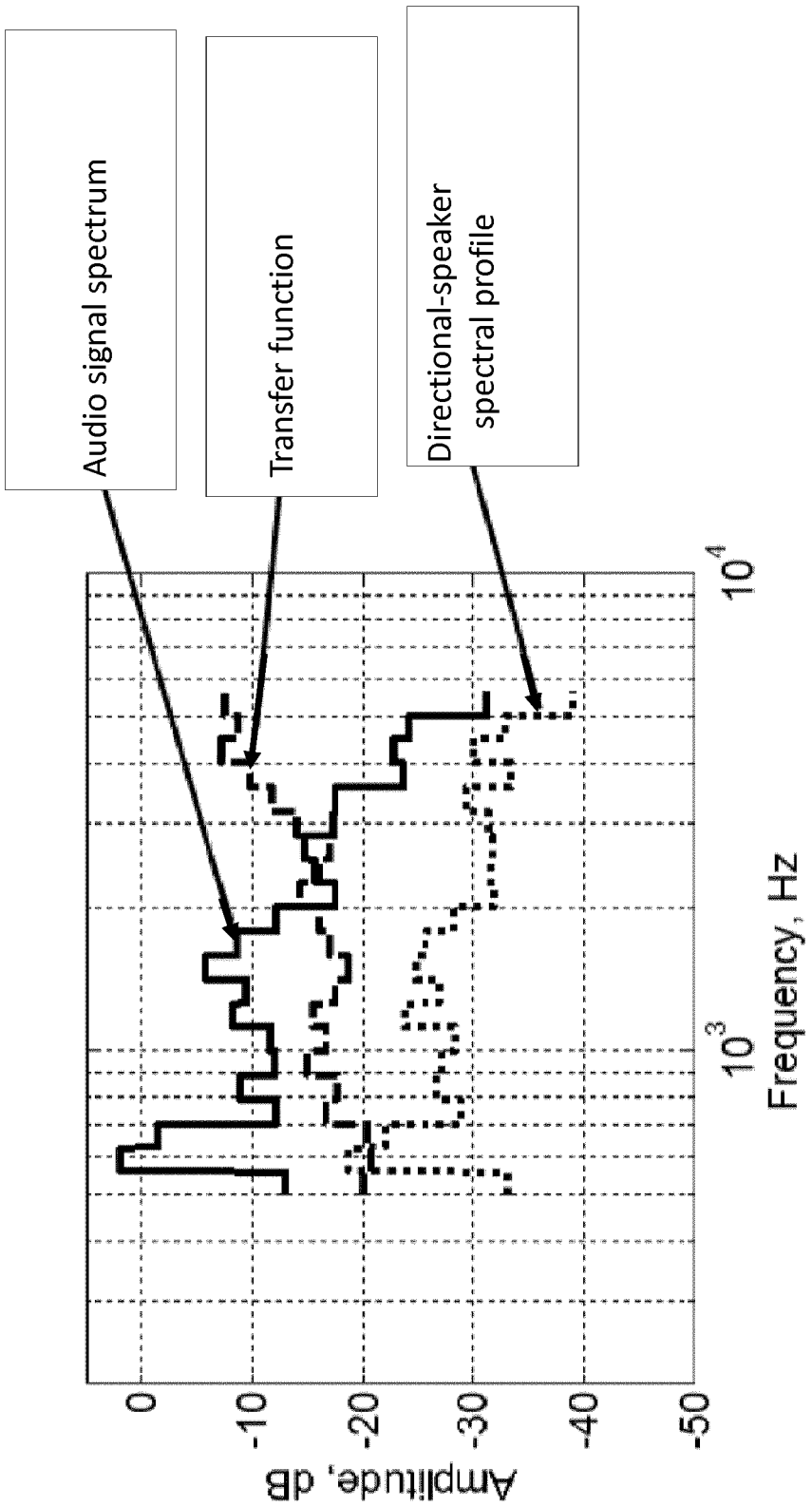
Audio signal spectrum
Transfer function
Directional-speaker spectral profile
Amplitude, dB
0
-10
-20
-30
-40
-50
10^3
10^4
Frequency, Hz

Fig. 5

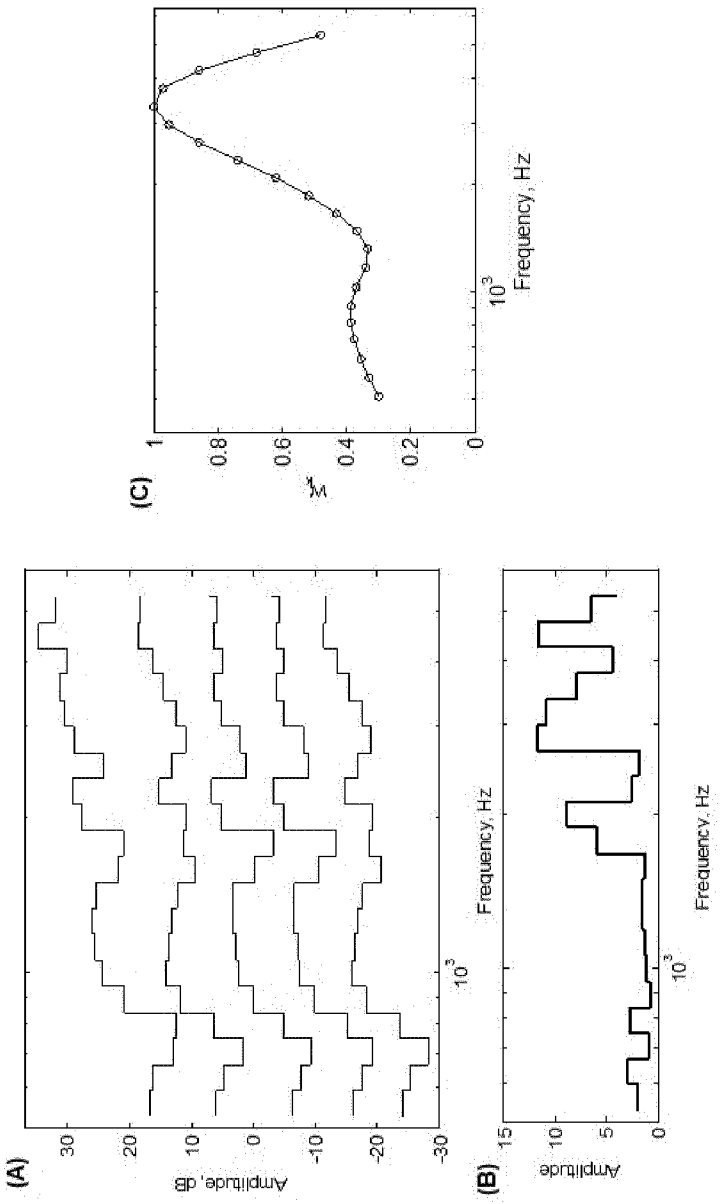
(A)
Amplitude, dB
30
20
10
0
-10
-20
-30
10^3
Frequency, Hz
(B)
Amplitude
15
10
5
0
10^3
Frequency, Hz
(C)
W_k
1
0.8
0.6
0.4
0.2
0
10^3
Frequency, Hz

Fig. 6

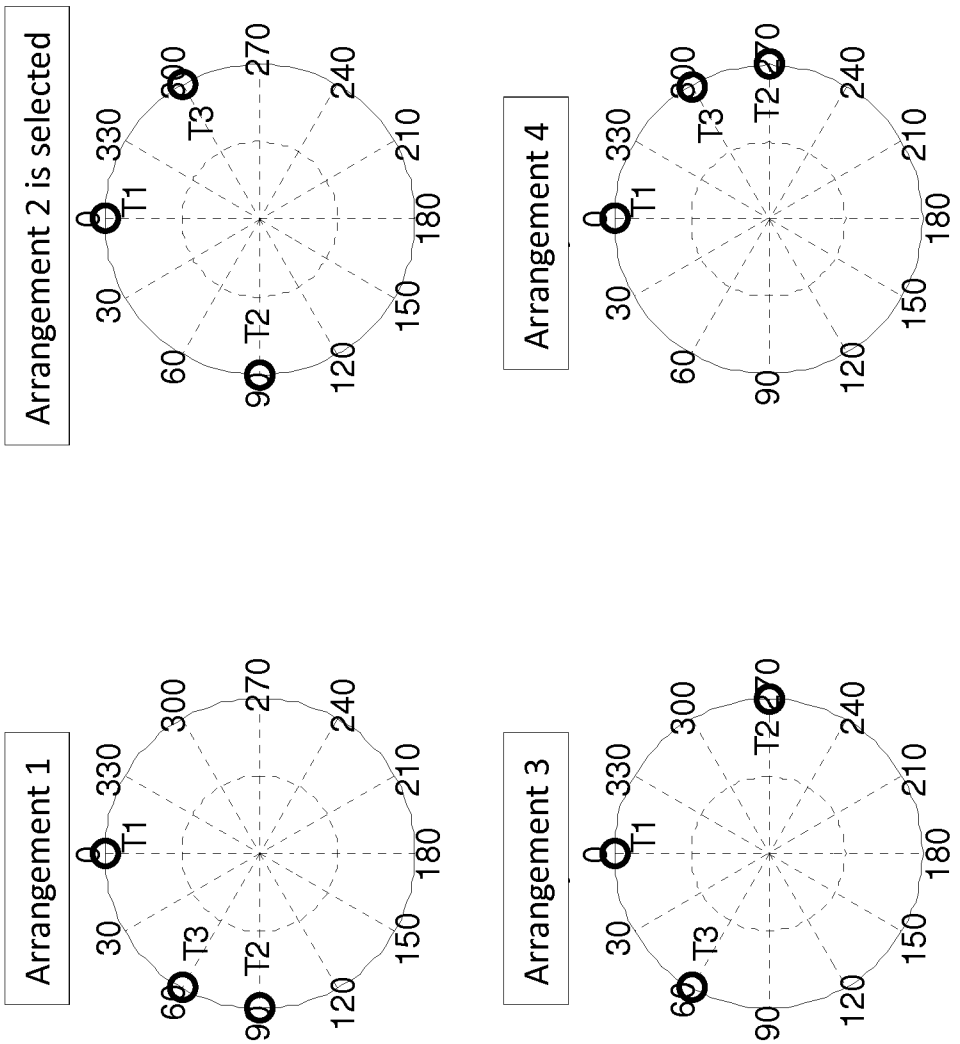
Arrangement 2 is selected
Arrangement 4
Arrangement 1
Arrangement 3
T1
T2
T3
0
30
60
90
120
150
180
210
240
270
300
330

Fig. 7

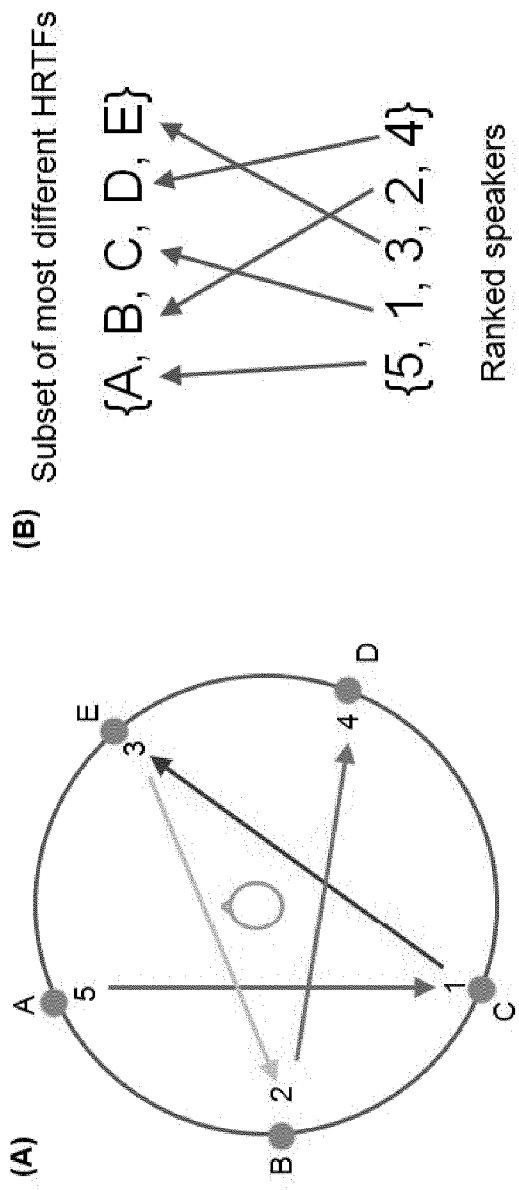
(A)
A
B
C
D
E
1
2
3
4
5
(B)
Subset of most different HRTFs
{A, B, C, D, E}
{5, 1, 3, 2, 4}
Ranked speakers

Fig. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2015/058694

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04S7/00 H04M3/56
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04S H04M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007/263823 A1 (JALAVA TEEMU [FI] ET AL) 15 November 2007 (2007-11-15) paragraphs [0010], [0011], [0026], [0027]; claims 1,5,6; figures 1-10	1-15
A	US 2009/112589 A1 (HISELIUS PER OLOF [SE] ET AL) 30 April 2009 (2009-04-30) paragraph [0067] - paragraph [0077]; figures 3,4	1-15
A	US 2007/217590 A1 (LOUPIA DAVID [FR] ET AL) 20 September 2007 (2007-09-20) paragraphs [0006] - [0011]; figures 2-6	1-15
A	EP 1 995 993 A1 (PANASONIC CORP [JP]) 26 November 2008 (2008-11-26) abstract; figure 1	1-15

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 December 2015	18/12/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Righetti, Marco

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2015/058694

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007263823 A1	15-11-2007	NONE	
US 2009112589 A1	30-04-2009	US 2009112589 A1	30-04-2009
		WO 2009056922 A1	07-05-2009
US 2007217590 A1	20-09-2007	NONE	
EP 1995993 A1	26-11-2008	CN 101422054 A	29-04-2009
		EP 1995993 A1	26-11-2008
		JP 4846790 B2	28-12-2011
		US 2009046865 A1	19-02-2009
		WO 2007119330 A1	25-10-2007