Audio Modeling and Components Separation using Physics and Machine Learning

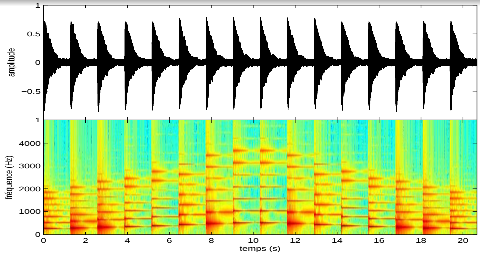
Dominique Fourer

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

February 10, 2026

Plan

Time-Frequency Analysis



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

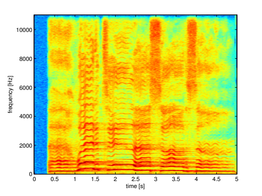
Non-stationary multicomponent signal processing

- 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] ⇔ music pitch)

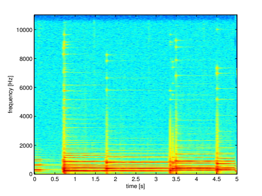
Observation model

$$F_x^h(t,\omega) = \int_{\mathbb{R}} x(\tau)h(t-\tau)^* e^{-j\omega\tau}\, d\tau \quad \text{, with } j^2 = -1 \tag{1}$$

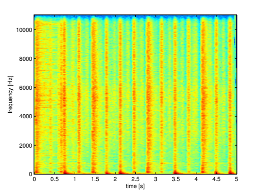
Examples of spectrograms $|F_x^h(t,\omega)|^2$



singing voice



piano (h)



drums (p)

Monophonic Instantaneous Mixture Model

$$x(t) = s_h(t) + s_p(t) \tag{2}$$

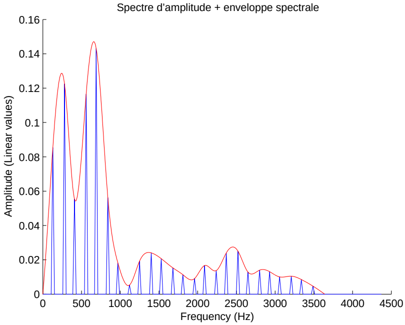
Purpose of this work : to blindly compute estimates $\hat{s}_h$ and $\hat{s}_p$, from their observed mixture x

Harmonic Signal Model

$$s_h(t) = \sum_k a_k(t) e^{j\phi_k(t)} \quad (3)$$

with a_k the time varying amplitude, ϕ_k the time-varying phase of the k-th component. $\frac{d\phi_k}{dt}(t) = f_k(t)$ is the instantaneous frequency (IF).

A pure harmonic sound perceived as a unique pitch is related to a fundamental frequency f_0 and its (almost) integer multiples (its harmonics).



Example : saxophone playing the note Db2 ($f_0 \approx 138.6$Hz)

Audio Parameters

Polyphony

A polyphonic source (eg. piano, guitar, vibraphone, etc.) can play several notes simultaneously. Each note can be modeled using the harmonic model

Timbre

The perceived timbre results from the time envelope (e.g., ADSR), the spectral envelope, the degree of harmonicity, and other features (see Timbre Toolbox [Peeters et al., 2011 ; Fourer et al., 2014]). [a]

a. Python implementation : `https://github.com/dfourer/timbre-descriptor-py`

Unpitched and inharmonic sounds

- Unpitched sounds (e.g., colored noise, percussion) and inharmonic sounds (e.g., bells) also exist.
- Noise cannot be efficiently represented by the presented harmonic model, as it requires a large number of components.

Assumption : The present work assumes non-harmonic sounds (noise and percussion) as residuals included in s_p.

Separation Model (2nd-order polynomial phase)

TF orthogonality assumption between sources

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

$$x(t) = \mathrm{e}^{\lambda_x(t)+j\phi_x(t)}, \quad \text{with } j^2 = -1, \tag{4}$$

- $\lambda_x(t) = l_x + \mu_x t + \nu_x \frac{t^2}{2}$, 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 :

$$\frac{dx}{dt}(t) = \left(\frac{d\lambda_x}{dt}(t) + j\frac{d\phi_x}{dt}(t)\right) x(t) = (q_x t + p_x)x(t) \tag{5}$$

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

Short-Time Fourier Transform (STFT)

Definition

$$F_x^h(t,\omega) = \int_{\mathbb{R}} x(u)h(t-u)^* \, e^{-j\omega u} \, \mathrm{d}u \tag{6}$$

$$= e^{-j\omega t} \int_{\mathbb{R}} x(t-u)h(u)^* \, e^{j\omega u} \mathrm{d}u \tag{7}$$

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

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

$$\frac{\partial F_x^h}{\partial t}(t,\omega) = \int_{\mathbb{R}} x(u) \underbrace{\frac{\mathrm{d}h}{\mathrm{d}t}(t-u)^*}_{Dh^*} \, e^{-j\omega u} \, \mathrm{d}u \tag{8}$$

$$= -j\omega F_x^h(t,\omega) + e^{-j\omega t} \int_{\mathbb{R}} \frac{\mathrm{d}x}{\mathrm{d}t}(t-u)h(u)^* \, e^{j\omega u} \mathrm{d}u \tag{9}$$

$$= (q_x t + p_x - j\omega) F_x^h(t,\omega) - q_x \, e^{-j\omega t} \int_{\mathbb{R}} x(t-u) \underbrace{uh(u)^*}_{Th^*} \, e^{j\omega u} \mathrm{d}u \tag{10}$$

STFT properties

$$\boxed{F_x^{\mathcal{D}h}(t,\omega) = -q_x F_x^{\mathcal{T}h}(t,\omega) + (q_x t + p_x - j\omega) F_x^h(t,\omega)} \quad (11)$$

with $\mathcal{D}h(t) = \frac{\mathrm{d}h}{\mathrm{d}t}(t)$ 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] :

$$\boxed{F_x^{\mathcal{D}^n h}(t,\omega) = -q_x F_x^{\mathcal{T}\mathcal{D}^{n-1} h}(t,\omega) + (q_x t + p_x - j\omega) F_x^{\mathcal{D}^{n-1} h}(t,\omega)} \quad (12)$$

derivatives with respect to ω at order n, $\forall n \geq 1$, using $\frac{\partial F_x^h}{\partial \omega}(t,\omega) = j(F_x^{\mathcal{T}h}(t,\omega) - t\,F_x^h(t,\omega))$:

$$F_x^{\mathcal{T}^{n-1}\mathcal{D}h}(t,\omega) + (n-1)\,F_x^{\mathcal{T}^{n-2}h}(t,\omega) = -\,q_x\,F_x^{\mathcal{T}^n h}(t,\omega) + (q_x t + p_x - j\omega)\,F_x^{\mathcal{T}^{n-1}h}(t,\omega) \quad (13)$$

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

Estimators

With Eqs. (11) and (12), we construct $\forall(t,\omega) \in \mathbb{R}^2$ a linear system with unknowns q_x and $\Psi_x = q_x t + p_x$:

$$\begin{pmatrix} F_x^{\mathcal{D}^{n-1}h} & -F_x^{\mathcal{T}\mathcal{D}^{n-1}h} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \Psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{D}^n h} + j\omega F_x^{\mathcal{D}^{n-1}h} \\ F_x^{\mathcal{D}h} + j\omega F_x^h \end{pmatrix} \qquad (14)$$

Solution

When (14) est reversible (*i.e.* $|F_x^h(t,\omega)|^2 > 0$), we obtain (tn) :

$$\hat{q}_x^{(tn)}(t,\omega) = \frac{F_x^{\mathcal{D}^n h} F_x^h - F_x^{\mathcal{D}^{n-1}h} F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}h} F_x^{\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}\mathcal{D}^{n-1}h} F_x^h} \qquad (15)$$

$$\hat{\Psi}_x^{(tn)}(t,\omega) = \frac{F_x^{\mathcal{D}h} F_x^{\mathcal{T}\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^n h}}{F_x^{\mathcal{T}\mathcal{D}^{n-1}h} F_x^h{}_{(t,\omega)} - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^{n-1}h}} + j\omega \qquad (16)$$

Estimators (ωn) are obtained by replacing Eq. (12), by Eq. (13) in the linear system in Eq. (14).

Estimator (ωn)

Similarly we obtain :

$$\begin{pmatrix} F_x^{\mathcal{T}^{n-1}h} & -F_x^{\mathcal{T}^n h} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \Psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h} + j\omega F_x^{\mathcal{T}^{n-1}h} \\ F_x^{\mathcal{D}h} + j\omega F_x^h \end{pmatrix} \quad (17)$$

Solution

$$\hat{q}_x^{(\omega n)}(t,\omega) = \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^h - F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^h}$$

$$\hat{\Psi}_x^{(\omega n)}(t,\omega) = \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^h} + j\omega \quad (18)$$

Signal parameters estimation

Model

$$x(t) = e^{\lambda_x(t)+j\phi_x(t)} \tag{19}$$

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

Estimators

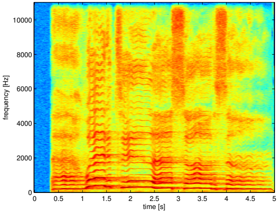
- Log-amplitude linear modulation : $\dot{\lambda}_x(t) = \frac{d\lambda_x}{dt}(t) = \mu_x + \nu_x t$
- Instantaneous frequency : $\dot{\phi}_x(t) = \frac{d\phi_x}{dt}(t) = \omega_x + \alpha_x t$

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 $\hat{q}_x^{(\omega n)}$.

$$\hat{\nu}_x(t,\omega) = \mathrm{Re}\Big(\hat{q}_x(t,\omega)\Big), \quad \hat{\alpha}_x(t,\omega) = \mathrm{Im}\Big(\hat{q}_x(t,\omega)\Big) \tag{20}$$

$$\hat{\dot{\lambda}}_x(t,\omega) = \mathrm{Re}\left(\hat{\Psi}_x(t,\omega)\right), \quad \hat{\dot{\phi}}_x(t,\omega) = \mathrm{Im}\left(\hat{\Psi}_x(t,\omega)\right) \tag{21}$$

Discretization and Implementation



Discrete-time transforms

- Rectangular approximation : $F_x^h[k,m] \approx F_x^h(\frac{k}{F_s}, 2\pi\frac{mF_s}{M})$
- 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

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

Corresponding audio separation features

$G_x[k,m] \in \{|\hat{\dot{\lambda}}_x[k,m]|, |\hat{\dot{\phi}}_x[k,m]|, C_x[k,m]\}$ computed from the observed mixture $x[k]$ using Eqs. (20) and (21)

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 :

$$\mathcal{Q}_{x[k,m]} = \left\{ \left. \frac{G_{x[k',m']} \left| F_{x[k',m']} \right|^2}{\sum_{k'} \sum_{m'} \left| F_{x[k',m']} \right|^2} \right|_{\substack{\forall k' \in [k-\Delta_k; k+\Delta_k] \\ \forall m' \in [m-\Delta_m; m+\Delta_m]}} \right\} \quad (22)$$

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 machine 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)

$$M_h^{(true)}[k,m] = \begin{cases} 1 & \text{if } |F_{s_h}^h[k,m]|^2 > |F_{s_p}^h[k,m]|^2 \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

- Percussive reference separation mask :

$$M_p^{(true)}[k,m] = 1 - M_h^{(true)}[k,m] \tag{24}$$

- 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 :

$$W = \frac{1}{n} \sum_{c=1}^{C} n_c W_c, \tag{25}$$

  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.

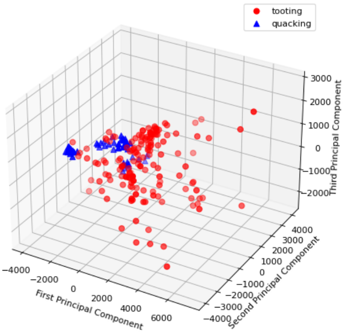
- we define B the inter-class variance-covariance matrix expressed as follows :

$$B = \frac{1}{n} \sum_{c=1}^{K} n_c (\mu_c - \mu)(\mu_c - \mu)^T, \tag{26}$$

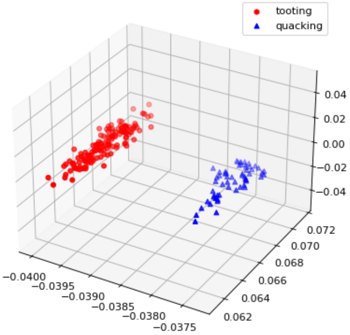
  where μ_c corresponds to the mean vector of class c and μ is the mean vector of the entire dataset.

- 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. (7).
- for each TF point, computes $\mathcal{Q}_x[k, m]$ using Eq. (22).
- computation of linear projections $P_\mathcal{Q}$ using the eigenvectors provided by LDA.
- compute the separation masks :

$$M_h[k,m] = \begin{cases} 1 & \text{if } ||P_\mathcal{Q}[k,m] - \mu_h|| < ||P_\mathcal{Q}[k,m] - \mu_p|| \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

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

- reconstruct each source signal through the inverse STFT :

$$\hat{s}_h = \text{TFCT}^{-1}(F_x^h[k,m]M_h[k,m]) \quad (28)$$

$$\hat{s}_p = \text{TFCT}^{-1}(F_x^h[k,m]M_p[k,m]) \quad (29)$$

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

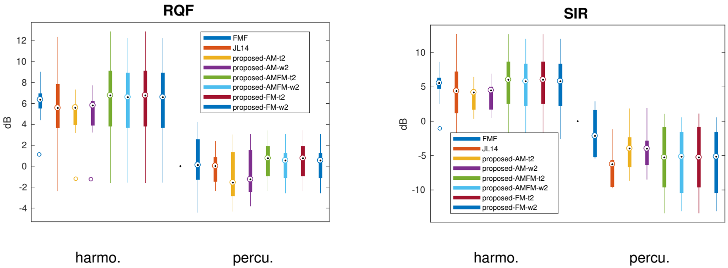
- FMF [Fitzgerald et al. 2014]
- JL14 [Jeong et al. 2014]
- (proposed) AM, FM, AM-FM ($t2$)
- (proposed) FM, AM-FM ($\omega 2$)

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

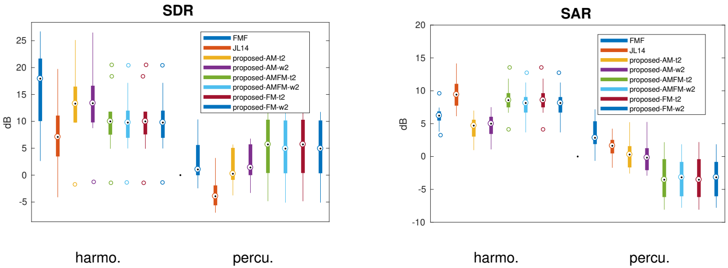


(c) (d)

Figure – Comparative results expressed using BssEval.

Audio results : `https://fourer.fr/publi/gretsi22/`

Results



(a) (b)

Figure – Comparative results expressed using BssEval.

Audio results : `https://fourer.fr/publi/gretsi22/`

Conclusion and future work

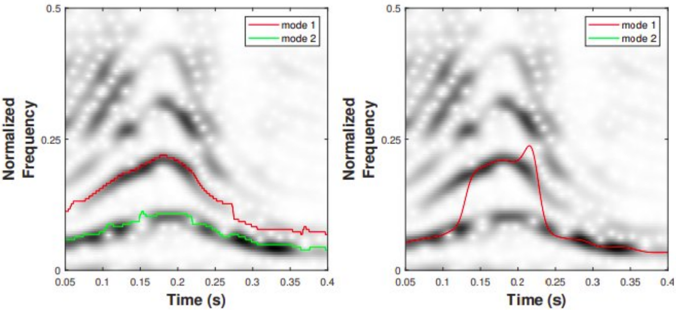
Contributions

- A physics-based model for separating harmonic and percussive components or noise 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)

Limitations

- Not very robust to noise AM/FM estimators (require regularization)
- Binary adaptive separation mask (require phase reconstruction for overlapping components)
- Non-optimal patch size and features for the computing the separation mask
- The timbre features are ignored

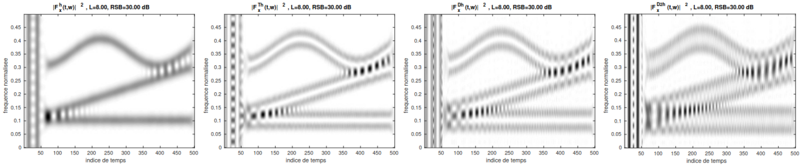
Robust Component Retrieval using Bayesian Approach



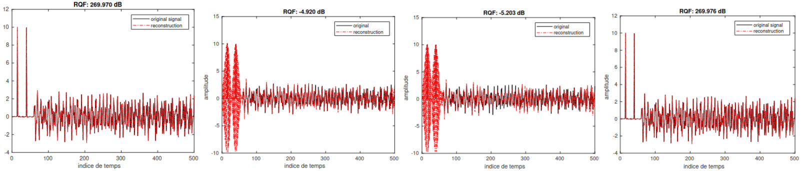
Estimation of the 2 first prominent components in a speech signal comparing (left-hand side) the proposed EM-Laplace method [2] with the Ridge Detector proposed by Laurent and Meignen in 2021 (IEEE TSP).

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

Futher investigation of the STFT properties



(a) $|F_x^h(t,\omega)|^2$ (b) $|F_x^{\mathcal{T}h}(t,\omega)|^2$ (c) $|F_x^{\mathcal{D}h}(t,\omega)|^2$ (d) $|F_x^{\mathcal{D}^2h}(t,\omega)|^2$



(e) $\frac{1}{h(0)^*}\int_{\mathbb{R}} F_x^h(t,\omega)e^{j\omega t}\frac{d\omega}{2\pi}$ (f) $\frac{1}{jh(0)^*}\int_{\mathbb{R}} \omega F_x^{\mathcal{T}h}(t,\omega)e^{j\omega t}\frac{d\omega}{2\pi}$ (g) $\frac{1}{j\mathcal{D}^2h(0)^*}\int_{\mathbb{R}} \omega F_x^{\mathcal{D}h}(t,\omega)e^{j\omega t}\frac{d\omega}{2\pi}$ (h) $\frac{1}{\mathcal{D}^2h(0)^*}\int_{\mathbb{R}} F_x^{\mathcal{D}^2h}(t,\omega)e^{j\omega t}\frac{d\omega}{2\pi}$

[3]

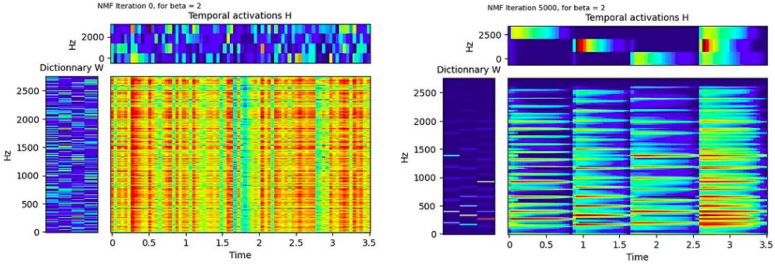
3. D. Fourer, F. Auger, E. Chassande-Mottin and P. Flandrin. Nouvelles formules de synthese de la transformee de Fourier a court terme avec une fenetre d'analyse modifiee . Proc. GRETSI 2025. Strasbourg, France.

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 :

$$V \approx WH \quad \text{s.t,} \;\; W \geq 0, \;\; H \geq 0 \tag{30}$$

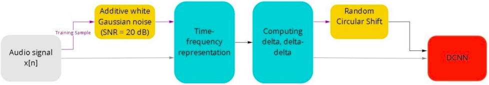
where $W, H = \arg\min_{W,H \geq 0} D(V|WH))$, D being an arbitrary distance or divergence function.

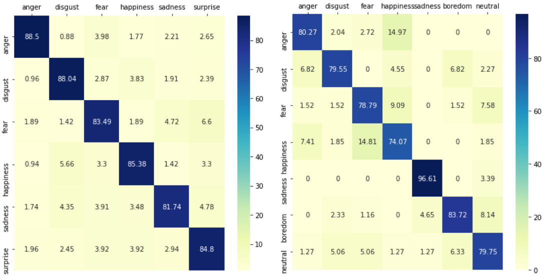


4

4. https://medium.com/@zahrahafida.benslimane/audio-source-separation-using-non-negative-matrix-factorization-nmf-a8b204490c7d

Deep Learning Applied on a time-frequency representation



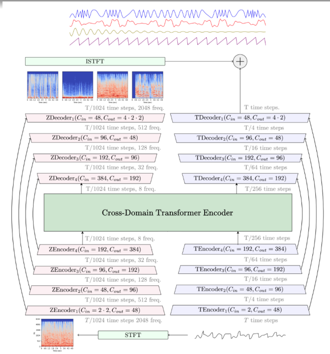


(a) eNTERFACE05, STFT-Alex+RCS41 (Acc. 85.33%) (b) EMO-DB, STFT-Alex+RCS19 (Acc. 81.82%)

[5]

5. 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.

Deep Learning Baseline : HDemucs v4 [Defossez et al, 2021]



HDemucs v4 is a state-of-the-art deep learning model for music source separation based on a **hybrid time-frequency architecture** combining :

- waveform-domain processing for accurate phase reconstruction,
- spectrogram-domain processing for long-term and harmonic structures.

The model estimates each source directly from the mixture using a deep encoder–decoder architecture with large receptive fields.

```
python3 -m pip install -U demucs
```

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

- [*Can*10] 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.
- [*JL*14] I.-Y. Jeong and K. Lee. Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints. 21(10) :1197-1200, 2014.
- [*VGF*06] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. 14(4) :1462-1469, July 2006.
- [*FAG*18] D. Fourer, F. Auger and G. Peeters. Local AM/FM parameters estimation : application to sinusoidal modeling and blind audio source separation. IEEE Signal Processing Letters. Vol. 25. Issue 10. DOI : 10.1109/LSP.2018.2867799. pp.1600-1604. oct. 2018..

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