PHASE RECOVERY IN NMF FOR AUDIO SOURCE SEPARATION: AN INSIGHTFUL BENCHMARK

Paul Magron *Roland Badeau* *Bertrand David*

Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France
<firstname>.<lastname>@telecom-paristech.fr

ABSTRACT

Nonnegative Matrix Factorization (NMF) is a powerful tool for decomposing mixtures of audio signals in the Time-Frequency (TF) domain. In applications such as source separation, the phase recovery for each extracted component is a major issue since it often leads to audible artifacts. In this paper, we present a methodology for evaluating various NMF-based source separation techniques involving phase reconstruction. For each model considered, a comparison between two approaches (blind separation without prior information and oracle separation with supervised model learning) is performed, in order to inquire about the room for improvement for the estimation methods. Experimental results show that the High Resolution NMF (HRNMF) model is particularly promising, because it is able to take phases and correlations over time into account with a great expressive power.

***Index Terms*—** Nonnegative matrix factorization, audio source separation, phase reconstruction, time-frequency analysis.

1. INTRODUCTION

The problem of separating polyphonic music mixtures into isolated sources has become very popular in the last 15 years. The family of techniques based on nonnegative factorizations, often applied to spectrogram-like representations, has proved to provide a successful and promising framework for this task [1].

NMF, originally introduced as a rank-reduction method [2], approximates a nonnegative data matrix V as a product of two low-rank nonnegative matrices W and H. In audio signal processing, V is often chosen as the magnitude or power spectrogram of the signal, whose factorization is interpretable intuitively: W is a dictionary of spectral templates and H is a temporal activation matrix. Usual alternative versions constrain NMF to enforce properties such as sparsity [1], smoothness or harmonicity [3, 4].

However, when it comes to resynthesize the separated time signals, the recovery of the phase of the corresponding Short-Time Fourier Transform (STFT) is necessary. Even if common practice consists in applying Wiener-like filtering (*e.g* soft masking of the complex-valued STFT of the original mixture), phase recovery is still an open issue, for this kind of filtering does not enforce phase *consistency*. That is, the obtained complex-valued matrix is not the STFT of a time signal. It is worth noting here that consistency can also refer to specific properties of the instantaneous phase of a sinusoidal component [6], but we will hereafter employ *consistency* in the first usage only.

Several extensions to NMF have been introduced, which include a phase model [7, 8, 9], but do not refer to phase consistency. Wiener-like filtering is used for instance in [7]. The separated components are then derived by inverting a TF representation whose phase is that of the STFT of the mixture. This technique ensures phase consistency as long as only one source is active within each TF bin. In order to handle the case of overlapping sources, iterative methods [10, 5] minimize the inconsistency of the reconstructed TF representation. On the other hand, some NMF-inspired models combine phase modeling and spectrogram factorization. The complex NMF model introduced in [8] was later improved by means of consistency constraints [11]. More recently, High Resolution NMF (HRNMF) has been introduced in [12]. It models a TF mixture as a sum of autoregressive components in the TF domain, thus dealing explicitly with a phase model which takes time dependencies from one TF bin to another into account.

All the above-mentioned models are suitable for blind source separation of audio signals since they factorize the spectrogram, reconstruct the phase and enforce its consistency. In this paper, we propose a methodology for assessing their potential and performance. This methodology is based on a comparison between two approaches: blind separation without prior information and oracle separation with supervised model learning. This comparison is performed in order to inquire about the room for improvement for the estimation methods. Algorithms are evaluated with BSS EVAL [14], a set of objective criteria dedicated to measuring source separation quality. Finally, the algorithms are tested on different data types. Since difficulties often arise when sources overlap in the TF domain, a particular emphasis has been put on the related tests.

The paper is organized as follows. Section 2 presents the considered NMF-based algorithms. Section 3 describes the methodology of this benchmark, through objectives and protocol. Section 4 presents results and interpretations of the tests conducted on a variety of data, and Section 5 draws some concluding remarks.

2. NMF-BASED SOURCE SEPARATION ALGORITHMS

2.1. NMF main principle

The NMF problem is expressed as follows: given a matrix V of dimensions $F \times T$ with nonnegative entries, find a factorization $V \approx WH$ where W and H are nonnegative matrices of dimensions $F \times K$ and $K \times T$. In order to reduce the dimension of data, K is chosen such that $K(F+T) \ll FT$. In audio source separation, V is generally the magnitude or the power spectrogram of a TF representation X of a mixture signal (most of the time an STFT). One can interpret W as a dictionary of spectral templates and H as a matrix of temporal activations. If W_k denotes the k-th column of W and H_k denotes the k-th line of H, then $V_k = W_k H_k$ is

This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02).

the magnitude or power spectrogram of the component indexed by k and $\hat{V} = \sum_{k=1}^{K} V_k$. Note that this result expresses an additivity property of spectrograms, which only approximately holds when sources overlap in the TF domain. This factorization is generally obtained by minimizing a cost function $D(V, \hat{V})$. Popular choices for D are the Euclidean distance, Kullback-Leibler divergence [2] and Itakura-Saito divergence [7]. Our benchmark uses multiplicative update rules (MUR) [15], in order to estimate a regular NMF with the Kullback-Leibler divergence (KLNMF).

2.2. Phase reconstruction

Estimating a complex TF representation X_k of a separated source by applying Wiener filtering [7] consists in computing:

$$X_k = \frac{W_k H_k}{\sum_{l=1}^{K} W_l H_l} X = \frac{V_k}{\hat{V}} X. \tag{1}$$

This method will be referred to as **NMF-Wiener**.

Alternatively, a regular NMF can be combined with a phase reconstruction algorithm based on the minimization of a cost function which penalizes inconsistency. The **Griffin-Lim** algorithm [10] is an iterative method described in Eq. (2) for recursively estimating the k-th component. For each iteration i:

$$X_k^i \longrightarrow Y_k^{i+1} = \mathcal{F}(X_k^i) \longrightarrow X_k^{i+1} = \frac{V_k}{|Y_k^{i+1}|} Y_k^{i+1} \tag{2}$$

where $\mathcal{F} = STFT \circ STFT^{-1}$. It has been shown to make the Euclidean distance between V_k and $|Y_k^i|$ decrease over iterations. This method will be referred to as **NMF-GL**.

The **LeRoux** algorithm [5] consists in explicitly calculating and minimizing the inconsistency defined as the Euclidean distance between X and $\mathcal{F}(X)$. Iterative optimization techniques then lead to update rules for the phase of the reconstructed source in the TF domain. This method will be referred to as **NMF-LR**.

In **NMF-GL** and **NMF-LR**, the magnitude is constant over iterations. The user can force it to be equal to V_k, obtained from the NMF. However, experiments show that initializing **Griffin-Lim** and **LeRoux** algorithms with the magnitude of X_k in Eq. (1) provides better results.

2.3. Complex NMF

Complex NMF [11] consists in factorizing a magnitude spectrogram while reconstructing a phase field for each source. The mixture TF representation is modeled as follows: for each TF bin (f, t),

$$X(f,t) = \sum_{k=1}^{K} X_k(f,t) = \sum_{k=1}^{K} W_k(f) H_k(t) e^{j\phi_k(f,t)}. \tag{3}$$

This method will be referred to as **CNMF**. An explicit phase consistency constraint [11] leads to a consistent TF representation. It will be referred to as **CNMF-LR**. The main advantage of this technique is to jointly estimate the magnitude and phase parameters, instead of deriving the phase from an imposed magnitude (as in **NMF-LR**).

2.4. High Resolution NMF

More recently, the HRNMF model has been introduced in [12]. It consists in modeling each frequency band of the TF representation by means of auto-regressive filtering. This technique naturally captures phase relationships and dependencies over time.

The mixture TF representation is modeled as follows:

$$X(f,t) = n(f,t) + \sum_{k=1}^{K} X_k(f,t) \tag{4}$$

where $n(f,t)$ is a white Gaussian noise. Each source $X_k(f,t)$ is obtained by autoregressive filtering of a non-stationary signal $b_k(f,t)$:

$$X_k(f,t) = b_k(f,t) + \sum_{p=1}^{P(k,f)} a_p(k,f) X_k(f, t-p) \tag{5}$$

where $P(k,f)$ is the order of the autoregressive filter for source k and frequency f, of coefficients $a_p(k,f)$. Finally, $b_k(f,t)$ follows a centered normal distribution of variance $V_k(f,t)$ such that $V_k = W_k H_k$, and all $b_k(f,t)$ are independent.

The model parameters can be estimated either by a regular EM algorithm, which is computationally costly, or by a variational Bayesian EM (VBEM) algorithm, allowing faster computation without significant quality loss. We conduct an experience to estimate the best HRNMF initialization and algorithm in Section 4.1. Note that recently, HRNMF has been extended to multichannel signals and convolutive mixtures, and is now able to model correlations over frequency [13].

3. METHODOLOGY

3.1. Objectives

In order to assess audio source separation quality, we use BSS EVAL [14], a set of objective criteria dedicated to this purpose. From the original sources x_k and the estimated sources $\hat{x}_k$, $k = 1, ..., K$, BSS EVAL computes various energy ratios: the SIR (signal to interference ratio) that measures the rejection of interferences, the SAR (signal to artifact ratio) for the rejection of artifacts, and the SDR (signal to distortion ratio) for the global quality.

In order to evaluate the room for improvement for these techniques, we compare the results obtained with a blind approach and an oracle approach. The blind approach consists in estimating the models directly from the mixture without using any prior information about the isolated sources. The oracle approach consists in evaluating, for each technique, the best performance possible: the parameters are learned from each isolated source. A comparison between those two approaches informs us about the opportunities for further enhancement of these methods.

Since phase recovery is a major issue in source separation, it is interesting to evaluate if the consistency constraint used in various methods (**NMF-GL**, **NMF-LR** and **CNMF-LR**) is related to audio quality.

Finally, we want to evaluate the expressive power of the models, that is to say their ability to represent a variety of signals observed in music analysis. We use both synthetic and real data, with and without TF overlap.

3.2. Datasets and protocol

We perform audio source separation on several datasets. Firstly, we synthesize 60 mixtures of two harmonic signals ($K = 2$) which consist of damped sinusoids whose amplitude, origin phase, frequency and damping coefficients are randomly-defined, and a 60 dB additive white noise. The damping coefficient is the same for all harmonics. One set of 30 mixtures does not include TF overlap while the other one does (see an example in Figure 1).

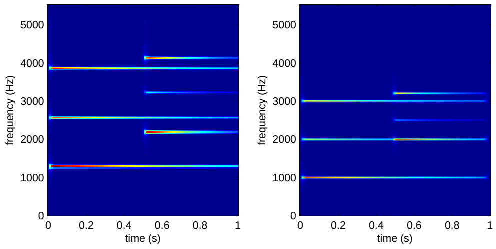

Fig. 1. Synthetic data spectrograms: without TF overlap (left) and with TF overlap (right)

The MAPS (MIDI Aligned Piano Sounds) dataset [16] provides various data to design mixtures of real piano sounds. For the tests on real data, we consider 30 mixtures of two piano notes, selected randomly in the MAPS database. We also enforce TF overlap in some data. Finally, we tested the benchmark on a 1.57 second-long MIDI audio excerpt. It is composed of several occurrences of three bass notes and one guitar chord ($K = 4$).

The data is sampled at $F_s = 11025$ Hz. It is important to note that HRNMF involves more diverse parameters than the regular NMF model. Indeed, correlations across time are taken into account by means of autoregressive filtering in each frequency subband of order $P(k, f)$. In our experiments, $P(k, f)$ was set to 1 for all (k, f). This means that the HRNMF model uses twice as many spectral parameters (W and a) as regular NMF (W only). In order to make a fair comparison, it is interesting to compare both models with the same total number of parameters. The STFT is thus calculated with a 512 sample-long normalized Hann window with 75% overlap for testing **CNMF**, **CNMF-LR** and **HRNMF** models, and with a 1024 sample-long window for testing **NMF-Wiener**, **NMF-GL** and **NMF-LR** models[1].

For both blind and oracle approaches, KLNMF and CNMF are estimated with 30 iterations of MUR algorithms, and phase reconstruction algorithms involve 50 iterations. HRNMF is initialized with a 30-iterations KLNMF and estimated with 30 iterations of the VBEM algorithm for the blind approach, and 10 iterations of the VBEM algorithm for each source learning (oracle approach). We compute BSS EVAL scores on the different mixtures (for synthetic and real data) and on 30 different initializations (for MIDI data).

The numbers of iterations are chosen such that the performance is not further improved beyond. Energy ratios are expressed in dB.

4. EXPERIMENTAL RESULTS

4.1. HRNMF initialization and estimation algorithm

HRNMF requires a well-chosen initialization to produce meaningful results (likely because of the great number of local minima of the cost function). The data to be processed is a mixture of real notes without frequency overlap. We consider the regular EM algorithm [12] and the VBEM algorithm [13]. Initializations can be random, KLNMF [15] or Itakura-Saito NMF (ISNMF, [7]), computed by means of MUR algorithms.

Table 1. Influence of HRNMF initialization and algorithm on source separation performance

Algorithm	Initialization	SDR	SIR	SAR	Time (s)
EM	Random	5.3	6.4	14.3	379
	ISNMF	15.0	21.2	17.0	376
	KLNMF	17.0	22.2	18.7	377
VBEM	Random	1.4	2.8	11.1	1.03
	ISNMF	**16.9**	**25.3**	**17.7**	**0.95**
	KLNMF	**16.9**	**24.5**	**17.8**	**0.89**

Results are presented in Table 1 (the best performance is highlighted in bold font). We observe that initializing HRNMF with a prior NMF algorithm provides significantly better results than applying the EM or VBEM algorithm directly on random parameters. The choice of the NMF (KL or IS) does not influence much the results. We also see that the VBEM algorithm provides results similar to the EM algorithm, with a dramatic reduction of the computational cost. We will thus use the VBEM algorithm with KLNMF initialization for our benchmark.

4.2. Synthetic data

Benchmark results for synthetic harmonics are presented in Figure 2. Box-plots compile data for blind approach. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles, and whiskers indicating the minimum and maximum values. The triangles and stars indicate the performance of the oracle approach.

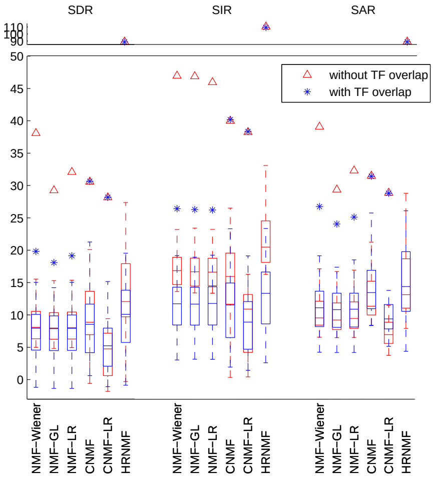

Fig. 2. Synthetic harmonics separation performance

[1] Note that the total number of parameters involved in the CNMF model is higher than the dimension of the TF data itself, because all phase coefficients are free. However, even if comparing CNMF with NMF or HRNMF using the same total number of parameters is not possible, the results in Section 4 will show that CNMF is most often outperformed by the other models.

These results show that **Griffin-Lim** and **LeRoux** phase reconstruction algorithms provide poor results in terms of audio quality. While consistency is increased in **NMF-GL** and **NMF-LR**, those methods lead to a decrease of the SDR and SAR scores compared to **NMF-Wiener**. Enforcing the magnitude to be constant over iterations seems too constraining to increase audio quality. **CNMF-LR** is supposed to be a response to the aforementioned problem, but it does not provide better results than **NMF-LR**. It also requires much more memory for storing the phase field of each source. We also note that **CNMF** provides better results than **CNMF-LR**, confirming that consistency may not be a good criterion for audio quality. Results generally drop when TF bins overlap, but not in terms of SAR: artifact rejection seems globally increased when overlap occurs in the blind benchmark.

Finally, blind separation with the **HRNMF** model provides slightly better results than the other models (except when overlap occurs in the TF domain: **CNMF** and **HRNMF** then lead to a similar SAR median). This model also provide the best performance in the oracle benchmark. **NMF-Wiener** is the fastest algorithm (40 ms), the other models are estimated in approximately 1.5 seconds. Similar computation times are obtained for real data.

The tests performed on synthetic harmonics with vibratos (that cannot be presented here because of a lack of room) lead to similar results: the **HRNMF** model significantly outperforms the other models in the oracle approach, demonstrating its ability to accurately represent a variety of signals.

4.3. Piano notes mixtures

Benchmark results for piano notes mixtures are presented in Figure 3. We note that the algorithms do not perform worse than in the synthetic data case. The blind benchmark shows that **HRNMF** results are similar to the other algorithms (or slightly better), but the oracle results confirm that it is the best model available in terms of potential for source separation. **NMF-Wiener** is also interesting, because it provides a fast and relatively accurate audio source separation. The analysis of the results for each mixture reveals that the quality of **NMF-Wiener** is slightly worse than **HRNMF** when there are overlapping TF bins.

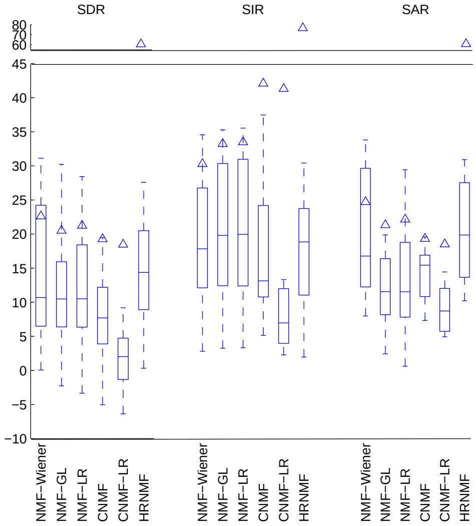

Fig. 3. Piano notes mixtures separation performance

4.4. MIDI song

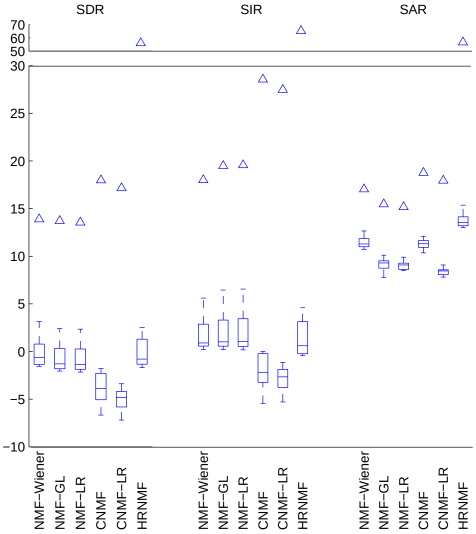

Fig. 4. MIDI audio excerpt separation performance

Figure 4 presents the results obtained with a realistic MIDI audio excerpt. It shows a dramatic reduction of blind source separation quality compared to the previous tests. This signal seems too complex to obtain an efficient factorization. Then, **HRNMF** estimation does not improve the result of the initial KLNMF. However, the oracle approach still shows that this method has a higher potential than the other models for this application. **NMF-Wiener** is computed in 60 ms and the others models are estimated in 3 to 4 seconds.

5. CONCLUSION

This benchmark presents HRNMF as a very promising model in terms of source separation quality. It is able to take both phases and correlations over time into account, and models a variety of signals frequently observed in music analysis. In particular, the oracle approach showed that HRNMF is likely to be particularly efficient when source separation is partially informed. The other models and algorithms appear to be less appealing for source separation, because sources often overlap in the TF domain, a common phenomenon in music. More generally, algorithms that take correlation over time and frequencies into account with a great expressive power should be considered with particular attention. Consistency has also been shown not to be an appropriate criterion for audio quality. The datasets and procedure described in this work can be a good basis for further evaluation of the potential of source separation models.

Besides, the experiments show that the VBEM algorithm used for estimating HRNMF is highly sensitive to initialization. Semi-supervised learning or prior information about the sources, such as harmonicity, sparsity or temporal smoothness should be introduced in order to address this issue. Alternative estimation methods, more robust and less sensitive to initialization, could be implemented in future research. Bayesian methods such as Markov Chain Monte Carlo (MCMC) and message passing algorithms might be an option. Alternatively, the algebraic principles used in High Resolution methods (such as the ESPRIT algorithm [17]) could also be exploited in order to address this estimation problem.

6. REFERENCES

[1] Paris Smaragdis and Judith C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2003, pp. 177–180.

[2] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[3] Nancy Bertin, Roland Badeau, and Emmanuel Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, March 2010.

[4] François Rigaud, Bertrand David, and Laurent Daudet, "A parametric model and estimation techniques for the inharmonicity and tuning of the piano," *Journal of the Acoustical Society of America (JASA)*, vol. 133, no. 5, pp. 3107–3118, May 2013.

[5] Jonathan Le Roux, Nobutaka Ono, and Shigeki Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)*, Brisbane, Australia, September 2008, pp. 23–28.

[6] Jean Laroche and Mark Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[7] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, March 2009.

[8] Hirokazu Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 3437–3440.

[9] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, March 2010.

[10] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[11] Jonathan Le Roux, Hirokazu Kameoka, Emmanuel Vincent, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, "Complex NMF under spectrogram consistency constraints," in *Proc. Acoustical Society of Japan Autumn Meeting*, Hukushima, Japan, September 2009, number 2-4-5.

[12] Roland Badeau, "Gaussian modeling of mixtures of nonstationary signals in the time-frequency domain (HR-NMF)," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2011, pp. 253–256.

[13] Roland Badeau and Mark D. Plumbley, "Multichannel high resolution NMF for modelling convolutive mixtures or nonstationary signals in the time-frequency domain," *IEEE Transactions on Audio Speech and Language Processing*, vol. 22, no. 11, November 2014.

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

[15] Daniel D. Lee and H. Sebastian Seung, "Algorithms for nonnegative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T.K. Leen, T.G. Dietterich, and V. Tresp, Eds., pp. 556–562. MIT Press, 2001.

[16] Valentin Emiya, Nancy Bertin, Bertrand David, and Roland Badeau, "MAPS - A piano database for multipitch estimation and automatic transcription of music," Tech. Rep. 2010D017, Télécom ParisTech, Paris, France, July 2010.

[17] Yingbo Hua, Alex B. Gershman, and Qi Cheng, *High-resolution and robust signal processing*, Signal processing and communications. Marcel Dekker, New York, 2004.