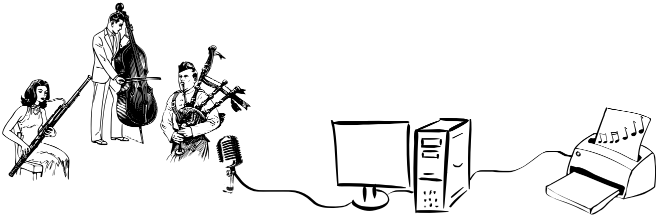
Probabilistic Latent Component Analysis and its adjustments to audio signals. Application to automatic music transcription and source separation.

Benoit Fuentes
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI

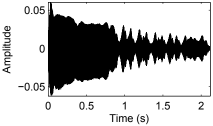
Ph.D. defense – Thursday, March 14$^{\text{th}}$ 2013

What is automatic transcription of music ?

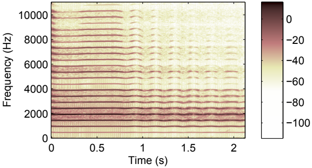


- The goal: a computer program analyzes an audio signal, and identify the notes.
- One notes: pitch, onset time and duration.
- A difficult problem: all the played notes are mixed.

Observing data: a note of trumpet



Waveform

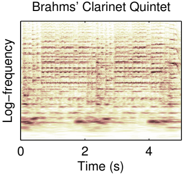


A time-freqency representation (TFR)

Observations:

- harmonic spectra,
- temporal evolutions: fundamental frequency and spectral envelope,
- presence of noise.

What solution? TFR factorizations



Input polyphonic TFR: $\mathbf{V}$.

- Put forward a TFR model $\hat{\mathbf{V}}$, depending on parameters Λ.
- Find algorithms to estimate Λ, such as:

$$\hat{\mathbf{V}}(\Lambda) \approx \mathbf{V}.$$

The transcription is deduced from Λ.

Deterministic vs probabilistic frameworks

- Deterministic: minimizing some distance between $\mathbf{V}$ and $\hat{\mathbf{V}}(\Lambda)$ [Lee and Seung 1999].

- Probabilistic:
  - $\mathbf{V}$ results from a generative process, depending on Λ,
  - Λ is estimated due to an estimator (*e.g.* ML).

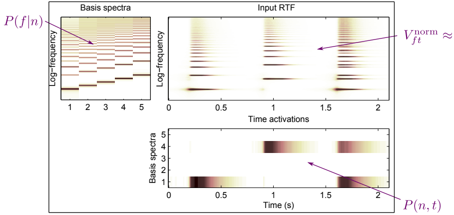
e.g. Probabilistic latent component analysis (PLCA) [Shashanka 2007].

PLCA: principle

- Generative process: drawing of many time-frequency bins $(f, t) \sim P(f, t)$.
- $\mathbf{V}$ is the histogram of the draws: $V_{ft}^{\text{norm}} = \frac{V_{ft}}{\sum_{ft} V_{ft}} \approx P(f, t)$.
- $P(f, t)$ is modeled and depends on Λ.
- Use of EM algorithm to estimate Λ.

How to model $P(f, t)$?

PLCA: basic model [Shashanka 2007]



Basis model

- A column of a CQT: weighted sum of basis spectra (atoms):

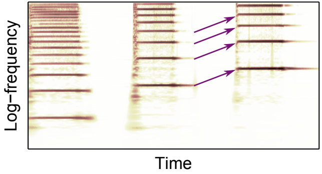
$$P(f,t) = \sum_n P(n,t)P(f|n) \qquad \Lambda = \{P(n,t), P(f|n)\}.$$

- n: a new variable representing an atom (note).

Cannot model notes with time-varying spectra !

Shift-invariant PLCA: introducing the CQT

CQT: constant-Q transform

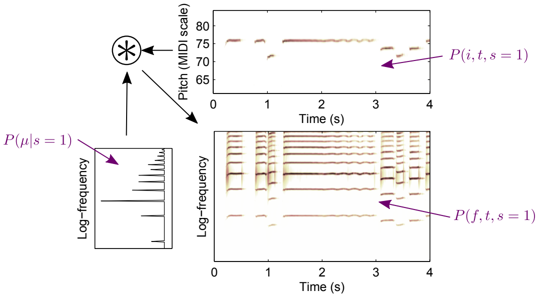


- Log. frequency scale: pitch modulation = translation of partials.

A single atom can be used to model different notes.

Shift-invariant PLCA [Smaragdis *et. al.* 2008]

- CQT = sum of sources: $P(f,t) = \sum_s P(f,t,s)$.
- Model of one source: $P(f,t,s) = \sum_i P(i,t,s)P(f-i|s)$.



Limitation: cannot model variations of spectral envelope.

Contributions

- Create new models of CQT that consider:
  - notes having pitch and spectral envelope variations,
  - robust to noise.

Contributions

- Create new models of CQT that consider:
  - notes having pitch and spectral envelope variations,
  - robust to noise.
- Proposing new tools to improve parameter estimation:
  - can be applied to any CQT model.

Contributions

- Create new models of CQT that consider:
  - notes having pitch and spectral envelope variations,
  - robust to noise.
- Proposing new tools to improve parameter estimation:
  - can be applied to any CQT model.
- Applications:
  - automatic transcription,
  - source separation.

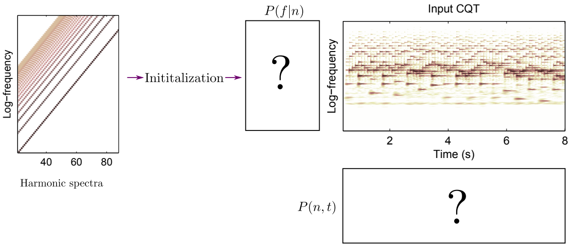
Outline

Addition of priors

- Account for prior information on observation, and therefore on parameters.
- Two advantages:
  - helping the EM algorithm to avoid local maxima,
  - making a model more identifiable.
- Four new priors:
  - sparseness,
  - temporal continuity,
  - resemblance,
  - monomodality.

Sparse priors

- Consider the following problem:



- The input signal does not necessarily contain all 88 notes.
- Order of the model overestimated.
- Idea: sparse prior on $P(n,t) = \theta_{nt}$.

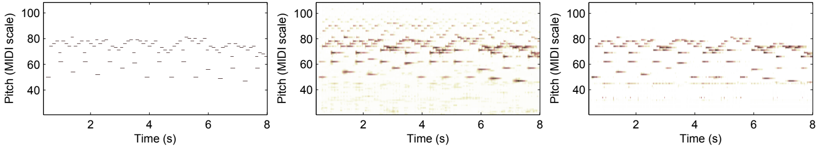
Sparse priors

- $l_{1/2}$-based sparse prior:

$$Pr\left(\boldsymbol{\theta}\right) \propto \exp\left(-2\beta_{\text{sparse}}||\boldsymbol{\theta}||_{1/2}\right) \quad \text{with} \quad ||\boldsymbol{\theta}||_{1/2} = \sum_{n,t} \sqrt{\theta_{nt}}.$$

- Rigorous proof for EM derivation.

Results: temporal activations of atoms

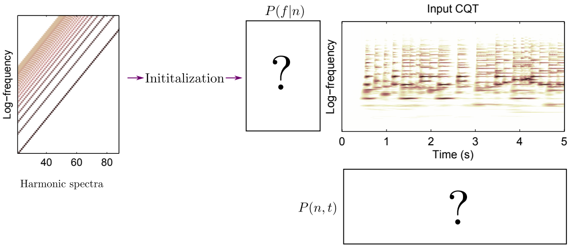


(a) No prior

(b) With sparse prior

Temporal continuity priors

- Consider the following problem:



- What if we suppose $P(n,t) \approx P(n,t-1)$?
- Could help the algorithm converge toward a more meaningful solution.
- Idea: temporal continuity prior on $P(n,t) = \theta_n^t$.

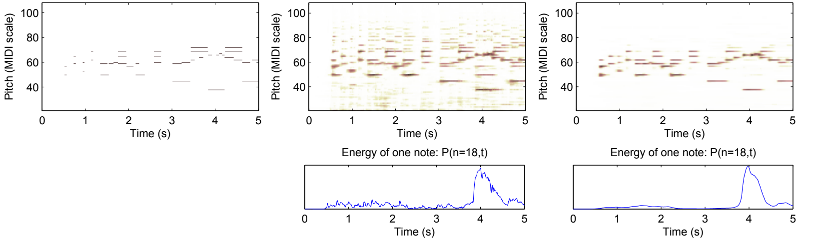
Temporal continuity prior

- Based on the ratio between geometric and arithmetic mean:

$$Pr(\boldsymbol{\theta}) \propto \left(\prod_n \prod_t 2 \frac{\sqrt{\theta_n^t \theta_n^{t-1}}}{\theta_n^t + \theta_n^{t-1}} \right)^{\beta_{\text{temp}}} .$$

- Fixed-point method for EM derivation.

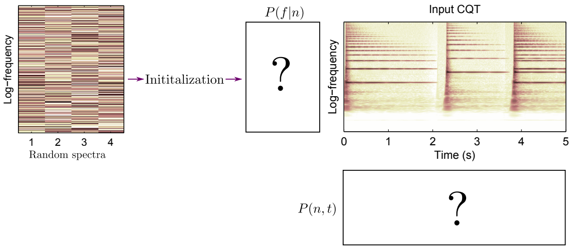
Results: temporal activations of atoms



(a) No prior (b) With temporal continuity prior

Resemblance prior

- Consider the following problem:



- Modeling notes with variations of spectral envelope: use several atoms per note.
- Cluster the atoms beforehand: atoms in one cluster are similar but not equal.
- Resemblance prior: applied to Z adjacent basis spectra $\{P(f|n=1),\dots,P(f|n=Z)\} = \{\theta_f^1,\dots,\theta_f^Z\}$.

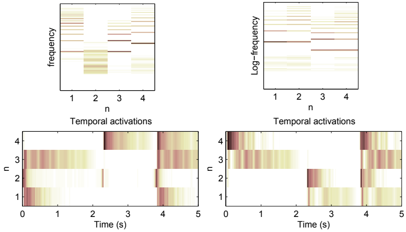
Resemblance prior

- Based on the ratio between geometric and arithmetic mean:

$$Pr(\boldsymbol{\theta}) \propto \left(\prod_f \frac{\sqrt[Z]{\prod_z \theta_f^z}}{\frac{1}{Z}\sum_z \theta_f^z} \right)^{Z\beta_{\text{res}}}.$$

- Fixed-point method for EM derivation.

Results: atoms and their temporal activations



(a) No prior (b) With resemblance prior

Slowing down the rate of convergence

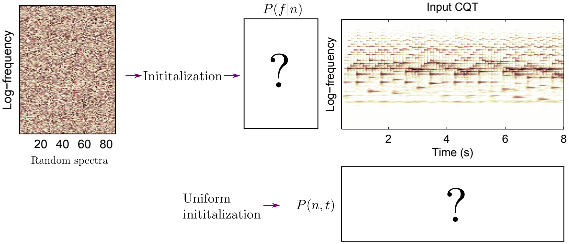
Apply a brake to the convergence of a subset of parameters:

- value at the end of the algorithm: closed to initialization,
- avoid local minima,
- make sparser the parameters that are not slowed down.

Simple to implement and effective.

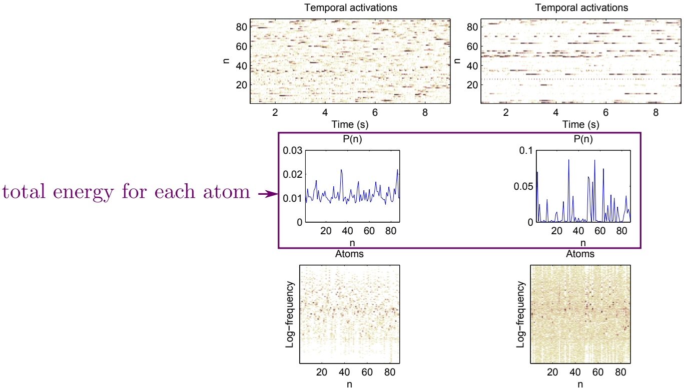
Slowing down the rate of convergence: example

- Consider the following problem:



- Brake on $P(f|n)$: makes $P(n,t)$ sparser, like the input.

Slowing down the rate of convergence: example



(a) No brake (b) Brake on $P(f|n)$

Improving parameter estimation: conclusion

- Tools to help the parameter estimations.
- Can be used with any PLCA-based model, applied to any set of parameters.
- Now: let us design new models of CQT.

Outline

Source-based model: HALCA

HALCA: Harmonic Adaptive Latent Component Analysis.

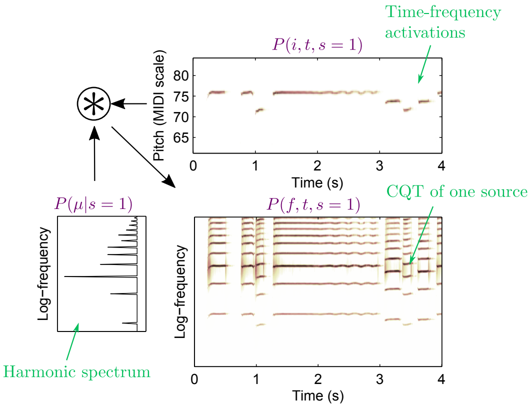
- Goal: modeling harmonic instruments having time-varying spectra:
  - pitch variations,
  - spectral envelope variations.
- Source-based model, inspired by:
  - shift invariant PLCA [Mysore and Smaragdis 2009],
  - model with harmonic constraint [Vincent *et al.* 2010].
- Model:

$$\text{CQT} = \quad \text{sum of sources} \quad + \quad \text{noise}$$

$$P(f,t) = P(c=h)\sum_s P_h(f,t,s) + P(c=b)P_b(f,t)$$

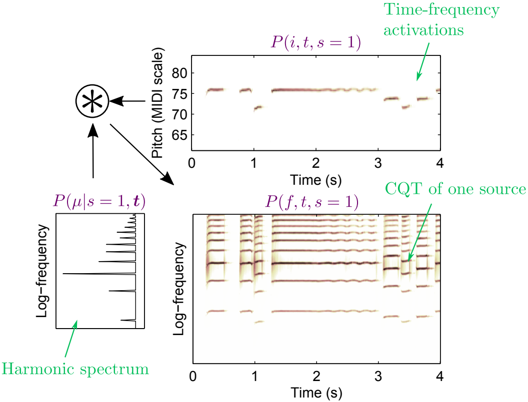
HALCA: source model

From shift-invariant PLCA to HALCA:



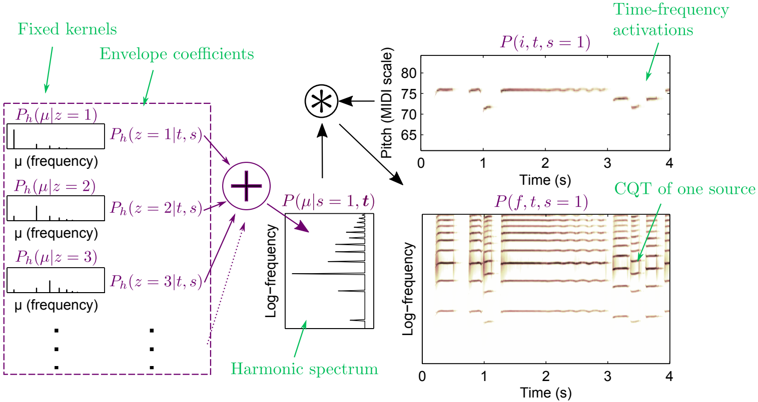
HALCA: source model

From shift-invariant PLCA to HALCA:



HALCA: source model

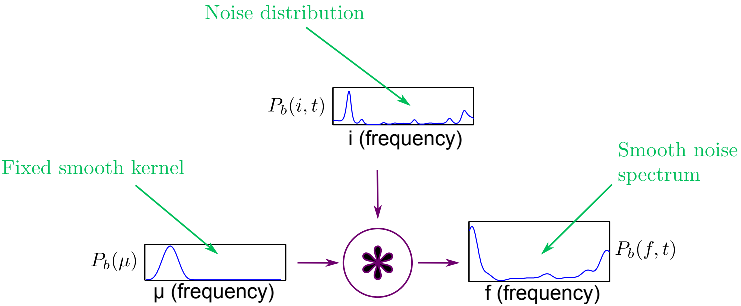
From shift-invariant PLCA to HALCA:



$$P_h(f,t,s) = \sum_{z,i} P_h(i,t,s) P_h(f-i|z) P_h(z|t,s).$$

HALCA: noise model

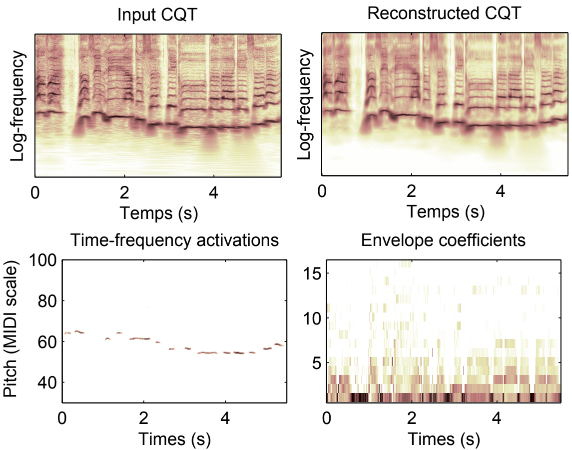
At time t:



$$P_b(f,t) = \sum_i P_b(i,t) P_b(f-i)$$

HALCA: example

Example on singing voice:

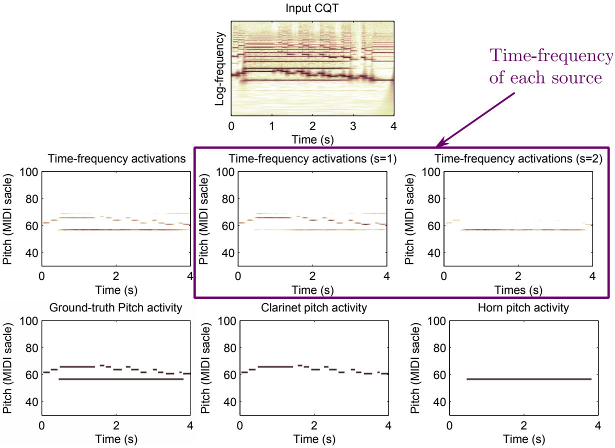


HALCA: addition of plugins

- Sparse prior on time-frequency activations $P_h(i, t, s)$.
- Temporal continuity prior on envelope coefficients $P_h(z|t, s)$: continuity of timbre.
- Brake on envelope coefficients $P_h(z|t, s)$: initialization is relevant.
- Resemblance prior: not applied here.

HALCA: discussion

Sources do not correspond to real instruments:



HALCA: conclusion

- Sources represent meta-instruments:
  - several sources are used to model a single instrument,
  - one source contributes to the modeling of several instruments.
- The number of sources can be fixed:
  - a fix number of sources can model an unknown number of instrument.
- Overall time-frequency activations: $P_h(i,t) = \sum_s P_h(i,t,s)$.

But is it relevant to keep the concept of source?

Note-based model: BHAD

BHAD: Blind Harmonic Adaptive Decomposition.

- Get rid of the concept of sources, but keep an expressive model.
- The noise component is kept.

Note-based model: BHAD

BHAD: Blind Harmonic Adaptive Decomposition.

- Get rid of the concept of sources, but keep an expressive model.
- The noise component is kept.
- From HALCA to BHAD:

$$P_h(f,t,s) = \sum_{z,i} P_h(i,t,s) P_h(f-i|z) P_h(z|t,s).$$

Note-based model: BHAD

BHAD: Blind Harmonic Adaptive Decomposition.

- Get rid of the concept of sources, but keep an expressive model.
- The noise component is kept.
- From HALCA to BHAD:

$$P_h(f,t,\cancel{s}) = \sum_{z,i} P_h(i,t,\cancel{s}) P_h(f-i|z) P_h(z|t,\cancel{s}).$$

Note-based model: BHAD

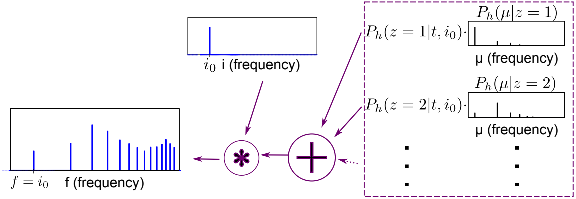
BHAD: Blind Harmonic Adaptive Decomposition.

- Get rid of the concept of sources, but keep an expressive model.
- The noise component is kept.
- From HALCA to BHAD:

$$P_h(f,t,\cancel{s}) = \sum_{z,i} P_h(i,t,\cancel{s}) P_h(f-i|z) P_h(z|t,\cancel{s},\mathbf{i}).$$

Note-based model: BHAD

- At time t, consider a comb spectrum, of fundamental frequency i_0:



$$P_h(f|i_0,t) = \sum_z P_h(f-i_0|z)P_h(z|t,i_0)$$

- All values of i considered to model a polyphonic spectrum:

$$P_h(f,t) = \sum_{z,i} P_h(i,t)P_h(f|i,t).$$

BHAD: addition of plugins

- Sparse prior on the time-frequency activations $P_h(i,t)$.
- Brake on envelope coefficients $P(z|t,i)$.
- Resemblance prior on envelope coefficients $P(z|t,i)$ for given i:
  - account for timbre redundancy of notes over time.
- Temporal continuity prior: not applied here.

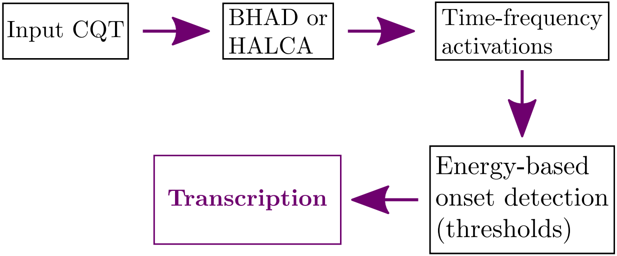
CQT models: conclusion

- Two new models to factorize CQTs of musical signals.
- Adaptive models: all parameters depend on time t.
- Possibility to add plugins (priors, brake).

We can now applied those algorithms to music transcription and source separation.

Outline

Transcription algorithm



Databases and metrics

- Three evaluation databases:
  - MAPS (piano) [Emiya 2008],
  - MIREX (woodwind quintet),
  - QUASI (rock, reggae, song,...).
- Metric to measure transcription quality:
  - Recall $\mathcal{R}$,
  - Precision $\mathcal{P}$,
  - F-measure $\mathcal{F}$.

Transcription systems

- HALCA ($S = 4$ sources):
  - H_4 : no plugins,
  - $H_4 - sb$: sparse prior + brake,
  - $H_4 - sbt$: sparse prior + brake + temporal prior.
- BHAD:
  - B: no plugins,
  - $B - sb$: sparse prior + brake,
  - $B - sbr$: sparse prior + brake + resemblance prior.

Results

Sparse prior and brake: improve performances

Algorithm	MAPS	MIREX	QUASI
H_4	57.8 [55.6, 61.5]	**62.4** [51.4, 79.4]	38.8 [38.1, 41.9]
$H_4 - sb$	**59.4** [52.3, 70.9]	59.3 [45.7, 84.6]	**41.5** [37.9, 50.3]
B	47.5 [56.1, 41.9]	61.5 [55.5, 69.0]	32.9 [39.7, 32.9]
$B - sb$	**60.0** [52.8, 71.7]	**63.6** [51.3, 83.7]	**43.1** [40.0, 52.0]

$\mathcal{F}$ $[\mathcal{R},\mathcal{P}]$ (%)

Temporal prior: depends on database

Algorithm	MAPS	MIREX	QUASI
$H_4 - sb$	59.4 [52.3, 70.9]	59.3 [45.7, 84.6]	**41.5** [37.9, 50.3]
$H_4 - sbt$	**61.8** [54.9, 73.6]	**64.2** [51.7, 84.6]	40.7 [36.8, 49.6]

$\mathcal{F}$ $[\mathcal{R},\mathcal{P}]$ (%)

Resemblance prior: no a good assumption

Algorithm	MAPS	MIREX	QUASI
$B - sb$	60.0 [52.8, 71.7]	**63.6** [51.3, 83.7]	**43.1** [40.0, 52.0]
$B - sbr$	**60.6** [51.6, 76.7]	61.6 [47.4, 88.2]	37.3 [34.3, 46]

$\mathcal{F}$ $[\mathcal{R},\mathcal{P}]$ (%)

Results: comparison

- Comparison with two reference algorithms:

Algorithm	MAPS	MIREX	QUASI
$H_4 - sb$	59.4 [52.3, 70.9]	59.3 [45.7, 84.6]	41.5 [37.9, 50.3]
$B - sb$	**60.0** [52.8, 71.7]	**63.6** [51.3, 83.7]	**43.1** [40.0, 52.0]
[Vincent *et al.* 2010]	**45.3** [67.0, 35.8]	**57.9** [81.1, 45.0]	20.3 [63.8, 12.3]
[Dessein *et al.* 2012]	45.1 [43.3,48.5]	52.0 [48.6, 55.9]	**20.9** [33.4, 17.0]

$$\mathcal{F}\ [\mathcal{R},\mathcal{P}]\ (\%)$$

- Robustness of our algorithms to musical genre.

Results: MIREX

- $B - sb$ has been submitted to MIREX 2012 international competition:

Algorithm	$\mathcal{R}$ (%)	$\mathcal{P}$ (%)	$\mathcal{F}$ (%)
BD2	52.4	38.1	43.0
BD3	46.8	38.2	41.1
CPG1	14.5	54.5	21.9
CPG2	15.1	54.0	22.5
CPG3	19.9	51.5	27.3
FBR2 ($B - sb$)	71.6	55.3	**61.3**
FT1	3.3	21.8	5.5
KD3 ([Dressler 2012])	65.2	64.7	**64.6**
SB5	63.5	42.3	49.8

Sound example

- Grieg, Violon Sonata:

Original Resynthized

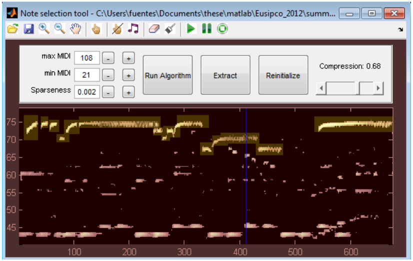
Melody extraction

- The goal: automatically extract the main melody.
- Hybrid model:

$$\begin{aligned}\text{input CQT} &= \text{melody} + \text{accompaniment},\\ &= \text{HALCA}_s + \text{PLCA}.\end{aligned}$$

- HALCA_s : source model of HALCA.
- After estimation of parameters, soft masks can be deduced and source temporal signals estimated.

Supervised source separation



- Source separation based on time-frequency masking.

Conclusion

- Two new models for musical signal analysis, HALCA and BHAD:
  - expressive models,
  - suitable for a large class of signals.
- Tools to help parameter estimations:
  - four new priors to account for prior knowledge on signals to analyze,
  - slowing down the convergence of a subset of parameters: cheap and effective,
- Applications:
  - new state of the art transcription algorithms, especially for complex music,
  - two source separation applications.

Perspectives

- Multiply semantic levels for spectrum modeling:
  - from mid-level to low-level representations: *e.g.* more realistic note spectra models,
  - from high-level to mid-level representations: *e.g.* MIDI notes←chroma←chords←tonality.
- Work on dynamic modeling:
  - MLCATS: modeling energy transitions between t and $t+1$,
  - modeling onsets/offsets.

The end

Publication:

B. Fuentes, R. Badeau and G. Richard: Harmonic adaptive latent component analysis of audio and application to music transcription. *IEEE TASLP* (accepted), 2013.

B. Fuentes, R. Badeau and G. Richard: Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In Proc. of EUSIPCO, Romania, 2012.

B. Fuentes, A. Liutkus, R. Badeau and G. Richard: Probabilistic Model for main melody extraction using constant-Q transform. *In Proc. of ICASSP*, Japan, 2012.

B. Fuentes, R. Badeau and G. Richard: Analyse des structures harmo-niques dans les signaux audio : modéliser les variations de hauteur et d'enveloppe spectrale. *In GRETSI*, France, 2011.

B. Fuentes, R. Badeau and G. Richard: Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. *In Proc. of ICASSP*, Czech Republic, 2011.

Thank you for you attention !