Harmonic/Percussive Separation Using Kernel Additive Modelling

Derry FitzGerald[1], **Antoine Liukus**[2], **Zafar Rafii**[3], **Bryan Pardo**[3] **and Laurent Daudet**[4]

[1]*Nimbus Centre*
Cork Institute of Technology

[3]*Northwestern University,*
Evanston, IL, USA

[4]*Institut Langevin,*
Paris Diderot Univ., France

[2]*Inria, Villers-lès-Nancy, France,*
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, France

E-mail: [1]derry.fitzgerald@cit.ie

Abstract — **Recently, Kernel Additive Modelling was proposed as a new framework for performing sound source separation. Kernel Additive Modelling assumes that a source at some location can be estimated using its values at nearby locations where nearness is defined through a source-specific proximity kernel. Different proximity kernels can be used for different sources, which are then separated using an iterative kernel backfitting algorithm. These kernels can efficiently account for features such as continuity, stability in time or frequency and self-similarity. Here, we show that Kernel Additive Modelling can be used to generalise, extend and improve on a widely-used harmonic/percussive separation algorithm which attempts to separate pitched from percussive instruments.**

Keywords — **Sound Source Separation, Kernel Additive Modelling, Harmonic/Percussive Separation**

I Introduction

Harmonic/Percussive (H/P) separation of mixed audio signals deals with attempting to separate pitched (harmonic) instruments from percussion instruments, which can be loosely modelled as different types of broadband noise. H/P separation has numerous applications including remixing and DJing, as well as a preprocessing tool for other tasks including automatic music transcription, chord estimation, and key signature detection. In these applications, the elimination of the percussive sources allows improved estimation of the pitched content, while elimination of the pitched sources allows improved results in applications such as rhythm analysis, beat tracking, and the automatic transcription of drum instruments.

In recent years, different techniques have been proposed for the purpose of H/P separation. Among these, a number of algorithms are based on the intuition that harmonic instruments form stable horizontal ridges across time in spectrograms, while percussive instruments form stable vertical ridges across frequency due to their broadband noise-based nature. This is illustrated in Figure 1, where the harmonics are clearly visible as horizontal lines, while the drums are visible as vertical lines in the spectrogram. Algorithms exploiting this include those based on anisotropic diffusion [1], as well as others based on Bayesian models of this assumption [2]. Of particular interest in our context is the algorithm proposed in [3], based on median filtering of spectrograms. In the present paper, we extend the algorithm presented in [3] in order to refine separation through several iterations and to account for multichannel mixtures. The proposed method fits in the recently proposed Kernel Additive Modelling (KAM, [4, 5]) framework for source separation and is consistently shown to improve over the original method [3].

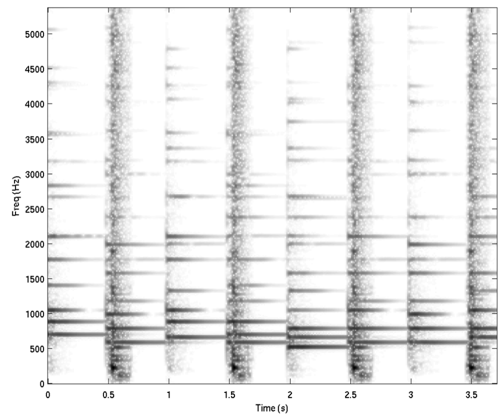

Fig. 1: Spectrogram of Pitched and Percussive Mixture.

The paper is structured as follows. In a background section II, we recall the original H/P separation procedure [3] and briefly present the recently KAM framework. In section III, we detail the proposed harmonic/percussive separation algorithm and finally evaluate it in section IV.

II Background

a) H/P Separation with median filtering

The algorithm proposed in [3] focuses on single channel mixtures. It assumes that vertical lines in a spectrogram correspond to percussion events, while horizontal lines are typically associated with the harmonics of pitched instruments. In this case, peaks due to pitched harmonics can be regarded as outliers on the vertical lines associated with percussion events. Similarly, peaks due to the percussion events can be regarded as outliers on the horizontal lines associated with pitched harmonic instruments. The algorithm goes as follows.

Let boldfaced $\boldsymbol{x}$ denote the power spectrogram of a monochannel mixture. It is a $N_f \times N_t$ matrix, where N_f and N_t respectively stand for the number of frequency bands and the number of frames. We define $\boldsymbol{x}(f,\cdot)$ as the f^{th} frequency slice containing the values of the f^{th} frequency bin across time. Similarly, we define $\boldsymbol{x}(\cdot,t)$ as the t^{th} time frame. As median filters are good at eliminating outliers, then median filtering each time frame will suppress harmonics in this frame resulting in a percussion enhanced frame $\boldsymbol{s}_P(\cdot,t)$, while median filtering each frequency slice will suppress percussion events in this slice, yielding a harmonic-enhanced slice $\boldsymbol{s}_H(f,\cdot)$:

$$\boldsymbol{s}_P(\cdot,t) = \mathsf{M}\{\boldsymbol{x}(\cdot,t), l_{perc}\} \quad (1)$$

$$\boldsymbol{s}_H(f,\cdot) = \mathsf{M}\{\boldsymbol{x}(f,\cdot), l_{harm}\} \quad (2)$$

where M denotes median filtering, and where l_{perc} and l_{harm} are the median filter lengths used to generate the percussion enhanced frames and harmonic enhanced slices respectively. $\boldsymbol{s}_P$ and $\boldsymbol{s}_H$ are then used to generate Wiener-filter type masks to apply to the original complex valued spectrograms before inversion to the time domain.

b) Kernel Additive Modelling

Kernel Additive Modelling (KAM) is a recently proposed framework for performing source separation [4, 5]. In contrast to well-established paradigms for separation, such as Non-negative Tensor Factorisations (NTF) [6], which perform a global decomposition based on superposition of fixed patterns or basis functions of the underlying sources, KAM focuses on the underlying regularities of the sources to separate them from mixtures. In the context of audio source separation, the human ability to discriminate between sources in a mixture has been shown to depend on local features such as repetitivity, common fate and continuity [7]. These dynamic features can be seen as depending on local regularities concerning the evolution of sources over time, frequency and space, rather than on fixed global patterns or basis functions corresponding to NTF-based approaches.

To model these reqularities within the spectrograms of the sources, KAM uses *kernel local parametric models* which have their roots in *local regression* [8]. In the audio case, it is assumed that the value $\boldsymbol{s}_j(f,t)$ of the spectrogram of a source j at a given TF point (f,t) is close to its values as other time-frequency bins given by a source-specific *proximity kernel* $\mathcal{I}_j(f,t)$ [5]:

$$\forall (f',t') \in \mathcal{I}_j(f,t), \boldsymbol{s}_j(f,t) \approx \boldsymbol{s}_j(f',t'), \quad (3)$$

where $\mathcal{I}_j(f,t)$ is a set containing the *nearest neighbours* of (f,t) from the perspective of source j. These kernels can be built using a variety of manners including the use of suitable feature spaces.

Different sources can then be modelled with different proximity kernels $\mathcal{I}_j$, and the KAM framework offers a large degree of flexibility in the incorporation of prior knowledge about the local dynamics of the sources to be separated. Separation of additive sources is then achieved through a variant on the *backfitting* algorithm [9]. Many popular audio separation algorithms can be viewed as special cases of the KAM framework including Adress [10], DUET [12], REPET [11] and the harmonic/percussive separation algorithm described in Section a). Furthermore, KAM provides an effective way to generate new source separation algorithms for sources which can be characterised by local features. A more detailed explanation of the framework can be found in [4, 5].

III Model and Method

a) Notation and Model

The mixture audio signal $\tilde{x}$ is a set of I time series, with $\tilde{x}(n,i)$ denoting the value of the i^{th} channel of the mixture at sample n. In most popular music $I=2$ for stereo signals, or 1 for mono recordings. The mixture is then assumed to be the sum of J sources $\tilde{s}_j$:

$$\tilde{x}(n,i)=\sum_{j=1}^{J}\tilde{s}_j(n,i) \tag{4}$$

We then define x and $\{s_j\}_{j=1\cdots J}$ as the Short-Time Fourier Transforms (STFT) of the mixture and the J sources respectively. These are all tensors of size $N_f \times N_t \times I$. $s_j(f,t)$ is then an $I \times 1$ vector containing the values of the STFTs of source s_j at time-frequency bin (f,t).

Assuming a Local Gaussian Model [13] each $s_j(f,t)$ are taken as independent vectors, each distributed according to a multivariate centered complex Gaussian distribution:

$$\forall(f,t),\ s_j(f,t) \backsim N_c(0,\boldsymbol{s}_j(f,t)R_j(f)) \tag{5}$$

where boldfaced $\boldsymbol{s}_j \geq 0$ is a nonnegative scalar that indicates the energy of source j at time-frequency bin (f,t). It is called the *spectrogram* of source j in the remaining of this paper[1]. R_j is a complex $I \times I$ positive semidefinite matrix, which is called the spatial covariance matrix of source j at frequency band f. It encodes the covariance between the different channels of s_j at frequency band f. Since the mixture $x(f,t)$ is the sum of J independent random Gaussian vectors, it is distributed as:

$$\forall(f,t),\ x(f,t) \backsim N_c\left(0,\sum_{j=1}^{J}\boldsymbol{s}_j(f,t)R_j(f)\right) \tag{6}$$

If estimates of $\boldsymbol{s}_j$ and R_j are available (termed $\hat{\boldsymbol{s}}_j$ and $\hat{R}_j$ respectively), then the Minimum Mean-Squared Error estimates $\hat{s}_j$ of the STFTs of the sources are obtained from:

$$\hat{s}_j(f,t)=\hat{\boldsymbol{s}}_j(f,t)\hat{R}_j\left[\sum_{j=1}^{J}\hat{\boldsymbol{s}}_j(f,t)\hat{R}_j(f)\right]^{-1}x(f,t) \tag{7}$$

The source time-domain signals can then be obtained via inverse STFT.

Prior knowledge about the sources to be separated is then encoded in terms of a proximity kernel $\mathcal{I}_j(f,t)$ for each source, which indicates which time-frequency points have values close to that of $\boldsymbol{s}_j$, as in (3). If we assume that $\boldsymbol{s}_j$ is not observed directly but only through possibly very noisy estimates $\boldsymbol{z}_j$ as is the case in practice, then $\boldsymbol{s}_j$ is estimated as:

Algorithm 1 Kernel backfitting for multichannel audio source separation with locally constant spectrogram models and binary proximity kernels.

```
1. Input:
     • Mixture STFT $x(f,t)$
     • Neighbourhoods $\mathcal{I}_j(f,t)$.
     • Number $N$ of iterations
2. Initialisation
     • $n \leftarrow 1$
     • $\forall j, \hat{\boldsymbol{s}}_j(f,t) \leftarrow x(f,t)^{\star} x(f,t)/IJ$
     • $R_j(f) \leftarrow I \times I$ identity matrix
3. Compute estimates $\hat{s}_j$ of all sources using (7)
4. For each source $j$:
     (a) $C_j(f,t) \leftarrow \hat{s}_j(f,t)\hat{s}_j(f,t)^{\star}$
     (b) $\hat{R}_j(f) \leftarrow \frac{I}{N_t}\sum_t \frac{C_j(f,t)}{\mathrm{tr}(C_j(f,t))}$
     (c) $\boldsymbol{z}_j(f,t) \leftarrow \frac{1}{I}\sum_t \mathrm{tr}\left(\hat{R}_j(f)^{-1}C_j(f,t)\right)$
     (d) $\hat{\boldsymbol{s}}_j(f,t) \leftarrow$
         median $\{\boldsymbol{z}_j(f',t') \mid (f',t') \in \mathcal{I}_j(f,t)\}$
5. If $n < N$ then set $n \leftarrow n+1$ and go to step 3
6. Output:
   sources spectrograms $\hat{\boldsymbol{s}}_j$ and spatial covariance matrices $\hat{R}_j(f)$ to use for filtering (7).
```

$$\hat{\boldsymbol{s}}_j(f,t)=\underset{\boldsymbol{s}_j(f,t)}{\operatorname{argmin}}\sum_{(f',t')\in\mathcal{I}_j(f,t)}\mathcal{L}_j(\boldsymbol{z}_j(f,t)\mid\boldsymbol{s}_j(f,t)) \tag{8}$$

where $\mathcal{L}_j(\boldsymbol{z}_j|u)$ is the *model cost function* for source j as defined in [4]. In our context, it is the cost of choosing $\boldsymbol{s}_j(f,t)=u$ when its noisy observation is $\boldsymbol{z}_j$. In this case we choose $\mathcal{L}_j$ to be the absolute deviation as the observations are likely to be contaminated by outliers during the iterative backfitting process described below, and the absolute deviation is known to yield estimates that are robust to the presence of outliers. We have:

$$\hat{\boldsymbol{s}}_j(f,t)= \underset{\boldsymbol{s}_j(f,t)}{\operatorname{argmin}}\sum_{(f',t')\in\mathcal{I}_j(f,t)}|\boldsymbol{z}_j(f,t)-\boldsymbol{s}_j(f,t)| \tag{9}$$

[1] $s_j(f,t)$ and $x(f,t)$ are hence complex $I \times 1$ vectors, while boldfaced $\boldsymbol{s}_j(f,t)$ is a nonnegative scalar. $R_j(f)$ is a complex $I \times I$ matrix. All estimates are denoted $\hat{\cdot}$

This cost function is minimised by:

$$\hat{\boldsymbol{s}}_j(f,t) = \mathrm{median}(\boldsymbol{z}_j(f,t)|(f',t') \in \mathcal{I}_j(f,t)) \qquad (10)$$

The kernel backfitting algorithm proposed in [4] for estimation of the source spectrograms $\hat{\boldsymbol{s}}_j$ proceeds in an iterative manner, with separation and estimation of the parameters performed alternatively. Here, the spectrograms $\boldsymbol{z}_j(f,t)$ of the current estimates of the source STFTs $\hat{s}_j$ are used as noisy observations of the true value with re-estimation of $\hat{\boldsymbol{s}}_j$ achieved through median filtering. The kernel backfitting procedure is summarised in algorithm 1, where $\cdot^{\star}$ denotes conjugate transpose and $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix. Details on the re-estimation procedure for the spatial covariance matrices can be found in [13].

b) Harmonic/Percussive Separation using KAM

The harmonic/percussive separation algorithm described in section II can be viewed as an instance of KAM. To describe this algorithm within the KAM framework we define the proximity kernel $\mathcal{I}_j(f,t)$ for the percussive source as:

$$\mathcal{I}_P\,(f,t) = \{(f+p,t) \mid p = -l,\ldots,l\} \qquad (11)$$

where $2l+1 = l_{perc}$, as defined in equation (1), is the total number of frequency bins in the neighbourhood. We similarly define the proximity kernel for the harmonic source as:

$$\mathcal{I}_H\,(f,t) = \{(f,t+p) \mid p = -k,\ldots,k\} \qquad (12)$$

where $2k+1 = l_{harm}$, as defined in equation (2), is the number of time frames in the neighbourhood.

It can clearly be seen that defining the proximity kernels in such a manner results in median filtering being applied to the same sets of time-frequency points as those in equations 1 and 2, and so the algorithm described in section II can be seen as a single iteration of KAM where the spatial covariance matrices $\hat{R}_j\,(f)$ are fixed to the identity matrix which in effect forces each of the mixture channels to be modelled independently. It can be seen that KAM generalises on this algorithm by allowing iterative updating of the sources, and by the inclusion of a spatial model for the sources. It is proposed to investigate if the iterative updating of sources and the inclusion of the spatial model results in improved harmonic/percussive separation.

IV Evaluation

In order to test the effectiveness of the proposed KAM algorithm for H/P separation, a database of 10 test signals was created using excerpts from multitrack recordings where the percussion and all other instruments were available as individual recordings. Percussion-only mixes and Pitched instrument mixes (including vocals) were prepared, and used to create an overall mix for each track. These were then separated using the proposed H/P KAM algorithm. These recordings had a sample-rate of 44.1kHz.

To allow direct comparison to the algorithm in [3] an FFT size of 4096 samples, a hopsize of 1024 samples, $l_{perc} = l_{harm} = 17$ were used, ensuring that the parameters used were the same as the original algorithm. To investigate the effects of the iterative backfitting procedure, the separation performance was evaluated after each iteration, with the total number of iterations set to 10. This number was used as previous work on KAM indicated that the backfitting procedure requires only a small number of iterations to converge.

Further, to identify how much improvement in performance was due to the backfitting procedure and how much was due to the spatial model, evaluation was also carried out on a version of the algorithm where $\hat{R}_j\,(f)$ is fixed to the identity matrix for all iterations. In this case, the results for iteration 1 are equivalent to those obtained from the algorithm proposed in [3], and so this represents a baseline from which improvements in performance using KAM can be measured against. Audio examples and MATLAB code for the algorithm can be found at the paper's webpage[2]

Separation performance was measured using metrics from the PEASS toolkit [14]. The metrics considered are the Overall Perceptual Score (OPS), which attempts to provide an overall indication of the separation quality, the Target-related Perceptual Score, which attempts to measure how the spatial position of the separated source corresponds to that of the original source, the Interference-related Perceptual Score (IPS), which attempts to measure the amount of interference from other sources present in the separated source, and the Artifact-related Perceptual Score, which measures the presence of artifacts in the separated source. All these metrics have ranges from 0-100 with higher scores being better.

These metrics are plotted in Figures 2-5, with red indicating the use of the spatial model and black indicating the spatial model was not used, with $\hat{R}_j\,(f)$ fixed to unity throughout all iterations. Figure 2 shows the average results obtained from all 10 excerpts for the OPS for percussive separation (left) and harmonic separation (right) plotted against iteration number.

It can be seen that initially the use of the iterated approach in KAM results in significant improvements in OPS over the baseline algorithm, which consisted of a single iteration. However, at higher iteration numbers performance falls off. This is particularly evident for the harmonic sources, where 2 iterations gives a large increase

[2] `www.loria.fr/~aliutkus/kamhp/`

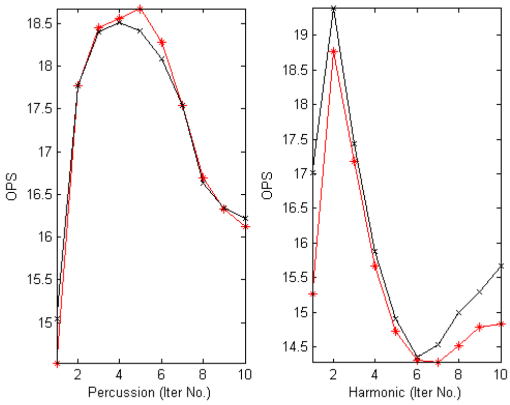

Fig. 2: Overall Perceptual Score for Percussive (left) and Harmonic (right) separations. Red indicates the use of the spatial model, while black indicates the spatial model was not used.

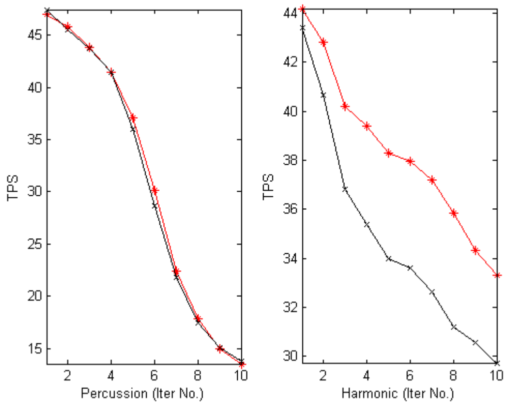

Fig. 3: Target related Perceptual Score for Percussive (left) and Harmonic (right) separations. Red indicates the use of the spatial model, while black indicates the spatial model was not used.

in performance, followed by a large drop in performance with subsequent iterations. In contrast, the performance for percussion separation keeps increasing up until iteration 5, before starting to fall off. This suggests that to achieve optimal separation of the percussive and harmonic parts, different numbers of iterations should be used, but that if a balance between separation quality of both sources is desirable, then 2 iterations is optimal. Also of interest is that the spatial model seems to have only minor effects on the OPS, resulting in slightly improved performance for the percussive sources for most iteration numbers, but giving slightly poorer performance for the harmonic sources.

It should also be noted that the phenomenon of falling OPS scores at higher number of iterations is not specific to KAM and also occurs with NTF-based methods [15]. This highlights a distinction between numerical convergence and perceptual quality of the separated sources obtained from the algorithms, and is an area that needs further research in sound source separation in general.

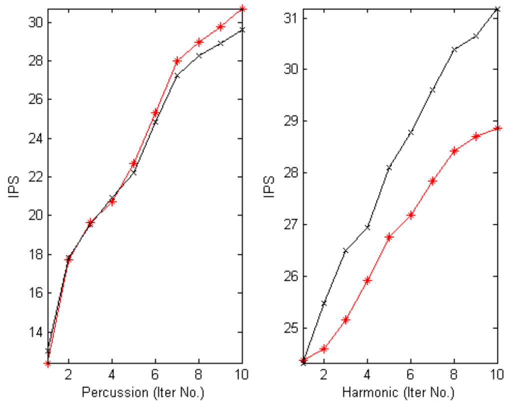

Fig. 4: Interference related Perceptual Score for Percussive (left) and Harmonic (right) separations. Red indicates the use of the spatial model, while black indicates the spatial model was not used.

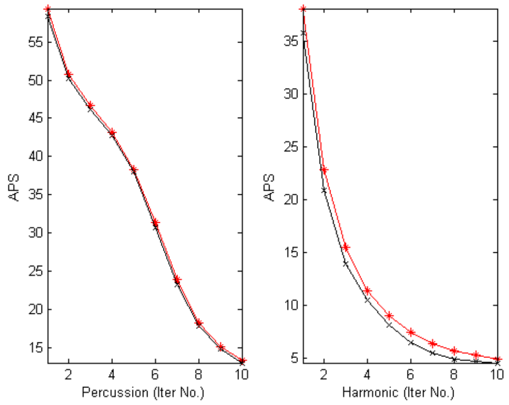

Fig. 5: Artifact related Perceptual Score for Percussive (left) and Harmonic (right) separations. Red indicates the use of the spatial model, while black indicates the spatial model was not used.

With respect to the TPS, shown in Figure 3, it can be seen that there is a constant drop with increasing iteration numbers, with the spatial model and non-spatial model results being very similar for the percussive sources. In contrast, the use of the spatial model results in a increase in TPS over the non-spatial model for harmonic sources. The IPS values, shown in Figure 4 show the exact opposite trends to those in Figure 3, with IPS increasing constantly with iteration number. Again, the results for percussive separation are very similar regardless of whether the spatial model is used or not, while for the harmonic separation, the spatial model results in decreased separation performance. This suggests there is a trade-off in the algorithm between accurate spatial estimation of the harmonic source and the amount of interference from the percussive source which remains in the harmonic source. Finally, the APS, shown in Figure 5 again shows a constant drop with iteration number, with similar results achieved regardless of the presence or otherwise of the spatial model. This shows that the spatial model has had little effect in reducing artifacts in the separated signals.

With respect to computation time, the KAM H/P Separation algorithm is still very computationally efficient, if 2 iterations are used, then runtime was approximately a third of the excerpt duration on a modern laptop computer, but was 1.7 times real-time for 10 iterations. It should be noted that this was using an unoptimised version of median filtering, and that increased speed can be achieved by implementing the median filtering in a manner similar to that described in [16].

V Conclusions

In this paper we have shown how a newly-proposed framework for source separation can be used to generalise and improve on an existing H/P Separation algorithm. This new framework, termed Kernel Additive Modelling models sources through local regularities in their spectrograms. Individual time-frequency bins in these spectrograms are assumed to be close in value to other bins nearby in the spectrogram, where nearness is defined through a source-specific proximity kernel. Separation is then performed using the kernel backfitting algorithm. The performance of the KAM H/P separator improves on the original median-filtering based H/P separation algorithm, which can be viewed as a single iteration of the KAM algorithm when the spatial model is omitted. This shows the utility of the KAM framework for the development of sound source separation algorithms.

References

[1] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama. "Comparative Evaluations of Various Harmonic/percussive Sound Separation algorithms based on anisotropic continuity of spectrogram". *Proc. ICASSP*, 465-468 2012.

[2] N. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama. "Multichannel Harmonic and Percussive Component Separation by Joint Modeling of Spatial and Spectral Continuity". *Proc. ICASSP*, 205-208 2011.

[3] D. FitzGerald. "Harmonic/Percussive Separation using Median Filtering". *13th International Conference on Digital Audio Effects (DAFX10)*, Graz, Austria, 2010.

[4] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo and L. Daudet. "Kernel Additive Models for Source Separation". *submitted to IEEE Transactions on Signal Processing*, 2014.

[5] A. Liutkus, Z. Rafii, B. Pardo, D. FitzGerald, and L. Daudet. "Kernel Spectrogram Models for Source Separation *4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014)*. 2014.

[6] A. Cichocki, R. Zdunek, A. Phan and S. Amari. "Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation". *Wiley Publishing*, 2009.

[7] A. Bregman. "Auditory Scene Analysis". *MIT Press*,1990.

[8] W. Cleveland and S. Devlin. "Locally weighted regression: An approach to regression analysis by local fitting". *Journal of the American Statistical Association.* vol. 83, pp. 596610, 1988.

[9] T. Hastie and R. Tibshirani. "Generalized Additive Models". *Statistical Science*, vol 1, pp 297–310, 1986.

[10] D. Barry, E. Coyle, and B. Lawlor, "Sound Source Separation: Azimuth Discrimination and Resynthesis", *Proc. 7th International Conference on Digital Audio Effects*, 2004

[11] A. Liutkus and Z. Rafii and R. Badeau and B. Pardo and G. Richard, *Adaptive filtering for music/voice separation exploiting the repeating musical structure*, In IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.

[12] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking", *IEEE Transactions on Signal Processing*, Vol. 52, No. 7, 2004.

[13] N. Duong, E. Vincent, and R. Gribonval. "Under-determined reverberant audio source separation using a full-rank spatial covariance model". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18 pp 1830–1840, 2010.

[14] 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*, 2011, 19 (7), pp. 2046-2057.

[15] D. Fitzgerald and R. Jaiswal "'On the use of Masking Filters in Sound Source Separation". *15th International Conference on Digital Audio Effects*, 2012.

[16] A. Robertson, A. Stark, and M. Davies "Percussive Beat tracking using real-time median filtering". *6th International Workshop on Machine Learning and Music*, 2013.