Melody Extraction from Polyphonic Music Signals

Gaël Richard
Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France

With help from J. Salamon, E. Gomez, D. Ellis, J-L Durrieu, B. Fuentes, A. Ozerov, A. Liutkus

International Workshop on Acoustic Signal Enhancement (IWAENC 2014)
Sept. 11th, 2014

Audio recordings

- **What is an audio recording ?**

Audio recordings

- **What is an audio recording ?**

  - It is composed of *audio objects* or *sources*…

    piano drums guitar …. (stop)

  - …. Which are mixed together into a *mixture (*i.e. the audio recording*)* which is possibly multichannel (stereo is the most common for music)

Audio recordings

- **What is an audio recording ?**

  - It is composed of *audio objects* or *sources…*

    piano drums guitar …. (stop)

  - …. Which are mixed together into a *mixture (*i.e. the audio recording*)* which is possibly multichannel (stereo is the most common for music)

- **In most cases only the mixture is available which limits *Active Listening* capabilities …**

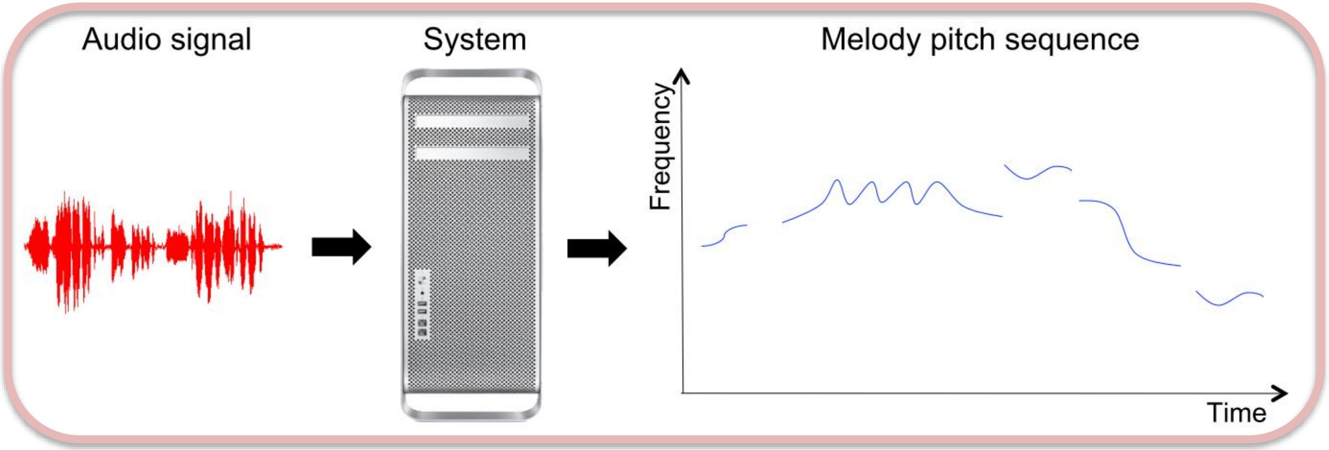
Applications

- **What could we do if we had the separated audio objects ?**
  - Active listening
  - Karaoke
  - Remixing
  - Music information retrieval
    - Cover song detection,
    - Music transcription (audio-to-midi, instrument recognition,…)
  - Audio Classification
  - ….

What is « melody extraction » ?

- **Also termed**
  - Audio melody extraction
  - Predominant melody extraction/estimation
  - Predominant fundamental frequency estimation

- **The aim: to obtain a sequence of frequency values representing the pitch of the dominant melodic line**



J. Salamon, E. Gomez, D. Ellis, G. Richard, *"Melody Extraction from Polyphonic Music Signals*", IEEE Signal Processing Magazine, March 2014.

The problem more precisely …

- **Definition:**

melody line = sequence of f_0 *(fundamental frequency values)* of the lead instrument or voice of a polyphonic music audio signal.

Polyphonic music audio signal:

- a music audio signal where two or more notes can sound simultaneously

Some difficulties of the problem

- **« Voicing detection »:** Determining when the leading voice is active

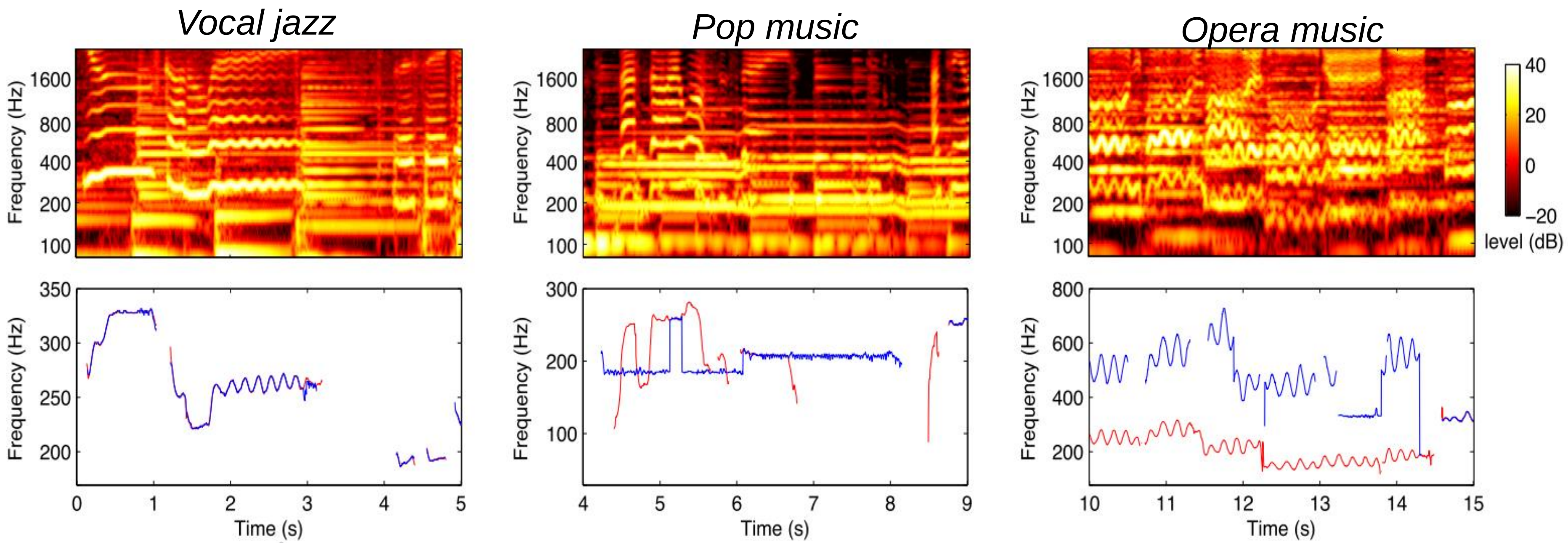
- **« Polyphony »:** Presence of multiple concurrent instruments

- **« Harmonicity »:** The notes of each instrument are often harmonically related

- **« Mixing effects »:** Presence of sound effects (reverberation, dynamic compression,…)

- **« Melody tracking »:** Associate the different fundamental frequencies obtained to the melody line

Some difficulties of the problem



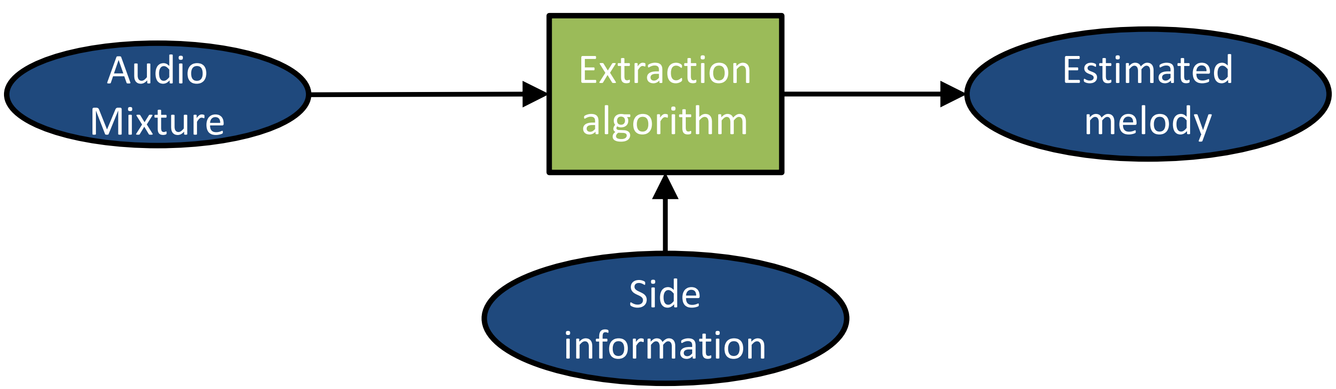
- Top: Spectrogram of the music signals (in dB)
- Bottom : extracted melody [Salomon2012] (blue) and ground truth (red).

[Salomon2012] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," IEEE Trans. on ASLP, vol. 20, no. 6, Aug. 2012.

From *“Blind” melody extraction* to Informed melody extraction

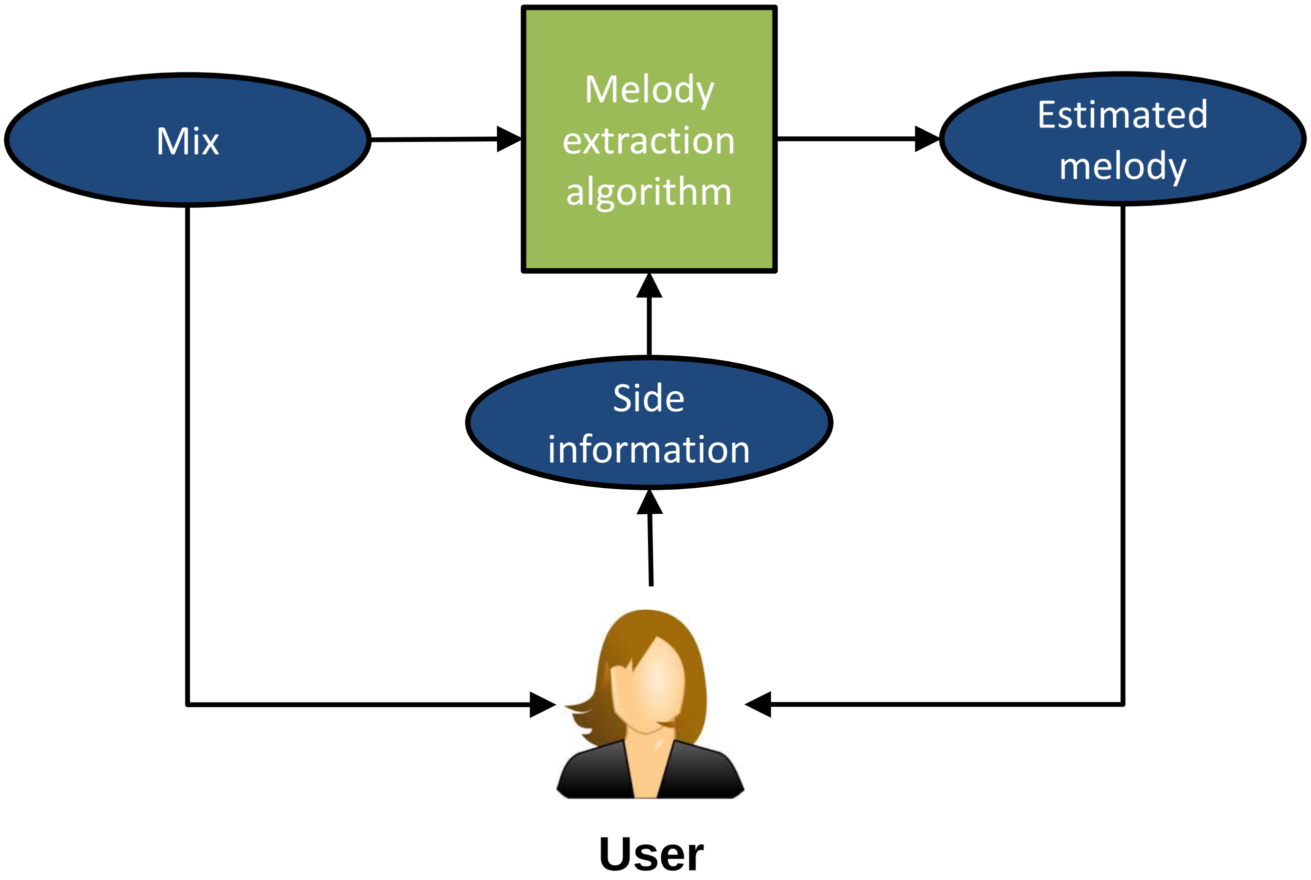
- **Strictly speaking, “Blind” melody extraction *is only done using the audio mixture.***
  - *In practice, some (limited) priors or assumptions are used, e.g.:*
    - *Harmonicity of the lead instrument*
    - *Production model of the lead instrument*

- ***Informed melody extraction***
  - Side information is transmitted to the extraction module
  - Extraction is done using the mixture and the side information

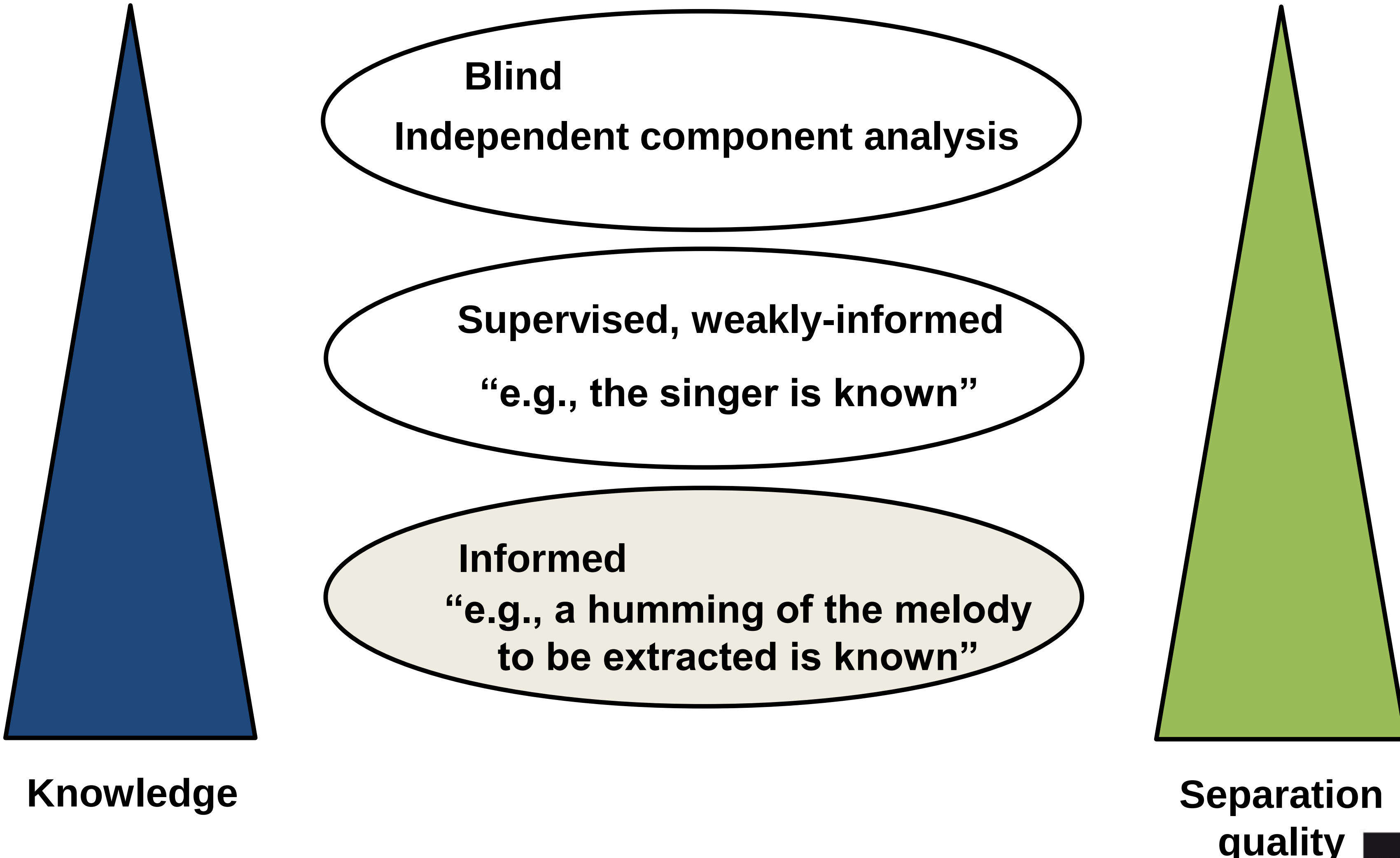


Informed melody extraction

User-guided melody extraction



From *“Blind”* to Informed



Content

- Introduction
- **(« blind ») Main melody extraction**
  - Salient based approaches
  - Source-separation-based approaches
  - Alternative approaches
  - Evaluation
- **Informed main melody extraction**
- **Conclusion**

From monopitch estimation to Main melody extraction

- … a task similar to monophonic pitch extraction ..

- Classically, monophonic pitch extraction estimates a sequence $\hat{f}_{mono}$ of pitch values as :

$$\hat{f}_{mono} = \arg\max_{f} \sum_{\tau} S_x(f_\tau, \tau) + C(f)$$

Sequence of pitch values

Temporal constraints

Function indicating the likelihood
Of the pitch candidates at each time trame τ

From monopitch estimation to Main melody extraction

- **In melody extraction**

$$y(t) = x(t) + n(t)$$

Input music signal — *Target monophonic melody signal* — *« noise » signal*

- But « noise » includes other periodic signals, potentially harmonically related to the melody
- The melody may not be always active or be the dominant source…

From monopitch estimation to Main melody extraction

- **Two main directions for main melody estimation**

  - *Salience-based* approaches, using a modified pitch salience function calculated over the mixed signal.

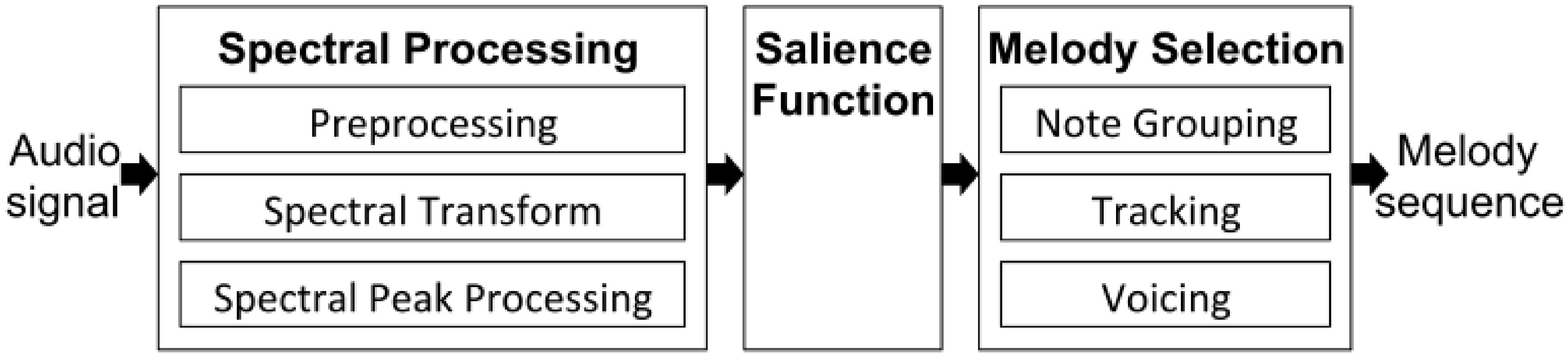
$$\hat{f}_{sal} = \arg\max_f \sum_\tau S'_y(f_\tau, \tau) + C'(f)$$

  - *Source-separation* approaches using an estimation of the separated leading voice component $\hat{x}(t)$

$$\hat{f}_{sep} = \arg\max_f \sum_\tau S_{\hat{x}}(f_\tau, \tau) + C'(f)$$

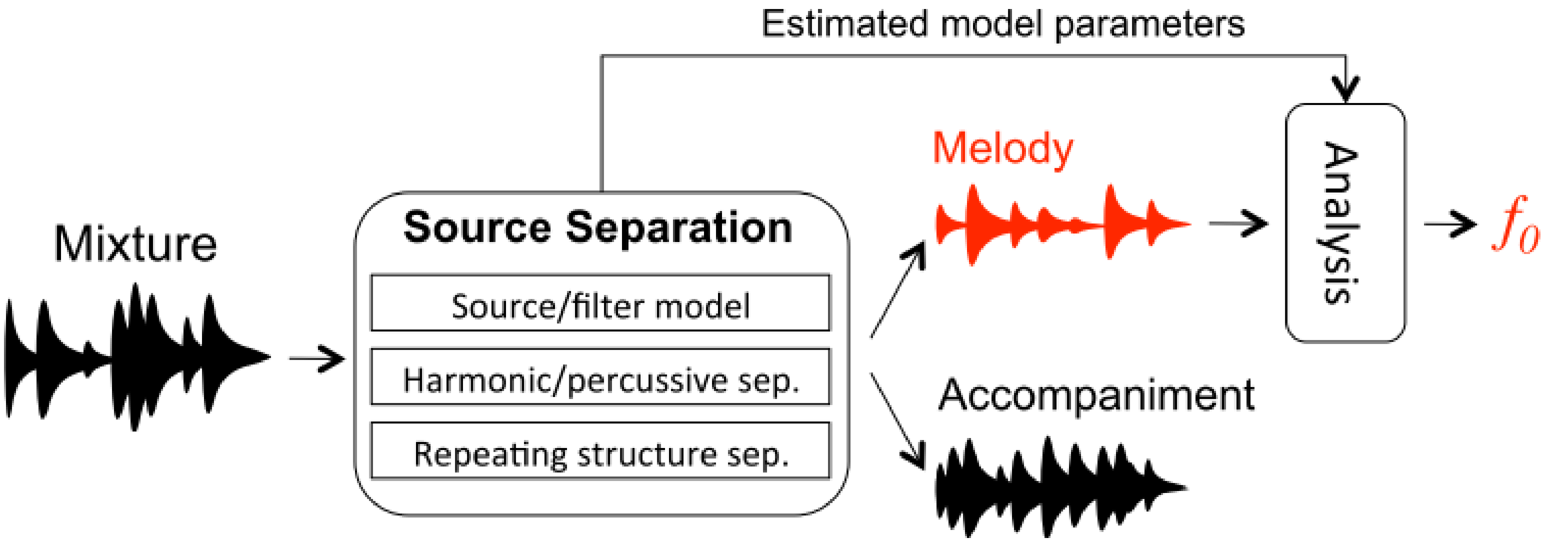
Salience-based approaches

- **Overview**



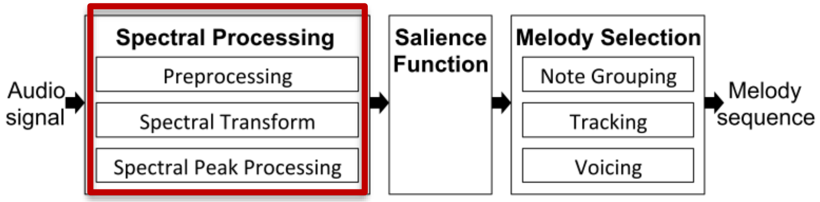
Source separation approaches

- **Overview**

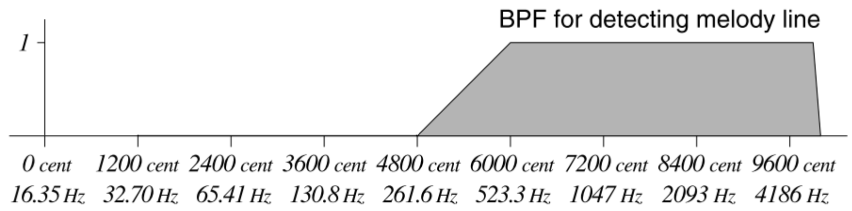


Salience based approaches

Salience-based approaches



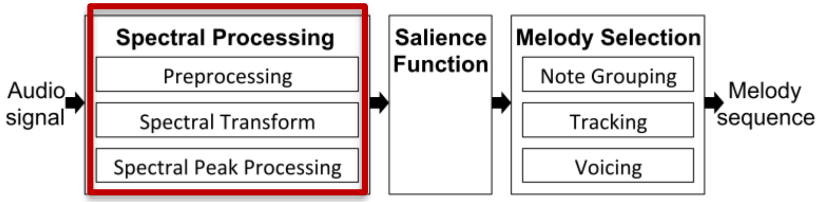
- **Pre-processing**
  - Use of a band-pass-pass filter (Goto2004)



  - Use of an equal-loudness filter (Salomon2012)
    - 10th order infinite impulse response (IIR) filter cascaded with a 2nd order Butterworth high pass filter, (Robinson,2013)

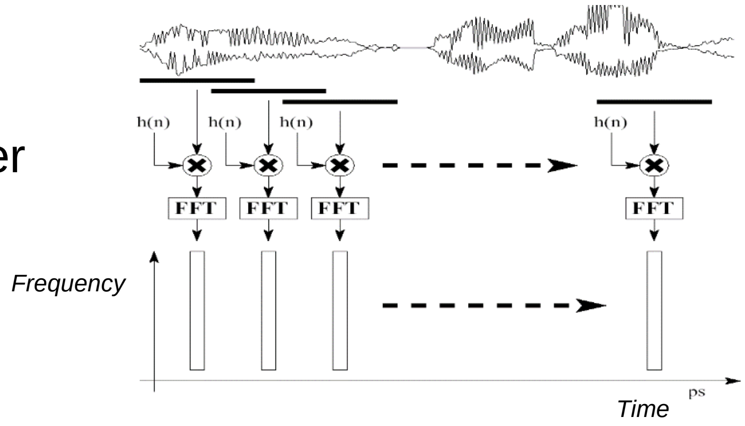
[Goto2004] Goto, M. (2004). A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. Speech Communication, 43, 311–329.
[Salomon2012] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," IEEE Trans. on ASLP, vol. 20, no. 6, Aug. 2012.

Salience-based approaches



Spectral Transform

- Typically based on Short-time Fourier Transform

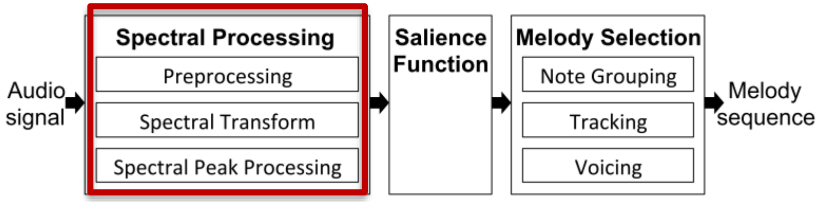


- Others: CQT (Cancela2004), Multi-resolution FFT (Dressler2006), Exploitation of Perceptual principles (Richard2013)

[Dressler2006] Dressler, K. (2006). Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In Proc. 9th Int. Conf. on Digital Audio Effects (DAFx-06), Montreal, Canada.

[Richard2013] G. Richard, S. Sundaram, S. Narayanan "An overview on Perceptually Motivated Audio Indexing and Classification", Proceedings of the IEEE, Sept. 2013.

Salience-based approaches

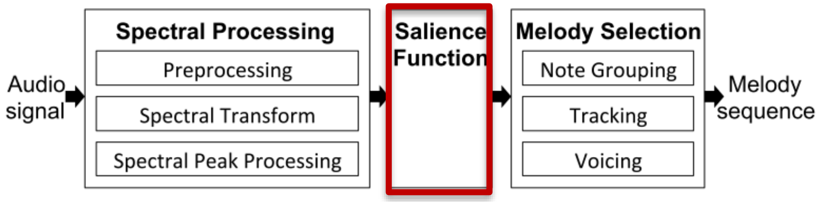


Spectral Peak Processing

- Objective:
  - Removing peaks which are not related to the lead voice (for ex. based on sinusoidality criteria [Goto2004])
  - Or Reducing the influence of timbre (spectral envelope whitening, see for exemple [Cancela2006])
  - Or Computing instaneous frequencies (see for example Dressler2006)

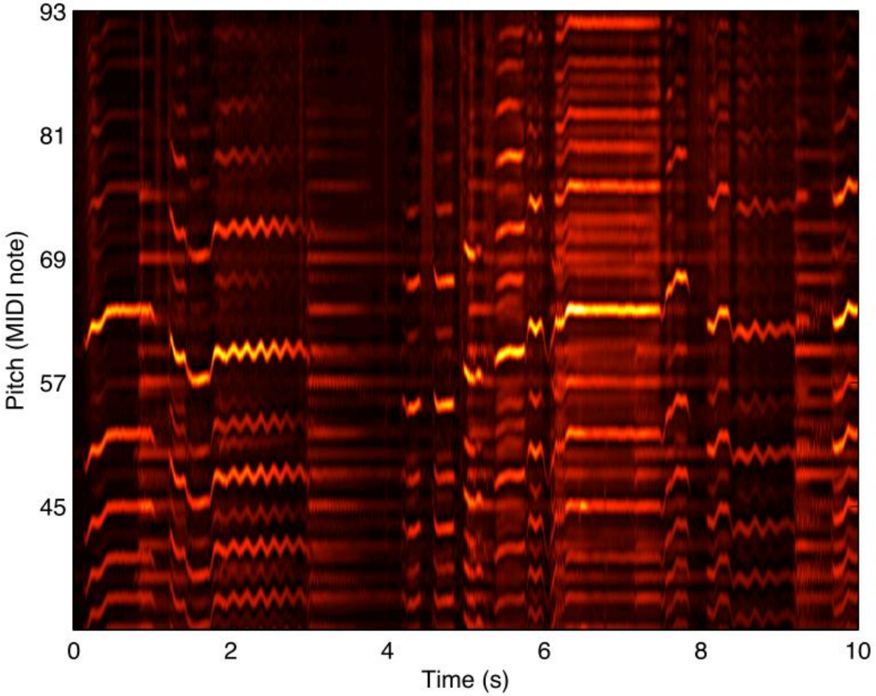
[Cancela2008] Cancela, P. (2008). Tracking melody in polyphonic audio. In 4th Music Inform. Retrieval Evaluation eXchange (MIREX).
[Dressler2006] Dressler, K. (2006). Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In Proc. 9th Int. Conf. on Digital Audio Effects (DAFx-06), Montreal, Canada.
[Goto2004] Goto, M. (2004). A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. Speech Communication, 43, 311–329.

Salience-based approaches



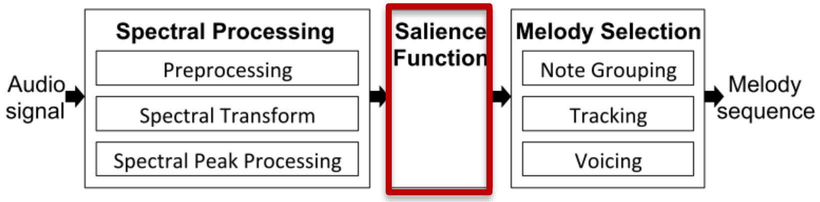
Salience Function

- Provides an estimate of the salience of each possible pitch value over time
- Many approaches …
  - Obtained as a weigthed sum of the amplitude of harmonic frequencies …
  - Use of tone models (Goto2004, Marolt2004)
  - Use of summary autocorrelation (Paiva2006),…



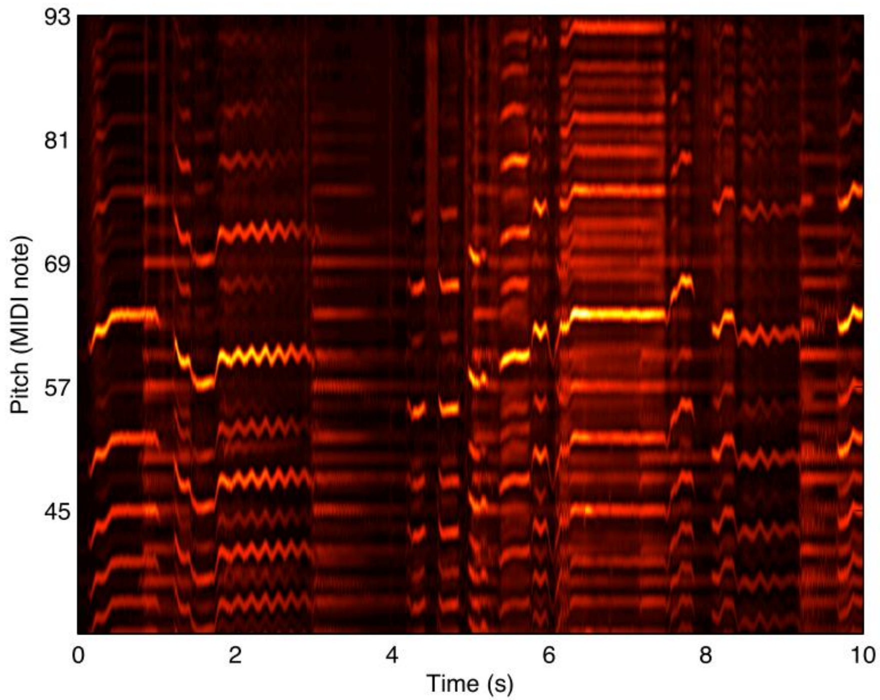
J. Salamon, E. Gomez, D. Ellis, G. Richard, *"Melody Extraction from Polyphonic Music Signals*", IEEE Signal Processing Magazine, March 2014.

Salience-based approaches



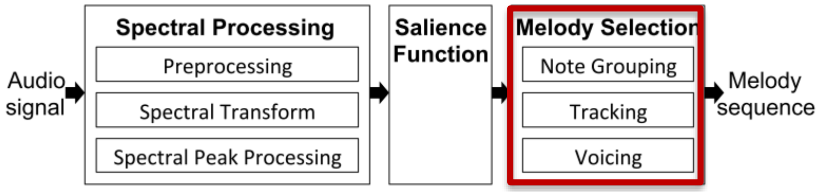
Salience Function

- Presence of ghost notes …
  - Use of « tricks » to reduce these « octave errors »
    - Ex: spectral smoothness,
  - Most errors are practically removed by the tracking stage



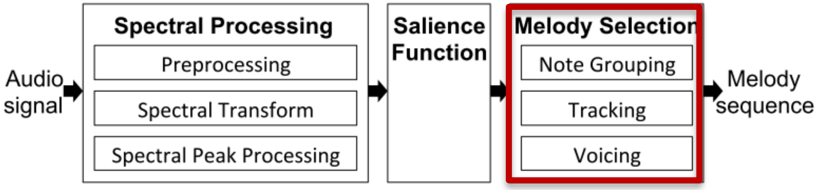
J. Salamon, E. Gomez, D. Ellis, G. Richard, *"Melody Extraction from Polyphonic Music Signals"*, IEEE Signal Processing Magazine, March 2014.

Salience-based approaches



- **Melody selection and tracking**
  - Tracking using:
    - Clustering (Marolt2004),
    - heuristic-based tracking agents (Dressler2006, Goto2004),
    - HMM (Ryynanen, Yeh2012), or
    - Dynamic Programming (Rao2010, Hsu2010)

Salience-based approaches

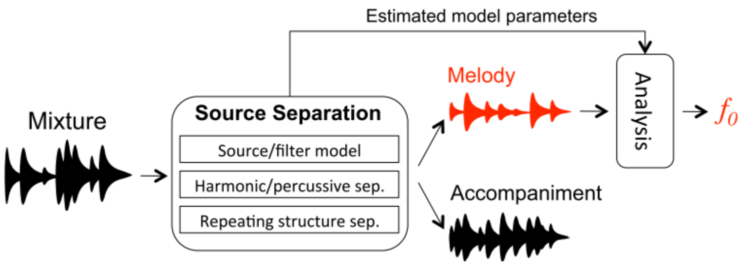


Voicing detection

- Aim:
  - To determine when the melody is present.
  - Usually applied at the very end…
  - Possible approaches: using a pre-frame fixed or dynamic threshold on the salience function
  - Using a « silence » state in HMM based approaches

Source separation based approaches

Source separation approaches



- **Numerous strategies exist:**
  - Exploiting prior information of the singing voice component (e.g. a source/filter model) [Durrieu2010]
  - Exploiting Harmonic / Percussive separation (singing voice is a temporally variable harmonic component) [Ono2010]
  - Exploiting the repeating structure of the background (and the on-repeating nature of the singing voice component) [Rafii2013, Liutkus2012]

An example of singing voice separation Source using Non-Negative Matrix factorization

- **General principle :**
  - The sources are recovered by filtering the mixtures

$$\underbrace{\hat{\mathbf{s}}}_{\text{sources}} = \underbrace{\mathcal{F}}_{\text{filtering technique}} \left\{ \underbrace{\mathbf{x}}_{\text{mixtures}}, \underbrace{\Theta}_{\text{parameters}} \right\}$$

A popular model for audio source separation : NMF

- **NMF = Non-negative Matrix Factorization**

"Templates or Atoms"

Original spectrogram

"Activations"

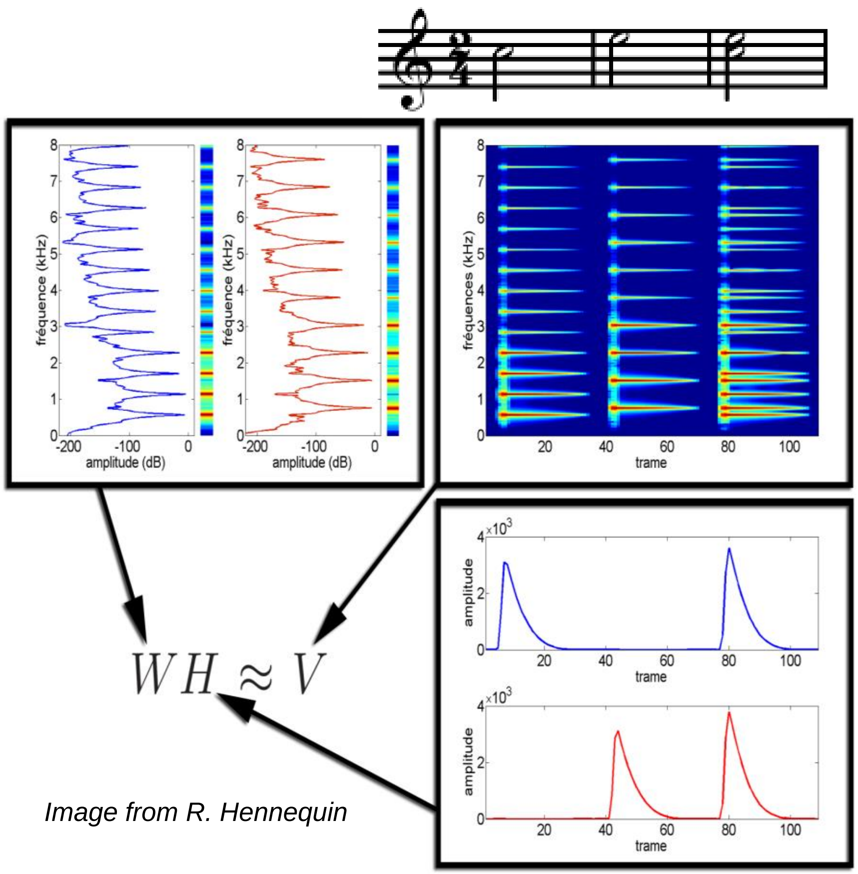
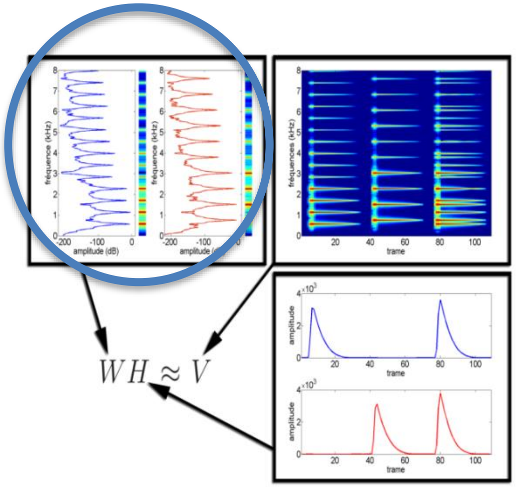


Image from R. Hennequin

A popular model for audio source separation : NMF

- **How the template matrix W and activation matrix H are obtained [Lee&al. 1999]?**



- **Minimization of**

$$D(\mathbf{V}|\hat{\mathbf{V}} = \mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn})$$

- **Typical distances and divergences used:**

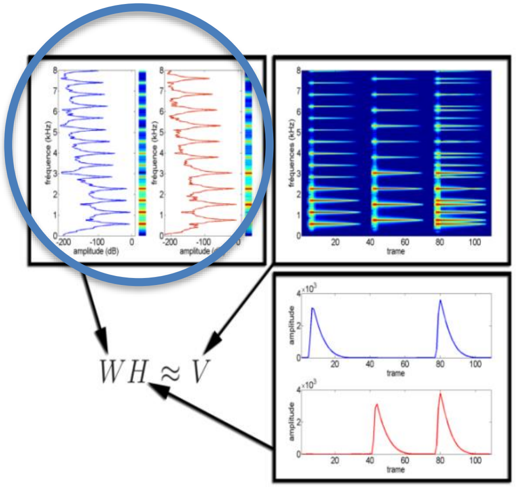
Euclidean	$d_{EUC}(a\|b) = (a-b)^2$
Kullback-Leibler divergence	$d_{KL}(a\|b) = a\log(\frac{a}{b}) - a + b$
Itakura-Saito divergence	$d_{IS}(a\|b) = \frac{a}{b} - \log(\frac{a}{b}) - 1$

β-divergence

$$d_\beta(a|b) = \begin{cases} \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1} & \beta \in \mathbb{R} \setminus \{0,1\} \\ a\log\frac{a}{b} + (b-a) & \beta = 1 \\ \frac{a}{b} - \log\frac{a}{b} - 1 & \beta = 0 \end{cases}$$

A popular model for audio source separation : NMF

- **How the template matrix W and activation matrix H are obtained [Lee&al. 1999]?**



- In general, the cost function is not convex in (**W,H**)…. However, it is **separately convex** in **W** and **H** (for Euclidean and Kullback-Leibler divergence)

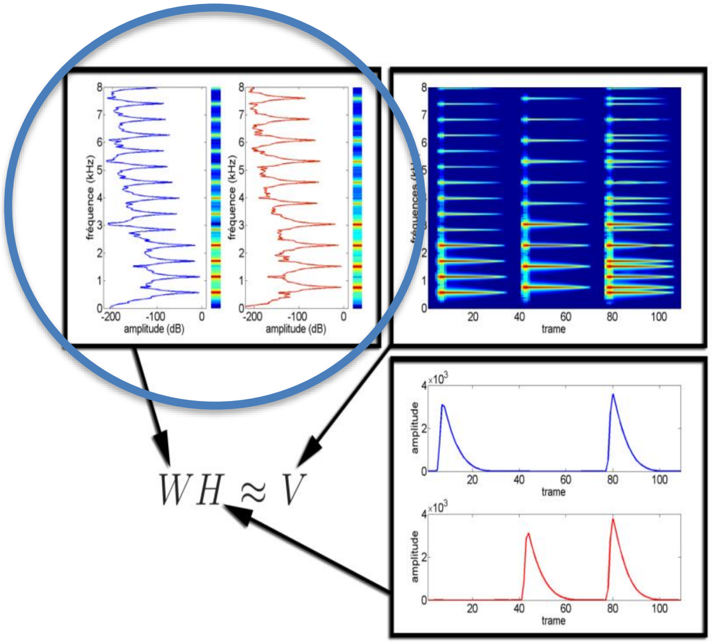
- **The solution is iteratively obtained by means of multiplicative update rules:**
  - **For example with the Euclidean distance:**

$$\begin{cases} \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T(\mathbf{W}\mathbf{H})} \\ \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\mathbf{V}\mathbf{H}^T}{(\mathbf{W}\mathbf{H})\mathbf{H}^T} \end{cases}$$

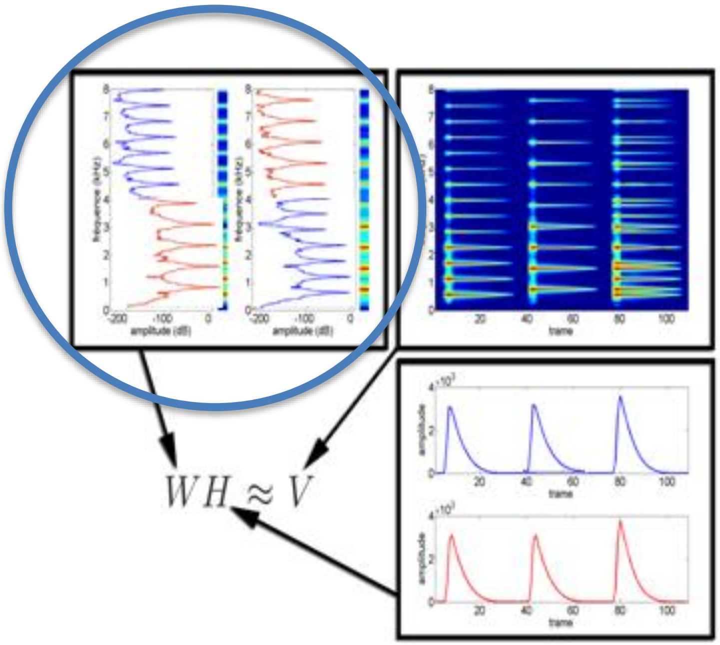
A popular model for audio source separation : NMF

- **NMF does not necessarily provides a semantically meaningful decomposition in absence of "constraints"**

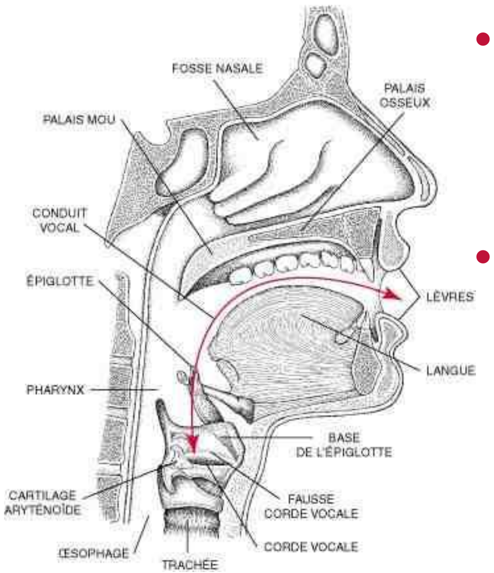
Templates correspond to musical notes



•Templates are built from half of each note and are less semantically meaningful
• Activations are less sparse
• Templates grouping for source recovery



An example of model-based constraints for main melody separation using NMF



- The model: **A**udio = **V**oice + **M**usic
  - The voice **V**oice follows a source filter production model : **V**oice = **S**ource * **F**ilter
  - Each component (Voice and Music) is represented by separate NMF

$$\underbrace{\mathbf{S}_{Audio}}_{\text{Spectrogram of the input audio signal}} = \underbrace{(\mathbf{W}^{F_0}\mathbf{H}^{F_0}) \odot (\mathbf{W}^{\phi}\mathbf{H}^{\phi})}_{\text{Spectrogram of the singing voice}} + \underbrace{(\mathbf{W}^{M}\mathbf{H}^{M})}_{\text{Spectrogram of the background music}}$$

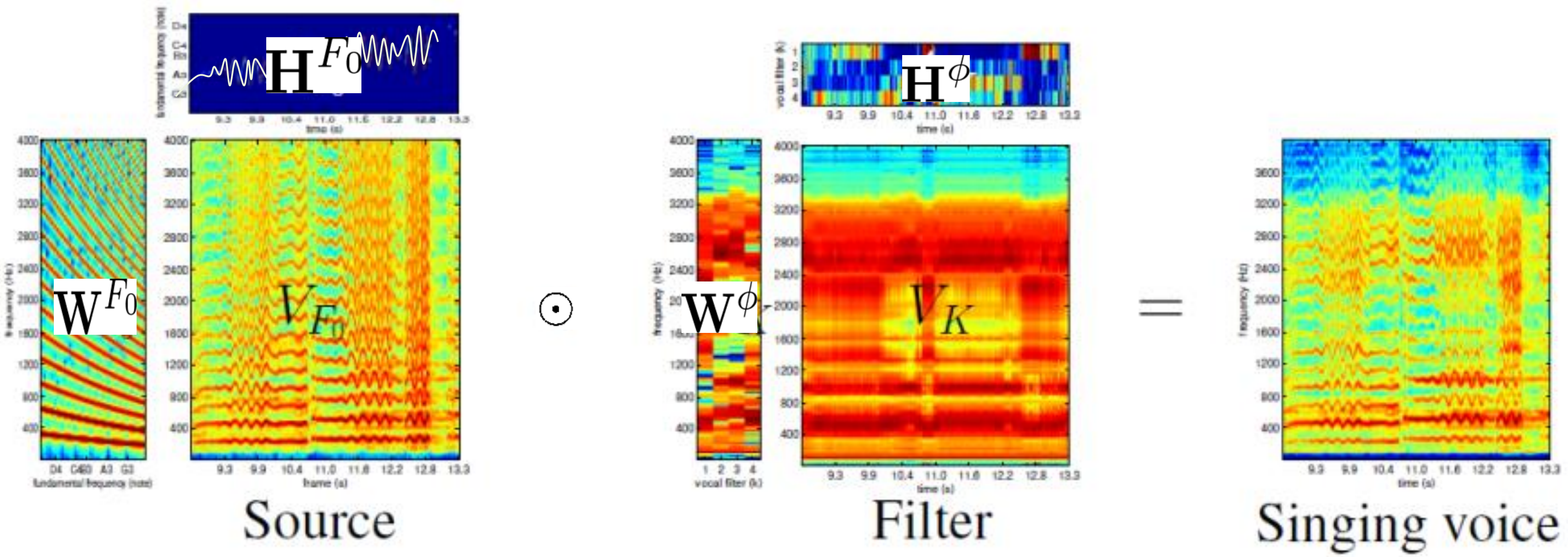
Spectrogram of the input audio signal

Spectrogram of the singing voice

Spectrogram of the background music

An example of model-based constraints for main melody separation using NMF

- **Illustration of the source/filter model with NMF**



J-L Durrieu & al. G, Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals, IEEE Trans. On ASLP, March 2010.
J-L Durrieu, & al. A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011*

An example of model-based constraints for main melody separation using NMF

- ***Exemple of Blind leading voice extraction** [Durrieu&al.2011]*

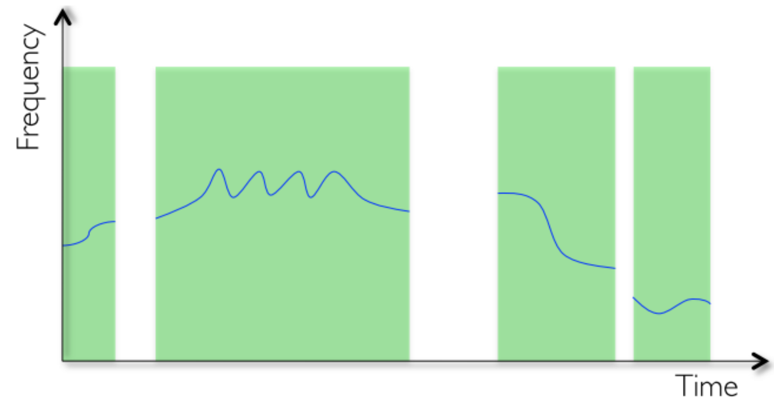
	Original	Backgrounds	Leading voice
Singing voice			
Trumpet			

J-L Durrieu, & al. A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011.*

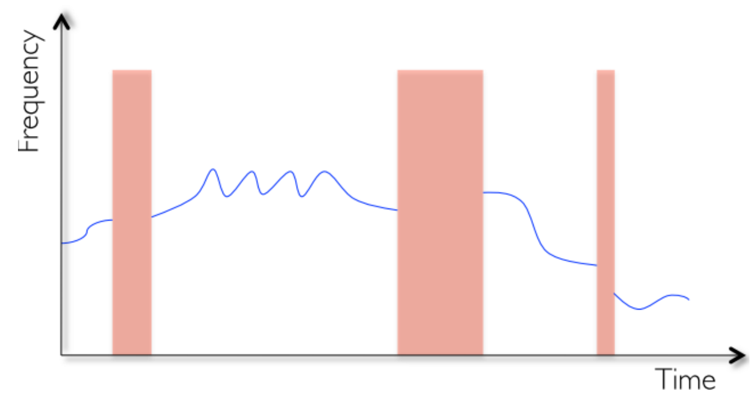
Evaluation

Evaluation: several measures

- **Voicing recall rate:** Proportion of frames labeled as melody frames in the ground truth and that are estimated as melody frames by the algorithm

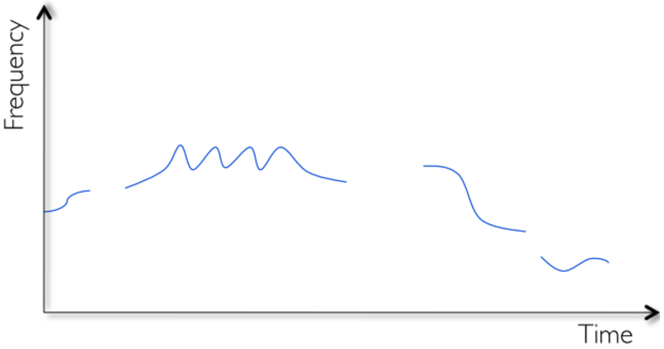


- **Voicing false alarm rate:** Proportion of frames labeled as non-melody in the ground truth and that are estimated as melody frames by the algorithm

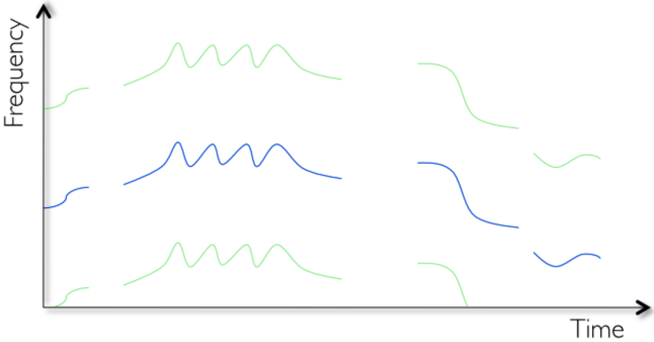


Evaluation: several measures

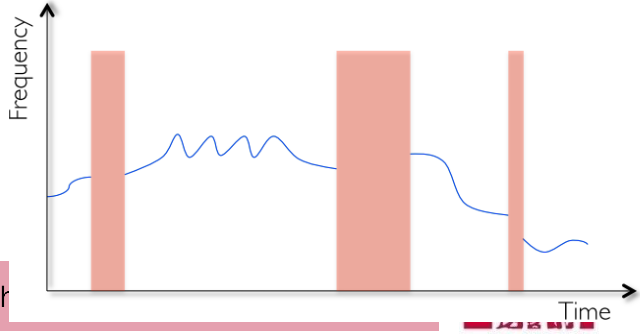
- **Raw pitch accuracy :** Proportion of melody frames in the ground truth for which the pitch estimation is considered correct (i.e. within half a semi-tone)



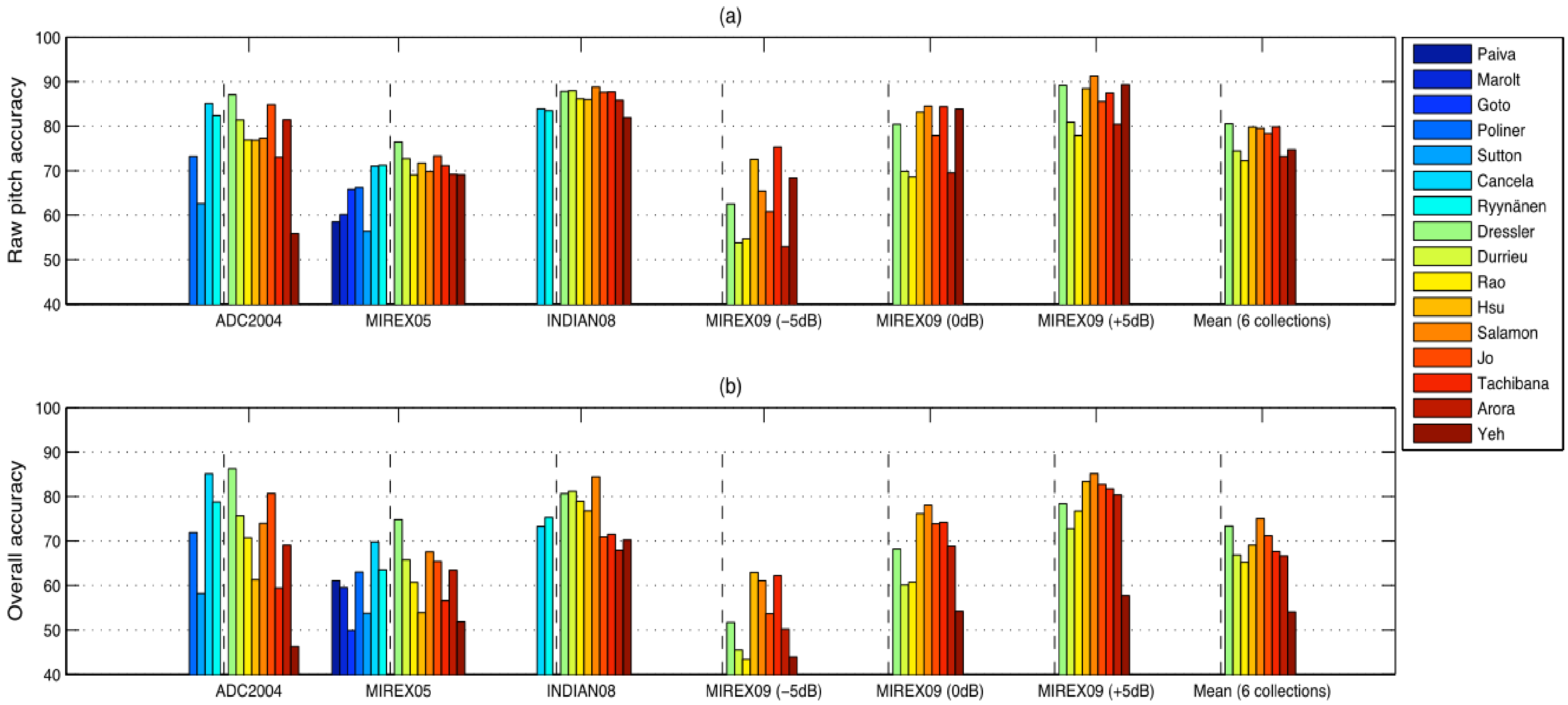
- **Raw chroma accuracy:** same as raw pitch accuracy but without counting octave errors



- **Overall accuracy :** combines pitch accuracy and voicing detection accuracy

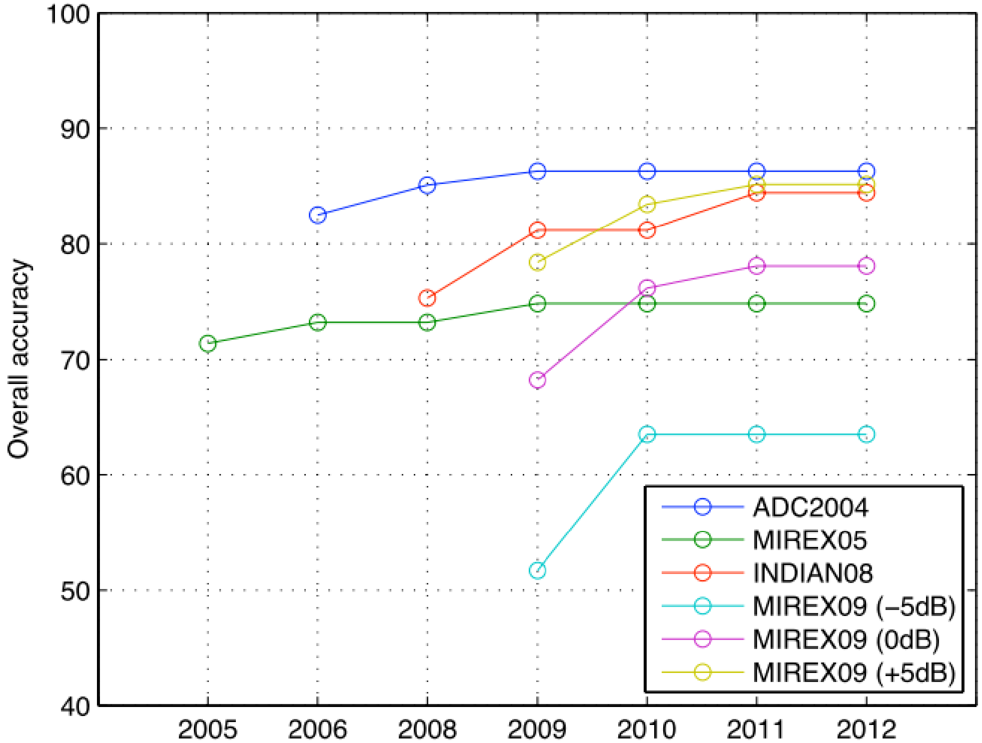


Evaluation


(a)
Raw pitch accuracy
100
90
80
70
60
50
40
ADC2004
MIREX05
INDIAN08
MIREX09 (−5dB)
MIREX09 (0dB)
MIREX09 (+5dB)
Mean (6 collections)
Paiva
Marolt
Goto
Poliner
Sutton
Cancela
Ryynänen
Dressler
Durrieu
Rao
Hsu
Salamon
Jo
Tachibana
Arora
Yeh
(b)
Overall accuracy
100
90
80
70
60
50
40
ADC2004
MIREX05
INDIAN08
MIREX09 (−5dB)
MIREX09 (0dB)
MIREX09 (+5dB)
Mean (6 collections)

Evaluation: are we improving ?

- **Evolution of the best overall accuracy result over the years (on 6 MIREX collections)**



Alternatives approaches

- Using machine learning approaches (Poliner)
- Using the repetitive structure of the music (and non-repeating structure of singing voice) [Rafii2013, Liutkus2012]

- Combining source separation (SS) and salience-based (SB) approaches:
  - SB can bring prior information for SS based approaches
  - And SS can bring a « lead voice enhanced » spectrogram for SB approaches

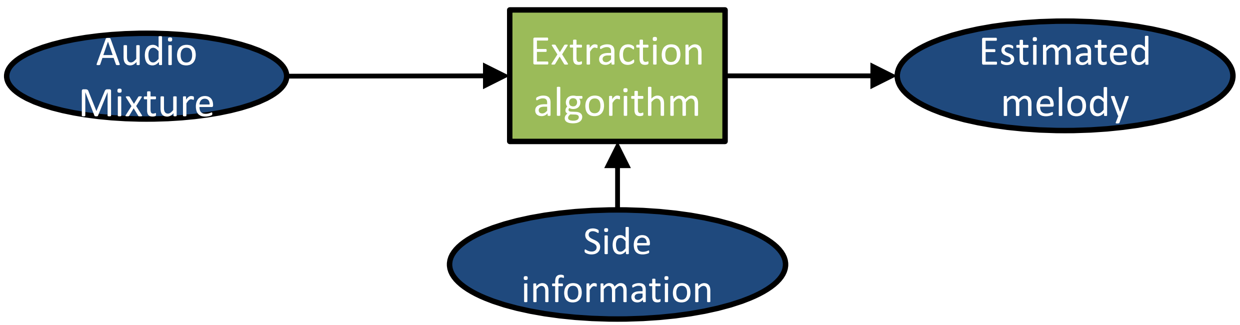
  - Towards informed methods..

[Poliner2006] G. Poliner and D. Ellis, “A classification approach to melody transcription,” in Proc. of ISMIR 2005.
[Rafii2013] Z. Rafii and B. Pardo, “Repeating pattern extraction technique (REPET): A simple method for music/voice separation,” IEEE Trans. on ASLP, vol. 21, no. 1, pp. 71–82, Jan. 2013.
[Liutkus 2012]] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, “Adaptive filtering for music/voice separation exploiting the repeating musical structure,” in IEEE -ICASSP 2012

Informed Source Separation

Towards Informed melody extraction …

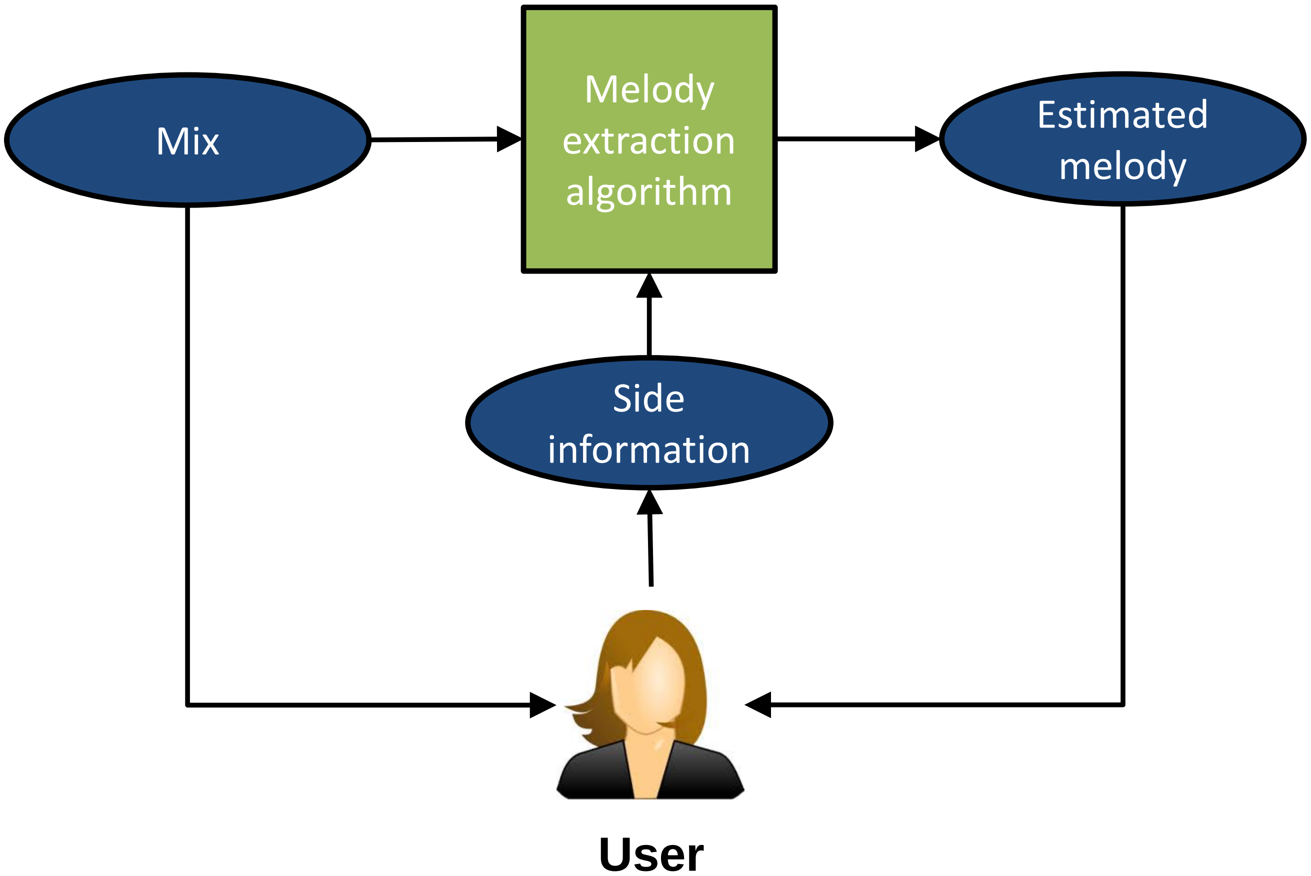
- **Significant performance gain can be obtained by using better prior information….**
- ***Informed melody extraction***
  - Side information is transmitted to the extraction module
  - Extraction is done using the mixture and the side information



A. Ozerov, A. Liutkus and G. Richard, "*ICASSP 2014 tutorial on "Informed Audio Source Separation: Trends, Approaches and Algorithms*" available online: www.loria.fr/~aliutkus/PDF/2014/ICASSP2014_ISS_TUTO.pdf
A. Liutkus, J.-L. Durrieu, L. Daudet and G. Richard, "An overview of informed audio source separation, in Proc. of WIAMIS, 2013, Paris France

Informed melody extraction

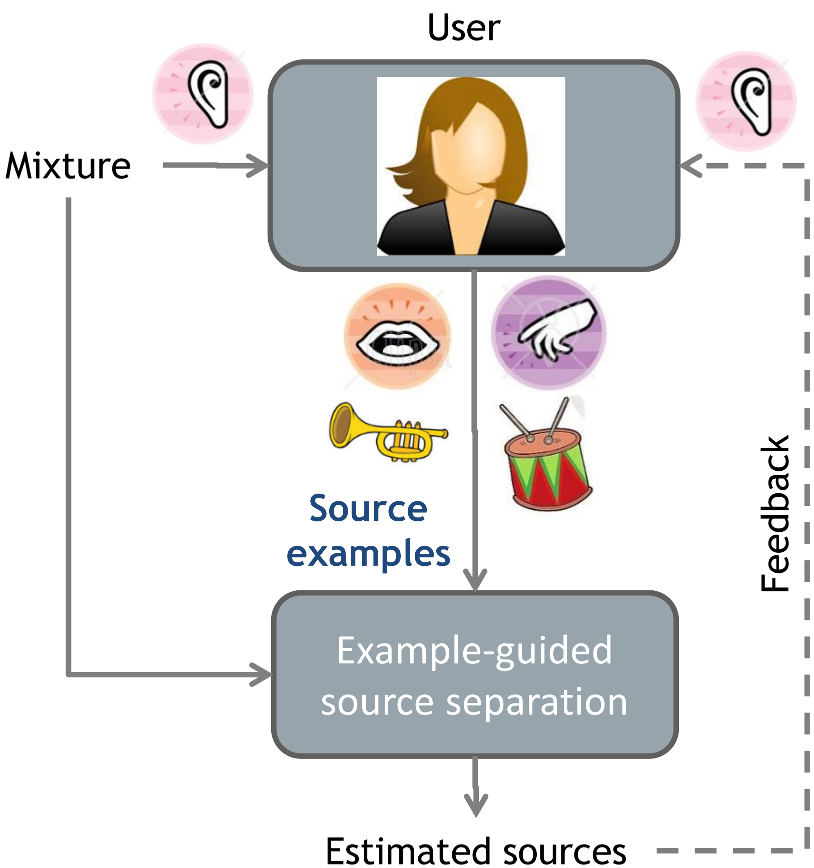
For example : User-guided melody extraction



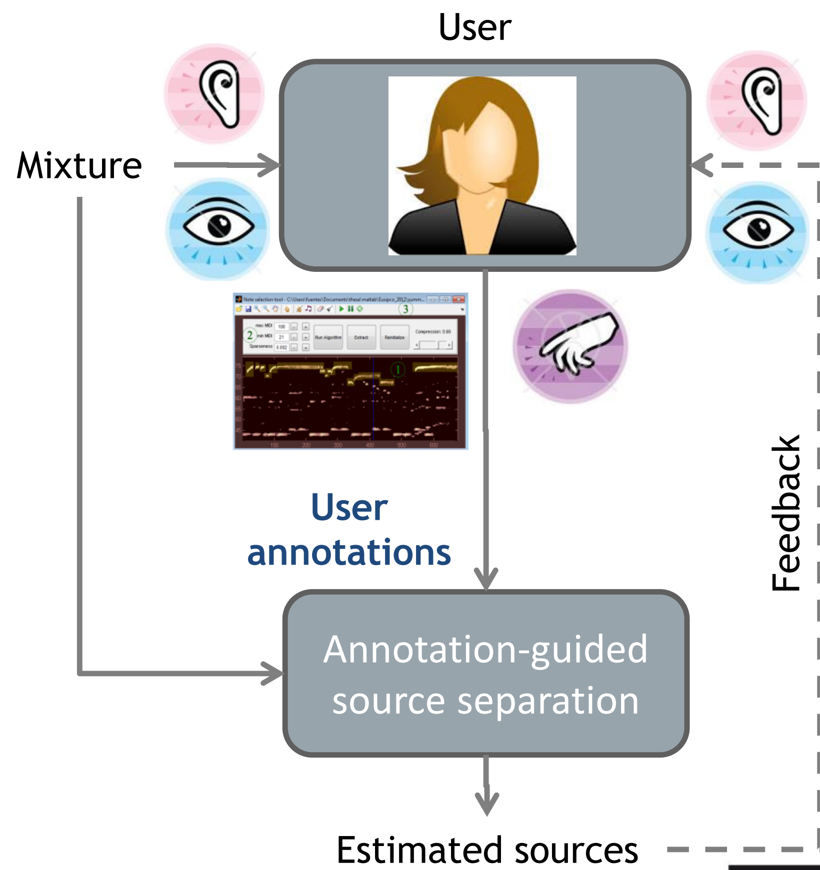
User-guided source separation

Main approaches

Example-based approaches



Annotation-based approaches

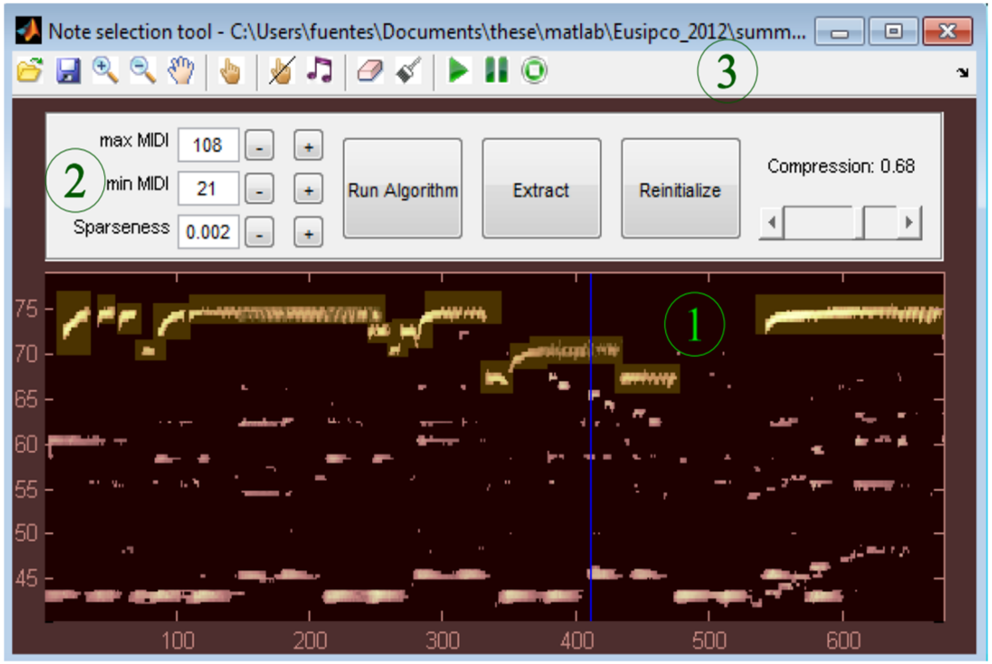


User-guided source separation

Interactive time-pitch annotation-informed separation

- The user paints the parts corresponding to the melody in the GUI
- Algorithm is re-run but with many zero values in the initial decomposition for the melody part
- Several iterations are possible

Demo with a GUI

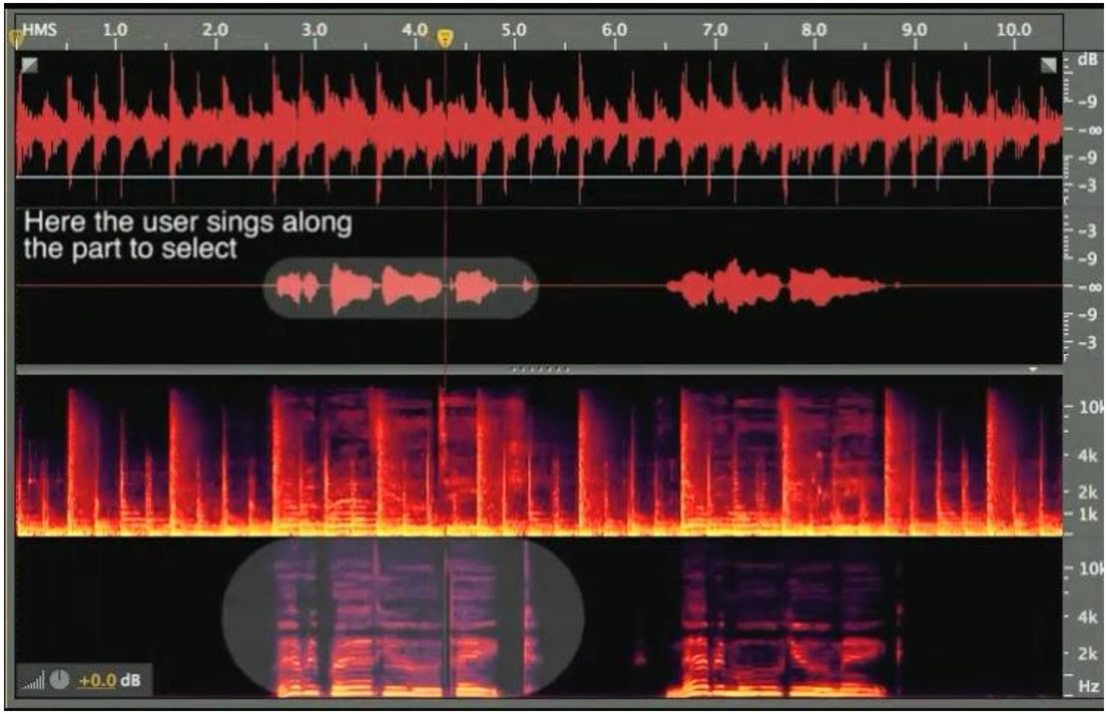


B. Fuentes, R. Badeau et G. Richard : Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In Proc. of EUSIPCO, Bucarest, Romania, 2012.

User-guided source separation
Example-based approaches

Separation by humming

- **Demonstration video [Smaragdis & Mysore 2009]**



Video from P. Smaragdis and G. Mysore, "Separation by Humming": User Guided Sound Extraction from Monophonic Mixtures" in Proc. WASPAA, New Paltz, NY. October 2009
http://www.cs.illinois.edu/~paris/demos/ai/user-guide.mp4

Conclusion

- Steady improvement in melody extraction in the last decade…
- Mainly targeted to singing voice melodies …

- **Challenges:**
  - Going from Singing voice to instrument music
  - Target higher polyphony (5+ music sources)
  - Target songs with backing vocals
  - Improving the voicing detection
  - Public access to larger annotated databases ….

Additional References

- [Hoyer04] P. Hoyer, “Non-negative Matrix Factorization with Sparseness Constraints”, Journal of Machine Learning Research 5 (2004) 1457–1469
- [Smaragdis08] P. Smaragdis , B. Raj et M.V. Shashanka : Sparse and shift-invariant feature extraction from non-negative data. In Proc. of ICASSP, pages 2069–2072, Las Vegas, Nevada, USA, 2008.
- [Virtanen2007] T. Virtanen : Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. on Audio, Speech and Language Processing, 15(3), 2007.
- [Bertin2010] N. Bertin , R. Badeau et E. Vincent : Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription. IEEE Trans. on ASLP, 18(3):538–549, 2010.
- [Raczinsky&al.2007] S. Raczinski, N. Ono, S. Sagayama, “Multipitch analysis with harmonic nonnegative matrix approximation”, in Proc. of ISMIR; Vienna, Austria, 2007
- [Robinson2013] Robinson, D. (2013). Equal loudness filter. Hydrogenaudio Knowledgebase. Online: http://replaygain.hydrogenaudio.org/proposal/equal_loudness.html. 57
- [Cancela2008] Cancela, P. (2008). Tracking melody in polyphonic audio. In 4th Music Inform. Retrieval Evaluation eXchange (MIREX).
- [Marolt2004] M. Marolt, “On finding melodic lines in audio recordings,” in 7th Int. Conf. on Digial Audio Effects (DAFx’04), Naples, Italy, Oct. 2004, pp. 217–221.
- [Hsu2010] C. Hsu and J. R. Jang, “Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion,” in ISMIR Utrecht, The Netherlands, Aug. 2010, pp. 525–530.
- [Yeh2012] T.-C. Yeh, M.-J. Wu, J.-S. Jang, W.-L. Chang, and I.-B. Liao, “A hybrid approach to singing pitch extraction based on trend estimation and hidden Markov models,” in ICASSP 2012
- [Rao2010] V. Rao and P. Rao, “Vocal melody extraction in the presence of pitched accompaniment in polyphonic music,” IEEE Trans. on Audio Speech and Language Processing, vol. 18, no. 8, pp. 2145–2154, Nov 2010.
- [Paiva2006] R. P. Paiva, T. Mendes, and A. Cardoso, “Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness,” Computer Music J., vol. 30, Dec. 2006.
- [Ryynanen2008] M. Ryynanen and A. Klapuri, “Automatic transcription of melody, bass line, and chords in polyphonic music,” Computer Music J., vol. 32, no. 3, pp. 72–86, 2008.