PROJET - SPATIAL AUDIO SEPARATION USING PROJECTIONS

Derry FitzGerald[1] *Antoine Liutkus*[2] *Roland Badeau*[3]

[1]NIMBUS Centre, Cork Institute of Technology, Ireland
[2]Inria, speech processing team, Villers-lès-Nancy, France
[3]LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France.

ABSTRACT

We propose a projection-based method for the unmixing of multichannel audio signals into their different constituent spatial objects. Here, spatial objects are modelled using a unified framework which handles both point sources and diffuse sources. We then propose a novel methodology to estimate and take advantage of the spatial dependencies of an object. Where previous research has processed the original multichannel mixtures directly and has been principally focused on the use of inter-channel covariance structures, here we instead process projections of the multichannel signal on many different spatial directions. These linear combinations consist of observations where some spatial objects are cancelled or enhanced. We then propose an algorithm which takes these projections as the observations, discarding dependencies between them. Since each one contains global information regarding all channels of the original multichannel mixture, this provides an effective means of learning the parameters of the original audio, while avoiding the need for joint-processing of all the channels. We further show how to recover the separated spatial objects and demonstrate the use of the technique on stereophonic music signals.

***Index Terms*—Sound Source Separation, α-stable, Spatial Projection**

I. INTRODUCTION

Demixing audio signals into their constitutive audio objects has attracted much attention in the audio research community under the name of *sound source separation* (SSS, see, e.g. [21], [32], [34] and references therein). Such an operation indeed enables many new interactions with audio content, such as upmixing [1], restoration [5] or active listening, notably including automatic karaoke applications [17], [14], [10], [30].

Broadly speaking, research in SSS has undergone two main directions. The first one concerns the devising of spectral models for the audio objects to separate. Indeed, provided that good spectrograms estimates are available, excellent performance for the separation can be obtained through Time-Frequency (TF) masking, that can be understood as Wiener filtering from a signal processing point of view: see e.g. [4], [3], [19], [20] for the Gaussian case and, more recently, [18] for some theory on TF masking with fractional power spectrograms. A lot of effort has hence been devoted in proposing flexible yet powerful models able to capture the specificities of audio objects, while easily fitting the data. Among them, the Nonnegative Matrix Factorization has been extremely popular in the last ten years (NMF [6], [32], [8]), while recent research has focused on deep neural networks [12], [13], [15], [36], [27] or on local models for audio [11], [23], [25].

Another very important direction for research concerns the design of separation techniques exploiting the so-called *spatial information* found in multichannel audio signals. It is indeed common for audio objects not to be present in the same way in the different channels of the recordings, so that separation based on such spatial diversity is possible. In the case of music, audio objects are routinely located at different *panning positions* in the left-right stereo plane. Techniques exploiting this diversity include DUET [37] and ADRESS [2], that build a binary TF mask based on the identified panning positions of the sources. They were showed to provide extremely robust performance when spatial diversity is indeed present in the mix. Recent research benefited from the well established beamforming techniques to generalise the single channel TF masking techniques to yield multichannel Wiener filtering [7], [29], [34], hence unifying these two research directions to create techniques that can exploit both the spectral and spatial diversity of the audio objects.

However, recent multichannel separation techniques suffer from several drawbacks. First, they are still limited to Gaussian models, which are the only ones we know of that provide a theoretically grounded separation procedure in the multichannel case. However, recent research [18], [22], [31] suggests that using non-Gaussian modelling may be preferable for audio. If some studies proposed multichannel separation procedures that depart from the Gaussian case [26], no probabilistic model we are aware of yet justifies the methodology even if the technique is effective. Second, all those methods require inversion of inter-channels covariance matrices for all TF bins of the mixture, which is computationally demanding as soon as we have more than two channels. Previous research addressed this issue [33], but it comes at the cost of assuming only one source to be active for each TF bin.

In this paper, we focus on exploiting spatial diversity to provide better separation performance, while allowing for non-Gaussian spectral modelling. We present a method that avoids the annoying artifacts due to the binary nature of DUET or ADRESS, and which is much more computationally effective than multichannel Wiener filtering. Indeed, our proposed technique, coined as PROJET, does not require the inversion of any covariance matrix. The main idea of the method is to combine the different channels of the mixture in a first *projection* step that yields different combined observations, and then to process these projections independently. Since each one contains different information from all channels of the original mixture, the spatial information of the audio objects is preserved and reconstruction is possible as a post-processing step.

The remaining of the paper is organized as follows. First, we present the probabilistic model we use in section II. Then, we discuss the parameter estimation technique in section III. Finally, we evaluate the method on the separation of stereo music signals in section IV.

II. SPATIAL MODEL

It is assumed that a multichannel audio signal, termed a *mixture*, composed of I channels is observed, where $I = 2$ corresponds to the typical stereophonic case. The Short Time Fourier Transform (STFT) of the mixture is denoted x and is a tensor of size $N_f \times N_t \times I$, where N_f and N_t are the numbers of frequency bins and time frames respectively. Then $x(f,t)$ is a $I \times 1$ vector, giving the value of the complex spectrum of each channel of the mixture (e.g. left

This work was partly supported by the research programmes EDiSon3D (ANR-13-CORD-0008-01) and KAMoulox (ANR-15-CE38-0003-01) funded by ANR, the French State agency for research.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

and right) at TF bin (f, t). We then assume that the mixture is the sum of J multichannel signal STFTs y_j, each of size $N_f \times N_t \times I$ which we term the *object images*:

$$\forall (f,t), x(f,t) = \sum_j y_j(f,t). \tag{1}$$

We now define a simple model for *punctual* sources (sound objects which appear to come from a definite direction) before extending it to handle *diffuse* sound objects (which appear to originate from multiple directions).

II-A. Punctual Model

For punctual objects, it is assumed that the object image y_j is generated from an underlying monophonic signal termed the object *source* whose STFT is denoted s_j and is a matrix of size $N_f \times N_t$. In the punctual model, each image is obtained by multiplying each source by a particular gain for each channel while mixing. For an I channel mixture we define[1]

$$\mathcal{P} \triangleq \mathcal{C}_I \cap \mathbb{R}_+^I, \tag{2}$$

as the *panning set*, which lies on the intersection of the unitary sphere $\mathcal{C}_I$ in $\mathbb{R}^I$ and the positive cone $\mathbb{R}_+^I$ (where all coordinates are positive). A *panning direction* $\theta \in \mathcal{P}$ is a nonnegative $I \times 1$ vector such that its norm $\|\theta\| = 1$. Then, y_j is given as:

$$\forall (f,t), y_j(f,t) = \theta s_j(f,t). \tag{3}$$

This can be seen as a generalization of the classic stereophonic panning law where the panning angle is in the range $\phi \in [0, \pi/2]$, and where $\theta = [\cos\phi \ \sin\phi]^\top$ with $\cdot^\top$ denoting matrix transpose.

II-B. Diffuse Spatial Model

However, in many cases, the punctual model is not sufficient for handling real-world signals. To this end, we propose an extension where the image y_j is a weighted sum of independent contributions coming from all panning directions in $\mathcal{P}$:

$$y_j(f,t) = \int_{\theta \in \mathcal{P}} \theta q_j(\theta) s_j(f,t,\theta) d\theta, \tag{4}$$

where all $\{s_j(f,t,\theta)\}_\theta$ are *object sources* and are all assumed to be independent and $q_j(\theta) \geq 0$ is a *panning gain* indicating the strength of the object source coming from direction θ. This can be further simplified by approximating the integral as a discrete sum over a fixed countable set $\overline{\mathcal{P}}$ of L positions in $\mathcal{P}$:

$$y_j(f,t) = \sum_{\theta \in \overline{\mathcal{P}}} \theta q_j(\theta) s_j(f,t,\theta), \tag{5}$$

where $\overline{\mathcal{P}} = \{\theta_1, \ldots, \theta_L\} \in \mathcal{P}^L$. This model can be viewed as a simplification of that proposed in [26] where the physical acoustic modelling has been dropped.

We choose to assume that all $\{s_j(f,t,\theta)\}_\theta$ are not only independent, they are distributed with respect to an isotropic complex α-stable distribution, written $S\alpha S_c$ (see [18]):

$$s_j(f,t,\theta) \sim S\alpha S_c(P_j(f,t)). \tag{6}$$

where $P_j(f,t)$ is a nonnegative scale parameter, called the fractional spectral density (α-PSD) of object j at TF bin (f,t). It corresponds to the classical PSD when $\alpha = 2$. In effect, the model assumes that sources are additive in the magnitude-to-the-power-α domain, an approximation which has found widespread use in audio source separation.

Algorithm 1 PROJET Algorithm for audio separation through projections.

```
1) Input
     • Panning set $\overline{\mathcal{P}}$ and projection matrix $\mathbf{M}$
     • Number of iterations
     • Mixture $x$
2) Initialization
     • Compute projections $c(f,t)$ through (10)
     • Compute vectors $k_m$ with (8)
     • Initialize parameters $\Theta$ to non-negative values.
3) Parameter fitting: for each object $j$,
     a) Update $\alpha$-PSD $P_j$according to (13)
     b) Update panning coefficients $Q_j$ according to (14)
4) If another iteration is needed, go back to 3)
5) Separation: for each object $j$:
     a) Estimate the $M \times 1$ projected images $\hat{y}_j^c(f,t)$
        through (18)
     b) Estimate object image $\hat{y}_j$ through (20)
     c) Apply inverse STFT to $\hat{y}_j$ to recover waveforms
```

II-C. Spatial Projections

Now, consider a point $n \in \mathcal{C}_I$, the unitary sphere in $\mathbb{R}^I$. It is an $I \times 1$ vector. The dot product of this vector and the mixture, assuming independence of the sources, is given by:

$$\begin{aligned} \langle n, x(f,t) \rangle &= \sum_j \langle n, y_j(f,t) \rangle \\ &\sim S\alpha S_c\left(\sum_j P_j(f,t) k(n)^\top Q_j\right). \end{aligned} \tag{7}$$

where $k(n)$ is an $L \times 1$ vector defined as:

$$k(n) \triangleq \left[|\langle n, \theta_1 \rangle|^\alpha, \ldots, |\langle n, \theta_L \rangle|^\alpha\right]^\top \tag{8}$$

and the $L \times 1$ vector Q_j as:

$$Q_j \triangleq \left[q_j^\alpha(\theta_1), \ldots, q_j^\alpha(\theta_L)\right]^\top,$$

and where $\langle \cdot \rangle$ indicates dot product. We now consider a set $\{n_1, \ldots, n_M\}$ of M points in $\mathcal{C}_I$. Gathering their transpose as the rows of the $M \times I$ matrix $\mathbf{M}$, the resulting signals $\langle n_m, x(f,t) \rangle$ are grouped into *projection* matrices c_m, each of dimension $N_f \times N_t$:

$$c_m(f,t) \triangleq \langle n_m, x(f,t) \rangle. \tag{9}$$

We then denote $c(f,t)$ as the $M \times 1$ vector gathering the various $c_m(f,t)$:

$$c(f,t) \triangleq [c_1(f,t), \ldots, c_M(f,t)]^\top,$$

This results in a *projection tensor* c of size $N_f \times N_t \times M$. Equation (9) then leads to:

$$c(f,t) = \mathbf{M} x(f,t). \tag{10}$$

Now, following (7), the marginal distribution of each entry of the projection tensor is given by:

$$c_m(f,t) \sim S\alpha S_c\left(\sum_j P_j(f,t) k_m^\top Q_j\right), \tag{11}$$

where k_m is taken as a short-hand notation for $k(n_m)$ and is computed only once through (8). The parameters to be estimated in this model are the $L \times J$ panning gains Q, as well as the $N_f \times N_t$ objects α-PSD P_j and are gathered into a parameter set denoted Θ.

[1] $\triangleq$ denotes a definition

III. PARAMETER ESTIMATION

Taking the entries of the projection tensor c as the observations, we now estimate the parameters Θ using a Fractional Lower Order Moments (FLOM) fitting strategy, that basically amounts to enforcing that the α-moment of the observations matches the model [24]. Here, this moment corresponds to the magnitude of c to the power α and in this paper, we take the data-fit criterion for this fitting as the generalized Kullback-Leibler (KL) divergence, given by $d_{KL}\left(a \mid b\right) = a \log \frac{a}{b} - a + b$. For the value $\alpha \approx 1$ that we will be considering in our evaluation, this leads to:

$$\hat{\Theta} \leftarrow \underset{\Theta}{\operatorname{argmin}} \sum_{f,t,m} d_{KL}\left(\left| c_m\left(f,t\right) \right| \mid \sum_j P_j\left(f,t\right) k_m^\top Q_j \right), \tag{12}$$

Adapting the now-classical approach to derive non-negative multiplicative updates for that cost function [16], [9] results in the following update equations for P_j and Q_j:

$$P_j \leftarrow P_j \cdot \frac{\sum_m k_m^\top Q_j \left(v_m / \sigma_m\right)}{\sum_m k_m^\top Q_j} \tag{13}$$

$$Q_j \leftarrow Q_j \cdot \frac{\sum_{f,t,m} k_m \left[\left(v_m / \sigma_m\right) \cdot \left(P_j\right)\right]_{ft}}{\sum_{f,t,m} k_m \left[\left(P_j\right) / \sigma_m\right]_{ft}} \tag{14}$$

where:

$$v_m\left(f,t\right) = \left| c_m\left(f,t\right) \right| \tag{15}$$

and

$$\sigma_m\left(f,t\right) = \sum_j P_j\left(f,t\right) k_m^\top Q_j, \tag{16}$$

while $a \cdot b$ and $\frac{a}{b}$ stand for element-wise multiplication and division respectively.

Having estimated the parameters, there remains the requirement to separate the original mixture x given the parameters estimated using the projection tensor c. To do this, we first decompose the projection entries $c\left(f,t\right)$ into J *projected images* of the objects $y_j^c\left(f,t\right)$such that:

$$c\left(f,t\right) = \sum_j y_j^c\left(f,t\right). \tag{17}$$

Again discarding dependencies between the different M projection channels y_{mj}^c, we estimate each through their marginal expected value given the mixture and parameters:

$$\hat{y}_{mj}^c\left(f,t\right) = \frac{P_j\left(f,t\right) k_m^\top Q_j}{\sum_{j'} P_{j'}\left(f,t\right) k_m^\top Q_{j'}} c_m\left(f,t\right). \tag{18}$$

To recover the original object images y_j, note that (10) leads to:

$$c\left(f,t\right) = \mathbf{M} \sum_j y_j\left(f,t\right) = \sum_j \mathbf{M} y_j\left(f,t\right), \tag{19}$$

implying through (17) that

$$y_j^c\left(f,t\right) = \mathbf{M} y_j\left(f,t\right).$$

Given this, we then estimate the corresponding image y_j using a least-squares strategy:

$$\hat{y}_j\left(f,t\right) = \mathbf{M}^\dagger \hat{y}_j^c\left(f,t\right), \tag{20}$$

with $\cdot^\dagger$ denoting pseudo-inversion.

The entire procedure, which we term PROJET (PROJection Estimation Technique) is outlined in Algorithm 1. The mixture signal is taken as an input, as well as the parameters $\overline{\mathcal{P}}$ and $\mathbf{M} = \left\{n_1, \ldots, n_M\right\}^\top$, which are required to construct the projection tensor and the elements of the dictionary k_m. The parameters are then iteratively estimated before being utilized for separation.

As the choice of $\overline{\mathcal{P}}$ and $\mathbf{M}$ are important for good performance of the technique we now discuss them. Regarding the panning set $\overline{\mathcal{P}}$, for the stereo case ($I = 2$), we have observed that having more panning directions than objects ($L > J$) is a good strategy, as well as choosing $\overline{\mathcal{P}}$ so that the stereo space is spanned equally by the L panning directions. Experiences with PROJET also suggest that separation is improved when a given projection direction, say n, is orthogonal to one of the elements of $\overline{\mathcal{P}}$, say θ, thereby ensuring that energy from direction θ is cancelled out in $\left\langle n, x\left(f,t\right)\right\rangle$.

IV. EVALUATION OF PROJET

In order to evaluate the PROJET method we created a test set from the MSD100 development set used in SiSEC 2015[2]. This consists of 50 full length songs created from mixtures of 4 objects, all with a sample rate of 44.1kHz. The original recordings for these objects are available as part of the development set. 30 second excerpts were taken from these recordings, with the same relative start and end points, for all objects. Using the panning model described in (3), stereo objects were created from these mono excerpts, and then summed to generate the mixture signals. The objects were mixed with an equal angle between them. To test the robustness of the algorithm with respect to the angle between the objects, the angle was varied from 10 degrees to 30 degrees in steps of 10, resulting in a total of 150 test mixtures.

The metrics chosen for evaluation were those defined and implemented in version 3 of the BSS Eval Toolbox [35]. These are Signal to Distortion Ratio (SDR), which measures overall sound quality of a separated object, Signal to Artifacts Ratio (SAR), which measures the presence of artifacts, Signal to Interference Ratio (SIR), which measures the presence of interference from other objects in the mixture, and Image to Spatial Distortion Ratio (ISR), which measures spatial distortions in the position of the separated object.

Two tests were run, the first tested the oracle performance of PROJET when it was given the correct spatial angle of the sources. In this case, the sources were assumed to be punctual and so Q becomes the identity matrix. Then the only parameters to be updated are the source fractional spectral densities P_j. The number of objects to separate was set at $J = 4$. The algorithm was ran for 1000 iterations as it was observed that perceptually better separation was achieved at higher iteration numbers. The STFTs of the original multichannel mixture were calculated using a window length of 4096 samples, and a hop size of 1024 samples, giving a 75% overlap between frames, using a Hann window. As a benchmark, PROJET was tested against the well known DUET algorithm, where again DUET was provided with the correct source angles. The time-frequency bins that fell within a given angle on either side of the actual objects positions were associated with the object, with the transition point between objects being the angle halfway between two adjacent objects. Figure 1 shows boxplots of the obtained results. It can be seen that PROJET outperforms DUET in terms of overall separation as measured by SDR, with on average 1.5 dBs of improvement over DUET, and that the performance of PROJET is essentially the same regardless of the angle between sources. With respect to SIR, DUET outperforms PROJET by on average 2 dB. This is to be expected as DUET is based on binary masking, which is known to reduce interference at the expense of increased artifacts, and indeed, PROJET outperforms DUET with respect to SAR by on average 2.5 dB, thereby highlighting this trade-off. With respect to ISR, DUET performs slightly better at low angles, but is outperformed by PROJET at the 30 degree source spacing.

The second set of tests run were blind tests, with only the number of objects to separate provided. The same STFT settings were used as in the oracle tests. The panning set $\overline{\mathcal{P}}$ for testing was chosen to have $L = 30$ equally spaced panning positions spanning the

[2] https://sisec.inria.fr/professionally-produced-music-recordings/

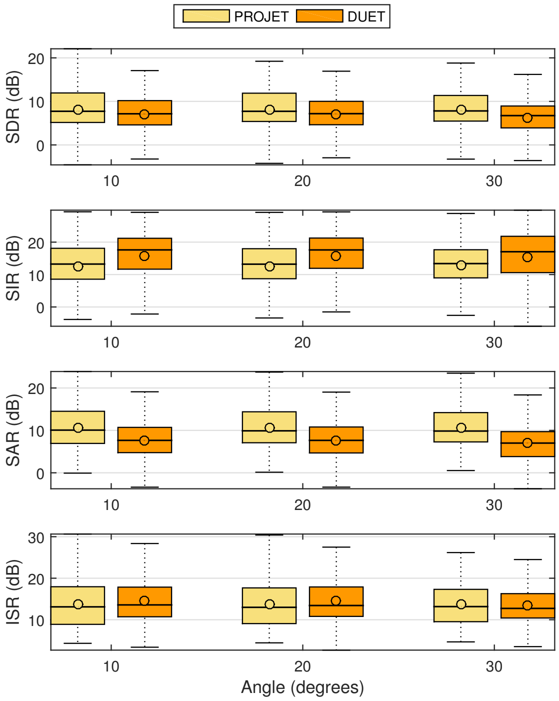

Fig. 1. Oracle separation results for PROJET vs. DUET (circle denotes mean of the data, line denotes median)

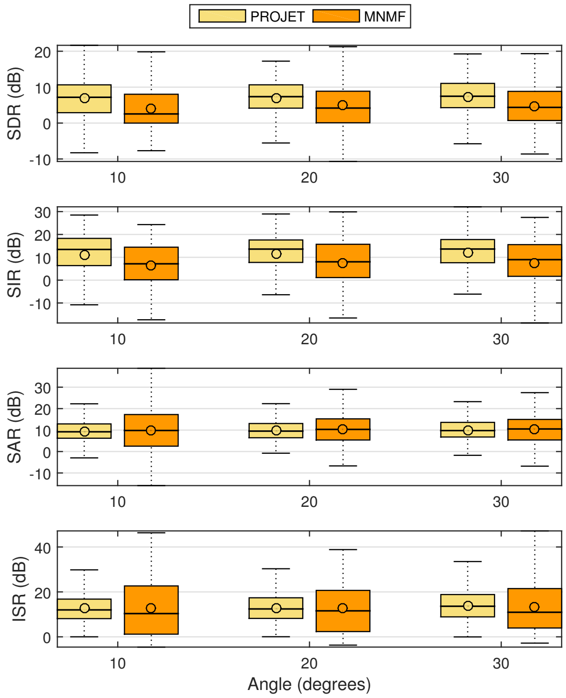

Fig. 2. Blind Separation results for PROJET vs. MNMF (circle denotes mean of the data, line denotes median)

range $[0, \pi/2]$, i.e. $\overline{\mathcal{P}} = \{\theta_1, \ldots, \theta_L\}$, with $\theta_l = [\cos\phi_l, \sin\phi_l]^\top$, where ϕ_l is the angle of the lth pan position. The projection matrix $\mathbf{M}$ was set at $M = 10$ projections, with $\mathbf{M} = \{n_1, \ldots, n_M\}^\top$ and $n_m = [\sin\omega_m, -\cos\omega_m]^\top$, where ω_m is the angle of the mth projection. Again, the projection angles were chosen to equally span the range $[0, \pi/2]$. In order to benchmark the spatial projection separation algorithms against the state of the art, we also tested the separation performance of the multichannel NMF (MNMF) algorithm described in [28], using the implementation of that algorithm as found in the FASST toolbox [29], on the same mixtures.

Figure 2 shows boxplots of results obtained in blind testing. It can clearly be seen that the PROJET method outperforms that of MNMF with respect to SDR, and that PROJET shows only very small decreases in performance with decreasing angle, with less than a 1dB difference in the average between the results for 10 degrees and 30 degrees for PROJET. This demonstrates that PROJET is robust with respect to the angle between the spatial objects. With respect to SIR, PROJET again considerably outperforms MNMF, with also only a small decrease in performance with decreasing angle. It can be observed that MNMF slightly outperforms PROJET with respect to SAR, with the maximum difference being approximately 1dB. Finally, with regards to ISR, PROJET is again better than MNMF. Slight increases in the average results for SAR and ISR can again be observed with increasing angle, but again this improvement is small, demonstrating the robustness of PROJET. On top of these results, PROJET shows much less variability in its performance compared to MNMF, suggesting a higher robustness. Finally, PROJET was also informally tested on a number of real-world commercial recordings. The results can be found online, along with a MATLAB implementation of the algorithm[3].

V. CONCLUSIONS

We have introduced PROJET, a novel technique for the separation of multichannel audio based on the use of spatial projections of the original multichannel mixture. In contrast to existing multichannel separation techniques which operate directly on the original mixture, PROJET projects the multichannel signal onto a range of spatial directions and then operates on this augmented observation set. The underlying mixing model assumes that spatial object images are represented using a weighted sum of independent contributions from all panning directions, and we show how to estimate the model parameters in an effective manner, as well as how to project the obtained results back into the original multichannel domain.

PROJET was then evaluated under oracle conditions, where the source positions were given a-priori to the algorithm, and it was demonstrated to outperform DUET, a well-known source separation technique. The algorithm was also evaluated under blind conditions where the source positions were not provided. In this case it was benchmarked against a well-known multichannel MNMF algorithm and was demonstrated to offer considerably improved and robust performance in comparison to the benchmark. This is remarkable in light of the fact that PROJET imposes no constraints on the spectro-temporal characteristics of the objects to be separated. The algorithm was also informally tested on a number of commercial recordings.

Future work will focus on extending the mixing model to allow complex-valued mixing to incorporate delays between the channels for the source object, and on the incorporation of other constraints on the source spectral densities, such as sparsity. It is also proposed to investigate developing an online version of the algorithm to allow real-time demixing of music.

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

VI. REFERENCES

[1] C. Avendano and J-M. Jot. Frequency domain techniques for stereo to multichannel upmix. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[2] D. Barry, B. Lawlor, and E. Coyle. Real-time sound source separation using azimuth discrimination and resynthesis. In *117th Audio Engineering Society (AES) Convention*, San Francisco, CA, USA, October 2004.

[3] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):191–199, January 2006.

[4] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. In *IEEE International Conference Acoustics Speech Signal Processing (ICASSP)*, pages 613–616, Hong-Kong, April 2003.

[5] D. Betts. Masked positive semi-definite tensor interpolation. In *Latent Variable Analysis and Signal Separation*, pages 446–453. Springer, 2015.

[6] 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, September 2009.

[7] N.Q.K. 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*, 18(7):1830 –1840, Sept. 2010.

[8] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[9] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011.

[10] D. FitzGerald. Vocal separation using nearest neighbours and median filtering. In *23nd IET Irish Signals and Systems Conference*, pages 583–588, Maynooth, 2012.

[11] D. Fitzgerald, A. Liutkus, Z. Rafii, B. Pardo, and L. Daudet. Harmonic/percussive separation using Kernel Additive Modelling. In *Proceedings of the 25th IET Irish Signals and Systems Conference,*, 2014.

[12] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1562–1566. IEEE, 2014.

[13] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. *International Society for Music Information Retrieval (ISMIR)*, 2014.

[14] 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)*, pages 57–60, Kyoto, Japan, March 2012.

[15] M. Kim and P. Smaragdis. Adaptive denoising autoencoders: A finetuning scheme to learn from test mixtures. In *Latent Variable Analysis and Signal Separation*, pages 100–107. Springer, 2015.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 556–562. The MIT Press, April 2001.

[17] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1475–1487, 2007.

[18] A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.

[19] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[20] A. Liutkus, R. Badeau, and G. Richard. Multi-dimensional signal separation with Gaussian processes. In *Proc. of IEEE Conf. on Statistical Signal Processing (SSP2011)*, Nice, France, June 2011.

[21] A. Liutkus, J-L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, Paris, France, July 2013.

[22] A. Liutkus, D. Fitzgerald, and R. Badeau. Cauchy Nonnegative Matrix Factorization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, October 2015.

[23] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16):4298–4310, Aug 2014.

[24] A. Liutkus, T. Olubanjo, E. Moore, and M. Ghovanloo. Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, United States, October 2015.

[25] A. Liutkus, Z. Rafii, B. Pardo, D. Fitzgerald, and L. Daudet. Kernel Spectrogram models for source separation. In *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[26] J. Nikunen and T. Virtanen. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 22(3):727–739, 2014.

[27] A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. Technical report, Inria, 2015.

[28] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563, March 2010.

[29] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, May 2012.

[30] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech & Language Processing*, 21(1):71–82, January 2013.

[31] U. Simsekli, A. Liutkus, and T. Cemgil. Alpha-Stable Matrix Factorization. *IEEE Signal Processing Letters*, page 5, September 2015.

[32] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[33] J. Thiemann and E. Vincent. A fast EM algorithm for Gaussian model-based source separation. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE, 2013.

[34] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot. From blind to guided audio source separation: How models and side information can improve the separation of sound. *IEEE Signal Processing Magazine*, 31(3):107–115, May 2014.

[35] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462 –1469, July 2006.

[36] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1381–1390, July 2013.

[37] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.