An Experimental Approach to Generalized Wiener Filtering in Music Source Separation

Christian Dittmar, Jonathan Driedger, Meinard Müller
International Audio Laboratories Erlangen
Am Wolfsmantel 33
91058 Erlangen, Germany
Email: christian.dittmar@audiolabs-erlangen.de

Jouni Paulus
Fraunhofer Institute for Integrated Circuits
Am Wolfsmantel 33
91058 Erlangen, Germany
Email: jouni.paulus@iis.fraunhofer.de

***Abstract*—Music source separation aims at decomposing music recordings into their constituent component signals. Many existing techniques are based on separating a time-frequency representation of the mixture signal by applying suitable modeling techniques in conjunction with generalized Wiener filtering. Recently, the term α-Wiener filtering was coined together with a theoretic foundation for the long-practiced use of magnitude spectrogram estimates in Wiener filtering. So far, optimal values for the magnitude exponent α have been empirically found in oracle experiments regarding the additivity of spectral magnitudes. In the first part of this paper, we extend these previous studies by examining further factors that affect the choice of α. In the second part, we investigate the role of α in Kernel Additive Modeling applied to Harmonic-Percussive Separation. Our results indicate that the parameter α may be understood as a kind of selectivity parameter, which should be chosen in a signal-adaptive fashion.**

I. Introduction

Music signals can be understood as superposition (mixture) of different sound sources (components), such as melodic instruments, singing voice, bass, and drums. Music source separation aims at recovering these constituent component signals from the mixture. The separated sources may be used for music retrieval tasks, automatic music transcription, as well as music production and restoration, see [1] for an overview. At the core, many source separation techniques try to extract the target component signal from the mixture by means of time-variant filtering. In practice, this filtering procedure is commonly realized by element-wise weighting of the mixture's short-time Fourier transform (STFT) with some kind of time-frequency (TF) mask. Besides the wide-spread use of *binary masks* [2], many approaches use so-called *soft masks*. The most common strategy to construct soft masks is to use generalized Wiener filtering [3]–[5]. Loosely speaking, this procedure consists of first estimating the spectrogram of the target source and subsequently taking its ratio to the sum of all source spectrogram estimates as the filter weight. In order to disambiguate the usage of the term spectrogram throughout the literature, we use the notion of an α-spectrogram as introduced in [5], meaning the modulus of the STFT raised to some arbitrary exponent $\alpha \in [0, 2]$. With this clarification, Wiener filtering relies on the rather strong assumption that the sources' α-spectrograms add up to the mixture's α-spectrogram. This completely neglects possible phase-related issues, such as destructive interference [6]. While some research effort has been dedicated to incorporating phase information [6]–[8], other authors have attempted to find more appropriate masking strategies. In [4], an alternative family of TF masks based on well-known divergence measures such as the Kullback-Leibler and Itakura-Saito was proposed. In a similar fashion, [9] tried to find an optimal magnitude exponent (among other parameters) in diverse source separation tasks. *Oracle* source separation experiments with known component signals were conducted in [10] in order to identify a domain satisfying the additivity assumption of spectral magnitudes. Similar settings were used in [5], where the authors also established a theoretic foundation for using magnitude instead of power spectra (i.e., $\alpha = 1$ instead of $\alpha = 2$) in Wiener filtering.

From the literature, we see that the magnitude exponent α is considered to be an important parameter, which is not fully understood yet. In this paper, we take a different perspective and investigate two aspects of α in an experimental fashion. In Section II, we extend the oracle experiments from [5] in order to assess the dependency of the additivity assumption of α-spectrograms on the signal type as well as other influential factors. In Section III, we asses the influence of α in Kernel Additive Modeling (KAM), a recently proposed [11] source separation procedure that strongly relies on iteratively applying Wiener filtering. As we will show, α can be understood as a selectivity parameter to trade off between interference reduction and artifacts depending on the target signal type.

II. Additivity of α-spectrograms revisited

In this section, we present the settings and results of oracle experiments on the additivity assumption of α-spectrograms. As in related studies [5], [10], we work with known source signals in order to create a controlled scenario that is independent of specific source separation methods. Our initial question is if it is beneficial to choose α differently depending on whether we want to separate a saxophone solo from accompanying instruments or if we want to separate drum instruments from a drum solo recording.

A. Notation and Signal Model

Let $x : \mathbb{Z} \to \mathbb{R}$ be the real-valued, discrete-time domain mixture signal that is based on the linear superposition $x :=$

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

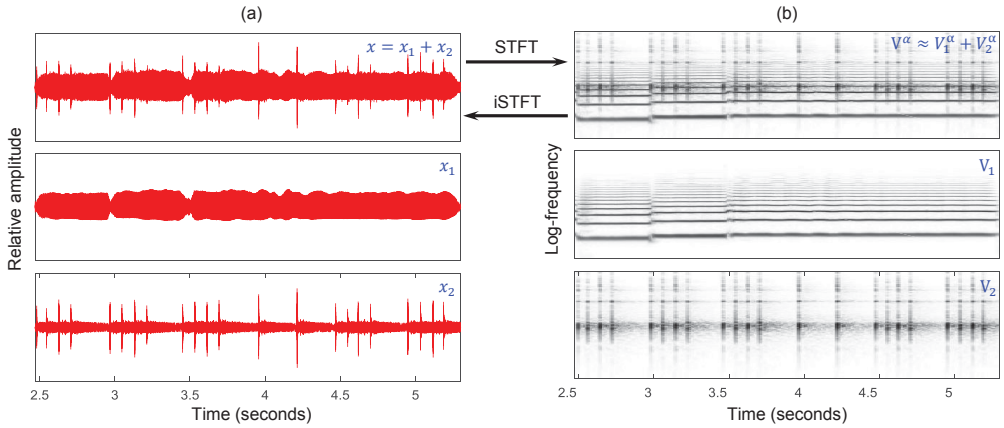

Fig. 1. Illustration of our signal model. **(a):** Mixture signal $x := x_1 + x_2$, which is the superposition of two source signals x_1 and x_2. **(b):** Magnitude spectrogram of the mixture V and the sources V_1 and V_2. For better visibility, we use a logarithmic frequency axis and logarithmic magnitude.

$x_1 + x_2$ of two component signals corresponding to the individual sources. Example component signals are illustrated in Figure 1(a), where x_1 is a harmonic melody instrument and x_2 is a percussive accompaniment. We will return to these specific music signal properties in Section III. As already discussed, we transition to the TF domain as depicted Figure 1(b). To this end, let $\mathrm{V}\,(k, m)$ be the non-negative modulus of the STFT at the k^{th} spectral bin and the m^{th} time frame. As shown in the top plot of Figure 1(b), we assume that the α-spectrograms V_1 and V_2 approximately add up to the mixture:

$$\mathrm{V}^\alpha \approx \mathrm{V}_1^\alpha + \mathrm{V}_2^\alpha. \tag{1}$$

Here, the magnitude exponent $\alpha \in [0, 2]$ is applied in an element-wise fashion. Although our notation easily extends to the more general scenario involving an arbitrary number $C \in \mathbb{N}$ of sources, we will restrict ourselves to the case $C = 2$ in the following discussion for the sake of clarity.

B. Signal-Dependency of α

First, we repeat a similar experiment as in [5], using multitrack recordings to quantify the additivity of α-spectrograms under varying α. The basic protocol is to create linear mixtures from oracle source signals and then switch to the TF domain to assess the additivity assumption. With respect to our running example (see Figure 1), this can be formalized as computing a suitable divergence D between the mixture's α-spectrogram and the sum of the sources' α-spectrograms as

$$D\left(\mathrm{V}^\alpha, \mathrm{V}_1^\alpha + \mathrm{V}_2^\alpha\right) \text{ for } \alpha \in [0.2, 2]\,. \tag{2}$$

As in [5], the metric D can be either the α-dispersion, Itakura-Saito, or Kullback-Leibler divergence. In our experiments, we use source signals from the "QUASI"[1] data set. This set consists of several full-length songs from diverse music genres, each providing single track recordings of the involved sources, such as singing voice, melodic instruments, bass, drums, or percussion. Thus, the set covers a broad range of different signals characteristics, in the sense that it contains harmonic as well as percussive instruments with varying degree of interdependence between them.

In particular, we are interested if the results reported in [5] generalize to other, more homogeneous types of music recordings. Thus, we extend the experiment with two additional data sets. The first consists of all single tracks of the "Bach10"[2] and "TRIOS"[3] corpora, which are dominated by harmonic instruments, such as violin, viola, bassoon, horn, clarinet, and piano. It should be noted that one piece in the TRIOS corpus contains three drum tracks, these have been excluded for our experiment. The second set uses drum only recordings from the "IDMT-SMT-Drums" data set[4], where the sources correspond to the three drum instruments kick, snare, and hi-hat [12]. Across all sets, the audio items are in uncompressed PCM WAV format with 44.1 kHz sampling rate, 16 Bit, mono. As for the STFT parameters, we adopt the settings from [5], using Hamming-windowed frames of approx. 80 ms duration and 80% overlap between them.

As is shown in Figure 2(b) we can not exactly replicate the curves reported in [5] (dashed lines) but the tendencies are similar. However, we can indeed see a different behavior of the curves in Figure 2(a) and (c). Besides different minimum positions, it is remarkable that the curves in (c) are much flatter. As a tendency, one might say that for drums, the range of possible quasi-optimal α is much broader than with harmonic instruments. As we will show in the next section, these results should be read with great care, since there are more factors involved than just the signal types.

C. Level-Dependency of α

In this second experiment, we basically repeat the same protocol as before. This time, the only difference is that each

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

[2] http://www.ece.rochester.edu/~zduan/resource/Resources.html

[3] http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27

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

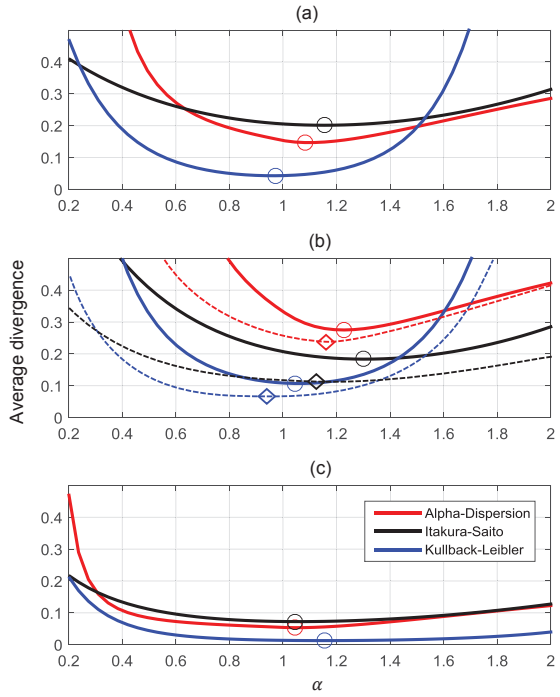

Fig. 2. Average α-dispersion, Itakura-Saito and Kullback-Leibler divergences as a function of α. Global minimum positions are are marked with a circle. The legend in **(c)** applies to all plots. **(a):** Results obtained with purely harmonic sources (Bach10 & TRIOS data sets). **(b):** Results obtained with harmonic and percussive sources (QUASI data set). The dashed lines provide the original results from [5] for comparison, diamond markers represent the respective minimum positions. **(c):** Results obtained with purely percussive sources (IDMT-SMT-DRUMS dataset).

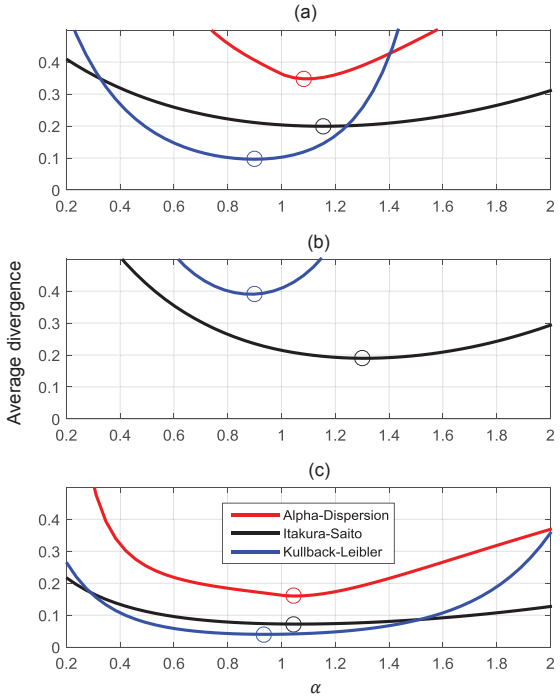

Fig. 3. Three different divergences are shown as a function of α, the same description applies as in Figure 2. **(a):** Normalized Bach10 & TRIOS sources. **(b):** Normalized QUASI sources. The curve for α-dispersion is outside the plot range. **(c):** Results from normalized IDMT-SMT-DRUMS sources.

source signal is normalized so that its absolute maximum value is 1.0 before adding them up to the linear mixture. Since the normalization factor depends on the properties of the respective signal, the normalization step is expressed by introducing modified sources signals $\overline{x}_1$ and $\overline{x}_2$ in

$$x := \overline{x}_1 + \overline{x}_2. \tag{3}$$

As can be seen in Figure 3, this simple modification affects the results quite a lot. As expected, the Itakura-Saito divergences (black curves) stay the same for all data sets, since they are less susceptible to level differences. The α-dispersion (red curves) look like scaled versions of the ones in Figure 2, a direct consequence of its calculation rule given in [5]. For the QUASI data set, the scaling is so pronounced that the curve is out the plot range in Figure 3(b). However, at least the minimizing α-values stay the same. In contrast, the Kullback-Leibler divergences (blue curves) look completely different and the minimizing α-values end up in different positions, so the choice of an optimal α for a certain signal type becomes questionable. From the empirical results in [5], one could get the impression that an $\alpha \approx 1.2$ is a sensible choice for general purpose music source separation (see Figure 2(b)). Our results rather indicate that the choice of α is sensitive to a number of additional factors. This is in line with the findings in [10], where also the number of sources C has been shown to be influential. Other factors, such as the mutual correlation between the sources remain completely nebulous and should be addressed in further studies.

III. Influence of α in Kernel Additive Modeling

After these oracle-based experiments, we now want to study the influence of α in a concrete source separation scenario. In particular, we consider the task of "Harmonic-Percussive Separation" (HPS) as a specific case study. HPS aims at splitting a music recording into harmonic (e.g., melodic instruments, tonal components) and percussive (e.g., drums and percussion, transient components) sources, see Figure 1 for an example. A high quality HPS is an important pre-requisite for advanced sources separation tasks such drum transcription and separation [12], [13].

A comprehensive overview of recent methods for HPS is given in [14]. Many works already discussed the problem that music recordings may consist of sounds that are neither clearly harmonic nor percussive [15], [16]. An example are harmonic tones, whose fundamental frequency is modulated over time as is typical in recordings of instruments with vibrato. We neglect this problem in our study for the sake of compactness.

A. KAM-Based HPS

Recently, a novel class of source separation approaches was proposed under the notion of Kernel Additive Model-

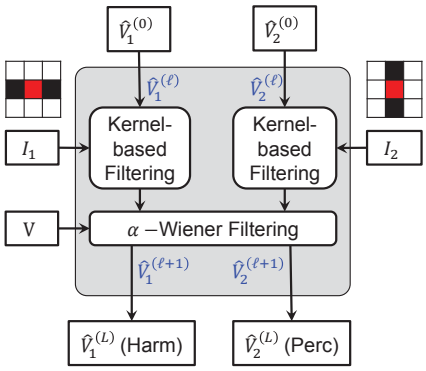

Fig. 4. Overview of the KAM-based algorithm for HPS. The gray box stands for iterative refinement.

ing (KAM) [11], [17]. In contrast to global decomposition paradigms (e.g., Non-negative Matrix Factorization), KAM exploits local regularities of the source spectrograms. An HPS variant based on KAM was introduced in [18], which is also the main technique used in our study. The method relies on iteratively modeling the component spectrograms and applying Wiener filtering as depicted in Figure 4. With respect to our two-component signal model, the α-Wiener mask for the two components is computed as

$$\mathrm{M}_c := \hat{\mathrm{V}}_c^{\alpha} \oslash \left(\hat{\mathrm{V}}_1^{\alpha} + \hat{\mathrm{V}}_2^{\alpha}\right) \text{ for } c \in \{1, 2\}, \tag{4}$$

where $\oslash$ denotes element-wise division, $\hat{\mathrm{V}}_1$ represents the harmonic and $\hat{\mathrm{V}}_2$ the percussive component estimate.
In our reimplementation of KAM-based HPS, we set the initial estimate $\hat{\mathrm{V}}_1^{(0)}$ and $\hat{\mathrm{V}}_2^{(0)}$ to the mixture α-spectrogram V^{α}. As in [18], we construct two kernels $\mathcal{I}_1$ and $\mathcal{I}_2$ for the enhancement of harmonic and percussive structures. As illustrated by Figure 4, the harmonic kernel is all zero accept one horizontal row and the percussive kernel shows a perpendicular structure. We introduce the iteration index $\ell = 0, 1, 2, \ldots, L \in \mathbb{N}$ and proceed with iterative refinements by first applying a kernel-based filtering to each of the components and subsequently applying (4). It should be noted that the original procedure in [18] applies 2D median filters in the kernel-based filtering stage. In contrast, we use 2D convolution with the kernels $\mathcal{I}_1$ and $\mathcal{I}_2$. The rationale behind replacing median filtering by convolution is that we want to eliminate any other nonlinear operations besides raising the STFT modulus to the magnitude exponent α in (4).

B. Evaluation in HPS Task

To investigate the influence of α-Wiener filtering in KAM-based HPS, we generate one test item by superimposing a real-world trumpet melody (harmonic) with castanets (percussive). The experimental settings are the same as in [18], using an STFT blocksize of 4096 samples (approx. 92 ms) and a hopsize of 1024 samples (75 overlap). Kernels of 17×17 elements are used for KAM, and the number of iterations is set to $L = 10$. At each iteration, the iSTFT for the harmonic $\hat{\mathrm{V}}_1^{(\ell)}$ and the percussive component estimate $\hat{\mathrm{V}}_2^{(\ell)}$ is computed using the mixture's phase spectrogram to yield the time-domain reconstructions $\hat{x}_1$ and $\hat{x}_2$ respectively. In accordance to the standards used in the literature on music source separation, we employ the Perceptual Evaluation Methods for Audio Source Separation (PEASS) [19], [20] in order to evaluate the quality of the reconstructed component signals. In contrast to the experiments in the original paper, we vary the α-parameter in (4) in order to assess its influence on this procedure.

In Figure 5, we show the evolution of four perceptually motivated PEASS metrics with increasing iteration count ℓ for the harmonic component in (a) and the percussive component in (b). The metrics comprise the artifact-related (APS), interference-related (IPS), target-related (TPS) and overall perceptual score (OPS), which can attain a maximal score of 100 in case of a perfect separation. In Figure 5(a), it can be seen that the OPS benefits from higher α but quickly saturates already after a few iterations. As expected, the artifacts-related APS drops with increasing ℓ, while the interference-related IPS improves. In Figure 5(b), the percussive component attains much lower APS even for $\alpha = 1$, probably due to pre-echos that occur when using large STFT blocksizes in conjunction with the mixture phase for reconstruction of transient signal components [21]. Informal listening tests confirmed that pre-echos are indeed an issue. However, it is interesting that the OPS and interference-related IPS can go much higher than for the harmonic component, and seem to be less susceptible to changing α. This indicates that it might be beneficial to use $\alpha \approx 1$ if we are interested in extracting the percussive component and $\alpha \approx 2$ for the harmonic component. We also want to stress how easily the PEASS scores can be increased or decreased by just changing α. In competitive evaluation campaigns such as SiSec[5], often a few score points can decide over the ranking of submitted source separation algorithms. The susceptibility to such basic parameters as the magnitude exponent α sheds a new light on the interpretation of these competition results.

IV. Conclusion

In this paper we reported results of exploratory experiments on the influence of the magnitude exponent α with respect to the additivity assumption as well as Wiener filtering in KAM-based HPS. Empirically, we show that the choice of α is sensible to several factors, such as the signal types, the relative mixing levels and the number of sources. In KAM-based HPS, where α-Wiener Filtering is applied iteratively, there is a delicate trade-off between softer separation with $\alpha = 1$ compared to stronger selectivity and faster convergence at the cost of undesired artifacts for $\alpha = 2$. Future work will be directed towards developing strategies for α-Wiener filtering that are adaptive to signal types, temporal evolution or even local spectral characteristics of the target sources.

[5]https://sisec.inria.fr/

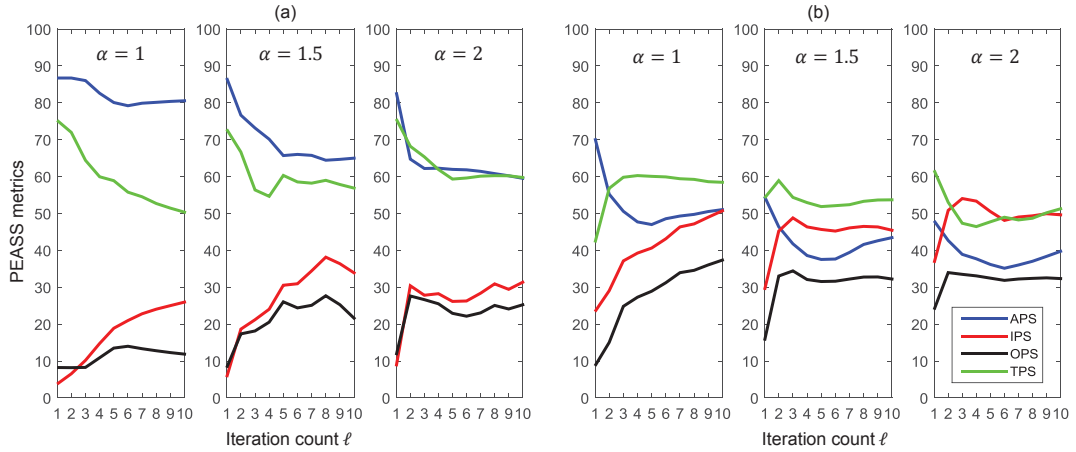

Fig. 5. Evolution of the PEASS measures plotted along the iteration count ℓ. **(a):** Results for the harmonic component signal. **(b):** Results for the percussive component signal.

ACKNOWLEDGMENT

The authors would like to thank Antoine Liutkus for fruitful discussions and support in validating the experimental results. This work has been supported by the German Research Foundation (DFG MU 2686/6-1). The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

REFERENCES

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

[2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Norwell, Massachusetts, USA: Kluwer Academic, 2005, pp. 181–197.

[3] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 613–616.

[4] D. FitzGerald and R. Jaiswahl, "On the use of masking filters in sound source separation," in *Proc. of the Intl. Conf. on Digital Audio Effects (DAFx)*, York, UK, September 2012.

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

[6] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[7] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3437–3440.

[8] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, "Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency," in *Proc. of the Intl. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)(Lecture Notes in Computer Science, Vol. 6365)*, St. Malo, France, 2010, pp. 89–96.

[9] P. S. Brian King, Cédric Fevotte, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *Proc. of the IEEE Intl. Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, September 2015.

[10] S. Voran, "Exploration of the additivity approximation for spectral magnitudes," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2015.

[11] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

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

[13] C. Dittmar and M. Müller, "Reverse Engineering the Amen Break – Score-informed Separation and Restoration applied to Drum Recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[14] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2059–2072, December 2014.

[15] J. Park and K. Lee, "Harmonic-percussive source separation using harmonicity and sparsity constraints," in *Proc. of the Intl. Society for Music Information Retrieval Conf. (ISMIR)*, Málaga, Spain, 2015, pp. 148–154.

[16] R. Füg, A. Niedermeier, J. Driedger, S. Disch, and M. Müller, "Harmonic-percussive-residual sound separation using the structure tensor on spectrograms," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016, to appear.

[17] T. Prätzlich, R. Bittner, A. Liutkus, and M. Müller, "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[18] D. FitzGerald, A. Liutkus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *Irish Signals and Systems Conf. (IET)*, Limerick, Ireland, 2014, pp. 35–40.

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

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

[21] C. Dittmar and M. Müller, "Towards transient restoration in score-informed audio decomposition," in *Proc. of the Intl. Conf. on Digital Audio Effects (DAFx)*, Trondheim, Norway, December 2015, pp. 145–152.