Harmonic/Percussive Components Separation using Local Linear AM-FM Estimators

Dominique Fourer

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

8 novembre 2023

Plan

The AQUA-RIUS project - ANR-22-CE23-0022

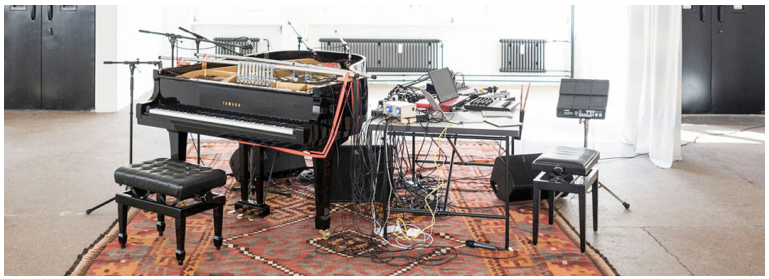
Audio QUality Analysis for Representing, Indexing and Unifying Signals

- from **jan. 2023** to **june 2026**
- 510 157.90 €
- 3 labs : IBISC (Univ. Evry) / UMR STMS (IRCAM) / LTCI (TelecomParis)
- Open positions : 1 postdoc IRCAM, several Master thesis

Goals

- Exhaustive investigation of "audio quality" : analysis/modeling with a focus to prediction/detection of the audio effects applied during the audio signal production and diffusion chain
- Simulation and synthesis of audio quality with possible applications to data augmentation and domain adaptation techniques using machine-learning-based methods
- Audio Quality control with application to audio effects canceling/reverting and signal enhancement/restoration

Audio Quality



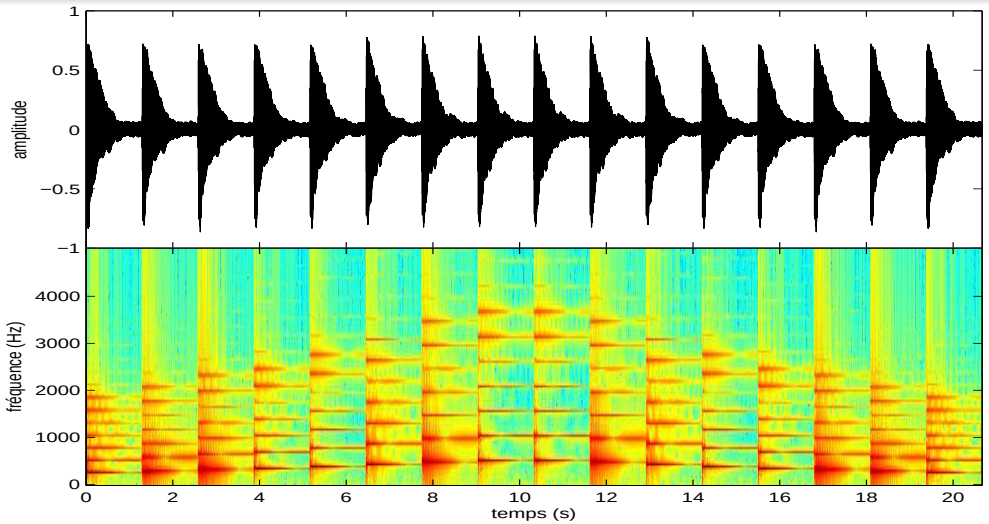
What is audio quality ?

As instrumental sound "timbre" is defined as all sound characteristics not related to pitch, loudness and duration, "audio quality" is everything related to the sound characteristics not related to the content sources. (ie. recording media, lossy compression, mastering, etc.)

Motivation :

- Detecting AQ is full of interest for audio indexing systems (AQ is related to listening experience)
- AQ can have an impact on the prediction accuracy of music information retrieval systems
- AQ results from recording conditions and studio practices
- Controlling AQ can enable robust machine learning methods (ie. data augmentation, models invariant to unwanted signal transformations)

Why Time-Frequency Analysis ?



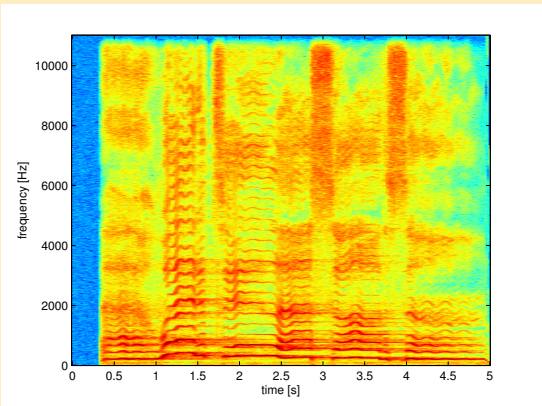
signal and spectrogram of a piano playing the C major scale.

Non-stationary multicomponent signal processing

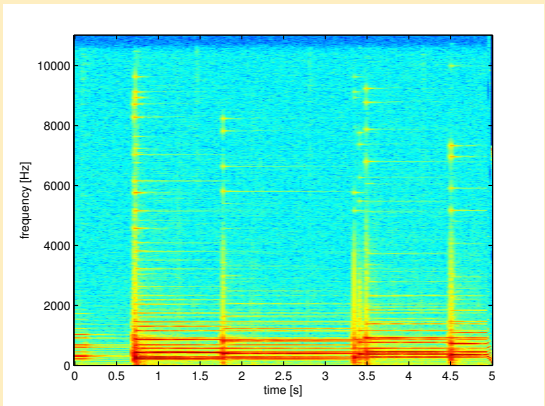
- Disentangle harmonic components (eg. piano, guitar, voice, etc.) from transients (eg. drums, percussion, etc.) to characterize the source and the effects (*i.e.* sinusoids, transients, noise, etc.)
- Computation of sharpen and sparse representations (*i.e.* data modeling, compression)
- Physics meaningful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

Observation model

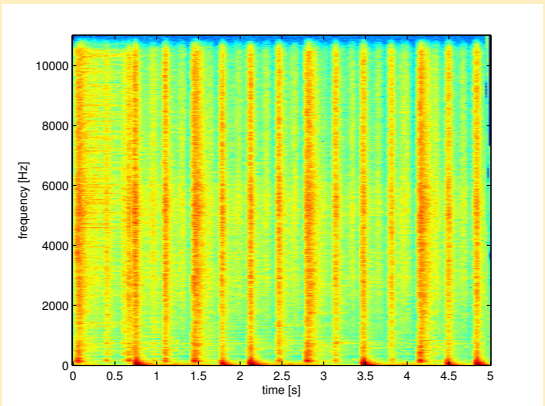
Examples of spectrograms



singing voice



piano (h)



drums (p)

Monophonic Instantaneous Mixture Model

$$x(t) = s_h(t) + s_p(t) \tag{1}$$

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. $F^h_{s_h}(t,\omega)F^h_{s_p}(t,\omega) = 0, \forall (t,\omega) \in \mathbb{R}^2$:

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

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

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

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

Short-Time Fourier Transform (STFT)

Definition

$$F_x^h(t,\omega) = \int_{\mathbb{R}} x(u)h(t-u)^* \, \mathbf{e}^{-j\omega u} \, \mathrm{d}u \tag{4}$$

$$= \mathbf{e}^{-j\omega t} \int_{\mathbb{R}} x(t-u)h(u)^* \, \mathbf{e}^{j\omega u} \mathrm{d}u \tag{5}$$

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

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

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

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

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

STFT properties

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

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] :

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

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

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

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

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

Solution

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

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

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

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

Estimator (ωn)

Similarly we obtain :

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

Solution

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

$$\hat{\Psi}_x^{(\omega n)}(t,\omega) = \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h}F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h}F_x^h} + j\omega \tag{16}$$

Signal parameters estimation

Model

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

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

Estimators

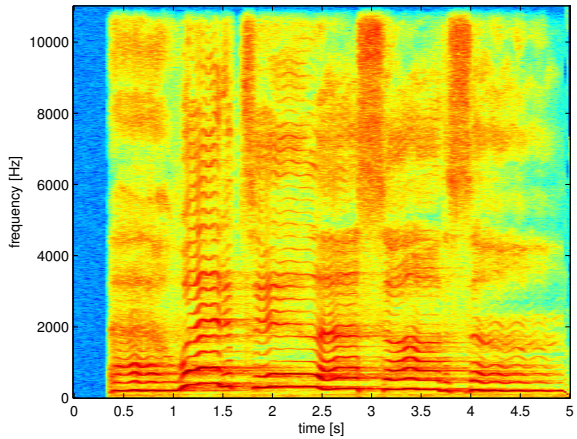
- Log-amplitude linear modulation : $\dot{\lambda}_x(t) = \frac{d\lambda_x}{dt}(t) = \mu_x + \nu_x t$
- Instantaneous frequency : $\dot{\phi}_x(t) = \frac{d\phi_x}{dt}(t) = \omega_x + \alpha_x t$

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

$$\hat{\nu}_x(t,\omega) = \mathrm{Re}\Big(\hat{q}_x(t,\omega)\Big), \qquad \hat{\alpha}_x(t,\omega) = \mathrm{Im}\Big(\hat{q}_x(t,\omega)\Big) \tag{18}$$

$$\hat{\dot{\lambda}}_x(t,\omega) = \mathrm{Re}\left(\hat{\Psi}_x(t,\omega)\right), \quad \hat{\dot{\phi}}_x(t,\omega) = \mathrm{Im}\left(\hat{\Psi}_x(t,\omega)\right) \tag{19}$$

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. (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 :

$$\mathcal{Q}_x[k,m] = \left\{ \frac{G_x[k',m']\left|F_x[k',m']\right|^2}{\sum_{k'}\sum_{m'}\left|F_x[k',m']\right|^2} \;\middle|\; \begin{array}{l} \forall \boldsymbol{k'} \in [\boldsymbol{k}-\Delta_{\boldsymbol{k}};\boldsymbol{k}+\Delta_{\boldsymbol{k}}] \\ \forall \boldsymbol{m'} \in [\boldsymbol{m}-\Delta_{\boldsymbol{m}};\boldsymbol{m}+\Delta_{\boldsymbol{m}}] \end{array} \right\} \tag{20}$$

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)

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

- Percussive reference separation mask :

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

- 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.

ALD 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 :

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

  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 :

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

  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.

Components separation

Algorithm

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

$$M_h[k,m] = \begin{cases} 1 & \text{if } ||P_{\mathcal{Q}}[k,m] - \mu_h|| < ||P_{\mathcal{Q}}[k,m] - \mu_p|| \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

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

- reconstruct each source signal through the inverse STFT :

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

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

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 ($t2$)
- (proposed) FM, AM-FM ($\omega 2$)

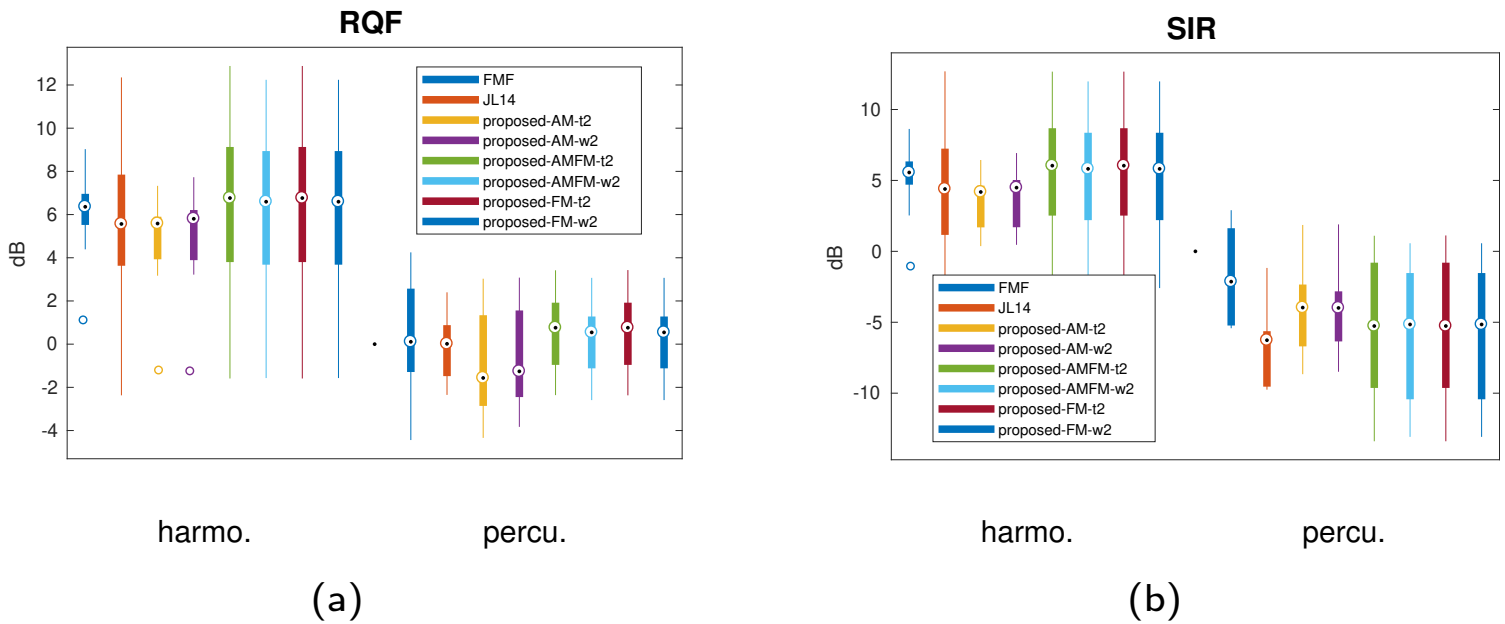
Metrics

- 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



(a) (b)

Figure : Comparative results expressed using BssEval.

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

Results

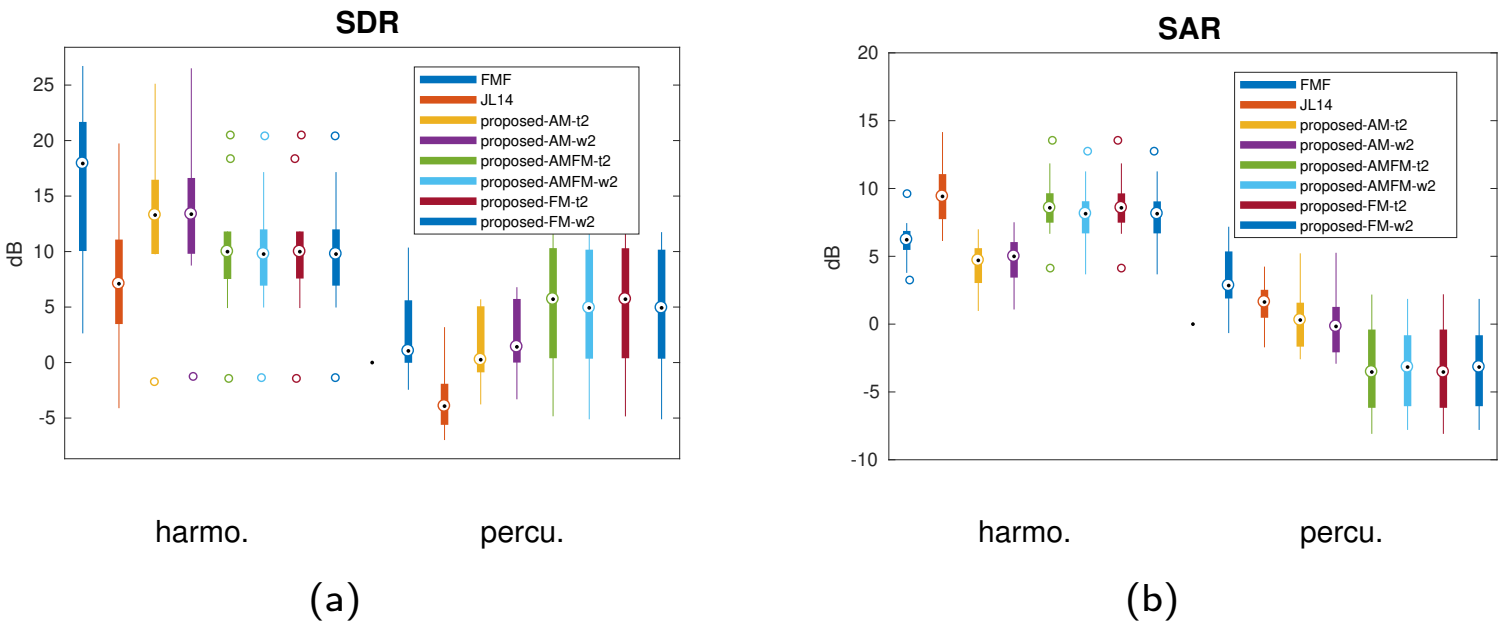


Figure : Comparative results expressed using BssEval.

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

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.

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