Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure

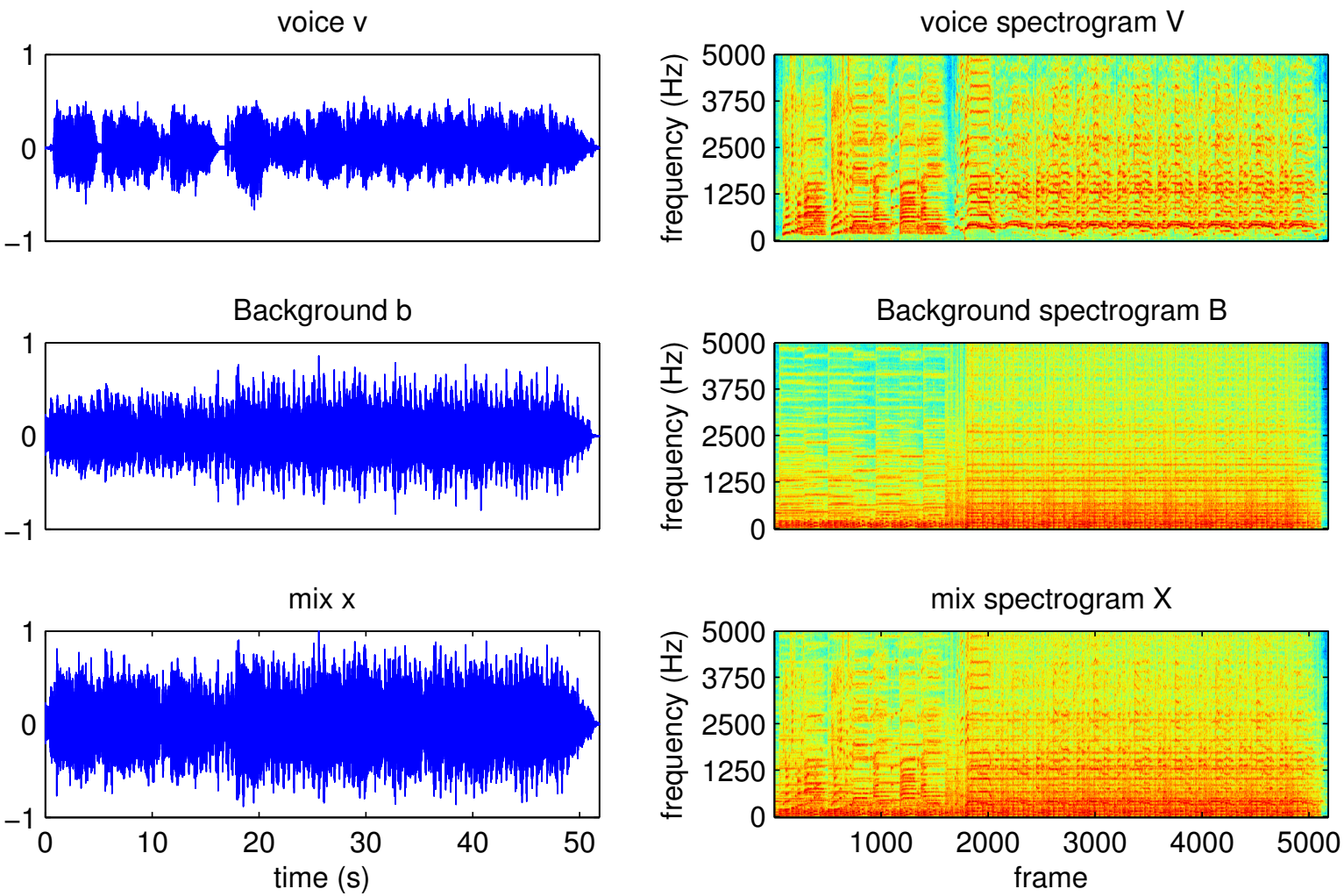
Adaptive REPET

Antoine Liutkus[1], Zafar Rafii[2], Roland Badeau[1], Bryan Pardo[2], Gaël Richard[1]

[1] *Telecom ParisTech, CNRS LTCI, Paris, France*

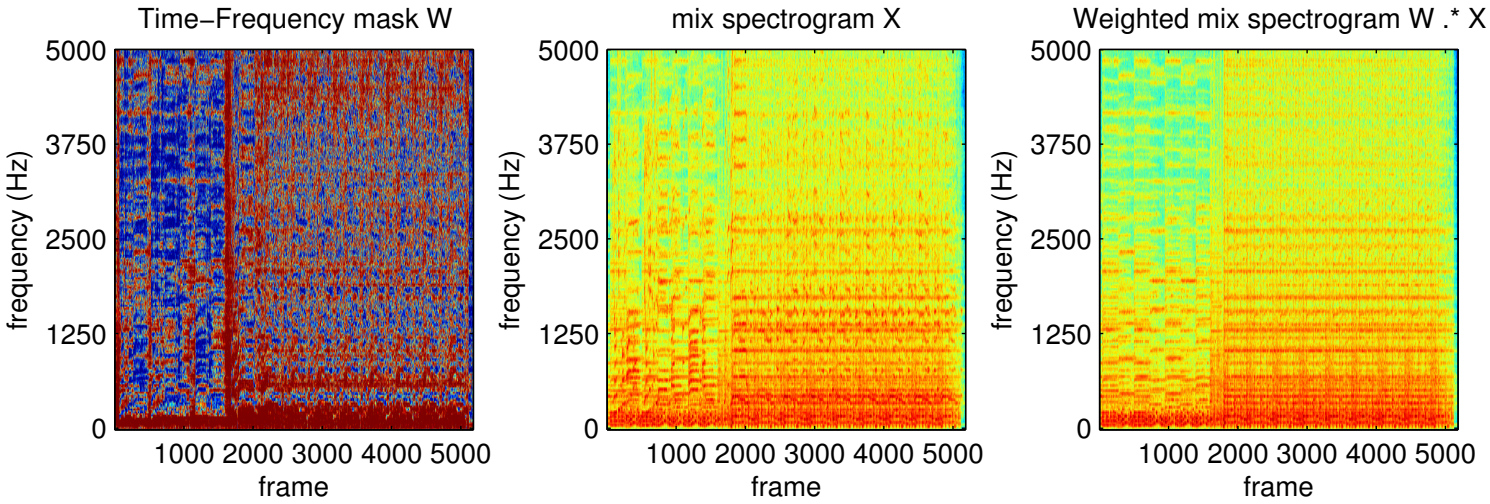
[2] *Northwestern University, EECS Department, Evanston, USA*

Source separation: notation



Separation as an adaptive filter

- Separating a source = filtering the mixture
- Time-varying filter w_t: different for each frame t
- Element-wise weighting of the STFT
- Here: $W \in [0\,1]$



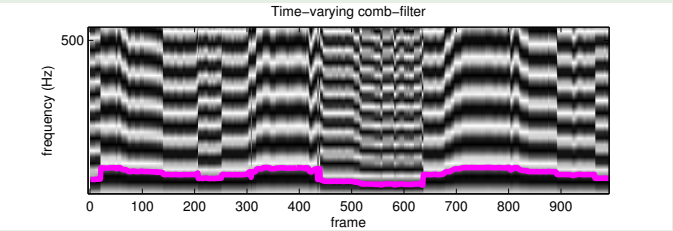
Time-Frequency masks

interpretation

- $W(f,t) \in [0\,1]$: **Proportion of the source of interest** in the mix.
- $W(f,t) \approx 1 \Rightarrow$ TF bin (f,t) **mostly** comes from source of interest
- $W(f,t) \approx 0 \Rightarrow$ TF bin (f,t) mostly comes from **other sources**

Comb filter

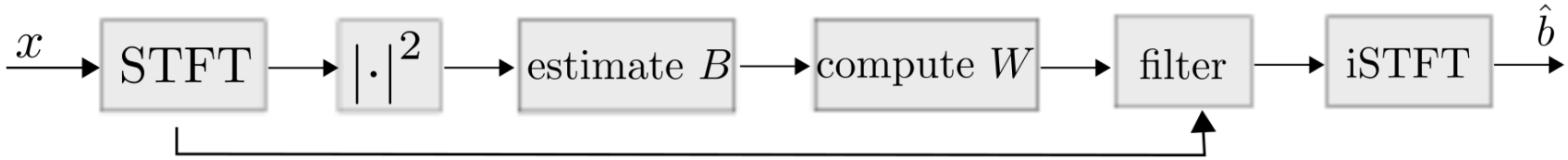
- Given a pitch contour $f_0(t)$, keep multiples of $f_0(t)$



Beyond the harmonic model

Modeling the accompaniement

- Most studies focus on **harmonic voice models**:
  - Voice assumed harmonic and predominant
  - pitch is estimated
  - Filtering e.g. through comb filters
- **Problems**:
  - breathy voices ? Consonants ?
  - Loud accompaniement ?
- We focus on **a model for the background** B !



Filtering given the model

From the B to the mask

Mask from B alone

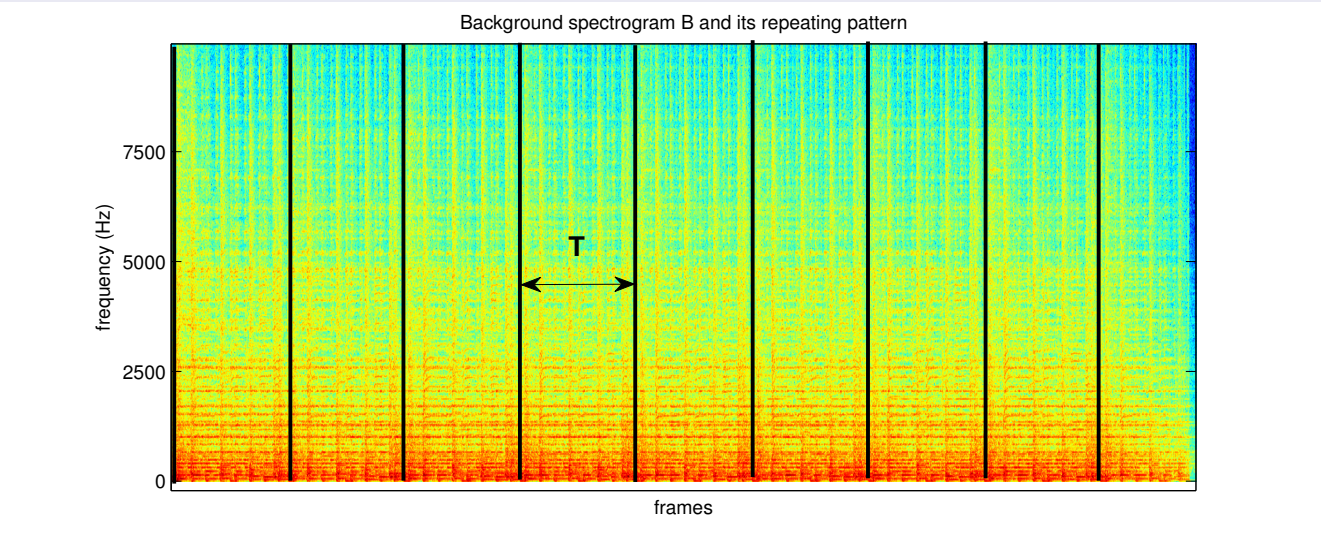
Imagine X and B are available. What is W ?

- $X(f,t)$ close to $B(f,t) \rightarrow W(f,t) \approx 1$
- $X(f,t)$ far from $B(f,t) \rightarrow W(f,t) \approx 0$
- Binary Mask: 0 or 1 based on a thresholding of $\frac{B}{X}$
- Soft mask:

$$W(f,t) = \exp\left(-\frac{(\log X(f,t) - \log B(f,t))^2}{\lambda^2}\right)$$

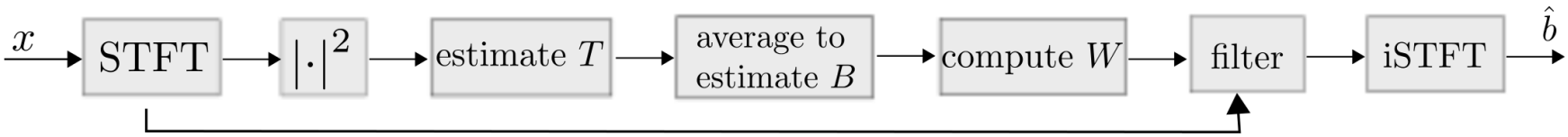
Repeating patterns in music
modeling B

- Musical background is **repetitive** !



- Given several repetitions, **average** to estimate B
- and filter it out !

REpeating Pattern Extraction Technique (REPET)



Original REPET algorithm

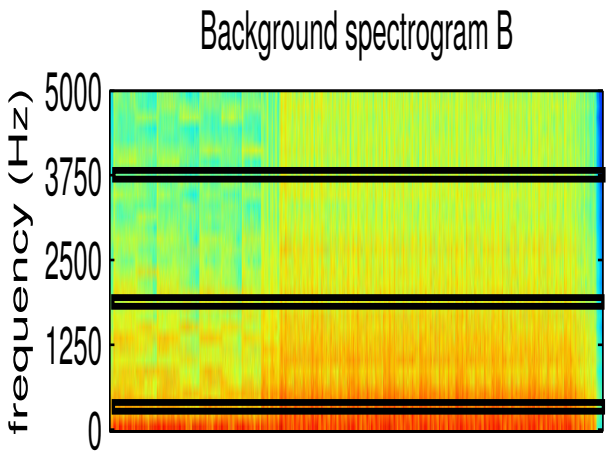
- Estimate a **fixed** repeating period T
- Estimate the **fixed** repeating pattern through **averaging**
- Compute W as a **binary mask**

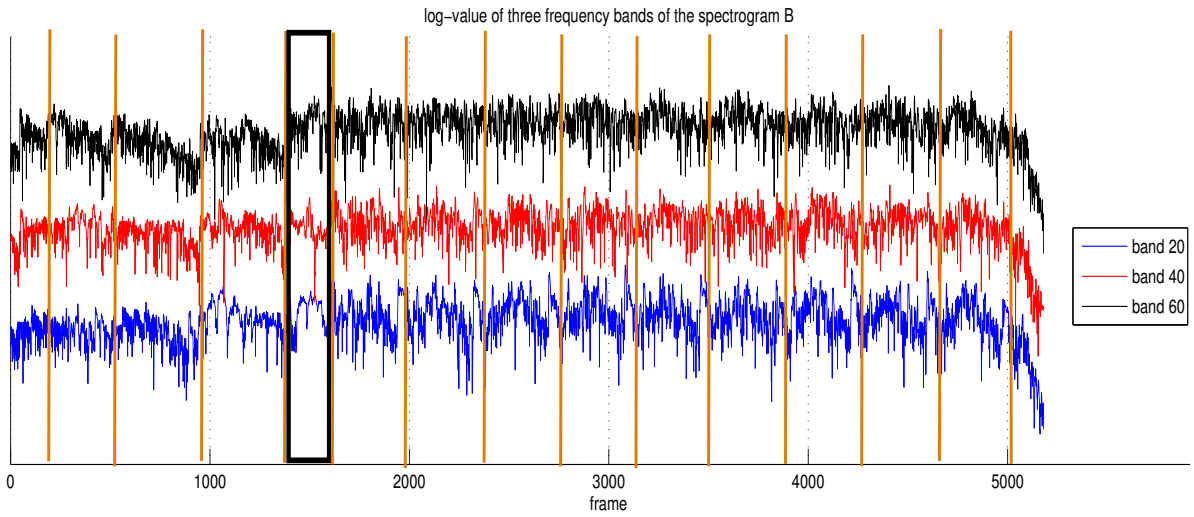
Advantages and limitations of REPET

- Advantages
  - Fast
  - Efficient for constant rythmic patterns (electro, short excerpts)
- Limitations
  - Repeating pattern is changing over time
  - Binary masking leads to artifacts
- We extend REPET to **varying repeating patterns**

Pseudo-periodic patterns

- Patterns are not fixed:
  - period may vary
  - pattern may vary
- Frequency bands of B are assumed pseudo periodic, with the **same** period

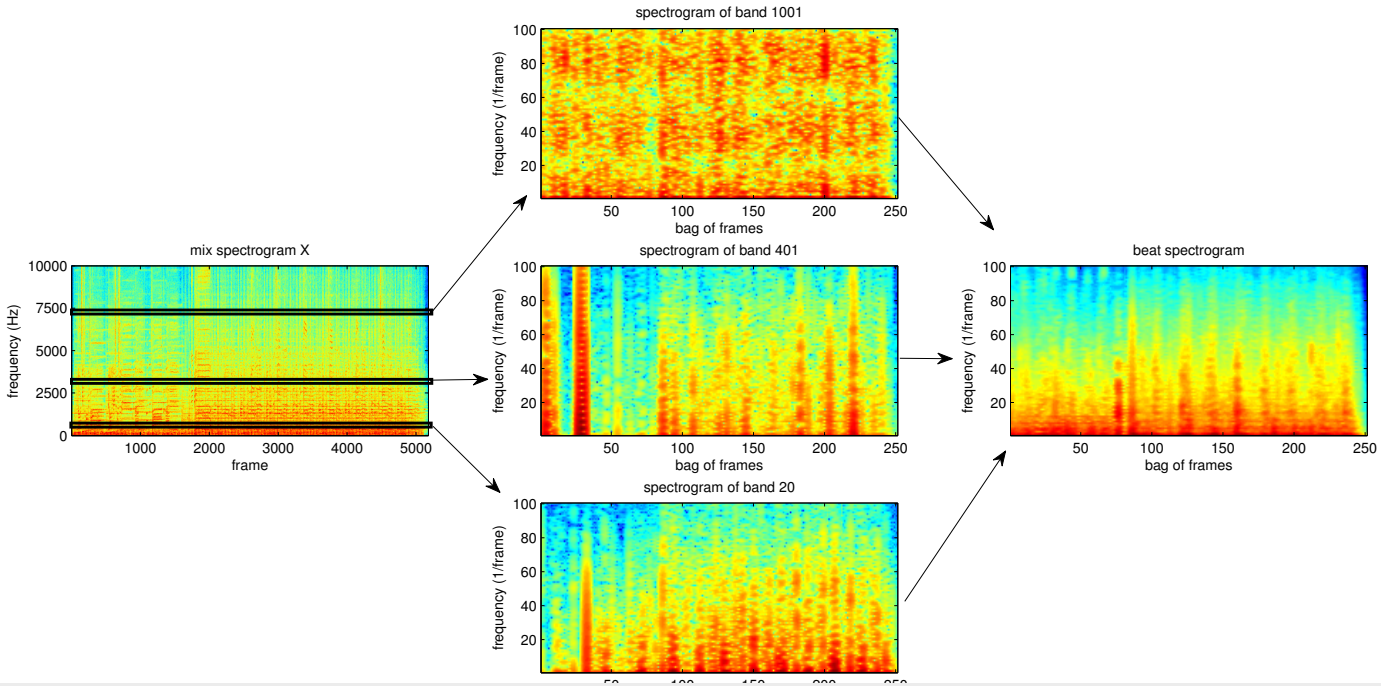




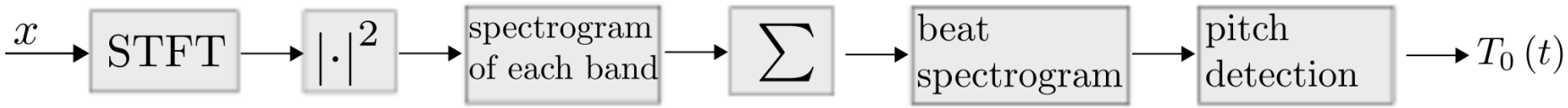
Beat-spectrum estimation

Estimating the period (1/2)

- Perform a short-term analysis of each band
- Add them all together
- **Beat spectrogram** : rythmic content of the signal



Pseudo-period estimation



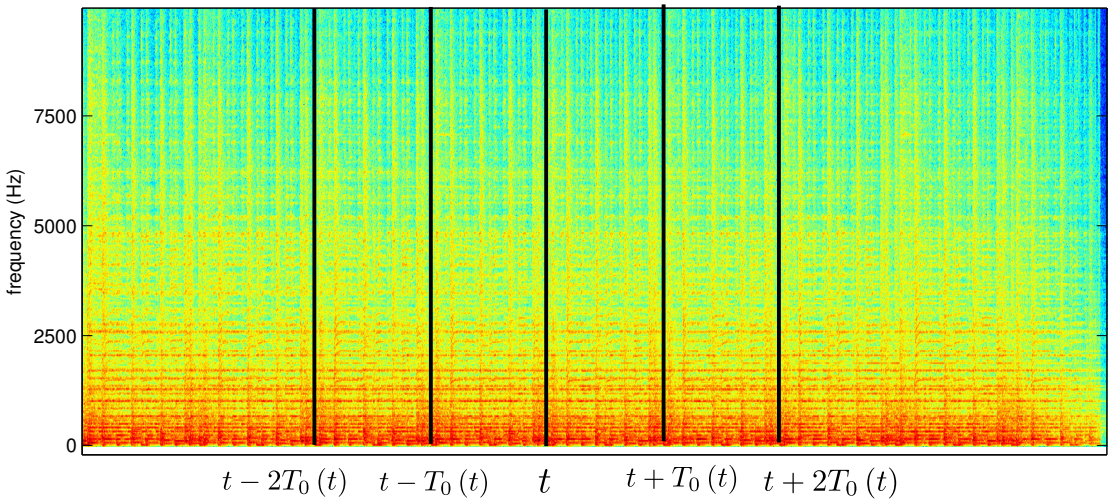
- Compute the beat spectrogram
- Estimate the time-varying repeating period
- Any frequency-based pitch detector will do !

Background model given $T_0(t)$

Background model

$\forall t$, accompaniement is periodic for $2K$ periods around t:

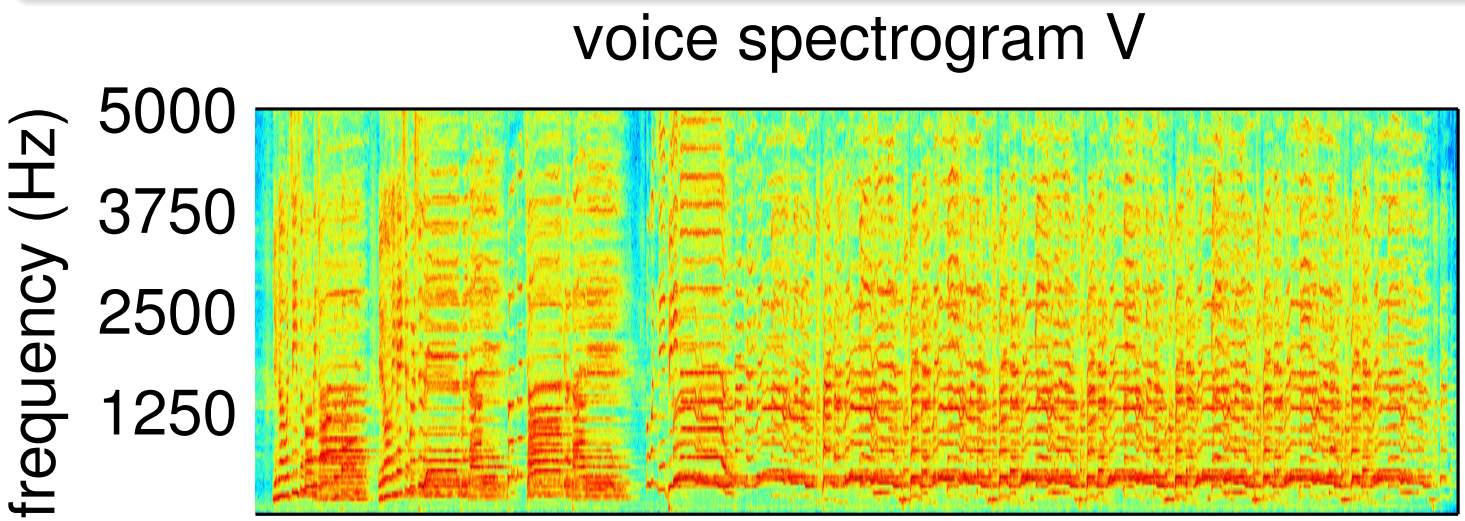
$$B(f,t) = B(f, t + kT_0(t)),\ k = -K \cdots K$$



Voice model

Voice model

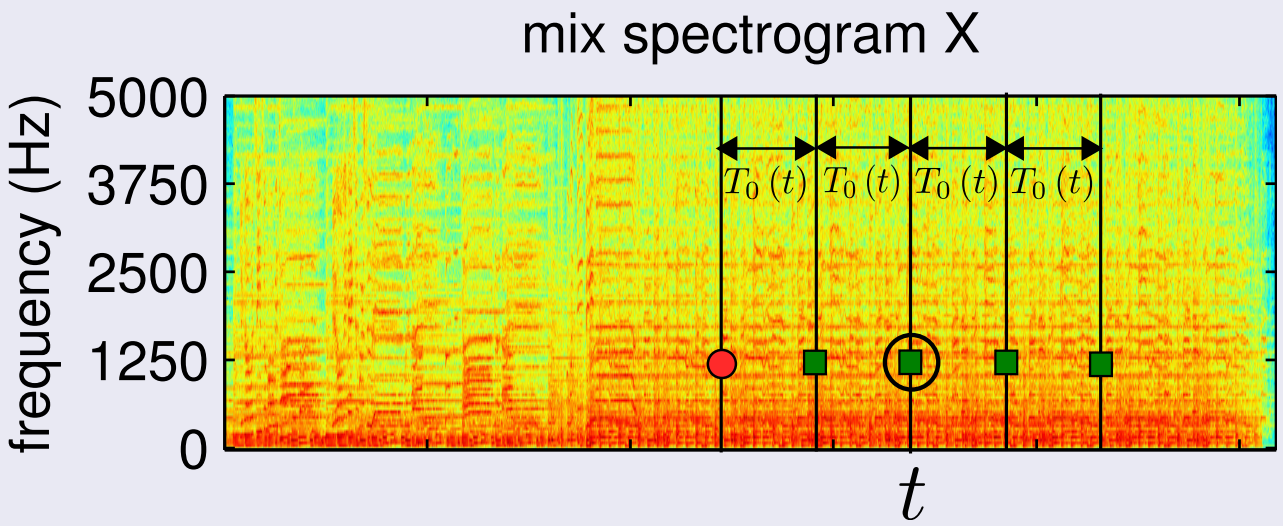
voice V is assumed to be sparse



Background estimation

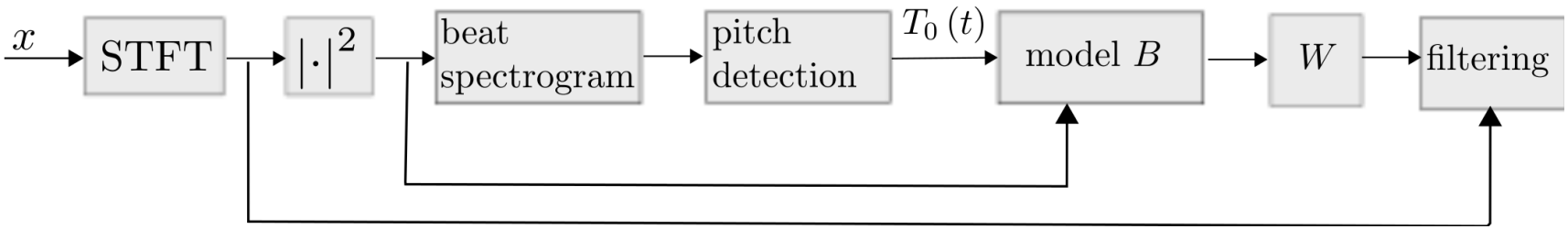
estimation of B given X and $T_0(t)$

- Sparsity of V
  - **Most of the time**, $V \approx 0 \Rightarrow X \approx B$
  - Sometimes, V active$\Rightarrow$ **outliers**



$$\hat{B}(f,t) = \text{median}\left[X\left(f, t + kT_0(t)\right)\right]_{k=-K\cdots K}$$

Adaptive REPET Block diagram



Demonstration

Demonstration on different musical genres

Conclusion

- Adaptive algorithms for complete recordings
- Fast (approx. reading time)
- Extensions : from repetitivity to self-similarity