TOWARDS TRANSIENT RESTORATION IN SCORE-INFORMED AUDIO DECOMPOSITION

Christian Dittmar, Meinard Müller

International Audio Laboratories Erlangen*,
Erlangen, Germany
{christian.dittmar,meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Our goal is to improve the perceptual quality of transient signal components extracted in the context of music source separation. Many state-of-the-art techniques are based on applying a suitable decomposition to the magnitude of the Short-Time Fourier Transform (STFT) of the mixture signal. The phase information required for the reconstruction of individual component signals is usually taken from the mixture, resulting in a complex-valued, modified STFT (MSTFT). There are different methods for reconstructing a time-domain signal whose STFT approximates the target MSTFT. Due to phase inconsistencies, these reconstructed signals are likely to contain artifacts such as pre-echos preceding transient components. In this paper, we propose a simple, yet effective extension of the iterative signal reconstruction procedure by Griffin and Lim to remedy this problem. In a first experiment, under laboratory conditions, we show that our method considerably attenuates pre-echos while still showing similar convergence properties as the original approach. A second, more realistic experiment involving score-informed audio decomposition shows that the proposed method still yields improvements, although to a lesser extent, under non-idealized conditions.

1. INTRODUCTION

Music source separation aims at decomposing a polyphonic, multitimbral music recording into component signals such as singing voice, instrumental melodies, percussive instruments, or individual note events occurring in a mixture signal [1]. Besides being an important step in many music analysis and retrieval tasks, music source separation is also a fundamental prerequisite for applications such as music restoration, upmixing, and remixing. For these purposes, high fidelity in terms of perceptual quality of the separated components is desirable. The majority of existing separation techniques work on a time-frequency (TF) representation of the mixture signal, often the Short-Time Fourier Transform (STFT). The target component signals are usually reconstructed using a suitable inverse transform, which in turn can introduce audible artifacts such as musical noise, smeared transients or pre-echos, as exemplified in Figure 1(c).

In order to better preserve transient signal components, we propose in this paper a simple, yet effective extension to the signal reconstruction procedure by Griffin and Lim [2]. The original method iteratively estimates the phase information necessary for time-domain reconstruction from a magnitude STFT (STFTM) by going back and forth between the STFT and the time-domain, only updating the phase information, while keeping the STFTM fixed. Our proposed extension manipulates the intermediate time-domain reconstructions in order to attenuate the pre-echos that potentially precede the transients.

We conduct two kinds of evaluations in an audio decomposition scenario, where our objective is to extract isolated drum sounds from polyphonic drum recordings. To this end, we use a publicly available test set that is enriched with all necessary side information, such as the true "oracle" component signals and their precise transient positions. In the first experiment, under laboratory conditions, we make use of all side-information in order to focus on evaluating the benefit of our proposed method for transient preservation in signal reconstruction. Under these idealized conditions, we can show that our proposed method considerably attenuates pre-echos while still exhibiting similar convergence properties as the original method. In the second experiment, we employ a state-of-the-art decomposition technique [3, 4] with score-informed constraints [1] to estimate the component signal's STFTM from the mixture. Under these more realistic conditions, our proposed method still yields improvements yet to a lesser extent than in the idealized scenario.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of related work before Section 3 introduces our new method. Section 4 details and discusses the experimental evaluation under laboratory conditions. Section 5 describes a more realistic application and evaluation of our proposed method in conjunction with score-informed audio decomposition. Finally, in Section 6 we conclude and indicate directions for future work.

2. RELATED WORK

Three research fields are important for our work: First, a number of publications on signal reconstruction and transient preservation are related and relevant for our proposed restoration method. Second, papers on score-informed audio decomposition (i.e., source separation) provide the basis for deploying our method in a real-world application.

2.1. Signal Reconstruction

The problem of signal reconstruction, also known as magnitude spectrogram inversion or phase estimation is a well researched topic. In their classic paper [2], Griffin and Lim proposed the so-called LSEE-MSTFTM algorithm (denoted as GL throughout this paper) for iterative, blind signal reconstruction from modified STFT magnitude (MSTFTM) spectrograms. In [5], Le Roux et al. developed a different view on this method by describing it using

* The International Audio Laboratories Erlangen is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS.

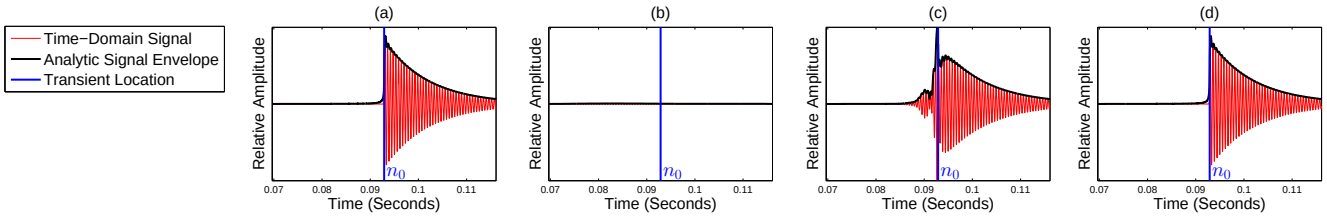

Figure 1: Illustration of the transient restoration. **(a):** Target component signal, an exponentially decaying sinusoid preceded by silence. **(b):** Reconstruction using zero phase. Due to destructive interference, the overall amplitude seemingly decreased to silence. **(c):** Reconstruction after 200 GL iterations, exhibiting pronounced transient smearing. **(d):** Reconstruction after 200 iterations of the proposed transient restoration method. The left hand legend applies to all plots, n_0 denotes the transient position.

a TF consistency criterion. By keeping the necessary operations entirely in the TF domain, several simplifications and approximations could be introduced that lower the computational load compared to the original procedure. Since the phase estimates obtained using GL can only converge to local optima, several publications were concerned with finding a good initial estimate for the phase information [6, 7]. Sturmel and Daudet [8] provided an in-depth review of signal reconstruction methods and pointed out unsolved problems. An extension of GL with respect to convergence speed was proposed in [9]. Other authors tried to formulate the phase estimation problem as a convex optimization scheme and arrived at promising results hampered by high computational complexity [10]. Another work [11] was concerned with applying the spectrogram consistency framework to signal reconstruction from wavelet-based magnitude spectrograms.

2.2. Transient Preservation

The problem of transient preservation has been extensively addressed in the field of perceptual audio coding, where pre-echo artifacts can occur ahead of transient signal components. Pre-echos are caused by the use of relatively long analysis and synthesis windows in conjunction with coding-related modification of TF bins such as quantization of spectral magnitudes according to a psycho-acoustic model. It can be considered as state-of-the-art to use block-switching to account for transient events [12]. An interesting approach was proposed in [13], where spectral coefficients are encoded by linear prediction along the frequency axis, automatically reducing pre-echos. Other authors proposed to decompose the signal into transient and residual components and use optimized coding parameters for each stream [14]. In [15], the authors proposed a scheme that unifies iterative signal reconstruction (see Section 2.2) and block-switching in the context of audio coding. Transient preservation has also been investigated in the context of time-scale modification methods based on the phase-vocoder [16]. In addition to an optimized treatment of the transient components, several authors follow the principle of phase-locking or re-initialization of phase in transient frames [17, 18].

2.3. Score-informed Audio Decomposition

The majority of music source separation techniques operate on a TF representation of the mixture signal. It is common practice to compute the mixture signal's STFT and apply suitable decomposition techniques (e.g., Non-Negative Matrix Factorization (NMF)) to the corresponding magnitude spectrogram. This yields an MSTFTM, ideally representing the isolated target signal component. The corresponding time-domain signal is usually derived by using the original phase information and applying signal reconstruction methods.

When striving for good perceptual quality of the separated target signals, many authors propose to impose score-informed constraints on the decomposition [19, 20, 1]. This has the advantage that the separation can be guided and constrained by information on the approximate location of component signals in time (onset, offset) and frequency (pitch, timbre). A few studies deal with source separation of strongly transient signals such as drums [21, 22]. Usage of the Non-Negative Matrix Factor Deconvolution (NMFD) for drum sound separation was first proposed in [3]. Later works applied it to drum sound detection using sparseness constraints [4] as well as regularisation in [23]. Others authors focus on the separation of harmonic vs. percussive components [24, 25, 26]. The importance of phase information for source separation quality is discussed in [27].

3. TRANSIENT RESTORATION

In the following, we first fix our notation and signal model and describe the employed signal reconstruction method. Afterward, we introduce our novel extension for transient preservation in the GL method and provide an illustrative example.

3.1. Notation and Signal Model

We consider the real-valued, discrete time-domain signal $x : \mathbb{Z} \to \mathbb{R}$ to be a linear mixture $x := \sum_{c=1}^{C} x_c$ of $C \in \mathbb{N}$ component signals x_c corresponding to individual instruments. As shown in Figure 2(a), each component signal contains at least one transient audio event produced by the corresponding instrument (in our case, by striking a drum). Furthermore, we assume that we have a symbolic transcription available that specifies the onset time (i.e., transient position) and instrument type for each of the audio events. From that transcription, we derive the total number of onset events S as well as the number of unique instruments C. Our aim is to extract individual component signals x_c from the mixture x as shown in Figure 2. For evaluation purposes (see Section 4), we assume to have the oracle component signals x_c available.

We decompose x in the TF-domain, to this end we employ STFT as follows. Let $\mathcal{X}(m,k)$ be a complex-valued TF coefficient at the m^{th} time frame and k^{th} spectral bin. The coefficient is computed by

$$\mathcal{X}(m,k) := \sum_{n=0}^{N-1} x(n+mH)w(n)\exp(-2\pi ikn/N), \qquad (1)$$

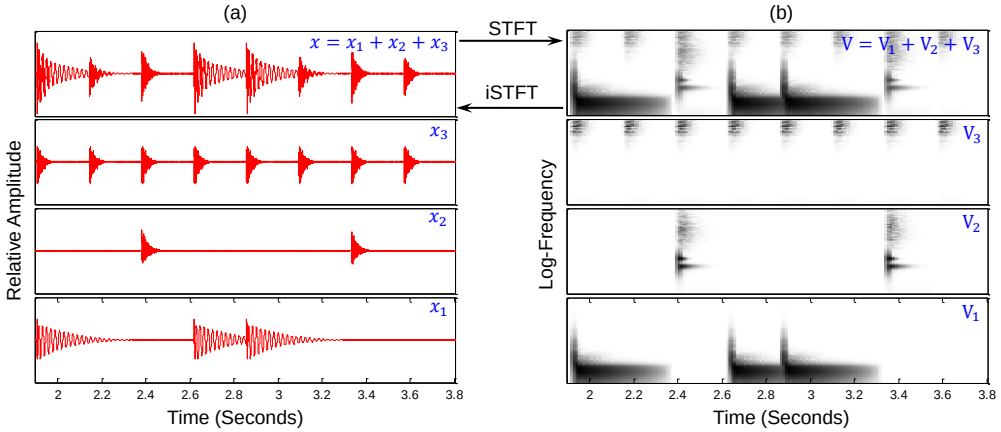

Figure 2: Illustration of our signal model. **(a):** Mixture signal x is the sum of $C = 3$ component signals x_c, each containing sequences of synthetic drum sounds sampled from a Roland TR 808 drum machine (x_1: kick drum, x_2: snare drum, x_3: hi-hat). **(b):** TF representation of the mixture's magnitude spectrogram V and $C = 3$ component magnitude spectrograms V_c. For better visibility, the frequency axis and the magnitudes are on a logarithmic scale.

where $w : [0 : N-1] \to \mathbb{R}$ is a suitable window function of blocksize $N \in \mathbb{N}$, and $H \in \mathbb{N}$ is the hop size parameter. The number of frequency bins is $K = N/2$ and the number of spectral frames $M \in [1 : M]$ is determined by the available signal samples. For simplicity, we also write $\mathcal{X} = \mathrm{STFT}(x)$. Following [5], we call $\mathcal{X}$ a consistent STFT since it is a set of complex numbers which has been obtained from the real time-domain signal x via (1). In contrast, an inconsistent STFT is a set of complex numbers that was not obtained from a real time-domain signal. From $\mathcal{X}$, the magnitude spectrogram $\mathcal{A}$ and the phase spectrogram φ are derived as

$$\mathcal{A}(m,k) := |\mathcal{X}(m,k)|\,, \tag{2}$$

$$\varphi(m,k) := \angle\mathcal{X}(m,k), \tag{3}$$

with $\varphi(m,k) \in [0, 2\pi)$.
Let $\mathrm{V} := \mathcal{A}^{\mathrm{T}} \in \mathbb{R}_{\geq 0}^{K \times M}$ be a non-negative matrix holding a transposed version of the mixture's magnitude spectrogram $\mathcal{A}$. Our objective is to decompose V into component magnitude spectrograms V_c that correspond to the distinct instruments as shown in Figure 2(b). For the moment, we assume that some oracle estimator extracts the desired $\mathcal{A}_c := \mathrm{V}_c^{\mathrm{T}}$. One possible approach to estimate the component magnitudes using a state-of-the-art decomposition technique will be described in Section 5. In order to reconstruct a specific component signal x_c, we set $\mathcal{X}_c := \mathcal{A}_c \odot \exp(i\varphi_c)$, where $\mathcal{A}_c = \mathrm{V}_c^{\mathrm{T}}$ and φ_c is an estimate of the component phase spectrogram. It is common practice to use the mixture phase information φ as an estimate for φ_c and to invert the resulting MSTFT via the LSEE-MSTFT reconstruction method from [2]. The method first applies the inverse Discrete Fourier Transform (DFT) to each spectral frame in $\mathcal{X}_c$, yielding a set of intermediate time signals y_m, with $m \in [0 : M-1]$, defined by

$$y_m(n) := \frac{1}{N} \sum_{k=0}^{N-1} \mathcal{X}_c(m,k) \exp(2\pi i k n / N), \tag{4}$$

for $n \in [0 : N-1]$ and $y_m(n) := 0$ for $n \in \mathbb{Z} \setminus [0 : N-1]$. Second, the least squares error reconstruction is achieved by

$$x_c(n) := \frac{\sum_{m\in\mathbb{Z}} y_m(n - mH) w(n - mH)}{\sum_{m\in\mathbb{Z}} w(n - mH)^2}, \tag{5}$$

$n \in \mathbb{Z}$, where the analysis window w is re-used as synthesis window. Please note that LSEE-MSTFT should not be confused with LSEE-MSTFTM (called GL in this work) that extends the signal reconstruction with iterative phase estimation (cf. Algorithm 3.2). In the following, for the sake of brevity, we will use $x_c = \mathrm{iSTFT}(\mathcal{X}_c)$ as short form for the application of (4) and (5).

3.2. Proposed Algorithm

Since we construct the MSTFT $\mathcal{X}_c$ in the TF domain, we have to consider that it may be an inconsistent STFT, i.e., there may not exist a real time-domain signal x_c fulfilling $\mathcal{X}_c = \mathrm{STFT}(x_c)$. Intuitively speaking, the complex relationship between magnitude and phase is likely corrupted as soon as the magnitude in certain TF bins is modified. In practice, this inconsistency can lead to transient smearing and pre-echos in x_c, especially for large N.
To remedy this problem, we propose to iteratively minimize the inconsistency of $\mathcal{X}_c$ by the following extension (denoted as TR) of the GL procedure [2]. For the moment, let's assume that $\mathcal{X}_c$ contains precisely one transient onset event, whose exact location in time n_0 is known. Now, we introduce the iteration index $\ell = 0, 1, 2, \ldots L \in \mathbb{N}$. Given $\mathcal{A}_c$ and some initial phase estimate $(\varphi_c)^{(0)}$, we introduce the initial STFT estimate of the target component signal $(\mathcal{X}_c)^{(0)} := \mathcal{A}_c \odot \exp(i(\varphi_c)^{(0)})$ and repeat for $\ell = 0, 1, 2, \ldots L$ the following steps

Transient Restoration (TR) Algorithm:

1. $(x_c)^{(\ell+1)} := \mathrm{iSTFT}\left((\mathcal{X}_c)^{(\ell)}\right)$ via (4) and (5)
2. Enforce $(x_c)^{(\ell+1)}(n) := 0$ for $n \in \mathbb{Z}, n < n_0$
3. $(\varphi_c)^{(\ell+1)} := \angle\mathrm{STFT}\left((x_c)^{(\ell+1)}\right)$ via (1) and (3)
4. $(\mathcal{X}_c)^{(\ell+1)} := \mathcal{A}_c \odot \exp\left(i(\varphi_c)^{(\ell+1)}\right)$

The crucial point of our proposed extension is the intermediate step

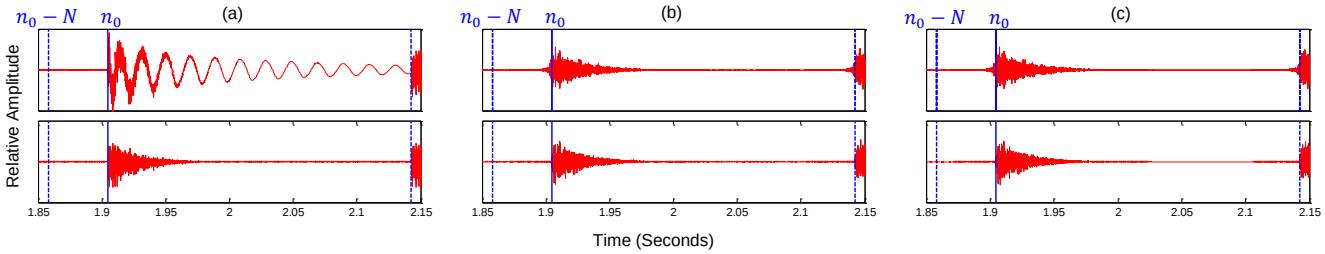

Figure 3: Different hi-hat component signals of our example drum loop. The transient position n_0 is given by the solid blue line, the excerpt boundaries by the dashed blue lines. **(a):** Mixture signal (top) vs. oracle hi-hat signal (bottom). **(b):** Hi-hat signal in Case 2, reconstruction after $L = 200$ iterations of GL (top) vs. TR (bottom). **(c):** Hi-hat signal in Case 4, reconstruction after $L = 200$ iterations of GL (top) vs. TR (bottom). Since the NMFD decomposition works very well for our example drum loop, there is almost no noticeable visual difference between (b) and (c).

2. which enforces transient constraints in the GL procedure. Figure 1 illustrates our proposed method with the target component signal in red, overlaid with the envelope of its analytic signal in Figure 1(a). The example signal exhibits transient behavior around n_0 (blue line) when the waveform transitions from silence to an exponentially decaying sinusoid. Figure 1(b) shows the time-domain reconstruction obtained from the iSTFT with $(\varphi_c)^{(0)} = 0$ (i.e., zero phase for all TF bins). Through destructive interference of overlapping frames, the transient is completely destroyed, the amplitude of the sinusoid is strongly decreased and the envelope looks nearly flat. Figure 1(c) shows the reconstruction with pronounced transient smearing after $L = 200$ GL iterations. Figure 1(d) shows that the restored transient after $L = 200$ iterations of the proposed method is much closer to the original signal. In real-world recordings, there usually exist multiple transient onsets event throughout the signal. In this case, one may apply the proposed method to signal excerpts localized between consecutive transients (resp. onsets) as shown in Figure 3.

4. EVALUATION UNDER LABORATORY CONDITIONS

For evaluation, we compared the conventional GL reconstruction with our proposed TR method under two different initialization strategies for $(\mathcal{X}_c)^{(0)}$. In the following, we describe the used data set, the test item generation, and our evaluation metrics.

4.1. Dataset

In principle, we follow the evaluation approach from [27]. In all our experiments, we use the publicly available "IDMT-SMT-Drums" dataset[1]. In the "WaveDrum02" subset, there are 60 drum loops, each given as perfectly isolated single track recordings (i.e., oracle component signals) of the three instruments kick drum, snare drum, and hi-hat. All 3×60 recordings are in uncompressed PCM WAV format with 44.1 kHz sampling rate, 16 Bit, mono. Mixing all three single tracks together, we obtain 60 mixture signals. Additionally, the onset times and thus the approximate n_0 of all onsets are available per individual instrument. Using this information, we constructed a test set of 4421 drum onset events by taking excerpts from the mixtures, each located between consecutive onsets of the target instrument. In doing so, we zero pad N samples ahead of each excerpt. The rationale is to deliberately prepend a section of silence in front of the local transient position. Inside that section, decay influence of preceding note onsets can be ruled out and potentially occurring pre-echos can be measured. In turn, this leads to a virtual shift of the local transient location to $n_0 + N$ (which we denote again as n_0 for notational convenience). In Figure 3, the adjusted excerpt boundaries are visualized by the dashed blue lines and the virtually shifted n_0 by the blue line. Since the drum loops are realistic rhythms, the excerpts exhibit varying degree of superposition with the remaining drum instruments played simultaneously. In Figure 3(a), the mixture (top) exhibits pronounced influence of the kick drum compared to the isolated hi-hat signal (bottom). For comparison, the two top plots in Figure 2(a) show a longer excerpt of the mixture x and the hi-hat component x_3 of our example signal. In the bottom plot in Figure 3(a), one can see the kick drum x_1 in isolation. It is sampled from a Roland TR 808 drum computer and resembles a decaying sinusoid.

Test case	Initial phase estimate	Fixed magnitude estimate
Case 1	$(\varphi_c)^{(0)} := \varphi^{\mathrm{Mix}}$	$\mathcal{A}_c := \mathcal{A}_c^{\mathrm{Oracle}}$
Case 2	$(\varphi_c)^{(0)} := 0$	$\mathcal{A}_c := \mathcal{A}_c^{\mathrm{Oracle}}$

Table 1: Configuration of the test cases in the experiment under laboratory conditions.

4.2. Evaluation Setting

For each mixture excerpt, we compute the STFT via (1) with $H = 512$ and $N = 2048$ and denote it as $\mathcal{X}^{\mathrm{Mix}}$. Since all test items have 44.1 kHz sampling rate, the frequency resolution is approx. 21.5 Hz and the temporal resolution is approx. 11.6 ms. We use a symmetric Hann window of size N for w. As a reference target, we take the same excerpt boundaries, apply the same zero-padding, but this time from the single track of each individual drum instrument, denoting the resulting STFT as $\mathcal{X}_c^{\mathrm{Oracle}}$. Subsequently, we define two different cases for the initialization of $(\mathcal{X}_c)^{(0)}$ as detailed in Table 1. Using these settings, we expect the inconsistency of the resulting $(\mathcal{X}_c)^{(0)}$ to be lower in case 1 compared to case 2. Knowing that there exists a consistent $\mathcal{X}_c^{\mathrm{Oracle}}$, we go through $L = 200$ iterations of both GL and our proposed TR method as described in Sec. 3.2.

[1] `http://www.idmt.fraunhofer.de/en/business_units/smt/drums.html`

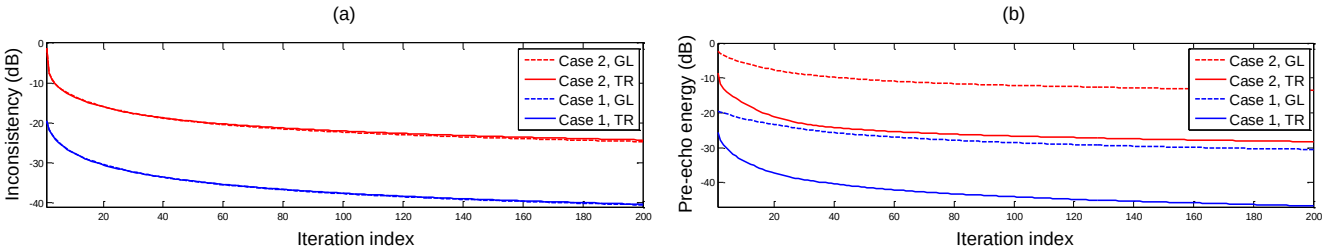

Figure 4: **(a):** Evolution of the normalized consistency measure vs. the number of iterations. **(b):** Evolution of the pre-echo energy vs. the number of iterations. The curves show the average over all test excerpts.

4.3. Quality Measures

We introduce $\mathrm{G}\left((\mathcal{X}_c)^{(\ell)}\right) := \mathrm{STFT}\left(\mathrm{iSTFT}\left((\mathcal{X}_c)^{(\ell)}\right)\right)$ to denote successive application of the iSTFT and STFT (core of the GL algorithm) on $(\mathcal{X}_c)^{(\ell)}$. Following [28], we compute at each iteration ℓ the normalized consistency measure (NCM) as

$$C\left((\mathcal{X}_c)^{(\ell)}, \mathcal{X}_c^{\mathrm{Oracle}}\right) := 10\log_{10}\frac{\left\|\mathrm{G}\left((\mathcal{X}_c)^{(\ell)}\right) - \mathcal{X}_c^{\mathrm{Oracle}}\right\|^2}{\|\mathcal{X}_c^{\mathrm{Oracle}}\|^2}, \tag{6}$$

for both test cases (see Table 1). As a more dedicated measure for the transient restoration, we compute the pre-echo energy as

$$E\left((x_c)^{(\ell)}\right) := \sum_{n=n_0-N}^{n_0}\left|(x_c)^{(\ell)}(n)\right|^2, \tag{7}$$

from the section between the excerpt start and the transient location in the intermediate, time-domain component signal reconstructions $(x_c)^{(\ell)} := \mathrm{iSTFT}\left((\mathcal{X}_c)^{(\ell)}\right)$ for both test cases (see Table 1).

4.4. Results and Discussion

Figure 4 shows the evolution of both quality measures from (6) and (7) with respect to ℓ. Diagram 4(a) indicates that, on average, the proposed TR method performs equally well as GL in terms of inconsistency reduction. In both test cases, the curves for TR (solid line) and GL (dashed line) are almost indistinguishable, which indicates that our new approach shows similar convergence properties as the original method. As expected, the blue curves (Case 1) start at much lower initial inconsistency than the red curves (Case 2), which is clearly due to the initialization with the mixture phase φ^{Mix}. Diagram 4(b) shows the benefit of TR for pre-echo reduction. In both test cases, the pre-echo energy for TR (solid lines) is around 15 dB lower and shows a steeper decrease during the first few iterations compared to GL (dashed line). Again, the more consistent initial $(\mathcal{X}_c)^{(0)}$ of Case 1 (blue lines) exhibit a considerable head start in terms of pre-echo reduction compared to Case 2 (red lines). From these results, we infer that it is sufficient to apply only a few iterations (e.g., $L < 20$) of the proposed method in cases where reasonable initial phase and magnitude estimates are available. However, we need to apply more iterations (e.g., $L < 200$) in case we have a good magnitude estimate in conjunction with a weak phase estimate and vice versa. In the following, we will assess if our preliminary findings obtained under laboratory conditions hold true in a more realistic scenario.

5. APPLICATION TO NMF-BASED AUDIO DECOMPOSITION

In this section, we describe how to apply our proposed transient restoration method in a score-informed audio decomposition scenario. As in Section 4, our objective is again the extraction of isolated drum sounds from polyphonic drum recordings with enhanced transient preservation. In contrast to the idealized laboratory conditions we used before, we now estimate the magnitude spectrograms of the component signals from the mixture. To this end, we employ NMFD [3, 4, 23] as decomposition technique. We briefly describe our strategy to enforce score-informed constraints on NMFD. Finally, we repeat the experiments described Section 4 under these more realistic conditions and discuss our observations.

5.1. Spectrogram Decomposition via NMFD

In this section, we briefly review the NMFD method that we employ for decomposing the TF-representation of x. As indicated in Section 2.3, a wide variety of alternative separation approaches exists. Previous works [3, 4, 23] successfully applied NMFD, a convolutive version of NMF, for drum sound detection and separation. Intuitively speaking, the underlying, convolutive model assumes that all audio events in one of the component signals can be explained by a prototype event that acts as an impulse response to some onset-related activation (e.g., striking a particular drum). In Figure 2(b), one can see this kind of behavior in the hi-hat component V_3. There, all instances of the 8 onset events look more or less like copies of each other that could be explained by inserting a prototype event at each onset position.

NMF can be used to compute a factorization $\mathrm{V} \approx \mathrm{W} \cdot \mathrm{H}$, where the columns of $\mathrm{W} \in \mathbb{R}_{\geq 0}^{K \times C}$ represent spectral basis functions (also called templates) and the rows of $\mathrm{H} \in \mathbb{R}_{\geq 0}^{C \times M}$ contain time-varying gains (also called activations). NMFD extends this model to the convolutive case by using two-dimensional templates so that each of the C spectral bases can be interpreted as a magnitude spectrogram snippet consisting of $T \ll M$ spectral frames. To this end, the convolutive spectrogram approximation $\mathrm{V} \approx \Lambda$ is modeled as

$$\Lambda := \sum_{\tau=0}^{T-1} \mathrm{W}_\tau \cdot \overrightarrow{\mathrm{H}}^{\tau}, \tag{8}$$

where $\overrightarrow{(\cdot)}^{\tau}$ denotes a frame shift operator. As before, each column in $\mathrm{W}_\tau \in \mathbb{R}_{\geq 0}^{K \times C}$ represents the spectral basis of a particular component, but this time we have T different versions of the component available. If we take lateral slices along selected columns of W_τ, we can obtain C prototype magnitude spectrograms as de-

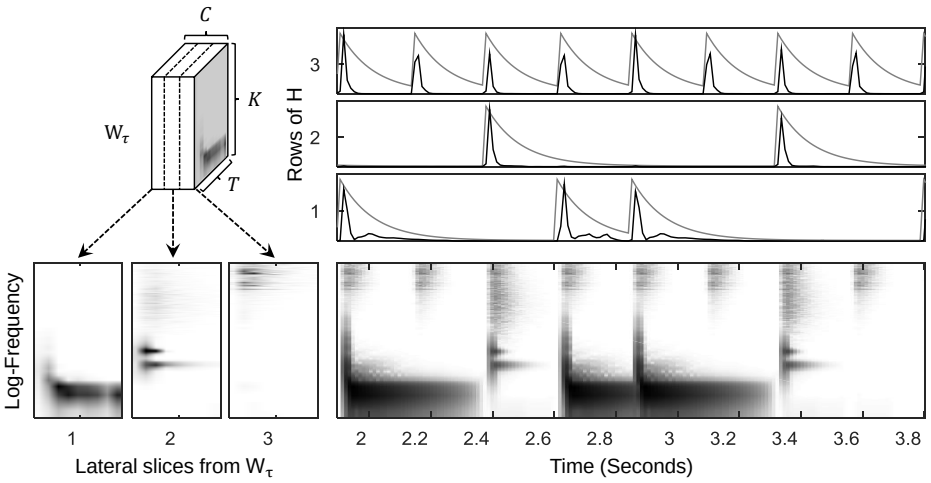

Figure 5: NMFD templates and activations computed for the example drum recording from Figure 2. The magnitude spectrogram V is shown in the lower right plot. The three leftmost plots are the spectrogram templates in W_τ that have been extracted via NMFD. Their corresponding activations in H are shown as black curves in the three top plots. The gray curves show the score-informed initialization $(\mathrm{H})^{(0)}$.

picted on the left hand side of Figure 5. NMFD typically starts with a suitable initialization of matrices $(\mathrm{W}_\tau)^{(0)}$ and $(\mathrm{H})^{(0)}$. Subsequently, these matrices are iteratively updated to minimize a suitable distance measure between the convolutive approximation Λ and V. In this work, we use the update rules detailed in [3], which we omit for brevity.

5.2. Score-Informed NMFD

Proper initialization of $(\mathrm{W}_\tau)^{(0)}$ and $(\mathrm{H})^{(0)}$ is an effective means to constrain the degrees of freedom in the NMFD iterations and enforce convergence to a desired, musically meaningful solution. One possibility is to impose score-informed constraints derived from a time-aligned, symbolic transcription [1]. To this end, the individual rows of $(\mathrm{H})^{(0)}$ are initialized as follows: Each frame corresponding to an onset of the respective drum instrument is initialized with an impulse of unit amplitude, all remaining frames with a small constant. Afterward, we apply a nonlinear exponential moving average filter to model the typical short decay of a drum event. The outcome of this initialization is shown in the top three plots of Figure 5 (gray curves).

In [19], best separation results were obtained by score-informed initialization of both the templates and the activations. For separation of pitched instruments (e.g., piano), prototypical overtone series can be constructed in $(\mathrm{W}_\tau)^{(0)}$. For drums, it is more difficult to model prototype spectral bases. Thus, it has been proposed to initialize the bases with averaged or factorized spectrograms of isolated drum sounds [21, 22, 4]. In this paper, we use a simple alternative that first computes a conventional NMF whose activations H and templates W are initialized by the score-informed $(\mathrm{H})^{(0)}$ and setting $(\mathrm{W})^{(0)} := 1$.

With these settings, the resulting factorization templates are usually a pretty decent approximation of the average spectrum of each involved drum instrument. Simply replicating these spectra for all $\tau \in [0 : T-1]$ serves as a good initialization for the template spectrograms. After some NMFD iterations, each template spectrogram typically corresponds to the prototype spectrogram of the corresponding drum instruments and each activation function corresponds to the deconvolved activation of all occurrences of that particular drum instrument throughout the recording. A typical decomposition result is shown in Figure 5, where one can see that the extracted templates (three leftmost plots) indeed resemble prototype versions of the onset events in V (lower right plot). Furthermore, the location of the impulses in the extracted H (three topmost plots) are very close to the maxima of the score-informed initialization.

In the following, we describe how to further process the NMFD results in order to extract the desired components. Let $\mathrm{H} \in \mathbb{R}_{\geq 0}^{C \times M}$ be the activation matrix learned by NMFD. Then, we define for each $c \in [1 : C]$ the matrix $\mathrm{H}_c \in \mathbb{R}_{\geq 0}^{C \times M}$ by setting all elements to zero except for the c^{th} row that contains the desired activations previously found via NMFD. We approximate the c^{th} component magnitude spectrogram by $\Lambda_c := \sum_{\tau=0}^{T-1} \mathrm{W}_\tau \cdot \overset{\tau\rightarrow}{\mathrm{H}_c}$.

Since the NMFD model yields only a low-rank approximation of V, spectral nuances may not be captured well. In order to remedy this problem, it is common practice to calculate soft masks that can be interpreted as a weighting matrix reflecting the contribution of Λ_c to the mixture V. The mask corresponding to the desired component can be computed as $\mathrm{M}_c := \Lambda_c \oslash \left(\epsilon + \sum_{c=1}^{C} \Lambda_c\right)$, 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 $\mathrm{V}_c := \mathrm{V} \odot \mathrm{M}_c$, with $\odot$ denoting element-wise multiplication. This procedure is referred to as α-Wiener filtering in [29].

Test case	Initial phase estimate	Fixed magnitude estimate
Case 3	$(\varphi_c)^{(0)} := \varphi^{\mathrm{Mix}}$	$\mathcal{A}_c := \mathrm{V}_c^{\mathrm{T}}$
Case 4	$(\varphi_c)^{(0)} := 0$	$\mathcal{A}_c := \mathrm{V}_c^{\mathrm{T}}$

Table 2: Configuration of the test cases in the second experiment involving score-informed audio decomposition.

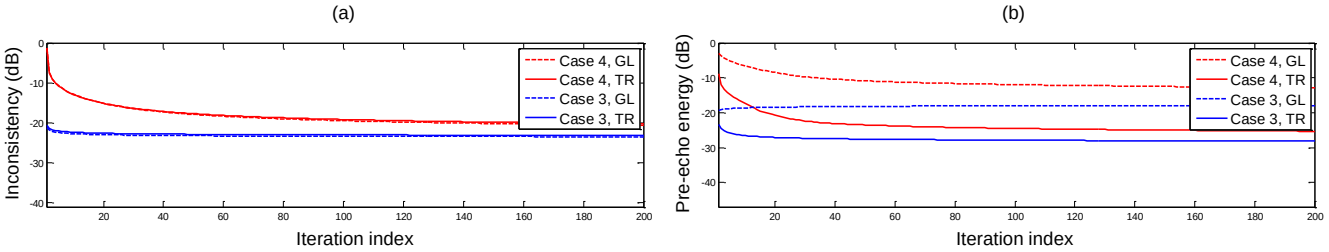

Figure 6: **(a):** Evolution of the normalized consistency measure vs. the number of iterations. **(b):** Evolution of the pre-echo energy vs. the number of iterations. The curves show the average over all test excerpts, the axis limits are the same as in Figure 4.

5.3. Evaluation Results

We now basically repeat the experiment from Section 4, keeping the STFT parameters and excerpt boundaries as described in Section 4.2. The component magnitude spectrograms are estimated from the mixture using $L = 30$ NMFD iterations and spectrogram templates with a duration of $T = 8$ frames (approx. 100 ms). Consequently, we introduce two new test cases as detailed in Table 2.

In Figure 6(a), we again observe that the inconsistency reduction obtained using TR reconstruction (solid lines) is indistinguishable from the GL method (dashed lines). The improvements are less significant compared to the numbers that can be obtained when using oracle magnitude estimates (compare Figure 4(a)). On average, the reconstructions in Case 3 (initialized with φ^{Mix}) seem to quickly get stuck in a local optimum. Presumably, this is due to imperfect NMFD decomposition of the onset related spectrogram frames, where all instruments exhibit a more or less flat magnitude distribution and thus show increased spectral overlap.

In Figure 6(b), we first see that pre-echo reduction with NMFD-based magnitude estimates $\mathcal{A}_c := \mathrm{V}_c^{\mathrm{T}}$ and zero phase (Case 4) works slightly worse than in Case 2 (compare Figure 4(b)). This supports our earlier findings, that weak initial phase estimates benefit the most from applying many iterations of the proposed method. GL reconstruction using φ^{Mix} (Case 3) slightly increases the pre-echno energy over the iterations. In contrast, applying the TR reconstruction decreases the pre-echo energy by roughly -3 dB, which amounts to approx. 15 % of the improvement achievable under idealized conditions (Case 1).

In Figure 3, different reconstructions of a selected hi-hat onset from our example drum loop is shown in detail. Regardless of the used magnitude estimate (oracle in (b) or NMFD-based in (c)), the proposed TR reconstruction (bottom) clearly exhibits reduced pre-echos in comparison to the conventional GL reconstruction (top). We provide example component signals from this drum loop and a few test items online[2]. By informal listening (preferably using headphones), one can clearly spot differences in the onset clarity that can be achieved with different combinations of MSTFT initializations and reconstruction methods. Even in cases, where imperfect magnitude decomposition leads to undesired cross-talk artifacts in the single component signals, our proposed TR method better preserves transient characteristics than the conventional GL reconstruction. Furthermore, usage of the mixture phase for MSTFT initialization seems to be a good choice since one can often notice subtle differences in the reconstruction of the drum events' decay phase in comparison to the oracle signals. However, timbre differences caused by imperfect magnitude decomposition are much more pronounced.

[2]Audio examples: `http://www.audiolabs-erlangen.de/resources/MIR/2015-DAFx-TransientRestoration/`

6. CONCLUSIONS

We proposed a simple, yet effective extension to Griffin and Lim's iterative LSEE-MSTFTM procedure (GL) for improved restoration of transient signal components in music source separation. The method requires additional side information about the location of the transients, which we assume as given in an informed source separation scenario. Two experiments with the publicly available "IDMT-SMT-Drums" data set showed that our method is beneficial for reducing pre-echos both under laboratory conditions as well as for component signals obtained using a state-of-the-art source separation technique. Future work will be directed towards automatic estimation of the required transient positions and application of this technique for polyphonic music recordings involving more than just drums and percussion.

7. ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG MU 268661). We would like to thank the colleagues from Fraunhofer IDMT for releasing the "IDMT-SMT-Drums" data set to the public.

8. REFERENCES

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

[2] Daniel W. Griffin and Jae 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.

[3] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. of the Intl. Conference for Independent Component Analysis and Blind Signal Separation (ICA)*, September 2004, pp. 494–499.

[4] Henry Lindsay-Smith, Skot McDonald, and Mark Sandler, "Drumkit transcription via convolutive NMF," in *Proc. of the Intl. Conference on Digital Audio Effects Conference (DAFx)*, York, UK, September 2012.

[5] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. of the ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, Brisbane, Australia, September 2008, pp. 23–28.

[6] Xinglei Zhu, Gerald T. Beauregard, and Lonce L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, July 2007.

[7] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Phase initialization schemes for faster spectrogram-consistency-based signal reconstruction," in *Proc. of the Acoustical Society of Japan Autumn Meeting*, September 2010, number 3-10-3.

[8] Nicolas Sturmel and Laurent Daudet, "Signal reconstruction from STFT magnitude: a state of the art," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011, pp. 375–386.

[9] Nathanaël Perraudin, Peter Balazs, and Peter L. Søndergaard, "A fast Griffin-Lim algorithm," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2013, pp. 1–4.

[10] Dennis L. Sun and Julius O. Smith III, "Estimating a signal from a magnitude spectrogram via convex optimization," in *Proc. of the Audio Engineering Society (AES) Convention*, San Francisco, USA, October 2012, Preprint 8785.

[11] Tomohiko Nakamura and Hiokazu Kameoka, "Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistency," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, September 2014, pp. 129–135.

[12] Bernd Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, vol. 43, no. 9, pp. 252–256, September 1989.

[13] Jürgen Herre and James D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," in *Proc. of the Audio Engineering Society (AES) Convention*, Los Angeles, USA, November 1996, Preprint 4384.

[14] Oliver Niemeyer and Bernd Edler, "Detection and extraction of transients for audio coding," in *Proc. of the Audio Engineering Society (AES) Convention*, Paris, France, May 2006, Preprint 6811.

[15] Volker Gnann and Martin Spiertz, "Inversion of short-time fourier transform magnitude spectrograms with adaptive window lengths," in *Proc. of the IEEE Intl. Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Taipei, Taiwan, April 2009, pp. 325–328.

[16] Jonathan Driedger, Meinard Müller, and Sebastian Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[17] Chris Duxbury, Mike Davies, and Mark B. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *Proc. of the Audio Engineering Society (AES) Convention*, Munich, Germany, May 2002, Preprint 5530.

[18] Axel Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, London, UK, September 2003, pp. 344–349.

[19] Sebastian Ewert and Meinard Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. of the IEEE Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kobe, Japan, March 2012, pp. 129–132.

[20] Umut Şimşekli and Ali Taylan Cemgil, "Score guided musical source separation using generalized coupled tensor factorization," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, August 2012, pp. 2639–2643.

[21] Eric Battenberg, *Techniques for Machine Understanding of Live Drum Performances*, Ph.D. thesis, University of California at Berkeley, 2012.

[22] Christian Dittmar and Daniel Gärtner, "Real-time transcription and separation of drum recordings based on nmf decomposition," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Erlangen, Germany, September 2014, pp. 187–194.

[23] Axel Röbel, Jordi Pons, Marco Liuni, and Mathieu Lagrange, "On automatic drum transcription using non-negative matrix deconvolution and itakura saito divergence," in *Proc. of the IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 414–418.

[24] Derry Fitzgerald, "Harmonic/Percussive Separation Using Median Filtering," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010, pp. 246–253.

[25] Jonathan Driedger, Meinard Müller, and Sascha Disch, "Extending harmonic-percussive separation of audio signals," in *Proc. of the Intl. Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, October 2014, pp. 611–617.

[26] Estefanía Cano, Mark Plumbley, and Christian Dittmar, "Phase-based harmonic percussive separation," in *Proc. of the Annual Conference of the Intl. Speech Communication Association (Interspeech)*, Singapore, September 2014, pp. 1628–1632.

[27] Estefanía Cano, Jakob Abeßer, Christian Dittmar, and Gerald Schuller, "Influence of phase, magnitude and location of harmonic components in the perceived quality of extracted solo signals," in *Proc. of the Audio Engineering Society (AES) Conference on Semantic Audio*, Ilmenau, Germany, July 2011, pp. 247–252.

[28] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. of the Intl. Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010, pp. 397–403.

[29] Antoine Liutkus and Roland Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Proc. of the IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 266–270.