US 20140114650A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2014/0114650 A1

(43) Pub. Date: Apr. 24, 2014

Hershey et al.

(54) METHOD FOR TRANSFORMING NON-STATIONARY SIGNALS USING A DYNAMIC MODEL

(71) Applicant: **MITSUBISHI ELECTRIC RESEARCH LABS, INC.**, Cambridge, MA (US)

(72) Inventors: **John R. Hershey**, Winchester, MA (US); **Cedric Fevotte**, Paris (FR); **Jonathan Le Roux**, Somerville, MA (US)

(73) Assignee: **Mitsubishi Electric Research Labs, Inc.**, Cambridge, MA (US)

(21) Appl. No.: **13/657,077**

(22) Filed: **Oct. 22, 2012**

Publication Classification

(51) **Int. Cl.**
G10L 19/02 (2006.01)

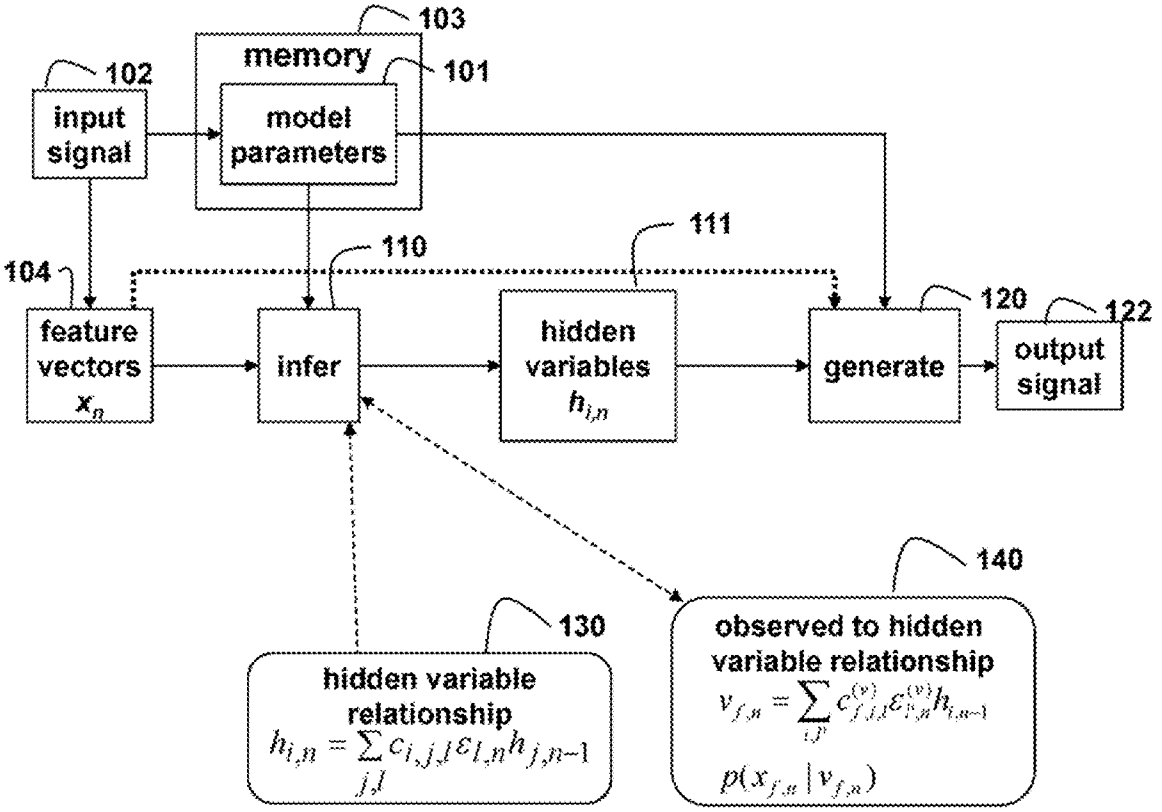
(52) **U.S. Cl.**
USPC **704/203**; 704/E19.001

(57) **ABSTRACT**

An input signal, in the form of a sequence of feature vectors, is transformed to an output signal by first storing parameters of a model of the input signal in a memory. Using the vectors and the parameters, a sequence of vectors of hidden variables is inferred. There is at least one vector h_n of hidden variables $h_{i,n}$ for each feature vector x_n, and each hidden variable is nonnegative. The output signal is generated using the feature vectors, the vectors of hidden variables, and the parameters. Each feature vector x_n is dependent on at least one of the hidden variables $h_{i,n}$ for the same n. The hidden variables are related according to

$$h_{i,n} = \sum_{j,l} c_{i,j,l} \epsilon_{l,n} h_{j,n-1},$$

where j and l are summation indices. The parameters include non-negative weights $c_{i,j,l}$, and $\epsilon_{l,n}$ are independent non-negative random variables.



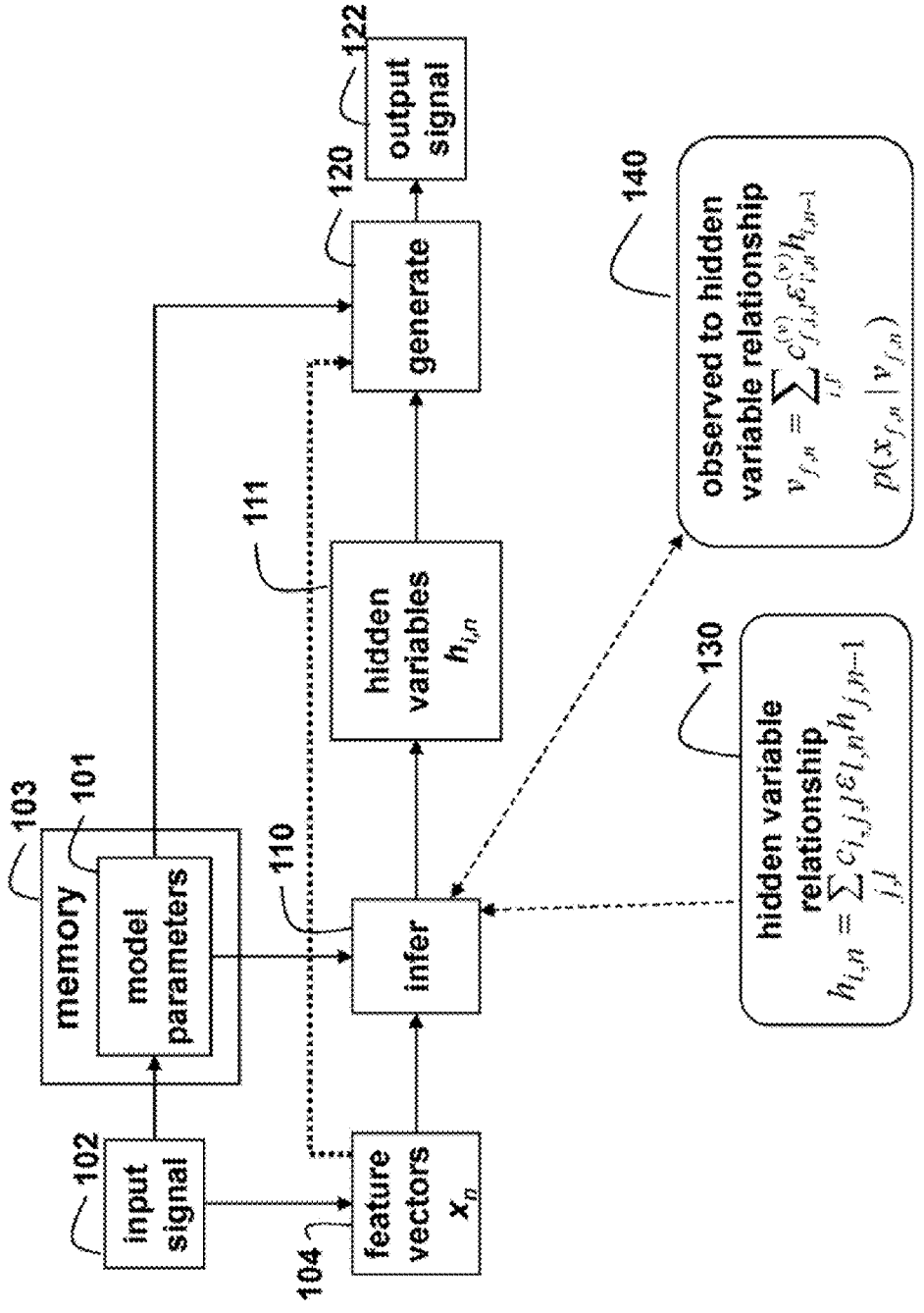

input signal
102
feature vectors x_n
104
memory
103
model parameters
101
infer
110
hidden variables $h_{i,n}$
111
generate
120
output signal
122
hidden variable relationship
$h_{i,n} = \sum_{j,l} c_{i,j,l} \varepsilon_{l,n} h_{j,n-1}$
130
observed to hidden variable relationship
$v_{f,n} = \sum_{i,l} c_{f,i,l}^{(v)} \varepsilon_{i,n}^{(v)} h_{i,n-1}$
$p(x_{f,n} | v_{f,n})$
140

Fig. 1

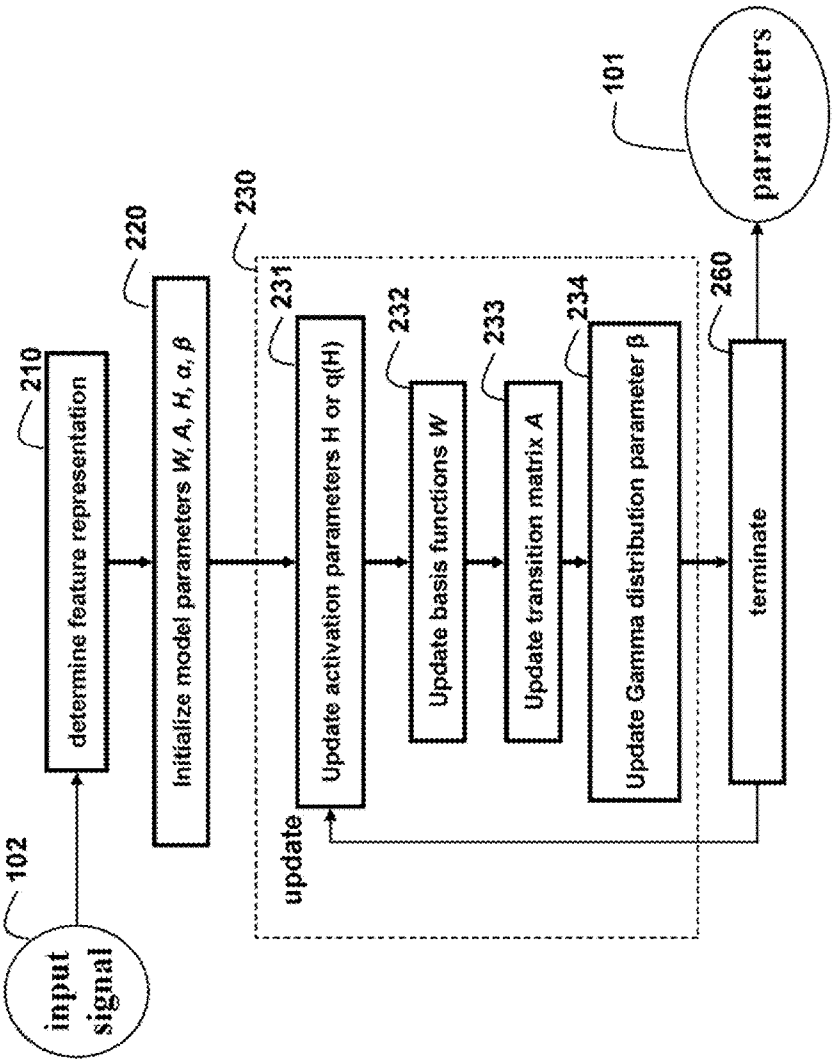

input signal
102
determine feature representation
210
Initialize model parameters W, A, H, α, β
220
230
update
Update activation parameters H or q(H)
231
Update basis functions W
232
Update transition matrix A
233
Update Gamma distribution parameter β
234
terminate
260
parameters
101

Fig. 2

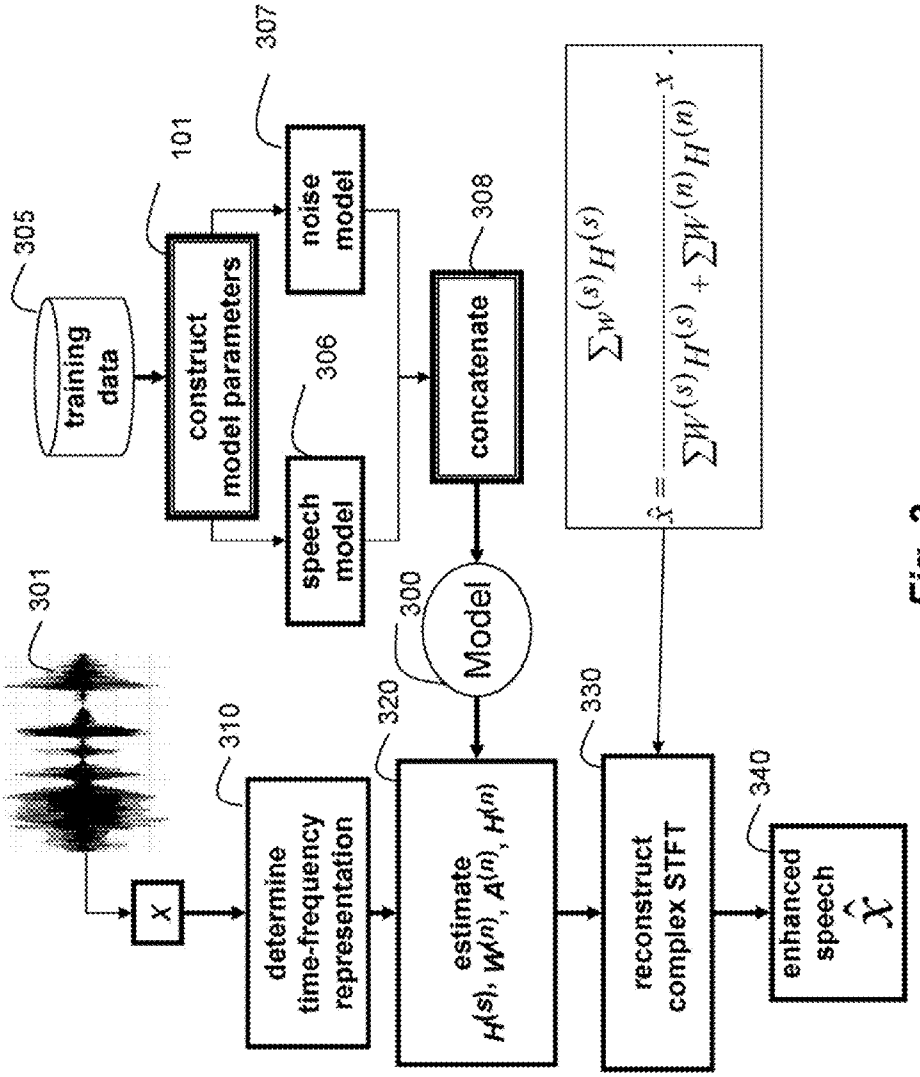

training data
305
construct model parameters
101
noise model
307
speech model
306
concatenate
308
Model
300
301
x
determine time-frequency representation
310
estimate $H^{(s)}$, $W^{(n)}$, $A^{(n)}$, $H^{(n)}$
320
reconstruct complex STFT
330
enhanced speech $\hat{x}$
340
$\hat{x} = \frac{\sum_w^{(s)} H^{(s)}}{\sum W^{(s)} H^{(s)} + \sum W^{(n)} H^{(n)}} x$

Fig. 3

METHOD FOR TRANSFORMING NON-STATIONARY SIGNALS USING A DYNAMIC MODEL

FIELD OF THE INVENTION

[0001] This invention relates generally to signal processing, and more particularly to transforming an input signal to an output signal using a dynamic model, where the signal is an audio (speech) signal.

BACKGROUND OF THE INVENTION

[0002] A common framework for modeling dynamics in non-stationary signals is a hidden Markov model (HMM) with temporal dynamics. The HMM is the de facto standard for speech recognition. A discrete-time HMM models a sequence of N observed (acquired) random variables

$$\{x_n\} \stackrel{def}{=} x_{1:N} \stackrel{def}{=} \{x_1, x_2, \ldots, x_N\},$$

i.e., signal samples, by conditioning probability distributions on the sequence of unobserved random state variables $\{h_n\}$. Two constraints are typically defined on the HMM.

[0003] First, the state variables have first-order Markov dynamics. This means that $p(h_n|h_{1:n-1})=p(h_n|h_{n-1})$, where the $p(h_n|h_{n-1})$ are known as transition probabilities. The transition probabilities are usually constrained to be time-invariant.

[0004] Second, each sample x_n, given the corresponding state h_n, is independent of all other hidden states $h_{n'}$, $n'\neq n$, so that $p(x_n|h_{1:N})=p(x_n|h_n)$, where the $p(x_n|h_n)$ are known as observation probabilities. In many speech applications, the states h_n are discrete, and observations x_n are F-dimensional vector-valued continuous acoustic features,

$$x_n \stackrel{def}{=} \{x_{f,(n)}\} \stackrel{def}{=} \{x_{1n}, x_{2n}, \ldots, x_{Fn}\},$$

where the parentheses indicate that n is not iterated. Typical frequency features are short-time log power spectra, where f indicates a frequency bin.

[0005] Defining initial probabilities

$$p(h_1 \mid h_0) \stackrel{def}{=} p(h_1),$$

the joint distribution of the random variables of the HMM is

$$p(\{x_n\}, \{h_n\}) = \prod_{n=1}^{N} p(x_n \mid h_n) p(h_n \mid h_{n-1}). \quad (1)$$

[0006] Linear Dynamical Systems

[0007] A related model is a linear dynamical system used in Kalman filters. The linear dynamical system is characterized by states and observations that are continuous, vector-valued, and jointly Gaussian distributed

$$h_n = Ah_{n-1} + \epsilon_n, \quad (2)$$

$$v_n = Bh_n + v_n, \quad (3)$$

where $h_n \in R^K$ (or $h_n \in C^K$) is the state at lime n, K the dimension of the state space, A is a state transition matrix, ϵ_n is additive Gaussian transition noise, $v_n \in R^F$ (or $v_n \in C^F$) is the observation at time n, F is the dimension of the observation (or feature) space, B is an observation matrix, v_n is additive Gaussian noise, and R is real.

[0008] Non-Negative Matrix Factorization

[0009] In the context of audio signal processing, the signal is typically processed using a sliding window and a feature vector representation that is often a magnitude or power spectrum of the audio signal. The features are nonnegative. In order to discover repeating patterns in the signal in an unsupervised way, nonnegative matrix factorization (NMF) is extensively used.

[0010] For a nonnegative matrix V of dimensions F×N, a rank-reduced approximation is

$$V \approx WH,$$

where W and H are nonnegative matrices of dimensions F×K and K×N, respectively. The approximation is typically obtained from a minimization

$$\min_{W,H\geq 0} D(V \mid WH) = \sum_{fn} d(v_{fn} \mid [WH]_{fn}),$$

where d(x|y) is a positive function scalar cost function with a unique minimum at x=y.

[0011] Itakura-Saito Nonnegative Matrix Factorization (IS-NMF)

[0012] For the audio signal, where the matrix V is the power spectrogram of a complex-valued short-time Fourier transform (STFT) matrix X, conventional methods have used the Itakura-Saito distance, which measures the difference between the actual and approximated spectrum, as the cost function, because the cost function implies a latent model of superimposed zero-mean. Gaussian components that is relevant for audio signals. More precisely, let x_{fn} be the complex-valued STFT coefficient at frame n and frequency f, and

$$x_{fn} = \sum_k c_{fkn},$$

Where

[0013]

$$c_{fkn} : N_c(0, w_{fk} h_{kn}).$$

[0014] Then,

$$-\log p(X \mid W, H) = \sum_{fn} \frac{v_{fn}}{\sum_k w_{fk} h_{kn}} + \log \sum_k w_{fk} h_{kn} \quad (4)$$

$$= D_{IS}(|X|^2 \mid WH) + cst, \quad (5)$$

where

$$v_{fn} = |x_{fn}^2|.$$

[0015] The model can also be expressed as

$$x_{fn}: N_c\left(0, \sum_k w_{fk} h_{kn}\right).$$

[0016] It is equivalent to assume that $|x|_{fn}^2$ is exponentially distributed with parameter $\Sigma_k w_{fk} h_{kn}$ and uniform phase

$$|x|_{fn}^2: \text{Exponential}\left(\sum_k w_{fk} h_{kn}\right), \quad (6)$$

$$\angle x_{fn}: \text{Uniform}(-\pi, +\pi). \quad (7)$$

[0017] Smooth IS-NMF

[0018] In smooth variants of IS-NMF, an inverse-gamma or gamma random walk is assumed for independent rows of H. More precisely, the following model has been considered:

$$h_{kn}=h_{k(n-1)}\circ\epsilon_{kn},$$

where ϵ_{kn} is nonnegative multiplicative innovation random variable with mode 1, such as

$$\epsilon_{kn}: G(\alpha,\alpha-1), \text{ or}$$

$$\epsilon_{kn}: IG(\alpha,\alpha+1),$$

where by convention gamma and inverse-gamma are

$$G(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp - \beta x, \quad (8)$$

and

$$IG(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp - \frac{\beta}{x}. \quad (9)$$

[0019] Models Combining HMMs and NMF

[0020] If HMMs and NMF are combined, then the restriction that only one discrete state can be active at a time is inherited from the HMMs. This means that multiple model are required for multiple source, leading to potential issues to computational tractability.

[0021] U.S. Pat. No. 7,047,047 describes denoising a speech signal using an estimate of a noise-reduced feature vector and a model of an acoustic environment. The model is based on a non-linear function that describes a relationship between the input feature vector, a clean feature vector, a noise feature vector and a phase relationship indicative of mixing of the clean feature vector and the noise feature vector.

[0022] U.S. Pat. No. 8,015,003 describes denoising a mixed signal, e.g., speech and noise, using a NMF constrained by a denoising model. The denoising model includes training basis matrices of a training acoustic signal and a training noise signal, and statistics of weights of the training basis matrices. A product of the weights of the basis matrix of the acoustic signal and the training basis matrices of the training acoustic signal and the training noise signal is used to reconstruct the acoustic signal.

[0023] In general, the prior art methods that focus on slow-changing noise, are inadequate for fast-changing nonstationary noise, such as experienced by using a mobile telephone in a noisy environment.

[0024] Although HMMs can handle speech dynamics, HMMs often lead to combinatorial issues due to the discrete state space, which is computationally complex, especially for mixed signals from several sources. In conventional HMM approaches it is also not straightforward to handle gain adaptation.

[0025] NMF solves both the computational and gain adaptation issues. However, NMF does not handle dynamic signals. Smooth IS-NMF attempts to handle dynamics. However, the independence assumption of the rows of H is not realistic, as the activation of a spectral pattern at frame n is likely to be correlated with the activation of other patterns at a previous frame n−1.

[0026] It is an object of the invention to solve inherent problems associated with signal and data processing using HMMs and NMF frameworks.

SUMMARY OF THE INVENTION

[0027] It is an object of the invention to transform an input signal to an output signal when the input signal is a non-stationary signal, and more specifically a mixture of signals. Therefore, the embodiments of the invention provide a non-negative linear dynamical system model for processing the input signal, particularly a speech signal that is mixed with noise. In the context of speech separation and speech denoising, our model adapts to signal dynamics on-line, and achieves better performance than conventional methods.

[0028] Conventional models for signal dynamics frequently use hidden Markov models (HMMs) or non-negative matrix factorization (NMF).

[0029] HMMs lead to combinatorial problems due to the discrete state space, are computationally complex, especially for mixed signals from several sources. In conventional HMM approaches it is also not straightforward to handle gain adaptation.

[0030] NMF solves both the computational complexity and gain adaptation problems. However, NMF does not take advantage of past observations of a signal to model future observations of that signal. For signals with predictable dynamics, this is likely to be suboptimal.

[0031] Our model has advantages of both the HMMs and the NMF. The model is characterized by a continuous non-negative state space. Gain adaptation is automatically handled during inference. The complexity of the inference is linear in the number of signal sources, and dynamics are modeled via a linear transition matrix.

[0032] Specifically the input signal, in the form of a sequence of feature vectors, is transformed to the output signal by first storing parameters of a model of the input signal in a memory.

[0033] Using the vectors and the parameters, a sequence of vectors of hidden variables is inferred. There is at least one vector h_n of hidden variables $h_{i,n}$ for each feature vector x_n, and each hidden variable is nonnegative.

[0034] The output signal is generated using the feature vectors, the vectors of hidden variables, and the parameters. Each feature vector X_n is dependent on at least one of the hidden variables $h_{i,n}$ or the same n. The hidden variables are related according to

$$h_{i,n} = \sum_{j,l} c_{i,j,l} \varepsilon_{l,n} h_{j,n-1},$$

where j and l are summation indices. The parameters include non-negative weights $c_{i,j,l}$, and $\epsilon_{l,n}$ are independent non-negative random variables.

BRIEF DESCRIPTION OF THE DRAWINGS

[0035] FIG. **1** is a flow diagram for transforming an input signal to an output signal;

[0036] FIG. **2** is a flow diagram of a method for determining parameters of a dynamic model according to embodiment of the invention; and

[0037] FIG. **3** is a flow diagram of a method for enhancing a speech signal using the dynamic model according to embodiments of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Introduction

[0038] The embodiments of our provide a model for transforming and processing dynamic (non-stationary) signal and data that has advantages of HMMs and NMF based models.

[0039] The model is characterized by a continuous non-negative state space. Gain adaptation is automatically handled on-line during inference. Dynamics of the signal are modeled using a linear transition matrix A. The model is a non-negative linear dynamical system with multiplicative non-negative innovation random variables ϵ_n. The signal can be a non-stationary linear signal, such as an audio or speech signal, or a multi-dimensional signal. The signal can be expressed in the digital domain as data. The innovation random variable is described in greater detail below.

[0040] The embodiments also provide applications for using the model. Specifically, the model can be used to process an audio signal acquired from several, sources, e.g., the signal is a mixture of speech and noise (or other acoustic interference) and the model is used to enhance the signal by, e.g., reducing noise. When we say "mixed," we mean that the speech and noise are acquired by a single sensor (microphone).

[0041] However, it is understood that the model can also be used for other non-stationary signals and data that have characteristics that vary over time, such as economic or financial data, network data and signals, or signals, medical signals, or other signals acquired from natural phenomena. The parameters include non-negative weights $c_{i,j,l}$, and $\epsilon_{l,n}$ are independent non-negative random variables, the distributions of which also have parameters. The indices i,j,l, and n are described below.

[0042] General Method

[0043] As shown in FIG. **1**, parameters **101** of a model of an input signal **102** are stored in a memory **103**.

[0044] The input signal is received as a feature vectors x_n **104** of salient characteristics of the signal. The features are of course application and signal specific. For example, if the signal is an audio signal, the features can be log power spectra. It is understood that the different type of features that can be used is essentially unlimited for many types of different signals and data that can be processed by the method according to the invention.

[0045] The method infers **110** a sequence of vectors of hidden variables **111**. The inference is based on the feature vector **104**, the parameters, a hidden variable relationship **130**, and a relationship **140** of observations to hidden variables. There is at least one vector h_n of hidden variables $h_{i,n}$ for each feature vector x_n. Each hidden variable is nonnegative.

[0046] An output signal **122** corresponding to the input signal is generated **120** to form the feature vectors, the vectors of hidden variables, and the parameters.

[0047] General Method Details

[0048] In our method, each feature vector x_n is dependent on at least one of the hidden variables $h_{i,n}$ for the same n. The hidden variables are related according to a hidden variable relationship

$$h_{i,n} = \sum_{j,l} c_{i,j,l}\epsilon_{l,n}h_{j,n-1}$$

130, where j and l are summation indices. The stored parameters include non-negative weights $c_{i,j,l}$, and $\epsilon_{l,n}$ are independent non-negative random variables. This formulation enables the model to represent statistical dependency over time in a structured way, so that the hidden variables for the current frame, n, are dependent on those of the previous frame, n−1 with a distribution that is determined by the combination of $c_{i,j,l}$, and the parameters of the distribution of the weights $\epsilon_{l,n}$. The weight $\epsilon_{l,n}$, for example, may be Gamma random variables with shape parameter α and inverse scale parameter β.

[0049] In one embodiment, $c_{i,j,l}=\delta(i,l)a_{i,j}$, where $a_{i,j}$ are non-negative scalars, so that

$$h_{i,n} = \left(\sum_j a_{i,j}h_{j,n-1}\right)\epsilon_{i,n},$$

where δ is a Kronecker delta. In this case, if the weights $\epsilon_{l,n}$, are Gamma random variables with shape parameter α and inverse scale parameter β, then the conditional distribution of $h_{i,n}$ given $\{h_{j,n-1}\}_{j=1}^{K}$, where K is a number of elements in the hidden states vector, is

$$p(h_{i,n} \mid h_{j,n-1}) = \text{Gamma}\left(h_{i,n} \mid \alpha, \frac{\beta}{\sum_j a_{i,j}h_{j,n-1}}\right),$$

where

$$\text{Gamma}(x \mid a, b) = \frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$$

is the gamma distribution for random variable x with shape a, inverse scale b, and

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt$$

is the gamma function. This embodiment is designed to conform to the simplicity of the basic structure of a conventional

linear dynamical system, but differs from prior art by the non-negative structure of the model, and the multiplicative innovation random variables.

[0050] In another embodiment, $c_{i,j,l}=\alpha(m(i,j),l),a_{i,j}$, where $a_{i,j}$ are non-negative scalars, δ is the Kronecker delta, $\delta(a,b)=\begin{cases}1: if\ a=b\\0: otherwise\end{cases}$, and m(i,j) is a one-to-one mapping from each combination of i and j to an index corresponding to l, (e.g., $m(i,j)=(i-1)K+j$, where K is a number of elements in the hidden variable h_n) so that

$$h_{i,n}=\sum_j a_{i,j}\varepsilon_m(i,\ j),\ n^h j,\ n-1.$$

This embodiment enables flexibility in modeling the signal, because each transition, can be inferred independently.

[0051] Another embodiment that is important to modeling multiple sources comprises partitioning hidden variables $h_{i,n}$ into S groups, where each group corresponds to one independent source in a mixture. Likewise, the non-negative random variables $\epsilon_{l,n}$ are partitioned according to the same S groups. This can be accomplished by a special case of the parameters $c_{i,j,l}$ where $c_{i,j,l}=0$ when $h_{i,n}$, and $h_{j,n}$ are not in the same group or when $h_{i,n}$ and $\epsilon_{l,n}$ are not associated with the same group. When the hidden variables are ordered accordingly, this gives $c_{i,j,l}$ block structure, where each block corresponds to the model for one of the signal sources.

[0052] In our embodiments, the hidden variables are related **140** to feature variables via a non-negative feature $v_{f,n}$, of the signal indexed by feature f and frame n. An observation model is based on

$$v_{f,n}=\sum_j c^{(v)}_{f,i,l}h_{i,n}\varepsilon^{(v)}_{l,n},$$

where $c_{f,i,l}^{(v)}$ is a non-negative scalar, and $\epsilon_{l,n}^{(v)}$ are independent non-negative random variables, and j, and l are indices of different components.

[0053] In a more constrained embodiment $c_{f,i,l}^{(v)}=\delta(i,l)w_{f,i}$, where $w_{f,i}$ are non-negative scalars, where δ is the Kronecker delta, and $\epsilon_{f,n}^{(v)}$ are the Gamma distributed random variables, so that the observation model based, at least in part, on

$$p(v_{f,n}\mid h_n)=\text{Gamma}\left(v_{f,n}\mid \alpha^{(v)},\ \beta^{(v)}\Big/\sum_i w_{f,i}h_{i,n}\right),$$

where $v_{f,n}$ is non-negative feature of the signal at frame n and frequency f, $\alpha^{(v)}$ and $\beta^{(v)}$ are positive scalars, and $w_{f,i}$ are non-negative scalars.

[0054] In applications where the features $x_{f,n}$ are complex spectrogram values of the input signal, a frame n and frequency f, the observation model can use $v_{f,n}=|x_{f,n}|^2$, which is the power in frame n, and frequency f. Thus, an observation model can be formed based on

$$x_{f,n}=(e^{\theta_{f,n}\sqrt{-1}})\sqrt{v_{f,n}},$$

where $\sqrt{-1}$ is the unit imaginary number, and $\theta_{f,n}=\angle x_{f,n}$ is a phase for a frame n and frequency f.

[0055] In another embodiment, we select the parameter $\alpha^{(v)}=1$, so that the gamma distribution reduces to an exponential distribution as a special case. In this case, if the phases $\theta_{f,n}$ are distributed uniformly, then we obtain the observation model

$$p(x_{f,n}\mid h_n)=N_c\left(0,\ \sum_i w_{f,i}h_{i,n}\right),$$

where N_C is a complex Gaussian distribution. This observation model corresponds to the Itakura-Saito nonnegative matrix factorization described above, and is combined in our embodiments with the non-negative dynamical system model.

[0056] Another embodiment uses an observation model for $v_{f,n}$ based on a cascade of transformations of the same type:

$$u_{i',n}=\sum_i c^{(u)}_{i',i,l}h_{i,n}\varepsilon^{(u)}_{l',n},$$

and

$$v_{f,n}=\sum_{i'} c^{(v)}_{f,i',l}u_{i',n}\varepsilon^{(v)}_{l'',n},$$

where $c_{i',i,l'}^{(u)}$ and $c_{f,i',l''}^{(v)}$ are non-negative scalars, and $\epsilon_{l',n}^{(u)}$ and $\epsilon_{l'',n}^{(v)}$ are independent non-negative random variables, and i, i', l', l'' are indices.

[0057] The method for inferring the hidden variables depends on the model parameterization for each embodiment.

[0058] Model Parameters

[0059] As shown in FIG. **2**, from the input signal **102**, we obtain the model parameters **101** as follows. The input signal can be considered a training signal, although it should be understood that the method can be adaptive to the signal, and "learn" the parameters on-line. The input signal can also be in the form of a digital signal or data.

[0060] For example, the training signal is a speech signal, or a mixed signal from multiple acoustic sources, perhaps including non-stationary noise, or other acoustic interference. The signal is processed as frames of signal samples. The sampling rate and number of samples in each frame is application specific. It is noted that the updating **230** described below for processing the current frame n is dependent on a previous frame n−1. For each frame we determine **210** a feature vector x_n representation. For an audio input signal, frequency features such as log power spectra could be used.

[0061] Parameters of the model are initialized **220**. The parameters can include basis functions W, a transition matrix A, activation matrix H, and a fixed shape parameter ca and an inverse scale parameter β of a continuous gamma distribution parameter, and various combinations of these parameters depending on the particular application. For example in some applications, updating H and β are optional. In a variational Bayes (VB) method, H is not used. Instead an estimate of the posterior distribution of H is used and updated. If a maximum a-posteriori (MAP) estimation, then updating β is optional.

[0062] During each iteration of the method, the activation matrix, the basis function, the transition matrix, and the

gamma parameter are updated **231-134**. It should again be noted that the set of parameters to be updated is also application specific.

[0063] A termination condition **260**, e.g., convergence or a maximum number of iterations, is tested after the updating **230**. If true, store the parameters in a memory, otherwise if false, repeat at step **230**.

[0064] The above steps of the general method and the parameter determination can be performed in a processor connected to a memory and input/output interfaces as know. Specialized microprocessors, and the like can also be used. It is understood that the signals processed by the method, e.g., speech or financial data, can be extremely complex. The method transforms the input signal into features which can be stored in the memory. The method also stores the model parameters and inferred hidden variables in the memory.

[0065] Model Parameters Details

[0066] For simplicity of this description, we limit the notation to the embodiment where $c_{f,i,l}^{(v)}=\delta(i,l)w_{f,i}$, the $w_{f,i}$ are non-negative scalars δ is a Kronecker delta, and $\epsilon_{f,n}^{(v)}$ are gamma distributed random variables, with parameter $\alpha^{(v)}=1$, and phases $\theta_{f,n}$ are distributed uniformly. In this case, our model is

$$x_{fn} : N_c\left(0, \sum_k w_{fk} h_{kn}\right), \tag{10}$$

$$h_n = (Ah_{n-1}) \circ \varepsilon_n, \tag{11}$$

where x_{fn} is the complex-valued STFT coefficient at frame n and frequency f, N_C is the complex Gaussian distribution, w_{fk} is the value of the k^{th} basis function for the power spectrum at frequency f, h_n and h_{n-1} are the n^{th} and the $(n-1)^{th}$ columns of the activation matrix H, respectively, A is the nonnegative K×K transition matrix that models the correlations between the different patterns in successive frames n−1 and n, ϵ_n is a nonnegative innovation random variable, e.g., a vector of dimension K, and ∘ denotes entry-wise multiplication. The smooth IS-NMF can be obtained as a particular case of our model by setting $A=I_K$, where I_K is the K×K identity matrix.

ADVANTAGES

[0067] A distinctive and advantageous property of our model is that more than one state dimension can be non-zero at a given time. This means that a signal simultaneously acquired from multiple sources by a single sensor can analyzed using a single model, unlike the prior art HMM which requires multiple models.

[0068] Gamma Model of Innovations

[0069] We use an independent gamma distribution for the innovation ϵ_{kn}, namely

$$p(\epsilon_{in}|\alpha,\beta)=G(\alpha_i,\beta_i).$$

[0070] It follows that h_n is conditionally gamma distributed, such that

$$p(h_n \mid Ah_{n-1}) = \prod_i G(h_{in} \mid \alpha_i, \beta_i / [Ah_{n-1}]_i),$$

and in particular

$$E(h_{in} \mid Ah_{n-1}) = \frac{\alpha_i}{\beta_i} \sum_j a_{ij} h_{j(n-1)}. \tag{12}$$

[0071] For h_1, we use an independent scale-invariant non-informative Jeffreys prior i.e.,

$$p(h_1) = \prod_k p(h_{k1}).$$

In Bayesian probability, the Jeffreys prior is a non-informative (objective) prior distribution on a parameter space that is proportional to the square root of the determinant of Fisher information.

[0072] MAP Inference in the Gamma Innovation Model

[0073] The maximum a-posteriori (MAP) objective function is

$$C(W, H, A, \beta) = \sum_{fn} \left(\frac{v_{fn}}{\sum_k w_{fk} h_{kn}} + \log \sum_k w_{fk} h_{kn} \right) +$$

$$\sum_{i=1}^{K} \sum_{n=2}^{N} \left(\alpha_i \log \sum_j a_{ij} h_{j(n-1)} + \beta_i \frac{h_{in}}{\sum_j a_{ij} h_{j(n-1)}} + (1-\alpha_i) \log h_{in} \right) +$$

$$(N-1) \sum_i (\log\Gamma(\alpha_i) - \alpha_i \log\beta_i) - \sum_i \log p(h_{i1})$$

[0074] Scales

[0075] Scale-Ambiguity Between A and β

[0076] A K×K nonnegative diagonal matrix with coefficients λ_i on its diagonal is Λ, thus,

$$C(W,H,\Lambda A,\Lambda\beta)=C(W,H,A,\beta),$$

which has a scale-ambiguity between A and β. When both A and β are estimated, the scale-ambiguity can be corrected in a number of ways, for example by fixing β to arbitrary values or by normalizing the rows of A at every iteration **230** and rescaling β accordingly. For example, we can normalize the rows of the transition matrix A such that the rows sum to 1, or so that the maximum coefficient in every row is 1. In some embodiments, $\beta_i=\alpha_i$, i.e., the model expectation of the innovation random variable is 1.

[0077] Ill-Posedness of MAP

[0078] The scales of W and H are related by

$$C(W\Lambda^{-1}, \Lambda H, A, \beta) = C(W, H, \Lambda^{-1} A \Lambda, \beta) + N \sum_i \log\lambda_i,$$

where λ_i is the i-th element of the diagonal of Λ.

[0079] Without further constraints, the minimization of the MAP objective leads to a degenerate solution such that $\|W\|\to\infty$ and $\|H\|\to 0$. If we assume that all the diagonal elements of Λ are equal, such that $\Lambda=\lambda I_K$, then

$$C(W\Lambda^{-1},\Lambda H,A)=C(W,H,A)+KN\log\lambda.$$

[0080] The MAP objective can be made arbitrarily small by decreasing the value of λ. Hence, the norm of W is controlled during optimization. This can be achieved by hard or soft

constraints. The hard constraint is a regular constraint that must be satisfied, and the soft constraint is a cost functions expressing a preference.

[0081] Hard Constraint

[0082] We solve

$$\min C(W,H,A) s.t\ W\geq 0, H\geq 0, \|w_k\|_{l1}=1$$

using the change of variable $\overline{W}=W\Lambda^{-1}$, $\overline{H}=\Lambda H$ with $\Lambda=\mathrm{diag}[\lambda_1, \ldots, \lambda_K]$, and $\lambda_k=Pw_kP_1$, the norm-constraint can be relaxed by solving

$$\min C(\overline{W},\overline{H},A)=D_{IS}(V|WH)+S(\Lambda H) s.t\ W\geq 0, H\geq 0.$$

[0083] Soft Constraint (Penalization)

[0084] Another way we can control the norm of W is to add an appropriate penalty to the objective function, e.g.,

$$\min C(W,H,A)+\lambda\|W\|_1 s.t\ W\geq 0, H\geq 0$$

[0085] The soft constraint is typically simpler to implement than the hard constraint, but requires the tuning of λ.

[0086] Learning and Inference Procedures for MAP Estimation

[0087] We describe a majorization-minimization (MM) procedure. The MM is an iterative optimization procedure that can be applied to a convex objective function to determine maximums. That is, MM is a way to construct the objective function. MM determines a surrogate function that majorizes the objective function by driving the function to a local optimum. In our embodiments, the matrices H, A, and W are updated conditionally on one and another. In the following, tildes () denote current parameter iterations.

[0088] Inequalities

[0089] For $\{\phi_k\}$ such that $\Sigma_k\phi_k=1$, we have

$$\frac{1}{\sum_k x_k} \leq \sum_k \frac{\phi_k^2}{x_k},$$

by Jensen s inequality. We can form an upper bound on $\log \alpha$ by linearization, at any point ϕ,

$$\log a \leq \log\phi + \frac{\partial \log a}{\partial a}(a-\phi) = (\log\phi - 1) + \frac{a}{\phi}.$$

[0090] In particular,

$$\log\sum_k a_k x_k \leq \left(\log\sum_k a_k \tilde{x}_k - 1\right) + \frac{1}{\sum_j a_j \tilde{x}_j}\sum_k a_k x_k,$$

and

$$\frac{1}{\sum_k a_k x_k} \leq \frac{1}{\left(\sum_j a_j \tilde{x}_j\right)^2}\sum_k a_k \frac{\tilde{x}_k^2}{x_k}.$$

[0091] Fit to Data

$$D_{IS}(V \mid WH) \leq \sum_{kn}\left(\tilde{p}_{kn}\frac{\tilde{h}_{kn}^2}{h_{kn}} + \tilde{q}_{kn}h_{kn}\right)$$

$$\tilde{p}_{kn} = \sum_f w_{fk}\frac{v_{fn}}{\tilde{v}_{fn}^2}$$

$$\tilde{q}_{kn} = \sum_f \frac{w_{fk}}{\tilde{v}_{fn}}$$

$$\tilde{v}_{fn} = \left[W\tilde{H}\right]_{fn}$$

$$D_{IS}(V \mid WH) \leq \sum_{fk}\left(\tilde{p}_{fk}\frac{\tilde{w}_{fk}^2}{w_{fk}} + \tilde{q}_{fk}w_{fk}\right)$$

$$\tilde{p}_{fk} = \sum_n h_{kn}\frac{v_{fn}}{\tilde{v}_{fn}^2}$$

$$\tilde{q}_{fk} = \sum_n \frac{h_{kn}}{\tilde{v}_{fn}}$$

$$\tilde{v}_{fn} = \left[\tilde{W}H\right]_{fn}$$

[0092] Penalty Terms

Let $g_{in} = \sum_j a_{ij}h_{j(n-1)}$.

Then,

$$\log(g_{i(n+1)}) \leq \log(\tilde{g}_{i(n+1)}) + \frac{1}{\tilde{g}_{i(n+1)}}\sum_j a_{ij}\left(h_{jn} - \tilde{h}_{jn}\right)$$

$$\log(g_{i(n+1)}) \leq \log(\tilde{g}_{i(n+1)}) + \frac{1}{\tilde{g}_{i(n+1)}}\sum_j h_{jn}(a_{ij} - \tilde{a}_{ij})$$

$$\frac{1}{g_{i(n+1)}} \leq \frac{1}{\tilde{g}_{i(n+1)}^2}\sum_j a_{ij}\frac{\tilde{h}_{jn}^2}{h_{jn}}$$

$$\frac{1}{g_{i(n+1)}} \leq \frac{1}{\tilde{g}_{i(n+1)}^2}\sum_j h_{jn}\frac{\tilde{a}_{ij}^2}{a_{ij}}$$

$$\left(\tilde{g}_{in} \text{ is either } \sum_j a_{ij}\tilde{h}_{j(n-1)} \text{ or } \sum_j \tilde{a}_{ij}h_{j(n-1)}\right)$$

[0093] Update Rules

[0094] The MM framework includes majorizing the terms of the objective function with the previous inequalities, providing an upper bound of the objective function that is tight at the current parameters, and minimizing the upper bound instead of the original objective. This strategy applied to the minimization of the MAP objective with the soft constraint on the norm of W leads to the following updates **230** as shown in FIG. **2**.

[0095] Update **231** Activation Matrix H

[0096] The columns of H are updated **231** sequentially. Left to right updates makes the update $h_n^{(l)}$ of h_n at iteration l dependent of $h_{n-1}^{(l)}$ and $h_{n+1}^{(l-1)}$. The update of h_{kn} involves rooting a polynomial of order 2, such that

$$h_{kn} = \frac{\sqrt{b^2 - 4ac} - b}{2a}$$

where the values of a, b, c are given in the next table.

n = 1	
a	$\tilde{q}_{kn} + \Sigma_i \alpha_i \frac{a_{ik}}{\tilde{g}_{i(n+1)}}$
b	1 (Jeffreys) or 0 (uniform)
c	$-\tilde{h}_{kn}^2 \left(\tilde{p}_{kn} + \Sigma_i \beta_i \frac{a_{ik} h_{i(n+1)}}{\tilde{g}_{i(n+1)}^2} \right)$
1 < n < N	
a	$\tilde{q}_{kn} + \Sigma_i \alpha_i \frac{a_{ik}}{\tilde{g}_{i(n+1)}} + \frac{\beta_k}{g_{kn}}$
b	$1 - \alpha_k$
c	$-\tilde{h}_{kn}^2 \left(\tilde{p}_{kn} + \Sigma_i \beta_i \frac{a_{ik} h_{i(n+1)}}{\tilde{g}_{i(n+1)}^2} \right)$
n = N	
a	$\tilde{q}_{kn} + \frac{\beta_k}{g_{kn}}$
b	$1 - \alpha_k$
c	$-\tilde{h}_{kn}^2 \tilde{p}_{kn}$

[0097] In particular, for the exponential innovation with expectation 1 ($\alpha_i=\beta_i=1$), we obtain the following multiplicative updates:

For $n = 1$,

$$h_{kn} = \tilde{h}_{kn} \sqrt{\frac{\tilde{p}_{kn} + \sum_i \frac{a_{ik} h_{i(n+1)}}{\tilde{g}_{i(n+1)}^2}}{\tilde{q}_{kn} + \sum_i \frac{a_{ik}}{\tilde{g}_{i(n+1)}} + \frac{1}{\tilde{h}_{kn}}}}.$$

For $1 < n < N$,

$$h_{kn} = \tilde{h}_{kn} \sqrt{\frac{\tilde{p}_{kn} + \sum_i \frac{a_{ik} h_{i(n+1)}}{\tilde{g}_{i(n+1)}^2}}{\tilde{q}_{kn} + \sum_i \frac{a_{ik}}{\tilde{g}_{i(n+1)}} + \frac{1}{g_{kn}}}}$$

For $n = N$,

$$h_{kn} = \tilde{h}_{kn} \sqrt{\frac{\tilde{p}_{kn}}{\tilde{q}_{kn} + \frac{1}{g_{kn}}}}.$$

[0098] Update **232** Basis Function W

$$w_{fk} = \tilde{w}_{fk} \sqrt{\frac{\tilde{p}_{fk}}{\tilde{q}_{fk} + \lambda_W}}$$

[0099] Update **233** Transition Matrix A

$$a_{ij} = \tilde{a}_{ij} \sqrt{\frac{\beta_i \sum_{n=2}^{N} \frac{h_{in} h_{j(n-1)}}{g_{in}^2}}{\alpha_i \sum_{n=2}^{N} \frac{h_{j(n-1)}}{\tilde{g}_{in}} + \lambda_A}}$$

[0100] Variational EM Procedure for Maximum Likelihood Estimation

[0101] The activation parameter H is a latent variable to integrate from the joint likelihood. For generality, we assume the gamma distribution parameters $\beta=\{\beta_i\}$ to be free. The shape parameters α_i are treated as fixed parameters. We minimize

$$C(W,A,\beta) = -\log p(V|W,A,\beta) = -\log \int_H p(V|W,H) p(H|A,\beta)\, dH.$$

[0102] This yields a better posed estimation problem because the set of parameters is of fixed-dimensionality w.r.t to the number of samples N. Furthermore, the objective is now better posed in terms of scales. For any positive diagonal matrix Λ, we have

$$C(W,A,\beta) = C(W\Lambda^{-1}, \Lambda A \Lambda^{-1}, \beta)$$

so that the renormalization of solution W* only induces a renormalization of A*. This is not true for the MAP approach.

[0103] For minimizing C(W,A,β), the EM procedure can be based on the complete dataset (V,H), and on the iterative minimization of

$$Q(\theta|\hat{\theta}) = -\int_H \log p(V,H|W) p(H|V,\hat{\theta})\, dH,$$

where $\theta=\{W,A,\beta\}$. We do not use the posterior probability $p(H|V,\theta)$. Instead, we use a variational EM procedure. For any probability density function q(H), the following inequality holds:

$$C(\theta) \le -\langle \log p(V|WH) \rangle_q - \langle \log p(H|A) \rangle_q + \langle \log q(H) \rangle_q = B_q(\theta),$$

where $\langle \bullet \rangle_q$ denotes the expectation under q(H). Variational EM minimizes $B_q(\theta)$, instead of $C(\theta)$. At each iteration, the bound is first evaluated and tightened, given W and A by minimizing $B_q(\theta)$ over q, or more precisely, over the shape parameters of q, given a specific parameterized form, and then minimized with respect to (θ) given q. Variational EM coincides with EM when $q(H)=p(H|\theta)$, in which case $C(\theta)$ is decreased at every iteration. In other cases, variational EM conducts approximate inference. The validity depends on how well q(H) approximates the true posterior probability $p(H|\theta)$.

[0104] Derivation of the Bound

[0105] The expressions of log p(V|WH) and log p(H|A) show that the coefficients of H are coupled through ratios or logarithms of linear combinations $\Sigma_k w_{fk} h_{kn}$ and $\Sigma_j a_{ij} h_{j(n-1)}$. This makes expectations of log p(V|WH) and log p(H|A) very difficult to determine independently of the specific form of q(H).

[0106] Therefore, we majorize log p(V|WH) and log p(H|A), to obtain a tractable bound. Using the above inequalities and assuming a factored form of the variational distribution, such that

$$q(H) = \prod_{kn} q(h_{kn})$$

is an upper bound of C(W,A,β), the function

$$B_{q,\xi}(W, A, \beta) = \sum_{fkn}\left(\phi_{fkn}^2 \frac{v_{fn}}{w_{fk}}\langle h_{kn}^{-1}\rangle + \frac{w_{fk}}{\psi_{fn}}\langle h_{kn}\rangle\right) + \sum_{fn}(\log\psi_{fn} - 1) +$$

$$\sum_{n=2}^{N}\sum_{i=1}^{K}\left(\begin{array}{c}(1-\alpha_i)\langle\log h_{in}\rangle + \\ \sum_{j=1}^{K}\left(\begin{array}{c}\alpha_i \frac{a_{ij}}{\rho_{in}}\langle h_{j(n-1)}\rangle + \\ \beta_i \frac{v_{ijn}^2}{a_{ij}}\langle h_{in}\rangle\langle h_{j(n-1)}^{-1}\rangle\end{array}\right)\end{array}\right) + \sum_{n=2}^{N}\sum_{i=1}^{K}\alpha_i(\log\rho_{in} - 1) +$$

$$(N-1)\sum_{i=1}^{K}(\log\Gamma(\alpha_i) - \alpha_i\log\beta_i) + \sum_{i=1}^{K}\langle\log h_{i1}\rangle + \sum_{kn}\langle\log q(h_{kn})\rangle$$

Where ϕ_{fkn} are nonnegative coefficients such that $\Sigma_k\phi_{fkn}=1$,

v_{ijn} are nonnegative coefficients such that $\Sigma_i v_{ijn}=1$,

ρ_{in}, ψ_{fn} are nonnegative coefficients,

ξ denotes the set of all tuning parameters $\{\phi_{fkn}, v_{ijn}, \rho_{in}, \psi_{fn}\}^{fknij}$,

$\langle\bullet\rangle$ denotes expectation w.r.t q, i.e., corresponds to $\langle\bullet\rangle_q$. We remove subscript q to alleviate notations.

[0107] The expression of the bound involves the expectation of h_{kn}, $1/h_{kn}$ and $\log h_{kn}$. These expectations are precisely the sufficient statistics of the generalized inverse-Gaussian (GiG), which is a practical convenience for q(H). We use

$$q(H) = \prod_{kn} GIG(h_{kn} \mid \overline{\alpha}_{kn}, \overline{\beta}_{kn}, \overline{\gamma}_{kn}),$$

where

$$GIG(x \mid \alpha, \beta, \gamma) = \frac{(\beta/\gamma)^{\alpha/2}}{2K_\alpha(2\sqrt{\beta\gamma})} x^{\alpha-1}\exp -\left(\beta x + \frac{\gamma}{x}\right),$$

and where K_α is a modified Bessel function of the second kind and x, β and γ are nonnegative scalars. Under the GIG distribution,

$$\langle x\rangle = \frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\sqrt{\frac{\gamma}{\beta}} \tag{13}$$

$$\langle x^{-1}\rangle^{-1} = \frac{K_\alpha(2\sqrt{\beta\gamma})}{K_{\alpha-1}(2\sqrt{\beta\gamma})}\sqrt{\frac{\gamma}{\beta}}. \tag{14}$$

[0108] For any α, $K_{\alpha+1}(x)=2(\alpha/x)K_\alpha(x)+K_{\alpha-1}(x)$, which leads to the alternative, implementation-efficient expression of

$$\langle x^{-1}\rangle = \frac{K_{\alpha+1}(2\sqrt{\beta\gamma})}{K_\alpha(2\sqrt{\beta\gamma})}\sqrt{\frac{\beta}{\gamma}} - \frac{\alpha}{\gamma}. \tag{15}$$

[0109] Optimization of the Bound

[0110] We give the conditional updates of the various parameters of the bound. Update orders are described below.

[0111] Updates

Tuning parameters v

$$\phi_{fkn} = \frac{w_{fk}\langle h_{kn}^{-1}\rangle^{-1}}{\sum_j w_{fj}\langle h_{jn}^{-1}\rangle^{-1}}, \tag{16}$$

$$\psi_{fn} = \sum_j w_{fj}\langle h_{jn}\rangle, \tag{17}$$

$$v_{ijn} = \frac{a_{ij}\langle h_{j(n-1)}^{-1}\rangle^{-1}}{\sum_k a_{ik}\langle h_{k(n-1)}^{-1}\rangle^{-1}}, \tag{18}$$

and

$$\rho_{in} = \sum_j a_{ij}\langle h_{j(n-1)}\rangle \tag{19}$$

Variational distribution q

n = 1	
$\overline{\alpha}_{kn}$	0 (Jeffreys) or 1 (uniform)
$\overline{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\psi_{fn}} + \Sigma_i \alpha_i \frac{a_{ik}}{\rho_{i(n+1)}}$
$\overline{\gamma}_{kn}$	$\Sigma_f \phi_{fkn}^2 \frac{v_{fn}}{w_{fk}} + \Sigma_i \beta_i \frac{v_{ik(n+1)}^2}{a_{ik}}\langle h_{i(n+1)}\rangle$
1 < n < N	
$\overline{\alpha}_{kn}$	α_k
$\overline{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\psi_{fn}} + \Sigma_i \alpha_i \frac{a_{ik}}{\rho_{i(n+1)}} + \beta_k \Sigma_j \frac{v_{kjn}^2}{a_{kj}}\langle h_{j(n-1)}^{-1}\rangle$
$\overline{\gamma}_{kn}$	$\Sigma_f \phi_{fkn}^2 \frac{v_{fn}}{w_{fk}} + \Sigma_i \beta_i \frac{v_{ik(n+1)}^2}{a_{ik}}\langle h_{i(n+1)}\rangle$
n = N	
$\overline{\alpha}_{kn}$	α_k
$\overline{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\psi_{fn}} + \beta_k \Sigma_j \frac{v_{kjn}^2}{a_{kj}}\langle h_{j(n-1)}^{-1}\rangle$
$\overline{\gamma}_{kn}$	$\Sigma_f \phi_{fkn}^2 \frac{v_{fn}}{w_{fk}}$

[0112] Parameters of Interest

$$w_{fk} = \sqrt{\frac{\sum_{n=1}^{N} \phi_{fkn}^2 v_{fn} \langle h_{kn}^{-1} \rangle}{\sum_{n=1}^{N} \psi_{fn}^{-1} \langle h_{kn} \rangle}} \tag{20}$$

$$a_{ij} \sqrt{\frac{\beta_i \sum_{n=2}^{N} v_{ijn}^2 \langle h_{in} \rangle \langle h_{j(n-1)}^{-1} \rangle}{\alpha_i \sum_{n=2}^{N} \rho_{in}^{-1} \langle h_{j(n-1)} \rangle}} \tag{21}$$

$$\beta_i = \alpha_i (N-1) \left(\sum_{n=2}^{N} \frac{\langle h_{in} \rangle}{\sum_j a_{ij} \langle h_{j(n-1)}^{-1} \rangle^{-1}} \right)^{-1} \tag{22}$$

[0113] Updating Order

[0114] We denote the set of tuning parameter for frame n by ξ_n, i.e., $\xi_n \{\{\phi_{fkn}\}_{fk}, \{v_{ijn}\}_{ij}, \{\rho_{in}\}_i, \{\psi_{fn}\}_f\}$.

[0115] As shown in FIG. **2**, the following order of updates **230** leads to an efficient implementation.

[0116] At iteration (l) do

[0117] For n=1, . . . , N,

[0118] Update **231** the activation parameters $[q(h_n)]^{(l)}$ as a function of $[q(h_{n-1})]^{(l)}$, $[q(h_n)]^{(l-1)}$, $[q(h_{n+1})]^{(l-1)}$, $\xi_n^{(2l-2)}$, $W^{(l-1)}$, $A^{(l-1)}$, $\beta^{(l-1)}$.

[0119] Update $\xi_n^{(2l-1)}$.

Update **232** the basis function $W^{(l)}$ as a function of $W^{(l-1)}$, $[q(H)]^{(l)}$, $\xi^{(2l-1)}$.

Update **233** the transition matrix $A^{(l)}$ as a function of $A^{(l-1)}$, $\beta^{(l-1)}$, $[q(H)]^{(l)}$, $\xi^{(2l-1)}$,

Update tuning parameters $\xi^{(2l)}$

Update **234** gamma distribution parameters $\beta^{(l)}$ as a function of the transition matrix $A^{(l)}$ and the activation parameters, $[q(H)]^{(l)}$.

[0120] Under this updating order, the VB-EM procedure is:

[0121] Update q(H).

n = 1	
$\bar{\alpha}_{kn}$	0 (Jeffreys) or 1 (uniform)
$\bar{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\Sigma_j w_{fj} \langle h_{jn} \rangle} + \Sigma_i \alpha_i \frac{a_{ik}}{\Sigma_j a_{ij} \langle h_{jn} \rangle}$
$\bar{\gamma}_{kn}$	$\langle h_{kn}^{-1} \rangle^{-2} \left(\Sigma_f w_{fk} \frac{v_{fn}}{(\Sigma_j w_{fj} \langle h_{jn}^{-1} \rangle^{-1})^2} + \Sigma_i \beta_i a_{ik} \frac{\langle h_{i(n+1)} \rangle}{(\Sigma_j a_{ij} \langle h_{jn}^{-1} \rangle^{-1})^2} \right)$
1 < n < N	
$\bar{\alpha}_{kn}$	α_k
$\bar{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\Sigma_j w_{fj} \langle h_{jn} \rangle} + \Sigma_i \alpha_i \frac{a_{ik}}{\Sigma_j a_{ij} \langle h_{jn} \rangle} + \frac{\beta_k}{\Sigma_j a_{kj} \langle h_{j(n-1)}^{-1} \rangle^{-1}}$
$\bar{\gamma}_{kn}$	$\langle h_{kn}^{-1} \rangle^{-2} \left(\Sigma_f w_{fk} \frac{v_{fn}}{(\Sigma_j w_{fj} \langle h_{jn}^{-1} \rangle^{-1})^2} + \Sigma_i \beta_i a_{ik} \frac{\langle h_{i(n+1)} \rangle}{(\Sigma_j a_{ij} \langle h_{jn}^{-1} \rangle^{-1})^2} \right)$

-continued

n = N	
$\bar{\alpha}_{kn}$	α_k
$\bar{\beta}_{kn}$	$\Sigma_f \frac{w_{fk}}{\Sigma_j w_{fj} \langle h_{jn} \rangle} + \frac{\beta_k}{\Sigma_j a_{kj} \langle h_{j(n-1)}^{-1} \rangle^{-1}}$
$\bar{\gamma}_{kn}$	$\langle h_{kn}^{-1} \rangle^{-2} \Sigma_f w_{fk} \frac{v_{fn}}{(\Sigma_j w_{fj} \langle h_{jn}^{-1} \rangle^{-1})^2}$

Update W, A, β

[0122]

$$w_{fk} = w_{fk} \sqrt{\frac{\sum_{n=1}^{N} \langle h_{kn}^{-1} \rangle^{-1} v_{fn} \left[\sum_j w_{fj} \langle h_{jn}^{-1} \rangle^{-1} \right]^{-2}}{\sum_{n=1}^{N} \langle h_{kn} \rangle \left[\sum_j w_{fj} \langle h_{jn} \rangle \right]^{-1}}}$$

$$a_{ij} = a_{ij} \sqrt{\frac{\beta_i \sum_{n=2}^{N} \langle h_{j(n-1)}^{-1} \rangle^{-1} \langle h_{in} \rangle \left[\sum_k a_{ik} \langle h_{k(n-1)}^{-1} \rangle^{-1} \right]^{-2}}{\alpha_i \sum_{n=2}^{N} \langle h_{j(n-1)}^{-1} \rangle \left[\sum_k a_{ik} \langle h_{k(n-1)} \rangle \right]^{-1}}}$$

$$\beta_i = \alpha_i (N-1) \left(\sum_{n=2}^{N} \frac{\langle h_{in} \rangle}{\sum_j a_{ij} \langle h_{j(n-1)}^{-1} \rangle^{-1}} \right)^{-1}$$

Determine the Bound

[0123]

$$B_{q,\xi}(W, A, \beta) = \sum_{fn} \left(\log \sum_j w_{fj} \langle h_{jn} \rangle + \frac{v_{fn}}{\sum_j w_{fj} \langle h_{jn}^{-1} \rangle^{-1}} \right) +$$

$$\sum_{n=2}^{N} \sum_{i=1}^{K} \left(\alpha_i \log \sum_j a_{ij} \langle h_{j(n-1)} \rangle + \beta_i \frac{\langle h_{in} \rangle}{\sum_j a_{ij} \langle h_{j(n-1)}^{-1} \rangle^{-1}} \right) +$$

$$(N-1) \sum_{i=1}^{K} (\log \Gamma(\alpha_i) - \alpha_i \log \beta_i) -$$

$$\sum_{n=1}^{N} \sum_{i=1}^{K} \left(\bar{\alpha}_{in} \log \sqrt{\frac{\bar{\gamma}_{in}}{\bar{\beta}_{in}}} + \log K_\alpha \left(2 \sqrt{\bar{\beta}_{in} \bar{\gamma}_{in}} \right) + \bar{\beta}_{in} \langle h_{in} \rangle + \bar{\gamma}_{in} \langle h_{in}^{-1} \rangle \right) -$$

$$KN \log 2$$

[0124] Speech Denoising with the Dynamic Model

[0125] As shown in FIG. **3** for one embodiment, we use our method and model for speech enhancement, e.g., denoising. We construct our model parameters **101** for speech **306** by

estimating bases W and the transition matrix A on some speech (audio) training data **305** as described above. We denote the trained bases and transition matrix as $W^{(s)}$ and $A^{(s)}$, where (s) is speech.

[0126] Similarly, we construct a noise model **307** with bases $W^{(n)}$ and transition matrix $A^{(n)}$, and combining the two models **306-307** into the single model **300** by concatenating $W^{(s)}$ and $W^{(n)}$ into $W=[W^{(s)},W^{(n)}]$, and $A^{(s)}$ and $A^{(n)}$ into A, where A is a block-diagonal matrix with $A^{(s)}$ and $A^{(n)}$ on the diagonal.

[0127] We can also train for noise on some noise training data, or we can fix the speech part of the model, and train for the noise part on the test data, thus making the noise part a general model that collects parts of the signal that cannot be modeled by the speech model. The simplest version of the later model uses a single basis for the noise, and uses an identity matrix as the transition matrix A.

[0128] After the model **300** is constructed, we can use the model to enhance an input audio signal x **301**. We determine **310** a time-frequency feature representation. We estimate **320** the parameters of the model **300** that vary, i.e., the activation matrix $H^{(s)}$ for the speech and $H^{(n)}$ for the noise (n), and the bases $W^{(n)}$ and transition matrix $A^{(n)}$ for the noise.

[0129] Thus, we obtain a single model that combines speech, $W^{(s)}H^{(s)}$ and noise $W^{(n)}H^{(n)}$, which we then use to reconstruct **330** the complex STFT of the enhanced speech $\hat{x}$ **340**, using

$$\hat{x}_{fn} = \frac{\sum_k w_{fk}^{(s)} H_{kn}^{(s)}}{\sum_k W_{fk}^{(s)} H_{kn}^{(s)} + \sum_k W_{fk}^{(n)} H_{kn}^{(n)}} x_{fn}. \quad (23)$$

[0130] The time-domain signal can be reconstructed using a conventional overlap-add method, which evaluates a discrete convolution of a very long input signal with a finite impulse response filter

[0131] Extensions

[0132] Other complex models can also generated based on the above embodiments.

[0133] Dirichlet Innovations

[0134] Instead of considering the innovation random variables ϵ_n to be gamma distributed, the innovation can be Dirichlet distributed, which is similar to a normalization of the activation parameter h_n.

[0135] HMM-Like Behavior

[0136] We can constrain h_n to be 1-sparse during inference.

[0137] Structured Variational Inference

[0138] Conventional variational inference assumes that the variational posterior probabilities $q(h_n)$ are independent of each other, which, given a strong dependency relation between h_n and h_{n-1}, is likely to be very wrong. We can model the posterior probability in terms of $q(h_n|h_{n-1})$. One possibility for such a q distribution uses a GIG distribution with parameters dependent on Ah_{n-1}.

[0139] Gamma Distribution of Innovation

[0140] The complex Gaussian model on the complex STFT coefficients in Eqn. (6) is equivalent to assuming that the power is exponentially distributed with parameter WH. We can extend the model by assuming that the power is gamma distributed, thus leading to a donut-shaped distribution for the complex coefficients.

[0141] Full Covariance of Innovation Random Variables

[0142] In linear dynamical systems, the innovation random variables can have a full-covariance. For positive random variables, one way to include the correlations is to transform an independent random vector with a non-negative matrix. This leads to the model,

$$h_n=(Ah_{n-1})\circ(Bf_n),$$

where f_n is a nonnegative random vector of size J×1 and B is a nonnegative matrix of dimension K×J. When $B=I_{K\times K}$, this simplifies to $f_n=\epsilon_n$. This can be accomplished in the more general form of the model

$$h_{i,n} = \sum_{j,l} c_{i,j,l}\epsilon_{l,n}h_{j,n-1}$$

by setting the parameters to a factorized form: $c_{i,j,l}=a_{i,j}b_{i,l}$, where $a_{i,j}$ are the elements of A, and $b_{i,j}$ are the elements of B.

[0143] Transition Innovations

[0144] It can also be useful to model the transition between each of the components of h_n and h_{n-1} using separate innovation random variables. This is analogous to the use of Dirichlet prior probabilities in discrete Markov models. One method would admit $h_n=(A\circ E_n)h_{n-1}$, where E_n is a nonnegative innovations matrix of dimension K×K. This can be accomplished in the more general form of the model

$$h_{i,n} = \sum_{j,l} c_{i,j,l}\epsilon_{l,n}h_{j,n-1}$$

by setting the parameters $c_{i,j,l}=\delta(m(i,j),l)a_{i,j}$, where $a_{i,j}$ are the elements of A and m(i,j) is a one-to-one mapping from each combination of i and j to an index corresponding to l. Then, the i, j-th element of E_n is $\epsilon_{m(i,j),n}$.

[0145] Considering Other Innovation Types Besides Gam Ma

[0146] A log-normal, Poisson distribution leads to yet different types of dynamical systems.

[0147] Considering Other Divergences

[0148] We so far only considered the Itakura-Saito divergence. We can also use the KL-divergence, and different divergences for $h_n|h_{n-1}$ and for v|h.

[0149] Online Procedure

[0150] For real-time applications, only the signal up to the current time is used, e.g., an application where only the activation matrix H are estimated, or another application where all parameters are optimized. In the later application, we can perform a "warm" start with pretrained bases W and transition matrix A.

[0151] Multi-Channel Version

[0152] Because our model relies on a generative model involving the complex STFT coefficients, the model can be extended to a multi-channel application. Optimization in this setting involves EM updates between mixing system and a source NMF procedure.

EFFECT OF THE INVENTION

[0153] The embodiments of the invention provide a non-negative linear dynamical system model for processing non-stationary signals, particularly speech signals mixed with noise. In the context of speech separation and speech denois-

ing, our model adapts to signal dynamics on-line, and achieves better performance than conventional methods.

[0154] Conventional models for signal dynamics frequently use hidden Markov models (HMMs) or non-negative matrix factorization (NMF). HMMs lead to combinatorial problems due to the discrete state space, are computationally complex, especially for mixed signals from several sources, and make it difficult to handle gain adaptation. NMF solves both the computational complexity and gain adaptation problems. However, NMF does not take advantage of past observations of a signal to model future observations of that signal. For signals with predictable dynamics, this is likely to be suboptimal.

[0155] Our model has advantages of both the HMMs and the NMF. The model is characterized by a continuous non-negative state space. Gain adaptation is automatically handled during inference. The complexity of the inference is linear in the number of sources, and dynamics are modeled via a linear transition matrix.

[0156] Although the invention has been described by way of examples of preferred embodiments, it is to be understood that various other adaptations and modifications can be made within the spirit and scope of the invention. Therefore, it is the object of the appended claims to cover all such variations and modifications as come within the true spirit and scope of the invention.

We claim:

1. A method for transforming an input signal, comprising the steps of:

storing parameters of a model of the input signal in a memory;

receiving the input signal as a sequence of feature vectors;

inferring, using the sequence of feature vectors and the parameters, a sequence of vectors of hidden variables, wherein there is at least one vector h_n of hidden variables $h_{i,n}$ for each feature vector x_n, and wherein each hidden variable is nonnegative;

generating an output signal corresponding to the input signal, using the feature vectors, the vectors of hidden variables, and the parameters,

wherein each feature vector x_n is dependent on at least one of the hidden variables $h_{i,n}$ for the same n, and the hidden variables are related according to

$$h_{i,n} = \sum_{j,l} c_{i,j,l} \epsilon_{l,n} h_{j,n-1},$$

where j and l are summation indices, the parameters include non-negative weights $c_{i,j,l}$, and $\epsilon_{l,n}$ are independent non-negative random variables, wherein the steps are performed in a processor.

2. The method of claim **1**, wherein $c_{i,j,l}=\delta(i,l)a_{i,j}$, where $a_{i,j}$ are non-negative scalars, and where δ is a Kronecker delta, so that

$$h_{i,n} = \left(\sum_j a_{i,j} h_{j,n-1}\right) \epsilon_{i,n}.$$

3. The method of claim **1**, wherein $c_{i,j,l}=\delta(m(i,j),l)a_{i,j}$, where $a_{i,j}$ are non-negative scalars, δ is a Kronecker delta and, m(i,j) is a one-to-one mapping from each combination of i and j to an index corresponding to l, so that

$$h_{i,n} = \sum_j a_{i,j} \epsilon_m(i,\ j),\ n^h j,\ n-1.$$

4. The method of claim **1**, wherein the random variables $\epsilon_{l,n}$ are gamma distributed.

5. The method of claim **1**, wherein an observation model used during the inferring is based at least in part on

$$v_{f,n} = \sum_j c^{(v)}_{f,i,l} h_{i,n} \epsilon^{(v)}_{l,n},$$

where $c_{f,i,l}^{(v)}$ is a non-negative scalar, and $\epsilon_{l,n}{}^{(v)}$ are independent non-negative random variables, $v_{f,n}$ is a non-negative feature of the input signal at a frame n and feature f and j, and l are indices.

6. The method of claim **5**, wherein $c_{f,i,l}{}^{(v)}=\delta(i,l)w_{f,j}$, where $w_{f,j}$ are non-negative scalars, where δ is a Kronecker delta, and $\epsilon_{f,n}{}^{(v)}$ are Gamma distributed random variables, so that the observation model based at least in part on

$$p(v_{f,n} \mid h_n) = \text{Gamma}\left(v_{f,n} \mid \alpha^{(v)},\ \beta^{(v)} \Big/ \sum_i w_{f,i} h_{i,n}\right),$$

where $v_{f,n}$ is a non-negative feature of the input signal at frame n, f is frequency, Gamma(.|a,b) is a gamma distribution with shape parameter a and inverse-scale parameter b, $\alpha^{(v)}$ and $\beta^{(v)}$ are positive scalars, and $w_{f,i}$ are non-negative scalars.

7. The method of claim **5**, further comprising:

obtaining the feature vectors $x_{f,n}$ as a complex spectrogram of the input signal, where $x_{f,n}$ is a value of the complex spectrogram for a frame n and frequency f, and

determining a non-negative feature $v_{f,n}=|x_{f,n}|^2$ as a power in frame n and frequency f so that the observation model is based at least in part on

$$x_{f,n} = (e^{\theta_{f,n}\sqrt{-1}})\sqrt{v_{f,n}},$$

where $\sqrt{-1}$ is a unit imaginary number, and $\theta_{f,n}$ is a random variable representing a phase for the frame n and the frequency f.

8. The method of claim **6**, further comprising:

setting the parameter $\alpha^{(v)}=1$, and where $\theta_{f,n}$ is a uniformly distributed random phase variable, so that

$$p(x_{f,n} \mid h_n) = N_c\left(0,\ \sum_i w_{f,i} h_{i,n}\right).$$

where N_C is a complex Gaussian distribution.

9. The method of claim **1**, wherein the inferring uses a maximum a-posteriori estimation.

10. The method of claim **1**, wherein the inferring uses a variational Bayes method.

11. The method of claim **1**, wherein the inferring is adaptive and performed on-line on the input signal.

12. The method of claim **1**, wherein the input signal is received simultaneously multiple channels.

13. The method of claim **1**, wherein an observation model used during the inferring is based at least in part on

$$u_{i',n} = \sum_{i} c^{(u)}_{i',i,l} h_{i,n} \varepsilon^{(u)}_{l',n},$$

and

$$v_{f,n} = \sum_{i'} c^{(v)}_{f,i',l} u_{i',n} \varepsilon^{(v)}_{l'',n},$$

where

$c^{(u)}_{i',i,l'}$

and

$c^{(v)}_{f,i',l''}$

are non-negative scalars, and $\epsilon_{l',n}{}^{(u)}$ and $\epsilon_{l'',n}{}^{(v)}$ are independent non-negative random variables, and i, i', l', l'', f, and n are indices.

14. The method claim **1**, where the hidden variables $h_{i,n}$ are partitioned into S groups, and the non-negative random variables $\epsilon_{l,n}$ are each associated with one of the groups, wherein $c_{i,j,l}=0$ when $h_{i,n}$, and $h_{j,n}$, or $h_{i,n}$ and $\epsilon_{l,n}$ are in different groups.

15. The method of claim **1**, wherein the model is dynamic, and the input signal is non-stationary.

16. The method of claim **1**, further comprising:

adapting to again of the input signal on-line during the inferring.

17. The method of claim **1**, wherein the input signal is a mixed signal of speech and noise, and the output signal is an enhanced speech signal.

18. The method of claim **1**, wherein the parameters include basis functions W, a transition matrix A, an activation matrix H, a fixed shape parameter α, an inverse scale parameter β of a continuous gamma distribution parameter, and various combinations thereof.

19. The method of claim **18** wherein updating H and β are optional.

20. The method of claim **18**, wherein updating β is optional in a maximum a-posteriori estimation used by the inferring.

21. The method of claim **1**, wherein the input signal is received simultaneously from multiple sources by a single sensor.

22. The method of claim **18**, wherein a posterior distribution of H is used in a variational Bayes method.

* * * * *