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

(11) **EP 3 121 811 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
25.01.2017 Bulletin 2017/04

(51) Int Cl.:
G10L 19/005 *(2013.01)* ***G10L 21/0208*** *(2013.01)*

(21) Application number: **15306212.0**

(22) Date of filing: **24.07.2015**

(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
Designated Extension States:
BA ME
Designated Validation States:
MA

(71) Applicant: **Thomson Licensing**
92130 Issy-les-Moulineaux (FR)

(72) Inventors:
- **Bilen, Cagdas**
  35000 Rennes (FR)
- **Ozerov, Alexey**
  35000 Rennes (FR)
- **Perez, Patrick**
  35200 Rennes (FR)

(74) Representative: **Huchet, Anne et al**
TECHNICOLOR
1-5, rue Jeanne d'Arc
92130 Issy-les-Moulineaux (FR)

(54) METHOD FOR PERFORMING AUDIO RESTAURATION, AND APPARATUS FOR PERFORMING AUDIO RESTAURATION

(57) A method for performing audio inpainting, wherein missing portions in an input audio signal are recovered and a recovered audio signal is obtained, comprises computing a Short-Time Fourier Transform (STFT) on portions of the input audio signal, computing conditional expectations of the source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$ and complex Short-Time Fourier Transform (STFT) coefficients of the input audio signals are used, iteratively re-calculating the variance tensor $\boldsymbol{V}$ from the estimated power spectra $\mathbf{P}(f, n, j)$ and re-calculating updated estimated power spectra $\boldsymbol{P}(f, n, j)$, computing an array of STFT coefficients $\hat{S}$ from the resulting variance tensor $\boldsymbol{V}$ according to $\hat{\boldsymbol{s}}(f,n,j) = E\{\boldsymbol{S}(f, n, j)|\boldsymbol{x}, \boldsymbol{I_S}, \boldsymbol{I_L}, \boldsymbol{V}\}$, and converting the array of STFT coefficients $\hat{S}$ to the time domain, wherein coefficients $\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, ..., \tilde{\boldsymbol{s}}_J$ of the recovered audio signal are obtained.

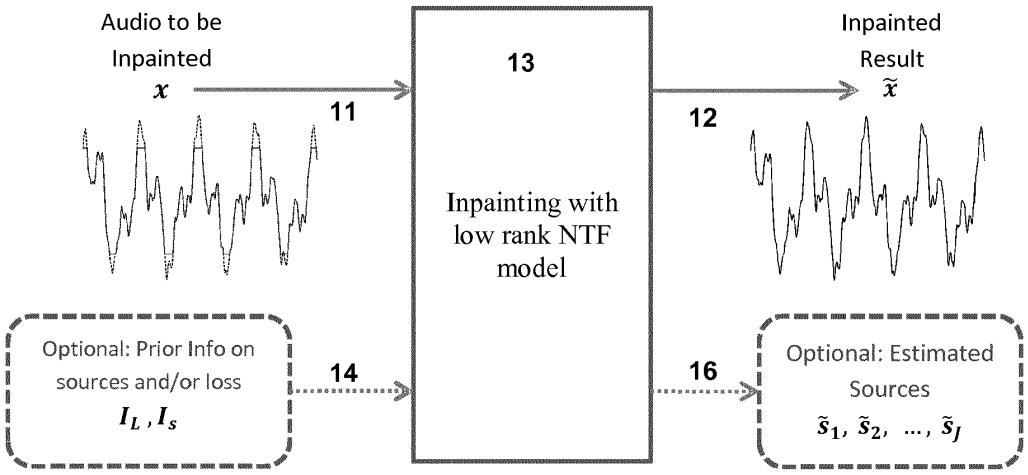

Fig.1

EP 3 121 811 A1

Description

Field of the invention

[0001] This invention relates to a method for performing audio restauration and to an apparatus for performing audio restauration. One particular type of audio restauration is audio inpainting.

Background

[0002] The problem of audio inpainting can be defined as the one of reconstructing the missing parts in an audio signal [1]. The name of "audio inpainting" was given to this problem to draw an analogy with image inpainting, where the goal is to reconstruct some missing regions in an image. A particular problem is audio inpainting in the case where some temporal samples of the audio are lost, ie. samples of the time domain. This is different from some known solutions that focus on lost samples in the time-frequency domain. This problem occurs e.g. in the case of saturation of amplitude (clipping) or interference of high amplitude impulsive noise (clicking). In such case, the samples need to be recovered (de-clipping or de-clicking respectively).

[0003] There exist methods for audio inpainting problems such as audio de-clipping [1], [2] and de-clicking [1]. In [1], audio inpainting is accomplished by enforcing sparsity of the audio signal in a Gabor dictionary which can be used both for audio de-clipping and de-clicking. For de-clipping, the approach proposed in [2] similarly relies on sparsity of audio signals in Gabor dictionaries while also optimizing for an adaptive sparsity pattern using the concept of social sparsity. Combined by the constraint of signal magnitude having to be greater than a clipping threshold, the method in [2] is shown to be much more effective than earlier works such as [1].

Summary of the Invention

[0004] The proposed solution is expected to not only perform better than these sparsity inducing approaches, but also be computationally less expensive. Furthermore the approaches based on time domain sparse dictionaries such as Gabor dictionary do not inherently result in phase invariant results, whereas the Non-negative Tensor Factorization (NTF) based model used herein is designed to be phase-invariant. This means that the models employed by the known methods need to be extended at the expense of performance in order to be near phase-invariant, whereas the proposed approach has no such drawback.

[0005] Existing methods [1], [2] usually rely on some sparse models (i.e., the signal is represented with few activation coefficients in some dictionary of elementary signals) [1] or locally-structured sparse models (ie., relations between activation coefficients are locally enforced) [2]. The models exploiting some global audio signal structure (e.g., long-term similarity of time or frequency patterns) were not applied for these problems. According to the present principles, an audio inpainting method applied to recover (short) missing temporal parts is based on a Non-negative Tensor Factorization (NTF) model. This method is more efficient than the known methods [1], [2], since the NTF model exploits some global audio signal structure (notably the long-term similarity of frequency patterns) in the time domain. NTF-like models were already used for missing audio reconstruction in the time-frequency domain [3]. The main difference with these approaches is that they assume the missing parts to be defined in some time-frequency domain, while here missing temporal parts (ie. in the time domain) are considered.

[0006] An additional problem considered herein and not considered by earlier works is performing audio inpainting jointly with source separation. Source separation problem can be defined as separating an audio signal into multiple sources often with different characteristics, for example separating a music signal into signals from different instruments. When the audio to be inpainted is known to be a mixture of multiple sources and some information about the sources is available (e.g. temporal source activity information [4], [5]), it can be easier to separate the sources while at the same time explicitly modeling the unknown mixture samples as missing. This situation may happen in many real-world scenarios, e.g. when one needs separating a recording that was clipped, which happens quite often. It was found that a sequential application of inpainting and source separation in one order or another is suboptimal, since the latter stage processing will suffer from the errors produced on the former stage processing, while within the joint processing these errors may be compensated. Moreover, distortion such as clipping may have quite harmful impact on the audio signal in the Short-Time Fourier Transform (STFT) domain, thus possibly destroying the low-rank signal structure and making the NTF modeling poorer. Treating the clipped values as missing within the joint approach should avoid this problem. Disclosed herein is a method for audio inpainting that uses a low-rank NTF model to model the audio signals. The disclosed method does not rely on a fixed dictionary but instead relies on a more general model representing global signal structure, which is also automatically adapted to the reconstructed audio signals. In addition to being naturally extendable to handle the joint inpainting and source separation problem, the disclosed method is also highly parallelizable for faster and more efficient computation.

[0007] In one embodiment, the present invention relates to a method for performing audio restauration, wherein missing coefficients of an input audio signal are recovered and a recovered audio signal is obtained. The method comprises steps of initializing a variance tensor ***V*** such that it is a low rank tensor that can be composed from component matrices ***H,Q,W*** (or initializing said component matrices ***H***,***Q***,***W*** to obtain the low rank variance tensor ***V***), iteratively applying the following steps, until convergence of the component matrices ***H***,***Q***,***W***: computing conditional expectations of source power spectra of the input audio signal, wherein estimated source power spectra ***P***(*f*,*n*,*j*) are obtained and wherein the variance tensor ***V,*** known signal values of the input audio signal and time domain information on loss ($\boldsymbol{I_L}$) are input to the computing, re-calculating the component matrices ***H***,***Q***,***W*** and the variance tensor ***V*** using the estimated source power spectra ***P***(*f*,*n*,*j*) and current values of the component matrices ***H,Q,W***, upon convergence of the component matrices ***H,Q,W***, computing a resulting variance tensor ***V'***, and computing from the resulting variance tensor ***V'***, from known signal values (**x,y**) of the input audio signal and from time domain information on loss ($\boldsymbol{I_L}$), an array of a posterior mean of Short Time Fourier Transform (STFT) samples (*S*) of the recovered audio signal, and converting coefficients of the array of the posterior mean of the STFT samples (*S*) to the time domain, wherein coefficients ($\tilde{\boldsymbol{s}}_1,\tilde{\boldsymbol{s}}_2,...,\tilde{\boldsymbol{s}}_J$) of the recovered audio signal are obtained.

[0008] In one embodiment, a computer readable medium has stored thereon executable instructions that when execution on a computer cause the computer to perform a method comprising steps of the method as disclosed in claim 1.

[0009] In one embodiment, an apparatus for performing audio inpainting comprises at least one of a hardware component and a hardware processor, and a non-transitory, tangible, computer-readable, storage medium tangibly embodying at least one software component, and the software component when executing on the at least one hardware component or hardware processor cause steps of the method of claim 1.

[0010] Further objects, features and advantages of the invention will become apparent from a consideration of the following description and the appended claims when taken in connection with the accompanying drawings.

Brief description of the drawings

[0011] Exemplary embodiments of the invention are described with reference to the accompanying drawings, which shows in

Fig.1 the structure of audio inpainting;
Fig.2 more details on an audio inpainting system;
Fig.3 a flow-chart of a method; and
Fig.4 elements of an apparatus.

Detailed description of embodiments

[0012] Fig.1 shows the structure of audio inpainting. It is assumed that the audio signal ***x*** to be inpainted is given with known temporal positions of the missing samples. For the problem with joint source separation, some prior information for the sources can also be provided. E.g. some samples from individual sources may be provided, simply because they were kept during the audio mixing step or because some temporal source activity information was provided by a user, e.g. as described in [4], [5]. Additionally, further information on the characteristics of the loss in the signal ***x*** can also be provided. E.g. for the de-clipping problem, the clipping threshold is given so that the magnitude of the lost signal can be constrained, in one embodiment. Given the signal x, the problem is to find the inpainted signal $\tilde{\boldsymbol{x}}$ for which the estimated sections are to be as close as possible to the original signal before the loss (ie. before clipping or clicking). If some prior information on the sources is available, the problem definition is extended to include joint source separation so that the individual sources are also estimated that are as close as possible to the original sources (before mixing and loss).

[0013] Throughout this specification, the time-domain signals will be represented by a letter with two primes, e.g. x", framed and windowed time-domain signals will be denoted by a letter with one prime, e.g. x', and complex-valued short-time Fourier transforms (STFT) coefficients will be denoted by a letter with no primes, e.g. x. The following is a single-channel mixing equation in the time domain:

$$x_t'' = \sum_{j=1}^{J} s_{jt}'' + a_t'', \qquad t=1,...,T \tag{1}$$

where t=1,...,T is the discrete time index, j=1,...,J is the source index, and x_t'', s_{jt}'' and a_{jt}'' denote respectively mixture, source and quantization noise samples. Moreover, it is assumed that the mixture is only observed on a subset of time indices $\Xi'' \subset \{1, ..., T\}$ called *mixture observation support* (MOS). For clipped signals this support indicates the indices with magnitude smaller than the clipping threshold.

[0014] The sources are unknown. It is assumed, however, that it is known which sources are active at which time periods. For example for a multi instrument music, this information corresponds to knowing which instruments are playing at any instant. Furthermore it is also assumed that if the mixture is clipped, the clipping threshold is known.

[0015] The time domain signals are converted into their windowed-time version using overlapping frames of length M. In this domain, mixing equation (1) reads

$$x'_{mn} = \sum_{j=1}^{J} s'_{jmn} + a'_{mn}, \qquad \text{m=1,...,M, n=1,...,N} \qquad (2)$$

where n=1,...,N is the frame index and m=1,...,M is an index within the frame. We also introduce the set $\Xi'_n \subset \{1,\dots,Mt\} \times \{1,\dots,Nt\}$ that is the MOS within the framed representation corresponding to Ξ" in the time domain, and its frame-level restriction Ξ' = {*m*|(*m*,*n*) ∈ Ξ'}. In this specification, the observed clipped mixture in the windowed time domain will be denoted as x'_{c} and its restriction to unclipped instants as $\bar{\mathrm{x}}'$, where $\bar{\mathrm{x}}'_n = [x'_{mn}]_{m\epsilon\Xi'_n}$.

[0016] Let $U \epsilon \mathbb{C}^{M\times F}$ be the complex-valued Hermitian matrix of the Discrete Fourier Transform (DFT). Applying this transform to eq.(2) yields the STFT domain model:

$$x_{fn} = \sum_{j=1}^{J} s_{jfn} + a_{fn}, \qquad \text{f=1,...,F, n=1,...,N} \qquad (3)$$

where *f*=1,...,F is the frequency bin index, $\mathrm{x_n} = \mathrm{Ux'_n}$, $\mathrm{s_{jn}} = \mathrm{Us'_{jn}}$ and $\mathrm{a_n} = \mathrm{Ua'_n}$ are STFT frames (F-length column vectors) obtained from the corresponding time frames (M-length column vectors). For example, $\mathrm{x_n} = [x_{fn}]_{f=1,\dots,F}$ is a mixture STFT frame and $\mathrm{x'_n} = [x'_{mn}]_{m=1,\dots,M}$ is a mixture time frame. The sources are modelled in the STFT domain with a normal distribution ($s_{jfn} \sim N_c(0, v_{ifn})$), where the variance tensor V = [v_{jfn}] has the following low-rank NTF structure

$$v_{jfn} = \sum_{k=1}^{K} q_{jk} w_{fk}\, h_{nk}, \qquad (4)$$

where k<max(J,F,N) and all the variables are non-negative reals. This model is parameterized by Θ = {**Q**, **W**, **H**}, with **Q** = $[q_{jk}]_{j,k}$, **W** = $[w_{fk}]_{f,k}$ and **H** = $[h_{nk}]_{n,k}$ being, respectively, $J \times K$, $F \times K$ and $N \times K$ non-negative matrices.

[0017] The assumed information on which sources are active at which time periods is captured by constraining certain entries of Q and H to be zero [5]. Each of the *K* components being assigned to a single source through $Q(\Psi_Q) \equiv 0$ for some appropriate set Ψ_Q of indices, the components of each source are marked as silent through $H(\Psi_H) \equiv 0$ with an appropriate set Ψ_H of indices.

[0018] Finally, for the sake of simplicity it is assumed that there is no mixture quantization ($a'_{mn} = 0$). Note however that assuming a complex valued normal distribution instead for this error only requires minor changes. The problem at hand is now the estimation of the model parameters Θ and of the unknown un-clipped sources $\{s_{jn}\}_n$, j = 1,...,J, given the observed clipped mixture x'_c.

[0019] Fig.2 shows more details on an exemplary audio inpainting system in a case where prior information on loss $\boldsymbol{I_L}$ and/or prior information on sources $\boldsymbol{I_S}$ are available.

[0020] In one embodiment, the invention performs audio inpainting by enforcing a low-rank non-negative tensor structure for the covariance tensor of the Short-Time

[0021] Fourier Transform (STFT) coefficients of the audio signal. It estimates probabilistically the most likely signal $\tilde{\boldsymbol{x}}$, given the input audio $\boldsymbol{x}$ and some prior information on the loss in the signal $\boldsymbol{I_L}$, based on two assumptions:

[0022] First assumption is that the sources are jointly Gaussian distributed in the Short-Time Fourier Transform (STFT) domain with window size F and number of windows *N*.

[0023] Second assumption is that the variance tensor of the Gaussian distribution, $\boldsymbol{V} \in \mathrm{R}_{+}^{F\times N\times J}$, has a low rank Non-Negative Tensor Decomposition (NTF) of rank *K* such that

$$\boldsymbol{V}(f,n,j) = \sum_{k=1}^{K} \boldsymbol{H}(n,k)\boldsymbol{W}(f,k)\boldsymbol{Q}(j,k), \qquad \boldsymbol{H} \in \mathrm{R}_{+}^{N\times K}, \boldsymbol{W} \in \mathrm{R}_{+}^{F\times K}, \boldsymbol{Q} \in \mathrm{R}_{+}^{J\times K} \quad (5)$$

[0024] Both assumptions are usually fulfilled. Further, estimation of the sources $\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \ldots, \tilde{\boldsymbol{s}}_J$ is further improved if some prior information on the sources $\boldsymbol{I}_S$ is given.

[0025] In the following, the most general case will be described, wherein samples from multiple sources are available. In the case that information on multiple sources are not provided, one can simply assume that there is a single source $J = 1$ and the known samples of the source coincide with the input audio signal. In an exemplary embodiment, an implementation of the invention can be summarized with the following steps:

1. Initialize the variance tensor $\boldsymbol{V} \in \mathrm{R}_{+}^{F\times N\times J}$ by random matrices $\boldsymbol{H} \in \mathrm{R}_{+}^{N\times K}$, $\boldsymbol{W} \in \mathrm{R}_{+}^{F\times K}$, $\boldsymbol{Q} \in \mathrm{R}_{+}^{J\times K}$ such that:

$$\boldsymbol{V}(f,n,j) = \sum_{k=1}^{K} \boldsymbol{H}(n,k)\boldsymbol{W}(f,k)\boldsymbol{Q}(j,k) \qquad (6)$$

2. Until convergence or maximum number of iterations reached, repeat:

2.1 Compute the conditional expectations of the source power spectra such that

$$\boldsymbol{P}(f,n,j) = \mathrm{E}\{|\boldsymbol{S}(f,n,j)|^2 | \boldsymbol{x}, \boldsymbol{I}_S, \boldsymbol{I}_L, \boldsymbol{V}\} \qquad (7)$$

where $\boldsymbol{S} \in C^{F\times N\times J}$ is the array of the STFT coefficients of the sources. This step can be performed for each STFT frame independently, hence providing significant gain by parallelism. More details on this posterior mean computation can be found below.

2.2 Re-estimate NTF model parameters $\boldsymbol{H} \in \mathrm{R}_{+}^{N\times K}$, $\boldsymbol{W} \in \mathrm{R}_{+}^{F\times K}$, $\boldsymbol{Q} \in \mathrm{R}_{+}^{J\times K}$ using the multiplicative update (MU) rules minimizing the Itakura-Saito divergence (IS divergence) [6] between the 3-valence tensor of estimated source power spectra $\boldsymbol{P}(f,n,j)$ and the 3-valence tensor of the NTF model approximation $\boldsymbol{V}(f,n,j)$ such that:

$$\boldsymbol{Q}'^{(j,k)} \leftarrow \boldsymbol{Q}(j,k) \left(\frac{\sum_{f,n} \boldsymbol{W}(f,k)\boldsymbol{H}(n,k)\boldsymbol{P}(f,n,j)\,\boldsymbol{V}(f,n,j)^{-2}}{\sum_{f,n} \boldsymbol{W}(f,k)\boldsymbol{H}(n,k)\boldsymbol{V}(f,n,j)^{-1}} \right) \quad (8)$$

$$\boldsymbol{W}'^{(f,k)} \leftarrow \boldsymbol{W}(f,k) \left(\frac{\sum_{j,n} \boldsymbol{Q}(j,k)\boldsymbol{H}(n,k)\boldsymbol{P}(f,n,j)\,\boldsymbol{V}(f,n,j)^{-2}}{\sum_{j,n} \boldsymbol{Q}(j,k)\boldsymbol{H}(n,k)\boldsymbol{V}(f,n,j)^{-1}} \right) \quad (9)$$

$$\boldsymbol{H}'^{(n,k)} \leftarrow \boldsymbol{H}(n,k) \left(\frac{\sum_{f,j} \boldsymbol{W}(f,k)\boldsymbol{Q}(j,k)\boldsymbol{P}(f,n,j)\,\boldsymbol{V}(f,n,j)^{-2}}{\sum_{f,j} \boldsymbol{W}(f,k)\boldsymbol{Q}(j,k)\boldsymbol{V}(f,n,j)^{-1}} \right) \quad (10)$$

Then update V by

$$\boldsymbol{V}'^{(f,n,j)} = \sum_{k=1}^{K} \boldsymbol{H}'^{(n,k)}\boldsymbol{W}'^{(f,k)}\boldsymbol{Q}'^{(j,k)} \qquad (11)$$

This can be repeated multiple times.

3. Compute the array of STFT coefficients $\boldsymbol{S} \in C^{F \times N \times J}$ as the posterior mean as

$$\hat{\boldsymbol{S}}(f,n,j) = \mathrm{E}\{\boldsymbol{S}(f,n,j)|\boldsymbol{x},\boldsymbol{I_s},\boldsymbol{I_L},\boldsymbol{V}\} \qquad (12)$$

and convert back into the time domain to recover the estimated sources $\tilde{\boldsymbol{s}}_1,\tilde{\boldsymbol{s}}_2,\ldots,\tilde{\boldsymbol{s}}_J$. Set the estimated signal as $\boldsymbol{x} = \sum_{j=1}^{J} \tilde{\boldsymbol{s}}_j$. More details on this posterior mean computation can be found below.

[0026] The following describes some mathematical basics on the above calculations.

[0027] A tensor is a data structure that can be seen as a higher dimensional matrix. a matrix is 2-dimensional, whereas a tensor can be N-dimensional. In the present case, V is a 3-dimensional tensor (like a cube) that represents the covariance matrix of the jointly Gaussian distribution of the sources.

[0028] A matrix can be represented as the sum of few rank-1 matrices, each formed by multiplying two vectors, in the low rank model. In the present case, the tensor is similarly represented as the sum of K rank one tensors, where a rank one tensor is formed by multiplying three vectors, e.g. h_i, q_i and w_i. These vectors are put together to form the matrices ***H, Q*** and ***W***. There are K sets of vectors for the K rank one tensors. Essentially, the tensor is represented by K components, and the matrices ***H, Q*** and ***W*** represent how the components are distributed along different frames, different frequencies of STFT and different sources respectively.

[0029] Similar to a low rank model in matrices, K is kept small because a small K better defines the characteristics of the data, such as audio data, e.g. music. Hence it is possible to guess unknown characteristics of the signal by using the information that V should be a low rank tensor. This reduces the number of unknowns and defines an interrelation between different parts of the data.

[0030] The steps of the above-described iterative algorithm can be described as follows. First, initialize the matrices ***H, Q*** and ***W*** and therefore ***V***. Note that it is also possible to initialize ***V*** and then obtain the initial matrices ***H, Q*** and ***W*** from it, since ***H, Q*** and ***W*** directly define ***V***. After the initialization, ***V*** always equals to the multiplied sum of ***H, Q*** and ***W***, so it is a low rank tensor. If there is only one source, then ***Q*** does not exist (or equivalently can be set to be a constant), so that ***V*** is a low rank matrix. Note further that ***H, Q*** and ***W*** may also be called "model parameters" or "low-rank components" herein.

[0031] Given ***V***, the probability distribution of the signal is known. And looking at the observed part of the signals (signals are observed only partially), it is possible to estimate the STFT coefficients $\hat{S}$, e.g. by Wiener filtering. This is the posterior mean of the signal. Further, also a posterior covariance of the signal is computed, which will be used below. This step is performed independently for each window of the signal, and it is parallelizable. This is called the expectation step (*E-step*).

[0032] The posterior mean $\hat{s}_{jn}$ and posterior covariance $\hat{\Sigma}_{s_{jn}s_{jn}}$ can be computed by

$$\hat{s}_{jn} = \sum_{\bar{x}_n' s_{jn}}^{H} \sum_{\bar{x}_n' \bar{x}_n'}^{-1} \bar{x}_n' \qquad (13)$$

$$\hat{\Sigma}_{s_{jn}s_{jn}} = \Sigma_{s_{jn}s_{jn}} - \sum_{\bar{x}_n' s_{jn}}^{H} \sum_{\bar{x}_n' \bar{x}_n'}^{-1} \sum_{\bar{x}_n' s_{jn}} \qquad (14)$$

given the definitions

$$\Sigma_{s_{jn}s_{jn}} = diag\left(\left[v_{jfn}\right]_f\right) \qquad (15)$$

$$\sum_{\bar{x}_n' s_{jn}} = U^H(\Xi_n')diag\left(\left[v_{jfn}\right]_f\right) \qquad (16)$$

$$\sum_{\vec{x}'_n \vec{x}'_n} = U^H(\Xi'_n) diag\left(\left[\sum_j v_{jfn}\right]_f\right) U(\Xi'_n) \qquad (17)$$

where $U(\Xi'_n)$ is the $M \times |\Xi'_n|$ matrix of columns from U with index in Ξ'_n. For a de-clipping application, it is also known that the estimated mixture must obey

$$\left|\mathrm{U}^H(\Xi'_n) \sum_j \hat{s}_{jn}\right| \geq \left|x'_{c,jn}(\Xi'_n)\right| \qquad (18)$$

[0033] This is difficult to enforce directly into the model since posterior distribution of the sources under this prior would no longer be Gaussian. In order to find a workaround, let's suppose that eq.(18) is not satisfied at the indices $\hat{\Xi}'_n$. A simple way to enforce eq.(18) can be directly scaling up the magnitude of the sources at window indices $\hat{\Xi}'_n$ so that eq.(18) is satisfied.

[0034] The clipping constraint can be handled as follows.

[0035] In order to update the model parameters, one needs to estimate the posterior power spectra of the signal defined as $\tilde{\mathrm{p}}_{jfn} = \mathbb{E}\left[\left|s_{jfn}\right|^2 \overline{\mathrm{x}}'_n; \Theta\right]$. For an audio inpainting problem without any further constraints, the posterior signal estimate $\hat{s}_n$ and the posterior covariance matrix $\hat{\Sigma}_{s_n s_n}$ would be sufficient to estimate $\tilde{p}_{fn}$ since the posterior distribution of the signal is Gaussian. However, in clipping, the original unknown signal is known to have its magnitude above the clipping threshold outside the OS, and so should have the reconstructed signal frames $\hat{s}'_n = U^H \hat{s}_n$:

$$\hat{s}'_{mn} \times sign(x'_{mn}) \geq |x'_{mn}|, \forall n, \forall m \notin \Xi'_n$$

[0036] This constraint is difficult to enforce directly into the model since the posterior distribution of the signal under it is no longer Gaussian, which significantly complicates the computation of the posterior power spectra. In the presence of such constraints on the magnitude of the signal, various ways can be considered to approach the problem:

[0037] Unconstrained: The simplest way to perform the estimation is to ignore completely the constraints, treating the problem as a more generic audio inpainting in time domain. Hence during the iterations, the "constrained" signal is taken simply as the estimated signal, i.e. $\tilde{s}_n = \hat{s}_n$, $n = 1, ..., N$, as is the posterior covariance matrix,

$$\tilde{\Sigma}_{s_n s_n} = \hat{\Sigma}_{s_n s_n}, n = 1, \ldots, N.$$

[0038] Ignored projection: Another simple way to proceed is to ignore the constraint during the iterative estimation process and to enforce it at the end as a post-processing of the estimated signal. In this case, the signal is treated the same as the unconstrained case during the iterations.

[0039] Signal projection: A more advanced approach is to update the estimated signal at each iteration so that the magnitude obeys the clipping constraints. Let's suppose eq. (18) is not satisfied at the indices in set $\hat{\Xi}'_n$. We can set $\tilde{s}'_n = \hat{s}'_n$ and then force $\tilde{s}'_n\left(\hat{\Xi}'_n\right) = x'_{c,n}\left(\hat{\Xi}'_n\right)$. However, this approach does not update the posterior covariance matrix, ie. $\hat{\Sigma}_{s_n s_n} = \hat{\Sigma}_{s_n s_n}$, $n = 1, ..., N$, which is needed to compute the posterior power spectra of the sources to update the NTF model.

[0040] Covariance projection: In order to update as well the posterior covariance matrix, we can re-compute the posterior mean and the posterior covariance by eq. (13) and (14) respectively. The posterior mean and the posterior

covariance are simply re-computed with the above equations respectively, by using $\Xi'_n \cup \hat{\Xi}'_n$ instead of Ξ'_n , and $x'_{c,jn}(\Xi'_n \cup \hat{\Xi}'_n)$ instead of $\bar{\mathrm{x}}'_n$ in eq.(13)-(17).

[0041] If the resulting estimation of the sources violate eq.(18) on additional indices, $\hat{\Xi}'_n$ is extended to include these indices and the computation is repeated.

[0042] As a result, final sources estimates s that satisfy eq.(18) and the corresponding posterior covariance matrix $\tilde{\Sigma}_{s_n s_n}$ are obtained. Note that in addition to updating the posterior covariance matrix, this approach also updates the entire estimated signal and not just the signal at the indices of violated constraints.

[0043] Therefore the posterior power spectra p, which will be used to update the NTF model as described in the following, can be computed as

$$\tilde{\mathrm{p}}_{fn} = \mathbb{E}\left[\left|s_{fn}\right|^2 \bar{\mathrm{x}}'_n; \Theta\right] \cong \left|\tilde{s}_{fn}\right|^2 + \tilde{\Sigma}_{s_n s_n}(f,f) \qquad (19)$$

[0044] Once the posterior mean and covariance are computed, these are used to compute the posterior power spectra ***p***. This is needed to update the earlier model parameters, ie. ***H, Q*** and ***W***.

[0045] NMF model parameters can be re-estimated using the multiplicative update (MU) rules minimizing the $\boldsymbol{I_S}$ divergence between the matrix of estimated signal power spectra $\tilde{P} = [\tilde{p}_{fn}]$ and the NMF model approximation $V=WH^T$:

$$D_{IS}(\tilde{P}||V) = \sum_{f,n} d_{IS}(\tilde{p}_{fn}||v_{fn}) \qquad (20)$$

where $d_{IS}(x||y) = \frac{x}{y} - \log(x/y) - 1$ is the $\boldsymbol{I_S}$ divergence, and $\tilde{p}_{fn}$ and v_{fn} are specified respectively by (19) and (12). Hence the model parameters can be updated as

$$w_{fk} \leftarrow w_{fk} \left(\frac{\sum_n h_{nk} \tilde{p}_{fn} {v_{fn}}^{-2}}{\sum_n h_{nk} {v_{fn}}^{-1}}\right) \qquad (21)$$

$$h_{nk} \leftarrow h_{nk} \left(\frac{\sum_f w_{fk} \tilde{p}_{fn} {v_{fn}}^{-2}}{\sum_f w_{fk} {v_{fn}}^{-1}}\right) \qquad (22)$$

[0046] It may be advantageous to repeat this step more than once in order to reach a better estimate (e.g. 2-10 times). This is called the maximization step (*M-step*). Once the model parameters ***H, Q*** and ***W*** are updated, all the steps (from estimating the STFT coefficients $\hat{S}$) can be repeated until some convergence is reached, in an embodiment. After the convergence is reached, in an embodiment the posterior mean of the STFT coefficients $\hat{S}$ is converted into the time domain to obtain an audio signal as final result.

[0047] The approximation of S and P, as described above, is based on the following basic idea. An exact computation of P normally relies on the assumption that the signal is Gaussian distributed with zero mean. When the distribution is Gaussian, posterior mean and posterior variance of the signal are enough to compute P. However, when some constraints exist, like information on loss I_L, the distribution is not Gaussian any more. With the true distribution, an exact computation of $\boldsymbol{P}(f,n,j) = E\{|\boldsymbol{S}(f,n,j)|^2 | \boldsymbol{x}, \boldsymbol{I_S}, \boldsymbol{I_L}, \boldsymbol{V}\}$ is computationally not viable. According to the present principles, the posterior estimate $\hat{\boldsymbol{S}}(f,n,j)$ is computed, and then the time domain signal is projected to the subspace satisfying the information on loss I_L. After that, it is assumed that the modified values (the values of $\hat{\boldsymbol{S}}$ not obeying I_L) are known for that iteration. When these values are assumed to be known to their current values, the rest of the unknowns can be assumed to be Gaussian again, and corresponding posterior mean and posterior variance can be computed. By using this, P can also be computed. Note that the values that are assumed to be known are only an approximation, so that P is also an approximation. However, P is altogether much more accurate than if the information on loss I_L would be ignored.

[0048] For information on loss $\boldsymbol{I_L}$, one example is the clipping threshold. If the clipping threshold thr is known, such that the unknown values of the time domain signal s_u is known to be s_u > thr if s_u >0, and s_u < -thr if s_u <0 for a known threshold thr. Other examples for information on loss $\boldsymbol{I_L}$ are the sign of the unknown value, an upper limit for the signal magnitude (essentially the opposite of the first example), and/or the quantized value of the unknown signal, so that there is the constraint $thr_2 < s_u < thr_1$. All these are constraints in the time domain. No other method is known that can enforce

them in a low rank NTF/NMF model enforced on the time frequency distribution of the signal. At least one or more of the above examples, in any combination, can be used as information on loss $\boldsymbol{I_L}$.

[0049] For information on sources $\boldsymbol{I_S}$, one example is information about which sources are active or silent for some of the time instants. Another example is a number of how many components each source is composed in the low rank representation. A further example is specific information on the harmonic structure of sources, which can introduce stronger constraints on the low rank tensor or on the matrix. These constraints are often easier to apply on the STFT coefficients or directly on the low rank variance tensor of the STFT coefficients or directly on the model, ie. on ***H, Q*** and ***W***.

[0050] One advantage of the invention is enabling efficient recovery of missing portions in audio signals that resulted from effects such as clipping and clicking.

[0051] A second advantage of the invention is the possibility of jointly performing inpainting and source separation tasks without the need for additional steps or components in the methodology. This enables the possibility of utilizing the additional information on the components of the audio signal for a better inpainting performance.

[0052] Further, a third advantage is making use of the NTF model and hence efficiently exploiting the global structure of an audio signal for an improved inpainting performance.

[0053] A fourth advantage of the invention is that it allows joint audio inpainting and source separation, as described below.

[0054] Further, the Non-negative tensor factorization (NTF) or Non-negative Matrix factorization (NMF) can be applied to improve dequantization of a quantized signal. As mentioned above, quantized signals can be handled by treating quantization noise as Gaussian. In a case where there are no other time domain losses, handling noisy signals with low rank NTF/NMF model is known. But since the present principles introduce a way to handle time domain constraints (with $\boldsymbol{I_L}$), this provides an opportunity to handle the quantized signals in a better way. More specifically, when the quantization step sizes are known, the quantized time domain signals are known to obey constraints such that

quant_level_low < s < quant_level_high

where the upper and lower bounds (quant_level_low/high) are known. Hence, it is possible to enforce this constraint while applying the low rank NMF/NTF model.

[0055] Fig.3 shows, in one embodiment, a flow-chart of a method 30 for performing audio inpainting, wherein missing portions in an input audio signal are recovered and a recovered audio signal is obtained. The method comprises initializing 31 a variance tensor $\boldsymbol{V}$ such that it is a low rank tensor that can be composed from component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ or initializing said component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ to obtain the low rank variance tensor $\boldsymbol{V}$, computing 32 of source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$, known signal values **x,y** of the input audio signal and time domain information on loss $\boldsymbol{I_L}$ are input to the computing, iteratively re-calculating 33 the component matrices $\boldsymbol{H,Q,W}$ and the variance tensor $\boldsymbol{V}$ using the estimated source power spectra $\boldsymbol{P}(f,n,j)$ and current values of the component matrices $\boldsymbol{H,Q,W,}$ and upon detecting convergence 34 of the component matrices $\boldsymbol{H,Q,W,}$ or upon reaching a predefined maximum number of iterations, computing 35 a resulting variance tensor $\boldsymbol{V'}$, and further computing 36 from the resulting variance tensor $\boldsymbol{V'}$, known signal values **x,y** of the input audio signal and time domain information on loss $\boldsymbol{I_L}$, an array of a posterior mean of Short Time Fourier Transform (STFT) samples *S* of the recovered audio signal, and converting 37 coefficients of the array of the posterior mean of the STFT samples *S* to the time domain, wherein coefficients $\tilde{\boldsymbol{s}}_1$, $\tilde{\boldsymbol{s}}_2$, ... , $\tilde{\boldsymbol{s}}_{\boldsymbol{J}}$ of the recovered audio signal are obtained.

[0056] In one embodiment, the estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained according to $\boldsymbol{P}(f,n,j) = E\{|\boldsymbol{S}(f,n,j)|^2|\boldsymbol{x},\boldsymbol{I_S},\boldsymbol{I_L},\boldsymbol{V}\}$, with $\boldsymbol{I_S}$ being time domain information on sources.

[0057] In one embodiment, the time domain information on sources $\boldsymbol{I_S}$ comprises at least one of: information about which sources are active or silent for a particular time instant, information about a number of how many components each source is composed in the low rank representation, and specific information on a harmonic structure of the sources.

[0058] In one embodiment, the time domain information on loss $\boldsymbol{I_L}$ comprises at least one of: a clipping threshold, a sign of an unknown value in the input audio signal, an upper limit for the signal magnitude, and the quantized value of an unknown signal in the input audio signal.

[0059] In one embodiment, the variance tensor $\boldsymbol{V}$ is initialized by random matrices $\boldsymbol{H} \in \mathrm{R}_+^{N\times K}$, $\boldsymbol{W} \in \mathrm{R}_+^{F\times K}$, $\boldsymbol{Q} \in \mathrm{R}_+^{J\times K}$, as explained above.

[0060] In one embodiment, the variance tensor $\boldsymbol{V}$ is initialized by values derived from known samples of the input audio signal.

[0061] In one embodiment, the input audio signal is a mixture of multiple audio sources, and the method further comprises receiving 38 side information comprising quantized random samples of the multiple audio signals, and performing 39 source separation, wherein the multiple audio signals from said mixture of multiple audio sources are separately obtained.

[0062] In one embodiment, the STFT coefficients are windowed time domain samples S. In one embodiment, the input

audio signal contains quantization noise, wherein wrongly quantized coefficients take the position of the missing coefficients, wherein the quantization levels are used as further constraints in said time domain information on loss $\boldsymbol{I_L}$, and wherein the recovered audio signal is a de-quantized audio signal.

[0063] Fig.4 shows, in one embodiment, an apparatus 40 for performing audio restauration, wherein missing portions in an input audio signal are recovered and a recovered audio signal is obtained. The apparatus comprises a processor 41 and a memory 42 storing instructions that, when executed on the processor, cause the apparatus to perform a method comprising initializing a variance tensor $\boldsymbol{V}$ such that it is a low rank tensor that can be composed from component matrices $\boldsymbol{H,Q,W}$, or initializing said component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ to obtain the low rank variance tensor $\boldsymbol{V}$, iteratively applying the following steps, until convergence of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$:

computing 32 conditional expectations of source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$, known signal values $\mathbf{x}$, $\mathbf{y}$ of the input audio signal and time domain information on loss ($\boldsymbol{I_L}$) are input to the computing, re-calculating 33 the component matrices $\boldsymbol{H,Q,W}$ and the variance tensor $\boldsymbol{V}$ using the estimated source power spectra $\boldsymbol{P}(f,n,j)$ and current values of the component matrices $\boldsymbol{H,Q,W}$,

upon convergence of the component matrices $\boldsymbol{H,Q,W\_}$, computing a resulting variance tensor $\boldsymbol{V'}$, and computing from the resulting variance tensor $\boldsymbol{V'}$, known signal values $\mathbf{x,y}$ of the input audio signal and time domain information on loss $\boldsymbol{I_L}$, an array of a posterior mean of Short Time Fourier Transform (STFT) samples S of the recovered audio signal, and converting 37 coefficients of the array of the posterior mean of the STFT samples S to the time domain, wherein coefficients $\tilde{\boldsymbol{s}}_1,\tilde{\boldsymbol{s}}_2$, ..., $\tilde{\boldsymbol{s}}_{\boldsymbol{J}}$ of the recovered audio signal are obtained.

[0064] In one embodiment, the estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained according to $\boldsymbol{P}(f,n,j) = \mathrm{E}\{|\boldsymbol{S}(f,n,j)|^2|\boldsymbol{x},\boldsymbol{I_S},I_L,\boldsymbol{V}\}$ with $\boldsymbol{I_S}$ being time domain information on sources.

[0065] In one embodiment, the time domain information on loss comprises at least one of: a clipping threshold, a sign of an unknown value in the input audio signal, an upper limit for the signal magnitude, and the quantized value of an unknown signal in the input audio signal.

[0066] In one embodiment, the input audio signal is a mixture of multiple audio sources, and the instructions when executed on the processor further cause the apparatus to receive 38 side information comprising quantized random samples of the multiple audio signals, and perform 39 source separation, wherein the multiple audio signals from said mixture of multiple audio sources are separately obtained. In one embodiment, the input audio signal contains quantization noise, wherein wrongly quantized coefficients take the position of the missing coefficients, wherein the quantization levels are used as further constraints in said time domain information on loss $\boldsymbol{I_L}$, and wherein the recovered audio signal is a de-quantized audio signal.

[0067] In one embodiment, the input audio signal contains quantization noise, wherein wrongly quantized coefficients take the position of the missing coefficients, wherein the quantization levels are used as further constraints in said time domain information on loss $\boldsymbol{I_L}$, and wherein the recovered audio signal is a de-quantized audio signal.

[0068] In one embodiment, an apparatus for performing audio restauration, wherein missing coefficients of an input audio signal are recovered and a recovered audio signal is obtained, comprises

first *computing means* for initializing 31 a variance tensor $\boldsymbol{V}$ such that it is a low rank tensor that can be composed from component matrices $\boldsymbol{H,Q,W}$, or for initializing said component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ to obtain the low rank variance tensor $\boldsymbol{V}$, second *computing means* for computing 32 conditional expectations of source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$, known signal values $\mathbf{x,y}$ of the input audio signal and time domain information on loss $\boldsymbol{I_L}$ are input to the computing, *calculating means* for iteratively re-calculating 33 the component matrices $\boldsymbol{H,Q,W}$ and the variance tensor $\boldsymbol{V}$ using the estimated source power spectra $\boldsymbol{P}(f,n,j)$ and current values of the component matrices $\boldsymbol{H,Q,W}$, *detection means* for detecting 34 convergence of the component matrices $\boldsymbol{H,Q,W}$ or for detecting that a predefined maximum number of iterations is reached,

third *computing means* for computing 35, upon said convergence of the component matrices $\boldsymbol{H,Q,W}$ or upon reaching said predefined maximum number of iterations, a resulting variance tensor $\boldsymbol{V'}$, fourth *computing means* for computing 36 from the resulting variance tensor $\boldsymbol{V'}$, known signal values $\mathbf{x,y}$ of the input audio signal and time domain information on loss $\boldsymbol{I_L}$, an array of a posterior mean of Short Time Fourier Transform (STFT) samples S of the recovered audio signal, and *converter means* for converting 37 coefficients of the array of the posterior mean of the STFT samples S to the time domain, wherein coefficients $\tilde{\boldsymbol{s}}_1$, $\tilde{\boldsymbol{s}}_2$, ... , $\tilde{\boldsymbol{s}}_{\boldsymbol{J}}$ of the recovered audio signal are obtained. The coefficients $\tilde{\boldsymbol{s}}_1$, $\tilde{\boldsymbol{s}}_2$, ..., $\tilde{\boldsymbol{s}}_{\boldsymbol{J}}$ of the recovered audio signal can be used e.g. to reproduce or store the recovered audio signal.

[0069] Usually, the invention leads to a low-rank tensor structure in the power spectrogram of the reconstructed signal.

[0070] The use of the verb "comprise" and its conjugations does not exclude the presence of elements or steps other than those stated in a claim. Furthermore, the use of the article "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. Several "means" may be represented by the same item of hardware. Furthermore, the invention resides in each and every novel feature or combination of features. As used herein, a "digital audio signal" or "audio signal" does not describe a mere mathematical abstraction, but instead denotes information embodied in or carried by a physical medium capable of detection by a machine or apparatus. This term includes recorded or transmitted

signals, and should be understood to include conveyance by any form of encoding, including pulse code modulation (PCM), but not limited to PCM.

[0071] While there has been shown, described, and pointed out fundamental novel features of the present invention as applied to preferred embodiments thereof, it will be understood that various omissions and substitutions and changes in the apparatus and method described, in the form and details of the devices disclosed, and in their operation, may be made by those skilled in the art without departing from the spirit of the present invention. It is expressly intended that all combinations of those elements that perform substantially the same function in substantially the same way to achieve the same results are within the scope of the invention. Substitutions of elements from one described embodiment to another are also fully intended and contemplated.

[0072] Each feature disclosed in the description and (where appropriate) the claims and drawings may be provided independently or in any appropriate combination. Features may, where appropriate be implemented in hardware, software, or a combination of the two. Connections may, where applicable, be implemented as wireless connections or wired, not necessarily direct or dedicated, connections. Reference numerals appearing in the claims are by way of illustration only and shall have no limiting effect on the scope of the claims. In one embodiment, an apparatus is at least partially implemented in hardware by using at least one silicon component.

Cited References

[0073]

[1] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting", IEEE Transactions on Audio, Speech and Language Processing, vol. 20, no. 3, pp. 922 - 932, 2012.

[2] Kai Siedenburg, Matthieu Kowalski and Monika Dörfler, "Audio Declipping with Social Sparsity" in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2014.

[3] Smaragdis, P., B. Raj, M. Shashanka. "Missing data imputation for time-frequency representations of audio signals", in the Journal of Signal Processing Systems. August 2010.

[4] A. Ozerov, C. Fevotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11), Prague, May 2011, pp. 257--260.

[5] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," Proc. IEEE International Conference on Consumer Electronics (ICCE-Berlin), Germany, Sept. 2014.

[6] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis", Neural Computation, vol. 21, no. 3, pp.793-830, Mar. 2009.

Claims

1. A method (30) for performing audio restauration, wherein missing coefficients of an input audio signal are recovered and a recovered audio signal is obtained, comprising steps of

   - initializing (31) a variance tensor $\boldsymbol{V}$ such that it is a low rank tensor that can be composed from component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$, or initializing said component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ to obtain the low rank variance tensor $\boldsymbol{V}$;
   - iteratively applying the following steps, until convergence of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$:

     i. computing (32) conditional expectations of source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$, known signal values $(\mathbf{x},\mathbf{y})$ of the input audio signal and time domain information on loss $(\boldsymbol{I_L})$ are input to the computing;
     ii. re-calculating (33) the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ and the variance tensor $\boldsymbol{V}$ using the estimated source power spectra $\boldsymbol{P}(f,n,j)$ and current values of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$;

   - upon convergence (34) of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}\_$, computing (35) a resulting variance tensor $\boldsymbol{V}'$, and computing (36) from the resulting variance tensor $\boldsymbol{V}'$, known signal values $(\mathbf{x},\mathbf{y})$ of the input audio signal and time domain information on loss $(\boldsymbol{I_L})$, an array of a posterior mean of Short Time Fourier Transform (STFT) samples (S) of the recovered audio signal; and
   - converting (37) coefficients of the array of the posterior mean of the STFT samples (S) to the time domain, wherein coefficients $(\tilde{\boldsymbol{s}}_1,\tilde{\boldsymbol{s}}_2, ..., \tilde{\boldsymbol{s}}_J)$ of the recovered audio signal are obtained.

2. The method according to claim 1, wherein in the step of computing (32) conditional expectations of the source power spectra of the input audio signal the estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained according to $\boldsymbol{P}(f,n,j) = E\{|\boldsymbol{S}(f,n,j)|^2|\boldsymbol{x},\boldsymbol{I_S},\boldsymbol{I_L},\boldsymbol{V}\}$, with $\boldsymbol{I_S}$ being time domain information on sources.

3. The method according to claim 2, wherein the time domain information on sources ($\boldsymbol{I_S}$) comprises at least one of: information about which sources are active or silent for a particular time instant, information about a number of how many components each source is composed in the low rank representation, and specific information on a harmonic structure of the sources.

4. The method according to one of the claims 1-3, wherein the time domain information on loss ($\boldsymbol{I_L}$) comprises at least one of: a clipping threshold, a sign of an unknown value in the input audio signal, an upper limit for the signal magnitude, and the quantized value of an unknown signal in the input audio signal.

5. The method according to one of the claims 1-4, wherein the variance tensor $\boldsymbol{V}$ is computed from matrices $\boldsymbol{H} \in \mathrm{R}_+^{N\times K}$, $\boldsymbol{W} \in \mathrm{R}_+^{F\times K}$, $\boldsymbol{Q} \in \mathrm{R}_+^{J\times K}$ of rank k according to $\boldsymbol{V}(f,n,j) = \sum_{k=1}^{K} \boldsymbol{H}(n,k)\boldsymbol{W}(f,k)\boldsymbol{Q}(j,k)$.

6. The method according to one of the claims 1-5, wherein the variance tensor $\boldsymbol{V}$ is initialized by random matrices $\boldsymbol{H} \in \mathrm{R}_+^{N\times K}$, $\boldsymbol{W} \in \mathrm{R}_+^{F\times K}$, $\boldsymbol{Q} \in \mathrm{R}_+^{J\times K}$, according to $\boldsymbol{V}(f,n,j) = \sum_{k=1}^{K} \boldsymbol{H}(n,k)\boldsymbol{W}(f,k)\boldsymbol{Q}(j,k)$.

7. The method according to one of the claims 1-6, wherein the variance tensor $\boldsymbol{V}$ is initialized by values derived from known samples of the input audio signal.

8. The method according to one of the claims 1-7, wherein the input audio signal is a mixture of multiple audio sources, further comprising steps of

    - receiving (38) side information comprising quantized random samples of the multiple audio signals; and
    - performing (39) source separation, wherein the multiple audio signals from said mixture of multiple audio sources are separately obtained.

9. The method according to one of the claims 1-8, wherein the STFT coefficients are windowed time domain samples ($\hat{S}$).

10. The method according to one of the claims 1-9, wherein the input audio signal contains quantization noise, wherein wrongly quantized coefficients take the position of the missing coefficients, wherein the quantization levels are used as further constraints in said time domain information on loss ($\boldsymbol{I_L}$), and wherein the recovered audio signal is a de-quantized audio signal.

11. An apparatus (40) for performing audio restauration, wherein missing coefficients of an input audio signal are recovered and a recovered audio signal is obtained, the apparatus comprising a processor (41) and a memory (42) storing instructions that, when executed on the processor, cause the apparatus to perform a method comprising

    - initializing a variance tensor $\boldsymbol{V}$ such that it is a low rank tensor that can be composed from component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$, or initializing said component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ to obtain the low rank variance tensor $\boldsymbol{V}$;
    - iteratively applying the following steps, until convergence of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$:

        i. computing (32) conditional expectations of source power spectra of the input audio signal, wherein estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained and wherein the variance tensor $\boldsymbol{V}$, known signal values ($\mathbf{x}$, $\mathbf{y}$) of the input audio signal and time domain information on loss ($\boldsymbol{I_L}$) are input to the computing;
        ii. re-calculating (33) the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$ and the variance tensor $\boldsymbol{V}$ using the estimated source power spectra $\boldsymbol{P}(f,n,j)$ and current values of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}$;

    - upon convergence of the component matrices $\boldsymbol{H},\boldsymbol{Q},\boldsymbol{W}\_$, computing a resulting variance tensor $\boldsymbol{V}'$, and computing from the resulting variance tensor $\boldsymbol{V}'$, known signal values ($\mathbf{x},\mathbf{y}$) of the input audio signal and time domain information on loss ($\boldsymbol{I_L}$), an array of a posterior mean of Short Time Fourier Transform (STFT) samples (S) of the recovered audio signal; and

- converting (37) coefficients of the array of the posterior mean of the STFT samples (S) to the time domain, wherein coefficients ($\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, ..., \tilde{\boldsymbol{s}}_J$) of the recovered audio signal are obtained.

12. The apparatus according to claim 11, wherein the estimated source power spectra $\boldsymbol{P}(f,n,j)$ are obtained according to $\boldsymbol{P}(f,n,j) = E\{\backslash S(f,n,j)\backslash 2\backslash \boldsymbol{x}, \boldsymbol{I}_S, \boldsymbol{I}_L, \boldsymbol{V}\}$ with $\boldsymbol{I}_S$ being time domain information on sources.

13. The apparatus according to one of the claims 11-12, wherein the time domain information on loss comprises at least one of: a clipping threshold, a sign of an unknown value in the input audio signal, an upper limit for the signal magnitude, and the quantized value of an unknown signal in the input audio signal.

14. The apparatus according to one of the claims 11-13, wherein the input audio signal is a mixture of multiple audio sources, the instructions when executed on the processor further cause the apparatus to

- receive (38) side information comprising quantized random samples of the multiple audio signals; and
- perform (39) source separation, wherein the multiple audio signals from said mixture of multiple audio sources are separately obtained.

15. The apparatus according to one of the claims 11-14, wherein the input audio signal contains quantization noise, wherein wrongly quantized coefficients take the position of the missing coefficients, wherein the quantization levels are used as further constraints in said time domain information on loss ($\boldsymbol{I}_L$), and wherein the recovered audio signal is a de-quantized audio signal.

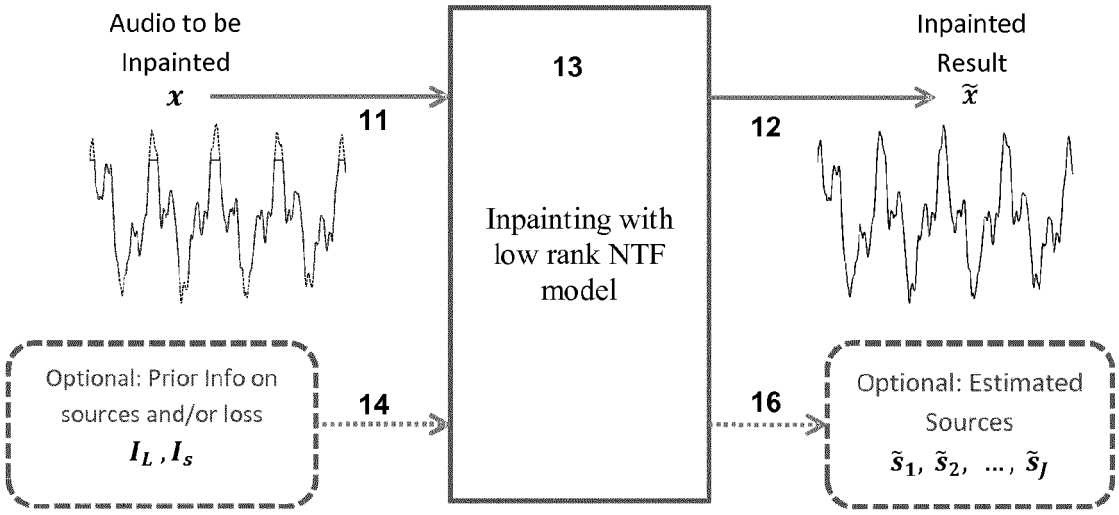
Audio to be Inpainted
x
11
13
Inpainting with low rank NTF model
12
Inpainted Result
$\tilde{x}$
Optional: Prior Info on sources and/or loss
I_L , I_s
14
16
Optional: Estimated Sources
$\tilde{s}_1$, $\tilde{s}_2$, ..., $\tilde{s}_J$

Fig.1

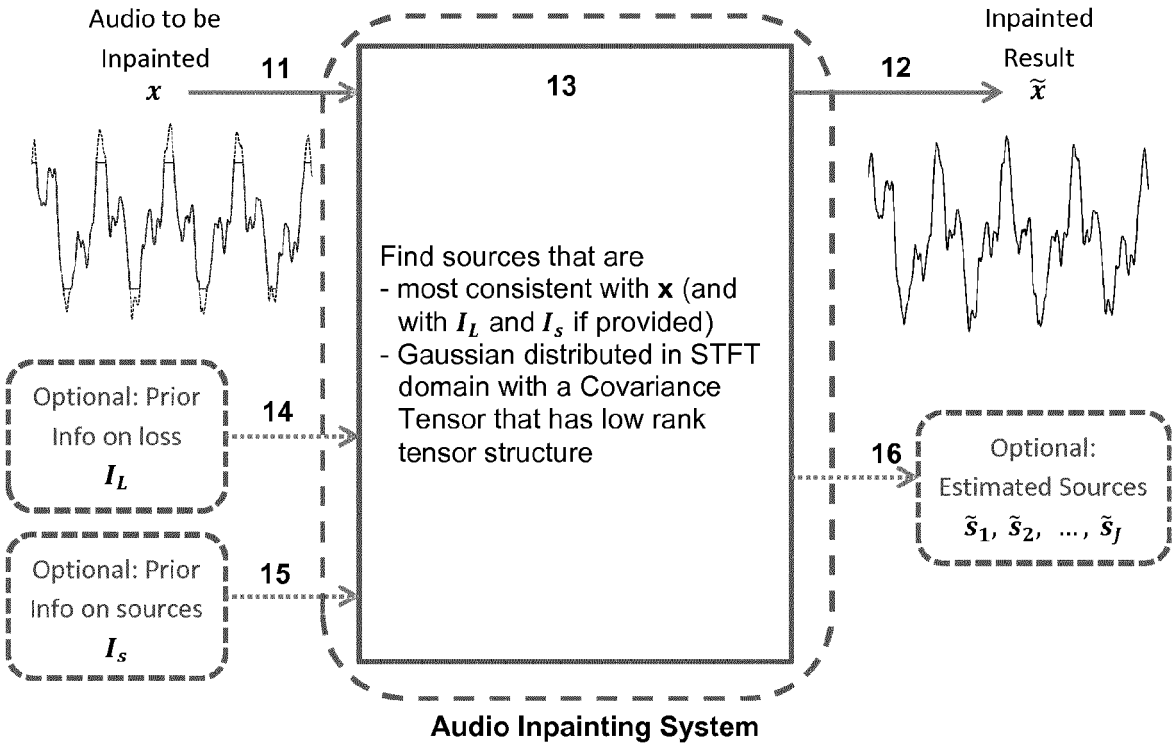
Audio to be Inpainted
x
11
13
12
Inpainted Result
$\tilde{x}$
Find sources that are
- most consistent with **x** (and with I_L and I_s if provided)
- Gaussian distributed in STFT domain with a Covariance Tensor that has low rank tensor structure
Optional: Prior Info on loss
I_L
14
Optional: Prior Info on sources
I_s
15
16
Optional: Estimated Sources
$\tilde{s}_1$, $\tilde{s}_2$, ..., $\tilde{s}_J$
Audio Inpainting System

Fig.2

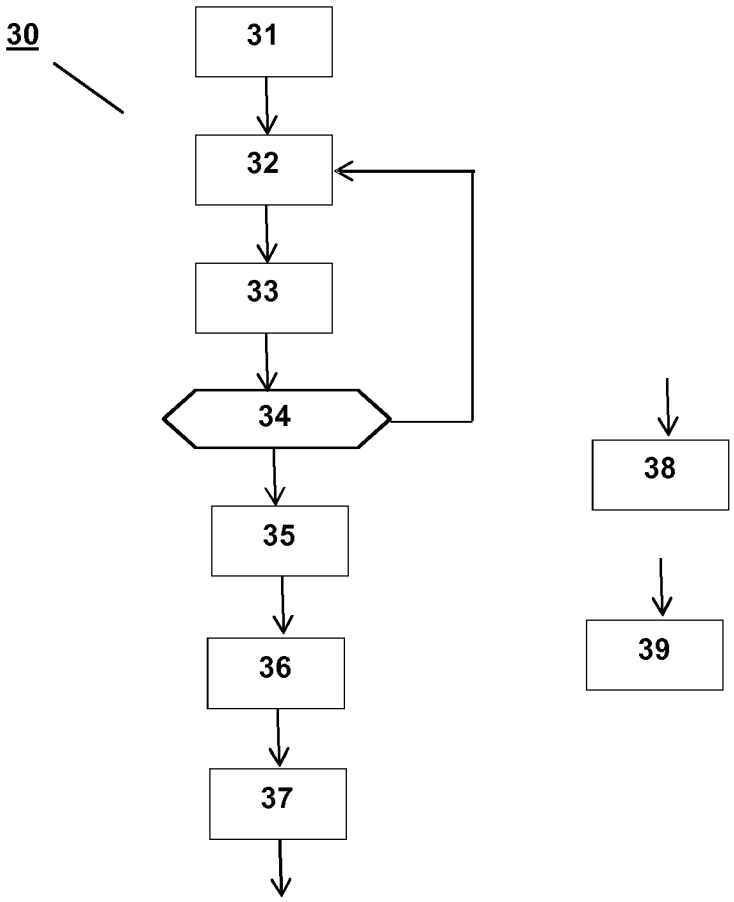

Fig.3

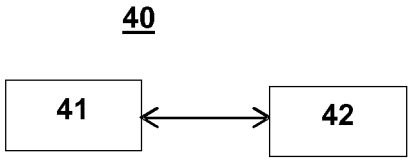

Fig.4

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 15 30 6212

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	Cagdas Bilen, Alexey Ozerov, Patrick Perez: "Joint Audio Inpainting and Source Separation", , 5 June 2015 (2015-06-05), XP002754817, Retrieved from the Internet: URL:https://hal.inria.fr/hal-01160438/document [retrieved on 2016-02-26] * the whole document *	1-15	INV. G10L19/005 G10L21/0208
X	Cagdas Bilen ET AL: "Audio Inpainting, Source Separation, Audio Compression. All with a Unified Framework Based on NTF Model", MissData 2015, 18 June 2015 (2015-06-18), pages 1-2, XP055216560, Retrieved from the Internet: URL:https://hal.inria.fr/hal-01171843/document [retrieved on 2015-09-28] * the whole document *	1-15	TECHNICAL FIELDS SEARCHED (IPC) G10L
A	Cagdas Bilen ET AL: "Audio Inpainting, Source Separation, Audio Compression. All with a Unified Framework Based on NTF Model", missDATA 2015, 18 June 2015 (2015-06-18), XP055216546, Rennes, France Retrieved from the Internet: URL:http://arxiv.org/abs/1502.06919 [retrieved on 2015-09-28] * pages 32,35,36 *	1-15	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	26 February 2016	Van Doremalen, Hans

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

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

- **A. ADLER ; V. EMIYA ; M. JAFARI ; M. ELAD ; R. GRIBONVAL ; M. D. PLUMBLEY.** Audio inpainting. *IEEE Transactions on Audio, Speech and Language Processing,* 2012, vol. 20 (3), 922-932 **[0073]**
- **KAI SIEDENBURG ; MATTHIEU KOWALSKI ; MONIKA DÖRFLER.** Audio Declipping with Social Sparsity. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* 2014 **[0073]**
- **SMARAGDIS, P. ; B. RAJ ; M. SHASHANKA.** Missing data imputation for time-frequency representations of audio signals. *the Journal of Signal Processing Systems,* August 2010 **[0073]**
- **A. OZEROV ; C. FEVOTTE ; R. BLOUET ; J.-L. DURRIEU.** Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11),* May 2011, 257-260 **[0073]**
- **N. Q. K. DUONG ; A. OZEROV ; L. CHEVALLIER.** Temporal annotation-based audio source separation using weighted nonnegative matrix factorization. *Proc. IEEE International Conference on Consumer Electronics (ICCE-Berlin),* September 2014 **[0073]**
- **C. FEVOTTE ; N. BERTIN ; J.-L. DURRIEU.** Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation,* March 2009, vol. 21 (3), 793-830 **[0073]**