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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 September 2017 (21.09.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/158105 A1

(51) International Patent Classification:
G10L 19/00 (2013.01) *G10L 21/00* (2013.01)
G10H 1/00 (2006.01)

(21) International Application Number:
PCT/EP2017/056263

(22) International Filing Date:
16 March 2017 (16.03.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16161230.4 18 March 2016 (18.03.2016) EP

(71) Applicants: **FRAUNHOFER-GESELLSCHAFT ZUR FÖRDERUNG DER ANGEWANDTEN FORSCHUNG E.V.** [DE/DE]; Hansastraße 27c, 80686 München (DE). **FRIEDRICH-ALEXANDER-UNIVERSITAET ERLANGEN-NUERNBERG** [DE/DE]; Schlossplatz 4, 91054 Erlangen (DE).

(72) Inventors: **NIEDERMEIER, Andreas**; Ungererstr. 80, 80805 München (DE). **FÜG, Richard**; Kirchenweg 26, 90419 Nürnberg (DE). **DISCH, Sascha**; Wilhelmstrasse 70, 90766 Fürth (DE).

(74) Agents: **SCHAIRER, Oliver** et al.; Schoppe, Zimmermann, Stöckeler, Zinkler, Schenk & Partner mbB, Radlkoferstr. 2, 81373 München (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, DJ, 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, KH, KN, KP, KR, KW, 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).

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

(54) Title: ENCODING BY RECONSTRUCTING PHASE INFORMATION USING A STRUCTURE TENSOR ON AUDIO SPECTROGRAMS

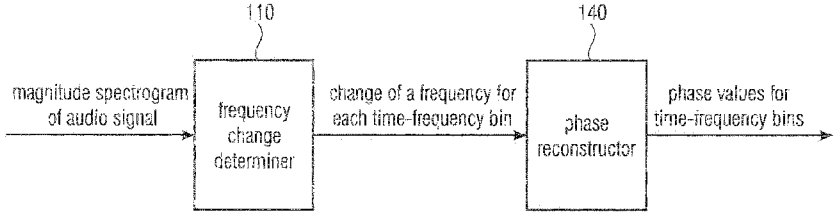


Fig. 7

(57) Abstract: An apparatus for phase reconstruction from a magnitude spectrogram of an audio signal is provided. The apparatus comprises a frequency change determiner (110) being configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal, and a phase reconstructor (140) being configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

WO 2017/158105 A1

ENCODING BY RECONSTRUCTING PHASE INFORMATION USING A STRUCTURE TENSOR ON AUDIO SPECTROGRAMS

Description

The present invention relates to audio signal processing, and, in particular, to an apparatus and method for harmonic-percussive-residual sound separation using the structure tensor on spectrograms.

Being able to separate a sound into its harmonic and percussive component is an effective preprocessing step for many applications.

While "Harmonic-Percussive(-Residual) Separation" is a common term, it is misleading as it implies a harmonic structure with sinusoidals having a frequency of an integer multiple of the fundamental frequency. Even though the correct term should be "Tonal-Percussive-(Residual) Separation", the term and "harmonic" instead of "tonal" is used in the following for easier understanding.

Using the separated percussive component of a music recording for example can lead to a quality improvement for beat tracking (see [1]), rhythm analysis and transcription of rhythm instruments. The separated harmonic component is suitable for the transcription of pitched instruments and chord detection (see [3]). Furthermore, harmonic-percussive separation can be used for remixing purposes like changing the level ratio between both signal components (see [4]), which leads to an either "smoother" or "punchier" overall sound perception.

Some methods for harmonic-percussive sound separation rely on the assumption that harmonic sounds have a horizontal structure in the magnitude spectrogram of the input signal (in time direction), while percussive sounds appear as vertical structures (in frequency direction). Ono et al. presented a method that first creates harmonically/percussively enhanced spectrograms by diffusion in time/frequency direction (see [5]). By comparing these enhanced representations afterwards, a decision if a sound is either harmonic or percussive could be derived.

A similar method was published by Fitzgerald, where the enhanced spectrograms were calculated by using median filtering in perpendicular directions instead of diffusion (see [6]), which led to similar results while reducing the computational complexity.

Inspired by the sines+transients+noise (S+T+N) signal model (see [7], [8], [9]), a framework that aims to describe the respective signal components by means of a small set of parameters. Fitzgerald's method was then extended to harmonic-percussive-residual (HPR) separation in [10]. As audio signals often consist of sounds that are neither clearly harmonic nor percussive, this procedure captures these sounds in a third, residual component. While some of these residual signals clearly have an isotropic, neither horizontal nor vertical, structure (as for example noise), there exist sounds that do not have a clear horizontal structure but nevertheless carry tonal information and may be perceived as harmonic part of a sound. An example are frequency modulated tones like they can occur in recordings of violin playing or vocals, where they are said to have "vibrato". Due to the strategy of recognizing either horizontal or vertical structures, the aforementioned methods are not always able to capture such sounds in their harmonic component.

A harmonic-percussive separation procedure based on non-negative matrix factorization that is capable of capturing harmonic sounds with non-horizontal spectral structures in the harmonic component was proposed in [11]. However it did not include a third residual component.

Summarizing the above, recent methods rely on the observation that in a spectrogram representation, harmonic sounds lead to horizontal structures and percussive sounds lead to vertical structures. Furthermore, these methods associate structures that are neither horizontal nor vertical (i.e., non-harmonic, non-percussive sounds) with a residual category. However, this assumption does not hold for signals like frequency modulated tones that show fluctuating spectral structures, while nevertheless carrying tonal information.

The structure tensor, a tool used in image processing (see [12], [13]), is applied there to grey scale images for edge and corner detection (see [14]) or to estimate the orientation of an object. The structure tensor has already been used for preprocessing and feature extraction in audio processing (see [15], [16]).

The object of the present invention is to provide improved concepts for audio signal processing. The object of the present invention is solved by an apparatus according to claim 1, by a system according to claim 18, by an encoder according to claim 19, by a method according to claim 20 and by a computer program according to claim 21.

An apparatus for phase reconstruction from a magnitude spectrogram of an audio signal is provided. The apparatus comprises a frequency change determiner being configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal, and a phase reconstructor being configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

Moreover, an encoder being configured to generate a magnitude spectrogram of an audio signal for the apparatus for phase reconstruction as described above is provided.

Moreover, a method for phase reconstruction from a magnitude spectrogram of an audio signal is provided. The method comprises:

- Determining a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal. And:

- Generating phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

Furthermore, a computer program is provided, wherein the computer program is configured to implement the above-described method when being executed on a computer or signal processor.

In the following, embodiments of the present invention are described in more detail with reference to the figures, in which:

Fig. 1 illustrates an apparatus for analysing a magnitude spectrogram of an audio signal according to an embodiment,

Fig. 2 illustrates a spectrogram of a mixture of a singing voice, castanets, and applause with zoomed in region according to an embodiment, wherein the orientation of the arrows indicates direction and wherein the length of the arrows indicates an anisotropy measure,

Fig. 3 illustrates a range of orientation/anisotropy values computed by using the structure tensor according to an embodiment,

Fig. 4 illustrates a comparison between the HPR-M and HPR-ST method for an excerpt of a synthetic input signal,

Fig. 5 illustrates an apparatus according to an embodiment, wherein the apparatus comprises a signal generator,

Fig. 6 illustrates an apparatus according to an embodiment, wherein the apparatus comprises one or more microphones for recording the audio signal,

Fig. 7 illustrates an apparatus for phase reconstruction according to an embodiment,

Fig. 8 illustrates an apparatus for phase reconstruction according to an embodiment comprising a signal generator,

Fig. 9 illustrates a system comprising an encoder and a decoder according to an embodiment, wherein the decoder is an apparatus for phase reconstruction according to an embodiment, and

Fig. 10 illustrates an audio encoder and an audio decoder according to embodiments using the magnitude spectrum for transmission.

Fig. 1 illustrates an apparatus for analysing a magnitude spectrogram of an audio signal according to embodiments.

The apparatus comprises a frequency change determiner 110. The frequency change determiner 110 is configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal.

Moreover, the apparatus comprises a classifier 120. The classifier 120 is configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups depending on the change of the frequency determined for said time-frequency bin.

According to an embodiment, the frequency change determiner 110 may, e.g., be configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on an angle $\alpha(b, k)$ for said time-frequency bin. The angle $\alpha(b, k)$ for said time-frequency bin depends on the magnitude spectrogram of the audio signal.

In an embodiment, the frequency change determiner 110 may, e.g., be configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins further depending on a sampling frequency f_s of the audio signal, and depending on a length N of an analysis window and depending on a hop size H of the analysis window.

According to an embodiment, the apparatus frequency change determiner 110 is configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on the formula

$$R(b, k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b, k))$$

(b, k) indicates a time-frequency bin of the plurality of time-frequency bins, wherein $R(b, k)$ indicates the change of the frequency for said time-frequency bin (b, k), wherein b indicates time, wherein k indicates frequency, wherein f_s indicates the sampling frequency of the audio signal, wherein N indicates the length of the analysis window, wherein H indicates the hop size of the analysis window, and wherein $\alpha(b, k)$ indicates the angle for said time-frequency bin (b, k), wherein the angle $\alpha(b, k)$ depends on the magnitude spectrogram.

In an embodiment, the frequency change determiner 110 may, e.g., be configured to determine a partial derivative $\mathbf{S}_b$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to a time index. In such an embodiment, the frequency change determiner 110 may, e.g., be configured to determine a partial derivative $\mathbf{S}_k$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to a time index.

Moreover, in such an embodiment, the frequency change determiner 110 is configured to determine a structure tensor $\mathbf{T}(b, k)$ for each time-frequency bin (b, k) of the plurality of time-frequency bins depending on the partial derivative $\mathbf{S}_b$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to the time index and depending on the partial derivative $\mathbf{S}_k$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to the frequency index.

Furthermore, in such an embodiment, the frequency change determiner 110 may, e.g., be configured to determine the angle $\alpha(b,k)$ for each time-frequency bin (b,k) of the plurality of time-frequency bins depending the structure tensor $\mathbf{T}(b,k)$ for said time-frequency bin (b,k).

According to an embodiment, the frequency change determiner 110 may, e.g., be configured to determine the angle $\alpha(b,k)$ for each time-frequency bin (b,k) of the plurality of time-frequency bins by determining two components of $v_1(b,k)$ and $v_2(b,k)$ an eigenvector $\mathbf{v}(b,k)$ of the structure tensor($\mathbf{T}(b,k)$) of said time-frequency bin (b,k), and by determining the angle ($\alpha(b,k)$) for said time-frequency bin ((b,k)) according to

$$\alpha(b,k) = \operatorname{atan}\left(\frac{v_2(b,k)}{v_1(b,k)}\right) \in [-\pi/2;\pi/2]$$

$\alpha(b,k)$ indicates the angle for said time-frequency bin ((b,k)), wherein b indicates time, wherein k indicates frequency, and wherein atan() indicates an inverse tangent function.

In an embodiment, the classifier 120 may, e.g., be configured to determine a measure of anisotropy for each time-frequency bin (b,k) of the plurality of time-frequency bins depending on at least one of the formulae:

$$\left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2$$

and

$$\mu(b,k)+\lambda(b,k) \geq e,$$

$\mu(b,k)$ is a first eigenvalue $\lambda(b,k)$ is a second eigenvalue of the structure tensor ($\mathbf{T}(b,k)$) of said time-frequency bin (b,k), and $e \in \mathbb{R}^{>0}$.

In such an embodiment, the classifier 120 may, e.g., be configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups further depending on the change of the measure of anisotropy.

According to an embodiment, the classifier 120 may, e.g., be configured to determine the measure of anisotropy for said time-frequency bin (b, k) depending on the formula:

$$C(b,k) = \begin{cases} \left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2 & , \mu(b,k)+\lambda(b,k) \geq e \\ 0 & , \text{ else} \end{cases}$$

$C(b, k)$ is the measure of anisotropy depending for said time-frequency bin (b, k), and wherein the classifier 120 is configured to assign said time-frequency bin (b, k) to a residual component group of the two or more signal component groups, if the measure of anisotropy $C(b, k)$ is smaller than a first threshold value c, or wherein the classifier 120 is configured to assign said time-frequency bin (b, k) to the residual component group of the two or more signal component groups, if the measure of anisotropy $C(b, k)$ is smaller than or equal to the first threshold c, wherein $c \in \mathbb{R}^{>0}$.

In an embodiment, the classifier 120 may, e.g., be configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups depending on the change $R(b, k)$ of the frequency determined for said time-frequency bin (b, k), so that the classifier 120 assigns a time-frequency bin of the plurality of time-frequency bins to a harmonic signal component group of the two or more signal component groups depending on whether an absolute value $|R(b, k)|$ of the change $R(b, k)$ of the frequency determined for said time-frequency bin (b, k) is smaller than a second threshold r_h, or depending on whether the absolute value $|R(b, k)|$ of the change $R(b, k)$ of the frequency determined for said time-frequency bin (b, k) is smaller than or equal to the second threshold r_h, wherein $r_h \in \mathbb{R}^{>0}$.

According to an embodiment, the classifier 120 may, e.g., be configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups depending on the change $R(b, k)$ of the frequency determined for said time-frequency bin (b, k), so that the classifier 120 assigns a time-frequency bin of the plurality of time-frequency bins to a percussive signal component group of the two or more signal component groups depending on whether the absolute value $|R(b, k)|$ of the change $R(b, k)$ of the frequency determined for said time-frequency bin (b, k) is greater than a third threshold r_p, or depending on whether the absolute value $|R(b, k)|$ of the change ($R(b, k)$) of the frequency determined for said time-frequency bin (b, k) is greater than or equal to the third threshold r_p, wherein $r_p \in \mathbb{R}^{>0}$.

In the following, a detailed description of embodiments is provided.

Embodiments provide improved concepts for Harmonic-percussive-residual (HPR) sound separation based on the structure tensor. Some embodiments capture frequency modulated sounds that hold tonal information in the harmonic component by exploiting the information about the orientation of spectral structures provided by the structure tensor.

Some embodiments are based on the finding that a strict classification into horizontal and vertical is inappropriate for these signals and might lead to leakage of tonal information into the residual component. Embodiments relate to a novel method that instead uses the structure tensor, a mathematical tool, to calculate predominant orientation angles in the magnitude spectrogram. Embodiments employ this orientation information to distinguish between harmonic, percussive, and residual signal components, even in the case of frequency modulated signals. Finally, the effectiveness of the concept of embodiments is verified by means of both objective evaluation measures as well as audio examples.

Moreover, some embodiments are based on the finding that the structure tensor can be considered a black box, where the input is a gray scale image and the outputs are angles n for each pixel corresponding to the direction of lowest change and a certainty or anisotropy measure for this direction for each pixel. The structure tensor additionally offers the possibility to be smoothed, which reduces the influence of noise for enhanced robustness. Furthermore the certainty measure can be used to determine the quality of the estimated angles. A low value of this certainty measure indicates that the pixel lays in a region of constant brightness without any clear direction.

A local frequency change may, e.g., be extracted from the angles obtained by the structure tensor. Form these angles, it can be determined, whether a time-frequency-bin in the spectrogram belongs to the harmonic (= low local frequency change) or the percussive (= high or infinite local frequency change) component.

Improved embodiments for Harmonic-Percussive-Residual Classification And Separation are provided.

Harmonic-percussive-residual sound separation is a useful preprocessing tool for applications such as pitched instrument transcription or rhythm extraction. Instead of searching only for strictly horizontal and vertical structures, some embodiments determine predominant orientation angles as well as the local anisotropy in the spectrogram by using the structure tensor known from image processing.

In embodiments, the provided information about the orientation of spectral structures can then be used to distinguish between harmonic, percussive, and residual signal components by setting appropriate thresholds, see Fig. 2.

Fig. 2 illustrates a spectrogram of a mixture of a singing voice, castanets, and applause with zoomed in region additionally showing direction (orientation of arrows) and anisotropy measure (length of arrows) obtained by the structure tensor. The color of the arrows indicate whether the respective time-frequency bin is assigned to the harmonic component (areas 210), to the percussive component (areas 230), or to the residual component (areas 220) based on the orientation and anisotropy information.

All bins having neither a high nor a low local frequency change rate or a certainty measure that indicates a constant region were assigned to belong to the residual component. An example for this separation of a spectrogram can be seen in Fig. 2. Embodiments perform better for audio signals comprising frequency modulated sounds than similar methods working on the magnitude spectrogram.

At first, a concept of structure tensor is described and this general concept is extended to be applicable in the context of audio processing.

In the following, matrices and vectors are written as bold letters for notational convenience. Furthermore, the $(\cdot)$ operator is used to index a specific element. In this case the matrix or vector is written as a non-bold letter to show its scalar use.

At first, calculation of a spectrogram according to embodiments is described. The audio signal may, e.g., be a (discrete) input audio signal.

The structure tensor may be applied to the spectrogram representation of a discrete input audio signal $\mathbf{x} \in \mathbb{R}^M$ with a sampling frequency of f_s. For the spectral analysis of $\mathbf{x}$, the short-time Fourier-transform (STFT)

$$X(b,k) := \sum_{n=0}^{N-1} w(n)x(n+Hb)\exp\left(-i2\pi nk/N\right) \tag{1}$$

is used, where $X(b,k) \in \mathbb{C}$, b denotes the frame index, k the frequency index and $w \in \mathbb{R}^N$ is a window function of length N (in other words: N is a length of an analysis

window). $H \in \mathbb{N}$, $H \leq N$ represents the analysis hop size of the window. It should be noted that since the STFT spectrum has a certain symmetry around the Nyquist point at $\frac{N}{2}$, the processing may, for example, be restricted to $0 \leq k \leq \frac{N}{2}$, as the symmetry can be reconstructed during the inverse STFT.

By using the above formula (1), a spectrogram can be obtained. A spectrogram comprises a plurality of spectra, wherein the plurality of spectra succeed each other in time. A second spectrum of the plurality spectra succeeds a first spectrum in time, if at least some second time domain samples exist that are used to generate the second spectrum and that are not used to generate the first spectrum, and that are time domain samples that refer to a later point-in-time than first time domain samples that are used to generate the first spectrum. Windows of time domain samples used for generating timely neighbored spectra may, for example, overlap.

In embodiments, the analysis window length N may, e.g., be defined to be:

$$256 \text{ samples} \leq N \leq 2048 \text{ samples}.$$

In some embodiments, the analysis window length may, e.g., be 2048. In other embodiments,the analysis window length may, e.g., be 1024 samples. In further embodiments, the analysis window length may, e.g., be 768 samples. In still further embodiments, the analysis window length may, e.g., be 256 samples.

In embodiments, the analysis hop size H may, e.g., be in a range between 25 % and 75 % of the analysis window. In such embodiments:

$$0.25\, N \leq H \leq 0.75\, N.$$

Thus, in such embodiments, if the analysis window has, e.g., 2048 samples (N = 2048), the analysis hop size may, e.g., be in the range:

$$512 \text{ samples} \leq H \leq 1536 \text{ samples}.$$

If the analysis window has, e.g., 256 samples (N = 256), the analysis hop size may, e.g., be in the range:

64 samples $\leq H \leq$ 192 samples.

In preferred embodiments, the analysis hop size may, e.g., be 50 % of the analysis window. This corresponds to a window overlap of two subsequent analysis windows of 50 %.

In some embodiments, the analysis hop size may, e.g., be 25 % of the analysis window. This corresponds to a window overlap of two subsequent analysis windows of 75 %.

In other embodiments, the analysis hop size may, e.g., be 75 % of the analysis window. This corresponds to a window overlap of two subsequent analysis windows of 25 %.

It should be noted that the concepts of the present invention are applicable for any kind of time domain to spectral domain transformation, such as for the MDCT (Modified Discrete Cosine Transform), MDST (Modified Discrete Sine Transform, DSTFT (Discrete Short-Time Fourier Transform), etc.

The real valued logarithmic spectrogram may, e.g., be calculated as:

$$S(b,k) = 20\log_{10}|X(b,k)| \tag{2}$$

The magnitude spectrogram of the audio signal may be referred to as $\mathbf{S}$ and a value of the magnitude spectrogram for a time-frequency bin (b,k) may be referred to as $S(b,k)$.

In the following, calculation of the structure tensor according to embodiments is described.

For the calculation of the structure tensor the partial derivatives of $\mathbf{S}$ are needed. The partial derivative with respect to time index b is given by

$$\mathbf{S}_b = \mathbf{S} * \mathbf{d} \tag{3}$$

while the partial derivative with respect to frequency index k is defined as

$$\mathbf{S}_k = \mathbf{S} * \mathbf{d}^T \tag{4}$$

where $\mathbf{d}$ is a discrete differentiation operator (for example, for central differences one could choose $\mathbf{d} = [-1, 0, 1] / 2$) and $*$ denotes the 2-dimensional convolution.

Furthermore, it may be defined:

$$\mathbf{T}_{11} = (\mathbf{S}_b \odot \mathbf{S}_b) * \mathbf{G} \quad (5)$$

$$\mathbf{T}_{21} = \mathbf{T}_{12} = (\mathbf{S}_k \odot \mathbf{S}_b) * \mathbf{G} \quad (6)$$

$$\mathbf{T}_{22} = (\mathbf{S}_k \odot \mathbf{S}_k) * \mathbf{G} \quad (7)$$

where $\odot$ is the point wise matrix multiplication, also known as the Hadamard product and $\mathbf{G}$ is a 2-D Gaussian smoothing filter having the standard deviation σ_b in time index direction and σ_k in frequency index direction. The structure tensor $\mathbf{T}(b, k)$ is then given by a 2×2 symmetric and positive semidefinite matrix

$$\mathbf{T}(b, k) = \begin{bmatrix} T_{11}(b, k) & T_{12}(b, k) \\ T_{21}(b, k) & T_{22}(b, k) \end{bmatrix} \quad (8)$$

The structure tensor comprises information about the dominant orientation of the spectrogram at position (b, k). It should be noted that in the special case where $\mathbf{G}$ is a scalar, $\mathbf{T}(b, k)$ does not comprise more information than the gradient at this position in the spectrogram. However in contrast to the gradient, the structure tensor can be smoothed by $\mathbf{G}$ without cancellation effects, which makes it more robust against noise.

It should be noted, that a structure tensor $\mathbf{T}(b, k)$ is defined for each time-frequency bin (b, k) of the plurality of time-frequency bins. So when a plurality of time frequency bins is considered, e.g., the time-frequency bins $(0, 0); (0, 1); (0, 2); \ldots (1, 0); (1, 1); (1, 2); \ldots$ then there exist a plurality of structure tensors $\mathbf{T}(0, 0); \mathbf{T}(0, 1); \mathbf{T}(0, 2); \ldots \mathbf{T}(1, 0); \mathbf{T}(1, 1); \mathbf{T}(1, 2); \ldots$ For example, for each time-frequency bin $(b, k)$of the plurality of time-frequency bins, one structure tensor $\mathbf{T}(b, k)$ is determined.

In the following, calculation of angles and an anisotropy measure according to embodiments is described.

The information about the orientation for each bin in the spectrogram is obtained by calculating the eigenvalues $\lambda(b,k)$, $\mu(b,k)$ with $\lambda(b,k) \leq \mu(b,k)$ and the corresponding eigenvectors $\mathbf{v}(b,k) = [v_1(b,k), v_2(b,k)]^T$ and $\mathbf{w}(b,k) = [w_1(b,k), w_2(b,k)]^T$ of the structure tensor $\mathbf{T}(b,k)$. It should be noted that $\mathbf{v}(b,k)$, the eigenvector corresponding to the smaller eigenvalue $\lambda(b,k)$, is pointing into the direction of lowest change in the spectrogram at index (b,k), while $\mathbf{w}(b,k)$ is pointing in to the direction of highest change. Thus, the angle of the orientation at a specific bin can be obtained by

$$\alpha(b,k) = \operatorname{atan}\left(\frac{v_2(b,k)}{v_1(b,k)}\right) \in [-\pi/2; \pi/2] \tag{9}$$

$v_1(b,k)$ and $v_2(b,k)$ are the components of the eigenvector $\mathbf{v}(b,k)$.

atan() indicates an inverse tangent function.

In addition, a measure of anisotropy

$$C(b,k) = \begin{cases} \left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2 & , \mu(b,k) + \lambda(b,k) \geq e \\ 0 & , \text{ else} \end{cases} \tag{10}$$

with $e \in \mathbb{R}^{>0}$ can be determined for each bin. It should be noted that $C(b,k) \in [0;1]$. Values of $C(b,k)$ close to 1 indicate a high anisotropy of the spectrogram at index (b,k), while a constant neighborhood leads to values close to 0. The threshold e, that defines a limit on what should be considered anisotropic, can be chosen to further increase the robustness against noise.

The physical meaning of angle $\alpha(b,k)$ can be understood by considering a continuous signal with a change of instantaneous frequency Δf during a time interval Δt. Thus the instantaneous frequency change rate R is denoted by

$$R = \frac{\Delta f}{\Delta t} \tag{11}$$

For example, according to embodiments, the angles (indicated by the direction of the arrows in Fig. 2) obtained by the structure tensor may, e.g., be translated into a local frequency change rate

$$R = \frac{\Delta f}{\Delta t}[Hz/s] \qquad (11a)$$

for each time-frequency-bin of the spectrogram.

The change of the frequency for each time-frequency bin may, e.g., be referred to as instantaneous frequency change rate.

Considering sample rate, length and hop-size of the applied STFT analysis, a relation between the angles in the spectrogram and the instantaneous frequency change rate $R(b,k)$ for each bin can be derived by

$$R(b,k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b,k)) \qquad (12)$$

Also the standard deviations of the smoothing filter $\mathbf{G}$ in the discrete domain σ_b and σ_k can be converted to the continuous physical parameters σ_t and σ_f by

$$\sigma_t = \frac{H}{f_s}\sigma_b, \qquad \sigma_f = \frac{f_s}{N}\sigma_k \qquad (13)$$

In the following harmonic-percussive-residual separation using the structural tensor is described.

The information obtained via the structure tensor can be applied to the problem of HPR separation, e.g., to classify each bin in the spectrogram as being part of either the harmonic, the percussive or the residual component of the input signal.

Embodiments are based on the finding that bins assigned to the harmonic components should belong to rather horizontal structures, while bins belonging to rather vertical structures should be assigned to the percussive component. Furthermore, bins that do not belong to any kind of oriented structure should be assigned to the residual component.

According to embodiments, a bin (b,k) may, e.g., be assigned to the harmonic component, if it satisfies the first of the following two constraints.

According to preferred embodiments, a bin (b,k) may, e.g., be assigned to the harmonic component, if it satisfies both of the following two constraints:

- The first constraint may, e.g., be that the absolute value of the angle $\alpha(b,k)$ is smaller than (or equal to) a threshold α_h. The threshold α_h may, e.g., be in the range $\alpha_h \in [0;\pi/2]$. This means, that the bin should be part of some spectral structure that does not have a slope bigger or smaller than α_h. This way also frequency modulated sounds can be considered to be part of the harmonic component, depending on the parameter α_h.

- The second constraint may, e.g., be that the measure of anisotropy $C(b,k)$ supports that the bin (b,k) is part of some directed, anisotropic structure, and therefore exceeds a further threshold c. It should be noted that for a given bin (b,k), the angle $\alpha(b,k)$ and the measure of anisotropy $C(b,k)$ together define a point in $\mathbb{R}^2$ given in polar coordinates.

Similarly, in embodiments, another angle threshold α_p is assigned to define when a bin should be assigned to the percussive component (areas 330 with vertical lines in Fig. 3).

Thus, according to embodiments, a bin (b,k) may, e.g., be assigned to the percussive component, if it satisfies the first of the following two constraints.

According to preferred embodiments, a bin (b,k) may, e.g., be assigned to the percussive component, if it satisfies both of the following two constraints:

- The first constraint may, e.g., be that the absolute value of the angle $\alpha(b,k)$ is greater than (or equal to) a threshold value α_p. The threshold α_p may, e.g., be in the range $\alpha_p \in [0;\pi/2]$. This means, that the bin should be part of some spectral structure that does not have a slope bigger or smaller than α_p. This way also frequency modulated sounds can be considered to be part of the harmonic component, depending on the parameter α_p.

- The second constraint may, e.g., be that the measure of anisotropy $C(b,k)$ supports that the bin (b,k) is part of some directed, anisotropic structure, and therefore exceeds a further threshold c. It should be noted that for a given bin (b,k), the angle $\alpha(b,k)$ and the measure of anisotropy $C(b,k)$ together define a point in $\mathbb{R}^2$ given in polar coordinates.

Finally, in embodiments, all bins that are assigned to neither the harmonic nor the percussive component may, e.g., be assigned to the residual component.

The above-described assignment process can be expressed by defining a mask for the harmonic component $\mathbf{M}_h$, a mask for the percussive component $\mathbf{M}_p$ and a mask for the residual component $\mathbf{M}_r$.

It should be noted, that instead of using the threshold value α_h and the threshold value α_p thresholds may in embodiments, e.g., be defined on the maximum absolute frequency change rate $r_h, r_p \in \mathbb{R}^{>0}$ with $r_p \geq r_h$ to give the choice of parameters a better physical interpretation. The masks are then given by:

$$M_h(b,k) = \begin{cases} 1 & , |R(b,k)| \leq r_h \wedge C(b,k) > c \\ 0 & , \text{else} \end{cases} \quad (14)$$

$$M_p(b,k) = \begin{cases} 1 & , |R(b,k)| > r_p \wedge C(b,k) > c \\ 0 & , \text{else} \end{cases} \quad (15)$$

$$M_r(b,k) = 1 - M_h(b,k) - M_p(b,k) \quad (16)$$

Finally, the STFT of the harmonic component $\mathbf{X}_h$, the percussive component $\mathbf{X}_p$ and the residual component $\mathbf{X}_r$ are obtained by

$$\mathbf{X}_h = \mathbf{M}_h \odot \mathbf{X} \quad (17)$$

$$\mathbf{X}_p = \mathbf{M}_p \odot \mathbf{X} \quad (18)$$

$$\mathbf{X}_r = \mathbf{M}_r \odot \mathbf{X} \quad (19)$$

The corresponding time signals can then be calculated via the inverse STFT.

Fig. 3 illustrates a range of orientation/anisotropy values computed by the structure tensor.

In particular, Fig. 3 depicts the subset of all points that lead to an assignment to the harmonic component. In particular, values in the areas 310 with wavy lines lead to an assignment to the harmonic component.

Values in the areas 330 with vertical lines lead to an assignment to the percussive component.

Values in the areas 320 that are dotted lead to an assignment to the residual component.

The threshold α_h defines line 301 in Fig. 3, and the threshold α_p defines line 302 in Fig. 3.

Fig. 5 illustrates an apparatus according to an embodiment, wherein the apparatus comprises a signal generator 130 being configured to generate an audio output signal depending on the assigning of the plurality of time-frequency bins to the two or more signal component groups.

For example, signal generator may filter the different components of the audio signal by applying different weighting factors on the magnitude values of the time-frequency bins of the different signal component groups. For example, the harmonic signal component group may have a first weighting factor w_h, the percussive signal component group may have a second weighting factor w_p, and the residual signal component group may have a first weighting factor w_r, and the magnitude value of each time-frequency bin of the plurality of time-frequency bins may, e.g., be weighted with the weighting factor of the signal component group, the time-frequency bin is assigned to.

For example, to emphasize harmonic signal components, in an embodiment, where the weighting factors are multiplied the with linear magnitude values e.g.,

$$w_h = 1.3,\ w_p = 0.7, \text{ and } w_r = 0.2$$

For example, to emphasize harmonic signal components, in an embodiment, where the weighting factors are added to the logarithmic magnitude values e.g.,

$$w_h = +0.26,\ w_p = -0.35, \text{ and } w_r = -1.61$$

For example, to emphasize percussive signal components, in an embodiment, where the weighting factors are multiplied the with linear magnitude values e.g.,

$$w_h = 0.7,\ w_p = 1.3, \text{ and } w_r = 0.2$$

For example, to emphasize percussive signal components, in an embodiment, where the weighting factors are added to the logarithmic magnitude values e.g.,

$$w_h = -0.35,\ w_p = +0.26, \text{ and } w_r = -1.61$$

Thus, the signal generator 130 is configured to apply a weighting factor on the magnitude value of each time-frequency bin of the plurality of time-frequency bins to obtain the audio output signal, wherein the weighting factor that is applied on said time-frequency bin depends on the signal component group to which said time-frequency bin is assigned.

In a particular embodiment of Fig. 5, the signal processor 130 may, e.g., be an upmixer being configured to upmix the audio signal to obtain the audio output signal comprising two or more audio output channels. The upmixer may, e.g., be configured to generate the two or more audio output channels depending on the assigning of the plurality of time-frequency bins to the two or more signal component groups.

For example, the two or more audio output channels may be generated from the audio signal filter the different components of the audio signal by applying different weighting factors on the magnitude values time-frequency bins of the different signal component groups as described above.

However, to generate the different audio channels different weights for the signal component groups may be used that may, e.g., be specific for each of the different audio output channels.

For example, for a first audio output channel, the weights to be added to the logarithmic magnitude values may, e.g., be

$$w_{1h} = +0.26,\ w_{1p} = -0.35, \text{ and } w_{1r} = -1.61.$$

And for a second audio output channel, the weights to be added to the logarithmic magnitude values may, e.g., be

$$w_{2h} = +0.35,\ w_{2p} = -0.26, \text{ and } w_{2r} = -1.61.$$

For example, when upmixing the audio signal to obtain five audio output channels front left, center, right, left surround and right surround:

- The harmonic weighting factor w_{1h} may be greater for generating the left, center and right audio output channels compared to the harmonic weighting factor w_{2h} for generating the left surround and right surround audio output channels.

- The percussive weighting factor w_{1p} may be smaller for generating the left, center and right audio output channels compared to the percussive weighting factor w_{2p} for generating the left surround and right surround audio output channels.

Individual weighting factors may be used for each audio output channel to be generated.

Fig. 6 illustrates an apparatus according to an embodiment, wherein the apparatus comprises one or more microphones 171, 172 for recording the audio signal.

In Fig. 6, the first microphone 171 records a first audio channel of the audio signal. The optional second microphone 172 records an optional second audio channel of the audio signal.

Moreover, the apparatus of Fig. 6 further comprises a magnitude spectrogram generator 180 for generating the magnitude spectrogram of the audio signal from the audio signal which comprises the first audio channel and optionally comprises the optional second audio channel. Generating a magnitude spectrogram from an audio signal is a well-known concept for a person skilled in the art.

In the following, evaluation of embodiments is considered.

To show the effectiveness of embodiments in capturing frequency modulated sounds in the harmonic component, the HPR method based on the structure tensor (HPR-ST) according to embodiments is compared with the non-iterative method based on median filtering presented in [10] (HPR-M). Additionally, the metrics are also computed for the separation results with ideal binary masks (IBM) that served as a reference for the maximal achievable separation quality.

Considering a system-under-test parameters, for both HPR-ST as well as HPR-M, the STFT parameters were chosen to be f_s = 22050Hz, N = 1024 and H = 256, using a sine window for $\mathbf{w}$. The separation parameters for HPR-M were chosen as in the experiments performed in [10]. According to embodiments, the structure tensor is calculated using a differential operator, for example, the Scharr-Operator [17] as discrete differentiation

operator **d**. The smoothing was performed using a 9×9 isotropic Gaussian filter with the standard deviations $\sigma_b = \sigma_k = 1.4$ which leads to $\sigma_t \approx 16ms$ and $\sigma_f \approx 30Hz$. Finally, the thresholds for the separation were set to $e = 20$, $c = 0.2$ and $r_h = r_p = 10000Hz/s$.

It should be noted that by the choice of r_h and r_p according to embodiments, even very steep structures in the spectrogram are assigned to the harmonic component. Embodiments employ the observations about real world vibrato sounds as for example shown in Fig. 2. Here, you can see that at some instances the vibrato in the singing voice has a very high instantaneous frequency change rate. Furthermore, it should be noted that by choosing $r_h = r_p$, an assignment of a bin in the spectrogram to the residual component is purely dependent on its anisotropy measure.

The effectiveness of HPR-ST according to embodiments was evaluated by comparing it to the state-of-art median filtering based method HPR-M presented in [10] by means of both objective evaluation measures as well as audio examples.

To compare the behavior of HPR-ST according to embodiments and HPR-M of the prior art when applied to signals comprising frequency modulated sounds to obtain objective results, two test items were generated.

Test item 1 consists of the superposition of purely synthetic sounds. The harmonic source was chosen to be a vibrato tone with a fundamental frequency of 1000Hz, a vibrato frequency of 3Hz, vibrato extent of 50Hz and 4 overtones. For the percussive source several impulses are used, while white noise represents the neither harmonic nor percussive residual source.

Test item 2 was generated by superimposing real world signals of singing voice with vibrato (harmonic), castanets (percussive), and applause (neither harmonic nor percussive).

Interpreting the HPR separation of these items as a source separation problem, standard source separation evaluation metrics have been computed (source to distortion ratio SDR, source to interference ratio SIR, and source to artifacts ratios SAR, as introduced in [18]) for the separation results of both procedures. The results are shown in table 1.

Table 1 depicts objective evaluation measures, where all values are given in dB:

		SDR			SIR			SAR		
		IBM	HPR-M	HPR-ST	IBM	HPR-M	HPR-ST	IBM	HPR-M	HPR-ST
Item 1	Vibrato	29.43	11.51	21.25	34.26	27.94	30.01	31.16	11.61	21.88
	Impulses	8.56	-10.33	-1.47	20.31	-7.96	12.03	8.90	2.02	-1.00
	Noise	8.49	-13.53	2.58	24.70	-11.99	14.12	8.61	3.97	3.06
Item 2	Vocals	14.82	6.48	9.18	22.75	20.83	15.61	15.60	6.68	10.42
	Castanets	8.48	3.79	2.37	21.59	16.09	17.96	8.73	4.16	2.56
	Applause	7.39	-2.03	-0.37	20.31	1.11	6.34	7.66	3.33	1.58

(Table 1)

For item 1 HPR-ST yields a SDR of 21.25dB for the vibrato tone, and is therefore closer to the optimal separation result of IBM (29.43dB) than to the separation result of HPR-M (11.51dB). This indicates that HPR-ST improves on capturing this frequency modulated sound in the harmonic component in comparison to HPRM. This is also shown in Fig. 4.

Fig. 4 illustrates a comparison between the HPR-M and HPR-ST method for an excerpt of the synthetic input signal (item 1). For enhanced visibility the spectrograms were calculated with different STFT parameters than used for the separation algorithms.

Fig. 4 (a) illustrates the frequency of the input signal with respect to time. In Fig. 4, the spectrograms of the harmonic components and the sum of the percussive and residual component computed for both procedures are plotted. It can be seen that for HPR-M the steep slopes of the vibrato tone leaked into the residual component (Fig. 4 (b) and (c)), while HPR-ST (Fig. 4 (d) and (e)) yields a good separation. This also explains the very low SIR values of HPRM for the residual component compared to HPR-ST (-11.99dB vs. 14.12dB).

It should be noted that the high SIR value of HPR-M for the harmonic component only reflects that there are little interfering sounds from the other components, not that the sound of the vibrato is well captured as a whole. In general most of the observations for item 1 are less pronounced, but also valid for the mixture of real world sounds in item 2. For this item, the SIR value of HPR-M for the vocals even exceeds the SIR value of HPR-ST (20.83dB vs. 15.61dB). Again, the low SIR value for the applause supports that portions of the vibrato in the vocals leaked into the residual component for HPR-M (1.11dB) while the residual component of HPR-ST comprises less interfering sounds

(6.34dB). This indicates that embodiments were capable of capturing the frequency modulated structures of the vocals much better than HPR-M.

Summarizing the results, for signals that comprise frequency modulated tones, the HPR-ST concept of embodiments provides much better separation results compared to HPR-M.

Some embodiments employ the structure tensor for singing voice detection. (Singing voice detection according to the prior art is described in [2]).

In the following, another aspect of embodiments is described. This further aspect relates to phase reconstruction from a magnitude spectrogram.

Fig. 7 illustrates an apparatus for phase reconstruction from a magnitude spectrogram of an audio signal according to an embodiment.

The apparatus comprises a frequency change determiner 110 being configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal.

Moreover, the apparatus comprises a phase reconstructor 140 being configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

According to embodiments, the phase reconstructor 140 may, e.g., be configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times (by integrating two times temporally). In other words, in embodiments, the phase reconstructor 140 is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating two times over time. So, in other words, the interval(s) over which integration is/are performed twice extend/s along the time axis of the spectrogram.

According to other embodiments, the phase reconstructor 140 is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating two times over frequency (by integrating two times spectrally). So, in other words, the interval(s) over which integration is/are performed twice extend/s along the frequency axis of the spectrogram. For example, the integration over frequency that is conducted twice may, e.g., be conducted based on formulae (30) and (31).

In embodiments, the phase reconstructor 140 may, e.g., be configured to generate phase values for the plurality of time-frequency bins depending on the formula

$$\phi(bH) = \phi(b_0 H) + \sum_{n=b_0}^{b} \left(\phi_b(b_0) + \sum_{m=b_0}^{n} \phi_{bb}(m) \right)$$

b_0 may, e.g., be an index indicating an analysis block of a plurality of blocks. b may, e.g., be a further index indicating a further analysis block of the plurality of blocks. H may, e.g., indicate a hop size. $\phi(bH)$ and $\phi_b(b_0)$ and $\phi_{bb}(m)$ may, e.g., indicate phase values.

In embodiments, the frequency change determiner 110 may, e.g., be configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on an angle $\alpha(b,k)$ for said time-frequency bin, wherein the angle $\alpha(b,k)$ for said time-frequency bin depends on the magnitude spectrogram of the audio signal.

According to embodiments, the frequency change determiner 110 may, e.g., be configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins further depending on a sampling frequency f_s of the audio signal, and depending on a length N of an analysis window and depending on a hop size H of the analysis window.

In embodiments, the apparatus frequency change determiner 110 may, e.g., be configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on the formula

$$R(b,k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b,k))$$

(b,k) indicates a time-frequency bin of the plurality of time-frequency bins, $R(b,k)$ indicates the change of the frequency for said time-frequency bin (b,k), b indicates time, k indicates frequency, f_s indicates the sampling frequency of the audio signal, N indicates the length of the analysis window, H indicates the hop size of the analysis window, and $\alpha(b,k)$ indicates the angle for said time-frequency bin (b,k), wherein the angle $\alpha(b,k)$ depends on the magnitude spectrogram.

According to embodiments, the frequency change determiner 110 may, e.g., be configured to determine a partial derivative $\mathbf{S}_b$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to a time index. Moreover, the frequency change determiner 110 may, e.g., be configured to determine a partial derivative $\mathbf{S}_k$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to a time index. Furthermore, the frequency change determiner 110 may, e.g., be configured to determine a structure tensor $\mathbf{T}(b, k)$ for each time-frequency bin (b, k) of the plurality of time-frequency bins depending on the partial derivative $\mathbf{S}_b$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to the time index and depending on the partial derivative $\mathbf{S}_k$ of the magnitude spectrogram $\mathbf{S}$ of the audio signal with respect to the frequency index. Moreover, the frequency change determiner 110 may, e.g., be configured to determine the angle $\alpha(b, k)$ for each time-frequency bin (b, k) of the plurality of time-frequency bins depending the structure tensor $\mathbf{T}(b, k)$ for said time-frequency bin (b, k).

In embodiments, the frequency change determiner 110 may, e.g., be configured to determine the angle $\alpha(b, k)$ for each time-frequency bin (b, k) of the plurality of time-frequency bins by determining two components of $v_1(b, k)$ and $v_2(b, k)$ an eigenvector $\mathbf{v}(b, k)$ of the structure tensor $\mathbf{T}(b, k)$ of said time-frequency bin (b, k), and by determining the angle $\alpha(b, k)$ for said time-frequency bin (b, k) according to

$$\alpha(b, k) = \operatorname{atan}\left(\frac{v_2(b, k)}{v_1(b, k)}\right) \in [-\pi/2; \pi/2]$$

$\alpha(b, k)$ indicates the angle for said time-frequency bin (b, k); b indicates time, k indicates frequency, and atan() indicates an inverse tangent function.

According to embodiments, the phase reconstructor 140 being configured to generate phase values for each time-frequency bin of the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bin

In embodiments, the phase reconstructor 140 may, e.g., be configured to determine a measure of anisotropy for each time-frequency bin (b, k) of the plurality of time-frequency bins depending on the change of the frequency determined for each time-frequency bin of a plurality of time-frequency bins.

According to embodiments, the phase reconstructor 140 may, e.g., be configured to determine the measure of anisotropy for each time-frequency bin (b, k) of the plurality of time-frequency bins depending on at least one of the formulae:

$$\left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2$$

and

$$\mu(b,k)+\lambda(b,k) \geq e,$$

$\mu(b,k)$ is a first eigenvalue $\lambda(b,k)$ is a second eigenvalue of the structure tensor ($\mathbf{T}(b,k)$) of said time-frequency bin (b,k), and $e \in \mathbb{R}^{>0}$. The phase reconstructor 140 may, e.g., be configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups further depending on the change of the measure of anisotropy.

In embodiments, the phase reconstructor 140 is configured to determine the measure of anisotropy for said time-frequency bin (b,k) depending on the formula:

$$C(b,k) = \begin{cases} \left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2 & , \mu(b,k)+\lambda(b,k) \geq e \\ 0 & , \text{ else} \end{cases}$$

$C(b,k)$ is the measure of anisotropy depending for said time-frequency bin (b,k). The phase reconstructor 140 may, e.g., be configured to assign said time-frequency bin (b,k) to a residual component group of the two or more signal component groups, if the measure of anisotropy $C(b,k)$ is smaller than an upper threshold value c, or wherein the phase reconstructor 140 may, e.g., be configured to assign said time-frequency bin (b,k) to the residual component group of the two or more signal component groups, if the measure of anisotropy $C(b,k)$ is smaller than or equal to the upper threshold value c. $c \in \mathbb{R}^{>0}$.

According to embodiments, the phase reconstructor 140 may, e.g., be configured to determine a phase reconstruction mode for each time-frequency bin of the plurality of time-frequency bins depending on the measure of anisotropy, whether phase reconstruction shall be conducted,

The phase reconstructor 140 may, e.g., be configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times, if the phase reconstruction mode determined by the phase reconstructor 140 for said time-frequency bin indicates a first mode. Moreover, the phase reconstructor 140 may, e.g., be configured to determine a phase for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times, if the phase reconstruction mode determined by the phase reconstructor 140 for said time-frequency bin indicates a second mode, being different from the first mode.

In embodiments, the phase reconstructor 140 may, e.g., be configured to determine the phase reconstruction mode for each time-frequency bin (b, k) of the plurality of time-frequency bins so that the phase reconstruction mode indicates the first mode, if the measure of anisotropy is greater than a lower threshold value c, and so that the phase reconstruction mode indicates the second mode, if the measure of anisotropy for said time-frequency bin (b, k) is smaller than or equal to the lower threshold value c.

In embodiments, the upper threshold value and the lower threshold value c may, e.g., be equal.

Or, the phase reconstructor (140) is configured to determine the phase reconstruction mode for each time-frequency bin (b, k) of the plurality of time-frequency bins so that the phase reconstruction mode indicates the first mode, if the measure of anisotropy for said time-frequency bin (b, k) is greater than or equal to the threshold value c, and so that the phase reconstruction mode indicates the second mode, if the measure of anisotropy is smaller than the threshold value c.

The first mode may, for example, indicate that the time-frequency bin belongs to the harmonic or percussive signal component group.

The first mode may, for example, indicate that the time-frequency bin belongs to the residual signal component group.

In the following, phase estimation according to particular embodiments is described in more detail.

Algorithms for the task of estimating the phase information from a given magnitude spectrogram have been proposed in [19] and [20]. However these algorithms either have

high computational complexity or lead to unsatisfactory perceptible quality for general audio signals.

Estimating the phase information from a magnitude spectrogram is a task, that could be used for example in audio coding, where the encoder could transmit only the magnitude while the phase is restored at the decoder. In comparison to state-of-the-art MDCT based coders, where the spectrum show fluctuations over time even for constant tonal signals, a magnitude spectrum (e.g. of the DFT, or CMDCT aka. MCLT) is much more stable which allows for bit rate efficient differential coding over time (see [21]) and effective multi-channel redundancy reduction.

In embodiments, based on the local frequency change estimation by the structure tensor a novel algorithm according to embodiments is based on a signal model considering an arbitrary input signal to be composed of several linear chirps. A linear chirp signal is given by

$$x(t) = sin(\phi(t)), \qquad \phi(t) = 2\pi f_0 t + \pi R t^2 \tag{20}$$

where t is the time variable, f_0 [Hz] is the constant starting frequency and R [Hz/s] is the constant frequency change rate. By deriving the phase argument $\phi(t)$ two times with respect to t it can be shown that

$$\frac{d^2}{dt^2}\phi(t) = 2\pi R. \tag{21}$$

Thus the current phase can be obtained by integrating the frequency change rate two times. For our algorithm working in the discrete domain this integration turns into a summation. With the assumption of each spectrogram bin corresponding to a local linear chirp signal, the local frequency change estimated by the structure tensor then corresponds to R for each local chirp. Furthermore using the anisotropy measure obtained via the structure tensor it is possible to perform the needed integration of the local frequency change rates without any explicit signal model or prior semantic classification.

As a high anisotropy measure corresponds do directed structures like tonal or percussive components in the spectrogram, it consequently corresponds to regions where it is possible to reconstruct a phase in contrast to isotrope, noisy regions, where random phase can be assumed.

This is now described in more detail:

As has been described above, the angles (for example, indicated by the direction of the arrows in Fig. 2) obtained by the structure tensor can be translated into a local frequency change rate

$$R(b,k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b,k)) = \frac{\Delta f}{\Delta t}(b,k)[Hz/s] \tag{22}$$

for each time-frequency-bin of the spectrogram, where $\alpha(b,k)$ are the angles for each bin extracted via the structure tensor.

In the following, concepts for using information, for example, of the structure tensor, for phase reconstruction from a magnitude spectrogram and its use for audio coding is explained. It is assumed

$$X(b,k) = |X(b,k)| \cdot \exp(\mathrm{i} \cdot \varphi(b,k)) \tag{23}$$

to be not only an STFT but also arbitrary time-frequency representation of $x(t)$. The problem of phase reconstruction can then be described as the task to extract an estimate $\hat{\varphi}(b,k) \approx \varphi(b,k)$ out of the magnitude spectrogram $|X(b,k)|$. Via an inverse transform of

$$Y(b,k) = |X(b,k)| \cdot \exp(\mathrm{i} \cdot \hat{\varphi}(b,k)) \tag{24}$$

it is then possible to generate the corresponding time domain signal $y(t)$ again.

In embodiments, based on the local frequency change estimation by the structure tensor an novel algorithm for phase reconstruction from a magnitude spectrogram is provided. The principal idea can be understood by considering a discrete linear chirp signal given by

$$x(t) = sin(\phi(t)), \qquad \phi(t) = 2\pi\frac{f_0}{f_s}t + \pi R_0\left(\frac{t}{f_s}\right)^2 + \phi_0 \tag{25}$$

where $t \in \mathbb{Z}$ is the time index variable, f_0 [Hz] is the constant starting frequency, R_0 [Hz/s] is the constant frequency change rate and ϕ_0 the initial phase. If this signal is analyzed by

using an overlapping block transform e.g. the STFT as a filter bank respectively time-frequency representation, the phase propagation because of the window hop size H from the previous analysis block $b-1$ to the current block b can be expressed by

$$\phi_b(b) = \phi(bH) - \phi((b-1)H) = 2\pi \frac{f_0}{f_s} H + \frac{\pi R_0}{f_s^2} \left(2bH^2 + H^2\right) \tag{26}$$

Thus $\phi_b(b)$ is the finite difference with respect to analysis block index b and can be interpreted as an estimate for the instantaneous angular frequency of $x(t)$. By using $\phi_b(b)$ one can express $\phi(bH)$ at any analysis block $b > b_0$ when $\phi_b(b_0)$ is known

$$\phi(bH) = \phi_b(b_0) + \sum_{n=b_0}^{b} \phi_b(n) \tag{27}$$

Now, the change of $\phi_b(b)$ from the previous analysis block $b-1$ to the current block b is considered:

$$\phi_{bb}(b) = \phi_b(b) - \phi_b(b-1) = \frac{2\pi H}{f_s^2} R_0 \tag{28}$$

$\phi_{bb}(b)$ is constant for the linear chirp and comprises the fixed constants as well as the frequency change rate R_0. It can be interpreted as the angular frequency increase from one analysis block to the next block. If $\phi_b(b_0)$ is known, formula (27) can be expressed even further using $\phi_{bb}(b)$

$$\phi(bH) = \phi(b_0 H) + \sum_{n=b_0}^{b} \left(\phi_b(b_0) + \sum_{m=b_0}^{n} \phi_{bb}(m) \right) \tag{29}$$

This means that we can calculate the phase $\phi(bH)$ at analysis block b if we know the phase and the instantaneous frequency at a previous analysis block and further the constant frequency change R_0 .

Note that the dual of these expressions and formulas can be calculated for transient signals e.g. impulses. However there the finite difference has to be done with respect to frequency index k in the spectral phase $\psi(k) = arg(X(b,k))$ of a certain analysis block b.

The temporal center of gravity for a transient signal is then obtained by

$$\psi_k(k) = \psi(k) - \psi(k-1) \tag{30}$$

and the change of temporal center of gravity for transient signal over frequency index k by

$$\psi_{kk}(k) = \psi_k(k) - \psi_k(k-1) \tag{31}$$

The dual expressions for formula (27) and formula (29) can be deduced for this case.

According to embodiments, an algorithm for phase reconstruction from a magnitude spectrum is provided as explained above using the estimate of the local frequency change rate obtained by the structure tensor.

With the assumption of each spectrogram bin corresponding to a local linear chirp signal respectively having a linear frequency change, the local frequency change $R(b,k)$ estimated by the structure tensor then corresponds to the chirp rate or local linear frequency change R_0 as explained in the previous subsection. This means that the predominant directions obtained by the structure tensor can be seen as a smoothed, robust estimate for the second derivative of $\phi_{bb}(b)$ with respect to the time index b and up to some multiplicative constants.

Using this estimate, the formula (29) can then be used to calculate the current phase of the signal. As a linear frequency change is only assumed in the scope of one bin even for signals with more complex frequency modulation a phase estimate can be obtained. It has to be noted that the instantaneous frequency as well as the initial phase have either to be known beforehand (e.g. transmitted as side information) or estimated with different methods. Especially the instantaneous frequency can be estimated using interpolation like QFFT or derivates of the magnitude spectrum.

It has to be noted that even though formula (29) shows sums strictly summing into analysis block direction b, in a more advanced algorithm the sum has to follow the trajectory of the signals main lobe in the spectrogram. Thus it might be necessary to

incorporate the directional information respectively local frequency change rate obtained via the structure tensor either implicitly or explicitly in the direction of summation. This could lead to the estimated phase being a superposition of the result of sums in frequency index k direction and analysis block index b direction.

Further it has to be noted, that the expression calculated via formula (29) corresponds to the argument of the sine as shown in formula (25). Even though this has a strong correspondence to the phase observed in the spectral representation, it might be necessary to use further knowledge (e.g. an analytic expression of the spectrum in dependence of the phase) to synthesize the phase for each bin (e.g. for the side lobes of the spectrum) correctly.

Furthermore using the anisotropy measure obtained via the structure tensor it is possible to perform the needed integration of the local frequency change rates without any explicit signal model or prior semantic classification. As a high anisotropy measure corresponds do directed structures like tonal or percussive components in the spectrogram, it consequently corresponds to regions where it is possible to reconstruct a phase in contrast to isotrope, noisy regions, where random phase can be assumed.

Furthermore it is not necessary to restrict the algorithm to monotimbral signals.

Embodiments achieve advantages over the prior art. For example, some embodiments exhibit moderate computational complexity (e.g., less than in [19]). Moreover, some embodiments, better phase estimation than in [20] for frequency modulated signals.

Some embodiments realize inherent classification of signal components:

For example, according to some embodiments, phase estimation is possible for harmonic signal components; phase estimation is possible for percussive signal components; but phase estimation is not possible for residual signal components and/or noisy signal components.

In the following, audio coding applications are considered.

The choice of filter bank is a critical step in designing an audio codec. State-of-the-art codecs often use the MDCT (Modified-Discrete-Cosine-Transform), as it offers 50% overlap while being critically sampled and perfect reconstructing in absence of quantization. In coding applications these properties reduce blocking artifacts while

keeping the amount of data for spectral coefficients that have to be transmitted low. A downside of the MDCT is its spectral fluctuation over time even for stationary signals. This leads to a loss of coding gain for e.g. differential coding of spectral coefficients, as information of previously transmitted MDCT spectrum have only limited use in the reconstruction of the current MDCT spectrum.

As the magnitude of a spectrum , e.g. the magnitude of a 50% overlapping MCLT (Modulated-Complex-Lapped-Transform) filter bank, is much more stable over time [21] especially for stationary signals, embodiments provide a design of a codec based on the above concepts for phase reconstruction.

According to embodiments, the encoder performs a time-frequency decomposition of the PCM input signal $x(t)$ using the analysis filter bank to obtain the complex spectrum $X(f)$ for a certain time interval e.g. one frame. $X(f)$ is used to extract side info.

In embodiments, the side info may, e.g., comprise a fundamental frequency and/or a time position of transient signals and/or initialization phases (e.g. in regular intervals) and/or information about the current signal class, etc.

Then the magnitude of $X(f)$ is quantized to $|Y(f)|$ and transmitted to the decoder together with the side info. The decoder then uses the side information as well as the quantized magnitude spectrum $|Y(f)|$ to estimate the phase of the original complex spectrum $X(f)$ as described above. Using the this estimated phase a complex spectrum $Y(f)$ can be obtained, which should resemble $X(f)$ closely. $Y(f)$ is then fed into the synthesis filter bank to obtain the time domain output signal $y(t)$. With the proposed MCLT filter bank a codec like this would still show the desirable features like overlap and critical sampling, while offering more efficient possibilities for differential coding of stationary signals.

Fig. 8 illustrates an apparatus for phase reconstruction according to one of the above-described embodiments comprising a signal generator 150. The signal generator 150 is configured to generate an audio output signal depending on the magnitude spectrogram of the audio signal and depending on the phase values for the plurality of time-frequency bins.

For example, the magnitude spectrogram provides the magnitude value for a particular time-frequency bin and the phase value for the particular time-frequency bin has been reconstructed by the phase reconstructor 140.

Fig. 9 illustrates a system comprising an encoder 210 and a decoder 220 according to an embodiment, wherein the decoder 220 is an apparatus for phase reconstruction according to one of the above-described embodiments.

The encoder 210 is configured to encode a magnitude spectrogram of an audio signal.

The decoder 220 is configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal.

Moreover, the decoder 220 is configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

Furthermore, the decoder 220 is configured to decode the audio signal using the magnitude spectrogram of the audio signal and using the phase values for the plurality of time-frequency bins.

Fig. 10 illustrates an audio encoder 210 and an audio decoder 220 according to embodiments using the magnitude spectrum for transmission.

The audio encoder 210 is configured to generate a magnitude spectrogram of an audio signal for the apparatus for phase reconstruction as described above. In Fig. 10, the decoder 220 may, e.g., be an apparatus for phase reconstruction as described above.

Although some aspects have been described in the context of an apparatus, it is clear that these aspects also represent a description of the corresponding method, where a block or device corresponds to a method step or a feature of a method step. Analogously, aspects described in the context of a method step also represent a description of a corresponding block or item or feature of a corresponding apparatus. Some or all of the method steps may be executed by (or using) a hardware apparatus, like for example, a microprocessor, a programmable computer or an electronic circuit. In some embodiments, one or more of the most important method steps may be executed by such an apparatus.

Depending on certain implementation requirements, embodiments of the invention can be implemented in hardware or in software or at least partially in hardware or at least partially in software. The implementation can be performed using a digital storage medium, for

example a floppy disk, a DVD, a Blu-Ray, a CD, a ROM, a PROM, an EPROM, an EEPROM or a FLASH memory, having electronically readable control signals stored thereon, which cooperate (or are capable of cooperating) with a programmable computer system such that the respective method is performed. Therefore, the digital storage medium may be computer readable.

Some embodiments according to the invention comprise a data carrier having electronically readable control signals, which are capable of cooperating with a programmable computer system, such that one of the methods described herein is performed.

Generally, embodiments of the present invention can be implemented as a computer program product with a program code, the program code being operative for performing one of the methods when the computer program product runs on a computer. The program code may for example be stored on a machine readable carrier.

Other embodiments comprise the computer program for performing one of the methods described herein, stored on a machine readable carrier.

In other words, an embodiment of the inventive method is, therefore, a computer program having a program code for performing one of the methods described herein, when the computer program runs on a computer.

A further embodiment of the inventive methods is, therefore, a data carrier (or a digital storage medium, or a computer-readable medium) comprising, recorded thereon, the computer program for performing one of the methods described herein. The data carrier, the digital storage medium or the recorded medium are typically tangible and/or non-transitory.

A further embodiment of the inventive method is, therefore, a data stream or a sequence of signals representing the computer program for performing one of the methods described herein. The data stream or the sequence of signals may for example be configured to be transferred via a data communication connection, for example via the Internet.

A further embodiment comprises a processing means, for example a computer, or a programmable logic device, configured to or adapted to perform one of the methods described herein.

A further embodiment comprises a computer having installed thereon the computer program for performing one of the methods described herein.

A further embodiment according to the invention comprises an apparatus or a system configured to transfer (for example, electronically or optically) a computer program for performing one of the methods described herein to a receiver. The receiver may, for example, be a computer, a mobile device, a memory device or the like. The apparatus or system may, for example, comprise a file server for transferring the computer program to the receiver.

In some embodiments, a programmable logic device (for example a field programmable gate array) may be used to perform some or all of the functionalities of the methods described herein. In some embodiments, a field programmable gate array may cooperate with a microprocessor in order to perform one of the methods described herein. Generally, the methods are preferably performed by any hardware apparatus.

The apparatus described herein may be implemented using a hardware apparatus, or using a computer, or using a combination of a hardware apparatus and a computer.

The methods described herein may be performed using a hardware apparatus, or using a computer, or using a combination of a hardware apparatus and a computer.

The above described embodiments are merely illustrative for the principles of the present invention. It is understood that modifications and variations of the arrangements and the details described herein will be apparent to others skilled in the art. It is the intent, therefore, to be limited only by the scope of the impending patent claims and not by the specific details presented by way of description and explanation of the embodiments herein.

References:

[1] Aggelos Gkiokas, Vassilios Katsouros, George Carayannis, and Themos Stafylakis, "*Music tempo estimation and beat tracking by applying source separation and metrical relations*", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 421–424.

[2] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner, "*On the reduction of false positives in singing voice detection*", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 7480–7484.

[3] Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama, "*HMM-based approach for automatic chord detection using refined acoustic features*", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, Texas, USA, 2010, pp. 5518–5521.

[4] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama, "*A real-time equalizer of harmonic and percussive components in music signals*", in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Philadelphia, Pennsylvania, USA, 2008, pp. 139–144.

[5] Nobutaka Ono, Kenichi Miyamoto, Jonathan LeRoux, Hirokazu Kameoka, and Shigeki Sagayama, "*Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram*", in European Signal Processing Conference, Lausanne, Switzerland, 2008, pp. 240–244.

[6] Derry Fitzgerald, "*Harmonic/percussive separation using median filtering*", in Proceedings of the International Conference on Digital Audio Effects (DAFX), Graz, Austria, 2010, pp. 246–253.

[7] Scott N. Levine and Julius O. Smith III, "*A sines+transients+noise audio representation for data compression and time/pitch scale modications*", in Proceedings of the AES Convention, 1998.

[8] Tony S. Verma and Teresa H.Y. Meng, "*An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio*", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, Washington, USA, May 1998, pp. 3573–3576.

[9] Laurent Daudet, "*Sparse and structured decompositions of signals with the molecular matching pursuit*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1808–1816, September 2006.

[10] Jonathan Driedger, Meinard Müller, and Sascha Disch, "*Extending harmonic-percussive separation of audio signals*", in Proceedings of the International Conference on Music Information Retrieval (ISMIR), Taipei, Taiwan, 2014, pp. 611–616.

[11] Jeongsoo Park and Kyogu Lee, "*Harmonic-percussive source separation using harmonicity and sparsity constraints*", in Proceedings of the International Conference on Music Information Retrieval (ISMIR), Málaga, Spain, 2015, pp. 148–154.

[12] Josef Bigun and Gösta H. Granlund, "*Optimal orientation detection of linear symmetry*", in Proceedings of the IEEE First International Conference on Computer Vision, London, UK, 1987, pp. 433–438.

[13] Hans Knutsson, "R*epresenting local structure using tensors*", in 6th Scandinavian Conference on Image Analysis, Oulu, Finland, 1989, pp. 244–251.

[14] Chris Harris and Mike Stephens, "*A combined corner and edge detector*", in Proceedings of the 4th Alvey Vision Conference, Manchester, UK, 1988, pp. 147–151.

[15] Rolf Bardeli, "*Similarity search in animal sound databases*", IEEE Transactions on Multimedia, vol. 11, no. 1, pp. 68–76, January 2009.

[16] Matthias Zeppelzauer, Angela S. Stöger, and Christian Breiteneder, "*Acoustic detection of elephant presence in noisy environments*", in Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data, Barcelona, Spain, 2013, pp4. 3–8.

[17] Hanno Scharr, "*Optimale Operatoren in der digitalen Bildverarbeitung*", Dissertation, IWR, Fakultät für Physik und Astronomie, Universität Heidelberg, Heidelberg, Germany, 2000.

[18] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "*Performance measurement in blind audio source separation*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] Daniel W. Griffin and Jae S. Lim, "*Signal estimation from modified short-time Fourier transform*", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 2, pp. 236–243, 1984.

[20] Paul Magron, Roland Badeau, and Bertrand David, *"Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration*", in Signal Processing Conference (EUSIPCO), 2015 23rd European. IEEE, 2015, pp. 1–5.

[21] Byung-Jun Yoon and Henrique S Malvar, "*Coding overcomplete representations of audio using the mclt*", in Data Compression Conference, 2008. DCC 2008. IEEE, 2008, pp. 152–161.

Claims

1. An apparatus for phase reconstruction from a magnitude spectrogram of an audio signal, comprising:

a frequency change determiner (110) being configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal, and

a phase reconstructor (140) being configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

2. An apparatus according to claim 1,

wherein the phase reconstructor (140) is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times.

3. An apparatus according to claim 1 or 2,

wherein the phase reconstructor (140) is configured to generate phase values for the plurality of time-frequency bins depending on the formula

$$\phi(bH) = \phi(b_0 H) + \sum_{n=b_0}^{b} \left(\phi_b(b_0) + \sum_{m=b_0}^{n} \phi_{bb}(m) \right)$$

wherein b_0 is an index indicating an analysis block of a plurality of blocks,

wherein b is a further index indicating a further analysis block of the plurality of blocks,

wherein H indicates a hop size, and

wherein $\phi(bH)$ and $\phi_b(b_0)$ and $\phi_{bb}(m)$ indicate phase values.

4. An apparatus according to one of the preceding claims,

wherein the frequency change determiner (110) is configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on an angle ($\alpha(b,k)$) for said time-frequency bin, wherein the angle ($\alpha(b,k)$) for said time-frequency bin depends on the magnitude spectrogram of the audio signal.

5. An apparatus according to claim 4,

wherein the frequency change determiner (110) is configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins further depending on a sampling frequency (f_s) of the audio signal, and depending on a length (N) of an analysis window and depending on a hop size (H) of the analysis window.

6. An apparatus according to claim 5,

wherein the apparatus frequency change determiner (110) is configured to determine the change of the frequency for each time-frequency bin of the plurality of time-frequency bins depending on the formula

$$R(b,k) = \frac{{f_s}^2}{HN} \cdot \tan(\alpha(b,k))$$

wherein (b,k) indicates a time-frequency bin of the plurality of time-frequency bins,

wherein $R(b,k)$ indicates the change of the frequency for said time-frequency bin (b,k),

wherein b indicates time,

wherein k indicates frequency,

wherein f_s indicates the sampling frequency of the audio signal,

wherein N indicates the length of the analysis window,

wherein H indicates the hop size of the analysis window, and

wherein $\alpha(b,k)$ indicates the angle for said time-frequency bin (b,k), wherein the angle $\alpha(b,k)$ depends on the magnitude spectrogram.

7. An apparatus according to one of claims 4 to 6,

wherein the frequency change determiner (110) is configured to determine a partial derivative ($\mathbf{S}_b$) of the magnitude spectrogram ($\mathbf{S}$) of the audio signal with respect to a time index,

wherein the frequency change determiner (110) is configured to determine a partial derivative ($\mathbf{S}_k$) of the magnitude spectrogram ($\mathbf{S}$) of the audio signal with respect to a time index, and

wherein the frequency change determiner (110) is configured to determine a structure tensor ($\mathbf{T}(b,k)$) for each time-frequency bin ((b,k)) of the plurality of time-frequency bins depending on the partial derivative ($\mathbf{S}_b$) of the magnitude spectrogram ($\mathbf{S}$) of the audio signal with respect to the time index and depending on the partial derivative ($\mathbf{S}_k$) of the magnitude spectrogram ($\mathbf{S}$) of the audio signal with respect to the frequency index, and

wherein the frequency change determiner (110) is configured to determine the angle ($\alpha(b,k)$) for each time-frequency bin ((b,k)) of the plurality of time-frequency bins depending the structure tensor ($\mathbf{T}(b,k)$) for said time-frequency bin ((b,k)).

8. An apparatus according to claim 7,

wherein the frequency change determiner (110) is configured to determine the angle ($\alpha(b,k)$) for each time-frequency bin ((b,k)) of the plurality of time-frequency bins by determining two components of $v_1(b,k)$ and $v_2(b,k)$ an eigenvector $\mathbf{v}(b,k)$ of the structure tensor($\mathbf{T}(b,k)$) of said time-frequency bin ((b,k)), and by determining the angle ($\alpha(b,k)$) for said time-frequency bin ((b,k)) according to

$$\alpha(b,k) = \operatorname{atan}\left(\frac{v_2(b,k)}{v_1(b,k)}\right) \in [-\pi/2;\pi/2]$$

wherein $\alpha(b,k)$ indicates the angle for said time-frequency bin ((b,k)),

wherein b indicates time,

wherein k indicates frequency, and

wherein atan() indicates an inverse tangent function.

9. An apparatus according to one of the preceding claims, wherein the phase reconstructor (140) being configured to generate phase values for each time-frequency bin of the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of time-frequency bins.

10. An apparatus according to claim 7 or 8, wherein the phase reconstructor (140) is configured to determine a measure of anisotropy for each time-frequency bin (b,k) of the plurality of time-frequency bins depending on the change of the frequency determined for each time-frequency bin of the plurality of time-frequency bins.

11. An apparatus according to claim 10,

wherein the phase reconstructor (140) is configured to determine the measure of anisotropy for each time-frequency bin (b,k) of the plurality of time-frequency bins depending on at least one of the formulae:

$$\left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2$$

and

$$\mu(b,k) + \lambda(b,k) \geq e,$$

wherein $\mu(b,k)$ is a first eigenvalue $\lambda(b,k)$ is a second eigenvalue of the structure tensor ($\mathbf{T}(b,k)$) of said time-frequency bin (b,k), and $e \in \mathbb{R}^{>0}$,

wherein the phase reconstructor (140) is configured to assign each time-frequency bin of the plurality of time-frequency bins to a signal component group of two or more signal component groups further depending on the change of the measure of anisotropy.

12. An apparatus according to claim 11,

wherein the phase reconstructor (140) is configured to determine the measure of anisotropy for said time-frequency bin (b, k) depending on the formula:

$$C(b,k) = \begin{cases} \left(\frac{\mu(b,k)-\lambda(b,k)}{\mu(b,k)+\lambda(b,k)}\right)^2 & , \mu(b,k)+\lambda(b,k) \geq e \\ 0 & , \text{ else} \end{cases}$$

wherein $C(b, k)$ is the measure of anisotropy depending for said time-frequency bin (b, k), and

wherein the phase reconstructor (140) is configured to assign said time-frequency bin (b, k) to a residual component group of the two or more signal component groups, if the measure of anisotropy $C(b, k)$ is smaller than an upper threshold value c, or wherein the phase reconstructor (140) is configured to assign said time-frequency bin (b, k) to the residual component group of the two or more signal component groups, if the measure of anisotropy $C(b, k)$ is smaller than or equal to the upper threshold value c,

wherein $c \in \mathbb{R}^{>0}$.

13. An apparatus according to one of claims 9 to 12,

wherein the phase reconstructor (140) is configured to determine a phase reconstruction mode for each time-frequency bin of the plurality of time-frequency bins depending on the measure of anisotropy, whether phase reconstruction shall be conducted,

wherein the phase reconstructor (140) is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times, if the phase reconstruction mode determined

by the phase reconstructor (140) for said time-frequency bin indicates a first mode, and

wherein the phase reconstructor (140) is configured to determine a phase for each time-frequency bin of the plurality of time-frequency bins by integrating the changes of the frequency two times, if the phase reconstruction mode determined by the phase reconstructor (140) for said time-frequency bin indicates a second mode, being different from the first mode.

14. An apparatus according to claim 13,

wherein the phase reconstructor (140) is configured to determine the phase reconstruction mode for each time-frequency bin of the plurality of time-frequency bins so that the phase reconstruction mode indicates the first mode, if the measure of anisotropy for said time-frequency bin ((b,k)) is greater than a lower threshold value (c), and so that the phase reconstruction mode indicates the second mode, if the measure of anisotropy for said time-frequency bin ((b,k)) is smaller than or equal to the lower threshold value (c), or

wherein the phase reconstructor (140) is configured to determine the phase reconstruction mode for each time-frequency bin of the plurality of time-frequency bins so that the phase reconstruction mode indicates the first mode, if the measure of anisotropy for said time-frequency bin ((b,k)) is greater than or equal to the lower threshold value (c), and so that the phase reconstruction mode indicates the second mode, if the measure of anisotropy for said time-frequency bin ((b,k)) is smaller than the lower threshold value (c).

15. An apparatus according to one of the preceding claims,

wherein the phase reconstructor (140) is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating two times over frequency.

16. An apparatus according to one of the preceding claims,

wherein the phase reconstructor (140) is configured to generate the phase value for each time-frequency bin of the plurality of time-frequency bins by integrating two times over time.

17. An apparatus according to one of the preceding claims, wherein the apparatus comprises a signal generator (150) being configured to generate an audio output signal depending on the magnitude spectrogram of the audio signal and depending on the phase values for the plurality of time-frequency bins.

18. A system, comprising:

an encoder (210) for encoding a magnitude spectrogram of an audio signal, and

a decoder (220) being an apparatus according to one of the preceding claims, for decoding the audio signal,

wherein the decoder (220) is configured to determine a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal,

wherein the decoder (220) is configured to generate phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins, and

wherein the decoder (220) is configured to decode the audio signal using the magnitude spectrogram of the audio signal and using the phase values for the plurality of time-frequency bins.

19. An encoder (210) being configured to generate a magnitude spectrogram of an audio signal for the apparatus for phase reconstruction according to one of claims 1 to 17.

20. A method for phase reconstruction from a magnitude spectrogram of an audio signal, comprising:

determining a change of a frequency for each time-frequency bin of a plurality of time-frequency bins of the magnitude spectrogram of the audio signal depending on the magnitude spectrogram of the audio signal, and

generating phase values for the plurality of time-frequency bins depending on the changes of the frequencies determined for the plurality of the time-frequency bins.

21. A computer program for implementing the method of claim 20 when being executed on a computer or signal processor.

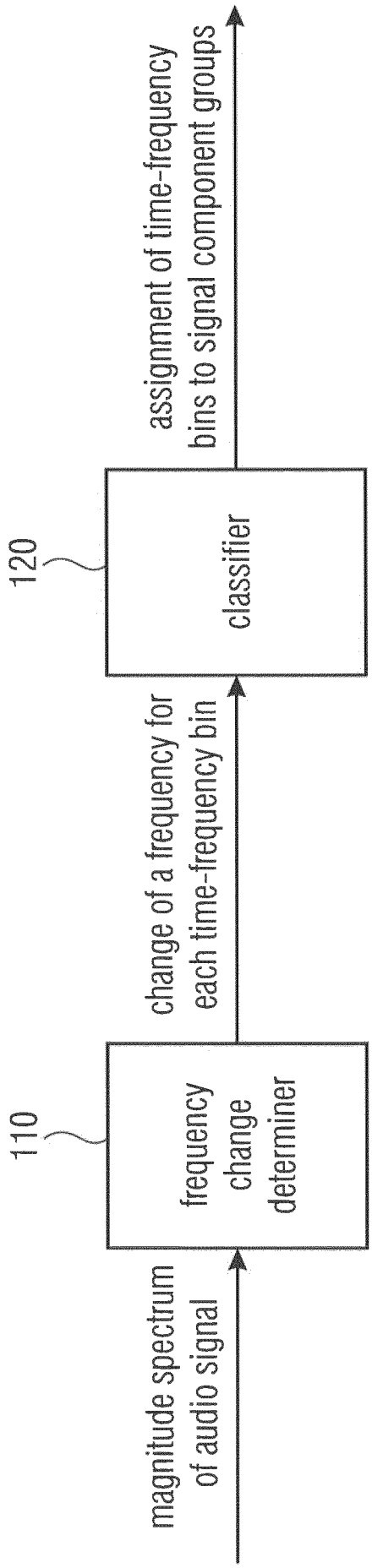

magnitude spectrum of audio signal
110
frequency change determiner
change of a frequency for each time-frequency bin
120
classifier
assignment of time-frequency bins to signal component groups

Fig. 1

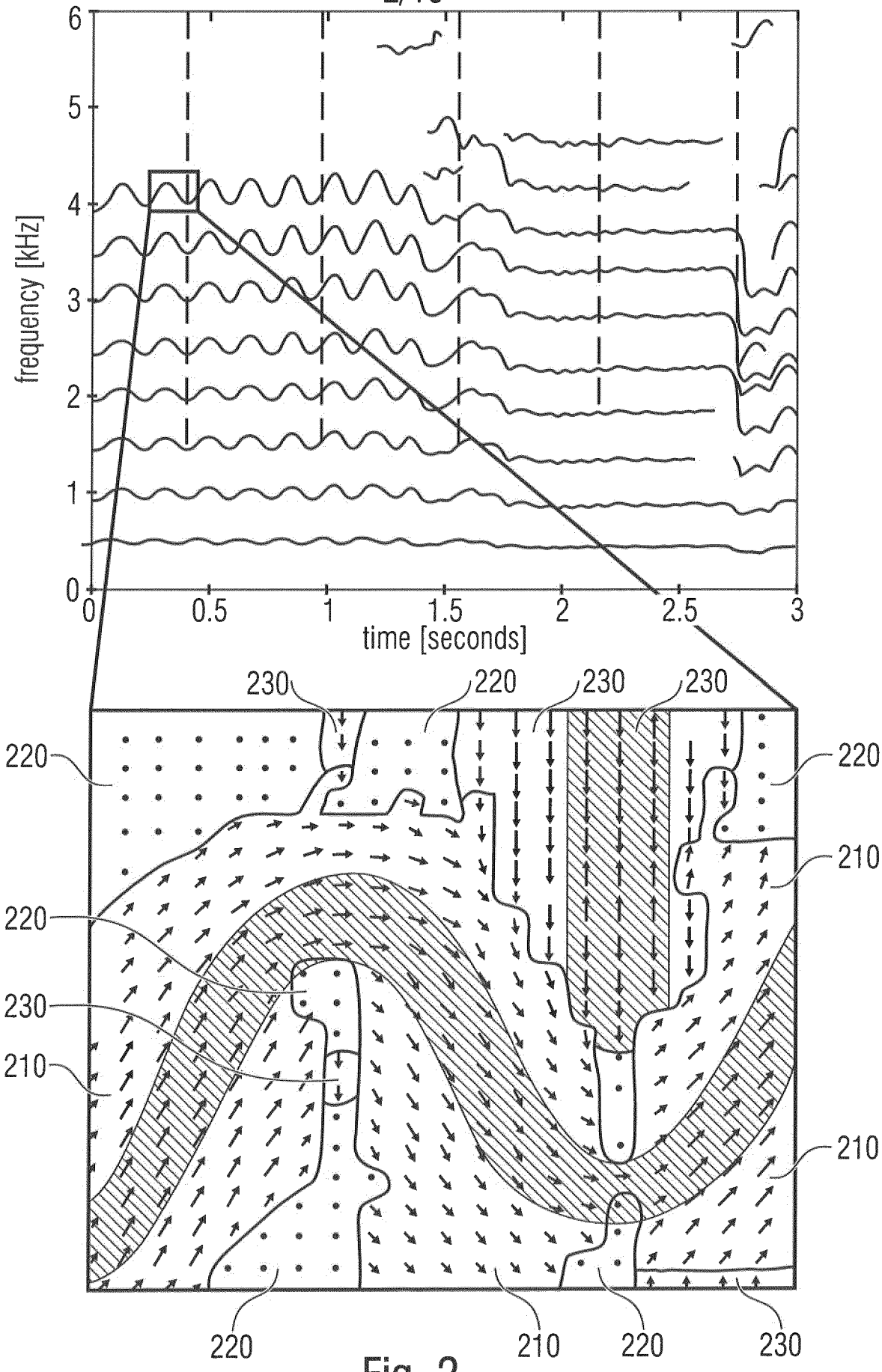

frequency [kHz]
time [seconds]
0
1
2
3
4
5
6
0
0.5
1
1.5
2
2.5
3
230
220
230
230
220
220
210
220
230
210
210
220
210
220
230

Fig. 2

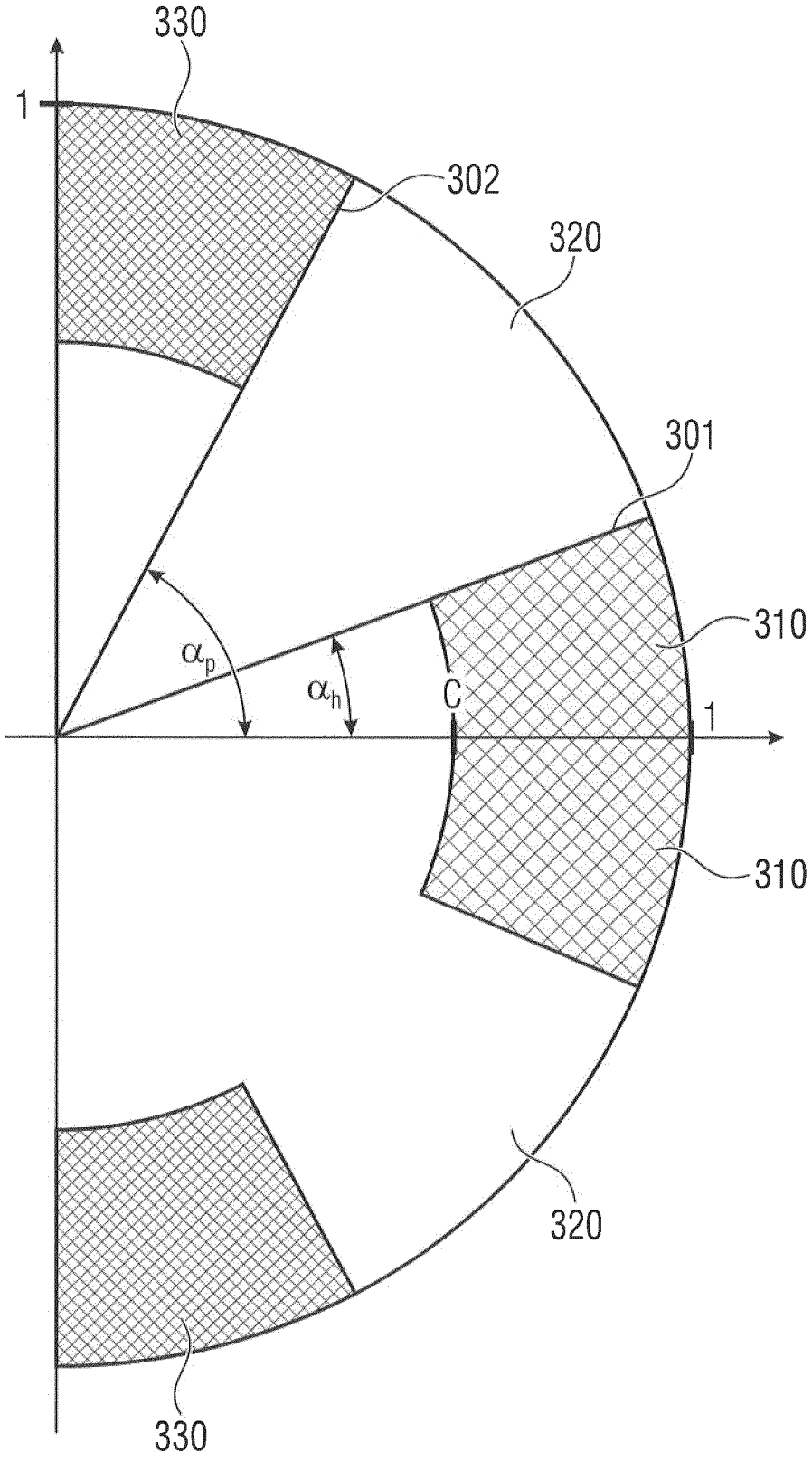


Fig. 3

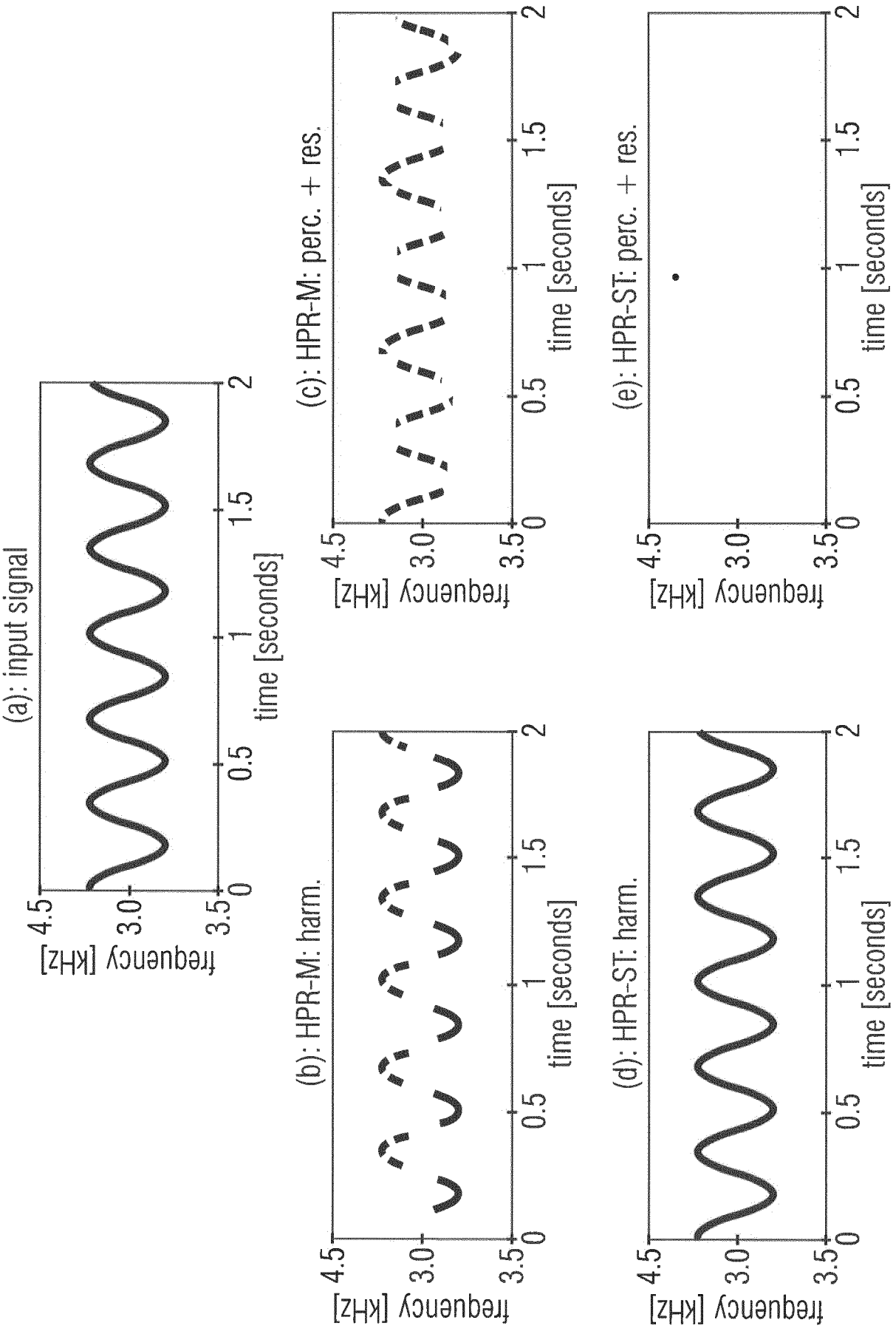

(a): input signal
(b): HPR-M: harm.
(c): HPR-M: perc. + res.
(d): HPR-ST: harm.
(e): HPR-ST: perc. + res.
frequency [kHz]
time [seconds]
4.5
3.0
3.5
0
0.5
1
1.5
2

Fig. 4

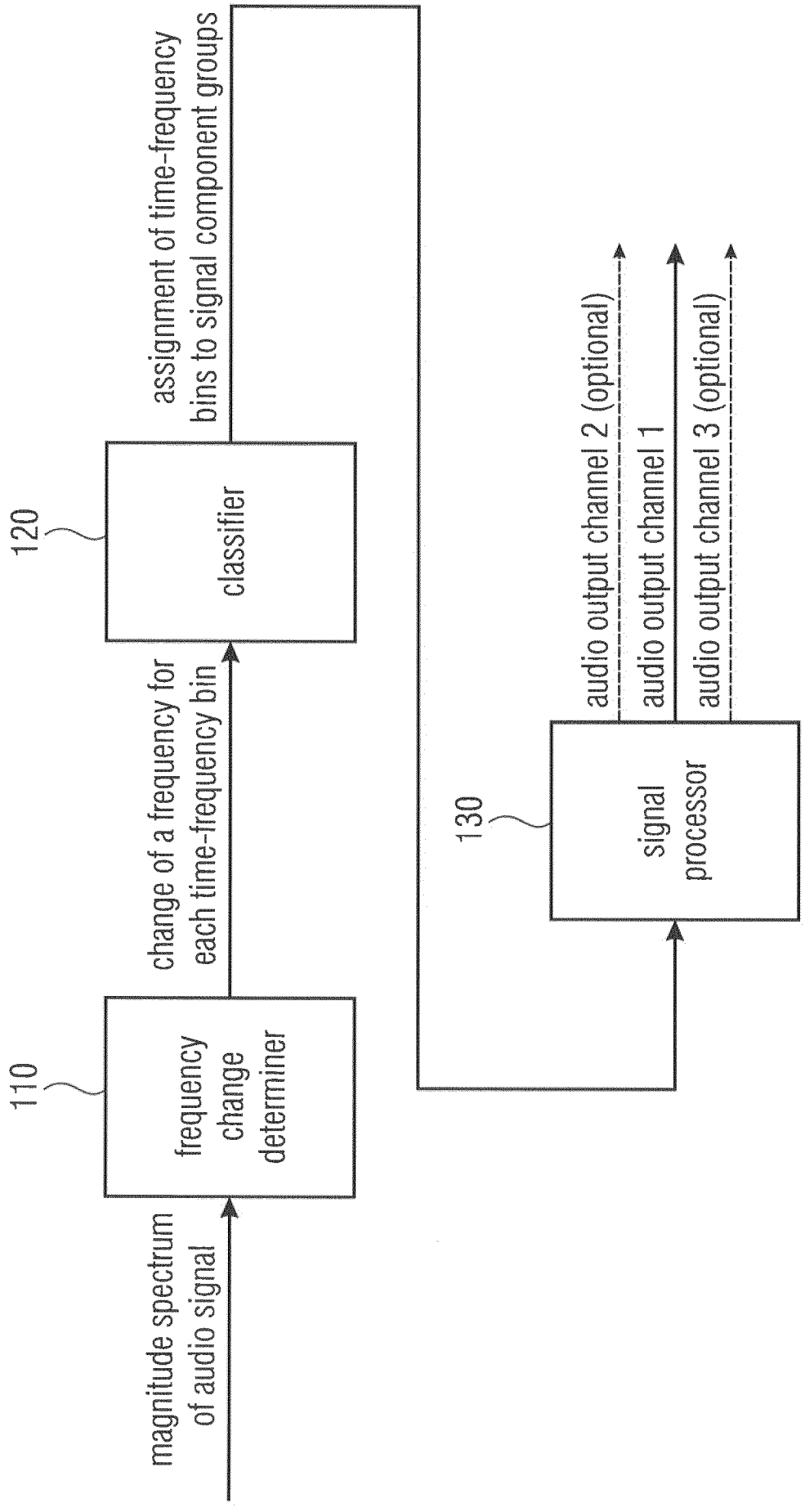

magnitude spectrum of audio signal
110
frequency change determiner
change of a frequency for each time-frequency bin
120
classifier
assignment of time-frequency bins to signal component groups
130
signal processor
audio output channel 2 (optional)
audio output channel 1
audio output channel 3 (optional)

Fig. 5

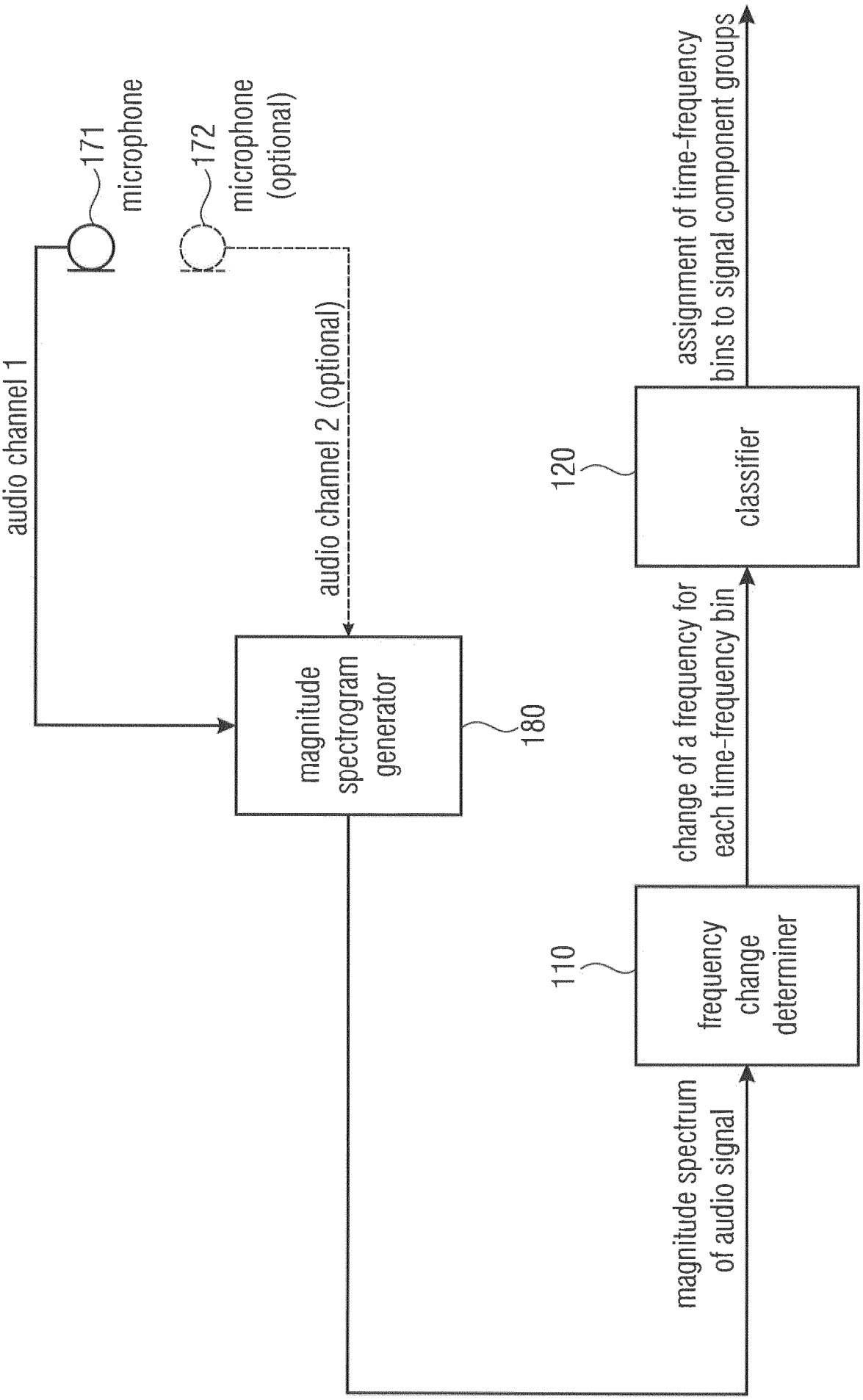

171
microphone
172
microphone
(optional)
audio channel 1
audio channel 2 (optional)
magnitude
spectrogram
generator
180
120
classifier
assignment of time-frequency
bins to signal component groups
change of a frequency for
each time-frequency bin
110
frequency
change
determiner
magnitude spectrum
of audio signal

Fig. 6

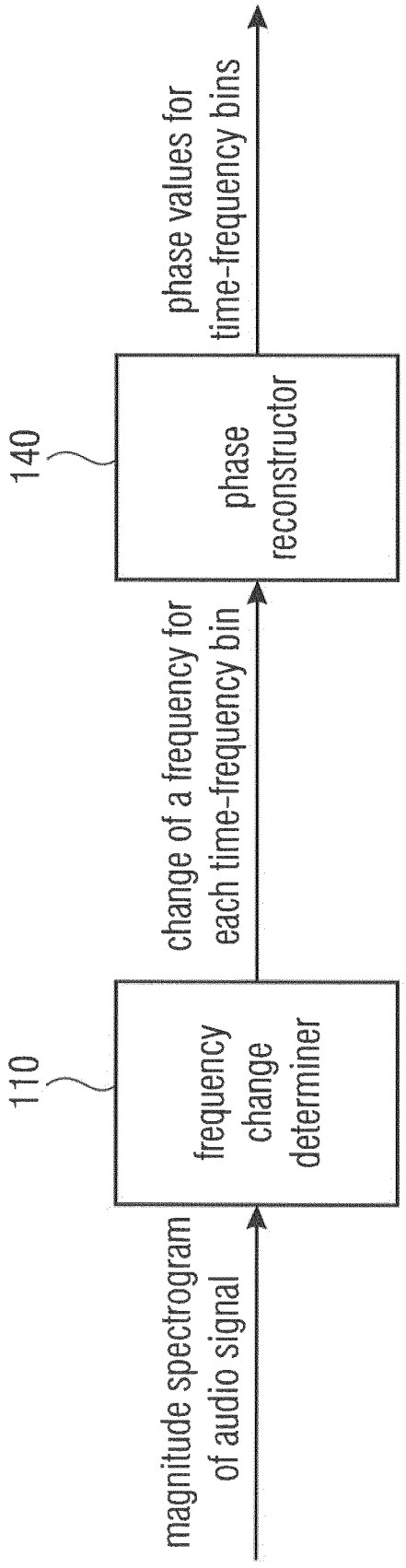

magnitude spectrogram of audio signal
110
frequency change determiner
change of a frequency for each time-frequency bin
140
phase reconstructor
phase values for time-frequency bins

Fig. 7

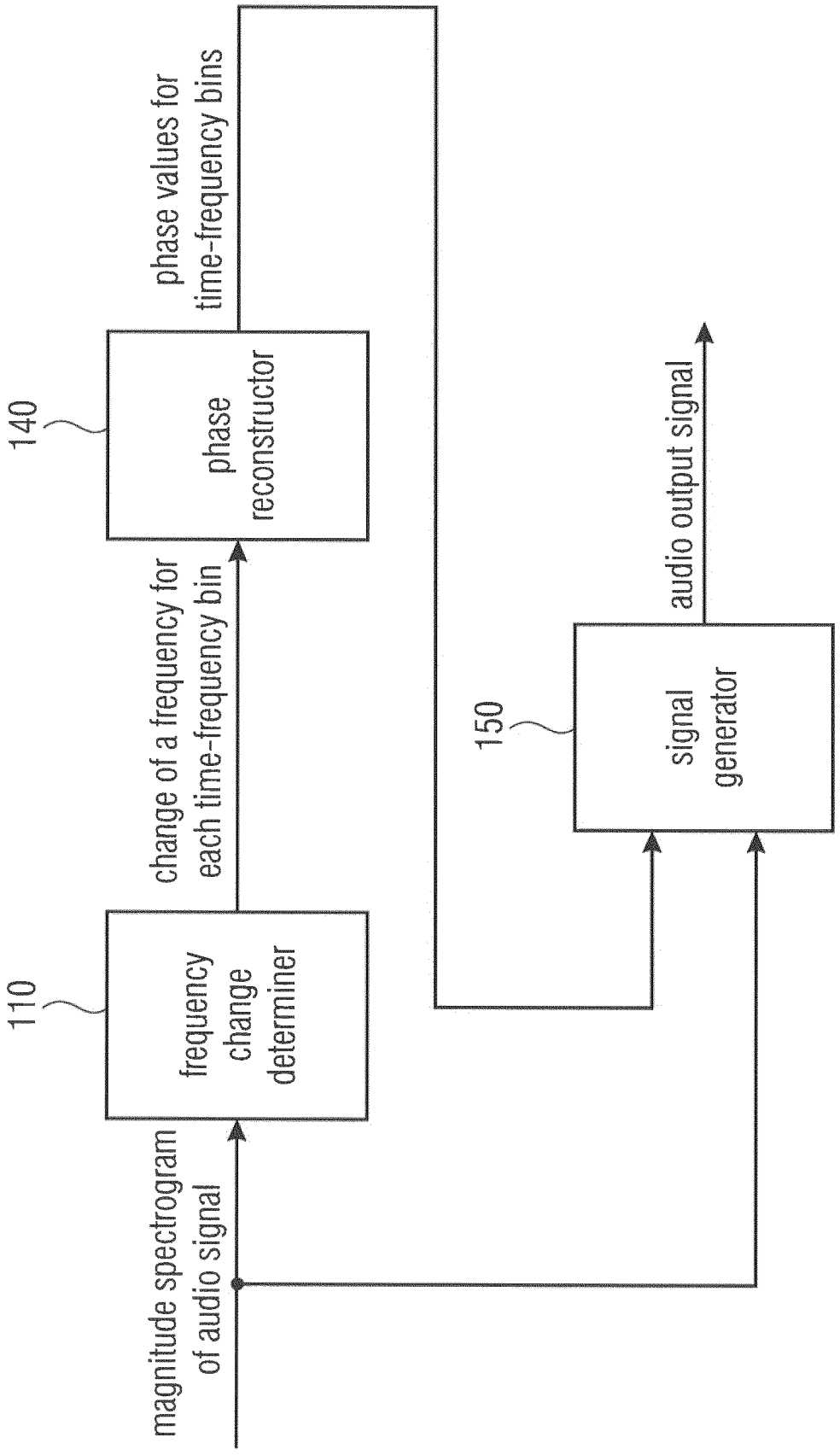

magnitude spectrogram of audio signal
110
frequency change determiner
change of a frequency for each time-frequency bin
140
phase reconstructor
phase values for time-frequency bins
150
signal generator
audio output signal

Fig. 8

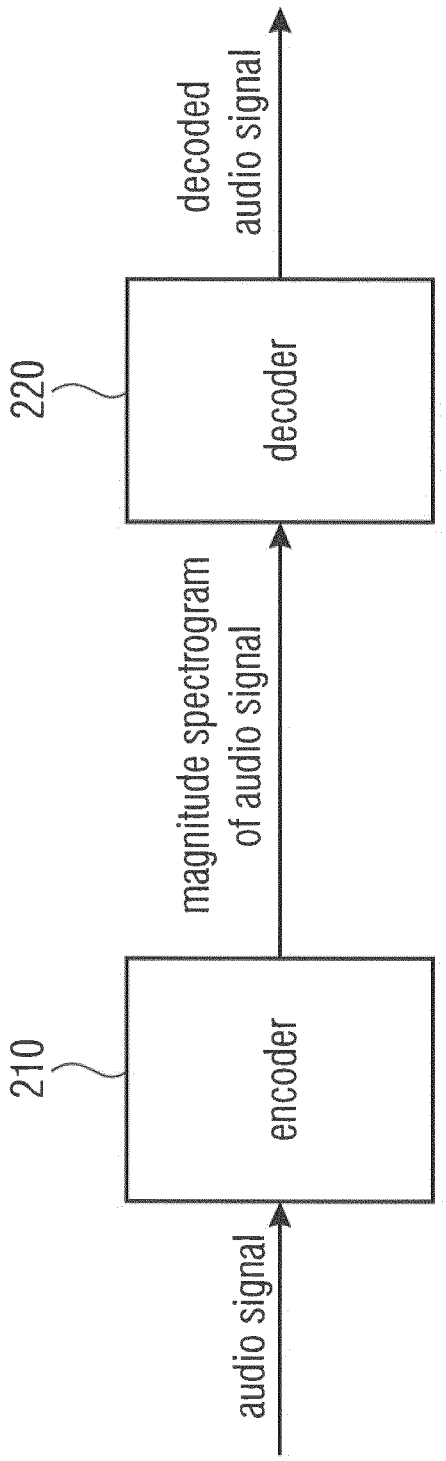

audio signal
210
encoder
magnitude spectrogram of audio signal
220
decoder
decoded audio signal

Fig. 9

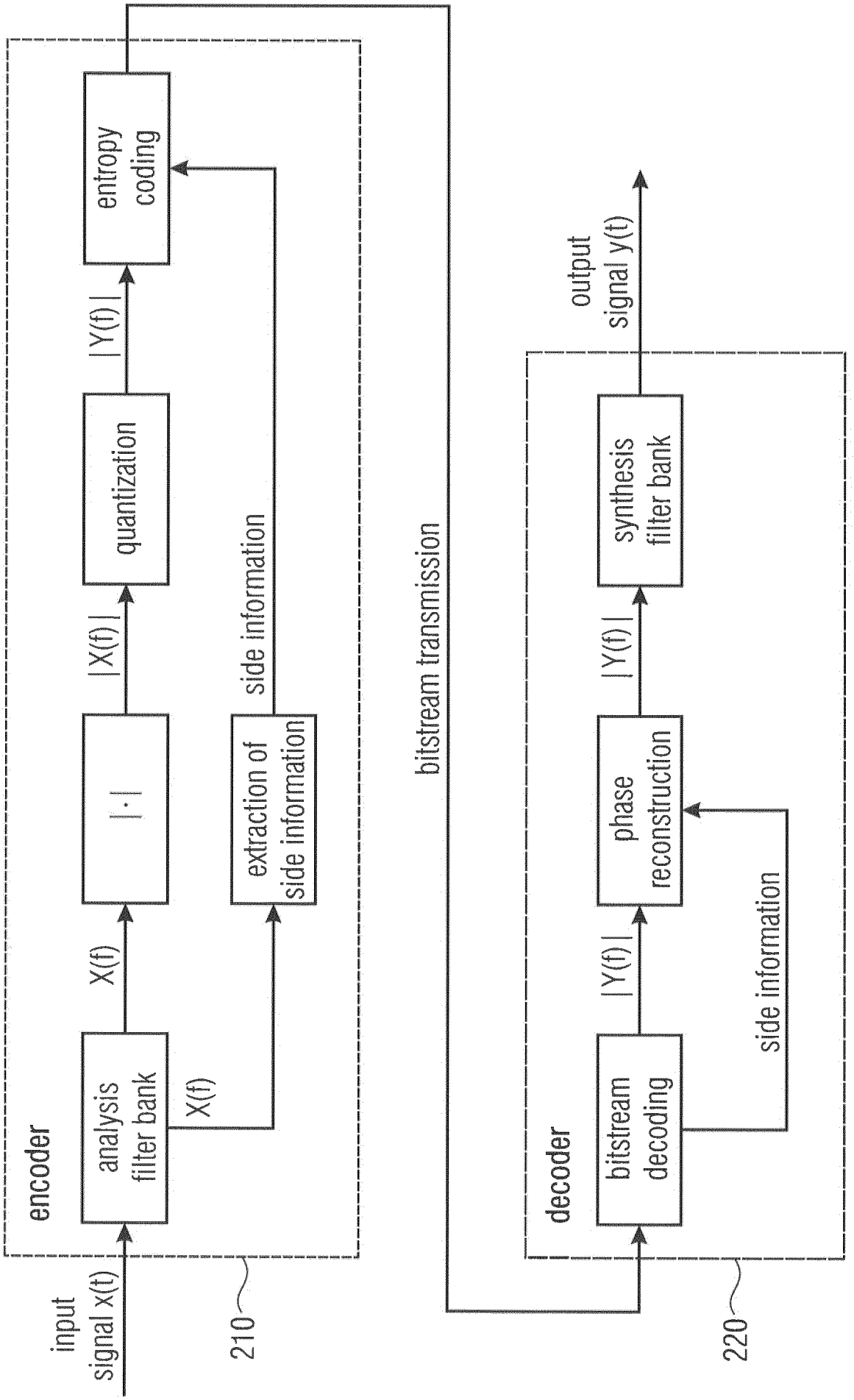

encoder
analysis filter bank
X(f)
|·|
|X(f)|
quantization
|Y(f)|
entropy coding
X(f)
extraction of side information
side information
bitstream transmission
input signal x(t)
210
decoder
bitstream decoding
|Y(f)|
phase reconstruction
|Y(f)|
synthesis filter bank
output signal y(t)
side information
220

Fig. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/056263

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L19/00
ADD. G10H1/00 G10L21/00

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)
G10L G10H G06F H04R H04N G11B H04S

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.
X	MAGRON PAUL ET AL: "Phase reconstruction of spectrograms with linear unwrapping: Application to audio signal restoration", 2015 23RD EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO), EURASIP, 31 August 2015 (2015-08-31), pages 1-5, XP032836338, DOI: 10.1109/EUSIPCO.2015.7362333 [retrieved on 2015-12-22] cited in the application	1,2,19
Y	computing phase;	18,20,21
A	page 2; table equation (6) phase reconstruction; paragraph [04.2] - paragraph [04.4] onset phase reconstruction; page 3, left-hand column, paragraph 3; tables Equations (10), (11)	3-17

-/--

[X] Further documents are listed in the continuation of Box C.

[X] 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
14 June 2017	26/06/2017

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	Glasser, Jean-Marc

4

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CHANG-CHUN BAO ET AL: "A blind bandwidth extension method for audio signals based on phase space reconstruction", EURASIP JOURNAL ON AUDIO SPEECH AND MUSIC PROCESSING, vol. 2014, no. 1, 1 December 2014 (2014-12-01), XP055381022, US ISSN: 1687-4714, DOI: 10.1186/1687-4722-2014-1	18,20,21
A	blind bandwidth extension (BWE) of audio signals.; paragraph [0005]	1-17

X	AU 2013 203 159 B2 (FRAUNHOFER GES FORSCHUNG) 17 September 2015 (2015-09-17) page 4; figure 18a Audio encoder generates magnitude spectrogram; page 4; claim 1	19

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2017/056263

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
AU 2013203159 B2	17-09-2015	NONE	