



(11)

EP 2 912 660 B1

(12)

EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:

11.01.2017 Bulletin 2017/02

(21) Application number: **12794680.4**

(22) Date of filing: **21.11.2012**

(51) Int Cl.:

G10L 21/0232 ^(2013.01)

(86) International application number:

PCT/EP2012/073149

(87) International publication number:

WO 2014/079484 (30.05.2014 Gazette 2014/22)

(54) **METHOD FOR DETERMINING A DICTIONARY OF BASE COMPONENTS FROM AN AUDIO SIGNAL**

VERFAHREN ZUM BESTIMMEN EIN WÖRTERBUCH VON BASISKOMPONENTEN AUS EINEM AUDIOSIGNAL

PROCEDE POUR DETERMINER UNE DICTIONNAIRE DE COMPOSANTS DE BASE D'UN SIGNAL AUDIO

(84) Designated Contracting States:

AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(43) Date of publication of application:

02.09.2015 Bulletin 2015/36

(73) Proprietor: **Huawei Technologies Co., Ltd.**

**Longgang District
Shenzhen, Guangdong 518129 (CN)**

(72) Inventors:

- **JODER, Cyril**
81825 München (DE)
- **WENNINGER, Felix**
80335 München (DE)
- **SCHULLER, Björn**
82205 Gilching (DE)
- **VIRETTE, David**
80992 München (DE)

(74) Representative: **Pfenning, Meinig & Partner mbB**

**Patent- und Rechtsanwälte
Theresienhöhe 11a
80339 München (DE)**

(56) References cited:

- **MANUEL MOUSSALLAM ET AL: "Audio source separation informed by redundancy with greedy multiscale decompositions", SIGNAL PROCESSING CONFERENCE (EUSIPCO), 2012 PROCEEDINGS OF THE 20TH EUROPEAN, IEEE, 27 August 2012 (2012-08-27), pages 2644-2648, XP032254797, ISBN: 978-1-4673-1068-0**
- **ROBERT PEHARZ ET AL: "Sparse nonnegative matrix factorization with -constraints", NEUROCOMPUTING, vol. 80, 15 March 2012 (2012-03-15), pages 38-46, XP028356707, ISSN: 0925-2312, DOI: 10.1016/J.NEUCOM.2011.09.024 [retrieved on 2011-11-11]**
- **NASSER MOHAMMADIHA ET AL: "Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints", SIGNAL PROCESSING AND INFORMATION TECHNOLOGY (ISSPIT), 2009 IEEE INTERNATIONAL SYMPOSIUM ON, IEEE, PISCATAWAY, NJ, USA, 14 December 2009 (2009-12-14), pages 418-423, XP031624854, ISBN: 978-1-4244-5949-0**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

Description

BACKGROUND OF THE INVENTION

[0001] The present invention relates to a method and a device for determining a dictionary of base components from an input signal. In particular, the present invention relates to the processing of an acoustic signal input for the estimation of a feature vector dictionary for describing acoustic sources.

[0002] Most audio signals are composed of a plurality of individual sound sources. Musical recordings, for example, comprise most of the time several instruments. In the case of speech communication, the signal often comprises, in addition to the speech itself, other interfering sounds which are recorded by the same microphone. Such interfering sounds can be for example ambient noise or other people talking in the same room.

[0003] Several applications would take advantage of the separation of audio signal into several parts. One of them is the reduction of acoustic noise in telephonic communication, especially in the case of hand-free system where the noise level is often high because of the distance between the microphone and the speaker. Another usage of source separation is the extraction of some target instrument from musical signals, for karaoke or remixing application.

[0004] Non-negative Matrix Factorisation (NMF) has been first proposed by Paatero: "Least Squares Formulation of Robust Non-Negative Factor Analysis", Chemometrics and Intelligent Laboratory Systems 37, pp. 23-35, 1997 and has been successfully applied to a wide variety of applications since then. In particular, this technique has become a standard method for audio source separation, where an input audio signal is to be separated into several signals corresponding to the different acoustic sources. It is based on a decomposition of the power spectrogram of the mixture into a non-negative combination of several spectral bases, each associated to one of the present sources. Non-negative Matrix Factorization (NMF) methods have been used in that context with relatively good results. Indeed, the non-negative constraint which is inherent to this technique complies with the structure of the audio spectrograms, and can allow for the decomposition of a sound into some meaningful components. These components form a dictionary of spectral bases which describe the signal. The decomposition typically aims to estimate spectral bases corresponding to different "parts" of the spectrogram, e.g. different sounds or speakers. A separation of these parts can then be performed by a partial reconstruction of the signal, considering only the wanted components. This technique has been applied by C. Joder, F. Weninger, F. Eyben, D. Virette, B. Schuller "Real-time Speech Separation by Semi-Supervised Nonnegative Matrix Factorization", Proc. International Conference on Latent Variable Analysis and Signal Separation, March 2012, in particular, to the separation of a target speaker from noisy recordings.

[0005] The basic principle of NMF-based audio processing 100 as schematically illustrated in Fig. 1 is to find a locally optimal factorization of a short-time magnitude spectrogram V 103 of an audio signal 101 into two factors W and H , of which the first one W represents the spectra of the events occurring in the signal 101 and the second one H their activation over time. The first factor W describes the component spectra of the source model 109. The second factor H describes the activations 107 of the signal spectrogram 103 of the audio signal 101. The first factor W and the second factor H are matched with the short-time magnitude spectrogram V 103 of the audio signal 101 by an optimization procedure. The source model 109 is pre-defined when applying supervised NMF and a joint estimation is applied for the source model 109 when using unsupervised NMF. The source signal or signals 113 can be derived from the source spectrogram 111.

[0006] The conventional formulation of NMF is defined as follows. The matrix V defines a $m \times n$ matrix of non-negative real values. The goal is to approximate this matrix by the product of two other non-negative matrices $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$, where $r \ll m, n$ holds. In mathematical terms, a cost function is minimized, measuring the so called "reconstruction error"

$$D(V, W \cdot H),$$

where the term D describes some distance or divergence function. When processing sounds, the input matrix V is given by the succession of short-time magnitude (or power) spectra of the input signal, each column of the matrix containing the values of the spectrum computed at a specific instance in time. In general, these features are given by a short-time Fourier transform of the input signal, after some window function is applied to it. This matrix contains only non-negative values, because of the kind of features used. Typically, the values of the matrices W and H which are estimated by the NMF are initialized by a random number generator and then updated by an iterative process. However, the initial values can also be set according to some prior knowledge of the signal. In particular for an implementation in an on-line system, several decompositions are performed on successive mid-term windows of the signal as shown by C. Joder, F. Weninger, F. Eyben, D. Virette, B. Schuller: "Real-time Speech Separation by Semi-Supervised Nonnegative Matrix Factorization", Proc. of LVA/ICA 2012, Springer, p. 322-329. Then, a faster convergence can be obtained by initializing the matrices

according to the output of the previous decomposition.

[0007] Similarly, some of the spectral basis can be set to a constant value, fixed by a prior learning. This can be beneficial if one of the sources is known and sufficient data is available to estimate the characteristic spectra of this source. In this case, the corresponding columns of W are not updated. The methods wherein the matrix W is entirely constant during the decomposition and the method in which the matrix W is entirely updated are called supervised NMF and unsupervised NMF, respectively. In the case where only a part of the spectral basis is updated, the method is called semi-supervised NMF.

[0008] The NMF decomposition is illustrated in Fig. 2 by a simple example. The figure represents a spectrogram 201 represented by the matrix V , a matrix of two spectral bases 202 represented by the matrix W and the corresponding temporal weights 203 represented by the matrix H . The greyscale of the spectrogram 201 represents the amplitude of the Fourier coefficients. The spectrogram defines an acoustic scene which can be described as the superposition of two so called "atomic sounds". By applying a two-component NMF to this spectrogram, the matrices W and H as defined in Fig. 2 can be obtained. Each column of W can be interpreted as a basis function for the spectra contained in V , when weighted with the corresponding values of H .

[0009] Since the spectral bases are non-negative, they correspond to proper magnitude spectra, which can then be used to reconstruct each of the so called "atomic sounds". The example of Fig. 2 is simplistic; however the NMF method can provide satisfactory results in separating different sound sources from realistic recordings. In these cases, a larger value of the order of decomposition r is used. Then, each "component", i.e. the product of one spectral basis with the corresponding temporal weights, is assigned to a specific source. The estimated spectrogram of each source is finally obtained by the sum of all the components attributed to the source.

[0010] However, in the conventional NMF method, the estimation of the dictionary of spectral bases often suffers from some inaccuracies and results in components representing several sources at the same time. Indeed, this method minimizes a reconstruction error between the original input and the decomposition, without taking into account the structure of the individual signals. As a result, the estimated bases can capture some unstructured so called "building blocks" which can be used to reconstruct several sources, whereas the goal is to match each basis to a specific source. In order to overcome this problem, several modifications of the standard NMF method have been proposed, which impose a structure by favoring some properties of the decomposition, such as temporal continuity or component sparsity.

[0011] The sparsity property relates to the fact that the proportion of elements with non-zero value or, more generally, of non-negligible value is very small. In particular, the sparsity of the component activations is often enforced. This property relates to the fact that few components are active at the same time.

[0012] A simple example of the usefulness of a sparsity constraint is represented in Fig. 3. The spectrogram 300 corresponds to the succession of two musical notes, the second one having a pitch one octave higher than the first one. The plots 301 and 302 are the respective spectrograms of these two notes. However, without any constraint on the structure of the decomposition, an NMF factorization with order $r=2$ applied to the spectrogram 300 can also result in the estimation of the spectrograms 303 and 304, since they yield the same perfect reconstruction of the original signal. Enforcing the sparsity property would favor the first decomposition.

[0013] This constraint is generally achieved by adding a penalty term in the cost function to be minimized. The cost function then becomes

$$D(V, W \cdot H) + \lambda f(H)$$

where λ is a real non-negative parameter and f is a function measuring the sparsity of the matrix H . The use of the "pure sparsity measure", that is the number of positive elements in the decomposition, as a penalty in the NMF generally leads to an intractable problem because of its lack of regularity. Thus, the common practice is to approximate this measure with the L^1 norm, also called the Manhattan distance according to A. Cichocki, R. Zdunek, S. Amari, "New Algorithms for Non-negative Matrix Factorization in Application to Blind Source Separation", Proc. of IEEE ICASSP 2006. Other variants of this criterion have also been employed, such as a normalized version of the L^1 norm according to T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria", IEEE Trans. on Audio, Speech and Signal Process., vol. 15(3), pp. 1066-1074, 2007 or the ratio between the L^1 and the L^2 norm, according to P. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints", Journal of Machine Learning Research, Vol. 5, pp. 1457-1469, 2004.

Document "Audio source separation informed by redundancy with greedy multiscale decompositions" (Munuel Mousallam et al, 2012-08-27, pages 2644-2648, XP032254797) describes an algorithm for audio source separation of repeated musical patterns. A Time-Frequency mask usually based on the power spectral density of the mixtures is constructed for the repeating musical background and the separation is performed by means of Wiener filtering relative to this mask.

Document "Sparse nonnegative matrix factorization with constraints" (Robert Peharz et al, 2012-03-15, XP028356707) discloses nonnegative matrix factorization to factorize a nonnegative matrix X into a product of nonnegative matrices W and H with ℓ^1 -constraints.

SUMMARY OF THE INVENTION

[0014] It is the object of the invention to provide a concept for improving sound source estimation when using Non-Negative Matrix Factorization decompositions.

[0015] This object is achieved by the features of the independent claims. Further implementation forms are apparent from the dependent claims, the description and the figures.

[0016] The invention is based on the finding that sound source estimation is improved when a Wiener entropy-constrained Non-negative Matrix Factorization (WNMF) is used for the factorization which identifies different components of an input signal. Applying this technique to non-negative features describing an input signal, such as magnitude spectra, the features are decomposed into a sparse combination of non-negative feature bases. The decomposition can be used to separate the input signal into several output signals corresponding to different components. The obtained dictionary of feature bases can also be used to separate the corresponding components from another signal, by decomposing this other signal according to the elements of the dictionary.

[0017] Aspects of the invention provide a novel method for enforcing a sparse decomposition, resulting in a dictionary of spectral bases which is more characteristic of the different parts of the signal, as will be presented in the following.

[0018] In order to describe the invention in detail, the following terms, abbreviations and notations will be used:

audio rendering: a reproduction technique capable of creating spatial sound fields in an extended area by means of loudspeakers or loudspeaker arrays,

NMF: Non-negative matrix factorization,

WNMF: Wiener entropy-constrained Non-negative Matrix Factorization.

Vector 1-norm: The vector 1-norm is the matrix norm of an m times n matrix A defined as the sum of the absolute values of its elements,

$$\|A\|_1 = \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|$$

Hadamard product: The Hadamard product is a binary operation that takes two matrices of the same dimensions, and produces another matrix where each element ij is the product of elements ij of the original two matrices.

[0019] According to a first aspect, the invention relates to a method for determining a dictionary of base components from an audio signal, the audio signal being represented by an input matrix which columns comprise features of the audio signal at different instances in time, the method comprising: decomposing the input matrix into a product of a non-negative base matrix and a non-negative weight matrix, the decomposing being constrained by a Wiener entropy measure with respect to elements of the non-negative weight matrix, wherein components of the non-negative base matrix represent the dictionary of base components of the audio signal.

[0020] The decomposing being constrained by a Wiener entropy measure, also denoted as Wiener entropy constraint is a new constraint providing a novel method for enforcing sparsity. The Wiener entropy or spectral flatness measures how flat the vector is. It is used as a sparsity penalty for NMF. By using that measure meaningful spectral patterns are estimated, speech separation quality, measured by both signal-based and perceptual criteria is improved. Compared to standard NMF the complexity increase is limited. The Wiener entropy constrained NMF (WNMF) can be integrated into any system using NMF.

[0021] In a first possible implementation form of the method according to the first aspect, the dictionary of base components represents a specific audio source of a plurality of audio sources of the audio signal.

[0022] Thus, a specific audio source of a multi-source audio signal can be extracted from the noisy multi-source audio signal.

[0023] In a second possible implementation form of the method according to the first aspect as such or according to the first implementation form of the first aspect, the decomposing is performed by using a Non-negative Matrix Factor-

ization.

[0024] The Wiener entropy measure can thus be adapted to a standard NMF factorization with only little overhead thereby saving computational complexity.

[0025] In a third possible implementation form of the method according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, the decomposing constrained by the Wiener entropy measure is configured to enforce a sparse decomposition of the non-negative base matrix.

[0026] Computing with sparse matrices improves speed and reduces complexity.

[0027] In a fourth possible implementation form of the method according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, the decomposing constrained by the Wiener entropy measure comprises: forming the non-negative weight matrix such that a Wiener entropy of each column of the non-negative weight matrix is close to zero.

[0028] By that specific forming of the H matrix, reconstruction of the original signal is improved.

[0029] In a fifth possible implementation form of the method according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, the decomposing constrained by the Wiener entropy measure comprises: minimizing a weighted sum of Wiener entropy values of columns of the non-negative weight matrix by using a cost function.

[0030] By using a cost function iterative or recursive adaptations can be applied which are computational efficient. Reconstruction of the original signal is improved.

[0031] In a sixth possible implementation form of the method according to the fifth implementation form of the first aspect, the cost function is according to

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the operation $\|\cdot\|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication and the symbol $\dot{\cdot}$ denotes the element-wise division, A is a real non-negative parameter, the symbol ϵ denotes a (small) positive real number and the operator $[\cdot]_{\epsilon}$ is defined by

$$[x]_{\epsilon} = \max(x, \epsilon).$$

[0032] Such a cost function provides an efficient reconstruction of the original signal.

[0033] In a seventh possible implementation form of the method according to the fifth implementation form or according to the sixth implementation form of the first aspect, the method comprises: updating the cost function by one of a multiplicative update rule and a gradient descent algorithm.

[0034] Multiplicative update rules are easy to implement and gradient descent algorithms converge to the locally optimum solution.

[0035] In an eighth possible implementation form of the method according to the seventh implementation form of the first aspect, the multiplicative update rule is according to:

$$W = W \otimes \frac{V}{W \cdot H} \cdot H^T \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H} + \lambda \frac{G}{rA \otimes A}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, λ is a real non-negative parameter, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise

multiplication, $\mathbb{I}_{m,n}$ denotes a matrix of dimension $m \times n$ whose elements are all equal to one, A denotes a matrix of dimension $r \times n$, defined by:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

and G denotes a matrix of dimension $r \times n$, defined by:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r}.$$

where the symbol ϵ denotes a positive real number and the operator $[\cdot]_{\epsilon}$ is defined by

$$[x]_{\epsilon} = \max(x, \epsilon).$$

[0036] These multiplicative update rules are easy to implement and fast converging.

[0037] In a ninth possible implementation form of the method according to the fifth implementation form of the first aspect, the cost function is according to

$$D(V, W \cdot H) + \lambda \left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the operation $\|\cdot\|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication and the symbol $\cdot$ denotes the element-wise division, λ is a real non-negative parameter, the symbol ϵ denotes a (small) positive real number and the operator $[\cdot]_{\epsilon}$ is defined by

$$[x]_{\epsilon} = \max(x, \epsilon).$$

[0038] Such a cost function provides an efficient reconstruction of the original signal and a homogeneous estimation of the components, regardless of the amplitude of the original signal

[0039] In a tenth possible implementation form of the method according to the ninth implementation form or according to the sixth implementation form of the first aspect, the method comprises: updating the cost function by one of a multiplicative update rule and a gradient descent algorithm.

[0040] Multiplicative update rules are easy to implement and gradient descent algorithms converge to the locally optimum solution.

[0041] In an eleventh possible implementation form of the method according to the tenth implementation form of the first aspect, the multiplicative update rule is according to:

$$W = W \otimes \frac{V}{\overline{W \cdot H} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \overline{V}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}},$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, λ is a real non-negative parameter, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication, $\mathbb{I}_{m,n}$ denotes a matrix of dimension $m \times n$ whose elements are all equal to one, A denotes a matrix of dimension $r \times n$, defined by:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

and G denotes a matrix of dimension $r \times n$, defined by:

$$G_{i,j} = \left(\prod_{i=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r}.$$

where the symbol ϵ denotes a positive real number and the operator $[\cdot]_{\epsilon}$ is defined by

$$[x]_{\epsilon} = \max(x, \epsilon).$$

[0042] These multiplicative update rules are easy to implement and fast converging.

[0043] In a twelfth possible implementation form of the method according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, the method comprises: reconstructing a plurality of output signals from the audio signal, the reconstruction being based on the input matrix, the non-negative base matrix and the non-negative weight matrix.

[0044] The reconstructed signals are noise-reduced and they indicate the source components of the original audio signal.

[0045] In a thirteenth possible implementation form of the method according to the twelfth implementation form of the first aspect, magnitude spectrograms S_k of the plurality of output signals are determined according to:

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V,$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix, $W_{:,k}$ denotes the column-vector constituted by the k-th column of W and $H_{k,:}$ denotes the row-vector constituted by the k-th row of H and the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication.

[0046] The output signals can be superposed in order to obtain signals corresponding to the combination of several components, for source separation applications.

[0047] In a fourteenth possible implementation form of the method according to the twelfth implementation form of the first aspect, magnitude spectrograms S_k of the plurality of output signals are determined by a product of a column-vector $W_{:,k}$ constituted by the k-th column of the non-negative base matrix W and a row-vector $H_{k,:}$ constituted by the k-th row of the non-negative weight matrix H.

[0048] When the output signals are directly reconstructed, computational complexity is reduced.

[0049] In a fifteenth possible implementation form of the method according to the thirteenth implementation form or according to the fourteenth implementation form of the first aspect, the method comprises: constructing output spectrograms by summing several of the magnitude spectrograms S_k of the plurality of output signals.

[0050] In a sixteenth possible implementation form of the method according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, the method comprises: determining a dictionary of base components from a training speech signal according to the first aspect as such or according to any of the preceding implementation forms of the first aspect, and forming a non-negative base matrix of a noisy speech signal by extending the non-negative base matrix of the training speech signal; and updating the non-negative base matrix of the noisy speech signal by using a semi-supervised Non-negative Matrix Factorization.

[0051] When using a training speech signal, source separation is improved as a speech signal which is not corrupted by noise is used for determining the dictionary of base components.

[0052] In a seventeenth possible implementation form of the method according to the sixteenth implementation form of the first aspect, the method comprises: reconstructing the speech signal based on the updated non-negative base matrix of the noisy speech signal.

[0053] Accuracy of the reconstruction is improved when the reconstruction is based on the updated base matrix of the noisy speech signal.

[0054] According to a second aspect, the invention relates to a device for determining a dictionary of base components from an input signal represented by an input matrix, the device comprising: a buffer for storing the input matrix; and means for decomposing the input matrix into a product of a non-negative base matrix and a non-negative weight matrix, wherein the decomposing is constrained by a Wiener entropy measure and wherein components of the non-negative base matrix represent the dictionary of base components of the input signal.

[0055] By using the Wiener entropy measure, meaningful spectral patterns are estimated and thus, speech separation quality, measured by both signal-based and perceptual criteria is improved. The complexity increase is not significant when compared to standard NMF implementations. The Wiener entropy constrained NMF can be integrated into any device using NMF.

[0056] The methods and systems described herein may be implemented as software in a Digital Signal Processor (DSP), in a micro-controller or in any other side-processor or as hardware circuit within an application specific integrated circuit (ASIC). The means for decomposing the input matrix may be implemented as software in a Digital Signal Processor (DSP), in a micro-controller or in any other side-processor or as a hardware unit, e.g. within an application specific integrated circuit (ASIC).

[0057] The invention can be implemented in digital electronic circuitry, or in computer hardware, firmware, software, or in combinations thereof, e.g. in available hardware or software of conventional mobile devices or hands-free communication systems or in new hardware or software dedicated for processing the methods described herein after.

[0058] Aspects of the invention provide a method for decomposing a signal according to a Wiener entropy-constrained Non-negative Matrix Factorization (WNMF). This method brings a modification to Non-negative Matrix Factorization (NMF) to enforce a sparse decomposition of a non-negative matrix.

[0059] The Wiener entropy, also called "spectral flatness", of a set of non-negative values is the ratio between the geometric mean and the arithmetic mean of these values. The Wiener entropy is always between zero and one, and it is equal to one if and only if all the values in the set are equal. Intuitively, a large value of the Wiener entropy corresponds to a "flat" plot and a small value corresponds to a "peaky" plot. Hence, to enforce the sparsity property of the NMF decomposition, it has to be ensured that the Wiener entropy of each column of H is small.

[0060] In the WNMF method, the penalty term used to measure the sparsity of the decomposition is given by a weighted sum of the Wiener entropy values of the columns of the matrix H. The cost function to be minimized is then

$$D(V, W \cdot H) + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}},$$

where $H_{i,j}$ is the value in the i-th row and j-th column of the matrix H. λ is a real non-negative parameter and the parameters ω_j are non-negative weighting parameters, which can depend on the matrix H. The symbol ϵ denotes a (small) positive real number and the operator $[\cdot]_{\epsilon}$ is defined by

$$[x]_{\epsilon} = \max(x, \epsilon).$$

[0061] This maximum ensures that the penalty term is positive, and that sparsity is enforced even when one of the weights is equal to zero. A variety of functions can be used for measuring the reconstruction error. In an implementation form, the reconstruction error is defined as

$$D(V, W \cdot H) = \left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1,$$

where the operation $\| \cdot \|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication and $\frac{\cdot}{\cdot}$ is the element-wise division.

[0062] In an implementation form, the weighting parameters ω_j used in the above defined cost function are all set to one.

[0063] In an implementation form, the optimization of this cost function is performed by multiplicative update rules, which enforce non-negativity without needing explicit constraints. In an implementation form, A and G are two matrices of dimensions $r \times n$, defined by:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

and

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r}.$$

[0064] The updates of the decomposition are performed according to:

$$W = W \otimes \frac{V}{\mathbb{W} \cdot \mathbb{H} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \frac{V}{\mathbb{W} \cdot \mathbb{H}} + \lambda \frac{G}{rA \otimes A}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

where $\mathbb{I}_{m,n}$ is a matrix of dimensions $m \times n$ whose elements are all equal to one. The optimization process stops when convergence is observed or when a sufficient number of iteration has been performed.

[0065] In an alternative implementation form, gradient-descent algorithms are applied instead of these multiplicative updates.

[0066] In an alternative implementation form, the weighting parameters ω_j used in the above defined cost function are all set to the mean value of the corresponding columns of the matrix H .

$$\omega_j = \frac{1}{r} \sum_{i=1}^r H_{i,j}$$

[0067] Hence, the sparsity penalty applied to each instance in time is approximately proportional to the amplitude of the input signal at the corresponding instance in time.

[0068] This ensures that the relative orders of magnitude of the sparsity term and the reconstruction error term are homogeneous over time. Thus, the relative importance of both constraints does not depend on the amplitude of the input signal. In this case, the cost function simplifies to:

$$D(V, W \cdot H) + \lambda \left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}.$$

[0069] In an implementation the optimization of this cost function is performed by multiplicative update rules. The updates of the decomposition are performed according to:

$$W = W \otimes \frac{V}{\mathbb{W} \cdot \mathbb{H} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \frac{V}{\mathbb{W} \cdot \mathbb{H}}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}}.$$

[0070] Another advantage of this setting is that the complexity of the parameter updates is reduced compared to the previous implementation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0071] Further embodiments of the invention will be described with respect to the following figures, in which:

Fig. 1 shows a schematic diagram 100 of a conventional non-negative Matrix Factorization (NMF) technique;

Fig. 2 shows three schematic diagrams 201, 202, 203 representing V , W and H matrices of a conventional Non-negative Matrix Factorization decomposition;

Fig. 3 shows exemplary spectrograms of two musical notes 301, 302, a succession of the two musical notes 300 and reconstructions 303, 304 of the two musical notes reconstructed by using a conventional NMF factorization;

Fig. 4 shows a schematic diagram of a method 440 for determining a dictionary of base components from an audio signal by performing a WNMF decomposition according to an implementation form;

Fig. 5 shows a schematic diagram of a method 500 for decomposing an audio signal into a dictionary of base components and reconstructing a set of audio signals according to an implementation form;

Fig. 6 shows a schematic diagram of a method 600 for decomposing an audio signal into a dictionary of base components applied to a noisy speech signal according to an implementation form;

Fig. 7 shows a schematic diagram of a de-noising system 700 according to an implementation form; and

Fig. 8 shows a schematic diagram of a device 800 for determining a dictionary of base components 804 from an audio signal 802 according to an implementation form.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

[0072] Fig. 4 shows a schematic diagram of a method 440 for determining a dictionary of base components from an audio signal by performing a WNMF decomposition according to an implementation form.

[0073] The method 440 performs a WNMF decomposition 400 from a digital single-channel acoustic signal 401. The digital input signal 401 is input to a short-time transform module 410, which performs a windowing into short-time frames and a transform, so as to produce non-negative feature vectors 411, e.g. magnitude spectra. A buffer 420 stores these features in order to produce the matrix V 421. The WNMF module 430 then performs a decomposition of the matrix V 421, representing the magnitude spectra of the input signal. The outputs of this module are the matrices W 431 and H 432 which represent respectively the dictionary of feature bases and the temporal weights of these bases.

[0074] Fig. 5 shows a schematic diagram of a system 500 for decomposing an audio signal into a dictionary of base components and reconstructing a set of audio signals according to an implementation form. The system 500 is adapted for separating a single-channel acoustic signal into several components (here $r = 3$). The system 500 comprises a factorization element 400 performing the WNMF decomposition 400 as described above with respect to Fig. 4 and a reconstruction element 510. The factorization element 400 takes as input an acoustic signal 401 and estimates a dictionary of feature bases 431 and the corresponding temporal weights 432 describing the signal. The result of the decomposition is input to the reconstruction module 510, which produces several output signals 511, 512 and 513.

[0075] In an implementation form, the reconstruction module 510 exploits a so-called "soft mask" approach as described in the following. For $k = 1 \dots 3$, $W_{:,k}$ is the column-vector constituted by the k -th column of W and $H_{k,:}$ is the row-vector constituted by the k -th row of H . A magnitude spectrogram S_k is calculated as:

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V$$

[0076] The three obtained matrices constitute the magnitude spectrograms of the three output signals. The time-domain signal are then obtained by a standard approach, involving an inverse Fourier transform exploiting the phase of the original complex spectrogram, followed by an overlap-add procedure.

[0077] In an implementation form, the output signals are then superposed in order to obtain signals corresponding to the combination of several components, for a source separation application.

[0078] In another implementation form, the magnitude spectrogram of the output signals are directly reconstructed as $S_k = W_{:,k} \cdot H_{k,:}$.

[0079] The components of the system 500 described above may also be implemented as steps of a method.

[0080] Fig. 6 shows a schematic diagram of a system 600 for decomposing an audio signal into a dictionary of base components applied to a noisy speech signal according to an implementation form. The decomposition is applied to the reduction of noise in a noisy speech signal. This system 600 involves a prior training phase 610 which comprises a factorization element 400 performing the WNMF decomposition 400 as described above with respect to Fig. 4. In the training phase, a training speech signal 601 is input to the factorization element 400, which computes a dictionary of feature bases 611 and a matrix of temporal weights 612 corresponding to the WNMF decomposition of the training signal.

[0081] The system 600 further comprises a short-time transform 630, a buffer 640, a semi-supervised NMF module 650 and a reconstruction module 660. A single-channel noisy speech signal 621 undergoes a short-time transform 630 which calculates non-negative features 631, similarly to the element 410 described above with respect to Fig. 4. The buffer 640 stores these features to produce a matrix V 641. This matrix undergoes a decomposition using semi-supervised NMF 650, where the feature bases corresponding to speech are set to the values of the dictionary 611 given by the training phase. The other bases are updated by the semi-supervised NMF. The outputs of this decomposition 650 are the dictionary W 651 and the corresponding weights H 652. These matrices are used by a reconstruction element 660, which produces the de-noised speech signal.

[0082] The reconstruction is performed by the "soft mask" method. In an implementation form, H' is the matrix extracted from the matrix H 652 comprising the weights corresponding to the speech bases W_s . The magnitude spectrogram S of the de-noised output signal is calculated as:

$$S = \frac{W_s \cdot H^s}{W \cdot H} \otimes V.$$

[0083] The time-domain signal is then obtained by the same approach as described above with respect to Fig. 5 for the reconstruction element 510.

[0084] In an implementation form, the semi-supervised NMF 650 is replaced by a WNMF decomposition 400 as described above with respect to Fig. 4.

[0085] In yet another implementation form, a noise training phase similar to 610 is performed to estimate a noise feature dictionary from a training recording of noise. In this case, the dictionary W 651 is defined as the concatenation of the speech dictionary 611 and the noise dictionary, and the semi-supervised NMF 650 is replaced by a supervised NMF.

[0086] The components of the system 600 described above may also be implemented as steps of a method.

[0087] Fig. 7 shows a schematic diagram of a de-noising system 700 according to an implementation form.

[0088] In a training phase, spectral components W_{speaker} 713 and W_{noise} 715 are estimated from clean speech W_{speaker} 701 and noise V_{noise} 703 separately using WNMF 707, 709. These spectral components W_{speaker} 713 and W_{noise} 715 are fed to a de-noising system 711, which exploits them to separate speech from noise. The noise components are estimated on the noisy speech V_{mix} 705 without noise training by the de-noising system 711 which provides the de-noised speech 717.

[0089] In an implementation form, the de-noising system 711 is a supervised system. In an implementation form the de-noising system 711 is a semi-supervised system. In an implementation form the de-noising system 711 is an unsupervised NMF de-noising system where no a priori knowledge of the speech and noise models is available.

[0090] Fig. 8 shows a schematic diagram of a device 800 for determining a dictionary of base components 804 from an input signal 802 according to an implementation form. The input signal 802 is represented by an input matrix V . The device 800 comprises a buffer 803 for storing the input matrix V . The device 800 further comprises decomposing means 801 for decomposing the input matrix V into a product of a non-negative base matrix W and a non-negative weight matrix H , wherein the decomposing is constrained by a Wiener entropy measure and wherein components of the non-negative base matrix W represent the dictionary of base components 804 of the input signal 802.

[0091] In an implementation form, the dictionary of base components represents a specific audio source of a plurality of audio sources of the audio signal. In an implementation form, the decomposing means is configured for decomposing the input matrix V by using a Non-negative Matrix Factorization. In an implementation form, the decomposing means is configured to enforce a sparse decomposition of the non-negative base matrix W . In an implementation form, the decomposing means comprises means for forming the non-negative weight matrix H such that a Wiener entropy of each column of the non-negative weight matrix H is close to zero. In an implementation form, the decomposing means comprises means for minimizing a sum of Wiener entropy values of columns of the non-negative weight matrix H by using a cost function. In an implementation form, the cost function is according to

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the operation $\|\cdot\|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication and the symbol $\frac{\cdot}{\cdot}$ denotes the element-wise division. In an implementation form, the device 800 comprises means for updating the cost function by one of a multiplicative update rule and a gradient descent algorithm. In an implementation form, the multiplicative update rule is according to:

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H} + \lambda \frac{G}{rA \otimes A}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication, $\mathbb{I}_{m,n}$ denotes a matrix of dimension $m \times n$ whose elements are all equal to one, A denotes a matrix of dimension $r \times n$, defined by:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

and G denotes a matrix of dimension $r \times n$, defined by:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r}.$$

[0092] In an implementation form, the device 800 comprises means for reconstructing a plurality of output signals from the audio signal, the reconstruction being based on the input matrix V , the non-negative base matrix W and the non-negative weight matrix H . In an implementation form, magnitude spectrograms S_k of the plurality of output signals are determined according to:

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V,$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix, $W_{:,k}$ denotes the column-vector constituted by the k -th column of W and $H_{k,:}$ denotes the row-vector constituted by the k -th row of H and the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication.

[0093] In an implementation form, magnitude spectrograms S_k of the plurality of output signals are determined by a product of a column-vector $W_{:,k}$ constituted by the k -th column of the non-negative base matrix W and a row-vector $H_{k,:}$ constituted by the k -th row of the non-negative weight matrix H .

[0094] In an implementation form, the device 800 comprises means for determining a dictionary of base components from a training speech signal according to the method 400 as described above with respect to Fig. 4; and means for forming a non-negative base matrix W of a noisy speech signal by using a semi-supervised Non-negative Matrix Factorization; and means for updating the non-negative base matrix W of the noisy speech signal with the non-negative base matrix W_S of the training speech signal. In an implementation form, the device 800 comprises means for reconstructing the speech signal based on the updated non-negative base matrix W of the noisy speech signal.

[0095] In another implementation form, the decomposing means comprises means for minimizing a weighted sum of

Wiener entropy values of columns of the non-negative weight matrix H by using a cost function, the weighting parameters of the sum being the mean values of the columns of the matrix H. In an implementation form, the cost function is according to

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}.$$

[0096] In an implementation form, the device 800 comprises means for updating the cost function by multiplicative update rule according to:

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot V}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}}.$$

[0097] From the foregoing, it will be apparent to those skilled in the art that a variety of methods, systems, computer programs on recording media, and the like, are provided.

[0098] The present disclosure also supports a computer program product including computer executable code or computer executable instructions that, when executed, causes at least one computer to execute the performing and computing steps described herein.

[0099] The present disclosure also supports a system configured to execute the performing and computing steps described herein.

[0100] Many alternatives, modifications, and variations will be apparent to those skilled in the art in light of the above teachings. Of course, those skilled in the art readily recognize that there are numerous applications of the invention beyond those described herein. While the present invention has been described with reference to one or more particular embodiments, those skilled in the art recognize that many changes may be made thereto. It is therefore to be understood that within the scope of the appended claims, the invention may be practiced otherwise than as specifically described herein.

Claims

1. A method (440) for determining a dictionary of base components (431) from an audio signal (401), the audio signal (401) being represented by an input matrix (V) which columns comprise features of the audio signal (401) at different instances in time, the method (440) comprising:

decomposing (430) the input matrix (V) into a product of a non-negative base matrix (W) and a non-negative weight matrix (H), the decomposing (430) being constrained by a Wiener entropy measure with respect to elements of the non-negative weight matrix (H), wherein components of the non-negative base matrix (W) represent the dictionary of base components (431) of the audio signal (401);

wherein the decomposing (430) constrained by the Wiener entropy measure comprises:

minimizing a weighted sum of Wiener entropy values of columns of the non-negative weight matrix (H) by using a cost function, and
wherein the cost function is according to

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the operation $\|\cdot\|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the

Hadamard product, i.e. element-wise multiplication, the symbol $\oslash$ denotes the element-wise division, λ is a real non-negative parameter, ω_j denote non-negative weighting parameters which can depend on the matrix H, the symbol ε denotes a positive real number and the operator $[\cdot]_\varepsilon$ is defined as

$$[x]_\varepsilon = \max(x, \varepsilon).$$

2. The method (440) of claim 1, wherein the dictionary of base components (431) represents a specific audio source of a plurality of audio sources of the audio signal (401).

3. The method (440) of claim 1 or claim 2, wherein the decomposing (430) uses a Non-negative Matrix Factorization.

4. The method (440) of one of the preceding claims, wherein the decomposing (430) constrained by the Wiener entropy measure is configured to enforce a sparse decomposition of the non-negative base matrix (W).

5. The method (440) of one of the preceding claims, wherein the decomposing (430) constrained by the Wiener entropy measure comprises:

forming the non-negative weight matrix (H) such that a Wiener entropy of each column of the non-negative weight matrix (H) is close to zero.

6. The method (440) of claim 1, comprising:

updating the cost function by one of a multiplicative update rule and a gradient descent algorithm.

7. The method (440) of claim 6, wherein the multiplicative update rule is according to:

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H} + \lambda \frac{G}{rA \otimes A}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, λ is a real non-negative parameter, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication, $\mathbb{I}_{m,n}$ denotes a matrix of dimension $m \times n$ whose elements are all equal to one, A denotes a matrix of dimension $r \times n$, defined by:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

and G denotes a matrix of dimension $r \times n$, defined by:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_\varepsilon \right)^{1/r}.$$

where the symbol ε denotes a positive real number and the operator $[\cdot]_\varepsilon$ is defined as

$$[x]_\varepsilon = \max(x, \varepsilon).$$

8. The method (440) of claim 6, wherein the multiplicative update rule is according to:

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{and} \quad H = H \otimes \frac{W^T \cdot V}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}},$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, λ is a real non-negative parameter, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication, $\mathbb{I}_{m,n}$ denotes a matrix of dimension $m \times n$ whose elements are all equal to one, G denotes a matrix of dimension $r \times n$, defined by:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r},$$

where the symbol ϵ denotes a positive real number and the operator $[\cdot]_{\epsilon}$ is defined as

$$[x]_{\epsilon} = \max(x, \epsilon).$$

9. The method (500) of one of the preceding claims, comprising:

reconstructing (510) a plurality of output signals (511, 512, 513) from the audio signal (401), the reconstruction (510) being based on the input matrix (V), the non-negative base matrix (W) and the non-negative weight matrix (H).

10. The method (500) of claim 9, wherein magnitude spectrograms S_k of the plurality of output signals (511, 512, 513) are determined according to:

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V,$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix, $W_{:,k}$ denotes the column-vector constituted by the k -th column of W and $H_{k,:}$ denotes the row-vector constituted by the k -th row of H and the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication; or wherein magnitude spectrograms S_k of the plurality of output signals (511, 512, 513) are determined by a product of a column-vector $W_{:,k}$ constituted by the k -th column of the non-negative base matrix W and a row-vector $H_{k,:}$ constituted by the k -th row of the non-negative weight matrix H .

11. The method (500) of claim 10, comprising:

constructing output spectrograms by summing several of the magnitude spectrograms S_k of the plurality of output signals (511, 512, 513).

12. The method (600) of one of the preceding claims, comprising:

determining (610) a dictionary of base components (611) from a training speech signal (601) according to one of the methods 1 to 13; and forming (651) a non-negative base matrix (W) of a noisy speech signal (621) by extending the non-negative base matrix (W_S) of the training speech signal (601); and

updating the non-negative base matrix (W) of the noisy speech signal (621) by using a semi-supervised Non-negative Matrix Factorization.

13. Device (800) for determining a dictionary of base components (804) from an input signal (802) represented by an

input matrix (V), the device (800) comprising:

a buffer (803) for storing the input matrix (V); and
means (801) for decomposing the input matrix (V) into a product of a non-negative base matrix (W) and a non-negative weight matrix (H), wherein the decomposing is constrained by a Wiener entropy measure and wherein components of the non-negative base matrix (W) represent the dictionary of base components (804) of the input signal (802);

wherein the decomposing constrained by the Wiener entropy measure comprises:

minimizing a weighted sum of Wiener entropy values of columns of the non-negative weight matrix (H) by using a cost function, and
wherein the cost function is according to

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

where V denotes the input matrix, W denotes the non-negative base matrix, H denotes the non-negative weight matrix with elements $H_{i,j}$, the operation $\|\cdot\|_1$ denotes the vector 1-norm, the symbol $\otimes$ denotes the Hadamard product, i.e. element-wise multiplication, the symbol $\cdot$ denotes the element-wise division, λ is a real non-negative parameter, ω_j denote non-negative weighting parameters which can depend on the matrix H, the symbol ϵ denotes a positive real number and the operator $[\cdot]_{\epsilon}$ is defined as

$$[x]_{\epsilon} = \max(x, \epsilon).$$

Patentansprüche

1. Verfahren (440) zum Bestimmen eines Wörterbuchs von Basiskomponenten (431) aus einem Audiosignal (401), wobei das Audiosignal (401) durch eine Eingabematrix (V) dargestellt ist, wobei Spalten Merkmale des Audiosignals (401) zu unterschiedlichen Zeitpunkten umfassen, wobei das Verfahren (440) Folgendes umfasst:

Zerlegen (430) der Eingabematrix (V) in ein Produkt einer nichtnegativen Basismatrix (W) und einer nichtnegativen Gewichtsmatrix (H), wobei das Zerlegen (430) durch ein Wiener-Entropiemaß in Bezug auf Elemente der nichtnegativen Gewichtsmatrix (H) eingeschränkt ist, wobei Komponenten der nichtnegativen Basismatrix (W) das Wörterbuch von Basiskomponenten (431) des Audiosignals (401) darstellen;
wobei das Zerlegen (430), das durch das Wiener-Entropiemaß eingeschränkt ist, Folgendes umfasst:

Minimieren einer gewichteten Summe von Wiener-Entropiewerten von Spalten der nichtnegativen Gewichtsmatrix (H) unter Verwendung einer Kostenfunktion, und
wobei die Kostenfunktion wie folgt lautet

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

wobei V die Eingabematrix bezeichnet, W die nichtnegative Basismatrix bezeichnet, H die nichtnegative Gewichtsmatrix mit Elementen $H_{i,j}$ bezeichnet, die Operation $\|\cdot\|_1$ die Vektor-1-Norm bezeichnet, das Symbol

$\otimes$ das Hadamard-Produkt, d. h. eine elementweise Multiplikation bezeichnet, das Symbol $\frac{\cdot}{\cdot}$ die elementweise Division bezeichnet, λ ein realer nichtnegativer Parameter ist, ω_j nichtnegative Gewichts-

parameter bezeichnen, die von der Matrix H abhängen können, das Symbol ϵ eine positive reale Zahl bezeichnet und der Operator $[\cdot]_{\epsilon}$ wie folgt definiert ist

$$[x]_{\epsilon} = \max(x, \epsilon).$$

2. Verfahren (440) nach Anspruch 1, wobei das Wörterbuch von Basiskomponenten (431) eine spezifische Audioquelle mehrerer Audioquellen des Audiosignals (401) darstellt.

3. Verfahren (440) nach Anspruch 1 oder 2, wobei das Zerlegen (430) eine nichtnegative Matrixfaktorisierung anwendet.

4. Verfahren (440) nach einem der vorhergehenden Ansprüche, wobei das Zerlegen (430), das durch das Wiener-Entropiemaß eingeschränkt ist, konfiguriert ist, eine dünnbesetzte Zerlegung der nichtnegativen Basismatrix (W) durchzusetzen.

5. Verfahren (440) nach einem der vorhergehenden Ansprüche, wobei das Zerlegen (430), das durch das Wiener-Entropiemaß eingeschränkt ist, Folgendes umfasst:

Bilden der nichtnegativen Gewichtsmatrix (H), sodass eine Wiener-Entropie jeder Spalte der nichtnegativen Matrix (H) nahe null ist.

6. Verfahren (440) nach Anspruch 1, umfassend:

Aktualisieren der Kostenfunktion durch einen einer multiplikativen Aktualisierungsregel und eines Gradientenabstiegsalgorithmus.

7. Verfahren (440) nach Anspruch 6, wobei die multiplikative Aktualisierungsregel wie folgt lautet:

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{und} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H} + \lambda \frac{G}{rA \otimes A}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

wobei V die Eingabematrix bezeichnet, W die nichtnegative Basismatrix bezeichnet, H die nichtnegative Gewichtsmatrix mit Elementen $H_{i,j}$ bezeichnet, λ ein realer nichtnegativer Parameter ist, das Symbol $\otimes$ das Hadamard-Produkt, d. h. eine elementweise Multiplikation bezeichnet, $\mathbb{I}_{m,n}$ eine Matrix der Dimension $m \times n$ bezeichnet, deren Elemente alle gleich eins sind, G eine Matrix der Dimension $r \times n$ bezeichnet, die wie folgt definiert ist:

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

und G eine Matrix der Dimension $r \times n$ bezeichnet, die wie folgt definiert ist:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r}.$$

wobei das Symbol ϵ eine positive reale Zahl bezeichnet und der Operator $[\cdot]_{\epsilon}$ wie folgt definiert ist

$$[x]_{\epsilon} = \max(x, \epsilon).$$

8. Verfahren (440) nach Anspruch 6, wobei die multiplikative Aktualisierungsregel wie folgt lautet:

$$W = W \otimes \frac{V \cdot H^T}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{und} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}},$$

wobei V die Eingabematrix bezeichnet, W die nichtnegative Basismatrix bezeichnet, H die nichtnegative Gewichtsmatrix mit Elementen $H_{i,j}$ bezeichnet, λ ein realer nichtnegativer Parameter ist, das Symbol $\otimes$ das Hadamard-Produkt, d. h. eine elementweise Multiplikation bezeichnet, $\mathbb{I}_{m,n}$ eine Matrix der Dimension $m \times n$ bezeichnet, deren Elemente alle gleich eins sind, G eine Matrix der Dimension $r \times n$ bezeichnet, die wie folgt definiert ist:

$$G_{i,j} = \left(\prod_{k=1}^r [H_{k,j}]_{\epsilon} \right)^{1/r},$$

wobei das Symbol ϵ eine positive reale Zahl bezeichnet und der Operator $[\cdot]_{\epsilon}$ wie folgt definiert ist

$$[x]_{\epsilon} = \max(x, \epsilon).$$

9. Verfahren (500) nach einem der vorhergehenden Ansprüche, umfassend:

Rekonstruieren (510) mehrerer Ausgabesignale (511, 512, 513) aus dem Audiosignal (401), wobei die Rekonstruktion (510) auf der Eingabematrix (V), der nichtnegativen Basismatrix (W) und der nichtnegativen Gewichtsmatrix (H) basiert.

10. Verfahren (500) nach Anspruch 9, wobei Betragsspektrogramme S_k der mehreren Ausgabesignale (511, 512, 513) wie folgt bestimmt werden:

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V,$$

wobei V die Eingabematrix bezeichnet, W die nichtnegative Basismatrix bezeichnet, H die nichtnegative Gewichtsmatrix bezeichnet, $W_{:,k}$ den Spaltenvektor bezeichnet, der durch die kte Spalte von W gebildet wird und $H_{k,:}$ den Reihenvektor bezeichnet, der durch die kte Reihe von H gebildet wird, und das Symbol $\otimes$ das Hadamard-Produkt, d. h. eine elementweise Multiplikation bezeichnet; oder wobei Betragsspektrogramme S_k der mehreren Ausgabesignale (511, 512, 513) durch ein Produkt eines Spaltenvektors $W_{:,k}$, der durch die kte Spalte der nichtnegativen Basismatrix W gebildet wird, und einen Reihenvektor $H_{k,:}$, der durch die kte Reihe der nichtnegativen Gewichtsmatrix H gebildet wird, bestimmt werden.

11. Verfahren (500) nach Anspruch 10, umfassend:

Konstruieren von Ausgabespektrogrammen durch Summieren mehrerer Betragsspektrogramme S_k der mehreren Ausgabesignale (511, 512, 513).

12. Verfahren (600) nach einem der vorhergehenden Ansprüche, umfassend:

Bestimmen (610) eines Wörterbuchs von Basiskomponenten (611) aus einem Trainingssprachsignal (601) gemäß einem der Verfahren 1 bis 13; und
Bilden (651) einer nichtnegativen Basismatrix (W) eines verrauschten Sprachsignals (621) durch Erweitern der nichtnegativen Basismatrix (W_s) des Trainingssprachsignals (601); und
Aktualisieren der nichtnegativen Basismatrix (W) des verrauschten Sprachsignals (621) durch Anwenden einer

halbüberwachten nichtnegativen Matrixfaktorisierung.

13. Vorrichtung (800) zum Bestimmen eines Wörterbuchs von Basiskomponenten (804) aus einem Eingangssignal (802), das von einer Eingabematrix (V) dargestellt wird, wobei die Vorrichtung (800) Folgendes umfasst:

einen Puffer (803) zum Speichern der Eingabematrix (V); und
Mittel (801) zum Zerlegen der Eingabematrix (V) in ein Produkt einer nichtnegativen Basismatrix (W) und einer nichtnegativen Gewichtsmatrix (H), wobei das Zerlegen durch ein Wiener-Entropiemaß eingeschränkt ist und wobei Komponenten der nichtnegativen Basismatrix (W) das Wörterbuch von Basiskomponenten (804) des Eingangssignals (802) darstellen;

wobei das Zerlegen, das durch das Wiener-Entropiemaß eingeschränkt ist, Folgendes umfasst:

Minimieren einer gewichteten Summe von Wiener-Entropiewerten von Spalten der nichtnegativen Gewichtsmatrix (H) unter Verwendung einer Kostenfunktion, und
wobei die Kostenfunktion wie folgt lautet

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

wobei V die Eingabematrix bezeichnet, W die nichtnegative Basismatrix bezeichnet, H die nichtnegative Gewichtsmatrix mit Elementen $H_{i,j}$ bezeichnet, die Operation $\|\cdot\|_1$ die Vektor-1-Norm bezeichnet, das Symbol $\otimes$ das Hadamard-Produkt, d. h. eine elementweise Multiplikation bezeichnet, das Symbol $\frac{\cdot}{\cdot}$ die elementweise Division bezeichnet; λ ein realer nichtnegativer Parameter ist, ω_j nichtnegative Gewichtungparameter bezeichnen, die von der Matrix H abhängen können, das Symbol ϵ eine positive reale Zahl bezeichnet und der Operator $[\cdot]_{\epsilon}$ wie folgt definiert ist

$$[x]_{\epsilon} = \max(x, \epsilon).$$

Revendications

1. Procédé (440) pour déterminer un dictionnaire de composants de base (431) provenant d'un signal audio (401), le signal audio (401) étant représenté par une matrice d'entrée (V) dont les colonnes comprennent des caractéristiques du signal audio (401) à différents points dans le temps, le procédé (440) comprenant :

la décomposition (430) de la matrice d'entrée (V) en un produit d'une matrice de base non négative (W) et une matrice de poids non négative (H), la décomposition (430) étant contrainte par une mesure de l'entropie selon Wiener par rapport à des éléments de la matrice de poids non négative (H), les composants de la matrice de base non négative (W) représentant le dictionnaire de composants de base (431) du signal audio (401) ;
dans lequel la décomposition (430) contrainte par la mesure de l'entropie selon Wiener comprend :

la minimisation d'une somme pondérée de valeurs de l'entropie selon Wiener de colonnes de la matrice de poids non négative (H) par une fonction économique, et
dans lequel la fonction économique est conforme à :

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r [H_{i,j}]_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

où V représente la matrice d'entrée, W représente la matrice de base non négative, H représente la matrice

de poids non négative avec des éléments $H_{i,j}$, l'opération $\|\cdot\|_1$ représente la norme vectorielle 1, le symbole $\otimes$ représente le produit de Hadamard, c'est-à-dire la multiplication par élément, le symbole $\div$ représente la division par élément, λ est un paramètre réel non négatif, ω_i représente des paramètres de pondération non négatifs qui peuvent dépendre de la matrice H , le symbole ε représente un nombre réel positif, et l'opérateur $\lfloor \cdot \rfloor_\varepsilon$ est défini par

$$\lfloor x \rfloor_\varepsilon = \max(x, \varepsilon).$$

2. Procédé (440) selon la revendication 1, dans lequel le dictionnaire de composants de base (431) représente une source audio spécifique d'une pluralité de sources audio du signal audio (401).

3. Procédé (440) selon la revendication 1 ou 2, dans lequel la décomposition (430) utilise une factorisation de matrice non négative.

4. Procédé (440) selon l'une des revendications précédentes, dans lequel la décomposition (430) contrainte par la mesure de l'entropie selon Wiener est configurée pour imposer une décomposition sporadique de la matrice de base non négative (W).

5. Procédé (440) selon l'une des revendications précédentes, dans lequel la décomposition (430) contrainte par la mesure de l'entropie selon Wiener comprend :

la formation de la matrice de poids non négative (H) de façon que l'entropie selon Wiener de chaque colonne de la matrice de poids non négative (H) soit proche de zéro.

6. Procédé (440) selon la revendication 1, comprenant :

la mise à jour de la fonction économique par l'un parmi une règle de mise à jour multiplicative et un algorithme de descente à gradient.

7. Procédé (440) selon la revendication 6, dans lequel la règle de mise à jour multiplicative est conforme à :

$$W = W \otimes \frac{V}{\frac{W \cdot H}{\mathbb{I}_{m,n}} \cdot H^\top} \quad \text{et} \quad H = H \otimes \frac{W^\top \cdot \frac{V}{\frac{W \cdot H}{\mathbb{I}_{m,n}}} + \lambda \frac{G}{rA \otimes A}}{W^\top \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rA \otimes H}},$$

où V représente la matrice d'entrée, W représente la matrice de base non négative, H représente la matrice de poids non négative avec des éléments $H_{i,j}$, λ est un paramètre réel non négatif, le symbole $\otimes$ représente le produit de Hadamard, c'est-à-dire la multiplication par élément, $\mathbb{I}_{m,n}$ représente une matrice de dimension $m \times n$ dont les éléments sont tous égaux à un, A représente une matrice de dimension $r \times n$, définie par :

$$A_{i,j} = \frac{1}{r} \sum_{k=1}^r H_{k,j}$$

et G représente une matrice de dimension $r \times n$, définie par :

$$G_{i,j} = \left(\prod_{k=1}^r \lfloor H_{k,j} \rfloor_\varepsilon \right)^{1/r}.$$

où le symbole ε représente un nombre réel positif, et l'opérateur $\lfloor \cdot \rfloor_\varepsilon$ est défini par

$$\lfloor x \rfloor_{\epsilon} = \max(x, \epsilon).$$

8. Procédé (440) selon la revendication 6, dans lequel la règle de mise à jour multiplicative est conforme à :

$$W = W \otimes \frac{V}{\mathbb{I}_{m,n} \cdot H^T} \quad \text{et} \quad H = H \otimes \frac{W^T \cdot \frac{V}{W \cdot H}}{W^T \cdot \mathbb{I}_{m,n} + \lambda \frac{G}{rH}},$$

où V représente la matrice d'entrée, W représente la matrice de base non négative, H représente la matrice de poids non négative avec des éléments $H_{i,j}$, λ est un paramètre réel non négatif, le symbole $\otimes$ représente le produit de Hadamard, c'est-à-dire la multiplication par élément, $\mathbb{I}_{m,n}$ représente une matrice de dimension $m \times n$ dont les éléments sont tous égaux à un, G représente une matrice de dimension $r \times n$, définie par :

$$G_{i,j} = \left(\prod_{k=1}^r \lfloor H_{k,j} \rfloor_{\epsilon} \right)^{1/r},$$

où le symbole ϵ représente un nombre réel positif, et l'opérateur $\lfloor \bullet \rfloor_{\epsilon}$ est défini par

$$\lfloor x \rfloor_{\epsilon} = \max(x, \epsilon).$$

9. Procédé (500) selon l'une des revendications précédentes, comprenant :

la reconstruction (510) d'une pluralité de signaux de sortie (511, 512, 513) provenant du signal audio (401), la reconstruction (510) étant basée sur la matrice d'entrée (V), la matrice de base non négative (W) et la matrice de poids non négative (H).

10. Procédé (500) selon la revendication 9, dans lequel des spectrogrammes d'amplitude S_k de la pluralité de signaux de sortie (511, 512, 513) sont déterminés conformément à :

$$S_k = \frac{W_{:,k} \cdot H_{k,:}}{W \cdot H} \otimes V,$$

où V représente la matrice d'entrée, W représente la matrice de base non négative, H représente la matrice de poids non négative, $W_{:,k}$ représente le vecteur de colonne constitué de la k-ième colonne de W et $H_{k,:}$ représente le vecteur de rangée constitué de la k-ième rangée de H et le symbole $\otimes$ représente le produit de Hadamard, c'est-à-dire la multiplication par élément ; ou

dans lequel les spectrogrammes d'amplitude S_k de la pluralité de signaux de sortie (511, 512, 513) sont déterminés par un produit d'un vecteur de colonne $W_{:,k}$ constitué de la k-ième colonne de la matrice de base non négative W et d'un vecteur de rangée $H_{k,:}$ constitué de la k-ième rangée de la matrice de poids non négative H.

11. Procédé (500) selon la revendication 10, comprenant :

l'élaboration de spectrogrammes de sortie par la somme de plusieurs des spectrogrammes d'amplitude S_k de la pluralité de signaux de sortie (511, 512, 513).

12. Procédé (600) selon l'une des revendications précédentes, comprenant :

la détermination (610) d'un dictionnaire de composants de base (611) à partir d'un signal vocal d'entraînement (601) selon l'un des procédés 1 à 13 ; et
la formation (651) d'une matrice de base non négative (W) d'un signal vocal bruyant (621) par extension de la

matrice de base non négative (W_s) du signal vocal d'entraînement (601) ; et
la mise à jour de la matrice de base non négative (W) du signal vocal bruyant (621) par utilisation d'une factorisation de matrice non négative semi-supervisée.

5 **13.** Dispositif (800) pour déterminer un dictionnaire de composants de base (804) provenant d'un signal d'entrée (802) représenté par une matrice d'entrée (V), le dispositif (800) comprenant :

un tampon (803) pour stocker la matrice d'entrée (V) ; et
des moyens (801) pour décomposer la matrice d'entrée (V) en un produit d'une matrice de base non négative
10 (W) et d'une matrice de poids non négative (H), la décomposition étant contrainte par une mesure de l'entropie selon Wiener, et les composants de la matrice de base non négative (W) représentant le dictionnaire de composants de base (804) du signal d'entrée (802) ;

dans lequel la décomposition contrainte par la mesure de l'entropie selon Wiener comprend :

la minimisation d'une somme pondérée de valeurs de l'entropie selon Wiener de colonnes de la matrice de poids non négative (H) par une fonction économique, et
dans lequel la fonction économique est conforme à :

20

$$\left\| V \otimes \ln \frac{V}{W \cdot H} - V + W \cdot H \right\|_1 + \lambda \sum_{j=1}^n \omega_j \frac{\left(\prod_{i=1}^r \lfloor H_{i,j} \rfloor_{\epsilon} \right)^{1/r}}{\frac{1}{r} \sum_{i=1}^r H_{i,j}}$$

25

30

où V représente la matrice d'entrée, W représente la matrice de base non négative, H représente la matrice de poids non négative avec des éléments $H_{i,j}$, l'opération $\|\cdot\|_1$ représente la norme vectorielle 1, le symbole $\otimes$ représente le produit de Hadamard, c'est-à-dire la multiplication par élément, le symbole $\div$ représente la division par élément, λ est un paramètre réel non négatif, ω_j représente des paramètres de pondération non négatifs qui peuvent dépendre de la matrice H , le symbole ϵ représente un nombre réel positif, et l'opérateur $\lfloor \cdot \rfloor_{\epsilon}$ est défini par

35

$$\lfloor x \rfloor_{\epsilon} = \max(x, \epsilon).$$

40

45

50

55

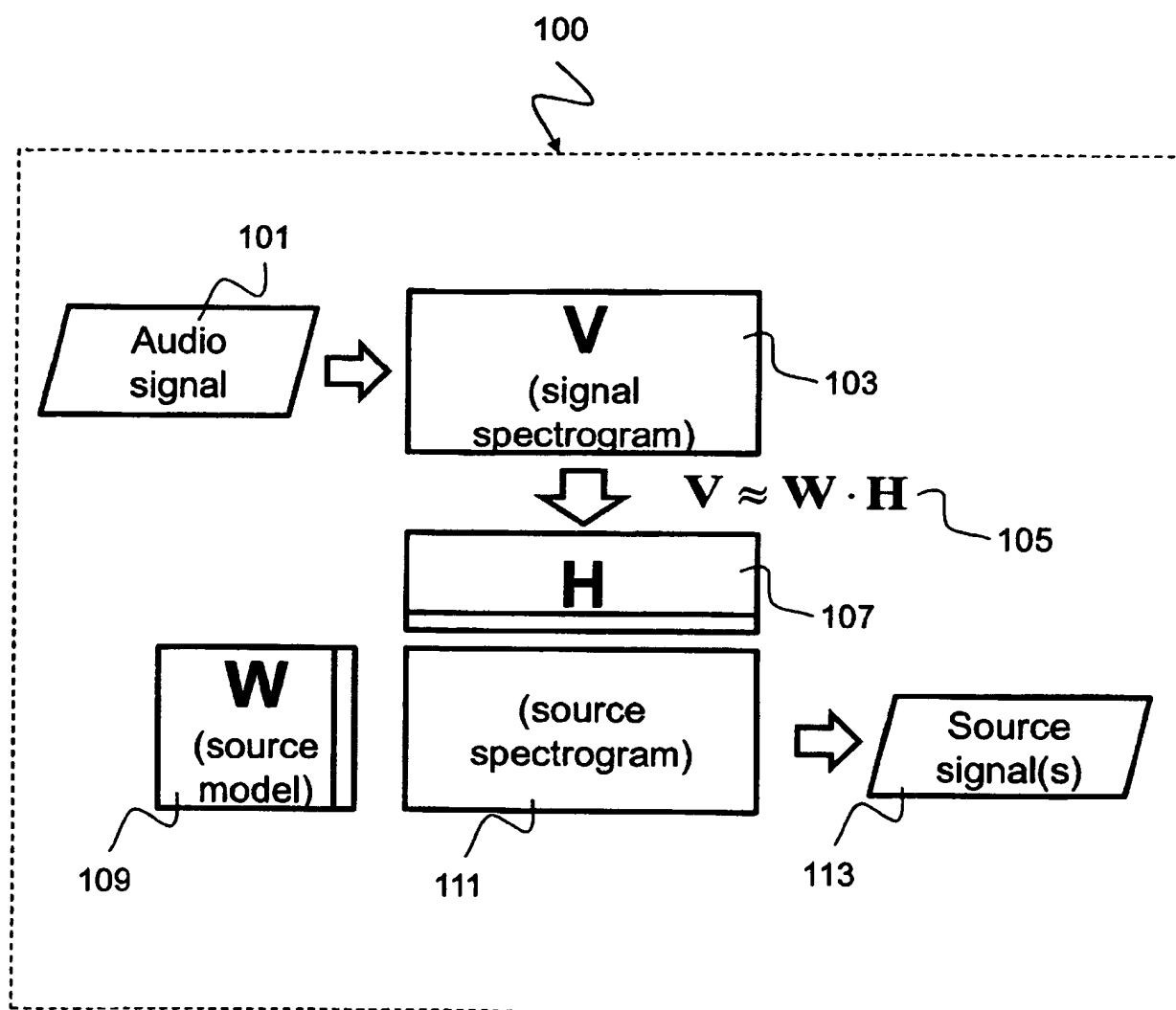


Fig. 1

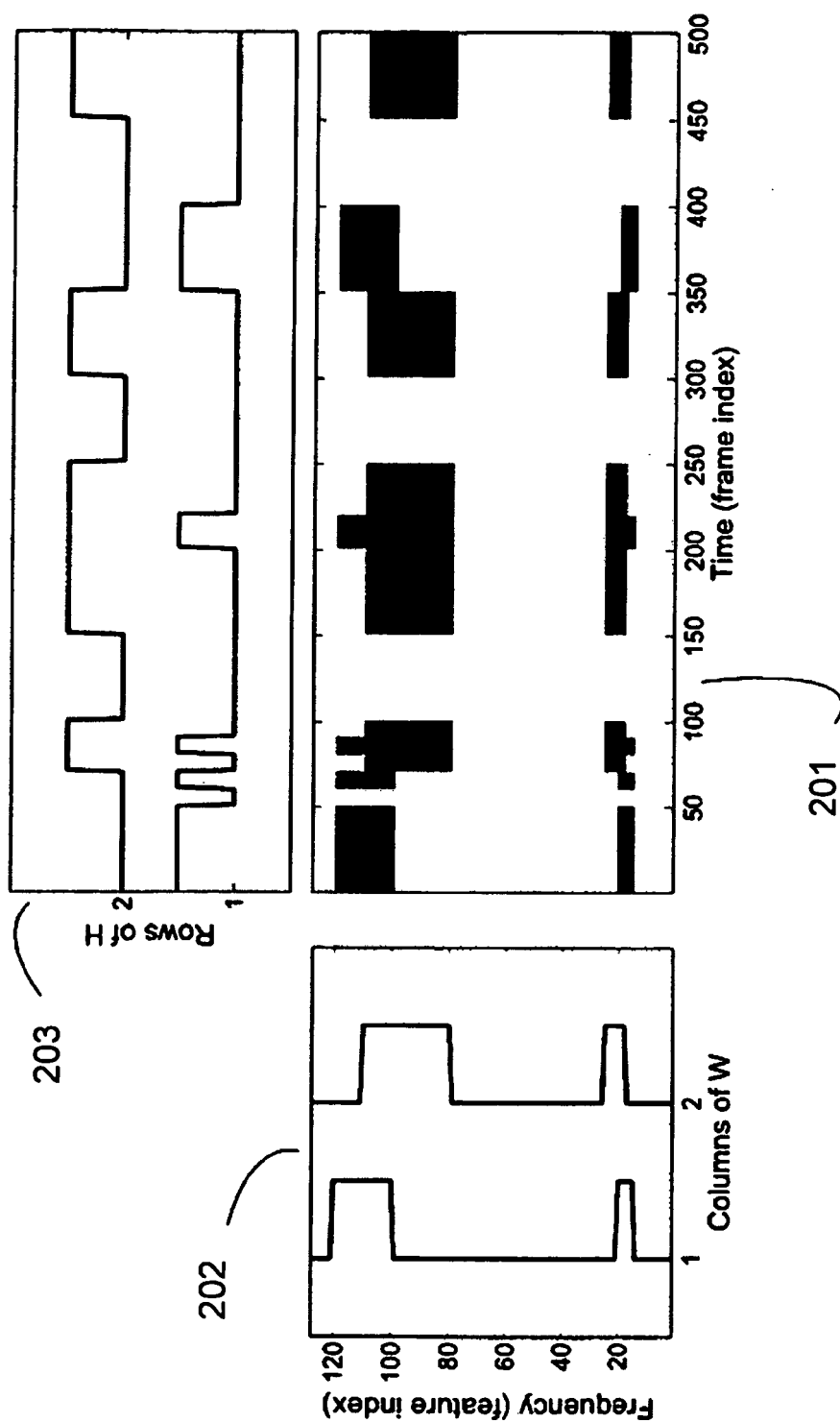


Fig. 2

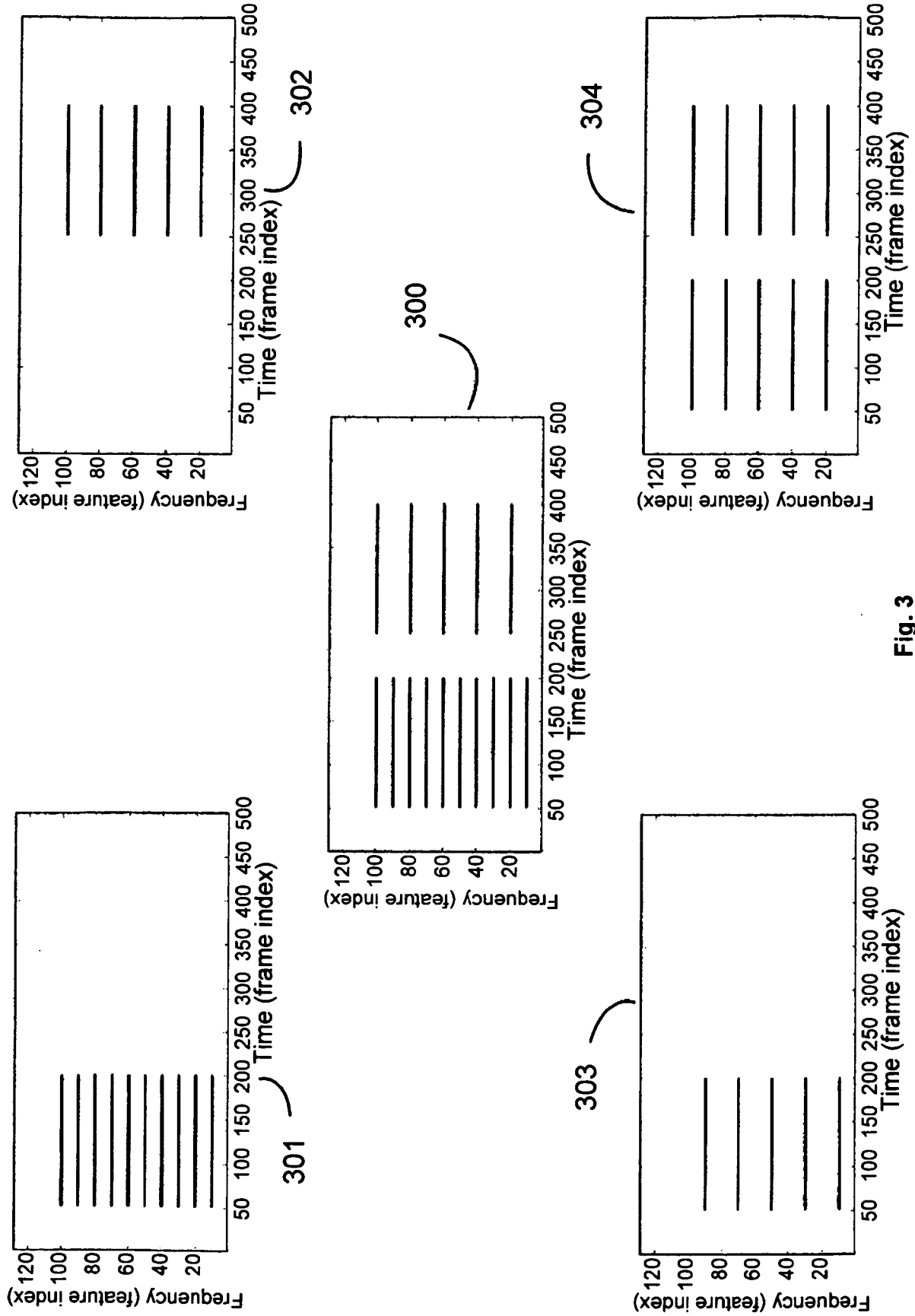


Fig. 3

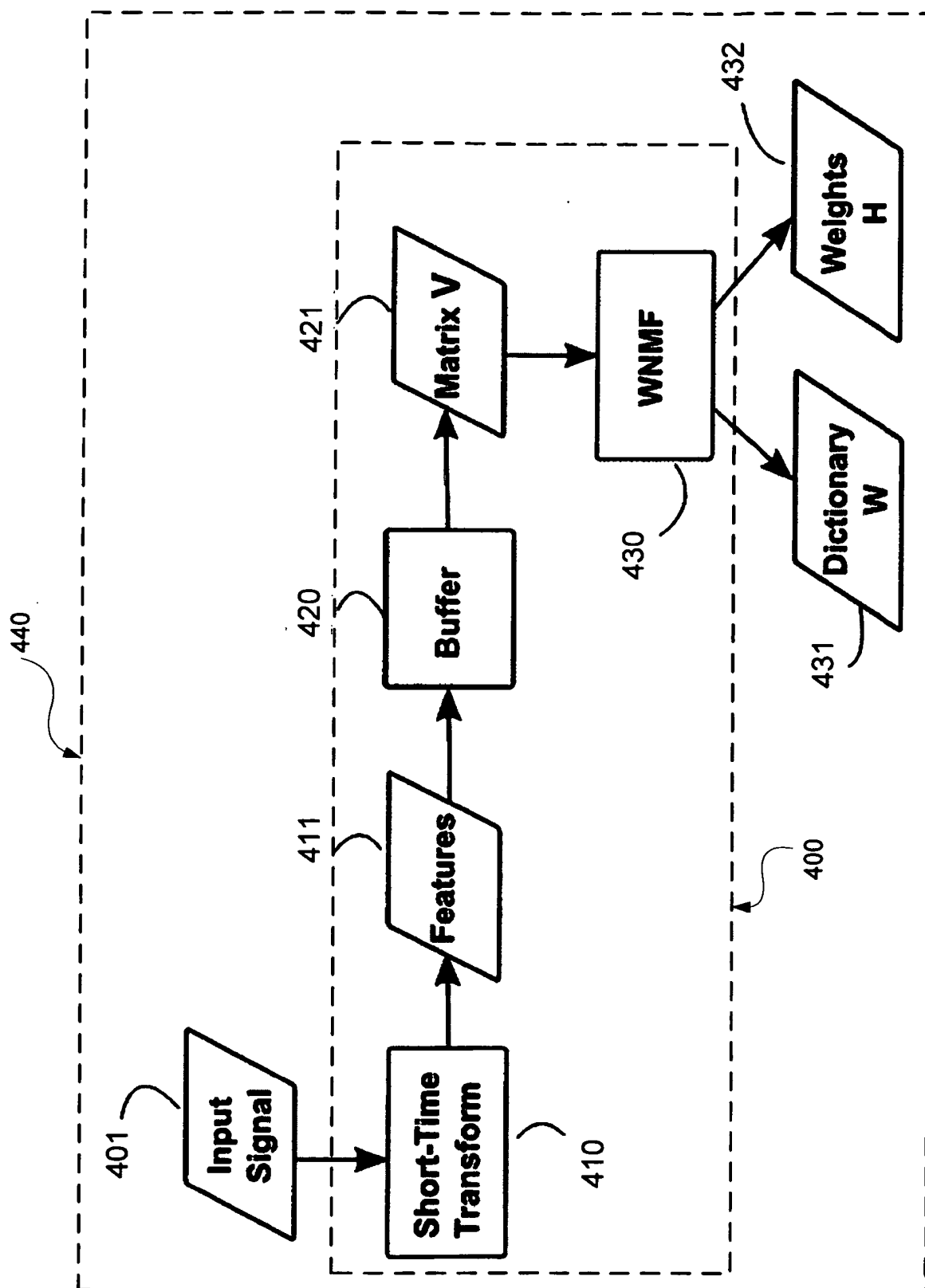


Fig. 4

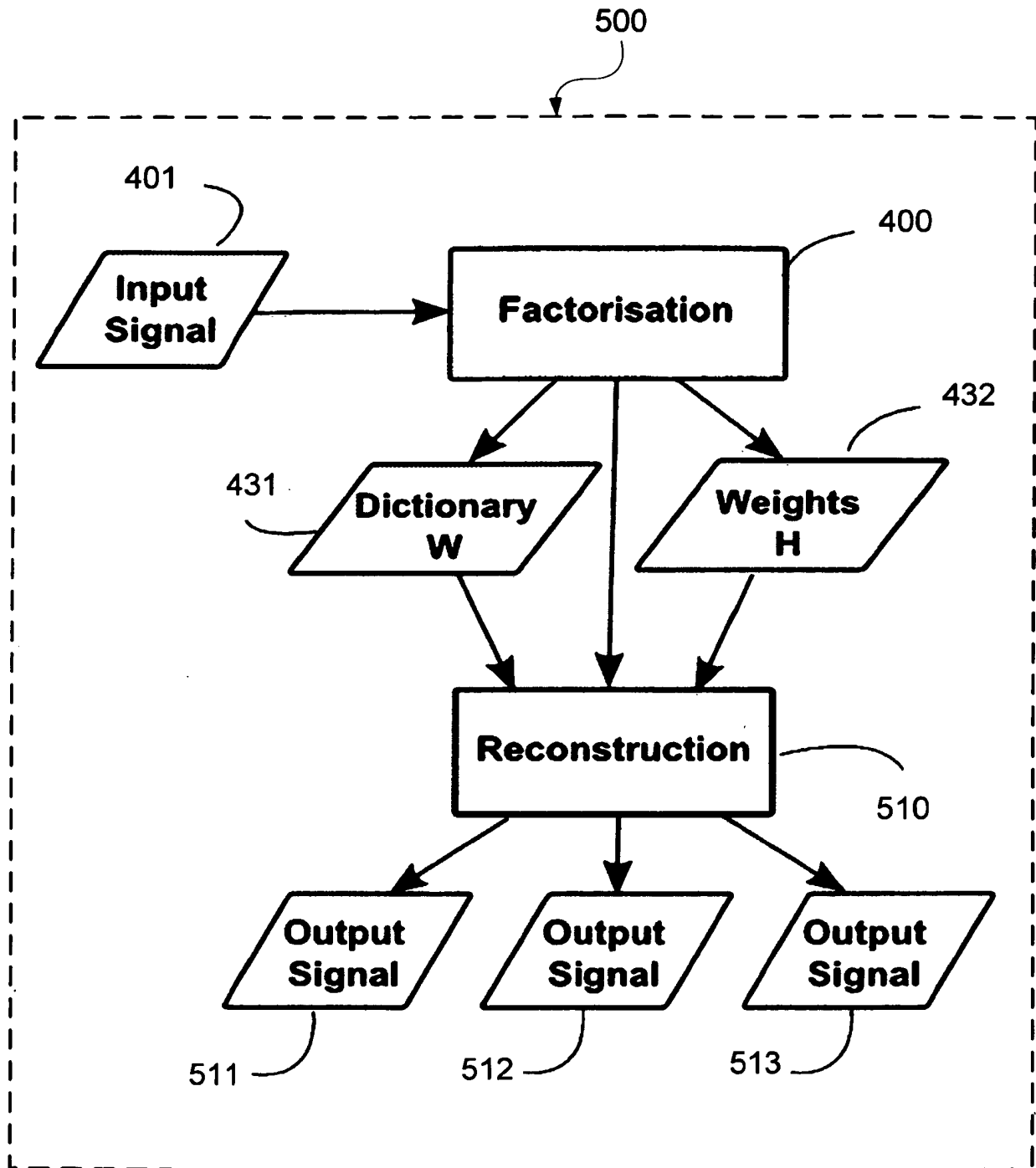


Fig. 5

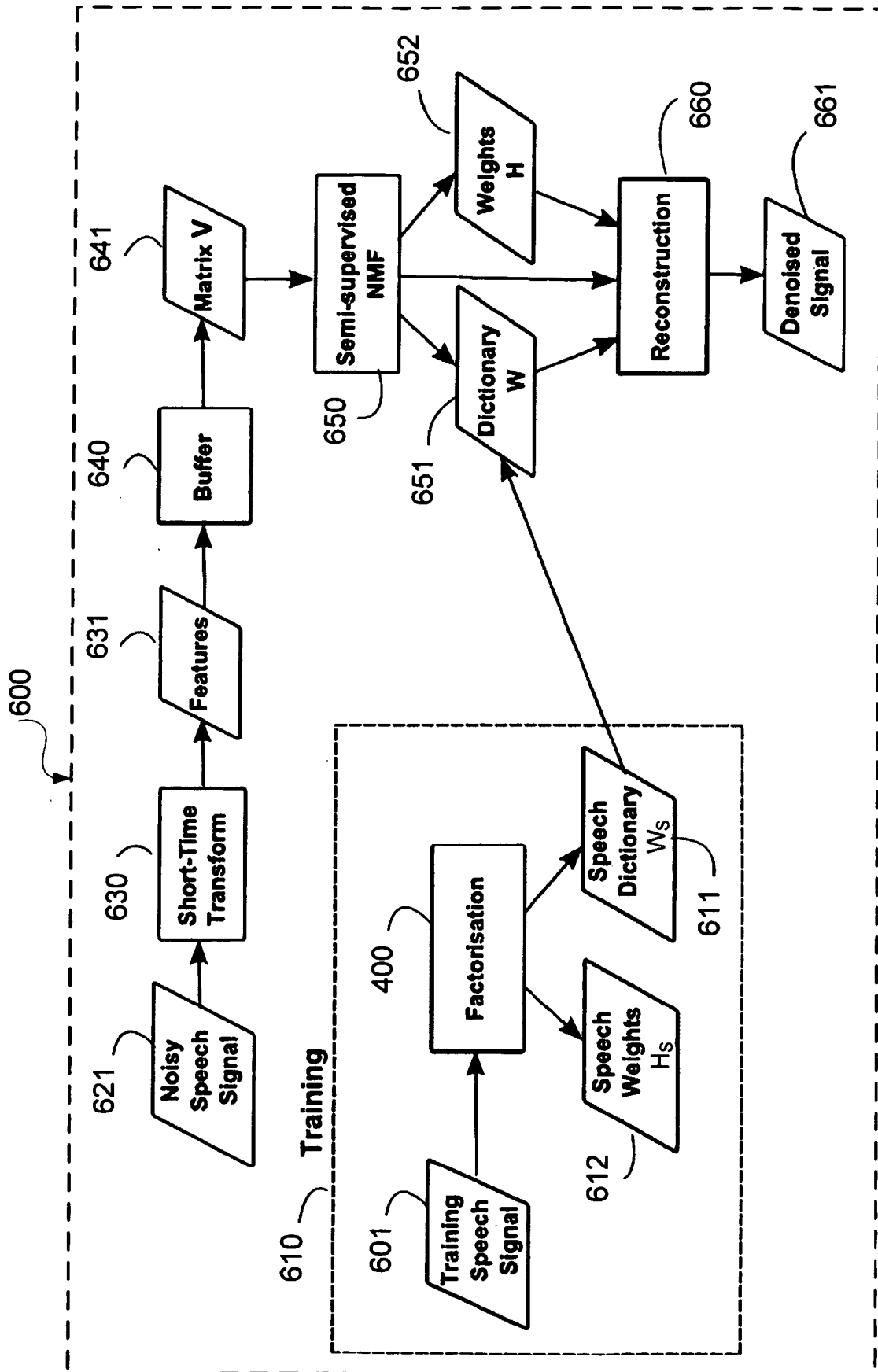


Fig. 6

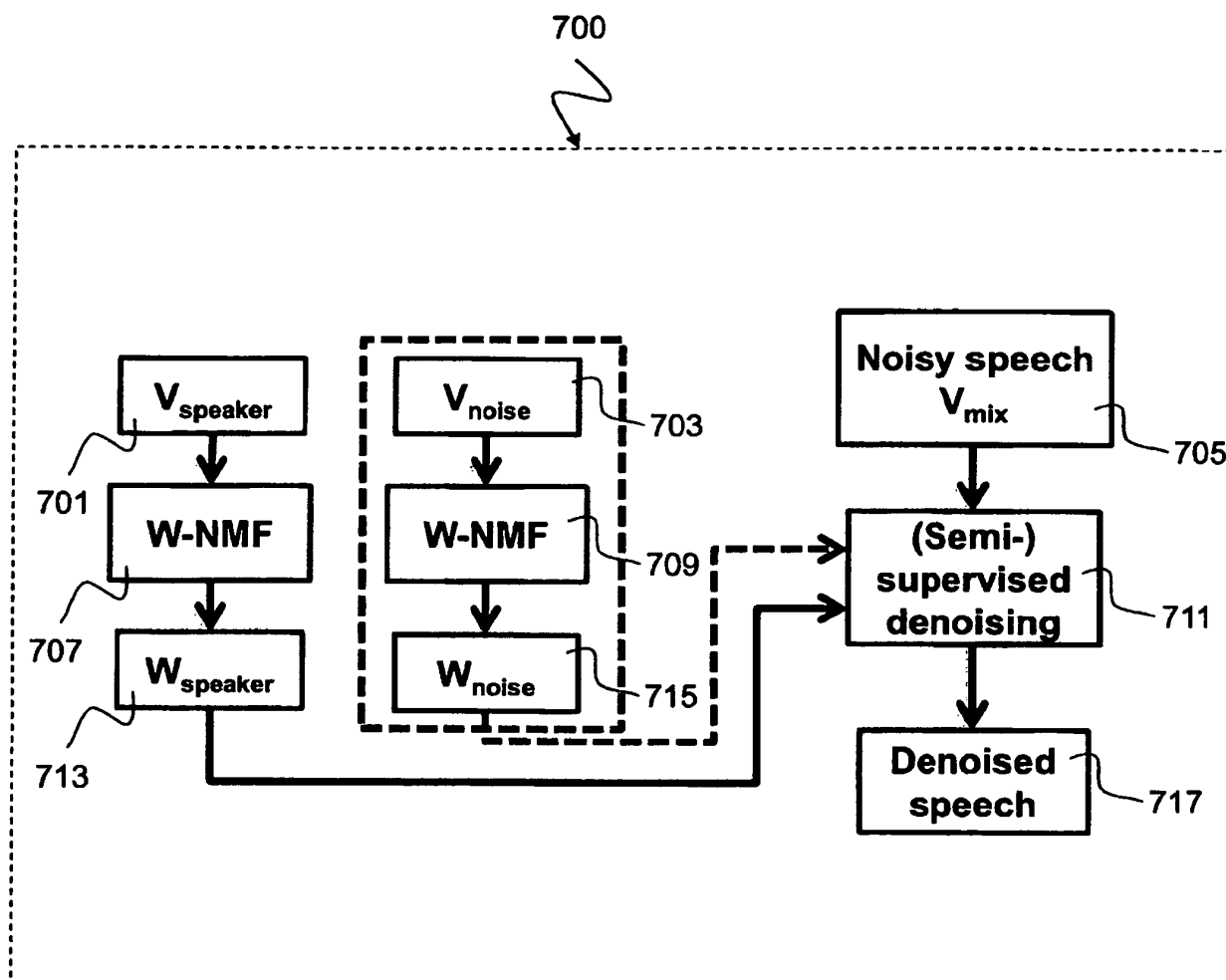


Fig. 7

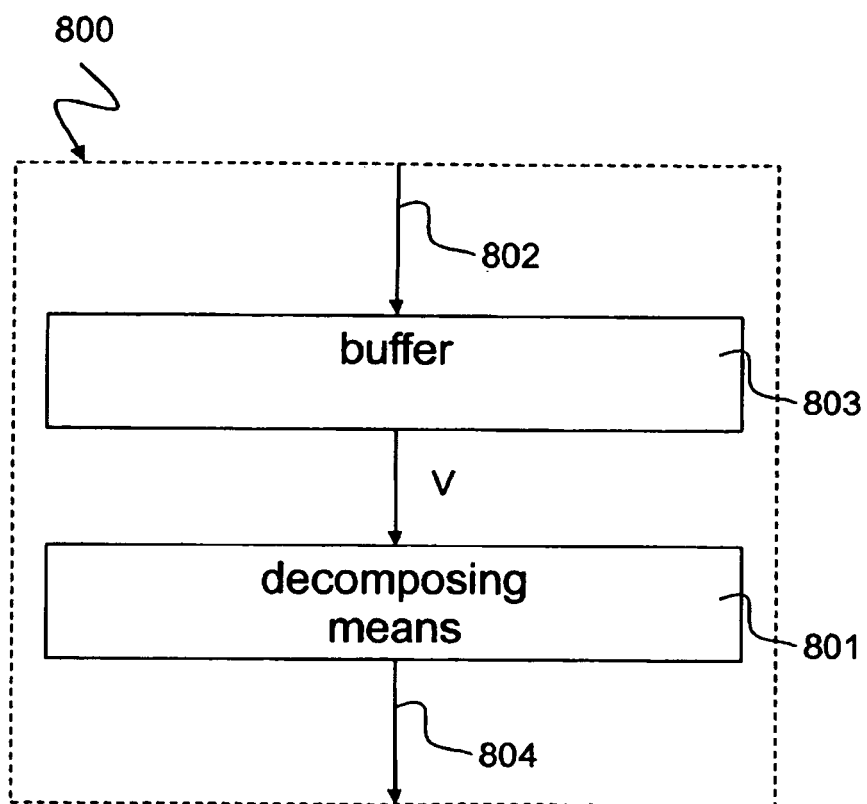


Fig. 8

REFERENCES CITED IN THE DESCRIPTION

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

Non-patent literature cited in the description

- **PAATERO.** Least Squares Formulation of Robust Non-Negative Factor Analysis. *Chemometrics and Intelligent Laboratory Systems*, 1997, vol. 37, 23-35 [0004]
- **C. JODER ; F. WENINGER ; F. EYBEN ; D. VIRETTE ; B. SCHULLER.** Real-time Speech Separation by Semi-Supervised Nonnegative Matrix Factorization. *Proc. International Conference on Latent Variable Analysis and Signal Separation*, March 2012 [0004]
- Real-time Speech Separation by Semi-Supervised Nonnegative Matrix Factorization. **C. JODER ; F. WENINGER ; F. EYBEN ; D. VIRETTE ; B. SCHULLER.** *Proc. of LVA/ICA*. Springer, 2012, 322-329 [0006]
- **A. CICHOCKI ; R. ZDUNEK ; S. AMARI.** New Algorithms for Non-negative Matrix Factorization in Application to Blind Source Separation. *Proc. of IEEE ICASSP*, 2006 [0013]
- **T. VIRTANEN.** Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria. *IEEE Trans. on Audio, Speech and Signal Process.*, 2007, vol. 15 (3), 1066-1074 [0013]
- **P. HOYER.** Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, 2004, vol. 5, 1457-1469 [0013]
- **MUNUEL MOUSSALLAM et al.** *Audio source separation informed by redundancy with greedy multi-scale decompositions*, 27 August 2012, 2644-2648 [0013]
- **ROBERT PEHARZ et al.** *Sparse nonnegative matrix factorization with constraints*, 15 March 2012 [0013]