Fast and adaptive blind audio source separation using recursive Levenberg-Marquardt synchrosqueezing

Dominique Fourer and Geoffroy Peeters
UMR STMS (IRCAM - CNRS - UPMC), Paris, France

Contributions summary

- We improve the performances of the DUET [1] algorithm using the synchrosqueezed STFT.
- Our recursive implementation [2] allows a real-time implemation.
- The Levenberg-Marquardt algorithm makes our method adaptive using a damping parameter μ.

The synchrosqueezed STFT

For any time t and any angular frequency ω, the STFT of a signal x using a differentiable analysis window h is defined as :

$$X^h(t,\omega) = \int_{\mathbb{R}} x(u)h(t-u)^* \,\mathbf{e}^{-j\omega u}\,\mathrm{d}u \qquad (1)$$

$$= \mathbf{e}^{-j\omega t}\int_{\mathbb{R}} x(t-u)\,\underbrace{h(u)^*\,\mathbf{e}^{j\omega u}}_{g(u,\omega)}\,\mathrm{d}u. \qquad (2)$$

The synchrosqueezed STFT can be defined by [3] :

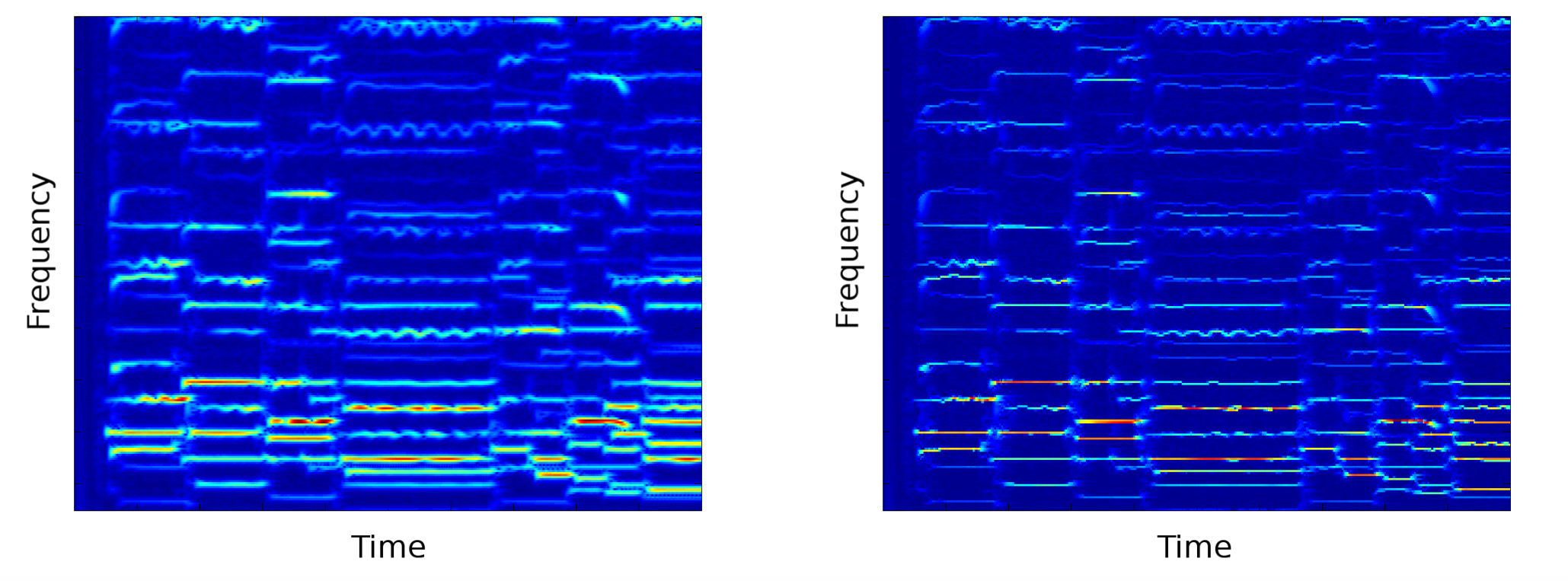
$$\mathcal{S}X^h{}_{(t,\omega)} = \int_{\mathbb{R}} X^h(t,\omega')\,\mathbf{e}^{j\omega'(t-t_0)}\delta(\omega-\hat{\omega}_{x(t,\omega')})\,\mathrm{d}\omega' \qquad (3)$$

which provides a sharpened time-frequency representation (TFR) $|\mathcal{S}X^h(t,\omega)|^2$ when an efficient local instantaneous frequency estimator is used such as [2] :

$$\hat{\omega}_x(t,\omega) = \omega + \mathrm{Im}\left(\frac{X^{Dh}(t,\omega)}{X^h(t,\omega)}\right), \quad \text{with } Dh(t) = \frac{dh}{dt}(t). \qquad (4)$$

The main advantage of synchrosqueezing over the reassignment method [2], is that it admits a signal reconstruction formula :

$$\hat{x}(t-t_0) = \frac{1}{h(t_0)^*}\int_{\mathbb{R}} \mathcal{S}X^h(t,\omega)\frac{\mathrm{d}\omega}{2\pi}. \qquad (5)$$



(a) spectrogram $|X^h(t,\omega)|^2$ (b) synchrosqueezing $|\mathcal{S}X^h(t,\omega)|^2$

The Levenberg-Marquardt synchrosqueezing

Make the synchrosqueezing adjustable and adaptive using a damping parameter μ.
This parameter could be locally matched to the signal content by a voice activity detector or by a noise only/signal+noise binary detector.

A new instantaneous frequency estimator is computed as :

$$\begin{pmatrix}\hat{t}_\mu(t,\omega)\\ \hat{\omega}_\mu(t,\omega)\end{pmatrix} = \begin{pmatrix}t\\ \omega\end{pmatrix} - \left(\nabla^t R_x^h(t,\omega) + \mu \mathrm{I}_2\right)^{-1} R_x^h(t,\omega) \qquad (6)$$

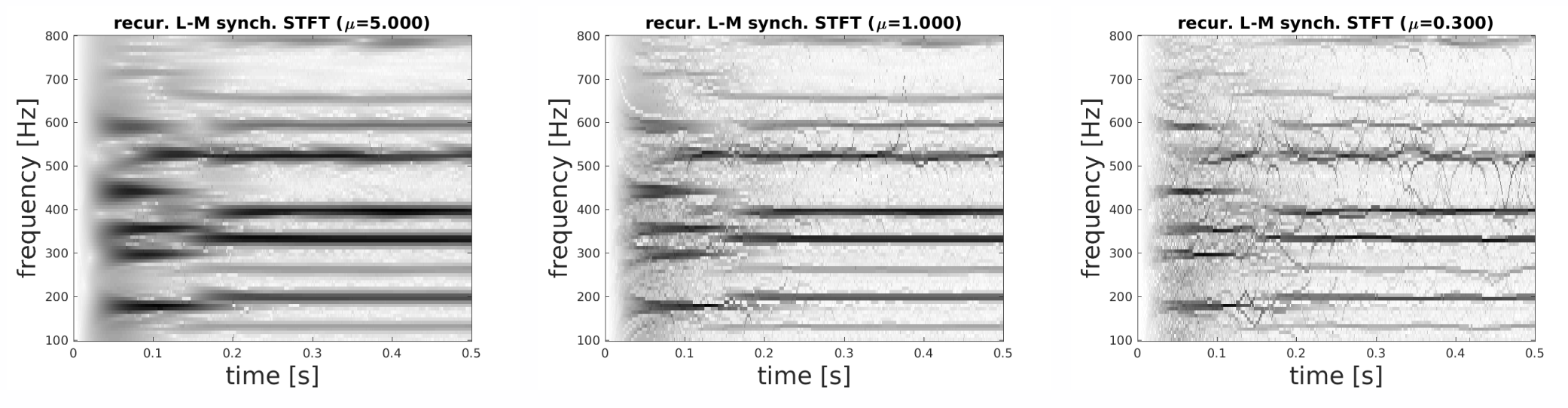
$$\text{with} \quad R_x^h(t,\omega) = \begin{pmatrix} t-\hat{t}_x(t,\omega)\\ \omega-\hat{\omega}_x(t,\omega)\end{pmatrix} \qquad (7)$$

$$\nabla^t R_x^h(t,\omega) = \begin{pmatrix}\frac{\partial R_x^h}{\partial t}(t,\omega) & \frac{\partial R_x^h}{\partial \omega}(t,\omega)\end{pmatrix} \qquad (8)$$

I_2 being the 2×2 identity matrix and $\hat{t}_x$ being the time reassignment operator computed as [2] :

$$\hat{t}_x(t,\omega) = t - \mathrm{Re}\left(\frac{X^{Th}(t,\omega)}{X^h(t,\omega)}\right), \quad \text{with } Th(t) = t\ h(t). \qquad (9)$$

The Levenberg-Marquardt synchrosqueezing transform is computed by replacing $\hat{\omega}_x$ in Eq. (3) by $\hat{\omega}_\mu$.



(c) $\mu = 5$ (d) $\mu = 1$ (e) $\mu = 0.3$

Recursive implementation

Use a specific analysis window $h_k(t) = \frac{t^{k-1}}{T^k(k-1)!}\mathbf{e}^{-t/T}\,U(t)$ ($k \geq 1$ being the filter order, T the time spread of the window, and $U(t)$ the Heaviside step function). This allows to implement the STFT in terms of recursive filtering operations [3].

When filters $g_k(t,\omega) = h_k(t)\,\mathbf{e}^{j\omega t} = \frac{t^{k-1}}{T^k(k-1)!}\mathbf{e}^{pt}\,U(t)$, with $p = j\omega - \frac{1}{T}$, are discretized under the impulse invariance assumption, we obtain :

$$G_k(z,\omega) = T_s\mathcal{Z}\left\{g_k(t,\omega)\right\} = \frac{\sum_{i=0}^{k-1} b_i z^{-i}}{1+\sum_{i=1}^{k} a_i z^{-i}}, \qquad (10)$$

with $b_i = \frac{1}{L^k(k-1)!}B_{k-1,k-i-1}\alpha^i$, $\alpha = \mathbf{e}^{pT_s}$, $L = T/T_s$, $\mathcal{Z}\left\{f(t)\right\} = \sum_{n=0}^{+\infty} f(nT_s)z^{-n}$, $a_i = A_{k,i}\left(-\alpha\right)^i$, T_s being the sampling period, $B_{k,i}$ denotes the Eulerian numbers and $A_{k,i}$ are the binomial coefficients. Hence, $X_k[n,m] \approx X^{h_k}(nT_s, \frac{2\pi m}{MT_s})\,\mathbf{e}^{j\frac{2\pi mn}{M}}$ can be computed from the sampled analyzed signal $x[n]$ by a standard difference equation :

$$X_k[n,m] = \sum_{i=0}^{k-1} b_i\,x[n-i] - \sum_{i=1}^{k} a_i\,X_k[n-i,m] \qquad (11)$$

where $n \in \mathbb{Z}$ and $m = 0, 1, ..., M-1$ are respectively the discrete time and frequency indices. The z transform of the other specific impulse responses related to $\mathcal{D}h_k(t)$ and $\mathcal{T}h_k(t)$, can simply be expressed as functions of $G_k(z,\omega)$ at different orders as detailed in [3]. A matlab implementation of this technique is freely available at [4] :
`https://github.com/dfourer/ASTRES_toolbox`.

The DUET blind source separation method

Mixture model :

$$x_1(t) = \sum_{i=1}^{I} s_i(t)$$

$$x_2(t) = \sum_{i=1}^{I} a_i s_i(t-\tau_i) \qquad (12)$$

Mixing parameters estimation :
When a non-overlapping source is active at coordinates (t,ω), and when $X_1(t,\omega) \neq 0$ (resp. $X_2(t,\omega) \neq 0$), we have :

$$\hat{a}_i(t,\omega) = \left|\frac{X_2^h(t,\omega)}{X_1^h(t,\omega)}\right| \qquad (13)$$

$$\hat{\tau}_i(t,\omega) = -\frac{1}{\omega}\arg\left(\frac{X_2^h(t,\omega)}{X_1^h(t,\omega)}\right), \quad \forall\omega \neq 0. \qquad (14)$$

An histogram $H(a,\tau)$ is thus computed from whole time-frequency plane and the source parameters can be deduced from detected peaks.

Sources estimation :
Each time-frequency coordinate allocated to the histogram $H(a,\tau)$ can be associated to the prominent source using its corresponding mixing parameters $(\hat{a}_k, \hat{\tau}_k)$ such as :

$$J(t,\omega) = \arg\min_k \left(\frac{\left|\hat{a}_k\,\mathbf{e}^{-j\omega\hat{\tau}_k}X_1^h(t,\omega) - X_2^h(t,\omega)\right|}{1+\hat{a}_k^2}\right) \qquad (15)$$

which allows the computation of the binary separation mask M_i of each source computed as :

$$M_i(t,\omega) = \begin{cases} 1 & \text{if} \quad J(t,\omega) = i\\ 0 & \text{otherwise}\end{cases}. \qquad (16)$$

Finally, the TFR of each source is simply recovered by :

$$\hat{S}_i(t,\omega) = M_i\,\frac{X_1^h(t,\omega) + \hat{a}_k\,\mathbf{e}^{j\omega\hat{\tau}_k}X_2^h(t,\omega)}{1+\hat{a}_k^2} \qquad (17)$$

for which the waveform is reconstructed using the corresponding synthesis formula (*i.e.* Eq. (5) when the synchrosqueezed STFT is used).

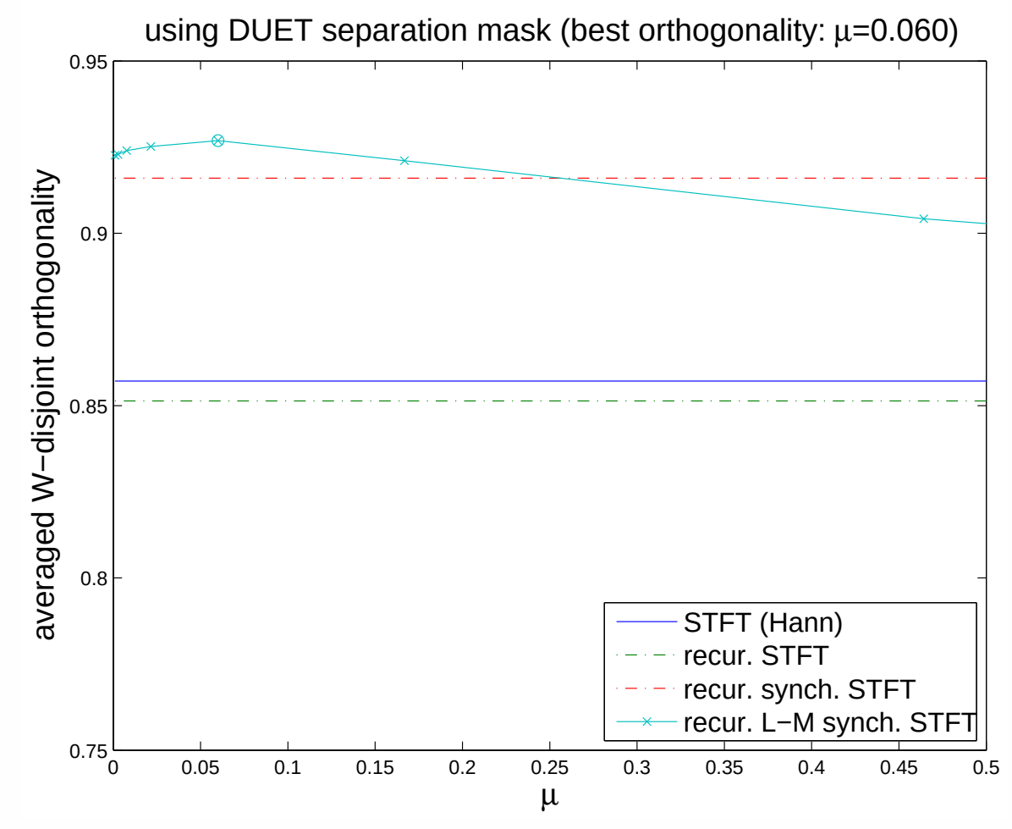
Numerical results

Effect on W-disjoint orthogonality :

Two sources s_1, s_2 are said W-disjoint orthogonal if their STFTs verify [1] : $S_1^h(t,\omega)S_2^h(t,\omega) = 0$.
For a given separation mask M_i, the W-disjoint orthogonality of a source i present in a mixture can be measured by [1] : $D_i(M_i) =$:

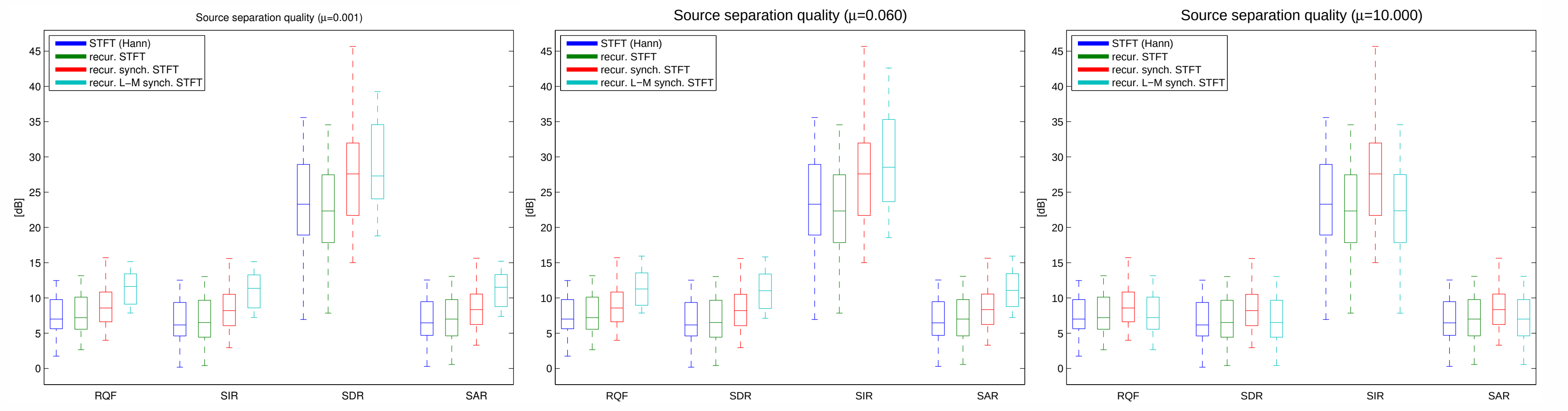
$$\frac{\iint_{\mathbb{R}^2} |M_{i(t,\omega)}S_i^h{}_{(t,\omega)}|^2\,dtd\omega - \iint_{\mathbb{R}^2} |M_{i(t,\omega)}Y_{i(t,\omega)}|^2\,dtd\omega}{\iint_{\mathbb{R}^2} |S_i^h{}_{(t,\omega)}|^2\,dtd\omega} \qquad (18)$$

where $Y_i(t,\omega) = \sum_{\forall j\neq i} S_j^h(t,\omega)$ denotes the sum of all the other sources present in the analyzed mixture.



Approximated W-disjoint orthogonality using different TFRs, as a function of μ. Results are averaged over 10 mixtures of 4 sources from the Bach10 dataset.

Experiment description :
We use the Bach10 research dataset freely available at : `http://music.cs.northwestern.edu/data/Bach10.html` , which contains 10 musical pieces made of 4 sources (pitched instruments). We generate random mixtures through Eq. (12). Comparative results assume that mixing parameters are known and identical for all the methods, in order to focus on the separation capability of each TFR.
Results :



(f) $\mu = 0.001$ (g) OPTIMAL value $\mu = 0.06$ (h) $\mu = 10$

Comparative source separation results measured in terms of Bss Eval [5], provided by the proposed methods and classical STFT applied on the Bach10 dataset. The results are obtained with different values of the damping parameter μ used by the recursive Levenberg-Marquardt synchrosqueezed STFT (other TFRs are not affected by μ).

Conclusions and future works

- We have proposed a new practical application to blind audio source separation, of our recently introduced time-frequency computation methods [3][4].
- Our methods could also be used to improve the results of any blind source separation methods based on time-frequency masking.
- Future works will investigate more complicated configuration such as convolutive mixture.

Bibliography

- [1] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals : Demixing N sources from 2 mixtures," in Proc. IEEE ICASSP, Istanbul, Turkey, June 2000, vol. 5, pp. 2985–2988.
- [2] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method' IEEE Trans. Signal Process., vol.43, no. 5, pp. 1068–1089, May 1995.
- [3] D. Fourer, F. Auger, and P. Flandrin, "Recursive versions of the Levenberg-Marquardt reassigned spectrogram and of the synchrosqueezed STFT," in Proc. IEEE ICASSP, Shanghai, China, May 2016, pp. 4880–4884.
- [4] D. Fourer, J. Harmouche, J. Schmitt, T. Oberlin, S. Meignen, F. Auger, and P. Flandrin, "The ASTRES toolbox for mode extraction of non-stationary multicomponent signals", in Proc. EUSIPCO, Kos island, Greece, Aug. 2017, pp. 1170–1174. `https://github.com/dfourer/ASTRES_toolbox`
- [5] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing (TASLP), vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

Horizon 2020
European Union funding
for Research & Innovation

This work was supported by the European Union's Horizon 2020 research and innovation program
under grant agreement n° 688122, as part of the ABC-DJ project (http://www.abcdj.eu)
and the ANR ASTRES Project (ANR-13-BS03-0002-01).

Artist-to-Business-to-Business-to-Consumer
Audio Branding System