Automatic Extraction of the Main Melody from Polyphonic Music Signals

With Application to Transcription and Separation

Jean-Louis Durrieu, PhD candidate

TSI Department

Institut Télécom; Télécom ParisTech; CNRS LTCI

04/12/09

Automatic Extraction of the Main Melody from Polyphonic Music Signals

With Application to Transcription and Separation

Automatic Extraction of the Main Melody from Polyphonic Music Signals

- Introduction
- Signal Model
- Transcription and Separation Systems
- Results
- Conclusion

Introduction: MIR and BASS

- **Music Information Retrieval (MIR):**

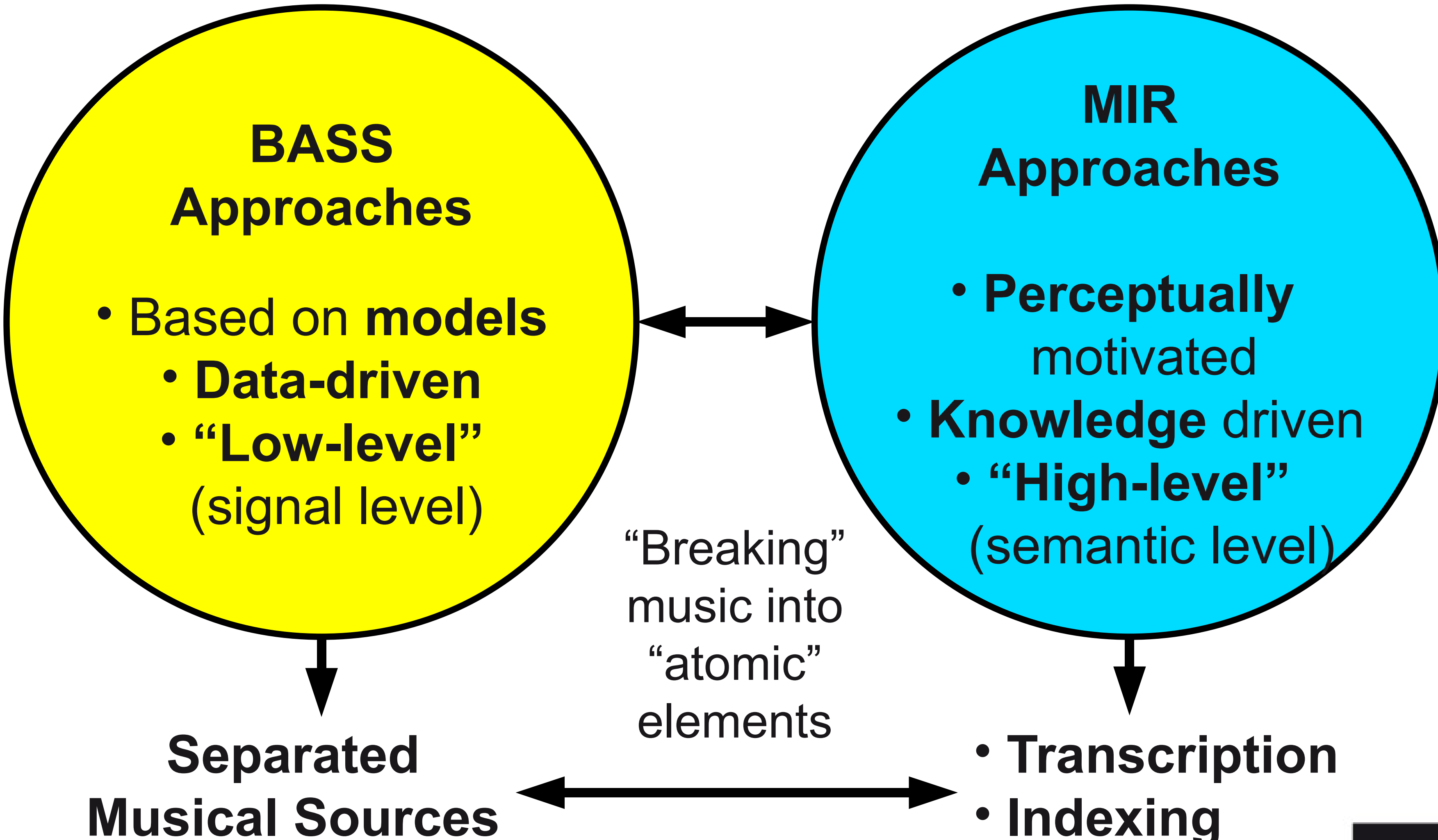
  Interdisciplinary: music libraries, HCI, retrieval, **music analysis** (symbolic, **audio**)

- **(Blind) Audio Source Separation (BASS):**

  Isolating the different "sources" that are present in an audio mixture.

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

Introduction: Bridging MIR and BASS



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**

Introduction: Applications

- **Transcribed Melody**
  - **Indexing** large music database,
  - **Musical transcription** into “human readable” score,
  - **Feature** (singing voice detection)

- **Separating the Main Instrument from the Accompaniment:**
  - **Generate accompaniments** for solo performers (“Karaoke”...)
  - **Pre-Processing** for MIR applications (chord detection, instrument classification, etc.)
  - ...

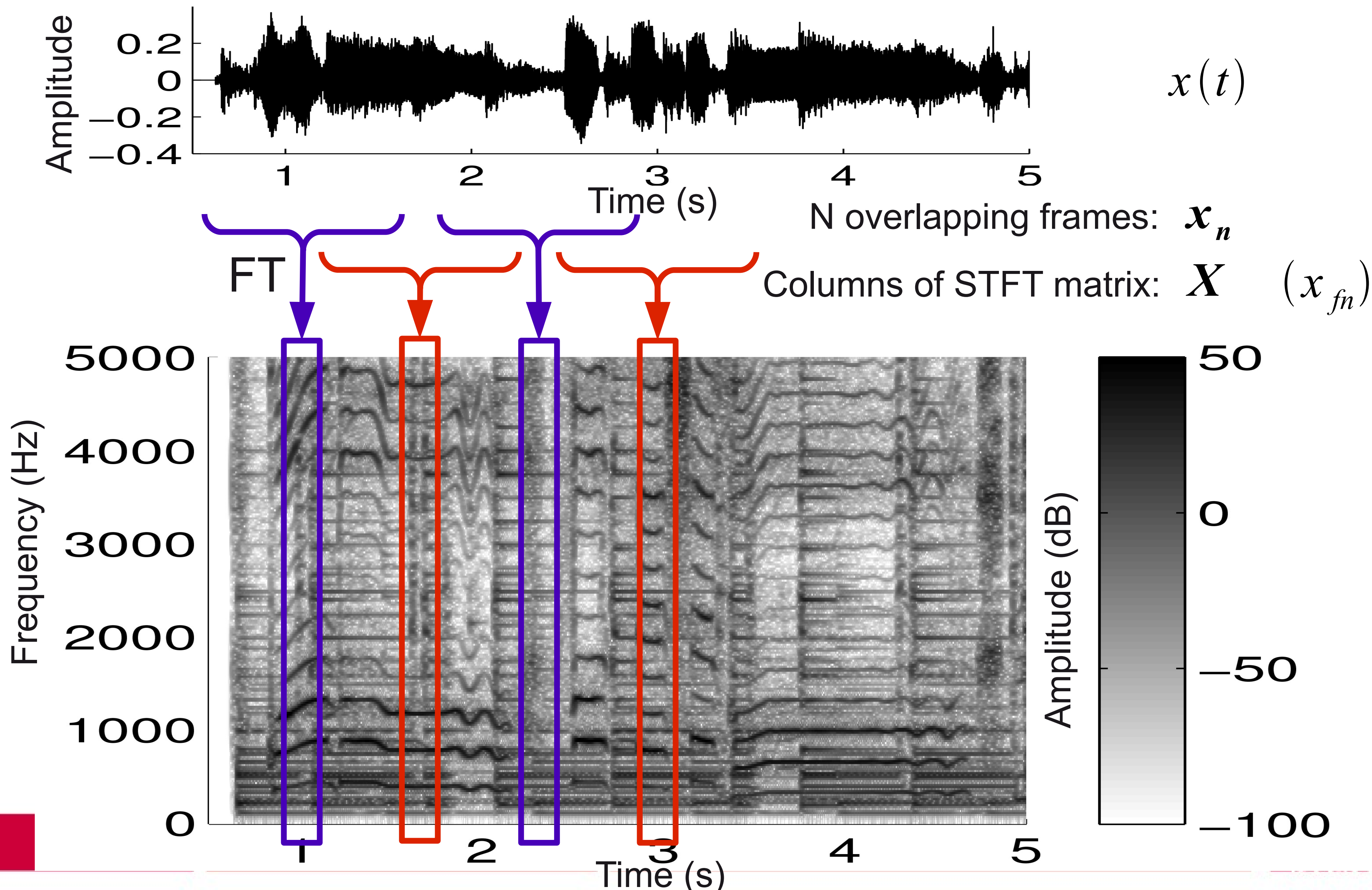
Organization of this presentation

- **Signal Model(s):**
  - **Source/Filter Gaussian Scaled Mixture Model (GSMM)**
  - **Instantaneous Mixture Model (IMM)**: “approximation”
  - **Temporal Constraint**
- **Systems** for **Transcription/Separation:**
  - **Parameter estimation** principle
  - **GSMM** based system
  - **IMM** system
- **Results:**
  - **Transcription**
  - **Separation**

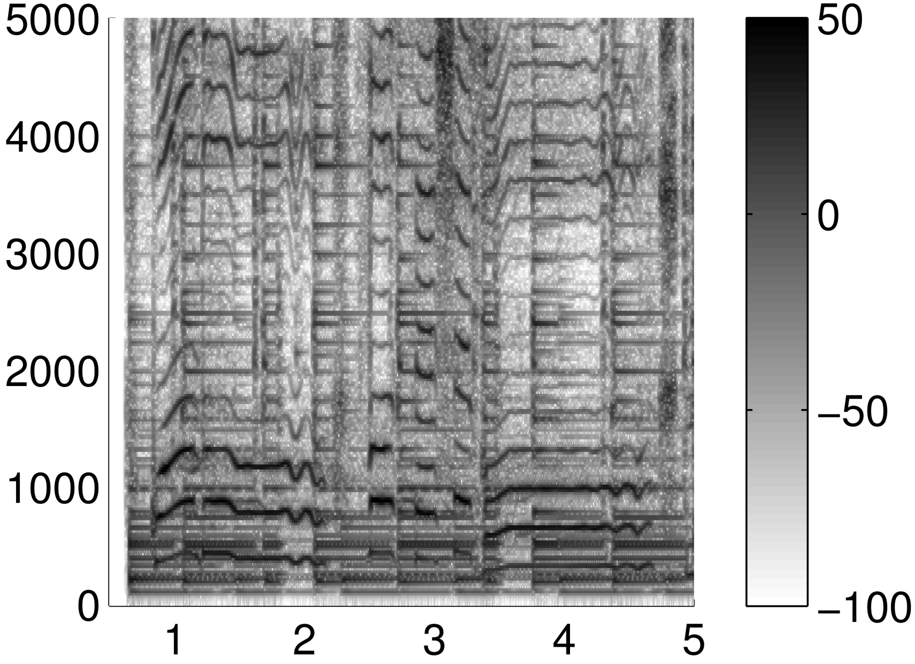
Signal Model

- Gaussian Scaled Mixture Model (GSMM)
- Instantaneous Mixture Model (IMM)
- Temporal Constraint

Signal Model: Short-Time Fourier Transform (STFT)



Signal Model: Framework (Benaroya)

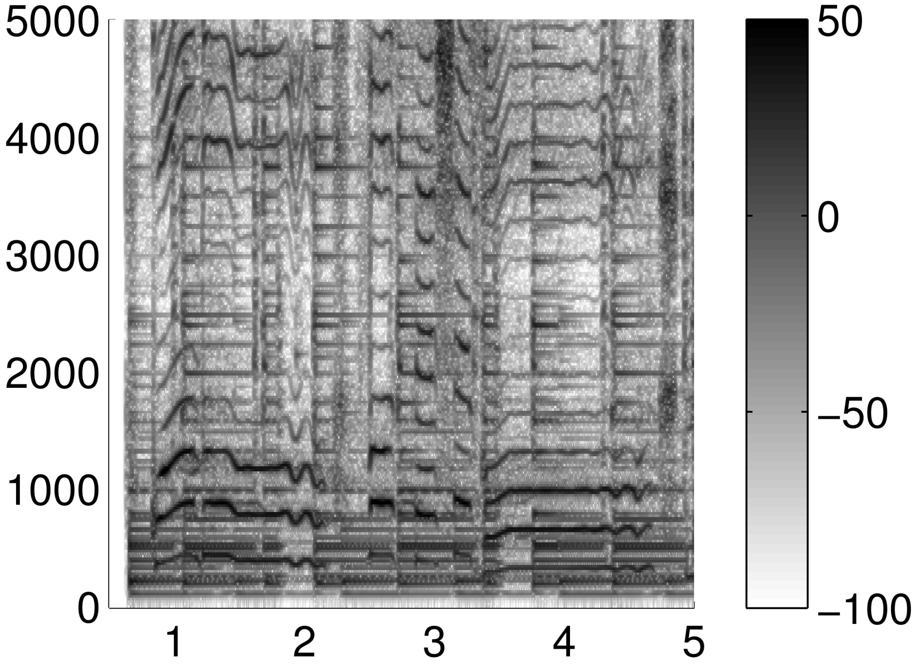


- **Spectral model:**
  - Time-Frequency representation**: STFT**
  - **Invertible**
  - **Limited resolution**
- **Statistical model:**
  - FT vector = **Gaussian** vector

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

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

Signal Model: mixture



- **Instantaneous Mixture:**
  - $X = V + M$
- **Contributions characterized by variance:**
  - $\boldsymbol{v}_n \sim N_c(\boldsymbol{0}_F, diag(\boldsymbol{s}_n^V))$
  - $\boldsymbol{m}_n \sim N_c(\boldsymbol{0}_F, diag(\boldsymbol{s}_n^M))$
  - **Independence:**

$$\boldsymbol{x}_n \sim N_c(\boldsymbol{0}_F, diag(\boldsymbol{s}_n^V + \boldsymbol{s}_n^M))$$

Signal Model: complex Gaussians

Advantages:

- **Interpretation** of parameters: $-\log N_c(\boldsymbol{v_n}; \boldsymbol{0_F}, diag(\boldsymbol{s_n^V}))$

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

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

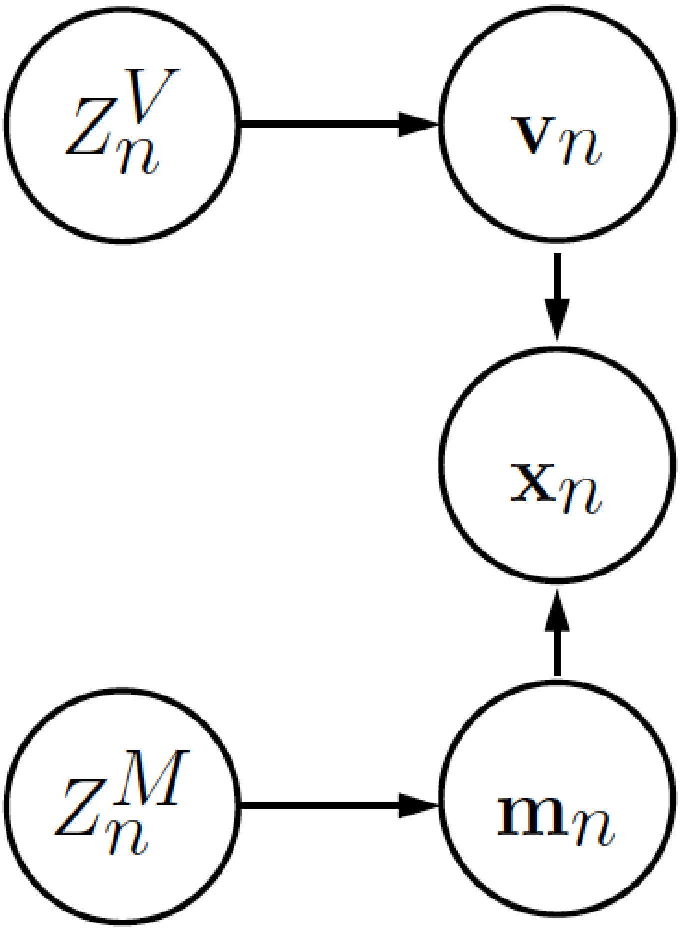
$$\hat{\boldsymbol{v}}_{\boldsymbol{n}} =^d E[\boldsymbol{v_n} | \boldsymbol{x_n} = \boldsymbol{v_n} + \boldsymbol{m_n}],\ s.t.\ \hat{v}_{fn} = \left(\frac{s_{fn}^V}{s_{fn}^V + s_{fn}^M}\right) x_{fn}$$

- **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**

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

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

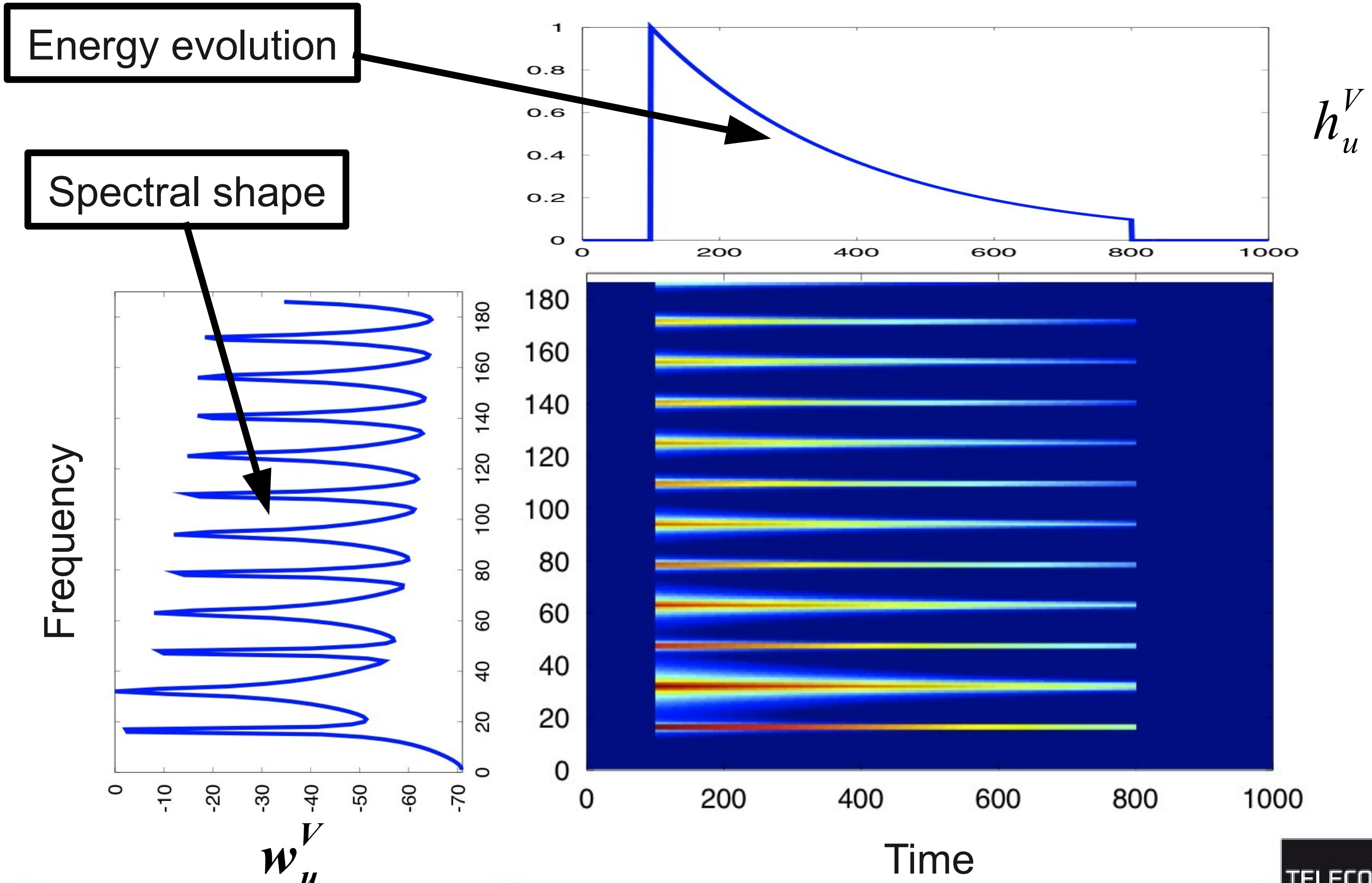
- **Hidden state mixture model:**

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

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

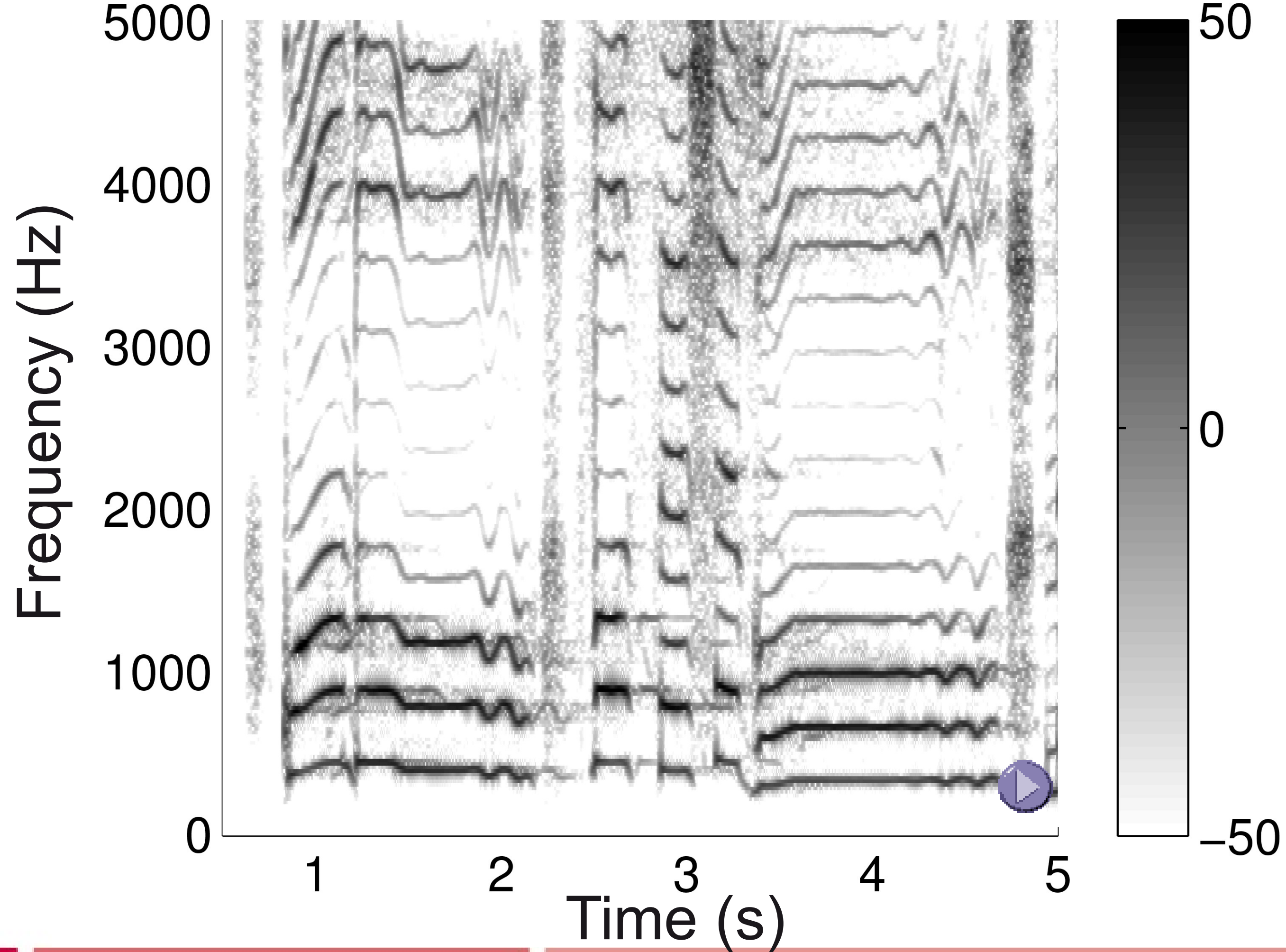
Signal Model: Interpretation for GSMM

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



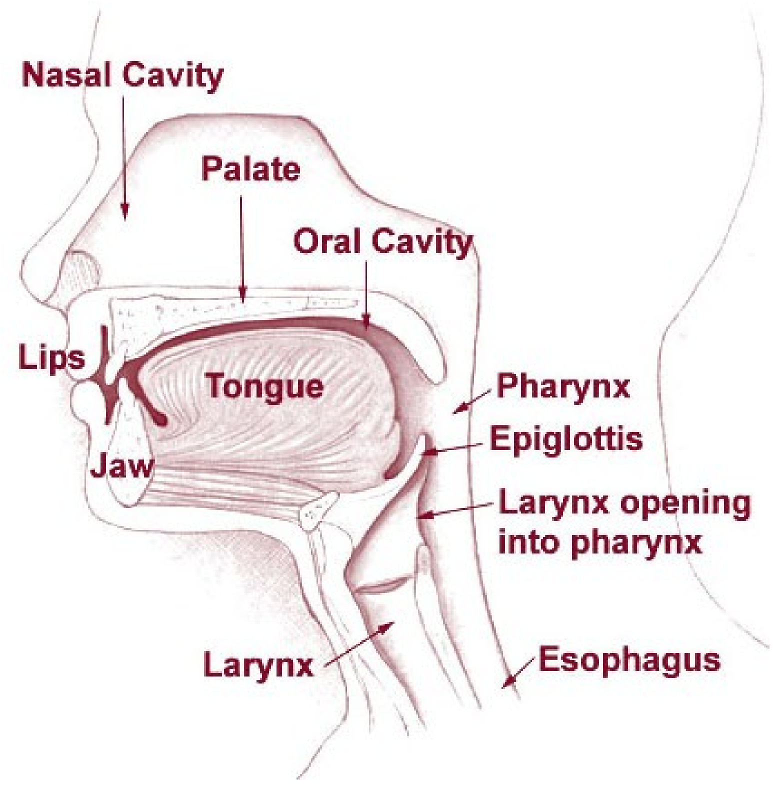
Signal Model:
Leading instrument, need for variability

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



Signal Model: Source/Filter for leading voice

- **Motivations:**
  - **Singing voice** often main instrument,
  - **Source/Filter** widely used, suitable for wide range of other instruments,
  - **Separately** modelling **pitched** aspects (source) from **timbre** aspects (filter).



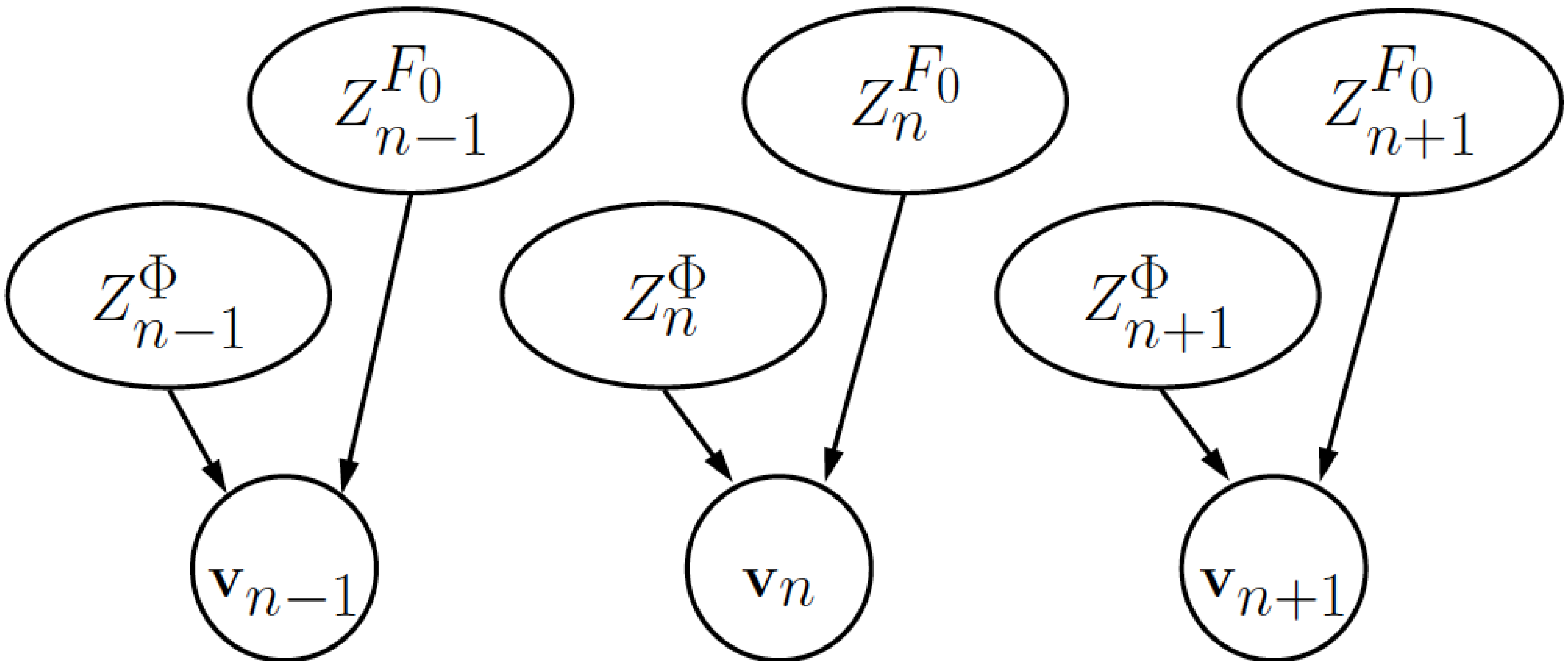
(Glottal) Source → (Vocal Tract) Filter → Voice Signal

Signal Model:
Lead instrument modelled as a singer

- **Human singer:**
  - **Independent evolution** of pitches and filters (vowel),
  - **Continuous pitch (F0)** variations,
  - **Limited set** of vowels,
  - **Unvoiced** parts...
- **Proposed Model** for **Main Instrument:**
  - **Discrete range** of possible F0 for **voiced** source component, log-spaced s.t. 96 F0 per octave,
  - Limited number of **"smooth" filters**,
  - **Unvoiced** source component (integrated **later** in the estimation process).

Signal Model: Source/Filter GSMM

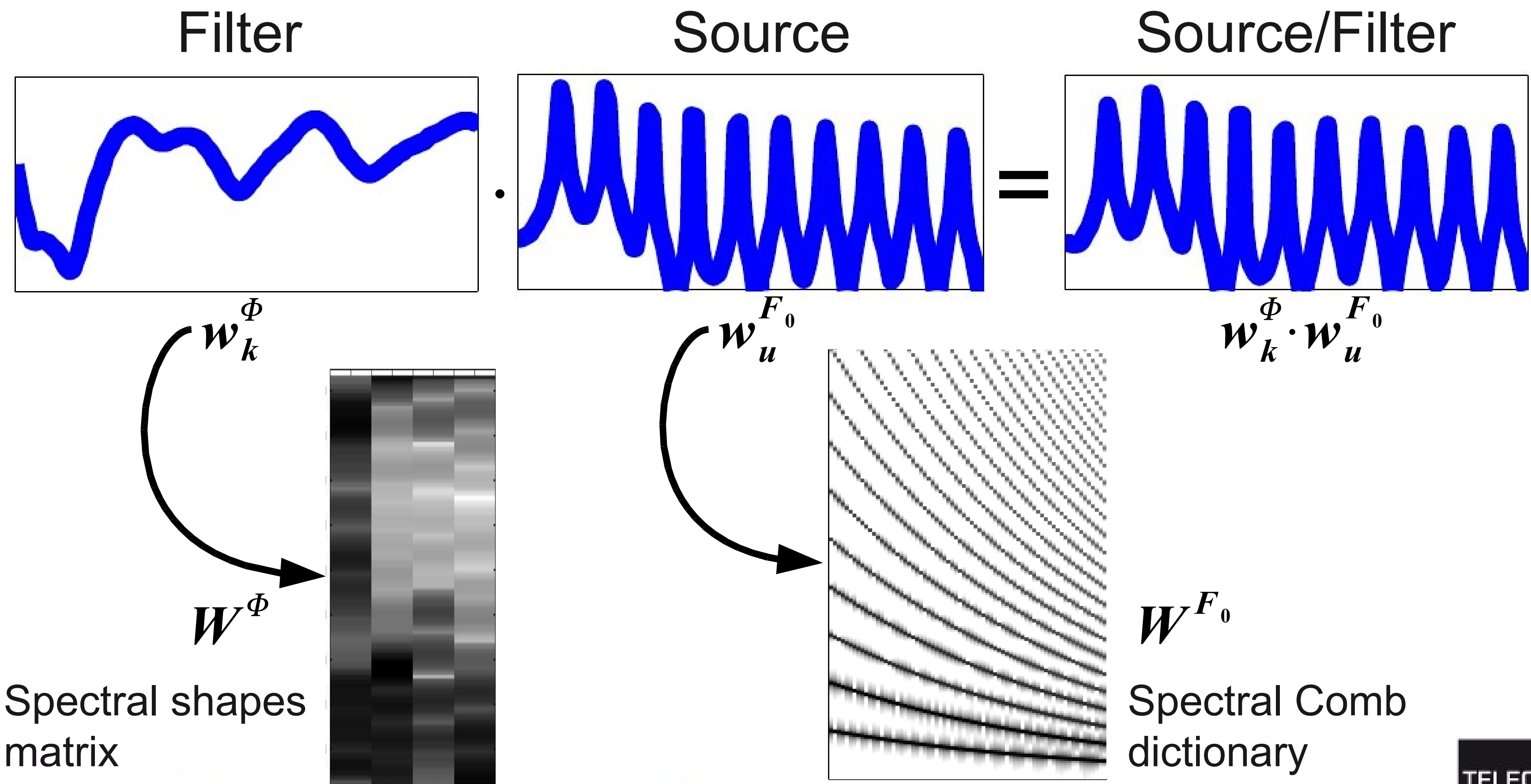
- **Proposed model** for leading voice, **source/filter** state:
  - Filter: **smooth spectral envelope**
  - Source: **harmonic comb, parameterized with F0**



$$\boldsymbol{v}_{\boldsymbol{n}} | (Z_n^{\Phi}=k, Z_n^{F_0}=u) \sim N_c(\boldsymbol{0}_{\boldsymbol{F}}, b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\boldsymbol{\Phi}} \cdot \boldsymbol{w}_{\boldsymbol{u}}^{\boldsymbol{F_0}}))$$

Signal Model: Source/Filter GSMM

$$\boldsymbol{v}_{\boldsymbol{n}} | (Z_n^{\Phi}=k,\, Z_n^{F_0}=u) \sim N_c(\mathbf{0}_{\boldsymbol{F}},\, b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\boldsymbol{\Phi}} \cdot \boldsymbol{w}_{\boldsymbol{u}}^{\boldsymbol{F}_0}))$$



Signal Model: Model details

- **Voiced source** component:
  - KLGLOTT88 (Glottal source) model, [Klatt90]: **spectral comb dictionary** $\boldsymbol{W}^{F_0}$, U "notes"
  - Freq. bin f, Pitch n. u : power spectrum $\boldsymbol{w}_u^{F_0}$

- **Unvoiced source**
  - In dictionary $\boldsymbol{W}^{F_0}$, "unvoiced" component such that:

$$w_{fu}^{F_0} \propto 1, \forall f$$

- **Smooth filters:**
  - Decomposition of filter part on dict. of smooth functions:

$$\boldsymbol{w}_k^{\Phi} = \sum_p h_{pk}^{\Gamma} \boldsymbol{w}_{fp}^{\Gamma} \Leftrightarrow \boldsymbol{W}^{\Phi} = \boldsymbol{W}^{\Gamma} \boldsymbol{H}^{\Gamma}$$

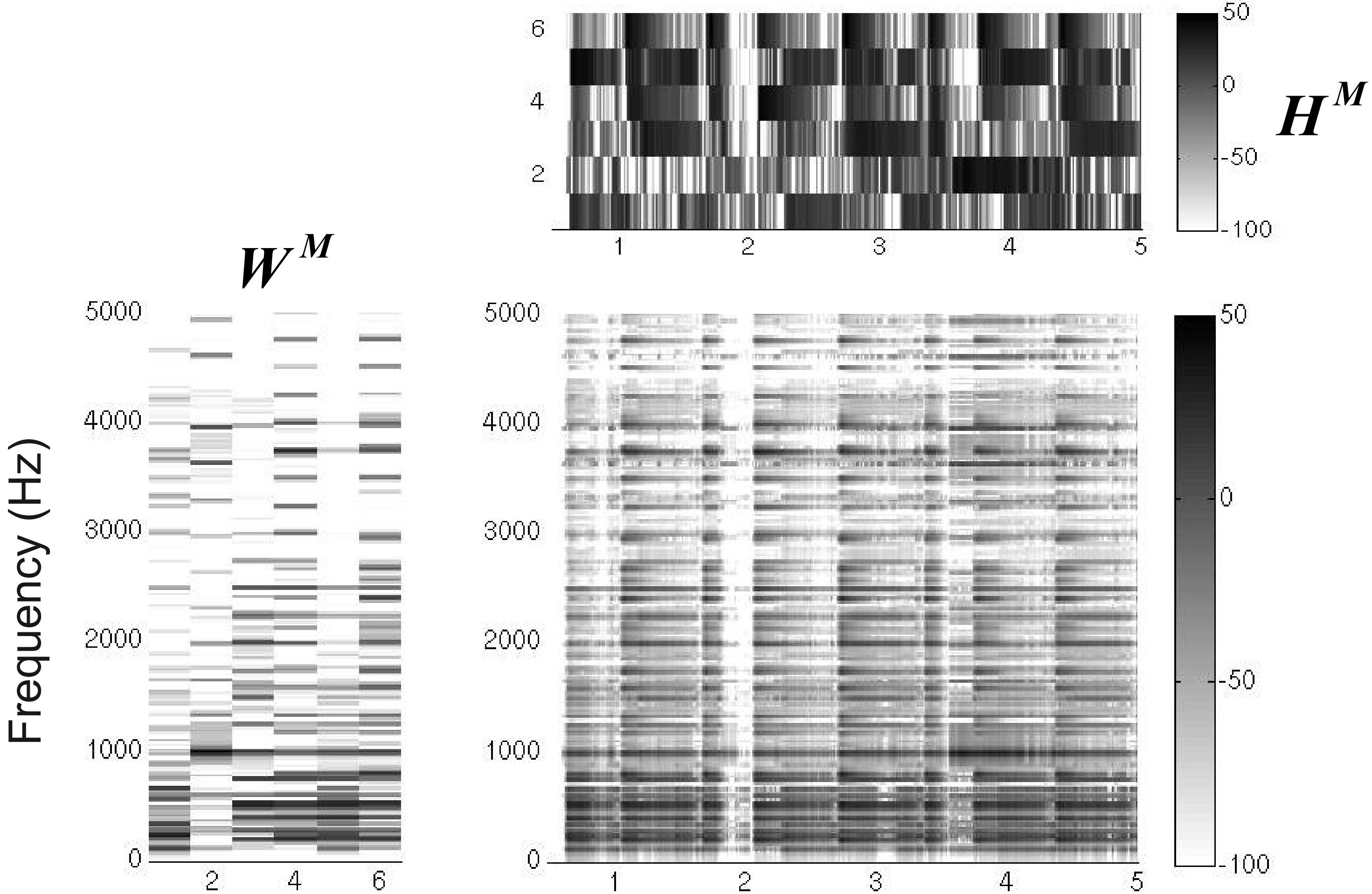
Signal Model: Accompaniment

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

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

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

Signal Model: NMF for the Accompaniment



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

Signal Model: GSMM for mixture

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

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

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

$$C_{GSMM}(\theta, \theta') = E_{\theta'}\left[\log p_{\theta}\left(\boldsymbol{X}, Z^{\Phi}, Z^{F_0}\right) | \boldsymbol{X}\right]$$

- Estimating optimal F0 sequence:

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

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)

$$x_n \sim \sum_{k,u} \pi_{ku} N_c\left(\mathbf{0}_F, b_{kun}\, diag\left(\boldsymbol{w}_k^{\Phi} \cdot \boldsymbol{w}_u^{F_0}\right) + \sum_r h_{rn}^{M}\, diag\left(\boldsymbol{w}_r^{M}\right)\right)$$

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

$$x_n \sim N_c\left(\mathbf{0}_F, \sum_{k,u} b_{kun}\, diag\left(\boldsymbol{w}_k^{\Phi} \cdot \boldsymbol{w}_u^{F_0}\right) + \sum_r h_{rn}^{M}\, diag\left(\boldsymbol{w}_r^{M}\right)\right)$$

  - Further modified into:

$$x_n \sim N_c\left(\mathbf{0}_F, \left(\sum_k h_{kn}^{\Phi}\, diag\left(\boldsymbol{w}_k^{\Phi}\right)\right) \cdot \left(\sum_u h_{un}^{F_0}\, diag\left(\boldsymbol{w}_u^{F_0}\right)\right) + \sum_r h_{rn}^{M}\, diag\left(\boldsymbol{w}_r^{M}\right)\right)$$

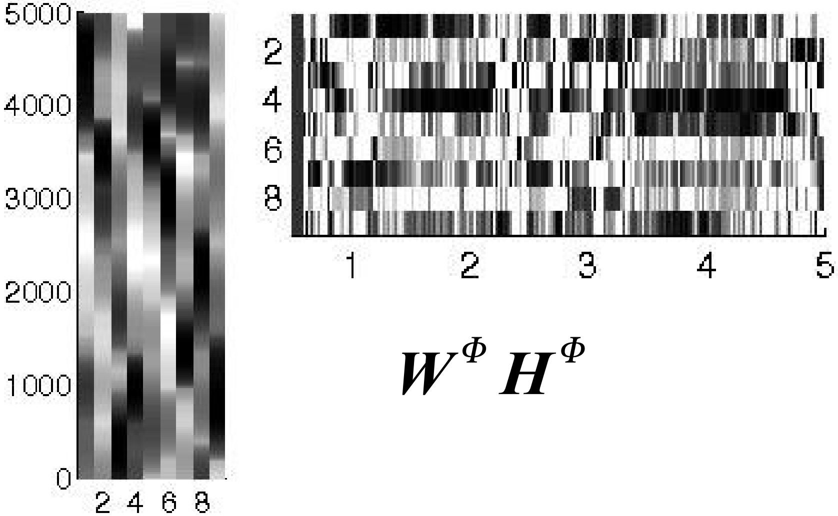
$$\Downarrow$$

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

TELECOM ParisTech

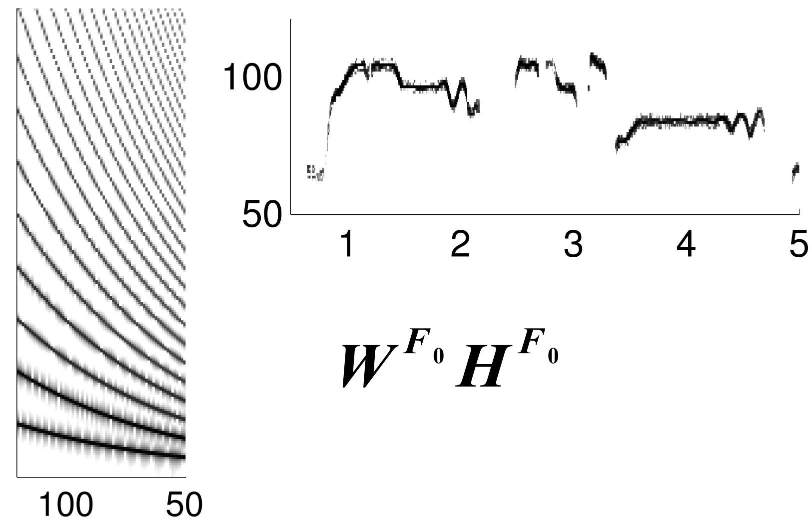
Signal Model: IMM

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



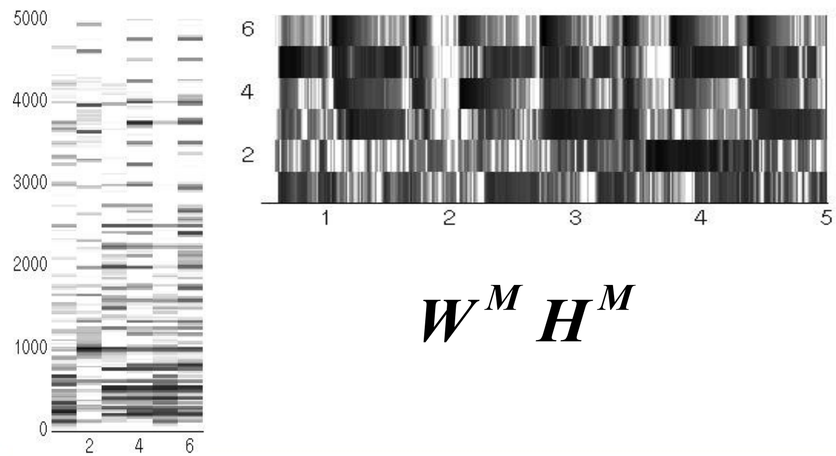
$W^{\Phi} H^{\Phi}$

$\cdot$



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

$+$



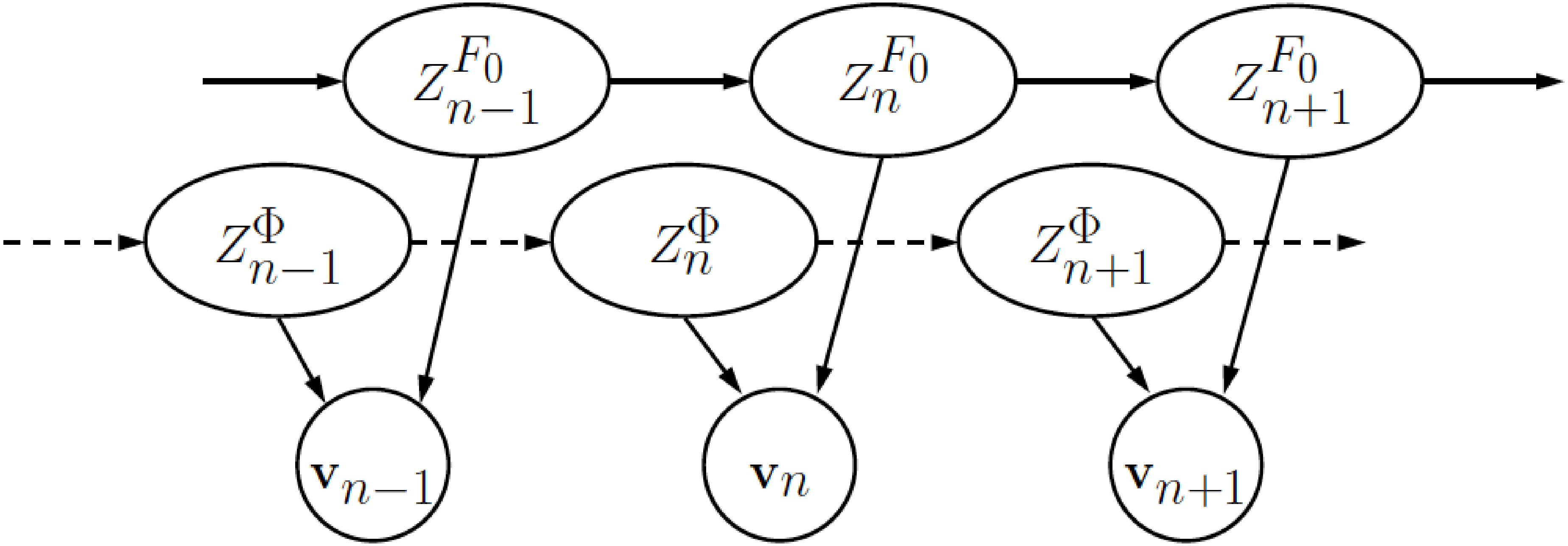
$W^{M} H^{M}$

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: Temporal constraint



- 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.

TELECOM ParisTech

Signal Model: HMM on F0 sequence

- Smoothness with **HMM** (Hidden Markov Model):
  - **Max. A Posteriori** (MAP):

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

where

$$p(Z^{F_0} | \boldsymbol{X}) \propto p(\boldsymbol{X} | Z^{F_0}) P(Z^{F_0})$$

with

$$p(\boldsymbol{X} | Z^{F_0}) = \prod_n p(\boldsymbol{x_n} | Z_n^{F_0})$$

$$P(Z^{F_0}) = p(Z_1^{F_0}) \prod_n p(Z_n^{F_0} | Z_{n-1}^{F_0})$$

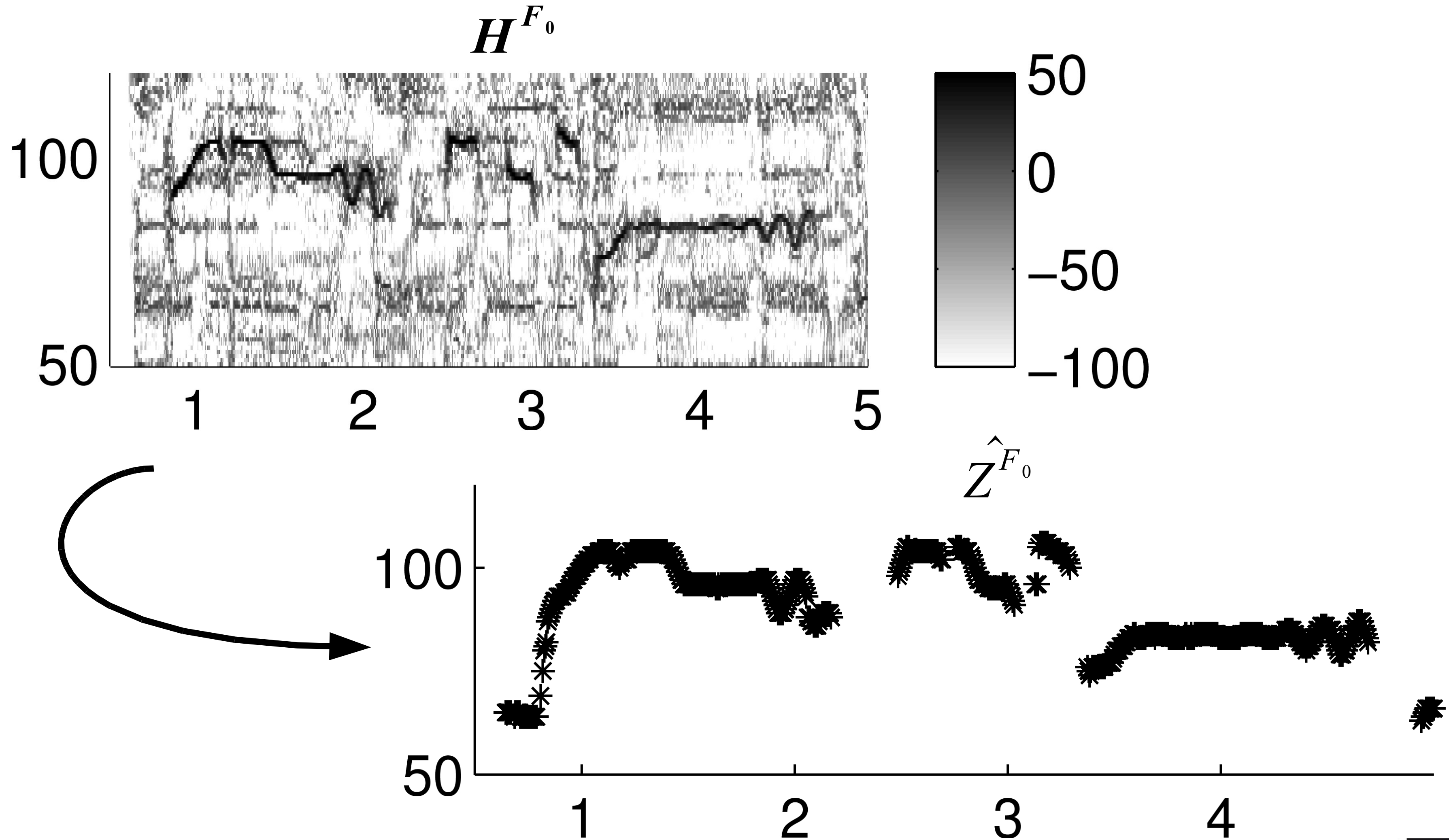
- **Transition probabilities** set to:

$$p(Z_n^{F_0} = u_1 | Z_{n-1}^{F_0} = u_2) \propto \exp(\alpha |u_1 - u_2|)$$

- **Approximation** for the IMM:

$$p(\boldsymbol{x_n} | Z_n^{F_0} = u) \propto h_{un}^{F_0}$$

Signal Model: Viterbi melody tracking



Signal Model: Summary

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

$$x_n \sim \sum_{k,u} \pi_{ku} N_c\left(\mathbf{0}_F, b_{kun}\, diag\left(\mathbf{w}_k^{\Phi} \cdot \mathbf{w}_u^{F_0}\right) + \sum_r h_{rn}^M\, diag\left(\mathbf{w}_r^{M}\right)\right)$$

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

$$x_n \sim N_c\left(\mathbf{0}_F, \sum_{k,u} b_{kun}\, diag\left(\mathbf{w}_k^{\Phi} \cdot \mathbf{w}_u^{F_0}\right) + \sum_r h_{rn}^M\, diag\left(\mathbf{w}_r^{M}\right)\right)$$

  - Less **realistic** than **GSMM**

- For both: **temporal smoothness** with HMM

Systems: Parameter estimation and algorithms

Systems: Parameter estimation

- **NMF** methodology both for GSMM and IMM:
  - **Partial derivatives** of criterion to minimize:

$$\nabla_\theta C = \nabla_\theta^+ C - \nabla_\theta^- C$$

  - **Multiplicative gradients**

$$\theta \leftarrow \theta \frac{\nabla_\theta^- C}{\nabla_\theta^+ C}$$

- **Drawbacks:**
  - **Slow "convergence",** if any...
  - **Initialization** sensitive
- **Advantages:**
  - **Fast** implementation
  - NMF in general: lots of other existing optimisations

Systems: Criteria for parameter estimation

■GSMM criterion:

- $$C_{GSMM}(\theta,\theta')=E_{\theta'}[\log p_\theta(\boldsymbol{X},Z^{\Phi},Z^{F_0})|\boldsymbol{X}]$$

- **EM algorithm:**
  - **-Compute** $p_{\theta'}(Z_n^{\Phi}=k,Z_n^{F_0}=u|\boldsymbol{x_n})$
  - **-Update parameters with multiplicative rules**

■IMM criterion:

- $$C_{IMM}(\theta)=\sum_{fn}\log s_{fn}^X+\frac{|x_{fn}|^2}{s_{fn}^X}$$

- **Update parameters with multiplicative rules**

Systems: workflow

1)Computing the **STFT**

2)Estimating the parameters, **EM for GSMM** or **simple multiplicative gradient for IMM**

3)Main melody inference

4)Re-iterate **parameter estimation,** if needed for lead/accompaniment separation – e.g. including unvoicing model

Signal Model and Algorithms: summary

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

$$\boldsymbol{x}_{\boldsymbol{n}} \sim \sum_{k,u} \pi_{ku} N_c(\boldsymbol{0}_{\boldsymbol{F}}, b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\Phi} \cdot \boldsymbol{w}_{\boldsymbol{u}}^{F_0}) + \sum_r h_{rn}^M\, diag(\boldsymbol{w}_{\boldsymbol{r}}^{\boldsymbol{M}}))$$

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

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

$$\boldsymbol{x}_{\boldsymbol{n}} \sim N_c(\boldsymbol{0}_{\boldsymbol{F}}, \sum_{k,u} b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\Phi} \cdot \boldsymbol{w}_{\boldsymbol{u}}^{F_0}) + \sum_r h_{rn}^M\, diag(\boldsymbol{w}_{\boldsymbol{r}}^{\boldsymbol{M}}))$$

  - 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: Performance measures

- **Correctly estimated pitch**: F0 within ¼ tone of groundtruth
- “raw pitch” = #(correct pitch) / #(voiced frames)
- “raw chroma” = #(correct chroma pitch) / #(voiced frames)
- “Overall accuracy” = #(correct pitch)/#(frames)

Transcription of the main melody: results

Particip ant	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

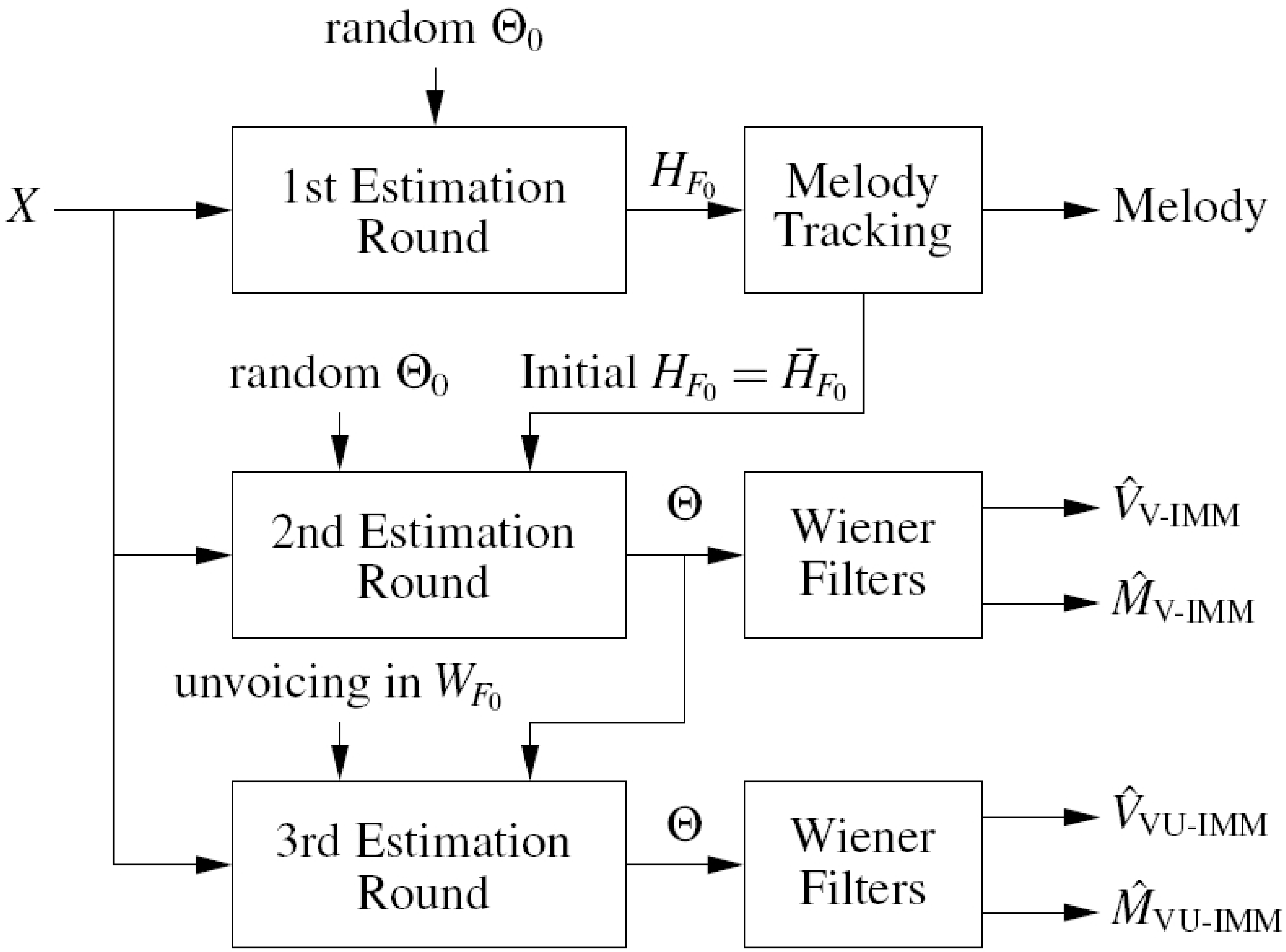
Partici pant	Raw Pitch (%)	Raw Chroma (%)	Overall Acc(%)
cl1	63.45	66.29	52.19
cl2	63.45	66.29	55.19
drd1	**74.45**	**76.82**	**66.86**
drd2	**72.09**	**75.72**	**66.17**
hjc1	66.12	72.58	50.49
hjc2	51.13	67.12	49.01
jjy	73.33	79.68	56.64
kd	80.58	82.52	73.35
mw	73.44	77.5	55.07
pc	64.1	65.84	62.88
rr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

MIREX 2009 AME Results

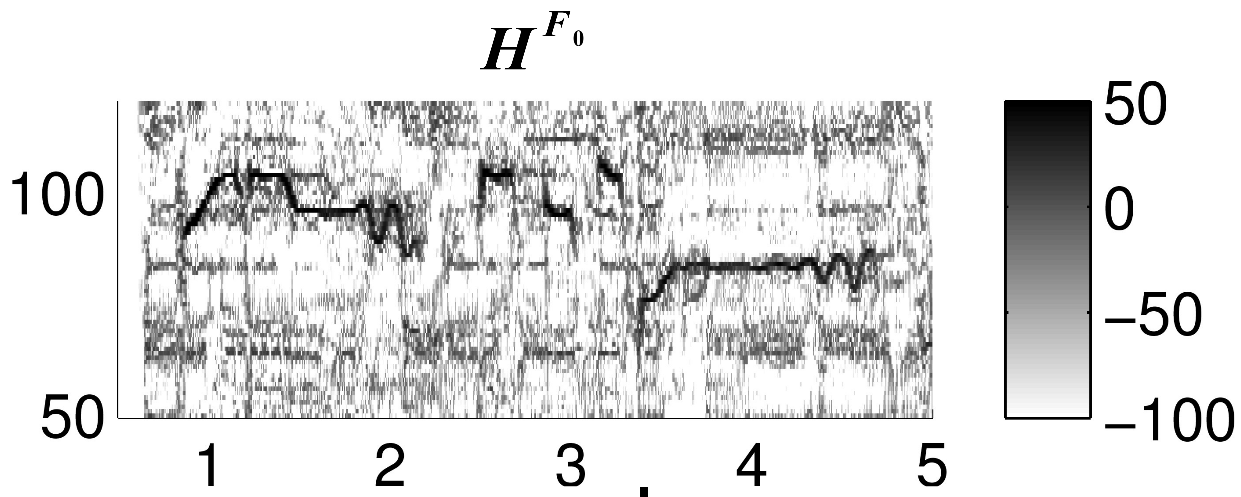
Leading Instrument separation

- **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**)

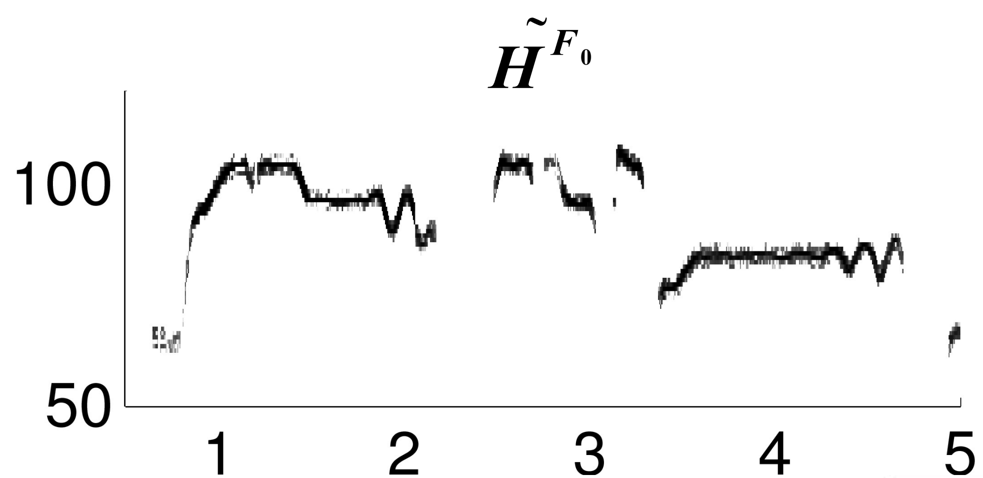
Leading Instrument separation: system



Leading Instrument separation: Parameter estimation knowing the melody

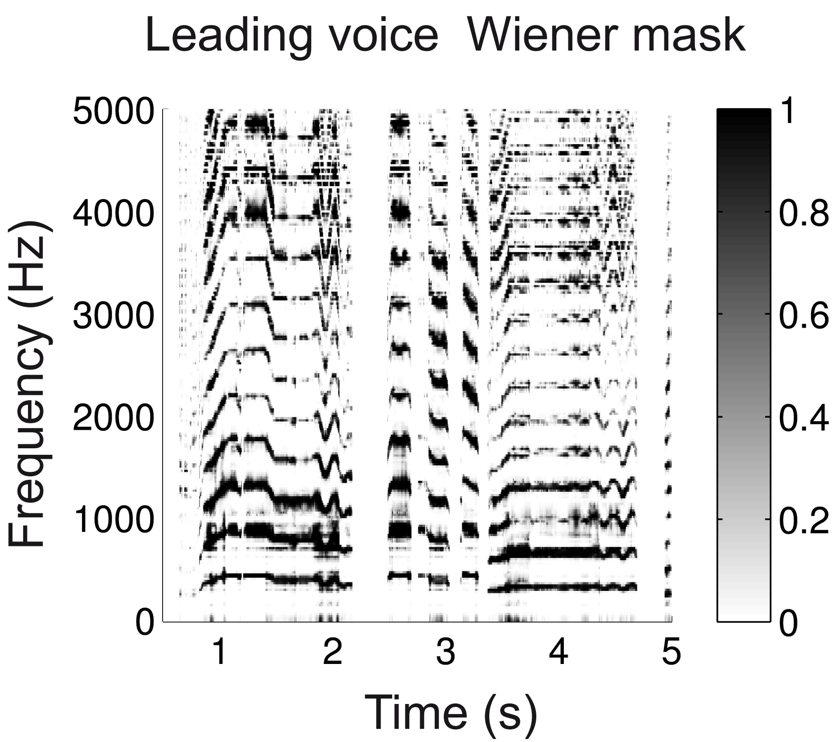


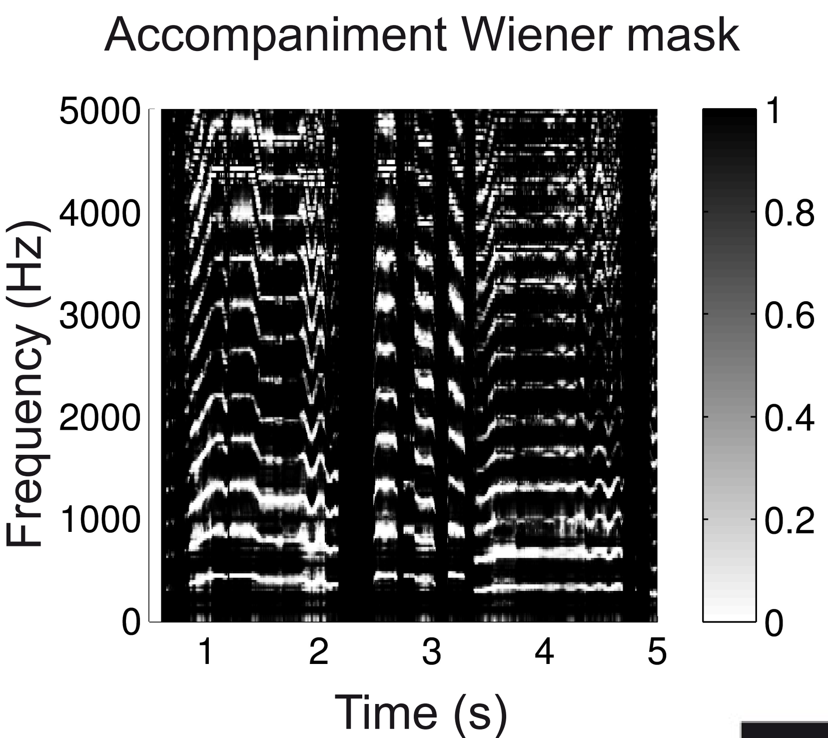
Melody estimation
+ hard constraint



Leading Instrument separation: Adaptive Wiener filtering

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





Leading 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.

To go further...

Introduction: State of the art

- **Audio Melody Extraction** (AME), at Music Information Retrieval Evaluation eXchange (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**

Signal Model: GSMM for mixture

$$\boldsymbol{x}_{\boldsymbol{n}}|k,u \sim N_c\Big(\boldsymbol{0}_{\boldsymbol{F}}, \overbrace{b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\Phi}\cdot\boldsymbol{w}_{\boldsymbol{u}}^{F_0}) + \sum\nolimits_r h_{rn}^M\, diag(\boldsymbol{w}_{\boldsymbol{r}}^{\boldsymbol{M}})}^{\boldsymbol{s}_{\boldsymbol{n},\boldsymbol{ku}}^{\boldsymbol{X}}}\Big)$$

$$\boldsymbol{x}_{\boldsymbol{n}} \sim \sum\nolimits_{k,u} \pi_{ku} N_c\big(\boldsymbol{0}_{\boldsymbol{F}}, b_{kun}\, diag(\boldsymbol{w}_{\boldsymbol{k}}^{\Phi}\cdot\boldsymbol{w}_{\boldsymbol{u}}^{F_0}) + \sum\nolimits_r h_{rn}^M\, diag(\boldsymbol{w}_{\boldsymbol{r}}^{\boldsymbol{M}})\big)$$

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

$$C(\theta,\theta') = \sum_{k,u,n}\left(\sum_f \log(s_{fn,ku}^X) + \frac{|x_{fn}|^2}{s_{fn,ku}^X}\right) p_{\theta'}(k,u|\boldsymbol{x}_{\boldsymbol{n}}) + \lambda\left(\sum_{k,u}\pi_{ku} - 1\right)$$

- Estimating optimal F0 sequence:

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