

Model-based STFT phase recovery for audio source separation

Paul Magron, Roland Badeau, Bertrand David

▶ **To cite this version:**

Paul Magron, Roland Badeau, Bertrand David. Model-based STFT phase recovery for audio source separation. IEEE Transactions on Audio, Speech and Language Processing, 2018, 26. hal-01718718v2

HAL Id: hal-01718718
https://hal.science/hal-01718718v2

Submitted on 30 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Model-based STFT Phase Recovery for Audio Source Separation

Paul Magron, Roland Badeau, *Senior Member, IEEE,* and Bertrand David, *Member, IEEE*

***Abstract*—For audio source separation applications, it is common to estimate the magnitude of the short-time Fourier transform (STFT) of each source. In order to further synthesizing time-domain signals, it is necessary to recover the phase of the corresponding complex-valued STFT. Most authors in this field choose a Wiener-like filtering approach which boils down to using the phase of the original mixture. In this paper, a different standpoint is adopted. Many music events are partially composed of slowly varying sinusoids and the STFT phase increment over time of those frequency components takes a specific form. This allows phase recovery by an unwrapping technique once a short-term frequency estimate has been obtained. Herein, a novel iterative source separation procedure is proposed which builds upon these results. It consists in minimizing the mixing error by means of the auxiliary function method. This procedure is initialized by exploiting the unwrapping technique in order to generate estimates that benefit from a temporal continuity property. Experiments conducted on realistic music pieces show that, given accurate magnitude estimates, this procedure outperforms the state-of-the-art consistent Wiener filter.**

***Index Terms*—Phase recovery, sinusoidal modeling, phase unwrapping, auxiliary function method, audio source separation.**

I. Introduction

AUDIO source separation [1] consists in extracting the underlying *sources* that add up to form an observable audio *mixture*. To address this issue, it is common to act on a time-frequency (TF) representation of the data, such as the short-term Fourier transform (STFT), since it provides a meaningful representation of audio signals.

Much research in audio has focused on the processing of nonnegative-valued TF representations, such as the magnitude of the STFT. These representations are usually structured by means of a model, such as nonnegative matrix factorization (NMF) [2], [3], kernel additive models [4] or deep neural networks [5], [6]. However, phase recovery has recently become a growing topic of interest [7], [8]. Indeed, obtaining the phase of the corresponding complex-valued STFT is necessary to resynthesize time signals. In the source separation framework, a common practice consists in applying a Wiener-like filtering [3], [9] to the original mixture: the phase of the mixture is then given to each extracted component. This technique builds upon the observation that the phase may appear as uniformly-distributed [10], which leads to modeling the complex-valued STFT coefficients as circularly-symmetric random variables (e.g., Gaussian [3] or stable [9]). In such a framework, this method yields a set of estimates that is optimal in a minimum mean square error (MMSE) sense. However, even if this filter leads to quite satisfactory results in practice [2], [3], it has been pointed out [11] that when sources overlap in the TF domain, it is responsible for residual interference and artifacts in the separated signals.

Improved phase recovery can be achieved with *consistency*-based approaches [12]. Consistency is an important property of the STFT since it directly originates from its redundancy property. Indeed, the STFT is usually computed with overlapping analysis windows, which introduces dependencies between adjacent TF bins. Consequently, not all complex-valued matrices are the STFT of an actual time-domain signal. The authors in [12] proposed an objective function called *inconsistency* that measures this mismatch, and minimizing this criterion results in computing a complex-valued matrix that is as close as possible to the STFT of a time signal. From the baseline Griffin-Lim (GL) algorithm [13], several developments have been proposed to design faster procedures [14]–[17]. For source separation applications, Wiener filtering and consistency-based approaches have been combined in a unified framework [18]–[21], among which consistent Wiener filtering [21] has proved to be the most promising candidate.

Another approach to reconstruct the phase from a spectrogram is to use a phase model based on the observation of fundamental signals that are mixtures of sinusoids [22]. This family of techniques exploits the natural relationship between adjacent TF bins that originates from signal modeling. It leads to a procedure called the *phase unwrapping* (PU) algorithm, which unwraps the phases over time frames, therefore ensuring the temporal coherence of the signal. The main difference between the PU algorithm and consistency-based approaches is that the former relies on a signal model while the latter exploit a property of the STFT. Such an approach has been used in the phase vocoder algorithm [23] for time stretching, and applied to speech enhancement [24], [25], audio restoration [26] and source separation [27]. In many cases, the mixtures are assumed to be in harmonic proportions, which means that the partial frequencies are integer multiples of a fundamental frequency, but the PU technique can be extended to signals which do not comply with this assumption [26].

In this paper, we introduce a novel source separation procedure which exploits the PU algorithm. Since we focus on the phase recovery issue, the magnitudes are assumed known or estimated beforehand. We address this problem by considering a cost function which measures the mixing error between the

P. Magron is with the Laboratory of Signal Processing, Tampere University of Technology, Finland (e-mail: firstname.lastname@tut.fi), but this work was conducted while he was a PhD student at Télécom ParisTech, Paris, France.

R. Badeau and B. David are with LTCI, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France (e-mail: firstname.lastname@telecom-paristech.fr).

observed and estimated complex mixtures. This function is minimized by means of the auxiliary function method, leading to an iterative procedure. The key idea is to give to the initial estimates the phase obtained with the PU technique. Indeed, those initial estimates are expected to be close to a local minimum. Besides, by doing so, the resulting estimates benefit from a temporal continuity property. Experiments conducted on realistic music songs show its potential for an audio source separation task. In particular, it performs similarly to, or better than, the state-of-the-art consistent Wiener filter, with improved interference rejection and a lower computational cost.

This paper is organized as follows. Section II presents the PU algorithm that is obtained from a sinusoidal model. Section III introduces an audio source separation framework which uses this technique. Section IV experimentally validates the potential of the PU algorithm for audio source separation. Finally, section V draws some concluding remarks.

II. The phase unwrapping algorithm

A. Sinusoidal modeling

Let us consider a mixture of P time-varying sinusoids:

$$\forall n \in \mathbb{Z},\ x(n) = \sum_{p=1}^{P} A_p(n) e^{2i\pi\nu_p(n)n + i\varphi_p}, \tag{1}$$

where $A_p(n) > 0$ are the amplitudes, $\nu_p(n) \in [0, \frac{1}{2}]$ are the normalized frequencies and $\varphi_p \in]-\pi, \pi]$ are the initial phases. Note that we use here complex sinusoids while audio signals are real-valued. However, this is a widely-used model [22], [23], [28] and it yields results similar to the real-valued model since we do not account for negative frequency components.

Let w be an analysis window of length N_w, which means that $w(n) = 0\ \forall n \notin \{0, ..., N_w - 1\}$. We assume that the mixture is locally stationary, i.e., that the sinusoidal parameters are constant within time segments of length N_w. In order words, for a given time frame t, we can note $\nu_p(n + St) = \nu_p(t)$ and $A_p(n + St) = A_p(t)$, where S is the hop size of the STFT. Therefore, the STFT of this mixture in a frequency channel f and time frame t is:

$$X(f,t) = \sum_{p=1}^{P} A_p(t) e^{2i\pi S\nu_p(t)t + i\varphi_p} W_{\nu_p(t)}(f), \tag{2}$$

where $W_{\nu_p(t)}$ is the discrete Fourier transform of the analysis window modulated by the p-th frequency in time frame t:

$$W_{\nu_p(t)}(f) = \sum_{n=0}^{N_w-1} w(n) e^{2i\pi\left(\nu_p(t) - \frac{f}{N_w}\right)n}. \tag{3}$$

Now, let us assume that there is at most one active sinusoid per frequency channel and per source. Drawing on [23], we propose to partition the whole frequency range into several regions called *regions of influence*. A region of influence $I_p(t) \subset \{0, ..., F-1\}$ corresponds to the set of frequency channels where the STFT is mainly determined by the p-th sinusoidal partial, i.e., the contributions of the other partials are negligible:

$$\forall f \in I_p(t),\ X(f,t) = A_p(t) e^{2i\pi S\nu_p(t)t + i\varphi_p} W_{\nu_p(t)}(f). \tag{4}$$

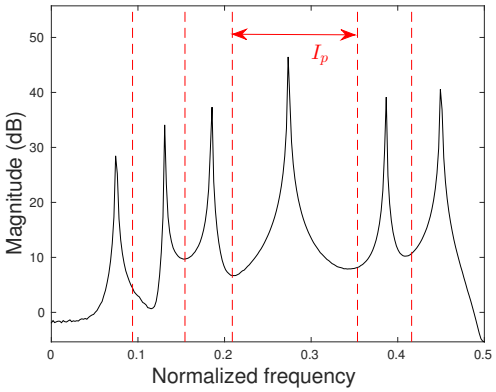


Fig. 1. Example of a spectrum (solid line) decomposed into regions of influence (dashed lines).

Several definitions of regions of influence exist in the literature, such as choosing their boundaries as the channels of lowest energy between the peaks [23]. Here we propose to adjust these boundaries so that the greater a magnitude peak is relatively to the neighboring peaks, the wider its region of influence becomes. The resulting regions, whose exact definition can be found in [26], are illustrated in Fig. 1.

From (4) we obtain the phase of the STFT $\phi = \angle X$, where $\angle$ denotes the complex argument:

$$\forall f \in I_p(t),\ \phi(f,t) = 2\pi S\nu_p(t)t + \varphi_p + \angle W_{\nu_p(t)}(f). \tag{5}$$

Finally, we assume that the sinusoids are *slowly-varying* [24], which means that $\nu_p(t) \approx \nu_p(t-1)$. This leads to the following recursive relationship between the phase of adjacent time frames:

$$\phi(f,t) \approx \phi(f,t-1) + 2\pi S\nu(f,t). \tag{6}$$

where $\nu(f,t) = \nu_p(t)\ \forall f \in I_p(t)$ with $p \in \{1, ..., P\}$. This relationship is called *phase unwrapping*.

B. Frequency estimation

In order to unwrap the phase through (6), one needs to estimate the frequencies $\nu(f,t)$. Many frequency estimation techniques exist, but they generally either require the phase of the STFT (e.g., as in the phase vocoder algorithm [23]) or are restricted to mixtures whose frequencies are in harmonic proportions (for instance, sophisticated versions of the harmonic sum or spectral product, such as the PEFAC algorithm [29]).

Therefore, we have chosen to use the Quadratic Interpolated FFT (QIFFT) [30]: we approximate the shape of the log-spectrum around a magnitude peak by a parabola, and the computation of the maximum of the parabola provides a frequency estimate, as illustrated in Fig. 2. This parabolic approximation is justified theoretically for Gaussian analysis windows, and used in practical applications for any window type. Since this approach provides overall fairly good results [31], [32], we will use it in our study.

C. The phase unwrapping algorithm

Algorithm 1 describes the PU procedure applied in one frame t. Note that the algorithm only reconstructs the phase

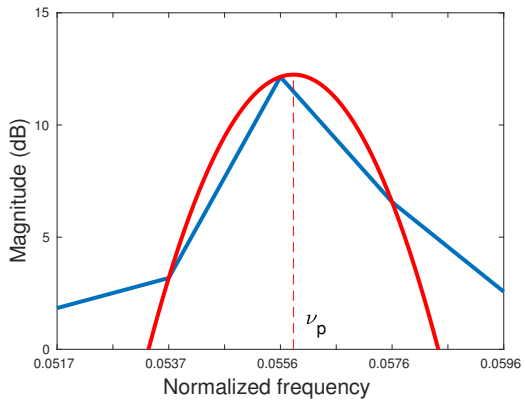


Fig. 2. Illustration of the QIFFT technique: a magnitude peak is approximated by a parabola, whose maximum leads to the frequency estimate.

Algorithm 1: Phase unwrapping

Inputs: Magnitude spectrum $v \in \mathbb{R}_+^F$,
phase in the previous time frame $\phi' \in \mathbb{R}^F$.
Peak localization f_p from v.
Frequencies ν_p with QIFFT on v around f_p.
Regions of influence I_p and $\forall f \in I_p$, $\nu(f) = \nu_p$.
Phase unwrapping $\phi = \phi' + 2\pi S\nu$.
Output: $\phi \in \mathbb{R}^F$

within non-onset frames, since the PU algorithm relies on a recursive relation between adjacent time frames (6). The phase within onset frames must be estimated differently. For instance, in the source separation framework, the mixture phase can be given to each component within onset frames, but alternative estimation techniques (e.g., [26], [33]) can be used.

III. Source separation procedure

In this section, we introduce a source separation procedure that exploits the PU algorithm. More details on the mathematical aspects related to this procedure are presented in a supporting document [34].

A. Problem setting

Source separation consists in extracting the K complex components X_k that form a mixture X. In this paper, we consider a linear, instantaneous and monaural mixture model: $X = \sum_{k=1}^{K} X_k$, and we assume that for all sources X_k, $k \in \{1, ..., K\}$, a magnitude estimate V_k is available. Indeed, in this work, we do not tackle the problem of magnitude estimation. Therefore, in our experiences (see Section IV), magnitudes will be assumed known or estimated beforehand on the isolated source spectrograms (as in *informed* source separation [35]).

We address this problem by minimizing the mixing reconstruction error, which is given by the following cost function:

$$\mathcal{C}(\theta) = \sum_{f,t} |X(f,t) - \sum_k \hat{X}_k(f,t)|^2, \tag{7}$$

under the constraint $|\hat{X}_k(f,t)| = V_k(f,t)$, and where $\theta = \{\hat{X}_k,\, k \in \{1, ..., K\}\}$. The Wiener filtering estimates:

$$\hat{X}_k(f,t) = \frac{V_k(f,t)^2}{\sum_{l=1}^{K} V_l(f,t)^2} X(f,t), \tag{8}$$

are not a solution to this problem since they do not verify $|\hat{X}_k| = V_k$. Thus, we propose here to introduce an iterative procedure which provides a novel set of estimates of the sources. Our idea is that a proper initialization of this procedure will provide estimates that benefit from some properties of the initial estimates (*cf.* Section III-E).

B. General framework

In order to minimize (7), we propose to use the auxiliary function method. Indeed, by decorrelating the variables, this technique makes it possible to update them in parallel rather than sequentially (as in the coordinate descent method). This leads to fast procedures and reduces the risk of local minima, so it is well-suited for addressing source separation problems [18], [36], [37]. Considering a cost function $h(\theta)$, the idea is to introduce a function $g(\theta, \tilde{\theta})$ which depends on some new parameters $\tilde{\theta}$, and verifies:

$$h(\theta) = \min_{\tilde{\theta}} g(\theta, \tilde{\theta}). \tag{9}$$

Such a function is called an *auxiliary function*. It can be shown (for instance in [36]) that h is non-increasing under the following update rules:

$$\tilde{\theta} \leftarrow \arg\min_{\tilde{\theta}} g(\theta, \tilde{\theta}) \text{ and } \theta \leftarrow \arg\min_{\theta} g(\theta, \tilde{\theta}). \tag{10}$$

C. Auxiliary function

Since all TF bins are treated independently, we remove the indexes (f,t) in what follows for more clarity, which results in seeking to minimize:

$$h(\theta) = |X - \sum_k \hat{X}_k|^2. \tag{11}$$

To obtain an auxiliary function for (11), we first introduce the auxiliary variables $\tilde{\theta} = \{Y_k, k \in k \in \{1, ..., K\}\}$ such that $\sum_k Y_k = X$. We have:

$$|X - \sum_k \hat{X}_k|^2 = |\sum_k (Y_k - \hat{X}_k)|^2. \tag{12}$$

We then introduce the following nonnegative weights:

$$\lambda_k = \frac{V_k^2}{\sum_l V_l^2}, \tag{13}$$

which leads to:

$$|X - \sum_k \hat{X}_k|^2 = \left| \sum_k \lambda_k \left(\frac{Y_k - \hat{X}_k}{\lambda_k} \right) \right|^2. \tag{14}$$

Since the weights defined in (13) verify $\sum_k \lambda_k = 1$, we can apply the Jensen inequality to the convex function $z \to z^2$:

$$|X - \sum_k \hat{X}_k|^2 \leq \sum_k \frac{|Y_k - \hat{X}_k|^2}{\lambda_k}. \tag{15}$$

Thus, $h(\theta) \leq g(\theta, \tilde{\theta})$ with:

$$g(\theta, \tilde{\theta}) = \sum_k \frac{|Y_k - \hat{X}_k|^2}{\lambda_k}, \tag{16}$$

and the problem becomes that of minimizing g under the constraints $\sum_k Y_k = X$ and $\forall k$, $|\hat{X}_k| = V_k$. Let us prove that g is an auxiliary function of the objective cost function h, i.e., that it satisfies (9). To do so, we aim to minimize g with respect to $\tilde{\theta}$ under the constraint $\sum_k Y_k = X$, which is equivalent to finding a saddle point for the functional $\mathcal{L}$:

$$\mathcal{L}(\theta, \tilde{\theta}, \gamma) = g(\theta, \tilde{\theta}) + \gamma(\sum_k \bar{Y}_k - \bar{X}) + \gamma'(\sum_k Y_k - X), \tag{17}$$

where $\bar{z}$ denotes the complex conjugate of z. Note that we need to introduce the constraint by means of the Lagrange multipliers twice: indeed, since the the function $\mathcal{L}$ is not real-valued, we have to treat separately the variables and their complex conjugates. We then calculate the partial derivatives of $\mathcal{L}$ with respect to the complex variables Y_k and $\bar{Y}_k$ (the so-called Wirtinger derivatives [38]).

$$\frac{\partial \mathcal{L}}{\partial Y_k}(\theta, \tilde{\theta}, \gamma) = \frac{1}{\lambda_k}(\bar{Y}_k - \bar{\hat{X}}_k) + \gamma', \tag{18}$$

and

$$\frac{\partial \mathcal{L}}{\partial \bar{Y}_k}(\theta, \tilde{\theta}, \gamma) = \frac{1}{\lambda_k}(Y_k - \hat{X}_k) + \gamma. \tag{19}$$

Setting those derivative at zero leads to equivalent conditions, therefore we have to solve:

$$Y_k = \hat{X}_k + \lambda_k \gamma. \tag{20}$$

Besides, summing (20) over k and using the constraint $\sum_k Y_k = X$ leads to:

$$\sum_k Y_k = \sum_k \hat{X}_k + \gamma \sum_k \lambda_k = X, \tag{21}$$

and since the weights λ_k add up to 1, we obtain:

$$\gamma = X - \sum_k \hat{X}_k, \tag{22}$$

which leads to:

$$Y_k = \hat{X}_k + \lambda_k (X - \sum_l \hat{X}_l). \tag{23}$$

Thus, $g(\theta, \tilde{\theta})$ is minimized for a set of auxiliary parameters defined by (23), and it is quite straightforward to observe that it is then equal to $h(\theta)$. This proves that g is an auxiliary function of h.

D. Derivation of the updates

In accordance with (10), we obtain the update rules on θ and $\tilde{\theta}$ by alternatively minimizing g with respect to these variables. As it has already been shown, the update rule on Y_k is given by (23). To obtain the update rule on $\hat{X}_k$, we introduce the constraints $|\hat{X}_k| = V_k$, $\forall k$, by means of the Lagrange multipliers:

$$\mathcal{H}(\theta, \tilde{\theta}, \delta_1, ..., \delta_K) = g(\theta, \tilde{\theta}) + \sum_k \delta_k (|\hat{X}_k|^2 - V_k^2), \tag{24}$$

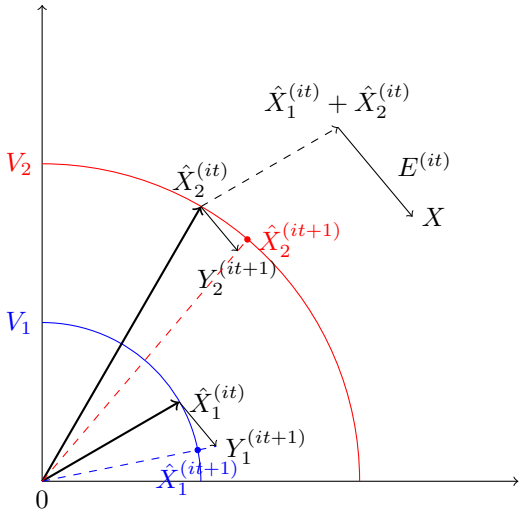


Fig. 3. Iterative estimation of two complex numbers of fixed magnitude and whose sum is known.

and we apply the same methodology as before: we compute the zeros of the partial derivatives of $\mathcal{H}$ with respect to the complex variables $\hat{X}_k$. This leads to:

$$\hat{X}_k = \frac{Y_k}{1 + \lambda_k \delta_k}. \tag{25}$$

By taking the modulus in (25) and using the constraints $|X_k| = V_k$, we have:

$$|1 + \lambda_k \delta_k| = \frac{|Y_k|}{V_k}, \tag{26}$$

and finally, by combining this relation and (25), we get:

$$\hat{X}_k = \pm V_k \frac{Y_k}{|Y_k|}. \tag{27}$$

To avoid any ambiguity on the sign in (27), we calculate the value of g for both cases and find that g is minimized with respect to θ when $\forall k$:

$$\hat{X}_k = V_k \frac{Y_k}{|Y_k|}. \tag{28}$$

Ultimately, the objective function h is minimized by alternatively applying the update rules (23) and (28). The procedure, illustrated in Fig. 3 for $K = 2$, actually appears as quite similar to the work in [18]: the key idea is to distribute the mixing error over the current estimates, and then normalizing. The choice for the weights λ_k in (13) is therefore consistent with this interpretation: the components of highest energy have more impact on the estimation error than the components of lowest energy.

E. Initialization

The keystone of our approach is that it enables us to incorporate some prior phase information about the components through a properly-chosen initialization. Indeed, the cost function $\mathcal{C}$ has many global minima (for $K \geq 3$, the problem has infinitely many solutions). Thus, our goal is to find a solution which benefits from some prior knowledge about the phase in order to lead to satisfactorily sounding results. Intuitively,

one could initialize the algorithm by giving the phase of the mixture to each source, but according to (23) and (28), the phase of the estimates would not be modified over iterations. Then, we propose to initialize this procedure with the PU algorithm: the corresponding initial components are expected to be close to a local minimum and the output estimates can benefit from some temporal continuity. The usefulness of such an initialization will be demonstrated experimentally in Section IV-D.

F. Source separation procedure

The procedure is summarized in Algorithm 2. Firstly, let us note that the set of onset frames for each source Ω_k and the corresponding onset phases ϕ_k^o are provided as inputs of the algorithm, since both onset frame detection and onset phase estimation are outside the scope of this article.

Outside onset frames, the phase is initialized by applying the PU technique (Algorithm 1). From this initial estimate, we apply the proposed iterative procedure. Finally, we move to the next time frame. A MATLAB implementation of this algorithm is available on the companion website for this paper [39].

Note that we propose here to apply the iterative procedure within a given time frame before moving to the next one. Alternatively, we can apply the PU algorithm to the whole spectrograms and then apply the iterative procedure in parallel for all TF bins, in order to reduce the computational cost of the method. However, we observed experimentally that the time gain is of about 5%, but the performance in terms of separation quality decreases more significantly. Therefore, it is better to get a good estimate of the phase within a time frame before unwrapping it to the next frame, since it allows us to avoid propagating (and amplifying) the estimation error over time.

Algorithm 2: Source separation procedure using PU

Inputs :
Mixture $X \in \mathbb{C}^{F\times T}$,
Spectrograms $V_k \in \mathbb{R}_+^{F\times T}$, $\forall k \in \{1,...,K\}$,
Onset frames sets Ω_k, $\forall k \in \{1,...,K\}$,
Onset phases $\phi_k^o(f,t)$, $\forall t \in \Omega_k$,
Number of iterations N_{it}.
for $t = 1$ *to* $T-1$ **do**
  % Initialization
  for $k = 1$ *to* K **do**
    if $t \in \Omega_k$ **then**
      Onset phase: $\phi_k(f,t) = \phi_k^o(f,t)$.
    else
      $\phi_k(f,t)$ = Phase unwrapping (*cf.* Algorithm 1).
    end
    $\hat{X}_k(f,t) = V_k(f,t)e^{i\phi_k(f,t)}$.
  end
  % Iterative procedure
  for $it = 1$ *to* N_{it} **do**
    Update $Y_k(f,t)$ with (23),
    Update $\hat{X}_k(f,t)$ with (28),
  end
end
Output : $\forall k \in \{1,...,K\}$, $\hat{X}_k \in \mathbb{C}^{F\times T}$.

IV. Experimental validation

In this section, we evaluate the potential of the PU algorithm for phase recovery and its usefulness for initializing the iterative procedure for source separation. Sound excerpts can be found on the companion website for this paper [39] to illustrate the experiments.

A. Setup

We consider 50 music song excerpts from the DSD100 database, a semi-professionally mixed set of music song used for the SiSEC 2016 campaign [40]. Each excerpt is 10 seconds-long and is made up of $K = 4$ sources: `bass`, `drums`, `vocals` and `other` (which may contain various instruments such as guitar, piano...). The signals are sampled at $F_s = 44100$ Hz and the STFT is computed with a Hann window, 75 % overlap and no zero-padding. The length of the analysis window will be discussed in Section IV-C.

The MATLAB Tempogram Toolbox [41] provides a fast and reliable onset frames detection from spectrograms (it estimates the onsets before several post-processing operations to find the tempo). The phases within onset frames will be either assumed known or estimated by assigning the mixture onset phase to each source (in Section IV-E). In Section IV-F, we will investigate the impact of the onset phase estimation on the separation quality.

Finally, the magnitude peaks f_p are tracked from the spectra by using the corresponding MATLAB function (`findpeaks`) in Algorithm 1, and the iterative procedure in Algorithm 2 uses 50 iterations.

In order to measure the performance of the methods, we use the BSS Eval toolbox [42] which computes various energy ratios: the signal-to-distortion, signal-to-interference, and signal-to-artifact ratios (SDR, SIR and SAR), which are expressed in dB and where only a rescaling (not a refiltering) of the reference is allowed [43].

B. Separation scenarios

In this paper, we do not address the task of blind magnitude estimation, therefore the magnitudes are either assumed known or estimated beforehand.

First, in the oracle scenario, we assume that the magnitudes V_k are equal to the ground truth.

Second, we consider an informed scenario, which corresponds to a *coding-based informed source separation* framework [35]: in this scenario, some side-information can be computed from the isolated sources (the *encoding* stage) and then used to enhance the separation performance (the *decoding* stage). A common approach consists of computing a nonnegative matrix [44], [45] or tensor [46]–[48] factorization on the isolated source spectrograms and then using the corresponding decomposition to estimate a Wiener filter at the decoding stage. Those approaches have shown very good

results in terms of separation quality at a very low bitrate. We therefore propose to apply an NMF with Kullback-Leibler divergence [2] to the spectrogram of each isolated source, in order to obtain an estimate of the magnitudes V_k. Each NMF uses 200 iterations of multiplicative update rules and a rank of factorization set at 50. This scenario informs us about the potential of the proposed method for an audio source separation task when the magnitude estimates differ from the ground truth, while still remaining of relatively good quality.

Third, we consider a blind scenario, in which the magnitude spectrograms are directly estimated from the mixture. We apply 200 iterations of multiplicative updates of Kullback-Leibler NMF with a rank of factorization set at 200 to the mixture's spectrogram $|X|$. Then, the NMF components are grouped into 4 sources by means of the source filter-based clustering method described in [49] which yields an estimate of the magnitudes V_k. This scenario informs us about the impact of phase recovery on source separation quality when the magnitudes are not accurately estimated.

C. Griffin-Lim vs. phase unwrapping

The goal of this experiment is to compare the performance of a consistency-based approach (the GL algorithm) and a model-based approach (the PU algorithm) for a blind phase retrieval task. Since the songs from the DSD100 database are made up of sources that overlap in the TF domain, we can no longer assume that there is at most one frequency component per channel in the mixtures, which is a key hypothesis in the PU technique (see Section II-A). Therefore, we report here the results for 30 piano pieces from the MAPS [50] database, where this scenario is less likely to occur (similar results have been obtained on guitar and speech signals, which we omit here for brevity). Note that for source separation applications, since the PU technique is performed on isolated sources, this will no longer be a problem.

First, a comparison between three analysis windows (Hann, Hamming and Blackman) showed no significant difference in terms of SDR. In addition, overlap ratios higher than 75 % did not improve the results, while they required more processing time. For those reasons, we chose a Hann window with 75 % overlap in our experiments. We study here the impact of the window length on the reconstruction quality measured by means of the SDR, and we also compute the inconsistency of the estimates defined as follows:

$$\mathcal{I}(X) = \sum_{f,t} |X(f,t) - \text{STFT} \circ \text{iSTFT}(X)(f,t)|^2 . \quad (29)$$

In this experiment, the onset phases are assumed known. We corrupt the complex STFT of the signals by setting the phases within non-onset frames to random values taken in $]-\pi;\pi]$. We then apply the algorithms in both oracle and informed scenarios. The GL algorithm uses 200 iterations (performance is not further improved beyond). We also report the scores (SDR and inconsistency) computed on the corrupted STFTs as a comparison reference. The results are presented in Fig. 4.

We observe that overall, the PU algorithm outperforms the traditional GL method in terms of SDR for most window

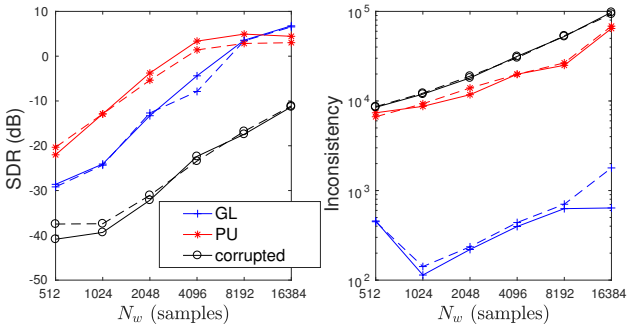


Fig. 4. Comparison between GL and PU in terms of SDR (left) and inconsistency (right) in the oracle (solid lines) and informed (dashed lines) scenarios for various window lengths.

lengths. Both algorithms are sensitive to the accuracy of the magnitude spectrogram, as suggested by the drop in SDR values when going from the oracle to the informed scenario for most analysis windows. However, when the spectrogram is no longer equal to the ground truth, our approach still provides overall better results than the consistency-based GL algorithm.

Both algorithms decrease the inconsistency compared to the corrupted reference, but the GL algorithm performs better than PU according to this criterion. This was expected since the GL algorithm is designed to directly minimize this criterion. However, a comparison between SDR and inconsistency shows that minimizing the inconsistency does not imply increasing the SDR. This suggests that the direct optimization of the inconsistency criterion may not be the most appropriate way of accounting for this property. Besides, the GL algorithm computes an estimate that is optimal (in terms of inconsistency) only locally, while a global minimum would correspond to a null inconsistency in the oracle scenario.

We note that the longer the window, the better the results. Actually, for very long analysis windows, GL performs better than PU in terms of SDR. This may be explained by some artifacts that appear in the signals estimated with the PU technique. Indeed, perceptually (sound examples are available in [39]), two phenomena characterize them: *musical noise*, which appears for short windows, when the frequency resolution is poor, and *phasiness* [23], which appears for long windows, when the temporal resolution is poor[1]. In the latter case, the local stationarity assumption, on which the PU algorithm is based, does no longer hold, thus leading to a decrease of its performance. Therefore, it is not obvious that the SDR is able to capture both the musical noise and the phasiness phenomena. Indeed, some informal listening tests showed that windows shorter than 16384 samples lead to more satisfactorily sounding results for both algorithms, while they make the SDR decrease according to Fig. 4. In particular, a 4096 sample-long analysis window leads to the best results in terms of perceptual quality, with a fairly high SDR.

Finally, for audio source separation applications, we tested different window lengths and we observed that a 4096 sample-

[1]To overcome the issue of looking for a compromise between temporal and frequency resolution, a multiple resolution framework could further be investigated, as in some improved versions of the phase vocoder [51], [52].

TABLE I
SOURCE SEPARATION PERFORMANCE (SDR, SIR AND SAR IN DB) FOR VARIOUS INITIALIZATIONS ON THE DSD100 DATASET.

Initialization	SDR	SIR	SAR
Mixture	7.5	13.7	8.9
Random	9.5	22.8	9.7
PU	**13.6**	**31.0**	**13.7**

long analysis window leads to the best results in terms of SDR for the consistent Wiener filtering technique and the proposed iterative procedure. Therefore, we use this value in the following experiments.

D. Initialization of the iterative procedure

Here, we investigate the influence of the initialization of the iterative procedure on the separation quality, as motivated in Section III-E. Let us consider 50 songs from the DSD100 dataset in the oracle scenario, and the onset phases are assumed known. We run the procedure with different initializations (line 13 in Algorithm 2): either PU or random values. We also test initializing with the mixture phase in order to obtain a comparison reference, even if the phase of the estimates will not be modified over iterations, as explained in Section III-E. The results provided in Table I show that the initialization with the PU algorithm significantly improves the separation quality over the other initializations.

We consider one mixture and we plot the error $\mathcal{C}$ in a TF bin where the sources overlap in Fig. 5. We see that the PU initialization leads to a better and faster convergence (in terms of error) than a random initialization. Besides, these two techniques reach a significantly lower error value than the initialization with the mixture phase. Perceptually (audio examples are available at [39]), we observe that the PU initialization yields estimates with less artifacts than the other techniques, especially in the `bass` and `drum` tracks (which overlap the most). We also note that there is no significant difference in terms of sounding quality between a random initialization and using the mixture's phase, while the corresponding output errors (see Fig. I) strongly differ. This suggests that this criterion does not perfectly retrieve perceptual criteria.

To illustrate this result, we consider a mixture composed of two piano notes from the MAPS database. In order to visualize the real parts of the reconstructed components, we synthesize time-domain signals and compute another STFT with a hop size of 1 sample. As illustrated in Fig. 6, the initialization with PU yields components that better fit the original signal compared to the other approaches. This confirms the usefulness of the PU algorithm to initialize Algorithm 2.

E. Comparison to other methods

In this experiment, the onset phases are estimated by giving the mixture phase to each component. We compare the following methods: Wiener filtering [3], consistent Wiener filtering [21][2], and Algorithm 2. Those methods will be respectively denoted **Wiener**, **Cons-W** and **PU-Iter**. The separation is performed on the 50 songs composing the dataset, and the results are represented with box-plots in Fig. 7. To complete these results, we also provide some sound excerpts [39] so that the interested reader can assess the sounding quality of the corresponding estimates.

[2]This technique depends on a weight parameter that promotes the consistency constraint. It is learned beforehand on 50 other songs from the dataset by choosing the value that maximizes the SDR, SIR and SAR.

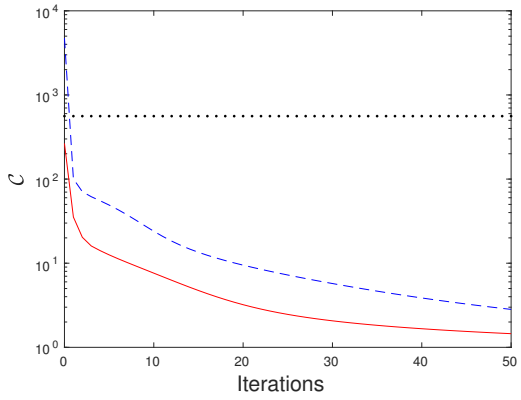


Fig. 5. Error $\mathcal{C}$ over iterations within a TF bin where the sources overlap. The dotted and solid lines respectively correspond to the initializations with the mixture phase and PU algorithm, and the dashed line corresponds to the average values over 10 random initializations.

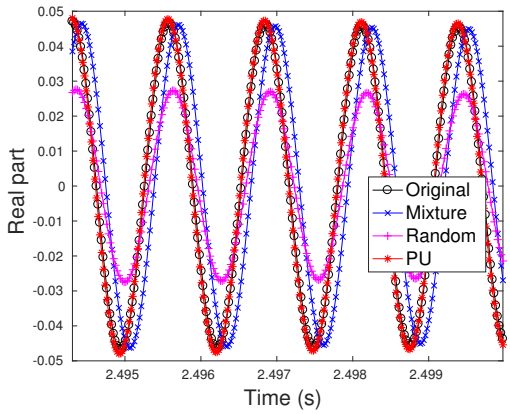


Fig. 6. Real part of the third partial (784 Hz) in the STFT channel at 786 Hz of a C4 piano note where it overlaps with another note (G4), for various initializations of Algorithm 2.

In the oracle scenario, **PU-Iter** outperforms **Wiener** and **Cons-W**, notably in terms of SIR. In addition to those indicators, we perceptually note that **Cons-W** and **PU-Iter** lead to similar `vocals` and `other` tracks (in terms of sounding quality), but the `bass` and `drum` tracks estimated with **PU-Iter** are of higher quality: the `bass` is neater and the musical noise artifacts in the `drum` track are reduced compared with the other approaches.

In the informed scenario, the proposed method yields slightly worse results than **Cons-W** in terms of SDR and SAR (it is still better than **Wiener**), but leads to an improvement in terms of interference rejection. This observation is consistent with previous works [53] on sinusoidal model-based phase recovery, where this approach has been shown useful to reduce the interference at the cost of more artifacts. However, when

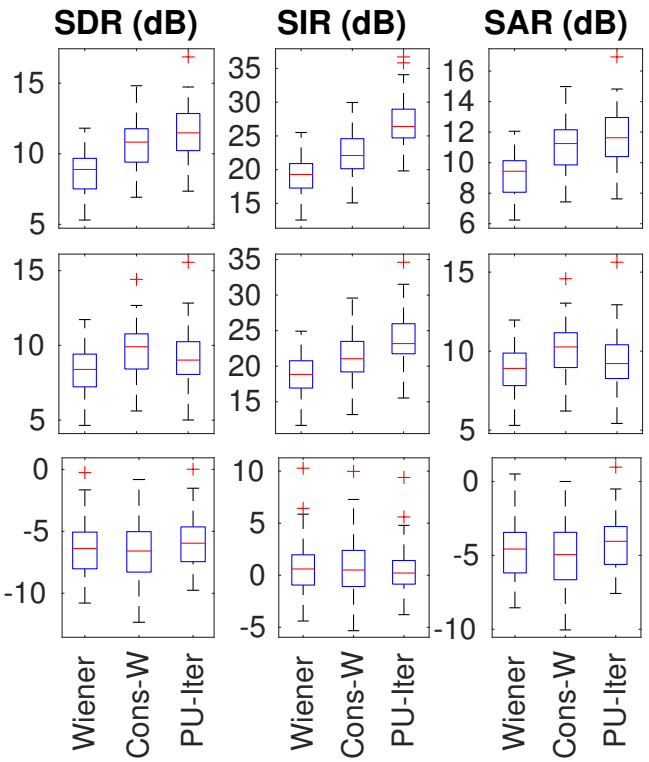


Fig. 7. Source separation performance of various methods in the oracle (top), informed (middle) and blind (bottom) scenarios. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles, whiskers indicating the minimum and maximum values, and crosses representing the outliers.

TABLE II
AVERAGE PEASS SCORES.

		Wiener	Cons-W	PU-Iter
Oracle	OPS	19.2	19.7	**27.4**
	TPS	28.4	30.4	**36.1**
	IPS	34.7	34.5	**39.8**
	APS	30.6	31.0	**36.3**
Informed	OPS	19.0	19.0	**24.5**
	TPS	25.8	26.9	**33.3**
	IPS	34.0	33.8	**41.5**
	APS	28.8	28.9	**33.5**
Blind	OPS	10.4	10.5	**12.5**
	TPS	11.6	12.4	**17.0**
	IPS	33.1	**32.6**	27.7
	APS	17.5	18.2	**25.4**

listening at the corresponding excerpts, we observe that the **PU-Iter** method still enhances the sounding quality of the `bass` track compared to the **Cons-W** technique, while other tracks are similar.

In the blind scenario, all methods' performance significantly decrease compared to the other scenarios, and yield overall similar results. Nonetheless, **PU-Iter** leads to a slight increase in SDR and SAR compared to the other methods, but this method's interest is greater when the magnitude spectrograms are reliably estimated.

Since our informal listening evaluation is somewhat inconsistent with the results in terms of objective criteria, we propose to compute the PEASS score [54], which provides a

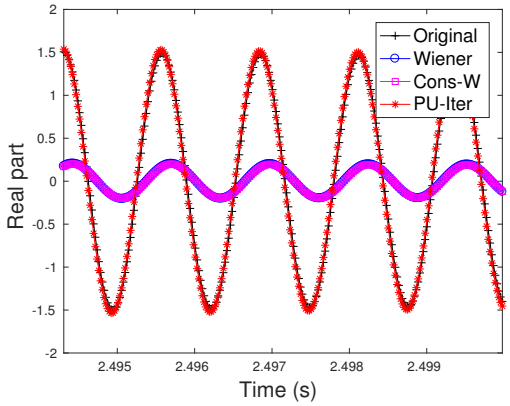


Fig. 8. Real part of the third partial (784 Hz) in the STFT channel at 786 Hz of a C4 piano note where it overlaps with another note (G4), reconstructed with several methods in the oracle scenario.

novel set of criteria that is built upon a subjective evaluation of source separation quality, and designed to better match perception than the SDR, SIR and SAR. The resulting criteria are the overall, target-related, interference-related and artifacts-related Perceptual Scores (OPS, TPS, IPS and APS). The corresponding results are presented in Table II. We observe that for all those criteria, in both the oracle and informed scenario, the proposed **PU-Iter** method outperforms **Wiener** and **Cons-W** by a large margin: in particular, the improvement when going from **Cons-W** to **PU-Iter** is more significant than the improvement of **Wiener** over **Cons-W** on average. This confirms the performance of the proposed approach and is consistent with our informal perceptive evaluation. In the blind scenario, **PU-Iter** still outperforms the other techniques, except in terms of IPS. Overall, the relative performance of **PU-Iter** increases when the magnitude estimates get close to the ground truth. Indeed, the OPS difference between **PU-Iter** and **Cons-W** is of 2, 5.5 and 7.7 in the blind, informed and oracle scenarios respectively. This confirms that while the proposed approach shows good results compared to other methods, its potential is fully exploited when the magnitude spectrograms are accurately estimated.

We illustrate these results on the same example as in the previous Section (mixture of overlapping piano notes). We observe in Fig. 8 that the **PU-Iter** estimate better fits the ground truth than the other methods. This is due to the fact that **Wiener** and **Cons-W** modify the target magnitude when sources overlap in the TF domain, which is not a desirable property if the magnitude has been reliably estimated.

Finally, it is important to note that **Cons-W** is computationally costly: for 10 seconds excerpts, the average phase retrieval time is 11.8 seconds with **Cons-W** vs 4.9 seconds with our method. The proposed approach then appears appealing for an efficient audio source separation task, notably in terms of interference rejection.

F. Impact of the onset phase

Finally, we evaluate the room for improvement of onset phase recovery. We run the **PU-Iter** procedure in the oracle

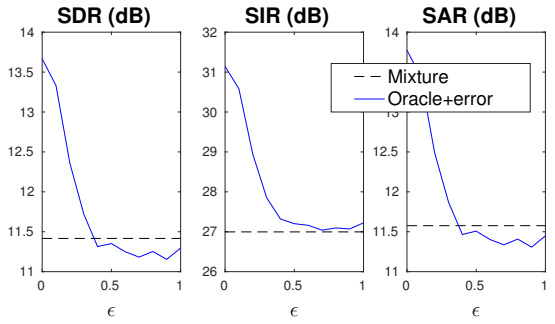


Fig. 9. Separation performance for various onset phases.

scenario considering two different settings: onset phases can be estimated by assigning the mixture phase to each component (as in the previous experiment), or alternatively, they are equal to the ground truth phase to which a random error is added under the form of a centered Gaussian white noise with standard deviation $2\pi\epsilon$, where ϵ ranges between 0 and 1.

From the results in Fig. 9 we remark that there is a gap in terms of SDR, SAR (≈ 2 dB) and SIR (≈ 4 dB) between using the mixture phase and the oracle phase within onset frames. When some perturbation is added to the oracle phase, the performance decreases. We also note that in terms of SDR and SAR, when the error is of approximately 40 %, the performance becomes similar to that of giving the mixture phase to each component within onset frames. This is consistent with the fact that the error between true phases and mixture phases within onset frames amounts to roughly 40 %. Note that this observation does not hold for the SIR. One possible explanation is that using the mixture phase yields estimates with more interferences because the mixture phase contains information that is relative to all sources. Thus, when the error between the mixture phase and the true phase is important (more than 50 %), using the mixture phase reduces the artifacts and distortion, but it may introduce some interferences.

Overall, giving the mixture phase to each component is fast and easy to implement, but onset phase reconstruction can be improved to fully exploit the potential of the PU technique.

V. Conclusion

The source separation procedure introduced in this paper exploits the PU algorithm in order to promote a form of temporal continuity in its output estimates. The experimental results have shown that such a procedure outperforms consistent Wiener filtering in a scenario in which the magnitude spectra are known. In a more realistic scenario, where the magnitudes are estimated beforehand, it reaches a performance similar to other methods in terms of objective criteria, with a significant improvement in terms of computational cost. Sound excerpts also show that in terms of perceptual quality, this approach compares favorably with the state-of-the-art consistency-based source separation approach, which is confirmed by the computation of perception-related metrics.

The proposed approach has shown good results when the magnitude spectrograms are reliably estimated: therefore, it could be further combined with deep neural networks since these models have demonstrated remarkably good performance for magnitude estimation [55]. Another promising future research direction is to combine consistency-based and model-based phase recovery techniques for exploiting the full potential of both approaches, as first attempted in [56]. Besides, as suggested by the last experiment, onset phase recovery is an interesting research direction for improved sounding quality. For instance, onsets can be modeled as impulses [26], [57], or one can use a model of repeated audio events within onset frames [33]. Finally, some phase information can be incorporated into a probabilistic model where the phase is no longer uniform, in order to yield conservative source estimates [58].

References

[1] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*. Academic press, 2010.

[2] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, March 2009.

[4] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, August 2014.

[5] P.-S. Huang, M. K. M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014.

[6] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, September 2016.

[7] T. Gerkmann, M. Krawczyk, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[8] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, July 2016.

[9] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[10] R. M. Parry and I. Essa, "Incorporating Phase Information for Source Separation via Spectrogram Factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007.

[11] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[12] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, September 2008.

[13] D. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[14] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-Time Signal Estimation From Modified Short-Time Fourier Transform Magnitude Spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

[15] N. Perraudin, P. Balazs, and P. L. Sondergaard, "A fast Griffin-Lim algorithm," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013.

[16] G. T. Beauregard, M. Harish, and L. L. Wyse, "Single Pass Spectrogram Inversion," in *Proc. IEEE International Conference on Digital Signal Processing (DSP)*, 2015.

[17] V. Gnann and M. Spiertz, "Improving RTISI phase estimation with energy order and phase unwrapping," in *Proc. International Conference on Digital Audio Effects (DAFx)*, September 2010.

[18] D. Gunawan and D. Sen, "Iterative Phase Estimation for the Synthesis of Separated Sources From Single-Channel Mixtures," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 421–424, May 2010.

[19] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March.

[20] ——, "Informed Source Separation Using Iterative Reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 178–185, January 2013.

[21] J. Le Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[22] R. J. McAuley and T. F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[23] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[24] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, December 2014.

[25] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. International Conference on Spoken Language Processing*, September 2012.

[26] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *Proc. European Signal Processing Conference (EUSIPCO)*, August 2015.

[27] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014.

[28] M. Lagrange and S. Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods," *Journal of the Audio Engineering Society*, vol. 55, pp. 385–399, May 2007.

[29] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb 2014.

[30] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Proc. Audio Engineering Society Convention*. Audio Engineering Society, May 2004.

[31] M. Betser, P. Collen, G. Richard, and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 505–517, February 2008.

[32] M. Abe and J. O. Smith, "Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolation," Stanford University, Department of Music, Tech. Rep. STAN-M-117, 2004.

[33] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms based on a model of repeated audio events," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2015.

[34] ——, "An iterative algorithm for recovering the phase of complex components from their mixture," Paris, France, Tech. Rep. hal-01325625, June 2016.

[35] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937–1949, 2012.

[36] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2009.

[37] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011.

[38] P. Bouboulis, "Wirtinger's calculus in general hilbert spaces," *arXiv preprint arXiv:1005.5170*, 2010.

[39] http://www.cs.tut.fi/~magron/demos/demo_PUITER.html.

[40] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 Signal Separation Evaluation Campaign," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Feb. 2017.

[41] P. Grosche and M. Müller, "Tempogram Toolbox: MATLAB tempo and pulse analysis of music recordings," in *Proc. International Society for Music Information Retrieval (ISMIR) Conference*, October 2011.

[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[43] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, September 2016.

[44] C. Rohlfing, J. M. Becker, and M. Wien, "NMF-based informed source separation," in *Prof. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

[45] C. Rohlfing and J. M. Becker, "Generalized constraints for NMF with application to informed source separation," in *Proc. European Signal Processing Conference (EUSIPCO)*, August 2016.

[46] C. Rohlfing, J. E. Cohen, and A. Liutkus, "Very low bitrate spatial audio coding with dimensionality reduction," in *Prof. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[47] A. Liutkus, R. Badeau, and G. Richard, "Low bitrate informed source separation of realistic mixtures," in *Prof. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

[48] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1699–1712, Aug 2013.

[49] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. International Conference on Digital Audio Effects (DAFx)*, September 2009.

[50] V. Emiya, N. Bertin, B. David, and R. Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Télécom ParisTech, Paris, France, Tech. Rep. 2010D017, July 2010.

[51] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. International Conference on Digital Audio Effects (DAFx)*, September 2003.

[52] N. Juillerat and B. Hirsbrunner, "Audio time stretching with an adaptive multiresolution phase vocoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[53] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: application to audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

[54] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, September 2011.

[55] N. Takahashi and Y. Mitsufuji, "Multi-scale Multi-band DenseNets for Audio Source Separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017.

[56] P. Magron, J. Le Roux, and T. Virtanen, "Consistent anisotropic Wiener filtering for audio source separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017.

[57] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Prof. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2013.

[58] P. Magron, R. Badeau, and B. David, "Phase-dependent anisotropic Gaussian model for audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.