A SCORE-INFORMED SHIFT-INVARIANT EXTENSION OF COMPLEX MATRIX FACTORIZATION FOR IMPROVING THE SEPARATION OF OVERLAPPED PARTIALS IN MUSIC RECORDINGS

Francisco J. Rodriguez-Serrano**, Sebastian Ewert*$^{+}$*, Pedro Vera-Candeas***, Mark Sandler*$^{+}$

* Universidad de Jaén
Spain

$^{+}$ Queen Mary University of London
United Kingdom

ABSTRACT

Similar to non-negative matrix factorization (NMF), complex matrix factorization (CMF) can be used to decompose a given music recording into individual sound sources. In contrast to NMF, CMF models both the magnitude and phase of a source, which can improve the separation of overlapped partials. However, the shift-invariance for spectral templates enabling NMF-based methods to efficiently model vibrato in music is not available with CMF. Further, the estimation of an entire phase matrix for each source results in a high number of parameters in CMF, which often leads to poor local minima. In this paper we show that score information provides a source of prior knowledge rich enough to stabilize the CMF parameter estimation, without sacrificing its expressive power. As a second contribution, we present a shift-invariant extension to CMF bringing the vibrato-modeling capabilities of NMF to CMF. As our experiments demonstrate our proposed method consistently improves the separation quality for overlapped partials compared to score-informed NMF.

Index Terms— Source separation, music processing, non-negative matrix factorization, overlapped partials.

1. INTRODUCTION

The decomposition of a given music recording into its constituent parts, a task also known as source separation, is one of the central topics in music information retrieval and processing. Possible applications range from stereo-to-surround up-mixing, remixing tools for DJs or producers to instrument-wise equalizing or karaoke systems. Also the use as an intermediate step in music analysis gains in importance, as it enables exploiting properties specific to each instrument class as part of the analysis [1].

With musical instruments typically being highly correlated in time and frequency, musical source separation remains highly challenging. In particular, techniques successfully used for speech separation, such as independent component analysis, typically fail for music as the assumptions made often do not hold. The introduction of non-negative matrix factorization (NMF), however, led to significant improvements during the last decade and most state-of-the-art methods can be considered as extensions. For example, previous methods target music-specific properties including the harmonicity and high temporal continuity of instruments in time-frequency representations [2–4] or the sound production process [5]. Further, many state-of-the-art NMF-based methods employ a *shift-invariant spectral modelling* extension to efficiently model vibrato in music [6–8]. Other approaches integrate additional prior knowledge into the separation process [9, 10], for example, by humming one of the sources into a microphone [11] or letting a user select parts of a sound source in a spectrogram of the recording [12]. Many of these extensions effectively decrease the degrees of freedom compared to standard NMF to obtain a more meaningful separation result, at the expense of flexibility and detail in the model.

Another approach presented in recent years is based on integrating information from a musical score [13–19]. While the availability of such a score is often a strong assumption, the gain in separation quality is typically considerable. Further, the structured information provided by the score allows for a straightforward and automated way to specify the separation targets. Many current score-informed source separation methods employ signal models that are quite similar to traditional methods, i.e. signal models with a low number of parameters compared to standard NMF. In other words, the rich information provided by the score is typically not translated back into an increased level of detail in the model, and thus it remains an open question what the best trade-off between flexibility and robustness in a score-informed model is. In this context, an interesting extension of NMF is the less often used complex matrix factorization (CMF) [20], which employs a factorization-type model for the magnitude of a time-frequency representation similar to NMF but additionally estimates a phase matrix for each source. The phase information can be used to improve the separation quality of overlapping partials as phase cancellations might be taken into consideration. However, due to the size and number of the phase matrices, the number of free parameters in CMF is considerably higher compared to NMF, which in practice can lead to poor local minima during model fitting [20].

A central goal in this paper is to identify if score information is typically rich enough to robustly make use of the expressive power of CMF. In particular, we present a score-informed variant of CMF and demonstrate in systematic experiments that replacing score-informed NMF with our variant consistently and reliably improves the separation quality, with a focus on overlapped partials. Compared to standard CMF, our variant additionally incorporates temporal continuity (or total variation) penalties for the phase which we found to further improve the results. As a second contribution, our variant additionally shows how shift-invariant spectral modelling can be incorporated into CMF, bringing the vibrato-modelling capabilities of NMF to CMF. This way, our CMF variant is a direct replacement for score-informed NMF without sacrificing this central and beneficial property.

The paper is organized as follows. In Section 2, we briefly review NMF and its variants, including CMF. Based on these foundations, we then present in Section 3 our score-informed, shift-invariant extension to CMF. In Section 4 we then demonstrate using systematic experiments that our proposed method indeed improves the separation

This work was supported by the Andalusian Business, Science and Innovation Council under project P2010- TIC- 6762 and (FEDER) the Spanish Ministry of Economy and Competitiveness under Project TEC2012-38142-C04-03, as well as EPSRC grants EP/J010375/1 and EP/L019981/1.

quality over a score-informed NMF, using the same settings for shared parameters to increase the comparability. Finally, our conclusions and prospects on future work are given in Section 5.

2. NON-NEGATIVE MATRIX FACTORIZATION AND VARIANTS

Non-negative matrix factorization has turned out to be a highly useful tool for decomposing a given music recording into its constituent parts [21]. To this end, one approximates the magnitude of a time-frequency representation of a given recording $\mathcal{S} \in \mathbb{C}^{M \times N}$ by a product of two non-negative matrices $W \in \mathbb{R}_{\geq 0}^{M \times K}$ and $H \in \mathbb{R}_{\geq 0}^{K \times N}$, i.e. $V := |\mathcal{S}| \approx W \cdot H$. In this context, the columns of W are often referred to as *template vectors* and the rows of H as the corresponding *activations*. The former provide information about the spectral energy distribution of a sound source, while the latter encode when and how intense a source is active. For an overview of algorithms for computing such an NMF factorization, we refer for example to [21]. By allowing the templates we learn using NMF to be shifted along the frequency axis, we obtain a slightly extended version of NMF often referred to as *shift-invariant NMF* [6]. To this end, we choose a number of possible shifts $S > 0$ we want to consider, and approximate V component-wise by $V(m,n) = \sum_{k=1}^{K} \sum_{s=1}^{S} W(m-s,k) H_s(k,n)$, where each H_s contains the activations for shift s. If a log-distributed frequency scale is used in V, the shift corresponds to slight changes of the fundamental frequency for a template and thus enables accounting for vibrato or tuning differences without requiring an excessive amount of template vectors for a given musical pitch [8]. For algorithms to compute a shift-invariant factorization, see for example [6, 7].

To incorporate score information into the NMF factorization process, various approaches have been proposed [13–19]. Most of them employ parametric signal models, where the templates (and sometimes the activations) are described using a few meaningful parameters which can be associated with and constrained using score information. While the integration of prior knowledge is often simplified using such models, the reduced number of parameters often limits the expressivity of the model. A different approach was used in [19] extending ideas from [3]. This approach exploits that certain constraints can easily be incorporated into NMF by setting specific entries to zero – since most algorithms use multiplicative update methods a value of zero will stay zero throughout the entire parameter estimation. This way, after associating each template with a specific musical pitch, one can use the information from the score to set the corresponding pitch activation to zero whenever the pitch is known be inactive. Similarly, given a musical pitch, it is possible to roughly estimate the fundamental frequency and the harmonics and one can set entries in a template to zero that are not in a neighborhood of these frequencies and thus can be expected to be zero. Beyond that no other structured prior is imposed on the activations or the templates such that the model retains a lot of freedom. As shown in [19], the score information is rich enough to obtain a robust yet highly detailed signal model. However, it is yet unclear if the score information might even allow for increasing the modelling detail even further. It is one goal of this paper to find out if it is rich enough to support a CMF-based signal model which has many more parameters than an NMF model.

A difference between complex matrix factorization as introduced in [20] and NMF is the introduction of a phase matrix for each source. More precisely, the complex time-frequency representation $\mathcal{S}$ is approximated using CMF as

$$\mathcal{S} \approx \sum_{k=1}^{K} W(m,k) H(k,n) e^{i\phi_k(m,n)}. \qquad (1)$$

As we can see, compared to the $K(M+N)$ parameters used in NMF, CMF uses KMN additional phase parameters which is typically several times higher. As an advantage, CMF models the complex spectrogram, which in contrast to its magnitude is truly additive w.r.t. the individual sources, enabling CMF to account for phase cancellations between overlapping sources, which can potentially increase the separation performance. A first algorithm to minimize a distance between $\mathcal{S}$ and the right hand side of Eq. 1 was presented in [20]. However, since the number of parameters in the phase matrix is high, it is often useful to constrain the ϕ_k to values that are plausible given knowledge of the underlying time-frequency representation. For example, the approach presented in [22] assumes that for each source, a precise track of the fundamental frequency is available, which enables constraining the phase advance $\phi_k(m,n+1) - \phi_k(m,n)$ between frames n and $n+1$ to plausible values w.r.t. the step size used using a temporal continuity or total variation penalty. In the next section, we will use a similar constraint, eliminating however the need for a precise track of the fundamental frequencies. In this context, we want to note that other non-magnitude models are available, for example the HR-NMF model [23]. This approach models certain parameters using ARMA processes, which is likely to be more stable than a simple total variation. We chose CMF instead of HR-NMF or other approaches for two reasons. First, the computational costs of these alternatives are significantly higher compared to CMF. Second, CMF provides even more modelling flexibility than HR-NMF and a main point of this paper is to find out if such a flexibility can be controlled by integrating score-information.

3. SCORE-INFORMED SHIFT-INVARIANT CMF

For our proposed method, we integrate both the idea of shift-invariance and the idea of score-based constraints into complex matrix factorization. This way, we can combine compact vibrato modelling capabilities with a full phase-aware signal model, while the score-information guides the parameter estimation to a meaningful result. Our full model and corresponding objective function to be minimized can be written as (see below for an explanation of the new terms):

$$\begin{aligned}
&\sum_{m,n} \left| \mathcal{S}(m,n) - \sum_{q,p,s} \overline{W}_q(m-s,p) H_{q,s}(p,n) e^{i\phi_q(m,n)} \right|^2 \\
&+\sigma \sum_{p,q,m,n,r} M_q(m,p) A_q(p,n) \left| e^{i\phi_q(m,n)} - e^{i\phi_q(m,n-1)} e^{i2\pi hr f_{q,p}(n)/F} \right|^2 \\
&\quad s.t.\ \ \overline{W} \geq 0,\ \ H \geq 0,\ \ H_{q,s}(p,n) > 0 \Leftrightarrow A_q(p,n) = 1, \\
&\quad \overline{W}_q(m,p) > 0 \Leftrightarrow M_q(m,p) = 1
\end{aligned} \qquad (2)$$

To explain the model, we assume that we have Q instruments, each playing up to P pitches and we consider S shifts for each template. The first term is a data fidelity term in the form of the square of the Frobenius norm between the given time-frequency representation $\mathcal{S}$ and our model, which is a shift-invariant version of CMF – compare also the shift invariant NMF model in the last section. The dashes above W are not operators but simply notation to make the relationship to other objects to be defined below clear. The second term is a regularization for the phase, which was originally introduced in a similar form in [22]. The idea behind this term is simply that the phase in frame n should not deviate much from the phase in frame $n-1$ after being advanced using the fundamental frequency for a certain pitch. More precisely, $f_{q,p}(n)$ is the fundamental frequency of instrument q playing pitch p in frame n and F is the sampling frequency, both in Hertz. Further, h is the hope size in samples between frames and r denotes the r-th harmonic. Since the theoretical phase advance associated with this r-th harmonic can only meaningfully be used to approximate the real phase advance for certain frequency bins m, we need to constrain the sum in the second term somehow to allow only certain combinations of r and m. To this end, we employ a slightly different mechanism compared to [22] based on two terms, A_q and M_q, which are assumed to be given.

More precisely, A_q encodes in a binary form information provided by the score: $A_q(p,n) = 1$ indicates that pitch p of instrument q is active in frame n – here, we assume that the score is temporally aligned to the given audio recording. If that is not the case, an alignment method such as [24] can be used. Resembling the concepts behind A_q, $M_q(m,p)$ is assumed to be 1, if frequency bin m is in a close vicinity of a harmonic of pitch p for instrument q, i.e. M_q expresses that we do not expect energy between the partials of a harmonic sound and is therefore referred to as the *harmonic mask* in the following, see also [19] for similar approaches. Using these two terms, we make sure that we only penalize unexpected phase advances in ϕ if we have a rough understanding of how it should be based on the information from the score. Furthermore, we also use the same A and M to apply constraints on the magnitude model, by specifying which entries in W and H are allowed to be positive and which must remain zero.

Before we present an algorithm to compute (W, H, ϕ) minimizing our distance, we first change the model slightly and make the shift of templates explicit in W, i.e. $W_{q,s}(m,p) = \overline{W}_q(m-s,p)$. The additional constraint ensures that this relationship holds during the parameter estimation process.

$$\begin{aligned}
&\sum_{m,n} \left| \mathcal{S}(m,n) - \sum_{q,p,s} W_{q,s}(m,p) H_{q,s}(p,n) e^{i\phi_q(m,n)} \right|^2 \\
&+\sigma \sum_{p,q,m,n,r} M_q(m,p) A_q(p,n) \left| e^{i\phi_q(m,n)} - e^{i\phi_q(m,n-1)} e^{i2\pi h r f_{q,p}(n)/F} \right|^2 \\
&\quad s.t. \;\; W \geq 0, \;\; H \geq 0, \;\; H_{q,s}(p,n) > 0 \Leftrightarrow A_q(p,n) = 1, \\
&\quad W_{q,s}(m,p) > 0 \Leftrightarrow M_q(m-s,p) = 1, \\
&\quad W_{q,s}(m,p) = W_{q,0}(m-s,p)
\end{aligned} \tag{3}$$

Using this slight change, we can now more easily define the iterative update rules for our *score-informed shift-invariant CMF*. For a lack of space, we cannot provide detailed derivations here. The general approach, however, is similar to the ones used for example in [20–22].

Algorithm 1 Score-Informed Shift-Invariant CMF Algorithm

```
1  Compute $\mathcal{S}$ from the input signal.
2  Initialize activation mask $A$ and $f$ using score information, $M$ based on $f$, $\phi$ with copies of $\mathrm{Arg}(\mathcal{S})$, $W$ and $H$ with random positive values.
3  for $J_1$ iterations do
4      Compute $B$ with Eq.(7).
5      Compute $Y$ with Eq.(6).
6      Update $\phi$ with Eq.(8).
7      Update $W$ with Eq.(4).
8      Project $W$ onto non-negative orthant.
9      Compute $\overline{W}$ via $\overline{W}_q(m,p) := \frac{1}{S}\sum_s W_{q,s}(m-s,k)$.
10     Apply harmonic mask $M$: $\overline{W} = \overline{W} \odot M$.
11     Normalize $\overline{W}$ such that $\sum_m \overline{W}_q(m,p) = 1$.
12     Derive shifted dictionary $W_{q,s}(m,p) := \overline{W}_q(m-s,p)$.
13     Update $H$ with Eq.(5).
14     Project $H$ onto non-negative orthant.
15 end for
16 Refine $f_k$ according to section 3.1.
17 Update harmonic mask $M$ using $f$.
18 Repeat steps (4-14) for $J_2$ iterations.
```

The entire algorithm is shown in Algor. 1, with some of its steps given in more detail below. In particular, our algorithm is split into two steps. During the first step, we initialize the fundamental frequencies f we need for the phase regularization using only rough estimates, i.e. we set $f_{q,p}(n)$ to the standard MIDI frequency corresponding to pitch p. Based on this f, we then derive the harmonic mask M which encodes the location of harmonics for each pitch. While this is not perfect, we can use these initial f and M to obtain a first estimate of all parameters. Next, based on the resulting initial model and the score information, we track the fundamental frequency of each note specified by the score. The details of this step are given below. After that, we use the refined f to update M and re-estimate the remaining model parameters. Using this two step process, we can eliminate the need for manually provided fundamental frequency estimates as required in [22]. The updates for the other parameters within each step are as follows:

$$W_{q,s}(m,p) = \frac{\sum\limits_n H_{q,s}(k,n) \Re\left(\left(\frac{Y_q(m,n)}{B_q(m,n)}\right) e^{-i\phi_q(m,n)}\right)}{\left(\frac{\sum\limits_n H_{q,s}(p,n)^2}{B_q(m,k,n)}\right)} \tag{4}$$

$$H_{q,s}(p,n) = \frac{\sum\limits_m W_{q,s}(m,p) \Re\left(\left(\frac{Y_q(m,n)}{B_q(m,n)}\right) e^{-i\phi_q(m,n)}\right)}{\left(\frac{\sum\limits_m W_{q,s}(m,p)^2}{B_q(m,n)}\right)} \tag{5}$$

where $\Re$ is the real part, and the reconstruction term $Y_q(m,n)$ and selection term $B_q(m,n)$ are defined as:

$$\begin{aligned}
Y_q(m,n) = &\sum_p W_{q,s}(m,p) H_{q,s}(p,n) e^{i\phi_q(m,n)} \\
&+ B_q(m,n)\left(\mathcal{S}(m,n) - \widetilde{\mathcal{S}}(m,n)\right)
\end{aligned} \tag{6}$$

$$B_q(m,n) = \frac{\sum_p W_{q,s}(m,p) H_{q,s}(p,n) A_q(p,n)}{\sum\limits_{\tilde{q},\tilde{p},\tilde{s}} W_{\tilde{q},\tilde{s}}(m,\tilde{p}) H_{\tilde{q},\tilde{s}}(\tilde{p},n) A_{\tilde{q}}(\tilde{p},n)} \tag{7}$$

where $\widetilde{\mathcal{S}}(m,n) := \sum_{q,p,s} W_{q,s}(m,p) H_{q,s}(p,n) e^{i\phi_q(m,n)}$. Further, the update rule for the phase parameter is similar to the one proposed in [22]:

$$\begin{aligned}
\phi_q(m,n) = \mathrm{Arg}\Bigg(&\sum_{p,s} \frac{Y_q(m,n)}{B_q(m,n)} W_{q,s}(m,p) H_{q,s}(p,n) \\
&+ \sigma \sum_{r,p} M_q(m,p) A_q(p,n) \Big(e^{i\phi_q(m,n-1)} e^{i2\pi f_{q,p}(n) r h/F} \\
&+ e^{i\phi_q(m,n+1)} e^{-i2\pi f_{q,p}(n) r h/F}\Big)\Bigg)
\end{aligned} \tag{8}$$

3.1. Refinement of Fundamental Frequencies

The phase regularizer proposed in [22] requires the fundamental frequency of each component in each frame as input to the method. Since our method is score informed, the stability gained this way allows us to employ this regularizer using only estimates of the fundamental frequencies, precise only within a semitone, and still get a meaningful result. However, the results improve if more accurate estimates are available. Therefore, after the initial step using the rough estimates, we refine the fundamental frequencies based on the initial model we have obtained so far based on a simple procedure. In particular, we have for each instrument q, pitch p and frame n the activations for each of the S possible shifts: $H_{q,1}(p,n),\dots,H_{q,S}(p,n)$. Using the shift having maximal activation among all shifts in this frame, $S_{\max}$, we analyze the template for pitch p in $W_{q,S_{\max}}$ to obtain a refined fundamental frequency estimate. To this end, for each fundamental frequency candidate, we simply compute a weighted sum of entries $W_{q,S_{\max}}(m,p)$ for all frequency bins m that correspond to that fundamental frequency or one of its harmonics. The candidate with the highest sum is used as the value for $f_{q,p}(n)$. After that, we can use the new f to update our harmonic mask M. In particular, since during the first step f was only a rough estimate, we use rather wide regions of ones around possible positions of partials. During the second step, f is more accurate and we can use less entries of ones, which sharpens the mask and its constraints.

Overall, we want to remark that the presented algorithm is not guaranteed to converge. In particular, using projections in steps 8 and 14 and the projection-like operations in steps 4, 6 and 10 effectively implement the constraint-compliance outside of the actual optimization. Further, we enforce a decoupling of dependent variables in the algorithm. Alltogether, this can make the algorithm slightly 'jump' in the vicinity of a local minimum. With this algorithm, we chose simplicity of the algorithm and its implementation over providing theoretical guarantees. Whether our method can improve the separation quality over a comparable score-informed shift-invariant NMF method despite this jumping behaviour, i.e. whether they have an actual influence on the result, will be shown in the next section.

4. EXPERIMENTS

For our experiments we used the dataset proposed in [25], consisting of 10 four-part chorales by J.S. Bach which are given as multitrack recordings of real recordings, each approximately 30 seconds long. Each music excerpt consists of an instrumental duet among these instruments: violin, clarinet, tenor saxophone and bassoon. Temporally aligned score information was obtained from MIDI files[1], where the alignment was manually checked. We mixed individual tracks from each chorale to create 60 duets. Note that other works using CMF, such as [20, 22], use signals consisting of two real instrument notes, so the number of components is highly limited and the fundamental frequency of each constituent signal is known in advance. Here we deal with a set of complete melodies where the score information (with a semitone frequency resolution) is used as input instead of the fundamental frequency ground truth. This way, we hope to provide a somewhat more realistic evaluation.

We used our proposed method as well as a score-informed NMF-based method similar to [19] to decompose these mixes into the two instruments. Overall, we tried to make the two approaches as comparable as possible – indeed the most distinctive, conceptual difference between the two approaches is the addition of the phase parameter tensor and the corresponding regularizer. In both cases, we use the log-frequency spectrogram proposed in [26] as our spectral front-end. This invertible transform employs a parameter providing control over the time-frequency trade-off for lower frequencies; to improve the temporal resolution of our method we set this to $\gamma = 20$, see [26] for details. Further, both methods use $S = 5$ shifts and one spectral template per pitch and instrument (as specified by the score). Further, for our proposed CMF method we set $\sigma = 0.1$ for the phase regularizer as proposed in [22]. Also, we use 50 initial iterations for the first step in our method, and let it run until convergence for the second step. Table 1 shows a set of measures over the dataset computed using the *PEASS Toolbox* [27]. This toolbox offers non-perceptual (SDR, ISR, SIR and SAR) and perceptual measures (OPS, TPS, IPS and APS).

Method	**SDR**	**ISR**	**SIR**	**SAR**
SISI-CMF	**11.51**	**18.35**	**17.55**	22.16
SISI-NMF	11.15	17.87	17.27	**23.99**

Method	***OPS***	***TPS***	***IPS***	***APS***
SISI-CMF	**38.65**	61.17	**56.10**	**41.65**
SISI-NMF	35.75	**63.86**	49.02	38.62

Table 1. SSS results of the proposed Score-Informed Shift-Invariant Complex-NMF method (SISI-CMF) compared to a Score-Informed Shift-Invariant NMF method (SISI-NMF). Listening examples are available at `http://anclas3.ujaen.es/cnmf_sss/`

Overall, the two methods are conceptually very similar and if any we expect a gain in separation quality only to be related to overlapped partials – and those contribute only to some degree to the overall signal energy. Looking at the results in table 1, we find indeed small improvements for most quality measures for our SISI-CMF methods compared to the comparable SISI-NMF method. In particular, our CMF variant aims at modeling the overlapped partials in such way that the interference between instruments could be reduced. And indeed, not only SIR, but also its corresponding perceptual measure IPS, indicate that this desirable effect has been achieved. Taking the overlap into the account, the magnitude estimates could also be improved, which typically has an effect on the overall separation quality. Looking at the SDR, which summarizes the whole frequency domain without any energy weighting, indeed shows a small gain. The perceptual OPS value, however, shows the improvement more clearly. A negative effect of the increase of the number of parameters to be estimated in CMF compared to NMF might be indicated by the SAR and TPS, where we see better values for NMF. However, overall, the results indicate that the score information is rich enough to contain the vast number of parameters in CMF and use the resulting freedom to improve the separation results.

5. CONCLUSIONS

We have presented a novel score-informed shift-invariant extension of complex matrix factorization. Our results indicate that incorporating and estimating phase information indeed leads to improved results in score-informed source separation. This also indicates, that the score information provides enough prior knowledge to control and guide the CMF parameter estimation process, despite the vast number of parameters it contains. For the future, we plan to investigate parametric dictionary learning processes and hybrid source separation systems, which treat overlapped and non-overlapped time-frequency areas in different ways. In this context, our proposed method could be an interesting base signal model.

6. REFERENCES

[1] N. Ono, K. Miyamoto, H. Kameoka, J. Le Roux, Y. Uchiyama, E. Tsunoo, T. Nishimoto, and S. Sagayama, "Harmonic and percussive sound separation and its application to MIR-related

[1] http://www.jsbchorales.net/index.shtml

tasks," in *Advances in Music Information Retrieval*. Springer, 2010, pp. 213–236.

[2] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication (ISCA Journal)*, vol. 43, no. 4, pp. 311–329, 2004.

[3] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 381–386.

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

[5] T. Heittola, A. P. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 327–332.

[6] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," *Independent Component Analysis and Blind Signal Separation*, pp. 700–707, 2006.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 2069–2072.

[8] B. Fuentes, R. Badeau, and G. Richard, "Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 401–404.

[9] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[10] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed audio source separation: A comparative study," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2397–2401.

[11] P. Smaragdis and G. J. Mysore, "Separation by humming: User guided sound extraction from monophonic mixtures," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2009, pp. 69–72.

[12] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[13] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," in *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 133–138.

[14] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, New York, USA, 2010, pp. 462–465.

[15] S. Ewert and M. Müller, "Estimating note intensities in music recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 385–388.

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

[17] U. Şimşekli and A. T. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2639–2643.

[18] F. J. Rodríguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J. J. Carabias-Orti, "Online score-informed source separation with adaptive instrument models," *Journal of New Music Research*, no. ahead-of-print, pp. 1–14, 2015.

[19] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, May 2014.

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

[21] A. Cichocki, R. Zdunek, and A. H. Phan, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley and Sons, 2009.

[22] J. Bronson and P. Depalle, "Phase constrained complex nmf: Separating overlapping partials in mixtures of harmonic musical sources," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7475–7479.

[23] R. Badeau and A. Drémeau, "Variational bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[24] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 1869–1872.

[25] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1205–1215, Oct 2011.

[26] C. Schörkhuber, A. Klapuri, and A. Sontacchi, "Audio pitch shifting using the constant-q transform," *Journal of the Audio Engineering Society*, pp. 562–572, 2013.

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