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
Part I

Introduction
Non-negative Matrix Factorization (NMF)

Musical score
Non-negative Matrix Factorization (NMF)

Musical score

Spectrogram V
Non-negative Matrix Factorization (NMF)

Musical score

Temporal activations $H$

Spectral templates $W$

Spectrogram $V$
Non-negative Matrix Factorization

- Factorization of a matrix $V \in \mathbb{R}^{F \times T}_+$ as a product $V \approx WH$

- Rank reduction: $W \in \mathbb{R}^{F \times S}_+$ and $H \in \mathbb{R}^{S \times T}_+$ 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**

- **Input signal**
- **Time-frequency representation**
- **Nonnegative decomposition**
  - Estimation of MIDI pitch
  - Detection of the attacks and ends of notes
- **Transcription**
- **MIDI file**

**Demo**

- Original signal (Liszt):
- Transcribed signal:

Score-based informed source separation

- Algorithm

Round Midnight (Thelonious Monk):

Algorithm

1) Coding

- S Source signals
- M Mixtures

Coder

Side info

2) Decoding

- M Mixtures

Decoder

Side info

- S Source signals

Mix Tape (Jim’s Big Ego):

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]

Gaussian model [Févotte, 2009]

\[ x_s(f, t) \sim \mathcal{N}(0, w_{fs}) \]

\[ \sigma_{x_s}^2(f, t) = w_{fs} h_{st} \]

All time-frequency bins are independent.
Gaussian model [Févotte, 2009]

\[
Y(f, t) = \sum_{s=0}^{S-1} X_s(f, t) \sim \mathcal{N}(0, W_{fs})
\]

\[
x_s(f, t) \sim \mathcal{N}(0, w_{fs} h_{st})
\]

\[
y(f, t) = \sum_{s=0}^{S-1} x_s(f, t) \sim \mathcal{N}(0, W_{fs})
\]

\[
\hat{v}_s = \sum_{s=0}^{S-1} \sigma_{x_s}^2(f, t)
\]

\[
\hat{v}_{ft} = \sum_{s=0}^{S-1} \sigma_{x_s}^2(f, t)
\]

all time-frequency bins are independent
Gaussian model [Févotte, 2009]

\[ y(f, t) = \sum_{s=0}^{S-1} x_s(f, t) \sim \mathcal{N}(0, W_{fs}) \]

\[ x_s(f, t) \sim \mathcal{N}(0, W_{fs}) \quad \sigma_{x_s}^2(f, t) = w_{fs} h_{st} \]

\[ \hat{V}_s \]

\[ \hat{V} = WH \]

\[ \max L(Y) \quad \Downarrow \]

\[ \min D_{ls}(V = |Y|^2 | \hat{V}) \]

all time-frequency bins are independent
Review of Itakura-Saito NMF (IS-NMF)

- Estimation of $W$ and $H$:
  - The maximum likelihood estimate is obtained by minimizing the IS divergence between the spectrogram $V = |Y|^2$ and $\hat{V}$
  - 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 $Y$ 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)

White noise

Filter

bank

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)

Real white noise

2D proper complex white noise

Figure: Complex TF transform of a real white noise
Perfect reconstruction (PR)

Figure: TF transform of a time series
Perfect reconstruction (PR)

Input signal

Filter bank

Analysis

Frequency

TF transform

Synthesis

Filter bank

Time

Figure: Perfect reconstruction filter bank
Solution of (PW) + (PR)

\[
E(z) - 1 + R(z) = \tilde{E}(z)
\]

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

Time series:

Window:

TF transform:

Uncorrelated time samples
Non-overlapping time frames
Non-adjacent columns are uncorrelated

Figure: TF transform of uncorrelated time samples
TF transform of uncorrelated time samples

Figure: TF transform of uncorrelated time samples with slowly varying power
TF transform of a WSS process with high stop-band rejection

Non-overlapping sub-bands

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
⇓
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

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

![Diagram](image)

**Fig. 1**: Applying a TF transformation to a TD signal
Solution: $\mathcal{T}_{TF}$ is represented in the larger frame in Figure 2, where the input is $x(f, t)$, the output is $y(f, t)$, and $\mathcal{T}_{TD}$ is the convolution by $g(n)$.

![Diagram of Convolution in TF domain](image-url)
Convolution in TF domain

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

\[ f \ast \tau \phi \] \( c_g(f, \varphi, \tau) \)

\[ x(f, t) \rightarrow y(f, t) \]

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

\[ \forall \varphi, \tau, f, a_g(f - \varphi, \tau) \ast c_g(f, \varphi, \tau) = b_g(f, \varphi, \tau) \]

Time-domain mixing model

Source signals $x_s(n)$
Linear filters $g_{ms}(n)$
Source images $y_{ms}(n)$
Noisy mixing $n_m(n)$
Mixture signals $y_m(n)$

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: $x_s(f, t) \sim \mathcal{N}(0, \sum_k w_{fk}^s h_{kt}^s)$

- 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 $n_m(f, t) \sim \mathcal{N}(0, \sigma_n^2)$

Dependency graph in the TF domain

- **TF innovation**: \( x_s(f, t) \)
- **TF state**: \( z_s(f, t) \)
- **TF observation**: \( y_{ms}(f, t) \)
Particular cases

The HRNMF model encompasses:

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

\[ h^s_{kt} \]

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

\[ h^s_{kt} \]
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

---


Variational EM algorithm

- **Goal**: estimate the parameter $\theta$ of a probabilistic model involving observations $y$ and latent variables $z$

- **Idea**: $\rho(z|y; \theta)$ is approximated by a distribution $q$

Decomposition of log-likelihood $L(\theta) = \ln(\rho(y; \theta))$:

$$L(\theta) = D_{KL}(q||\rho(z|y; \theta)) + \mathcal{L}(q; \theta),$$

where

- $D_{KL}(q||\rho(z|y; \theta)) = \langle \ln \left( \frac{q(z)}{\rho(z|y; \theta)} \right) \rangle_q$ (KL divergence)
- $\mathcal{L}(q; \theta) = \langle \ln \left( \frac{\rho(y,z; \theta)}{q(z)} \right) \rangle_q$ (variational free energy)

Since $D_{KL} \geq 0$, $\mathcal{L}(q; \theta)$ is a lower bound of $L(\theta)$

- **Method**: maximize $\mathcal{L}(q; \theta)$: at each iteration $i$,
  - E-step (update $q$): $q^* = \arg\max_{q \in \mathcal{F}} \mathcal{L}(q; \theta_{i-1})$
  - M-step (update $\theta$): $\theta_i = \arg\max_{\theta} \mathcal{L}(q^*; \theta)$
Variational EM for multichannel HR-NMF

Parameters: $\theta = \{ a_s(f, \tau), b_{ms}(f, \varphi, \tau), \sigma_n^2, w_{sk}, h_{kt}^s \}$

Mean field approximation: $q(z) = \prod_{s,f,t} q_{sft}(z_s(f, t))$

Complexity: $4 \text{MSFT} (1 + 2P_b)(1 + \max(Q_b, Q_a))$ (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 $S > 1$

Part IV

Application to piano sounds
Application to piano tones

Spectrogram of the original mixture signal

Spectrogram of the input piano sound (C4 + C3)

(\(\mathbb{F} = \mathbb{C}, \ S = 2, \ M = 1, \ F_s = 8600\text{kHz}\))
Source separation

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

Spectrogram of the original mixture signal

Spectrogram of the input piano sound (C4 + C3)

C4+C3:
C4 alone:
IS-NMF:
HR-NMF:
Audio inpainting (mono)

Masked spectrogram

C4+C3:
C4 alone:
IS-NMF:
HR-NMF:

Masked spectrogram of the input piano sound
Audio inpainting (mono)

Restored spectrogram

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)

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

![Graph showing SDR, SIR, and SAR metrics for different methods with and without TF overlap.](image-url)
Piano notes mixtures

![Graph showing SDR, SIR, and SAR performance metrics for different methods such as NMF-Wiener, NMF-GL, NMF-LR, CNMF, CNMF-LR, and HRNMF.]

June 19, 2015
MIDI audio excerpt
Oracle separation of a MIDI audio excerpt

<table>
<thead>
<tr>
<th></th>
<th>Mix</th>
<th>NMF-Wiener</th>
<th>HRNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bass</td>
<td>🎷️</td>
<td>🎷️</td>
<td>🎷️</td>
</tr>
<tr>
<td>Guitar</td>
<td>🎷️</td>
<td>🎷️</td>
<td>🎷️</td>
</tr>
<tr>
<td>Keyboard</td>
<td>🎷️</td>
<td>🎷️</td>
<td>🎷️</td>
</tr>
</tbody>
</table>
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


