Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration

Paul Magron, Roland Badeau, Bertrand David

Phase reconstruction of spectrograms with linear unwrapping : application to audio signal restoration

Reconstruction de phases de spectrogrammes par déroulé linéaire : application à la restauration de signaux audio

---

Paul Magron,
Roland Badeau
Bertrand David

2015D002

avril 2015

Département Traitement du Signal et des Images
Groupe AAO : Audio, Acoustique et Ondes

Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration

Reconstruction de phases de spectrogrammes par déroulé linéaire : application à la restauration de signaux audio

Paul Magron Roland Badeau Bertrand David
Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France
<firstname>.<lastname>@telecom-paristech.fr *

Abstract

This paper introduces a novel technique for reconstructing the phase of modified spectrograms of audio signals. From the analysis of mixtures of sinusoids we obtain relationships between phases of successive time frames in the Time-Frequency (TF) domain. To obtain similar relationships over frequencies, in particular within onset frames, we study an impulse model. Instantaneous frequencies and attack times are estimated locally to encompass the class of non-stationary signals such as vibratos. These techniques ensure both the vertical coherence of partials (over frequencies) and the horizontal coherence (over time). The method is tested on a variety of data and demonstrates better performance than traditional consistency-based approaches. We also introduce an audio restoration framework and observe that our technique outperforms traditional methods.

Key words

Phase reconstruction, sinusoidal modeling, linear unwrapping, phase consistency, audio restoration.

Résumé

Ce rapport présente une nouvelle technique pour la reconstruction de phases de spectrogrammes modifiés. A partir de l'analyse de mélanges de sinusoïdes, on obtient des relations entre les phases trames successives dans le plan temps-fréquence (TF). Pour obtenir des relations similaires entre fréquences, en particulier au sein des trames d'attaque, nous étudions un modèle d'impulsion. Les fréquences instantanées et les temps d'attaque sont estimés localement afin de pouvoir représenter des signaux non stationnaires, tels que les vibratos. Ces techniques permettent d'assurer à la fois une cohérence verticale entre les partiels (à travers les fréquences) et horizontale (au cours du temps). Cette méthode est testée sur des données expérimentales, et montre de meilleurs résultats que l'approche traditionnelle basée sur la consistance. Nous proposons également d'introduire cette technique dans un contexte de restauration de signaux audio, dans lequel une meilleure performance qu'avec les méthodes traditionnelles est observée.

Mots clés

Reconstruction de phase, mélanges de sinusoïdes, déeroulé linéaire, consistance de phase, restauration de signaux audio

*This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02).

1 Introduction

A variety of music signal processing techniques act in the TF domain, exploiting the particular structure of music signals. For instance, the family of techniques based on Nonnegative Matrix Factorization (NMF) is often applied to spectrogram-like representations, and has proved to provide a successful and promising framework for source separation [1]. Magnitude-recovery techniques are also useful for restoring missing data in corrupted signals [2].

However, when it comes to resynthesize time signals, the phase recovery of the corresponding Short-Time Fourier Transform (STFT) is necessary. In the source separation framework, a common practice consists in applying Wiener-like filtering (soft masking of the complex-valued STFT of the original mixture). When there is no prior on the phase of a component (*e.g.* in the context of audio restoration), a consistency-based approach is often used for phase recovery [3]. That is, a complex-valued matrix is iteratively computed to be close to the STFT of a time signal. A recent benchmark has been conducted to assess the potential of source separation methods with phase recovery in NMF [4]. It points out 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. Thus, phase recovery of modified audio spectrograms is still an open issue. The High Resolution NMF (HRNMF) model [5] has shown to be a promising approach, since it models a TF mixture as a sum of autoregressive (AR) components in the TF domain, thus dealing explicitly with a phase model.

Another approach to reconstruct the phase of a spectrogram is to use a phase model based on the observation of fundamental signals that are mixtures of sinusoids. Contrary to consistency-based approaches using the redundancy of the STFT, this model exploits the natural relationship between adjacent TF bins due to the model. This approach is used in the phase vocoder algorithm [6], although it is mainly dedicated to time stretching and pitch modification of signals, and it requires the phase of the original STFT. More recently, [7] proposed a complex NMF framework with phase constraints based on sinusoidal modeling. Although promising, this approach is limited to harmonic and stationary signals, and requires prior knowledge on fundamental frequencies and numbers of partials.

In this paper, we propose a generalization of this approach that consists in estimating the phase field of mixtures of sinusoids from its explicit calculation. We then obtain an algorithm which unwraps the phases *horizontally* (over time frames) to ensure the temporal coherence of the signal, and *vertically* (over frequency channels) to enforce spectral coherence between partials, which are naturally observed in musical acoustics. Our technique is suitable for a variety of pitched music signals, such as piano or guitar sounds. A dynamic estimation (at each time frame) of instantaneous frequencies extends the validity of this technique to non-stationary signals such as cellos and speech. This technique is tested on a variety of signals and integrated in an audio restoration framework.

The paper is organized as follows. Section 2 presents the horizontal phase unwrapping model. Section 3 is dedicated to phase reconstruction on onset frames. Section 4 presents a performance evaluation of this technique through various experiments. Section 5 introduces an audio restoration framework using this phase recovery method. Finally, section 6 draws some concluding remarks.

2 Horizontal phase reconstruction

2.1 Sinusoidal modeling

Let us consider a sinusoid of normalized frequency $f_0 \in [-\frac{1}{2}; \frac{1}{2}]$, origin phase $\phi_0 \in [-\pi; \pi]$ and amplitude $A > 0$:

$$\forall n \in \mathbb{Z},\ x(n) = Ae^{2i\pi f_0 n + i\phi_0}. \tag{1}$$

The expression of the STFT is, for each frequency channel $k \in [\![-\frac{F-1}{2}; \frac{F-1}{2}]\!]$ (with F the odd-valued Fourier transform length) and time frame $t \in \mathbb{Z}$:

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

where w is a N sample-long analysis window and S is the time shift (in samples) between successive frames. Let $W(f) = \sum_{n=0}^{N-1} w(n)e^{-2i\pi fn}$ be the discrete time Fourier transform of the analysis window for each normalized frequency $f \in [-\frac{1}{2}; \frac{1}{2}]$. Then the STFT of the sinusoid (1) is:

$$X(k,t) = Ae^{2i\pi f_0 St + i\phi_0} W\left(\frac{k}{F} - f_0\right). \tag{3}$$

The unwrapped phase of the STFT is then:

$$\phi(k,t) = \phi_0 + 2\pi S f_0 t + \angle W\left(\frac{k}{F} - f_0\right) \tag{4}$$

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

$$\phi(k,t) = \phi(k,t-1) + 2\pi S f_0. \tag{5}$$

More generally, we can compute the phase of the STFT of a frequency-modulated sinusoid. If the frequency variation is low between two successive time frames, we can generalize the previous equation:

$$\phi(k,t) = \phi(k,t-1) + 2\pi S f_0(t). \tag{6}$$

Instantaneous 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).

2.2 Instantaneous frequency estimation

Quadratic interpolation FFT (QIFFT) is a powerful tool for estimating the instantaneous frequency near a magnitude peak in the spectrum [8]. 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 instantaneous frequency estimate. Note that this technique is suitable for signals where only one sinusoid is active per frequency channel.

The frequency bias of this method can be reduced by increasing the zero-padding factor [9]. 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.

2.3 Regions of influence

When the mixture is composed of one sinusoid, the phase must be unwrapped in all frequency channels according to (5) using the instantaneous frequency f_0. When there is more than one sinusoid, frequency estimation is performed near each magnitude peak. Then, the whole frequency range must be decomposed in several regions (*regions of influence* [6]) to ensure that the phase in a given frequency channel is unwrapped with the appropriate instantaneous frequency.

At time frame t, we consider a magnitude peak A_p in channel k_p. The magnitudes (resp. the frequency channels) of neighboring peaks are denoted A_{p-1} and A_{p+1} (resp. k_{p-1} and k_{p+1}). We define the region of influence I_p of the p-th peak as follows:

$$I_p = \left[\frac{A_p k_{p-1} + A_{p-1} k_p}{A_p + A_{p-1}}; \frac{A_p k_{p+1} + A_{p+1} k_p}{A_p + A_{p+1}}\right]. \tag{7}$$

The greater A_p is relatively to A_{p-1} and A_{p+1}, the wider I_p is. Note that other definitions of regions of influence exist, such as choosing the limit between two peaks as the channel of lowest energy [6].

3 Onset phase reconstruction

3.1 Impulse model

Impulse signals are useful to obtain a relationship between phases over frequencies (vertical unwrapping) [10]. Although they do not accurately model attack sounds, they provide simple equations that can be further improved for more complex signals. The model is:

$$\forall n \in \mathbb{Z},\ x(n) = A\delta_{n-n_0} \tag{8}$$

where δ is equal to one if $n = n_0$ (the so-called *attack time*) and zero elsewhere and $A > 0$ is the amplitude. Its STFT is equal to zero except within attack frames:

$$X(k,t) = Aw(n_0 - St)e^{-2i\pi\frac{k}{F}(n_0 - St)}. \tag{9}$$

We can then obtain a relationship between the phases of two successive frequency channels within an onset frame, assuming that $w \geq 0$:

$$\phi(k,t) = \phi(k-1,t) - \frac{2\pi}{F}(n_0 - St). \tag{10}$$

The similarity between (10) and (5) was expected because the impulse is the dual of the sinusoid in the TF domain. This comparison naturally leads to estimating parameter n_0 (the "instantaneous" attack time) in each frequency channel as we previously estimated f_0 (the instantaneous frequency) in each time frame (cf. equation (6)). This leads to the following vertical unwrapping equation:

$$\phi(k,t) = \phi(k-1,t) - \frac{2\pi}{F}(n_0(k) - St). \tag{11}$$

3.2 Attack time estimation

In order to estimate $n_0(k)$, we look at the magnitude of the STFT of the impulse in a frequency channel k:

$$|X(k,t)| = Aw(n_0(k) - St). \tag{12}$$

We then choose n_0 such that the STFT magnitude of the impulse over onset frames has a shape similar to that of the analysis window. For instance, a least-squares estimation method can be used. Synthetic mixtures of impulses are perfectly reconstructed with this technique. Alternatively, we can also estimate $n_0(k)$ with a temporal QIFFT and update the phase with (11).

4 Experimental evaluation

4.1 Protocol and datasets

The MATLAB Tempogram Toolbox [11] provides a fast and reliable onset frames detection from spectrograms. We use several datasets in our experiments:

A: 30 mixtures of piano notes from the Midi Aligned Piano Sounds (MAPS) database [12],

B: 30 piano pieces from the MAPS database,

C: 12 string quartets from the SCore Informed Source Separation DataBase (SCISSDB) [13],

D: 40 speech excerpts from the Computational Hearing in Multisource Environments (CHiME) database [14].

The data is sampled at $F_s = 11025$ Hz and the STFT is computed with a 512 sample-long Hann window and 75 % overlap. The Signal to Distortion Ratio (SDR) is used for performance measurement. It is computed with the **BSS Eval** toolbox [15] and expressed in dB. The popular consistency-based Griffin and Lim (GL) algorithm [3] is also used as a reference. We run 200 iterations of this algorithm (performance is not further improved beyond). It is initialized with known phase values if any or random values if not, and results are averaged on 30 initializations. Simulations are run on a 3.60GHz CPU processor and 16Go RAM computer.

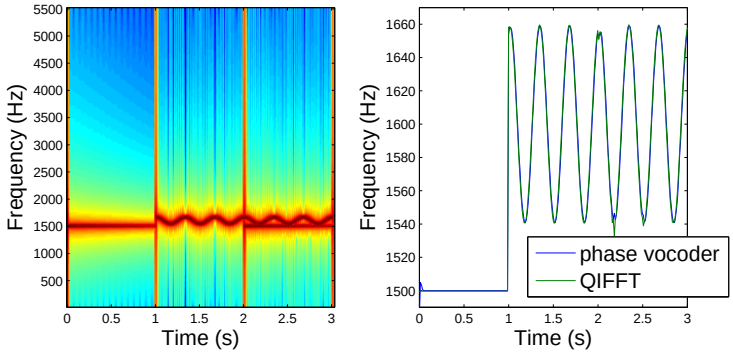

Figure 4.1: Spectrogram of a mixture with vibrato (left) and instantaneous frequencies in the 2800 Hz channel (right)

Dataset	Error	GL	PU
A	0.38	−6.9	**2.5**
B	0.36	−12.6	**1.7**
C	0.41	−9.7	**5.3**
D	0.52	−0.4	**0.5**

Table 1: Frequency estimation error (%) and reconstruction performance (SDR in dB) for various audio datasets

4.2 Horizontal phase reconstruction

A first experiment consists in estimating instantaneous frequencies on synthetic mixtures of damped sinusoids, which parameters (in particular the frequencies) are user-defined. Frequency estimation error with QIFFT is below the threshold of 0.2 %, commonly referred to as the maximal human auditory resolution.

Figure 4.1 illustrates the instantaneous frequencies estimated with the phase vocoder technique [6], used as a reference, and with our algorithm on a vibrato. Identical results are obtained. Our method is thus suitable for estimating variable instantaneous frequency signals as well as stationary components. We computed the average frequency error between phase vocoder and QIFFT estimates for the datasets presented in section 4.1. The results presented in the first column of Table 1 confirm that QIFFT provides an accurate frequency estimation.

Table 1 also presents reconstruction performance (assuming onset phases are known) for both Griffin and Lim (GL) and our Phase Unwrapping (PU) algorithms. Our approach significantly outperforms the traditional GL method: both stationary and variable frequency signals are reconstructed accurately. In addition, our algorithm is faster than the GL technique: on a 3min 48s piano piece, the reconstruction is performed in 18s with our approach and in 623s with GL algorithm.

4.3 Onset phase reconstruction

Onset phases can be reconstructed with n_0-estimation using the impulse magnitude (**PU-Impulse**) or with QIFFT (**PU-QIFFT**). We also test random phases values (**PU-Rand**, no vertical coherence), zero phases (**PU-0**, partials in phase) and alternating partial phases between 0 and π (**PU-Alt**, phase-opposed partials). These choices are justified by the observation of the relationship between partials in musical acoustics [16]. The phase of the partials is then fully recovered with horizontal unwrapping. We test these methods on dataset A. Results presented in Table 2 show that all our approaches provide better results than GL algorithm on this class of signals. Onset phase unwrapping with n_0-estimation based on QIFFT provides the best result, ensuring some form of vertical coherence. In particular, we perceptually observe that this approach provides a neat percussive attack.

Method	SDR (dB)
GL	−7.9
PU-Impulse	−4.0
PU-QIFFT	**−2.6**
PU-Rand	−4.3
PU-0	−4.7
PU-Alt	−3.5

Table 2: Signal reconstruction performance of different methods on dataset A

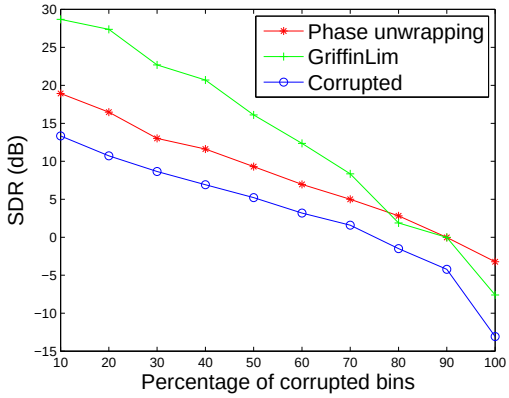

Figure 4.2: Reconstruction performance of different methods and percentages of corruption on dataset A

4.4 Complete phase reconstruction

We consider unaltered magnitude spectrograms from dataset A. A variable percentage of the STFT phases is randomly corrupted. We evaluate the performance of our algorithm to restore the phase both on onset and non-onset frames.

Figure 4.2 confirms the potential of this technique. Our method produced an average increase in SDR of 6dB over the corrupted data. It also performs better than the GL algorithm when a high percentage of the STFT phases must be recovered.

However, note that this experiment consists in phase reconstruction of *consistent* spectrograms (*i.e* positive matrices that are the magnitude of the STFT of a time signal): GL algorithm is then naturally advantaged in this case. Realistic applications (cf. next section) involve the restoration of both phase and magnitude, which leads to inconsistent spectrograms.

The goal of audio inpainting is to restore corrupted or missing value of a signal. Since corruption can be done in the temporal domain or in the TF domain, we will study those two approaches.

First, we will partially recover phase from a consistent spectrogram (the STFT magnitude is not modified). Secondly, we will reconstruct whole parts of STFT (both magnitude and spectrogram). Finally, a time signal is corrupted with clicks and we compare restoration with a temporal method, our approach, and HR NMF algorithm [5].

5 Application of phase reconstruction to audio restoration

%subsectionTemporal corruption: click removal

A common alteration of music signals is the presence of noise on short time periods (a few samples) called clicks. We corrupt time signals with clicks that represent less than 1 % of the total duration. Clicks are obtained by differentiating a 10 sample-long Hann window.

Magnitude restoration of missing bins is performed by linear interpolation of the log-magnitudes in each frequency channel. Figure 5.1 illustrates this technique. Phase recovery is then performed with our method

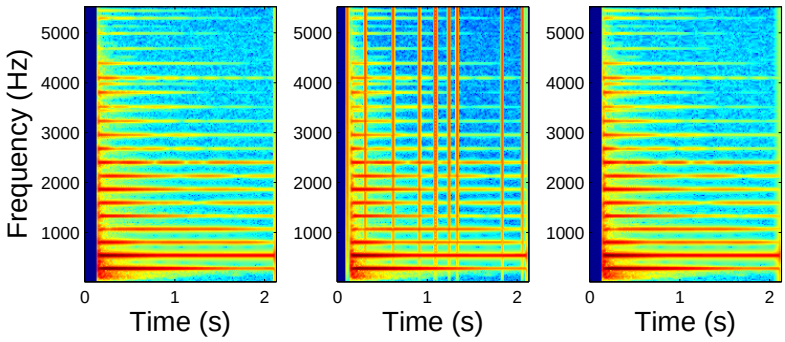

Figure 5.1: Restoration of spectrogram by linear interpolation of the log-magnitudes on a piano notes

Dataset	AR	HRNMF	GL	PU
A	11.4	**16.9**	8.6	**11.7**
B	4.3	**10.9**	5.9	**7.1**
C	8.2	**10.6**	6.6	**7.1**
D	8.3	**10.9**	8.9	**9.4**

Table 3: Signal restoration performance (SDR in dB) for various methods and datasets

(PU) or alternatively with the GL algorithm. We compare those results to the traditional restoration method based on autoregressive (AR) modeling of the time signal [18], and with HRNMF [5].

Table 3 presents results of restoration. HRNMF provides the best results in terms of SDR. Though, our approach outperforms the traditional method and GL algorithm. Besides, we underline that the HRNMF model uses the phase of the non-corrupted bins, while our algorithm is blind. Lastly, our technique remains faster than HRNMF: for a 3min55s piano piece, restoration is performed in 99s with our algorithm and in 222s with HRNMF.

6 Conclusion

The new phase reconstruction technique introduced in this work appears to be an efficient and promising method. The analysis of mixtures of sinusoids leads to relationships between successive TF bins phases. Physical parameters such as instantaneous frequencies and attack times are estimated dynamically, encompassing a variety of signals such as piano and cellos sounds. The phase is then unwrapped in all frequency channels for onset frames and over time for partials. Experiments have demonstrated the accuracy of the unwrapping method, and we integrated it in an audio restoration framework. Better results than with traditional methods have been reached.

The reconstruction of onset frames still needs to be improved as suggested by the variety of data. Further work will focus on exploiting known phase data for reconstruction: missing bins can be inferred from observed phase values. Alternatively, time-invariant parameters such as phase offsets between partials [19] can be used. Such developments will be introduced in an audio source separation framework, where the phase of the mixture can be exploited.

References

[1] Paris Smaragdis and Judith 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.

[2] Derry Fitzgerald and Dan Barry, "On inpainting the adress algorithm," in *Proc. IET Irish Signals and Systems Conference (ISSC)*, Maynooth, Ireland, June 2012, pp. 1–6.

[3] Daniel Griffin and Jae 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.

[4] Paul Magron, Roland Badeau, and Bertrand David, "Phase reconstruction 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.

[5] Roland Badeau and Mark D. Plumbley, "Multichannel high resolution NMF for modelling convolutive mixtures or 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.

[6] Jean Laroche and Mark 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.

[7] James Bronson and Philippe 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.

[8] Mototsugu Abe and Julius 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, May 2004, Audio Engineering Society.

[9] Mototsugu Abe and Julius O. Smith, "Design criteria for the quadratically interpolated FFT method (i): Bias due to interpolation," Tech. Rep. STAN-M-117, Stanford University, Department of Music, 2004.

[10] Akihiko Sugiyama and Ryoji 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.

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

[12] Valentin Emiya, Nancy Bertin, Bertrand David, and Roland Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Tech. Rep. 2010D017, Télécom ParisTech, Paris, France, July 2010.

[13] Romain Hennequin, Roland Badeau, and Bertrand 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.

[14] Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, and Phil Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, Feb. 2013.

[15] Emmanuel. Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[16] Antoine Chaigne and Jean Kergomard, *Acoustique des instruments de musique*, Broché, November 2008.

[17] Vincent D. Blondel, Ngoc-Diep Ho, , and Paul van Dooren, "Algorithms for weighted non-negative matrix factorization," *Image and Vision Computing*, 2007.

[18] Simon J. Godsill and Peter J. W. Rayner, *Digital Audio Restoration - A Statistical Model-Based Approach*, Springer-Verlag, 1998.

[19] Holger Kirchhoff, Roland Badeau, and Simon 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.

Dépôt légal : 2015 – 2e trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut Mines-Télécom - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.fr

Département TSI