Kernel Spectrogram models for source separation

Antoine Liutkus, Zafar Rafii, Bryan Pardo, Derry Fitzgerald, Laurent Daudet

KERNEL SPECTROGRAM MODELS FOR SOURCE SEPARATION

Antoine Liutkus[*1,2,3], *Zafar Rafii*[4], *Bryan Pardo*[4], *Derry Fitzgerald*[5], *Laurent Daudet*[6]

[1]Inria, Villers-lès-Nancy, F-54600,France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[4]Northwestern University, Evanston, IL, USA
[5]NIMBUS Centre, Cork Institute of Technology, Ireland [6]Institut Langevin, Paris Diderot Univ., France

ABSTRACT

In this study, we introduce a new framework called Kernel Additive Modelling for audio spectrograms that can be used for multichannel source separation. It assumes that the spectrogram of a source at any time-frequency bin is close to its value in a neighbourhood indicated by a source-specific *proximity kernel*. The rationale for this model is to easily account for features like periodicity, stability over time or frequency, self-similarity, etc. In many cases, such local dynamics are indeed much more natural to assess than any global model such as a tensor factorization. This framework permits one to use different proximity kernels for different sources and to estimate them blindly using their mixtures only. Estimation is performed using a variant of the kernel backfitting algorithm that allows for multichannel mixtures and permits parallelization. Experimental results on the separation of vocals from musical backgrounds demonstrate the efficiency of the approach.

***Index Terms*—audio source separation, spatial filtering, spectrogram models**

I. INTRODUCTION

Source separation is a field of research that gathered much attention during the last 20 years [1]. Its objective is to recover several unknown signals called *sources* that were mixed together into observable *mixtures*. In audio signal processing, the sources are typically understood as different auditory streams [2] that make sense perceptually. In music processing for instance, they correspond to different instruments playing in a song. In spoken speech enhancement, one source may be the target voice whereas others correspond to background noise to filter out.

One of the dominating paradigm today for the separation of audio waveforms is the use of generalized Wiener filtering [3], [4] under Gaussian assumptions. In practice, this approach requires good models of the spectrograms of each source along with its spatial characteristics and permits very good separation provided these parameters are well estimated. The main challenge in achieving good separation then mainly becomes the devising of good spectrogram models that catch the main features of the sources to separate while requiring few parameters. Techniques such as Nonnegative Tensor Factorizations (NTF, see [5]) are often used to this purpose. Their principle is to assume that the spectrogram of each source may be decomposed as the sum of only a few spectral templates activated over time. In spite of their appealing tractability, NTF models often come with some limitations. First, they often fail at efficiently decomposing in a concise way many sources such as voice that exhibit a great variety of spectra. Second, they typically assume that different sources are characterized by different sets of spectra, which may not be realistic, e.g. for mixtures of speech.

In this study, instead of decomposing the spectrograms of the sources as a combination of fixed patterns, we rather focus on their *regularities* to identify them from the mixtures. Auditory Scene Analysis [2] indeed demonstrated on perceptual grounds that apart from the commonly used *harmonic* property that refers to an *absolute* feature of many auditory sources, local features such as *repetitivity*, *continuity or common fate* are fundamental in our ability to discriminate them within a mixture. These dynamic features can be seen not to depend on any particular spectral absolute *template* modelled by NTF, but rather on local regularities concerning their evolution over time, frequency and space.

In order to model dependencies within the spectrograms of the sources, we use *kernel local parametric models*, that are deeply rooted in the *local regression* approach [6]. Basically, the value of the spectrogram of a source at some time-frequency (TF) bin is supposed to be close to its values *nearby*. There are often sophisticated ways to decide whether two TF bins have similar values. In full generality, *proximity kernels* are introduced which give the proximity of two TF points from the perspective of a source. There are several ways of building such kernels, including direct analytical expressions or through the use of *feature*

This work is supported by LABEX WIFI (Laboratory of Excellence within the French Program "Investments for the Future") under references ANR-10-LABX-24 and ANR-10-IDEX-0001-02 PSL. Correspondance should be addressed to antoine@liutkus.net

*space*s. In any case, the value of a source spectrogram is supposed to be correctly estimated using its values at locations whose proximity is high. Separation of additive sources in this context can be performed using a variant of the *backfitting* algorithm [7]. Different sources are modelled through different proximity kernels. The approach, coined in as Kernel Additive Modelling (KAM), is flexible enough to permit taking prior knowledge about the dynamics of many kinds of signals into account. The proposed methodology encompasses many popular methods for audio source separation, such as DUET [8], ADRESS [9], REPET and REPET-SIM [10], [11], [12], median filtering for drums removal [13], etc. Moreover, it provides an efficient way to devise new specific separation algorithms for sources that are characterized by local features, rather than by a global additive model such as NTF. We show its performance on music/voice separation and provide a complete MATLAB implementation.

II. MODEL AND METHOD

II-A. Notations and model

Let the *mixture* $\tilde{x}$ be a set of I time series, where $\tilde{x}(n,i)$ denotes the value of the i^{th} channel of the mixture at sample n. In music processing, we often have $I = 2$ in the stereo case. We assume that the mixture is the sum of J *sources* $\tilde{s}_j$: $\tilde{x}(n,i) = \sum_{j=1}^{J} \tilde{s}_j(n,i)$.

Let $\{s_j\}_{j=1\cdots J}$ and x be the Short Term Fourier Transforms (STFTs) of the J sources and of the mixture, respectively. They are all $N_f \times N_t \times I$ tensors, where N_f is the number of frequency bands and N_t the number of frames. $s_j(f,t)$ is the $I \times 1$ vector that gives the value of the STFT s_j for all channels (e.g. left and right) at TF bin (f,t).

Under the Local Gaussian Model [4], the vectors $s_j(f,t)$ for all TF bins of a multichannel audio signal are assumed to be independent, each one of them being distributed with respect to a multivariate centered complex Gaussian distribution:

$$\forall (f,t),\ s_j(f,t) \sim \mathcal{N}_c\left(0, \mathbf{s}_j(f,t) R_j(f)\right). \tag{1}$$

In expression (1), boldfaced $\mathbf{s}_j(f,t) \geq 0$ indicates the *spectrogram* of source j at TF bin (f,t). It is a nonnegative scalar that basically accounts for the *energy* of that source at TF bin (f,t). $R_j(f)$ is a $I \times I$ positive semidefinite matrix that 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 that frequency[1]. Such a model notably encompasses the popular linear instantaneous and convolutive cases [1], that correspond to a rank-1 $R_j(f)$ [4].

Since the mixture $x(f,t)$ is the sum of J independent random Gaussian vectors $s_j(f,t)$, it also has a Gaussian distribution. If the parameters $\mathbf{s}_j$ and R_j are known or estimated as $\hat{\mathbf{s}}_j$ and $\hat{R}_j$, it can be shown that the Minimum Mean-Squared Error (MMSE) estimates $\hat{s}_j$ of the STFTs of the sources are readily obtained through generalized spatial Wiener filtering [3], [14], [15], [4], using:

$$\hat{s}_j(f,t) = \hat{\mathbf{s}}_j(f,t) \hat{R}_j(f) \left[\sum_{j'=1}^{J} \hat{\mathbf{s}}_{j'}(f,t) \hat{R}_{j'}(f)\right]^{-1} x(f,t). \tag{2}$$

The waveforms of the sources in the time domain are then easily obtained through inverse STFT.

II-B. Kernel constant models for spectrograms

In many source separation studies, the spectrograms $\mathbf{s}_j$ of the sources are taken as the activation over-time of a few K spectral templates $W_j(f,k)$:

$$\mathbf{s}_j(f,t) = \sum_{k=1}^{K} W_j(f,k) H_j(k,t), \tag{3}$$

where H_j gives the activation gains of these templates over time. This approach leads to the popular Nonnegative Matrix Factorization (NMF) framework [16], [17], [18], [19] for audio source separation.

Here, we do not assume that the spectrogram $\mathbf{s}_j$ of a source is properly described using a parametric model such as (3). Instead, we will draw from the ideas of local regression [6] to model spectrograms only *locally*. More specifically, prior knowledge about the source comes as neighbourhoods $\mathcal{I}_j(f,t)$, called the *proximity kernel* of source j, which indicates the TF points where the spectrogram has a value equal to $\mathbf{s}_j(f,t)$:

$$\forall (f',t') \in \mathcal{I}_j(f,t),\ \mathbf{s}_j(f',t') \approx \mathbf{s}_j(f,t).$$

In musical signals for instance, percussive elements are known to be self-similar along the frequency axis, while harmonic stable sounds are self-similar along time [13], leading to proximity kernels that are either vertical or horizontal, as depicted in figure 1(a) and 1(b) respectively. Alternatively, source j may be known to be repetitive at period T_j, so that $\mathcal{I}_j(f,t)$ includes $\{(f, t+kT_j)\}_{k\in\mathbb{Z}}$, as in figure 1(c).

This approach is flexible enough to take prior knowledge about the dynamics of many kinds of signals into account. It encompasses a large number of recently proposed methods for source separation [13], [20], [21], [10], [12], [8] and provides an efficient way to devise new specific separation algorithms for sources that are characterized by local features, instantiated by the definition of the neighbours $\mathcal{I}_j(f,t)$ of any point (f,t), rather than by a global model such as NTF.

II-C. The kernel backfitting algorithm

Assume that $\mathbf{s}_j$ is not observed exactly but only through a noisy observation $\mathbf{z}_j$ whose likelihood $p(\mathbf{z}_j(f,t) \mid \mathbf{s}_j(f,t))$ is known. This likelihood accounts for the fact that even if

[1] Thus, $x(f,t)$ and $s_j(f,t)$ are $I \times 1$ vectors, boldfaced $\mathbf{s}_j(f,t)$ is a scalar and $R_j(f)$ is a $I \times I$ matrix. Estimates are denoted $\hat{s}_j$, $\hat{\mathbf{s}}_j$ and $\hat{R}_j(f)$.

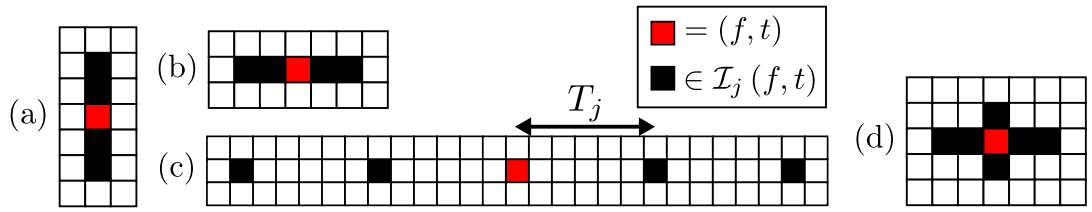

Fig. 1. Examples of proximity kernels to account for prior knowledge about sources. (a) *vertical*, for percussive elements, (b) *horizontal*, for stable harmonic elements, (c) *periodic*, for repetitive elements, (d) *cross-like*, for smoothly varying spectrograms such as vocals.

$\mathbf{z}_j(f,t)$ is likely to be close to $\mathbf{s}_j(f,t)$, important discrepancies may occur during iterations. This can for example be taken into account by choosing a Laplacian likelihood, leading to: $-\log p(\mathbf{z}_j(f,t) \mid \mathbf{s}_j(f,t)) = |\mathbf{z}_j(f,t) - \mathbf{s}_j(f,t)|$. With such a likelihood and supposing that $\mathbf{z}_j(f,t)$ are all independent, $\mathbf{s}_j(f,t)$ may be estimated through maximum likelihood as:

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

which is readily shown to be equivalent to:

$$\hat{\mathbf{s}}_j(f,t) = \operatorname{median}\{\mathbf{z}_j(f',t') \mid (f',t') \in \mathcal{I}_j(f,t)\}, \quad (4)$$

so that $\hat{\mathbf{s}}_j$ is readily estimated through a median filtering of $\mathbf{z}_j$, which can be achieved in linear complexity thanks to efficient implementations found in most numerical computing libraries.

The *kernel backfitting algorithm* we propose for estimation of the sources spectrograms $\mathbf{s}_j$ is strongly inspired by the original backfitting procedure proposed in the context of nonparametric additive modelling [7], [22]. This algorithm proceeds in an iterative fashion, where separation and re-estimation of the parameters are performed alternatively. In this procedure, the spectrograms $\mathbf{z}_j(f,t)$ of the current estimates $\hat{s}_j$ of the sources STFT are used as noisy observations of their true value, and re-estimation of $\hat{\mathbf{s}}_j$ from $\mathbf{z}_j$ is achieved through median filtering (4).

The whole procedure is summarized in algorithm 1, where $\cdot^\star$ denotes conjugate transpose and $\operatorname{tr}(\cdot)$ stands for the trace of a square matrix. For more details about re-estimation of the spatial covariance matrices $R_j(f)$, the reader is referred to [4], [23]. Remarkably, all sources can be handled in parallel during step 3, leading to a computationally efficient technique for source separation. Typical computing time is about 5 times slower than real time on a modern desktop computer. Computational complexity furthermore scales linearly with the duration of the audio to process and the number of iterations (typically 5).

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 $\mathcal{I}_j(f,t)$ as in figure 1.
   - Number L of iterations
2) **Initialization**
   - $l \leftarrow 1$
   - $\forall j, \hat{\mathbf{s}}_j(f,t) \leftarrow x(f,t)^\star x(f,t)/IJ$
   - $R_j(f) \leftarrow I \times I$ identity matrix
3) Compute estimates $\hat{s}_j$ of all sources using (2)
4) For each source j:
   a) $C_j(f,t) \leftarrow \hat{s}_j(f,t)\hat{s}_j(f,t)^\star$
   b) $\hat{R}_j(f) \leftarrow \frac{I}{T}\sum_t \frac{C_j(f,t)}{\operatorname{tr}(C_j(f,t))}$
   c) $\mathbf{z}_j(f,t) \leftarrow \frac{1}{I}\sum_t \operatorname{tr}\left(\hat{R}_j(f)^{-1} C_j(f,t)\right)$
   d) $\hat{\mathbf{s}}_j(f,t) \leftarrow \operatorname{median}\{\mathbf{z}_j(f',t') \mid (f',t') \in \mathcal{I}_j(f,t)\}$
5) If $l < L$ then set $l \leftarrow l+1$ and go to step 3
6) **Output**:
   sources spectrograms $\hat{\mathbf{s}}_j$ and spatial covariance matrices $\hat{R}_j(f)$ to use for filtering (2).

III. EVALUATION

Separating vocals from the musical background in popular music is a very challenging task that has many applications in the entertainment industry and in the automatic indexing and querying of musical databases [24]. In the recent years, it has been the topic of numerous research studies and many different techniques were devised for this purpose [25], [26], [21], [27], [12], [11], [10]. In the following, we detail and evaluate a voice/music separation procedure based on KAM.

III-A. Data and metrics

In our experiments, the processed data consists of 10 complete stereo tracks from the album *The Pet Sounds* by the popular band THE BEACH BOYS. This band published an extensive set of studio recordings for this album in 1997 as a commercial release[2], which includes separated vocals and background as stereo tracks. After some manual synchronization, they were mixed down so as to produce the full-length stereo mixtures to separate.

[2]THE BEACH BOYS, *The Pet Sounds Sessions*, Capitol rec. 1997.

For the purpose of evaluation, all separated full-length vocals and backgrounds tracks from each technique are segmented into $10s$ excerpts, yielding 168 such excerpts, for which separation performance is evaluated on both the separated vocals and background music. The metrics considered are the classical Source to Distorsion Ratio (SDR) from the BSSEVAL toolkit [28]. It is given in dB and is higher for better separations.

Since different excerpts may yield very different separation difficulties, it is known that directly averaging BSSEVAL metrics is not meaningful [29]. For this reason, the delta-metric Δ_{SDR} is considered instead. It gives the difference of the performance with those obtained through oracle Wiener separation [30], which uses the true spectrograms of the sources in(2). This delta-metric were shown to be more reliable for averaging over a corpus [29].

III-B. Techniques and parameters

For performance comparison, each full track of the corpus was separated using the techniques IMM [25], RPCA [27], REPET-SIM [11], [12], adaptive REPET [21] and adaptive REPET with a further DUET processing [8].

For KAM separation, background and vocals were both modelled as having locally constant spectrograms as described in section II-B. In one so called *KAM multirepet* setting, the musical accompaniment is modelled as the sum of 5 repeating patterns as in figure 1 (c). In another *KAM multirepet+harm* setting, a further stable harmonic source is included in the background model as in figure 1 (b) and its length corresponds to 2s. In all cases, the vocal part was modelled using the cross-like kernel of figure 1 (d), whose height and length were respectively set to 50Hz and 0.4s. Frames of 90ms with an overlap of 85% were considered for the computation of STFTs and the periods of the patterns were estimated by a peak-picking of the beat-spectrum [20].

Running time is approximately 5 times slower than real time and varies linearly with the number of iterations (typically 5) and the number of sources. A full MATLAB implementation of the proposed method is made publicly available on the companion webpage of this paper[3], along with audio examples.

III-C. Results

Considering the results given on figure 2, we note that objective performance of the proposed KAM setup is at the state of the art level. More precisely, the background is consistently shown to be better estimated than with other competitive methods, which is very interesting for karaoke applications and is an encouraging result. However, the scores tend to show that other techniques such as adaptive REPET [21] give better estimates for the vocals. Several remarks may be done considering this evaluation.

First, the lack of a large full-tracks audio corpus prevents giving separation performance for different kinds of music.

[3] www.loria.fr/~aliutkus/kam/

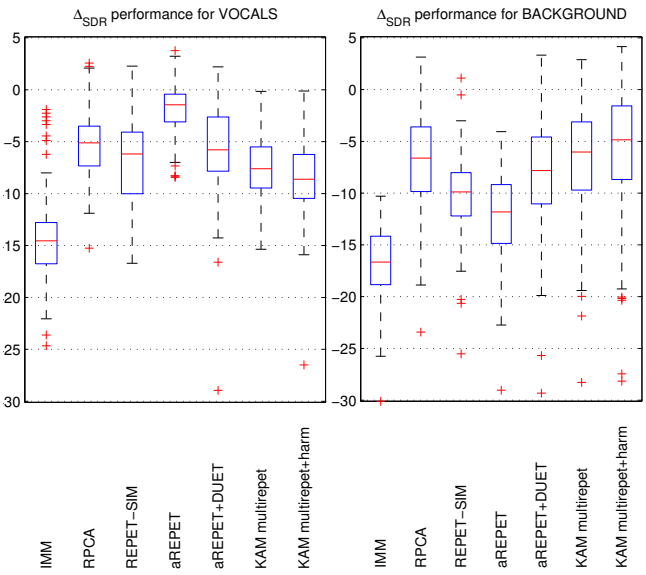

Fig. 2. Distribution of the Δ_{SDR} score over 10s excerpts. Higher is better. The background is shown to be well separated.

The *Pet Sounds* album is indeed characterized by mostly center-panned vocals, which fits well the assumptions of a DUET approach [8] and explains the very high scores of the *aREPET+DUET* method. Extensive informal perceptual testing has shown that KAM is very robust to different kinds of music, from black metal to jazz to electro-pop music. The reader is strongly encouraged to listen to the separated signals on the companion webpage of this paper and to run the provided MATLAB script on his own sound examples.

Second, the scores given here only hold for the particular choice of proximity kernels we made in this voice/music separation task. KAM may be used in many other settings or yield improved performance with more adequate proximity kernels depending on the track considered. Remarkably, 5 out of the 7 techniques evaluated here can be understood as particular instances of KAM.

IV. CONCLUSION

In this paper, we have proposed a new framework for audio source separation, where each source is modelled through the local regularities of its spectrogram. The spectrogram taken at some time-frequency bin is supposed to be close to its values nearby, where *nearness* is defined through a source-specific *proximity kernel*. Separation is performed using a variant of the *backfitting* algorithm, coined in as *kernel backfitting*. The proposed method comes as a unifying framework for many state-of-the-art techniques for source separation and yields an easy and principled way to combine local models in order to build sophisticated mixture models. The corresponding algorithms are easy to implement and provide good performance.

V. REFERENCES

[1] P. Comon and C. Jutten, eds., *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[2] A. Bregman, *Auditory Scene Analysis, The perceptual Organization of Sound*. MIT Press, 1994.

[3] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.

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

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

[6] 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. 596–610, 1988.

[7] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical Science*, vol. 1, pp. 297–310, 1986.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] D. Barry, B. Lawlor, and E. Coyle, "Real-time sound source separation using azimuth discrimination and resynthesis," in *Proc.of 117th Audio Engineering Society Convention*, 2004.

[10] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 1, pp. 71–82, 2013.

[11] D. FitzGerald, "Vocal separation using nearest neighbours and median filtering," in *23rd IET Irish Signals and Systems Conference (ISSC2012)*, (NUI Maynooth, Ireland), June 2012.

[12] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *ISMIR* (F. Gouyon, P. Herrera, L. G. Martins, and M. Müller, eds.), pp. 583–588, FEUP Edições, 2012.

[13] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, (Graz, Austria), Sept. 2010.

[14] A. Cemgil, P. Peeling, O. Dikmen, and S. Godsill, "Prior structures for Time-Frequency energy distributions," in *Proc. of the 2007 IEEE Workshop on. App. of Signal Proc. to Audio and Acoust. (WASPAA'07)*, (NY, USA), pp. 151–154, Oct. 2007.

[15] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3155 –3167, July 2011.

[16] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. of Irish Sig. and Systems Conf. (ISSC'08)*, 2008.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[18] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, pp. 550–563, Mar. 2010.

[19] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *7th International Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, 2010.

[20] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 221 –224, may 2011.

[21] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 53–56, mars 2012.

[22] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. London: Chapman & Hall, 1990.

[23] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[24] E. Gómez, F. Cañadas, J. Salamon, J. Bonada, P. Vera, and P. Cabañas, "Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing," in *13th Int. Soc. for Music Info. Retrieval Conf.*, (Porto, Portugal), Oct. 2012.

[25] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, pp. 1180 –1191, oct. 2011.

[26] J. Durrieu and J. Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. of International Conference on Latent Variable Analysis and Signal Separation*, (Tel-Aviv, Israel), March 12-15 2012.

[27] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462 –1469, July 2006.

[29] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937 – 1949, 2012.

[30] E. Vincent, R. Gribonval, and M. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, pp. 1933 – 1950, Aug. 2007.