Local AM/FM parameters estimation: application to sinusoidal modeling and blind audio source separation

Dominique Fourer, François Auger, and Geoffroy Peeters

***Abstract*—This letter extends our recently introduced method which was designed to estimate instantaneous frequency and chirp rate of linearly modulated signals. Indeed, we derive several new estimators related to our previous ones which provide in the time-frequency plane all the signal parameters of the investigated model: amplitude, frequency, and their local modulations (AM/FM). Our estimators are first introduced and compared in terms of statistical efficiency with theoretical bounds and with other state-of-the-art estimators. Then, they are used to improve spectral analysis applied to audio sinusoidal modeling. Finally, they lead to a new source separation technique based on coherent amplitude and frequency modulation that is evaluated on real-world music signals.**

***Index Terms*—time-frequency analysis, sinusoidal modeling, source separation, audio processing.**

I. Introduction

ANALYSIS and transformation of non-stationary signals is an underlying task in audio processing with many applications in Music Information Retrieval (MIR) and source separation [1]. To this end, time-frequency and time-scale analysis [2] provide efficient frameworks for disentangling time-varying multicomponent signals such as audio signals [3]–[5]. The well-known Short-Time Fourier Transform (STFT) [6] is an interesting tool, but is however limited, due to the uncertainty principle and the resulting blurry Time-Frequency (TF) representations [2], [7], [8]. Reassignment [9] provides an efficient solution which improves the readability of a non-reversible Time-Frequency Representation (TFR). Another solution is offered by the synchrosqueezing method [10], [11], a variant of the reassignment technique [9], [12], which admits a signal reconstruction formula. A complementary approach, sinusoidal modeling [13], [14] focuses on local estimation of signal spectral parameters to allow transformations, signal reconstruction and denoising. Several works [15]–[17] have improved the efficiency of parameters estimation to allow applications such as audio synthesis [18], audio coding [19], [20] or blind- [1], [21] and informed-source separation [22]. This paper extends our previous work [23] where we proposed several new Instantaneous Frequency (IF) and Chirp Rate (CR) estimators applied to synchrosqueezing. As we promised in future work perspectives, we now extend this approach for spectral analysis to develop a new sinusoidal modeling framework applied to audio processing and blind source separation. Our contributions are threefold:

- We derive new spectral parameter estimators for all the parameters of a non-stationary signal model (Section II). These estimators generalize our previous results presented in [23].
- We propose an application of our new estimators to spectral analysis, which leads to an enhanced signal sinusoidal modeling method (Section III).
- We propose a new blind source separation method based on our estimators, that is evaluated by numerical simulations on real-world audio signals (Section IV).

II. Local Signal Parameters Estimation

A. Signal Model and Properties

We aim at estimating at every point of a TFR the signal parameters of an amplitude- and frequency-modulated signal. Thus, we consider the following second-order model and we recall its properties [23]:

$$x(t) = \mathbf{e}^{\lambda_x(t)+j\phi_x(t)} \tag{1}$$

$$\text{with} \quad \lambda_x(t) = l_x + \mu_x t + \nu_x \frac{t^2}{2} \tag{2}$$

$$\text{and} \quad \phi_x(t) = \varphi_x + \omega_x t + \alpha_x \frac{t^2}{2} \tag{3}$$

where j is the imaginary unit such that $j^2=-1$. $\lambda_x(t)$ and $\phi_x(t)$ are respectively the log-amplitude and the phase, both depending on the time instant t. This signal satisfies:

$$\frac{dx}{dt}(t) = \left(\frac{d\lambda_x}{dt}(t) + j\frac{d\phi_x}{dt}(t)\right) x(t) = (q_x t + p_x)x(t) \tag{4}$$

with $q_x = \nu_x + j\alpha_x$ and $p_x = \mu_x + j\omega_x$. We define the STFT of this signal using a differentiable analysis window h as:

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

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

with z^* the complex conjugate of z. Differentiating $F_x^h(t,\omega)$ with respect to t leads to:

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

$$= -j\omega F_x^h(t,\omega) + \mathbf{e}^{-j\omega t} \int_{\mathbb{R}} \frac{\mathrm{d}x}{\mathrm{d}t}(t-u)h(u)^* \, \mathbf{e}^{j\omega u} \mathrm{d}u. \tag{8}$$

This work was funded by the European H2020 ABC-DJ (688122) and the early stages of this research were funded until 2017 by the French ANR ASTRES Project (ANR-13-BS03-0002-01).

Replacing $\frac{\mathrm{d}x}{\mathrm{d}t}(t-u)$ by $\left(q_x\,(t-u) + p_x\right)\,x(t-u)$ leads to

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

where $F_x^{\mathcal{D}h}(t,\omega)$ and $F_x^{\mathcal{T}h}(t,\omega)$ are two STFTs using the analysis windows $\mathcal{D}h(t) = \frac{\mathrm{d}h}{\mathrm{d}t}(t)$ and $\mathcal{T}h(t) = t\,h(t)$. A second-order derivative with respect to t leads to:

$$F_x^{\mathcal{D}^2h}(t,\omega) = -q_x F_x^{\mathcal{T}\mathcal{D}h}(t,\omega) + (q_x t + p_x - j\omega) F_x^{\mathcal{D}h}(t,\omega) \quad (10)$$

and more generally for $n \geq 1$ [23]:

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

On the other hand, differentiating Eq. (9) $n-1$ times (for $n \geq 2$) with respect to ω leads to:

$$\begin{aligned} &F_x^{\mathcal{T}^{n-1}\mathcal{D}h}(t,\omega) + (n-1)\,F_x^{\mathcal{T}^{n-2}h}(t,\omega) = \\ &-q_x\,F_x^{\mathcal{T}^n h}(t,\omega) + (q_x t + p_x - j\omega)\,F_x^{\mathcal{T}^{n-1}h}(t,\omega). \end{aligned} \quad (12)$$

B. Overall parameters estimation

In order to recover the signal parameters estimators, we build linear systems of equations thanks to the previously introduced properties. Thus, combining Eqs. (9) and (11) for $n \geq 2$, leads to a linear system where q_x and $\Psi_x = q_x t + p_x$ are unknown ((t,ω) was omitted for the sake of clarity):

$$\begin{pmatrix} F_x^{\mathcal{D}^{n-1}h} & -F_x^{\mathcal{T}\mathcal{D}^{n-1}h} \\ F_x^h & -F_x^{\mathcal{T}h} \end{pmatrix} \begin{pmatrix} \Psi_x \\ q_x \end{pmatrix} = \begin{pmatrix} F_x^{\mathcal{D}^n h} + j\omega F_x^{\mathcal{D}^{n-1}h} \\ F_x^{\mathcal{D}h} + j\omega F_x^h \end{pmatrix}. \quad (13)$$

When (13) is reversible (*i.e.* $|F_x^h(t,\omega)|^2 > 0$), we obtain the following equality:

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

which leads to the estimator called (tn) since it implies n-order derivatives with respect to t:

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

$$\hat{\Psi}_x^{(tn)}(t,\omega) = j\omega + \frac{F_x^{\mathcal{D}h} F_x^{\mathcal{T}\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^n h}}{F_x^h F_x^{\mathcal{T}\mathcal{D}^{n-1}h} - F_x^{\mathcal{T}h} F_x^{\mathcal{D}^{n-1}h}}. \quad (15)$$

Eq. (15) can be reworded as a function of $\hat{q}_x^{(tn)}$ as follows:

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

where $\tilde{t}$ and $\tilde{\omega}$ are the complex reassignment operators, from which the reassignment operators $\hat{t}$ and $\hat{\omega}$ can be deduced as in [9], [11], [24]:

$$\hat{t}(t,\omega) = \mathrm{Re}\left(\tilde{t}(t,\omega)\right), \text{with} \quad \tilde{t}(t,\omega) = t - \frac{F_x^{\mathcal{T}h}(t,\omega)}{F_x^h(t,\omega)} \quad (17)$$

$$\hat{\omega}(t,\omega) = \mathrm{Im}\left(\tilde{\omega}(t,\omega)\right), \text{with} \quad \tilde{\omega}(t,\omega) = j\omega + \frac{F_x^{\mathcal{D}h}(t,\omega)}{F_x^h(t,\omega)}. \quad (18)$$

Thus, the signal definition in Eq. (1) allows to express the instantaneous log-amplitude derivative and frequency as $\dot{\lambda}_x(t) = \frac{d\lambda_x}{dt}(t) = \mu_x + \nu_x t$ and $\dot{\phi}_x(t) = \frac{d\phi_x}{dt}(t) = \omega_x + \alpha_x t$. These parameters can be estimated using $\Psi_x(t) = \dot{\lambda}_x(t) + j\dot{\phi}_x(t) = q_x t + p_x$, which can be estimated through Eq. (16). This expression can be generalized by replacing $\hat{q}_x^{(tn)}$ by any modulation estimator $\hat{q}$ as proposed in [11], [23] as:

$$\hat{\Psi}_x(t,\omega) = \tilde{\omega}(t,\omega) + \hat{q}_x(t,\omega)(t - \tilde{t}(t,\omega)). \quad (19)$$

Finally, we can derive the following estimators for the signal model provided by Eq. (1) as:

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

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

and the log-amplitude and the phase of x at $t=0$, can be estimated by:

$$\hat{l}_x(t,\omega) = \log\left(\left|\frac{F_x^h(t,\omega)}{G_h(t,\omega,\hat{\Psi}_x(t,\omega),\hat{q}_x(t,\omega))}\right|\right) \quad (22)$$

$$\hat{\varphi}_x(t,\omega) = \arg\left(\frac{F_x^h(t,\omega)}{G_h(t,\omega,\hat{\Psi}_x(t,\omega),\hat{q}_x(t,\omega))}\right) \quad (23)$$

$$\text{with:} \quad G_h(t,\omega,\Psi,q) = \int_{\mathbb{R}} h(t-u)^*\,\mathbf{e}^{(\Psi - j\omega)u - q\frac{u^2}{2}}\,du \quad (24)$$

since we have $F_x^h(t,\omega) = \mathbf{e}^{l_x + j\varphi_x}\;G_h(t,\omega,\Psi_x,q_x)$.

New estimators can be deduced from Eqs. (20)-(23) when an arbitrary local modulation estimator $\hat{q}_x$ is used in Eq. (19). For example, n-order derivatives of $F_x^h(t,\omega)$ with respect to ω lead to a new family of estimators involving $\hat{q}_x^{(\omega n)}$, which is obtained from Eqs. (9) and (12) with $n \geq 2$ [23]:

$$\hat{q}_x^{(\omega n)}(t,\omega) = \frac{(F_x^{\mathcal{T}^{n-1}\mathcal{D}h} + (n-1)F_x^{\mathcal{T}^{n-2}h})F_x^h - F_x^{\mathcal{T}^{n-1}h} F_x^{\mathcal{D}h}}{F_x^{\mathcal{T}^{n-1}h} F_x^{\mathcal{T}h} - F_x^{\mathcal{T}^n h} F_x^h}. \quad (25)$$

III. Sinusoidal Modeling

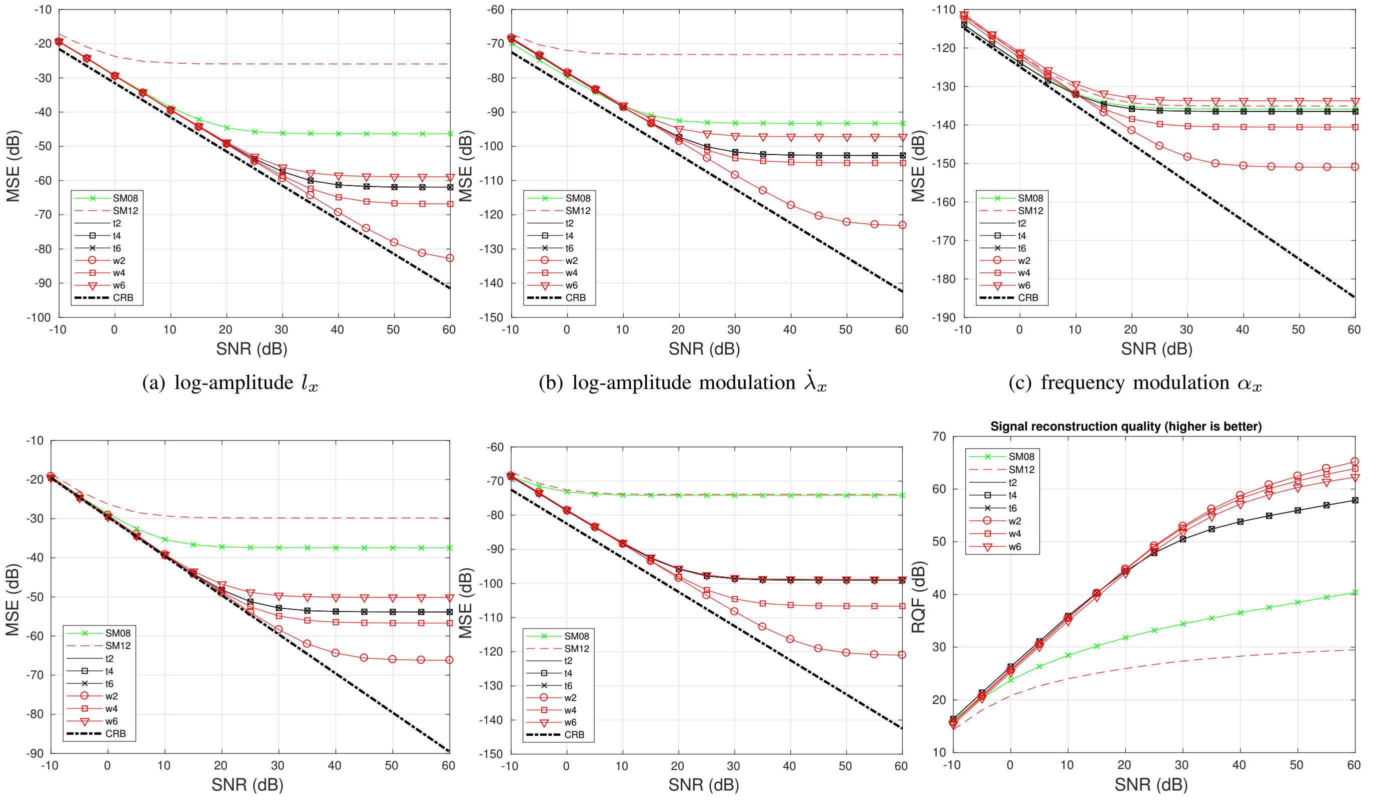
Sinusoidal modeling [13], [14] provides a parametric representation of a signal which allows to apply transformations or signal synthesis [18]. We present here a description of our analysis-synthesis algorithm, before completing a comparative evaluation of the proposed estimators when they are applied to synthetic signals in the presence of noise.

A. Proposed algorithm

We consider here a noisy multicomponent signal expressed as:

$$x(t) = \sum_{i \in \mathcal{I}} x_i(t) + \epsilon(t) = \sum_{i \in \mathcal{I}} \mathbf{e}^{\lambda_i(t) + j\phi_i(t)} + \epsilon(t) \quad (26)$$

$\epsilon(t)$ being an additive noise signal. The discretization process leads to $F_x^h[k,m] = F_x^h(kT_s, \frac{2\pi m}{MT_s})$, with $T_s = \frac{1}{F_s}$ the sampling period, $k \in \mathbb{Z}$ and $m = 0, 1, \cdots, M-1$. We consider an analysis window h of length L with a step $\Delta k = \lfloor (1-\rho)L \rfloor$ (where $\rho \in [0,1[$ corresponds to the overlap ratio between two adjacent analyzed signal frames). The discrete-time versions of the analysis windows involving signal derivatives (*i.e.* $\mathcal{D}^n h$, $\mathcal{T}\mathcal{D}^n h$) are directly computed from their continuous-time expressions. Our analysis-synthesis algorithm assumes that the noise can be neglected when the signal is detected (located at a local maximum) and that there is no more than one sinusoidal component active at each time-frequency point.



(a) log-amplitude l_x

(b) log-amplitude modulation $\dot{\lambda}_x$

(c) frequency modulation α_x

(d) phase φ_x

(e) frequency $\dot{\phi}_x$

(f) RQF of model using all parameters

Fig. 1. MSE (a)-(e) and RQFs comparison between (SM08), (SM12), (tn) and (ωn) at $n \in \{2,4,6\}$, for estimating signal parameters of a AM/FM-modulated sinusoid merged with an additive white Gaussian noise.

a) Analysis:

1) For each frame centered at time instant k, local maxima $m \in]0, M/2[$ are detected (*i.e.* m satisfies: $|F_x^h[k,m]| > |F_x^h[k,m+1]|$ and $|F_x^h[k,m]| > |F_x^h[k,m-1]|$).
2) For each local maximum m, we estimate the vector $P_m[k] = (l_x[m], \varphi_x[m], \dot{\lambda}_x[m], \nu_x[m], \dot{\phi}_x[m], \alpha_x[m])^T$.
3) In descending order of $l_x[m]$, each component associated to m is reconstructed from P_m using Eq. (1) considering that $t = 0$ is located at the center of the current frame. If the residual energy increases when a component is subtracted from the analyzed signal, it is ignored. Otherwise, P_m is kept and the residual signal is considered for processing other detected components.
4) We increase time index by $k \leftarrow k + \Delta k$ and we iterate from step 1, while kT_s is lower than the length of the entire signal.

b) Synthesis: We synthesize using the overlap-add method [25] each frame of signal $\hat{x}$ centered at time instant k moving by step Δk, using the estimated $P_m[k]$ and Eq. (26).

B. Application to synthesized signals

To further assess the efficiency of our proposed estimators, we compare them with (SM08) [15] and (SM12) [16], and with the Cramér-Rao Bounds (CRB) which were derived by Zhou *et al.* in [26]. We consider an about 23 ms-long signal sampled at $F_s = 44.1$ kHz (1023 samples), which contains one sinusoid synthesized from Eq. (1) using uniformly distributed parameters except for the amplitude which is constant. The log-amplitude is fixed at $l_x = 0.18$ and other parameters are chosen as $\varphi_x \in [-\pi, +\pi]$, $\mu_x \in [-100, 100]$, $\omega_x \in [0, 2\pi\frac{F_s}{2}]$ rad.s^{-1} and $\alpha_x \in [-10^4, 10^4]$ rad.s^{-2} ensuring that $0 \leq \omega_x + \alpha_x t \leq 2\pi\frac{F_s}{2}$. This signal is merged with an additional white Gaussian noise with Signal-to-Noise Ratio (SNR) values going from -10 dB to $+60$ dB. In Fig. 1 (a)-(e), we compute the Mean Squared Error (MSE) expressed in dB for each estimated parameter, except for ν_x. In Fig. 1 (f), we compute the Reconstruction Quality Factor (RQF) given by [12]: $\text{RQF}(x, \hat{x}) = 10\log_{10}\left(\frac{||x||^2}{||x-\hat{x}||^2}\right)$ measured between the reference signal x and the synthesized one $\hat{x}$ using all the estimated parameters. Thus, for each SNR value, 10,000 random signals are analyzed using a Hann window of length $L = 1023$. According to Fig. 1, our results show that estimators with higher orders ($n \geq 2$) have a negligible effect on the accuracy estimation for (tn). For (ωn), higher orders obtain poorer results than $(w2)$ which provides the overall best results. (tn) estimators also obtain good performances and significantly outperform the state-of-the art methods (SM08) and (SM12), as evidenced at high SNR values.

C. Application to real-world signals

Table I shows the RQFs obtained on real-world audio signals using the proposed estimators $(t2)$, $(\omega 2)$, $(\omega 6)$, and two state-of-the-art methods respectively called (SM08) [15] and (SM12) [16]. Each analyzed signal has a duration of about

TABLE I
RQFs EXPRESSED IN dB OBTAINED BY THE PROPOSED ALGORITHM APPLIED ON REAL-WORLD AUDIO SIGNALS.

	SM08	SM12	t2	$\omega 2$	$\omega 6$
speech	8.08	7.52	7.82	**8.21**	6.89
singing voice	14.40	14.02	15.05	**15.13**	13.62
saxophone	28.56	27.89	27.15	**29.90**	23.71
drums	6.54	6.52	**6.72**	6.63	4.29

5 seconds and is sampled at $F_s = 22.05$ kHz. Analysis uses Hann windows with a length of about 46 ms except for the drums signal (containing more transients) which is analyzed with a window of 23 ms. Our results show that the $(\omega 2)$ and $(t2)$ (only for the drums) methods obtain the best results when they are compared to the state-of-the-art methods. The audio samples used for this experiment can be found in [27].

IV. Application to source separation

Now, we consider the blind source separation problem [1] in the single-channel case, where the observed mixture contains $C \geq 2$ sources. Thus, we aim at recovering the sources s_c using the observed mixture expressed as:

$$x(t) = \sum_{c=1}^{C} s_c(t) = \sum_{c=1}^{C} \left(\sum_{i_c \in \mathcal{I}_c} \mathrm{e}^{\lambda_{i_c}(t) + j\phi_{i_c}(t)} \right). \qquad (27)$$

We propose to solve this problem under the assumption that each sinusoidal component i is assigned to only one source c, characterized by the set $\mathcal{I}_c$. Hence, blind source separation consists here in a clustering problem which should be solved using the component parameters directly estimated from $x(t)$.

A. Proposed method

Computational Auditory Scene Analysis (CASA) [28] suggests that a set of components whose parameters evolve in a coherent way tend to be perceived as one source. Thus, we propose to group the components of each source through the Coherent Frequency Modulation (CFM) [29] and the new proposed Coherent Amplitude Modulation (CAM) descriptors which can be computed for a signal x as:

$$\mathrm{CFM}_x(t,\omega) = \frac{\hat{\alpha}_x(t,\omega)}{\hat{\dot{\phi}}_x(t,\omega)}, \quad \mathrm{CAM}_x(t,\omega) = \frac{\hat{\dot{\lambda}}_x(t,\omega)}{\hat{l}_x(t,\omega)}. \qquad (28)$$

These descriptors measure the linear modulation factor in frequency and in amplitude. They are assumed to be almost identical at each instant for the components of the same source [28]. This idea has already been investigated in several state-of-the-art methods such as [29], [30]. Our proposed blind source separation algorithm can be formulated as follows:

1) Computation of parameters $P_i[k]$ from the mixture x as $P_i[k] = (l_i, \varphi_i, \nu_i, \dot{\lambda}_i, \dot{\phi}_i, \alpha_i)^T$ associated to the component i (detected by a local maximum of $|F_x^h[k,m]|$), estimated at time instant $t = kT_s$ (*cf.* Section III-A).
2) Computation of $\mathrm{CFM}_i[k]$ and $\mathrm{CAM}_i[k]$ for each component using Eq. (28).
3) At each time instant k, we compute the sets $\mathcal{I}_{c'}$ associated to a sound source, by applying the *k-means* algorithm [31] on the components i, represented by the couple $(\mathrm{CFM}_i, \mathrm{CAM}_i)$, for a maximal number of clusters equal to C.
4) Modeling of each source c at each time instant by a representing vector computed as $v_c[k] = \left(\frac{\sum_{i \in \mathcal{I}_c} l_i^2 \ \mathrm{CFM}_i[k]}{\sum_{i \in \mathcal{I}_c} l_i^2}, \frac{\sum_{i \in \mathcal{I}_c} l_i^2 \ \dot{\phi}_i}{\sum_{i \in \mathcal{I}_c} l_i^2} \right)^T$.
5) If $k > 1$, we affect each cluster c' to the source c through $\arg\min_{c'} ||v_{c'}[k] - v_c[k-1]||$.
6) We synthesize each estimated source $\hat{s}_c$ from parameters $P_i[k]$ for $i \in \mathcal{I}_c$ using Eq. (26).

B. Numerical experiments on real-world musical signals

We analyze an excerpt of 3 seconds of a musical mixture sampled at $F_s = 44.1$ kHz, made of 2 sources (voice/guitar) from MedleyDb [32]. Estimator $(\omega 2)$ is compared to (SM12) using a 23 ms-long Hann window with an overlap between adjacent frames equal to $\rho = \frac{11}{12}$ (this configuration empirically provides the best RQF for the sinusoidal modeling for both methods). Table II shows the source separation scores: RQF [33], Signal-to-Interference Ratio (SIR), Signal-to-Distortion Ratio (SDR) and Signal-to-Artifact Ratio (SAR) [34] for each approach. The Oracle method provides the optimal clustering results obtained by matching each component to the closest one estimated from the reference source signals assumed known. In the blind case, our results show a clear advantage of method $(\omega 2)$ over (SM12), particularly when using the solely CFM descriptor, which provides the best balanced results. The novel descriptor CAM can also be of interest since it can lead to the best source isolation for source 2 (best SIR), but unfortunately with poorer RQF and SDR results. Examples on more audio excerpts from the MedleyDb dataset can be found online at [27].

TABLE II
COMPARISON OF THE VOICE/GUITAR SEPARATION RESULTS FOR THE MUSIC PIECE ALEXANDERROSS VELVETCURTAIN [32].

(a) $(\omega 2)$

Method	RQF (dB)	SIR (dB)	SDR (dB)	SAR (dB)
Oracle	8.70/9.66	19.02/23.21	8.08/9.24	8.50/9.44
CFM + k-means	**5.88/6.02**	9.77/11.14	**4.58/4.84**	**6.58**/6.32
CAM + k-means	2.47/2.62	2.57/11.70	0.23/1.49	0.84/**9.96**
CFM/CAM + k-means	2.99/3.16	3.65/**12.16**	0.63/1.93	1.20/8.36

(b) (SM12)

Method	RQF (dB)	SIR (dB)	SDR (dB)	SAR (dB)
Oracle	9.05/9.60	19.14/23.58	8.50/9.18	8.94/9.36
CFM + k-means	5.43/5.38	8.57/8.63	3.97/3.92	6.38/6.27
CAM + k-means	2.48/2.57	11.19/2.53	-0.14/1.33	0.50/9.41
CFM/CAM + k-means	2.38/2.47	**11.53**/2.46	-0.45/1.27	0.12/9.44

V. Conclusion and future works

We proposed several new estimators which were applied to audio sinusoidal modeling and to blind source separation. Our new proposed estimators have a significantly better accuracy than other state-of-the-art methods when they are used for spectral analysis in both simulations and real-world application scenarios. Our future works will further investigate the source separation method for a better understanding of how the proposed local modulation estimator can be optimally exploited. Moreover, the poorer results in comparaison to $(\omega 2)$ provided by higher-order estimators of (ωn) with $n > 2$ should be investigated from a theoretical point of view.

References

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

[2] P. Flandrin, *Time-Frequency/Time-Scale analysis*. Academic Press, 1998.

[3] 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 2017*, Kos Island, Greece, Aug. 2017, pp. 1170–1174.

[4] K. Abratkiewicz, K. Czarnecki, D. Fourer, and F. Auger, "Estimation of time-frequency complex phase-based speech attributes using narrow band filter banks," in *Proc. IEEE Signal Processing Symposium (SPS'17)*, Warsaw Poland, Sep. 2017, pp. 251–263.

[5] K. Czarnecki, D. Fourer, F. Auger, and M. Rojewski, "A fast time-frequency multi-window analysis using a tuning directional kernel," *Elsevier Signal Processing*, vol. 147, pp. 110–119, Jun. 2018.

[6] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, Jun. 1977.

[7] L. Cohen, *Time-Frequency Analysis: Theory and Applications*. Prentice Hall, 1995.

[8] F. Hlawatsch and F. Auger, Eds., *Time-Frequency Analysis: Concepts and Methods*. ISTE-Wiley, 2008.

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

[11] R. Behera, S. Meignen, and T. Oberlin, "Theoretical analysis of the second-order synchrosqueezing transform," *Applied and Computational Harmonic Analysis*, Nov. 2016.

[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, Mar. 2016, pp. 4880–4884.

[13] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[14] J. Smith and X. Serra, "PARSHL an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. ICMC*, Urbana, Illinois, USA, Aug. 1987, pp. 290–297.

[15] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *Proc. DAFx'08*, Espoo, Finland, Sep. 2008, pp. 281–288.

[16] S. Marchand, "The simplest analysis method for non-stationary sinusoidal modeling," in *Proc. DAFx'12*, York, UK, Sep. 2012, pp. 23–26.

[17] B. Hamilton and P. Depalle, "A unified view of non-stationary sinusoidal parameter estimation methods using signal derivatives," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 369–372.

[18] U. Zölzer, *DAFX: digital audio effects*. John Wiley & Sons, 2011.

[19] H. Purnhagen and N. Meine, "HILN – the MPEG-4 parametric audio coding tools," in *Proc. IEEE ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 201–204.

[20] E. Schuijers, W. Oomen, B. Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th Conv. Audio Eng. Soc. (AES)*, Mar. 2003, pp. 201–204.

[21] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. IEEE ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 765–768.

[22] D. Fourer and S. Marchand, "Informed spectral analysis: audio signal parameters estimation using side information," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 178, Dec. 2013.

[23] D. Fourer, F. Auger, K. Czarnecki, S. Meignen, and P. 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.

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

[25] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 389–406, Sep. 1997.

[26] G. Zhou, G. Giannakis, and A. Swami, "On polynomial phase signal with time-varying amplitudes," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 848–860, Apr. 1996.

[27] D. Fourer, G. Peeters, and F. Auger, "Local AM/FM modulations estimation: application to audio sinusoidal modeling and blind source separation," http://www.fourer.fr/publi/spl18, accessed: October 9, 2018.

[28] A. S. Bregman, *Auditory scene analysis*. MIT Press: Cambridge, MA, 1990.

[29] E. Creager, N. D. Stein, R. Badeau, and P. Depalle, "Nonnegative tensor factorization with frequency modulation cues for blind audio source separation," in *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, New York, USA, Aug. 2016.

[30] F. R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *Proc. IEEE ICASSP*, Shanghai, China, Mar. 2016, pp. 126–130.

[31] G. A. F. Seber, *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[32] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, Taipei, Taiwan, Oct. 2014, pp. 155–160.

[33] D. Fourer and G. Peeters, "Fast and adaptive blind audio source separation using recursive Levenberg-Marquardt synchrosqueezing," in *Proc. IEEE ICASSP*, Calgary, Canada, Apr. 2018.

[34] 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.