TOWARDS COMPLEX MATRIX DECOMPOSITION OF SPECTROGRAMS BASED ON THE RELATIVE PHASE OFFSETS OF HARMONIC SOUNDS

Holger Kirchhoff [*1] *Roland Badeau*[†2] *Simon Dixon*[1]

[1] Centre for Digital Music, Queen Mary University of London, UK
[2] Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France

ABSTRACT

In this paper we study the relative phase offsets between partials in the sustained part of harmonic sounds and investigate their suitability for complex matrix decomposition of spectrograms. We formally introduce this property in a sinusoidal model and visualise the phase relations of a musical instrument. A model of complex matrix decomposition in the time-frequency domain is derived and equations for the estimation of the model parameters are provided in the monophonic case. We illustrate the model with the analysis of a monophonic saxophone signal. The results suggest that the phase offset is able to capture inherent time-invariant phase properties of harmonic sounds and outline its potential use for complex matrix decomposition.

Index Terms— harmonic signals, relative phase offsets of partials, complex matrix decomposition, nonnegative matrix factorisation.

1. INTRODUCTION

Spectrogram factorisation techniques — particularly nonnegative matrix factorisation (NMF) [1] — have proven to be useful methods for the analysis of instrument sounds and have been successfully applied for tasks such as music transcription [2], source separation [3] and instrument recognition [4]. Most spectrogram factorisation techniques rely on the assumption that magnitude spectra of sound mixtures can be approximated by the superposition of the magnitude spectra of the sound sources. Although this assumption provides reasonable analysis results in practice, the linearity only holds for the *complex* coefficients of the short-time Fourier transform (STFT).

Phase information is often discarded in the analysis stage, not only because the human auditory system is considered insensitive to absolute phase shifts of harmonic partials [5], but also because the magnitude spectrogram is often considered more intuitive and easier to model. For all applications in which sounds have to be synthesised from a time-frequency representation, however, the correct estimation of phase values is crucial in order to avoid artefacts due to phase-incoherent overlap of consecutive time frames. For the task of instrument separation, for example, the phase information for the synthesis of each sound source either has to be estimated from the magnitude spectrogram [6], or the phases of the original mixture have to be employed for each source [3]. Using the mixture phases can lead to reasonable results when the number of sources is small and when most time-frequency bins are mainly influenced by a single source. For higher numbers of sources and larger time-frequency overlap, however, it can lead to cross-talk artefacts.

The STFT analysis phase of harmonic partials is not constant over time. Therefore the instantaneous phase cannot be simply embedded in a matrix factorisation framework, where the basis functions include prototypical properties of the underlying instrument spectra that vary little over time. Nevertheless, several approaches have been proposed to consider phase information in matrix factorisation frameworks. Parry and Essa [7] propose a phase-aware nonnegative matrix factorisation. The authors model the STFT bins as complex random variables, thereby assuming the phase to be uniformly distributed, and derive iterative update rules based on this assumption. The update rules, however, still estimate the matrices based on the magnitude spectrogram only. In a similar way, Févotte et al. [8] show that Itakura-Saito NMF is equivalent to a maximum-likelihood parameter estimation of a sum of complex Gaussian components. The Gaussian components have zero mean and a diagonal covariance matrix, which also assumes a uniformly distributed phase. An attempt to explicitly estimate the phase values of the individual sources was made by Kameoka et al. [9]. Their complex NMF algorithm combines the outer product of each NMF basis function and gain vector with a phase spectrogram with the same dimensions as the original spectrogram. In [10], complex NMF was shown to outperform NMF for speech separation. Complex NMF is not a complex matrix factorisation technique, but a combination of NMF with time-frequency phase estimates. The algorithm is heavily overparameterised and it can be shown that an initialisation with the original phase values leaves the phase parameters unaltered (up to $\pm\pi$). Lastly, a high resolution NMF framework was introduced in [11, 12], in order to model both the magnitude and phase of complex or real-valued time-frequency representations. However this framework does not take the phase relations of the partials into account.

In this paper, we exploit the relative phase offsets between partials in the sustained part of the sounds of harmonic instruments as a step towards complex matrix decomposition. The concept will be reviewed and illustrated in Section 2, where we also present a mathematical formulation. In Section 3 we derive the model for complex matrix decomposition and present the parameter estimation equations for the *monophonic* case. An example analysis of a monophonic signal is provided in Section 4, and we conclude this work in Section 5.

2. PHASE RELATIONS OF HARMONIC PARTIALS

2.1. Concept

Pitched musical instruments generally produce harmonic sounds which can be represented by a superposition of P sinusoids at in-

[*]This work was funded by a Queen Mary University of London CDTA studentship.

[†]This work was undertaken while Roland Badeau was visiting the Centre for Digital Music, partly funded by EPSRC Platform Grant EP/K009559/1.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

teger multiples of a fundamental frequency. Each harmonic partial can be described by its angular frequency $\omega_p > 0$, its amplitude $a_p \geq 0$ and an absolute phase shift $\phi_p \in [-\pi, \pi)$: $\forall t \in \mathbb{Z}$,

$$s(t) = \sum_{p=1}^{P} a_p e^{j[\omega_p t + \phi_p]}. \quad (1)$$

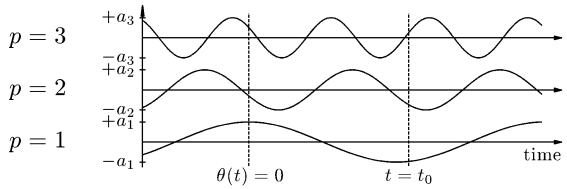
For strictly harmonic sounds, the frequency of each harmonic is given as the p-th multiple of the fundamental frequency: $\omega_p = p\,\omega_1$. Complex exponentials are used here to reflect the fact that we only consider the baseband of the DFT in our model.

In this paper we are interested in the relation between the absolute phase shifts of the harmonic partials. To capture this relation, we express the phase shift of each partial in relation to the instantaneous phase of the fundamental frequency ω_1:

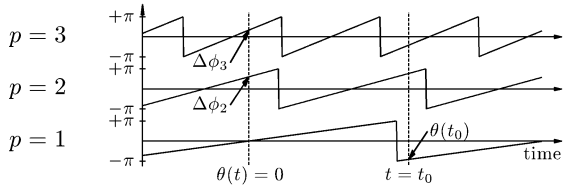
$$s(t) = \sum_{p=1}^{P} a_p e^{j[p\cdot\theta(t) + \Delta\phi_p]}, \quad (2)$$

where $\theta(t) = \omega_1 t + \phi_1$ denotes the instantaneous phase of the fundamental and $\Delta\phi_p = \phi_p - p\,\phi_1$ represents the phase offset between the p-th partial and the fundamental (with $\Delta\phi_1 = 0$).

Figure 1 shows a graphical illustration of the parameters in Eq. (2). The upper part (Fig. 1a) displays the waveform of the first three partials of a harmonic sound, and the lower part (Fig. 1b) the instantaneous phases. The phase offsets $\Delta\phi_p$ correspond to the instantaneous phases of the partials at the time where $\theta(t) = 0$. Modifying $\Delta\phi_p$ translates the p-th partial relative to the fundamental along the time axis. Since a translation by $\Delta\phi_p$ is equivalent to a translation by $\Delta\phi_p + c \cdot 2\pi$ with $c \in \mathbb{Z}$, $\Delta\phi_p$ is uniquely defined in the range $[-\pi, \pi)$. Given all phase offsets $\Delta\phi_p$ of the partials, the instantaneous phase of each partial can be computed at any given time t_0 based on the instantaneous phase of the fundamental $\theta(t_0)$ at that time. Note that even though we can only measure the *wrapped* phase of $\theta(t)$ (i.e. in the interval $[-\pi, \pi)$), the correct wrapped phase of each partial can still be calculated.



(a) Waveform of the first three harmonics.

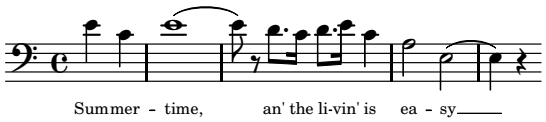


(b) Instantaneous phases of the first three harmonics.

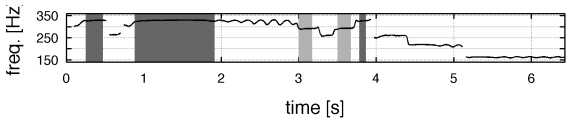
Fig. 1: Illustration of the model parameters.

2.2. Example

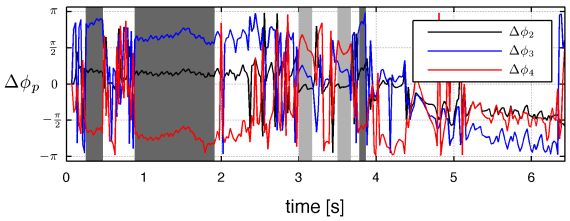
To illustrate the phase relations, we display the phase offsets $\Delta\phi_p$ of the partials with indices $p \in \{2, 3, 4\}$ in a monophonic saxophone recording of "Summertime" by G. Gershwin in Figure 2. The first four bars of this small excerpt are displayed in Fig. 2a, and Fig. 2b



(a) First four bars of "Summertime" by G. Gershwin.



(b) Fundamental frequencies of saxophone performance.



(c) Phase differences $\Delta\phi_p$ over time.

Fig. 2: Visualisation of the phase relations of a musical instrument.

displays the fundamental frequency of the saxophone performance measured by the YIN algorithm [13]. In Fig. 2c, the partial offsets $\Delta\phi_p$ are plotted over time. Phase offsets were obtained from the STFT and computed as the wrapped difference between the measured instantaneous phases of each partial and p times the instantaneous phase of the fundamental. It can be seen that the partial offsets exhibit little variation during the steady state of each note — which is not surprising given the fact that the sound is harmonic. In addition to that, however, the same phase offsets occur at different notes of the same pitch. The area shaded in dark grey highlights all renditions of the note E4 and the light grey area highlights all occurrences of the note D4. These observations make this property suitable for the use in a complex matrix decomposition framework as we will illustrate in the next section. It should be noted that the relative phase offsets can only be defined if the partial frequencies are in a strictly harmonic relation. For instruments with slightly inharmonic frequency relations — such as the piano — a constant phase offset does not exist.

3. PARAMETER ESTIMATION

3.1. Frequency domain model

We aim at estimating the parameters of the model in Eq. (2) from the STFT which is given by

$$X(n,k) = \sum_{t=-K}^{K-1} x(t + n\cdot m) \cdot h(t) \cdot e^{-j\Omega_k t}, \quad (3)$$

where $x(t)$ is the signal under analysis and n and k represent the time frame and frequency index, respectively. $h(t)$ denotes the analysis window of time support $[-K \ldots K-1]$. The distance between consecutive audio frames in samples (hopsize) is denoted by m. $\Omega_k = \frac{2\pi k}{N}$ is the normalised angular frequency of the k-th frequency index.

The STFT of the signal $s(t)$ from Eq. (2) is given by

$$S(n,k) = \sum_{p=1}^{P} a_p H(\Omega_k - p\omega_1) e^{j[p\Theta(n) + \Delta\phi_p]}. \quad (4)$$

In this equation, $H(\Omega) = \sum_{t=-K}^{K-1} h(t) \cdot e^{-j\Omega t}$ denotes the Fourier spectrum of the window function $h(t)$ and $\Theta(n) = \theta(n \cdot m)$.

To simplify the monophonic model in Eq. (4), we assume that each partial can be represented by the main lobe of the window function only. This assumption holds fairly well if the side lobe attenuation of the window spectrum $H(\Omega)$ is sufficiently high and if the frequency resolution of the STFT is high enough so that the main lobes of adjacent partials do not overlap. We denote the partial index belonging to frequency bin k by p_k. We set $p_k = 0$ for all k that lie outside the main lobes of the partials, and set $a_0 = 0$. This allows us to drop the sum in Eq. (4):

$$S'(n,k) = a_{p_k} H(\Omega_k - p_k\omega_1) e^{j[p_k\Theta(n) + \Delta\phi_{p_k}]}. \quad (5)$$

We additionally introduce a real time-varying gain factor $g(n) > 0$ that enables a uniform scaling of the magnitudes in order to accommodate loudness variations (similar to the gains in NMF):

$$\hat{B}(n,k) = g(n) \cdot S'(n,k). \quad (6)$$

Scaling ambiguities between $g(n)$ and a_p can be resolved by normalising a_p. Finally, the model can be extended to incorporate *multiple* harmonic sounds. We denote the index of each harmonic sound by r and append it to the quantities in Eq. (6):

$$\hat{V}(n,k) = \sum_{r=1}^{R} g_r(n) \cdot S'_r(n,k) \quad (7)$$

By substituting Eq. (5) into Eq. (7), we finally obtain:

$$\hat{V}(n,k) = \sum_{r=1}^{R} w_r(k) \cdot h_r(n,k) \quad (8)$$

where $w_r(k) = a_{p_{k,r}} H(\Omega_k - p_{k,r}\omega_{1,r}) e^{j\Delta\phi_{p_k,r}}$ and $h_r(n,k) = g_r(n) e^{j p_{k,r}\Theta_r(n)}$. The term $w_r(k)$ is not time-dependent and is therefore referred to as a *complex basis function*. Accordingly, the term $h_r(n,k)$ is referred to as a *complex activation*. Note that $h_r(n,k)$ is a 2-dimensional function. Eq. (8) is therefore not a complex matrix *factorisation*, but a *decomposition* of a complex spectrogram $V(n,k)$ into a matrix of complex basis functions $w_r(k)$, a matrix of real-valued gain factors $g_r(n)$ and a matrix of real-valued instantaneous phases of the fundamentals $\Theta_r(n)$.

In this paper we will not investigate the case of multiple concurrent sounds. Our goal is to prove that phase offsets between partials are a viable concept for sound analysis purposes. The model parameters will thus be estimated in the monophonic case of Eq. (6) only.

3.2. Parameter estimation

The parameters in Eq. (6) can be estimated by minimizing the error between the original complex spectrogram $B(n,k)$ and the model approximation $\hat{B}(n,k)$ for all $n \in [1 \dots N]$ and $k \in [1 \dots K]$ with $N > 0$ and $K > 0$. We choose to minimise the following cost function:

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} \left|\ln(B(n,k)) - \ln(\hat{B}(n,k))\right|^2 \quad (9)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \left[\ln\left(\frac{|B(n,k)|}{g(n) \cdot a_{p_k} H(\Omega_k - p_k\omega_1)}\right)\right]^2 + [\angle B(n,k) - \Delta\phi_{p_k} - p_k\Theta(n) + 2\pi q(n,k)]^2 \quad (10)$$

where $\angle B(n,k)$ denotes the argument of the complex number $B(n,k)$. The term $q(n,k) \in \mathbb{Z}$ stems from the fact that the logarithm of a complex number has an infinite number of solutions which are obtained by adding integer multiples of 2π to the imaginary part of the solution [14]. The integer $q(n,k)$ is here treated as an additional parameter that has to be estimated. In Eq. (10), $H(\Omega)$ is assumed positive, since we only consider the main lobe of the window function. The model parameters are estimated by means of a coordinate descent (J is minimized w.r.t. each parameter):

$$g(n) = \left(\prod_{k=1}^{K} \frac{|B(n,k)|}{a_{p_k} H(\Omega_k - p_k\omega_1)}\right)^{\frac{1}{K}} \quad (11)$$

$$a_p = \left(\prod_{n=1}^{N} \prod_{\{k|p_k=p\}} \frac{|B(n,k)|}{g(n) H(\Omega_k - p\omega_1)}\right)^{\frac{1}{N \cdot \#\{k|p_k=p\}}} \quad (12)$$

$$\Theta(n) = \frac{\sum_{k=1}^{K} p_k[\angle B(n,k) - \Delta\phi_{p_k} + 2\pi q(n,k)]}{\sum_{k=1}^{K} p_k^2} \quad (13)$$

$$\Delta\phi_p = \frac{\sum_{n=1}^{N}\sum_{\{k|p_k=p\}} \angle B(n,k) - p\Theta(n) + 2\pi q(n,k)}{N \cdot \#\{k|p_k=p\}} \quad (14)$$

$$q(n,k) = \text{round}\left(-\tfrac{1}{2\pi}\left[\angle B(n,k) - \Delta\phi_{p_k} - p_k\Theta(n)\right]\right) \quad (15)$$

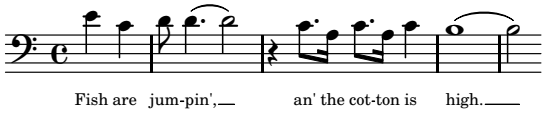
In these equations, the expression $\{k|p_k = p\}$ denotes the set of frequency indices k at which $p_k = p$ and the operator $\#\{\dots\}$ denotes the cardinality of the set. The function round() denotes the rounding of a real number to the nearest integer.

4. ANALYSIS OF AN EXAMPLE SIGNAL

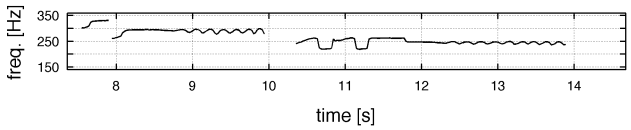
In this section we apply the estimation method to an example signal and illustrate how it can be used for a transcription task. The signal is the same monophonic saxophone recording of "Summertime" that we used to illustrate the phase relations in Figure 2. It has a sample rate of 44.1 kHz and we use the first eight bars of the recording.

The recording is split into two parts. The first part contains the first four bars (cf. Fig. 2a) and is used as *training material* in which prototypical partial amplitudes a_p and phase offsets $\Delta\phi_p$ are learned for different pitches ω_1. The spectrogram with $K = 2049$ frequency bins and $N = 5380$ time frames is manually segmented in time into the different notes. All spectrogram parts with the same nominal pitch are concatenated, the fundamental frequency is estimated by employing the YIN algorithm and ω_1 is computed as the average across all frames for each nominal pitch. $g(n)$ is estimated from the original spectrogram by taking the mean of the magnitudes in each time frame. In order to compute a_p we alternately apply Eq. (12) and (11) for 10 iterations. For the computation of $\Delta\phi_p$, $\Theta(n)$ is initialised by the instantaneous phase value of the frequency bin corresponding to the fundamental in each frame. An initial estimate for $\Delta\phi_p$ is obtained by replacing the terms in the summation in the numerator of Eq. (14) by $\text{wrap}(\angle B(n,k) - p\Theta(n))$, where $\text{wrap}(\alpha)$ calculates the principal argument of α. $q(n,k)$ is computed according to Eq. (15) and we iterate over Eq. (13)–(15) until $q(n,k)$ converges. $\Delta\phi_p$ is eventually given by result of Eq. (14) in the last iteration.

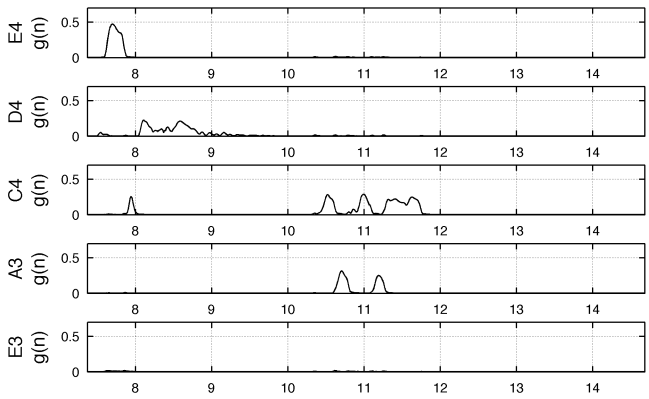
The second part of the recording contains the remaining four bars (cf. Fig. 3a) and is used as *test material*. The learned prototype amplitudes a_p and $\Delta\phi_p$ for all pitches ω_1 that occurred in the first



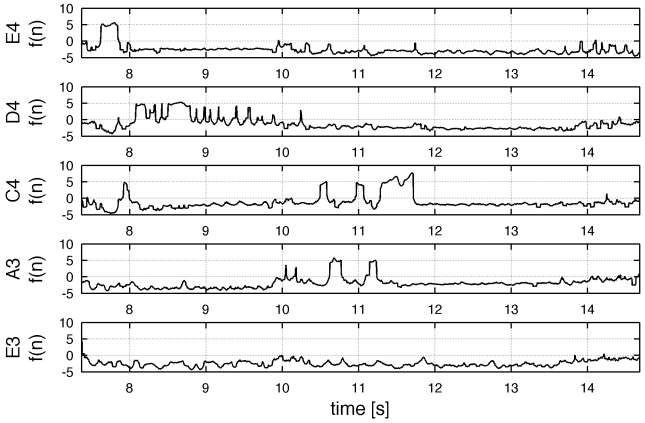
(a) Bars 5–8 of "Summertime" by G. Gershwin.



(b) Fundamental frequencies for bars 5–8.



(c) Gains $g(n)$ for bars 5–8 for each pitch from bars 1–4.



(d) Phase activations $f(n)$ for bars 5–8 for each pitch from bars 1–4.

Fig. 3: Example analysis of a monophonic saxophone example.

part are employed to estimate $g(n)$ and $\Theta(n)$ in the following way. First, $\Theta(n)$ is initialised with the instantaneous phase values at the frequency bins corresponding to ω_1. Then $q(n,k)$ is estimated according to Eq. (15). Finally, $g(n)$ and $\Theta(n)$ are estimated according to Eqs. (11) and (13).

Active pitches can be estimated from both $g(n)$ and $\Theta(n)$. While for $g(n)$ this is obvious — high values indicate activity, low values indicate inactivity —, the instantaneous phase $\Theta(n)$ of the fundamental can also be used as an activity detector. We here use a measure inspired by the phase-based onset detection function described in [15]. The measure is based on the unwrapped phase, which can be assumed to be linear when the note is active and non-linear when the note is inactive. We will denote the unwrapped phase of $\Theta(n)$ by $\Theta_u(n)$. The *second phase difference* can be used as a measure of phase-linearity. It is given by

$$\Delta\Theta_u(n) = \Theta_u(n) - 2\Theta_u(n-1) + \Theta_u(n-2). \qquad (16)$$

If the unwrapped phase is strictly linear, $\Delta\Theta_u(n)$ will be close to zero, if it is non-linear $\Delta\Theta_u(n)$ is likely to take on values with larger magnitudes. Additionally, $\Delta\Theta_u(n)$ is likely to take on low values in several *consecutive active* frames and more random values in *consecutive inactive* frames. We therefore compute the mean square of $\Delta\Theta_u(n)$ over a sliding window as

$$\sigma(n) = \frac{1}{M} \sum_{n'=-\frac{M}{2}}^{\frac{M}{2}-1} \Delta\Theta_u^2(n+n'), \qquad (17)$$

and define the phase-based activity measure as $f(n) = -\ln(\sigma(n))$. In our simulations a window length of 50 ms ($M = 37$) was used.

The results of the estimation are displayed in Figure 3. In Fig. 3b, the measured fundamental frequencies of the four bar excerpt are shown. Fig. 3c shows the gains $g(n)$ and Fig. 3d the results for the phase-based activity measure $f(n)$. The gains clearly show the activity of the different pitches and are very much reminiscent of activity measurements in NMF analyses. The results of the phase-based activity measure also reveal the active pitches very well, which confirms that the phase relations between harmonic partials can actually be used to characterise pitches of certain instruments and distinguish between them. Note that the pitch E3 does not occur in bars 5–8, and that the note B3, the last note in Fig. 3a, is missing because it did not occur in bars 1–4.

5. CONCLUSIONS

In this paper we have investigated the relative phase relations within the sustained part of harmonic sounds and their potential use for complex matrix decomposition. The phase relations between harmonic partials have been expressed as relative phase offsets of the partials w.r.t. the fundamental. Equations for the estimation of the model parameters have been presented based on a complex logarithmic cost function between the original spectrogram and the model approximation. With the analysis of an example signal, we demonstrated the potential of the phase coupling property to capture inherent time-independent phase characteristics of harmonic sounds.

In future work the method should be extended to deal with mixtures of harmonic sounds in order to obtain a complex matrix decomposition that can be used to unmix spectral components in the complex domain. A formulation of such a complex matrix decomposition framework has been provided in Section 3.1. For the monophonic case, the complex logarithmic cost function proved to be useful, not only because logarithmic amplitudes better match the human perception than linear amplitudes, but also because it separates the modulus and argument of the model, which allowed us to treat them separately. In the polyphonic case however, a complex matrix decomposition framework would need to deal with magnitudes and phases jointly, since the sum of two complex time-frequency components depends on both their modulus and phase.

6. 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, Oct. 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2003, IEEE, pp. 177–180.

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[4] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *10th International Conference on Music Information Retrieval*, Kobe, Japan, Oct. 2009, pp. 327–332.

[5] J.-C. Risset and D. L. Wessel, "Exploration of timbre by analysis and synthesis," in *The Psychology of Music*, D. Deutsch, Ed., pp. 113–169. Academic Press, 1999.

[6] D.W. Griffin and J.S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 4, pp. 986–998, Aug. 1983.

[7] R. M. Parry and I. Essa, "Phase-aware non-negative spectrogram factorization," in *7th International Conference on Independent Component Analysis and Signal Separation*, London, UK, Sept. 2007, pp. 536–543, Springer.

[8] 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, Mar. 2009.

[9] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 3437–3440.

[10] B. J. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 8, pp. 2591–2597, Nov 2011.

[11] Roland Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2011, IEEE, pp. 253–256.

[12] Roland Badeau, "High resolution NMF for modeling mixtures of non-stationary signals in the time-frequency domain," Tech. Rep. 2012D004, Télécom ParisTech, Paris, France, July 2012.

[13] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[14] D. Sarason, *Complex Function Theory*, American Mathematical Society, 2nd edition, Dec. 2007.

[15] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sept. 2005.