Model-based STFT phase recovery for audio source separation

Paul Magron, Roland Badeau, David Bertrand

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 Time-Frequency (TF) representation of each source. In order to recover a time-domain signal from a spectrogram for instance, it then becomes necessary to recover the phase of the corresponding complex-valued Short-Time Fourier Transform (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 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 whole iterative source separation procedure is proposed which builds upon these results. It is tested on a variety of data, both synthetic and realistic, and also with different source separation scenarios, oracle or non oracle. In terms of SIR, SAR and SDR, the method achieves better performance than consistency-based approaches. To complete the experimental analysis, sound examples are provided which allow the reader to assess the interest of the method regarding the improvement of sound quality.**

***Index Terms*—Phase recovery, sinusoidal modeling, linear unwrapping, audio source separation.**

I. Introduction

A variety of music signal processing techniques acts in the Time-Frequency (TF) domain, since it provides a meaningful representation of audio signals. For instance, the family of techniques based on Nonnegative Matrix Factorization (NMF) [1] is often applied to nonnegative TF representations, such as the magnitude of the STFT. It has been shown promising for various musical applications, such as automatic transcription [2] and source separation [3], [4].

However, when it comes to resynthesizing time signals, obtaining the phase of the corresponding complex-valued STFT is necessary. In the source separation framework, a common practice consists in applying a Wiener-like filtering [3] to the original mixture: the phase of the mixture is then given to each extracted component. Alternatively, a consistency-based approach can be used for phase recovery [5]: a complex-valued matrix is iteratively computed in order to maximize its consistency, that is, to bring it as close as possible to the STFT of a time signal. It has however been pointed out [6] that consistency-based approaches provide poor results in terms of audio quality. Besides, Wiener filtering fails to provide good results when sources overlap in the TF domain. There were some attempts [7]–[10] to overcome the limitations of those two approaches by combining them in a unified framework. Consistent Wiener filtering [10] has proved to be the most promising candidate for this task, although it is computationally costly. Thus, the phase recovery of STFT components is still a challenging and open issue [11], [12].

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 [13]. This family of techniques exploits the natural relationship between adjacent TF bins that originates from signal modeling. Such an approach has been used in the phase vocoder algorithm [14], where it is mainly dedicated to time stretching and pitch shifting, but requires the phase of the original STFT. More recently, it has been applied to speech signal reconstruction [15], [16] and source separation [17] based on a Complex NMF (CNMF) framework, which is dedicated to jointly estimating both the magnitudes and phases [18]. Although promising, these techniques are limited to harmonic and stationary signals, which means that they consider mixtures of sinusoids whose frequencies are integer multiples of a fundamental frequency that is constant over time. Besides, the phase-constrained CNMF approach [17] requires that the fundamental frequencies and numbers of partials are known.

Drawing on a preliminary work [19], we propose in this paper a generalization of this approach that consists in exploiting the phase of mixtures of sinusoids. We then obtain an algorithm which unwraps the phases over time frames, ensuring the temporal coherence of the signal. Our technique, called the *Phase Unwrapping* (PU) algorithm, is suitable for a variety of pitched music signals, such as piano or guitar sounds, but percussive signals are outside the scope of this research. A local estimation (at each time frame) of frequencies extends the validity of this technique to non-stationary signals such as cellos and speech. This enables us to overcome the limitations of the previous approaches [15], [17] that were restricted to harmonic and stationary signals. We further introduce a novel source separation procedure which exploits the prior information about the phase that is provided by the PU algorithm. Unlike CNMF methods, this technique assumes that the magnitude spectrograms of the sources are estimated beforehand (e.g. after a preliminary NMF [1]), and only tackles the phase recovery issue. This technique is tested on a variety of realistic music signals, which points out its potential for an audio source separation task.

This paper is organized as follows. Section II describes the most commonly used phase reconstruction techniques in audio. Section III presents the PU algorithm, and Section IV introduces an audio source separation framework which uses

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

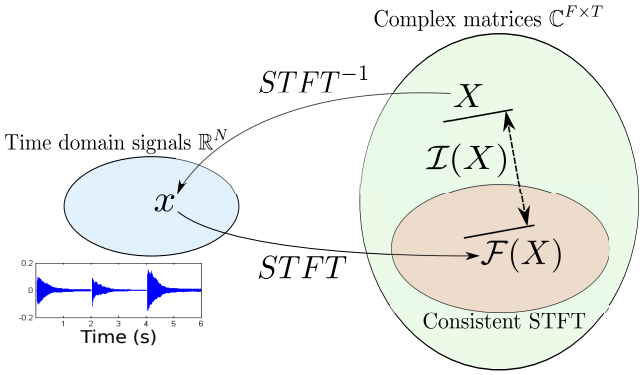

Fig. 1. The concept of inconsistency, which measures the difference between X and $\mathcal{F}(X)$.

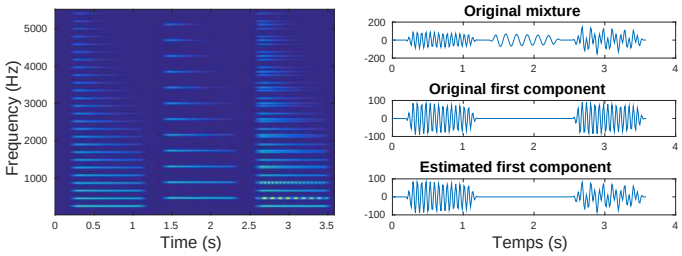

Fig. 2. Spectrogram of a synthetic signal composed of two overlapping sources (left). Real part of several partials in the 430 Hz frequency channel: original mixture (upper right), original first component (middle right) and estimated first component with Wiener filtering (lower right). The beating phenomenon due to the TF overlap is captured in the mixture phase and thus retrieved in the components when applying Wiener filtering.

this technique. Section V experimentally validates the potential of the PU algorithm, notably for an audio source separation task. Finally, section VI draws some concluding remarks.

II. Relation to prior work

Much research in audio has focused on the processing of nonnegative TF representations, such as magnitude or power spectrograms. Indeed, the phase recovery issue has been considered of minor importance, especially in the speech enhancement community [20]. However, some recent studies pointed out its importance [21], [22]. Thus, it has become a growing topic of interest [11], [12]. In this section, we describe the main phase reconstruction approaches that are specifically used for audio applications. We highlight the limitations of these techniques, and we show how the proposed approach can overcome some of their issues.

A. Consistency-based approaches

A common approach for recovering the phase of an estimated complex matrix $X \in \mathbb{C}^{F\times T}$ consists in minimizing its *inconsistency* $\mathcal{I}(X)$ [23], defined as follows:

$$\mathcal{I}(X) = \sum_{f,t} \left(X - \mathcal{F}(X)\right)^2_{(f,t)}, \tag{1}$$

where $\mathcal{F} = STFT \circ STFT^{-1}$ and $STFT^{-1}$ denotes the inverse STFT, computed with a standard overlap-add method [23]. It is illustrated in Fig. 1.

The Griffin Lim (GL) algorithm [5] consists in iteratively applying the operator $\mathcal{F}$ to a complex matrix, while enforcing the magnitude to be constant over iterations. Various strategies have been proposed in order to increase the speed and efficiency of this procedure: explicit consistency constraints [23], real-time implementation [24], new formulation [25], better initialization [26], [27] etc. Consistency is an important property of the STFT since it is closely related [23] to its redundancy property: indeed, the STFT is generally computed with overlapping analysis windows. However, it has appeared [6] that a direct optimization of the inconsistency criterion may not necessarily be the best way of accounting for it, since it does not lead to satisfactorily sounding signals.

B. Time-frequency masking

Alternatively, in a source separation framework, where the phase of the mixture is known, it is usual to apply a soft mask to the TF representation of the mixture, which leads to assigning the phase of the mixture to each extracted component. Let X be the complex-valued STFT of a mixture of K sources. Let us assume that an estimate V_k of the magnitude spectrogram (or equivalently of the power spectrum $V_k^{\odot 2}$, where $.^{\odot}$ denotes the element-wise matrix power) is available for each source $k \in \{1, ..., K\}$. For instance, such estimate can be obtained by a preliminary NMF [1]. The Wiener filtering technique provides the following complex component estimates:

$$\hat{X}_k = \frac{V_k^{\odot 2}}{\sum_{l=1}^{K} V_l^{\odot 2}} \odot X, \tag{2}$$

where $\odot$ (resp. $\div$) denotes the element-wise matrix multiplication (resp. division). Though this technique is optimal in a least-square sense [28], it fails to provide good results when the sources overlap in the TF domain [6], as illustrated in Fig. 2.

C. Consistent Wiener filtering

In order to overcome the limitations of both Wiener filtering and consistency-based approaches, there were attempts [7]–[10] to combine them in a source separation framework. In [7], the error between the mixture and the GL-estimates $\hat{X}_k$ is distributed over the sources in order to enforce the mixing constraint $X = \sum_k \hat{X}_k$. In [8], [9], the TF domain is decomposed into regions depending on whether a given source is dominant (in which case it is assigned the mixture phase) or if it overlaps with other sources (in which case the GL algorithm is locally applied). Finally, consistent Wiener filtering [10] outperforms the previous approaches, although it is computationally costly.

D. Sinusoidal models

Consistency-based approaches rely on properties of the complex representation itself: it is based on overlapping time frames, which introduces an amount of redundancies from one frame to another. Conversely, other techniques take the phase relationships induced by the characteristics of the signals into

account. For instance, the sinusoidal model of MacAulay and Quatiery [13] has been widely used in the literature. It is exploited in the phase vocoder algorithm [14] and it has also been popular in the speech enhancement community [15], [22], [29], where the sinusoids are assumed to be in a harmonic relationship. The fundamental frequency is estimated by means of the PEFAC algorithm [30], which is only suitable for harmonic mixtures. Then the estimation error is propagated and amplified through partials and time frames.

We proposed in [19] a generalization of this approach in order to extend its validity to non-stationary and non-harmonic signals, while avoiding propagating the frequency estimation error over partials and time frames. Besides, it did not require any prior knowledge (such as the numbers of partials) about the components other than a magnitude estimate. In this paper, we propose a detailed description of this algorithm and an extensive experimental evaluation that complete the preliminary work [19]. In particular, we assess the potential of our method when the magnitude spectrograms are no longer equal to the ground truth. In addition, a novel source separation framework is introduced, in which this PU technique is exploited, and it is compared to the consistent Wiener filtering [10].

However, this paper does not address the problem of onset phase reconstruction. Indeed, as it will be shown in the next Section, the PU algorithm relies on a recursive relation between adjacent time frames, therefore it must be initialized within onset frames with another technique. The interested reader can refer to several other papers (e.g. [19], [31]) that address this issue.

III. The phase unwrapping algorithm

In this section, we detail the sinusoidal model that leads to the PU algorithm.

A. Sinusoidal modeling

Let us consider a sinusoid of normalized frequency $\nu_0 \in]-\frac{1}{2};\frac{1}{2}]$, initial phase $\phi_0 \in]-\pi;\pi]$ and amplitude $A_0 > 0$:

$$\forall n \in \mathbb{Z},\ x(n) = A_0 e^{2i\pi\nu_0 n + i\phi_0}. \tag{3}$$

The expression of the STFT is, for each frequency channel $f \in \{0,...,F-1\}$ (with F the number of frequency channels) and time frame $t \in \mathbb{Z}$:

$$X(f,t) = \sum_{n=0}^{N_w-1} x(n+tS)w(n)e^{-2i\pi\frac{f}{F}n}, \tag{4}$$

where w is an N_w sample-long analysis window and S is the time shift (in samples) between successive frames. For every normalized frequency $\nu \in]-\frac{1}{2};\frac{1}{2}]$, let $W(\nu) = \sum_{n=0}^{N_w-1} w(n)e^{-2i\pi\nu n}$ be the discrete time Fourier transform of the analysis window. Then the STFT of the sinusoid (3) is:

$$X(f,t) = A_0 e^{2i\pi\nu_0 St + i\phi_0} W\left(\frac{f}{F} - \nu_0\right). \tag{5}$$

The phase of the STFT is then:

$$\phi(f,t) = \angle X(f,t) = \phi_0 + 2\pi S\nu_0 t + \angle W\left(\frac{f}{F} - \nu_0\right), \tag{6}$$

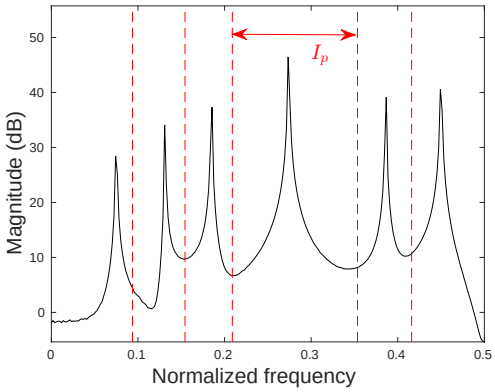

Fig. 3. Example of a spectrum (solid line) decomposed into regions of influence (dashed lines).

where $\angle$ denotes the complex argument. This leads to a relationship between two successive time frames:

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu_0. \tag{7}$$

B. Mixtures of sinusoids

When the signal x is a mixture of P sinusoids, (5) becomes:

$$X(f,t) = \sum_{p=1}^{P} A_p e^{2i\pi\nu_p St + i\phi_{p,0}} W\left(\frac{f}{F} - \nu_p\right). \tag{8}$$

We now assume that there is at most one active sinusoid per frequency channel and per source. Drawing on [14], we propose to decompose the whole frequency range into several regions called *regions of influence*. A region of influence $I_p \subset \{0,...,F-1\}$ corresponds to the set of frequency channels where the STFT X is mainly determined by the p-th sinusoidal partial (i.e. the contributions of the other partials are negligible). Within a time frame t, we consider the magnitude spectrum $v(f) = |X(f,t)|$. The frequency channels corresponding to the peaks of v are denoted f_p. We define the boundaries of the regions of influence as follows:

$$\forall p \in \{2,...,P\},\ l_p = \frac{v(f_p)f_{p-1} + v(f_{p-1})f_p}{v(f_p) + v(f_{p-1})}, \tag{9}$$

and $l_1 = 0$, $l_{P+1} = F$. Thus, the p-th region of influence is:

$$I_p = \{l_p,...,l_{p+1} - 1\}. \tag{10}$$

Such a definition ensures that the set of regions of influence forms a partition of the whole frequency range (they are pair-wise disjoint and they cover the whole interval $\{0,...,F-1\}$), as illustrated in Fig. 3. Note that other definitions of regions of influence exist, such as choosing their boundaries as the channels of lowest energy between the peaks [14].

Now, if we consider a frequency channel in the p-th region of influence, (8) becomes:

$$\forall f \in I_p,\ X(f,t) = A_p e^{2i\pi\nu_p St + i\phi_{p,0}} W\left(\frac{f}{F} - \nu_p\right), \tag{11}$$

which leads to:

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu_p. \tag{12}$$

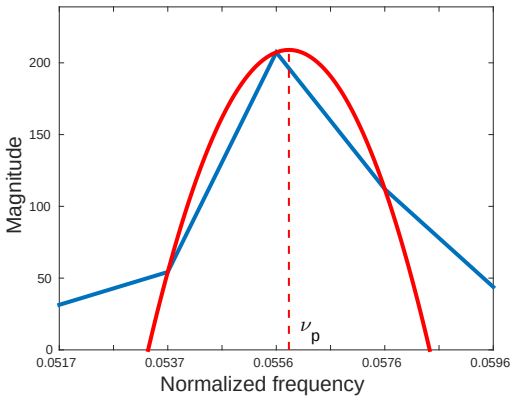

Fig. 4. Illustration of the QIFFT technique: a magnitude peak is approximated by a parabola, whose maximum leads to the frequency estimate.

We can then generalize this so-called *phase unwrapping* equation as follows:

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f), \tag{13}$$

where $\forall p \in \{1,...,P\}$, $\forall f \in I_p$, $\nu(f) = \nu_p$.

C. Slowly-varying sinusoids

More generally, we can compute the phase of the STFT of a frequency-modulated sinusoid. If the frequency variation is small between two successive time frames, we can generalize the previous equation, as demonstrated in [15] :

$$\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f,t). \tag{14}$$

The frequency must then be estimated at each time frame to encompass variable frequency signals such as vibratos, which commonly occur in music signals (singing voice or cello signals for instance).

D. Frequency estimation

In order to apply the PU equation (14), one needs to estimate the frequencies $\nu(f,t)$. Most frequency estimation techniques in the TF domain require the phase of the STFT. For instance, the phase vocoder algorithm [14] uses the phase difference between adjacent TF bins to estimate the frequency. Since our goal is rather to reconstruct the phase of the STFT, we have chosen to use a technique that requires only the magnitude: the Quadratic Interpolated FFT (QIFFT) [32], which is a powerful tool for estimating the frequency near a magnitude peak in the spectrum. It consists in approximating the shape of a spectrum near a magnitude peak by a parabola. This parabolic approximation is justified theoretically for Gaussian analysis windows, and used in practical applications for any window type. The computation of the maximum of the parabola leads to the frequency estimate, as illustrated in Fig. 4. Note that this technique is suitable for signals where only one sinusoid per source is active per frequency channel.

The frequency bias of this method can be reduced by increasing the zero-padding factor [33]. For a Hann window without zero-padding, the frequency estimation error is less than 1 %, which is hardly perceptible in most music applications according to the authors. Note that alternative frequency estimation techniques exist, such as the harmonic spectral sum or product, or more sophisticated versions of those methods (such as the PEFAC algorithm [30]), but these methods are restricted to harmonic mixtures.

Algorithm 1 Phase unwrapping

Inputs:
Magnitude spectrogram $V \in \mathbb{R}_+^{F\times T}$,
Onset frames t_m, $\forall m \in \{0,...,M\}$,
Onset phases $\phi(f,t_m)$, $\forall m \in \{0,...,M-1\}$.
for $m = 0$ to $M-1$ **do**
  for $t = t_m + 1$ to $t_{m+1} - 1$ **do**
    Compute $v(f) = V(f,t)$.
    Peak localization f_p from $v(f)$.
    Frequencies ν_p with QIFFT on f_p.
    Regions of influence I_p from (10),
    $\forall f \in I_p$, $\nu(f,t) = \nu_p$.
    Phase unwrapping $\phi(f,t) = \phi(f,t-1) + 2\pi S\nu(f,t)$.
  end for
end for
Outputs: $\phi \in \mathbb{R}^{F\times T}$

E. The phase unwrapping algorithm

Algorithm 1 describes the PU procedure. Note that the algorithm only reconstructs the phase within non-onset frames. The onset frames can be computed with the Tempogram toolbox [34] for instance, since it estimates the onsets in order to find the tempo. Then, the phases in onset frames must be estimated with another approach. For instance, in the source separation framework, the mixture phase can be given to each component within onset frames. Finally, we tracked the peaks f_p from the spectra v by using the corresponding MATLAB function (`findpeaks`).

IV. Source separation procedure

In this section, we introduce a source separation procedure that exploits the PU algorithm.

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: $X = \sum_k X_k$, and we assume that the magnitudes V_k of the components are fixed (either known or estimated beforehand). We address this problem by minimizing the cost function

$$\mathcal{C}(\theta) = \sum_{f,t} |E(f,t)|^2, \tag{15}$$

under the constraint $|\hat{X}_k| = V_k$, with $E(f,t) = X(f,t) - \sum_k \hat{X}_k(f,t)$ and $\theta = \{\hat{X}_k, k \in \{1,...,K\}\}$. Since all TF bins are treated independently, we remove the indexes (f,t) in what follows for more clarity.

The Wiener filtering estimates (2) are not a solution of this problem since they do not verify $|\hat{X}_k| = V_k$. Thus, we

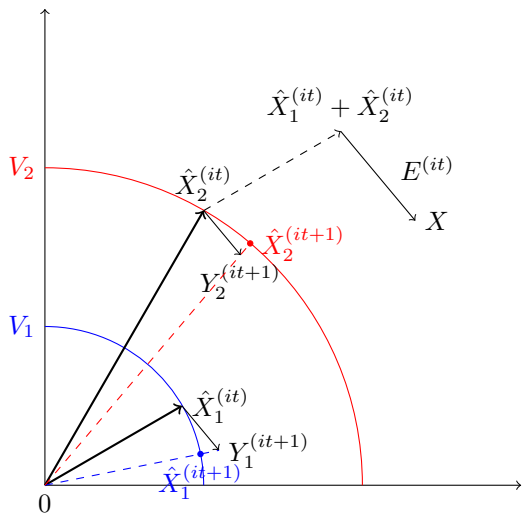

Fig. 5. Iterative estimation of two complex numbers of fixed magnitude and whose sum is known.

introduce an iterative procedure which provides a novel set of estimates of the sources, as motivated in the supporting document [35]. This document also contains all the mathematical aspects related to this procedure.

B. General procedure

The proposed approach is inspired from the work in [7]. At iteration (it), we have an estimate of the complex numbers $\hat{X}_k^{(it)}$. The mixing error $E^{(it)} = X - \sum_k \hat{X}_k^{(it)}$ is distributed over the estimates:

$$Y_k^{(it+1)} = \hat{X}_k^{(it)} + \lambda_k E^{(it)}, \text{ with } \lambda_k = \frac{V_k^2}{\sum_l V_l^2}. \quad (16)$$

As explained in [35], this definition of the weights λ_k is motivated by the fact that the components of highest energy have more impact on the estimation error than the components of lowest energy. Finally, the components $Y_k^{(it+1)}$ are normalized: their magnitude is set equal to the objective values V_k, which leads to the new estimates:

$$X_k^{(it+1)} = \frac{Y_k^{(it+1)}}{|Y_k^{(it+1)}|} V_k. \quad (17)$$

The procedure is illustrated in Fig. 5 for $K = 2$ sources and summarized in Algorithm 2. We provide in [35] the proof that $|E|$ (and by extension the cost function $\mathcal{C}$) is non-increasing under the corresponding update rules.

Even though the procedure is introduced quite intuitively here, it can be properly obtained by using the auxiliary function method. The full derivation of the procedure using this technique can be found in the supporting document [35].

C. The usefulness of phase unwrapping

The keystone of our approach is that it enables us to incorporate some prior phase information about the components through a properly-chosen initialization, as detailed in [35]. 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. However, those initial components would not be modified over iterations, as proved in [35]. Then, we propose to initialize this procedure with the PU algorithm: the corresponding estimates are expected to be close to a local minimum and to have some temporal continuity. Note that this initialization is performed for non-onset frames only. Indeed, as explained in Section II-D, onset phase initialization must be performed with another technique.

Algorithm 2 Estimation of complex components from their mixture

Inputs:
Mixture $X \in \mathbb{C}$, magnitudes $V_k \in \mathbb{R}_+$, weights λ_k, and initial values $\hat{X}_k \in \mathbb{C}$, $\forall k \in \{1, ..., K\}$,
Number of iterations N_{it}.
Compute initial error $E = X - \sum_k \hat{X}_k$.
for $it = 1$ to N_{it} **do**
  for $k = 1$ to K **do**
    $Y_k \leftarrow \hat{X}_k + \lambda_k E$,
    $\hat{X}_k \leftarrow \frac{Y_k}{|Y_k|} V_k$.
  end for
  $E \leftarrow X - \sum_k \hat{X}_k$.
end for
Ouputs: $\forall k \in \{1, ..., K\}$, $\hat{X}_k \in \mathbb{C}$.

V. Experimental validation

We propose here to experimentally assess the potential of the PU algorithm, notably for a source separation task.

A. Datasets

We use several datasets in our experiments:

- A: 30 piano pieces from the Midi Aligned Piano Sounds (MAPS) database [36];
- B: 6 guitar pieces from the IDMT-SMT-GUITAR database [37];
- C: 12 string quartets from the SCore Informed Source Separation DataBase (SCISSDB) [38];
- D: 40 speech excerpts from the Computational Hearing in Multisource Environments (CHiME) database [39];
- E: 50 music songs of various genres from the Demixing Secrets Database (DSD100), a remastered version of the database used for the SiSEC 2015 campaign [40].

The signals are sampled at $F_s = 44100$ Hz and the STFT is computed with a 92 ms long (4096 samples) Hann window, 75 % overlap and no zero-padding.

B. Protocol

Two scenarios are considered: an Oracle scenario, in which the magnitude spectrograms are assumed to be known (i.e. equal to the ground truth), and a more realistic scenario, in which the spectrograms are estimated from the Oracle values by means of an NMF with Kullback-Leibler divergence [4],

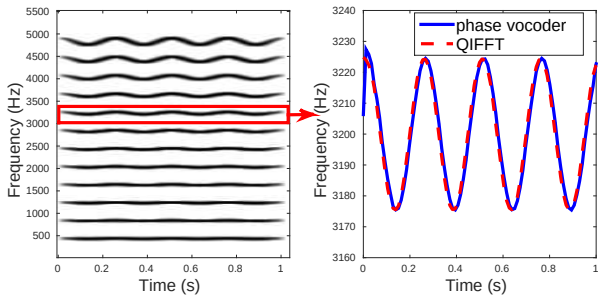

Fig. 6. Spectrogram of a synthetic mixture of sinusoids with vibratos (left) and frequency of the partial oscillating around 3220 Hz (right).

which uses 50 iterations of multiplicative update rules and a rank of factorization of 10. Note that this is not a fully blind scenario, since the NMFs are performed on the isolated spectrograms, but this will inform us about the performance of the methods when dealing with spectrograms that are not equal to the ground truth.

The MATLAB Tempogram Toolbox [34] 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 can be initialized by giving the mixture phase $\angle X(f,t)$ to each component, or alternatively it can be assumed known.

The popular consistency-based Griffin Lim (GL) algorithm [5] is also tested as a reference. We run 200 iterations of this algorithm (performance is not further improved beyond). It is initialized with random values, except in onset frames when it is assumed known.

In order to measure the performance of the methods, we use the BSS EVAL toolbox [41] which computes various energy ratios: the Signal to Distortion, Interference and Artifact Ratios (SDR, SIR and SAR), which are expressed in dB.

Sound excerpts can be found on the companion website for this paper [42] to illustrate the experiments.

C. Frequencies estimation

This experiment aims at assessing the potential of the QIFFT technique in the Oracle scenario. We compute the average frequency error between the phase vocoder [14] estimate ν^*, used as a reference, and the QIFFT estimate ν:

$$\epsilon = \frac{1}{|\Upsilon|} \sum_{(f,t)\in\Upsilon} \frac{|\nu^*(f,t) - \nu(f,t)|}{\nu(f,t)}, \qquad (18)$$

where Υ is the set of TF bins corresponding to the detected magnitude peaks, and $|\Upsilon|$ is the number of elements in Υ. Note that the phase vocoder estimates are not equal to the ground truth: the goal of this experiment is to compare estimates that use either magnitude information only (QIFFT) or magnitude and phase information (phase vocoder).

Fig. 6 illustrates the frequencies estimated with the phase vocoder technique and with our algorithm on a signal which contains some vibratos. The average error is computed on datasets A to D introduced in section V-A and the results are presented in Table I. We observe that the two frequency estimates are very similar. It shows that not accounting for phase information for performing frequency estimation does not lead to results that significantly differ from a phase-aware approach. Even if the phase vocoder estimate does not correspond to the true frequency, it has been shown quite accurate for this task [43]. We can then consider that the QIFFT method provides good estimates of the frequencies, and we will measure its impact on the phase reconstruction in the next experiment.

TABLE I
AVERAGE ERROR BETWEEN QIFFT AND PHASE VOCODER FREQUENCY ESTIMATES.

Dataset	A	B	C	D
Error ϵ (%)	0.48	0.62	0.58	0.35

TABLE II
RECONSTRUCTION PERFORMANCE FOR SEVERAL PHASE RECONSTRUCTION METHODS (SDR IN DB).

Dataset	Oracle		Non-Oracle	
	GL	PU	GL	PU
A	0.4	**5.8**	−0.2	**4.7**
B	−0.5	**2.2**	−11.2	**−9.7**
C	−6.5	**0.4**	−8.9	**−4.7**
D	**1.1**	−1.8	−11.8	**−11.6**

D. Griffin Lim vs Phase Unwrapping

The aim of this experiment is to compare the performance of the GL and PU algorithms. The onset phases are assumed known. We corrupt the complex STFT of the signals by setting the phases within non-onset frames at random values taken in $]-\pi;\pi]$. We then apply the algorithms in both Oracle and non-Oracle scenarios. The results are presented in Table II.

In the Oracle scenario, our approach significantly outperforms the traditional GL method on most datasets: both stationary and variable frequency signals are reconstructed accurately. In the non-Oracle scenario, the PU algorithm outperforms the GL algorithm on all datasets. 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 non-Oracle scenario. However, when the spectrogram is not longer equal to the ground truth, our approach still provide better results than the consistency-based GL algorithm.

For each mixture in dataset B in the Oracle scenario and for both GL and PU algorithms, we measured the inconsistency of the estimates as defined by (1). The GL algorithm leads to an average inconsistency of 2×10^2 vs 1×10^5 for the PU algorithm. Thus, the STFTs reconstructed with the GL algorithm are significantly more consistent than the PU estimates, though they lead to poor results in terms of SDR. This suggests that the direct optimization of an inconsistency criterion may not be the most appropriate way of accounting for this property.

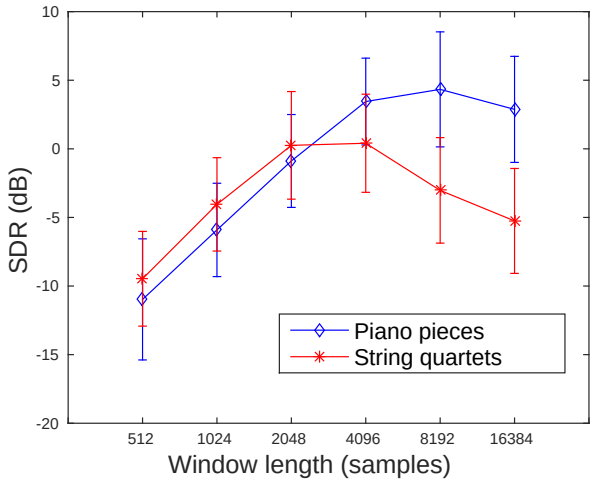

Fig. 7. Influence of the analysis window length on reconstruction quality for datasets A and C. The central marks (resp. the whiskers) represent the mean value (resp. the standard deviation).

E. Influence of the STFT parameters

We evaluate here the influence of some of the STFT parameters on the PU algorithm performance in the Oracle scenario. We have first investigated the influence of the window type and overlap ratio. A comparison between three analysis windows (Hann, Hamming and Blackman) showed no significant difference in terms of SDR over our datasets. In addition, overlap ratios higher than 75 % did not improve the results, while they were more time-consuming than this value. For those reasons, we chose a Hann window with 75 % overlap in our experiments. We propose to analyze how the analysis window length N_w impacts the PU algorithm performance for datasets A and C. The results are presented in Fig. 7. We observe that the window length has a great impact on the SDR. Perceptually (sound examples are available in [42]), two phenomena characterize the reconstructed signals:

- Musical noise, which appears when the analysis window is short. With such windows, the frequency resolution is low, thus the frequency is poorly estimated, which leads to audible artifacts.
- Loss of presence, a phenomenon also known as *phasiness* or *reverberation*, that is a challenging issue in the phase vocoder algorithm [14]. For long analysis windows, the temporal resolution is low. Then, the PU algorithm is not able to retrieve the phase of transients.

Intuitively, one can assume that the observed SDR peak corresponds to a compromise between those phenomena. However, 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 a value different from this optimum leads to more satisfactorily sounding results.

One possible way to overcome this issue can be to use zero-padding with a short analysis window, since the zero-padding increases the frequency precision (even if the resolution is not modified). One can expect this could refine the frequency estimation, and then reduce musical noise. We will not detail the experiment here due to a space constraint, but the conclusion is that the benefit of this method is not as significant as expected, while it is computationally demanding. Alternatively, we could treat differently onset and non-onset frames in order to preserve transients' phase coherence, as proposed in some improved versions of the phase vocoder algorithm [44]. More generally, a multiple resolution framework could overcome the issue of looking for a compromise between temporal and frequency resolution, although those approaches are outside the scope of this paper.

TABLE III
SOURCE SEPARATION PERFORMANCE (SDR, SIR AND SAR IN DB) FOR VARIOUS INITIALIZATIONS ON DATASET E.

Initialization	SDR	SIR	SAR
Random	10.4	20.6	10.9
Unwrapping	**14.0**	**27.0**	**14.2**

F. Application to source separation

Lastly, we assess the potential of the source separation procedure introduced in Section IV. We consider the songs from the dataset E. They are made up of $K = 4$ sources: `bass`, `drums`, `vocals` and `other` (which may contain various instruments such as guitar, piano...).

1) Influence of the initialization: Firstly, we investigate the influence of the initialization in Algorithm 2 on the separation quality. We consider 10 songs from the dataset E in the Oracle scenario. The onset phases are assumed to be known for all sources and the partial phases are estimated by means of 10 iterations of Algorithm 2. It is initialized with random values or alternatively with the PU algorithm. The results of this experiment are provided in Table III.

The initialization with the PU algorithm significantly improves the results (approx 3.5 dB in SDR and SAR and 6.5 in SIR) over a random initialization. To illustrate this result, we consider a mixture composed of two piano notes (C4 and G4) overlapping in the TF domain. We plot the error $|E|$ in a TF bin where the sources overlap in Fig. 8. We see that the PU initialization leads to a better and faster convergence (in terms of error) than a random initialization. As illustrated in Fig. 9, this initialization technique also leads to reconstruct components that better fit the original signal. This confirms the usefulness of the PU algorithm as motivated in Section IV-C.

2) 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 [10][1], the PU algorithm applied to each separated source without accounting for the mixture phase, and 10 iterations of Algorithm 2 initialized with the PU technique. Those methods will be respectively denoted **Wiener**, **Cons-W**, **Unwrap** and **Iter**. These methods are tested on the 50 songs composing the dataset E. The results are represented with box-plots in Fig. 10.

[1]Note that the consistent Wiener filtering technique depends on a weight parameter that promotes the consistency constraint, which is learned beforehand on 50 other songs from dataset E.

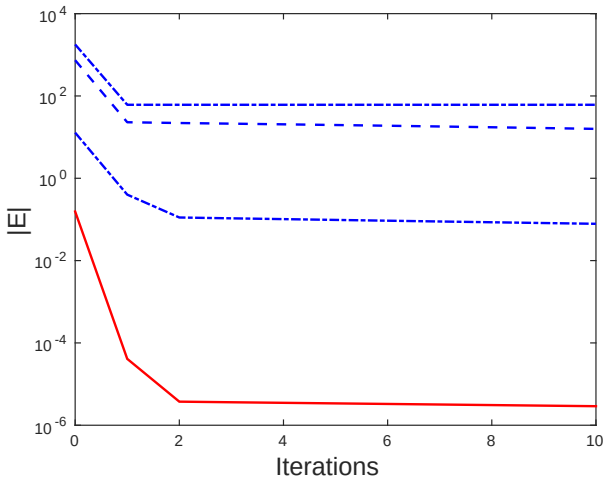

Fig. 8. Error $|E|$ over iterations within a TF bin where two piano notes (C4 and G4) overlap. The dashed lines correspond to the maximum, minimum and average values over 30 random initializations, and the solid line corresponds to the initialization with the PU algorithm.

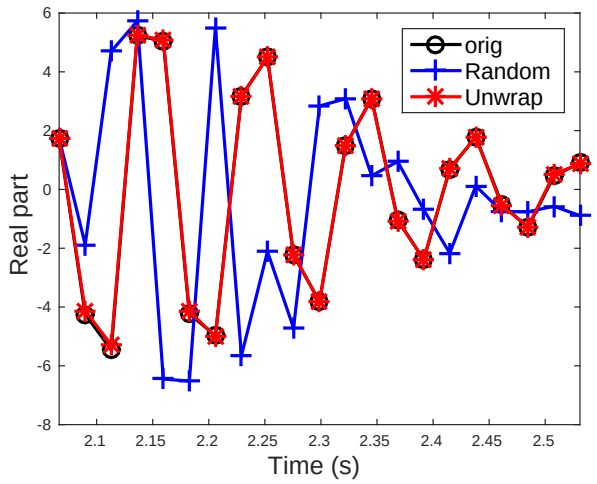

Fig. 9. Real part of a C4 piano note STFT in the 796 Hz frequency channel, when the sources (C4 and G4) overlap, for random and PU initializations of Algorithm 2.

In the Oracle scenario, **Iter** leads to a similar performance than **Cons-W** in terms of SDR and SAR, and better results in terms of interference rejection. This method, however, leads to worse results than **Wiener** and **Cons-W** in terms of SDR and SAR in the non-Oracle case. This is explained by the fact that this method imposes the magnitude of the reconstructed components to be equal to a target value that is no longer the ground truth. The isolated unwrapping **Unwrap** tends to decrease the performance over the traditional methods in both scenarios: since this technique does not use the mixture phase, the PU errors are propagated over time frames, while **Iter** uses the phase of the mixture, which reduces this error. We illustrate these results on a simple example. Let us consider a mixture composed of two piano notes (C4 and G4) overlapping in the TF domain. In Fig. 11, we plot the real part of a partial of the C4 note reconstructed with various methods. In particular,

TABLE IV
IMPACT OF THE ONSET PHASE ESTIMATION ON THE **ITER** PERFORMANCE ON DATASET E.

Onset phase	SDR	SIR	SAR
Mixture	11.2	22.3	11.7
Known	**13.2**	**25.4**	**13.6**

the **Iter** estimate better fits the ground truth than the other methods. We noted that property on a variety of songs from dataset E in both Oracle and non-Oracle settings. In addition, we perceptually observed (sounds excerpts available in [42]) that the **Iter** method tends to reduce the artifacts in the `bass` track compared to the **Cons-W** technique.

Finally, it is important to note that **Cons-W** is computationally costly: for a 10 seconds excerpt, the separation is performed in 26 seconds with **Cons-W** vs 4 seconds with our method. The proposed approach then appears appealing for an efficient audio source separation task, notably in terms of interference rejection.

3) Onset phase: Finally, we propose to evaluate the room for improvement of onset phase recovery. We run the **Iter** procedure (which uses 10 iterations) in the Oracle scenario considering two different settings: onset phases can be either assumed known or estimated by giving the mixture phase to each component (as in the previous experiment). From the results in Table IV we remark that there is a gap in terms of both SDR and SAR ($\approx$ 2 dB) and SIR ($\approx$ 3 dB) between the two considered settings. This means that onset phase reconstruction needs to be improved in order to fully exploit the potential of the PU technique. Though giving the mixture phase to each component is fast and easy to implement, there is a significant room for further enhancement of the onset phase estimation method.

VI. CONCLUSION

The PU technique introduced in this paper is a promising and efficient method for recovering the STFT phase of audio signals. The analysis of mixtures of sinusoids leads to a relationship between the phases of successive TF bins. Frequencies are estimated on a frame-by-frame basis, encompassing a variety of signals such as piano and cello sounds. The phase is then unwrapped over time frames, ensuring some form of temporal continuity. Experiments have demonstrated the accuracy of this method and investigated the impact of several parameters on the reconstruction quality, which allows us to propose an optimal tuning of the STFT parameters. The PU algorithm has also been integrated into a source separation framework. The experimental results show that such a procedure yields better results than state-of-the-art approaches in a scenario in which the magnitude spectra are known. In a more realistic scenario, it reached a performance similar to other methods, with a significant improvement in terms of computational cost.

As suggested by the last experiment, the reconstruction of onset frames can be an interesting research direction for an improved sounding quality. For instance, onsets can be

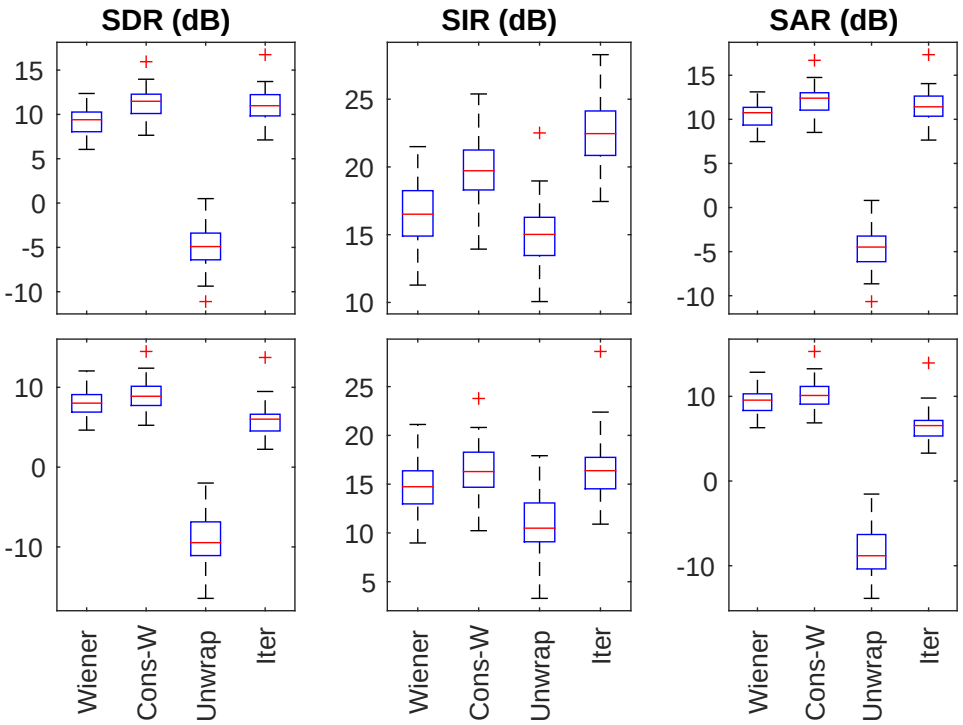

Fig. 10. Source separation performance (SDR, SIR and SAR in dB) of various methods on dataset E for the Oracle (top) and non-Oracle (bottom) scenarios. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles, whiskers indicating the minimum and maximum values, and crosses representing the outliers.

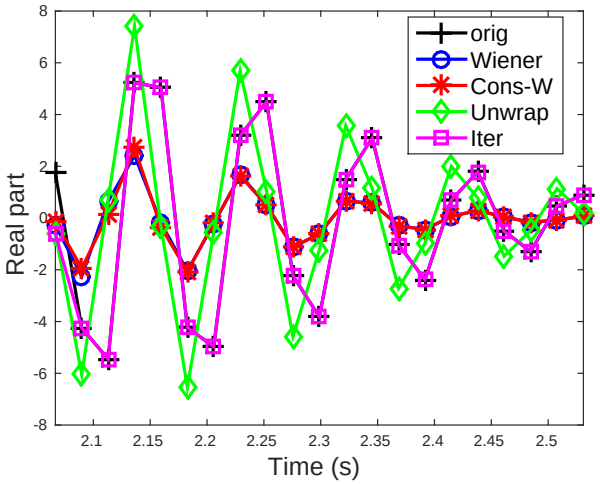

Fig. 11. Real part of a C4 piano note STFT in the 796 Hz frequency channel, when the sources (C4 and G4) overlap. Several reconstruction methods are compared in the Oracle scenario.

represented by an impulse model, which has applications in transient detection [45] and phase reconstruction [19]. One can also use a model of repeated audio event for modeling the phase within onset frames [31]. Alternatively, time-invariant parameters such as phase offsets between partials [46] can be used. In addition, frequency estimation from magnitude spectra can be refined, as the STFT inherently comes with a limited frequency resolution. Finally, future work can focus on exploiting known phase data for reconstruction: missing bins can be inferred from observed phase values, exploiting structured models such as Markov chains or autoregressive modeling in the TF domain [47].

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2003.

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

[5] 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.

[6] 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)*, Brisbane, Australia, April 2015, pp. 81–85.

[7] 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.

[8] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012.

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

[10] 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.

[11] 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.

[12] 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, phase-Aware Signal Processing in Speech Communication.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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)*, Florence, Italy, May 2014.

[18] 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)*, Taipei, Taiwan, April 2009.

[19] 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)*, Nice, France, August 2015.

[20] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[21] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[22] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement - Unimportant, important, or impossible?" in *IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, November 2012, pp. 1–5.

[23] 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)*, Brisbane, Australia, September 2008, pp. 23–28.

[24] 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.

[25] 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)*, New Paltz, NY, USA, October 2013, pp. 1–4.

[26] G. T. Beauregard, M. Harish, and L. L. Wyse, "Single Pass Spectrogram Inversion," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 427–431.

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

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[29] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1265–1277, July 2011.

[30] 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.

[31] 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)*, New Paltz, NY, USA, October 2015.

[32] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Audio Engineering Society Convention 117*. Berlin, Germany: Audio Engineering Society, May 2004.

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

[34] 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*, Miami, FL, USA, October 2011.

[35] P. Magron, R. Badeau, and B. David, "An iterative algorithm for recovering the phase of complex components from their mixture," Paris, France, Tech. Rep. hal-01325625, June 2016.

[36] 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.

[37] C. Kehling, A. Jakob, D. Christian, and S. Gerald, "Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters," in *Proc. of the International Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, September 2014.

[38] R. Hennequin, R. Badeau, and B. David, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 45–48.

[39] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, Feb. 2013.

[40] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 387–395.

[41] 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.

[42] "Demo Webpage," "http://www.cs.tut.fi/~magron/demos/demo_TASLP2017.html".

[43] 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.

[44] A. Röbel, "A new approach to transient processing in the phase vocoder," in *6th International Conference on Digital Audio Effects (DAFx)*, London, United Kingdom, September 2003, pp. 344–349.

[45] 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)*, New Paltz, NY, USA, October 2013, pp. 1–4.

[46] H. Kirchhoff, R. Badeau, and S. Dixon, "Towards complex matrix decomposition of spectrogram based on the relative phase offsets of harmonic sounds," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[47] R. Badeau and M. D. Plumbley, "Multichannel High resolution NMF for modelling Convolutive Mixtures of Non-Stationary signals in the time-frequency domain," *IEEE Transactions on Audio Speech and Language Processing*, vol. 22, no. 11, pp. 1670–1680, November 2014.