Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

Jean-Louis Durrieu, PhD candidate

TSI Department, Telecom ParisTech

http://perso.telecom-paristech.fr/durrieu/en/

07/09/09

Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

1. Introduction
2. Signal Models
3. Transcription of the Melody
4. “Solo/Accompaniment” Separation

Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

1. Introduction
2. Signal Models
3. Transcription of the Melody
4. “Solo/Accompaniment” Separation

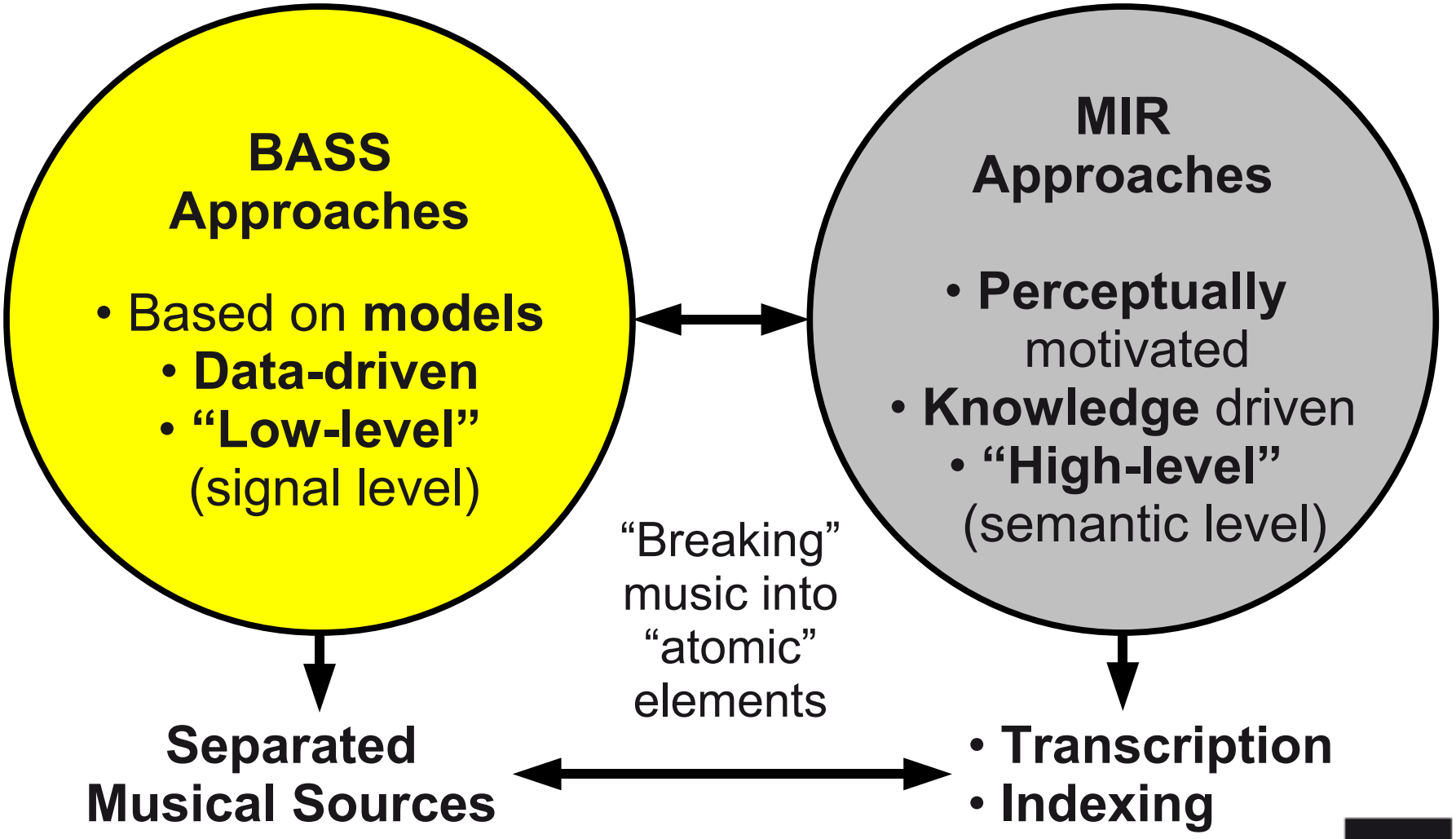
Introduction

- **Blind Audio Source Separation (BASS)** for music and **Music Information Retrieval (MIR)**:
  → **Inter-related Fields**

- Polyphonic music recordings:
  a **BASS/MIR hybrid** approach to **main melody transcription/separation**

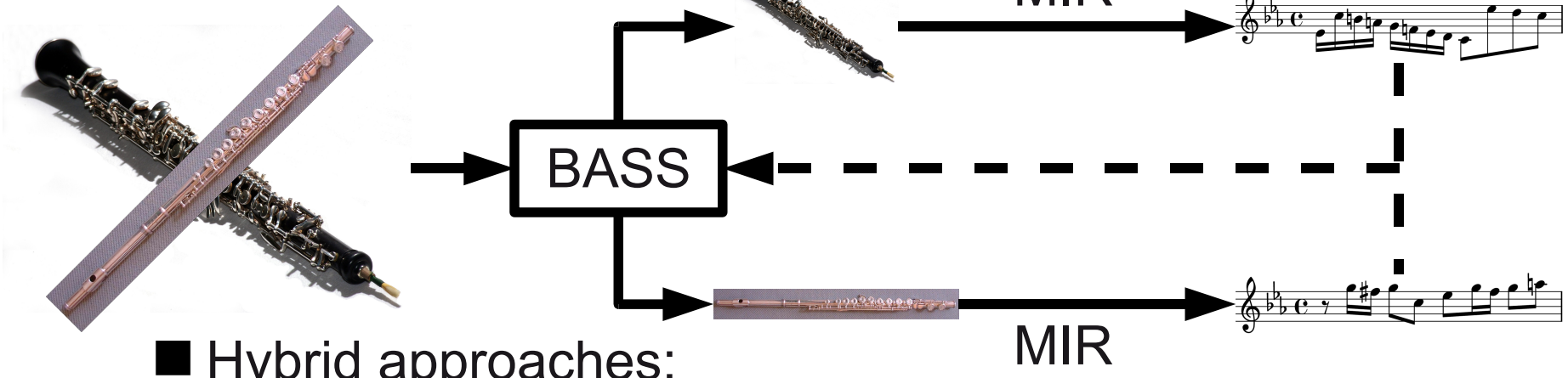
- **Applications**

Introduction: link between BASS/MIR



Introduction: Bridging BASS/MIR "gap"

- **Improving** BASS with MIR, and MIR with BASS
- 2 instruments transcription/separation example:



- Hybrid approaches:
  - **E. Vincent**, "Musical Source Separation Using Time-Frequency Source Priors", IEEE Transactions on Audio, Speech and Language Processing, vol. 14, No 1
  - **Singing voice signals?**

Introduction: Main Melody Transcription, Main Instrument Separation

- **Definitions:**
  - [MIREX] “Audio Melody Extraction”: *extract the main melody from polyphonic audio signals.*
  - [Paiva2007]: *“[The Main] Melody is the dominant individual pitched line in a musical ensemble”.*
- **Addressing 2 tasks:**
  - **Main Melody Transcription:** identify and transcribe the sequence of fundamental frequencies played by the main instrument in a polyphonic music signal (mono or stereo),
  - **Main Instrument/Accompaniment Separation:** separate the instrument playing the main melody from the other accompaniment instruments.

Introduction: Applications

- **Transcribed Melody**
  - **Indexing** large music database,
  - **Musical transcription** into “human readable” score,
  - ...
- **Separating the Main Instrument from the Accompaniment:**
  - **Generate accompaniments** for solo performers
  - **Pre-Processing** for MIR applications (chord detection, instrument classification, etc.)
  - ...

Introduction: Presentation Outline

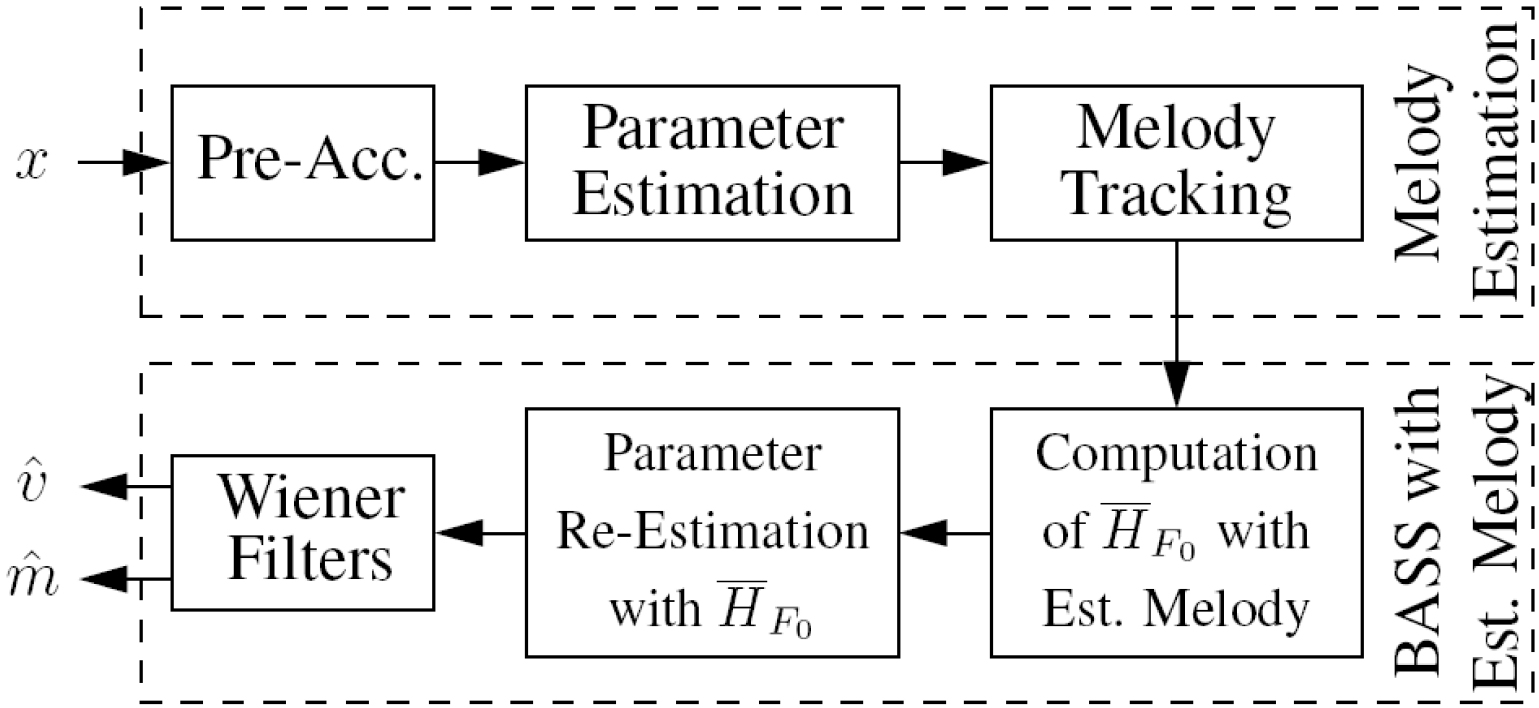
- **Signal Models**
  Source/Filter model for the main instrument, **NMF** for the other instruments; **estimation** algorithm for the corresponding parameters,

- **Melody transcription**
  Viterbi smoothing of the melody sequence,

- **Main Instrument/Accompaniment Separation** (also referred to as Solo/Accompaniment Separation)
  Wiener filters to estimate the separated sources,

- **Conclusion/Discussions**

Introduction: System Outline (ICASSP09)



Introduction: Contributors at Telecom ParisTech

- **Supervisors:**
  - **Bertrand DAVID,**
  - **Gaël RICHARD.**

- **Team members:**
  - **Nancy BERTIN,**
  - **Cédric FEVOTTE,**
  - **Alexey OZEROV,**
  - **And all the other Audiosig project team members...**

Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

1. Introduction
2. **Signal Models**
3. Transcription of the Melody
4. “Solo/Accompaniment” Separation

Signal Models

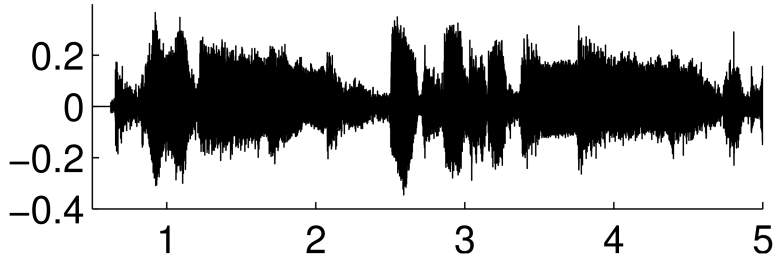
- **Audio signals:**
  - **Time-Frequency** Representation,
  - Statistical modeling.

- **Mixture** model
- **Source/Filter** model for the main instrument
  - **Motivations**
  - **Characterizing** the main melody instrument
- **NMF**-based model for the accompaniment
  - **Decomposition** on limited dictionary
  - Link between **NMF** and our framework

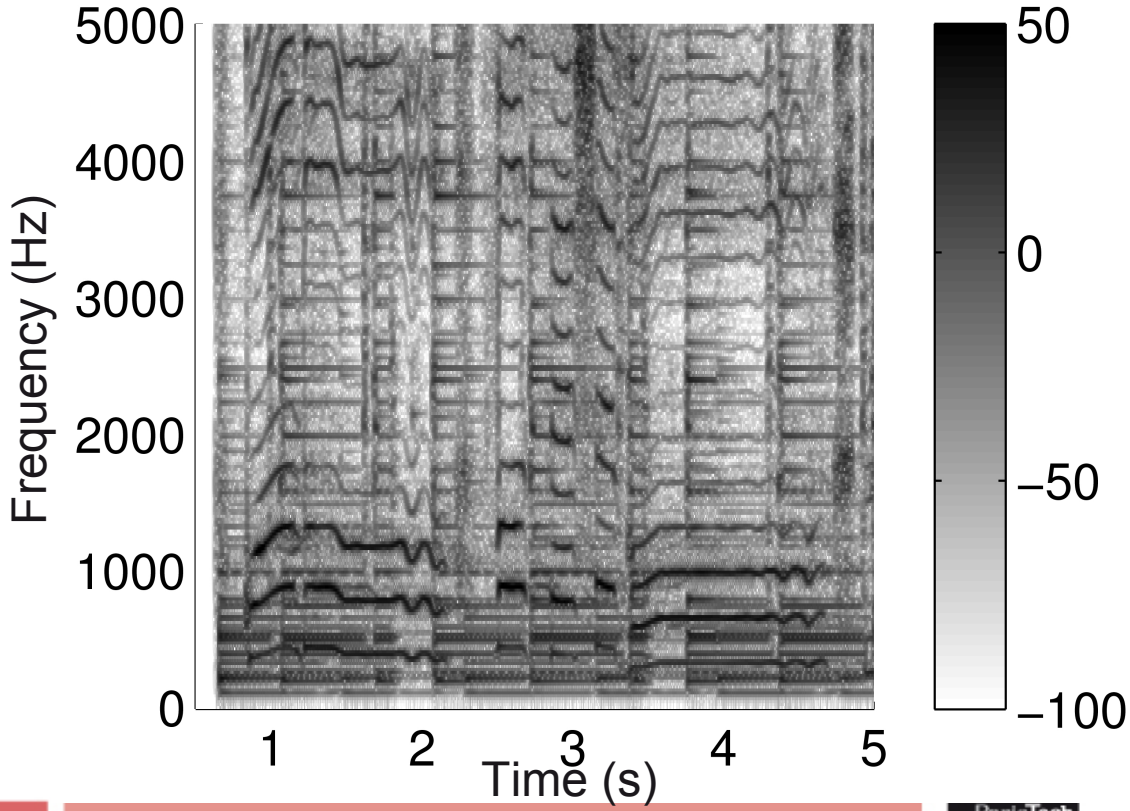
- **Parameter estimation**
  - **NMF**-like algorithm: multiplicative gradient approach

Signal Models: Time-Frequency Representation

- Digital audio: **waveform**



- **Time-frequency** representation:
  - **Evolution** of **frequency content,**
  - **Human auditory system.**



- **Short-Time Fourier Transform (STFT):**

$$X_n(f)$$

Signal Models: Complex Proper Gaussians

- Model for **complex spectrum:** $X_n(f) \sim \mathcal{N}_c\left(0, S_{X_n}(f)\right)$

$$N_c(X_n(f); 0, S_{X_n}(f)) = \frac{|X_n(f)|}{\pi S_{X_n}(f)} \exp\left(-\frac{|X_n(f)|^2}{S_{X_n}(f)}\right)$$

- **Independence** across time and frequencies:

$$\mathrm{p}(X) = \sum_{f,n} N_c(X_n(f); 0, S_{X_n}(f))$$

- For stationary processes, $S_{X_n} =$ **power spectrum density (PSD)** of X_n
- **Variance/PSD matrix** S_X s.t. $S_X(f,n) = S_{X_n}(f)$

Signal Models: Mixture Model

- **Mixture** X = **Solo Voice** V + **Accompaniment Music** M
- Each signal centered-Gaussian, with resp. variances : S_X, S_V, S_M
- **Independence** between V and M:

$$S_X = S_V + S_M$$

Source/Filter model

NMF decomposition of the power spectrum

Signal Models: Mixture Model

- **Mixture** X = **Solo Voice** V + **Accompaniment Music** M
- Each signal centered-Gaussian, with resp. variances : S_X, S_V, S_M
- **Independence** between V and M:

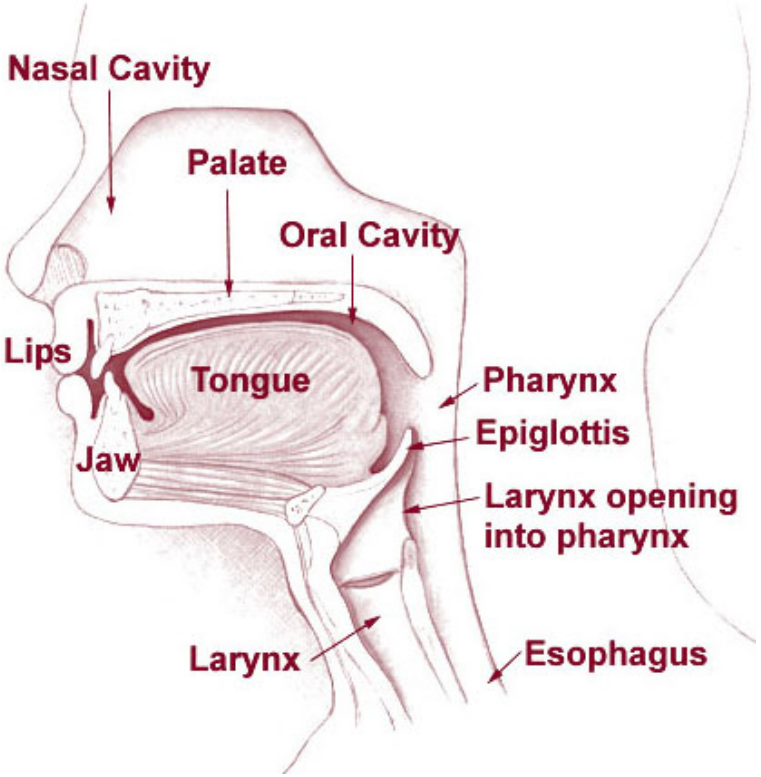
$$S_X = S_V + S_M$$

Source/Filter model

NMF decomposition of the power spectrum

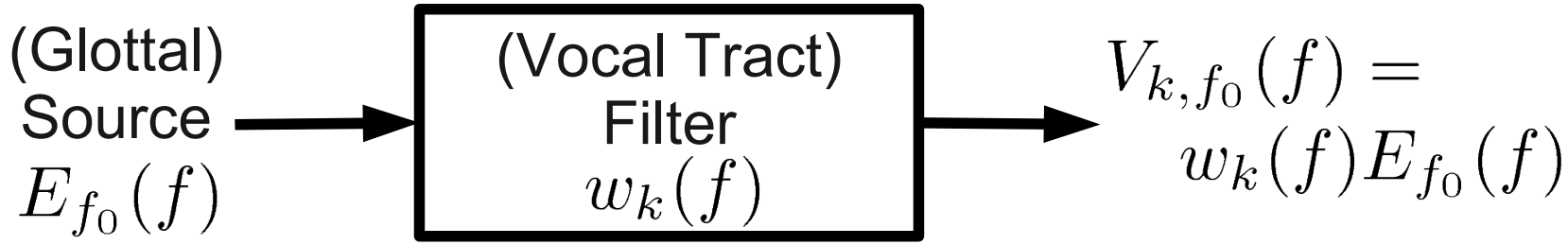
Signal Models: Source/Filter Model for the Main Instrument

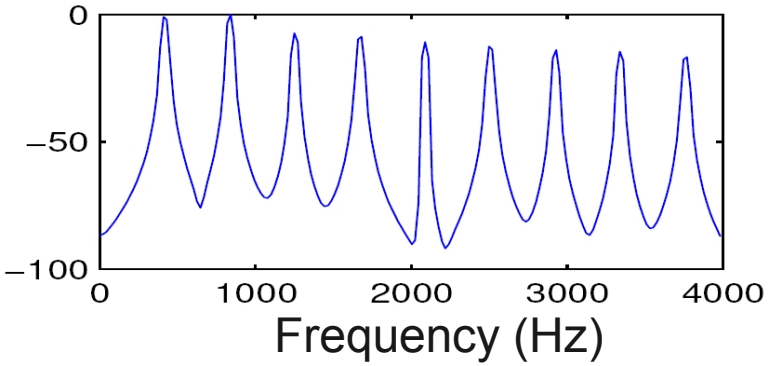
- **Motivations:**
  - **Singing voice** often main instrument,
  - **Source/Filter** widely used, suitable for wide range of other instruments,
  - **Separately** modeling **pitched** aspects (source) from **timbre** aspects (filter).

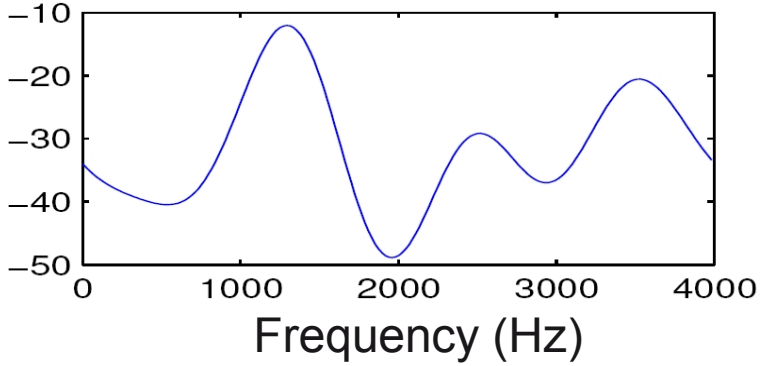


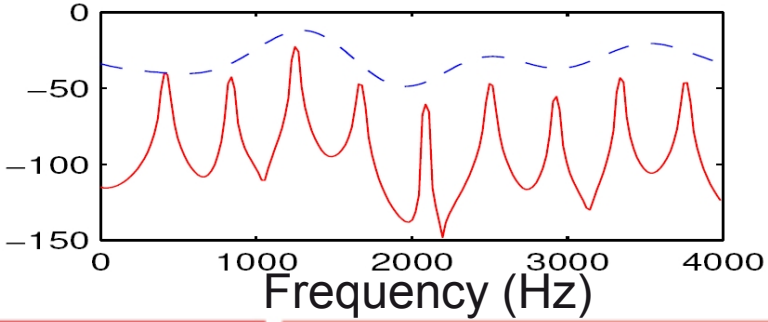
Human vocal tract (from Wikipedia)

Signal Models: Source/Filter Principle



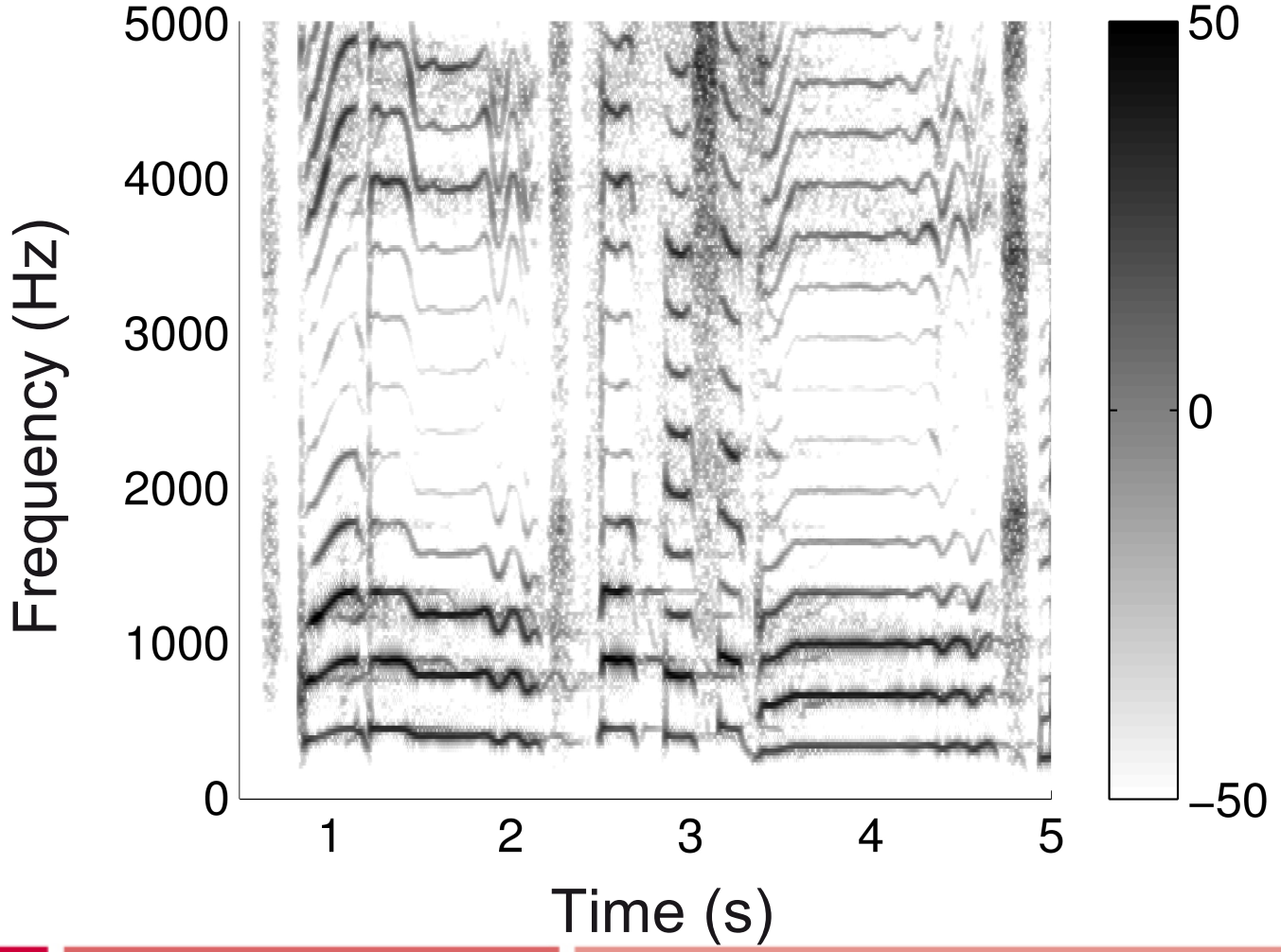






Signal Models: Source/Filter Variability

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



Signal Models: Source/Filter Variability

- **Human singer:**
  - **Independent evolution** of pitches f_0 and filters (vowel),
  - **Continuous pitch** variations,
  - **Limited set** of vowels (smooth filters),
  - **Unvoiced** parts...
- **Proposed Model** for **Main Instrument:**
  - Discrete range of possible f_0 for **voiced** source component, log-spaced s.t. 96 f_0 per octave,
  - Limited number of **"smooth" filters**,
  - **Unvoiced** source component integrated **later** in the estimation process.

Signal Models: Source Component (1/2)

- **Voiced source** component:
  - KLGLOTT88 (Glottal source) model, [Klatt90]: **spectral comb dictionary** W_{F_0}, N_{F_0} "notes",
  - Freq. f , Pitch f_0 : power spectrum $W_{F_0}(f, f_0)$,
  - Pitch f_0 , Frame n : activation coefficients $H_{F_0}(f_0, n)$,
  - Nonnegative combination of the element of the dictionary

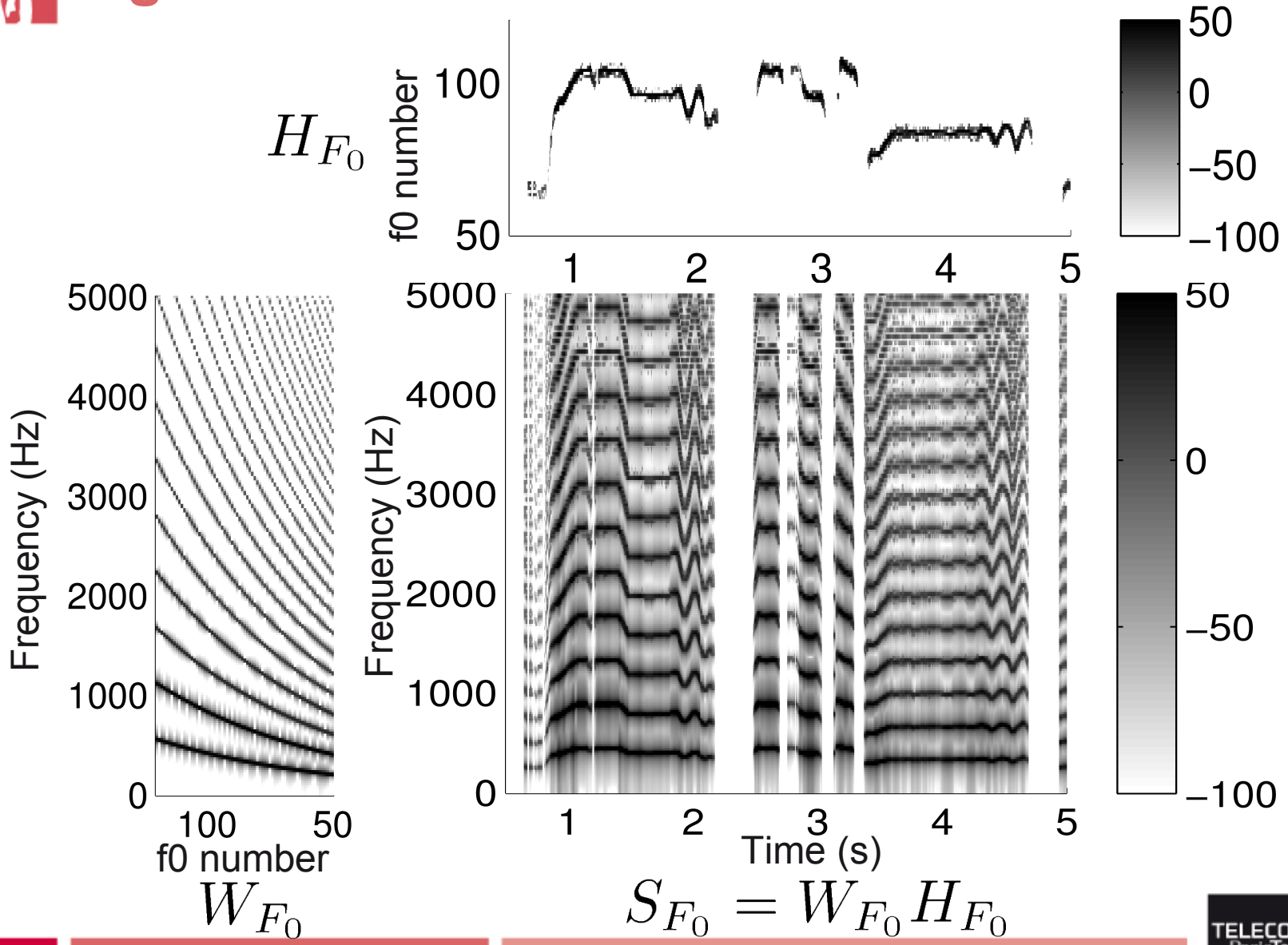
$$S_{F_0}(f, n) = \sum_{f_0} W_{F_0}(f, f_0) H_{F_0}(f_0, n)$$

- **Unvoiced source**
  - In dictionary W_{F_0}, "unvoiced" component such that:

$$W_{F_0}(f, f_0) = 1, \forall f,$$

  - Activation coefficient estimated only after filter part.

Signal Models: Source Component (2/2)



Signal Models: Filter Component (1/2)

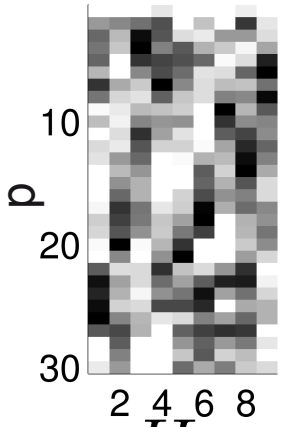
- **Filter component:**
  - **Dictionary of K filters W_Φ,**
  - Freq. f, Filter number k : freq. response $W_\Phi(f,k)$,
  - Filter k, Frame n : activation $H_\Phi(k,n)$,
  - Combination:

$$S_\Phi = W_\Phi H_\Phi$$

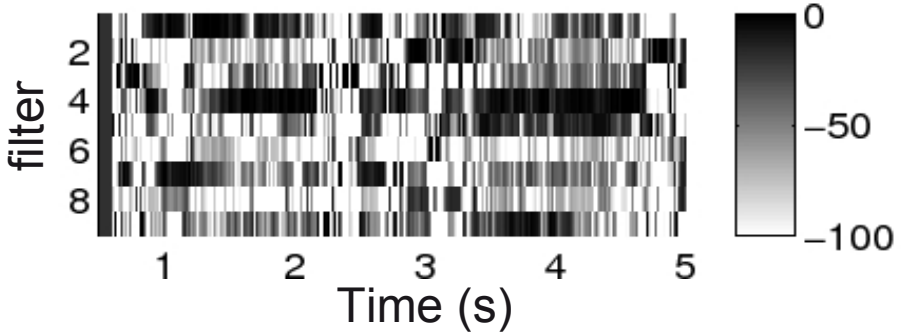
- **Filter smoothness:**
  - Decomposition on spectral dictionary of P **smooth "atomic" elements** W_Γ , activations H_Γ ,
  - $W_\Phi(f,k) = \sum_p W_\Gamma(f,p) H_\Gamma(p,k), \forall k$

    That is to say: $W_\Phi = W_\Gamma H_\Gamma$

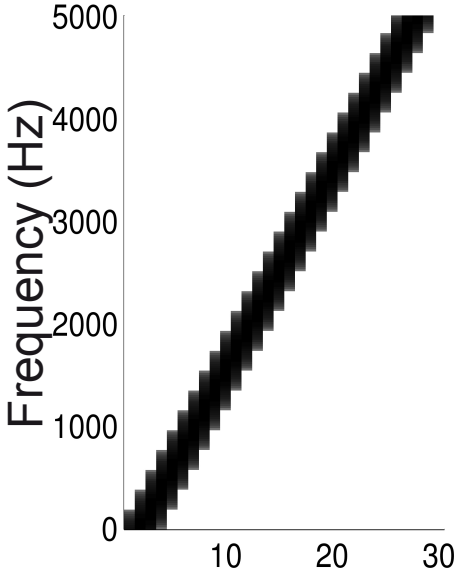
Signal Models: Filter Component (2/2)



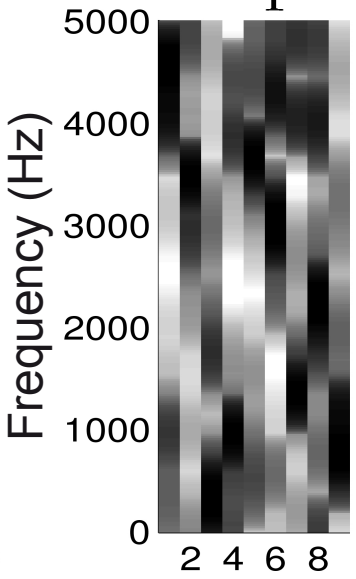
H_Γ



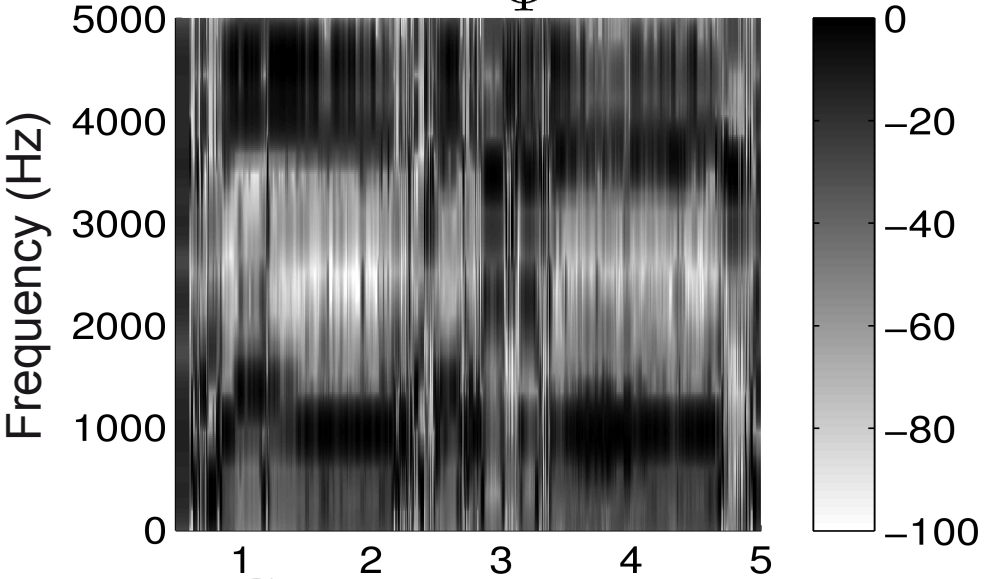
H_Φ



W_Γ



$W_\Phi = W_\Gamma H_\Gamma$



$S_\Phi = W_\Gamma H_\Gamma H_\Phi$

Signal Models: Source/Filter Summary

- **Source** contribution:

$$S_{F_0} = W_{F_0} H_{F_0}$$

- **Filter** contribution:

$$S_\Phi = W_\Phi H_\Phi = W_\Gamma H_\Gamma H_\Phi$$

- **Main Instrument** contribution to the mixture power spectrum:

$$S_V = S_\Phi .* S_{F_0} = (W_\Gamma H_\Gamma H_\Phi) .* (W_{F_0} H_{F_0})$$

- **Parameters:**
  - **Fixed** parameters: dictionaries W_{F_0} and W_Γ
  - To **estimate**: $\{H_\Gamma, H_\Phi, H_{F_0}\}$

Signal Models: Mixture Model

- **Mixture** X = **Solo Voice** V + **Accompaniment Music** M
- Each signal centered-Gaussian, with resp. variances : S_X, S_V, S_M
- **Independence** between V and M:

$$S_X = S_V + S_M$$

Source/Filter model

NMF decomposition of the power spectrum

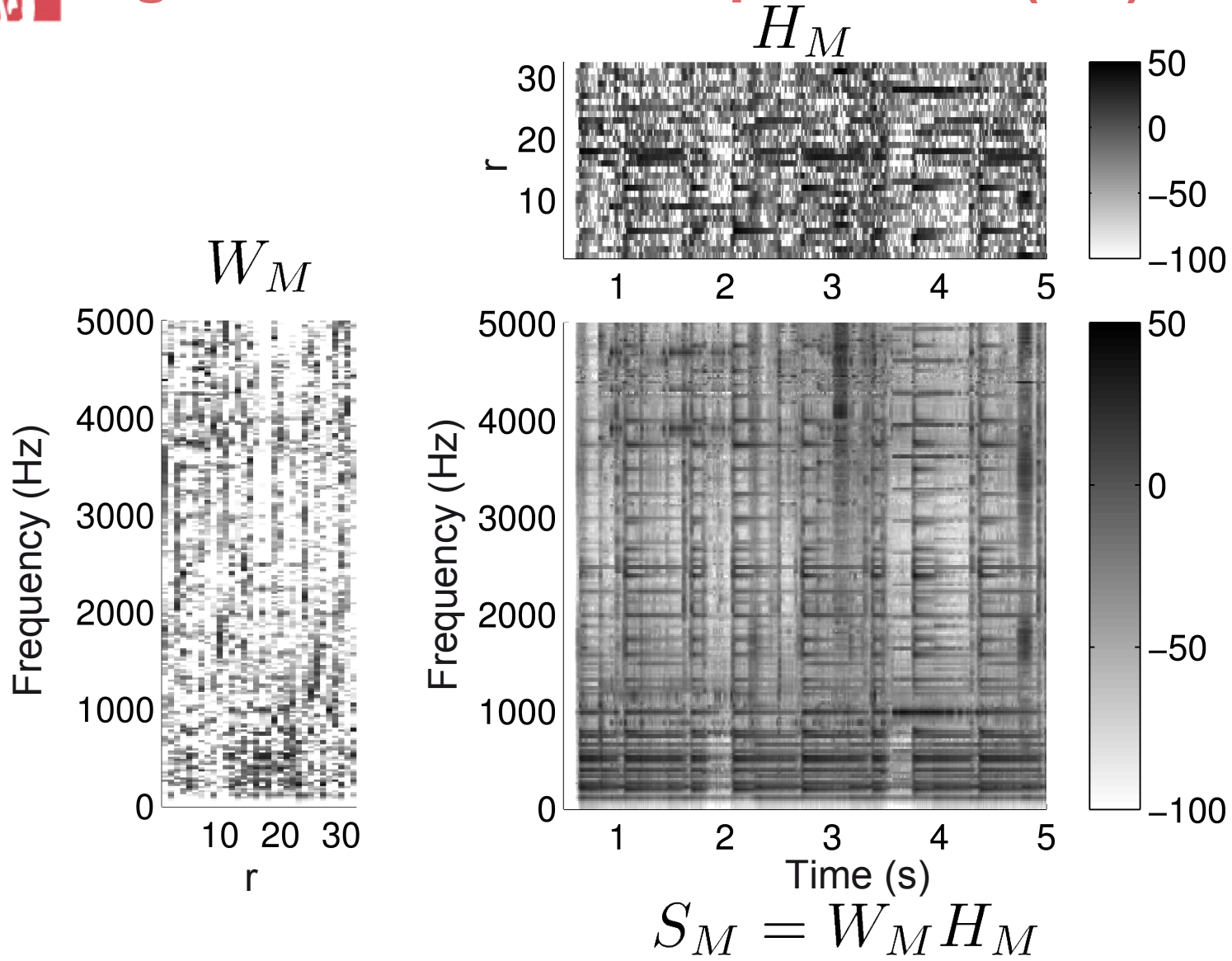
Signal Models: Accompaniment (1/2)

- **Accompaniment/Background Music** component:
  - **Power spectrum dictionary** W_M, with R elements,
  - **Activation** matrix H_M,
  - **Nonnegative combination** of the element of the dictionary

$$S_M = W_M H_M$$

- **Equivalence between [Fevotte09]:**
  - **Maximum Likelihood (ML)** estimation of W_M and H_M with $M \sim \mathcal{N}_c(0, W_M H_M)$
  - **NMF** minimizing the **Itakura-Saito divergence** between $|M|^2$ and the matrix product $W_M H_M$

Signal Models: Accompaniment (2/2)



$$S_M = W_M H_M$$

Signal Models: Mixture model summary

- **Mixture variance/PSD matrix:**
  - Main Instrument:
    $$S_V = S_\Phi .* S_{F_0} = (W_\Gamma H_\Gamma H_\Phi) .* (W_{F_0} H_{F_0})$$
  - Accompaniment:
    $$S_M = W_M H_M$$
  - Mixture:
    $$S_X = (W_\Gamma H_\Gamma H_\Phi) .* (W_{F_0} H_{F_0}) + W_M H_M$$
- **Parameters:**
  - Fixed Parameters: $\{W_\Gamma, W_{F_0}\}$
  - To be estimated: $\{H_\Gamma, H_\Phi, H_{F_0}, W_M, H_M\}$

Signal Models: Parameter Estimation

- **Maximum Likelihood (ML) estimation:**
  - Log-likelihood of the observations X :

$$\log \mathrm{p}(X) = \sum_{fn} -\log S_{X_n}(f) - \frac{|X_n(f)|^2}{S_{X_n}(f)} + \mathrm{cst}$$

  - With the parameterized variance:

$$S_X = (W_\Gamma H_\Gamma H_\Phi) .* (W_{F_0} H_{F_0}) + W_M H_M$$

- **NMF** inspired algorithm:
  - **Itakura-Saito divergence** between $|X|^2$ and S_X
  - **Multiplicative updates** for parameter estimation

Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

1. Introduction
2. Signal Models
3. **Transcription of the Melody**
4. "Solo/Accompaniment" Separation

Transcription of the Melody

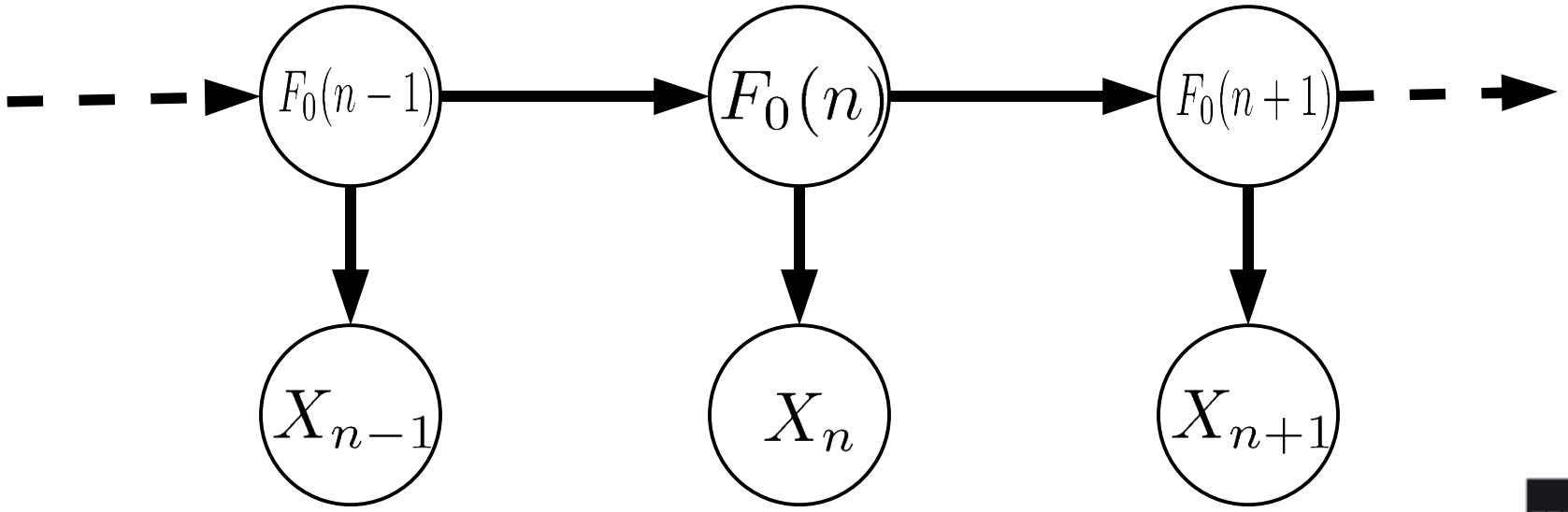
- Application **definition** and **scope**
- **Model** to estimate a smooth melody
- **Dynamic Programming** (Viterbi algorithm)

Transcription: Definition and scope

- **Definition:**
  - *“[The Main] Melody is the dominant individual pitched line in a musical ensemble.”* [Paiva2007],
  - Transcribe the fundamental frequencies played by the predominant instrument in a polyphonic music recording.
- **Scope:**
  - “low-level” transcription: **sequence of pitches,**
  - Various genres and musical ensembles (MIREX 2004 and 2005 database),
  - Participation to an international evaluation campaign: **MIREX 2008** (“audio melody extraction”)

Transcription: Modeling a Smooth Melody

- Assumptions on the melody line $F_0(n)$:
  - **Smooth**,
  - **Predominant** as concerns the **energy**,
  - Realistic melody line: **trade-off** between the smoothness and the energy of the line.
- **Hidden Markov Model** :



Transcription: Melody Tracking

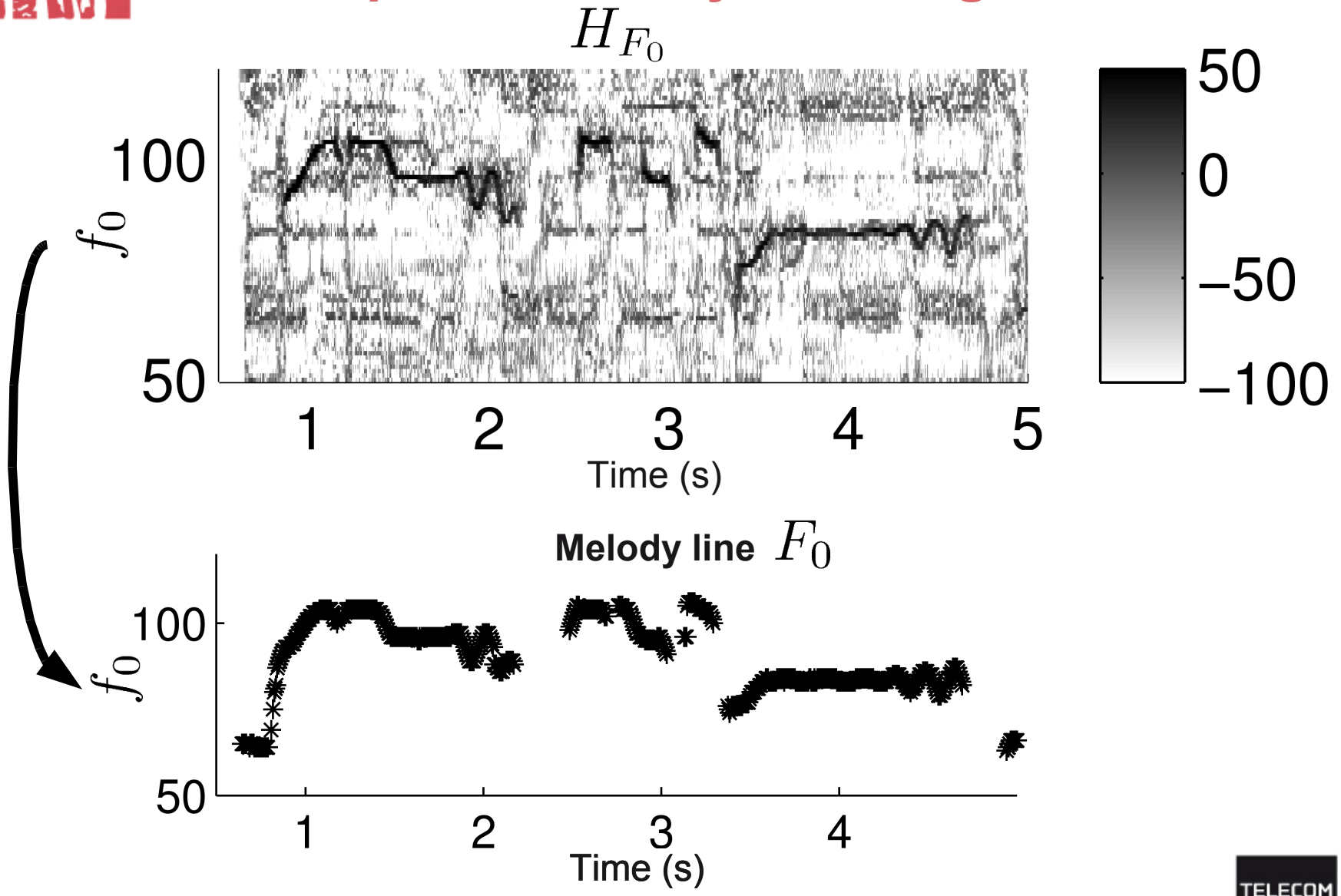
- Maximum Likelihood estimation:
  - $\mathrm{p}(F_0|X) \propto \mathrm{p}(X|F_0)\mathrm{p}(F_0)$
  - $\mathrm{p}(X|F_0) = \prod_n \mathrm{p}(X_n|F_0(n))$,

    concerns the energy: $\mathrm{p}(X_n|f_0) \propto H_{F_0}(f_0, n)$
  - Likelihood of the sequence of pitches:

    $$\mathrm{p}(F_0) = \mathrm{p}(F_0(1)) \prod_n \mathrm{p}(F_0(n+1)|F_0(n))$$

    where we chose: $\mathrm{p}(f_2|f_1) \propto \exp(\alpha|\log_2 \frac{f_2}{f_1}|)$
- Viterbi Tracking algorithm
  - Dynamic Programming,
  - Modified to deal with silences in the main melody.

Transcription: Melody Tracking



Transcription: Results

$$\text{Accuracy} = \frac{\#\text{correctly estimated frames}}{\#\text{frames}}$$

Audio Melody Extraction (ADC 2004 Dataset)

Rank	Participant	Accuracy
1	Cancela, P.	85.1%
2	Durrieu, Richard & David (imm)	81.5%
3	Ryynänen & Klapuri	78.8%
4	Rao & Rao	70.1%
5	Cao, Li, Liu & Yan (2)	68.0%
6	Durrieu, Richard & David (gmm)	59.6%
7	Cao, Li, Liu & Yan (1)	50.2%

Audio Melody Extraction (MIREX 2005 Dataset)

Rank	Participant	Accuracy
1	Cancela, P.	69.8%
2	Durrieu, Richard & David (imm)	66.0%
3	Rao & Rao	64.9%
4	Ryynänen & Klapuri	63.5%
5	Cao, Li, Liu & Yan (2)	61.4%
6	Durrieu, Richard & David (gmm)	52.2%
7	Cao, Li, Liu & Yan (1)	48.9%

Audio Melody Extraction (MIREX 2008 Dataset)

Rank	Participant	Accuracy
1	Durrieu, Richard & David (gmm)	76.0%
2	Ryynänen & Klapuri	75.3%
3	Durrieu, Richard & David (imm)	75.0%
4	Cancela, P.	73.3%
5	Rao & Rao	66.7%
6	Cao, Li, Liu & Yan (1)	51.4%
7	Cao, Li, Liu & Yan (2)	49.7%

Automatic Transcription and Separation of the Main Melody from Polyphonic Music Signals

1. Introduction
2. Signal Models
3. Transcription of the Melody
4. **"Solo/Accompaniment" Separation**

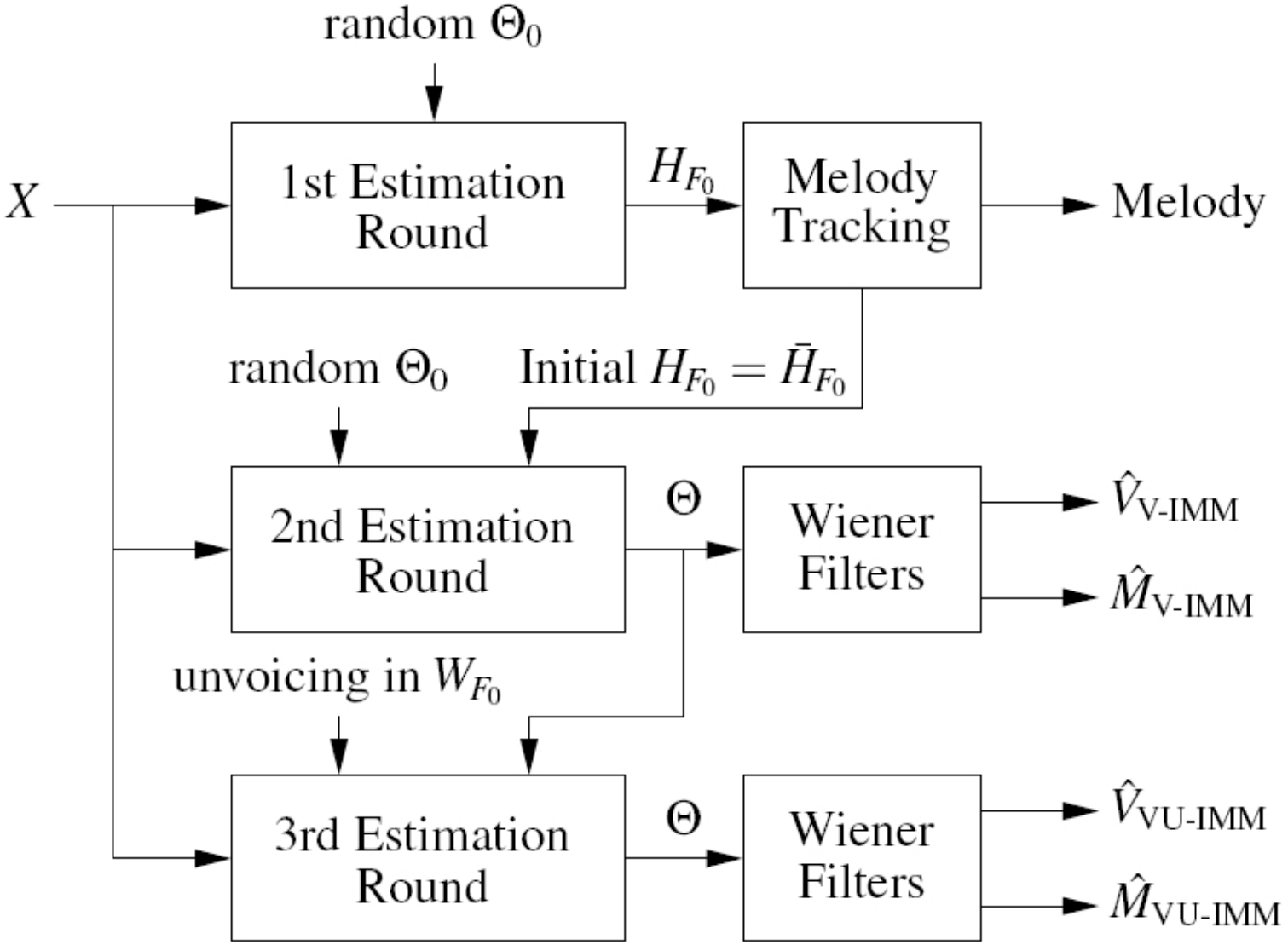
Solo/Accompaniment Separation

- **Definition**
- **Estimation** of the separated signals
- **Results**
- **Applications** and **Extensions**

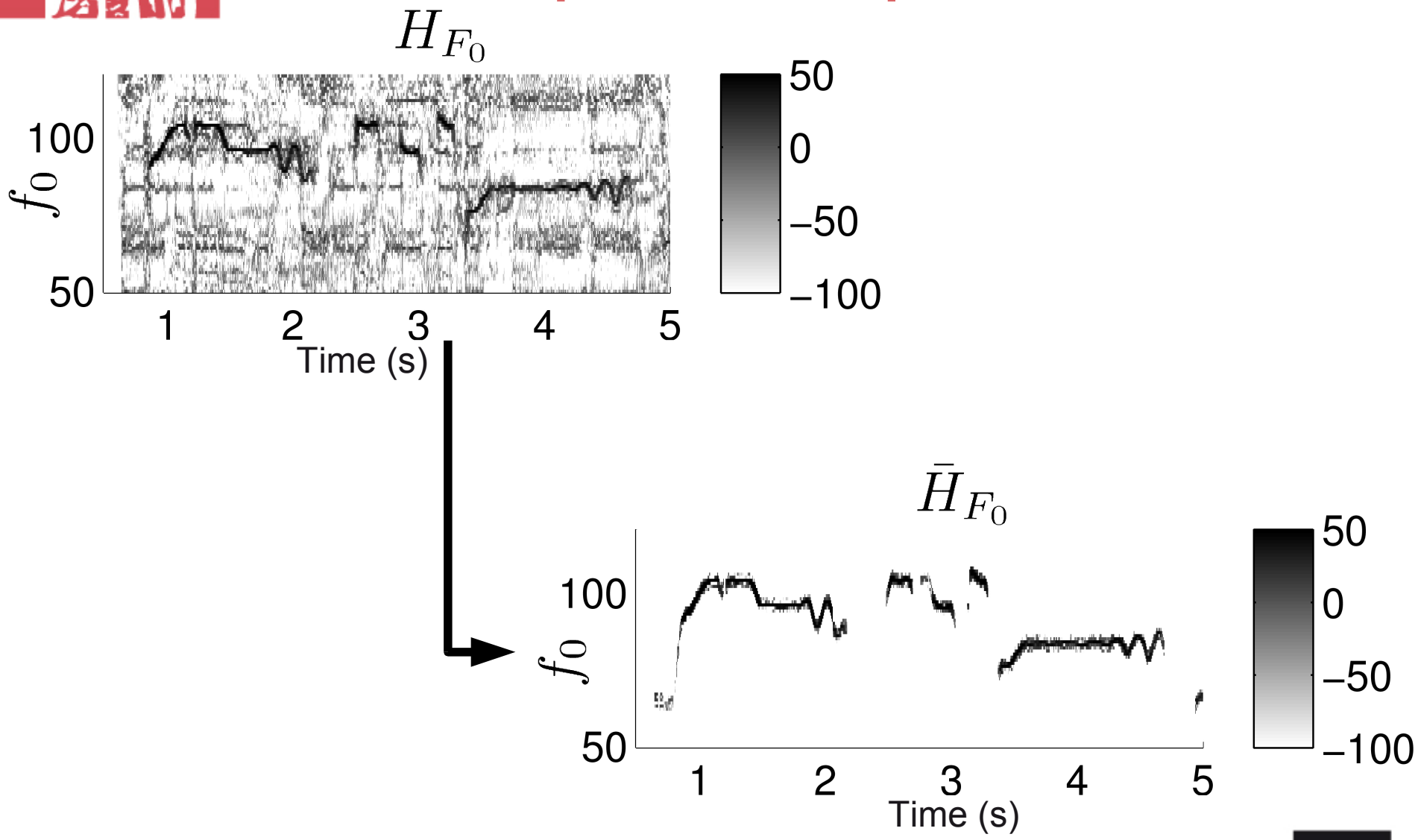
Solo/Accompaniment Separation

- **Definition:**
  - **“Solo”:** 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**,
  - Second step: re-estimation of the parameters **knowing the melody**,
  - (Third step: re-estimation including **unvoiced parts**)

Solo/Accompaniment Separation



Solo/Accompaniment Separation



Solo/Accompaniment 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/)
  - 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/
- Some other sounds here...

Solo/Accompaniment Separation: Applications/Extensions

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

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

Conclusions/Discussions

- **Conclusions:**
  - **Hybrid Framework BASS/MIR**,
  - **State-of-the-art** for “audio melody transcription” (MIREX08) and “solo/accompaniment” separation (SiSEC),
  - Techniques suitable for other applications: multipitch, background music enhancement, indexing, etc.
- **Extensions:**
  - Better formalism for **multichannel** signals**,**
  - Transcription into musical notes/musical score.

Conclusions/Discussions

- **Conclusions:**
  - **Hybrid Framework BASS/MIR**,
  - **State-of-the-art** for “audio melody transcription” (MIREX08) and “solo/accompaniment” separation (SiSEC),
  - Techniques suitable for other applications: multipitch, background music enhancement, indexing, etc.
- **Extensions:**
  - Better formalism for **multichannel** signals**,**
  - Transcription into musical notes/musical score.

Any questions?