SINGER MELODY EXTRACTION IN POLYPHONIC SIGNALS USING SOURCE SEPARATION METHODS

Jean-Louis Durrieu, Gaël Richard, Bertrand David
TELECOM ParisTech / CNRS LTCI - 46, rue Barrault - 75634 Paris Cedex 13 - France - e-mail: durrieu@enst.fr

Introduction

- Estimating the main melody in polyphonic music signals: transcribing the sequence of fundamental frequencies played by the dominant instrument;
- Proposed method: separating the desired source thanks to a source-filter model;
- Viterbi smoothing algorithm to find the best melody path.

SIGNAL MODEL

- 2 sources: singer voice V and background music M, observed signal X such that: $X = V + M$,
- Decomposition based on short time Fourier transform (STFT): we consider the N_f first frequency bins of the FFT, the number of frames is denoted T,
- Gaussian modelling of the Fourier transform (FT) Z of the signal $z(t)$. For a centered signal z, wide sense stationary, with FT $Z = \rho\exp(i\theta)$, we have: $Z \sim \mathcal{N}_c(0, \sigma^2) \iff \mathrm{p}(\rho, \theta) = \frac{\rho}{\pi\sigma^2}\exp\left(-\frac{\rho^2}{\sigma^2}\right)$

Singer voice model

- **Gaussian Mixture Model (GMM)** with a **source-filter** modelling. Frame t: one filter σ_k^2 in a dictionary Σ_K and one source $\sigma_{f_0}^2$ in Σ_{F_0}. Conditionally upon state (k, f_0):

$$V(f,t)|k, f_0 \sim \mathcal{N}_c\left(0, a_k^2(t)\sigma_k^2(f)a_{f_0}^2(t)\sigma_{f_0}^2(f)\right) \text{ with}$$

a_k^2 and $a_{f_0}^2$ amplitude coefficients for filter k and source f_0 at frame t.

- **Extended model:** multiple filters and multiple sources to allow more than one note at a time. Every state active at the same time:

$$V(f,t) \sim \mathcal{N}_c(0, \underbrace{\sum_k a_k^2(t)\sigma_k^2(f)}_{V_K(f,t)} \times \underbrace{\sum_{f_0} a_{f_0}^2(t)\sigma_{f_0}^2(f)}_{V_{F_0}(f,t)})$$

- **Dictionary:** $N_f \times K$ filter matrix Σ_K such that $\Sigma_K(f,k) = \sigma_k^2(f)$, $N_f \times N_{\text{notes}}$ source matrix Σ_{F_0} such that $\Sigma_{F_0}(f, f_0) = \sigma_{f_0}^2(f)$; **amplitude matrices** A_K and A_{F_0} such that $A_K(k,t) = a_k^2(t)$ and $A_{F_0}(f_0, t) = a_{f_0}^2(t)$; filter and source contribution respectively denoted $V_K = \Sigma_K A_K$ and $V_{F_0} = \Sigma_{F_0} A_{F_0}$.

Background music model

Instantaneous mixture of R centered Gaussian sources, with variances σ_r:

$$M(f,t) \sim \mathcal{N}_c(0, \underbrace{\sum_{r=1}^{R} a_r^2(t)\sigma_r^2(f)}_{D_R(f,t)})$$

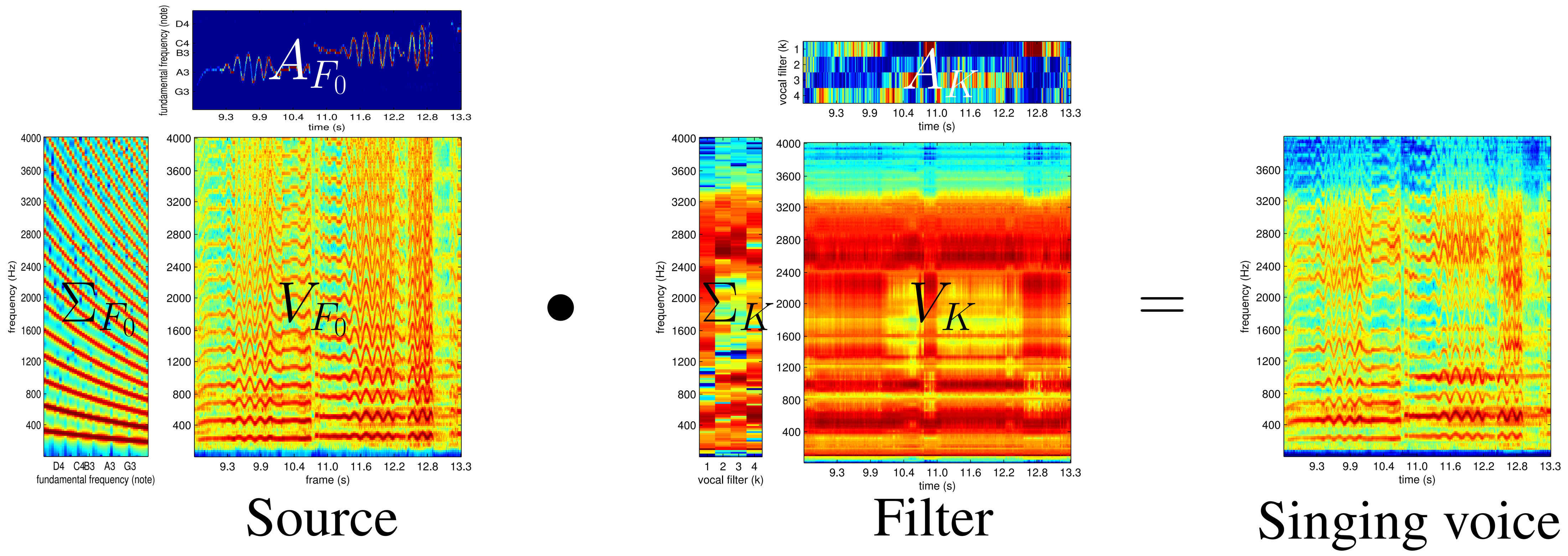
with $\Sigma_R(f,r) = \sigma_r^2(f)$ and $A_R(r,t) = a_r^2(t)$; $D_R = \Sigma_R A_R$.

Mixture signal

$X = V + M \implies$

$$X(f,t) \sim \mathcal{N}_c(0, D(f,t)) \text{ with:}$$
$$D = (\Sigma_K A_K) \bullet (\Sigma_{F_0} A_{F_0}) + \Sigma_R A_R$$

where $\bullet$ is the Hadamard product.



Proposed model for the source-filter GMM: an instantaneous mixture model.

PARAMETER ESTIMATION

Set of parameters to be estimated:
$\theta = \{\Sigma_K, A_K, A_{F_0}, \Sigma_R, A_R\}$

Maximum likelihood criterion

- Problem close to a **non-negative matrix factorisation (NMF)** problem, but solved here in a maximum likelihood framework,
- **Criterion** to be minimized:

$$C(\theta) = -\log(\mathrm{p}_\theta(X)) - \ldots = \sum_{f,t} \log\left(D(f,t)\right) + \frac{|X(f,t)|^2}{D(f,t)}$$

- $\sigma_{f_0}^2$ generated with KLGLOTT88.

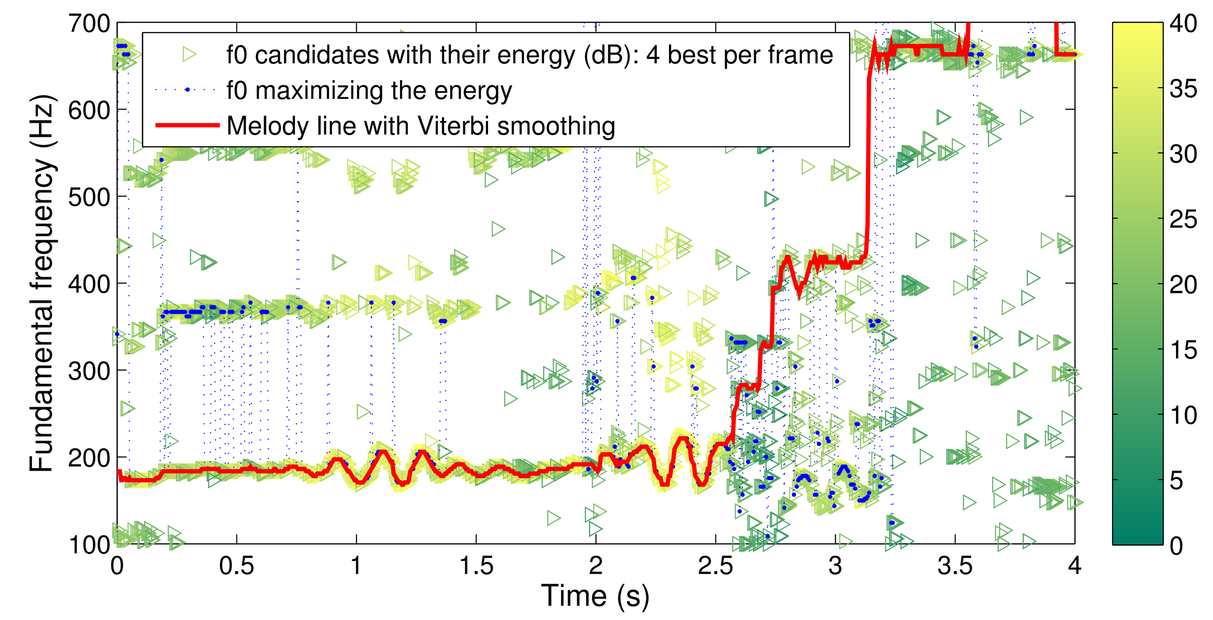
Iterative algorithm

- Estimating θ: finding the zeros of $\frac{\partial C(\theta)}{\partial \theta_i}$, $\theta_i \in \theta$,
- No closed-form solution $\Longrightarrow$ a multiplicative gradient approach is used.

Viterbi smoothing

Main path finding: dynamic programming with trade-off between:

- maximizing the "energy" of the singer voice signal ($\approx A_{F_0}$),
- minimizing the distance between the f_0.

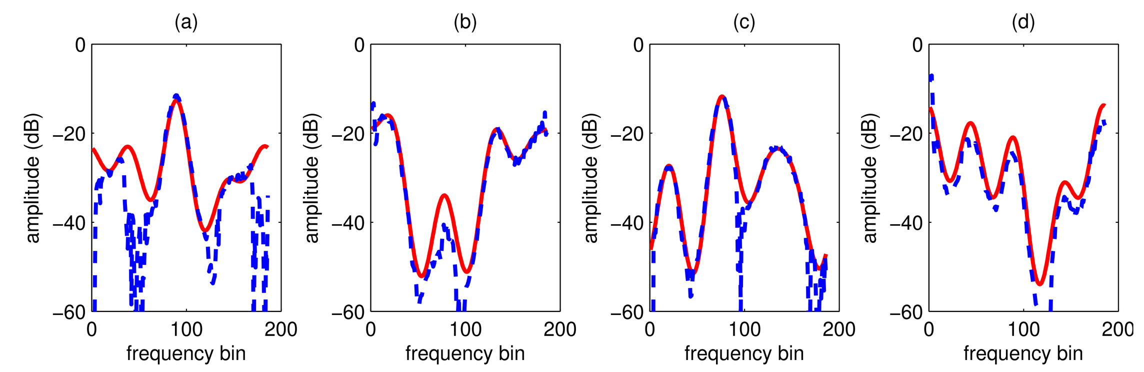


Main path finding on ISMIR 2004 song "opera_male5"

RESULTS

Synthetic data

- Synthetic matrix: random σ_k^2, only one active state (k, f_0) at each frame, with a simulated melody (chirp and natural singing melody).
- Resulting filters: original filters in red, estimated ones in dashed blue lines.



Main melody estimation on real data

ISMIR'04 Opera Songs	Raw Pitch Acc.	Overall Acc.
Proposed Method	**81.2%**	**70.1%**
Dressler	63.0%	64.1%
Poliner	42.6%	47.3%
Ryynänen	64.2%	61.9%
ISMIR'04 Vocal Songs	**Raw Pitch Acc.**	**Overall Acc.**
Proposed Method	**82.6%**	70.5%
Dressler	80.4%	**80.6%**
Poliner	70.7%	70.1%
Ryynänen	81.3%	78.6%

Results of our system compared to MIREX'06 participants

Conclusions

- Novel source separation approach for the main melody extraction task;
- Results at the state of the art for main fundamental frequency estimation;
- Promising results in blind audio source separation (main source extraction and "de-soloing"), results on *http://www.tsi.enst.fr/~durrieu/en/results_en.html*;
- A Bayesian framework that allows to consider several enhancements such as ARMA modelling of the vocal tract filters, HMM smoothing of the transition between states; the system would also profit from silence detection or vocal/non-vocal segmentation.

2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, March 30 - April 4, 2008