Score-Informed Source Separation for Musical Audio Recordings: An Overview

Sebastian Ewert[⋆] *Bryan Pardo*[†] *Meinard Müller*[⊤] *Mark D. Plumbley*[⋆]

[⋆]Queen Mary University of London, London, United Kingdom
[†]Northwestern University, Evanston, IL, USA
[⊤] International Audio Laboratories Erlangen, Erlangen, Germany

In recent years, source separation has been a central research topic in music signal processing, with applications in stereo-to-surround up-mixing, remixing tools for DJs or producers, instrument-wise equalizing, karaoke systems, and pre-processing in music analysis tasks. Musical sound sources, however, are often strongly correlated in time and frequency, and without additional knowledge about the sources a decomposition of a musical recording is often infeasible. To simplify this complex task, various methods have been proposed in recent years which exploit the availability of a musical score. The additional instrumentation and note information provided by the score guides the separation process, leading to significant improvements in terms of separation quality and robustness. A major challenge in utilizing this rich source of information is to bridge the gap between high-level musical events specified by the score and their corresponding acoustic realizations in an audio recording. In this article, we review recent developments in score-informed source separation and discuss various strategies for integrating the prior knowledge encoded by the score.

1 Introduction

In general, audio source separation methods often rely on assumptions such as the availability of multiple channels (recorded using several microphones) or the statistical independence of the source signals, to identify and segregate individual signal components. In music, however, such assumptions are not applicable in many cases. For example, musical sound sources often outnumber the information channels, such as a string quartet recorded in two-channel stereo. Also, sound sources in music are typically highly correlated in time and frequency: Instruments follow the same rhythmic patterns and play notes which are harmonically related. Purely statistical methods such as Independent Component Analysis (ICA) or Non-negative Matrix Factorization (NMF) therefore often fail to completely recover individual sound objects from music mixtures [1].

High-quality source separation for general music remains an open problem. One

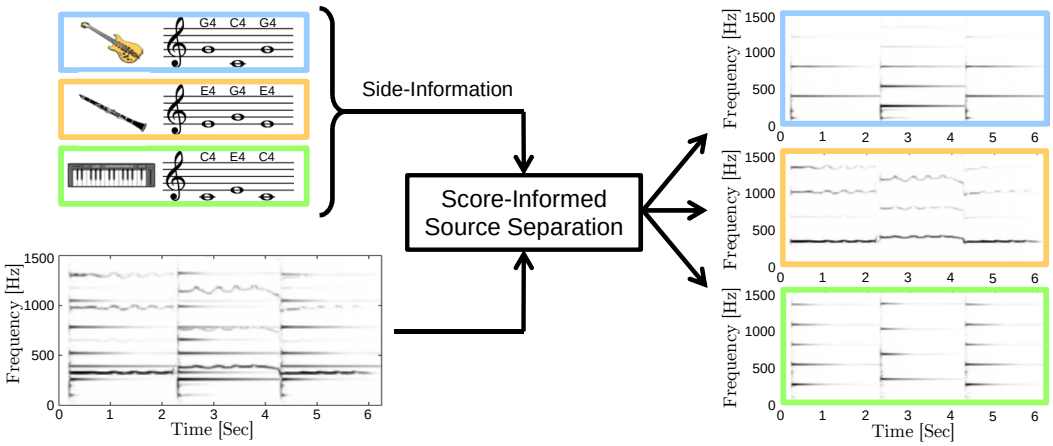


Figure 1: Score-informed source separation: Instrument lines as specified by a musical score (upper left) are employed as prior knowledge for the decomposition of a mixture audio recording (lower left) into individual instrument sounds (right). The mixture consists of a guitar (blue), a clarinet (orange) and a piano (green).

approach is to exploit known spectro-temporal properties of the sources to facilitate the segregation [1,2]. For example, in a time-frequency representation, percussive instruments typically exhibit structures in the frequency direction (short bursts of broadband energy) while harmonic instruments usually lead to structures in the time direction (slowly changing harmonics). Many instruments, however, emit similar energy patterns and thus they are hard to distinguish based on spectro-temporal characteristics alone. To overcome these problems, various approaches presented in recent years exploit (user-generated) annotations of a recording as additional prior knowledge. For example, to simplify the separation process, one can specify the fundamental frequency of instruments [3], manually assign harmonics in a spectrogram to a specific source [4], or provide timing information for instruments [5, 6]. However, while such annotations typically lead to a significant increase in separation performance, their creation can be a laborious task.

In this article, we focus on a natural and particularly valuable source of prior knowledge which exists for many pieces: a musical score. The score contains information about the instruments and notes of the musical piece, and can be used to guide and simplify the separation process even if the sources are hard to distinguish based on their spectro-temporal behaviour. In particular, information about pitch and timing of note events can be used to locate and isolate corresponding sound events in the audio mixture (Fig. 1). For example, note events for a guitar, clarinet and piano (Fig. 1, upper left) can be used to direct the extraction of corresponding instrument sounds from a given recording (Fig. 1, right). Knowledge about the instrumentation can also aid in

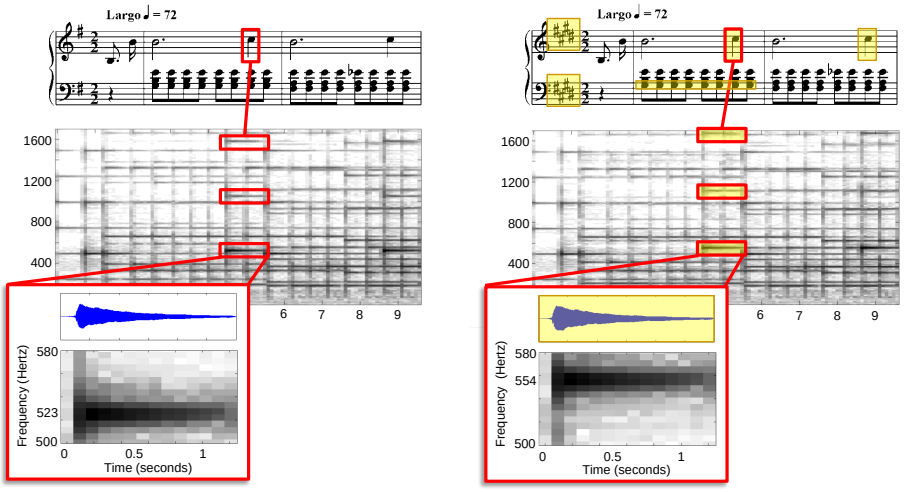


Figure 2: Score-informed audio editing (see [7]). (Left): For each note in the score, the corresponding sound is extracted from a recording of Chopin's Op. 28 No. 4. (Right): By applying pitch-shifting techniques to the individual notes, the piece is changed from minor to major.

selecting appropriate source models or training data. For example, the spectro-temporal characteristics of the clarinet (Fig. 1, right middle) are different from those of the piano, and should be modelled accordingly.

The score also gives an intuitive and user-friendly representation for musically experienced users to specify the target sources to be separated. For example, by partitioning the score into groups of note events, one can easily specify that the main melody should be separated from the accompaniment, or that all string instruments should be separated from the wind instruments. This concept led to novel ideas and application scenarios in the context of instrument-wise equalization [8], personal music remixing [9], music information retrieval [10], and intelligent audio editing [7]. Fig. 2 gives an example, where a user can easily specify the desired audio manipulation within the score simply by editing some of the notes. These manipulations are then automatically transferred to a given audio recording using score-informed audio parametrization techniques [7][1]. Additionally, applications such as singing voice removal for karaoke [11] or parametric coding of audio objects [12] can significantly benefit from the increase in separation robustness resulting from the integration of score.

While integrating score information bears the potential for a significant gain in separation quality, dealing with real data remains a major issue[2]. In particular, score-

[1] Demo website with videos: `http://www.audiolabs-erlangen.de/resources/2013-ACMMM-AudioDecomp/`

[2] Demo websites using non-synthetic data: `http://www.ece.rochester.edu/~zduan/jstsp2011/examples.html` [13], `http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF/`

informed separation methods often have only been tested on recordings synthesized from the score, such that many practical issues are not reflected in the test data. In a real world scenario, a score specifies relative positions for note events on a musical time and pitch grid using an abstract, high-level language with a lot of leeway for interpretation by a performer. The score specifies neither exact frequencies nor the precise timing and duration of the musical tones. Also, the timbre and the loudness are only specified in terms of coarse instructions such as "forte" meaning "loud". Additionally, a musician may deviate from the score by adding extra notes (ornaments and grace notes), or there may be playing errors or even structural differences such as skipped sections. Further, while full scores are freely available for many classical pieces as a result of substantial digitization efforts[3], there are often only so-called lead sheets available for pop music, which only specify parts of the score including the melody, lyrics and harmony. Altogether, such issues and uncertainties lead to significant challenges in score-informed source separation, which current approaches have just started to address.

In the following, we begin with a description of issues in applying standard source separation techniques, such as Non-Negative Matrix Factorization (NMF), to music signals and we explain how score-information can be integrated into NMF-based procedures. We then discuss methods for time-aligning the score and corresponding audio data, and strategies for dealing with frequency changes such as vibrato and frequency drifts. After presenting a strategy for separating instruments based on sound examples that are synthesized from the score, we discuss further extensions to these approaches and conclude with a look at potential future research directions.

2 Using NMF for Source separation

Among the various methods for blind source separation, Non-Negative Matrix Factorization (NMF) has been one of the most successful [16]. The method is easy to implement, is computationally efficient, and has been successfully applied to various problem areas, ranging from computer vision to text mining and audio processing. Let us see how NMF-based techniques can be used for musical audio source separation, by factoring the spectrogram into note spectra templates and note activations.

[14], `http://www.eecs.qmul.ac.uk/~jga/eusipco2012.html` [15].

[3] International Music Score Library Project `http://imslp.org`

2.1 Classic NMF

Let $Y \in \mathbb{R}_+^{M \times N}$ denote the magnitude spectrogram of a music recording, where $M \in \mathbb{N}$ and $N \in \mathbb{N}$ denote the number of frequency bins and number of time-frames, respectively. Given a parameter $K \in \mathbb{N}$, NMF derives two non-negative matrices $W \in \mathbb{R}_+^{M \times K}$ and $H \in \mathbb{R}_+^{K \times N}$ such that $WH \approx Y$, or more precisely, such that a distance function between Y and WH is minimized. This distance is often a modified Kullback-Leibler divergence [16]. To compute a factorization, the matrices W and H are first initialized with random values and then iteratively updated using multiplicative update rules [16]. After the update process, each column of W (also referred to as *template vector*) corresponds to the prototype spectrum of a certain sound component (e.g. a C4 note played on a piano), and the corresponding row of H (also called *activation*) encodes when that sound was active and its volume. When using NMF to separate musical sound sources, we assume that each pair of template vector and activation describes a sound that was produced by a single instrument, and that this instrument can easily be identified, to allow all the sounds from that instrument to be grouped together.

However, there are various issues with this approach. Consider Fig. 3(a) showing a spectrogram of a music recording of a piano and a guitar. The piano plays the notes C4, E4, C4 and, at the same time, the guitar plays the notes G4, C4, G4 (see also the box *Reading a Musical Score* A). Fig. 3(b) shows an NMF-based decomposition of the spectrogram, with the parameter K manually set to four allowing for one template for each of the two different musical pitches used by the two instruments. Looking at the template matrix W and the activation matrix H, some problems become apparent. It is not clear to which sound, pitch or instrument a given template vector corresponds. Furthermore, the activation patterns in H indicate that the templates correspond to mixtures of notes (and instruments). The first two templates seem to represent the note combinations piano-C4/guitar-G4 and piano-E4/guitar-C4, while the last two templates seem to correspond to short-lived broadband sounds that occur at the beginning of these notes. Based on such a factorization, the two instruments cannot readily be separated.

2.2 Score-Informed Constraints

To overcome these issues, most NMF-based musical source separation methods impose certain constraints on W and H. A typical approach is to enforce a harmonic structure

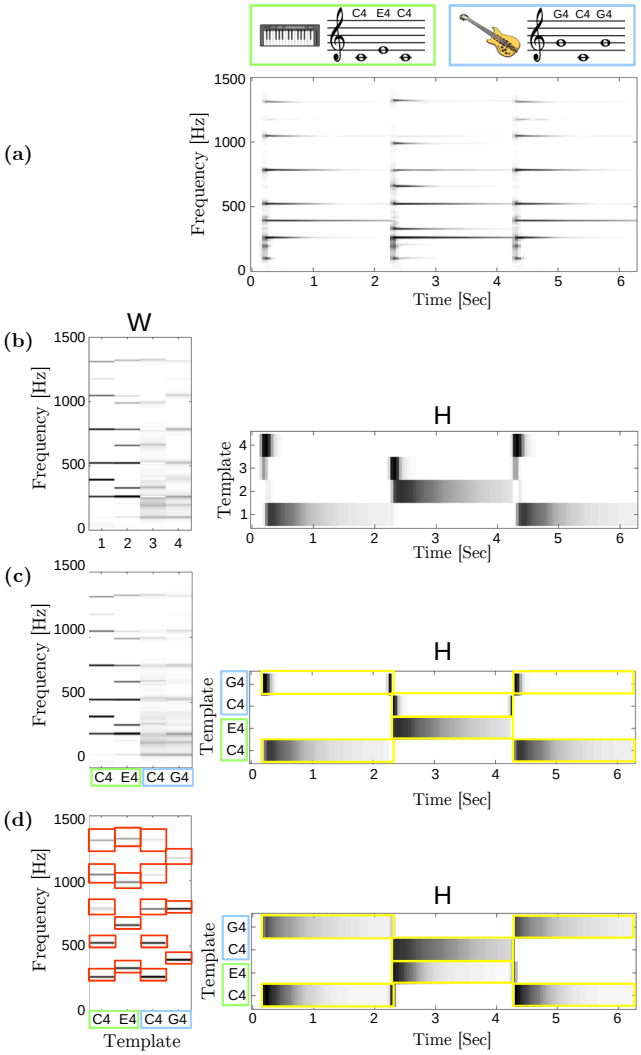


Figure 3: Integrating score information into NMF. **(a)** Spectrogram of a recording of a piano and a guitar. **(b)** Factorization into a template matrix W and an activation matrix H resulting from standard NMF. **(c)** Factorization result after applying constraints to H. **(d)** Factorization result after applying constraints to W and H. The red/yellow boxes indicate areas that were initialized with non-zero values.

in each template in W, and temporal continuity in each activation in H [1, 17]. Further, if the instruments occurring in a recording are known, one can use monophonic training material to learn meaningful templates [17]. While such extensions typically lead to a significant gain in separation quality over classic NMF, they do not fully solve the problem.

Therefore, if strong prior knowledge is available, it should be exploited to further increase the separation performance. In this context, a musical score is particularly valuable. On a coarse level, we can extract *global* information from the score, such as which instruments are playing or which and how many pitches occur over the course of a piece of music. In our example, this information can be used to set the number of

templates automatically to $K = 4$ (two instruments each with two different pitches). We can also assign an instrument and pitch attribute to each template (Fig. 6(c)). On a finer level, one may also exploit *local* information on when notes are actually played. Suppose we could assume that a score pre-aligned to a corresponding audio recording is available, i.e. that the note events specified by the score are aligned to the time positions where they occur in the audio recording. Using this score information, one can impose constraints on the times that certain templates may become active by initializing those activation entries with zero, where a certain instrument and pitch are known to be inactive. Once an entry in W or H is initialized to zero, it will remain set to zero during the subsequent multiplicative update steps [16]. As an example, consider Fig. 3(c), where all entries in H outside the yellow rectangles were initialized with zero values.

In some cases, such an approach will be sufficient to separate many of the notes. However, in our example, the resulting factorization is almost identical to the unconstrained one, compare Fig. 3(b) and (c). Since the piano-C4/guitar-G4 and piano-E4/guitar-C4 combinations always occur together, the constraints on the time activations H have no significant effect, and the first two templates still represent these note combinations. Indeed, individual sounds in music recordings often only occur in certain combinations, which limits also for real recordings the benefits of applying constraints on H alone.

To overcome this problem, we can apply dual-constraints, where both templates and activations are constrained in parallel [6, 14]. The idea to constrain the templates W is based on the observation that most instruments written in a score produce harmonic sounds, and that the templates should reflect this structure. In general, a *harmonic sound* is one whose energy in a time-frequency representation is concentrated around integer multiples of the so called *fundamental frequency*. These energy concentrations are also referred to as *harmonics*. To enforce such a structure in the templates, we can constrain the spectral energy between harmonics to be zero [18]. More precisely, after assigning an instrument and musical pitch to each template vector using the score information, we can use the standard frequency associated with each pitch as an estimate of the fundamental frequency (see Box A), and the rough positions for the harmonics can then be derived. As the exact frequencies are not known, a neighborhood around these positions can then be initialized with non-zero values in the templates, while setting the remaining entries to zero, see [14, 18] for details. Fig. 3(d) shows the resulting

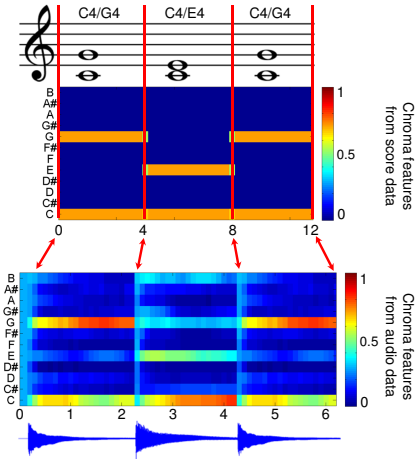


Figure 4: Score-audio synchronization: Positions in the score are aligned (red arrows) to positions in the audio recording based on a comparison of chroma features, which were derived from both representations.

factorization, with the non-zero neighbourhoods around the harmonics indicated by red rectangles in W. All four template vectors in W have now a clearly defined harmonic structure and most disturbing interferences from other sounds have been eliminated, such that the two instruments can finally be separated based on this factorization. Listening examples using full-length piano recordings and publicly available score-data can be found on a website[4].

3 Aligning Audio and Score Data

In the previous section, we assumed that we had a temporal alignment between the score's note events and the physical time position where they actually occur in a given audio recording. While musical scores are available for many songs, they are rarely aligned to a given recording and aligning them manually is very laborious. To automate this process, there are various methods for computing a temporal alignment between score and audio representations, a task also referred to as *score-audio synchronization*.

Rather than giving strict specifications, a score is rather a guide for performing a piece of music leaving scope for different interpretations (Box A). Reading the instructions in the score, a musician shapes the music by varying the tempo, dynamics, and articulation, thus creating a personal interpretation of the piece. The goal of score-audio synchronization is to automatically match the musical timing as notated in the score to the physical timing used in audio recordings. Automatic methods typically proceed in two steps: Feature extraction from both audio and score, followed by temporal alignment [19].

[4] http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF/

The feature representations should be robust to irrelevant variations, yet should capture characteristic information that suffice to accomplish the subsequent synchronization task. Chroma-based music features have turned out to be particularly useful [20]. Capturing the short-time energy distribution of a music representation across the 12 pitch classes (Box A), chroma features closely correlate to the harmonic progression while showing a large degree of robustness to variations in timbre and dynamics. Thanks to this property, chroma features allow for a comparison of score and audio data, where most acoustic properties in the audio that are not reflected in the score are ignored. Fig. 4 illustrates chroma feature sequences derived from score data (top) and audio data (bottom).

In the second step, the derived feature sequences are brought into temporal correspondence, using an alignment technique such as Dynamic Time Warping (DTW) or Hidden Markov Models (HMM) [19]. Intuitively, as indicated by the red bidirectional arrows shown in Fig. 4, the alignment can be thought of a structure, which links corresponding positions in the score and the audio and thus annotates the audio recording with available score data.

Various extensions to this basic scheme have been proposed. For example, additional onset cues extracted from the audio can be used to significantly improve on the temporal accuracy of the alignment [21, 22]. Other approaches address the problem of computing an alignment in real-time while the audio is recorded [19, 23]. Furthermore, methods have been proposed for computing an alignment in the presence of structural variations between the score and the audio version, such as the omission of repetitions, the insertion of additional parts (soli, cadenzas), or differences in the number of stanzas [24]. Such advanced score-audio synchronization methods are an active area of current research [21, 23].

4 Dealing with Vibrato and Frequency Drift

While the approach outlined in Section 2 yields good results in many cases, it relies on the assumption that the fundamental frequency associated with a musical pitch is approximately constant over time, since the frequency position of harmonics in each template is fixed and cannot move up or down. While this assumption is valid for some instruments such as a piano it is not true in general. Fig. 5 shows an audio recording of a piano and a clarinet. The piano (green) indeed exhibits stable horizontal frequency

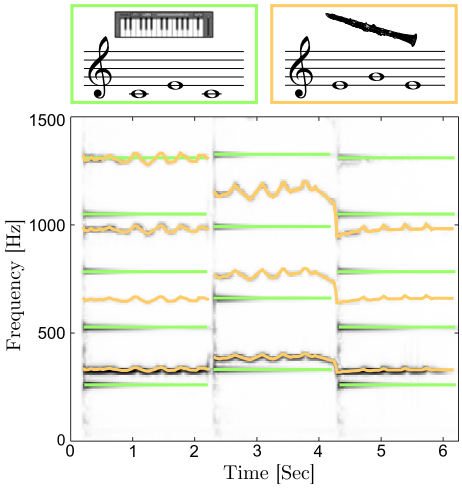


Figure 5: Spectrogram of a recording of a piano and a clarinet. The position of the fundamental frequency and the harmonics is illustrated for the piano (in green) and for the clarinet (in orange).

trajectories, whereas the clarinet produces strong frequency modulations due to the way it is played ("vibrato"). These are clearly visible, for example, between seconds 3 and 4 in a spectral band around 1200 Hz. Additionally, the clarinet player continuously glides from one note to the next, resulting in smooth transitions between the fundamental frequencies of notes (e.g. between second 4 and 5). As a result, while a single note in the score is associated with a single musical pitch, its realization in the audio can be much more complex, involving a whole range of frequencies.

To deal with such fluctuating fundamental frequencies, parametric signal models have been considered as extensions to NMF [17, 25]. In these approaches, the musical audio signal is modelled using a family of parameters capturing, for example, the fundamental frequency (including its temporal fluctuation), the spectral envelope of instruments or the amplitude progression. Such parameters often have an explicit acoustic or musical interpretation, and it is often straightforward to integrate available score information.

As an example for such a parametric approach, we consider a simplified version of the *Harmonic Temporal Structured Clustering (HTC)* strategy [17, 26]. Variants of this model have been widely employed for score-informed source separation [8–10, 27]. In an HTC-based approach, specialized model components replace NMF template vectors and activations. Each HTC template consists of several Gaussians, which represent the partials of a harmonic sound (Fig. 6(a)). To adapt the model to different instruments and their specific spectral envelopes, the height of each Gaussian in an HTC template can be scaled individually using a set of parameters ($\gamma_1, \ldots, \gamma_5$ in Fig. 6(a)). An additional

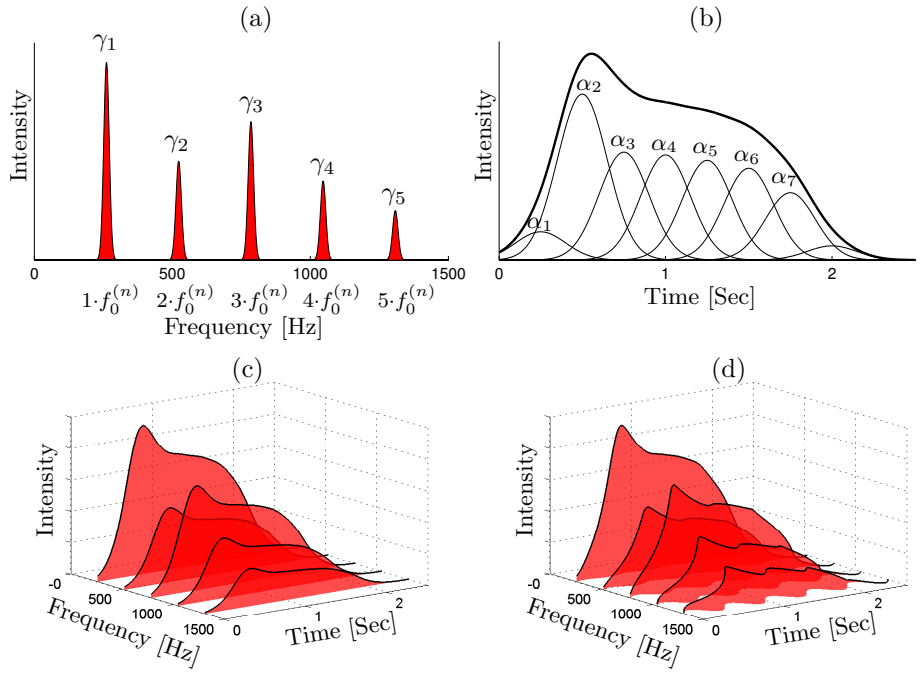


Figure 6: Simplified HTC model. **(a)** HTC template with parameters. **(b)** HTC activation with parameters. **(c)/(d)** Illustrations of the full spectrogram model combining the submodels shown in (a) and (b), using a constant and a fluctuating fundamental frequency in (c) and (d), respectively.

parameter $f_0^{(n)}$ specifies the fundamental frequency of an HTC template in each time frame n. Assuming a harmonic relationship between the partials, the parameter $f_0^{(n)}$ also controls the exact location of each Gaussian (Fig. 6(a)).

HTC activations are also constructed using Gaussians. Their position is typically fixed such that only some height parameters can be adapted (parameters $\alpha_1, \ldots, \alpha_7$ in Fig. 6(b)). By choosing suitable values for the variance of these Gaussians, one can enforce a significant overlap between them, which leads to an overall smooth activation progression.

Combining the HTC templates and activations in a way similar to NMF yields a spectrogram model which suppresses both non-harmonic elements in frequency direction and spurious peaks in time direction (Fig. 6(c)), see [17, 26]. HTC-based approaches model the spectral envelope independently from the fundamental frequency, such that both can be adapted individually. As an illustration, we used a constant fundamental frequency parameter in Fig. 6(c), and a fluctuating fundamental frequency in Fig. 6(d).

The explicit meaning of most HTC parameters enables a straightforward integration of score information [8–10, 27]. For example, after assigning a musical pitch to an HTC template, the fundamental frequency parameter can be constrained to lie in a small

interval around the standard frequency of the pitch [9, 10]. Using the score's instrument information, the γ-parameters can be initialized using sound examples for the specific instrument [8, 27]. Finally, using the position and duration of note events specified by the score, constraints on the activity parameters α can be imposed by setting them to zero whenever the corresponding instrument and pitch are known to be inactive [8, 9].

To model a given recording using the HTC approach, most methods minimize a distance between the spectrogram and the model to find suitable values for the parameters. To this end, most approaches employ minimization methods that are also used in the NMF context: multiplicative updates [9], expectation-minimization [8, 27], or interior points methods [10]. Constraints on the parameters are typically expressed using priors [8, 27] (in probabalistic models) or penalty terms [10] (in deterministic methods).

Many other parametric models are possible. For example, several score-informed source separation methods have used variants of the *Source/Filter (S/F) model* as their underlying signal model [25, 28]. In the S/F-model a sound is produced by an excitation source, which is subsequently filtered. When applied in speech processing, the source corresponds to the vocal chords while the filter models the vocal tract. Applied to musical instruments, the source typically corresponds to a vibrating element, e.g. the strings of a violin, and the filter corresponds to the instrument's resonance body. Since the parameters used to model the filter and the excitation source have an explicit meaning, they can often be initialized or constrained based on score information [29, 30].

5 Example-based Source Separation

The approaches discussed in previous sections were based on the assumption that all instruments notated in a score produce purely harmonic sounds. However, this assumption is not perfectly true for many instruments, including the piano or the guitar. Percussive instruments, such as drums or bongos, also exhibit complex broadband spectra instead of a set of harmonics. As an alternative to enforcing a harmonic structure in the signal model, we can use a data-driven approach, and guide the separation based on examples for the sound of the segregated sources [5,15]. Using the score information, we can provide these examples by employing a high-quality synthesiser to render a separate instrument audio track for each instrumental line specified by the score. For each instrument track, an NMF decomposition of the corresponding magnitude spectrogram can be computed, resulting in an *instrument template matrix* and an *instrument activation matrix*. Finally,

by horizontally stacking the instrument template matrices, one large *prior template matrix* $\tilde{W}$ can be created. Similarly, a large *prior activation matrix* $\tilde{H}$ can be built up by vertically stacking all instrument activation matrices. These two prior matrices essentially give an example of how a meaningful factorization of the magnitude spectrogram of the real audio recording could look like. Therefore, the separation of the real recording can be guided by employing the matrices $\tilde{W}$ and $\tilde{H}$ as Bayesian priors for the template matrix W and the activation matrix H within the Probabilistic Latent Component Analysis (PLCA) framework, a probabilistic formulation of NMF [3, 31]. This way, the matrices W and H tend to stay close to $\tilde{W}$ and $\tilde{H}$.

While such an example-based approach to separation enables non-harmonic sounds to be modelled, there are drawbacks if the synthetic examples are not sufficiently similar to the real sounds. For example, if the fundamental frequency of a synthesised harmonic sound is different from the corresponding frequency in the real audio recording, the matrices $\tilde{W}$ and $\tilde{H}$ impose false priors, for the position of the fundamental frequency as well as for the position of the harmonics, such that separation may fail. However, combining example-based source separation with harmonic constraints in the signal model (as discussed in Section 2.2) can mitigate these problems, often resulting in a significant increase in separation quality [32, 33].

6 Further Extensions and Future Work

In this article, we showed how information provided by a musical score can be used to facilitate the separation of musical sound sources, which are typically highly correlated in time and frequency in a music recording. We demonstrated how score and audio data can automatically be aligned, and how score information can be integrated into NMF. Further extensions addressed fluctuating fundamental frequencies or enabled the separation of instruments based on example sounds synthesized from the score.

The general idea of score-informed source separation leaves room for many possible extensions. For example, all of the approaches discussed above operate *offline*, where the audio recording to be processed is available as a whole. For streaming scenarios, the audio stream can only be accessed up to a given position, and the computational time is also limited to allow the separation result to be returned shortly after the audio data has been streamed. As a first approach to *online* score-informed separation, Duan and Pardo [13] combine a real-time score-audio alignment method with an efficient

score-informed separation method.

Besides information obtained from a score, various other sources of prior knowledge can be integrated. Examples include spatial information obtained from multi-channel recordings [6, 34], or side information describing the mixing process of the sources [35]. A distant goal could be a general framework where various different kinds of prior knowledge can be plugged in as they are available.

Since the prior knowledge provided by a score stabilizes the separation process significantly, one could use this stability to increase the level of detail used to model sound sources. For example, most current signal models typically do not account for the fact that the energy in higher partials of a harmonic sound often decays faster than in lower partials. Also room acoustics or time varying effect filters applied to the instruments are often not considered in separation methods. In such cases, score-informed signal models might be stable enough to robustly model even such details.

Further, since it is not always realistic to assume that an entire score is available for a given recording (in particular for pop music), exploiting partially available score information will be a central challenge. For example, so called lead sheets often do not encode the entire score but only the main melody and some chords for the accompaniment. Furthermore, the score could be available only for a specific section (e.g. the chorus) and not for the rest of the recording, such that suitable approaches to integrating partial prior knowledge, such as [4], have to be developed. Also, lyrics are often available as pure text without any information about notes or timing. Addressing these scenarios will lead to various novel approaches and interesting extensions of the strategies discussed in this article.

References

[1] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, 2010.

[2] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. Hershey: IGI Global, 2010, pp. 162–185.

[3] P. Smaragdis and G. J. Mysore, "Separation by humming: User guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. to Audio Acoust. (WASPAA)*, 2009, pp. 69–72.

[4] A. Lefevre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2012, pp. 115–120.

[5] U. Simsekli and A. T. Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *Proc. European Signal Process. Conf. (EUSIPCO)*, 2012, pp. 2639–2643.

[6] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 257–260.

[7] J. Driedger, H. Grohganz, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proc. ACM Int. Conf. Multimedia (ACM-MM)*, 2013, pp. 541–544.

[8] 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 *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2008, pp. 133–138.

[9] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 45–48.

[10] S. Ewert and M. Müller, "Estimating note intensities in music recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 385–388.

[11] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2012, pp. 57–60.

[12] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H.-O. Oh, "MPEG Spatial Audio Object Coding - The ISO/MPEG standard for efficient coding of interactive audio scenes," *Jour. Audio Engineering Soc.*, vol. 60, no. 9, pp. 655–673, 2012.

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

[14] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 129–132.

[15] J. Ganseman, P. Scheunders, G. J. Mysore, and J. S. Abel, "Source separation by score synthesis," in *Proc. Int. Computer Music Conf. (ICMC)*, 2010, pp. 462–465.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Systems (NIPS)*, 2000, pp. 556–562.

[17] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 982–994, 2007.

[18] S. A. Raczynski, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2007, pp. 381–386.

[19] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Commun. ACM, Special Iss.: Music information retrieval*, vol. 49, no. 8, pp. 38–43, 2006.

[20] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[21] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2385–2397, 2011.

[22] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009, pp. 1869–1872.

[23] Z. Duan and B. Pardo, "A state space model for online polyphonic audio-score alignment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 197–200.

[24] M. Müller and D. Appelt, "Path-constrained partial music synchronization," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2008, pp. 65–68.

[25] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 3, pp. 564–575, 2010.

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

[27] Y. Han and C. Raphael, "Informed source separation of orchestra and soloist," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2010, pp. 315–320.

[28] T. Heittola, A. P. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2009, pp. 327–332.

[29] P. Sprechmann, P. Cancela, and G. Sapiro, "Gaussian mixture models for score-informed instrument separation," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 49–52.

[30] C. Joder and B. Schuller, "Score-informed leading voice separation from monaural audio," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, 2012, pp. 277–282.

[31] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations (article id 947438)," *Comput. Intell. Neurosc.*, vol. 2008, p. 9, 2008.

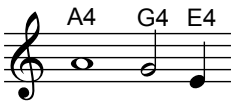
[32] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 888–891.

[33] J. Fritsch, J. Ganseman, and M. D. Plumbley, "A comparison of two different methods for score-informed source separation," in *Proc. Int. Workshop Machine Learning Music (MML)*, 2012, p. 2.

[34] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 314–319.

[35] 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 *Proc. European Signal Process. Conf. (EUSIPCO)*, 2012, pp. 2397–2401.

A Reading a Musical Score



Modern music notation uses an abstract language to specify musical parameters. Pitch is indicated by the vertical placement of a note on a *staff*, which consists of five horizontal lines. Each musical pitch is associated with a name, such as A4 (corresponding to the note between the second and the third line from below in the figure), and a standard frequency in Hz (440 Hz for the A4). If the standard frequency of a pitch is twice as high compared to another, they are said to differ by an *octave*. In this case, the two pitches share the same letter in their name, also referred to as *chroma*, and only differ in their number (e.g. A3 with 220 Hz is one octave below the A4). In most Western music, a system referred to as *equal temperament* is used that introduces twelve different chromas by the names C, $C^{\#}$, D, ..., B, which subdivide each octave equidistantly on a logarithmic frequency scale. A special symbol at the beginning of a staff, the *clef*, is used to specify which line corresponds to which pitch (e.g. the first symbol in the figure specifies that the second line from below corresponds to G4). Temporal information is specified in a score using different shapes for the note, which encode the relative duration of a note. For example, a whole note or semibreve (denoted by the symbol 𝅝) is played twice as long as a half note or minim (𝅗𝅥), which again is played twice as long as a quarter note or crotchet (♩). Additional information on music notation can be found under `http://en.wikipedia.org/wiki/Musical_notation`.

Acknowledgments

S. E. is supported by EPSRC Grant EP/J010375/1. M. D. P. is supported by EPSRC Leadership Fellowship EP/G007144/1 and EPSRC Grant EP/H043101/1.