Advanced Audio Processing
2. Audio Source Separation

Nicolas Obin
nicolas.obin@upmc.fr

Université Pierre et Marie Curie

(2014-2016)

A classical example : the cocktail party effect

In a cocktail party :

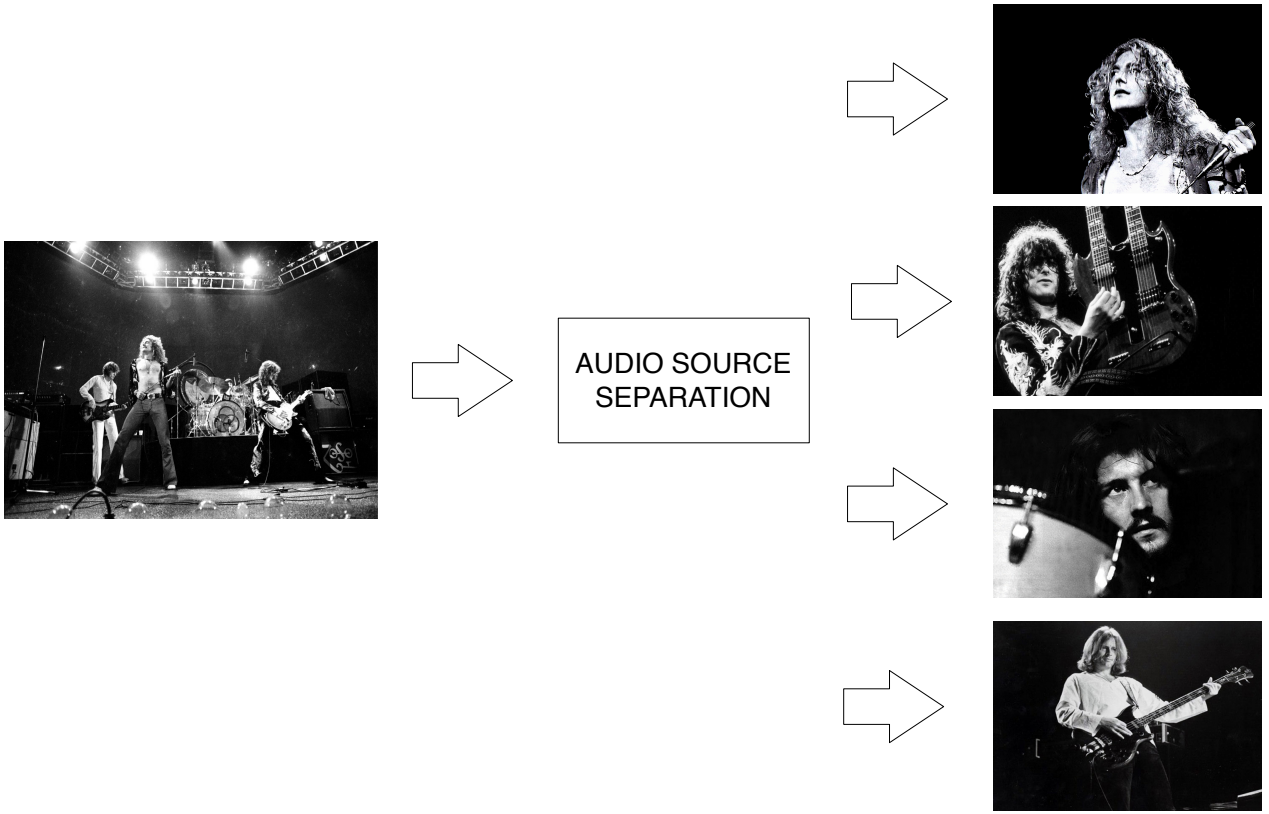
- Humans are able to focus on a specific audio source
- Artificial listening machines should be able to do the same

This is the objective of audio source separation

Principle of Audio Source Separation

The principle of audio source separation is to retrieve the original audio sources from a mixture of audio sources

- 4 sources (voice, guitar, drums, bass)
- 1/2 audio channels (mono/stereo audio track)



Formulation

Definition

- Unknown : K sources $s_k(t)$, concatenated into a source vector $s(t)$
- Observed : N mixtures $m_n(t)$, concatenated into a mixture vector $m(t)$

The general mixture model can be expressed as :

$$m(t) = A_t(s(\tau)), t \geq \tau$$

Assumptions

- Stationnarity

$$A_t = A$$

- Linearity : A is a linear application

Typologies of Mixture Models

Memory

- Instantaneous mixture

$$\mathbf{m(t)} = \mathbf{A}\ \mathbf{s(t)}$$

where : $\mathbf{A}$ is the $(N \times K)$ mixture matrix

For the n-th microphone :

$$\mathbf{m_n(t)} = \sum_{\mathbf{k}} \mathbf{a_{nk} s_k(t)}$$

Instantaneous mixtures assume that there is no time-delay between sources and microphones

Typologies of Mixture Models

Memory

- Convolutive mixture

$$m(t) \quad = \quad (A * s)(t) \tag{1}$$

For the n-th microphone :

$$m_n(t) \quad = \quad \sum_{k=1}^{K} \sum_{l=0}^{L-1} a_{nk}(l) s_k(t-l) \tag{2}$$

where : a_{nk} is a filter with impulse response $[a_{nk}(0), \ldots, a_{nk}(L-1)]$ that relates source k with microphone n

Convolutive mixtures assume that there are time-delays between sources and microphones

Typologies of Mixture Models

Inversibility

- Determined mixture

$$N \text{ (microphones)} = K \text{ (sources)}$$

- Over-determined mixture

$$N \text{ (microphones)} > K \text{ (sources)}$$

- Under-determined mixture

$$N \text{ (microphones)} < K \text{ (sources)}$$

Instantaneous Linear Mixture 1

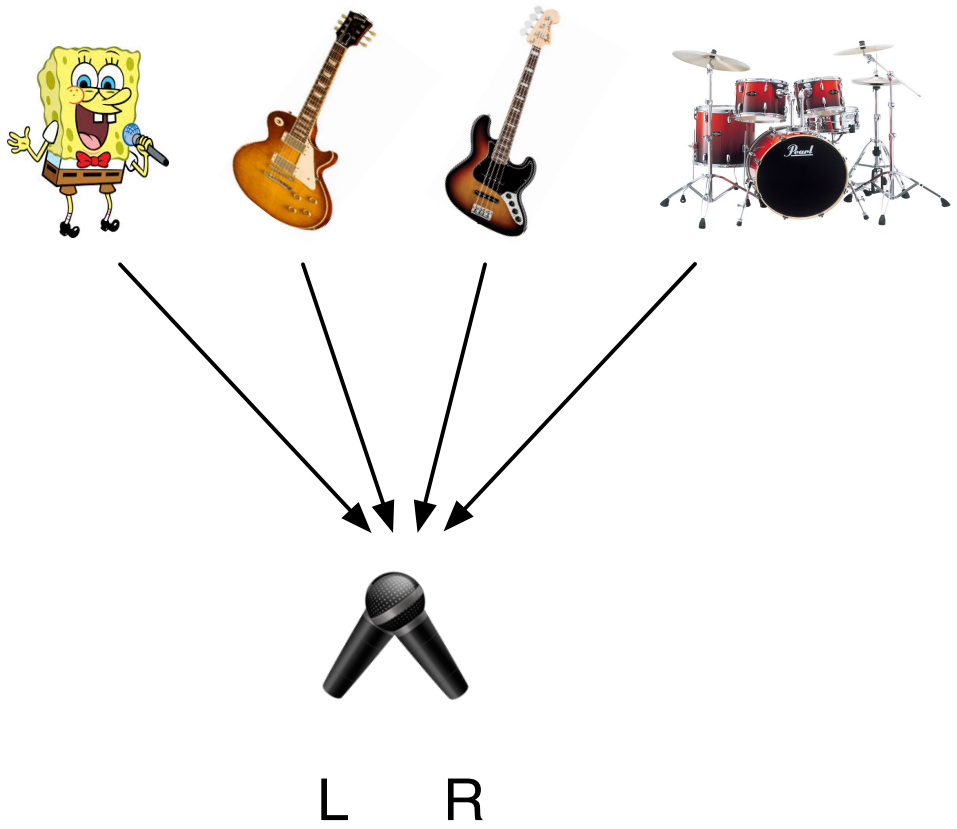


FIGURE: Stereo recording : XY configuration.

Convolutive Linear Mixture 1

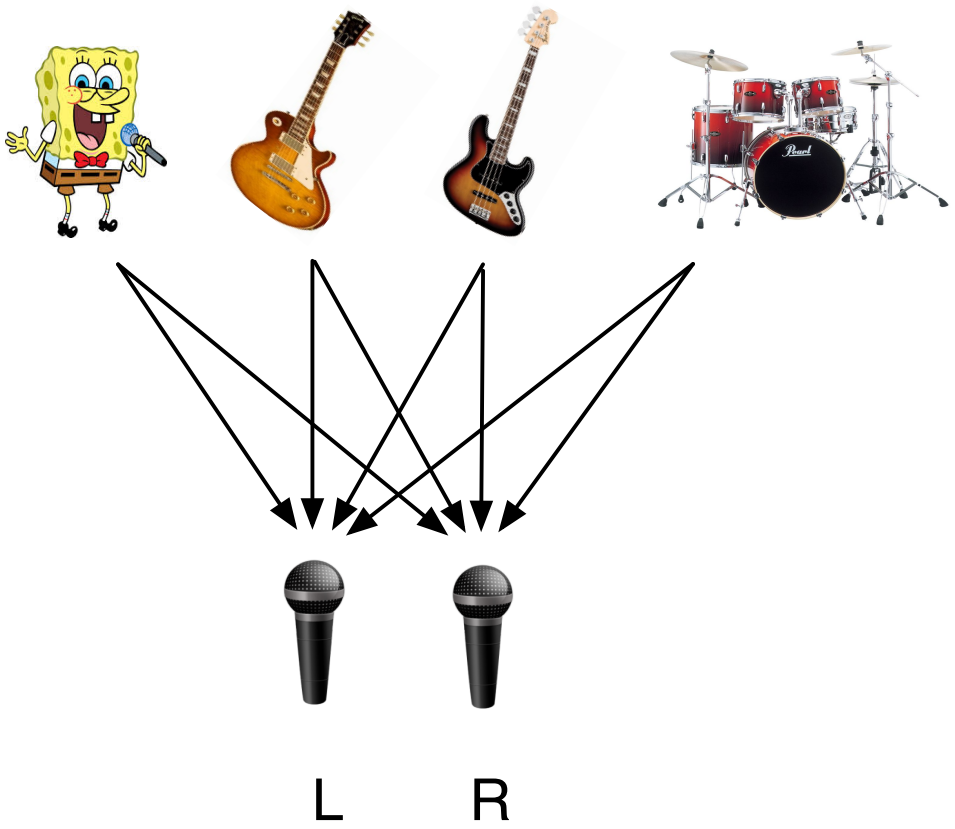


FIGURE: Stereo recording : AB configuration.

Convolutive Linear Mixture 2

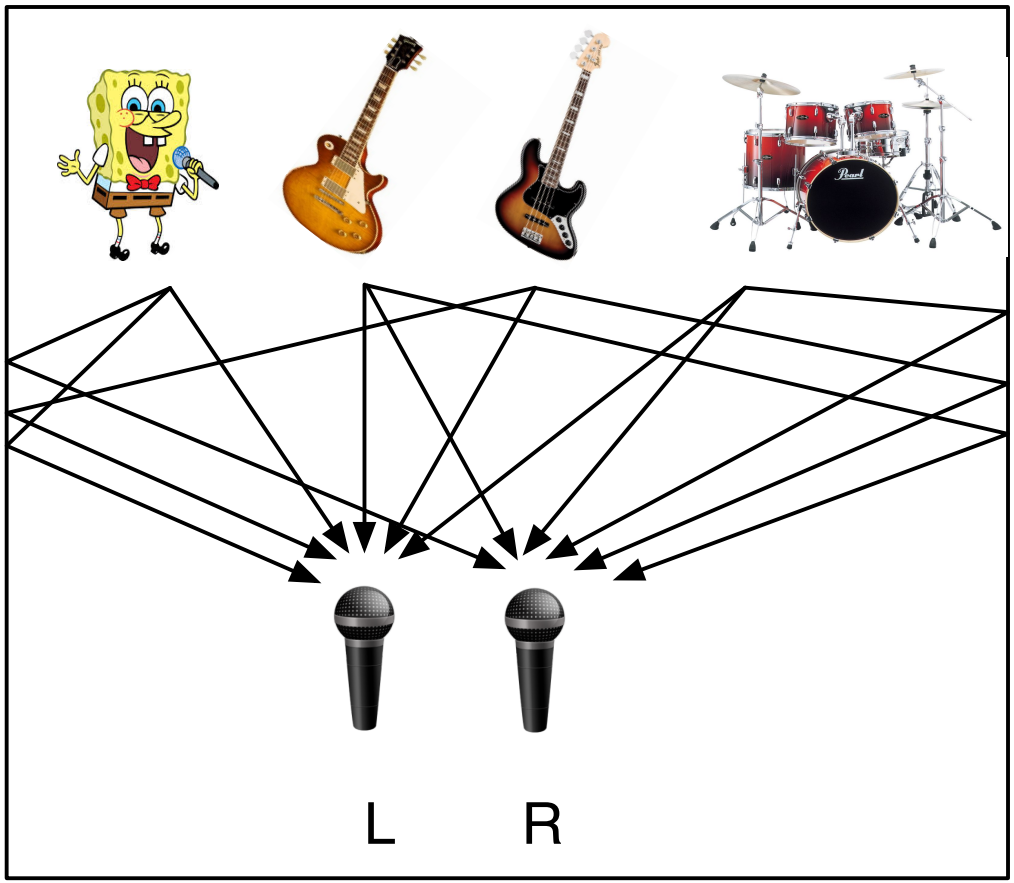


FIGURE: Stereo recording : AB configuration + room reverberation.

Typology of audio source separation methods

Information

- Blind source separation
  Source separation is processed without any knowledge about the nature of the sources

- Informed source separation
  Source separation is processed with knowledge about the nature of the sources

Supervision

- Unsupervised source separation
  Source separation is processed without any training

- Semi-supervised source separation
  Source separation is trained for a limited number of sources of interest

- Supervised source separation
  Source separation is trained for all sources

Resume

Audio source separation consists in solving a mixture model :

$$m(t) = As(t)$$

where :

- $m(t)$ is observed
- A and $s(t)$ are unknown

Concretely, one needs to estimate from the observed signals $m(t)$:

- the number of sources K
- the mixture A
- the source signals $s(t)$

In the remaining, we will assume : a stationary, linear, and instantaneous mixture model

Non-Negative Matrix Factorization (*NMF*)

Definition

The non-negative matrix factorization of a (KxN) matrix $\mathbf{V}$ ($\mathbf{V} \geq \mathbf{0}$) is written as :

$$\begin{cases} \mathbf{V} \simeq \mathbf{WH} \\ \mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0} \end{cases} \tag{3}$$

Notations

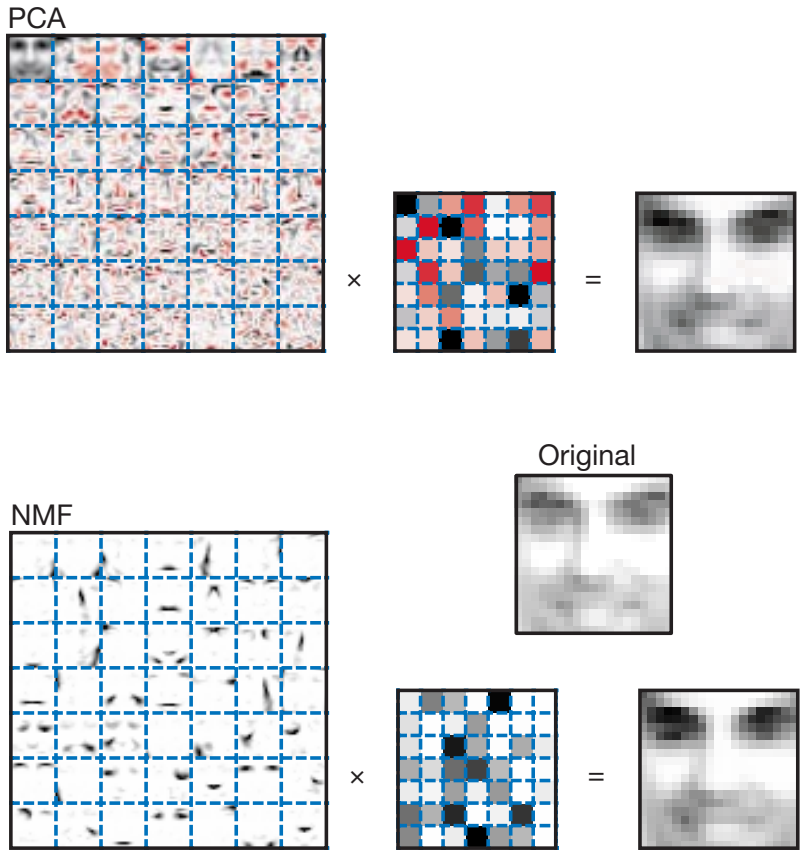
- $\mathbf{W}$ is the (KxS) *dictionary matrix*
- $\mathbf{H}$ is the (SxN) *activation matrix*

Properties

- NMF is a rank reduction algorithm (PCA,...)
- The positivity constraint blocks the creation of *black energy* (vs. PCA)

History

First, NMF has been introduced in image processing [Lee and Seung, 1999]

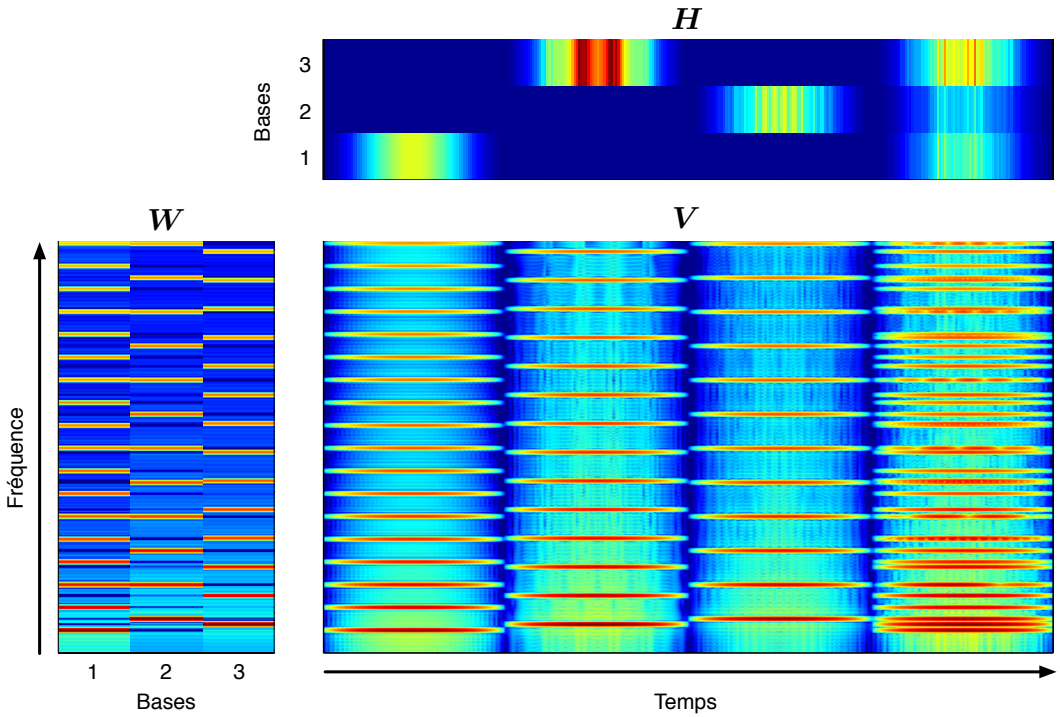


NMF provides a decomposition of a signal into elementary and interpretative parts

NMF of audio signals

The extension to audio signals is straightforward, by using the amplitude of the short-term Fourier transform (STFT) of the audio signal :

$$|X(k,n)| = STFT(x[n]) \tag{4}$$



Phases are ignored during NMF...

Estimation of NMF parameters

Problem : Estimating the model parameters $\theta = \{W, H\}$ that minimizes the cost function $\mathcal{C}(\mathbf{V}|\mathbf{WH})$

$$(\hat{W}, \hat{H}) = \underset{\mathbf{W},\mathbf{H}}{\arg\min}\ \mathcal{C}(\mathbf{V}|\mathbf{WH}) \tag{5}$$

subject to :

$$\mathbf{W}, \mathbf{H} \geq \mathbf{0} \tag{6}$$

This estimation requires optimization under constraints : the gradient descent algorithm.

At iteration i :

$$\theta^{(i+1)} = \theta^{(i)} - \boldsymbol{\eta}\, \nabla_\theta \mathcal{C} \tag{7}$$

subject to :

$$\theta^{(i+1)} \geq 0 \tag{8}$$

Multiplicative Update Rules

Solution : A particular solution is obtained by iteratively updating the gradient step $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \frac{\theta^{(i)}}{\nabla^{+}_{\theta^{(i)}} \mathcal{C}} \tag{9}$$

Here, the gradient of the cost function $\nabla_\theta \mathcal{C}$ is decomposed into a positive part $\nabla^{+}_{\theta} \mathcal{C}$ and a negative part $\nabla^{-}_{\theta} \mathcal{C}$:

$$\nabla_\theta \mathcal{C} = \nabla^{+}_{\theta} \mathcal{C} - \nabla^{-}_{\theta} \mathcal{C} \tag{10}$$

This leads to the multiplicative update rules :

$$\theta^{(i+1)} = \theta^{(i)} \otimes \frac{\nabla^{-}_{\theta} \mathcal{C}}{\nabla^{+}_{\theta} \mathcal{C}} \tag{11}$$

Proof 1 : the positivity constraint is guaranteed by the positivity of $\nabla^{+}_{\theta} \mathcal{C}$ and $\nabla^{-}_{\theta} \mathcal{C}$

Proof 2 : this solution is proved to converge to a stationary point (but not necessarily to a minimum) for the euclidean distance, the Kullback-Leibler divergence [Lee and Seung, 2001], and the β-divergence [Nakano et al., 2010].

We only need to derive the gradient of the cost function, depending on the cost function.

Usual Cost functions

β-divergence ($\beta \in \mathbb{R} \setminus \{0,1\}$)	$d_\beta(x\|y)$	$=$	$\frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}\right)$
Itakura-Saito divergence	$d_{IS}(x\|y)$	$=$	$\frac{x}{y} - \log\frac{x}{y} - 1$
		$=$	$\lim_{\beta\to 0} d_\beta(x\|y)$
Kullback-Leiber divergence	$d_{KL}(x\|y)$	$=$	$x\log\frac{x}{y} + (y-x)$
		$=$	$\lim_{\beta\to 1} d_\beta(x\|y)$
Euclidean distance	$d_{EUC}(x\|y)$	$=$	$\frac{1}{2}(x-y)^2$
		$=$	$d_\beta(x\|y)$ pour $\beta = 2$

Example 1 : Euclidean Distance

The euclidean distance between A and B is defined as :

$$D_{EUC}(\boldsymbol{A} \mid \boldsymbol{B}) = \sum_{i,j} (\boldsymbol{A}_{i,j} - \boldsymbol{B}_{i,j})^2 \tag{12}$$

In the case where $A = V$ and $B = WH$:

$$D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = \sum_{i,j} (\boldsymbol{V}_{i,j}^2 + \boldsymbol{WH}_{i,j}^2 - 2\boldsymbol{V}_{i,j} \times \boldsymbol{WH}_{i,j}) \tag{13}$$

Then, the gradient is :

$$\nabla_\theta D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = \frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial \theta} \tag{14}$$

This can be computed term by term as :

$$\frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial \theta_{p,q}} = \sum_{i,j} (0 + 2[WH]_{i,j} \frac{\partial [WH]_{i,j}}{\partial \theta_{p,q}} - 2V_{i,j} \frac{\partial [WH]_{i,j}}{\partial \theta_{p,q}}) \tag{15}$$

Example 1 : Euclidean Distance

Solution for W

This solution is obtained by computing the gradient :

$$\frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial W_{p,q}} = \sum_{i,j}(2[WH]_{i,j}\frac{\partial [WH]_{i,j}}{\partial W_{p,q}} - 2V_{i,j}\frac{\partial [WH]_{i,j}}{\partial W_{p,q}}) \quad (16)$$

By definition :

$$[WH]_{i,j} = \sum_{k} W_{i,k}H_{k,j} \quad (17)$$

Thus :

$$\frac{\partial [WH]_{i,j}}{\partial W_{p,q}} = \sum_{k}\frac{\partial W_{i,k}}{\partial W_{p,q}}H_{k,j} \quad (18)$$

This partial derivative is non zero for $i = p$ and $k = q$

Thus :

$$\frac{\partial [WH]_{i,j}}{\partial W_{p,q}} = \delta_{i,p}\delta_{k,q}H_{k,j} = \delta_{i,p}H_{q,j} \quad (19)$$

Example 1 : Euclidean Distance

Finally :

$$\frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial W_{p,q}} = 2\sum_{i,j} \delta_{i,p}[WH]_{i,j} H_{q,j} - 2\sum_{i,j} V_{i,j} H_{q,j} \delta_{i,p} \quad (20)$$

$$= 2\sum_{j} [WH]_{p,j} H^{\top}_{j,q} - 2\sum_{j} V_{p,j} H^{\top}_{j,q} \quad (21)$$

$$= \left[2(WH)H^{\top} - 2VH^{\top}\right]_{p,q} \quad (22)$$

In a matrix form :

$$\nabla_W D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2(WH)H^{\top} - 2VH^{\top} \quad (23)$$

The positive and negative parts of the gradients are :

$$\nabla^{+}_{\theta} D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2(WH)H^{\top} \quad (24)$$

$$\nabla^{-}_{\theta} D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2VH^{\top} \quad (25)$$

The multiplicative update for W is :

$$\boxed{W^{(i+1)} = W^{(i)} \otimes \frac{VH^{\top}}{(WH)H^{\top}}} \quad (26)$$

Example 1 : Euclidean Distance

Solution for H

This solution is obtained by computing the gradient :

$$\frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial H_{p,q}} = \sum_{i,j} (2[WH]_{i,j} \frac{\partial [WH]_{i,j}}{\partial H_{p,q}} - 2V_{i,j} \frac{\partial [WH]_{i,j}}{\partial H_{p,q}}) \quad (27)$$

By definition :

$$[WH]_{i,j} = \sum_{k} W_{i,k} H_{k,j} \quad (28)$$

Thus :

$$\frac{\partial [WH]_{i,j}}{\partial H_{p,q}} = \sum_{k} W_{i,k} \frac{\partial H_{k,j}}{\partial H_{p,q}} \quad (29)$$

(30)

This partial derivative is non zero for $k = p$ and $j = q$

Thus :

$$\frac{\partial [WH]_{i,j}}{\partial H_{p,q}} = W_{i,p} \delta_{j,q} \delta_{k,p} = W_{i,p} \delta_{j,q} \quad (31)$$

Example 1 : Euclidean Distance

Finally :

$$\frac{\partial D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial H_{p,q}} = 2\sum_{i,j} \delta_{j,q} W_{i,p} [WH]_{i,j} - 2\sum_{i,j} V_{i,j} W_{i,p} \delta_{j,q} \quad (32)$$

$$= 2\sum_{i} W^{\top}_{p,i} [WH]_{i,q} - 2\sum_{i} W^{\top}_{p,i} V_{i,q} \quad (33)$$

$$= \left[2W^{\top}(WH) - 2W^{\top}V\right]_{p,q} \quad (34)$$

In a matrix form :

$$\nabla_H D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2W^{\top}(WH) - 2W^{\top}V \quad (35)$$

The positive and negative parts of the gradients are :

$$\nabla^{+}_{\theta} D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2W^{\top}(WH) \quad (36)$$

$$\nabla^{-}_{\theta} D_{EUC}(\boldsymbol{V} \mid \boldsymbol{WH}) = 2W^{\top}V \quad (37)$$

The multiplicative update for H is :

$$H^{(i+1)} = H^{(i)} \otimes \frac{W^{\top}V}{W^{\top}(WH)} \quad (38)$$

Example 2 : β-divergence

The β-divergence distance between A and B is defined as :

$$D_\beta(\boldsymbol{A} \mid \boldsymbol{B}) = \frac{1}{\beta(\beta-1)} \sum_{i,j} \left(\boldsymbol{A}_{i,j}^\beta + (\beta-1)\boldsymbol{B}_{i,j}^\beta - \beta \boldsymbol{A}_{i,j} \boldsymbol{B}_{i,j}^{\beta-1} \right) \quad (39)$$

In the case where $A = V$ and $B = WH$:

$$D_\beta(\boldsymbol{V} \mid \boldsymbol{WH}) = \frac{1}{\beta(\beta-1)} \sum_{i,j} \left(\boldsymbol{V}_{i,j}^\beta + (\beta-1)[\boldsymbol{WH}]_{i,j}^\beta - \beta \boldsymbol{V}_{i,j} [\boldsymbol{WH}]_{i,j}^{\beta-1} \right) \quad (40)$$

Then, the gradient is :

$$\nabla_\theta D_\beta(\boldsymbol{V} \mid \boldsymbol{WH}) = \frac{\partial D_\beta(\boldsymbol{V} \mid \boldsymbol{WH})}{\partial \theta} \quad (41)$$

This can be computed term by term as :

$$\begin{aligned} \frac{\partial D_\beta}{\partial \theta_{pq}} &= \frac{1}{\beta(\beta-1)} \sum_{i,j} \left(\beta(\beta-1)[WH]_{i,j}^{\beta-1} \frac{\partial [WH]_{i,j}}{\partial \theta_{p,q}} - \beta \mathbf{V_{i,j}} (\beta-\mathbf{1}) [\mathbf{WH}]_{\mathbf{i,j}}^{\beta-\mathbf{2}} \frac{\partial [\mathbf{WH}]_{\mathbf{i,j}}}{\partial \theta_{\mathbf{p,q}}} \right) \\ &= \underbrace{\sum_{i,j} \left([WH]_{i,j}^{\beta-1} \frac{\partial [WH]_{i,j}}{\partial \theta_{p,q}} \right)}_{\nabla^+_{\theta_{p,q}}} - \underbrace{\sum_{i,j} \left(\mathbf{V_{i,j}} [\mathbf{WH}]_{\mathbf{i,j}}^{\beta-\mathbf{2}} \frac{\partial [\mathbf{WH}]_{\mathbf{i,j}}}{\partial \theta_{\mathbf{p,q}}} \right)}_{\nabla^-_{\theta_{p,q}}} \end{aligned}$$

Example 2 : β-divergence

Multiplicative updates of the β-divergence [Nakano et al., 2010]

The multiplicative update for W is :

$$W^{(i+1)} = W^{(i)} \otimes \left(\frac{\left[(WH)^{\beta-2} V \right] H^\top}{(WH)^{\beta-1} H^\top} \right)^{\phi(\beta)} \tag{42}$$

The multiplicative update for H is :

$$H^{(i+1)} = H^{(i)} \otimes \left(\frac{W^\top \left[(WH)^{\beta-2} V \right]}{W^\top (WH)^{\beta-1}} \right)^{\phi(\beta)} \tag{43}$$

Exercise !

Algorithm

- initialization

$$\begin{cases} W^{(0)} \longleftarrow W_0 \\ H^{(0)} \longleftarrow H_0 \end{cases} \tag{44}$$

The initialization is generally random

- update

$$\begin{cases} W^{(i+1)} \longleftarrow W^{(i)} \otimes \dfrac{\nabla_W^- \mathcal{C}}{\nabla_W^+ \mathcal{C}} \\ H^{(i+1)} \longleftarrow H^{(i)} \otimes \dfrac{\nabla_H^- \mathcal{C}}{\nabla_H^+ \mathcal{C}} \end{cases} \tag{45}$$

The update requires to compute the gradient of the cost function

- convergence

$$|\Delta \mathcal{C}| \leq threshold \tag{46}$$

Example : NMF in action !

Original audio signal

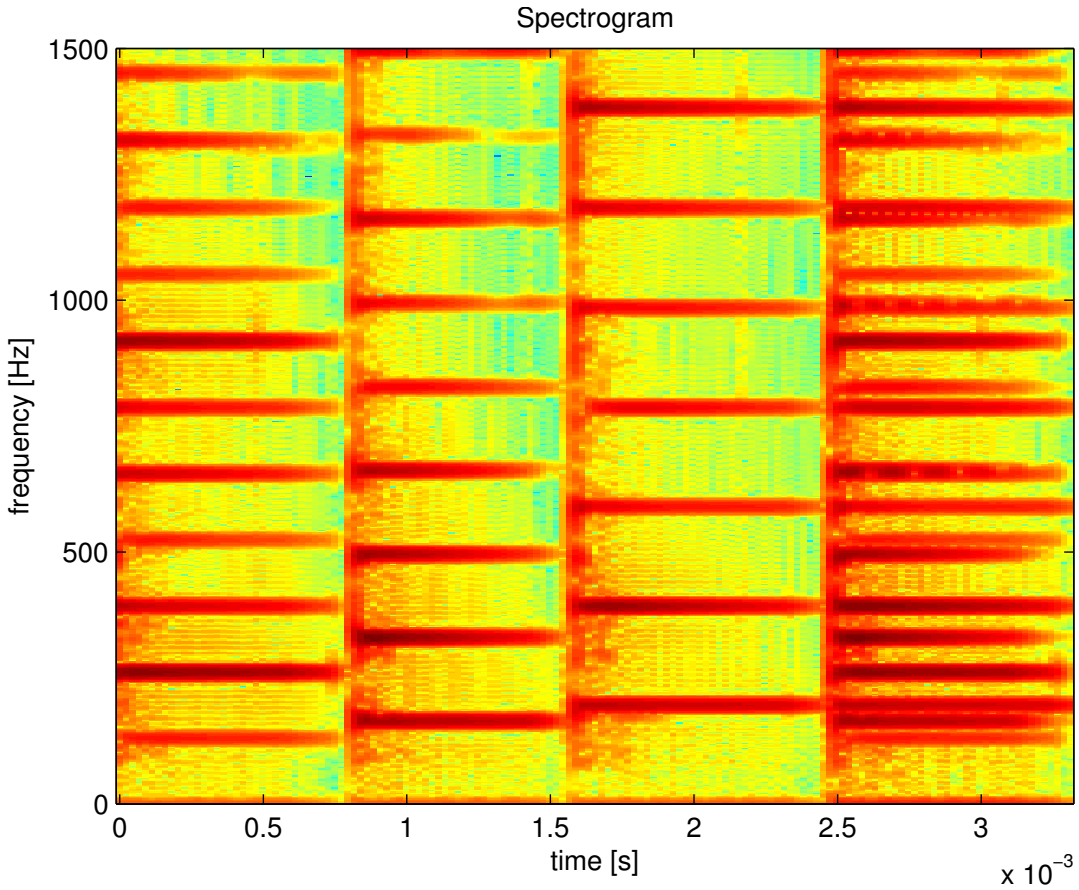
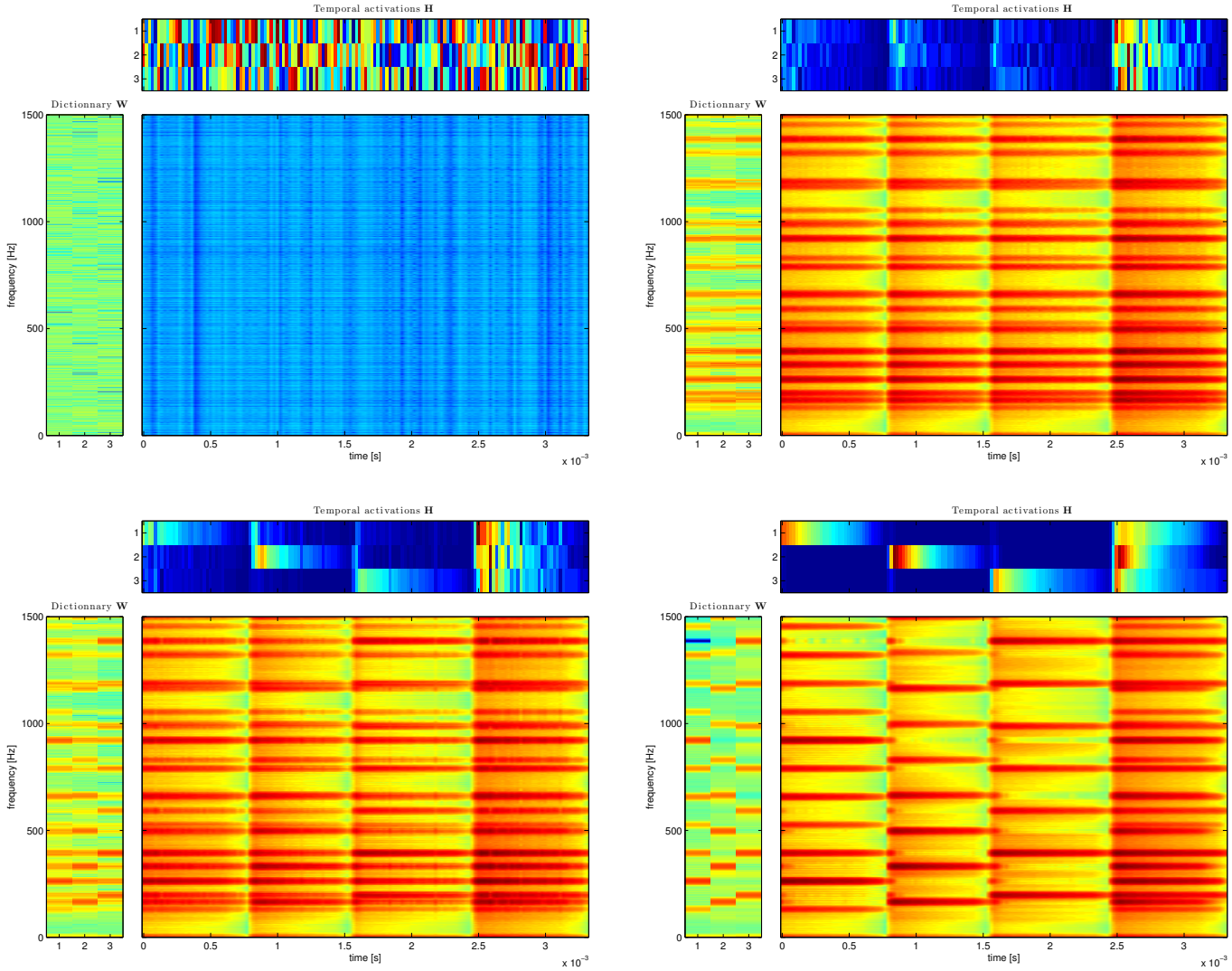


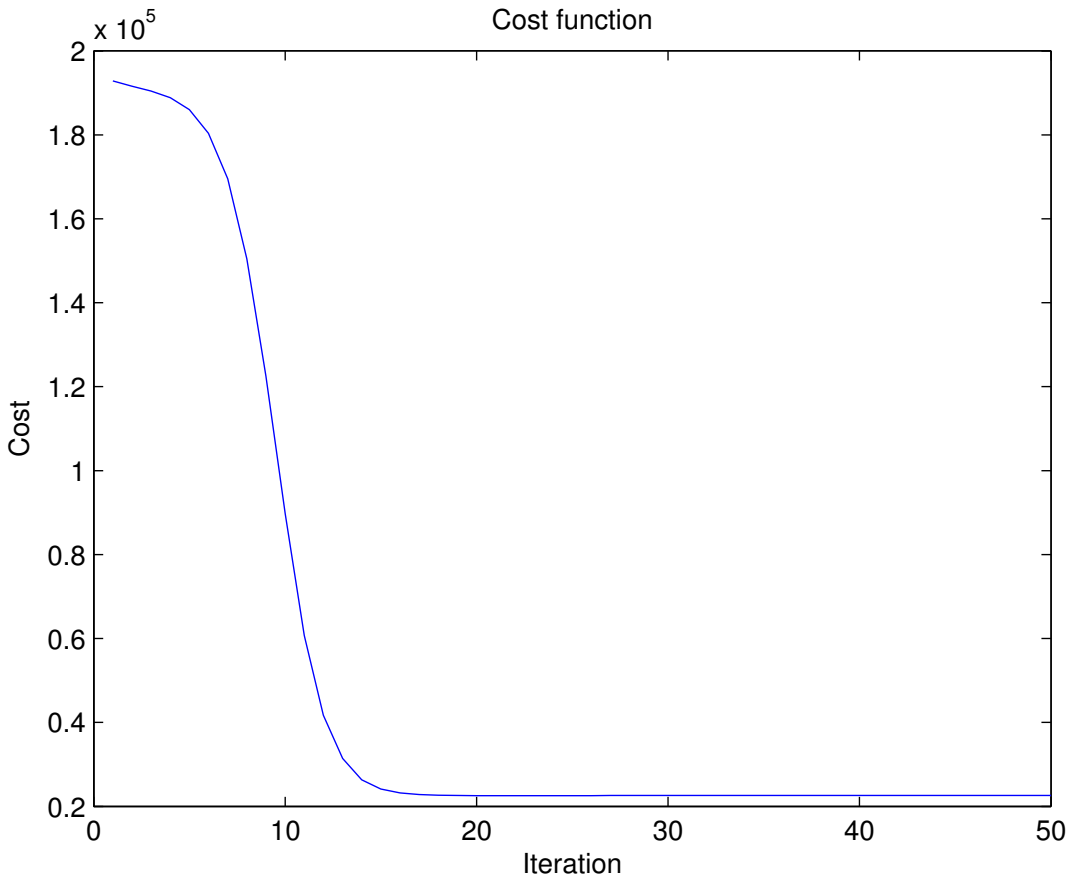
FIGURE: a sequence of 3 musical notes (do/mi/sol)

Example : NMF in action !



Example : NMF in action !

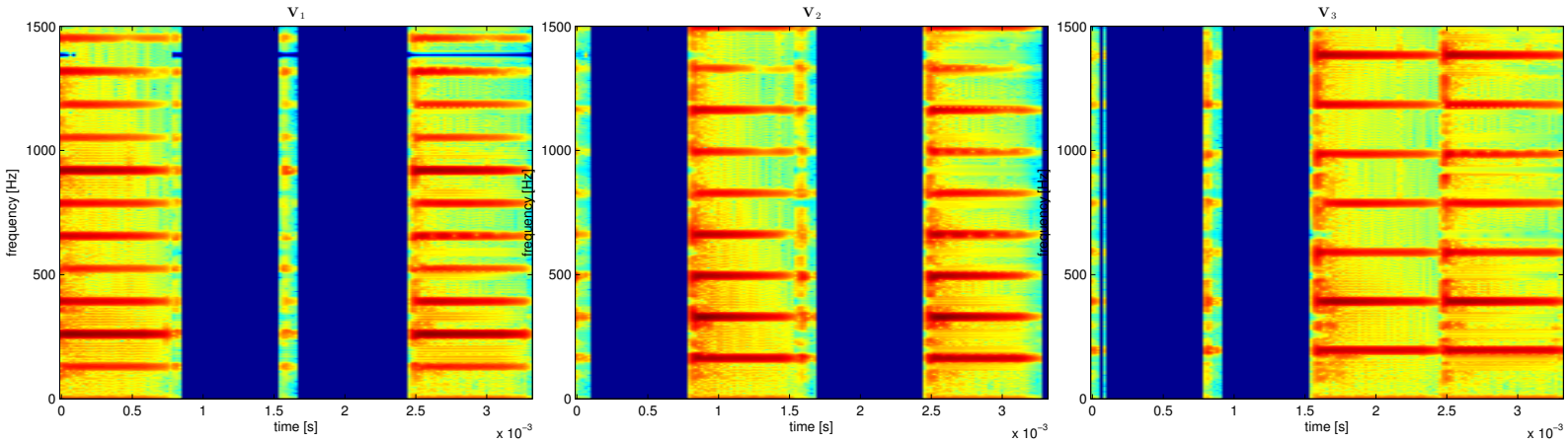
Cost function



Reconstruction

After NMF, each audio source S can be expressed as a frequency mask over time :

$$|X_S[k,n]| = W_S H_S \tag{47}$$



The signal of the audio source S can be then reconstructed by :

- Wiener filtering [Benaroya and Bimbot, 2003]
- inverse short-term Fourier transform (iSTFT)

Phase

What about the phases ? (required for iSTFT)

- Phases generally remain preserved from the original audio mix

Limitations of standard NMF

Limitations

Originally, NMF is :

- blind : no information is available about the audio sources
- unsupervised : separation is processed without any prior training

Solutions

Robust audio source separation requires information and supervision

Advanced NMF faces :

- information : add of knowledge and constraints about the sources and the separation
- supervision : separation is processed with prior training of a set of dictionaries

Advanced NMF

Many variants of the NMF have been proposed to exploit additional information about audio sources

Informed NMF

- contrained NMF [Bertin, 2009] : adding constraints to the NMF
- SF-NMF [Durrieu et al., 2009] : NMF based on a source/filter model
- text information [Le Magoarou et al., 2014] : exploiting prior knowledge about text (sequence of audio events)

Supervised/Semi-Supervised NMF

- fully supervised NMF [Virtanen et al., 2013]
- NMF-HMM [Mysore and Smaragdis, 2012] : NMF is the observation model combined with the language model of a HMM
- Universal Speaker Model [Sun and Mysore, 2013] : train speaker-dependent dictionaries on a large number of speakers
- Deep NMF [Le Roux et al., 2015] : deep learning of NMF parameters

Constrained NMF

The behavior of the NMF can be controlled by adding constraints

Constraints are simply integrated by adding a regularization term to the cost function :

$$\underbrace{\mathcal{C}(\boldsymbol{V}|\boldsymbol{\Lambda}(\Theta))}_{\text{cost}} = \underbrace{D(\boldsymbol{V}|\boldsymbol{\Lambda}(\Theta))}_{\text{divergence measure}} + \underbrace{\lambda\, d_c(\Theta)}_{\text{constraint measure}} \tag{48}$$

Usual Constraints [Bertin, 2009]

- Sparsity : force the NMF to have a minimum number of activation at the same time
- Decorrelation : avoid the NMF to have correlated temporal activation (respectively, frequency templates)
- Continuity : force the temporal continuity of activation (respectively, the frequency continuity of templates).

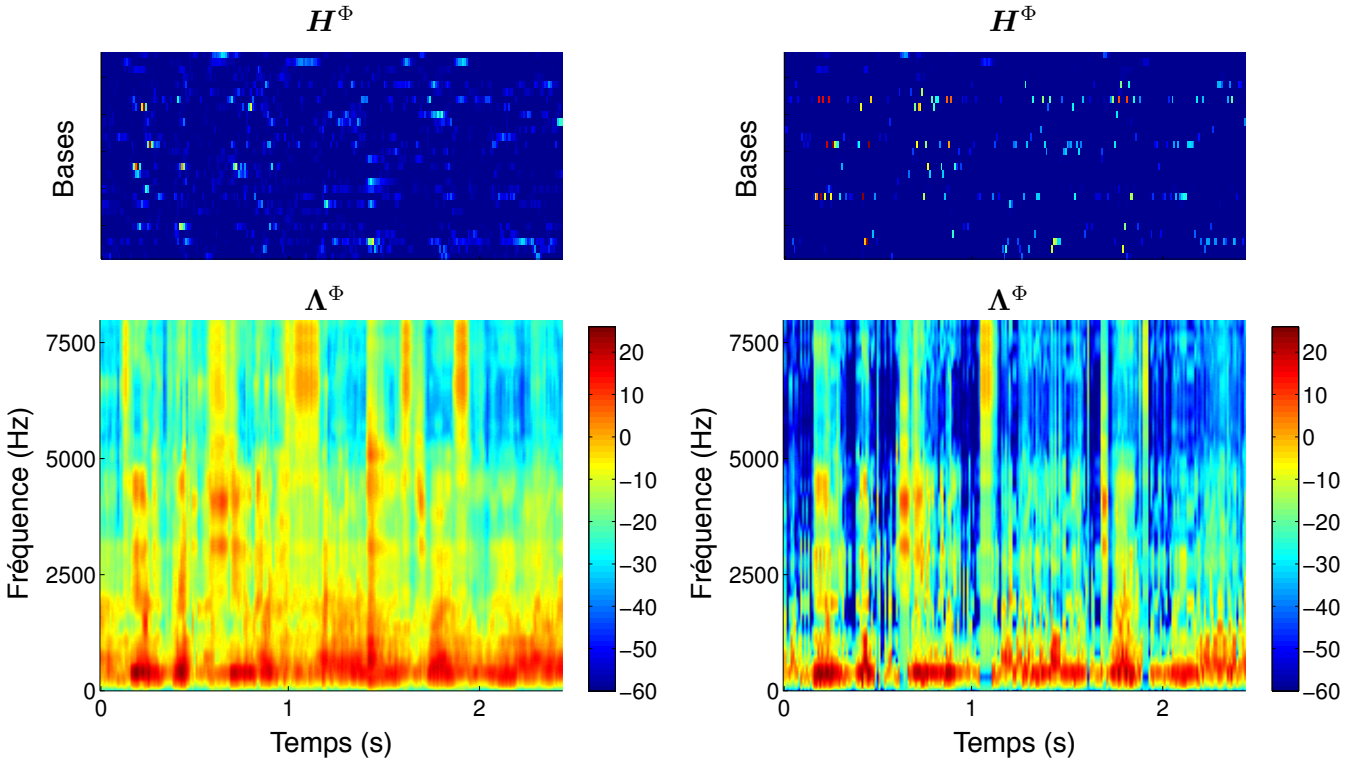
Estimation

- Update rules must be rewritten, with respect to the constraints
- The convergence to a stationary point is not guaranteed (but generally observed)

Sparsity Constraint

The sparsity of the activation matrix is defined as :

$$d_p(\mathbf{H}) = \left(\frac{\|\mathbf{H}\|_{\ell 1}}{\|\mathbf{H}\|_{\ell 2}} \right)^2 = \left(\frac{\sum_{i,j} |\, \mathbf{H_{ij}} \,|}{\sqrt{\sum_{i,j} \mathbf{H_{ij}^2}}} \right)^2 \tag{49}$$

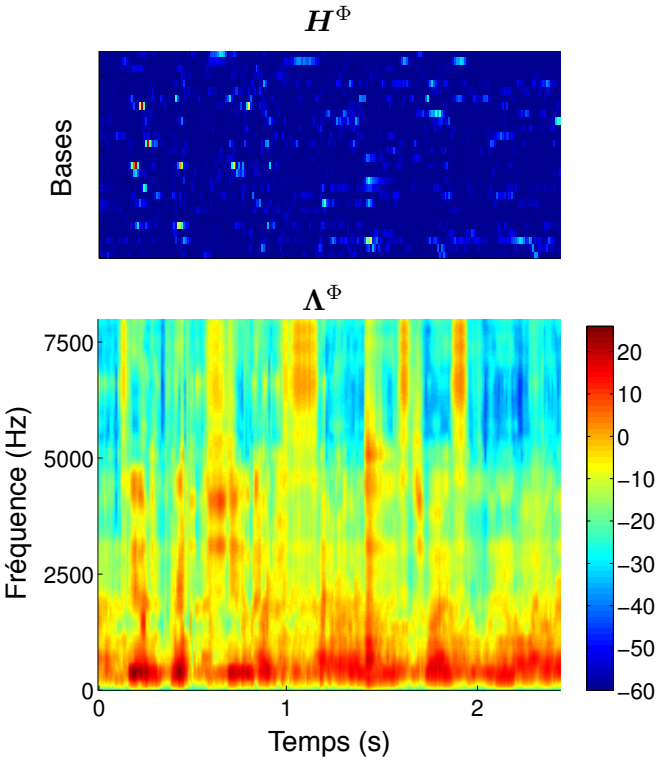


(a) w/o sparsity (b) w sparsity

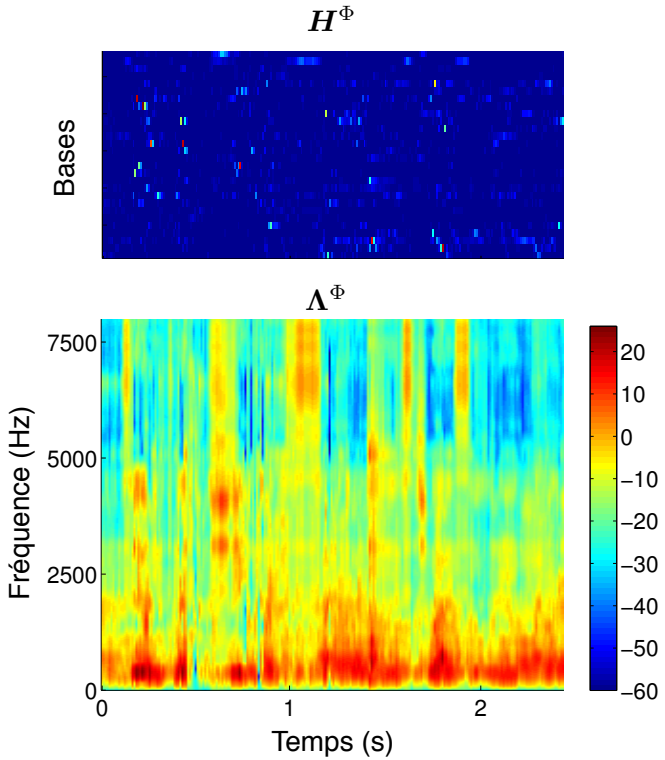
Decorrelation Constraint

The decorrelation of the activation matrix is defined as :

$$d_d(\mathbf{H}) = \sum_{\mathbf{j \neq k}} \left[\mathbf{H}\mathbf{H}^{\mathbf{T}}\right]_{\mathbf{jk}} \tag{50}$$



(c) w/o decorrelation

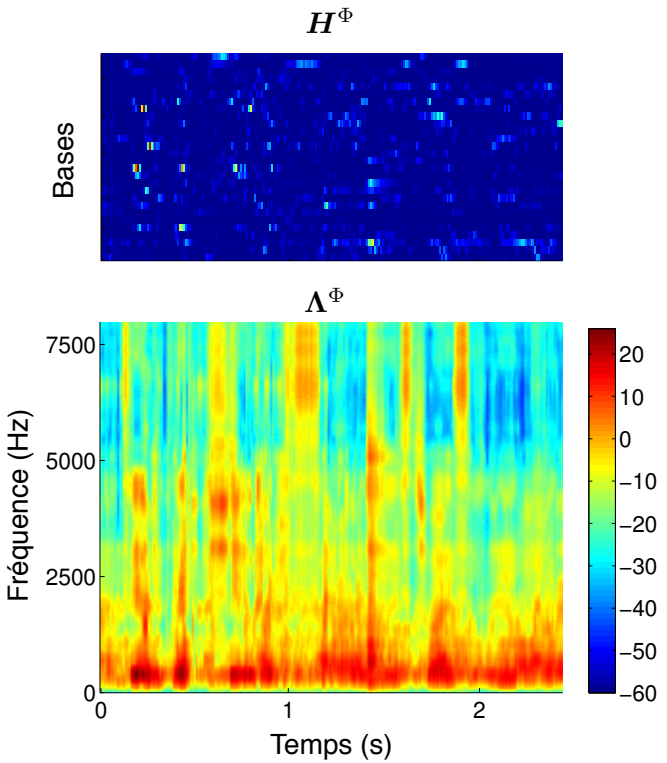


(d) w decorrelation

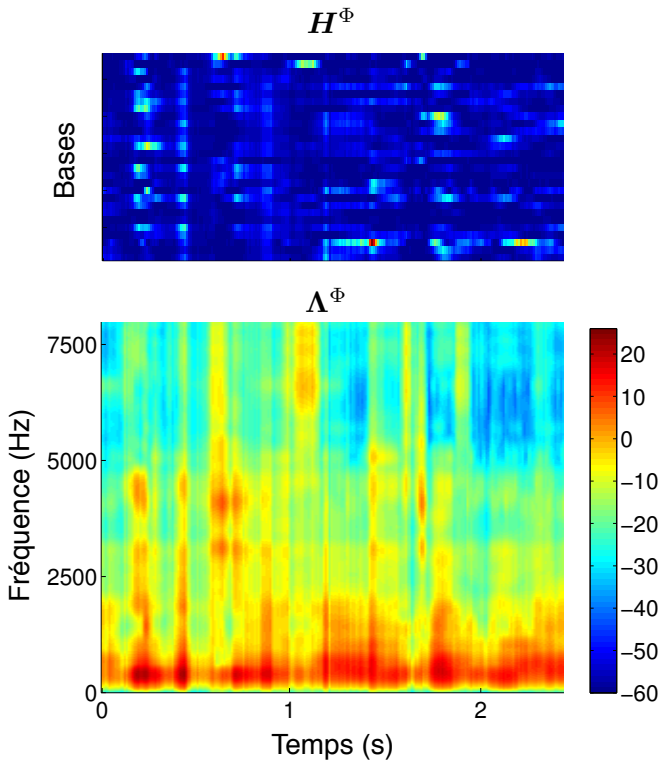
Continuity Constraint

The continuity of the activation matrix is defined as :

$$d_c(\mathbf{H}) = \sum_{\mathbf{i=1}}^{\mathbf{I}} \frac{\mathbf{1}}{\sigma_{\mathbf{i}}^{\mathbf{2}}} \sum_{\mathbf{j=2}}^{\mathbf{J}} \left(\mathbf{H_{ij}} - \mathbf{H_{ij-1}}\right)^{\mathbf{2}}, \ \sigma_i^2 = \sqrt{\frac{1}{J} \sum_j H_{ij}^2} \tag{51}$$



(e) w/o continuity



(f) w continuity

Source/Filter Model

Definition

The source/filter model can be used to provide information about the nature of the audio sources [Durrieu et al., 2009] (speech/music) :

$$\begin{aligned} \mathbf{V} &= \mathbf{V}^{\text{ex}} \otimes \mathbf{V}^{\mathbf{\Phi}} \\ &\simeq \mathbf{W}^{\text{ex}}\mathbf{H}^{\text{ex}} \otimes \mathbf{W}^{\mathbf{\Phi}}\mathbf{U}^{\mathbf{\Phi}}\mathbf{H}^{\mathbf{\Phi}} \end{aligned} \tag{52}$$

$\mathbf{W}^{\text{ex}}$	$F \times L$	source dictionary	imposed
$\mathbf{H}^{\text{ex}}$	$L \times N$	source activation	free
$\mathbf{W}^{\mathbf{\Phi}}$	$F \times P$	elementary filter dictionary	imposed
$\mathbf{U}^{\mathbf{\Phi}}$	$P \times K$	filter dictionary	free
$\mathbf{H}^{\mathbf{\Phi}}$	$K \times N$	filter activation	free

Estimation

- Multiplicative updates can be derived
- No proof of convergence to a stationary point

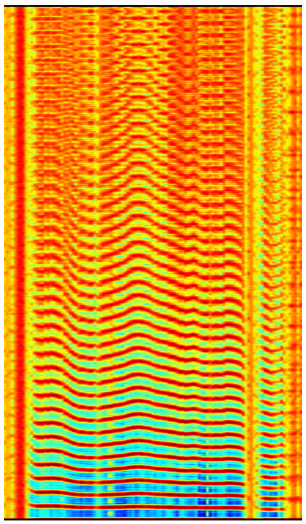
Source/Filter Model

source $*$ filter $=$ sound

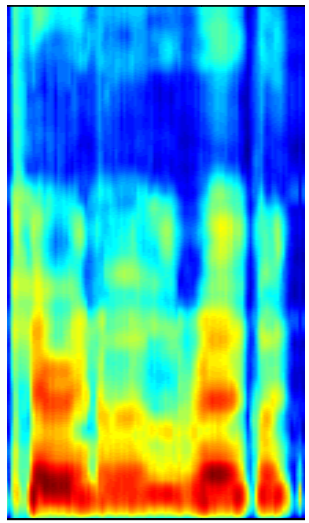
Source/Filter Model

source * filter = sound

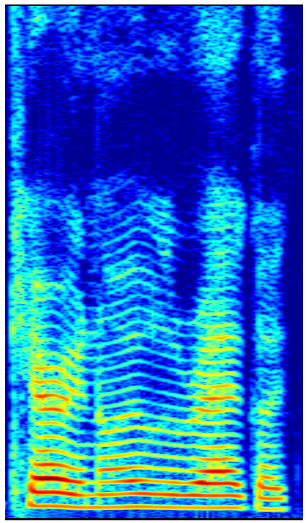
glottal source → vocal tract → voice



.*



=

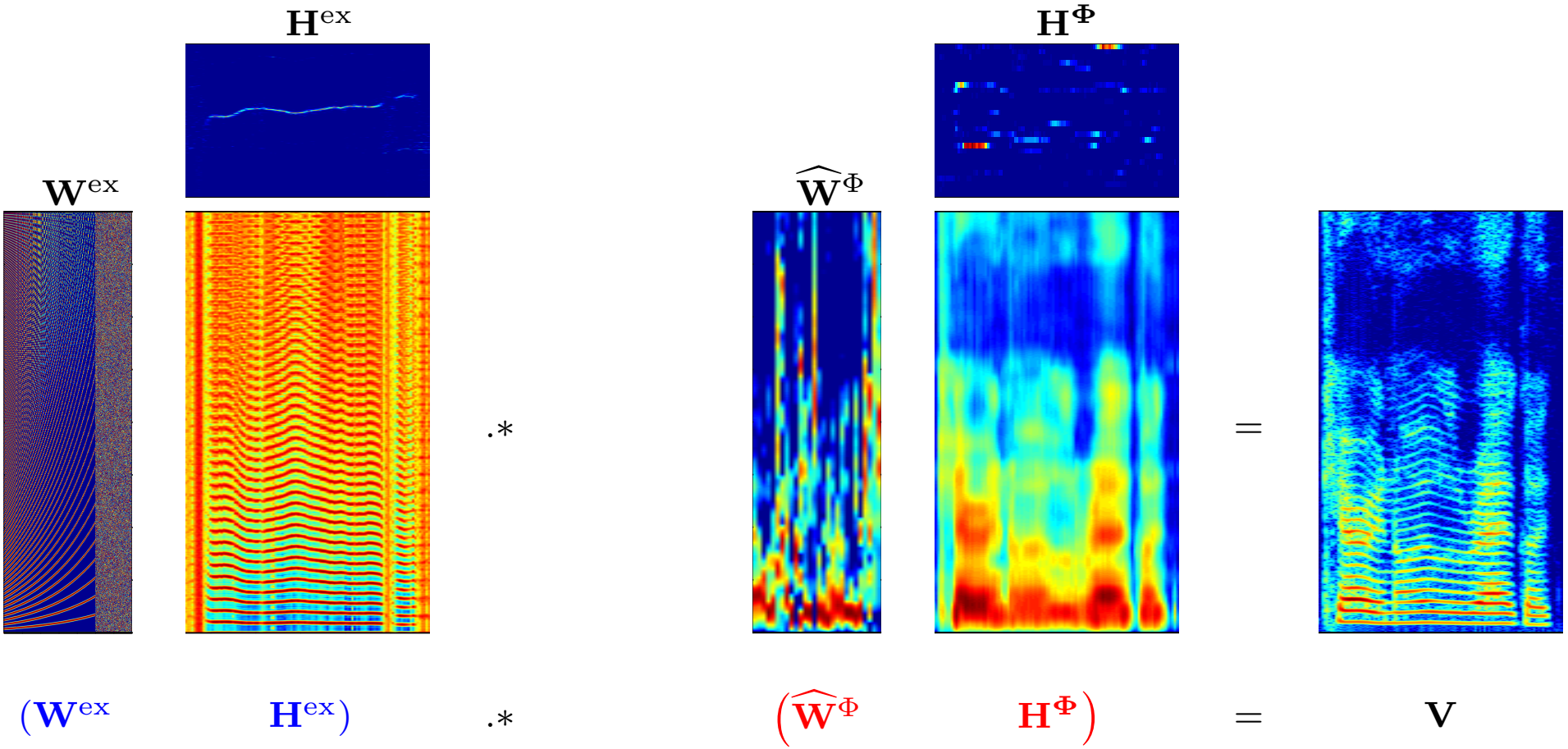


$\mathbf{V}^{\text{ex}}$.* $\mathbf{V}^{\Phi}$ = $\mathbf{V}$

Source/Filter Model

source * filter = sound

glottal source → vocal tract → voice

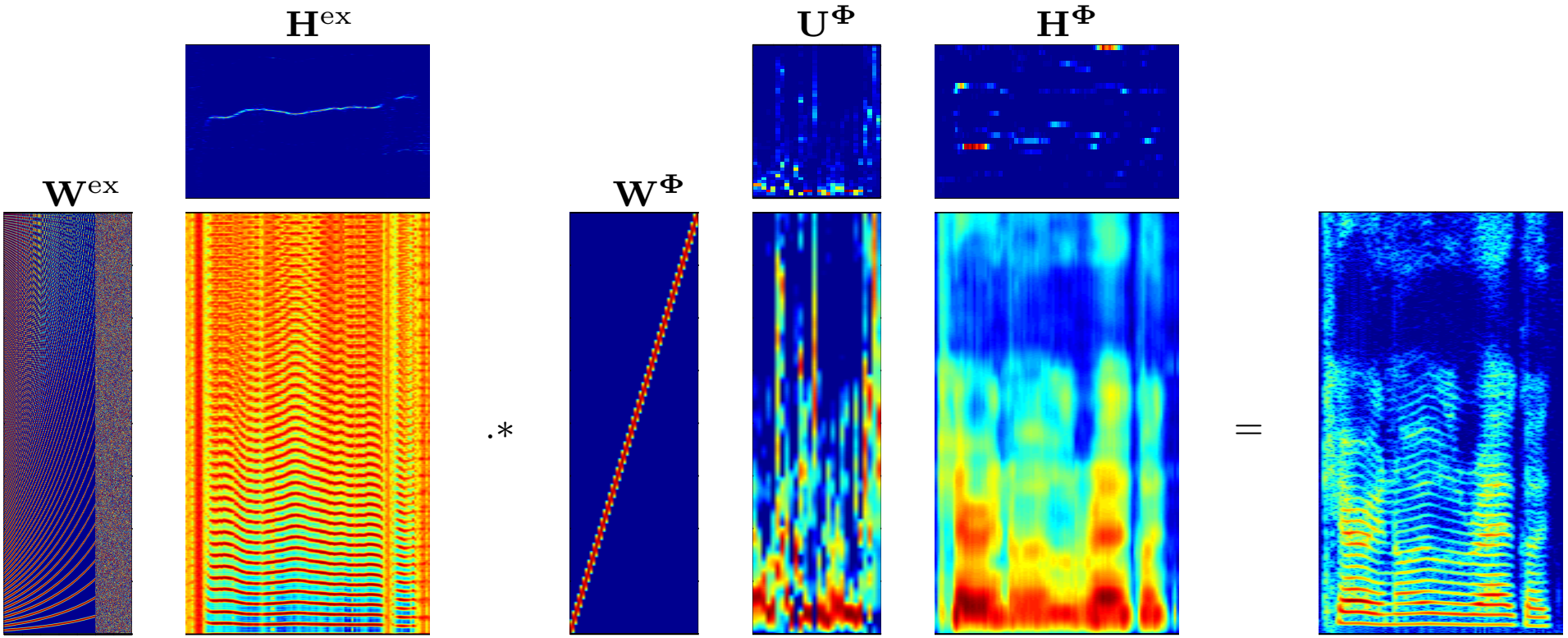


$(\mathbf{W}^{\text{ex}} \quad \mathbf{H}^{\text{ex}})$.* $(\widehat{\mathbf{W}}^{\Phi} \quad \mathbf{H}^{\Phi})$ = $\mathbf{V}$

Source/Filter Model

source * filter = sound

glottal source → vocal tract → voice

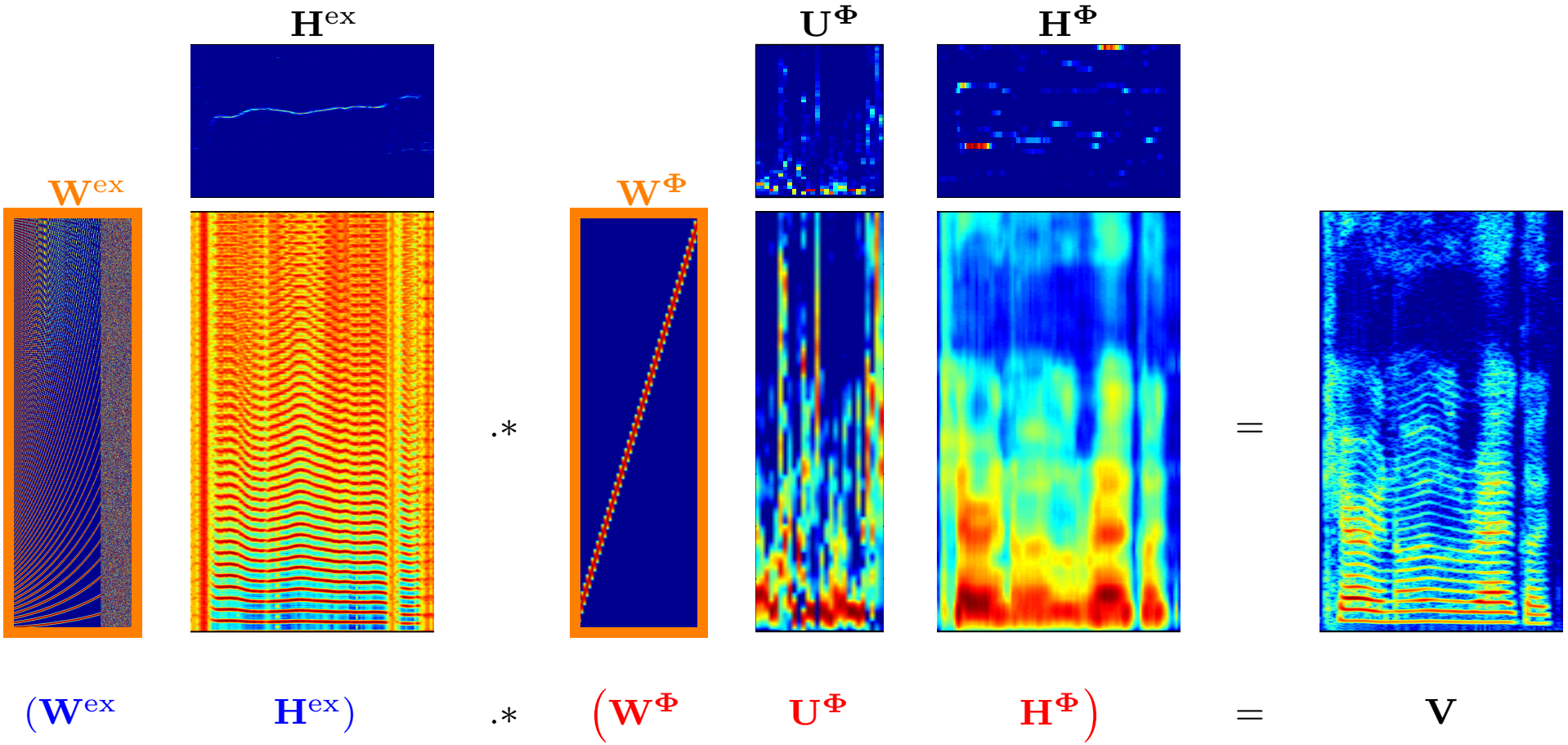


$(\mathbf{W}^{ex} \quad \mathbf{H}^{ex}) \; .* \; (\mathbf{W}^{\Phi} \quad \mathbf{U}^{\Phi} \quad \mathbf{H}^{\Phi}) = \mathbf{V}$

Source/Filter Model

source * filter = sound

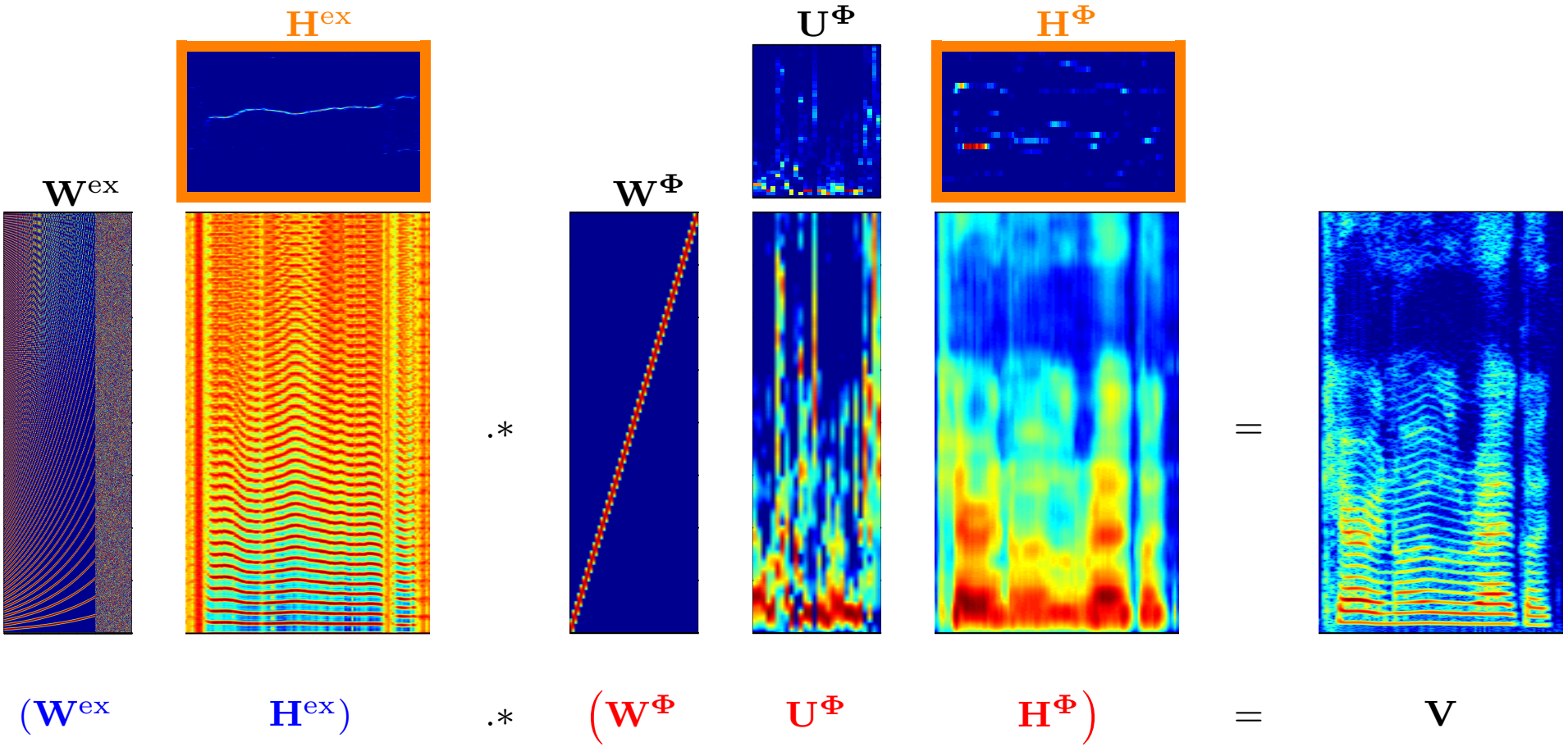
glottal source → vocal tract → voice



Source/Filter Model

source $*$ filter $=$ sound

glottal source → vocal tract → voice

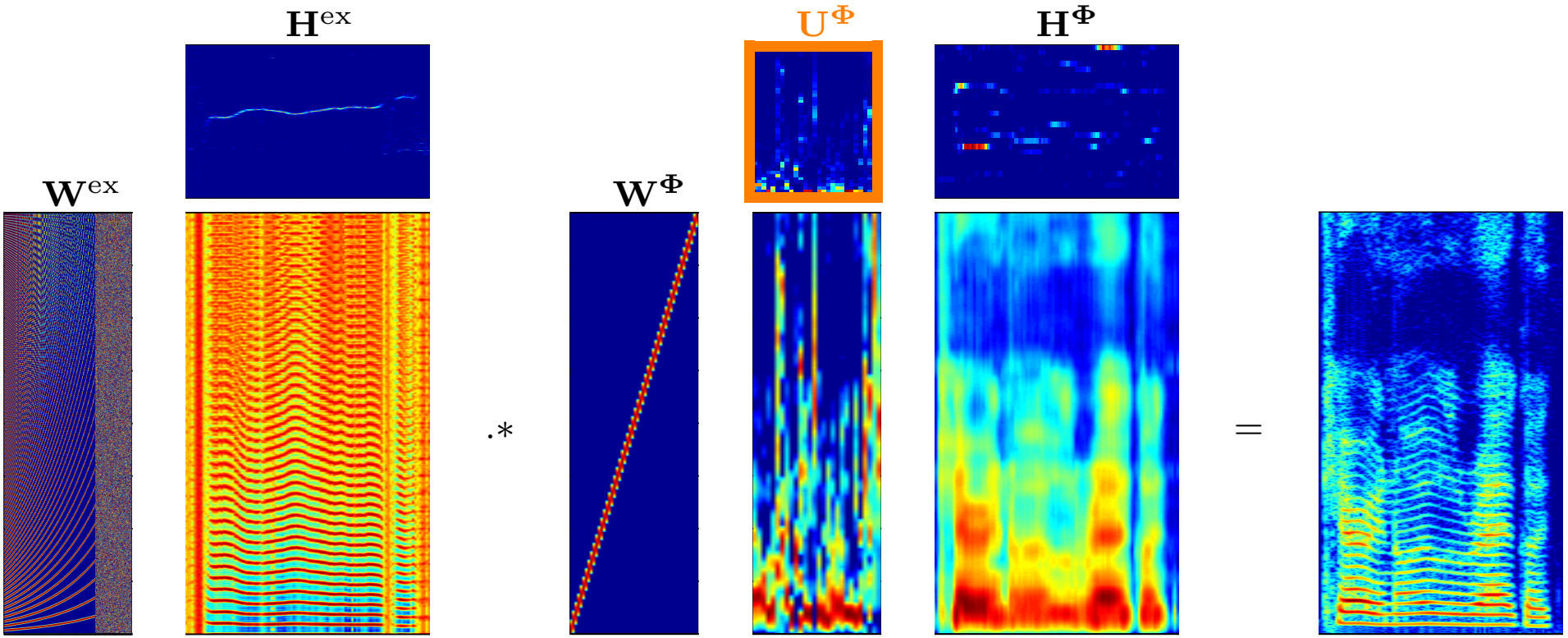


$(\mathbf{W}^{ex}\ \mathbf{H}^{ex})$.* $\left(\mathbf{W}^{\Phi}\ \mathbf{U}^{\Phi}\ \mathbf{H}^{\Phi}\right)$ $=$ $\mathbf{V}$

Source/Filter Model

source * filter = sound

glottal source → vocal tract → voice



$(\mathbf{W}^{\text{ex}} \quad \mathbf{H}^{\text{ex}})$.* $\left(\mathbf{W}^{\Phi} \quad \mathbf{U}^{\Phi} \quad \mathbf{H}^{\Phi}\right)$ = $\mathbf{V}$

Supervised NMF

Definition

Supervising the NMF consists of :

- estimating a set of dictionaries W_S
- from available databases of each audio source S

The dictionary matrix W is simply obtained by concatenating audio sources dictionaries W_S :

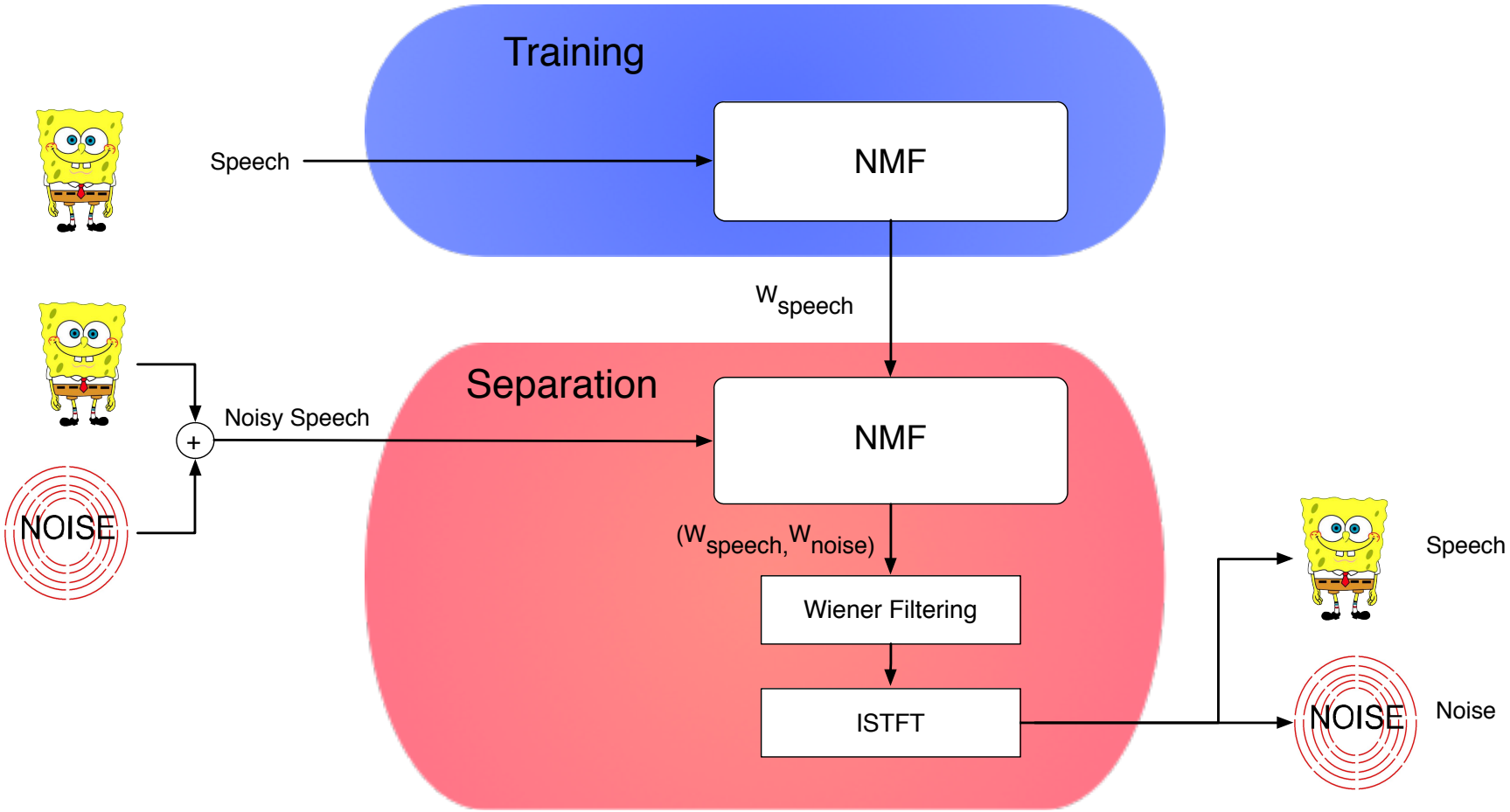
$$W = [W_{S_1}, W_{S_2}, \ldots, W_{S_S}] \tag{53}$$

Degree of Supervision

- Semi-supervised : some audio sources are known in advance and can be trained, the others remain unknown and free (e.g, speech/background separation)
- Supervised : all audio sources are known and can be trained (e.g., musical instruments separation)

Supervision is absolutely required in most audio source separation applications

Supervised NMF



References I

Benaroya, L. and Bimbot, F. (2003). Wiener based source separation with hmm/gmm using a single sensor. In *International Confererence on Independent Component Analysis Blind Source Separation (ICA)*, page 957–961.

Bertin, N. (2009). *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, Télécom ParisTech.

Durrieu, J.-L., Ozerov, A., Févotte, C., Richard, G., and David, B. (2009). Main instrument separation from stereophonic audio signals using a source/filter model. In *EUSIPCO*, pages 15–19.

Le Magoarou, L., Ozerov, A., and Duong, N. Q. (2014). Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. *Journal of Signal Processing Systems*, pages 1–15.

Le Roux, J., Hershey, J. R., and Weninger, F. (2015). Deep NMF for speech separation. In *Proceedings of the ICASSP 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 66–70.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

References II

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Mysore, G. J. and Smaragdis, P. (2012). A non-negative approach to language informed speech separation. In *Latent Variable Analysis and Signal Separation*, pages 356–363. Springer.

Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N., and Sagayama, S. (2010). Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence. *International Workshop on Machine Learning for Signal Processing, In Proc. IEEE*, 10 :283–288.

Sun, D. L. and Mysore, G. J. (2013). Universal speech models for speaker independent single channel source separation. In *Proceedings of the ICASSP 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 141–145.

Virtanen, T., Gemmeke, J. F., and Raj, B. (2013). Active-set newton algorithm for overcomplete non-negative representations of audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(11) :2277–2289.