ITERATIVE STRUCTURED SHRINKAGE ALGORITHMS FOR STATIONARY/TRANSIENT AUDIO SEPARATION

Kai Siedenburg and Simon Doclo*

Dept. of Medical Physics and Acoustics
Cluster of Excellence Hearing4All
University of Oldenburg, Germany
kai.siedenburg@uni-oldenburg.de
simon.doclo@uni-oldenburg.de

ABSTRACT

In this paper, we present novel strategies for stationary/transient signal separation in audio signals in order to exploit the basic observation that stationary components are sparse in frequency and persistent over time whereas transients are sparse in time and persistent across frequency. We utilize a multi-resolution STFT approach which allows to define structured shrinkage operators to tune into the characteristic spectrotemporal shapes of the stationary and transient signal layers. Structure is incorporated by considering the energy of time-frequency neighbourhoods or modulation spectrum regions instead of individual STFT coefficients, and shrinkage operators are employed in a dual-layered Iterated Shrinkage/Thresholding Algorithm (ISTA) framework. We further propose a novel iterative scheme, *Iterative Cross-Shrinkage* (ICS). In experiments using artificial test signals, ICS clearly outperforms the dual-layered ISTA and yields particularly good results in conjunction with a dynamic update of the shrinkage thresholds. The application of the novel algorithms to recordings from acoustic musical instruments provides perceptually convincing separation of transients.

1. INTRODUCTION

Among the primitives that constitute music signals, quasi-stationary sinusoidal components and short-lived transients are of prime importance. This becomes intuitively clear when considering spectrograms of music signals, which oftentimes feature horizontal and vertical lines, corresponding to stationary components that are sparse in frequency and persistent over time and transient components that are sparse in time and persistent over frequency. Whereas modeling sinusoidal components is a well-established field [1], transient estimation as such remains relatively unexplored, despite its numerous applications. Examples include audio restoration where short clicks and crackles must be removed from signals [2], beat tracking where the availability of a transient layer may improve onset detection algorithms [3], or psychoacoustics where robust transient separation could allow for refined investigations into the role of acoustic features in musical timbre perception [4]. The goal of this study is to exploit the distinctive properties of *sparsity* and *persistence* in order to propose robust schemes for stationary/transient separation.

It is important to note that stationary/transient separation is a different if not more fine-grained problem than drum separation or harmonic/percussive separation [5, 6]. Although sounds from drums and percussive instruments are mostly impulsively excited and often inharmonic, percussive sounds may comprise components that extend over similar time scales as those of non-percussive musical instruments (think of the sustained and tonal portions of a snare drum or an open bass-drum sound). This distinction also illustrates the problem that transients are notoriously hard to define in semantic terms, because defining features such as short-livedness and stochastic nature can be easily contested in the context of complex audio mixtures. Consequently, ground truth for the stationary/transient separation task is only available for synthesized test signals.

In order to individually characterize stationary and transient components, our approach is related to several studies using *multi-layered* (or *hybrid*) audio representations that decompose the signal with at least two distinct dictionaries [7]. Early research used orthogonal bases [8]. More recently, a combination of a Modified Discrete Cosine Transform (MDCT) and a wavelet basis was proposed in [9, 10, 11], and Févotte and colleagues further modelled dependencies between coefficients of dual-layered MDCT expansions using a Bayesian framework for the simultaneous estimation of both layers [12]. Our current approach has the same goals as the aforementioned work, utilizing sparsity and inter-coefficient dependencies, but takes a different formal pathway. Here we follow up on the well-known Iterative Shrinkage/Thresholding Algorithm (ISTA) for solving ℓ_1-regularized minimization problems [13], generalized to multi-frame mixed-norm regularization in [14, 15]. Further work extended the involved shrinkage operators with neighborhood-weighting in order to take into account the correlation between adjacent time-frequency coefficients [16, 17]. However, the structured shrinkage framework has not yet been applied to dual-layer signal decomposition with redundant STFT dictionaries.

The goal of the current study is to explore several strategies for structured shrinkage as part of a dual-layered framework with STFT dictionaries of different resolutions. Within each layer, the structured shrinkage operators are tailored towards the distinct spectrotemporal orientations of the stationary and transient components. In order to increase separation robustness, we further propose a novel iteration scheme and update rule for the threshold parameters. In Sec. 2, we provide a framework that allows us to formulate the structured shrinkage operators. Iteration rules are described in Sec. 3, before a comprehensive evaluation of the algorithmic variants using both artificial test signals as well as recordings is provided in Sec. 4 and Sec. 5.

* This work was supported by the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant ITN-GA-2012-316969 and by a postdoc grant from the University of Oldenburg.

2. STRUCTURED SHRINKAGE

This section provides a background on structured shrinkage. We outline the basic signal model in Sec. 2.1, before Sec. 2.2 reviews the idea of extending shrinkage operators with neighbourhood weighting, and Sec. 2.3 reformulates this as an operation in the modulation domain.

2.1. Formal framework

We assume the observed time-domain audio signal $\mathbf{y} \in \mathbb{R}^M$ of length M to be an additive combination of a stationary layer $\mathbf{s}_S$ and a transient layer $\mathbf{s}_T$. We further posit that each layer can be sparsely represented using appropriately chosen time-frequency dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ (practically realized below via STFTs with long and short analysis window lengths, respectively). That is, the layers are of the form $\mathbf{s}_S = \mathbf{\Phi}\boldsymbol{\alpha}$ and $\mathbf{s}_T = \mathbf{\Psi}\boldsymbol{\beta}$ with only few non-zero elements in the coefficient vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Specifically, the layer $\mathbf{s}_S$ (analogous for $\mathbf{s}_T$) can be written as

$$\mathbf{s}_S = \mathbf{\Phi}\boldsymbol{\alpha} = \sum_{\gamma} \alpha_\gamma \boldsymbol{\varphi}_\gamma \tag{1}$$

where for the sake of notational convenience $\gamma = (k, l)$ denotes a double labelling with $k = 1, \ldots, K$ and $l = 1, \ldots, L$ as frequency and time indices, respectively. The collection of time-localized atoms $\boldsymbol{\varphi}_\gamma \in \mathbb{R}^M$ constitute the dictionary $\mathbf{\Phi} \in \mathbb{R}^{(M, K\times L)}$, and α_γ are the STFT expansion coefficients. Here we use regular STFTs $\mathbf{\Phi}$ and $\mathbf{\Psi}$ with perfect reconstruction, corresponding to one form of tight Gabor dictionaries [18].

Because there is no access to the separate layers, the general problem is to estimate the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from an additive mixture with Gaussian white noise $\mathbf{e}$ (e.g., corresponding to measurement error):

$$\mathbf{y} = \mathbf{s}_S + \mathbf{s}_T + \mathbf{e} = \mathbf{\Phi}\boldsymbol{\alpha} + \mathbf{\Psi}\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

which can be stated as a dual-layer sparse regression problem [14],

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \Big\{ \frac{1}{2} \|\mathbf{y} - \mathbf{\Phi}\boldsymbol{\alpha} - \mathbf{\Psi}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 + \mu \|\boldsymbol{\beta}\|_1 \Big\}, \tag{3}$$

with sparsity parameters $\lambda, \mu > 0$. The solution of this problem is approximated by the *Iterative Shrinkage/Thresholding Algorithm* (ISTA) [13, 19], and more specifically its multi-layered extension [14, 15]:

$$\begin{cases} \boldsymbol{\alpha}^{(n+1)} & = \mathbb{S}_\lambda\big(\boldsymbol{\alpha}^{(n)} - \mathbf{\Phi}^*(\mathbf{y} - \mathbf{\Phi}\boldsymbol{\alpha}^{(n)} - \mathbf{\Psi}\boldsymbol{\beta}^{(n)})\big) \\ \boldsymbol{\beta}^{(n+1)} & = \mathbb{S}_\mu\big(\boldsymbol{\beta}^{(n)} - \mathbf{\Psi}^*(\mathbf{y} - \mathbf{\Phi}\boldsymbol{\alpha}^{(n)} - \mathbf{\Psi}\boldsymbol{\beta}^{(n)})\big), \end{cases} \tag{4}$$

with iteration index n and initializations $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = 0$. Here and in the following, the notation $(x)_+ = \max(x, 0)$ denotes the positive part of any $x \in \mathbb{R}$ and all operations are understood component-wise, i.e., per time-frequency index $\gamma = (k, l)$. For a complex-valued vector $\boldsymbol{\alpha}$, the operator $\mathbb{S}$ then refers to a shrinkage operation that is usually called *soft-thresholding*,

$$\mathbb{S}_\lambda(\boldsymbol{\alpha}) = \begin{Bmatrix} e^{i \arg \boldsymbol{\alpha}}(|\boldsymbol{\alpha}| - \lambda) & : |\boldsymbol{\alpha}| \geq \lambda \\ 0 & : |\boldsymbol{\alpha}| < \lambda \end{Bmatrix} = \boldsymbol{\alpha}\Big(1 - \frac{\lambda}{|\boldsymbol{\alpha}|}\Big)_+, \tag{5}$$

Whereas the optimization problem in (3) together with ISTA in (4) provide a theoretical footing for sparse multilayer decomposition, this approach has several drawbacks in practical situations. Most importantly, the independent handling of time-frequency coefficients does not utilize the full gamut of structure of stationary and transient layers. This is where the idea of *social sparsity* becomes useful.

2.2. Structured shrinkage via neighbourhoods

The approach of social sparsity extends classical shrinkage operators by the aspect of neighbourhood dependencies, yielding solutions of low computational cost that respect structural dependencies but are not strictly attached to known minimization functionals any more [17]. By generalizing the soft-thresholding operator in (5), we here focus on shrinkage operators of the form,

$$\mathbb{S}_{\lambda,\star}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}\left(1 - \left[\frac{\lambda}{\|\boldsymbol{\alpha}\|_\star}\right]^\tau\right)_+ \tag{6}$$

The placeholder $\star$ refers to a norm that allows to take into account neighbourhood structures and thus helps to orient the shrinkage operator towards stationary or transient components (e.g., by letting neighbourhoods extend across time for stationary components or frequency for transient components). By choosing a vector of non-negative time-frequency neighbourhood weights $\mathbf{w} = w_{\gamma,\gamma'}$, we can define the neighbourhood-based norm as

$$\|\boldsymbol{\alpha}\|_\star = \big(\|\boldsymbol{\alpha}\|_\star\big)_\gamma = \sqrt{\sum_{\gamma'} w_{\gamma\gamma'} |\boldsymbol{\alpha}_{\gamma'}|^2} \tag{7}$$

In practice we use sliding neighbourhoods, i.e., $w_{\gamma,\gamma'} = w_{0,\gamma-\gamma'}$, such that the norm $\|\cdot\|_\star$ can be efficiently computed via convolution:

$$\begin{aligned} \|\boldsymbol{\alpha}\|_\star^2 &= \textstyle\sum_{\gamma'} w_{\gamma,\gamma'} |\alpha_{\gamma'}|^2 \\ &= \textstyle\sum_{\gamma'} w_{0,\gamma-\gamma'} |\alpha_{\gamma'}|^2 \\ &= \big(\mathbf{w} * |\boldsymbol{\alpha}|^2\big)_\gamma \end{aligned} \tag{8}$$

The generic choice $\tau = 1$ leads to the *Windowed Group Lasso* [16], which constitutes a natural extension of the classic *Least Absolute Shrinkage/Selection Operator* (LASSO) [20], for which $\|\cdot\|_\star = |\cdot|$. Here, we focus on $\tau = 2$, which withdraws less energy from the signal compared to $\tau = 1$ and has been called *empirical Wiener* operator [21] or non-negative garotte shrinkage [22]. For single-layered expansions, previous research has shown that both the inclusion of neighbourhoods that extend across a few coefficients in time and the choice of $\tau = 2$ significantly improve audio noise removal [23] and declipping [24].

2.3. Structured shrinkage via modulation-filtering

Instead of defining the neighbourhood weights directly, they can also be defined in terms of their effect on the modulation spectrum of the shrinkage operation in (6) [25]. Let $\mathcal{F}_2$ denote the two-dimensional discrete Fourier transform on $\mathbb{C}^{K\times L}$, then (8) can be directly reformulated as

$$\|\boldsymbol{\alpha}\|_\star^2 = \Big[\mathcal{F}_2^{-1}\Big(\mathcal{F}_2(\mathbf{w}) \cdot \mathcal{F}_2(|\boldsymbol{\alpha}|^2)\Big)\Big] \tag{9}$$

The x-axis of the resulting modulation spectrum $\mathcal{F}_2(|\boldsymbol{\alpha}|^2)$ corresponds to temporal modulation measured in Hz, the y-axis to spectral modulation with unit cycles per Hz. This means the choice of the neighborhood $\mathbf{w}$ is equivalent to the choice of desired temporal and spectral modulation frequencies to be captured by the

modulation filter $\mathbf{W} := \mathcal{F}_2(\mathbf{w})$. For example, the usage of a rectangular neighbourhood corresponds to modulation filtering with a two-dimensional sinc-function centered at zero, i.e., a form of modulation low-pass filtering.

We here follow previous suggestions [26] and use an additional log-nonlinearity for computing the modulation spectrum:

$$\|\boldsymbol{\alpha}\|_{\sim} := \exp\left(\mathcal{F}_2^{-1}\left[\mathbf{W} \cdot \mathcal{F}_2\big(\log(|\boldsymbol{\alpha}| + \kappa)\big)\right]\right) - \kappa \qquad (10)$$

with a compression constant $\kappa = 1$. The log-nonlinearity may be justified by its resemblance to cepstral analysis, potentially separating the contributions of a signal's source and filter into additive contributions [26]. In conclusion, instead of shrinking coefficients in dependence of their neighbourhood's energy, an alternative perspective is to shrink these coefficients according to the energy retained by the modulation filter.

3. ALGORITHMS FOR STATIONARY/TRANSIENT SEPARATION

3.1. Iterative shrinkage/thresholding and cross-shrinkage

The point of departure of this study was to use structured shrinkage operators in iterative schemes such as the multilayered ISTA (4) for stationary/transient separation. Although a fast version of this algorithm, FISTA, has been proposed in [19], we could not observe improvements over ISTA for the considered application, and thus here only report on the regular ISTA.

When using ISTA in practice, we often encountered the problem that the transient layer was swallowed by the stationary layer, unless tedious tuning of the thresholds λ and μ was undertaken. For that reason, we also explored a related scheme, which follows a simple rationale: Assuming the estimate of the stationary layer ($\boldsymbol{\alpha}$) is accurate, the residual $\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\alpha} = \boldsymbol{\Psi}\boldsymbol{\beta} + \mathbf{e}$ mainly comprises components from the transient layer ($\boldsymbol{\beta}$) and thus allows for a more precise estimation of $\boldsymbol{\beta}$ than from the mixture. Due to the iterative estimation from the residual of the respective alternate layer, this yields a novel scheme which we call *Iterative Cross-Shrinkage* (ICS):

$$\begin{cases} \boldsymbol{\alpha}^{(n+1)} & = \mathbb{S}_{\lambda}\big(\boldsymbol{\Phi}^*\big(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta}^{(n)}\big)\big) \\ \boldsymbol{\beta}^{(n+1)} & = \mathbb{S}_{\mu}\big(\boldsymbol{\Psi}^*\big(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}^{(n)}\big)\big) \end{cases} \qquad (11)$$

with initializations $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = 0$. Essentially, this corresponds to ISTA with zero contribution of the respective previous iterate of each layer (e.g., with $\boldsymbol{\alpha}^{(n)}$ set to zero in the estimation of $\boldsymbol{\alpha}^{(n+1)}$).

3.2. Choice of thresholds

The right selection of the thresholds λ and μ is critical in applications. In scenarios with strong additive noise, the optimal thresholds naturally depend on the noise level [27]. In the stationary/transient separation scenario, however, additive noise does not play a similarly as crucial role such that alternative strategies for selecting the thresholds can be sought. In addition to using fixed thresholds, we here explored so-called *warm-start* strategies [28, 24]. These strategies start out conservatively with relatively high thresholds which are successively reduced and thus allow for more liberal estimates towards the end.

In the first warm-start strategy, we chose $\lambda^{(n)} = \lambda$ and $\mu^{(n)} = \mu$ as piece-wise constant sequences that decreased after every 10th iteration. Specifically, the thresholds were set equal to the P% quantiles q_P of the distribution of the magnitudes of the initial analysis coefficients of the respective layer, i.e., $\lambda^{(n)} = q_P(|\boldsymbol{\Phi}^*\mathbf{y}|)$ and $\mu^{(n)} = q_P(|\boldsymbol{\Psi}^*\mathbf{y}|)$. P was set equal to 99% and linearly decreased after every 10th iteration by around 2.11 percentage points, reaching 80% at iteration no. 91 (yielding 100 iterations in total). See Fig. 4 (right panel) for an example trajectory.

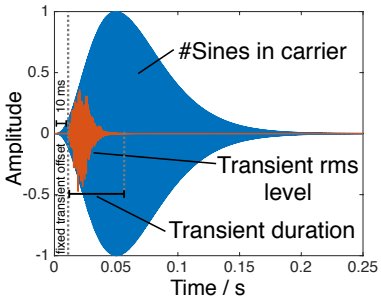

Figure 1: Exemplary test signal. Transients varied according to duration and level. Stationary parts varied by the number of sinusoidal components of their carrier.

A shortcoming of this strategy is that it is based on an inherently imperfect estimate of the magnitude distribution of the expansion coefficients of individual layers, because $|\boldsymbol{\Phi}^*\mathbf{y}| = |\boldsymbol{\Phi}^*\mathbf{s}_S + \boldsymbol{\Phi}^*\mathbf{s}_T|$ obviously contains both layers. Say we were aiming to adjust the threshold λ according to the true underlying magnitude distribution $q_P(|\boldsymbol{\Phi}^*\mathbf{s}_S|)$, it should be beneficial to update this threshold according to presumably more precise estimates of the individual layers that only arise at later iterations. For that reason, we also tested a second warm-start strategy where the thresholds were adjusted dynamically. That means, $\lambda^{(n)}$ and $\mu^{(n)}$ were updated after every 10th step according to the magnitude of the argument of the shrinkage operators in ISTA (4) and ICS (11). For ICS, for instance, this would correspond to $\lambda^{(n)} = q_P(|\boldsymbol{\Phi}^*\big(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta}^{(n)}\big)|)$. As before, P started at 99% and linearly decreased after every 10th iteration to reach 80% for the last 10 iterations (see Fig. 4).

4. EXPERIMENTS WITH SYNTHESIZED SOUNDS

In this section, we describe simulation experiments with signals for which the stationary and transient parts were artificially synthesized. We tested the ISTA and ICS algorithms in conjunction with three types of threshold selection strategies (cf., Sec. 3.2) and three ways of incorporating inter-coefficient structure (cf., Sec. 2.2 and 2.3).

4.1. Test stimuli and factors

We synthesized a set of fixed-frequency sinusoids and Gamma-shaped noise bursts as test components. For the stationary component, sinusoidal frequencies were randomly and uniformly chosen between 100 and 10,000 Hz and shaped with an Gamma-type amplitude envelope of 245 ms effective duration. The transient components were generated with white Gaussian noise, which was shaped by a much shorter Gamma-envelope. Specifically, we varied three factors (using four levels for each of them):

i) The transient Gamma-shaped white noise bursts had effective durations between 4.9 and 49 ms.

ii) The level of the transient relative to the stationary component was adjusted between -30 and 0 dB.

iii) For the stationary component, harmonic tone complexes comprised 1 to 50 sinusoids (i.e., corresponding to the sparsity of the stationary signal).

The Gamma-envelope was of the form

$$e(t) = t^{(Q-1)} \exp(-2\pi bt),$$

where the order $Q = 4$ was fixed and t denotes time in seconds. The scale parameter b was set as $b = Q - 1/(2\pi\eta)$, where η is the underlying Gamma distribution's mode (or the tone's rise time). From this, it can be inferred that the effective duration of any resulting sound (with a threshold of -60 dB) amounts to 4.9η.

Transients were delayed by 10 ms in order to ensure even for short transients a significant overlap with the energy of the slower rising envelope of the stationary components (i.e., to prohibit the possibility of overly simplistic separation). The additive combination of the stationary and transient components was used in conjunction with a white Gaussian noise floor at a signal-to-noise ratio of 40 dB. Throughout all numerical simulations reported in this paper, audio signals were sampled at 44.1 kHz. Fig. 1 shows an illustration of the stationary and transient components of a test signal.

4.2. Algorithm settings

We chose Gabor dictionaries with a tight Hann (raised cosine) window and a hop size of a quarter of the window length. In order to capture stationary components, $\mathbf{\Phi}$ was chosen with a window length of 2048 samples (46 ms). For capturing transient components, $\mathbf{\Psi}$ was chosen with a window length of 128 samples (3 ms). All simulations were performed with shrinkage exponent $\tau = 2$.

As outlined in Sec. 3.2, three strategies for choosing the thresholds $\lambda^{(n)}$ and $\mu^{(n)}$ were considered: i) A fixed threshold at the 80% quantile of each layer's initial analysis coefficients (denoted as *fix*), ii) a sequence of quantiles, linearly decreasing from the 99% quantile to the 80% quantile of the initial analysis coefficients (*quant*), and iii) dynamically decreasing thresholds (*dyn*).

The utility of exploiting inter-coefficient structure was investigated by comparing neighbourhood-based shrinkage (denoted as *Neigh.*) and modulation-based shrinkage (*Modul.*) to shrinkage with independent-coefficients (*Indep.*). For shrinkage of the stationary layer $\mathbb{S}_\lambda$, neighbourhoods comprised two coefficients forward and two coefficients backward of the centre coefficient. For shrinkage of the transient layer $\mathbb{S}_\mu$, the neighbourhood extended for three coefficients above and three coefficient below the centre coefficient. For modulation-based shrinkage, we chose W as a separable two-dimensional Gaussian distribution centred at the origin of the modulation spectrum. In order to tune in on spectral information for shrinkage of the stationary layer, the Gaussian's standard deviations were set to (1, 0.1) for the spectral scale and temporal rate axes, respectively, and the distribution was evaluated across the range $[-1, 1] \times [-1, 1]$. For the transient layer, we chose the reverse settings (0.1, 1), mainly tuning in on temporal information.

4.3. Results

We measured the performance in terms of the estimation accuracy of the 100th iterate of the above presented algorithms using the signal-to-distortion ratio (SDR). For the stationary layer $\mathbf{s}_S$, for instance, this corresponds to

$$\mathrm{SDR}(\mathbf{s}_S, \mathbf{\Phi}\hat{\boldsymbol{\alpha}}) = 20 \log_{10} \big(\frac{\|\mathbf{s}_S\|}{\|\mathbf{s}_S - \mathbf{\Phi}\hat{\boldsymbol{\alpha}}\|} \big).$$

We present results in terms of the SDR improvement compared to the unprocessed signal, i.e., $\Delta\mathrm{SDR} = \mathrm{SDR}(\mathbf{s}_S, \mathbf{\Phi}\hat{\boldsymbol{\alpha}}) - \mathrm{SDR}(\mathbf{s}_S, y)$.

Fig. 2 shows the mean ΔSDR values for all $2 \times 3 \times 3$ algorithmic variants, averaged across the three stimulus factors of transient duration, level, and number of sinusoids in the stationary carrier signal (each with four levels, such that every data point corresponds to a mean across 64 test signals). Both for the stationary and transient components, the ISTA and ICS methods yield similar performance in conjunction with fixed thresholds. At the same time, fixed thresholds yield negative ΔSDR values for the stationary layer, which indicates that this strategy has significant shortcomings. Whereas ISTA works best in conjunction with the quantile-based update of the thresholds and fails for the dynamic update, ICS seems to particularly profit from this latter strategy. Furthermore, ΔSDR is generally higher with neighbourhoods compared to the independent handling of coefficients and best performance is reached for modulation filtering.

Overall, the figure clearly illustrates the superior performance of ICS *dyn* with gains of around 10 dB SDR compared to the best-performing variant of ISTA (*quant*) for both stationary and transient layers. Compared to the initial reference algorithm—the dual-layered ISTA with independent coefficients originally proposed in [14, 15]—we were thus able to achieve improvements of more than 20 dB SDR.

In order to additionally compare our algorithm to an orthogonal transform, we used the dual-layered expansion of the Modified Discrete Cosine Transform, which also serves as a dual-layer decomposition example of the Large Time Frequency Analysis Toolbox [29]. This configuration achieved an average $\Delta\mathrm{SDR}_S = -2.8$ and $\Delta\mathrm{SDR}_T = 4.5$, thus markedly worse for transient estimation compared to the most rudimentary variant of our Gabor-dictionary-based ISTA approach with fixed thresholds and independent coefficients.

Fig. 3 depicts a more detailed picture of the performance of the three best-performing algorithmic variants: ISTA and ICS with quantile-based thresholds, and ICS with dynamic threshold update. These figures demonstrate that performance drops as the durations of the transients grow, which is expected, given that the differences in time-scale of stationary and transient components increasingly vanish. Yet, most algorithmic variants are still able to achieve a substantial SDR improvement even for the longest transients of 49 ms, which span more than a whole Gabor atom of the stationary layer (46 ms length). The dependency of ΔSDR on the transient level appears to be fairly linear. As expected, the biggest improvements of the stationary ΔSDR occur for conditions with strongest transients, and vice versa for the transient ΔSDR. Finally, when it comes to the number of sinusoidal components in the stationary carrier signal, that is, its sparsity, the dynamic ICS appears to be particularly robust, whereas the rest of the algorithms yields a sharp drop in performance already for four sinusoids.

Regarding the iterative behavior of the algorithms, Fig. 4 shows an example of the three best-performing algorithmic variants (using modulation filtering) across iterations as well as the corresponding evolution of thresholds (rightmost panel). It is clearly

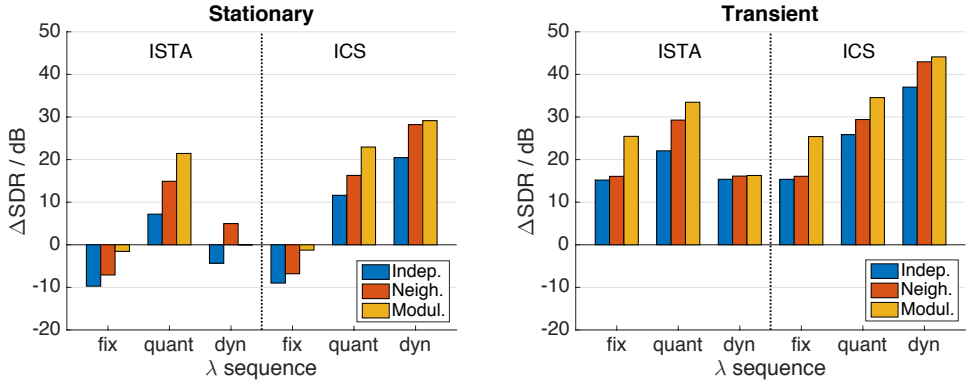

Figure 2: Mean SDRs improvement across all acoustic conditions for different algorithms (ISTA, ICS), different shrinkage operators (Indep., Neigh., Modul.), and different threshold update strategies (fix, quant, dyn).

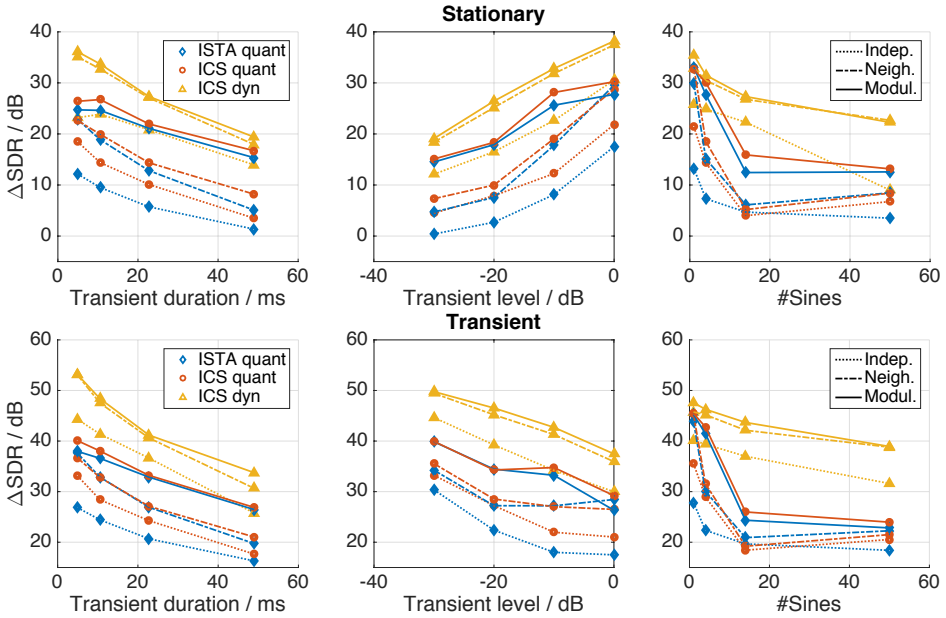

Figure 3: SDR improvement for estimates of stationary and transient components as a function of the algorithm type (indicated by symbols and line color, see left legend) and inter-coefficient structure (indicated by linetype, see right-hand-side legend).

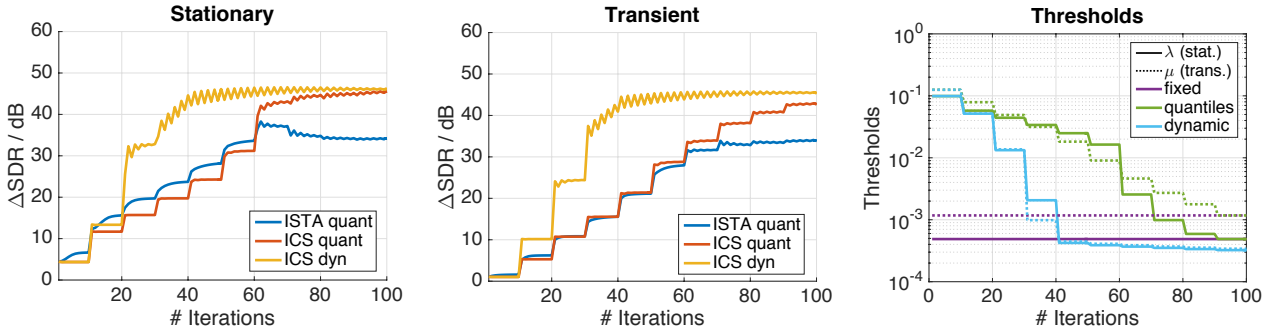

Figure 4: Exemplary iteration. Left and center panel: SDR improvement across 100 iterations for modulation filtering and operator types (see color legend). The sound's stationary part comprised 4 sinusoids and the transient was of 10 ms duration at an RMS level of 0 dB. Rightmost panel: Evolution of thresholds of stationary components (solid lines) and transient components (dotted lines). The threshold update is color-coded, see legend.

visible how after every 10 iterations, decreases of thresholds are accompanied by increases in ΔSDR. Notably, the ICS with dynamic threshold update appears to converge much faster, likely due to a steeper slope of thresholds across iterations for the dynamic update. Also note that the dynamic threshold update reaches the same absolute threshold values for both layers, which stands in contrast to the other two methods for which the 80% quantile of the stationary layer is much smaller compared to that of the transient layer.

Due to their rather heuristic nature, our findings motivate further mathematical and experimental inquiries into the formal roots of the proposed algorithms. For instance, it is currently unclear why ISTA appears to be incompatible with the dynamic threshold update whereas ICS profits substantially from it. It is also questionable whether this gain in performance (from ICS *quant* to ICS *dyn*) is due to the update's actual adaptivity or solely based on a steeper decrease of thresholds over iterations. Finally, it is left completely open whether it could be beneficial to replace the STFT dictionary with short window lengths by a wavelet basis, which may be better suited to account for the stochastic nature of transients [10]. On the other hand, the proposed algorithms appear to be robust enough already in order to be useful for stationary/transient separation in practical situations, as described in the next section.

5. EXAMPLES WITH RECORDED AUDIO

We considered natural recorded instrumental tones produced by a violoncello, a vibraphone, and a harpsichord. Each sound was of 500 ms duration, fundamental frequency 311 Hz, and of equal perceptual loudness as used in [30]. Due to its continuous mode of excitation, the violoncello is a quasi-harmonic sound without marked attack transient, yet with low-energy noise components stemming from the bow. As an impulsively excited sound, the vibraphone features strong attack transient at its onset. Interestingly, the harpsichord comprises one transient component at its onset, but also one at the release of its hopper while other sinusoidal components sustain. Fig. 5 presents the waveform of these three sounds in sequence, their spectrograms, as well as the separation provided by ICS with dynamic threshold update and modulation filtering.[1]

For this example, the stationary layer of the ISTA algorithm swallows the transient layers, i.e., the separation fails. On the contrary, the ICS algorithm provides non-zero estimates of transients for all three sounds, even for the onset noise components of the cello bow. As visible in the figure, the vibraphone contains the strongest transient component, which is clearly separated from the remaining sinusoidal components. Finally, the two transients of the harpsichord sound are well separated, even though they fully overlap in time. The latter sound once again illustrates that separation performance is not at all relying on temporal separation, but on distinct spectrotemporal shapes of stationary and transient signal components.

6. CONCLUSION

In this paper, we presented novel strategies for stationary/transient signal separation. Several shrinkage operators were defined by considering either the energy of time-frequency neighbourhoods or modulation spectrum regions instead of individual coefficients. These shrinkage operators were specifically tuned to the presumed sparsity and persistence properties of stationary and transient components, exploiting the basic observation that stationary components are sparse in frequency and persistent over time, and vice versa for transients. This step extends the usage of structured shrinkage operators to the context of dual-layer decomposition. We also proposed a novel iteration scheme, *Iterative Cross-Shrinkage*, which appears to work particularly well in conjunction with a dynamic update of the thresholds. In experiments with artificial test signals, the proposed scheme improved stationary/transient separation by surprisingly large margins by about 10 dB SDR compared to the dual-layered ISTA with neighbourhood/modulation persistence. Compared to the dual-layered ISTA with independent coefficients, we were able to achieve improvements of more than 20 dB SDR. In addition, the application of Iterative Cross-Shrinkage to recorded sounds from acoustic musical instruments provided a perceptually convincing separation of transients.

[1] These and additional audio examples can be accessed via `http://www.uni-oldenburg.de/en/mediphysics-acoustics/sigproc/research/audio-demos/`.

7. REFERENCES

[1] J. W. Beauchamp, "Analysis and synthesis of musical instrument sounds," in *Analysis, Synthesis, and Perception of Musical Sounds*, J. W. Beauchamp, Ed. Springer, 2007, pp. 1–89.

[2] S. J. Godsill and P. J. Rayner, *Digital audio restoration*. New York, NY: Springer, 2002.

[3] M. Müller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[4] S. Handel, "Timbre perception and auditory object identification," in *Hearing*, ser. Handbook of Perception and Cognition, B. C. Moore, Ed. San Diego, CA: Academic Press, 1995, vol. 2, pp. 425–461.

[5] C. Laroche, M. Kowalski, H. Papadopoulos, and G. Richard, "A structured nonnegative matrix factorization for source separation," in *Proc. of the 23rd European Signal Processing Conference (EUSIPCO)*. Nice, France: IEEE, 2015, pp. 2033–2037.

[6] C. Laroche, H. Papadopoulos, M. Kowalski, and G. Richard, "Genre specific dictionaries for harmonic/percussive source separation," in *Proc. of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, NY, 2016.

[7] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source seperation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[8] G. Evangelista, "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3313–3330, 1993.

[9] P. Polotti and G. Evangelista, "Analysis and synthesis of pseudo-periodic1/f-like noise by means of wavelets with applications to digital audio," *EURASIP Journal on Applied Signal Processing*, no. 1, pp. 1–14, 2001.

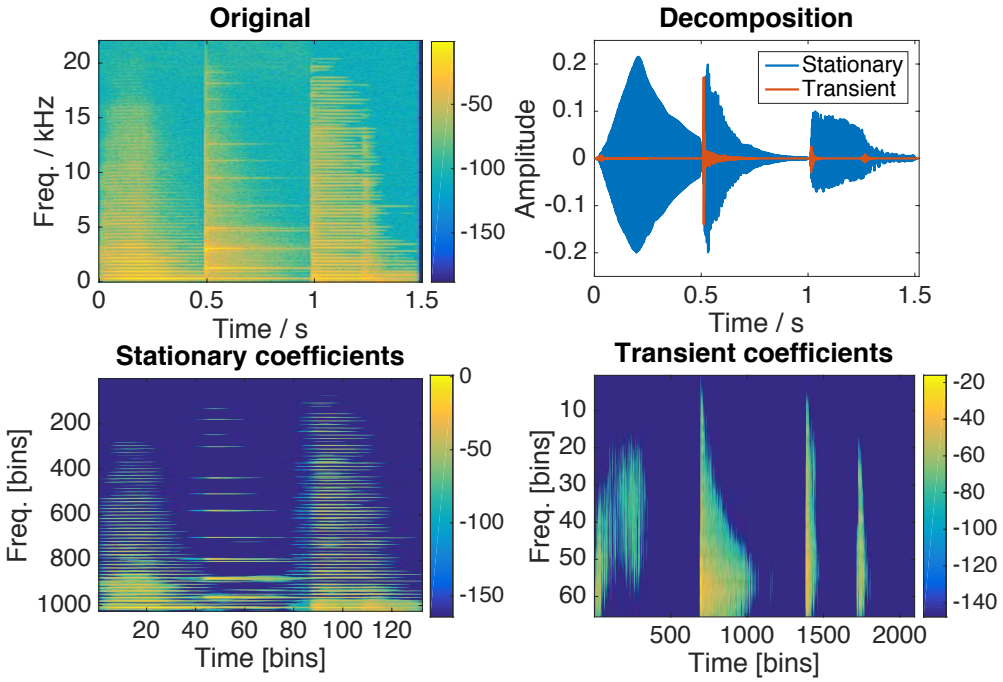

Figure 5: The ICS algorithm with dynamic threshold update and modulation filtering for a sequence of three recorded musical instrument tones (from left to right: violoncello, vibraphone, harpsichord). Top left figure shows the original signal spectrogram. Top right figure shows the time-domain representation of the separation. Bottom left and right figures show the estimated magnitude coefficients for the stationary (left) and transient (right) components. Axes units are in bins to highlight the different time-frequency resolutions.

[10] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002.

[11] S. Molla and B. Torresani, "A hybrid scheme for encoding audio signal using hidden markov models of waveforms," *Applied and Computational Harmonic Analysis*, vol. 18, pp. 137–166, 2005.

[12] C. Févotte, B. Torresani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 174–185, 2008.

[13] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[14] G. Teschke, "Multi-frame representations in linear inverse problems with mixed multi-constraints," *Applied and Computational Harmonic Analysis*, vol. 22, pp. 43–60, 2007.

[15] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303 – 324, 2009.

[16] M. Kowalski and B. Torresani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 251–264, 2008a.

[17] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social sparsity! Neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[18] M. Dörfler, "Time-frequency analysis for music signals a mathematical approach," *Journal of New Music Research*, vol. 30, no. 1, pp. 3–12, 2001.

[19] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM, Journal of Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.

[21] S. P. Ghael, A. M. Sayeed, and R. G. Baraniuk, "Improved wavelet denoising via empirical Wiener filtering," in *Proceedings of SPIE*, vol. 3169. San Diego, CA, 1997, pp. 389–399.

[22] A. Antoniadis *et al.*, "Wavelet methods in statistics: Some recent developments and their applications," *Statistics Surveys*, vol. 1, pp. 16–55, 2007.

[23] K. Siedenburg and M. Dörfler, "Persistent time-frequency shrinkage for audio denoising," *Journal of the Audio Engineering Society (AES)*, vol. 61, no. 1/2, pp. 29–38, 2013.

[24] K. Siedenburg, M. Kowalski, M. Dörfler *et al.*, "Audio declipping with social sparsity," in *Proc. of the IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014, pp. 1577–1581.

[25] K. Siedenburg and P. Depalle, "Modulation filtering for structured time-frequency estimation of audio signals," in *Proc. of the IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, pp. 1–4.

[26] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computational Biology*, vol. 5, no. 3, p. e1000302, 2009.

[27] K. Siedenburg, "Persistent empirical Wiener estimation with adaptive threshold selection for audio denoising," in *Proceedings of the 9th Sound and Music Computing Conference*, Copenhagen, DK, 2012.

[28] I. Loris, "On the performance of algorithms for the minimization of l1-penalized functionals," *Inverse Problems*, vol. 25, 2009.

[29] P. L. Søndergaard, B. Torrésani, and P. Balazs, "The Linear Time Frequency Analysis Toolbox," *International Journal of Wavelets, Multiresolution Analysis and Information Processing*, vol. 10, no. 4, 2012.

[30] K. Siedenburg, K. Jones-Mollerup, and S. McAdams, "Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds," *Frontiers in Psychology*, vol. 6, no. 1977, p. doi: 10.3389/fpsyg.2015.01977, 2016.