Automatic separation and transcription of the main melody from polyphonic music signals

Jean-Louis Durrieu, PhD candidate

TSI Department, Telecom ParisTech

11/30/09

Automatic separation and transcription of the main melody from polyphonic music signals

- **Introduction**
- **Signal model**:
  - Gaussian Scaled Mixture Model (GSMM)
  - Instantaneous Mixture Model (IMM)
- **Applications:**
  - **Transcription** of the main melody
  - **Separation** of the leading instrument
- **Conclusion**

Introduction

- **Music Information Retrieval (MIR)** and **Blind Audio Source Separation (BASS)**

- **Linking** these fields for **Leading Instrument transcription and separation**

- Related **tasks**:
  - **Predominant F0 estimation**
  - **"De-soloing"** - lead / accompaniment separation

- **Applications**: indexing (QBH), feature (classification), "Karaoke", etc.

Introduction: state of the art

- **Audio Melody Extraction** (AME), at **MIREX:**
  - Goto (2000): **PreFEst**
  - Ryynänen (2006): Note event model
  - Dressler (2009): Peak picking and "Auditory Streaming"

- **BASS:**
  - Benaroya (2006), Ozerov (2007): **spectral models, Wiener filtering**

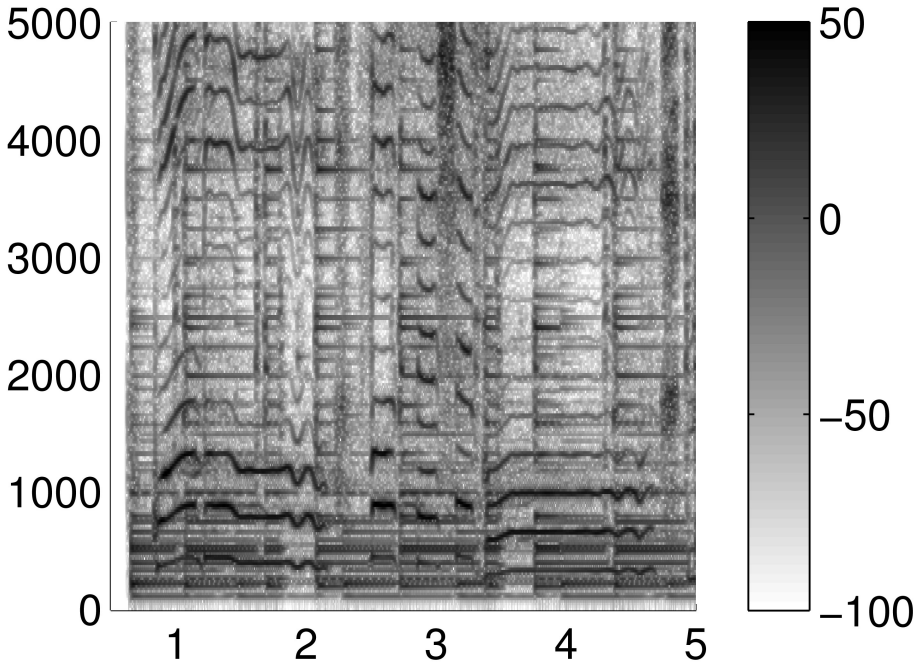
- ... or a bit of both:
  - Vincent (2006): **Instrument models, Wiener filtering**
  - Li YP (2007): pitch detection + **binary masking**

Introduction: transcription and separation

- **Transcription of the main melody:**
  - Main melody: predominant, individual sequence of fundamental frequencies
- **Separation of the leading instrument:**
  - Separate the signal corresponding to the estimated main melody – the **leading instrument**...
  - ... and its (complementary) **accompaniment**

Signal Model

Signal Model: framework

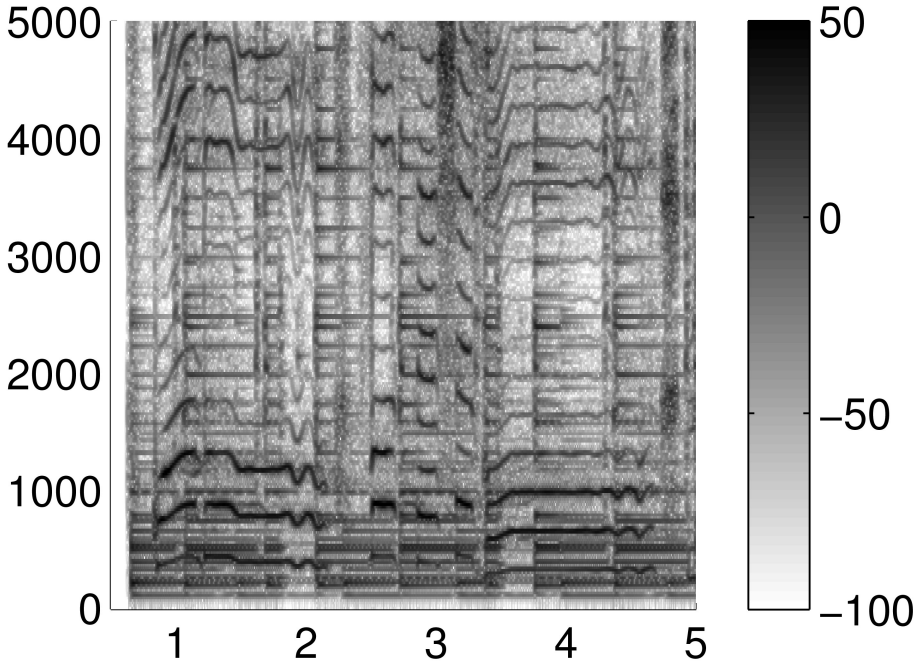


- **Spectral model:**
  - Time-Frequency representation: **STFT**
  - **Invertible**
  - **Limited resolution,** but...
- **Statistical model:**
  - FT vector = **Gaussian** vector

$$\boldsymbol{y_n} \sim N_c(\mathbf{0_F}, diag(\boldsymbol{s_n^Y}))$$

$$\Rightarrow y_{fn} \sim N_c(0, s_{fn}^Y)$$

Signal Model: mixture



- **Instantaneous Mixture:**
  - $X = V + M$
- **Contributions characterized by variance:**
  - $v_n \sim N_c(\mathbf{0}_F, diag(s_n^V))$
  - $m_n \sim N_c(\mathbf{0}_F, diag(s_n^M))$
- **Independence:**

$$x_n \sim N_c(\mathbf{0}_F, diag(s_n^V + s_n^M))$$

Signal Model: complex Gaussians

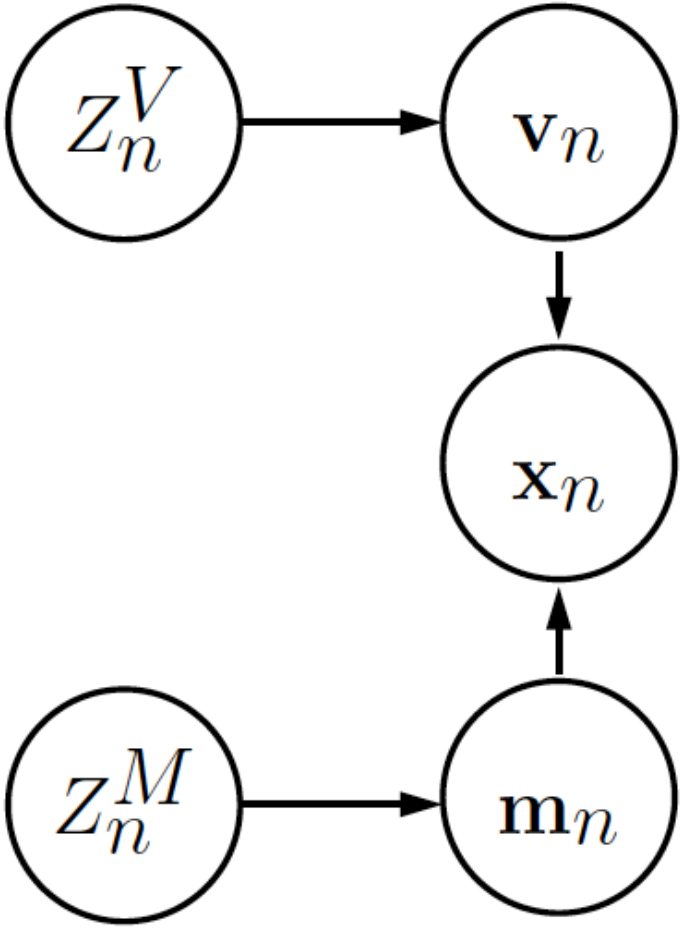
- **Advantages:**
  - **Interpretation** of parameters: $-\log N_c(\boldsymbol{v}_n; \mathbf{0}_F, diag(\boldsymbol{s}_n^V))$

$$\sum_f -\log\left(\frac{|v_{fn}|}{\pi s_{fn}^V}\right) + \frac{|v_{fn}|^2}{s_{fn}^V} =^c \sum_f -\log\left(\frac{|v_{fn}|^2}{s_{fn}^V}\right) + \frac{|v_{fn}|^2}{s_{fn}^V} - 1 = D_{IS}(|\boldsymbol{v}_n|^2, \boldsymbol{s}_n^V)$$

  - "Natural" expression of **Wiener filters**
  - **Easy** calculations

- **Drawbacks:**
  - **Realistic** from **generative** point of view?
  - **Phase uniformly distributed...**

Signal Model: Gaussian Scaled Mixture Model (GSMM)



- **Benaroya: GSMM to separate voice/music**

$$\boldsymbol{v}_n | Z_n^V = u \sim N_c(\mathbf{0}_F, h_{un}^V \, diag(\boldsymbol{w}_u^V))$$

$$\boldsymbol{m}_n | Z_n^M = r \sim N_c(\mathbf{0}_F, h_{rn}^M \, diag(\boldsymbol{w}_r^M))$$

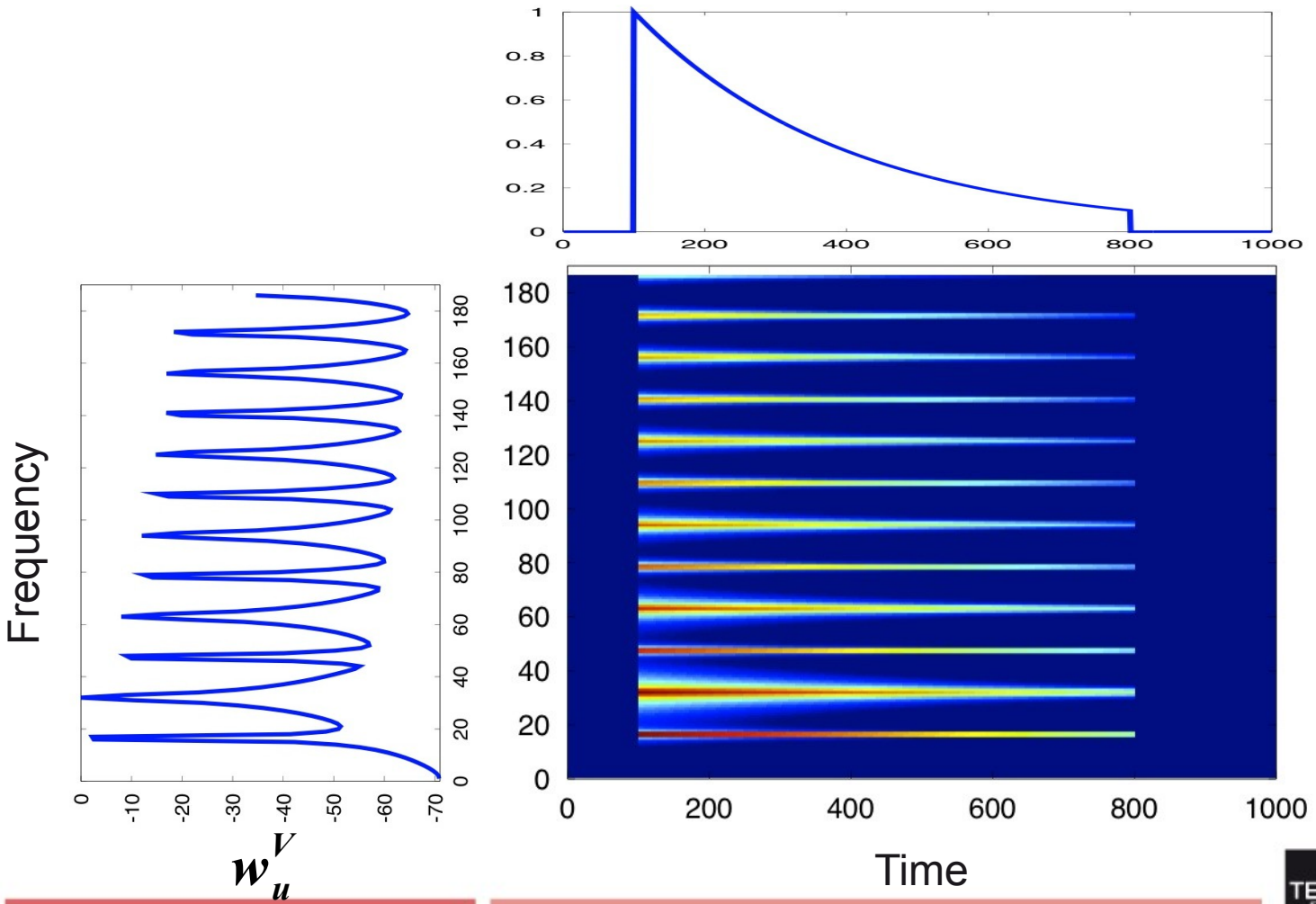
- **Hidden state mixture model:**

$$\boldsymbol{v}_n \sim \sum_{u=1}^{U} \pi_u N_c(\mathbf{0}_F, h_{un}^V \, diag(\boldsymbol{w}_u^V))$$

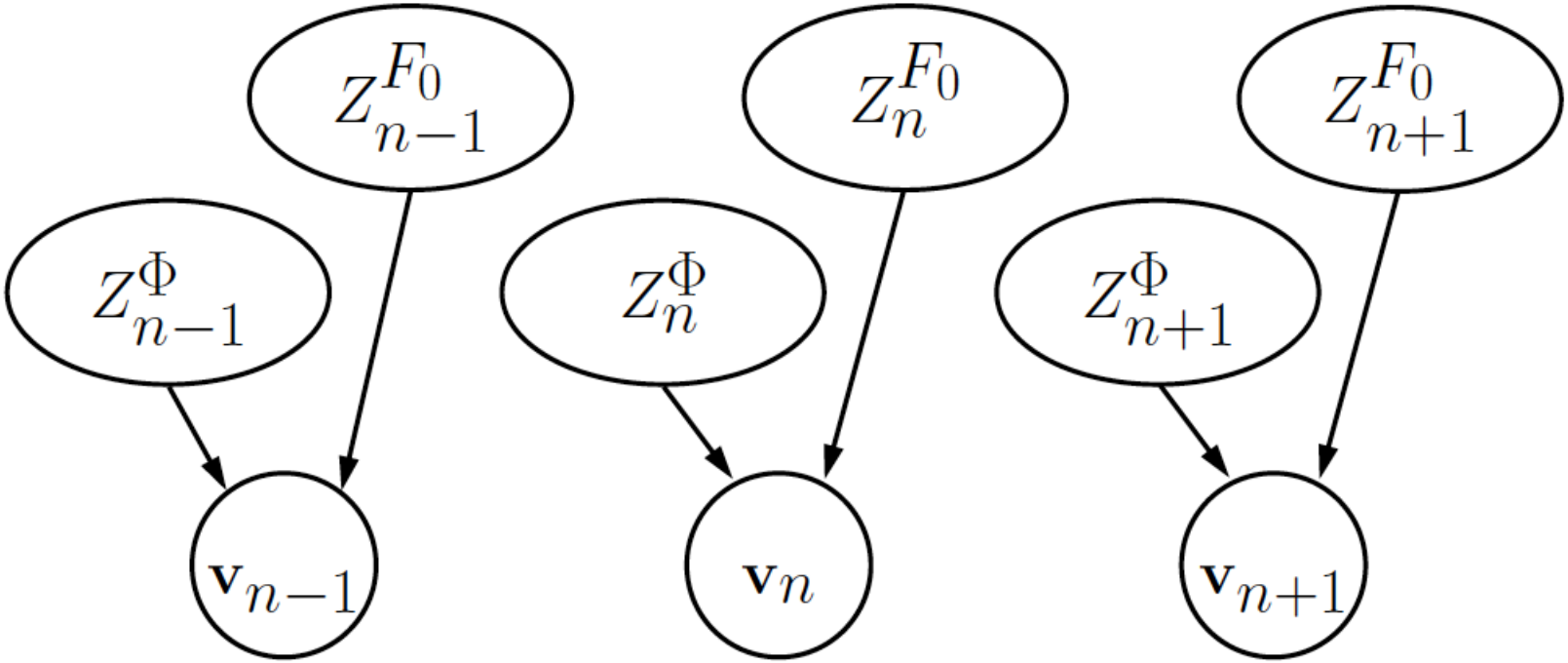
$$\boldsymbol{m}_n \sim \sum_{r=1}^{R} \pi_r N_c(\mathbf{0}_F, h_{rn}^M \, diag(\boldsymbol{w}_r^M))$$

Signal Model: Gaussian Scaled Mixture Model (GSMM)

$$\boldsymbol{v}_n | Z_n^V = u \sim N_c(\mathbf{0}_F, h_{un}^V \, diag(\boldsymbol{w}_u^V))$$



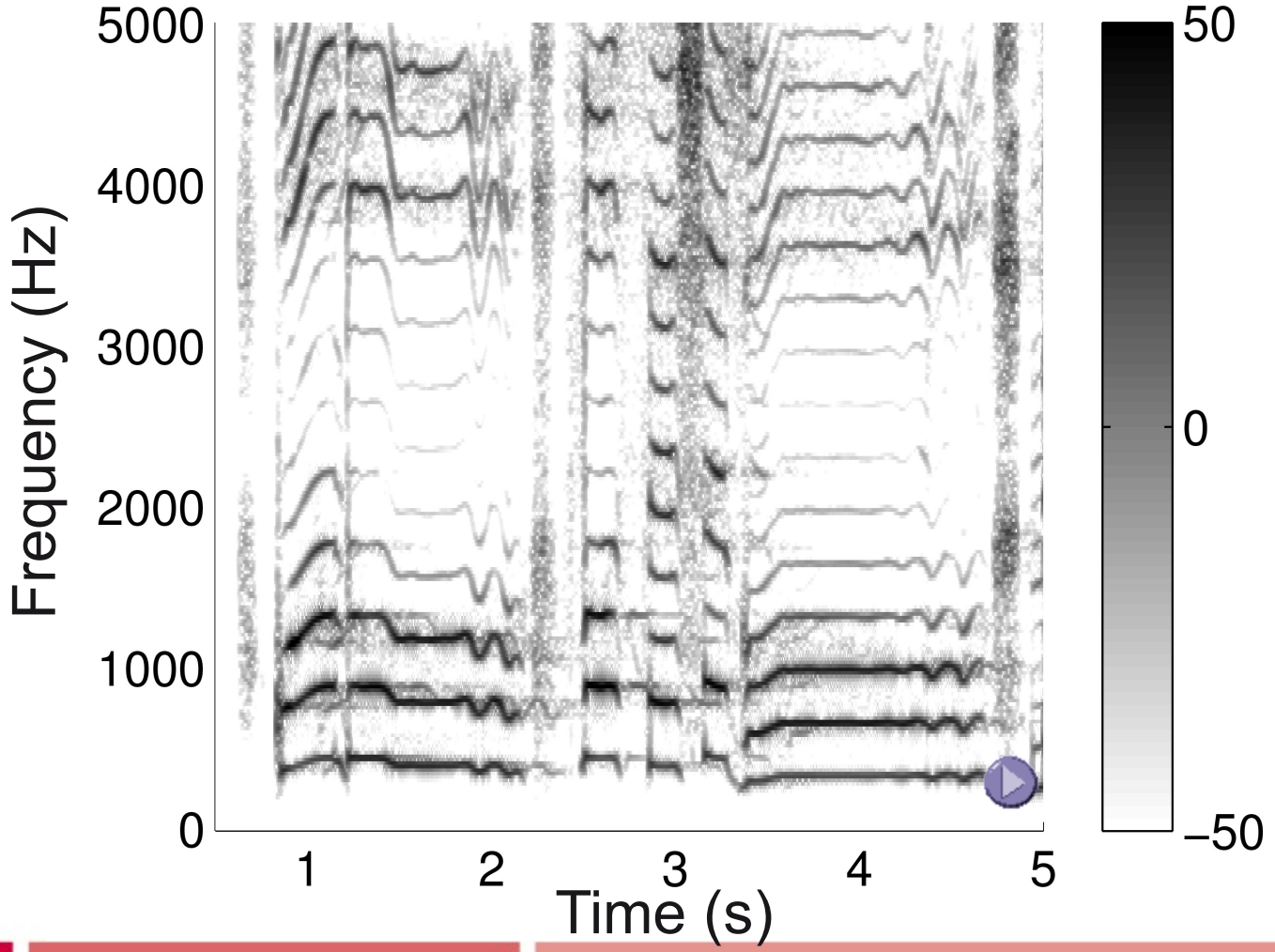
Signal Model: Source/Filter GSMM



- **Proposed model** for leading voice, **source/filter** state:
  - Filter: **smooth spectral envelope**
  - Source: **harmonic comb, parameterized with F0**

Signal Model: need for variability

A Vocal Signal (by *Tamy* - from MTG MASS database)



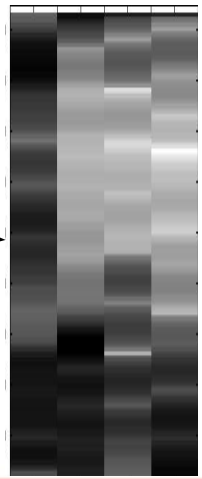
Signal Model: Source/Filter GSMM

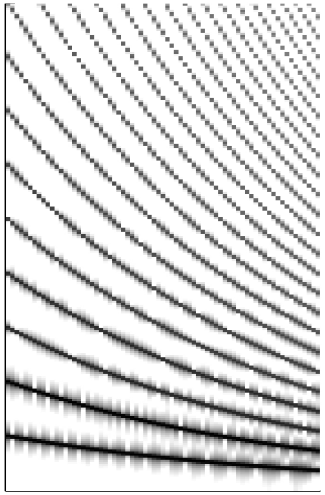
$$\boldsymbol{v}_n \left| \left(Z_n^{\Phi} = k , Z_n^{F_0} = u \right) \sim N_c \left(\mathbf{0}_F , b_{kun} \, diag \left(\boldsymbol{w}_k^{\Phi} \cdot \boldsymbol{w}_u^{F_0} \right) \right) \right.$$

Filter · Source = Source/Filter

$\boldsymbol{w}_k^{\Phi}$ $\boldsymbol{w}_u^{F_0}$ $\boldsymbol{w}_k^{\Phi} \cdot \boldsymbol{w}_u^{F_0}$

$\boldsymbol{W}^{\Phi}$ $\boldsymbol{W}^{F_0}$



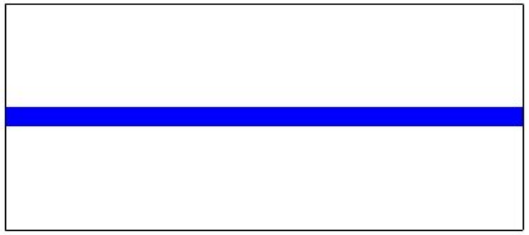


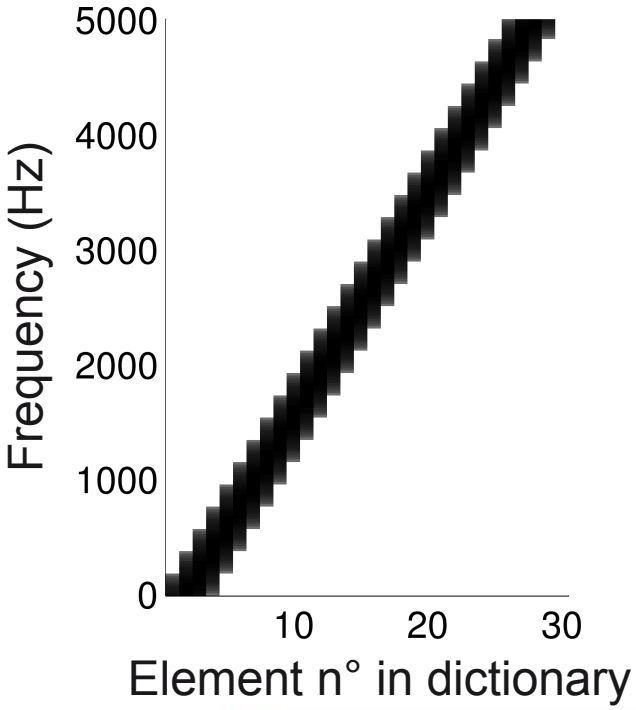
Spectral shapes matrix

Spectral Comb dictionary

Signal Model: Refining the model

- Source part:
  - Including **unvoiced** component





- Filter part:
  - **Smoothness of shapes,** decomposition on **dictionary of smooth shapes**

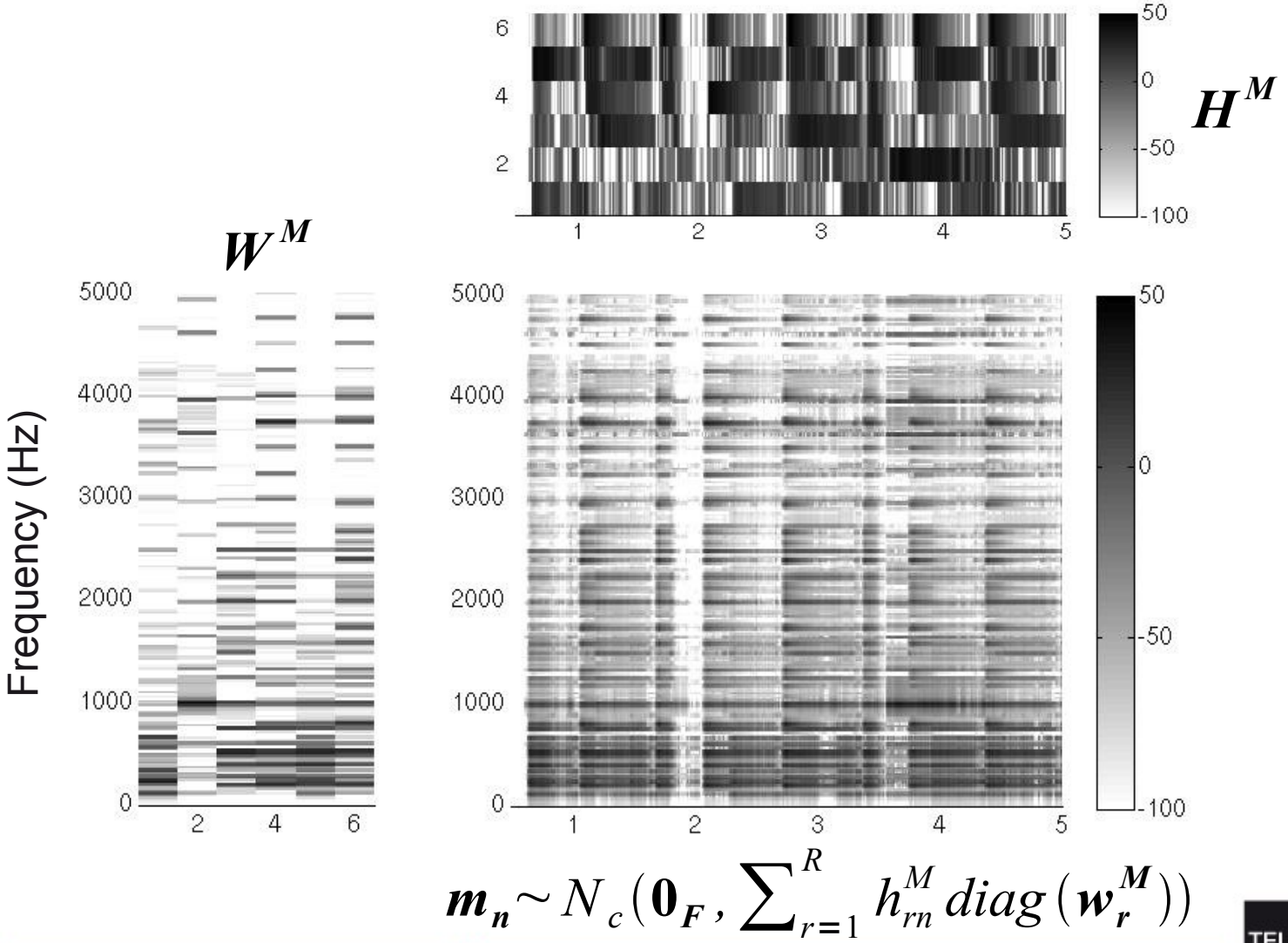
Signal Model: Accompaniment

- **Variability** of accompaniment:
  - Relatively **unconstrained** model
  - **Polyphonic/poly-instrumental** nature
  - **Repetitive** structure

- **Instantaneous mixture of Gaussian components:**
  - $$\boldsymbol{m}_n^r \sim N_c(\mathbf{0}_F, h_{rn}^M \, diag(\boldsymbol{w}_r^M))$$
    $$\boldsymbol{m}_n \sim N_c(\mathbf{0}_F, \sum_{r=1}^{R} h_{rn}^M \, diag(\boldsymbol{w}_r^M))$$

  ⇔ **Non-negative Matrix Factorisation (NMF)** with **Itakura-Saito** divergence!

Signal Model: NMF for the Accompaniment



$$\boldsymbol{m}_n \sim N_c(\mathbf{0}_F, \sum_{r=1}^{R} h_{rn}^M \, diag(\boldsymbol{w}_r^M))$$

Signal Model: GSMM for mixture

$$\boldsymbol{x_n}|k,u \sim N_c(\mathbf{0_F}, b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^M diag(\boldsymbol{w_r^M}))$$

$$\boldsymbol{x_n} \sim \sum_{k,u} \pi_{ku} N_c(\mathbf{0_F}, b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^M diag(\boldsymbol{w_r^M}))$$

- **EM** Algorithm:
  - **Maximum Likelihood** parameter estimation
  - $p(\boldsymbol{X}|\Theta), \Theta = \{\boldsymbol{W^{\Phi}}, \boldsymbol{B}, \boldsymbol{H^M}, \boldsymbol{W^M}\}$

- Estimating optimal F0 sequence:

$$\hat{Z^{F_0}} = argmax_{Z^{F_0}} p(Z^{F_0}|\boldsymbol{X})$$

Signal Model: GSMM "issues"

- **Implementation** issues:
  - Big "feature" space: **numerical problems** when computing the posterior probability
  - Consequence: **very long runtime**

- Need for a **new model:**
  - Allowing **faster** estimation
  - More **flexible** than the GSMM (constant pitch too restrictive?)
  - Keeping the **realistic** interpretation

Signal Model: Instantaneous Mixture Model (IMM)

$$\boldsymbol{x_n} \sim \sum_{k,u} \pi_{ku} N_c(\mathbf{0}_{\boldsymbol{F}}, b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^{M} diag(\boldsymbol{w_r^{M}}))$$

- Replacing **leading voice** model:
  - **Mixture of all possible states**, for any frame:

$$\boldsymbol{x_n} \sim N_c(\mathbf{0}_{\boldsymbol{F}}, \sum_{k,u} b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^{M} diag(\boldsymbol{w_r^{M}}))$$

  - Further modified into:

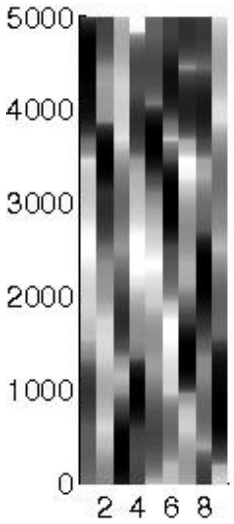
$$\boldsymbol{x_n} \sim N_c\left(\mathbf{0}_{\boldsymbol{F}}, \left(\sum_k h_{kn}^{\Phi} diag(\boldsymbol{w_k^{\Phi}})\right) \cdot \left(\sum_u h_{un}^{F_0} diag(\boldsymbol{w_u^{F_0}})\right) + \sum_r h_{rn}^{M} diag(\boldsymbol{w_r^{M}})\right)$$

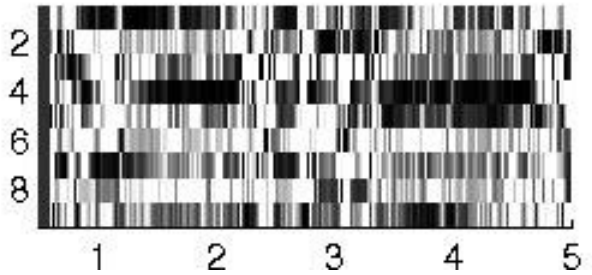
$$\downarrow$$

$$\boldsymbol{X} \sim N_c\left(\mathbf{0}_{\boldsymbol{F}\times N}, \left(\boldsymbol{W^{\Phi}}\,\boldsymbol{H^{\Phi}}\right) \cdot \left(\boldsymbol{W^{F_0}}\,\boldsymbol{H^{F_0}}\right) + \boldsymbol{W^{M}}\,\boldsymbol{H^{M}}\right)$$

Signal Model: IMM

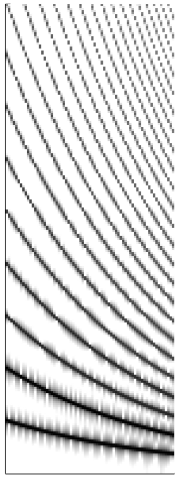
$$X \sim N_c\left(\mathbf{0}_{F\times N}, \left(W^{\Phi} H^{\Phi}\right)\cdot\left(W^{F_0} H^{F_0}\right) + W^{M} H^{M}\right)$$

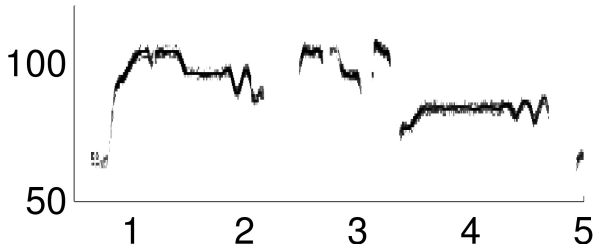




$W^{\Phi} H^{\Phi}$

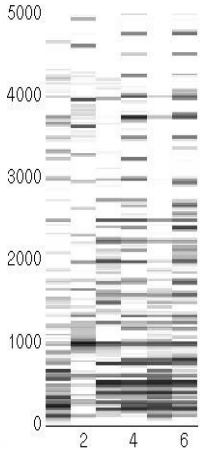
·

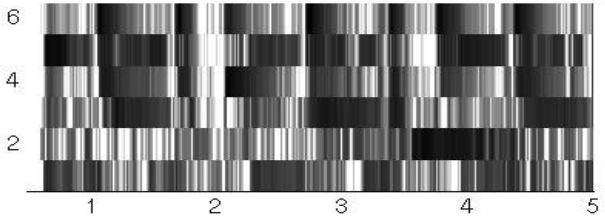




$W^{F_0} H^{F_0}$

+





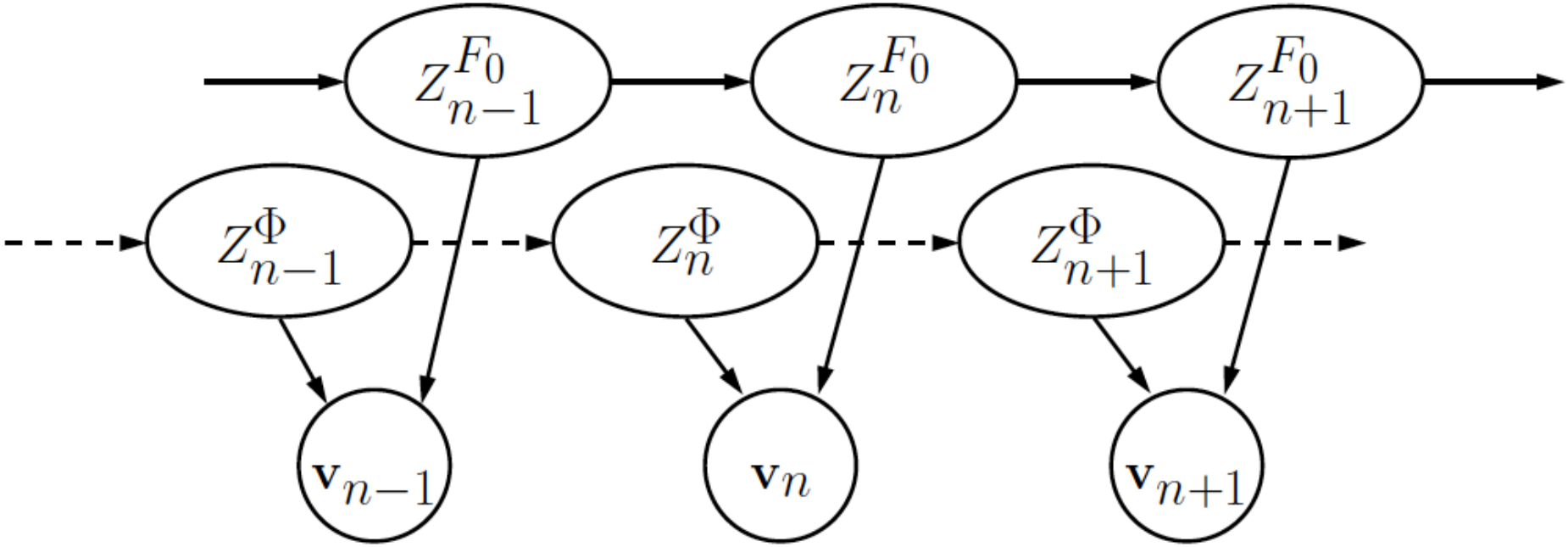
$W^{M} H^{M}$

Signal Model: IMM and GSMM

- **Advantages over GSMM:**
  - No hidden state: no **EM** needed!
  - Fast, but still **good interpretation** of parameters H^{F_0}
  - Link with GSMM
- **IMM limitations:**
  - Modelling **monophonic** signal with **polyphonic** model...
  - **Octave errors** easier: redundant representation

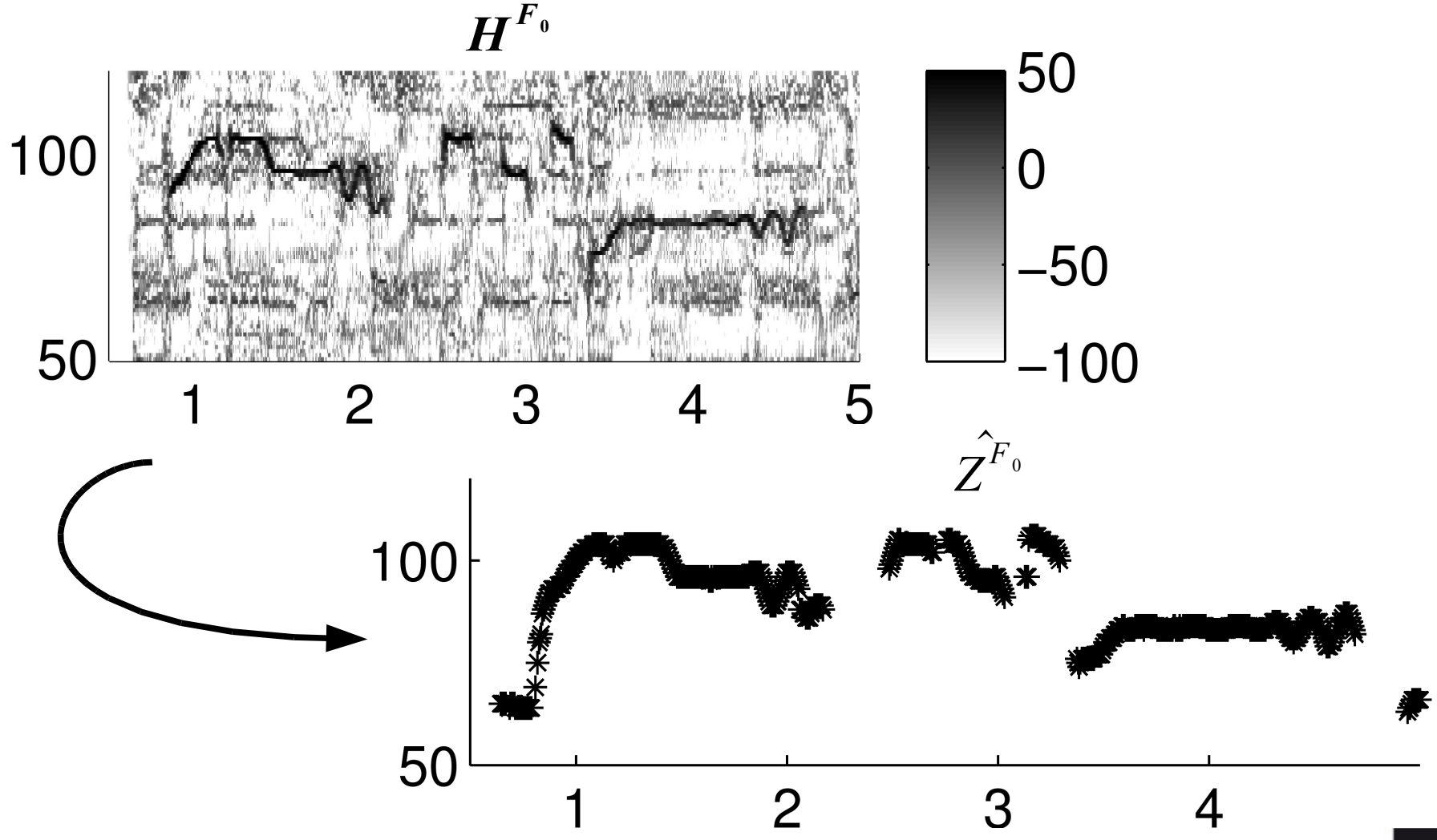
- Addressing these issues:
  - Defining an **HMM** (hidden Markov model) on
  - **Re-weighting**, favoring lower octave

Signal Model: HMM to model smoothness



- Assumptions on the melody F0 sequence:
  - **Smooth**,
  - **Predominant** as concerns the **energy**, $\boldsymbol{B}$ or $\boldsymbol{H}^{F_0}$
  - Realistic melody line: **trade-off** between the smoothness and the energy of the line.

Signal Model: Viterbi melody tracking



Signal Model: Parameter estimation

- **NMF** methodology both for GSMM and IMM:
  - **Partial derivatives**
  - **Multiplicative gradients**

- **Drawbacks:**
  - **Slow "convergence"**
  - **Convergence?**
  - **Initialization** sensitive

- **Advantages:**
  - Fast implementation
  - Easy extension to other types of optimisation

Signal Model: conclusion

- **Source/filter GSMM:**
  - **Mixture model:**

$$\boldsymbol{x_n} \sim \sum_{k,u} \pi_{ku} N_c(\mathbf{0_F}, b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^{M} diag(\boldsymbol{w_r^{M}}))$$

  - **EM** algorithm: **slow, unstable**

- **IMM:**
  - **Composite model:**

$$\boldsymbol{x_n} \sim N_c(\mathbf{0_F}, \sum_{k,u} b_{kun} diag(\boldsymbol{w_k^{\Phi}} \cdot \boldsymbol{w_u^{F_0}}) + \sum_r h_{rn}^{M} diag(\boldsymbol{w_r^{M}}))$$

  - Less **realistic** than **GSMM**
  - **Fast** algorithm

- For both: Viterbi algorithm to **track main melody**

Applications: Transcription and separation

Transcription of the main melody

- Estimation of $\hat{Z}^{F_0}$, frame-wise result.

- **MIREX 2008/2009** results on **Audio Melody Extraction (AME):**
  - **drd1 = GSMM, drd2 = IMM**
  - MIREX **2008: without smooth filters**
  - MIREX **2009: with smoothness**

- **Database:**
  - **ADC 2004 (various), MIREX 2005 (pop), MIREX 2008 (indian), MIR-1K (Chinese karaoke)**

Transcription of the main melody: results

Participant	Avg. Overall Acc.
clly1	49.80%
clly2	62.10%
drd1	**58.60%**
drd2	**73.20%**
pc	76.10%
rk	71.10%
vr	67.10%

MIREX 2008 AME Results

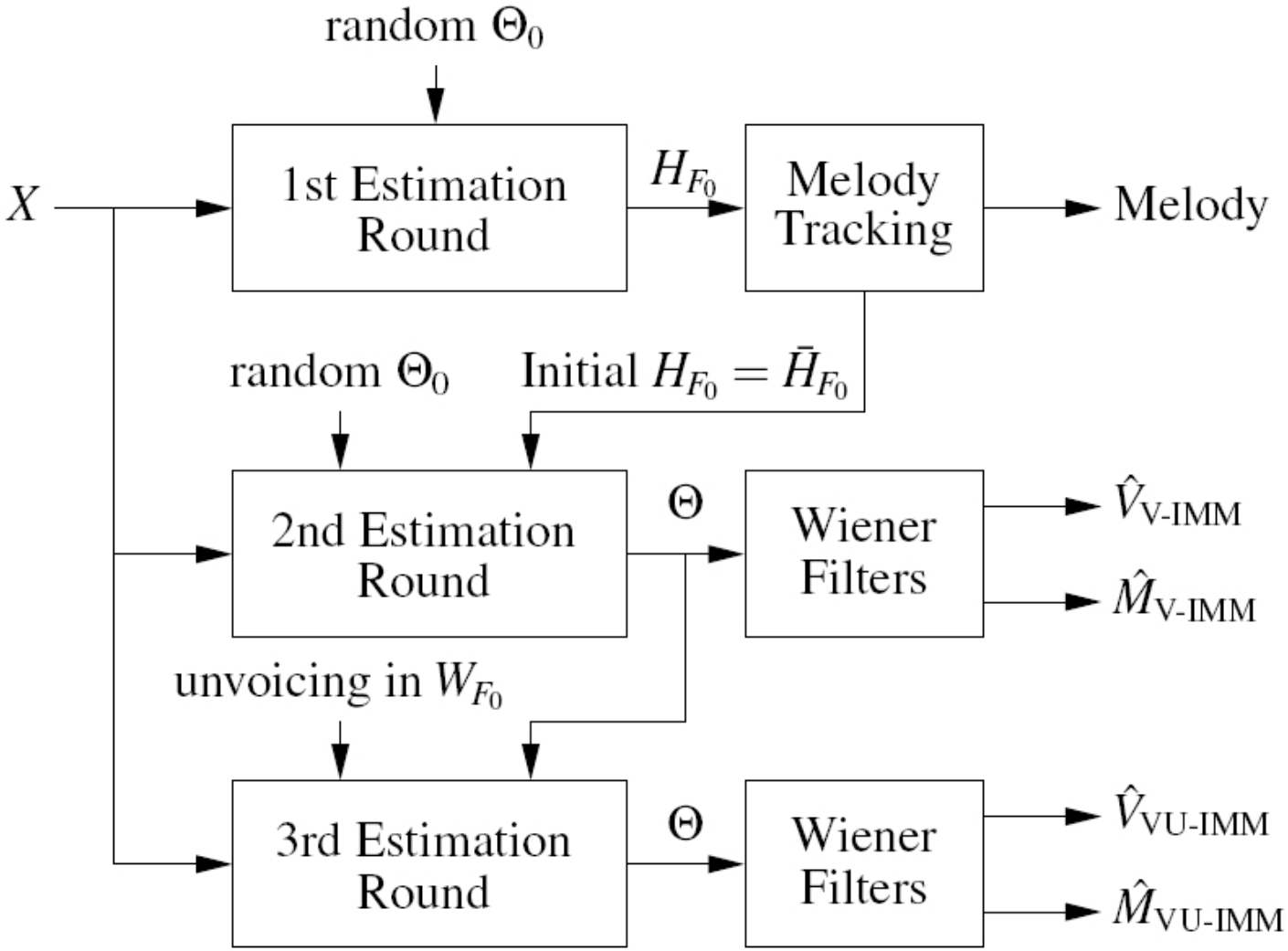
Participant	Raw Pitch (%)	Raw Chroma (%)	Overall Acc(%)
cl1	63.45	66.29	52.19
cl2	63.45	66.29	55.19
drd1	**74.45**	**76.82**	**66.86**
drd2	**72.09**	**75.72**	**66.17**
hjc1	66.12	72.58	50.49
hjc2	51.13	67.12	49.01
jjy	73.33	79.68	56.64
kd	80.58	82.52	73.35
mw	73.44	77.5	55.07
pc	64.1	65.84	62.88
rr	72.21	76.33	65.22
toos	75.05	80.34	55.08

MIREX 2009 AME Results

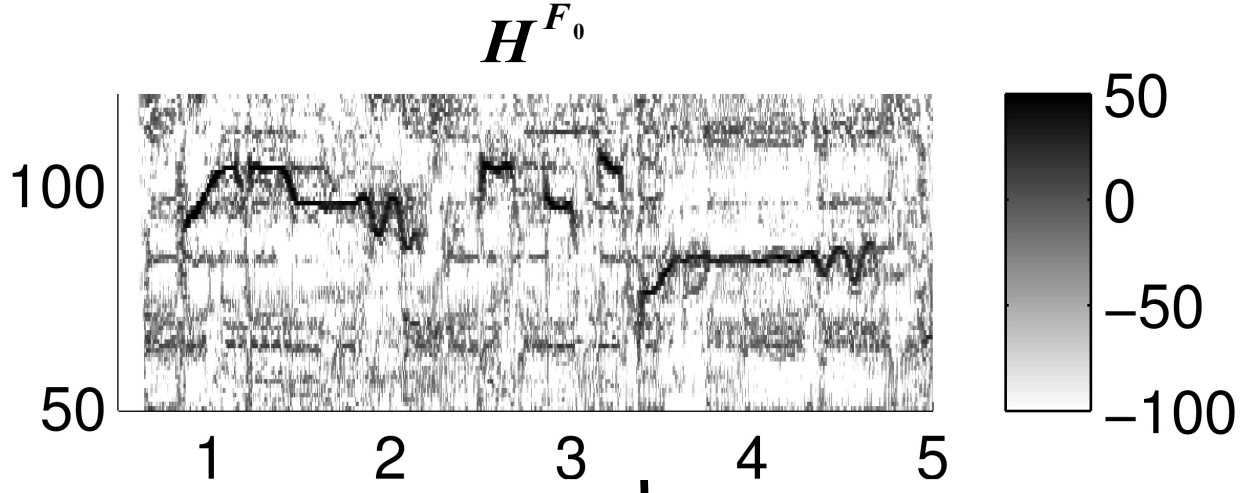
Leading Instrument separation

- **Definition:**
  - **“leading instrument”:** the track played by the main instrument, with the main melody,
  - **“Accompaniment”:** the remaining other background instruments.
  - Separate these two contributions and obtain their images.
- **MIR-aided** approach:
  - First step: **melody tracking**, using **IMM**,
  - Second step: re-estimation of the parameters **knowing the melody**,
  - (Third step: re-estimation including **unvoiced parts**)

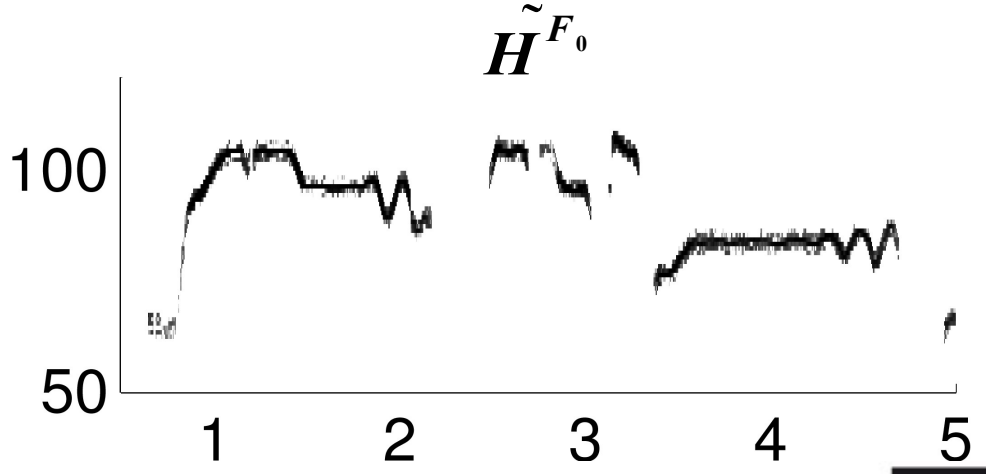
Leading Instrument separation: system



Leading Instrument separation: Parameter estimation knowing the melody



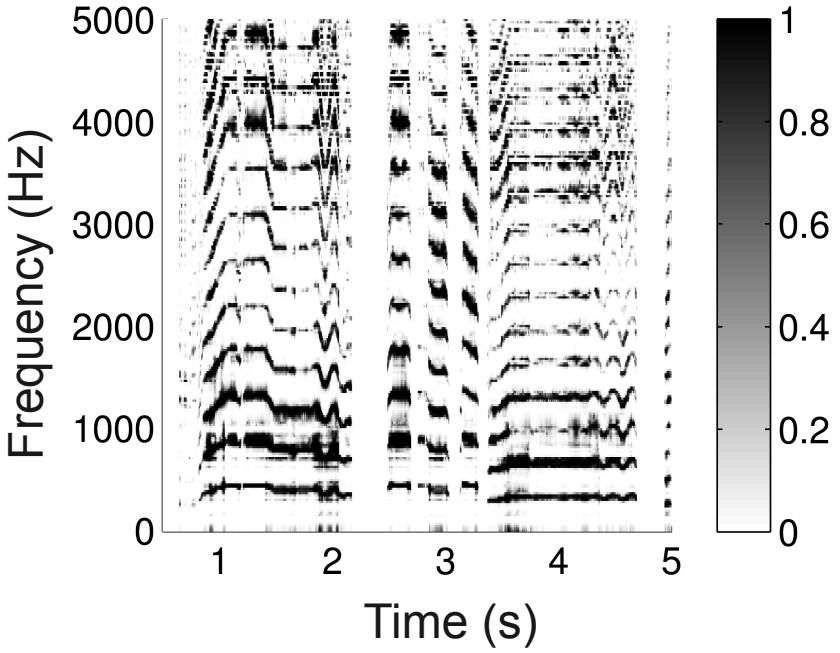
Melody estimation
+ hard constraint



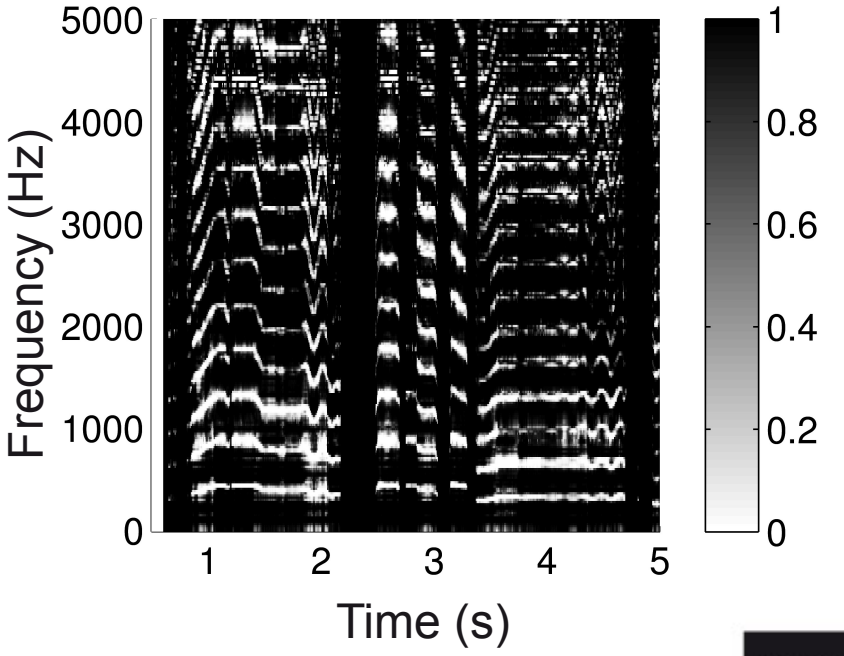
Leading Instrument separation: Adaptive Wiener filtering

■**Wiener filtering:** $\hat{V} = \underbrace{\frac{S^V}{S^V + S^M}}_{Wiener\ mask} X$

Leading voice Wiener mask



Accompaniment Wiener mask



Leading Instrument separation: results

- **ICASSP 2009:**
  - + 8 dB SDR for the estimated singing voice,
  - + 2 dB SDR for the accompaniment extraction.
- **SiSEC** “Professionally produced music recordings” (http://sisec.wiki.irisa.fr/) $\hat{V}$: $\hat{M}$:
  - Interesting result: on the excerpt by “Tamy”, flute+guitar, **best results** for algorithms who **first estimate the melody**.
- Some **sound examples** on:
  - http://perso.enst.fr/durrieu/en/results_en.html
  - http://perso.enst.fr/durrieu/en/icassp09/
  - http://perso.enst.fr/durrieu/en/eusipco09/

Lead/Accompaniment separation: Applications/Extensions

- **MIR applications** (MIREX 2008):
  - Pre-processing for multipitch estimation,
  - Accompaniment enhancement for Chord detection,

- Other **extensions**:
  - **Stereophonic** signals: article at Eusipco 2009,
  - Enhancing **discrimination of main instrument** by classification methods,
  - Adding **constraints** (*priors*) to the parameters, avoiding several steps to achieve separation.

Conclusions

- **GSMM:**
  - **Well suited** for main melody transcription
  - Results in **source separation** to be assessed
- **IMM:**
  - **Robust** main melody transcription
  - **State-of-the-art results** in **separation**
  - Model less **realistic** than **GSMM**

Conclusions, perspectives

- Conclusions:
  - **Source/filter model** for **singing voice**
  - **NMF** based parameter estimation
  - **State-of-the-art** for transcription and separation

- Perspectives:
  - **Constraints on the parameters:** smoothness, sparseness, regularity, etc.
  - **Joint estimation** of parameters and sequences
  - Better estimation algorithms?

Publications

- Journal articles:
  - C. Févotte, N. Bertin and **J.-L. Durrieu**, *"Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis"*, Neural Computation, 2009,
  - **J.-L. Durrieu,** G. Richard, B. David and C. Févotte, *"Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals"*, accepted to IEEE Trans. on Audio, Speech and Language Processing.
- Conference papers:
  - **J.-L. Durrieu**, G. Richard and B. David, *"Singer melody extraction in polyphonic signals using source separation methods"*, ICASSP 2008,
  - **J.-L. Durrieu**, G. Richard and B. David, *"An Iterative Approach to Monaural Musical Mixture De-Soloing"*, ICASSP 2009,
  - **J.-L. Durrieu**, A. Ozerov, C. Févotte, G. Richard and B. David, "*Main instrument separation from stereophonic audio signals using a source/filter model"*, EUSIPCO 2009.