ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/257873670

Score-Informed Source Separation for Music Signals

Chapter · January 2012

CITATIONS
15

READS
1,531

2 authors, including:

Meinard Müller
Friedrich-Alexander-Universität of Erlangen-Nürnberg
357 PUBLICATIONS **9,957** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

FMP (Fundamentals of Music Processing) Notebooks View project

METRUM View project

All content following this page was uploaded by Meinard Müller on 02 June 2016.

The user has requested enhancement of the downloaded file.

Score-Informed Source Separation for Music Signals

Sebastian Ewert*[1] and Meinard Müller†[2]

1 **Institute for Computer Science III, University of Bonn**
Römerstr. 164, 53117 Bonn, Germany
ewerts@iai.uni-bonn.de

2 **Saarland University and MPI Informatik**
Campus E1-4, 66123 Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Abstract

In recent years, the processing of audio recordings by exploiting additional musical knowledge has turned out to be a promising research direction. In particular, additional note information as specified by a musical score or a MIDI file has been employed to support various audio processing tasks such as source separation, audio parameterization, performance analysis, or instrument equalization. In this contribution, we provide an overview of approaches for score-informed source separation and illustrate their potential by discussing innovative applications and interfaces. Additionally, to illustrate some basic principles behind these approaches, we demonstrate how score information can be integrated into the well-known non-negative matrix factorization (NMF) framework. Finally, we compare this approach to advanced methods based on parametric models.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Audio processing, music signals, source separation, musical score, alignment, music synchronization, non-negative matrix factorization, parametric models

Digital Object Identifier 10.4230/DFU.Vol3.11041.73

1 Introduction

The decomposition of a mixture of superimposed acoustic sound sources into its constituent components, a task also known as *source separation*, is one of the central research topics in digital audio signal processing. For example, in speech signal processing, an important task is to separate the voice of a specific speaker from a mixture of conversations of multiple speakers and background noises ("Cocktail party scenario"), see for example [29]. Also in the field of musical signal processing, there are many related issues that are commonly subsumed under the notion of source separation. In the musical context, a source might correspond to a melody, a bassline, a drum track, or an instrument track. To extract such sources, various elaborate processing and analysis methods have been developed, which have led to significant improvements for tasks such as instrument recognition [22], harmonic analysis [47], or melody estimation [12]. Most of these methods exploit certain spectral and temporal

* Sebastian Ewert has been funded by the German Research Foundation (DFG CL 64/6-1).
† Meinard Müller has been supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). He is now with Bonn University, Department of Computer Science III, Germany.

© Sebastian Ewert and Meinard Müller;
licensed under Creative Commons License CC-BY-ND
Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 73–94

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

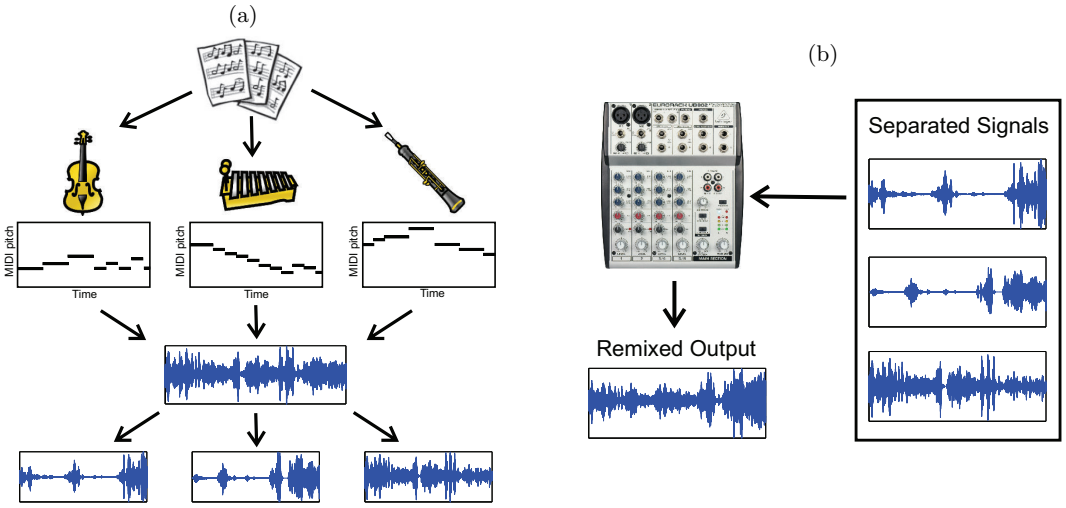

Figure 1 Score-informed source separation: **(a)** Instrument tracks as specified by a given score are employed for the separation of instrument sounds from a polyphonic audio recording (figure inspired by [24]). **(b)** Separated signals corresponding to instrument tracks can be remixed by the user in real-time (figure inspired by [27]).

properties of the sound sources to be extracted. For example, the melody is often the leading voice characterized by its dominance in dynamics and by its temporal continuity [3, 9]. The track of a bass guitar may be identified by specifically looking at the lower part of the frequency spectrum [19]. Furthermore, when extracting the drum track, one often relies on the assumption that the other sources are of harmonic nature. Then one can exploit that percussive elements (vertical spectral structures) are fundamentally different from harmonic elements (horizontal spectral structures) [36]. Last but not least, a human singing voice can often be distinguished from other musical sources because of the presence of vibrato and portamento (sliding voice) effects [40].

In the last years, also multimodal, score-informed source separation strategies have been employed where one assumes the availability of a score representation along with the music recording. The score provides valuable information in two respects. On the one hand, pitch and timing of note events provide a rough guidance within the separation process. On the other hand, the score provides a natural way of specifying what and how sound sources are to be separated. For example, in [24] the score's natural partition into instrument tracks is exploited to extract each individual instrument from a given audio recording, see Figure 1a for an illustration. Here, the score provides additional cues on the sources' spectral and temporal properties. In [27], it was demonstrated that this concept can be incorporated into an intuitive and easy-to-use interface. Here, the user can adjust the volume of each instrument in real-time using an interactive instrument equalizer, see Figure 1b. Developing this idea further, one can extend the instrument equalizer to a more general voice or note equalizer [14], where the user can not only emphasize or attenuate whole instrument tracks but also specific note groups played by different or the same instrument. Here, a group of notes might correspond to a motif, a voice, the left or the right hand of a piano score, or a staff as illustrated in Figure 2a. Incorporating these concepts into multimodal music players [5, 6], one can intuitively select note groups in the score and separate or enhance them in the audio recording in real-time, see Figure 2b.

In this contribution, we give an overview of strategies that employ score information

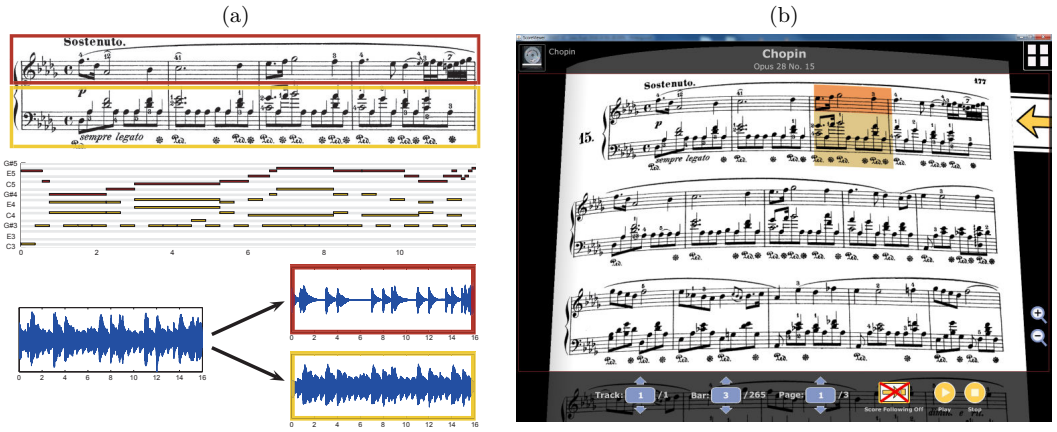

Figure 2 Score-informed voice separation: **(a)** Decomposition of a piano recording into two sound sources corresponding to the left and right hand as specified by a musical score. Shown are the first four measures of Chopin's Prélude "Raindrop" (Op. 28 No. 15). **(b)** Prototypical implementation of a voice equalizer based on the multimodal music player proposed in [5]. By selecting a staff/hand in the scanned score image the corresponding group of notes is separated/enhanced in real-time.

for separating musically meaningful sound sources from polyphonic music recordings. In Section 2, we summarize available score-informed source separation methods. Here, we focus on conceptual differences between the individual approaches rather than giving technical details. Then, using the well-known non-negative matrix factorization (NMF) framework as an example, we demonstrate in Section 3 how score information can be employed to guide the separation process. Finally, as an alternative to NMF-based approaches, we discuss in Section 4 advanced source separation methods based on parametric models. Conclusions and prospects on future work are given in Section 5.

2 Methods for Score-Informed Source Separation

In general, separating sound sources from polyphonic music recordings requires an understanding of many musical and technical aspects. For example, one has to account for the complexity of musical sound sources, the interaction and superposition of such sources in polyphonic mixtures, room acoustics, and recording conditions. Additionally, in many studio productions, numerous digital effect filters are applied to the recording thus making the task even more complex. However, although being extremely difficult, source separation is mostly pursued in a blind fashion, where as little prior knowledge as possible is used.

A natural idea to facilitate the separation process is to incorporate additional musical cues, for example, employing available musical score data. In this context, music synchronization methods are of particular importance [7, 8, 16, 28, 33]. Given a MIDI file representing the score and an audio recording representing an interpretation of a piece of music, the goal is to determine for each MIDI note event its corresponding time position in the audio recording. By adjusting the onset position and duration of each MIDI event, one can use the computed alignment to transform the original *score-like MIDI file* to a *synchronized MIDI file*, which runs synchronously to the audio, see Figure 3. Each score-informed source separation approach treats this problem differently. Some approaches consider or even account for typical differences between the score and a given interpretation, for example, in terms of structure, ornamentation, the interpretation of trills and arpeggios as well as additional

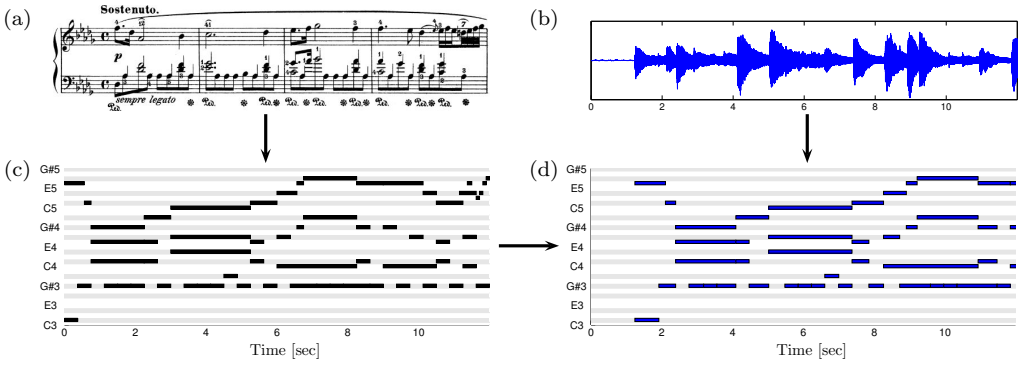

Figure 3 Music synchronization for a score and an audio recording of Chopin's Op. 28 No. 15: **(a)** Musical score. **(b)** Audio recording of an interpretation taken from the SMD database [34]. **(c)** Score-like MIDI file generated from the score shown in (a). **(d)** Synchronized MIDI file.

and missed notes. Other approaches simply assume that *perfectly synchronized MIDI files* are available. This assumption, however, is often not realistic. In real-world scenarios, one typically has to adjust a MIDI file to a given audio recording so that perfect synchronicity can not be guaranteed.

Early approaches adopt score and MIDI information only for evaluation purposes, for example, to investigate the influence of a pitch estimation step in a complex separation system [37]. One of the first approaches focusing on the conceptual benefits of incorporating score information was proposed in [42]. Here, the task consists in separating a single instrument specified by a given score-like MIDI file from a polyphonic music recording. The main idea is based on designing a filter, which in some sense optimally extracts the instrument from the recording. To compute the MIDI-audio synchronization, the authors refer to a procedure previously proposed in [41]. While presenting a novel application idea, this early work has several conceptual limitations. First of all, the proposed filter design procedure models all non-target sound sources as Gaussian noise. Therefore, in cases where the target instrument is accompanied by other instruments, this assumption is obviously violated. Furthermore, the proposed method assumes that the score provides an exact specification of the fundamental frequency for the target instrument for each analysis frame. This assumption is not realistic, since the score usually provides only high-level note information of the piece of music without specifying tuning or small pitch deviations of the respective music recording.

Subsequently proposed systems were not subject to such strict limitations. In [54], the authors integrate score information into a system for blind source separation previously described in [53] (an extended version was presented in [52]). Here, the goal is to extract individual instruments from a music recording, which then enables a user to create new music by remixing the extracted sound sources. In this approach, stereo information is employed in a first step to determine for each analysis frame the number of concurrent sources. Frames identified to contain only a single source are used as cues in the consecutive pitch-tracking step to support the separation in frames with multiple sources. The authors incorporate score information into this process as a rough guidance for the pitch-tracking. The underlying MIDI-audio alignment is based on a procedure proposed by Hu et al. [25]. A technical limitation of the approach is its dependency on reliable stereo information to identify the sources. This is problematic for many commercial studio productions, where spatial information contained in the stereo recordings is often corrupted by digital effect filters and virtual room acoustics. Furthermore, the influence of the alignment step is hard to

assess from the experimental results, as the method is only evaluated on a dataset consisting of four-second snippets of synthetically created MIDI sonifications.

While score information is used in [54] mostly as an add-on to an existing source separation system, Han and Raphael presented in [20, 21] a model that completely relies on available score data. In their contribution, the authors aim at removing the soloist from orchestral music recordings to generate recordings that can be used as a basis for automated accompaniment systems [7]. Relying on score information at an early stage of their algorithmic pipeline allowed for innovative computational concepts. On the one hand, the method represents a given input spectrogram as a compound of note-event based models. This allows for effectively using the score information to specify the temporal and spectral extent in which a note-event is permitted to be active. On the other hand, the score is used to identify the instruments occurring in a given music recording. This way, some instrument-dependent model parameters such as overtone energy distributions can simply be learnt from monophonic training material in advance and fixed afterwards. A benefit of this approach is that the parameter estimation process becomes efficient (as only a small set of parameters needs to be adjusted) and robust (as unreasonable parameter values are prevented by the model). However, a drawback is that the model can be imprecise, in particular when the training instruments differ strongly from the ones used in the given recording.

Roughly at the same time, Itoyama et al. presented a system, which explored novel application scenarios based on score-informed source separation [27]. This system allows a user to adjust the volume of each instrument in a polyphonic music recording in real-time. To this end, the system separates the individual instrument tracks in a preprocessing step as follows. In a first step, a MIDI synthesizer is employed to create one audio representation for each of the instrument tracks contained in a given MIDI file. This audio data is used as prior knowledge to initialize a note-based spectrogram model. Next, the model parameters are adapted to a given audio recording by minimizing a Kullback-Leibler distance between the given and the model spectrogram. Here, to allow only musically meaningful values for the model parameters, strong deviations form the initial values set in the first step are penalized. In a final step, the spectrogram model is employed to isolate the individual instrument tracks as specified by the MIDI file. Technically, the model is based on the harmonic-temporal-structured clustering (HTC) model proposed in [30], which will be discussed in more detail in Section 4. To control the influence of their percussion related submodel on the remaining system, the authors have to resort to smoothing and regulation techniques [26], which further increase the complexity of the system. Furthermore, alignment issues are not considered in this approach, hence it is not clear how the system behaves in real-world scenarios starting with score-like MIDI files.

Using MIDI-synthesized audio material for initialization purposes was also proposed by Gansemann et al. in [17, 18]. Given a MIDI file and an audio recording for a piece of music, the approach starts by sonifying the MIDI instrument tracks using a wavetable synthesizer similar to [27]. In a next step, probabilistic latent component analysis (PLCA) [43] is employed to identify the most important spectral components for each sonification. Here, PLCA is a probabilistic formulation of the well-known non-negative matrix factorization (NMF) method, which will be discussed in more detail in Section 3. In a last step, the instrument-wise spectral components are used as initialization and additional knowledge for a prior-based PLCA analysis of the original audio recording [45]. The results of this final analysis are subsequently used to extract each instrument from the original recording. Incorporating an alignment procedure by Turetsky and Ellis [46], the authors aim at using full-length score-like MIDI files as they can be found in real-world scenarios. While this approach presents a

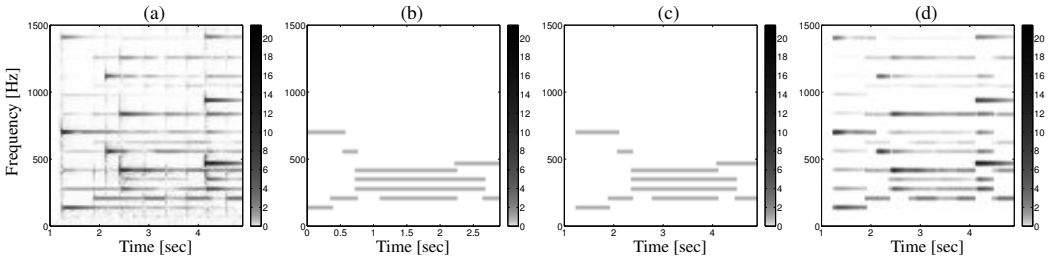

Figure 4 Score-informed parametric spectrogram model as employed in [13]. **(a)** Original magnitude spectrogram for a recording of Chopin's Op. 28 No. 15. **(b)** Model spectrogram after initialization with note events from a score-like MIDI file. **(c)** Model spectrogram after the synchronization step. **(d)** Model spectrogram after the estimation of remaining model parameters.

novel computational concept, the approach suffers from several weaknesses. Similar to all approaches relying on synthetic audio material as prior knowledge, this method's separation quality depends on the spectral similarity between the MIDI instruments and the actual target instruments. Moreover, this method also requires that the MIDI instruments have a similar tuning as the instruments in the given audio recording. For large tuning deviations, the separation quality might be significantly reduced.

An alternative way of using MIDI information for initialization purposes was presented in [24]. Here, instead of generating synthetic audio, the MIDI file is used to directly instruct the underlying spectrogram model when a given instrument is active with a certain pitch. This way, the separation performance does not depend on the quality of an underlying MIDI synthesizer. However, as a drawback, no expectations about the spectral shape of an instrument are incorporated, which may lead to a less robust separation process. As a novel contribution, the method employs a parametric NMF variant [23], which significantly enhances the modeling accuracy for instruments with vibrato and glissando. A technical limitation of this model is that all harmonic sounds in an analysis frame are assumed to be a compound of stationary sinusoidals. To evaluate the instrument separation quality of this approach, the authors neglect the alignment step and employ synthetic MIDI sonifications of Bach, Beethoven and Boccherini pieces.

While most score-informed source separation techniques aim at re-synthesizing the separation results with the goal to produce acoustically appealing sound sources, the method proposed in [13] employs these techniques for analysis purposes. Given a MIDI file and an audio recording for a piece of music, the task consists of estimating an intensity for each MIDI note event as occurring in the recording. On the one hand, this enables a user to analyze and compare different interpretations of a piece in terms of dynamics on a note-level. On the other hand, it allows for enriching a given score-like MIDI representation with performance-specific subtleties. The approach employs a parametric model that describes the spectrogram of a given recorded performance as a sum of note-event spectrograms, see Fig. 4. In a first step, the model is initialized with pitch, onset and duration information obtained from a given score-like MIDI file, see Fig. 4b. After that, music synchronization techniques are employed to determine for each note event the corresponding position in the audio recording, see Fig. 4c. In a next step, additional model parameters are iteratively refined such that the model spectrogram approximates the original spectrogram as accurately as possible, see Fig. 4d. In a final step, the individual note intensities are estimated using the adapted note-event spectrograms described by the model. The approach is evaluated based on audio and MIDI velocity values recorded via a Yamaha Disklavier. The influence of

the synchronization step is evaluated by artificially distorting the MIDI time information, which only roughly indicates the methods' performance for real-world score-like MIDI files.

As demonstrated in [14], a similar model can also be used to create acoustically appealing separation results. Here, the separation system is embedded into a multimodal music interface [6] to create a voice equalizer, see Figure 2b, which allows the user to intuitively select arbitrary note groups and attenuate or emphasize them in real-time. To demonstrate the applicability of this approach in real-world scenarios, the authors employ score-like MIDI files from the Mutopia Project[1] in combination with real audio recordings taken from the SMD [34] and European archive[2] databases and make their separation results available on a website[3]. One of the drawbacks of this system and the one proposed in [27] is that the separation has to be performed in advance, while the remixing step can be performed in real-time.

As demonstrated by Duan and Pardo in [10], the separation step can be performed in a low-delay real-time fashion. To this end, the authors replace the usually employed offline synchronization step by an online approach [11], which aligns a given MIDI file and a corresponding audio recording in real-time, a task often referred to as score-following [4, 7]. For each analysis frame, their separation system first estimates the exact fundamental frequency of each pitch using the aligned MIDI file as a guidance. In a next step, each pitch is extracted using a harmonic mask and assigned to one of the instruments as specified by the MIDI file. To make this process feasible in real-time, the mask is computed using a fixed overtone model, which is not adapted to a given recording.

Overall, while source separation has been a field of research for decades, using score information to guide the separation process is a relatively recent approach. As demonstrated by the contributions discussed in this section, score guidance allows for novel and innovative applications of source separation techniques. Furthermore, the additional musical cues provided by the score often allow for a gain in separation quality, which is difficult to achieve otherwise. Here, robust music synchronization techniques allow for using score-informed source separation methods in real-world scenarios, where usually no perfectly aligned MIDI file is available. In the next section, we give an impression of how score-informed source separation can be performed in practice.

3 Score-Informed Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) has turned out to be a powerful tool for modeling, analyzing and separating the constituent parts of polyphonic music recordings. For example, NMF variants form the basis of methods for pitch estimation [2, 44], source separation [50], and pattern and motive identification [51]. However, using classic NMF it is often hard to predict which properties of the input are captured after the learning process. In this section, we show how the classical NMF framework can be extended in a straightforward way using available score data. As we will see, the basic idea is to replace the standard NMF initialization without changing the established and computationally efficient NMF learning process. This way, a musically meaningful factorization structure can be enforced, which stabilizes NMF-based source separation.

[1] http://www.mutopiaproject.org
[2] http://www.europarchive.org
[3] http://www.mpi-inf.mpg.de/resources/MIR/2011-ISMIR-VoiceSeparation/

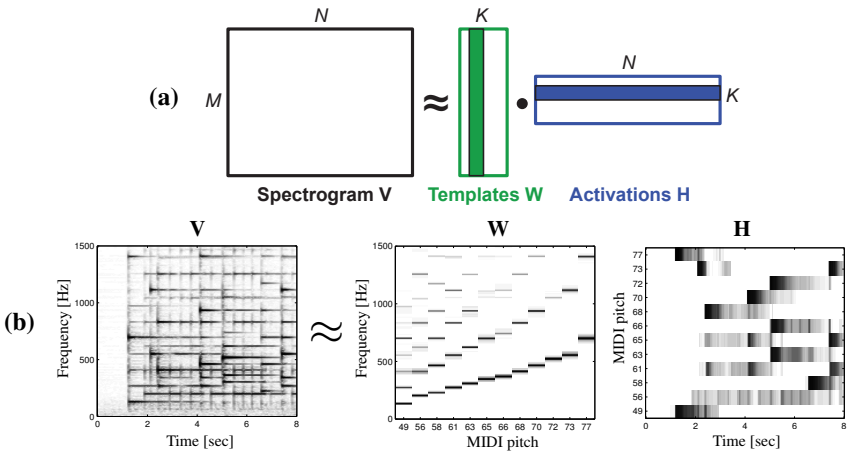

Figure 5 Non-negative matrix factorization (NMF). **(a)** A given non-negative matrix V is approximated as a product of two non-negative matrices W and H typically having a much smaller rank. **(b)** Example factorization of a magnitude spectrogram for an audio recording of Chopin's Op. 28 No. 15 taken from the SMD database [34].

3.1 Non-Negative Matrix Factorization

In classic non-negative matrix factorization, one approximates a spectral representation of a given recording by a product of two non-negative matrices. More exactly, given a magnitude spectrogram $V \in \mathbb{R}_{\geq 0}^{M \times N}$ of a music recording, NMF seeks to find non-negative matrices $W \in \mathbb{R}_{\geq 0}^{M \times K}$ and $H \in \mathbb{R}_{\geq 0}^{K \times N}$ such that $V \approx W \cdot H$, see Figure 5a. In this context, the columns of W are often referred to as *template vectors* and the rows of H as the corresponding *activations*. As an example, Figure 5b shows a factorization for a recording of Chopin's Op. 28 No. 15. Here, the free parameter K is set to the number of pitches that occur in the corresponding part of the piece. In this case, the activation matrix H is similar to a pianoroll representation and shows when these pitches become active.

In the classical approach for computing such a factorization, one employs some form of gradient descent to minimize a distance measure $D(V, W \cdot H)$ with respect to W and H, where D is typically based on the Euclidean norm or a variant of the Kullback-Leibler divergence, see [31]. However, to account for the non-negativity constraints for W and H, one usually has to resort to rather complex optimization algorithms [35]. As an easy-to-implement alternative, Lee and Seung proposed multiplicative update rules, which are derived from gradient descent by choosing a specific step size [31]. Using the popular Kullback-Leibler variant as a distance measure, these rules can be written as

$$H_{kn} \leftarrow H_{kn} \frac{\sum_i W_{ik} V_{in} / (WH)_{in}}{\sum_j W_{jk}} \quad \text{and} \quad W_{mk} \leftarrow W_{mk} \frac{\sum_i H_{ki} V_{mi} / (WH)_{mi}}{\sum_j H_{kj}},$$

where $m \in [1:M] := \{1, 2, \ldots, M\}$, $n \in [1:N]$, and $k \in [1:K]$. For vectorized programming languages such as Matlab it is useful to express these rules in matrix notation:

$$H \leftarrow H \odot \frac{W^\top \cdot (\frac{V}{W \cdot H})}{W^\top \cdot J} \quad \text{and} \quad W \leftarrow W \odot \frac{(\frac{V}{W \cdot H}) \cdot H^\top}{J \cdot H^\top},$$

where the $\cdot$ operator denotes the usual matrix product, the $\odot$ operator denotes the Hadamard product (point-wise multiplication), $J \in \mathbb{R}^{M \times N}$ denotes the matrix of ones, and the division is understood pointwise. These multiplicative update rules have several interesting

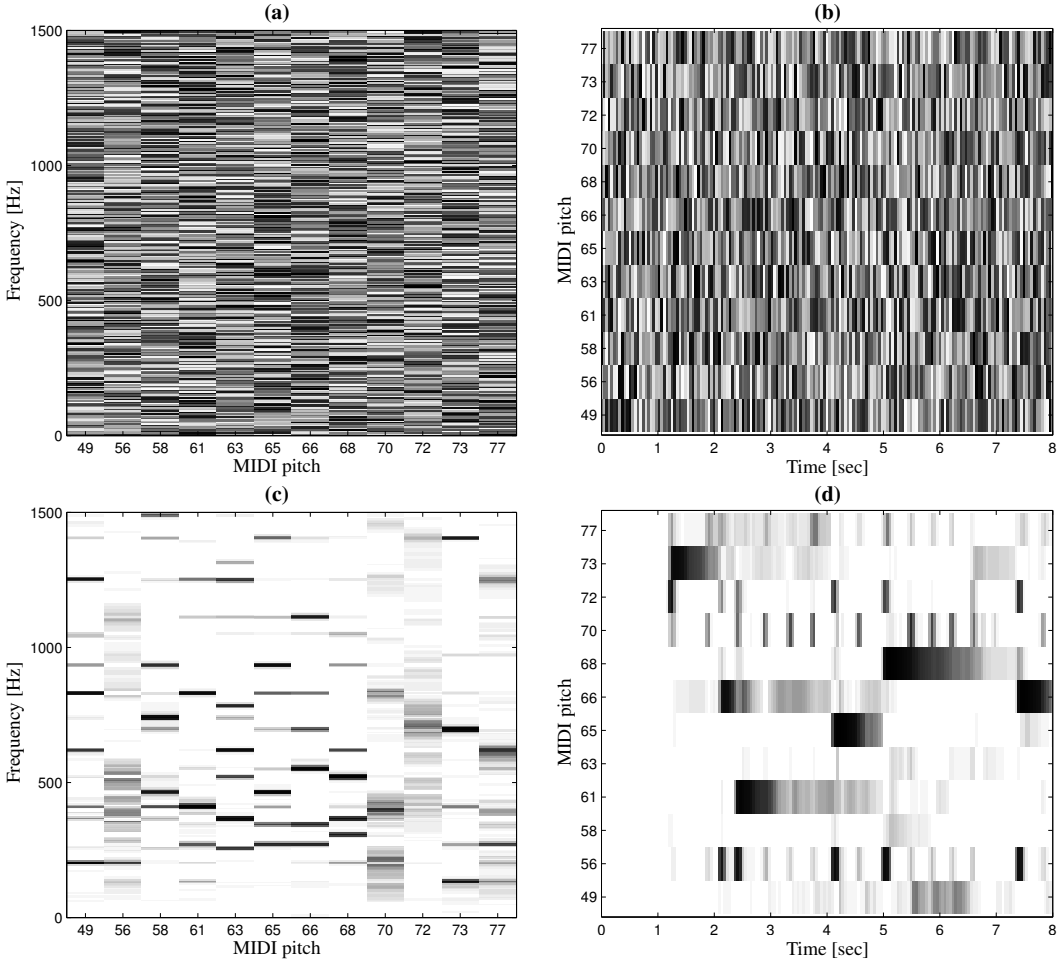

Figure 6 Classical NMF factorization for the magnitude spectrogram shown in Figure 5. **(a)** Random initialization of W. **(b)** Random initialization of H. **(c)** Learnt W. **(d)** Learnt H.

properties. First, the Kullback-Leibler distance measure is non-increasing under these rules[4]. Furthermore, initializing W and H with non-negative random values, these rules guarantee that W and H remain non-negative during the entire learning process.

In general, however, NMF factorizations computed in this classical way can not be as easily interpreted as the example shown in Figure 5b. For example, Figure 6 shows a factorization based on the classical NMF algorithm for the magnitude spectrogram shown in Figure 5b (again using $K = 12$). Here, the initialization of W and H with random values does not lead to a musically meaningful structure in the computed factorization. Furthermore, the free parameter K is usually set according to simple rules of thumb that usually do not account for any musical prior knowledge. As a result, the factorization often becomes completely unpredictable and lacks clear musical semantics.

Another important property of multiplicative update rules is that zero-valued entries remain zero during the entire learning process. Combined with musically informed initialization

[4] As pointed out by several authors [1, 32, 56], however, multiplicative rules do not guarantee in general convergence to a local minimum of the employed distance measure.

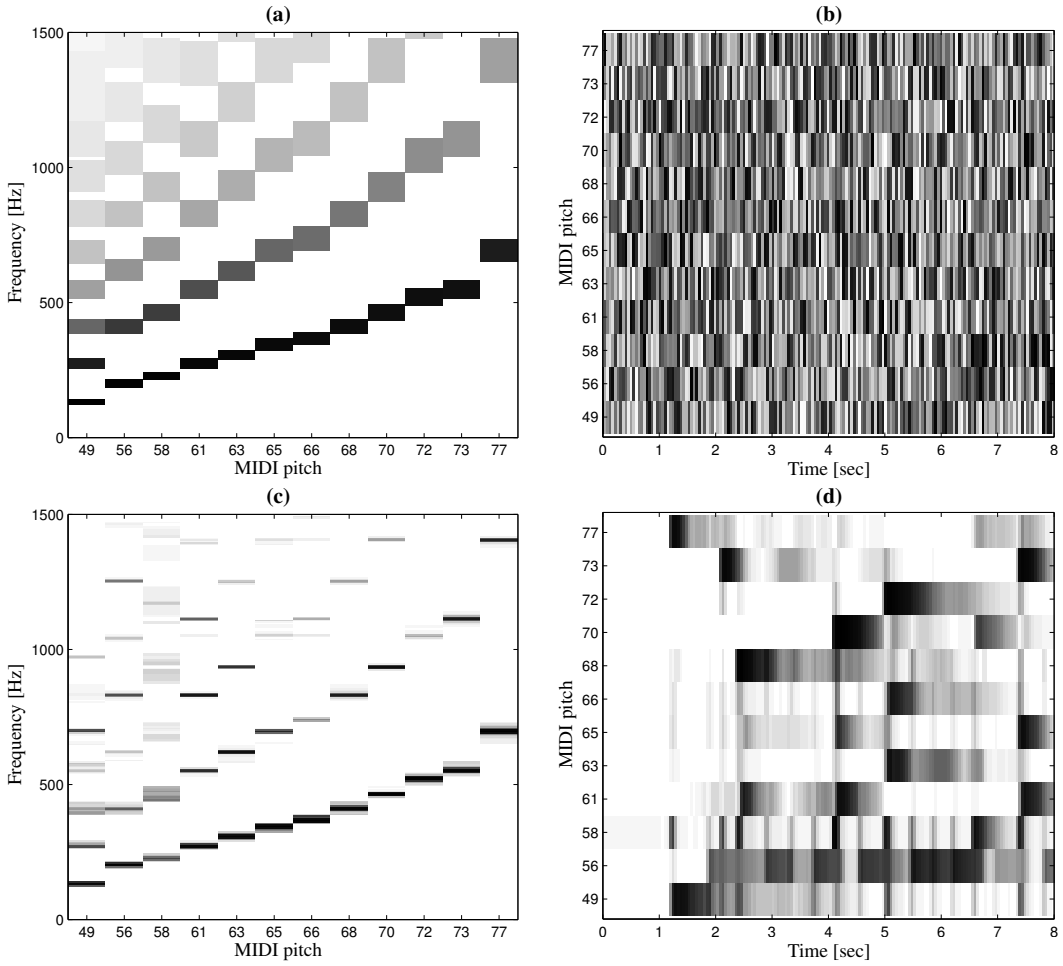

Figure 7 NMF factorization resulting from harmonic initialization of the template vectors for the magnitude spectrogram shown in Figure 5. **(a)** Harmonic initialization of W. **(b)** Random initialization of H. **(c)** Learnt W. **(d)** Learnt H.

schemes, this yields a straightforward way to enforce a specific structure of a factorization as proposed in [39, 49]. Here, one first creates one template vector for each possible MIDI pitch. Then, a harmonic structure is imposed by inserting zero-valued entries into the template initialization at positions where no partial is expected for a given pitch, see Figure 7a. The remaining entries are initialized according to a simplified overtone model. As we see in Figure 7c, the learning process based on multiplicative rules not only retains this harmonic structure but further refines it such that each template vector has a clear pitch association. This is a significant gain in structure compared to the unpredictable results computed via standard NMF as shown in Figure 6. However, looking at the resulting factorization in Figure 7c/d reveals that template vectors are still often 'misused', for example to represent onsets. This becomes particularly apparent in the template for MIDI pitch 58, where energy is distributed over a larger number of frequency bands compared to the other templates (Figure 7c). Here, instead of representing harmonic components of the spectrum, the template is misused to explain parts of the broadband energy distribution related to onsets. This is also reflected by the short-term intensity bursts in the corresponding activation row (Figure 7d).

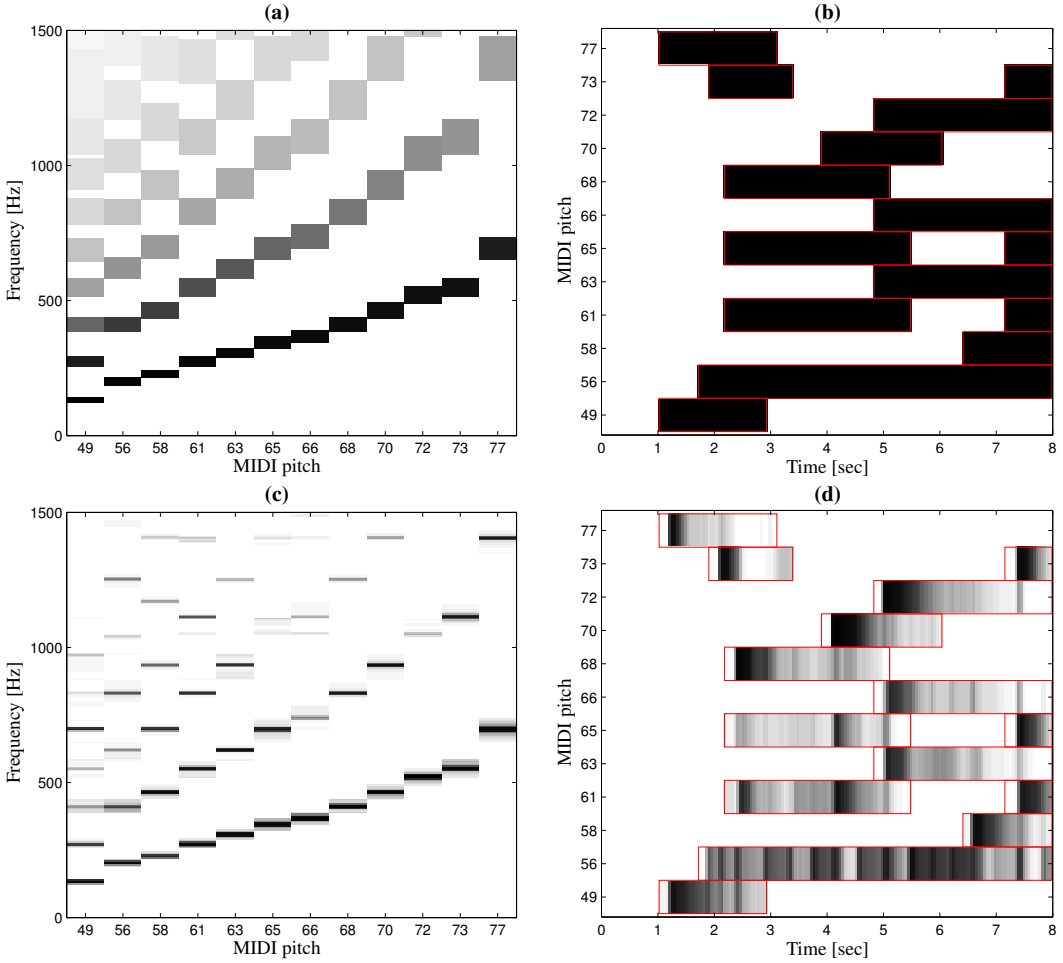

Figure 8 NMF factorization resulting from harmonic initialization of the template vectors and score-informed activation constraints for the magnitude spectrogram shown in Figure 5. **(a)** Harmonic initialization of W. **(b)** Score-informed initialization of H. **(c)** Learnt W. **(d)** Learnt H.

3.2 Integrating Score Information

Possible ways to further stabilize the factorization by incorporating additional score information were investigated in [15]. For example, in addition to the constraints on the template vectors, one can also impose constraints on the activations by incorporating note timing information. To generate such information, one employs music synchronization techniques in a first step to determine for each MIDI note event its corresponding position in the audio recording [16]. Next, based on the synchronized MIDI information, one marks suitable regions in H to determine where a given pitch can be active, see Figure 8b. The remaining entries are set to zero. To account for possible alignment inaccuracies, the temporal boundaries for these regions can be chosen relatively generous. As a result, the activation matrix H can be interpreted as a coarse piano roll representation of the synchronized MIDI file. As to be expected, combining these activation constraints with those for the template vectors further stabilizes the factorization. For example, most of the activation onset noise, which was present in Figure 7d, is suppressed in Figure 8d. Furthermore, almost all template vectors now have a well-defined harmonic structure. In some sense, the synchronization step can be

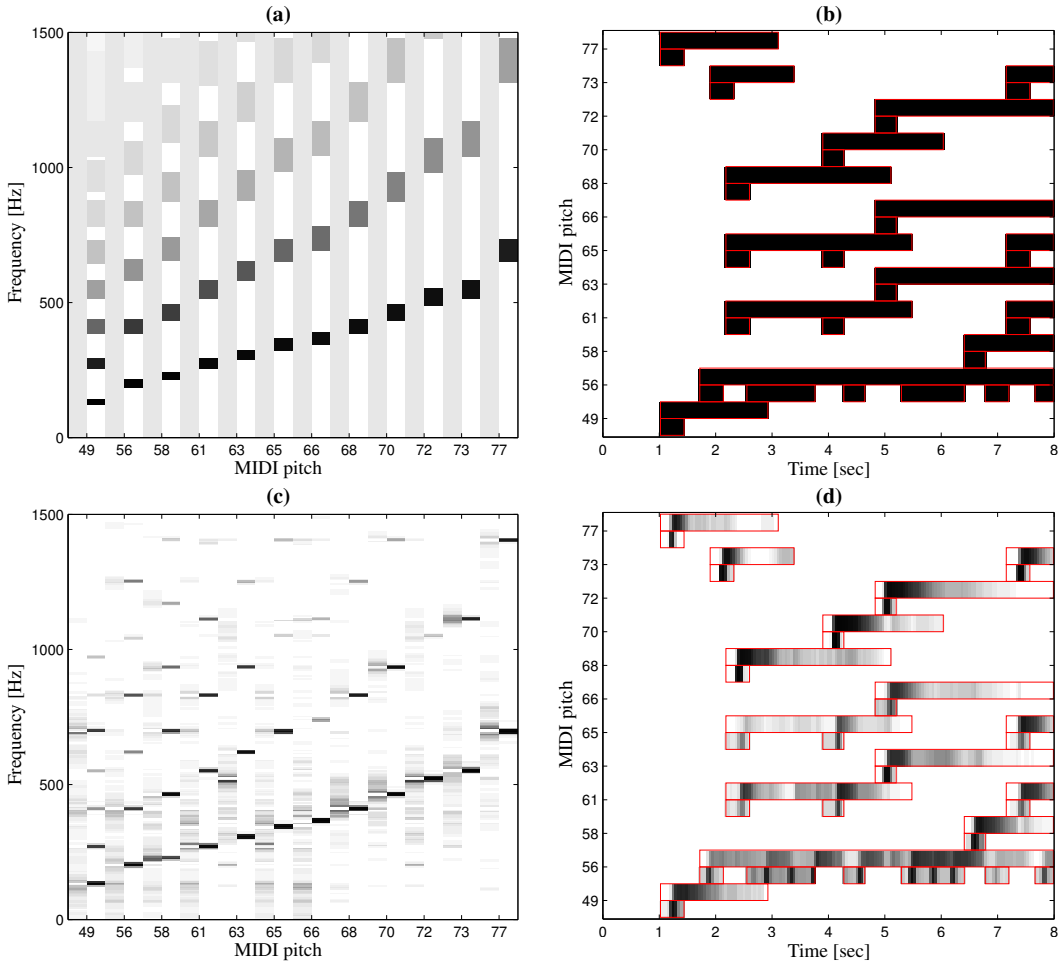

Figure 9 Extended NMF model with additional onset templates for the magnitude spectrogram shown in Figure 5. **(a)** Initialization of harmonic and onset template vectors in W. **(b)** Score-informed initialization of the corresponding activations in H. **(c)** Learnt W. **(d)** Learnt H.

seen to yield a first rough factorization, which is then refined by the NMF-based learning procedure.

So far, the model only represents harmonic parts of the signal and does not account for percussive elements such as onsets. Making again use of the score information, we extend the model by incorporating dedicated onset template vectors, see Figure 9a. Here, opposed to many other approaches, we take into account that the spectral shape for onsets is for many instruments (including the piano) not the same as for white noise but depends on the respective pitch. Therefore, instead of using one onset template jointly for all pitches as for example in [55], we use one onset vector for each pitch as suggested in [15]. Contrary to the harmonic templates, we do not enforce here any spectral constraints but initialize the onset templates uniformly and let the learning process derive their shape.

While the onset templates are hard to constrain in a meaningful way, the ephemeral nature of percussive sounds allows for imposing strict constraints on their activations. Using the synchronized MIDI, one has a rough estimate of the position of each onset. Initializing a small neighborhood around these positions to the value one in the corresponding activation while leaving all remaining entries at zero strongly restrains the time points where onset

templates are allowed to be active, see Figure 9b. Again, a tolerance should be used to compensate for possible synchronization inaccuracies. Looking at the resulting factorization shown in Figure 9c and 9d, we see that the learnt harmonic vectors have the clearest harmonic structure compared to all previous factorizations. Here, a reason is that percussive broadband energy is now captured by the onset templates, with the result that onsets now have a significantly less disturbing influence on the harmonic templates. Furthermore, the impulse-like activations of most onset templates at the start of note events indicate that these templates indeed represent onsets.

In summary, one can say that a combination of template and activation constraints leads to meaningful and robust matrix factorizations. Here, as for the case of the onset templates, constraints on the activation side can compensate for using relatively loose or even no constraints on the template side and vice versa. Furthermore, even though all constraints are hard in the sense that zero-entries in W and H remain zero throughout the learning process, one can use rather generous constraint regions to account for synchronization errors and retain some degree of flexibility. As one major advantage, the extended NMF model using hard constraints allows for using exactly the same multiplicative update rules as in classical NMF, thus it inherits the ease of implementability and computational efficiency.

3.3 Separation Process

By means of the initial constraint regions, a factorization as shown in Figure 9 describes how each note event of a given MIDI file manifests in the spectrogram of a corresponding audio recording. We now describe how this spectrogram model can be employed to separate note groups such as a melody line, the staff of the right hand, a specific motive, or the accompaniment from the recording. The only requirement is that the notes to be considered are somehow specified by the user or by some labeling of the score. As an illustrating example, we consider here the task of separating the left from the right hand staff as specified by a given score, see Figure 10a. While staffs do not always correspond to musically meaningful note groups, it demonstrates how note groups could be easily specified in a natural way.

For the separation, we exploit that every non-zero entry in H is associated with a specific note event, see Figure 9d. Therefore, we can partition H into two new matrices H_L and H_R, which contain either the activations for the left or the right hand, see Figure 10a/b. A straightforward way to create an audible separation result could be to multiply these two matrices with the template matrix W, shown in Figure 9c, and to invert the resulting spectrogram. However, as NMF-based models are typically used to compute a rough approximation of the original magnitude spectrogram spectral nuances in a given recording are usually not captured. Therefore, the resulting audio recording would sound rather unnatural.

An alternative to this direct sonification is commonly referred to as masking. Here, one first derives masking matrices via

$$M_L := \frac{WH_L}{WH+\varepsilon} \qquad \text{and} \qquad M_R := \frac{WH_R}{WH+\varepsilon},$$

where the division is understood pointwise and ε is a small positive constant to avoid a potential division by zero, see Figure 10c/d. M_L and M_R have the same size as the original spectrogram V and, having values between 0 and 1, indicate how strongly each entry in V belongs to either the left or the right hand. Multiplying these masking matrices point-wisely with V, one obtains a separated spectrogram for the left and the right hand, see Figure 10e/f. Finally, to obtain the separated audio signals, one applies an inverse

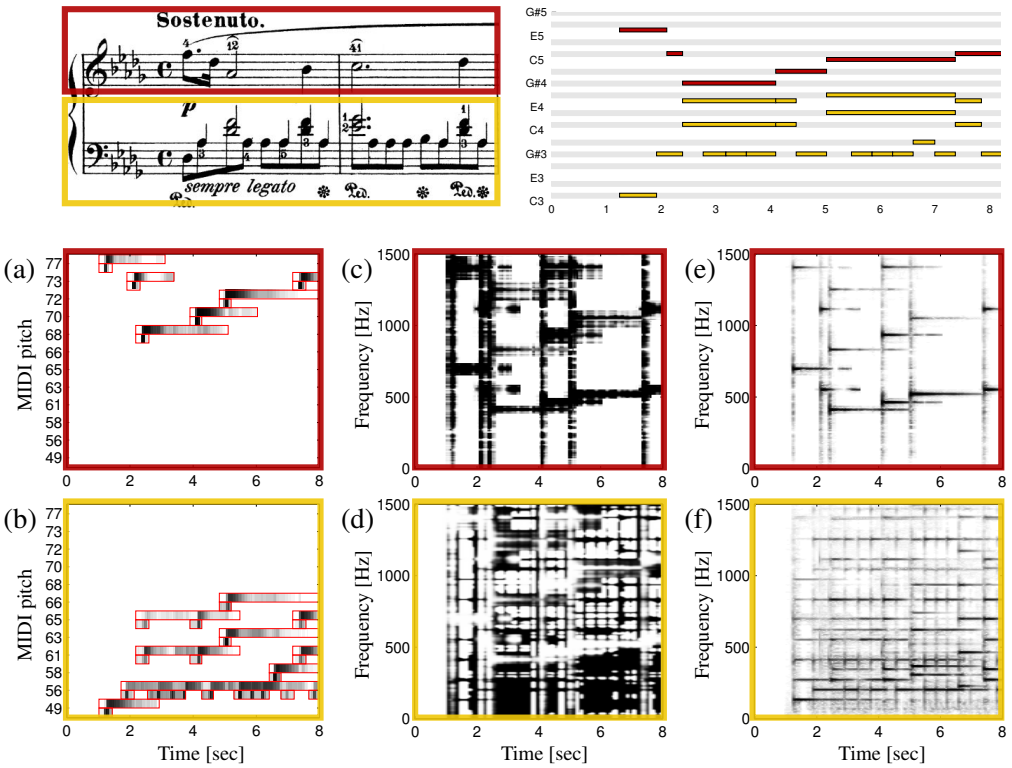

Figure 10 Illustration of the separation process for the left and the right hand. **(a)**/**(b)**: Partition of the activation matrix H (Figure 9d) into H_L and H_R. **(c)**/**(d)**: Masking matrices M_L and M_R. **(e)**/**(f)**: Separated spectrograms.

discrete Fourier transform in combination with an overlapp-add technique to the separated spectrograms. The necessary phase information is provided by the original spectrogram. This way, masking-based separation allows for preserving most spectral details of the original recording, which is important to create acoustically appealing results. However, by filtering the original audio data, masking may also retain more non-target spectrogram components compared to a direct sonification.

The quality of a separation result is often measured in terms of signal-to-distortion ratios (SDR) as proposed in [48]. While illustrating some general tendencies, these measures often do not capture the overall perceptual separation quality. In particular, in combination with synthetic audio material, one does not get an impression of the separation quality in real-world scenarios. To allow for a subjective, perceptual evaluation of their score-informed NMF variant, the authors in [15] provide a website[5] with separation results using real audio recordings and score-like MIDI files. Here, using full-length pieces by Bach, Beethoven and Chopin, most the audio material was taken from the SMD [34] dataset, while the MIDI files were provided by the Mutopia Project[6]. Some additional historical recordings were also taken from the European Archive[7]. To roughly indicate general quality differences between the NMF variants in a quantitative fashion, the authors also conducted experiments

[5] `http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF/`
[6] `http://www.mutopiaproject.org`
[7] `http://www.europarchive.org`

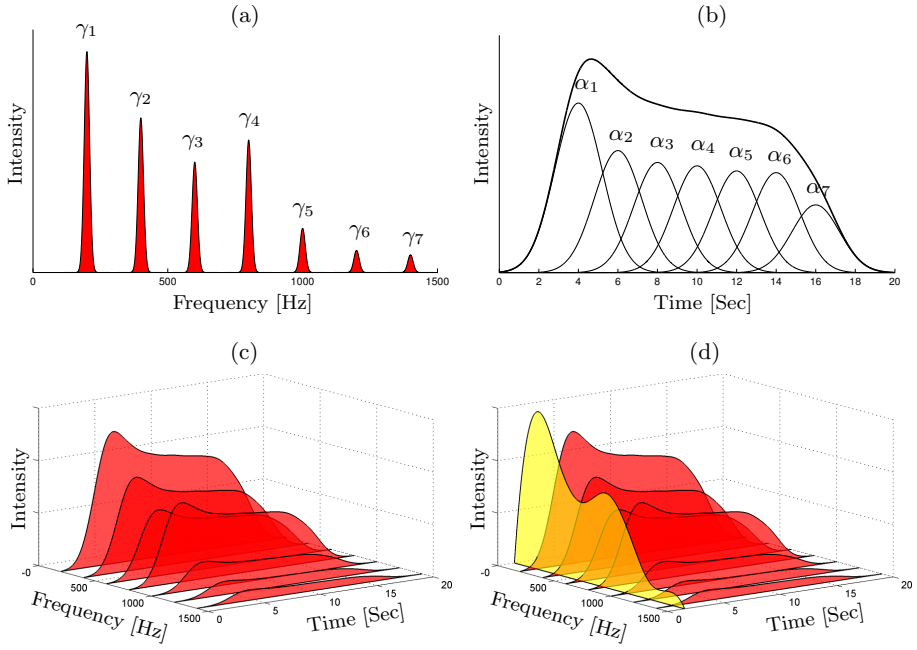

Figure 11 Harmonic-Temporal-Structured Clustering (HTC): **(a)** Template vector composed of several Gaussians. **(b)** Activation described by smooth, overlapping Gaussians. **(c)** Spectrogram model resulting from a combination of template vectors and activations similar to NMF. **(d)** Advanced HTC variant with an additional transient submodel. Figures are inspired by [55].

based on synthetic audio and the SDR measure. Here, on average, the strategy based on the harmonic initialization of W yielded the lowest SDR value. Combining this strategy with the score-informed initialization of H as in Figure 8 leads to a significant SDR-gain of roughly 1.5 dB. Finally, additionally integrating onset templates leads to another gain of roughly 1.2 dB.

4 Parametric Models

In addition to NMF, there are numerous other classical source separation methods which allow for the integration of score information. Many of the approaches discussed in Section 2 are based on so called *parametric models* [13, 14, 20, 24, 27], which have been widely used for blind source separation and music transcription. While these approaches differ strongly in their details, the common idea is to adapt a set of parameters such that the underlying model explains the spectrogram of a given recording as accurately as possible. Here, typical parameters are related to acoustical and musical properties such as pitch, amplitude and timbre. In this section, we exemplarily discuss some aspects of the *harmonic-temporal-structured clustering model (HTC)* [55], which was employed in [26, 27] for score-informed source separation. After a brief description of the main ideas underlying the HTC approach, we summarize some conceptual differences to the NMF model.

4.1 Harmonic-Temporal-Structured Clustering (HTC)

Harmonic-Temporal-Structured Clustering (HTC) employs a parametric model to approximate the magnitude spectrogram of a given audio recording. Compared to NMF, specialized

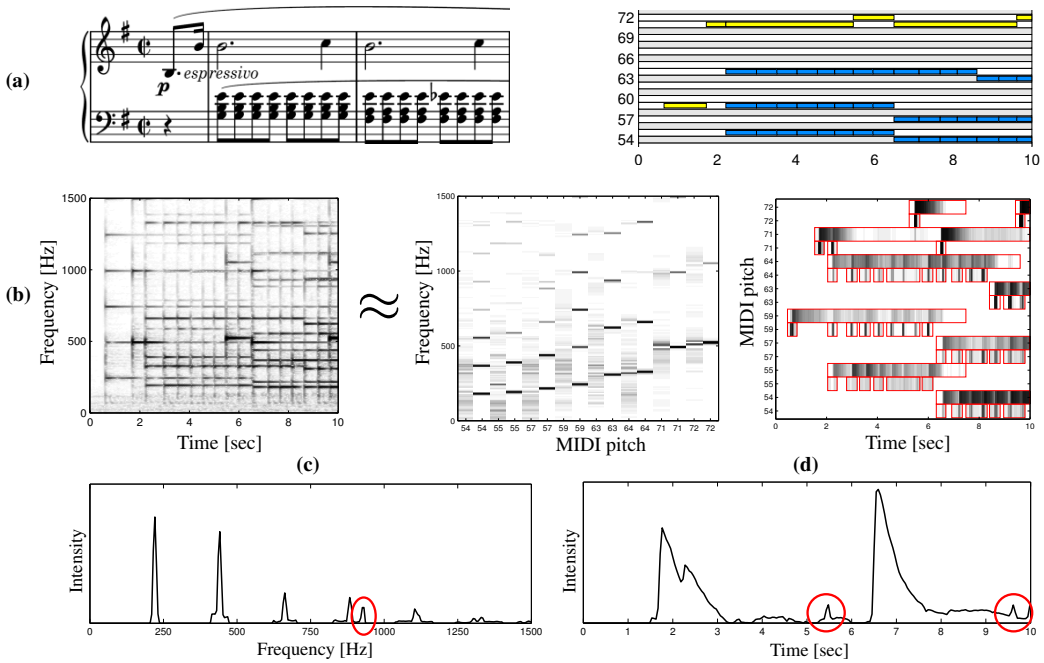

Figure 12 Score-informed NMF factorization for a recording of Chopin's Op. 28 No. 4 taken from the SMD database [34]. **(a)** Score and MIDI representation. **(b)** NMF factorization computed using the method presented in Section 3. **(c)** Zoomed template vector for pitch 57. **(d)** Zoomed activation for pitch 71. The red markers indicate positions discussed in the text.

model components take over the role of template vectors and activations. For example, each HTC template consists of several Gaussians, which represent the partials of a harmonic sound, see Figure 11a. To adapt the model to different instruments and their specific overtone energy distribution, the HTC model allows for scaling the height of each Gaussian individually using a set of parameters ($\gamma_1, \ldots, \gamma_7$ in Figure 11a). An additional parameter f_0 specifies the fundamental frequency for the template. Assuming a harmonic relationship between the overtones, this parameter controls the exact location of each partial.

Gaussians are also used in HTC to represent the activations, see Figure 11b. The position of these Gaussians is typically fixed such that only suitable height parameters can be adapted (parameters $\alpha_1, \ldots, \alpha_7$ in Figure 11b). By choosing suitable values for the variances and positions of the Gaussians, one obtains an overall smooth activation progression. Combining the HTC templates and activations similar to NMF, one obtains a spectrogram model as shown in Figure 11c. Recently proposed extensions of this model even allow for an integration of transient and onset models [27, 55], see Figure 11d. Again using a smoothed representation based on Gaussians, these additional models represent the broadband energy distribution usually found at onset positions.

Since the HTC model follows similar ideas as NMF, one can also employ similar strategies to incorporate score information. For example, note timing information can be used to restrict the use of the activation parameters. Furthermore, MIDI pitch information can be used to set the number of templates in the HTC model to the actual number of pitches in the piece. This is similar to setting the value of the free parameter K in NMF.

4.2 Comparison between HTC and NMF

To compare the HTC model with NMF, we consider an NMF factorization for an audio recording of Chopin's Prélude No. 4. Using the method presented in Section 3, one obtains a factorization as shown in Figure 12b. Here, similar to Figure 9, we see that almost all learnt harmonic template vectors have a well-defined harmonic structure. For a closer inspection of an exception, we plot the template for pitch 57 as a function over frequency in Figure 12c. We see a small peak at 930 Hz (see red marking), which does not fit into the harmonic pattern. Enforcing a meaningful distance between partials, the Gaussian-based HTC model offers here a straightforward way to enforce a clear harmonic relationship. However, this additional robustness against spurious peaks comes at the cost of model inaccuracies. One reason is that partials almost never perfectly take the form of a Gaussian, see [38], such that the HTC model leads to an additional inevitable approximation error.

Furthermore, the approximation accuracy does not only depend on the templates but also on the activations. To give an example, we plot the activation for pitch 71 as a function over time in Figure 12d. Here, we see three distinct peaks at 1.8, 2.3 and 6.6 seconds, respectively, which correspond to the three middle B notes, see Figure 12a. However, there are additional, smaller peaks at 5.4 and 9.6 seconds (marked in red), which do not seem to make any musical sense. Using Gaussians spanning several frames to model the activation, such short-time irregular peaks are smoothed out. However, whether this is meaningful depends on the application. In Figure 12a, we see that a note event with pitch 71 (middle B) is played after 2.3 seconds and is held afterwards. Then, after 5.4 seconds, a note event with pitch 72 is played. Since in this recording all piano dampers are up, the consequence is that the onset of pitch 72 also results in excitations of the neighboring pitches, in particular of the strings of pitch 71. Therefore, the small peak at 5.4 seconds in the NMF activation is indeed a physical fact rather than an extraction error.

5 Conclusion

Music signals possess specific characteristics that are not shared by spoken speech or audio signals from other domains. For example, for sound mixtures of polyphonic music, the general assumption that sources are somehow orthogonal in the spectral domain is often violated. This makes the separation of musical sources or voices very difficult. To remedy this problem, various approaches have been suggested that use additional cues as specified by a musical score.

In this paper, we have given a comprehensive overview of state-of-the-art source separation techniques that exploit additional score information in various ways. In particular, we discussed in detail a score-informed variant of NMF, where the integration of constraints can be done in a straightforward manner already at the initialization stage. We showed that by constraining both the template vectors and the activations, one obtains robust and musically meaningful separation results. Opposed to parametric models, where the integration of additional priors often leads to an increase in the computational complexity, score-informed NMF variants employ the same update rules as the original NMF and inherit its computational efficiency.

Besides stabilizing the separation process, the availability of score information also facilitates a natural and user-friendly way of specifying the voices or note groups to be separated. This opens up new ways for audio editing applications, where a user can simply mark certain note groups within a visual representation of the score, which are then separated, removed, amplified, or attenuated in a corresponding music recording. For the future, we plan to develop multimodal interfaces that realize such functionalities.

So far, we have conducted experiments mainly on piano music. In this context, we showed how the score-informed NMF framework can be extended by integrating additional onset templates without sacrificing robustness. A promising research direction is to further expand the NMF model to account for other musical aspects such as timbre or instrumentation and then to apply the NMF framework to other types of music.

6 Acknowledgment

We would like to express our gratitude to Björn Schuller, Joan Serrà, and Steve Tjoa for their helpful and constructive feedback.

References

1 Roland Badeau, Nancy Bertin, and Emmanuel Vincent. Stability analysis of multiplicative update algorithms for non-negative matrix factorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2148–2151, Prague, Czech Republic, 2011.

2 Nancy Bertin, Roland Badeau, and Emmanuel Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):538–549, 2010.

3 Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

4 Arshia Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):974–987, 2010.

5 David Damm, Christian Fremerey, Frank Kurth, Meinard Müller, and Michael Clausen. Multimodal presentation and browsing of music. In *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI)*, pages 205–208, Chania, Crete, Greece, 2008.

6 David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, and Meinard Müller. A digital library framework for heterogeneous music collections—from document acquisition to cross-modal interaction. *International Journal on Digital Libraries: Special Issue on Music Digital Libraries*, 2011, to appear.

7 Roger B. Dannenberg and Christopher Raphael. Music score alignment and computer accompaniment. *Communications of the ACM, Special Issue: Music information retrieval*, 49(8):38–43, 2006.

8 Simon Dixon and Gerhard Widmer. MATCH: A music alignment tool chest. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, GB, 2005.

9 Karin Dressler. An auditory streaming approach for melody extraction from polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 19–24, Miami, USA, 2011.

10 Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1205–1215, 2011.

11 Zhiyao Duan and Bryan Pardo. A state space model for online polyphonic audio-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 197–200, Prague, Czech Republic, 2011.

12 Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):564–575, 2010.

13 Sebastian Ewert and Meinard Müller. Estimating note intensities in music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 385–388, Prague, Czech Republic, 2011.

14 Sebastian Ewert and Meinard Müller. Score-informed voice separation for piano recordings. In *Proceedings of the International Society for Music Information Retrieval Conference(ISMIR)*, pages 245–250, Miami, USA, 2011.

15 Sebastian Ewert and Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

16 Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, 2009.

17 Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel. Evaluation of a score-informed source separation system. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 219–224, Utrecht, The Netherlands, 2010.

18 Joachim Ganseman, Paul Scheunders, Gautham J. Mysore, and Jonathan S. Abel. Source separation by score synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 462–465, New York, USA, 2010.

19 Masataka 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)*, 43(4):311–329, 2004.

20 Yushen Han and Christopher Raphael. Desoloing monaural audio using mixture models. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 145–148, Vienna, Austria, 2007.

21 Yushen Han and Christopher Raphael. Informed source separation of orchestra and soloist. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, Utrecht, The Netherlands, 2010.

22 Toni Heittola, Anssi P. Klapuri, and Tuomas 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)*, pages 327–332, Kobe, Japan, 2009.

23 Romain Hennequin, Roland Badeau, and Bertrand David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, 2010.

24 Romain Hennequin, Bertrand David, and Roland 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)*, pages 45–48, Prague, Czech Republic, 2011.

25 Ning Hu, Roger B. Dannenberg, and George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2003.

26 Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages I–57–I–60, Hawaii, USA, 2007.

27 Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Instrument equalizer for query-by-example retrieval: Improving sound source separ-

ation based on integrated harmonic and inharmonic models. In *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*, pages 133–138, Philadelphia, USA, 2008.

28 Cyril Joder, Slim Essid, and Gaël Richard. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. In *Proceedings of the 35nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010.

29 Cyril Joder, Felix Weninger, Florian Eyben, David Virette, and Björn Schuller. Real-time speech separation by semi-supervised nonnegative matrix factorization. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, 2012.

30 Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Harmonic-temporal-structured clustering via deterministic annealing EM algorithm for audio feature extraction. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 115–122, London, GB, 2005.

31 Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 556–562, Denver, CO, USA, 2000.

32 Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18:1589–1596, 2007.

33 Meinard Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

34 Meinard Müller, Verena Konz, Wolfgang Bogler, and Vlora Arifi-Müller. Saarland music data (SMD). In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session*, 2011.

35 Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer (Springer Series in Operations Research and Financial Engineering), 2006.

36 Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 139–144, Philadelphia, Pennsylvania, USA, 2008.

37 Mark D. Plumbley, Samer A. Abdallah, Juan Pablo Bello, Mike E. Davies, Giuliano Monti, and Mark B. Sandler. Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6):603–627, 2002.

38 John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processsing*. Prentice Hall, 1996.

39 Stanislaw Andrzej Raczynski, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 381–386, 2007.

40 Lise Regnier and Geoffroy Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1685–1688, Taipei, Taiwan, 2009.

41 Shai Shalev-Shwartz, Shlomo Dubnov, Nir Friedman, and Yoram Singer. Robust temporal and spectral modeling for query by melody. In *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 331–338, Tampere, Finland, 2002.

42 Adiel Ben Shalom, Shai Shalev-Shwartz, Michael Werman, and Shlomo Dubnov. Optimal filtering of an instrument sound in a mixed recording using harmonic model and score alignment. In *Proceedings of the International Computer Music Conference (ICMC)*, Miami, USA, 2004.

43 Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis. Probabilistic latent variable models as nonnegative factorizations (article id 947438). *Computational Intelligence and Neuroscience*, 2008.

44 Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, 2003.

45 Paris Smaragdis and Gautham 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)*, pages 69–72, New Paltz, NY, USA, 2009.

46 Robert J. Turetsky and Daniel P.W. Ellis. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 135–141, Baltimore, USA, 2003.

47 Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5518–5521, Dallas, USA, 2010.

48 Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

49 Tuomas Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, 2006.

50 Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

51 Ron J. Weiss and Juan Pablo Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 123–128, Utrecht, The Netherlands, 2010.

52 John Woodruff and Bryan Pardo. Using pitch, amplitude modulation, and spatial cues for separation of harmonic instruments from stereo music recordings (article id 86369). *EURASIP Journal on Advances in Signal Processing*, 2007.

53 John Woodruff and Bryan Pardo. Active source estimation for improved source separation. Technical Report NWU-EECS-06-01, Electrical Engineering and Computer Science Department, Northwestern University, 2006.

54 John Woodruff, Bryan Pardo, and Roger B. Dannenberg. Remixing stereo music with score-informed source separation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 314–319, 2006.

55 Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. Multipitch estimation by joint modeling of harmonic and transient sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 25–28, Prague, Czech Republic, 2011.

56 Shangming Yang and Zhang Yi. Convergence analysis of non-negative matrix factorization for BSS algorithm. *Neural Processing Letters*, 31:45–64, 2010.

View publication stats