A SEPARATE AND RESTORE APPROACH TO SCORE-INFORMED MUSIC DECOMPOSITION

Christian Dittmar, Jonathan Driedger, Meinard Müller

International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany,
{christian.dittmar;jonathan.driedger;meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Our goal is to improve the perceptual quality of signal components extracted in the context of music source separation. Specifically, we focus on decomposing polyphonic, mono-timbral piano recordings into the sound events that correspond to the individual notes of the underlying composition. Our separation technique is based on score-informed Non-Negative Matrix Factorization (NMF) that has been proposed in earlier works as an effective means to enforce a musically meaningful decomposition of piano music. However, the method still has certain shortcomings for complex mixtures where the tones strongly overlap in frequency and time. As the main contribution of this paper, we propose a restoration stage based on refined Wiener filter masks to score-informed NMF. Our idea is to introduce notewise soft masks created from a dictionary of perfectly isolated piano tones, which are then adapted to match the timbre of the target components. A basic experiment with mixtures of piano tones shows improvements of our novel reconstruction method with regard to perceptually motivated separation quality metrics. A second experiment with more complex piano recordings shows that further investigations into the concept are necessary for real-world applicability.

Index Terms— Score-informed music processing, source separation, music decomposition, signal reconstruction.

1. INTRODUCTION

The goal of music source separation is to decompose a music recording into its constituent signal components [1, 2]. We focus on the special case of polyphonic, mono-timbral piano recordings, where we aim to extract sound events that correspond to the composition's individual notes. We assume that each sound event corresponding to a musical note can be characterized as a harmonic tone with constant pitch as well as a sharp attack, a stable sustain, and a release phase. Moreover, the mixture signal is assumed to be a linear superposition of the isolated tones. This is, of course, a simplification since we neglect acoustic effects such as room impulse responses and sympathetic string resonances in the piano. With these assumptions, the decomposition of piano music into isolated tones constitutes a limited, yet challenging source separation task, which may pave the way to more complex scenarios.

One of the challenges is to improve the perceptual quality of the separated signals which may suffer from audible artifacts, depending on the complexity of the music, the recording conditions, as well as the decomposition technique. In this paper, we follow the paradigm of score-informed Non-Negative Matrix Factorization (NMF) to *separate* the mixture magnitude spectrogram as described in [3, 4]. This procedure involves soft masks used to derive the targeted component magnitude spectrograms by Wiener filtering. The soft masks are usually obtained from multiplying suitable NMF templates and activations. In contrast to that, we propose to refine the soft masks on the basis of a timbre-adapted dictionary of isolated tones. We refer to this extension of the conventional method as *restore* approach and show that it can be beneficial for certain types of mixtures.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of related work, Section 3 reviews score-informed NMF, Section 4 describes our proposed restore approach, Section 5 discusses the experimental results and indicates directions for future work.

2. RELATED WORK

Music source separation using score information has first been introduced in [5] and [6]. Related approaches used tensor factorization [7] or synthesized music for component initialization [8]. An important starting point for our work is the procedure for score-informed music decomposition described in [3], where the authors describe how to impose musically meaningful constraints on the components of a Non-Negative Matrix Factorization (NMF) without the need for a dedicated component training. The same principle is extended to note-wise decomposition in [4]. Other authors devised elaborate source-filter models to account for the time-varying spectral envelope of components with fixed pitch [9, 10]. In [11], a semi-adaptive NMF variant, which allows to efficiently capture the temporal evolution of component spectrograms, was proposed.

In [12], Ewert discussed the problems inherent to NMF decomposition in case of overlapping partials of the targeted components. He proposed to use the activations and gains computed by a first NMF stage to infer a time-frequency (TF) dependent weighting of the mixture magnitude spectrogram accounting for possible phase cancellations. He could show that this leads to more meaningful decompositions in a second NMF stage. The use of a weighted NMF had already been proposed by Virtanen in [13], where it was used as a means to fill gaps in the magnitude spectrogram that occurred due to binary masking of predominant pitched signal components. In [14], Cano et al. investigated the complex mutual influence of magnitude and phase on the quality of separated signals in source separation. Cano proposed to soften the additivity constraint in source separation and suggested to use instrument specific resynthesis approaches in [15].

*This work has been supported by the German Research Foundation (DFG MU 2686/6-1). The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

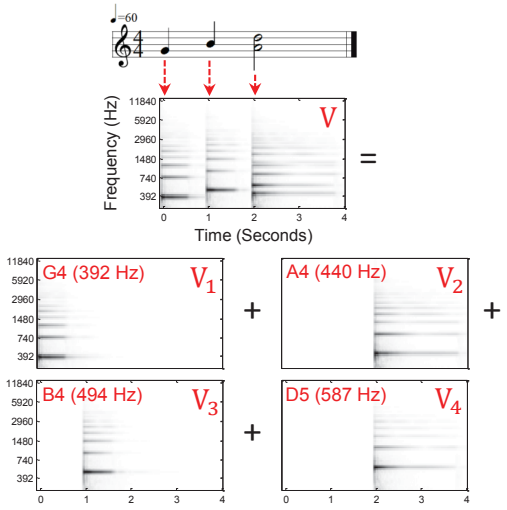

Figure 1: Artificial piano score used as illustrative example throughout the paper. Our target is the extraction of the isolated magnitude spectrograms corresponding to the individual notes G4, A4, B4, and D5 (ordered by pitch instead of onset time).

3. SEPARATE

In this section, we summarize the score-informed NMF approach as described in [3, 4]. In our signal model, we assume that the given piano recording x is a linear mixture of notewise audio events x_s, $s \in [1:S]$, where $S \in \mathbb{N}$ is the number of musical notes specified in the musical score (see our example score in Figure 1) such that $x := \sum_s x_s$. Let $\mathcal{X}(m,k)$, $m,k \in \mathbb{Z}$, be a complex-valued TF coefficient at the m^{th} time frame and k^{th} frequency bin of the Short-Time Fourier Transform (STFT) of our mixture signal x. Let $\mathrm{V} := |\mathcal{X}|^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K \times M}$ be a transposed version of the mixture signal's magnitude spectrogram. Our objective is to decompose V into component magnitude spectrograms V_s that correspond to the individual note events x_s. Ignoring possible phase issues, we assume that the additive relationship $\mathrm{V} := \sum_s \mathrm{V}_s$ is fulfilled (see Figure 1).

3.1. Music Decomposition via NMF

NMF can be used to decompose the magnitude spectrogram V into spectral basis functions (also called templates) encoded by the columns of $\mathrm{W} \in \mathbb{R}_{\geq 0}^{K \times R}$ and time-varying gains (also called activations) encoded by the rows of $\mathrm{H} \in \mathbb{R}_{\geq 0}^{R \times M}$ such that $\mathrm{V} \approx \mathrm{WH}$. NMF typically starts with a suitable initialization of matrices $\mathrm{W}^{(0)}$ and $\mathrm{H}^{(0)}$. Subsequently, these matrices are iteratively updated to adapt to V with regard to a suitable distance measure. In this work, we use the well-known update rules for minimizing the Kullback-Leibler Divergence [16] given by

$$\mathrm{W}^{(\ell+1)} = \mathrm{W}^{(\ell)} \odot \frac{\frac{\mathrm{V}}{\mathrm{W}^{(\ell)}\mathrm{H}^{(\ell)}}\mathrm{H}^{(\ell)\mathrm{T}}}{\mathrm{J}\mathrm{H}^{(\ell)\mathrm{T}}} \quad (1)$$

$$\mathrm{H}^{(\ell+1)} = \mathrm{H}^{(\ell)} \odot \frac{\mathrm{W}^{(\ell+1)\mathrm{T}}\frac{\mathrm{V}}{\mathrm{W}^{(\ell+1)}\mathrm{H}^{(\ell)}}}{\mathrm{W}^{(\ell)\mathrm{T}}\mathrm{J}} \quad (2)$$

for $\ell = 0, 1, 2, \ldots, L$ for some $L \in \mathbb{N}$. The symbol $\odot$ denotes element-wise multiplication and the division is also understood element-wise. Furthermore, $\mathrm{J} \in \mathbb{R}^{K \times M}$ denotes an all-one matrix.

3.2. Constraint Components via Score-Informed NMF

Proper initialization of $\mathrm{W}^{(0)}$ and $\mathrm{H}^{(0)}$ is an effective means to constrain the degrees of freedom in the NMF iterations and enforces convergence to a desired, musically meaningful solution. One possibility is to impose constraints derived from a time-aligned, symbolic representation (i.e., machine readable score) of the recording [7]. Three constraints can be obtained from the musical score. First, the rank R of the decomposition is chosen according to the number of unique musical pitches. Second, each column of $\mathrm{W}^{(0)}$ is initialized with a prototype harmonic overtone series reflecting the expected nature of a musical tone corresponding to the assigned musical pitch. Third, the rows of $\mathrm{H}^{(0)}$ are initialized as follows: A binary constraint matrix $\mathrm{C}_s \in \mathbb{R}^{R \times M}$ is constructed for each $s \in [1:S]$, where C_s is 1 at entries that correspond to the pitch and temporal position of the s^{th} aligned note event and 0 otherwise. The union (OR-sum) of all C_s is then used as initialization of $\mathrm{H}^{(0)}$. With this initialization, each template obtained from iteration (1) typically corresponds to an average spectrum (usually ℓ^1-normalized [2]) of the corresponding musical pitch and each activation function obtained from (2) corresponds to the temporal amplitude envelope of all occurrences of that particular pitch throughout the recording.

4. RESTORE

Score-informed NMF as described in Section 3.2 yields a decomposition of V into musically meaningful templates $\mathrm{W}^{(L)}$ and activations $\mathrm{H}^{(L)}$. In the following, we discuss the issues inherent to the restoration of our targeted V_s from the components and introduce our extension to the conventional procedure.

4.1. Component Magnitude Spectrogram Reconstruction

As shown on the left hand side of Figure 2(a), we can use the results of score-informed NMF to reconstruct magnitude spectrograms corresponding to individual note objects as $\mathrm{V}_s^{\mathrm{NMF}} \approx \mathrm{W}^{(L)}\left(\mathrm{H}^{(L)} \odot \mathrm{C}_s\right)$. In the conceptual illustration, the template and activation corresponding to the s^{th} tone are indicated by a hatched column and row, respectively. The binary activation in the corresponding C_s is visualized by a black box inside the hatched row. In order to obtain a time-domain signals from $\mathrm{V}_s^{\mathrm{NMF}}$, it is common practice to use the mixture phase information of the original STFT $\mathcal{X}$ and to invert the resulting modified STFTs via the signal reconstruction method from [17]. However, NMF-based models typically yield only a rough approximation of the original magnitude spectrogram, where spectral nuances may not be captured well. Therefore, the audio components reconstructed in this way may contain a number of audible artifacts. In order to better capture the temporal evolution of the spectral nuances, it is common practice to calculate soft masks that can be interpreted as a weighting matrix reflecting the contribution of the s^{th} tone to the original mixture V. The mask corresponding to the desired note event can be computed as $\mathrm{M}_s^{\mathrm{NMF}} := \mathrm{V}_s^{\mathrm{NMF}} \oslash \left(\mathrm{W}^{(L)}\mathrm{H}^{(L)} + \epsilon\right)$, where $\oslash$ denotes element-wise division and ϵ is a small positive constant to avoid division by zero. We obtain the masking-based estimate of the component magnitude spectrogram as $\mathrm{V}_s^{\mathrm{Mask}} := \mathrm{V} \odot \mathrm{M}_s^{\mathrm{NMF}}$. This procedure is also often referred to as Wiener filtering.

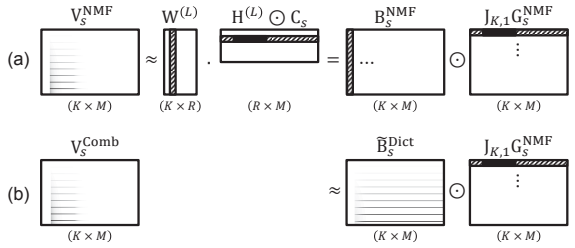

Figure 2: Conceptual illustration of **(a)** the procedures described in Section 4.1 and Section 4.2, and **(b)**, the restore approach described in Section 4.3.

4.2. Difficult Mixtures

As discussed in [12], even the integration of score information might not suffice to separate certain mixtures. This is especially true in the case of mutually overlapping harmonics and transients. Our artificial example exhibits some of these problems. The decay of the two quarter notes (G4 and B4) is interfered by the attack transient of the subsequent notes. The last two notes (A4 and D5) are played simultaneously, so their attack transients overlap and some of their harmonics collide since they have very close center frequencies due to the harmonic relationship between the two notes (fourth interval). As can be seen in Figure 3(b) this leads to corrupted NMF templates. The note $s = 2$ (A4) exhibits a spurious peak around 600 Hz in a spectrogram frame that lies within the attack phase of $\mathrm{V}_s^{\mathrm{NMF}}$. The artifact is caused by "crosstalk" of the first partial of D5 whose center frequency is 587 Hz and leads to deteriorated separation quality. For the following considerations, we introduce an alternative representation $\mathrm{B}_s \in \mathbb{R}_{\geq 0}^{K \times M}$ of the targeted V_s defined by

$$\mathrm{B}_s(k,m) := \frac{\mathrm{V}_s(k,m)}{\mathrm{G}_s(m)} \tag{3}$$

where $\mathrm{G}_s \in \mathbb{R}^{1 \times M}$ contains element-wise the ℓ^1-norm of each spectrogram frame of the original V_s. On the right hand side of Figure 2(a), we show $\mathrm{B}_s^{\mathrm{NMF}}$ derived by application of (3) to $\mathrm{V}_s^{\mathrm{NMF}}$. Each spectrogram frame in $\mathrm{B}_s^{\mathrm{NMF}}$ is depicted as a replicate of the ℓ^1-normalized template corresponding to s. This seemingly redundant representation will become useful in Section 4.3, where we replace the potentially imperfect NMF templates by timbre-adapted and ℓ^1-normalized spectral templates taken from a dictionary.

4.3. Component Restoration using a Note Dictionary

Inspired by the work of [7] and [8], we construct a dictionary of crosstalk-free magnitude spectrograms $\mathrm{V}_s^{\mathrm{Dict}}$ obtained from isolated piano tones. Via the score information, we select the appropriate tone and place it in the TF domain according to the onset position of the target note. However, the benefit of having an artifact-free $\mathrm{V}_s^{\mathrm{Dict}}$ comes at the expense that the dictionary tone is likely to differ in timbre from the target tone in the mixture. In order to adapt the timbral qualities of the dictionary without propagating potential errors in the NMF decomposition, we propose to transfer the spectral envelope of the target tone to the corresponding dictionary tone. Figure 3 illustrates this procedure for the note A4 ($s = 2$) in our artificial example. Similar to [12], we take $\mathrm{V}_s^{\mathrm{NMF}}$ as best estimate of the target component regardless of potential decomposition artifacts. Furthermore, we assume that a single estimate for the spectral

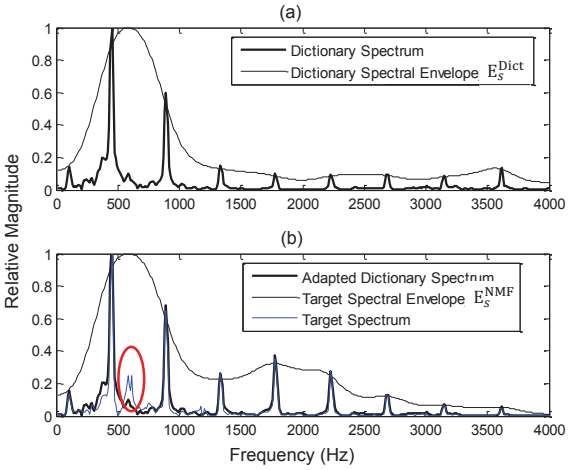

Figure 3: **(a):** Spectrogram frame (bold black curve) located in the attack phase of the dictionary tone representing the note A4. Its spectral envelope $\mathrm{E}_s^{\mathrm{Dict}}$ (thin black curve) is extracted via the true envelope method [18]. **(b):** Due to an imperfect NMF decomposition, a spurious peak is present around 600 Hz (marked by the red oval) in the target spectrum (blue curve). After an envelope transfer, the adapted dictionary spectrum follows nicely the target envelope $\mathrm{E}_s^{\mathrm{NMF}}$, but does not contain the artifact.

envelope can be applied to the complete tone spectrogram. The estimate could e.g., be derived from an average spectrum. In our case, we expect the target component to dominate over potential crosstalk components in a spectral frame located in the attack phase of the tone. From that particular frame, we extract the spectral envelope $\mathrm{E}_s^{\mathrm{NMF}} \in \mathbb{R}^{K \times 1}$ using the so-called true envelope method as described by Röbel et al. in [18]. This procedure iteratively refines an estimate for the spectral envelope obtained by conventional cepstral liftering. Using the same method, we extract an estimate for the spectral envelope $\mathrm{E}_s^{\mathrm{Dict}}$ of the dictionary spectrogram. Then, we introduce the timbre-adapted dictionary note spectrogram as

$$\tilde{\mathrm{V}}_s^{\mathrm{Dict}} := \mathrm{V}_s^{\mathrm{Dict}} \odot \frac{\left(\mathrm{E}_s^{\mathrm{NMF}} \mathrm{J}_{1,M}\right)}{\left(\epsilon + \mathrm{E}_s^{\mathrm{Dict}} \mathrm{J}_{1,M}\right)} \tag{4}$$

where the division is understood element-wise and $\mathrm{J}_{1,M} \in \mathbb{R}^{1 \times M}$ denotes an all-one matrix needed to replicate both spectral envelopes across all frames. Subsequently, we apply (3) to $\tilde{\mathrm{V}}_s^{\mathrm{Dict}}$ in order to obtain $\tilde{\mathrm{B}}_s^{\mathrm{Dict}}$ that we now use to replace the potentially corrupted $\mathrm{B}_s^{\mathrm{NMF}}$ as shown in Figure 2(b). This way, we obtain novel estimates for the component spectrogram and corresponding soft mask as

$$\mathrm{V}_s^{\mathrm{Comb}} := \tilde{\mathrm{B}}_s^{\mathrm{Dict}} \odot \left(\mathrm{J}_{K,1} \cdot \mathrm{G}_s^{\mathrm{NMF}}\right) \tag{5}$$

$$\mathrm{M}_s^{\mathrm{Comb}} := \mathrm{V}_s^{\mathrm{Comb}} \oslash \left(\epsilon + \sum_s \mathrm{V}_s^{\mathrm{Comb}}\right) \tag{6}$$

where $\mathrm{J}_{K,1} \in \mathbb{R}^{K \times 1}$ denotes an all-one matrix needed to replicate the corresponding note activation across all frequency bins. In short, the sequence of (4),(5), and (6) allows us to combine the NMF-based activations with ℓ^1-normalized dictionary spectra that have been adapted to match the timbre captured in the NMF-based tem-

plates. The resulting note component spectrogram is again obtained by Wiener filtering as $\mathrm{V}_s^{\mathrm{Prop}} := \mathrm{V} \odot \mathrm{M}_s^{\mathrm{Comb}}$.

5. EXPERIMENTS

We conducted two source separation experiments using wellknown quality metrics to assess the possible improvements achievable with our proposed approach. First, we evaluated the separation of simplistic piano tone mixtures. Second, we tried to decompose more complex piano recordings into bass and treble component signals.

5.1. Dataset

Our first test set consisted of pair-wise piano tone combinations. We assigned MIDI pitch P_1 to the first tone in the mixture and defined it to be the interfering signal. Consequently, we assigned MIDI pitch P_2 to the second tone and defined it to be the target signal. Both P_1, P_2 were varied from MIDI pitch $P = 21$ (A0, 27.5 Hz) to $P = 108$ (C8, 4186 Hz) resulting in 7569 tone pairs (including unison intervals). The underlying single tone signals were recorded from a real piano, while the tone dictionary used for separation was synthesized using the Pianoteq[1] physical modeling plugin. Each tone pair was treated as individual test item, i.e. only $R = 2$ components were used.

The second test uses a subset of 11 MIDI files from the Saarland Music Data (SMD[2]) collection. SMD contains MIDI files for various classical piano pieces which were performed by students of the Hochschule für Musik Saar on a Yamaha Disklavier. The Disklavier stores all key and pedal movements performed by the pianist in an interpreted MIDI file that is suitable for synthesizing piano performances with expressive dynamics and timing. Following the experimental design in [12], we split the note events in each of the interpreted MIDI files into a bass (MIDI pitch $P < 60$) and a treble set (MIDI pitch $P \geq 60$). We again used Pianoteq to synthesize the note sequences for the bass and treble set individually. The bass tones were defined to be the interferer and the treble tones the target, respectively. The superposition of both yielded our mixture signal.

All test files had 44.1 kHz sampling rate, the STFT was computed with a blocksize of approx. 46.4 ms and a hopsize of approx. 5.8 ms. The number of NMF iterations was set to $L = 30$. For each test item, we used $\mathrm{V}_s^{\mathrm{Mask}}$ as magnitude spectrogram representing the conventional approach and $\mathrm{V}_s^{\mathrm{Prop}}$ as magnitude spectrogram representing proposed reconstruction. We employed the PEASS Toolkit [19, 20] in order to evaluate the quality of the separated audio signals obtained from application of the conventional and the proposed method. From the available metrics, we focused on the perceptually-motivated Overall Perceptual Score (OPS) and used the objective Source-Distortion Ratio (SDR) to complement the evaluation.

5.2. Results and Discussion

From the first experiment, we derived two interval related evaluations. First, we aggregated the quality measures to the interval $\Delta P_{1,2} := P_2 - P_1$ between the two piano tones in each mixture and averaged the respective measurements. Second, we applied the same result aggregation, this time mapping to 12 absolute, octave-agnostic interval classes $\tilde{\Delta} P_{1,2} := (\Delta P_{1,2} \mod 12)$. Figure 4(a) shows that the restore approach surpasses the conventional Wiener filtering approach in terms of OPS mostly for positive intervals, i.e., where the MIDI pitch of the target is above the MIDI pitch of the interferer. Interestingly, this trend can not be observed for the SDR. Instead, the improvements are more evenly distributed and only decrease for very wide intervals regardless if they are positive or negative. Figure 4(b) shows that the proposed restore approach is approx. 9.5 OPS points ahead of the conventional approach if we ignore octave information. The average SDR improvement in that case amounts to approx. 0.7 dB.

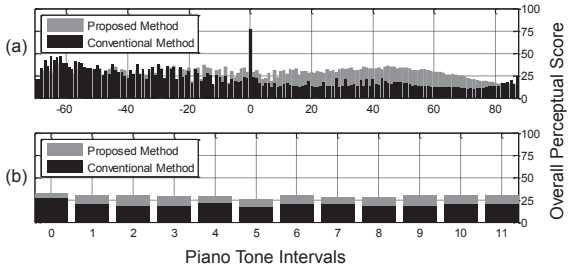

Figure 4: The Overall Perceptual Score (OPS) [19, 20] computed for separated piano tone mixtures with the conventional method (black bars) vs. the proposed method (gray bars). **(a):** OPS averaged over interval classes. **(b):** OPS averaged over absolute, octave-wrapped interval classes.

We obtained very mixed results in our second experiment with realistic piano performances. On average, we achieved an OPS of 38.06 (SDR 10.16 dB) using conventional Wiener filtering, while our proposed restore approach yielded a tiny OPS increase to 38.32 (SDR 10.25 dB). Unfortunately, there is no consistent improvement across the test items, roughly half of them exhibit lower quality metrics compared to the conventional approach. From inspection of selected examples, we believe that this might be related to inferior separation of tones with a long sustain phase. Since we transfer the spectral envelope taken from the tone's attack, the sustain phase might diverge from the desired target over time.

Still, we see potential in further developing the principal concept of our separate and restore approach. One obvious possibility would be an interval-selective application of the proposed method. Another possible direction is to investigate dedicated processing of percussive and harmonic NMF components to remedy some of the remaining problems related to unsatisfactory separation of complex mixtures. Audio examples covering positive and negative separation results are available online[3].

5.3. Conclusions and Future Work

We presented a method for post-processing score-informed music decomposition by means of refined soft masks based on a dictionary of timbre-adapted piano tone spectrograms. In simple tone mixtures, this step attenuates the mutual interference between components. In the future, we want to further enhance this concept and investigate its applicability to other source separation tasks, such as drum sound separation from drum loops.

[1] https://www.pianoteq.com/

[2] http://resources.mpi-inf.mpg.de/SMD/SMD_MIDI-Audio-Piano-Music.html

[3] http://www.audiolabs-erlangen.de/resources/MIR/2015-WASPAA-SeparateAndRestore/

6. REFERENCES

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

[2] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, London, UK, September 2007, pp. 414–421.

[3] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kobe, Japan, 2012, pp. 129–132.

[4] J. Driedger, H. Grohganz, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proceedings of the ACM International Conference on Multimedia (ACM-MM)*, Barcelona, Spain, 2013, pp. 541–544.

[5] B. P. J. Woodruff and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, October 2006, pp. 314–319.

[6] J. W. Y. Li and D. L. Wang, "Monaural musical sound separation using pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1361–1371, 2009.

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

[8] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 888–891.

[9] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[10] J.-L. Durrieu, B. David, and G. Richard, "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[11] C. Dittmar and D. Gärtner, "Real-time transcription and separation of drum recordings based on NMF decomposition," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, September 2014, pp. 187–194.

[12] S. Ewert, M. D. Plumbley, and M. Sandler, "Accounting for phase cancellations in non-negative matrix factorization using weighted distances," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 7480–7484.

[13] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music." in *Proceedings of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, Brisbane, Australia, September 2008, pp. 17–22.

[14] E. Cano, J. Abeßer, C. Dittmar, and G. Schuller, "Influence of phase, magnitude and location of harmonic components in the perceived quality of extracted solo signals," in *Proceedings of the Audio Engineering Society (AES) Conference on Semantic Audio*, Ilmenau, Germany, July 2011, pp. 247–252.

[15] E. Cano, C. Dittmar, and G. Schuller, "Re-thinking sound separation: Prior information and additivity constraint in separaton algorithms," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Maynooth, Ireland, 2013.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the Neural Information Processing Systems (NIPS)*, Denver, CO, USA, 2000, pp. 556–562.

[17] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[18] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Madrid, Spain, September 2005, pp. 344–349.

[19] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[20] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, March 2012, pp. 430–437.