Main Instrument Separation from Stereophonic Audio Signals Using a Source/Filter Model

J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard and B. David

Institut Télécom; Télécom ParisTech; CNRS LTCI

25/08/09

Presentation overview

- **Introduction**
- **Proposed method**
  - Signal model
  - System overview
- **Results**
- **Conclusion**

Introduction: MIR and Source Separation

- **Music Information Retrieval (MIR) and Audio Source Separation: e.g. [Vincent, 06] or [Gillet and Richard, 08].**

- **Using the *main melody* to separate the corresponding instrument:**
  - *"Audio Melody Extraction"* at MIREX since 2004.
  - Source separation inspired by [Ozerov *et al.*, 07].

- **Contributions:**
  - Stereo extension of our system for mono.
  - Unvoiced part.
  - Smoothed filters.

Task definition 1/2

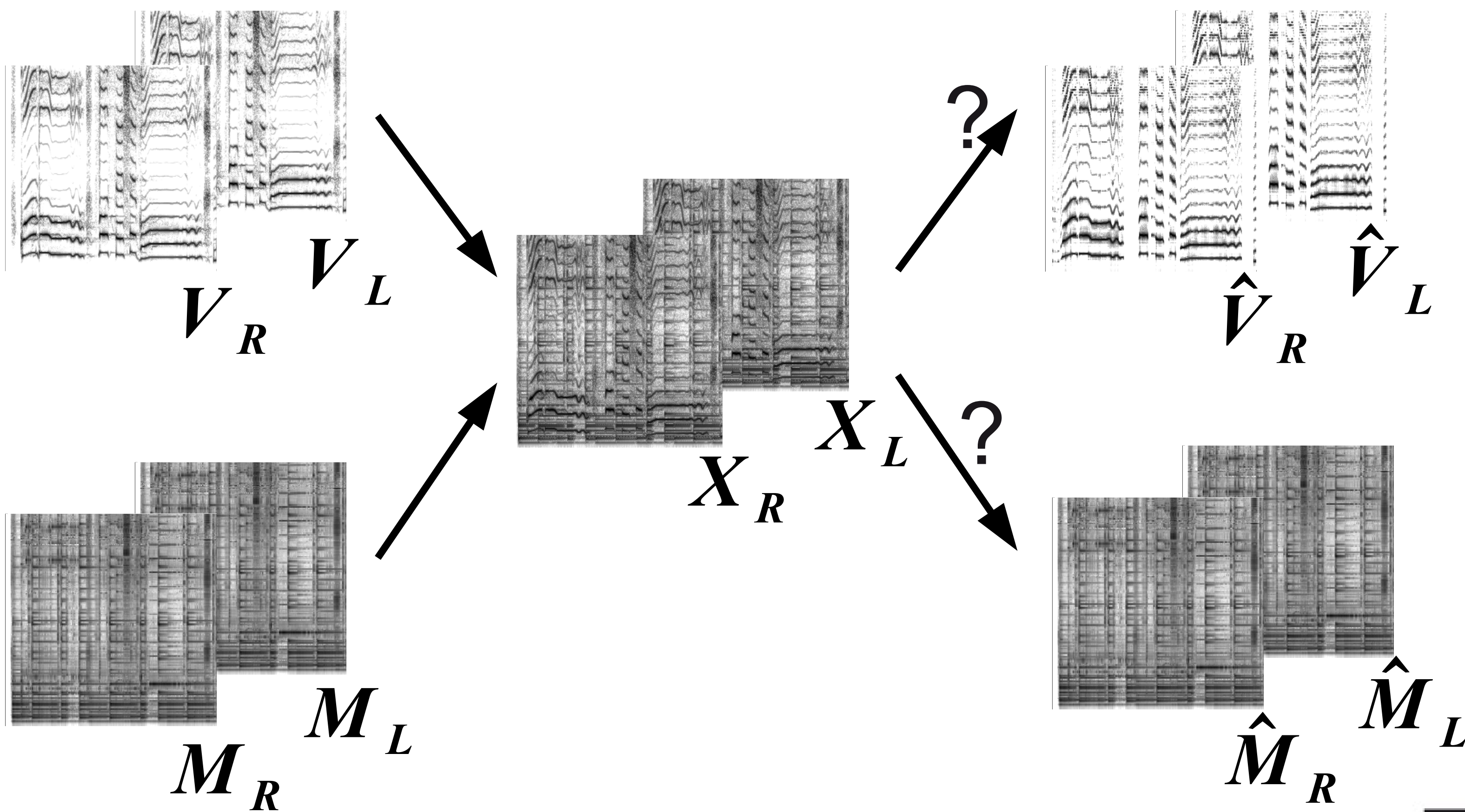
- **Mixture** $X = V + M$

  - V : “Main instrument”, i.e. leading instrument.
    - Harmonic and mono-pitch,
    - Predominant energy,
    - Continuous melody line.
  - M : accompaniment.
    - Multiple sources, multiple pitches, percussive sounds...

TELECOM ParisTech

Task definition 2/2

- **Stereo mixture:** $X = [X_R, X_L]$



Model: stereophonic signal

- **Stereo mixture:** $X = [X_R, X_L]$
  - Instantaneous, "panning" effect.

- **Assumption on STFT distribution:**

$$\begin{cases} \boldsymbol{X}_{\boldsymbol{R},\boldsymbol{fn}} \sim N_c\left(0,\ \alpha_R^2 \boldsymbol{S}_{\boldsymbol{V},\boldsymbol{fn}} + \sum_{j=1}^{J} \beta_{Rj}^2 \boldsymbol{S}_{\boldsymbol{M},\boldsymbol{j},\boldsymbol{fn}}\right) \\ \boldsymbol{X}_{\boldsymbol{L},\boldsymbol{fn}} \sim N_c\left(0,\ \underbrace{\alpha_L^2 \boldsymbol{S}_{\boldsymbol{V},\boldsymbol{fn}}}_{\textit{Solo}} + \underbrace{\sum_{j=1}^{J} \beta_{Lj}^2 \boldsymbol{S}_{\boldsymbol{M},\boldsymbol{j},\boldsymbol{fn}}}_{\textit{accompaniment}}\right) \end{cases}$$

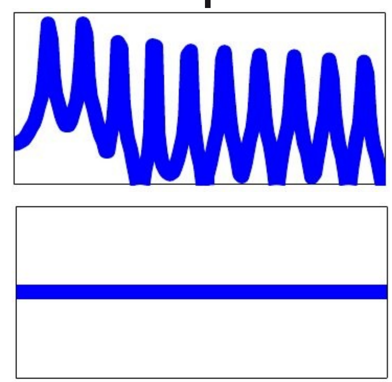
- $p(\boldsymbol{X}) = p(\boldsymbol{X}_{\boldsymbol{R}})\, p(\boldsymbol{X}_{\boldsymbol{L}})$

Model: unvoiced part in solo source/filter 1/2

- S_V product of 2 contributions:

$$S_V = S_\Phi . S_{F_0} = S_\Phi . [W_{F_0} H_{F_0}]$$

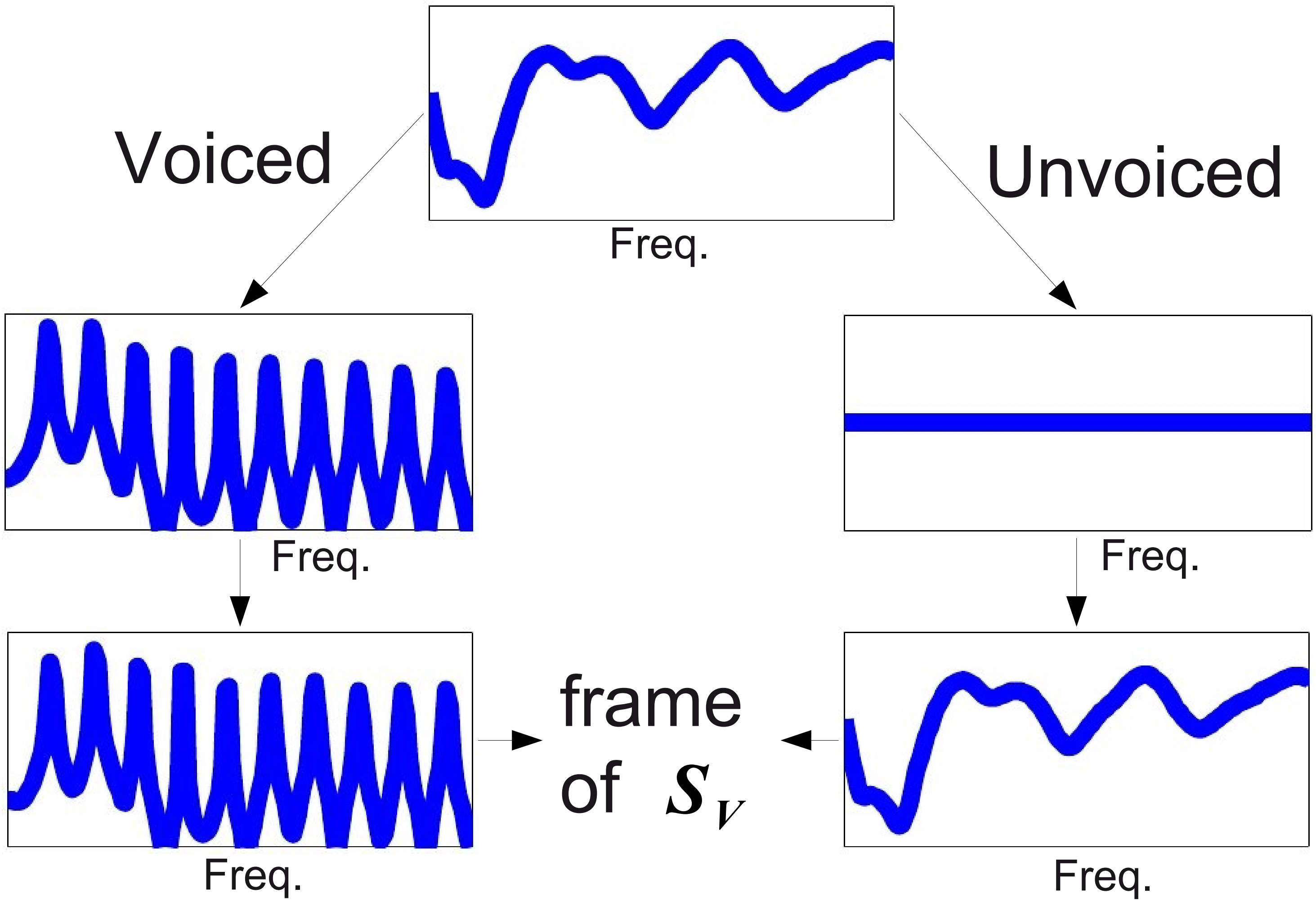
- W_{F_0} fixed dictionary of source spectra:
  - Voiced: spectral combs.
  - Unvoiced: “white noise”



- Assumption: unvoiced and voiced parts filtered by *same* spectral shapes.

TELECOM ParisTech

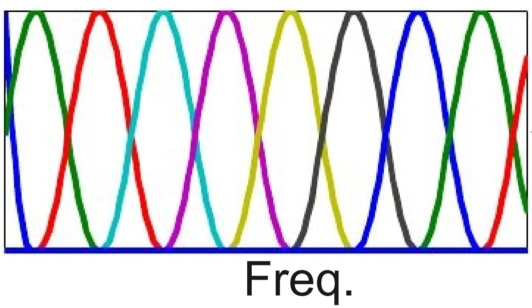
Model: unvoiced part in solo source/filter 2/2



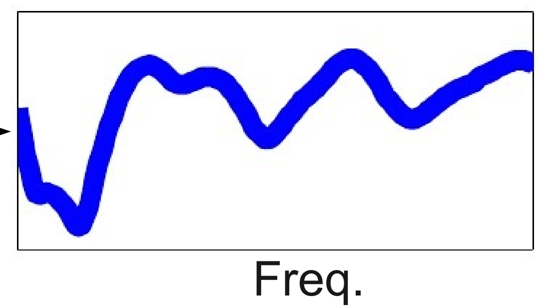
Model: smooth filters

- $S_V = S_\Phi . S_{F_0} = [\underbrace{W_\Gamma H_\Gamma}_{W_\Phi} H_\Phi] . S_{F_0}$
  - W_Γ fixed dictionary of smooth elements:

W_Γ



$\xrightarrow{H_\Gamma}$



$\in W_\Phi$

Model: accompaniment

- **J components:**

$$\begin{cases} \boldsymbol{S}_{\boldsymbol{M}_R,\boldsymbol{fn}} = \sum_{j=1}^{J} w_{fj}\,\beta_{Rj}^{2}\,h_{jn} = [\boldsymbol{W}_{\boldsymbol{M}}\boldsymbol{B}_{\boldsymbol{R}}\boldsymbol{H}_{\boldsymbol{M}}]_{\boldsymbol{fn}} \\ \boldsymbol{S}_{\boldsymbol{M}_L,\boldsymbol{fn}} = \sum_{j=1}^{J} w_{fj}\,\beta_{Lj}^{2}\,h_{jn} = [\boldsymbol{W}_{\boldsymbol{M}}\boldsymbol{B}_{\boldsymbol{L}}\boldsymbol{H}_{\boldsymbol{M}}]_{\boldsymbol{fn}} \end{cases}$$

with

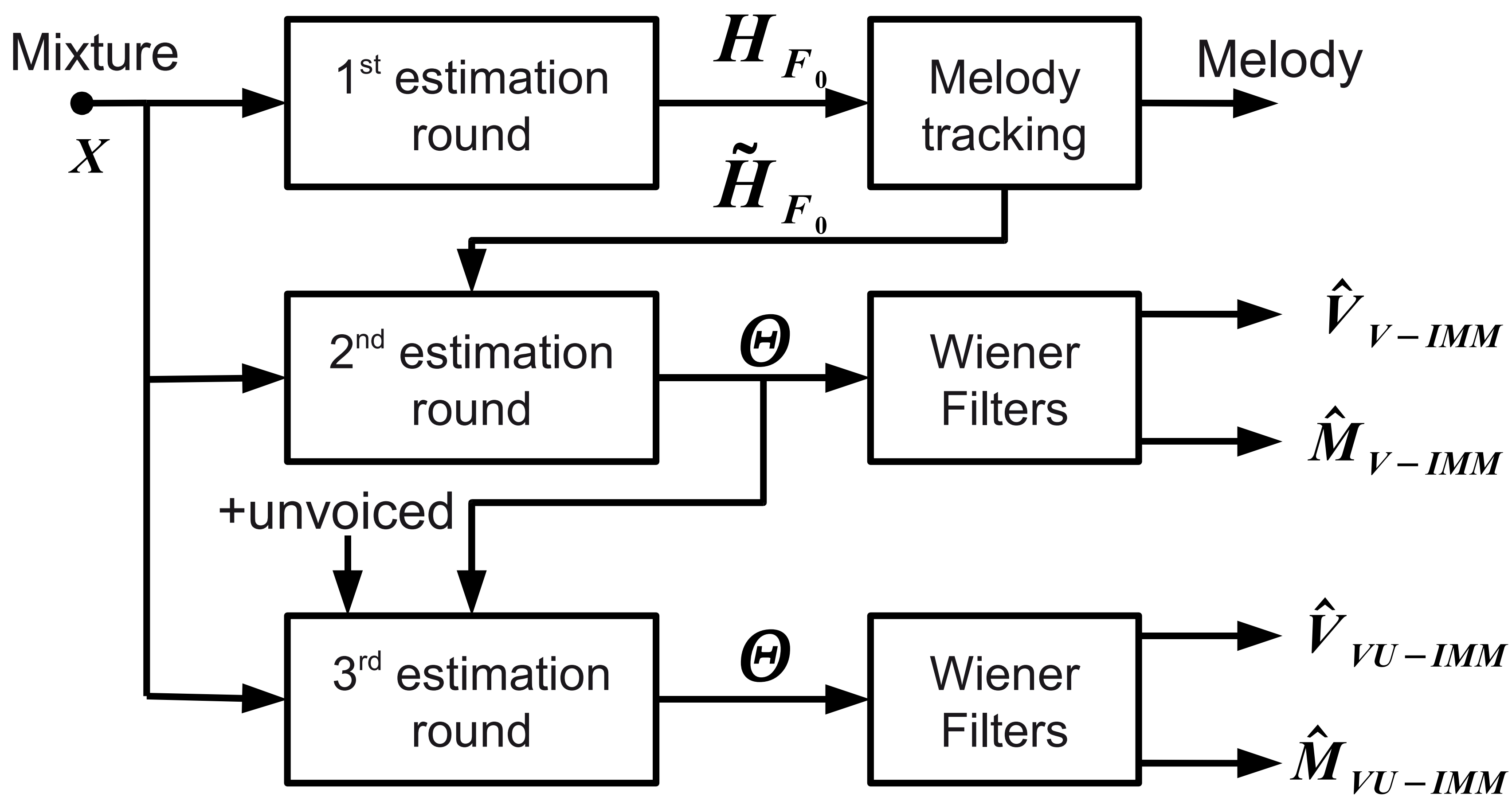
$$\begin{cases} \boldsymbol{B}_{\boldsymbol{R}} = diag([\beta_{R1}^2, \cdots, \beta_{Rj}^2, \cdots, \beta_{RJ}^2]) \\ \boldsymbol{B}_{\boldsymbol{L}} = diag([\beta_{L1}^2, \cdots, \beta_{Lj}^2, \cdots, \beta_{LJ}^2]) \end{cases}$$

Estimation of parameters

$$\begin{cases} \boldsymbol{S}_{X_R} = \alpha_R^2 (W_\Gamma H_\Gamma H_\Phi) . (W_{F_0} H_{F_0}) + W_M B_R H_M \\ \boldsymbol{S}_{X_L} = \alpha_L^2 (W_\Gamma H_\Gamma H_\Phi) . (W_{F_0} H_{F_0}) + W_M B_L H_M \end{cases}$$

- **ML Criterion:**
  - $\ln p(X) = \ln p(X_R) + \ln p(X_L)$
- **Estimation of parameters:**
  - Gradient method, analogous to classic NMF rules.

Proposed method: system overview



Results: database and evaluation criteria

Synthetic database:

- MTG MASS database, by M. Vinyes, http://www.mtg.upf.edu/static/mass/resources.
- 13 synthetic instantaneous mixtures from separated tracks.

Evaluation criteria:

- *BSS Eval* criteria for SiSEC: ***SDR***, *ISR*, *SIR*, *SAR*.

Results: effects of the contributions

■ Stereo vs. Mono

- ~2dB SDR gain
- Resulting audio signal more “coherent”.

■ Smooth filters

- Practically no improvement,
- More realistic, suitable for other applications.

■ Unvoiced model

- ~0.3dB SDR gain,
- Catches drum signals...
- … and misses some consonants.

Results: SiSEC08

- **SiSEC'08, *Professionally Produced Music Recordings***
  - Development set: 2 stereo signals, 2 artists,
  - Test set: 2 stereo signals (from the same songs).

- **1 result submitted: *Tamy*.**
  - Female singer + guitar,

- **Results:**
  - Success of algorithms based on *melody tracking*,
  - Music sound difficult because of guitar signal.

Conclusion

- **Improvements over previous de-soloing system:**
  - Explicit stereo model,
  - Smoothness of filter part,
  - Taking into account unvoiced parts.

- **http://perso.telecom-paristech.fr/durrieu/en/eusipco09/**
- **Perspectives:**
  - Use in applications such as lyrics recognition,
  - Better unvoiced model,
  - Take into account more mutual information between the channels [Ozerov and Févotte, 09].

TELECOM ParisTech

Conclusion

- **Improvements over previous de-soloing system:**
  - Explicit stereo model,
  - Smoothness of filter part,
  - Taking into account unvoiced parts.

- **http://perso.telecom-paristech.fr/durrieu/en/eusipco09/**
- **Perspectives:**
  - Use in applications such as lyrics recognition,
  - Better unvoiced model,
  - Take into account more mutual information between the channels [Ozerov and Févotte, 09].

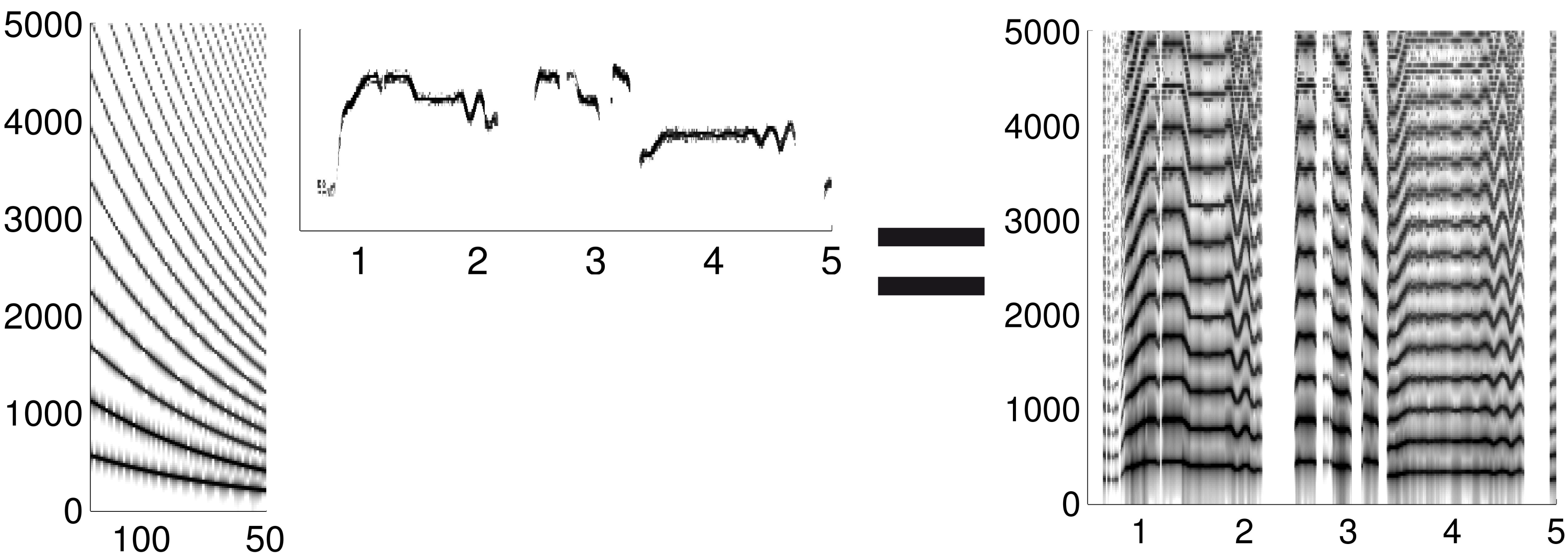
Some examples + questions?

Additional material

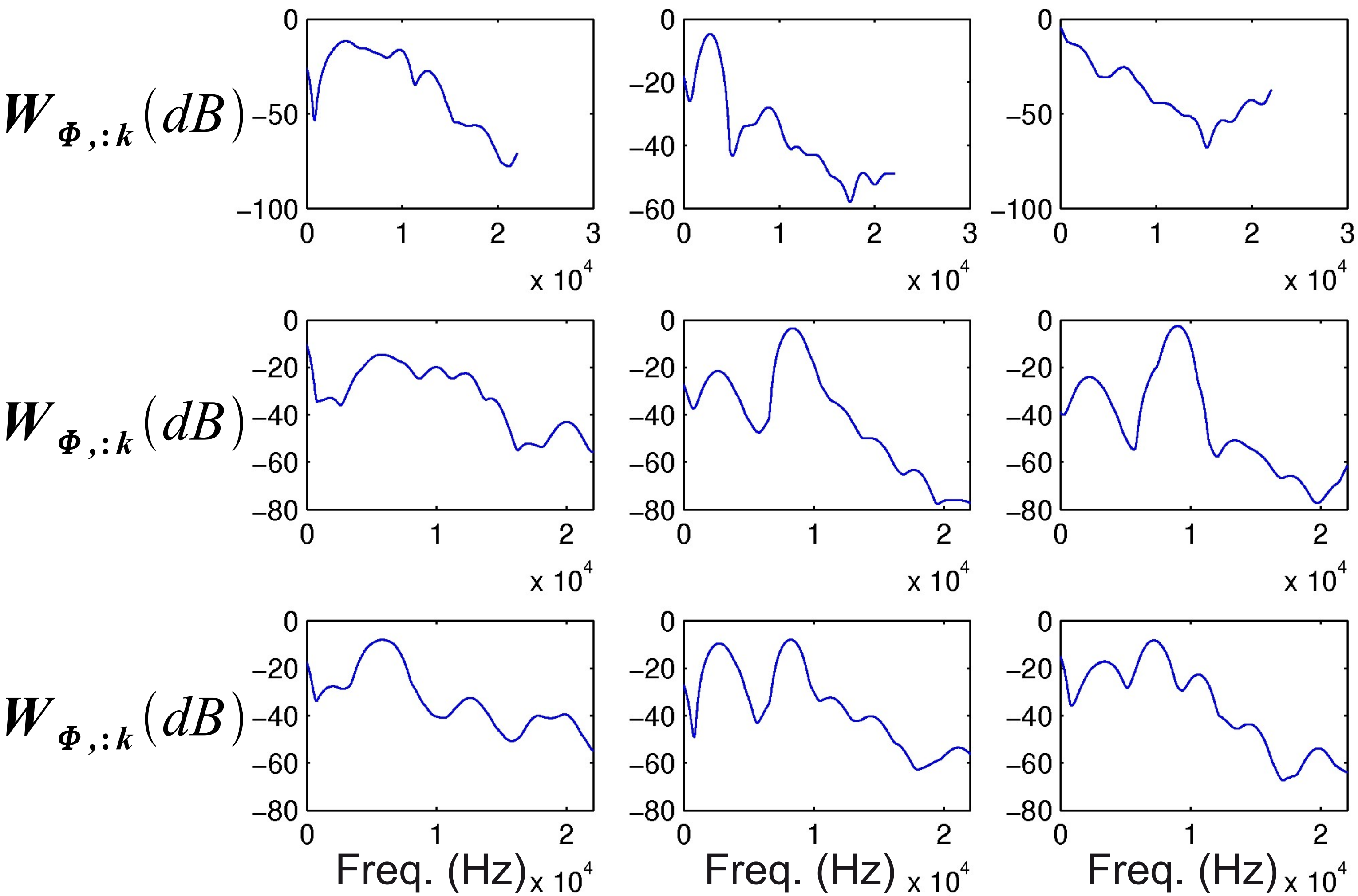
Main instrument/Accompaniment model [Durrieu *et al.*, ICASSP'09]

$$S_X = S_\Phi . S_{F_0} + S_M$$

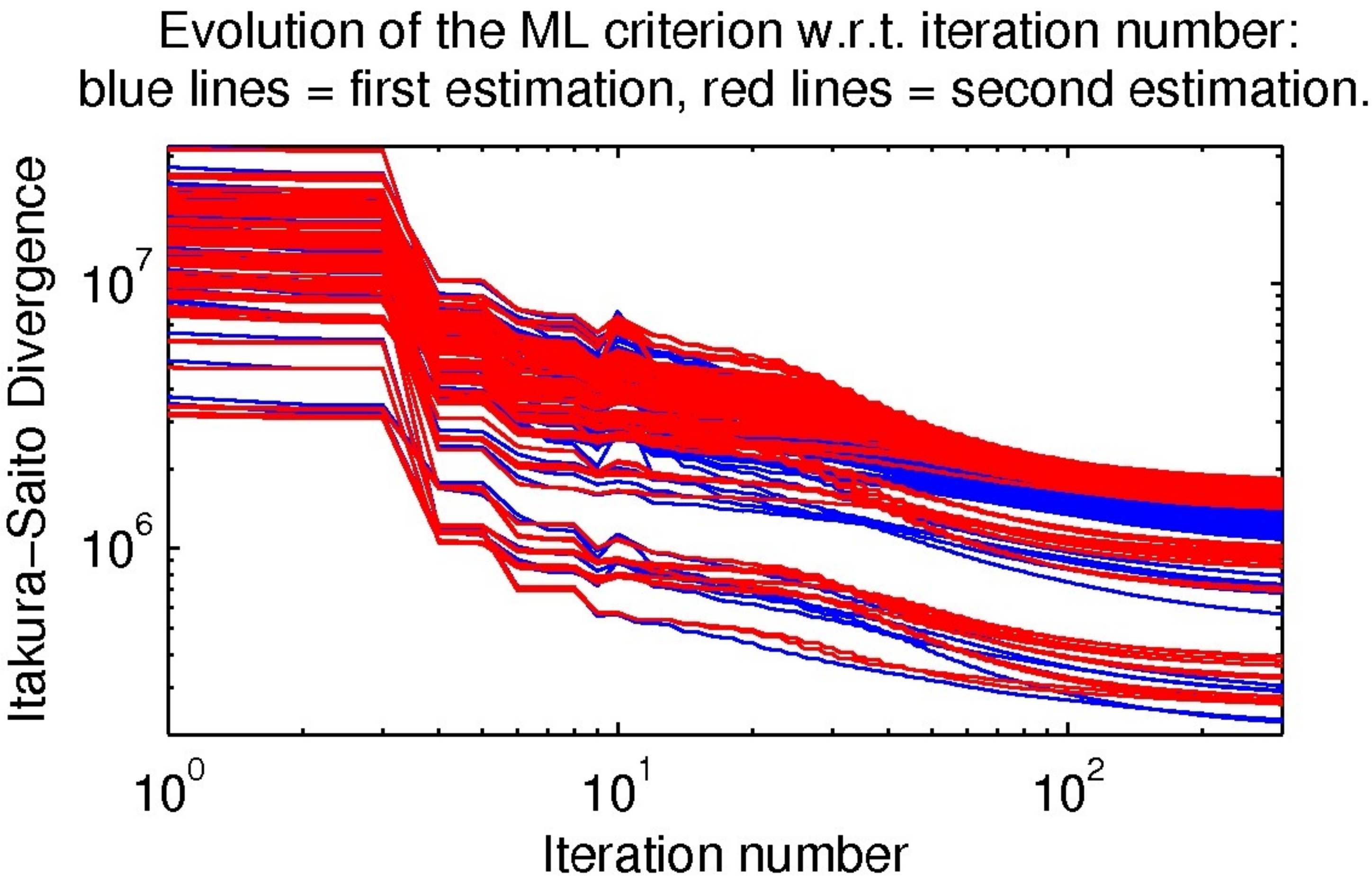
$$S_X = \underbrace{(W_\Phi H_\Phi)}_{Filter} . \underbrace{(W_{F_0} H_{F_0})}_{Source} + \underbrace{(W_M H_M)}_{Accompaniment}$$



Spectral shapes for the estimated filters



“Convergence”: evolution of Itakura Saito criterion over the iterations.



Audio source separation

- **Wiener filter:**

$$\hat{V}_R = \frac{\alpha_R^2 \boldsymbol{S}_V}{\alpha_R^2 \boldsymbol{S}_V + \sum_{j=1}^{J} \beta_{Rj}^2 \boldsymbol{S}_{M,j}} \boldsymbol{X}_R$$

$$\hat{V}_R = \frac{\alpha_R^2 (\boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma \boldsymbol{H}_\Phi).(\boldsymbol{W}_{F_0} \boldsymbol{H}_{F_0})}{\alpha_R^2 (\boldsymbol{W}_\Gamma \boldsymbol{H}_\Gamma \boldsymbol{H}_\Phi).(\boldsymbol{W}_{F_0} \boldsymbol{H}_{F_0}) + \boldsymbol{W}_M \boldsymbol{B}_R \boldsymbol{H}_M} \boldsymbol{X}_R$$

Results: details, MTG MASS

- **Average on our database: for each criterion, results given as solo/accompaniment:**

Method	SDR	ISR	SIR	SAR	gSDR	gSIR
Mono	5.8/6.9	9.0/21.8	16.8/9.6	5.8/11.5	6.9/5.8	17.8/8.5
V-IMM0	7.9/8.9	12.1/23.0	19.2/12.6	8.2/12.5	8.9/7.9	20.2/11.6
V-IMM1	7.9/8.9	12.5/22.1	18.4/12.8	8.3/11.6	8.9/7.9	19.4/11.8
VU-IMM0	**8.2/9.3**	12.4/**23.3**	**19.9**/12.9	8.7/**12.7**	**9.3/8.2**	**20.9**/11.8
VU-IMM1	**8.2/9.3**	**13.0**/21.8	18.6/**13.2**	**8.8**/12.0	**9.3/8.2**	19.6/**12.2**

SiSEC'08

System	Singer SDR	Guitar SDR
Cancela2	**9.7**	8.6
VU-IMM	7.8	**9.4**
Cancela1	8.7	8.0
V-IMM	6.9	8.6
Cobos	6.4	8.0
Ozerov	5.1	6.7
Ozerov/Févotte	3.6	5.3
Vinyes Raso	4.9	4.2
Ideal Binary Mask	*10.1*	*11.8*