FAST AND ADAPTIVE BLIND AUDIO SOURCE SEPARATION USING RECURSIVE LEVENBERG-MARQUARDT SYNCHROSQUEEZING

Dominique Fourer *Geoffroy Peeters*

UMR STMS (IRCAM - CNRS - UPMC)
dominique@fourer.fr, geoffroy.peeters@ircam.fr

This paper revisits the Degenerate Unmixing Estimation Technique (DUET) for blind audio separation of an arbitrary number of sources given two mixtures through a recursively computed and adaptive time-frequency representation. Recently, synchrosqueezing was introduced as a promising signal disentangling method which allows to compute reversible and sharpen time-frequency representations. Thus, it can be used to reduce overlaps between the sources in the time-frequency plane and to improve the sources' sparsity which is often exploited by source separation techniques. Furthermore, synchrosqueezing can also be extended using the Levenberg-Marquardt algorithm to allow a user to adjust the energy concentration of a time-frequency representation which can be efficiently implemented without the FFT algorithm. Hence, we show that our approach can improve the quality of the source separation process while remaining suitable for real-time applications.

Index Terms— blind source separation, time-frequency analysis, synchrosqueezing, Levenberg-Marquardt algorithm.

1. INTRODUCTION

Source separation is the task which aims at estimating the source components present in a mixture [1]. It could allow a user to freely manipulate each isolated instrument in a polyphonic audio mixture and could find many other practical applications (*e.g.* karaoke, remixing, denoising, etc.). The blind degenerate case (where the number of sources is greater than the number of observed mixtures) remains the most challenging. State-of-the-art methods should use strong assumptions over the sources, such as sparsity [2], harmonicity [3], or disjoint orthogonality [4] which can only be revealed thanks to a proper signal representation. Nowadays, these signal properties are also exploited by promising recent methods based on deep neural networks [5] or Kernel Additive Modeling (KAM) [6] which have shown their ability to capture source-specific features from a time-frequency representation (TFR).

The synchrosqueezing transform [7, 8] was introduced as a sharpening method which contrarily to the reassignment [9], provides reversible TFRs. This method was recently extended to frequency modulated signals [10, 11] and was combined with the Levenberg-Marquardt algorithm [12] to become adaptive thanks to a damping parameter. These methods have shown their interest for disentangling multicomponent signals or to obtain physically interpretable components [13].

Thus, we propose to revisit a well known and efficient algorithm for blind source separation called Degenerate Unmixing Estimation Technique (DUET) [4, 14], which operates in the time-frequency plane using a disjoint orthogonality assumption between the sources. This approach can estimate an arbitrary number of sources from a stereophonic mixtures in a blind configuration and has shown its efficiency on real-world audio signals. Due to its simplicity and its robustness, the DUET algorithm is suitable for an investigation of the TFR role in a source separation algorithm in realistic blind audio source separation scenarios. We also show that our proposed approaches can be implemented in terms of recursive filtering to allow real-time applications without using the Fast Fourier Transform (FFT) algorithm.

The paper is organized as follows. The Levenberg-Marquardt algorithm applied to the recursive synchrosqueezed Short-Time Fourier Transforms (STFT) is introduced in Section 2. A blind source separation method which combines DUET and synchrosqueezing is then proposed in Section 3 and evaluated by numerical experiments in Section 4. Finally, results and future works are discussed in Section 5.

2. THE RECURSIVE LEVENBERG-MARQUARDT SYNCHROSQUEEZING TRANSFORM

Let $X^h(t,\omega)$ denote for any time t and any angular frequency ω, the STFT of a signal $x(t)$ using a differentiable analysis window $h(t)$, defined as:

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

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

z^* being the complex conjugate of z and $j^2 = -1$. Since $|X^h(t,\omega)|^2$ provides a TFR called spectrogram, a signal reconstruction can be provided using [12]:

$$x(t-t_0) = \frac{1}{h(t_0)^*}\int_{\mathbb{R}} X^h(t,\omega)\,\mathrm{e}^{j\omega(t-t_0)}\,\frac{\mathrm{d}\omega}{2\pi} \tag{3}$$

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement n^o 688122.

for any time delay $t_0 \geq 0$ verifying $h(t_0) \neq 0$. In [12, 15], we showed that Eq. (2) can be viewed as a convolution product of x with a filter $g(t,\omega) = h(t)\,\mathrm{e}^{j\omega t}$, which can be efficiently implemented in terms of a recursive filtering process if we use a specific analysis window $h_k(t) = \frac{t^{k-1}}{T^k(k-1)!}\,\mathrm{e}^{-\frac{t}{T}}U(t)$, with $k \geq 1$ the filter order, T the time spread of the window and $U(t)$ the Heaviside step function. Computation details and a Matlab implementation of this approach is freely available as a part of ASTRES toolbox in [13].

2.1. Synchrosqueezed STFT

Synchrosqueezing is a post-processing technique which enhances the time-frequency localization of a transform while allowing signal reconstruction. For the STFT, the synchrosqueezing transform is based on the synthesis formula given by Eq. (3) which leads to the following definition [12]:

$$\mathcal{S}X^{h}{}_{(t,\omega)} = \int_{\mathbb{R}} X^h(t,\omega')\,\mathrm{e}^{j\omega'(t-t_0)}\delta(\omega - \hat{\omega}_{x(t,\omega')})\,\mathrm{d}\omega' \quad (4)$$

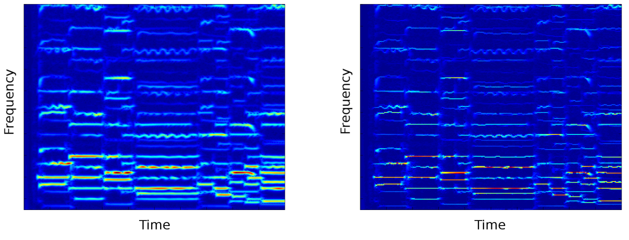
where $\delta(t)$ denotes the Dirac distribution. This transform provides a sharpened TFR computed as $|\mathcal{S}X^h(t,\omega)|^2$ (*cf.* illustration in Fig. 1), when an efficient local Instantaneous Frequency Estimator (IFE) is used for $\hat{\omega}_x$ [11]. Usually, the frequency reassignment operator is used such as [9]:

$$\hat{\omega}_x(t,\omega) = \omega + \mathrm{Im}\left(\frac{X^{Dh}(t,\omega)}{X^h(t,\omega)}\right), \text{ with } Dh(t) = \frac{dh}{dt}(t). \quad (5)$$

An enhanced second-order IFE could also be used to compute the vertical synchrosqueezed STFT as in [11]. Thus, the original signal x can be recovered from the synchrosqueezed STFT using the following reconstruction formula [12]:

$$\hat{x}(t-t_0) = \frac{1}{h(t_0)^*}\int_{\mathbb{R}} \mathcal{S}X^h(t,\omega)\frac{\mathrm{d}\omega}{2\pi} \quad (6)$$

where the integration interval can profitably be restricted to the vicinity of the signal ridge for components extraction.



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

Fig. 1. Comparison of the TFRs provided by the recursively computed spectrogram (a) and the squared modulus of the recursive synchrosqueezed STFT (b). The analysed signal is a mixture made of 4 audio sources from the Bach10 dataset.

2.2. Levenberg-Marquardt synchrosqueezing

Reflection on synchrosqueezing was continued, and by analogy, the Levenberg-Marquardt root finding algorithm has been used to compute new reassignment operators to adjust the energy localization in the time-frequency plane through a damping parameter μ [16]. This parameter could be locally matched to the signal content by a voice activity detector [17] or by a noise only/signal+noise binary detector [17, 18]. The new operators are computed as:

$$\begin{pmatrix}\hat{t}_\mu(t,\omega)\\ \hat{\omega}_\mu(t,\omega)\end{pmatrix} = \begin{pmatrix}t\\ \omega\end{pmatrix} - \left(\nabla^t R_x^h(t,\omega) + \mu \mathrm{I}_2\right)^{-1} R_x^h(t,\omega) \quad (7)$$

$$\text{with } R_x^h(t,\omega) = \begin{pmatrix} t - \hat{t}_x(t,\omega)\\ \omega - \hat{\omega}_x(t,\omega)\end{pmatrix} \quad (8)$$

$$\nabla^t R_x^h(t,\omega) = \begin{pmatrix}\frac{\partial R_x^h}{\partial t}(t,\omega) & \frac{\partial R_x^h}{\partial \omega}(t,\omega)\end{pmatrix} \quad (9)$$

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

$$\hat{t}_x(t,\omega) = t - \mathrm{Re}\left(\frac{X^{Th}(t,\omega)}{X^h(t,\omega)}\right), \text{ with } Th(t) = t\,h(t). \quad (10)$$

Thus, the Levenberg-Marquardt synchrosqueezing transform can be computed by replacing $\hat{\omega}$ in Eq. (4) by $\hat{\omega}_\mu$ This leads to new adaptive and reversible TFRs which can also be efficiently computed through recursive filtering [12, 15].

3. BLIND SOURCE SEPARATION

Now, let's consider a two-channel mixture made of $I \geq 2$ sources s_i. Each active source in the first channel x_1, is attenuated by a factor a_i and delayed by a duration τ_i (expressed in seconds), in the second channel x_2. Thus, the resulting mixture can be modeled as:

$$x_1(t) = \sum_{i=1}^{I} s_i(t)$$
$$x_2(t) = \sum_{i=1}^{I} a_i s_i(t - \tau_i) \quad (11)$$

which can be expressed in the time-frequency domain as:

$$\begin{pmatrix}X_1^h{}_{(t,\omega)}\\ X_2^h{}_{(t,\omega)}\end{pmatrix} = \begin{pmatrix}1 & \dots & 1\\ a_1\,\mathrm{e}^{-j\omega\tau_1} & \dots & a_I\,\mathrm{e}^{-j\omega\tau_I}\end{pmatrix}\begin{pmatrix}S_1^h(t,\omega)\\ \vdots\\ S_I^h(t,\omega)\end{pmatrix}. \quad (12)$$

Hence, our proposal consists in replacing the computed STFTs S_i by the desired TFR (*i.e.* a synchrosqueezed version) in the whole proposed source separation algorithm.

3.1. Mixing parameters estimation

DUET algorithm [4, 14] can recover both the mixing parameters and estimates of the original sources by assuming that only one source is active at each time-frequency coordinate (t, ω). This allows us to write the following expressions when a source i is active at a given (t, ω) coordinate:

$$\begin{pmatrix} X_1^h(t,\omega) \\ X_2^h(t,\omega) \end{pmatrix} = \begin{pmatrix} 1 \\ a_i\, \mathbf{e}^{-j\omega\tau_i} \end{pmatrix} S_i^h(t, \omega). \qquad (13)$$

Thus, the mixing parameters can be estimated when $X_1(t, \omega) \neq 0$ (resp. $X_2(t, \omega) \neq 0$) such as:

$$\hat{a}_i(t, \omega) = \left| \frac{X_2^h(t, \omega)}{X_1^h(t, \omega)} \right| \qquad (14)$$

$$\hat{\tau}_i(t, \omega) = -\frac{1}{\omega} \arg \left(\frac{X_2^h(t, \omega)}{X_1^h(t, \omega)} \right), \quad \forall \omega \neq 0. \qquad (15)$$

For the sake of enhancing the robustness, DUET algorithm computes a smoothed 2D histogram using the symmetric attenuation instead of $\hat{a}$, computed as:

$$\hat{\alpha}_i(t, \omega) = \hat{a}_i(t, \omega) - \frac{1}{\hat{a}_i(t, \omega)} \qquad (16)$$

The signal energy is then distributed according to the estimates $\hat{\alpha}$ and $\hat{\tau}$ over the corresponding axes with a resolution Δ_α and Δ_τ. Hence, the histogram is computed as:

$$H(\alpha, \tau) = \iint_{(t,\omega) \in \mathcal{I}(\alpha,\tau)} |X_1^h(t, \omega) X_2^h(t, \omega)|^2 \, dt d\omega \qquad (17)$$

with:
$\mathcal{I}(\alpha, \tau) = \{(t, \omega) : |\alpha - \hat{\alpha}(t,\omega)| < \Delta_\alpha, |\tau - \hat{\tau}(t,\omega)| < \Delta_\tau\}$

The parameters associated to each source can be deduced from the prominent detected peaks in $H(\alpha, \tau)$ for which the mixing parameter $\hat{a}_i$ can be recovered from $\hat{\alpha}_i$ using:

$$\hat{a}_i = \frac{\hat{\alpha}_i + \sqrt{\hat{\alpha}_i^2 + 4}}{2} \qquad (18)$$

which can be deduced after inverting Eq. (16).

3.2. Sources estimation

Each time-frequency coordinate associated to the histogram given by Eq. (17) can be associated to the prominent source using its corresponding mixing parameters such as:

$$J(t, \omega) = \arg \min_k \left(\frac{\left| \hat{a}_k\, \mathbf{e}^{-j\omega\tau_k} X_1^h(t, \omega) - X_2^h(t, \omega) \right|}{1 + \hat{a}_k^2} \right) \qquad (19)$$

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

$$M_i(t, \omega) = \begin{cases} 1 & \text{if} \quad J(t, \omega) = i \\ 0 & \text{otherwise} \end{cases}. \qquad (20)$$

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

$$\hat{S}_i(t, \omega) = M_i \, \frac{X_1^h(t, \omega) + \hat{a}_k\, \mathbf{e}^{+j\omega\tau_k} X_2^h(t, \omega)}{1 + \hat{a}_k^2} \qquad (21)$$

for which the waveform is reconstructed using the corresponding synthesis formula (*i.e.* Eq. (3) when STFT is used or Eq. (6) for the synchrosqueezed STFT).

4. NUMERICAL EXPERIMENTS

4.1. Audio dataset

Our experiments use the Bach10 dataset[1] which contains 10 musical excerpts of classical music for which the isolated tracks are availables. Each musical piece is made of 4 sources (string instruments) resampled at $F_s = 8000$ Hz and truncated at the 5 first seconds. The simulated stereophonic mixtures are computed using random mixing parameters $a_i \in \{0.4, 0.6, 0.8, 1\}$ and $\tau_i \in [-\frac{2}{F_s}; +\frac{2}{F_s}]$ ensuring that all of the original sources have distinct parameters.

4.2. TFR and W-disjoint orthogonality

Two sources s_1, s_2 are said window-disjoint orthogonal if their windowed Fourier transforms (or STFTs) verify [19]:

$$S_1^h(t, \omega) S_2^h(t, \omega) = 0 \quad \forall (t, \omega). \qquad (22)$$

We now propose to extend this definition to any TFR since the source separation quality of DUET depends on how the sources can overlap in the time-frequency plane. In [14], the author proposes to measure the W-disjoint orthogonality of a source i in a mixture, for a given separation mask M_i using: $D_i(M_i) =$

$$\frac{\iint_{\mathbb{R}^2} |M_i(t,\omega) S_i^h(t,\omega)|^2 \, dt d\omega - \iint_{\mathbb{R}^2} |M_i(t,\omega) Y_i(t,\omega)|^2 \, dt d\omega}{\iint_{\mathbb{R}^2} |S_i^h(t,\omega)|^2 \, dt d\omega} \qquad (23)$$

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.

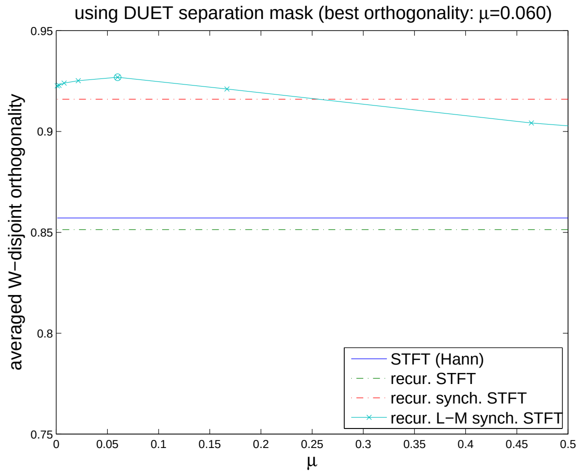
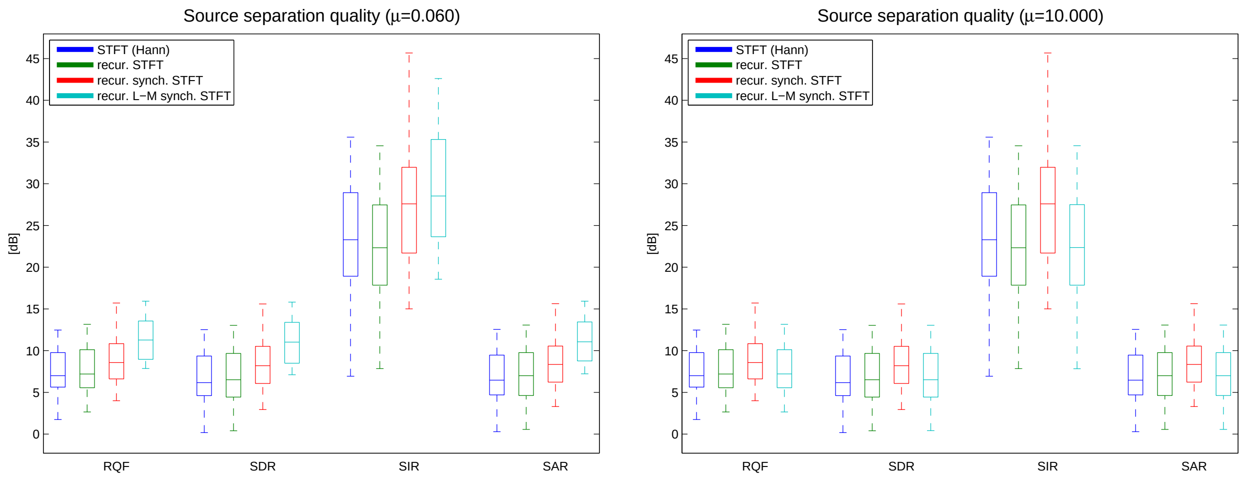


Fig. 3. Approximated W-disjoint orthogonality using different TFRs, as a function of μ. Results are averaged over 10 mixtures made of 4 sources from the Bach10 dataset.

[1] http://music.cs.northwestern.edu/data/Bach10.html



(a) optimal value ($\mu = 0.06$)

(b) arbitrary high value ($\mu = 10$)

Fig. 2. Comparison of the source separation results provided by the proposed methods applied on the Bach10 dataset. (a) and (b) use different values of the damping parameter μ used by the recursive Levenberg-Marquardt synchrosqueezed STFT (other TFRs are not affected by μ).

A high value for D_i corresponds to a better expected source separation quality provided by DUET. Thus, a comparison of the averaged W-disjoint orthogonality measured using Eq. (3) for each TFR computed on the Bach10 dataset, is displayed in Fig. 3. Here, D_i of all the sources is also averaged between all the pieces of the dataset. Each TFR is computed using $M = 1024$ discrete frequency bins. A classical STFT computed using FFT with a Hann window of length 128 ms (1024 samples at $F_s = 8000$ Hz) with an $\frac{1}{2}$-overlap between adjacent frames is used as a baseline method. For the recursively computed TFRs, an Infinite Impulse Response (IIR) filter of order $k = 5$ using a window time spread equal to $L = TF_s = 100$ is used. For the reconstruction a delay equal to $n_0 = (k-1)L$ samples, which corresponds to the maximum of the causal window, is considered (*cf.* [12] for details). Our results show that the synchrosqueezed STFT provides a better W-disjoint orthogonality than the STFT (recursive or classical). When combined with the Levenberg-Marquardt algorithm, a maximal D_i is reached with $\mu = 0.06$. As theoretically investigated in [16, 12], a lower value of μ improves the time-frequency energy localization and converges to the recursive synchrosqueezed STFT results and higher values for μ converge to the recursive STFT results due to a poorer time-frequency localization. Interestingly, the best results are provided by a trade-off with a value of μ which is not too small. The validity of this choice is also confirmed by the source separation results presented in the next section.

4.3. Blind source separation results

Now, we compare the source separation results provided by the proposed methods using the same configuration as in Section 4.2, which are applied to the Bach10 dataset. For each compared method, the mixing parameters are assumed to be known and identical The source separation quality is measured in terms of Reconstruction Quality Factor (RQF) computed as [12]: RQF $= 10\log_{10}\left(\frac{\sum_n |x[n]|^2}{\sum_n |x[n]-\hat{x}[n]|^2}\right)$ and in terms of Signal-to-Interference Ratio (SIR), Signal-to-Distortion Ratio (SDR) and Signal-to-Artifact Ratio (SAR) which are commonly used measures computed through BSS Eval[2] [20]. According to the results displayed in Fig. 2, best separation results (in particular with a higher SIR) are reached using the recursive Levenberg-Marquardt synchrosqueezed STFT using $\mu = 0.06$. This result confirm the expectation provided by the W-disjoint orthogonality in Fig. 3. As also expected, the recursive synchrosqueezed STFT outperforms the STFT. However, the recursive version of the STFT obtains slightly better results than the STFT computed using the FFT, despite a lower averaged value of D_i. For the comparison, Fig. 2(b) displays results for $\mu = 10$ and shows that Levenberg-Marquardt synchrosqueezing obtains results comparable to those provided by STFT when μ is too high.

5. CONCLUSION

In this paper, we have proposed new extensions of the DUET source separation algorithm using a recursive implementation of the Levenberg-Marquardt synchrosqueezed STFT. We have shown that synchrosqueezing allows to compute a sharpen and reversible TFR which can improve the disjoint orthogonality between the sources. This results in a significant improvement of the source separation results in comparison with classical TFRs (ΔSIR≈+5dB in average). Future works will consist in investigating new source separation methods based on TFR masking using the synchrosqueezing technique.

[2] BSS Eval: `http://bass-db.gforge.inria.fr/bss_eval/`

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[2] C. Chenot, J. Bobin, and J. Rapin, "Robust sparse blind source separation," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2172–2176, Nov. 2015.

[3] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 766–778, May 2008.

[4] 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.

[5] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1652–1664, June 2016.

[6] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4298–4310, June 2014.

[7] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous wavelet transform," *Wavelets in Medecine and Bio.*, pp. 527–546, 1996.

[8] F. Auger, P. Flandrin, Y.T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.T. Wu, "TF reassignment and synchrosqueezing: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 32–41, Nov. 2013.

[9] 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.

[10] T. Oberlin, S. Meignen, and V. Perrier, "Second-order synchrosqueezing transform or invertible reassignment? Towards ideal time-frequency representations," *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1335–1344, Mar. 2015.

[11] D. Fourer, F. Auger, K. Czarnecki, Meignen, and Flandrin, "Chirp rate and instantaneous frequency estimation: Application to recursive vertical synchrosqueezing," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1724–1728, Nov. 2017.

[12] 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.

[13] 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.

[14] S. Rickard, *Blind Speech Separation*, chapter The DUET blind source separation algorithm, pp. 217–241, Springer, 2007.

[15] D. Fourer and F. Auger, "Recursive versions of the Levenberg-Marquardt reassigned scalogram and of the synchrosqueezed wavelet transform," in *Proc. IEEE DSP*, London, UK, Aug. 2017, pp. 1–5.

[16] F. Auger, E. Chassande-Mottin, and P. Flandrin, "Making reassignment adjustable: the Levenberg-Marquardt approach," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 3889–3892.

[17] Q.-H. Jo, J.-H. Chang, J.W. Shin, and N.S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, May 2009.

[18] J. Huillery, F. Millioz, and N. Martin, "On the description of spectrogram probabilities with a chi-squared law," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2249–2258, June 2008.

[19] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE ICASSP*, Orlando, FL, USA, May 2002, vol. 1, pp. 529–532.

[20] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.