

Audio Modeling and Components Separation using Local Linear AM-FM Estimators

Dominique Fourer

IBISC (SIAM) - Univ. Evry / Paris-Saclay

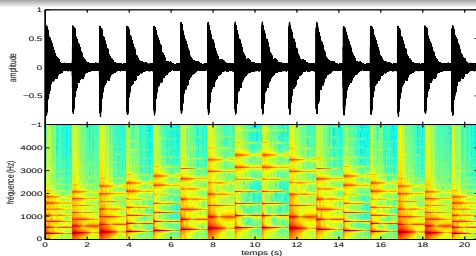
July 7, 2025



Plan

- ① Introduction
- ② Local AM-FM estimators
 - Signal properties
 - Parameters estimation
- ③ Harmonic/Percussive Components Separation
 - Discriminant Analysis of the Local Modulation Rate
 - Separation masks computation
- ④ Numerical results
 - Experimental protocol
 - Comparative evaluation
- ⑤ Discussion

Time-Frequency Analysis



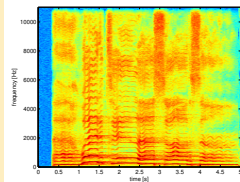
waveform and spectrogram of a piano playing the C major scale.

Non-stationary multicomponent signal processing

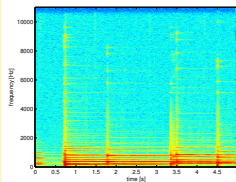
- Disentangle harmonic components (eg. piano, guitar, voice, etc.) from transients (eg. drums, percussion, etc.) to characterize the source and the effects (*i.e.* sinusoids, transients, noise, etc.)
- Computation of sharpen and sparse representations (*i.e.* data modeling, compression)
- Physics meaningful parameters estimation
- Music meaningful representation for transcription ("instantaneous fundamental frequency" [Ville, 48] $\Leftrightarrow$ music pitch)

Observation model

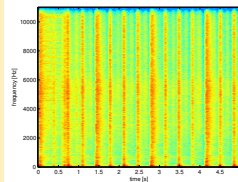
Examples of spectrograms



singing voice



piano (h)



drums (p)

Monophonic Instantaneous Mixture Model

$$x(t) = s_h(t) + s_p(t) \quad (1)$$

Purpose of this work : to blindly compute estimates $\hat{s}_h$ and $\hat{s}_p$, from their observed mixture x

Model

TF orthogonality assumption between sources

Only one and unique source is assumed active at each time-frequency coordinate, i.e. $F_{s_h}^h(t, \omega) F_{s_p}^h(t, \omega) = 0, \forall (t, \omega) \in \mathbb{R}^2$:

$$x(t) = e^{\lambda_x(t) + j\phi_x(t)}, \quad \text{avec } j^2 = -1, \quad (2)$$

- $\lambda_x(t) = l_x + \mu_x t + \nu_x \frac{t^2}{2}$, time-varying log-amplitude
- $\phi_x(t) = \varphi_x + \omega_x t + \alpha_x \frac{t^2}{2}$, time-varying phase (IF being $\frac{d\phi_x(t)}{dt}$).

Signal properties

Derivative of x with respect to time t :

$$\frac{dx}{dt}(t) = \left(\frac{d\lambda_x}{dt}(t) + j \frac{d\phi_x}{dt}(t) \right) x(t) = (q_x t + p_x) x(t) \quad (3)$$

with $q_x = \nu_x + j\alpha_x$ et $p_x = \mu_x + j\omega_x$.

Short-Time Fourier Transform (STFT)

Definition

$$F_x^h(t, \omega) = \int_{\mathbb{R}} x(u) h(t-u)^* e^{-j\omega u} du \quad (4)$$

$$= e^{-j\omega t} \int_{\mathbb{R}} x(t-u) h(u)^* e^{j\omega u} du \quad (5)$$

z^* being the complex conjugate of $z \in \mathcal{C}$.

The derivative of the STFT with respect to t leads to :

$$\frac{\partial F_x^h}{\partial t}(t, \omega) = \int_{\mathbb{R}} x(u) \underbrace{\frac{dh}{dt}(t-u)^*}_{Dh^*} e^{-j\omega u} du \quad (6)$$

$$= -j\omega F_x^h(t, \omega) + e^{-j\omega t} \int_{\mathbb{R}} \frac{dx}{dt}(t-u) h(u)^* e^{j\omega u} du \quad (7)$$

$$= (q_x t + p_x - j\omega) F_x^h(t, \omega) - q_x e^{-j\omega t} \int_{\mathbb{R}} x(t-u) \underbrace{u h(u)^*}_{Th^*} e^{j\omega u} du \quad (8)$$

STFT properties

$$F_x^{\mathcal{D}h}(t, \omega) = -q_x F_x^{\mathcal{T}h}(t, \omega) + (q_x t + p_x - j\omega) F_x^h(t, \omega) \quad (9)$$

with $\mathcal{D}h(t) = \frac{dh}{dt}(t)$ et $\mathcal{T}h(t) = t h(t)$.

generalization of the derivatives with respect to t at order n , $\forall n \in \mathbb{N}^*$ [Fourer, Auger et al, 2017] :

$$F_x^{\mathcal{D}^n h}(t, \omega) = -q_x F_x^{\mathcal{T}^{\mathcal{D}^{n-1}h}}(t, \omega) + (q_x t + p_x - j\omega) F_x^{\mathcal{D}^{n-1}h}(t, \omega) \quad (10)$$

derivatives with respect to ω at order n , $\forall n \geq 1$, using

$$\frac{\partial F_x^h}{\partial \omega}(t, \omega) = j(F_x^{\mathcal{T}h}(t, \omega) - t F_x^h(t, \omega)) :$$

$$\begin{aligned} F_x^{\mathcal{T}^{n-1}\mathcal{D}h}(t, \omega) + (n-1) F_x^{\mathcal{T}^{n-2}h}(t, \omega) = & -q_x F_x^{\mathcal{T}^n h}(t, \omega) \\ & + (q_x t + p_x - j\omega) F_x^{\mathcal{T}^{n-1}h}(t, \omega) \end{aligned} \quad (11)$$

with $\mathcal{D}^n h(t) = \frac{d^n h}{dt^n}(t)$ et $\mathcal{T}^n h(t) = t^n h(t)$

Estimators

With Eqs. (9) and (10), we construct $\forall (t, \omega) \in \mathbb{R}^2$ a linear system with unknowns q_x and $\Psi_x = q_x t + p_x$:

$$\begin{pmatrix} F_x^{\mathcal{D}^{n-1}h} & -F_x^{\mathcal{T}\mathcal{D}^{n-1}h} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \Psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{D}^n h} + j\omega F_x^{\mathcal{D}^{n-1}h} \\ F_x^{\mathcal{D}^h} + j\omega F_x^h \end{pmatrix} \quad (12)$$

Solution

When (12) est reversible (i.e. $|F_x^h(t, \omega)|^2 > 0$), we obtain (tn) :

$$\hat{q}_x^{(tn)}(t, \omega) = \frac{F_x^{\mathcal{D}^n h} F_x^h - F_x^{\mathcal{D}^{n-1}h} F_x^{\mathcal{D}^h}}{F_x^{\mathcal{T}h} F_x^{\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}\mathcal{D}^{n-1}h} F_x^h} \quad (13)$$

$$\hat{\Psi}_x^{(tn)}(t, \omega) = \frac{F_x^{\mathcal{D}^h} F_x^{\mathcal{T}\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^n h}}{F_x^{\mathcal{T}\mathcal{D}^{n-1}h} F_x^h(t, \omega) - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^{n-1}h}} + j\omega \quad (14)$$

Estimators (ωn) are obtained by replacing Eq. (10), by Eq. (11) in the linear system in Eq. (12).

Estimator (ωn)

Similarly we obtain :

$$\begin{pmatrix} F_x^{\mathcal{T}^{n-1}h} & -F_x^{\mathcal{T}^nh} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h} + j\omega F_x^{\mathcal{T}^{n-1}h} \\ F_x^{\mathcal{D}h} + j\omega F_x^h \end{pmatrix} \quad (15)$$

Solution

$$\begin{aligned} \hat{q}_x^{(\omega n)}(t, \omega) &= \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^h - F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^nh}F_x^h} \\ \hat{\psi}_x^{(\omega n)}(t, \omega) &= \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^nh}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^nh}F_x^h} + j\omega \end{aligned} \quad (16)$$

Signal parameters estimation

Model

$$x(t) = e^{\lambda_x(t) + j\phi_x(t)} \quad (17)$$

- $\lambda_x(t) = l_x + \mu_x t + \nu_x \frac{t^2}{2}$, time-varying log-amplitude
- $\phi_x(t) = \varphi_x + \omega_x t + \alpha_x \frac{t^2}{2}$, time-varying phase.

Estimators

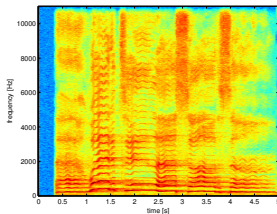
- Log-amplitude linear modulation : $\dot{\lambda}_x(t) = \frac{d\lambda_x}{dt}(t) = \mu_x + \nu_x t$
- Instantaneous frequency : $\dot{\phi}_x(t) = \frac{d\phi_x}{dt}(t) = \omega_x + \alpha_x t$

that can be estimated using $\Psi_x(t) = \dot{\lambda}_x(t) + j\dot{\phi}_x(t) = q_x t + p_x$ with $\hat{q}_x^{(tn)}$ or $\hat{q}_x^{(\omega n)}$.

$$\hat{\nu}_x(t, \omega) = \text{Re}(\hat{q}_x(t, \omega)), \quad \hat{\alpha}_x(t, \omega) = \text{Im}(\hat{q}_x(t, \omega)) \quad (18)$$

$$\hat{\lambda}_x(t, \omega) = \text{Re}(\hat{\Psi}_x(t, \omega)), \quad \hat{\phi}_x(t, \omega) = \text{Im}(\hat{\Psi}_x(t, \omega)) \quad (19)$$

Discretization and Implementation



Discrete-time transforms

- Rectangular approximation : $F_x^h[k, m] \approx F_x^h\left(\frac{k}{F_s}, 2\pi \frac{mF_s}{M}\right)$
- Time index : $k \in \mathbb{Z}$
- Frequency index : $m \in [-M/2 + 1; M/2]$
- Sampling Rate : F_s
- Number of frequency bins : M
- Number of signal samples : N

⇒ Each STFT is considered as a complex-valued matrix of dimension $M \times N$.

Time-Frequency Plane Clustering 1/2

Principle

- Each time-frequency point is associated to a unique source source (orthogonality assumption)
- Local modulation parameter estimation of the mixture x

Estimators

- AM : $\hat{\lambda}_x[k, m]$
- FM : $\hat{\alpha}_x[k, m]$
- AM-FM : $G_x[k, m] = \sqrt{\hat{\lambda}_x[k, m]^2 + \hat{\alpha}_x[k, m]^2}$

Corresponding audio separation features

$G_x[k, m] \in \{|\hat{\lambda}_x[k, m]|, |\hat{\phi}_x[k, m]|, C_x[k, m]\}$ computed from the observed mixture $x[k]$ using Eqs. (18) and (19)

Time-Frequency Plane Clustering 2/2

- Each time-frequency (TF) point $[k, m]$ is described by a set of audio separation features
- We consider a weighted vicinity around the considered TF point :

$$Q_x[k, m] = \left\{ \frac{G_x[k', m'] |F_x[k', m']|^2}{\sum_{k'} \sum_{m'} |F_x[k', m']|^2} \middle| \begin{array}{l} \forall k' \in [k - \Delta_k; k + \Delta_k] \\ \forall m' \in [m - \Delta_m; m + \Delta_m] \end{array} \right\} \quad (20)$$

Components are separated by associating each TF point $[k, m]$ to a source label (i.e. harmonic / percussive) used to compute a separation mask.

Supervised learning

Training

- Linear Discriminant Analysis (LDA) is used to discriminate harmonic from percussive sources.
- Computation of the reference ground truth harmonic separation mask (used only for training)

$$M_h^{(true)}[k, m] = \begin{cases} 1 & \text{if } |F_{sh}^h[k, m]|^2 > |F_{sp}^h[k, m]|^2 \\ 0 & \text{otherwise} \end{cases}, \quad (21)$$

- Percussive reference separation mask :

$$M_p^{(true)}[k, m] = 1 - M_h^{(true)}[k, m] \quad (22)$$

- Computation of the source centroid (i.e. μ_h or μ_p) in the discriminant space from the coefficients computed from the signal mixture.

The trained model correspond to the eigenvectors and the source centroids μ_h or μ_p obtained using the LDA.

LDA in a nutshell

Goal : Finding the best discriminant linear projections of the individuals features (minimize intra-class distance and maximize inter-class distance). We assume that each individual (rows in a given matrix M) is a member of a unique class $c \in [1, C]$.

- Construction of the intra-class variance-covariance matrix :

$$W = \frac{1}{n} \sum_{c=1}^C n_c W_c, \quad (23)$$

where W_c is the variance-covariance matrix computed from the $n_c \times p$ sub-matrix of M made of the n_c individuals included into the class c .

- we define B the inter-class variance-covariance matrix expressed as follows :

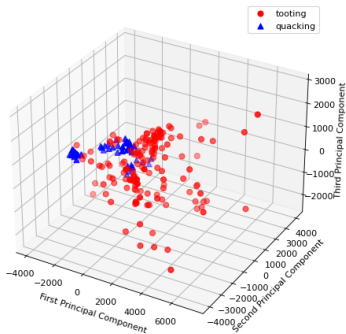
$$B = \frac{1}{n} \sum_{c=1}^K n_c (\mu_c - \mu)(\mu_c - \mu)^T, \quad (24)$$

where μ_c corresponds to the mean vector of class c and μ is the mean vector of the entire dataset.

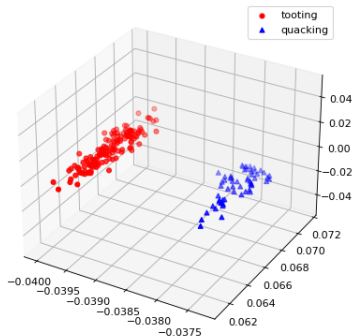
- The eigenvectors of matrix $D = (B + W)^{-1}B$ solve this optimization problem.

Example : [Fourer, Orlowska 2022] D. Fourer and A. Orlowska, Detection and Identification of Beehive Piping

Audio Signals. Proc. DCASE 2022.



(a) PCA



(b) LDA

Components separation

Algorithm

- compute the mixture STFT $F_x^h[k, m]$ using Eq. (5).
- for each TF point, computes $Q_x[k, m]$ using Eq. (20).
- computation of linear projections P_Q using the eigenvectors provided by LDA.
- compute the separation masks :

$$M_h[k, m] = \begin{cases} 1 & \text{if } \|P_Q[k, m] - \mu_h\| < \|P_Q[k, m] - \mu_p\| \\ 0 & \text{otherwise} \end{cases} \quad (25)$$

$$M_p[k, m] = 1 - M_h[k, m].$$

- reconstruct each source signal through the inverse STFT :

$$\hat{s}_h = \text{TFCT}^{-1}(F_x^h[k, m]M_h[k, m]) \quad (26)$$

$$\hat{s}_p = \text{TFCT}^{-1}(F_x^h[k, m]M_p[k, m]) \quad (27)$$

Data

- Public open dataset [E. Cano, 2010]¹
- 10 professional recordings of about 25 seconds
- Reference Harmonic and Percussive signals are available of isolated tracks.

Experimental protocol

- Each sound is resampled at $F_s = 22,05$ kHz
- Mixture x is created according to the instantaneous model ($x = s_h + s_p$).
- STFT are computed using the Hann analysis window :

$$h[n] = \frac{1}{2}(1 - \cos(2\pi \frac{n}{L})), \forall n \in [0; L]$$
- Overlap between successive audio frames (50%) with $\frac{L}{F_s} = 92,9ms$ with a stride $\Delta_n = 1024$ samples.
- $\mathcal{Q}_x[k, m]$ computed with $\Delta_k = \Delta_m = 1$ (3×3 patch size)
- LDA training is completed once using the 300,000 first TF points of the first musical excerpt (about 10 seconds of sound).

1. https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/phase_based_harmonic_percussive_separation.html

Results

Compared methods

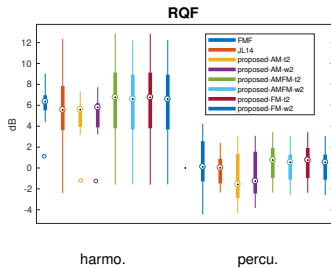
- FMF [Fitzgerald et al. 2014]
- JL14 [Jeong et al. 2014]
- (proposed) AM, FM, AM-FM (t_2)
- (proposed) FM, AM-FM (ω_2)

Metrics

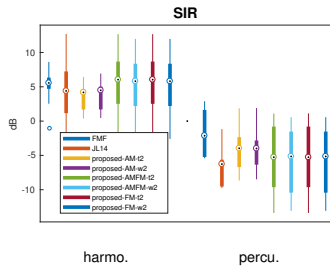
- RQF [Fourer et al. 2016] : $20 \log_{10} \left(\frac{||\hat{x}||}{||\hat{x} - x||} \right)$
- SIR : Interferences (BssEval^a)
- SAR : Artifacts (BssEval)
- SDR : Distortion (BssEval)

a. E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation", IEEE Transactions on Audio, Speech, and Language Processing, 14(4), pp 1462-1469, 2006.

Results



(c)



(d)

Figure : Comparative results expressed using BssEval.

Audio results : <https://fourer.fr/publi/gretsi22/>

Results

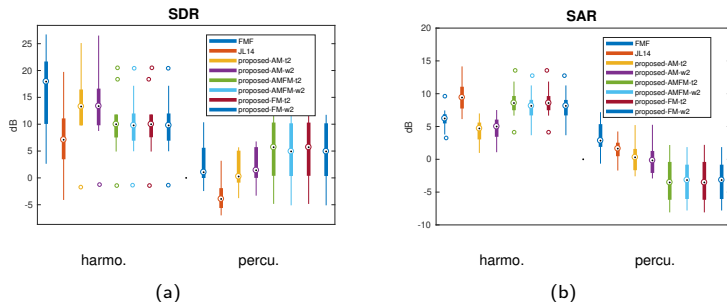


Figure : Comparative results expressed using BssEval.

Audio results : <https://fourer.fr/publi/gretsi22/>

Conclusion and future work

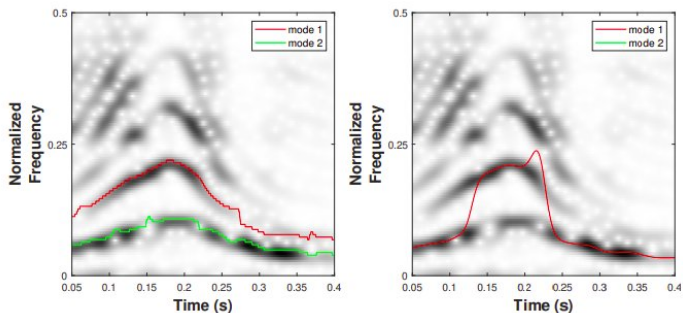
Contributions

- A new approach for separating harmonic and percussive components based on local AM-FM parameters
- Operate blindly in the monaural case (underdetermined degenerated case)
- Physics meaningful estimated parameters are used for separation
- Promising results when compared to the state of the art (blind approach)

Future work

- Propose more robust to noise AM/FM estimators (using regularization)
- Propose non-binary adaptive separation mask (phase reconstruction)
- Optimize the size of the patch used for estimating the separation mask

Robust Component Retrieval using Bayesian Approach



Estimation of the 2 first prominent components in a speech signal comparing (left-hand side) the proposed EM-Laplace method² with the Ridge Detector proposed by Laurent and Meignen in 2021 (IEEE TSP).

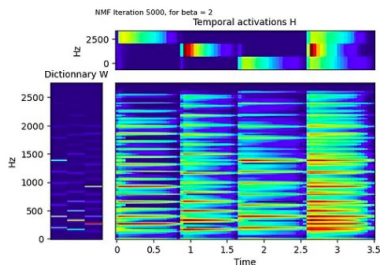
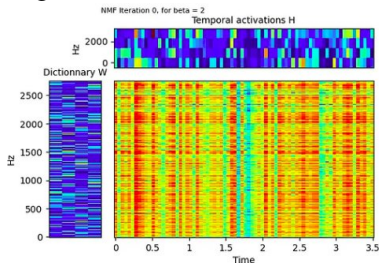
2. Q. Legros, D. Fourer, S. Meignen and M. Calomina, Instantaneous Frequency and Amplitude Estimation in Multi-Component Signals Using an EM-based Algorithm. IEEE Transactions on Signal Processing.10.1109/TSP.2024.3361713

Non-Negative Matrix Factorization (NMF) [Lee, D. D., & Seung, H. S. (1999)]

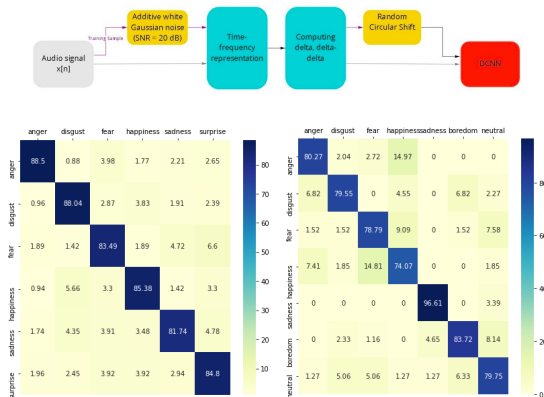
Decompose the Spectrogram $V = |F_x^h|^2$ as a product of two non-negative matrices :

$$V \approx WH \quad \text{s.t., } W \geq 0, H \geq 0 \quad (28)$$

where $W, H = \arg \min_{W, H \geq 0} D(V|WH)$, D being an arbitrary distance or divergence function.



Deep Learning Applied on a time-frequency representation



4

4. S. Xia, D. Fourer, L. Audin-Garcia, J-L. Rouas and T. Schochi. Speech Emotion Recognition using Time-frequency Random Circular Shift and Deep Neural Networks. Proc. Speech Prosody 2022.

Merci !

Article GRETSI 2022 :

Dominique Fourer. Séparation de Sources harmoniques/percussives utilisant des estimateurs locaux de modulation linéaire AM-FM.

GRETSI'22. Nancy, France. Sep. 2022.

Biblio

- [Can10] E. Cano. Phase based harmonic percussive separation. https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/phase_based_harmonic_percussive_separation.html, 2010.
- [FLR + 14] D. FitzGerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet. Harmonic/percussive separation using kernel additive modelling. In Irish Signals Systems Conference and China-Ireland International Conference on Information and Communications Technologies (ISSC'14/CIICT'14), pages 35-40, June 2014.
- [JL14] I.-Y. Jeong and K. Lee. Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints. 21(10) :1197-1200, 2014.
- [VGF06] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. 14(4) :1462-1469, July 2006.

dataset / results : <https://fourer.fr/publi/gretsi22/>