An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement

Joachim Thiemann, Emmanuel Vincent

AN EXPERIMENTAL COMPARISON OF SOURCE SEPARATION AND BEAMFORMING TECHNIQUES FOR MICROPHONE ARRAY SIGNAL ENHANCEMENT

*Joachim Thiemann**

Carl-von-Ossietzky University Oldenburg
D-26129 Oldenburg, Germany
joachim.thiemann@uni-oldenburg.de

Emmanuel Vincent

Inria, Centre de Nancy - Grand Est
F-54600 Villers-lès-Nancy, France
emmanuel.vincent@inria.fr

ABSTRACT

We consider the problem of separating one or more speech signals from a noisy background. Although blind source separation (BSS) and beamforming techniques have both been exploited in this context, the former have typically been applied to small microphone arrays and the latter to larger arrays. In this paper, we provide an experimental comparison of some established beamforming and post-filtering techniques on the one hand and modern BSS techniques involving advanced spectral models on the other hand. We analyze the results as a function of the number of microphones, the number of speakers and the input Signal-to-Noise Ratio (iSNR) w.r.t. multichannel real-world environmental noise recordings. The results of the comparison show that, provided that a suitable post-filter or spectral model is chosen, beamforming performs similar to BSS on average in the single-speaker case while in the two-speaker case BSS exceeds beamformer performance. Crucially, this claim holds independently of the number of microphones.

Index Terms: Source separation, beamforming, FASST, MVDR, post-filtering, evaluation

1. INTRODUCTION

In the signal processing community, blind source separation (BSS) [1–4] and beamforming [5, 6] are typically regarded as distinct families of techniques. Although they share the goal of isolating a target signal from a mixed signal, beamforming relies on the spatial location of the target source, while BSS relies on more general statistical properties of the source signals.

With few exceptions, e.g., [7, 8], BSS has typically been applied to noiseless single-channel or two-channel recordings in the past. In contrast, beamforming has often been employed with larger microphone arrays in noisy conditions, with each increase in the number of channels improving the ability to enhance the target. Recently, however, the frontier between these application domains has blurred and both families of techniques have been tried on mixed setups, e.g., two-channel mixtures of speech and real-world noise [9–11].

This raises the question of the relative performance of BSS and beamforming as a function of the number of microphones, the amount of noise, and other acoustic conditions. A few comparisons have been made that focus on the separation of determined noiseless mixtures of speech via early Independent Component Analysis (ICA)-based techniques [12, 13] or separation of recordings with a slightly larger but fixed number of channels [14]. Yet, to the best of our knowledge, no experiments have been performed to compare modern beamforming techniques involving post-filtering [15–17] with state-of-the-art BSS techniques including advanced spectral modelling [18–20] for different numbers of channels and acoustic conditions.

This paper aims to provide such an experimental comparison for the extraction of one or two target sources mixed with real-world environmental noise recorded by 2 to 8 microphones. We consider the case where the spatial location of the target source(s) is precisely known. For beamforming, this is essential for steering the beam in the correct direction. In BSS, this is used to initialize the estimation process so that separation can be performed without prior knowledge about the spectra of the speech sources. However, we show that the addition of constraints over these spectra improves performance over the use of spatial information alone. Such a setup is sometimes called Semi-Blind Source Separation (SBSS) [10].

In Section 2, we describe the signal models and the estimation algorithms underlying the chosen beamforming and BSS algorithms. In Section 3, we present the experimental setup and the resulting source separation performance. We conclude in Section 4.

2. SOURCE SEPARATION AND BEAMFORMING

We consider the problem of separating the signals of J sound sources from a multichannel signal $\mathbf{x}(t)$ recorded by M microphones. Using the Short-Time Fourier Transform (STFT), the $M \times 1$ vector $\mathbf{x}_{fn}$ of mixture STFT coefficients in time frame n and frequency bin f is classically expressed as [4]

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{j,fn}. \quad (1)$$

where $\mathbf{y}_{j,fn}$ denotes the contribution of the jth source to the mixture. For point sources, $\mathbf{y}_{j,fn}$ can further be expressed in terms of the product of a single-channel source signal $s_{j,fn}$ and some mixing coefficients [4].

2.1. Beamforming and postfiltering

2.1.1. The MVDR beamformer

Beamforming isolates a target source j by making a set of microphones act as a single receptor that targets a specific direction in space, while suppressing signals from other directions.

Given the spatial location of the target, a steering vector

$$\mathbf{d}_{j,f} = [1\ e^{if(\tau_{j2}-\tau_{j1})} \dots e^{if(\tau_{jM}-\tau_{j1})}]^H \quad (2)$$

*The work presented here was performed while Joachim Thiemann was at CNRS, IRISA-UMR6074, in Rennes, France.

is computed that accounts for the Time Differences Of Arrival (TDOA) of a given source j between the microphones of the array, where $\tau_{jm}, m = 1, \ldots, M$ is the time it takes for the sound to travel from source j to microphone m on a direct path. A set of weights $\mathbf{w}_{j,fn}^H$ are then computed from $\mathbf{x}_{fn}$ and $\mathbf{d}_{j,f}$ and applied to the mixture such that

$$\widehat{s}_{j,fn} = \mathbf{w}_{j,fn}^H \mathbf{x}_{fn}. \tag{3}$$

In the following, we consider the Minimum Variance Distortionless Response (MVDR) beamformer [21], which minimizes the energy of the interfering sources and noise under the constraint of unit response in the direction pointed by the steering vector. The weights are derived from $\mathbf{d}_{j,f}$ and from an estimate of the covariance matrix of the mixture $\mathbf{R}_{\mathbf{x},fn} = \widehat{\mathbb{E}}[\mathbf{x}_{fn}\mathbf{x}_{fn}^H]$ as

$$\mathbf{w}_{j,fn} = \frac{(\mathbf{R}_{\mathbf{x},fn} + \alpha\mathbf{I})^{-1}\mathbf{d}_{j,f}}{\mathbf{d}_{j,f}^H(\mathbf{R}_{\mathbf{x},fn} + \alpha\mathbf{I})^{-1}\mathbf{d}_{j,f}}, \tag{4}$$

where $\mathbf{I}$ is the $M \times M$ identity matrix and the diagonal loading factor α prevents instabilities [22].

2.1.2. Postfiltering for beamformers

Since beamforming is only capable of spatial filtering, an additional single-channel post-filter is often applied which modifies (3) to

$$\widehat{s}_{j,fn}^{\text{post}} = p_{nf}\mathbf{w}_{j,fn}^H \mathbf{x}_{fn}, \tag{5}$$

where the gain factor p_{nf} is real-valued and typically limited to $0 \leq p_{nf} \leq 1$ [23]. Such post-filtering provides a fair comparison of beamforming and source separation algorithms which also utilize some form of spectral filtering. The combination of MVDR beamforming with the ideal post-filter equal to the ratio of the (unknown) short-term power spectrum of the target and that of the beamformer output has been shown to estimate the target source in the Minimum Mean Square Error (MMSE) sense [15].

To provide a contrasted comparison, we evaluate three very diverse choices:

- no post-filtering (**MVDR no post**)
- Zelinski's post-filter [16], which approximates the ideal post-filter by estimating the noise power spectrum from the off-diagonal values of the observed cross-channel covariances under a diffuse noise assumption (**MVDR Zelinski**),
- a state-of-the-art spectral filter based on the optimally-modified log-spectral amplitude (OM-LSA) speech estimator with improved minima controlled recursive averaging (IMCRA) noise estimation [24, 25] (**MVDR OMLSA**).

Alternative post-filters have since been proposed [15, 17, 23] which we evaluated in preliminary experiments. McCowan's post-filter performed more poorly than Zelinski's in the considered acoustic setting. Ito's post-filter achieved a significant gain over Zelinski's, but it is less established and did not reach the performance of OM-LSA/IMCRA. Due to space restrictions, we only provide the results of OM-LSA/IMCRA and Zelinski's post-filter in the following experiments.

2.2. Variance model-based source separation

In contrast to beamforming, BSS estimates all sources jointly. A number of techniques have been proposed that are based either on spatial or on spectral models of the sources [2]. We select here the state-of-the-art flexible variance model in [20] which, together with [26], is one of the few models able to jointly exploit spatial and spectral cues. Its advantage compared to [26] is that it can impose constraints on the spectra of the sources while learning them from the mixture signal.

This method relies on the local Gaussian model

$$\mathbf{y}_{j,fn} \sim \mathcal{N}(\mathbf{0}, v_{j,fn}\mathbf{R}_{j,f}), \tag{6}$$

where $v_{j,fn}$ represents the short-term power spectrum of the jth source and $\mathbf{R}_{j,f}$ its spatial covariance matrix. In order to provide a contrasted comparison again, we consider two different models for the power spectra of the speech sources:

- either $v_{j,fn}$ is unconstrained (**BSS unconst.**),
- or it is modelled via harmonic Nonnegative Matrix Factorization (NMF) [27] (**BSS harmonic**)

$$v_{j,fn} = \sum_{k=1}^{K}\sum_{l=1}^{L} w_{j,fl}u_{j,lk}p_{j,kn} \tag{7}$$

where $w_{j,fl}$ are fixed narrowband spectral patterns with either voiced or unvoiced fine structure and $u_{j,lk}$ and $p_{j,kn}$ are spectral envelope and time activation coefficients to be estimated from the observed signal.

Analogous to the steering vector for the beamforming methods, the spatial covariance matrices of the speech sources are computed from $\mathbf{d}_{j,f}$, as described in more detail in the following section. For the noise source, the spatial covariance matrix is initialized with the identity matrix. All spectral parameters are randomly initialized. The noise spatial covariance matrix and all spectral parameters are estimated in the Maximum Likelihood (ML) sense via an iterative Expectation-Maximization (EM) algorithm. The final estimates of the sources are computed in the MMSE sense via the Wiener filter

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,f}\left(\sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}\right)^{-1}\mathbf{x}_{fn}. \tag{8}$$

3. COMPARISON FRAMEWORK AND RESULTS

3.1. Acoustic setup

Our experimental setup is intended to simulate a somewhat realistic usage scenario. One or two target speech signals are convolved by room impulse responses simulated via the source image technique and added to field recordings of environmental noise. The microphone array setup, the room dimensions and the reverberation time in the simulated environment mirrors the physical array with which the background noise was recorded. The use of simulated room impulse responses is widespread in source separation evaluations and has been shown to result in similar separation performance as actual microphone recordings [14].

For the purpose of this evaluation, we collected a database of 16-channel noise recordings (DEMAND) [28] from 18 real-world environments. This study uses a subset of the original recorded channels corresponding to a planar array of 8 microphones in three staggered rows such that the distance from each microphone to its immediate neighbour is 5 cm. The simulated target signal sources are placed one meter away from microphone one of the array, separated by 45 degrees, in the same plane as the array, 1.5 m off the ground (Fig. 1). To gauge how the algorithms scale with the number of available channels, proper subsets of the array are also evaluated.

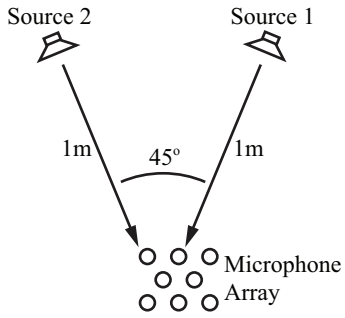

Fig. 1. Relative position of the simulated sources to the microphone array and geometry of the microphones within the array (not to scale). The microphone array and sources are located 1.5 m off the ground.

Environment	Principal noise character	Room sim. T_{60}
NFIELD	stationary	open
NPARK	periodic	open
OHALLWAY	stationary	0.16 s
OOFFICE	periodic	0.16 s
PRESTO	babble	0.76 s
PSTATION	stationary	0.76 s
SPSQUARE	mixed	open
STRAFFIC	stationary	open
TCAR	stationary	0.04 s
TMETRO	mixed	0.05 s

Table 1. Summary of noise recordings used in this study, the characteristics of the noises and the reverberation time of the corresponding room simulation. "Periodic" indicates repeated changes (e.g., typing noises), "mixed" indicates stationary noise interspersed with brief different noises. For the room simulation characteristics, "open" areas are characterized by having direct path reflections only, without significant diffuse reverberation.

In the two-channel case, only the two microphones of the center row are used. In the four-channel case, the four microphones forming a diamond in the center of the array are used.

Of the database described in [28], we use 5 s excerpts from the noise recordings listed in Table 1. The simulated sources are mixed with the noise recordings to have an input SNR (iSNR) of -6, 0, 6, 12 and 18 dB for all channels combined. In the case of two targets, the iSNR is measured between the sum of the target sources and the noise recording. All signals are sampled at 16 kHz.

3.2. Algorithm parameters

A STFT window size of 1024 samples is used. Based on preliminary experiments, the other parameters of the algorithms are set as follows.

Regarding the beamforming algorithms, the MVDR beamformer estimates $\mathbf{R}_{\mathbf{x},fn}$ by averaging $\mathbf{x}_{fn}\mathbf{x}_{fn}^H$ temporally over all frames, with the diagonal loading factor set to $\alpha = 0.01 \cdot \mathrm{tr}(\mathbf{R}_{\mathbf{x},fn})$. For Zelinski's post-filter, the noise estimate uses a moving average over 25 frames centered on the current frame, similar to [23]. OM-LSA/IMCRA relies on the reference implementation by its author[1].

[1]http://webee.technion.ac.il/Sites/People/IsraelCohen/Download/omlsa.m

Regarding the BSS algorithms, we use the FASST toolbox[2] up to some modifications required to account for the increased number of channels compared to the reference two-channel implementation. In all experiments, the noise is modelled as a single source with a full-rank adaptive spatial covariance matrix and spectral parameters constrained to NMF with 16 components. For the target sources, the spatial covariance matrices are fixed to $\mathbf{R}_{j,f} = \mathbf{d}_{j,f}\mathbf{d}_{j,f}^H + \sigma^2\mathbf{\Omega}_f$, where $\mathbf{\Omega}_f$ is the theoretical covariance matrix of a diffuse noise, as detailed in [29]. The reverberant-to-direct ratio σ^2 is set to a small value (0.001) since the algorithm is not given any prior information about the reverberant condition of the signal. For the BSS harmonic experiments, the harmonic constraint (7) is implemented by fixing the narrowband spectra $w_{j,fl}$ as in [20]. To ensure the EM algorithm has fully converged, 100 iterations per channel are used resulting in computation complexity several magnitudes higher than the beamforming methods. Typically a much lower number of iterations can be used with little loss of performance, but a detailed analysis of complexity is outside the scope of the current study.

3.3. Evaluation of separation performance

Rejection of interfering sources and noise was evaluated in terms of the Signal-to-Interference Ratio (SIR) as computed by BSS Eval [30], where the background noise and the second source (if present) are both considered as interfering sources[3]. We compare the various algorithms in terms of the SIR improvement (SIRI) with respect to an appropriate "null" algorithm, a metric similar to the noise reduction rate in [13]. In the beamforming case this "null" algorithm simply outputs the signal from the first microphone (located at the left centre of the array). In the source separation case, the "null" algorithm produces the input signals unmodified. Both "null" algorithms provide very similar SIR. This provides a consistent baseline for all algorithms in both the one and two target source cases.

To evaluate signal distortion, the Mel-cepstral distance (MCD) [31] measure is used, defined as

$$\mathrm{MCD}_{j,n} = \sqrt{\sum_{i=1}^{12}(\mathrm{MC}_{\hat{s}_{j,in}} - \mathrm{MC}_{s_{j,in}})^2}, \qquad (9)$$

where $\mathrm{MC}_{\hat{s}_{j,in}}$ and $\mathrm{MC}_{s_{j,in}}$ represent the ith Mel-frequency cepstral coefficient of the estimated and clean speech source j in time frame n respectively, computed using [32]. In the case of BSS, only the first channel of the estimates $\hat{\mathbf{y}}_{j,fn}$ is used to compute the MCD. The per-signal value is averaged over all time frames and all sources. The overall scores shown in the figures below are obtained by averaging SIRI and MCD respectively over all 10 noise environments.

3.4. SIRI for a single target in a noise field

Fig. 2a shows the SIRI resulting from the various algorithms in the single target scenario for a 4-channel array at different iSNRs. We find in this scenario that the best performance can be achieved with either MVDR OMLSA or BSS harmonic. BSS unconst. performs favourably compared with MVDR and MVDR Zelinski, but cannot approach the methods exploiting spectral cues.

In Fig. 2b we show how the SIRI varies with respect to the number of channels, for an iSNR of 0dB. We find that the BSS algo-

[2]http://bass-db.gforge.inria.fr/fasst/

[3]This metric is also sometimes termed Signal-to-Interference-plus-Noise Ratio (SINR).

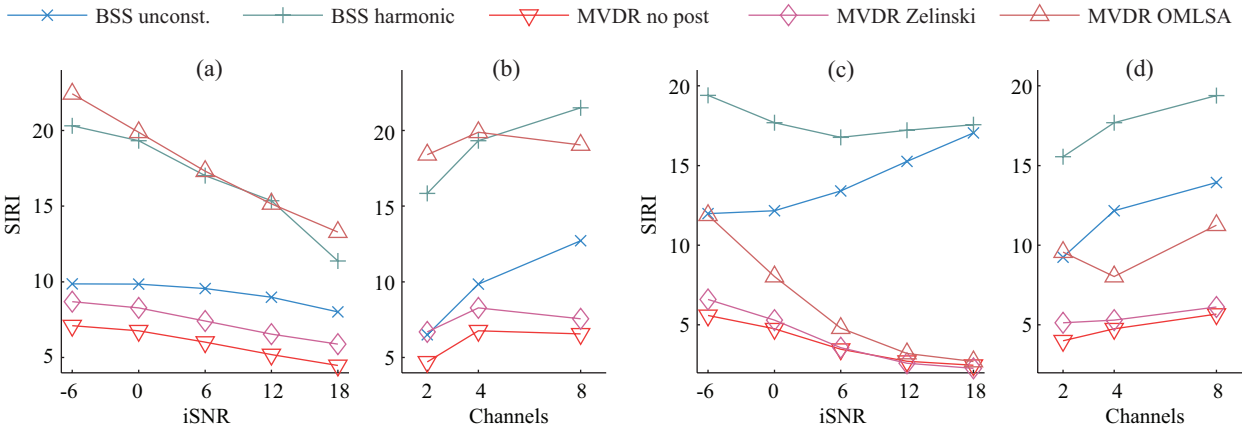

Fig. 2. SIRI for one (left) or two (right) target sources. Plots (a) and (c) show the SIRI for a 4-channel array at different iSNRs. Plots (b) and (d) show the effect of changing the number of channels for a fixed iSNR of 0 dB.

rithms exploit the additional channels very effectively. In contrast, the beamformer performance levels off past 4 channels.

3.5. SIRI for two simultaneous targets in a noise field

In the two-target scenario, the BSS algorithms perform significantly better than beamforming, as shown in Figs. 2c and 2d. Again, additional channels confer a great benefit to BSS. Since the beamforming algorithms only consider one direction (of one target), the other target is considered part of the background noise and cannot be suppressed as effectively. The performance of MVDR OMLSA is drastically reduced when compared to the single target case, due to the non-stationary nature of the interference. Overall, the best results are achieved by BSS harmonic for all iSNRs and all numbers of channels.

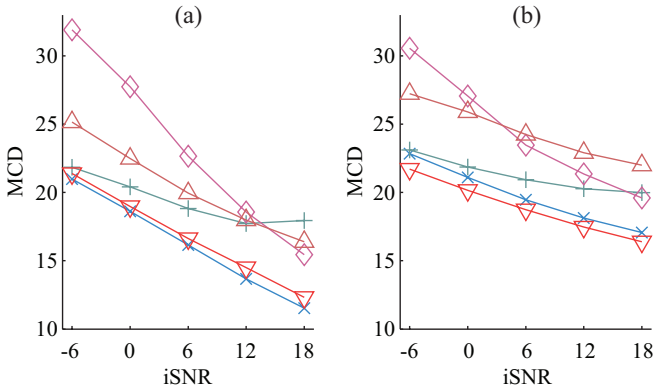

Fig. 3. Average MCD for two-channel recordings. For labels, see Fig. 2.

3.6. Signal distortion

The MCD for the resulting speech signals is shown in Figs. 3a and 3b for the two-channel case. With the exception of MVDR Zelinski, the results do not vary significantly as the number of channels is increased. We find that the distortion incurred by BSS unconst. and the plain MVDR beamformer is about equivalent for both the one- and two-target case. BSS harmonic results in larger distortion especially at high iSNRs since speech sources fit the imposed voiced/unvoiced spectral constraints only to a certain extent. MVDR OMLSA adds further signal distortion due to the aggressive removal of the noise that approaches a binary mask, but can also add musical noise like artifacts. Finally, MVDR Zelinski performed surprisingly poorly, which we attribute to the strong low-pass filtering effect that is observed in the resulting signal for most of the environments. When using 4 or 8 channels, the low-pass effect was even stronger, increasing the MCD further.

4. CONCLUSION

We provide a comparison of some recent source separation and beamforming techniques in a scenario involving either one or two target speech sources in a noise field. We analyze the results as a function of the number of channels, the number of speakers and the iSNR. We find that beamforming with a very effective post-filter performs similar to source separation with a harmonic spectral model only in the single target case. In the case where there are two speech signals interfering with each other, source separation has a clear advantage. However, we note that the iterative nature of the EM algorithm incurs a high computational cost, limiting the scenarios in which these algorithms can be applied. Future work will consider a wider range of techniques and analyze complexity and the impact of additional algorithm parameters.

5. REFERENCES

[1] P. O'Grady, B. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005.

[2] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, 2007.

[3] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[4] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for au-

dio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010, ch. 7, pp. 162–185.

[5] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4 –24, april 1988.

[6] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[7] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, Jan. 2007, article ID 86484.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[9] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 837–850, 2013.

[10] F. Nesta and M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 703–725, 2013.

[11] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, 2013.

[12] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[13] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ica and beamforming," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 2733 –2736 vol.5.

[14] E. Vincent, S. Araki, F. J. Theis, G. Nolte *et al.*, "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[15] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer Verlag, 2010, ch. 3, pp. 39–60.

[16] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, apr 1988, pp. 2578 –2581 vol.5.

[17] I. A. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, May, pp. I–905–I–908.

[18] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008, article ID 872425.

[19] A. Ozerov and C. 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, 2010.

[20] 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*, vol. 20, no. 4, pp. 1118 –1133, may 2012.

[21] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer Verlag, 2010, ch. 2, pp. 19–38.

[22] X. Mestre and M. Lagunas, "On diagonal loading for minimum variance beamformers," in *Proc. IEEE Int. Symp. on Signal Processing and Information Technology (ISSPIT)*, dec. 2003, pp. 459 – 462.

[23] N. Ito, N. Ono, E. Vincent, and S. Sagayama, "Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra," in *Proc. 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 2818–2821. [Online]. Available: http://hal.inria.fr/inria-00544104

[24] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168401001281

[25] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466 – 475, sept. 2003.

[26] T. Nakatani, S. Araki, T. Yoshioka, and M. Fujimoto, "Joint unsupervised learning of hidden Markov source models and source location models for multichannel source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 237–240.

[27] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[28] J. Thiemann, N. Ito, and E. Vincent, "The DEMAND database of multichannel environmental noise recordings," in *Proc. Int. Congress on Acoustics 2013*, to appear.

[29] N. Duong, E. Vincent, and R. Gribonval, "An acoustically-motivated spatial prior for under-determined reverberant source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 9–12.

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

[31] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1, May, pp. 125–128 vol.1.

[32] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/