Probabilistic modeling of non-stationary signals in time-frequency domain Application to music signals

Roland Badeau
Télécom ParisTech / CNRS LTCI

Currently visiting the CIRMMT
McGill University, Montreal

June 19, 2015

TELECOM ParisTech

Part I

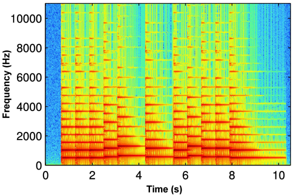
Introduction

Non-negative Matrix Factorization (NMF)

Musical score

Non-negative Matrix Factorization (NMF)

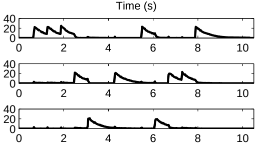
Musical score



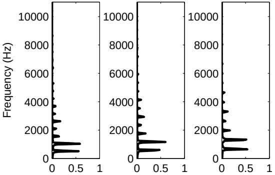
Spectrogram $\boldsymbol{V}$

Non-negative Matrix Factorization (NMF)

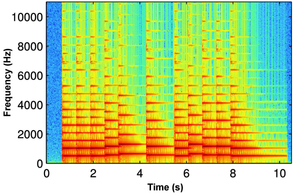
Musical score



Temporal activations ***H***



Spectral templates ***W***



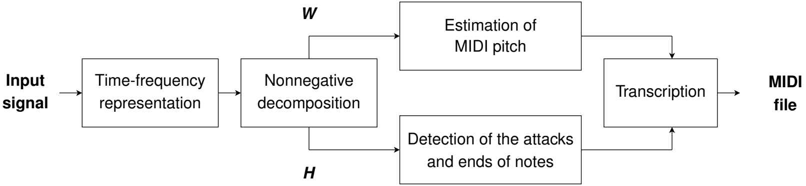
Spectrogram ***V***

Non-negative Matrix Factorization

- Factorization of a matrix $\boldsymbol{V} \in \mathbb{R}_+^{F \times T}$ as a product $\boldsymbol{V} \approx \boldsymbol{W}\,\boldsymbol{H}$
- Rank reduction: $\boldsymbol{W} \in \mathbb{R}_+^{F \times S}$ and $\boldsymbol{H} \in \mathbb{R}_+^{S \times T}$ where $S < \min(F, T)$
- Usual applications:
  - Image analysis, data mining, spectroscopy, finance, *etc.*
  - Audio signal processing:
    - Multi-pitch estimation, onset detection
    - Automatic music transcription
    - Musical instrument recognition
    - Source separation
    - Audio inpainting

NMF-based automatic transcription

- Algorithm

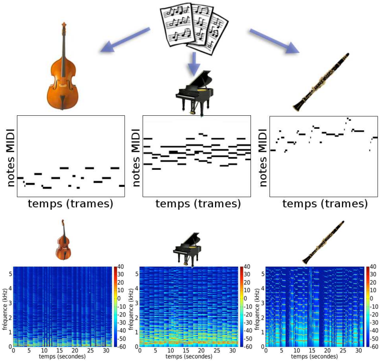


- Demo
  - Original signal (Liszt):
  - Transcribed signal:

N. Bertin, R. Badeau, and E. Vincent. "Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription". *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 18(3): 538–549, Mar 2010.

Score-based informed source separation

- Algorithm

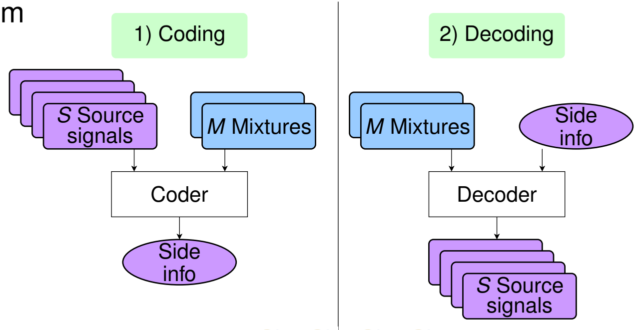


- Round Midnight (Thelonious Monk):

R. Hennequin, B. David, and R. Badeau. "Score informed audio source separation using a parametric model of non-negative spectrogram". In *ICASSP*, May 2011.

Watermarking-informed source separation

- Algorithm



- Mix Tape (Jim's Big Ego) :

A. Liutkus, R. Badeau, and G. Richard. "Informed source separation using multichannel NMF". In *LVA/ICA*, Sep 2010.

NMF probabilistic models

- Mixture models with (hidden) latent variables
  - + can exploit a priori knowledge
  - + can use well-known statistical inference techniques
- Probabilistic models of time-frequency distributions:
  - Magnitude-only models (phase is ignored)
    - Additive Gaussian noise [Schmidt, 2008],
    - Probabilistic Latent Component Analysis [Smaragdis, 2006],
    - Mixture of Poisson components [Virtanen, 2008],
  - Phase-aware models (theoretical ground for Wiener filtering)
    - Mixture of Gaussian components [Févotte, 2009],
    - Mixture of alpha-stable components [Liutkus & Badeau, 2015]

[1] A. Liutkus, R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *ICASSP*, Apr 2015, pp. 266–270.

Gaussian model [Févotte, 2009]

$$\underbrace{x_s(f,t)}_{\boldsymbol{X}_s} \sim \mathcal{N}\left(\mathbf{0},\ \underbrace{\sigma^2_{x_s}(f,t) = w_{fs}\, h_{st}}_{\hat{\boldsymbol{V}}_s}\right)$$

all time-frequency

bins are independent

Gaussian model [Févotte, 2009]

$$\underbrace{x_s(f,t)}_{\boldsymbol{X}_s} \sim \mathcal{N}\left(\mathbf{0},\; \underbrace{\sigma^2_{x_s}(f,t) = w_{fs}\, h_{st}}_{\hat{\boldsymbol{V}}_s}\right)$$

all time-frequency bins are independent

$$\Downarrow$$

$$\underbrace{y(f,t) = \sum_{s=0}^{S-1} x_s(f,t)}_{\boldsymbol{Y}} \sim \mathcal{N}\left(\mathbf{0},\; \underbrace{\hat{v}_{ft} = \sum_{s=0}^{S-1} \sigma^2_{x_s}(f,t)}_{\hat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}}\right)$$

w_{fs} ($\boldsymbol{W}$), h_{st} ($\boldsymbol{H}$)

Gaussian model [Févotte, 2009]

$$\underbrace{x_s(f,t)}_{\boldsymbol{X}_s} \sim \mathcal{N}\left(\mathbf{0},\ \underbrace{\sigma^2_{x_s}(f,t) = w_{fs}\, h_{st}}_{\hat{\boldsymbol{V}}_s}\right)$$

all time-frequency bins are independent

$$\Downarrow$$

$$\underbrace{y(f,t) = \sum_{s=0}^{S-1} x_s(f,t)}_{\boldsymbol{Y}} \sim \mathcal{N}\left(\mathbf{0},\ \underbrace{\hat{v}_{ft} = \sum_{s=0}^{S-1} \sigma^2_{x_s}(f,t)}_{\hat{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{H}}\right)$$

w_{fs}: $\boldsymbol{W}$ — h_{st}: $\boldsymbol{H}$

$$\max L(\boldsymbol{Y})$$

$$\Updownarrow$$

$$\min D_{IS}(\boldsymbol{V} = |\boldsymbol{Y}|^2 \,|\, \hat{\boldsymbol{V}})$$

Review of Itakura-Saito NMF (IS-NMF)

- Estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$:
  - The maximum likelihood estimate is obtained by minimizing the IS divergence between the spectrogram $\boldsymbol{V} = |\boldsymbol{Y}|^2$ and $\hat{\boldsymbol{V}}$
  - Methods: multiplicative update rules or SAGE algorithm
- Advantages of IS-NMF:
  - The MMSE estimation of $x_s(f,t)$ leads to Wiener filtering
  - The existence of phases is taken into account
- Drawbacks of IS-NMF:
  - $x_s(f,t)$ for all s, f, t are assumed uncorrelated
  - The values of phases in the STFT matrix $\boldsymbol{Y}$ are ignored

Questions

- Can we design time-frequency (TF) transforms such that the assumption of uncorrelated TF bins is best satisfied?
- For which class of stochastic processes can this assumption be satisfied? (TF bins of sinusoidal and impulse signals will always be correlated anyway)
- For stochastic processes whose TF correlations cannot be withdrawn, is it possible to extend the IS-NMF model in order to best take these correlations into account?
- What kind of improvement can we expect from modeling these correlations in applications such as source separation and audio inpainting?

Part II

Designing appropriate TF transforms

Preservation of whiteness (PW)

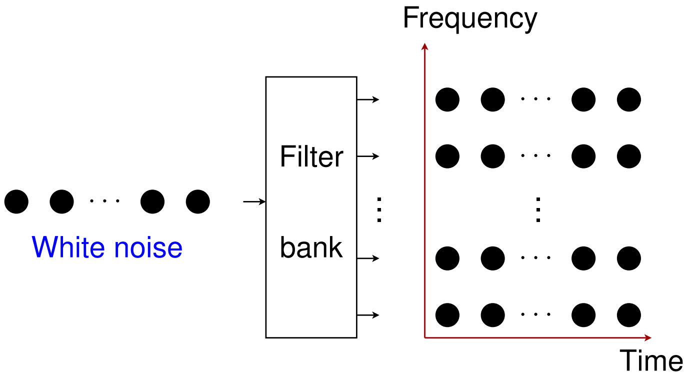


Figure: TF transform of a (proper complex) white noise

Preservation of whiteness (PW)

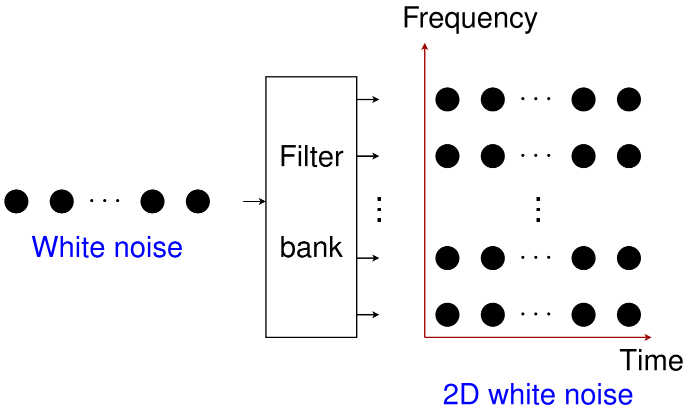


Figure: TF transform of a (proper complex) white noise

Preservation of whiteness (PW)

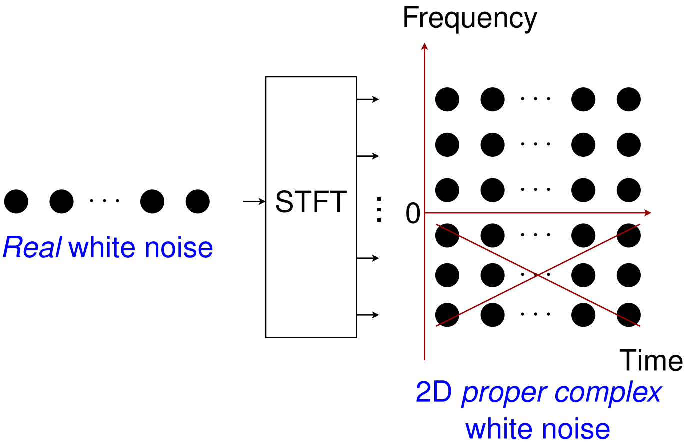


Figure: Complex TF transform of a real white noise

Perfect reconstruction (PR)

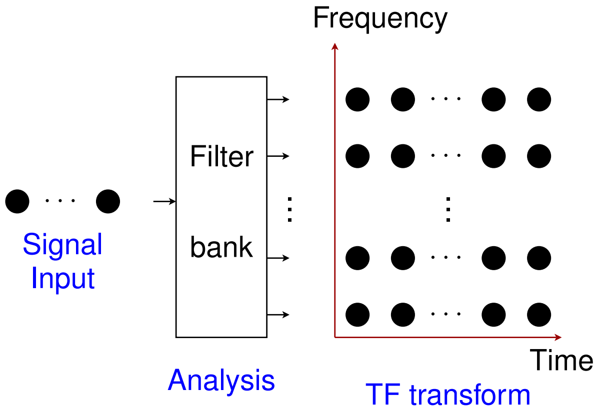


Figure: TF transform of a time series

Perfect reconstruction (PR)

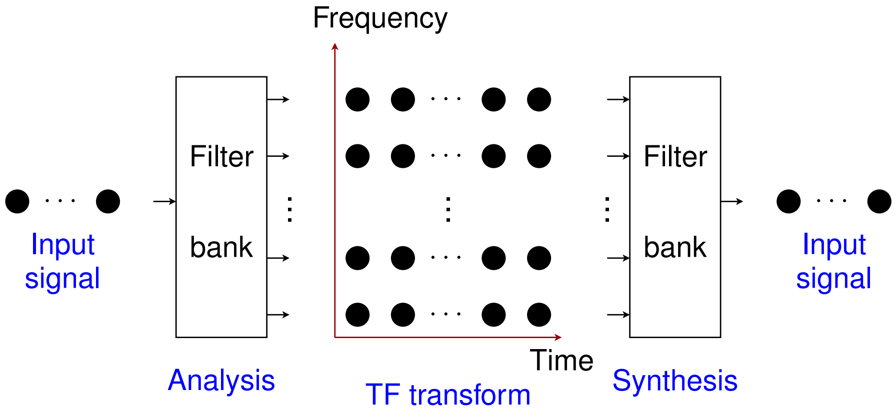


Figure: Perfect reconstruction filter bank

Solution of (PW) + (PR)

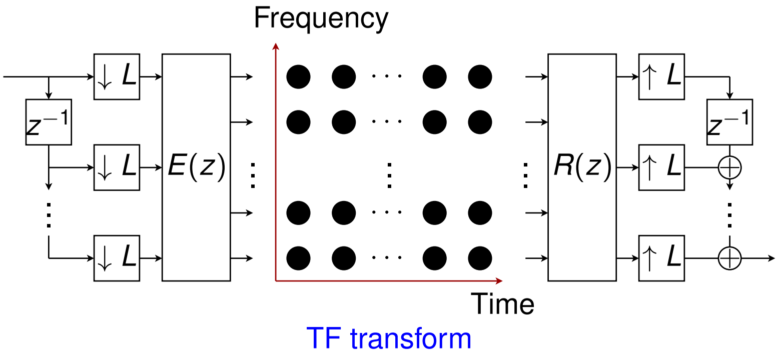


Figure: **Critically sampled paraunitary filter banks**: $R(z) = \widetilde{E}(z)$

Examples of solutions

- *Real TF transform of real signals:* **MDCT filter banks**

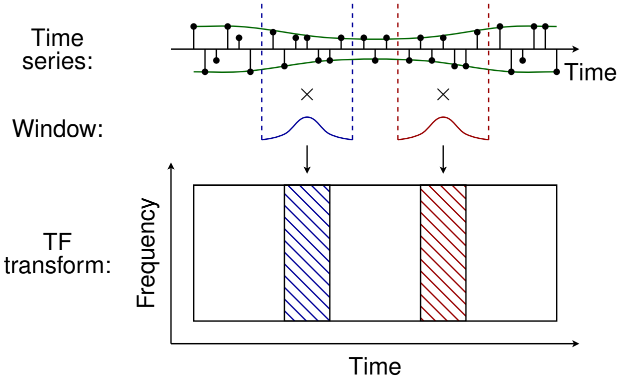
$$X_{f,t} = \sum_{n\in\mathbb{Z}} w_n \, x_{Ft-n} \cos\left(\frac{\pi}{F}\left(f+\frac{1}{2}\right)\left(n+\frac{F+1}{2}\right)\right)$$

- *Complex TF transform of complex signals:* PR critically decimated **GDFT filter banks** with matched analysis and synthesis filters:

$$X_{f,t} = \sum_{n\in\mathbb{Z}} w_n \, x_{Ft-n} \exp\left(+\frac{i2\pi}{F}(f+\phi)(n+\tau)\right)$$

- *Complex TF transform of real signals:* same **GDFT filter banks**, with F even and $\phi = \frac{1}{2}$

TF transform of uncorrelated time samples



Uncorrelated time samples

Non-overlapping time frames

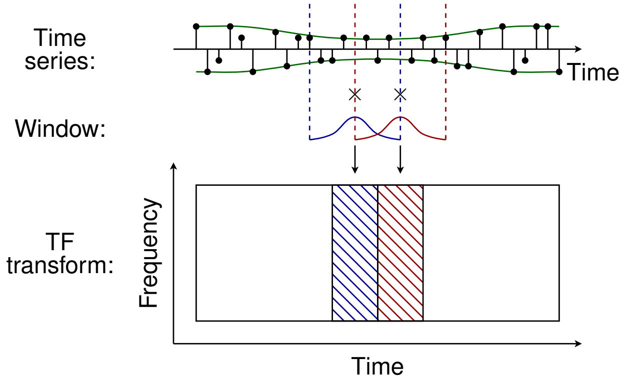
$\Downarrow$

Non-adjacent columns

are uncorrelated

Figure: TF transform of uncorrelated time samples

TF transform of uncorrelated time samples



Uncorrelated time samples

Slowly varying power

Preservation of whiteness

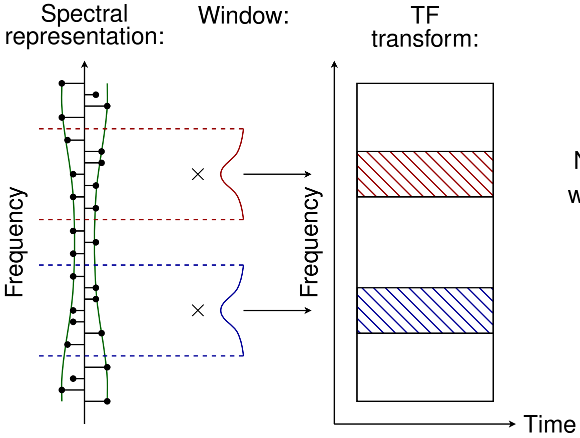
$\Downarrow$

Adjacent columns

are approx. uncorrelated

Figure: TF transform of uncorrelated time samples with slowly varying power

TF transform of a WSS process



WSS process

Non-overlapping sub-bands

with high stop-band rejection

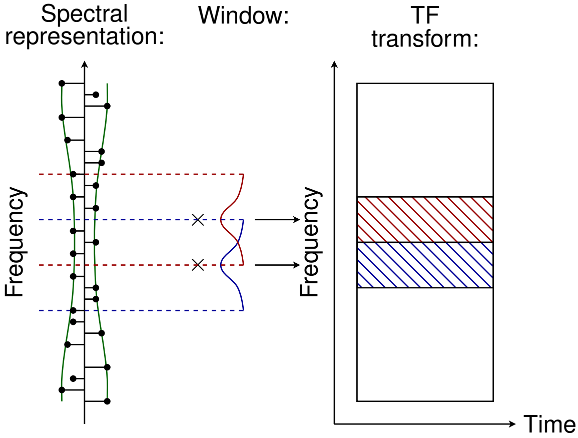
⇓

Non-adjacent rows

are approx. uncorrelated

Figure: TF transform of a WSS process with high stop-band rejection

TF transform of a WSS process



WSS process

Smooth PSD

Preservation of whiteness

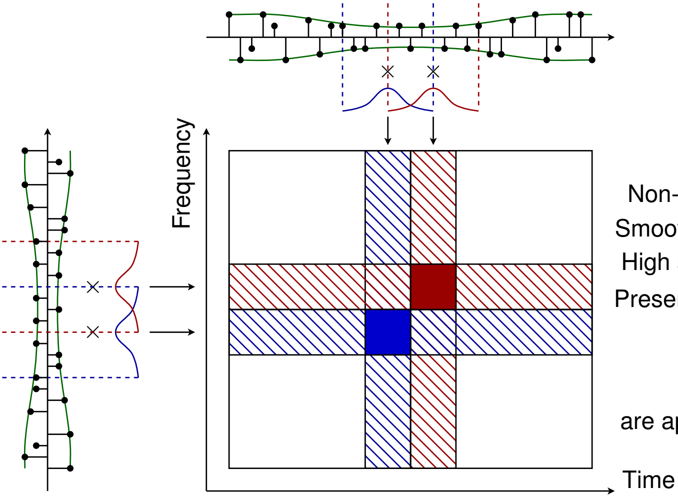
$\Downarrow$

Adjacent rows

are approx. uncorrelated

Figure: TF transform of a WSS process with smooth PSD

TF transform of a nonstationary process



Non-stationary process
Smooth power TF density
High stop-band rejection
Preservation of whiteness

$\Downarrow$

All TF bins
are approx. uncorrelated

Figure: TF transform of nonstationary signal with smooth TF density

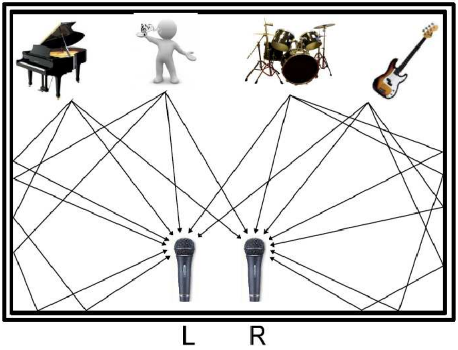
Take-home message

- Advantages of whiteness-preserving TF transforms:
  - **The assumption of uncorrelated TF bins holds approximately** for a wide range of nonstationary signals with smooth TF density.
  - **No need to care for the *consistency* of the TF transform**, since it is bijective (no redundancy in the TF domain).
  - Preliminary results, in a source separation application involving NMF modeling and Wiener filtering, showed **no performance loss** when using an MDCT instead of an STFT with 75% overlap
- Drawbacks:
  - Designing paraunitary STFT filter banks is constrained: solutions involve non-overlapping rectangular windows or recursive filters.
  - The assumption of uncorrelated TF bins does not hold for sinusoids and impulses: such correlations still need to be properly modeled.

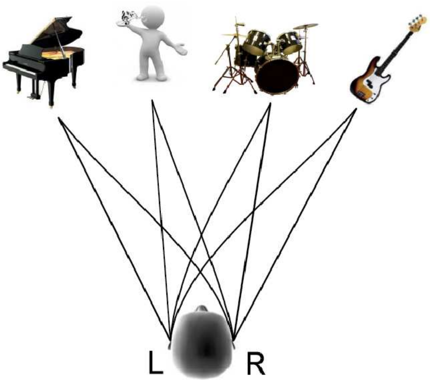
Part III

Modeling correlations in the TF domain

Linear convolutive mixtures modeling



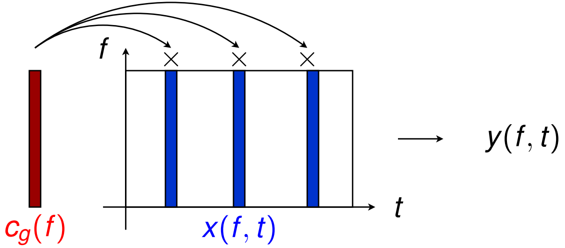
(a) Convolutive mixture.



(b) Binaural mixture.

Convolution in TF domain

- **Purpose**: implement $y(n) = (g * x)(n)$ in TF domain
- **Standard approach**: column-wise multiplication of the STFT $x(f,t)$ by the frequency response $c_g(f)$ of filter $g(n)$



- **Advantage**: $y(f,t)$ are uncorrelated if $x(f,t)$ are uncorrelated
- **Drawbacks**: Approximation, holds if $g(n)$ is much shorter than time frames (unrealistic). Approach restricted to the STFT.

Convolution in TF domain

- **Purpose**: implement $y(n) = (g * x)(n)$ in TF domain
- **Problem**: find transformation $\mathcal{T}_{\mathrm{TF}}$ in Figure 1 such that the output is $y(n)$ when the input is $x(n)$

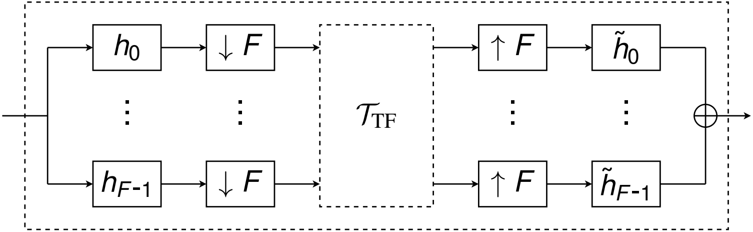


Fig. 1: Applying a TF transformation to a TD signal

Convolution in TF domain

Solution: $\mathcal{T}_{\mathrm{TF}}$ is represented in the larger frame in Figure 2, where the input is $x(f,t)$, the output is $y(f,t)$, and $\mathcal{T}_{\mathrm{TD}}$ is the convolution by $g(n)$

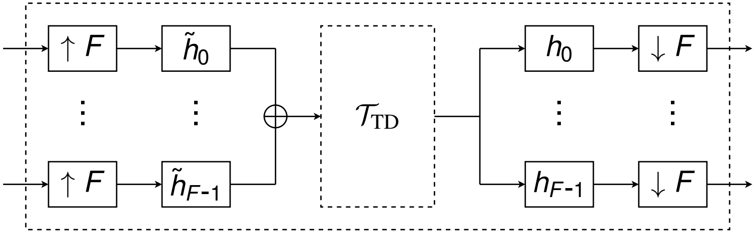
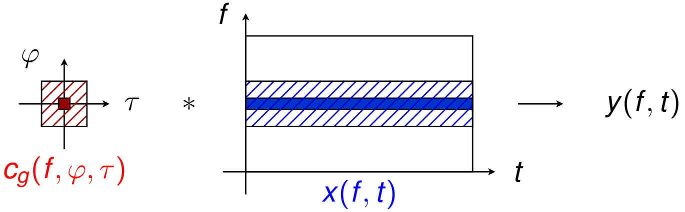


Fig. 2: Applying a TD transformation to TF data

Convolution in TF domain

- This solution can be implemented as a 2D filter:

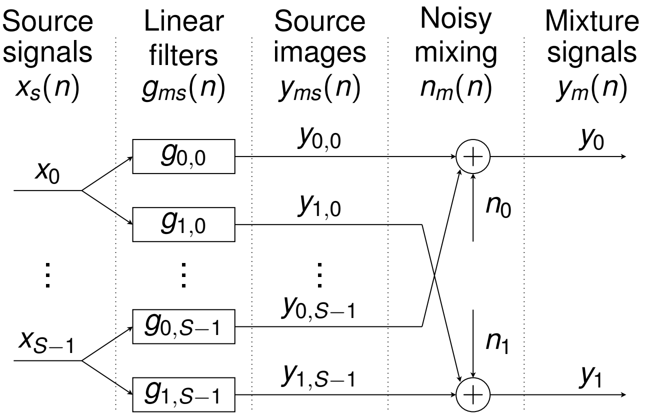


- **ARMA parametrisation**: if $g(n)$ is a causal and stable recursive (ARMA) filter then $c_g(f, \varphi, \tau)$ can be parametrised as

$$\forall \varphi, \tau, f,\ a_g(f - \varphi, \tau) \underset{\tau}{*} c_g(f, \varphi, \tau) = b_g(f, \varphi, \tau)$$

[2] R. Badeau and M.D. Plumbley, "Probabilistic Time-Frequency Source-Filter Decomposition of Non-Stationary Signals," in *EUSIPCO*, Sep 2013.

Time-domain mixing model



Example of a stereophonic setting ($M = 2$)

Multichannel HR-NMF model

- The TF transforms of the source signals $x_s(f,t)$ follow a regular IS-NMF model: $x_s(f,t) \sim \mathcal{N}(0, \sum_k w^s_{fk} h^s_{kt})$
- ARMA filtering is implemented via a state-space representation:

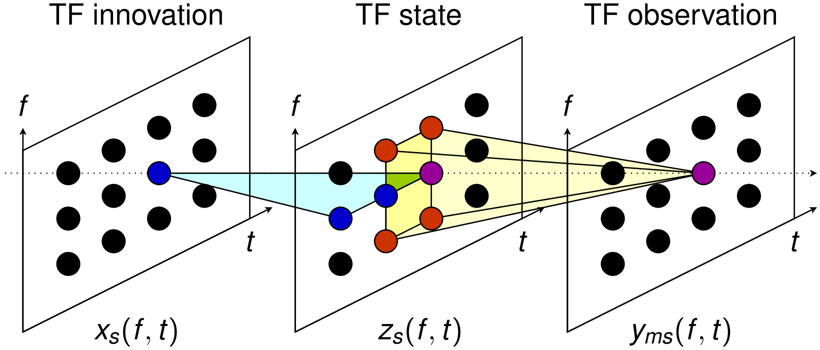
$$z_s(f,t) = x_s(f,t) - \sum_{\tau=1}^{Q_a} a_s(f,\tau) z_s(f,t-\tau)$$

$$y_{ms}(f,t) = \sum_{\varphi=-P_b}^{P_b} \sum_{\tau=0}^{Q_b} b_{ms}(f,\varphi,\tau)\, z_s(f-\varphi, t-\tau)$$

- Output: $y_m(f,t) = n_m(f,t) + \sum_{s=0}^{S-1} y_{ms}(f,t)$ with $n_m(f,t) \sim \mathcal{N}(0, \sigma_n^2)$

[3] R. Badeau and M.D. Plumbley, "Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain" in *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 22, no. 11, Nov 2014, pp. 1670–1680.

Dependency graph

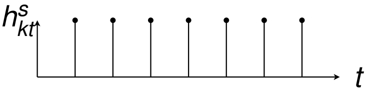


Dependency graph in the TF domain

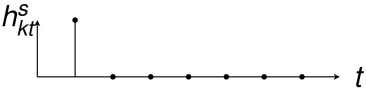
Particular cases

The HRNMF model encompasses:

- Multichanel NMF [Ozerov & Févotte, 2010] (if $Q_a = Q_b = P_b = 0$)
- ARMA processes (if $K = 1$ and h^s_{kt} is flat)



- Mixtures of damped sinusoids (if $K = 1$ and h^s_{kt} is an impulse)



Estimation of HR-NMF

Various approaches (initially developed for the mono $M = 1$ case)

- **EM algorithm** with Kalman filtering [Badeau, 2011]: slow convergence, high computational complexity
- **Multiplicative updates** [Badeau & Ozerov, 2013]: fast convergence but numerical stability issues
- **Variational EM algorithm** [Badeau & Drémeau, 2013] : low computational complexity

[4] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *WASPAA*, Oct 2011, pp. 253–256.
[5] R. Badeau and A. Ozerov, "Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain," in *EUSIPCO*, Sep 2013.
[6] R. Badeau and A. Drémeau, "Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *ICASSP*, May 2013.

Variational EM algorithm

- **Goal**: estimate the parameter θ of a probabilistic model involving observations y and latent variables z
- **Idea**: $p(z|y;\theta)$ is approximated by a distribution q
  Decomposition of log-likelihood $L(\theta) = \ln(p(y;\theta))$:
  $$L(\theta) = D_{\mathrm{KL}}(q||p(z|y;\theta)) + \mathcal{L}(q;\theta), \text{ where}$$
  - $D_{\mathrm{KL}}(q||p(z|y;\theta)) = \left\langle \ln\left(\frac{q(z)}{p(z|y;\theta)}\right)\right\rangle_q$ (KL divergence)
  - $\mathcal{L}(q;\theta) = \left\langle \ln\left(\frac{p(y,z;\theta)}{q(z)}\right)\right\rangle_q$ (variational free energy)

  Since $D_{\mathrm{KL}} \geq 0$, $\mathcal{L}(q;\theta)$ is a lower bound of $L(\theta)$
- **Method**: maximize $\mathcal{L}(q;\theta)$: at each iteration i,
  - E-step (update q): $q^\star = \underset{q\in\mathcal{F}}{\operatorname{argmax}}\, \mathcal{L}(q;\theta_{i-1})$
  - M-step (update θ): $\theta_i = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q^\star;\theta)$

Variational EM for multichannel HR-NMF

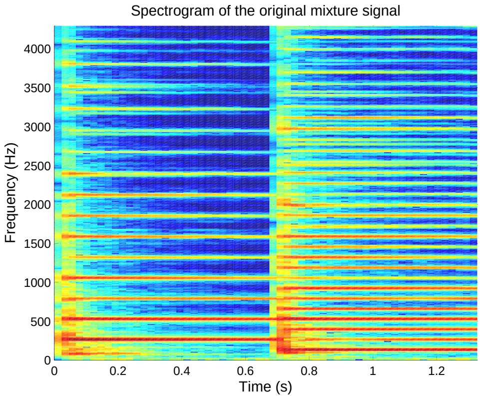
- Parameters: $\theta = \{a_s(f,\tau), b_{ms}(f,\varphi,\tau), \sigma_n^2, w_{fk}^s, h_{kt}^s)\}$
- Mean field approximation: $q(z) = \prod_{s,f,t} q_{sft}(z_s(f,t))$
- Complexity: $4MSFT(1+2P_b)(1+\max(Q_b,Q_a))$ (linear w.r.t. all model dimensions)
- Parallel implementation
- Application to real audio data:
  - Always converges to a relevant solution when $S=1$
  - Needs proper initialization or semi-supervised learning when $S>1$

[3] R. Badeau and M.D. Plumbley, "Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain" in *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 22, no. 11, Nov 2014, pp. 1670–1680.

Part IV

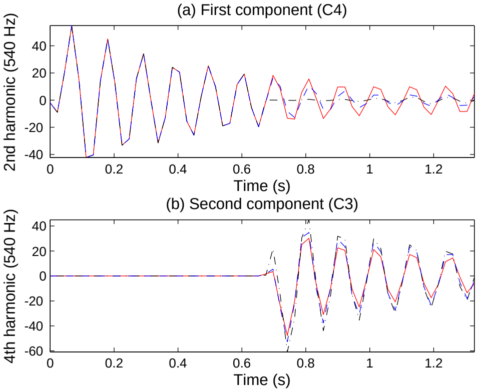
Application to piano sounds

Application to piano tones



Spectrogram of the input piano sound (C4 + C3)
($\mathbb{F} = \mathbb{C}$, $S = 2$, $M = 1$, $F_s = 8600$kHz)

Source separation



IS-NMF:

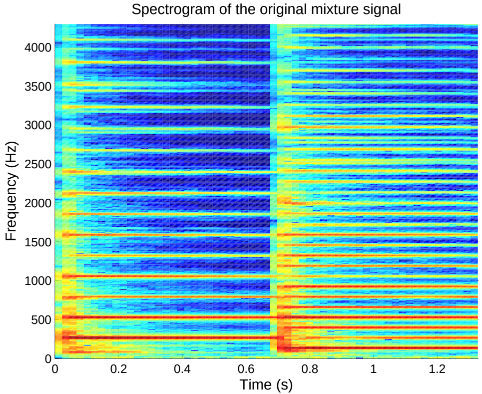
HR-NMF:

IS-NMF:

HR-NMF:

Separation of two sinusoidal components (real parts of STFT subband signals)

Audio inpainting (mono)



C4+C3:

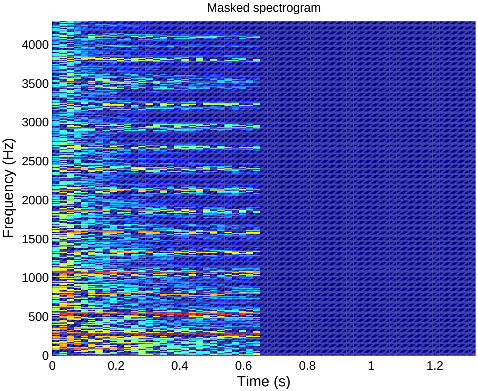
C4 alone:

IS-NMF:

HR-NMF:

Spectrogram of the input piano sound (C4 + C3)

Audio inpainting (mono)



Masked spectrogram of the input piano sound

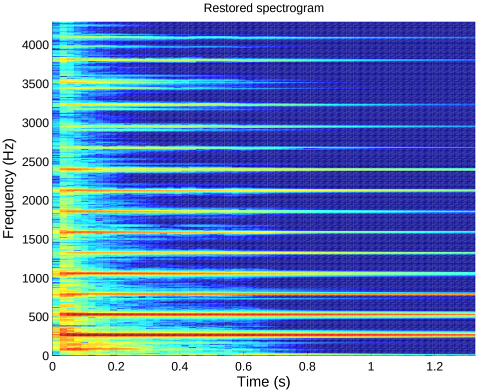
C4+C3:

C4 alone:

IS-NMF:

HR-NMF:

Audio inpainting (mono)



C4+C3:

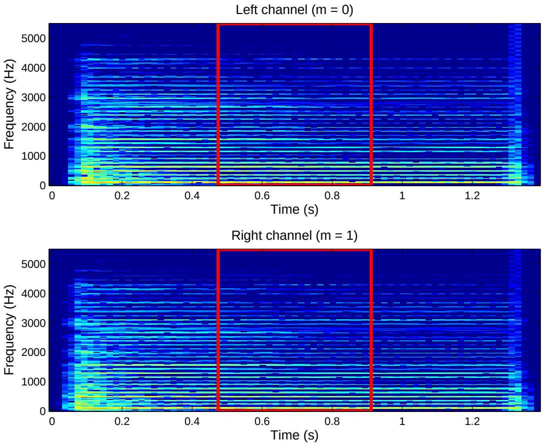
C4 alone:

IS-NMF:

HR-NMF:

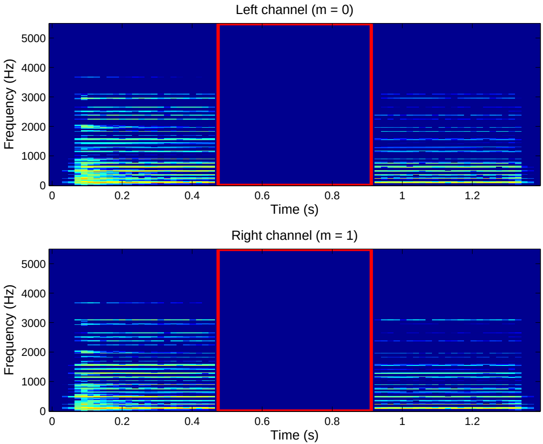
Recovery of the full C4 piano tone

Audio inpainting (stereo)



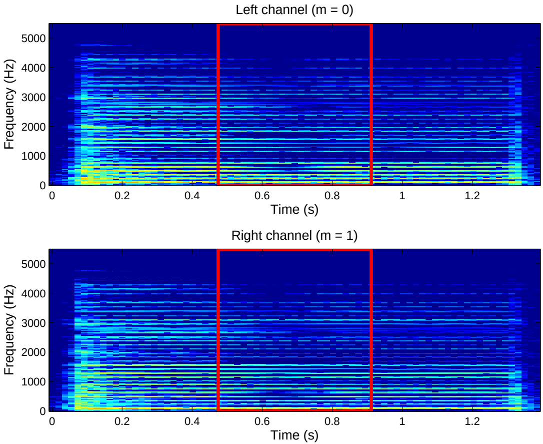
Input stereo piano MDCT $y_m(f,t)$ ($\mathbb{F}=\mathbb{R}$, $S=1$, $M=2$, $F_s=11$kHz)

Audio inpainting (stereo)



Stereo image $\hat{y}_{ms}(f,t)$ estimated with $Q_a = Q_b = P_b = 0$

Audio inpainting (stereo)



Stereo image $\hat{y}_{ms}(f, t)$ estimated with $Q_a = 2$, $Q_b = 3$, $P_b = 1$

Overview of the HR-NMF model

- Able to accurately represent multichannel, underdetermined mixtures of sound sources in presence of reverberation
- Achieved via an accurate TF implementation of ARMA filtering
- Compatible with any filter bank (either real or complex)
- Accounts for phases and correlations over time and frequency
- Able to separate overlapping sinusoids within the same frequency band (high spectral resolution)
- Able to restore missing observations (synthesis capability)

Part V

Source separation benchmark

Source separation benchmark

Benchmark of several NMF-based methods involving phase recovery:

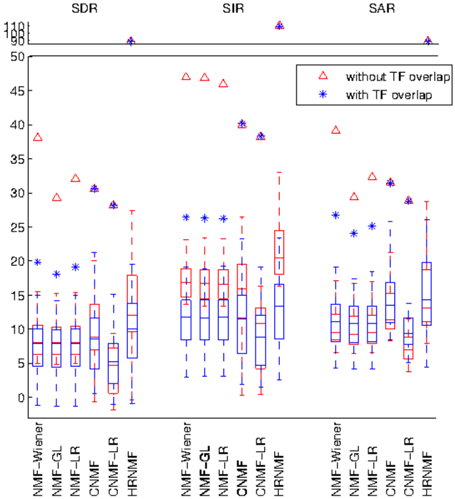
- **NMF-Wiener**: Wiener filtering with NMF models of spectrograms
- Phase reconstruction based on spectrogram consistency:
  - **NMF-GL**: NMF models with GL algorithm [Griffin & Lim, 1984]
  - **NMF-LR**: NMF models with LR algorithm [Leroux, 2008]
- Complex NMF (CNMF) estimation of the STFTs of the sources:
  - **CNMF**: without any phase constraint [Kameoka, 2009]
  - **CNMF-LR**: with consistency phase constraints [Leroux, 2009]
- **HR-NMF** (with a reduced frequency resolution in order to compensate for the extra ARMA parameters)

[7] P. Magron, R. Badeau, B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *ICASSP*, Apr 2015, pp. 81–85.

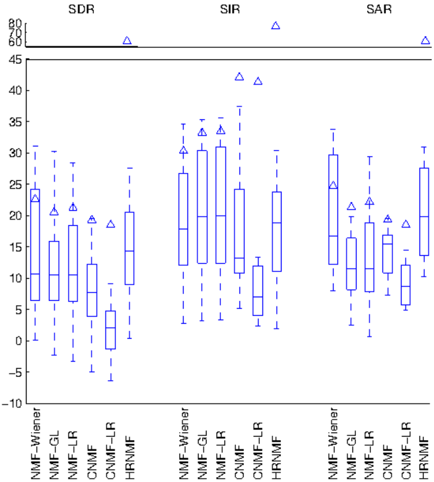
Source separation benchmark

- Datasets:
  - Synthetic mixtures of two harmonic signals with additive white noise
  - Piano notes mixtures from the MAPS database [Emiya, 2010]
  - MIDI audio excerpt (bass and guitar)
- Blind vs. Oracle approaches:
  - **Blind**: model parameters are estimated from the mixtures
  - **Oracle**: model parameters are learned from the isolated sources
- Evaluation criteria: BSS EVAL Toolbox [Vincent, 2006]
  - **SDR**: Source to Distortion Ratio
  - **SIR**: Source to Interference Ratio
  - **SAR**: Source to Artifact Ratio

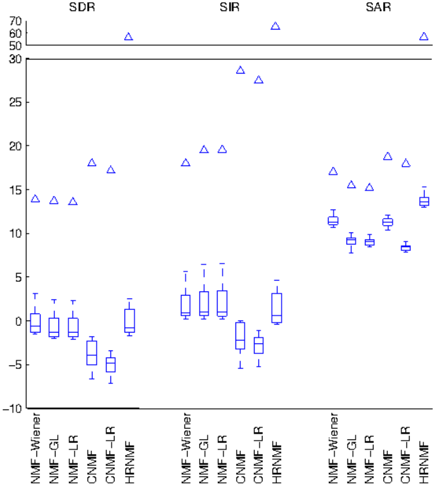
Synthetic mixtures of two harmonic signals



Piano notes mixtures



MIDI audio excerpt



Oracle separation of a MIDI audio excerpt

Mix	NMF-Wiener	HRNMF
Bass		
Guitar		
Keyboard		

Conclusions of the benchmark

- Spectrogram consistency may not be relevant for audio quality
- Oracle results show the potential of the HR-NMF model in source separation applications
- Blind results show the difficulty of estimating this model without a proper initialization
- Solutions could involve:
  - Semi-supervised learning,
  - A priori information (harmonicity, smoothness, sparsity...),
  - New estimation methods (MCMC, belief propagation, high resolution methods,...)

Part VI

Conclusion

Conclusions

- Take-home message:
  - Possibility of designing TF transforms that better fit the assumption of uncorrelated TF bins
  - Importance of modeling phases and correlations in the TF domain
- Outlooks of the HRNMF model:
  - Introduce high temporal resolution (to model sharp transients)
  - 2D linear prediction of TF state $z_s(f,t)$) (to model vibrato, chirps)
  - Correlations between components (to model sympathetic vibration)
  - Non-stationary filters (to model attack-decay-sustain-release)
- Applications:
  - Source coding, source separation, audio inpainting...

Thank you!

Bibliography

A. Liutkus and R. Badeau, “Generalized wiener filtering with fractional power spectrograms,” in *ICASSP*, Apr. 2015, pp. 266–270.

R. Badeau and M. D. Plumbley, “Probabilistic time-frequency source-filter decomposition of non-stationary signals,” in *EUSIPCO*, Sep. 2013.

——, “Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 11, Nov. 2014, 1670–1680.

R. Badeau, “Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF),” in *WASPAA*, Oct. 2011, pp. 253–256.

R. Badeau and A. Ozerov, “Multiplicative updates for modeling mixtures of non-stationary signals in the time-frequency domain,” in *EUSIPCO*, Sep. 2013.

R. Badeau and A. Drémeau, “Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF),” in *ICASSP*, May 2013, pp. 6171–6175.

P. Magron, R. Badeau, and B. David, “Phase recovery in NMF for audio source separation: an insightful benchmark,” in *ICASSP*, Apr. 2015, pp. 81–85.