UNIFYING LOCAL AND GLOBAL METHODS FOR HARMONIC-PERCUSSIVE SOURCE SEPARATION

Christian Dittmar, Patricio López-Serrano, Meinard Müller

International Audio Laboratories Erlangen, Germany
christian.dittmar@audiolabs-erlangen.de

ABSTRACT

This paper addresses the separation of drums from music recordings, a task closely related to harmonic-percussive source separation (HPSS). In previous works, two families of algorithms have been prominently applied to this problem. They are based either on local filtering and diffusion schemes, or on global low-rank models. In this paper, we propose to combine the advantages of both paradigms. To this end, we use a local approach based on Kernel Additive Modeling (KAM) to extract an initial guess for the percussive and harmonic parts. Subsequently, we use Non-Negative Matrix Factorization (NMF) with soft activation constraints as a global approach to jointly enhance both estimates. As an additional contribution, we introduce a novel constraint for enhancing percussive activations and a scheme for estimating the percussive weight of NMF components. Throughout the paper, we use a real-world music example to illustrate the ideas behind our proposed method. Finally, we report promising BSS Eval results achieved with the publicly available test corpora ENST-Drums and QUASI, which contain isolated drum and accompaniment tracks.

Index Terms— harmonic percussive source separation, music decomposition, signal reconstruction.

1. INTRODUCTION

The general goal of music source separation is to decompose a recording into its constituent signal components. Considering the challenging scenario of percussive and non-harmonic sound sources, we aim to extract drum sound events in a perceptually convincing quality that allows remixing and repurposing [1]. Going beyond our previous paper, we focus on drum recordings with moderate interference from melodic instruments. In particular, we are interested in decomposing break sections that appear in pop, jazz, soul, and funk recordings of the 1960's to 1980's [2]. Such instrumental passages are often characterized by a pronounced drum beat interspersed with melodic instruments (e. g., bass, guitar, organ, saxophone, trumpet) and, rarely, singing voice.

In Figure 1, we introduce an idealized example for our source separation task using an excerpt of the 1964 recording of "I Got You (I Feel Good)," by James Brown & The Famous Flames. This short instrumental break features Maceo Parker playing alto sax over a four-bar drum beat played by his brother Melvin Parker. This running example appears several times throughout the paper. For now, it is sufficient to realize that we are interested in removing the sound events of the alto sax (corresponding to the notes in the upper staff of Figure 1a).

Harmonic-percussive source separation (HPSS) seems like a natural choice for attenuating the interference of melodic instruments into the drum part. State-of-the-art HPSS methods are compared in detail in [3] and in [4]. Generally, there are two conceptually different HPSS approaches: local and global methods. Local HPSS methods emphasize localized time-frequency (TF) characteristics that distinguish drums from melodic instruments. In Section 2.1, we will briefly recapitulate a local HPPS method based on Kernel Additive Modeling (KAM) [5, 6]. Global HPSS methods decompose the mixture spectrogram into low-rank components according to a global optimization criterion. As we explain in Section 2.2, we use Non-Negative Matrix Factorization (NMF) to do so. Especially for HPSS, many authors advocate to apply constraints to NMF [4, 7, 8, 9, 10], exploiting the different TF characteristics of drums and melodic instruments.

In this paper, we propose a novel two-stage approach, unifying local and global methods. In a first stage, KAM finds initial estimates for the percussive and harmonic parts (see Algorithm 1 in Section 2.1). The second stage then jointly refines these estimates using NMF with soft activation constraints (see Algorithm 2 in Section 2.4). As an additional contribution, we introduce the notion of percussive weight, an adaptive measure implicitly classifying NMF components according to their contribution to the drums. We explain how the percussive weight can be easily derived in our framework. Furthermore, we introduce a soft constraint for NMF activations that emphasizes drum-specific temporal characteristics. The experiments in Section 3 show that our proposed method yields improved BSS Eval measures on the ENST-Drums and QUASI corpora. Finally, in Section 4, we discuss strengths and weaknesses of our approach and point out remaining challenges.

2. PROPOSED SYSTEM

As is common in music source separation, we decompose the mixture signal in the Short-time Fourier Transform (STFT) domain. To this end, let $\mathrm{A}\,(k, m)$ be the non-negative STFT magnitude at the k^{th} frequency bin and the m^{th} time frame, with $k \in [0 : K-1]$ and $m \in [0 : M-1]$. The number of available bins $K \in \mathbb{N}$ and frames $M \in \mathbb{N}$ determines the dimension of our mixture spectrogram matrix $\mathrm{A} \in \mathbb{R}_{\geq 0}^{K \times M}$. Our objective is to split the mixture A in two complementary magnitude spectrograms $\mathrm{A_p}$ (drum part) and $\mathrm{A_h}$ (melodic part), such that $\mathrm{A} = \mathrm{A_p} + \mathrm{A_h}$. As shown in Figure 2, our main idea is to decompose A using KAM and NMF in a cascade. KAM serves to find initial estimates $\mathrm{A_p^{KAM}}$ and $\mathrm{A_h^{KAM}}$ which

Christian Dittmar and Meinard Müller are supported by the German Research Foundation (DFG-MU 2686/10-1). Patricio López-Serrano is supported in part by CONACYT-DAAD. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

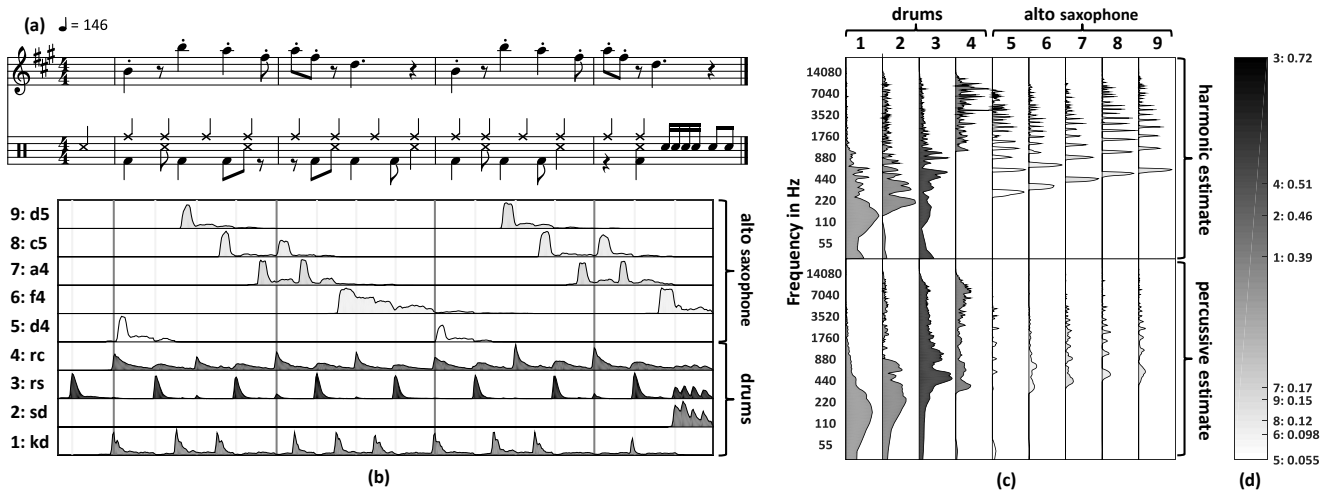

Fig. 1. Instrumental break from "I Got You (I Feel Good)". **(a)** Score notation of alto sax and drums. **(b)** NMF activations of alto sax and drums. Bar and beat grid are indicated in the background and aligned to the score notation. **(c)** NMF templates corresponding to the activations. **(d)** Percussive weight $\mathrm{p}\,(r)$ encoded in the gray scale, with indices referring to the NMF components.

are then jointly refined by NMF, yielding the final $\mathrm{A_p^{NMF}}$ and $\mathrm{A_h^{NMF}}$.

2.1. Local HPSS using KAM

Local HPSS methods rely on filtering and diffusion operations and are surprisingly effective for separating percussion instruments from melodic instruments. These methods exploit local characteristics apparent in TF representations of the music mixture (often the magnitude spectrogram). Typically, tonal TF areas are assigned to the harmonic component, and transient TF areas to the percussive component [5, 11, 12, 13]. These simplifications are strongly related to classic signal processing paradigms such as sinusoids & transient & noise modeling [14]. However, they are also problematic, since many drum instruments such as kick drums, tom toms, and cymbals exhibit strong tonal (yet inharmonic) components with relatively slow decay. Since harmonicity is a concept based on relationships between sinusoids positioned within a harmonic series (rather than a narrow frequency neighborhood), local methods are usually not suited to emphasize these characteristics. Thus, it may happen that TF components belonging to the percussion are erroneously treated as tonal. In practice, this often leads to audible leakage of the drums' decay into the harmonic signals. In contrast, the separated drum signals may sound unnatural and exhibit severe audible artifacts. As we will explain in Section 2.4, our novel approach can help to recover from these errors to a certain extent.

In Algorithm 1, we detail our variant of KAM-based HPSS [6], been originally proposed in [5] as a generalization of the median-filtering method [12]. The gist of this iterative procedure is to alternate between localized enhancement of percussive and harmonic structures [5] and generalized Wiener filtering [15, 6]. To this end, the estimates $\mathrm{A_h^{(0)}}$ and $\mathrm{A_p^{(0)}}$ are initialized with identical copies of the the mixture spectrogram A. Two filter kernels $\mathcal{I}_\mathrm{p} \in \mathbb{R}_{\geq 0}^{\kappa \times 1}$ and $\mathcal{I}_\mathrm{h} \in \mathbb{R}_{\geq 0}^{1 \times \kappa}$ are used to enhance percussive and harmonic structures, respectively. Kernel $\mathcal{I}_\mathrm{p}$ is Hann-shaped and oriented vertically (column vector). Kernel $\mathcal{I}_\mathrm{h}$ holds the same coefficients in perpendicular orientation (row vector). The kernel width $\kappa \in \mathbb{N}$ determines the smoothing strength (potentially, it could be tuned individually for each kernel). In Algorithm 1, the operator $*$ denotes 2D convolution with appropriate zero padding. This operation yields the percussively and harmonically enhanced estimates $\mathrm{B_p}$ and $\mathrm{B_h}$, respectively. Both $\mathrm{B_p}$ and $\mathrm{B_h}$ are used for Wiener filtering, where the multiplication $\odot$ and division $\oslash$ are performed element-wise.

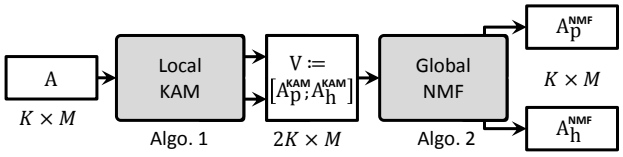

Fig. 2. Overview of our proposed method.

Algorithm 1: KAM-based HPSS.

Input: $\mathrm{A_p^{(0)}} := \mathrm{A}$ and $\mathrm{A_h^{(0)}} := \mathrm{A}$ with $L = L^{\mathrm{KAM}}$
for $\ell = 0, 1, 2, \ldots, L-1$ **do**
  $\left.\begin{aligned} \mathrm{B_p} &:= \mathrm{A_p^{(\ell)}} * \mathcal{I}_\mathrm{p} \\ \mathrm{B_h} &:= \mathrm{A_h^{(\ell)}} * \mathcal{I}_\mathrm{h} \end{aligned}\right\}$ 2D convolution
  $\left.\begin{aligned} \mathrm{A_p^{(\ell+1)}} &:= \mathrm{A} \odot \mathrm{B_p} \oslash (\mathrm{B_p} + \mathrm{B_h}) \\ \mathrm{A_h^{(\ell+1)}} &:= \mathrm{A} \odot \mathrm{B_h} \oslash (\mathrm{B_p} + \mathrm{B_h}) \end{aligned}\right\}$ Wiener filtering
end
Output: $\mathrm{A_p^{KAM}} := \mathrm{A_p^{(L)}}$ and $\mathrm{A_h^{KAM}} := \mathrm{A_p^{(L)}}$.

2.2. Global HPSS using NMF

Global HPSS approaches factorize the mixture spectrogram into low-rank components and cluster them into percussive and harmonic subsets. As an early example, components uncovered via Independent Subspace Analysis were classified as either percussive or harmonic via spectral and temporal low-level features in [16]. Later works employed more suitable factorization techniques such

as NMF [4, 7, 9, 10, 17, 18], Non-Negative Tensor Factorization [19], or Non-negative Matrix Factor Deconvolution [1, 20].

NMF is a popular algorithm [21] for iteratively computing a global, low-rank approximation $\mathrm{V} \approx \mathrm{W} \cdot \mathrm{H}$. In the context of music source separation, $\mathrm{V} \in \mathbb{R}_{\geq 0}^{K \times M}$ is usually the magnitude spectrogram, the columns of $\mathrm{W} \in \mathbb{R}_{\geq 0}^{K \times R}$ are interpreted as spectral basis functions (also called templates), and the rows of $\mathrm{H} \in \mathbb{R}_{\geq 0}^{R \times M}$ are interpreted as time-varying gains (also called activations). Different cost functions can be used to quantify the approximation quality, giving rise to specific iterative update rules. Following the recent paper by Park et al. [4], we also use the Kullback-Leibler divergence.

Since the optimal rank $R \in \mathbb{N}$ is generally unknown and dependent on the content in V, it is commonly set to a sufficiently high number (e. g., to 750 in [4]). In practice, this is problematic since components learned by NMF may represent atomic parts of the individual sources of interest. Automatically clustering them to form musically meaningful parts can be very challenging. For now, let us assume that we already know to which extent each of the R components contributes to the percussive part. Formally, we express this by introducing a percussive weight vector $\mathrm{p} \in \mathbb{R}_{\geq 0}^{1 \times R}$, with entries $0 \leq \mathrm{p}(r) \leq 1$, $r \in [1 : R]$. The values of $\mathrm{p}(r)$ define a continuum between the percussive ($\mathrm{p}(r) = 1$) and harmonic ($\mathrm{p}(r) = 0$) extremes. With this pre-requisite, it is straightforward to construct the percussive and harmonic estimates as:

$$\begin{aligned} \mathrm{V}_{\mathrm{p}}^{\mathrm{NMF}} &:= \mathrm{W} \cdot (\mathrm{P} \odot \mathrm{H}), \\ \mathrm{V}_{\mathrm{h}}^{\mathrm{NMF}} &:= \mathrm{W} \cdot ((1 - \mathrm{P}) \odot \mathrm{H}), \end{aligned} \tag{1}$$

where the percussive weight matrix $\mathrm{P} \in \mathbb{R}_{\geq 0}^{R \times M}$ just replicates the elements of p over all columns.

In Figure 1b and Figure 1c, we show activations and templates for our running example, the shades of gray encoding $\mathrm{p}(r)$ as given in Figure 1d. We will explain how to determine $\mathrm{p}(r)$ in Section 2.4. Note that the first four drum-related components exhibit decaying impulses at the time instances corresponding to drum hits. In contrast, the remaining sax-related components exhibit plateau-like activations when notes are being played.

2.3. Soft Activation Constraints

Many authors proposed to apply constraints to NMF in order to guide the iterative factorization process towards a meaningful solution. Usually, these constraints are applied in between the regular NMF update rules by manipulating both W and H. For example, gamma priors were used in [19] to encourage temporal continuity in the activations of harmonic components that are assumed to vary slowly in time. Similarly, constraints promoting smoothness of the activations and sparseness of the templates were used in [7]. A method for emphasizing harmonic structures in some templates was proposed in [9], while spectral and temporal continuity constraints were introduced via a weighted sum scheme in [4].

In this paper, we propose two non-linear operations which are applied exclusively to the individual rows of H prior to the NMF updates. Analogous to the HPSS method by Fitzerald [12], harmonically enhanced activations H_h are computed by median filtering along the horizontal axis as:

$$\mathrm{H}_h(r, m) := \mathrm{median}\left(\mathrm{H}(r, m - \tau), \ldots, \mathrm{H}(r, m + \tau)\right), \tag{2}$$

for $\tau \in \mathbb{N}$ with $2\tau + 1$ being the length of the median filter. In the following, we will use the operator $\mathcal{H}(\cdot)$ to denote this operation. In Figure 1b, $\mathcal{H}(\cdot)$ would emphasize the plateau-like shapes in the activations of the alto sax while suppressing impulse-like patterns.

In contrast, percussively enhanced activations H_{p} are derived by applying a non-linear exponential moving average filter (NEMA). Setting the first element $\mathrm{H}_{\mathrm{p}}(r, 0) := \mathrm{H}(r, 0)$, the NEMA operation applied to the r^{th} row can be described as follows:

$$\mathrm{H}_{\mathrm{p}}(r, m) := \max \begin{cases} \mathrm{H}(r, m) \\ \lambda \cdot \mathrm{H}_{\mathrm{p}}(r, m - 1) + (1 - \lambda) \cdot \mathrm{H}(r, m) \end{cases} \tag{3}$$

for $m \in [1 : M - 1]$. The weight $\lambda \in \mathbb{R}$ with $0 < \lambda < 1$ controls the decay of this recursive filter, which we denote by $\mathcal{P}(\cdot)$ in the following. As illustrated in Figure 1b, $\mathcal{P}(\cdot)$ promotes the development of sharp peaks followed by a moderate decay. Contrary to the common belief that impulse-like NMF activations are suited to model drum sound events, we showed in [1] that decaying impulses are more appropriate for music source separation.

Algorithm 2: NMF-based HPSS.

Input:
Concatenate KAM results $\mathrm{V} := \left[\mathrm{A}_{\mathrm{p}}^{\mathrm{KAM}}; \mathrm{A}_{\mathrm{h}}^{\mathrm{KAM}}\right]$
Initialize all-ones matrix $\mathrm{J} \in \mathbb{R}_{\geq 0}^{2K \times M}$
Initialize $\mathrm{H}^{(0)}$ and $\mathrm{W}^{(0)}$ with non-negative random values
Set $L = L^{\mathrm{NMF}}$
for $\ell = 0, 1, 2, \ldots, L - 1$ **do**
  Define P from $\mathrm{W}^{(\ell)}$ via equation (4)
  $\mathrm{B} := \mathrm{P} \odot \mathcal{P}\left(\mathrm{H}^{(\ell)}\right) + (1 - \mathrm{P}) \odot \mathcal{H}\left(\mathrm{H}^{(\ell)}\right)$
  $\mathrm{H}^{(\ell+1)} := \mathrm{B} \odot \frac{{\mathrm{W}^{(\ell)}}^{\mathrm{T}} \frac{\mathrm{V}}{\mathrm{W}^{(\ell)}\mathrm{B}}}{{\mathrm{W}^{(\ell)}}^{\mathrm{T}}\mathrm{J}}$
  $\mathrm{W}^{(\ell+1)} := \mathrm{W}^{(\ell)} \odot \frac{\frac{\mathrm{V}}{\mathrm{W}^{(\ell)}\mathrm{B}}\mathrm{B}^{\mathrm{T}}}{\mathrm{J}\mathrm{B}^{\mathrm{T}}}$
  } NMF updates
end
Set $\mathrm{W} := \mathrm{W}^{(L)}$ and $\mathrm{H} := \mathrm{H}^{(L)}$
Binarize p according to threshold p_{thr}
Compute $\mathrm{V}_{\mathrm{p}}^{\mathrm{NMF}}$ and $\mathrm{V}_{\mathrm{h}}^{\mathrm{NMF}}$ via (1)
Output:
$\mathrm{A}_{\mathrm{p}}^{\mathrm{NMF}} := \mathrm{A} \odot \left(\Lambda \cdot \mathrm{V}_{\mathrm{p}}^{\mathrm{NMF}} \oslash \left(\mathrm{V}_{\mathrm{p}}^{\mathrm{NMF}} + \mathrm{V}_{\mathrm{h}}^{\mathrm{NMF}}\right)\right)$
$\mathrm{A}_{\mathrm{h}}^{\mathrm{NMF}} := \mathrm{A} \odot \left(\Lambda \cdot \mathrm{V}_{\mathrm{h}}^{\mathrm{NMF}} \oslash \left(\mathrm{V}_{\mathrm{p}}^{\mathrm{NMF}} + \mathrm{V}_{\mathrm{h}}^{\mathrm{NMF}}\right)\right)$

2.4. Unifying KAM and NMF

In Algorithm 2, we detail the unification of KAM-based HPSS with our soft-constrained NMF. In contrast to prior approaches, we start by stacking the KAM-based estimates of percussive and harmonic parts into a concatenated matrix $\mathrm{V} \in \mathbb{R}_{\geq 0}^{2K \times M}$, with $\mathrm{V} := \left[\mathrm{A}_{\mathrm{p}}^{\mathrm{KAM}}; \mathrm{A}_{\mathrm{h}}^{\mathrm{KAM}}\right]$. This matrix is then used as the target for NMF decomposition. Consequently, our NMF bases can be imagined as stacked templates of dimension $\mathrm{W} \in \mathbb{R}_{\geq 0}^{2K \times R}$. This core idea of our novel approach serves the following two purposes.

First, it enables the redistribution of TF magnitude that had been assigned to the wrong part by KAM. The rationale is that in our framework, a single NMF template can be interpreted as a coupling between two templates. The first (corresponding to the lower K frequency bins) can only "see" the percussive estimate, while the second (corresponding to the upper K frequency bins) can only model spectral patterns contained in the harmonic estimate. Since the coupled templates share one activation, they both can collect

Algorithm	Details
PRK: Constrained NMF [4]	$L^{\mathrm{NMF}} = 100$, $R = 750$, continuity parameters from [4]
KAM: KAM-based HPSS [6]	$L^{\mathrm{KAM}} = 30$, $\kappa = 9$, Hann-shaped kernel
NMF: Proposed Method	$L^{\mathrm{NMF}} = 60$, $R = 30$, $p_{\mathrm{thr}} = 0.25$, Median: $\tau = 4$, NEMA: $\lambda = 0.75$
ORC: Oracle [15]	Wiener Filtering using true source spectrograms

Table 1. Configuration of the test cases in our comparative performance evaluation.

spectral contributions according to the activation. In Figure 1c, this effect is illustrated by the fact that the drum templates have been assigned considerable contributions from the harmonic part while the sax templates have been assigned transient spectra (note the smeared harmonic structure) from the percussive estimate.

Second, our approach enables straightforward estimation of the percussive weight p from the relationship between the lower half (percussive) and the upper half (harmonic) of the NMF templates. Formally, $\mathrm{p}\,(r)$ is given as:

$$\mathrm{p}\,(r) := \frac{\sum_{k=0}^{K-1} \mathrm{W}(k,r)}{\sum_{k=0}^{2K-1} \mathrm{W}(k,r)}, \tag{4}$$

for each of the r^{th} NMF components. In Algorithm 2, replication of p over all columns yields the percussive weight matrix P. This matrix is then used to construct the weighted superposition B of the latest percussively and harmonically enhanced activations. The NMF updates use B instead of the regular $\mathrm{H}^{(\ell)}$. Since the percussive weight vector p depends on the templates W, it implicitly classifies the NMF components. In contrast to [4], this classification is not pre-defined — it is soft, and it adapts to the components as they evolve during the NMF iterations.

Finally, after the iteration limit L^{NMF} has been reached, we achieve the final classification of the components by binarizing p according to a pre-defined threshold $p_{\mathrm{thr}} \in [0, 1]$. This step is necessary to achieve a good separation between the refined NMF components. It remains to be seen whether more elaborate classification schemes would be beneficial.

A final Wiener filtering step then delivers the desired percussive part $\mathrm{A}_{\mathrm{p}}^{\mathrm{NMF}}$ and harmonic part $\mathrm{A}_{\mathrm{h}}^{\mathrm{NMF}}$. To make this work, we need to revert the earlier spectrogram stacking by multiplication with an aggregation matrix $\Lambda \in \mathbb{R}_{\geq 0}^{K \times 2K}$, constructed as $\Lambda := [\mathrm{I}, \mathrm{I}]$, with $\mathrm{I} \in \mathbb{R}_{\geq 0}^{K \times K}$ being the identity matrix.

3. EVALUATION

In this section, we present some HPSS experiments to compare our proposed approach to other state-of-the-art methods.To this end, we composed 74 music mixtures using two datasets, the ENST-Drums corpus [22] and the QUASI[1] corpus, both containing multi-track music recordings with ground-truth drum parts. We use an STFT blocksize of 2048 samples (approx. 46 ms) and a hopsize of 512 samples (75 % overlap). Table 1 summarizes the comparison algorithms and their parameters. The chosen settings for the kernel width κ and the NEMA decay λ are based on findings from our previous papers [6, 1]. Note that the outcomes of KAM serve as initial estimates of our proposed method. All algorithms deliver estimates for the percussive and harmonic magnitude spectrograms. We reconstruct the corresponding time-domain signals via inverse STFT using the mixture phase. We measure the separation quality using the median BSS Eval metrics SDR, SIR, and SAR [23].

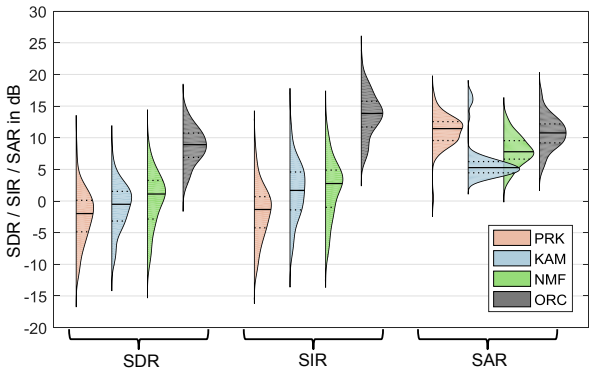

Fig. 3. BSS Eval comparison of the HPSS methods listed in Table 1. The test set comprises the corpora ENST-Drums and QUASI. The thick solid line in each distribution shows the median value over all items, and the dashed lines delimit the corresponding interquartile range.

As can be seen in Figure 3, our proposed method surpasses the comparison methods KAM [6] and PRK [4] in the majority of metrics. However, it still falls short of the oracle Wiener filtering (ORC) presented in Figure 3c.

4. DISCUSSION AND OUTLOOK

Since BSS Eval provides objective measures, we recommend listening to the audio examples on our accompanying website[2] to get a better impression of the capacities and limitations of our proposed method. The examples are taken both from real-world music recordings (including the excerpt in Figure 1), as well as from our test set. With KAM, one can hear that the decay and reverb of the drum part is often assigned to the harmonic component. The method by Park et al. [4] better preserves the drum characteristics but often has considerable leakage of the melodic instruments' attack into the drums. With our method, we are able to improve the quality of the drum signal considerably, while still achieving moderate separation. On the downside, the audio examples reveal that our method has difficulties to model quickly varying melodic signals, such as singing voice. Moreover, it is susceptible to distorted guitars, which produce spectra that look more broadband and noise-like than pure melodic tones.

Future work will be concerned with more thorough, data-driven parameter optimization. We plan to investigate using additional side-information (e. g., drum-specific templates), which can be easily integrated to guide the NMF updates [1, 24]. Also, it might be beneficial to tune the decay parameter λ depending on the underlying instrument (e.g., longer decay for cymbals and kick drums).

[1] http://www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

[2] https://www.audiolabs-erlangen.de/resources/MIR/2018-ICASSP-HPSS_KAM_NMF/

5. REFERENCES

[1] Christian Dittmar and Meinard Müller, "Reverse engineering the Amen break – score-informed separation and restoration applied to drum recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1531–1543, 2016.

[2] Patricio López-Serrano, Christian Dittmar, and Meinard Müller, "Finding drum breaks in digital music recordings," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Porto, Portugal, September 2017, pp. 68–79.

[3] Hideyuki Tachibana, Nobutaka Ono, and Shigeki Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, January 2014.

[4] Jeongsoo Park, Jaeyoung Shin, and Kyogu Lee, "Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1061–1074, May 2017.

[5] Derry FitzGerald, Antoine Liutkus, Zafar Rafii, Bryan Pardo, and Laurent Daudet, "Harmonic/percussive separation using Kernel Additive Modelling," in *Irish Signals and Systems Conference (IET)*, Limerick, Ireland, 2014, pp. 35–40.

[6] Christian Dittmar, Jonathan Driedger, Meinard Müller, and Jouni Paulus, "An experimental approach to generalized wiener filtering in music source separation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, 2016.

[7] Francisco J. Cañadas-Quesada, Pedro Vera-Candeas, Nicolás Ruiz-Reyes, Julio J. Carabias-Orti, and Pablo Cabañas Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 26, 2014.

[8] Jonathan Driedger, Thomas Prätzlich, and Meinard Müller, "Let It Bee – Towards NMF-inspired audio mosaicing," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 350–356.

[9] Jeongsoo Park and Kyogu Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 148–154.

[10] Francisco J. Cañadas-Quesada, Derry FitzGerald, Pedro Vera-Candeas, and Nicolás Ruiz-Reyes, "Harmonic-percussive sound separation using rhythmic information from non-negative matrix factorization in single-channel music recordings," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Edinburgh, UK, September 2017, pp. 276–282.

[11] Nobutaka Ono, Kenichi Miyamoto, Jonathan LeRoux, Hirokazu Kameoka, and Shigeki Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Lausanne, Switzerland, 2008, pp. 240–244.

[12] Derry FitzGerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010, pp. 246–253.

[13] Jonathan Driedger, Meinard Müller, and Sascha Disch, "Extending harmonic-percussive separation of audio signals," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014, pp. 611–616.

[14] Xavier Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds. Swets & Zeitlinger, 1997.

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

[16] Christian Uhle, Christian Dittmar, and Thomas Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," *Proceedings of the International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, pp. 843–847, 2003.

[17] Marko Helén and Tuomas Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2005.

[18] Minje Kim, Jiho Yoo, Kyeongok Kang, and Seungjin Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

[19] Derry FitzGerald, Eugene Coyle, and Matt Cranitch, "Using tensor factorisation models to separate drums from polyphonic music," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Camo, Italy, September 2009.

[20] Clément Laroche, Hélène Papadopoulos, Matthieu Kowalski, and Gaël Richard, "Drum extraction in single channel audio signals using multi-layer non negative matrix factor deconvolution," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017, pp. 46–50.

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

[22] Olivier Gillet and Gaël Richard, "Enst-drums: an extensive audio-visual database for drum signals processing," *Proceedings of the International Society for Music Information Retrieval Conference*, 2006.

[23] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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